Statistics

Analysis of Variance

Alexandra Posekany
alexandra.posekany@wu.ac.at

ANOVA

(One-way) analysis of variance

Test for the difference of means between two or more samples

model: one qualitative explanatory variable X with levels $1, \ldots, I$ models one quantitative explained variable Y

Hypotheses:

$$\begin{aligned} \mathrm{H}_0 : &\quad \mu_1 = \mu_2 = \mu_3 = \cdots \\ \mathrm{H}_1 : &\quad \text{At least one mean differs} \end{aligned}$$

- generalisation of the "Independent samples t-test for more than two groups"
- Variance test (F test) of the within-sample variance against the between sample variance

ANOVA Requirements/assumptions

1. Samples are independent and identically distributed (iid) random variables.
2. Populations are categorized only in one way (for one way ANOVA).
3. The mean and variance within each row are fixed.
4. Populations are approximately normally distributed (tests!).
5. Populations have the same variance σ^2 (loose).

ANOVA Designs

"**Balance**" of the design

- balanced design: the same number of observations in every row
- unbalanced design: different numbers of observations in every row

number of **explanatory categories**

- "one-way" refers to the fact that there is only one categorical variable, defining the means
  generalisation of the t-test for as many "samples" to compare as the categorical variable has categories
- "two-way" 2 categories splitting up the means
  comparing means within tables
- "k-way" generally k categories defining the means
  generalised high.dimensional tables and arrays considering serveral categorical variables at once

one-way ANOVA model

Analysis of variance (ANOVA) specifies the following simple model:

$$Y_{ij} = \underbrace{\mu}_{\text{total mean}} + \underbrace{\alpha_i}_{\text{sample mean}} + \underbrace{\epsilon_{ij}}_{\text{error terms}}$$

residuals

- μ and the α_i are unknown parameters of the population
- ϵ_{ij} are iid errors with mean 0 and a common unknown variance σ^2
- for identifiability the constraint $\sum_i \alpha_i = 0$ is added

ANOVA Estimates

The system of equations has the following **least squares** solutions:

$$\hat{\alpha}_i = \underbrace{\bar{y}_{i\cdot}}_{\text{sample mean}} - \underbrace{\bar{y}_{\cdot\cdot}}_{\text{total mean}}$$

$$\hat{\mu} = \underbrace{\bar{y}_{\cdot\cdot}}_{\text{total mean}}$$

with $\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ and $\bar{y}_{\cdot\cdot} = \frac{1}{\sum_{i=1}^{I} n_i} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}$.

The **least squares** estimates $\hat{y}_{ij}$ are then

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i$$

ANOVA Calculations

these estimates of the population parameters are obtained by minimising the **"sum of squared residuals"**

$$\sum_{ij}(Y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2.$$

where the **residuals** r_{ij} are defined as

$$r_{ij} = Y_{ij} - \hat{\mu} - \hat{\alpha}_i,$$

ANOVA Residual Sums of Squares (RSS)

We have a "sum of squares law":

$$\underbrace{RSS_{total}}_{\text{total residual sum of squares}} = \underbrace{RSS_1}_{\text{within-sample-variance}} + \underbrace{RSS_1}_{\text{between-sample-variance}}$$

where RSS_1 and RSS_0 are uncorrelated.

ANOVA Residual Sums of Squares (RSS)

The model under the null hypothesis (R formula notation: $Y \sim 1$) corresponds to the following **residual sum of squares** ignoring the categories ($i = 1, \ldots, I$), $n = \sum_{i=1}^{I} n_i$

$$RSS_0 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\hat{y}_{ij} - \bar{y}_{..})^2 = (n-1)s^2_{n-1}$$

The model under the alternative hypothesis (R formula notation: $Y \sim X$) correspond to the following RSS

$$RSS_1 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{I} (n_i - 1)s^2_{n_i-1}$$

ANOVA Mean squares

Sum of Squares	DF	Mean square
RSS_{total}	$\sum_{i=1}^{I} n_i - 1$	$RSS_{total}/(\sum_{i=1}^{I} n_i - 1)$
RSS_1	$I - 1$	$RSS_1/(I-1)$
RSS_0	$(\sum_{i=1}^{I} n_i - I)$	$RSS_0/(\sum_{i=1}^{I} n_i - I)$

For the special balanced case, replace n_i with n.

As these are variance estimates, the corresponding variance test (F test) with test statistic $F = \frac{RSS_1/(I-1)}{RSS_0/(I(n-1))}$ tests the **model fit**.

F has a F distribution with $(m-1, m(n-1))$ degrees of freedom.

Balanced two-way ANOVA

- test for the difference of means
- two qualitative explanatory variables (two factors) X_1, X_2 observation, one quantitative explained variable Y.
- The first factor X_1 has levels $1, \ldots, I_1$ the second factor X_2 levels $1, \ldots, I_2$. Thus, there exist $I_1 I_2$ distinct combinations of factor levels.
- equally many observations for each subcategory, defined by the combinations of X_1, X_2 exist

two-way ANOVA models

The base model in the two-way case is the **additive model**

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

- where the ϵ_{ij} are iid random variables with mean 0 and standard deviation σ.
- The following constraints are necessary for identifiability: $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$
- In total, the following models are possible: $Y \sim 1$, $Y \sim X_1$, $Y \sim X_2$, $Y \sim X_1 + X_2$, $Y \sim X_1 * X_2$

Example for motivating models: plot yield

As Fisher's original application, we look at different crop species and fertilizers and their influence on plot yield. We use an example, where one feature is the Aztec corn vs. Monsanto as an alternative. The second variable is the fertilizers for the products. All data are hypothetical exmaples.

	X_1 (Design)	
X_2 (Apps)	Aztec	Monsanto
Monsanto fertilizer		
organic fertilizer		

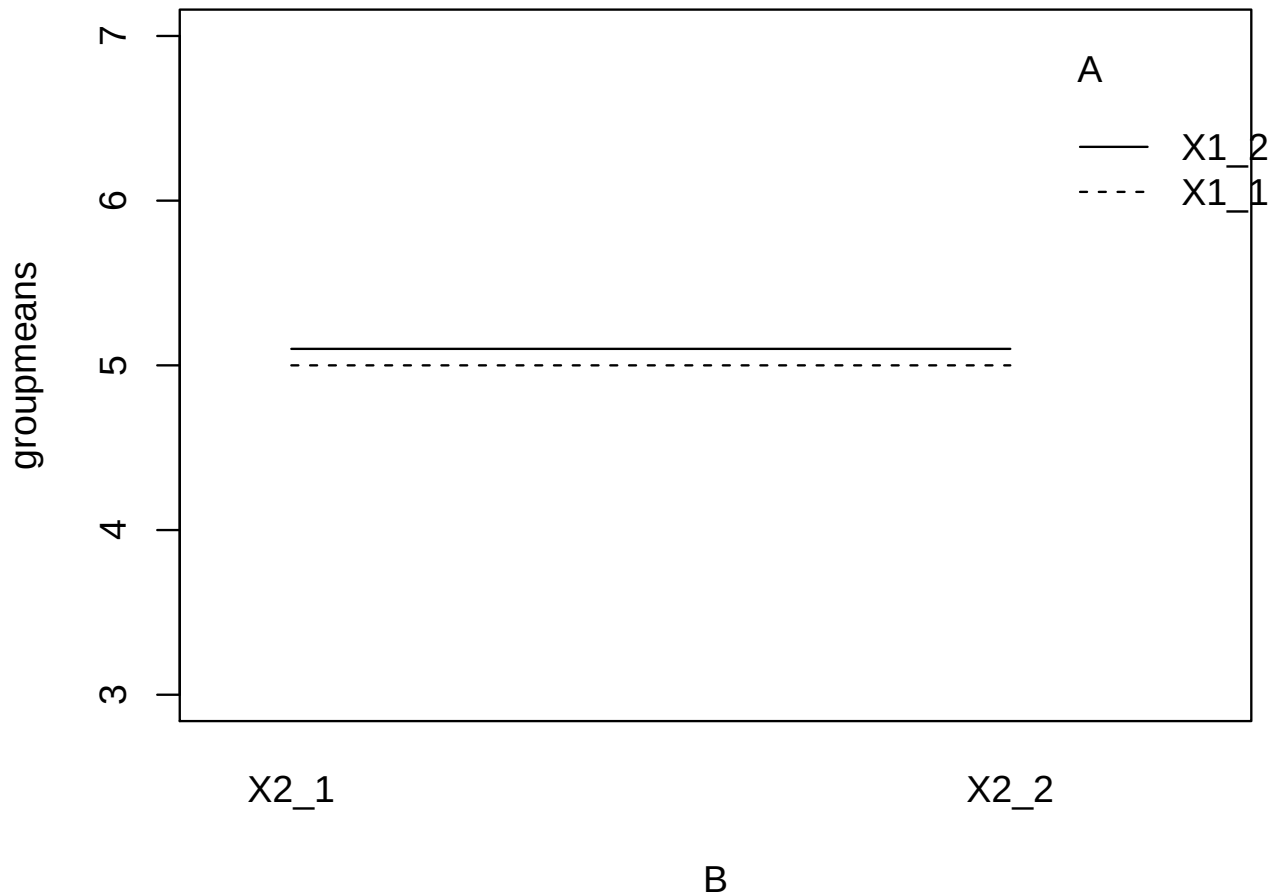
1^{st} case: no effect of factors X_1 and X_2

	X_1 (Design)	
X_2 (Apps)	Aztec	Monsanto
Monsanto fertilizer	5	5
organic fertilizer	5	5

All combinations yield the same outcome.

Thus, neither X_1 nor X_2 are required for the model (**null-model**).
R: $Y \sim 1$.

1^{st} case: no effect of factors X_1 and X_2



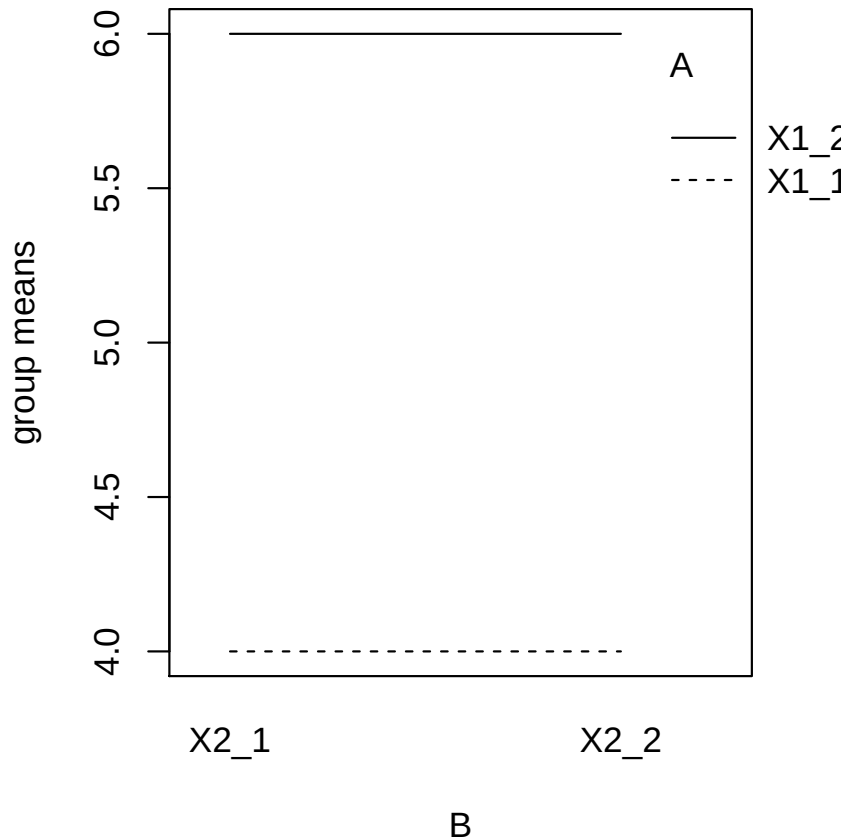
2^{nd} case: effect of X_1 or X_2 only

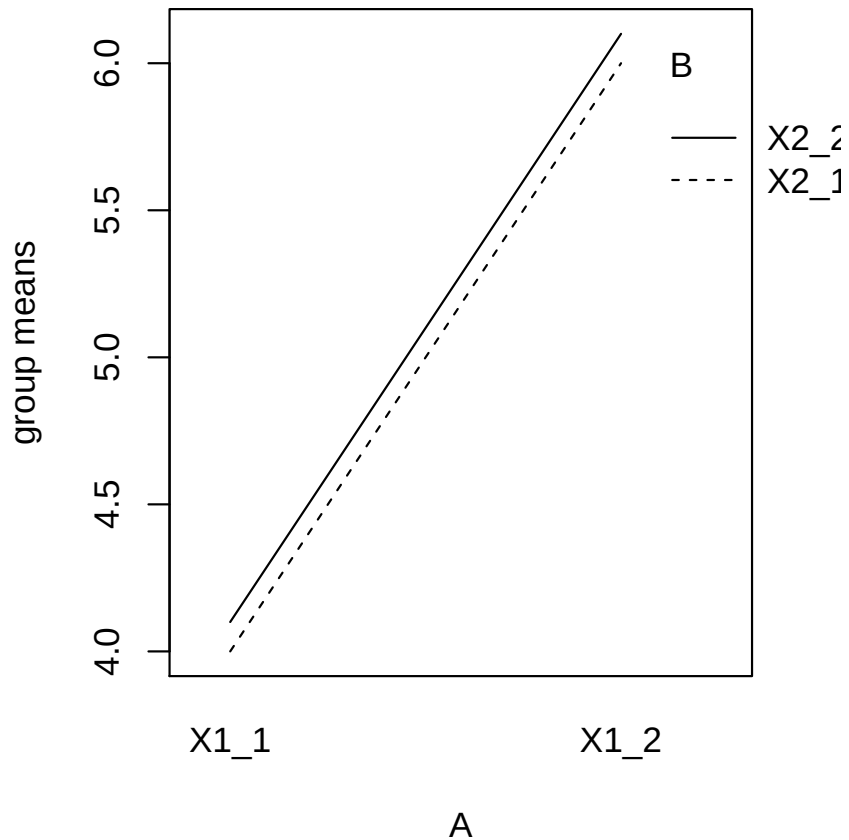
	X_1 (Design)	
X_2 (Apps)	Aztec	Monsanto
Monsanto fertilizer	4	6
organic fertilizer	4	6

Changes on means depend only on one of the categorical variables, not the other. The expected values thus do not change when the second factor is changed.

R: $Y \sim X_1$ or $Y \sim X_2$

2^{nd} case: effect of X_1 or X_2 only





3^{rd} case: additive effects $X_1 + X_2$

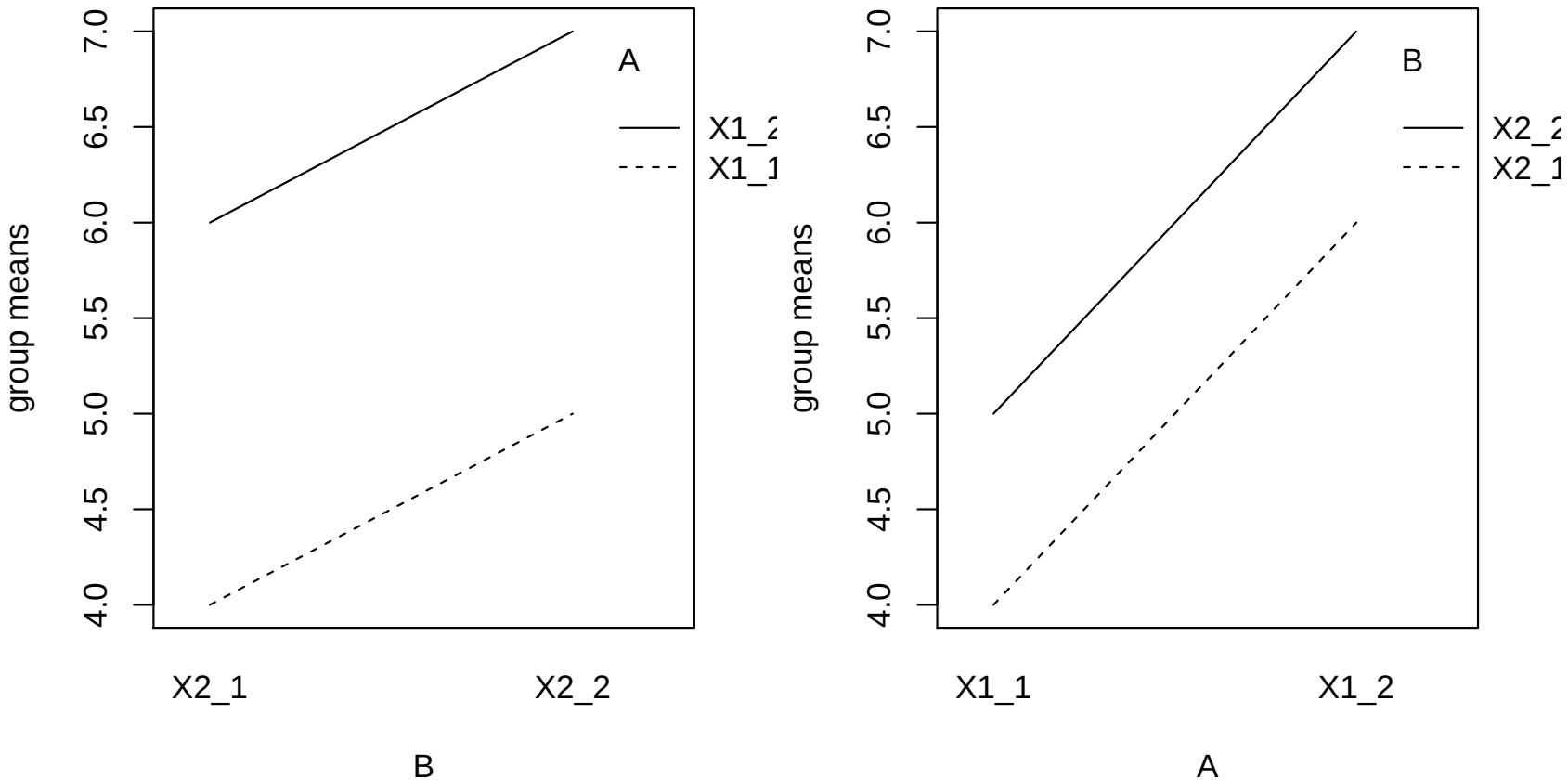
	X_1 (Design)	
X_2 (Apps)	Aztec	Monsanto
Monsanto fertilizer	4	6
organic fertilizer	5	7

Mean of joint categories depend on both factor X_1 (design) and factor X_2 (Apps) independently.

Base model: independent additive effects of factors X_1 **and** X_2.

R: $Y \sim X_1 + X_2$.

3^{rd} case: additive effects $X_1 + X_2$



4^{th} case: Interaction between X_1 and X_2

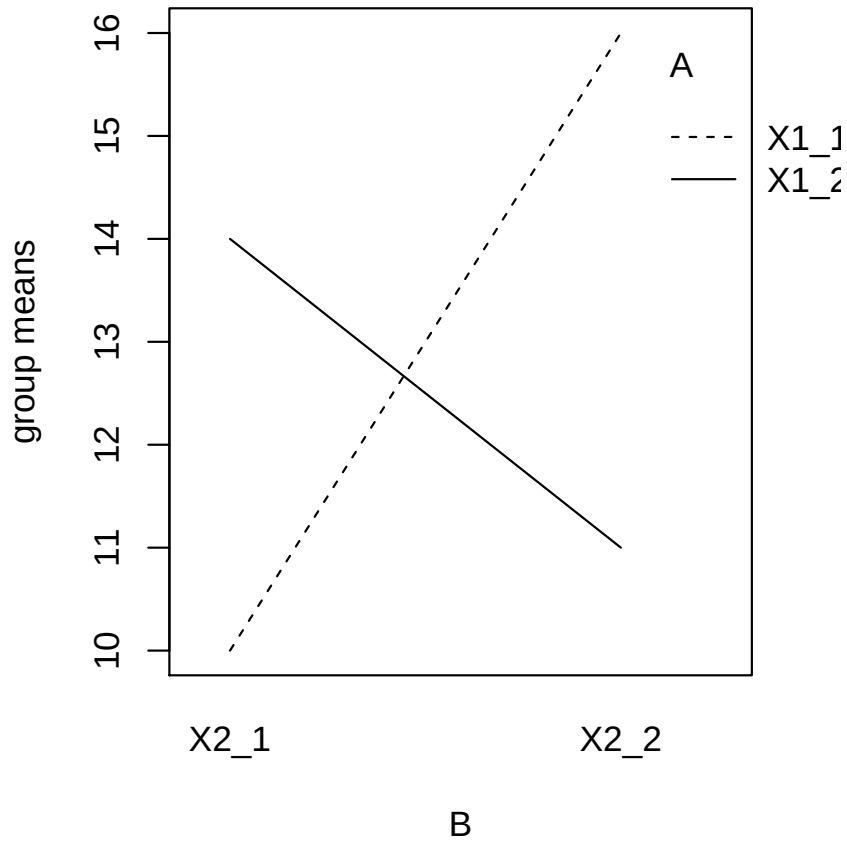
	X_1 (Design)	
X_2 (Apps)	Aztec	Monsanto
Monsanto fertilizer	4	6.5
organic fertilizer	5.5	4.5

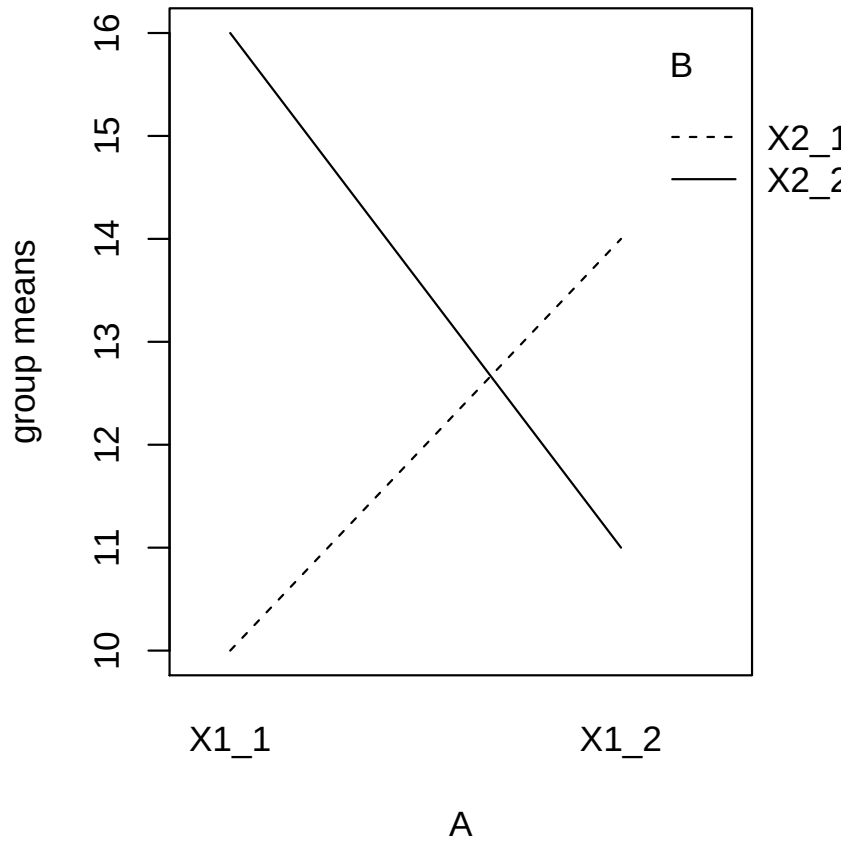
The change on group means depends on both factors X_1 and X_2 which do interact and thus yield different results in combination than each margin.

The corresponding model is the full model with interactions.

R: $Y \sim X_1 * X_2$.

4^{th} case: Interaction between X_1 and X_2





two-way ANOVA estimates

The least squares parameter estimates are:

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\alpha}_i &= \bar{y}_{i.} - \bar{y}_{..} \\ \hat{\beta}_j &= \bar{y}_{.j} - \bar{y}_{..} \end{aligned}$$

The fitted values are

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..},$$

two-way ANOVA, residual sums of squares

additivity results in $RSS_{1+2} = RSS_1 + RSS_2$, where

$$\mathrm{RSS}_1 = I_2 \sum_i \hat{\alpha}_i^2$$

and

$$\mathrm{RSS}_2 = I_1 \sum_j \hat{\beta}_j^2.$$

$$RSS_0 = \sum_{ij} (y_{ij} - \hat{\mu})^2$$

ANOVA Residual sums of squares

Model	Formula	degrees of freedom df
RSS_0	$Y \sim 1$	$\sum_{i=1}^{I} n_i - 1$
RSS_1	$Y \sim X_1$	$I_1 - 1$
RSS_2	$Y \sim X_2$	$I_2 - 1$
RSS_{1+2}	$Y \sim X_1 + X_2$	$I_1 + I_2 - 2$
RSS_{1*2}	$Y \sim X_1 * X_2$	$I_1 + I_2 - 4$

The F test for **model selection** can now compare any two nested models, i. e. that the more complex model fully contains the simpler model.

H_0 : simpler model suffices

H_1 : at least one parameter of the more complex model is required

Model selection

A visual approach towards model selection is the tree diagram related to top-down selection, which is the recommended form of model selection via ANOVA. Starting from the top, i. e. the most complex model, ANOVA comparisons against the closest less complex model are performed stepwise.