Linear Regression (NMSA407)

Test **Version** – MM1

- Solutions can be worked out in either of languages: English, Czech, Slovak.
- Although the answer may be very short (e.g. only one number, or one word), it should be clear how this answer was derived.

Task 1 (16 points)

We would like to investigate how the occipital angle of humans depends on gender (`fgender`) and population type (`fpopul`). There are **two levels for gender** considered (males and females) and **three different groups (levels) for the population** (Australian, Berg in Austria, and Burjati in Siberia).

Using the available data the following model is fitted

```
oca ~ fpopul * fgender
```

and the ANOVA table of **type I** is obtained:

```
                 Df Sum Sq Mean Sq F value  Pr(>F)  
fpopul            ?      ?       ?  3.0497 0.04926 *
fgender           ?      ?       ?  3.6888 0.05599 .
fpopul:fgender    ?      ?       ?  3.8722 0.02216 *
Residuals       234 5789.6       ?                  
```

(i) If possible, replace the question marks in the output above by the appropriate values.

(ii) Explain, how the p-values in the last column of the output above are calculated.

Task 2 (24 points)

We are interested how the time of the operation (time) depends on the gender and the age of the patient (gender, age), size of the kidney stone (stone) and the surgeon who performed the operation (fsurgeon). For better interpretation purposes the **age and size covariates were both lowered by their median values, 60 and 15 respectively** (age60, size15). Further, there are four surgeons (labelled subsequently as 1,2,3,4) and **the contrast sum parametrization** (contr.sum) was used with the following parametrization matrix:

```
           fsurgeon1 fsurgeon2 fsurgeon3
surgeon 1      1         0         0
surgeon 2      0         1         0
surgeon 3      0         0         1
surgeon 4     -1        -1        -1
```

Based on the observed data we estimated the following model

```
time ~ gender + fsurgeon + size15 + age60
```

and we get the following output

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.7063     3.5481  16.264  < 2e-16 ***
gendermale    9.0778     3.5935   2.526  0.01263 *
fsurgeon1     0.7036     3.3568   0.210  0.83428
fsurgeon2   -13.1560     3.2404  -4.060 8.08e-05 ***
fsurgeon3    11.5257     3.7075   3.109  0.00227 **
size15        0.8211     0.1567   5.239 5.71e-07 ***
age60        -0.2272     0.1290  -1.761  0.08032 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.06 on 142 degrees of freedom
Multiple R-squared:  0.3108,Adjusted R-squared:  0.2817
F-statistic: 10.67 on 6 and 142 DF,  p-value: 8.991e-10
```

(i) Describe the estimated effect of the gender on the expected time of operation.

(ii) Describe the estimated effect of the size of the kidney stone on the expected time of operation.

(iii) Compare the estimated expected times of operations for surgeons 2 and 3. Is this difference significant? Provide the corresponding p-value, if possible.

(iv) Let Y_i (for $i = 1, \ldots, n$) be the time of the operation and $\widehat{Y}_i$ the corresponding fitted value (based on the model above). Calculate $\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$.

Task 3 (24 points)

We are interested how the salary of associate professors (salary.assoc) depends on the number of professors (n.prof), number of associate professors (n.assoc) and the type of the school (type). To prevent heteroscedasticity the **logarithmic transformation of the salary of associate professors** (lsalary.assoc) was considered, that is lsalary.assoc $= \log($salary.assoc$)$. The number of professors and the number of associate professors were both lowered by 40 (roughly the median value for both) introducing so two new covariates n.prof40 and n.assoc40. Further, the variable type has 3 levels labelled subsequently as I, IIA, IIB and for this variable **the standard (R default) parametrization** (contr.treatment) was used.

Based on the observed data we estimated the following model

```
lsalary.assoc ~ type * n.prof40 + n.assoc40
```

and we get the following output

```
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       6.144e+00  1.777e-02 345.673  < 2e-16 ***
typeIIA          -9.035e-02  1.914e-02  -4.720 2.66e-06 ***
typeIIB          -1.675e-01  1.908e-02  -8.779  < 2e-16 ***
n.prof40          4.401e-05  6.595e-05   0.667  0.50469    
n.assoc40         1.284e-04  1.035e-04   1.241  0.21471    
typeIIA:n.prof40  2.933e-04  9.699e-05   3.024  0.00255 ** 
typeIIB:n.prof40  3.469e-03  2.472e-04  14.031  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1296 on 1118 degrees of freedom
Multiple R-squared:  0.4409,Adjusted R-squared:  0.4379 
F-statistic:   147 on 6 and 1118 DF,  p-value: < 2.2e-16
```

(i) Interpret the intercept parameter estimate, value 6.144 in the output above.

(ii) Describe the estimated effect of the number of professors (n.prof) on the salary of associate professors (salary.assoc) at the school of type IIA.

(iii) Estimate the expected salary of an associate professor (salary.assoc) at the school of type I with 50 professors (n.prof) and 100 associate professors (n.assoc).

(iv) Compare the salaries of associate professors (salary.assoc) at two different schools both with 100 professors (n.prof) and 100 associate professors (n.assoc) but one is of type I and the other one of type IIB.

(v) Explain in detail how the p-value on the line starting with n.assoc40 is calculated?

Task 4 (36 points)

Let$\{(X_i, Z_i, Y_i)^\mathsf{T};\ i = 1, \ldots, n\}$ be a random sample from some joint density function $f(x, z, y)$. Suppose, that $f(x, z, y) = f(y|x, z) \cdot f(x, z)$ where the conditional distribution for $Y_i|(X_i, Z_i)$ is $\mathsf{N}(\beta_1 X_i + \beta_2 Z_i, \sigma^2)$, for $\sigma^2 = 1$ and the marginal distribution of (X, Z) is some continuous distribution such that $\mathsf{P}(X_i > 0) = 1$ and $\mathsf{P}(Z_i > 0) = 1$. Let $\boldsymbol{\theta} = (\beta_1, \beta_2)^\mathsf{T}$ be the vector of unknown parameters and function $f(x, z)$ (marginal density of (X, Z)) does not depend on $\boldsymbol{\theta}$.

(i) Find the maximum likelihood estimator of the unknown vector parameter $\boldsymbol{\theta}$.

(ii) Find the confidence interval for the parameter β_1.

(iii) Derive a test of the null hypothesis that $\beta_2 = 0$.