TMA4267: Assignment 7

One-way ANOVA

We use table 13.6 in the textbook, which represents blood sugar levels (mg/100g) that were measured on 10 animals from each of five breeds (one factor, 5 levels, 10 observations per level). We perform one-way ANOVA. The model summary is given below.

(a) Explain how the model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ is related to the model summary. The design matrix used to fit the model presented below has 50 rows and 5 columns. Write out a row corresponding to an animal of breed E and a row corresponding to an animal of breed A. What is the estimated 'effect' of Breed E?

(b) Use the model summary to perform a test of breed, i.e. $H_0 : \alpha_i = 0$ for all i vs $H_1 : \alpha_i \neq 0$ for at least one i, $i = 1, \ldots, 5$.

(c) Explain the steps in a test for investigating wether the two breeds A and B are different (hint: $H_0 : \mathrm{C}\beta = 0$).

```
Call:
lm(formula = blood_sugar ~ breed, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-18.500  -8.675   1.000   9.075  18.100

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.680      1.561  78.571  < 2e-16 ***
breedA        -5.680      3.123  -1.819  0.07559 .
breedB        -4.680      3.123  -1.499  0.14094
breedC         7.220      3.123   2.312  0.02541 *
breedD        -8.680      3.123  -2.780  0.00792 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.04 on 45 degrees of freedom
Multiple R-squared:  0.3694, Adjusted R-squared:  0.3134
F-statistic: 6.591 on 4 and 45 DF,  p-value: 0.0002906
```

Two-way ANOVA

We use table 14.7 in the book, which presents measurements of hemoglobin in the blood of brown trout after treatment with sulferazine. The experiment is described in question 14.34.

(a) Identify the number of factors and the number of levels per factor. Write down the two-way ANOVA model with interaction terms. Make sure you understand what all the terms represent.

(b) The ANOVA table is given below. Fill inn the question marks. Is the interaction significant? Are any of the factors significant? Use a Bonferroni correction.

```
            Df  Sum Sq Mean Sq F value    Pr(>F)
Rate         ?  90.972 30.3241 19.7771 1.872e-09 ***
Method       ?   2.278  ?      1.4858    0.2269
Rate:Method  ?   4.971  1.6571  ?        0.3628
Residuals    ? 110.397  1.5333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Exam 2016, modified

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 used:

- 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 the chosen value of D, F_i is the chosen value of F, and T_i is the 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 are 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.

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

```
# 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
```

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.

(b) How would you, based on Figures 2 and 3, 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?

The researchers want to test the hypothesis

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

(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 in the model summary.

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 are independent and identically normally distributed with mean 0 and variance σ^2. Output from fitting the reduced model is given in Figure 7.

(d) 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 $D = 1$, $F = 1$, $T = -1$.

```
> 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
```

Other relevant exam questions for Part 3

- 2015: Problem 2
- 2017: Problem 3
- Summer 2025: Problem 4

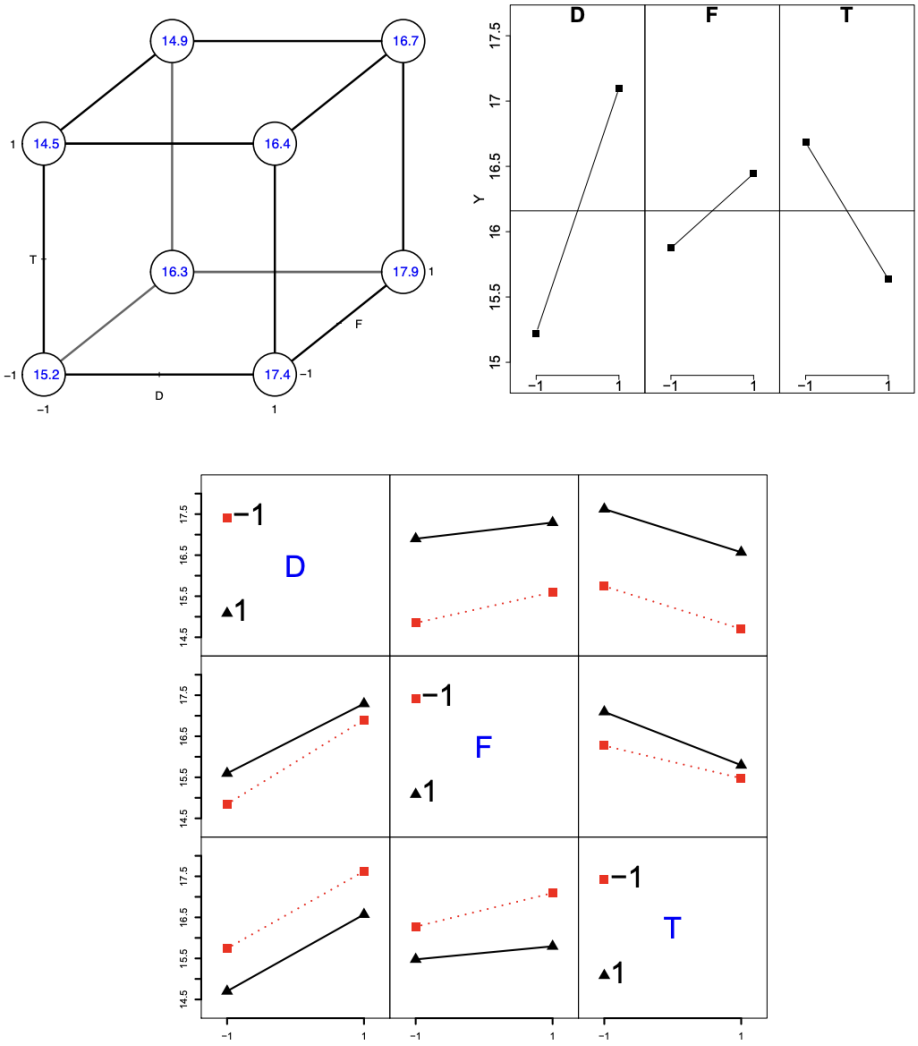

Figure 2: Cube plot (upper left panel), main effects plot (upper right panel) and interaction effects plot (lower panel) for the full model fitted to the plant stress data set.

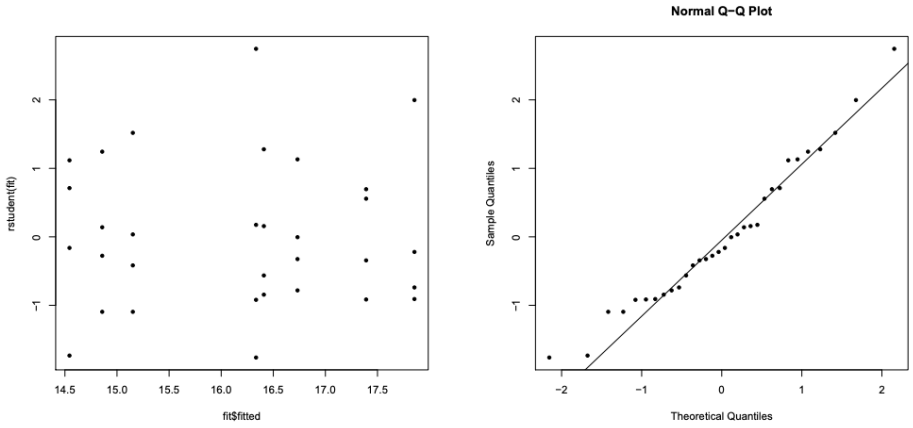

Figure 3: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for the full model fitted to the plant stress data set.