Linear Regression (NMSA407)

Test **Version** – Sample tasks

Solutions can be worked out in English, Czech, or Slovak.

Although the answer may be very short (e.g. only one number, or one word), it must be clear how this answer was derived.

Not all questions can be answered, based on the given input. If a question cannot be answered, provide a reason for it.

Task

We want to predict the expected yield (covariate `yield`) of grain given the observed concentration of magnesium (`Mg`, in mg) and nitrogen (`N`, in mg) in their leafs. The covariate `Mg` is continuous and considered after a logarithmic transformation (`lMg = log2(Mg)`); The covariate `N` is included as a categorical predictor `flN` with three levels `low`, `medium`, and `high`, according to the numerical values of `N`. The categorical covariate is parametrized by standard contrasts `contr.treatment` and the logarithmic transformation of the response is used in the model (`lyield = log(yield)`).
The following model is fitted

```
m1 <- lm(lyield ~ lMg * flN, data = Dris)
```

and the following summary table is obtained:

```
Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)    -0.4618     0.4324  -1.068   0.2863
lMg             0.6133     0.1301   4.714 3.48e-06 ***
flNmedian       0.8689     0.5730   1.517   0.1303
flNhigh         1.1336     0.5133   2.208   0.0279 *
lMg:flNmedium  -0.2834     0.1694  -1.673   0.0952 .
lMg:flNhigh    -0.3732     0.1504  -2.481   0.0136 *

Residual standard error: 0.2239 on 362 degrees of freedom
Multiple R-squared:  *** , Adjusted R-squared:  ***
F-statistic: 8.477 on 5 and 362 DF,  p-value: 1.331e-07
```

(i) Interpret the effect of magnesium (`Mg`) on the expected yield (`yield`).

(ii) Explain in detail how the test statistic and the p-value in the row that starts with `flNmedium` is computed. What is being tested there, and what is the conclusion of that test?

(iii) Is the nitrogen concentration a significant modifier of the effect of magnesium on the expected logarithmic yield? Provide a p-value of a formal test.

(iv) Compare the difference in the expected logarithmic yield of grain between low and high level of nitrogen concentration in the leafs if the underlying concentration of magnesium is 1 mg. If possible, provide the 95% confidence interval for this difference.

(v) If possible, fill in the values for Multiple R-squared and Adjusted R-squared.

(vi) Consider the magnesium transformation `lMg100 = log2(100*Mg)` and the model $m2$ analogous to model $m1$:

```
m2 <- lm(lyield ~ lMg100 * flN, data = Dris).
```

Which rows of the summary table above will be unaffected?

(vii) If possible, find the point estimates of the regression coefficients in $m2$ from part (vi).

Task

We want to predict the mean salary of an associate professor given the number of full professors (`n.prof`), the number of associate professors (`n.assoc`) and the university type I, IIA and IIB (`type`). The **contrast sum parametrization** (`contr.sum`) is used for the factor covariate and the continuous covariates are lowered by 40 (obtaining covariates `n.prof40` and `n.assoc40`).

The following model is fitted

```
salary.assoc ~ (n.prof40 + n.assoc40) * type
```

and the corresponding summary output is obtained:

```
Coefficients:
                Estimate  Std. Error t value Pr(>|t|)
(Intercept)     438.36946    2.91249 150.514  < 2e-16 ***
n.prof40          0.33196    0.04861   6.829 1.40e-11 ***
n.assoc40         0.40828    0.06650   6.139 1.15e-09 ***
type1            45.04159    5.17765   8.699  < 2e-16 ***
type2           -13.80097    3.59080  -3.843 0.000128 ***
n.prof40:type1   -0.19529    0.05161  -3.784 0.000162 ***
n.prof40:type2   -0.28168    0.05509  -5.113 3.73e-07 ***
n.assoc40:type1  -0.60532    0.07290  -8.303 2.90e-16 ***
n.assoc40:type2  -0.09509    0.08071  -1.178 0.238970

Residual standard error: 52.67 on 1116 degrees of freedom
Multiple R-squared:  0.4618,Adjusted R-squared:  0.458
F-statistic: 119.7 on 8 and 1116 DF,  p-value: < 2.2e-16
```

(i) Interpret the point estimate on the line that starts with `n.assoc40`.

(ii) Describe the effect of the number of full professors (`n.prof40`) on the expected salary of an associate professor at all three university types (I, IIA, and IIB).

(iii) Compare the expected salary of an associate professor at the university of type IIA and the university of type IIB if there are 60 full professors and 60 associate professors at both universities. Is this difference statistically significant? Provide the corresponding p-value if possible.

(iv) Can we say, that the university type is a significant modifier of the effect of the number of associate professors on the salary of an associate professor? Provide a formal test and provide the corresponding p-value if possible.

(v) Where possible, complete the following ANOVA table of type II.

	Sum Sq	Df	F value	Pr(>F)
n.prof40	.	.	.	.
n.assoc40	.	.	.	.
type	.	.	.	.
n.prof40:type	.	.	.	.
n.assoc40:type	.	.	.	.
Residuals	.	.		

(vi) Where possible, complete the following ANOVA table of type III.

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	.	.	.	.
n.prof40	.	.	.	.
n.assoc40	.	.	.	.
type	.	.	.	.
n.prof40:type	.	.	.	.
n.assoc40:type	.	.	.	.
Residuals	.	.		

(vii) Which lines of the ANOVA tables of type II and III coincide?

(viii) Which lines of the ANOVA tables of types II and III above will change, if we consider:

(a) the original covariates `n.prof` and `n.assoc` instead of `n.prof40` and `n.assoc40`?

(b) standard (contr.treatment) parametrization instead of the contrast sum?

(c) logarithmic transformation of the response (`salary.assoc`)?

Task

We want to predict the expected yield (covariate `yield`) of grain given the observed concentration of magnesium (`Mg`, in mg), calcium (`Ca`, in mg), and nitrogen (`N`, in mg) in their leafs. Covariates `Mg` and `Ca` are continuous and considered after a logarithmic transformation (`lMg = log(Mg)`, `lCa = log(Ca)`); covariate `N` is included as a categorical predictor `flN` with three levels `low`, `medium`, and `high`, according to the numerical values of `N`. The categorical covariate is parametrized by standard contrasts `contr.treatment`. We have $n = 368$ independent observations.
We fitted the model

```
m <- lm(yield ~ (lMg + lCa)*flN)
```

The following table is obtained by calling `Anova(m,type="II")`:

```
Response: yield
          Sum Sq  Df F value    Pr(>F)
lMg            .   .       . 1.286e-09 ***
lCa            .   .       .  0.127618
flN            .   .       .  0.001287 **
lMg:flN        .   .       .  0.779287
lCa:flN     1.85   .       .  0.425659
Residuals 387.93   .
```

(i) Where possible, fill in the blanks in the table above.

(ii) Explain in detail how the F-value and the p-value in the row that starts with `flN` are computed. What is being tested there, and what is the conclusion of that test? Specify the null and the alternative model in the test.

(iii) If possible, compute the F-statistic of the test of the null hypothesis that the conditional expectation of `yield` is independent of all the regressors, and provide the critical region of this test.

(iv) Is it possible to say whether `flN` is an effect modifier of `lCa`? State the null and the alternative hypothesis of this test, and find the value of the test statistic and the p-value of this test.

(v) Consider variables `lMg100 = log(100*Mg)` and `lCa100 = log(100*Ca)`, and the model

```
m2 <- lm(yield ~ (lMg100 + lCa100)*flN, contr=list(flN=contr.SAS)).
```

Which rows in the output from `Anova(m2,type="II")` below coincide with those from the table above?

```
            Sum Sq  Df F value    Pr(>F)
lMg100           .   .       .         .
lCa100           .   .       .         .
flN              .   .       .         .
lMg100:flN       .   .       .         .
lCa100:flN       .   .       .         .
Residuals        .   .
```

Task

We want to estimate the mean percentage of body fat of police applicants (`fat`, in %). We have information about the height of the applicants (`height`, in cm).

The following model is fitted

```
m <- lm(fat ~ I(height-180) + I((height-180)^2)+ I(height<170), data = Policie)
```

and the corresponding summary output is obtained:

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         13.639582   1.246149  10.945 2.13e-14 ***
I(height - 180)      0.359311   0.181064   1.984   0.0532 .
I((height - 180)^2) -0.002209   0.022380  -0.099   0.9218
I(height < 170)TRUE  0.737430   4.329414   0.170   0.8655
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.621 on 46 degrees of freedom
Multiple R-squared:  0.1186,Adjusted R-squared:  0.06113
F-statistic: 2.063 on 3 and 46 DF,  p-value: 0.1181
```

(i) Based on model `m`, specify the function that describes the conditional expectation of the applicants' fat given their height.

(ii) Describe the effect of the height on the percentage of body fat.

(iii) Based on model `m`, test whether the true relation of the expected body fat and height is linear, i.e. whether it holds that $\mathsf{E}\,(\texttt{fat}|\texttt{height}) = \alpha + \beta\,\texttt{height}$ for some $\alpha, \beta \in \mathbb{R}$. Specify the null and the alternative hypothesis, and provide a p-value if possible. Is it possible to use Bonferroni's correction?

(iv) Find a prediction interval for the percentage of body fat of an applicant whose height is 180 cm.

(v) Complete the following ANOVA table of type III.

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	.	.	.	.
I(height - 180)	.	.	.	.
I((height - 180)^2)	.	.	.	.
I(height < 170)	.	.	.	.
Residuals	.	.		

Task

We want to predict the mean weight of the roots of certain plants (`weight`) based on the knowledge of the sugar percentage in a nutrient solution where the plant was grown (`fpercentage`), and the pH of the soil (`pH`). Variable `pH` is continuous; `fpercentage` is a factor with four levels 1–4.

We fit the model

```
lm(log(weight) ~ fpercentage*pH, data=Koreny, contr=list(fpercentage=contr.sum))
```

The following summary table is obtained:

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.24590    0.17943  -6.944 1.11e-08 ***
fpercentage1    -0.34478    0.18425  -1.871   0.0677 .
```

```
fpercentage2     0.07292    0.22779   0.320   0.7503
fpercentage3     0.14256    0.28435   0.501   0.6185
pH               0.03551    0.04098   0.866   0.3907
fpercentage1:pH  0.06582    0.05893   1.117   0.2699
fpercentage2:pH  0.06399    0.07263   0.881   0.3829
fpercentage3:pH -0.02874    0.06846  -0.420   0.6766
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2281 on 46 degrees of freedom
Multiple R-squared:  0.6345,Adjusted R-squared:  0.5789
F-statistic: 11.41 on 7 and 46 DF,  p-value: 2.757e-08
```

(i) Interpret the regression coefficient that is estimated in line `pH`.

(ii) Explain how multiple R-squared is computed, and interpret this value.

(iii) Estimate the expected weight of a root if `fpercentage`=3 and `pH`=6.

(iv) Test whether it can be asserted that the expected value of the logarithm of `weight` given `fpercentage`=3 and `pH`=6 is negative. Provide a test statistic, critical region, and the p-value of this test.

Task

We model the median house value in the suburbs of Boston (covariate `medv`) given the level of NOx concentration in the area (`nox`, in parts per 10 million), the indicator of criminality in the neighbourhood (`fcrime`), and the median age of houses (`fage`). Covariate `nox` is continuous; covariates `fcrime` and `fage` are categorical (with levels `None`, `Some`, and `High` for `fcrime`; and levels `Old` and `New` for `fage`. Both categorical covariates are parametrized by standard contrasts `contr.treatment`).
We fitted the model

```
m1 = lm(medv ~ nox*(fcrime + fage) + I(nox^2):fage)
```

The following table is obtained by calling `anova(m1)`:

```
               Df Sum Sq Mean Sq  F value    Pr(>F)
nox             .   7800       . 115.9001 < 2.2e-16 ***
fcrime          .    746       .   5.5405  0.004171 **
fage            .     51       .   0.7534  0.385826
nox:fcrime      .    371       .   2.7536  0.064672 .
nox:fage        .     70       .   1.0402  0.308268
fage:I(nox^2)   .    298       .   2.2144  0.110295
Residuals     496  33381       .
```

(i) Where possible, fill in the blanks in the table above (columns `Df` and `Mean Sq`).

(ii) Explain in detail how the F-value and the p-value in the row that starts with `fage` are computed. What is being tested there, and what is the conclusion of that test? Specify the null and the alternative model in the test.

(iii) If possible, find the number of observations, residual degrees of freedom, (an estimate of) the residual variance, and the total sum of squares in the model.

(iv) If possible, find the coefficient of determination and the adjusted coefficient of determination in `m1`. Compute the F-statistic of the test of the null hypothesis that the conditional expectation of `medv` is independent of all the regressors, and provide the critical region of this test.

(v) Which p-values in the output from `drop1(m1,test="F")` below can be recovered from the table above?

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			.	.		
nox:fcrime	.	.	.	.	.	.
nox:fage	.	.	.	.	.	.
fage:I(nox^2)	.	.	.	.	.	.

Task

We want to estimate the mean salary of associate professors across different universities in the United States (`salary.assoc`). For this purpose we obtain the number of professors (`n.prof`) at each university. We recognize three university types (`type`): I, IIA, and IIB.

For the factor covariate `type` a **contrast sum parametrization** (`contr.sum`) is used.

The following model is fitted

```
salary.assoc~type*(n.prof + I(n.prof^2))
```

and the corresponding summary output is obtained:

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.952e+02  5.093e+00  77.594  < 2e-16 ***
type1              5.615e+01  9.143e+00   6.141 1.14e-09 ***
type2              1.787e+01  6.231e+00   2.867  0.00422 **
n.prof             1.121e+00  8.564e-02  13.087  < 2e-16 ***
I(n.prof^2)       -4.941e-03  6.891e-04  -7.171 1.36e-12 ***
type1:n.prof      -9.560e-01  9.423e-02 -10.145  < 2e-16 ***
type2:n.prof      -7.705e-01  9.920e-02  -7.767 1.81e-14 ***
type1:I(n.prof^2)  4.814e-03  6.903e-04   6.973 5.32e-12 ***
type2:I(n.prof^2)  4.480e-03  6.985e-04   6.414 2.10e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.97 on 1116 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.4556,Adjusted R-squared:  0.4517
F-statistic: 116.8 on 8 and 1116 DF,  p-value: < 2.2e-16
```

(i) Interpret the estimated intercept parameter.

(ii) Describe the effect of the number of professors on the salary of associate professors.

(iii) Estimate the difference in the mean salary of an associate professor at universities of type IIA and IIB. Does the difference depend on the number of professors? Can you assess whether it is statistically significant? Provide a p-value if possible.

(iv) We conjecture that the mean salary of an associate professor at a university of type IIB with 100 professors is at least 400. Formulate the null and the alternative hypothesis, and provide a test for this hypothesis (i.e. the test statistic and the critical region).

Task

We want to predict the mean grain yield (`yield`) based on the knowledge of the chemical composition of the soil. We observe the concentration of phosphorus (`P`) and magnesium (`Mg`).

We fit the model

```
log(yield) ~ P + Mg
```

The following summary table is obtained:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.572843   0.084305  18.657  < 2e-16 ***
P           -0.007886   0.001516  -5.203 3.28e-07 ***
Mg           0.031465   0.005731   5.490 7.54e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2226 on 365 degrees of freedom
Multiple R-squared:  0.1079,Adjusted R-squared:  0.103
F-statistic: 22.08 on 2 and 365 DF,  p-value: 8.863e-10
```

(i) Describe the effect of the concentration of magnesium on the yield.

(ii) Provide a confidence interval for the difference in mean log(yield) if the concentration of magnesium increases by 1.

(iii) Can we conclude that the effect of magnesium on the mean yield is positive? That is, based on our model, can we say that increasing magnesium concentration results in increased yield? Formulate the null and the alternative hypothesis, and conduct a formal test (i.e. provide the test statistic and the critical region).

(iv) Consider the model

```
log(yield) - mean(log(yield)) ~ I(P-50) + I(Mg-10)
```

Which lines of the summary table above will remain the same for this new model? Calculate the point estimates of the regression coefficients in the new model.

Task

We would like to model a relationship between the occipital angle (`oca`) and two covariates: a gender (`fgender`) and population (`fpopul`) to which the subject belongs. The gender covariate is a factor with two levels `female` and `male`. The population covariate has three levels labelled as `AUSTR`, `BERG`, `BURIAT`. Using the available data the following model was fitted:

```
oca ~ fpopul * fgender
```

For the fitted model the Anova of type I table was calculated:

```
Anova Table (Type I tests)

Response: oca
               Sum Sq  Df F value  Pr(>F)
fpopul          330.0   ?   ?       0.005
fgender         120.0   ?   4.0     0.020
fpopul:fgender  160.0   ?   ?       0.072
Residuals      6000.0 200
```

(i) Instead of the question marks in the Anova table above fill in the appropriate numbers.

(ii) Using the Anova table above, can we conclude that the differences in mean occipital angles for males and females are statistically significantly different across different populations (using a significance level of $\alpha = 0.05$)? Provide the corresponding p-value and the test statistic value.

(iii) If possible, calculate $\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$.

Task

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

We would like to model the surgery time (`time`) given the kidney stone size lowered by its sample median value, which is 15 (`size15`), and the information on which of four surgeons performed the surgery (four level factor covariate `fsurgeon`).

Using the available data and considering **the logarithm transformation of the response** (`ltime`) the following model is fitted

```
ltime ~ fsurgeon + size15
```

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

```
           Sum Sq  Df F value    Pr(>F)
fsurgeon   4.1776   3  12.511 2.568e-07 ***
size15     2.4785   1  22.268 5.559e-06 ***
Residuals 16.0281 144
```

Next, the ANOVA table of **type III** is calculated. If possible, replace the question marks in the table below with the appropriate values from the Anova II table above.

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	?	?	?	?
fsurgeon	?	?	?	?
size15	?	?	?	?
Residuals	?	?		

Task

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

We are interested in the number of fires (`fire`) in some specific town district given the percentage of minority decreased by 25 (`minor25`) and the location of the district (covariate `fside` with two levels - `North` and `South`). Considering the given data we obtain the following model:

```
log(fire) ~ minor25 * fside.
```

The north district `fside = North` was considered to be a reference category (`contr.treatment` in R). The following ANOVA table of **type II** was obtained:

```
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
minor25       1 13.0      13.0    43.3       7.2e-08 ***
fside         1  1.0      1.0      3.3       0.075   .
minor25:fside 1  1.5      1.5      5.0       0.031   *
Residuals    40 12.0      0.3
```

Suppose that we are interested in ANOVA table of **type I**. Fill in the question marks in the following table where possible.

```
               Df  Sum Sq  Mean Sq  F value    Pr(>F)
minor25         ?    ?       ?         ?         ?
fside           ?    ?       ?         ?         ?
minor25:fside   ?    ?       ?         ?         ?
Residuals       ?    ?       ?
```

Task

We are interested in how does the ground area of a yard (`ground`) of a house depend on its location (`locat`, which is a categorical covariate with two levels `North`, `South`) and the number of floors of that house (`floor`, which is a categorical covariate with three levels `1`, `2`, `3 and more`). We use the following linear model

```
ground ~ floor + locat + floor:locat
```

The corresponding ANOVA table (type I) was calculated:

```
Analysis of Variance Table

Response: ground
               Df Sum Sq Mean Sq F value Pr(>F)
floor           @   300     150     6.0   0.003
locat           @    @      122.5   4.9   0.029
floor:locat     @    *       *      4.0   0.019
Residuals     300   7500     ?
```

(i) Fill in the values in the table instead of `@` (4×).

(ii) Fill in the values in the table instead of (*, 2×).

(iii) Can we, based on the ANOVA table above decide, whether the model

`ground ~ floor + locat`

is significantly better than model

`ground ~ locat,`

if we based our decision on a statistical test at significance level 5 %? If we can, provide an appropriate p-value, and the value of the test statistic. If not, explain why.

Task

We are interested in how the occipital angle (`oca`) is influenced by the gender (`fgender`) and population (`fpopul`). We have two levels of the gender (`female`, `male`) and three levels of the population (`AUSTR`, `BERG`, `BURIAT`). The number of the observations in each of the groups are

```
 female:AUSTR   female:BERG female:BURIAT  male:AUSTR   male:BERG male:BURIAT
      49              53         54            22           56       55
```

Given the available data the following linear regression model was fitted:

```
oca ~ fpopul + fgender + fpopul:fgender
```

and ANOVA table of **type I** was calculated:

```
Analysis of Variance Table

Response: oca
               Df  Sum Sq   Mean Sq   F value   Pr(>F)
fgender         1    73       73       2.9     0.091
fpopul          2   372      186       7.3     0.001
fgender:fpopul  2   165       83       3.3     0.040
Residuals     283  7161       25
```

Suppose you are now interested in ANOVA table of **type II**. Where possible replace the question marks and fill in the appropriate numbers in the ANOVA table of **type II** that is given below.

```
Anova Table (Type II tests)

Response: oca
               Sum Sq  Df  F value   Pr(>F)
fgender          ?      ?    ?         ?
fpopul           ?      ?    ?         ?
fpopul:fgender   ?      ?    ?         ?
Residuals        ?      ?
```

Task

We would like to model a relationship between the occipital angle (`oca`) and two covariates: a gender (`fgender`) and population (`fpopul`) to which the subject belongs. The gender covariate is a factor with two levels `female` and `male`. The population covariate has three levels labelled as `AUSTR`, `BERG`, `BURIAT`. Using the available data the following model was fitted:

```
oca ~ fpopul * fgender
```

For the fitted model the Anova of type II table was calculated:

```
Anova Table (Type II tests)

Response: oca
               Sum Sq  Df F value  Pr(>F)
fpopul          330.0   ?  5.50     0.005
fgender         120.0   ?   ?       0.020
fpopul:fgender  160.0   ?   ?       0.072
Residuals      6000.0 200
```

(i) Give an explanation on how the p-value in the first line of of the Anova table is obtained (the line which corresponds to `fpopul` with p-value 0.005).

(ii) Instead of the question marks in the Anova table above fill in the appropriate numbers. (2 points)

(iii) Using the Anova of type II table above, can we conclude that the differences in mean occipital angles for males and females are statistically significantly different across different populations (using a significance level of $\alpha = 0.05$)? Provide the corresponding p-value and the test statistic value.

Task

We are interested how the heart rate (`HR`) is affected by the type of disturbing stimulus (`stimulus`) and the amount of time that the horse spends outside (`outside`). The variable

`stimulus` has seven levels and the variable `outside` has three levels. Based on the observed data the following model was fitted

```
HR ~ fstimulus * foutside
```

and based on this model the following ANOVA table of type I was calculated:

```
Analysis of Variance Table

Response: HR
                   Df Sum Sq  Mean Sq  F value  Pr(>F)
stimulus            6  90000    15000    15.0    < 0.001
outside             ?    ?         ?      7.0      0.001
stimulus:outside    ?    ?         ?      2.0      0.025
Residuals         245 245000       ?
```

In the table above replace the question marks and fill in the appropriate numbers.

Task

We are interested in how the amount of yield (`yield`) depends on magnesium (Mg) and nitrogen (N). For this reason we introduce the following variables: `lMg` which is the logarithm of Mg (i.e. `lMg` = log(`Mg`)), `lN` - logarithm of N (i.e. `lN` = log(`N`)), and `lMg2` = $\big[\log(\texttt{Mg})\big]^2$, `lN2` = $\big[\log(\texttt{N})\big]^2$. In addition, to limit heteroscedasticity in the model we use **the logarithmic transformation of the response** (`lyield`).

Based on available observations we estimated the following model

```
lyield ~ lN + lN2 + lMg + lMg2 + lN:lMg.
```

and we got the following output

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.3644    13.1323   0.485  0.62823
lN           -3.6275     4.6769  -0.776  0.43848
lN2           0.2783     0.4636       ?  0.54861
lMg           5.9233     3.8450   1.541  0.12430
lMg2         -0.9281     0.3497  -2.654  0.00829
lN:lMg       -0.1288     0.7744  -0.166  0.86794
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2178 on 362 degrees of freedom
Multiple R-squared:  0.1523,  Adjusted R-squared:  0.1406
F-statistic: 13.01 on 5 and 362 DF,  p-value: 1.184e-11
```

Based on the output above answer the following questions.

(i) Specify the effect of magnesium on the yield.

(ii) Is magnesium a statistically significant modifier of the effect of the nitrogen on the yield?
(Do not forget to explain your answer).

(iii) If possible, give the p-value of the test of the submodel
`lyield ~ lN + lN2 + lMg + lN:lMg.`

against the model
`lyield ~ lN + lN2 + lMg + lMg2 + lN:lMg.`

(iv) Replace the question mark (`?`) in the summary output above with the appropriate value. Explain, how the p-value on the same line is calculated.

(v) Let Y_i (for $i = 1, \ldots, n$) be the corresponding value of the logarithm of the yield for i-th observation 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

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

We are interested how time of the operation (`time`) depends on the gender of the patient (`gender`), size of the kidney stone (`stone`) and the surgeon who performed the operation (`fsurgeon`). To prevent heteroscedasticity the **logarithmic transformation of the time** was considered. Further, there are four surgeons (labelled subsequently as 1,2,3,4) and `contr.treatment` parametrization was used and we included also the quadratic term for the size of the kidney stone (i.e. `I(size^2)`).

Based on the observed data we estimated the following model

```
log(time) ~ gender + fsurgeon + size + I(size^2)
```

and we get the following output

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.7070     0.1085 34.1800   0.0000
gendermale     0.1213     0.0559  2.1689   0.0318
fsurgeon2     -0.2783     0.0628 -4.4303   0.0000
fsurgeon3      0.1250     0.0747  1.6739   0.0964
fsurgeon4      0.0778     0.1456  0.5342   0.5940
size           0.0298     0.0089  3.3653   0.0010
I(size^2)     -0.0003     0.0002 -2.0576   0.0415

Residual standard error: 0.3266 on 142 degrees of freedom
Multiple R-squared:  0.3252,    Adjusted R-squared:  0.2967
F-statistic: 11.41 on 6 and 142 DF,  p-value: 2.189e-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 4.

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

Task

We would like to model the logarithm of the number of fires (`fire`) in some town district given the percentage of a minority (`minor`) in this district (the locality where the district is situated is the factor variable `fside` with two levels - `North` and `South`) and the number of insurance policies (variable `insur`). Note that a **logarithmic** transformation of the number of fires is used as a response and a **contrast sum parametrization** (`contr.sum`) with the following contrast-matrix

```
      fside
North    1
South   -1
```

is used for the variable `fside`.

Based on the observed data we estimated the following model

```
log(fire) ~ minor + fside + insur + minor:fside + insur:fside
```

and we obtained the summary output below:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.579046   0.111366  14.179  < 2e-16 ***
minor         0.010833   0.003512   3.085  0.00364 **
fside1       -0.034162   0.111366  -0.307  0.76058
insur         0.559766   0.162037   3.455  0.00129 **
minor:fside1  0.006471   0.003512   1.843  0.07259 .
insur:fside1 -0.127027   0.162037  -0.784  0.43758
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4626 on 41 degrees of freedom
Multiple R-squared:  0.6774,    Adjusted R-squared:  0.6381
F-statistic: 17.22 on 5 and 41 DF,  p-value: 3.722e-09
```

Using the output above answer the following questions.

(i) Describe the effect of the percentage of minority (`minor`) on **the number of fires**.

(ii) Describe the effect of the locality of the district (`fside`) on **the number of fires**.

(iii) If possible give the sample size.

(iv) Let $\mathbb{X}$ be the regression matrix. If possible calculate what is the last diagonal element of the matrix $(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$.

Task

We would like to model the logarithm of the number of fires (`fire`) in some town district given the percentage of a minority in this district (the locality where the district is situated is the factor variable `fside` with two levels - `North` and `South`) and the number of insurance policies (variable `insur`). Note that a **logarithmic** transformation of the number of fires is used as a response and a **contrast sum parametrization** (`contr.sum`) with the following contrast-matrix

```
      fside
North    1
South   -1
```

is used for the variable `fside`.

Based on the observed data we estimated the following model

```
log(fire) ~ minor + fside + insur + minor:fside + insur:fside
```

and we obtained the summary output below:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.579046   0.111366  14.179  < 2e-16 ***
minor         0.010833   0.003512   3.085  0.00364 **
fside1       -0.034162   0.111366  -0.307  0.76058
insur         0.559766   0.162037   3.455  0.00129 **
minor:fside1  0.006471   0.003512   1.843  0.07259 .
insur:fside1 -0.127027   0.162037  -0.784  0.43758
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4626 on 41 degrees of freedom
Multiple R-squared:  0.6774,    Adjusted R-squared:  0.6381
F-statistic: 17.22 on 5 and 41 DF,  p-value: 3.722e-09
```

Using the output above answer the following questions.

(i) What is the average effect (an average over both possible district locations) of the logarithm of the percentage of minority on **the logarithm of the number of fires**?

(ii) Describe the effect of the percentage of minority (`minor`) on **the number of fires**.

(iii) Is the district location (`fside`) statistically significant modifier of the effect of the number of insurance policies (`insur`) on the **number of fires**?

(iv) Interpret the estimate of the intercept.
If possible, interpret this estimate with respect to ***the number of fires****. If this is not possible, interpret this number with respect to* ***the logarithm of the number of fires****.*

Task

We want to investigate a relationship between the occipital angle (`oca`) and two covariates: gender (`fgender`) and population (`fpopul`) to which the subject belongs. The gender covariate has two levels labelled as `female` and `male`. The population covariate has three levels labelled as `AUSTR`, `BERG`, and `BURIAT`. For both factor covariate the parametrization used in the model is the one based on a reference category (`contr.treatment`) where the reference category for the gender covariate are females (`female`) and the reference category for the population is Australia (`AUSTR`). In addition a logarithmic transformation of the occipital angle was considered. Given the available data the following linear regression model was fitted:

```
log(oca) ~ fpopul * fgender
```

and the following model summary table was obtained:

```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.741      0.006 759.636   <0.001
fpopulBERG               0.019      0.009     ?      0.026
fpopulBURIAT             0.020      0.009     ?      0.018
fgenderM                -0.006      0.011  -0.557    0.578
fpopulBERG:fgenderM      0.008      0.014   0.569    0.570
fpopulBURIAT:fgenderM   -0.022      0.014  -1.557    0.121

Residual standard error: 0.04369 on 283 degrees of freedom
Multiple R-squared:  0.07723,   Adjusted R-squared:  0.06092
F-statistic: 4.737 on 5 and 283 DF,  p-value: 0.0003583 .
```

Using the available information above answer the questions below.

(i) What is the number of all observations available in the dataset?

(ii) Compare the expected occipital angles for males and females from population `BERG`.

(iii) Using the information in the table above can we conclude that the difference in the occipital angle of females from populations `BERG` and `BURIAT` is significantly different? If yes, provide the corresponding p-value and the test statistic value. If not, explain why.

(iv) Let u_i $(i = 1, \ldots, n)$ be the residuum of Y_i ($u_i = Y_i - \hat{Y}_i$, where Y_i is the logarithm of the occipital angle of the i-th observation). Using the model summary output above express the value of $\sum_{i=1}^{n} u_i$ (the sum of all residuals).

Task

We are interested how salary of associate professors (`salary.assoc`) depends on the number of professors (`n.prof`), number of associate professors (`n.assoc`) and type of the school (`type`). To prevent heteroscedasticity the logarithmic transformation of the salary of associate professors (`lsalary.assoc`) was considered, that is $\texttt{lsalary.assoc} = \log\big(\texttt{salary.assoc}\big)$. Further, the variable `type` has 3 levels labelled subsequently as `I`, `IIA`, `IIB` and for this variable `contr.sum` parametrization with the following contrast-matrix

```
    type1 type2
I       1     0
IIA     0     1
IIB    -1    -1
```

was used.

Based on the observed data we estimated the following model

```
lsalary.assoc ~ type * n.prof + n.assoc
```

and we get the following output

```
Coefficients:
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)   10.61     0.0089    1192.1   < 0.001
type1         0.1361    0.0135      10.1   < 0.001
type2         0.0340    0.0010       3.4   < 0.001
n.prof        0.0013    0.0001      13.0   < 0.001
n.assoc       0.0001    0.0001         ?     0.318
type1:n.prof -0.0013    0.0001     -13.0   < 0.001
type2:n.prof -0.0010    0.0001     -10.0   < 0.001

---

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) 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.

(ii) 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`).

(iii) 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`.

(iv) Explain in detail how the p-value on the line starting with `n.assoc` is calculated?

(v) What is the total number of observations?

Task

We would like to investigate an effect of a disturbing stimulus (`fstimulus`) and the age of the horse given in years (`age`) on the heart rate of the horse (`HR`). The stimulus covariate is considered to be a factor with four levels labelled as *none*, *quiet*, *medium*, and *loud*) which reflect an increasing disturbance intensity. For the heart rate we additionally considered a logarithmic transformation (`lHR` = *log* `HR`) and the factor covariate was implemented using a sum contrast matrix (`contr.sum`). Using the given data the following model was fitted

```
lHR ~ fstimulus * age
```

and the following parameter estimates were obtained:

```
Coefficients:
   (Intercept)      fstimulus1      fstimulus2      fstimulus3
      4.303982       -0.402377       -0.056883       -0.004995

       age        fstimulus1:age  fstimulus2:age  fstimulus3:age
    -0.019772         0.001886       -0.009977        0.009743
```

Let us recall, that the contrast matrix for `contr.sum` parametrization takes the form:

```
        fstimulus1 fstimulus2 fstimulus3
none         1          0          0
quiet        0          1          0
medium       0          0          1
loud        -1         -1         -1
```

Using the output above answer all questions below:

(i) What is the expected logarithm of the horse heart rate `lHR` if we apply the most disturbing stimulus (`loud`) and we know that the age of the horse (`age`) is 10 years?

(ii) Specify the effect of the horse age (`age`) on the heart rate of the horse (`HR`) when applying no stressful stimulus (`none`).

(iii) Provide a vector for a linear combination of regression parameters which will be an estimable parameter estimating the difference between the effects of the age covariate (`age`) on the horse heart rate (`HR`) given two different stimuli: `quiet` and `loud`.

Task

We want to model the salary of associate professors (`salary.assoc`) given the number of professors (`n.prof`), the number of associate professors (`n.assoc`), and the university type (`type`). The university type is a factor covariate with three levels labelled as *I*, *IIA*, and *IIB*. For a better interpretation purposes a transformed regressors are used where the number of professors (`n.prof`) and the number of associate professors (`n.assoc`) are both lowered by 40 (covariates `n.prof40`, `n.assoc40`, where `n.prof40` = `n.prof` − 40 and `n.assoc40` = `n.assoc` − 40).
Beside that, there is a logarithmic transformation of the associate professors salary considered (`lsalary.assoc` = $\log($`salary.assoc`$)$) and for the factor covariate (`type`) there was a reference group parametrization used (`contr.treatment`) for the reference category set to type *I*. The following linear regression model was fitted:

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

Given the available data the following model summary was obtained (the values are all rounded to three decimal places and values 0.000 in the table below refer to quantities smaller than 0.0005).

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.118      0.028  219.40   <0.001
typeIIA             -0.077      0.029   -2.67    0.008
typeIIB             -0.121      0.029   -4.18   <0.001
n.prof40             0.005      0.001    4.40   <0.001
n.assoc40            0.000      0.000    0.28    0.776
typeIIA:n.prof40    -0.001      0.001   -1.84    0.066
typeIIB:n.prof40     0.002      0.000    4.79   <0.001
typeIIA:n.assoc40    0.001      0.000    3.25    0.001
typeIIB:n.assoc40    0.003      0.000    6.52   <0.001
n.prof40:n.assoc40   0.001      0.000   -2.46    0.014
```

```
Residual standard error: 0.1255 on 1115 degrees of freedom
Multiple R-squared:  0.4767,  Adjusted R-squared:  0.4724
F-statistic: 112.8 on 9 and 1115 DF,  p-value: < 2.2e-16
```

Using the available information about the fitted model answer the following questions.

(i) Assume, that the first four observations in the dataset are

Observation no. 1: school of type I with 40 professors and 50 associate professors;

Observation no. 2: school of type IIA with 50 professors and 50 associate professors;

Observation no. 3: school of type IIB with 50 professors and 40 associate professors;

Observation no. 4: school of type IIB with 50 professors and 50 associate professors;

Define the first four rows of the model matrix.

(ii) What is the expected difference in associate professor salaries between two universities of type IIA and IIB if the number of professors and the number of associate professors are both equal to 40?

(iii) Considering the summary table above and preserving a well-defined hierarchical structure of the model, can exclude the number of associate professors lowered by 40 (`n.assoc40`) form the model? If yes, provide the corresponding p-value of the test and the test statistic value. If no, explain why.

Task

We are interested how the associate professor's salary depends on the university type (`type`) and the **number of professors and associate professors, both lowered by 40** which is roughly the sample median value for both (`n.prof40` and `n.assoc40`). There are three university types considered (I, IIA and IIB) and **the contrast sum parametrization** (`contr.sum`) with the following parametrization matrix

```
    type1 type2
  I     1     0
IIA     0     1
IIB    -1    -1
```

was used. Based on the observed data we estimated the following model

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

and we get the following output

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    433.37029    2.91432 148.703  < 2e-16
type1           37.98301    5.12269   7.415  2.4e-13
type2           -3.61827    3.40707  -1.062    0.288
```

```
n.prof40          0.52629    0.04127  12.751  < 2e-16
n.assoc40         0.02167    0.04332       ?        @
type1:n.prof40   -0.48774    0.03816       ?        @
type2:n.prof40   -0.36960    0.04192  -8.816  < 2e-16
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.24 on 1118 degrees of freedom
Multiple R-squared:  0.4282,  Adjusted R-squared:  0.4251
F-statistic: 139.5 on 6 and 1118 DF,  p-value: < 2.2e-16
```

(i) Interpret the intercept parameter, value 433.37 in the output above.

(ii) Describe the estimated effect of the university type on the expected salary.

(iii) What is the expected salary of the associate professor at the university type IIB if there are 50 professors and 40 associate professors?

(iv) Compare the estimated expected salaries of associate professors for universities of type I and IIA if there are 40 professors and 50 associate professors. If possible, provide the p-value of the test whether this difference is significant, and if not, state some limits in which the p-value should be obtained.

(v) Instead of the question marks in the output above fill in the appropriate values.

Task

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

We want to model the associate professors salary (`salary.assoc`) given the number of professors (`n.prof`), the number of associate professors (`n.assoc`) and the country (`state`) where the university is located. The country covariate is a factor with three levels labelled as `CA` (California), `NY` (New York) and `TX` (Texas). For the parametrization of the factor `state` the `contr.sum()` parametrization was used with the corresponding contrast matrix

```
    state1 state2
CA     1      0
NY     0      1
TX    -1     -1
```

For better interpretation purposes the number of professors and the number of associate professors were both lowered by 40 (new covariates `n.prof40` and `n.assoc40`, where `n.prof40` = `n.prof` − 40 and `n.assoc40` = `n.assoc` − 40). The following linear regression model was considered

```
salary.assoc ~ state + n.prof40 + n.assoc40 + state:n.assoc40
```

and using the available data the following model summary output was obtained:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      432.877      4.836  89.504    0.000
state1            28.459      7.381   3.856    0.000
```

```
state2              2.103      6.306   0.333    0.739
n.prof40            0.057      0.054   1.053    0.294
n.assoc40           0.254      0.113   2.256    0.025
state1:n.assoc40   -0.103      0.087  -1.194    0.234
state2:n.assoc40    0.182      0.075   2.438    0.016

Residual standard error: 55.78 on 180 degrees of freedom
Multiple R-squared:  0.3497,    Adjusted R-squared:  0.328
F-statistic: 16.13 on 6 and 180 DF,  p-value: 8.077e-15
```

Using the available information answer the questions below.

(i) Interpret the intercept parameter estimate (number `432.877`).

(ii) Interpret the estimated regression coefficient corresponding to `n.assoc40` (number `0.254`).

(iii) What is the expected difference in the associate professor salaries at two universities in California (`CA`) and Texas (`TX`) if both have 40 professors a 40 associate professors?

(iv) What is the expected difference in associate professor salaries at two universities in California and (`CA`) and New York (`NY`) if both universities have 45 professors and 45 associate professors?

Task

We would like to investigate an effect of a disturbing stimulus (`fstimulus`) and the age of the horse given in years (`age`) on the heart rate of the horse (`HR`). The stimulus covariate is considered to be a factor with four levels labelled as *none*, *quiet*, *medium*, and *loud*) which reflect an increasing disturbance intensity. For the heart rate we additionally considered a logarithmic transformation (`lHR` = *log* `HR`) and the factor covariate was implemented using a sum contrast matrix (`contr.sum`). Using the given data the following model was fitted

```
lHR ~ fstimulus * age
```

and the following parameter estimates were obtained:

```
Coefficients:
   (Intercept)      fstimulus1      fstimulus2      fstimulus3
      4.303982       -0.402377       -0.056883       -0.004995

       age        fstimulus1:age  fstimulus2:age  fstimulus3:age
    -0.019772          0.001886       -0.009977        0.009743
```

Let us recall, that the contrast matrix for `contr.sum` parametrization takes the form:

```
        fstimulus1 fstimulus2 fstimulus3
none         1          0          0
```

```
quiet       0          1           0
medium      0          0           1
loud       -1         -1          -1
```

Using the output above answer all questions below:

(i) What is the expected logarithm of the horse heart rate `lHR` if we apply the most disturbing stimulus (`loud`) and we know that the age of the horse (`age`) is 10 years?

(ii) Specify the effect of the horse age (`age`) on the heart rate of the horse (`HR`) when applying no stressful stimulus (`none`).

(iii) Provide a vector for a linear combination of regression parameters which will be an estimable parameter estimating the difference between the effects of the age covariate (`age`) on the horse heart rate (`HR`) given two different stimuli: `quiet` and `loud`.

Task

We want to model the associate professors salary (`salary.assoc`) given the number of professors (`n.prof`), the number of associate professors (`n.assoc`) and the country (`state`) where the university is located. The country covariate is a factor with five levels labelled as `CA`, `NY`, `OH`, `PA`, and `TX`. For the parametrization there was `contr.sum`) parametrization used with the corresponding contrast matrix

```
    state1 state2 state3 state4
CA     1      0      0      0
NY     0      1      0      0
OH     0      0      1      0
PA     0      0      0      1
TX    -1     -1     -1     -1 .
```

For better interpretation purposes the number of professors and the number of associate professors were both lowered by 40 (new covariates `n.prof40` and `n.assoc40`, where `n.prof40` = `n.prof` − 40 and `n.assoc40` = `n.assoc` − 40). The following linear regression model was fitted

```
salary.assoc ~ state + n.prof40 + n.assoc40 + state:n.prof40,
```

and using the available data the following model summary output was obtained:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      426.22       3.577 119.144   <0.001
state1            36.00       8.537   4.222   <0.001
state2             8.60       6.406   1.350    0.178
state3           -33.40       7.306  -4.568    0.000
state4            11.90       6.054   1.969    0.050
n.prof40           0.07       0.062   1.188    0.236
n.assoc40          0.29       0.098   2.951    0.003
state1:n.prof40   -0.06       0.046  -1.234    0.218
```

```
state2:n.prof40    0.09       0.047   1.982    0.048
state3:n.prof40   -0.06       0.058  -0.967    0.334
state4:n.prof40    0.11       0.047   2.383    0.018

Residual standard error: 55.67 on 312 degrees of freedom
Multiple R-squared:  0.3914,    Adjusted R-squared:  0.3719
F-statistic: 20.07 on 10 and 312 DF,  p-value: < 2.2e-16 .
```

Using the available information answer the questions below.

(i) Interpret the intercept parameter estimate (number `426.22`).

(ii) What is the expected difference in the associate professor salaries at two universities in California (`CA`) and Texas (`TX`) if both have 40 professors a 40 associate professors?

(iii) What is the expected difference in associate professor salaries at two universities in California and (`CA`) a New York (`NY`) if both universities have 50 professors and 50 associate professors?

Task

We are interested in how the amount of yield (`yield`) depends on magnesium (`Mg`) and nitrogen (`N`). For this reason we introduce the following variables: `lMg` which is the logarithm of Mg (i.e. `lMg = log(Mg)`), `lN` - logarithm of N (i.e. `lN = log(N)`), and `lMg2` = $\big[$`log(Mg)`$\big]^2$, `lN2` = $\big[$`log(N)`$\big]^2$. Based on the available observations we estimated the following model

```
yield ~ lN + lN2 + lMg + lMg2 + lN:lMg.
```

and we got the following output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.4    61.22       0.448   0.655
lMg          18.7    12.42       1.509   0.132
lN          -11.5    15.11      -0.764   0.446
lN2           0.6     1.04       0.597   0.551
lMg2         -2.0     0.78      -2.505   0.013
lMg:lN       -0.3     1.73      -0.194   0.846

Residual standard error: 1.015 on 362 degrees of freedom
Multiple R-squared:  0.1525,    Adjusted R-squared:  0.1408
F-statistic: 13.03 on 5 and 362 DF,  p-value: 1.131e-11
```

Based on the output above answer the following questions.

(i) Describe the estimated effect of the logarithm of magnesium (`lMg`) on the yield.

(ii) Is magnesium a statistically significant modifier of the effect of the nitrogen on the yield?
(Do not forget to explain your answer).

(iii) If possible, give the p-value of the test of the submodel

```
yield ~ lN + lN2 + lMg + lN:lMg.
```

against the model

```
yield ~ lN + lN2 + lMg + lMg2 + lN:lMg.
```

(iv) Let Y_i (for $i = 1, \ldots, n$) be the value of the yield for the i-th observation 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

We want to model the amount of yield (`yield`) given the concentration of magnesium (Mg) and calcium (Ca). The following model was considered:

```
yield ~ Mg + Mg^2 + Ca ,
```

and using the available data the model summary output was obtained:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.10       1.169   0.082    0.935
Mg             0.80       0.189   4.121   <0.001
Mg^2          -0.02       0.007  -3.358    0.001
Ca            -0.02       0.005  -3.467    0.001 .
```

For better interpretation purposes a new model was refitted where the magnesium concentration was lowered by 1 (`Mg1` = `Mg` − 1) and the calcium concentration was lowered by 5 (`Ca5` = `Ca` − 5). The same model structure was fitted again

```
yield ~ Mg1 + (Mg1)^2 + Ca5.
```

(i) Briefly describe the difference in the interpretation of the intercept parameter in the first and the second model.

(ii) If possible, use the values in the first summary output to fill in missing values in the second model summary output below (2 points).

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    ?          ?        ?        ?
Mg1            ?          ?        ?        ?
(Mg1)^2        ?          ?        ?        ?
Ca5            ?          ?        ?        ?
```

Task

We want to model the amount of yield (`yield`) given the concentration of magnesium (Mg) and calcium (Ca). The following model was considered:

```
yield ~ Mg + Mg^2 + Ca
```

and using the available data the model summary output was obtained:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.10       1.169   0.082    0.935
Mg             0.80       0.189   4.121   <0.001
Mg^2          -0.02       0.007  -3.358    0.001
Ca            -0.02       0.005  -3.467    0.001 .
```

For better interpretation purposes a new model was refitted where the magnesium concentration was lowered by 1 (`Mg1` = `Mg` − 1) and the calcium concentration was lowered by 5 (`Ca5` = `Ca` − 5). The same model structure was fitted again

```
yield ~ Mg1 + (Mg1)^2 + Ca5.
```

(i) Briefly describe the difference in the interpretation of the intercept parameter in the first and the second model.

(ii) If possible, use the values in the first summary output to fill in missing values in the second model summary output below (2 points).

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     ?          ?        ?        ?
Mg1             ?          ?        ?        ?
(Mg1)^2         ?          ?        ?        ?
Ca5             ?          ?        ?        ?
```

Task

We are interested how the salary of associate professors (`salary.assoc`) depends on the number of professors (`n.prof`) and the number of associate professors (`n.assoc`).

Based on the observed data we estimated the following model

```
salary.assoc ~ sqrt(n.prof) + n.prof + sqrt(n.assoc) + n.assoc
               + sqrt(n.prof):sqrt(n.assoc)
```

where `sqrt(n.prof)` and `sqrt(n.assoc)` stand for the square roots (odmocnina) of the number of professors (`n.prof`) and the number of associate professors (`n.assoc`) respectively.

We got the following model summary output

```
Coefficients:
                        Estimate Std. Error t value  Pr(>|t|)
(Intercept)             277.7024     6.4634  42.966  < 0.0001
sqrt(n.prof)              9.9423     1.7929   5.545  < 0.0001
n.prof                    0.4410     0.1477       x    0.0029
sqrt(n.assoc)            15.3493     2.6522   5.787  < 0.0001
n.assoc                   0.3863     0.3254   1.187    0.2355
```

```
sqrt(n.prof):sqrt(n.assoc)  -1.5750     0.4121  -3.822    0.0001
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.75 on 1119 degrees of freedom
Multiple R-squared:  0.4589,    Adjusted R-squared:  0.4565
F-statistic: 189.8 on 5 and 1119 DF,  p-value: < 2.2e-16
```

Based on the output above answer the following questions.

(i) Describe the estimated effect of the number of professors on the salary of associate professors.

(ii) Is the number of associate professors a statistically significant modifier of the effect of the number of professors on the salary of associate professors?
Do not forget to explain your answer.

(iii) If possible, give the p-value of the test of the submodel

```
salary.assoc ~ sqrt(n.prof) + n.prof + sqrt(n.assoc)
               + sqrt(n.prof):sqrt(n.assoc)
```

against the model

```
salary.assoc ~ sqrt(n.prof) + n.prof + sqrt(n.assoc) + n.assoc
               + sqrt(n.prof):sqrt(n.assoc)
```

(iv) Explain how the p-value on the line starting with `n.assoc` is calculated.
The answer should be in the form that $0.0029 = f(x)$ *and you explain what is the function* f *and how the number* x *is numerically calculated from the output above.*

Task

We want to model the amount of yield (`yield`) given the concentration of magnesium (`Mg`) and nitrogen (`N`). The following model was considered:

```
yield ~ N + Mg + Mg^2
```

and using the available data the model summary output was obtained:

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)    0.400      1.135    0.35     0.725
N             -0.006      0.001   -5.90   < 0.001
Mg             0.996      0.189    5.26   < 0.001
I(Mg^2)       -0.031      0.007   -4.33   < 0.001
```

For better interpretation purposes a new model was refitted where the magnesium concentration was lowered by 10 (`Mg10` = `Mg` − 10) and the nitrogen concentration was lowered by 500 (`N500` = `N` − 500). The same model structure was fitted again

```
yield ~ N500 + Mg10 + (Mg10)^2
```

(i) Describe the difference in the interpretation of the intercept parameter in the first and the second model.

(ii) If possible, use the values in the first summary output to fill in missing values in the second model summary output below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	?	?	?	?
N500	?	?	?	?
Mg10	?	?	?	?
(Mg10)^2	?	?	?	?