# Advanced Statistics, Summer 2022

Homework 2, 01.06.2022

Data files are provided for you individually in a zip-file numbered with your homework ID. You find your homework ID in the file "Homework ID.pdf". All datasets are in CSV format (comma separated values) with the first line as heading, and as Excel sheets.

You can use any programming language or statistical package that you are familiar with (e.g. MATLAB, python, R, SPSS, …).

Please double-check your solutions. Wrong solutions will be sent back for corrections.

Please, submit the **solutions and the codes** that you used in **one PDF file** labeled "Your Name - Homework 2" to me at: steffen.gais@uni-tuebingen.de.

*Please use "ADVSTAT" in the subject line of your emails.*

Send your solutions before the next course on 15.06.2022.

Please solve these tasks on your own. Copy/paste solutions from others will not be accepted!

1. You believe that a new drug will improve the memory of rats. You have two groups (*group:* 1 = placebo and 2 = drug). You measure how long they take to find the goal in a labyrinth (*y*) as a measure of learning. You find the data in Table2.csv / Table2.xlsx.

   a. Generate a suitable **design matrix** $X$ for the effect using dummy coding.
   Use the formulas provided in the lecture to estimate $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{y}}$, $\hat{\boldsymbol{\varepsilon}}$, and the **residual sum of squares** ($SSE = \hat{\varepsilon}'\hat{\varepsilon}$) for the full model.
   **Generate a design matrix for the null hypothesis** that the factor *group* is irrelevant (restricted model). For this you remove the corresponding column of $X$ (or use a suitable $C$ matrix if you wish). Determine $\widehat{\boldsymbol{y}}$ and the **residual sum of squares** ($SSE_{H_0}$) for the restricted model.
   To test for significance, calculate **F** using $F = \frac{n-p}{r}\frac{SSE_{H_0}-SSE}{SSE}$ and provide the corresponding **p value**.
   (If you use Matlab, the following functions might be helpful: readtable, fcdf.)

   b. Import the data into a statistics package (e.g. SPSS, SAS, JASP, R ...). Use its **GLM function** to determine **beta values**, the **residual sum of squares**, **F**, **degrees of freedom**, **p value, and effect sizes** for the group difference.
   (In SPSS, the GLM->Univariate function has options that provide parameter estimates.)

   - *Is there a significant effect of your drug?*

   You notice that some animals seem to run faster than others, therefore you also measure running speed (*cov*) and test whether it has an influence on the time the animals take (*y*).

   c. Generate a suitable **design matrix** $X$ for running speed only.
   As above: Estimate $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{y}}$, $\hat{\boldsymbol{\varepsilon}}$, and the **residual sum of squares** for the **full model**.
   **Generate a design matrix for the null hypothesis** that the covariate is irrelevant (restricted model). For this you remove the corresponding column of $X$. Determine $\widehat{\boldsymbol{y}}$ and the **residual sum of squares** ($SSE_{H_0}$) for the restricted model.
   As above: To test for significance, calculate **F** and provide the corresponding **p value**.

d. Use your statistics package's **GLM function** to determine **beta values**, the **residual sum of squares**, **F**, **degrees of freedom**, **p value, and effect sizes** for the covariate.

- *Does running speed significantly influence total running time?*

You want to avoid drawing wrong conclusions. Because running speed may confound your measure of memory, you control for it by including it into your model as a covariate together with the main effect of factor *group*.

e. Generate a suitable **design matrix** with both the group effect and the covariate.
As above: Estimate $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{y}}$, $\hat{\boldsymbol{\varepsilon}}$, and the **residual sum of squares** for the **full model**.
**Generate design matrices for the following two null hypotheses:** 1. the drug has no effect, 2. running speed does not influence the result. Determine $\widehat{\boldsymbol{y}}$ and the **residual sum of squares** ($SSE_{H_0}$) for both and test for significance. Provide **F** and **p value**.
*(ATTN: Interactions between main effects and covariates are habitually not included.)*

f. Use your statistics package's **GLM function** to determine **beta values**, **F**, **degrees of freedom**, **p values, and effect sizes** for the group difference and the covariate.

- *Does your drug have a significant effect? Is there a significant effect of running speed? How do you interpret your findings?*

2. You have two new drugs that claim to improve intelligence (*y*) significantly. You test them in a placebo-controlled experiment with three groups. (Factor *A: 1=drug1, 2=drug2, 3=placebo)* You find the data in Table3.csv / Table3.xlsx. Answer the following questions

a. Generate a suitable design matrix $X$ for the main effect.
As above: Estimate $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{y}}$, $\hat{\boldsymbol{\varepsilon}}$, and the residual sum of squares for the full model.
**Generate a design matrix for the null hypothesis** that the factor drug is irrelevant (restricted model). For this you remove the corresponding column of $X$. Determine $\widehat{\boldsymbol{y}}$ and the **residual sum of squares** ($SSE_{H_0}$) for the restricted model.

- *Is there a significant effect of the factor drug?*

You notice that only females report improvements after receiving the drug. You therefore include sex as a second factor in your model (Factor *B*: *1=male, 2=female*).

b. Generate a suitable design matrix $X$ for the two main effects and the interaction. ***ATTN:*** If the design includes an interaction, you must use effect coding to get the correct main effects! As above: Estimate $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{y}}$, $\hat{\boldsymbol{\varepsilon}}$, and the residual sum of squares for the full model.
**Generate the three design matrices for the null hypotheses** that 1. drug, 2. sex, 3. the interaction drug x sex are irrelevant. Determine their $\widehat{\boldsymbol{y}}$ and **residual sum of squares.**

- *Is there a significant main effect of drug, main effect of sex, interaction drug x sex?*
- *Plot the data. How do you interpret the findings?*
- Check your results using your statistics package's **GLM function.**