TMA4267 Linear Statistical Models
Part 3: Hypothesis testing and ANOVA
Recommended exercise 5 - V2017

March 3, 2017

Keywords: multiple linear regression F-test, analysis of variance, effect coding, sum-zero-constraint, ANOVA, multiple testing, familywise error rate, false discovery rate.

- Problem 1: Classical multiple regression, only new in Part 3 is e) (all else known from Part 2), but the concept of using effect coding (and all possible combinations of three factors with two levels) will be a core part of Part 3 and Part 4. In particular, question b) on the interaction plot and the start of e) on comparing two models (and observing that the estimates are equal) will be discussed more in Part 4 (but you have the background to answer also now).

- Problem 2: Additional multiple testing problem related to Problem 1. This is new on this years reading list (not part of the reading list on previous years).

- Problem 3: Alternative theoretical outline of F-tests in regression.

- Problem 4: First step from linear regression to analysis of variance with effect coding (sum-zero constraints).

- Problem 5: Classical analysis of variance problem, formulated without an obvious link to multiple linear regression.

Problem 1: Plant stress

(Exam TMA4267 V2016, slightly modified to reflect some changes in the reading list - e.g. lasso not on reading list in 2017 because will be part of new course TMA4268 Statistical Learning V2018).

At the Department of Biology at NTNU researchers use the model plant *Arabidopsis thaliana* to study the response of a plant to different sources of stress. In an experiment *Arabidopsis thaliana* seedlings were subject to a stress situation. The following factors were fitted:

- D (damage): $D = 1$ means that the plant was damaged mechanically by cutting into the leaves of the plant by a pair of scissors. $D = -1$ means damage was not inflicted (no cutting).

- F (flagellin): $F = 1$ means that the pathogen-derived peptide flagellin was sprayed on the leaves of the plant. $F = -1$ means water (not flagellin) was sprayed.

- T (time): Plants were harvested at two different time points after the stress situation. $T = 1$ means that the plant was harvested 60 minutes after the stress situation and $T = -1$ means that the plant was harvested 30 minutes after the stress situation.

Thus, we have three factors, D, F and T, each at two levels. In the study experiments for all possible combinations of the three factors were performed four times yielding 32 experiments in total.

The response measured in the experiment, was the observed gene activity level (a continuous measurement) of each of around $40\,000$ genes. We will only focus on the gene activity level of one of these genes, the AT1G32920 gene, and we denote the gene activity level of this gene by Y. It is known that this gene is active in response to wounding.

For experiment number i (where $i = 1, \ldots, 32$): Y_i is the observed response, D_i is chosen value of D, F_i is chosen value of F, and T_i is chosen value of T. A multiple regression model with all main effects, and two- and three-way interactions, was considered,

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \beta_{D:T} D_i T_i + \beta_{F:T} F_i T_i + \beta_{D:F:T} D_i F_i T_i + \varepsilon_i,$$

where $i = 1, \ldots, 32$, and we assume ε_i independent and identically normally distributed with mean 0 and variance σ^2. We refer to this as the *full model*. Note that the interactions are simply products of the factors. The vector of regression parameters is
$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_D & \beta_F & \beta_T & \beta_{D:F} & \beta_{D:T} & \beta_{F:T} & \beta_{D:F:T} \end{pmatrix}^{\mathrm{T}}$, and the ith row of the design matrix $\boldsymbol{X}$
is $\begin{pmatrix} 1 & D_i & F_i & T_i & D_i F_i & D_i T_i & F_i T_i & D_i F_i T_i \end{pmatrix}$.

In Figure 1 you find R-commands and print-out from fitting the full model.

a) In the print-out from `summary(fit)` in Figure 1 *four* numerical values are replaced by question marks. Calculate numerical values for each of these, and explain what each of the values means.

A so called cube plot is given in the upper left panel of Figure 2. In the cube plot the fitted values from the multiple regression for the possible 8 combinations of the three factors are given. Plots of the main effects (upper right panel) and the interaction effects (lower panel) are found in Figure 2. In Figure 3 you find residual plots. See Figure 4 for the accompanying R-code and print-out.

b) How would you, based on Figures 2–4, evaluate the fit of the model?

How would you explain to a biologist what the estimated main effect of damage means in practice? How would you explain the estimated interaction effect between damage and flagellin?

Let $\gamma = 2^{\beta_F - \beta_D}$ be a new parameter of interest.

Suggest an estimator, $\hat{\gamma}$, for γ. Use approximate methods to find the expected value and variance of this estimator, that is, $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$. Use results in Figure 1 to calculate numerical value for $\hat{\gamma}$, and estimated numerical values for $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$.

Hint: You may use that $2^x = \exp(x \ln 2)$, where ln is the natural logarithm.

The researchers want to test the hypothesis

$$H_0 \colon \beta_{D:T} = \beta_{F:T} = \beta_{D:F:T} = 0 \qquad \text{vs.}$$
$$H_1 \colon \text{at least one of } \beta_{D:T},\ \beta_{F:T},\ \beta_{D:F:T} \text{ is different from 0.}$$

```
# data is in "standard order" in data frame with name "ds"
> ds %showing only rows 1-6 and 27-32 for space considerations
          Y  D  F  T
1  15.45169 -1 -1 -1
2  15.15908 -1 -1 -1
3  14.93064 -1 -1 -1
4  15.06569 -1 -1 -1
5  14.51032 -1 -1  1
6  14.76922 -1 -1  1
...
27 18.23645  1  1 -1
28 17.70327  1  1 -1
29 16.66523  1  1  1
30 16.96046  1  1  1
31 16.73133  1  1  1
32 16.57248  1  1  1
> fit=lm(Y~D*F*T,data=ds)
> model.matrix(fit)%only showing rows 1-6 and 27-32
   (Intercept)  D  F  T D:F D:T F:T D:F:T
1            1 -1 -1 -1   1   1   1    -1
2            1 -1 -1 -1   1   1   1    -1
3            1 -1 -1 -1   1   1   1    -1
4            1 -1 -1 -1   1   1   1    -1
5            1 -1 -1  1   1  -1  -1     1
6            1 -1 -1  1   1  -1  -1     1
...
27           1  1  1 -1   1  -1  -1    -1
28           1  1  1 -1   1  -1  -1    -1
29           1  1  1  1   1   1   1     1
30           1  1  1  1   1   1   1     1
31           1  1  1  1   1   1   1     1
32           1  1  1  1   1   1   1     1
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04140  ?         < 2e-16
D            0.93739    0.04140  22.644    < 2e-16
F            0.28546    0.04140   6.896   3.93e-07
T           -0.52354    0.04140 -12.647   4.18e-12
D:F         -0.08878    0.04140  -2.145   0.04231
D:T         -0.00242    ?         -0.058   0.95386
F:T         -0.12614    0.04140  -3.047   0.00555
D:F:T        0.09099    0.04140   2.198   ?
Residual standard error: 0.2342 on 24 degrees of freedom
Multiple R-squared:      ?, Adjusted R-squared:  0.9594
F-statistic: 105.6 on 7 and 24 DF,  p-value: < 2.2e-16
```

Figure 1: Printout from R-commands and statistical analyses for the plant stress data set. Four numbers are replaced by question marks.
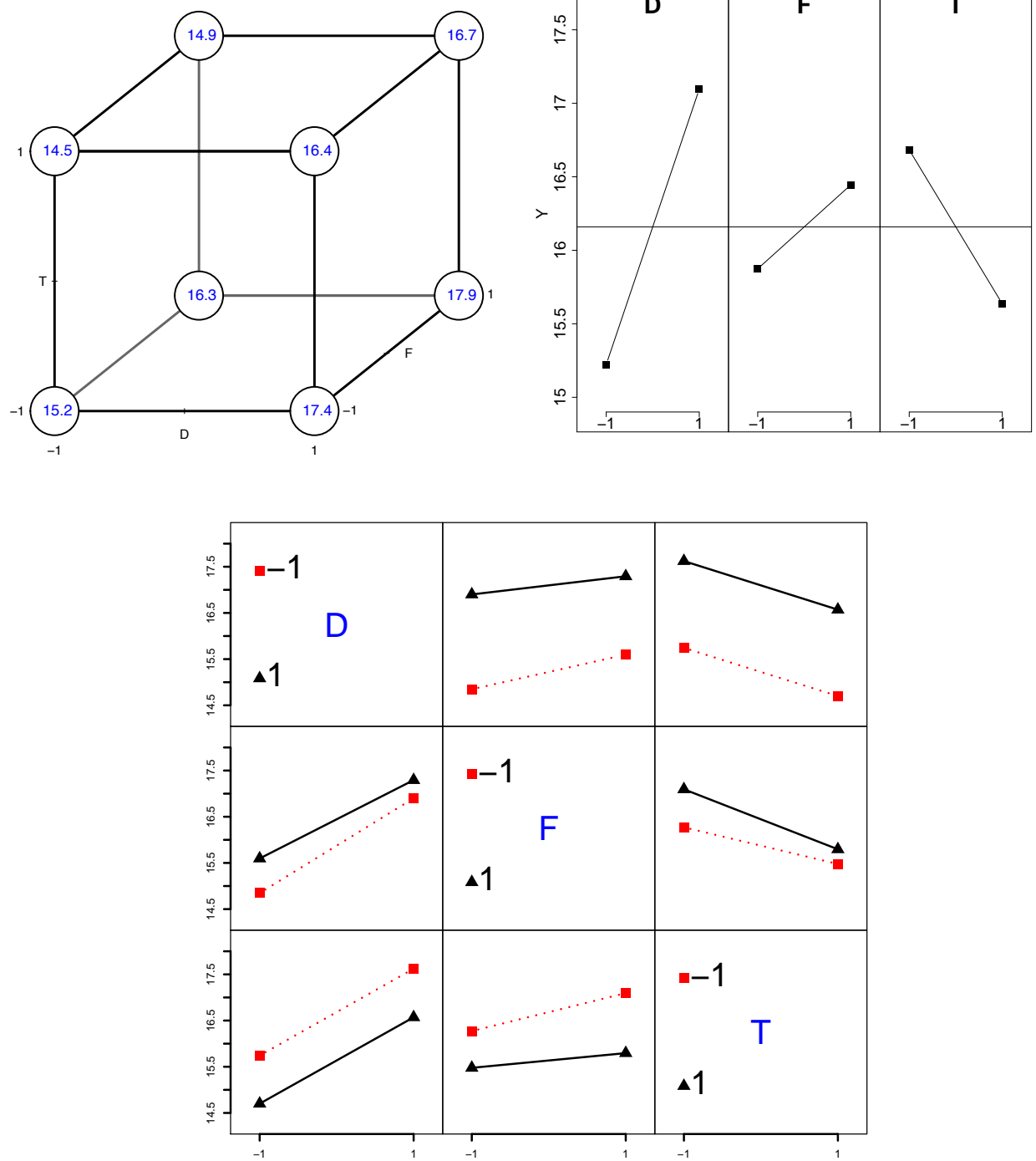
Figure 2: Cube plot (upper left panel), main effects plot (upper right panel) and interaction effects plot (lower panel) for the full model fitted to the plant stress data set.
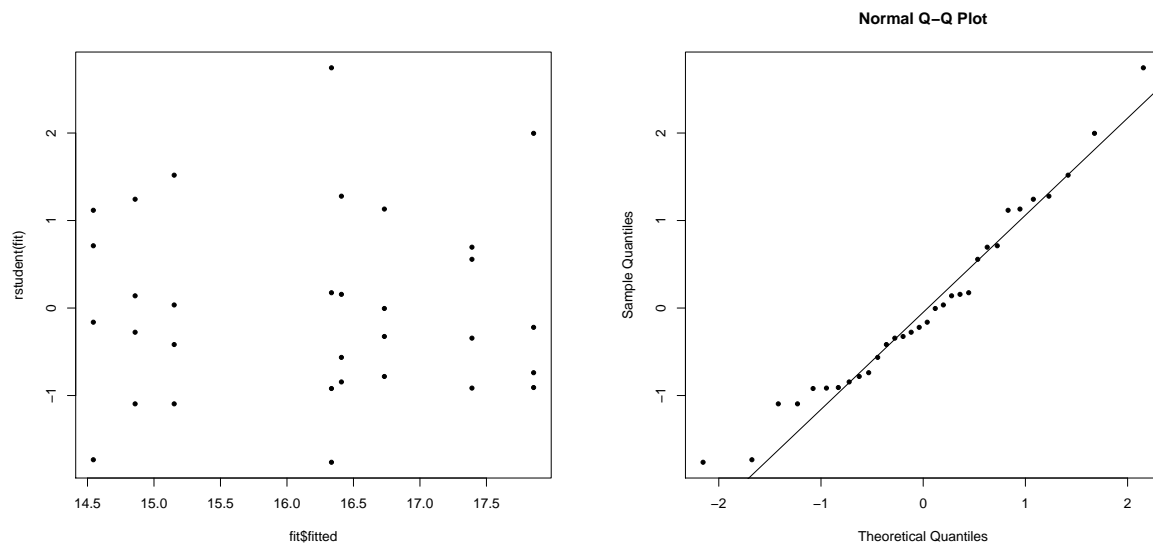
Figure 3: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for the full model fitted to the plant stress data set.

```
> library(FrF2)
> MEPlot(fit)
> IAPlot(fit)
> cubePlot(fit,"D","F","T",round=1,size=0.33,main="")
> plot(fit$fitted,rstudent(fit),pch=20)
> qqnorm(rstudent(fit),pch=20)
> qqline(rstudent(fit))
> ad.test(rstudent(fit))
        Anderson-Darling normality test
data:  rstudent(fit)
A = 0.43191, p-value = 0.2869
```

Figure 4: Print-out from R-commands and statistical analyses for the full model fitted to plant stress data set.

```
> x <- model.matrix(fit)[,-1]; dim(x)
[1] 32  7
> y <- ds$Y
> library(leaps)
> bests <- regsubsets(x,y)
> sumbests=summary(bests)
> sumbests
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         D   F   T   D:F D:T F:T D:F:T
1  ( 1 ) "*" " " " " " " " " " " " "
2  ( 1 ) "*" " " "*" " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" " " " " "*" " "
5  ( 1 ) "*" "*" "*" " " " " "*" "*"
6  ( 1 ) "*" "*" "*" "*" " " "*" "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
> plot(bests,scale="adjr2",col=gray(seq(0.6,0.9,length=20)))
> round(sumbests$adjr2,3)
[1] 0.661 0.874 0.938 0.950 0.955 0.961 0.959
```

Figure 5: Print-out from R performing best subset selection on the plant stress data.

c) Perform the hypothesis test at a significance level of your own choice. All the numerical values you need for the calculations are found in Figure 1.

The researchers want to use the data to fit a prediction model, and want to consider reduced versions of the full model, using best subset model selection. Results are presented in Figure 5.

d) Explain briefly what is done in the best subset model selection, and choose a good model based on the R^2_{adj}-criterion. Write down the fitted regression model for the model you choose.

The researchers choose to use the following *reduced model* for prediction:

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \varepsilon_i,$$

where $i = 1, \ldots, 32$, and we assume ε_i independent and identically normally distributed with mean 0 and variance σ^2. Output from fitting the reduced model is given in Figure 7.

e) Compare the estimated regression parameters and the estimated standard deviations of the estimated regression parameters for the full model (Figure 1) and the reduced model (Figure 7), and explain what you observe.

Based on the reduced model (Figure 7), provide a prediction and a 95% prediction interval for the gene activity level for the factor combination $D = 1, \ F = 1, \ T = -1$.

Hint: In a multiple linear regression with $n \times p$ design matrix $\boldsymbol{X}$, estimated regression coefficients $\hat{\boldsymbol{\beta}}$ and unbiased estimated error variance s^2, a $(1 - \alpha)100\%$ prediction interval at $\boldsymbol{x}_0$ is given as

$$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{\frac{\alpha}{2}, n-p} \, s \, \sqrt{1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0},$$

where $t_{\alpha/2, n-p}$ denotes the value in the t-distribution with $n - p$ degrees of freedom that has area $\frac{\alpha}{2}$ to the right. See Figure 7 for some possible values for $t_{\alpha/2, n-p}$.
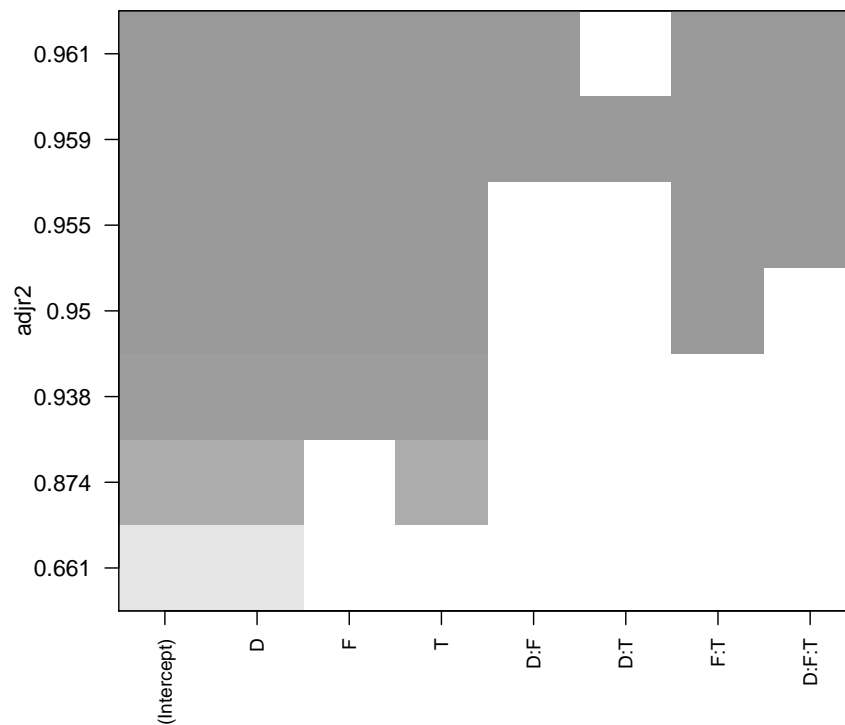
6

Figure 6: Figure from model selection, with R-code in Figure 5.

```
> fitRED=lm(Y~D+F+T+D:F,data=ds)
> summary(fitRED)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04919 328.528  < 2e-16
D            0.93739    0.04919  19.057  < 2e-16
F            0.28546    0.04919   5.804 3.56e-06
T           -0.52354    0.04919 -10.644 3.66e-11
D:F         -0.08878    0.04919  -1.805   0.0822
Residual standard error: 0.2782 on 27 degrees of freedom
Multiple R-squared:  0.95,        Adjusted R-squared:  0.9426
F-statistic: 128.4 on 4 and 27 DF,  p-value: < 2.2e-16
> qt(0.025,32,lower.tail=FALSE)
[1] 2.036933
> qt(0.025,27,lower.tail=FALSE)
[1] 2.051831
> qt(0.025,24,lower.tail=FALSE)
[1] 2.063899
```

Figure 7: Print-out from R performing linear regression on the reduced model for the plant stress data set.

Problem 2: Multiple testing with plant stress

In this problem we work with the same biological problem as in Problem 1, but we will here focus on multiple hypothesis testing.

We want to study the gene activity level of each of $m = 10\,000$ genes (in Problem 1 we only looked at one of the genes), and test for relationship with damage. Let μ_j denote the regression coefficient of damage for gene j $(j = 1, \ldots, m)$ (previously denoted β_D for one gene). Our aim is now to identify genes that respond to damage, and we do this by testing the two-sided hypotheses

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

for $j = 1, \ldots, m$ (all genes). Assume that we have performed the hypothesis tests and we have calculated two-sided p-values $p_1, \ldots, p_m$.

Note: for each gene we have performed one regression, with the gene activity level of that gene as the response, so the μ_js are from m different regression models with different responses. (Not to be confused with different regression parameters.)

These p-values are available as an m-dimensional vector from the course www-page, to be read into R as:

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

a) We want to test the m hypotheses stated above, and want to set a cut-off on the p-values to declare that a test is significant (i.e. decide when we want to reject each of the null hypotheses). To do that we want to control some overall version of the Type I error rate. Explain with words and a formula what the familywise error rate (FWER) and the false discovery rate (FDR) is.

b) The FWER can be controlled at level α by the Bonferroni method, by producing a local significance level α_{loc} to be used as cut-off for the p-values.

We choose to control the FWER at level $\alpha = 0.05$. Calculate the Bonferroni α_{loc} for our data. How many null hypotheses will you reject? Are there any requirements for when we can use the Bonferroni method? Why do people say that the Bonferroni method is conservative?

c) The p-values given above have been artificially generated and the truth is that the 9000 first p-values are from true null hypotheses and the last 1000 are from the false null hypotheses. Based on your rule in **b)** fill in numerical values for the quantities denoted S, T, U, V, R, m and m_0 in the following table. What is the number of false positives?

	Not reject H_0	Reject H_0	Total
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
Total	$m - R$	R	m

In a multiple hypothesis setting where the true nature of the hypotheses are not known, which of the entries in the table above are actually known?

d) What if we choose to say that all p-values below 0.05 gave a significant result. What would then S, T, U, V, R, m and m_0 become. What is the number of false positives now?

In a scientific research situation it is important to avoid false positive results (fake news). That is why a cut-off of 0.05 never should be used in a research situation where more than one hypothesis test is formally conducted.

Problem 3: F test and partial F test in multiple linear regression

Consider a multiple linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\beta} = (\beta_1 \ \cdots \ \beta_p)^T$ and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. We want to test the null hypothesis $\beta_{r+1} = \beta_{r+2} = \cdots = \beta_p = 0$. In effect, we want to compare the full model, involving p covariates and p parameters, with a reduced model, involving only r covariates and r parameters. Let $\boldsymbol{X}_0$ be the design matrix consisting only of the first r columns of $\boldsymbol{X}$. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{H}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T$ be the projection matrices of the full and the reduced model, respectively.

a) Show that in the partition $\boldsymbol{I} = (\boldsymbol{I} - \boldsymbol{H}) + (\boldsymbol{H} - \boldsymbol{H}_0) + \boldsymbol{H}_0$, all three terms on the right are idempotent. (Hint for $\boldsymbol{H} - \boldsymbol{H}_0$: View $\boldsymbol{H}$ and $\boldsymbol{H}_0$ as projections onto the column space of $\boldsymbol{X}$ and $\boldsymbol{X}_0$, respectively, to see that $\boldsymbol{H}\boldsymbol{H}_0 = \boldsymbol{H}_0$ and that $\boldsymbol{H}_0(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$.)

b) Show that under the null hypothesis, $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and that
$$(\boldsymbol{H} - \boldsymbol{H}_0)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y}.$$

c) What are the ranks of $\boldsymbol{I} - \boldsymbol{H}$ and $\boldsymbol{H} - \boldsymbol{H}_0$ (assuming $\boldsymbol{X}$ has full rank)? Use Definition B.14 and Theorem B.8 in Fahrmeir et al. to show that if the null hypothesis is true, then
$$F = \frac{\boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y}/(p - r)}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}/(n - p)} \sim F_{p-r, n-p}$$
where $\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ is the error sum of squares. The quadratic form $\boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y}$ is sometimes called a *sequential sum of squares.* A large sequential sum of squares compared to the error sum of squares, thus a large F, indicates that the null hypothesis is false.

d) Use the acid rain data (see course www-page) to test whether all coefficients except the intercept and the coefficients of `x1` and `x3` are zero.

e) Show that if the model includes an intercept, and the null hypothesis is that all other parameters are zero, the projection matrix $\boldsymbol{H}_0 = \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, so that in that case the sequential sum of squares is equal to what is called the regression sum of squares. For the acid rain data, test whether all coefficients except the intercept are zero.

Problem 4: One- and two-way ANOVA – and the linear model

In this problem we will consider a data set where income is explained by the two factors place and sex.

	Place A	Place B	Place C
Male	300 350 370 360	400 370 420 390	400 430 420 410
Female	300 320 310 305	350 370 340 355	370 380 360 365

Hence two categorical predictors: sex with 2 levels (Male and Female) and place with 3 levels (named A, B and C).

a) Enter the data in R as a data.frame with three columns, income, gender and place. Make gender and place to be factors. Call this `data`.

Examine the data visually with e.g.

```
pairs(data)
plot(income~place,data=data)
plot(income~gender,data=data)
interaction.plot(data$gender, data$place, data$income)
plot.design(income~place+gender, data = data)
```

b) (One-way anova).
Consider first a model with a factor α_i occurring at $i = 1, \ldots, I$ levels, with K observations per level. Use the model

$$y_{ik} = \mu + \alpha_i + e_{ik}, \quad i = 1, \ldots, I \quad k = 1, \ldots, K \tag{1}$$

where $e_{11}, \ldots, e_{IK}$ iid and $e_{ik} \sim N(0, \sigma^2)$, $\forall i, k$. Assume place is the only factor. The design matrix X can be found by:

```
X = cbind(rep(1,length(data$income)),data$place=="A",
    data$place=="B",data$place=="C")
```

What is the rank of $X^T X$? Why do we need $X^T X$ to have full rank. How can we solve rank problems?

Hint: `qr(matrix)$rank`.

c) Fit the model:

```
model = lm(income~place-1,data=data,x=TRUE)
```

where `x=TRUE` tells the function to calculate the design matrix X, which is stored as `model$x`. Examine the results with `summary` and `anova`.

What parametrization is used? What is the interpretation of the parameters? Which null hypothesis is tested in the anova-call? What is the result of the hypothesis test?

d) Fit the models:

```
options(contrasts=c("contr.treatment","contr.poly"))
model1 = lm(income~place,data=data,x=TRUE)

options(contrasts=c("contr.sum","contr.poly"))
model2 = lm(income ~ place,data=data,x=TRUE)
```

We have talked about dummy- and effect encoding of categorical covariates. What are the parametrizations used here? What is the interpretation of the parameters and how do the parameter interpretations differ between the models in **c)** and **d)**?

(You can read about `contr.treatment` by typing `?contr.treatment`, and ignore `contr.poly`.)

e) Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, and $\boldsymbol{X}$ has dimension $n \times p$. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ be estimators in this model.

We want to test the linear hypotheses

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \text{ vs. } H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}.$$

Here $\boldsymbol{C}$ is a $r \times p$ matrix. This can be done using F_{obs}, which under the null has a Fisher-distribution with r and $n - p$ degrees of freedom.

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

We want to test the one-way ANOVA null hypothesis that there is no factor effect of place. Use the F_{obs} to do this both using the dummy-variable and the effect coding of the place-factor. Compare the results from the two coding strategies.

f) (Two-way anova) Suppose now that there are two factors: α_i at I levels and β_j at J levels, with K observations per level. Then a general model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \quad i = 1, \ldots, I \quad j = 1, \ldots, J \tag{2}$$

where $e_{111}, \ldots, e_{IJK}$ iid and $e_{ijk} \sim N(0, \sigma^2)$, $\forall i, j, k$.

Fit the model without interactions effects:

```
options(contrasts=c("contr.treatment","contr.poly"))
model3 = lm(income~place+sex,data=data,x=TRUE)
summary(model3)
anova(model3)

options(contrasts=c("contr.sum","contr.poly"))
model4 = lm(income~place+sex,data=data,x=TRUE)
summary(model4)
anova(model4)
```

What are the parameterizations? What is the interpretation of the parameters? Does the ANOVA table look different for the two parametrizations?

Finally, fit a model with interactions (model formula is `place*sex` for both the contrasts and check if the interaction effect is significant. Do this also using the F_{obs} from **e)**.

Problem 5: Teaching reading

In a randomized study the aim was to compare three methods for teaching reading, one method currently in use (A), and two new methods (B and C). A total of 66 pupils were randomly assigned to one of the three teaching methods, with 22 pupils for each method.

We will look at data on reading score. Reading score is a numerical value, and high value for the reading score is preferred. A box plot of the data is presented in Figure 8, and summary statistics are given in Table 1. (The standard deviation is calculated using the unbiased estimator.)
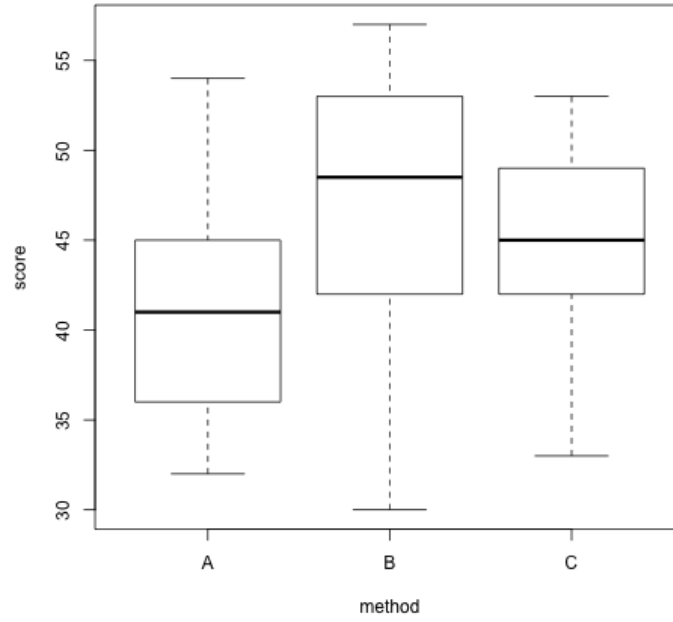
Figure 8: Box plot for reading data

Method	Sample size	Average	Standard deviation
A	22	41.05	5.636
B	22	46.73	7.388
C	22	44.27	5.767
Total	66	44.02	

Table 1: Summary statistics for the reading data

a) We would like to investigate if the expected reading score varies between the teaching methods. Write down the null and alternative hypotheses and perform a single hypothesis test based on the summary statistics in Table 1.
What are the assumptions you need to make to use this test?
What is the conclusion from the test?

b) The hypothesis test in **a)** can also be performed using results for linear hypotheses ($C\beta = d$) in a regression model ($Y = X\beta + \varepsilon$). What is the dimension of C in this case? Is there a simple way to calculate the F-statistic, $F_{obs} = \frac{1}{r}(C\hat{\beta} - d)^T(\hat{\sigma}^2 C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta} - d)$ in this case?

c) We will now compare the two new teaching methods, method B and C. Let μ_B and μ_C be the expected scores for teaching methods B and C. We would like to study the ratio, γ, between these two expected scores,

$$\gamma = \frac{\mu_B}{\mu_C}.$$

Suggest an estimator, $\hat{\gamma}$, for γ.

Use Taylor methods to approximate the expected value and standard deviation of this estimator, that is, $E(\hat{\gamma})$ and $SD(\hat{\gamma}) = \sqrt{Var(\hat{\gamma})}$.
Use the relevant data in Table 1 to calculate $\hat{\gamma}$, and the estimated approximate values for $E(\hat{\gamma})$ and $SD(\hat{\gamma})$ numerically.