Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4267 Linear statistical models
Recommended exercises 8

Problem 1 One- and two-way ANOVA – and the linear model

We consider a data set where income is explained by the two factors *place*, having levels *A*, *B* and *C*, and *gender*, having levels *male* and *female*.

	Place		
Gender	A	B	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

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

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)
```

One-way ANOVA. Consider first a model with one factor α_i occurring at levels $i = 1, \ldots,$ I, with K observations per level, that is,

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

where the e_{ik} are independent $N(0, \sigma^2)$.

Assume that `place` is the only factor α_i. We consider a design matrix X defined by the following R code:

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

b) What is the rank of $X^{\mathrm{T}}X$? Why do we need $X^{\mathrm{T}}X$ to have full rank? How can we solve rank problems? Hint: `qr(matrix)$rank` gives rank of `matrix`.

c) Fit the model using the following R code:

```
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 models using the following R code:

```
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`.)

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

Then, the linear hypothesis $H_0\colon C\boldsymbol{\beta} = \boldsymbol{d}$, with C an $r \times p$ matrix of rank r and $\boldsymbol{d}$ a vector of length r, can be tested against $H_1\colon C\boldsymbol{\beta} \neq \boldsymbol{d}$ by using

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

which under the null hypothesis has a Fisher distribution with r and $n-p$ degrees of freedom.

We want to test the one-way ANOVA null hypothesis that there is no factor effect of *place*.

e) Use F above to do this both using the dummy and the effect coding of the factor *place*. Compare the results from the two coding strategies.

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

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

where the e_{ijk} are independent $N(0, \sigma^2)$.

Assume that `place` is the factor α_i and `gender` the factor β_j

f) Fit the model using the following R code:

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

```
summary(model3)

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

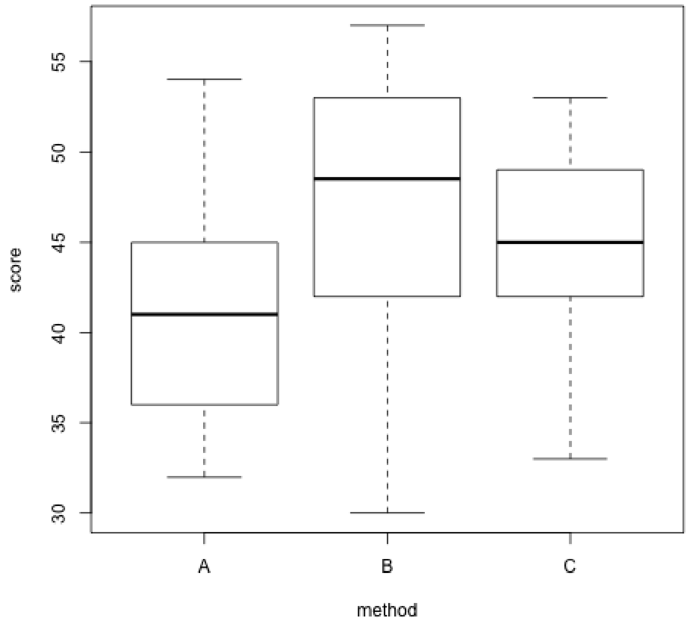
What are the parametrizations? What is the interpretation of the parameters? Does the ANOVA table look different for the two parametrizations?

Finally, fit a model with interactions (model formula `place + sex + place:sex`, or `place*sex`) and check if the interaction effect is significant. Do this also using the F above.

Problem 2 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.

Reading score is a numerical value, and high value for the reading score is preferred. A box plot and summary statistics of the data is given below. (The sample standard deviation is the square root of the unbiased variance estimate s^2.)



Method	Sample size	Average	Sample s.d.
A	22	41.05	5.636
B	22	46.73	7.388
C	22	44.27	5.767
Total	66	44.02	

We want to investigate whether the expected reading score varies between the teaching methods.

a) Write down the null and alternative hypotheses and perform a single hypothesis test based on the summary statistics above. What are the assumptions you need to make to use this test? What is the conclusion from the test?

Let $\gamma = \mu_\mathrm{B}/\mu_\mathrm{C}$ be the ratio between the expected scores μ_B and μ_C for teaching methods B and C, respectively.

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

Use a first-order Taylor expansion to approximate the expected value and standard deviation of this estimator, that is, $E\hat{\gamma}$ and $\mathrm{SD}\,\hat{\gamma} = \sqrt{\mathrm{Var}\,\hat{\gamma}}$. Use the relevant data in the table above to calculate $\hat{\gamma}$ and the estimated approximate values for $E\hat{\gamma}$ and $\mathrm{SD}\,\hat{\gamma}$ numerically.