Linear Regression (NMSA407)

Test

Version – V1 | 19.12.2017

- 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 estimate the mean salary of associate professors across different universities in the United States (`assoc.salary`). For this purpose we obtain the number of professors (`n.prof`), the number of associate professors (`n.assoc`), and the number of assistant professors (`n.assist`) at each university. In addition, we also recognize for three university types (`type`): I, IIA, and IIB university type. Moreover, for some better interpretation purposes the number of professors, number of associate professors, and the number of assistent professors were all lowered by 40 (roughly the median value for each). A set of correspoding new covariates was introduced in addition (`n.prof40`, `n.assoc40`, and `n.assist40`).

Using the available data (1125 independent observations) the following model is fitted

```
salary.assoc ~ type * (n.prof40 + n.assoc40 + n.assist40)
```

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

```
                  Df  Sum Sq Mean Sq  F value    Pr(>F)
type               ? 1825984       ?        ? < 2.2e-16 ***
n.prof40           ?  129544       ?        ? 8.324e-12 ***
n.assoc40          ?    1775       ?        ?  0.419011
n.assist40         ?    3563       ?        ?  0.252246
type:n.prof40      ?  513550       ?        ? < 2.2e-16 ***
type:n.assoc40     ?  223511       ?        ? < 2.2e-16 ***
type:n.assist40    ?   32569       ?        ?  0.002568 **
Residuals          ? 3022526       ?
```

(i) If possible, replace the question marks in the output above with the missing values. [6]

(ii) Explain, how the p-values in the last column of the output above are calculated. [4]

(iii) Is it possible to simplify the model above such that only non-significant parameters will be deleted and the resulting model will be hierarchically well formulated? Explain on which p-values from the output above can you base your conclusion on. [4]

(iv) If possible, calculate the quantities $\sum_{i=1}^{1125}(Y_i - \overline{Y})^2$ and $\sum_{i=1}^{1125}(Y_i - \widehat{Y})^2$. [2]

Task 2 (24 points)

We would like to predict the proportion of unregistered bike rental users (covariate `ratio`) given the outside temperature (`temp`) and two factor covariates – four level covarate `season` (coded as 1 for winter, 2 for spring, 3 for summer, and 4 for autumn) and two level covariate `holiday` (value one for holiday, value zero otherwise). A **standard contrast parametrization** was used (`contr.treatment`) and due to some heteroskedasticity in the data the **logarithmic transformation of the response** was used (`lratio` = $\log($`ratio`$)$).
Based on the observed data the following model was fitted

```
lratio ~ temp * season + holiday
```

and the following summary output was obtained:

```
              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   1.837397   0.064561  28.460  < 2e-16 ***
temp          0.062647   0.008359   7.495 1.95e-13 ***
season2       0.886250   0.143910   6.158 1.22e-09 ***
season3       0.693418   0.308871   2.245  0.02507 *
season4       0.213181   0.121877   1.749  0.08069 .
holiday1      0.645160   0.120199   5.367 1.08e-07 ***
temp:season2 -0.050666   0.010880  -4.657 3.82e-06 ***
temp:season3 -0.045549   0.014534  -3.134  0.00179 **
temp:season4 -0.026841   0.011585  -2.317  0.02079 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5411 on 722 degrees of freedom
Multiple R-squared:  0.3264,Adjusted R-squared:  0.3189
F-statistic: 43.72 on 8 and 722 DF,  p-value: < 2.2e-16
```

(i) Interpret the intercept parameter estimate, value 1.8374 in the output above. [4]

(ii) Describe the estimated effect of the temperature (`temp`) on the proportion of unregistered bike rental users (`ratio`) in summer. Is this effect significantly different from zero (provide the corresponding p-value if possible)? [5]

(iii) Provide an estimate for the expected proportion of the ratio of unregistered users (`ratio`) in winter season, if the outside temperature (`temp`) is 10 degrees (of Celsius) and there is no holiday that day. [5]

(iv) Compare the expected proportion of unregistered bike users (`ratio`) for a winter day and some summer day if there is a holiday (`holiday` = 1) that day and the outside temperature is 20°C (`temp` = 20). Is this difference statistically significant? Provide the corresponding p-value if possible. [6]

(v) Explain in detail how the p-value on the line starting with `holiday1` is calculated? [4]

Task 3 (24 points)

We would like to investigate whether some specific locality in Morava region and a high school student's English language skills play some important role for the overall performance of this student. The overall performance is measured as an average mark on the student's certificate (covariate `avg.mark`) and the English language skills are assesed using the average grade from the English classes (`mark.eng`). In addition, we distinguish four specific localities in Morava region (covariate `fregion`) which are ordered as *Brno*, *Olomouc*, *Zlín*, and *Ostrava*.

For the interpretation purposes the English grade was lowered by one (new covariate `mark.eng1`) and for the factor covariate a **contrast sum parametrization** (`contr.sum`) was used.

Using the available data the following model was fitted

```
avg.mark ~ fregion + mark.eng1 + fregion:mark.eng1
```

and the corresponding summary output was obtained:

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.486804   0.010930 136.028  < 2e-16 ***
fregion1            0.001228   0.015208   0.081 0.935649
fregion2           -0.067341   0.024651  -2.732 0.006358 **
fregion3            0.043312   0.020029   2.162 0.030708 *
mark.eng1           0.139692   0.007410  18.851  < 2e-16 ***
fregion1:mark.eng1  0.002125   0.010259   0.207 0.835905
fregion2:mark.eng1  0.055852   0.016079   3.474 0.000525 ***
fregion3:mark.eng1 -0.039290   0.014260  -2.755 0.005920 **
```

Recall, that the corresponding contrast matrix takes the form

```
          fregion1 fregion2 fregion3
Brno             1        0        0
Olomouc          0        1        0
Zlín             0        0        1
Ostrava         -1       -1       -1
```

Using the information provided above, answer the questions below.

(i) Interpret the estimated intercept parameter, value 1.48 in the output above. [4]

(ii) What is the student's expected average grade (`avg.mark`) if the student comes from Ostrava region and his/her English grade (`mark.eng1`) is equal to 2? [4]

(iii) What are the corresponding effects of the English grade (`mark.eng1`) on the overall average grade in Brno and Ostrava? Is the difference between these two effects statistically significant (provide the corresponding p-value if possible)? [6]

(iv) Define a vector of coefficients for a linear combination of the parameter estimates such that you will obtain an estimable parameter specifying a difference between the effect of the English grade (`mark.eng1`) on the the overall average grade (`avg.mark`) in Ostrava region versus Brno region. [6]

(v) Consider two high school students where both of them have the English grade one on their certificate, however, the first one is from Brno and the other from Olomouc. Can we say that their expected overall average grades are significantly different when using the critical level $\alpha = 0.05$? Provide the corresponding p-value if possible. [4]

Task 4 (36 points)

Suppose that you observe a random sample (independent and identically distributed random vectors) $(Z_1, Y_1)^\mathsf{T}, \ldots, (Z_n, Y_n)^\mathsf{T}$, where the conditional distribution of $Y_i|Z_i$ is assumed to be normal $\mathsf{N}(\zeta Z_i, \sigma^2 Z_i)$ and $P(Z_i > 0) = 1$. Let us also assume, that the distribution of the random variables Z_i, for $i = 1, \ldots, n$, does not depend on $\boldsymbol{\theta} = (\zeta, \sigma^2)^\top \in \mathbb{R} \times (0, \infty)$.

(i) Find the maximum likelihood estimator for the vector of unknown vector of parameters $\boldsymbol{\theta} = (\zeta, \sigma^2)^\top$.

(ii) For $\alpha = 0.05$ find the corresponding confidence region for the unknown parameter $\sigma^2 > 0$.

(iii) Derive a test (Wald test, likelihood ratio test and Score test) of the null hypothesis $H_0 : \zeta = 0$.

The calculations you provide for this task should form a well formulated mathematical text with a properly defined notation all supplied with some necessary explanation, if needed.