Statistisk Modellering (ST523,ST813)
Exercise Session 7

7 Exercises

Please consider/prepare all the exercises BEFORE the exercise session.

Exercise 7.1

Oneway ANOVA

Use a oneway model together with the first-category-baseline- and sum-to-zero-contrasts-parametrizations in **R** to analyze the dataset systolic.csv.[1]

This dataset contains 58 patients, suffering from one of three diseases (*disease*), that were randomly assigned to different treatments (*drug*) aiming at lowering systolic blood pressure. The variable *systolic* records the change in blood pressure.

Although this is a twoway layout (because *disease* and *drug* are two qualitative/factor variables) related to the outcome, we disregard the variable *disease* for the moment.

- Load the data and declare the factor variables in **R** .
- Get an overview about the grouping of the two factors in the data, e.g. by creating a suitable table. Is the data balanced?
- Fit a linear model using the two different parametrizations for the oneway layout.

```
Data=read.csv(”Systolic.csv”, header=TRUE)

Data$drug=factor(Data$drug)
Data$disease=factor(Data$disease)

# parametrization corresponding to dummy coding
# corresponds to constraining beta_1=0
 fit1 =lm(systolic˜drug, data=Data)

# one parametrization satisfying Ex. 1.11 from Agresti (=deviation coding)
 fit2 =lm(systolic˜drug, data=Data, contrasts=list(drug=contr.sum))

#corresponding model matrices
model.matrix(fit1)
model.matrix(fit2)
```

Interpret and compare the different model estimates.
Do these two statistical models coincide? How can you see this from the output in **R** ?

[1] data from Afifi, A. A., and S. P. Azen. 1979. Statistical Analysis: A Computer Oriented Approach.

- Proceed with these models and conduct a oneway ANOVA to investigate the null-hypothesis whether there is a relation between the different drugs and the observed change in blood pressure. Interpret the results.
- Discuss also how the given null-hypothesis and its alternative can be formulated exactly in terms of the parameter vector β?
- How is the test statistics distributed if H_0 holds? And what are the corresponding degrees of freedom?

Exercise 7.2

Twoway ANOVA

This exercise considers a twoway ANOVA to analyze the systolic blood pressure dataset.

1. Start out and make meaningful plots to visualize the data and relations between the two factors and the outcome. (in the same style as for the *Warpbreaks* example in the lecture)
2. Conduct a twoway ANOVA to analyze the blood pressure dataset. Especially, investigate whether the two factors interact relative to the outcome.
3. Interpret the corresponding ANOVA tables and the results of the different F-tests.
4. Do results depend on the order of the factors? Why? Clarify the (chain of) hypotheses underlying the different F-tests in the ANOVA table.
5. What are the different means that are estimated for the different factor combinations? How much does the blood pressure change on average for a patient with the second disease obtaining drug 3?
6. Follow up the test for main effects and apply Tukey's method of multiple comparisons. Calculate adjusted 95% confidence intervals and p-values and interpret the results.

Exercise 7.3

Twoway ANOVA, continued

The file *SystolicReplicationStudy.csv* contains data from another later study which tried to replicate the effects of the original study in a similar but different patient group.

Analyze this new dataset with a twoway ANOVA and answer again the subquestions 1.-3. and 5. from the last exercise for the new data.

What are your conclusions for this dataset? Is the change in systolic blood pressure associated with type of disease or type of treatment?

Exercise 7.4

ANCOVA (Analysis of Covariances)

Note: Ancova combines analysis of variance and linear regression. It examines the effect of one or more categorical covariates on a continuous response, while controlling for the effect of one or more continuous covariates.

You are a data analyst for a jewelry retailer. Your task is to model the price of diamonds based on their cut quality, color, and carat weight. The retailer suspects that:

- Both *cut* and *color* of the diamond influence price independent of each other,
- The effect of *carat* weight on price may depend on the *cut*,
- Other features of the diamond do not impact the price further.

1. Load the `diamonds` dataset from the `ggplot2` package.
2. Provide a summary of the main variables: `price`, `cut`, `color`, and `carat`.
3. How many different values (=levels) do the factor variables `cut` and `color` have?
4. Generate boxplots of `price` by `cut` and `color` to visualize group differences.
5. Fit an ANCOVA model, i.e. regress `price` onto `cut`, `color` and `carat`, also involving the interaction between `cut` and `carat`. (Why is it reasonable to include the mentioned interaction?)
6. Interpret the model output.
7. How does `carat` weight influence `price` differently for different cuts according to your estimates?
8. Carry out corresponding hypothesis tests using the *anova* command and investigate
   - the presence of the interaction between `cut` and `carat`, and
   - (if applicable) the presence of additive effects of `cut` and `carat`.
9. To test whether `color` has an impact onto `price` in addition to `cut` and `carat`, use a partial F-test which compares the current fitted model to a model without an effect of `color`.
   - Interpret your results and the estimated effects.
   - Why is this test different from the one listed in the anova table when applying the anova command to the fitted model?
10. Visualize the effects and interaction as follows:
    - Create a scatter plot of `price` versus `carat`, colored by `cut`.
    - Add your estimated regression lines for each `cut` level to visualize how the relationship between `carat` and `price` differs by `cut`. You can e.g. take a weighted average over the fitted lines corresponding to the different values of `color` for plotting, where the weights reflect the proportions of the different colors in the sample.

Exercise 7.5

Parametrization in twoway layout (optional)*

Consider a twoway layout together with the following additive model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

where ε_{ijk}, for $i = 1, \dots, r$, $j = 1, \dots, c$ and $k = 1, \dots, n$, are i.i.d. with zero mean and variance σ^2. Note, that the data is *balanced* having an equal sample size in each of the rc cells, and the model assumes an absence of interaction between the two factors in their effect on Y.

1. For the model as stated, is the parameter vector identifiable? Why or why not?

2. Give an example for a quantity that is (i) not estimable, (ii) estimable. Explain your reasoning.

Suppose now $r = 2$, $c = 3$ and $n = 2$.

3. Show the form of a full-rank model matrix $\mathbf{X}$ and corresponding parameter vector for the model, constraining $\alpha_1 = \beta_1 = 0$ to achieve identifiability.

4. Show the form of a full-rank model matrix and corresponding parameter vector when you constrain $\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{c} \beta_j = 0$. Explain how to interpret the parameters.

5. In the full rank case, what is the rank of $\mathbf{X}$?

6. Fit a correspondingly parametrized model in **R** using the above blood pressure data considering both factors. Interpret/verify your above considerations in **R** .