University of Padova

Master Degree in *Computer Science*

a.y. 2017/2018

Data Mining

Teacher: Annamaria Guolo

Written assessment: June, 20, 2018

INSTRUCTIONS: The examination takes 1 hour. You are asked to reply using these papers. In case you need other papers, you can use them but they will not be corrected. Do not use pencil. Do not use corrector tape.

Name:_______________ Surname:_______________ Enrolment number:_______________

Questions with multiple choice.

Only one response is the correct one. Mark the right response. Wrong or missing replies take 0 points.

1) The hypotheses on the errors in the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ include

(a) normality (b) mean equal to one (c) increasing variance
(d) correlation with X

2) In a linear regression model, one factor X with 3 levels gives rise to a number of dummy variables (indicators) equal to

(a) 3 (b) 4 (c) 2 (d) it depends on the response Y

3) In a linear regression model, which one of the following values for a p-value indicates the lack of association between response Y and covariate X?

(a) 0.05 (b) 0.83 (c) 0.003 (d) it depends on the sample size

4) Does the term spurious correlation refer to a nonsensical relationship between two variables?

(a) no (b) yes (c) it depends on the sample size
(d) it depends on the variances of the variables

5) The residual deviance

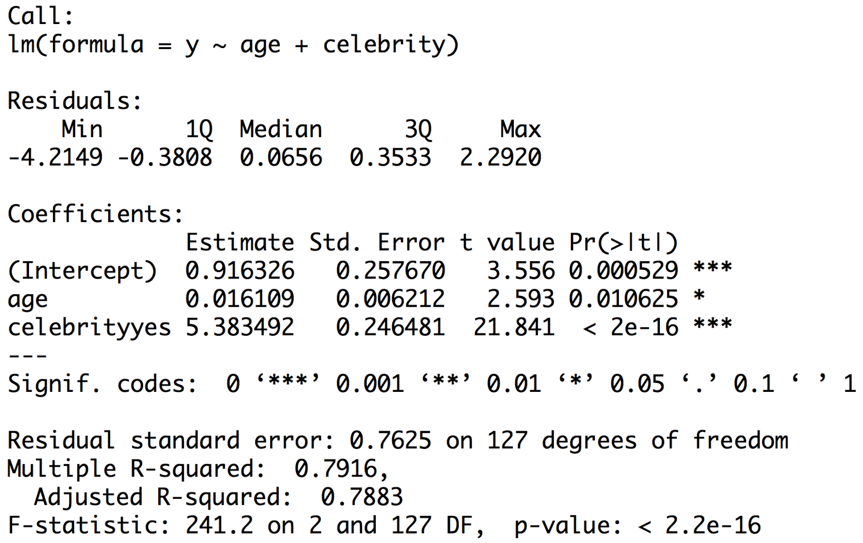
(a) increases as R^2 increases (b) increases as the total deviance increases
(c) always increases (d) none of the previous choices

Exercise.

Consider the data about the annual income of 130 subjects. Data include the following information:

- `earnings`: the natural logarithm of the annual personal earnings (originally measured in 10,000$)
- `age`: age (years)
- `gender`: female/male
- `celebrity`: is the subject a celebrity? yes/no

a) We estimate a linear regression model to explain the relationship between the logarithm of annual earnings and the age and the gender. This is the output from `R`

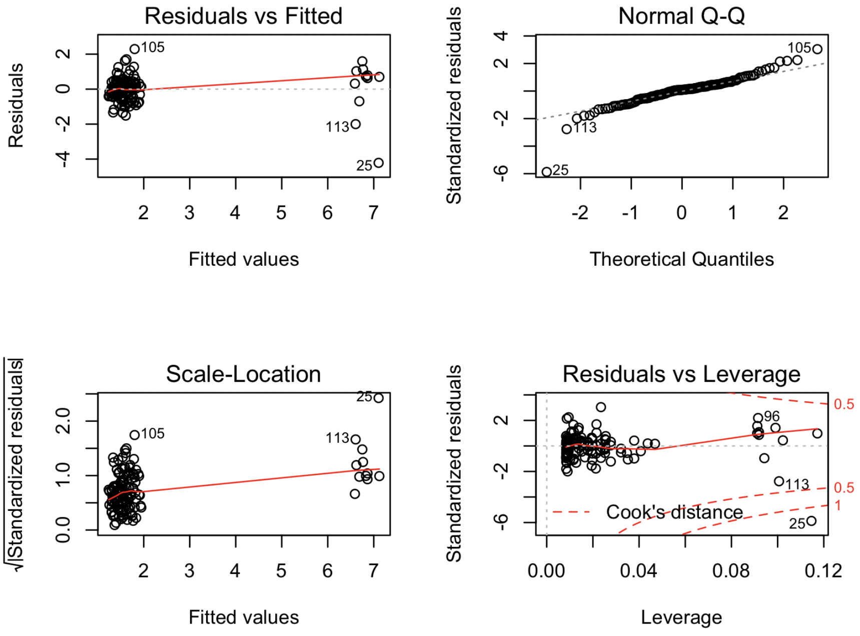


```
Call:
lm(formula = y ~ age + celebrity)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2149 -0.3808  0.0656  0.3533  2.2920 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.916326   0.257670   3.556 0.000529 ***
age          0.016109   0.006212   2.593 0.010625 *  
celebrityyes 5.383492   0.246481  21.841  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7625 on 127 degrees of freedom
Multiple R-squared:  0.7916,
  Adjusted R-squared:  0.7883 
F-statistic: 241.2 on 2 and 127 DF,  p-value: < 2.2e-16
```

a.1) Write the expression of the estimated model. Describe how `R` handles the qualitative variable `discipline` and which level is the baseline level.

a.2) Discuss the output of the model paying attention to i) the significance of the coefficients, ii) the possibility to simplify the model, iii) the accuracy of the model using R^2.

a.3) The following plot represents the residuals analysis of the fitted model. Comment on the plot and discuss whether the model is accurate, or whether the residuals suggest any modification of the model, or explaining whether there is indication of additional analyses.



a.4) Explain what the quantity *Multiple R-squared* in the output represents and how it is computed.

b) The extension of the model including variable `gender` provides the following output

```
Call:
lm(formula = y ~ age + gender + celebrity)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0746 -0.2930  0.0539  0.3309  2.1673 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.789541   0.260858   3.027   0.0030 ** 
age          0.015519   0.006132   2.531   0.0126 *  
gendermale   0.283988   0.132281   2.147   0.0337 *  
celebrityyes 5.399540   0.243167  22.205   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7519 on 126 degrees of freedom
Multiple R-squared:  0.799,
  Adjusted R-squared:  0.7942 
F-statistic: 166.9 on 3 and 126 DF,  p-value: < 2.2e-16
```

b.1) Does it make sense to maintain the interaction in the model? Can we simplify the model? Why?

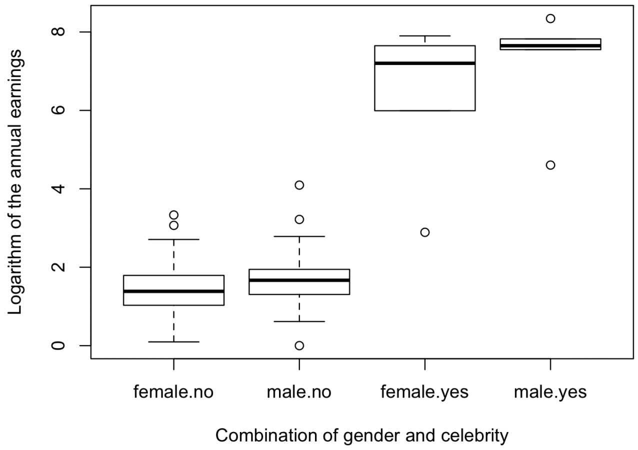
b.2) The comparison of the two models using function `anova()` provides the following output. Comment the output, explaining what it represents and what we can conclude from it.

```
Analysis of Variance Table

Model 1: y ~ age + celebrity
Model 2: y ~ age + gender + celebrity
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)  
1    127 73.840                             
2    126 71.234  1    2.6057 4.609 0.03372 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

b.3) Using the best model, provide a confidence interval at level 0.95 for the coefficient associated to `age`. Explain assumptions, if any.

b.4) Using the last model, predict the earnings (on the original scale) for a 30-years old famous male and then the earnings (on the original scale) for a non-famous 30-years old male.

c) The following plot shows the distribution of earnings for different levels of `gender` and `celebrity`



c.1) Does the plot suggest to add an interaction between the two covariates in the model? Why?

Useful information

Quantiles of a standard Normal distribution

$$z_{0.01} = -2.33 \;\; z_{0.025} = -1.96 \;\; z_{0.05} = -1.64 \;\; z_{0.95} = 1.64 \;\; z_{0.975} = 1.96 \;\; z_{0.99} = 2.33$$

Quantiles of F distribution

$$F_{0.025;1,127} = 0.00099 \;\; F_{0.05;1,127} = 0.0039 \;\; F_{0.975;1,127} = 5.146 \;\; F_{0.95;1,127} = 3.9163$$