Linear Regression (NMSA407)

Test

Version – Sample test

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 1 (40 points)

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. [4]

(ii) Describe the effect of the height on the percentage of body fat. [4]

(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? [8]

(iv) Find a prediction interval for the percentage of body fat of an applicant whose height is 180 cm. [6]

(v) Where possible, complete the following ANOVA table of type III. [8]

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

(vi) Where possible, complete the following ANOVA table of type III for model without the shift in `height`, i.e. for model given by

```
lm(fat ~ height + I(height^2)+ I(height<170), data = Policie)|.
```

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

[10]

Task 2 (24 points)

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`).
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:flNmedian  -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`). [4]

(ii) Explain in detail how the test statistic and the p-value in the row that starts with `flNmedian` is computed. What is being tested there, and what is the conclusion of that test? [4]

(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. [4]

(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. [4]

(v) If possible, plug in the values for Multiple R-squared and Adjusted R-squared. [4]

(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? [4]

Task 3 (36 points)

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 intercept parameter. [4]

(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). [6]

(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. [6]

(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. [4]

(v) Where possible, complete the following ANOVA table of type III. [8]

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

(vi) Which lines of the ANOVA table of type III above will change, if we consider: [8]

(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`)?