TMA4267 Linear Statistical Models
Compulsory Exercise Part 3: Hypothesis testing and ANOVA

Due date: 24 March 2017 at 6pm

Two students can hand in one solution together. One student uploads the (handwritten) solutions in Blackboard with the name of both students on the front page, the other only submits "I have handed in my solution together with NN" in Blackboard.

Please include your R code in the submission.

Problem 1: Diabetes progression and F-test for linear regression

Please refer to the problem of Compulsory Exercise 2, where a full linear regression model with 10 covariates was fitted. Then, in problem **c)** a reduced model was found and then fitted. Let us assume that the reduced model was the model including the five covariates `sex, bmi, map, hdl, ltg` and an intercept. This means that the following five covariates from the full model was not included in the reduced model: `age, tc, ldl, tch, glu`.

We want to investigate the following null and alternative hypotheses:

$$H_0 : \beta_{age} = \beta_{tc} = \beta_{ldl} = \beta_{tch} = \beta_{glu} = 0 \text{ vs. } H_1 : \text{ at least one } \neq 0$$

that is, we want to investigate if we with hypothesis testing would confirm that the reduced model is preferred over the full model.

```
>ds=read.csv("https://web.stanford.edu/~hastie/CASI_files/DATA/diabetes.csv",
sep=",")
>n=dim(ds)[1]
>full=lm(prog~.,data=ds)
>reduced=lm(prog~sex+bmi+map+hdl+ltg,data=ds)
```

a) Rewrite the hypothesis problem as a linear hypothesis of the form $\boldsymbol{C\beta} = \boldsymbol{d}$ by specifying what $\boldsymbol{C}$ and $\boldsymbol{d}$ are in this context.

b) The test statistic for performing the test is called F_{obs} and can be formulated in two ways:

$$F_{obs} = \frac{\frac{1}{r}(SSE_{H_0} - SSE)}{\frac{SSE}{n-p}} \tag{1}$$

$$F_{obs} = \frac{1}{r}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})^{\mathrm{T}}[\hat{\sigma}^2\boldsymbol{C}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C}^{\mathrm{T}}]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \tag{2}$$

What is the interpretation and the formula for SSE, SSE_{H_0}, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$? What is n, p and r?

c) Perform the hypothesis test in R using the F_{obs} test statistic, and decide yourself which version (1) or (2) of the test statistic you could like to use. Write down the steps that you do. Report the value of F_{obs} calculated and the corresponding p-value. Will you reject the null hypothesis (choose significance level yourself before you perform the test)? Does this mean that you prefer the full or the reduced model?

Problem 2: Multiple testing

In genome-wide association studies the aim is to test if there is an association between a genetic marker and a trait. This means that an hypothesis test is performed for each marker. We have data from 1000 markers and for each marker, $j = 1, \ldots, 1000$, we perform an hypothesis test:

$$H_0 : \mu_j = 0 \text{ vs. } H_1 : \mu_j \neq 0$$

where $\mu_j = 0$ means that there is no association between the marker and trait. From hypothesis test i we calculate a p-value p_i (based on some continuous test statistic). P-values are available to read into R as

```
pvalues=scan("https://www.math.ntnu.no/emner/TMA4267/2017v/CompEx3P2pvalues.txt")
```

a) Assume that we reject all null-hypotheses with corresponding p-values below 0.05. How many null-hypotheses will we then reject? What is a false positive finding? Do we know the number of false positive findings in our data?

b) Let the number of false positive findings for our data be called V. What is the definition of the familywise error rate FWER?

What does it mean to "control the FWER at level 0.05"? The Bonferroni method will control the FWER. What cut-off on p-values should we use if we want to control the FWER at level 0.1 for our data with the Bonferroni method? Call this cut-off α_B. How many null-hypotheses will we reject with this new cut-off?

To see the effect of choosing different cut-offs on p-value on the number of false positive findings we need to know which null hypotheses are true and which are false. Let us assume that all the 1000 p-values come from true null hypotheses. What does this imply about the number of rejection in **a)** and **b)**? What if only the first 500 p-values come from true null hypotheses?

c) Describe briefly what is meant by the following two terms: p-hacking and reproducibility crises. What is the relationship between these two terms and multiple testing.