University of Padova

Master Degree in *Computer Science*

a.y. 2017/2018

Data Mining

Teacher: Annamaria Guolo

Example of intermediate assessment

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 takes 0 points.

1) In the estimated linear regression model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ we have $\hat{\beta}_1 = 0$. Thus

(a) $R^2 = 0$ (b) $R^2 = 1$ (c) $R^2 = -1$ (d) none of the above

2) In the hypothesis testing, the observed significance level (p-value) is

(a) between 0 and $+\infty$ (b) between -1 and 1 (c) the type II error
(d) none of the above

3) In a linear regression model, the accuracy of the least squares estimates is measured using

(a) standard error (b) correlation (c) bias (d) sum of the residuals

4) When the sample size n increases, the width of the confidence intervals for the parameters in a linear regression model

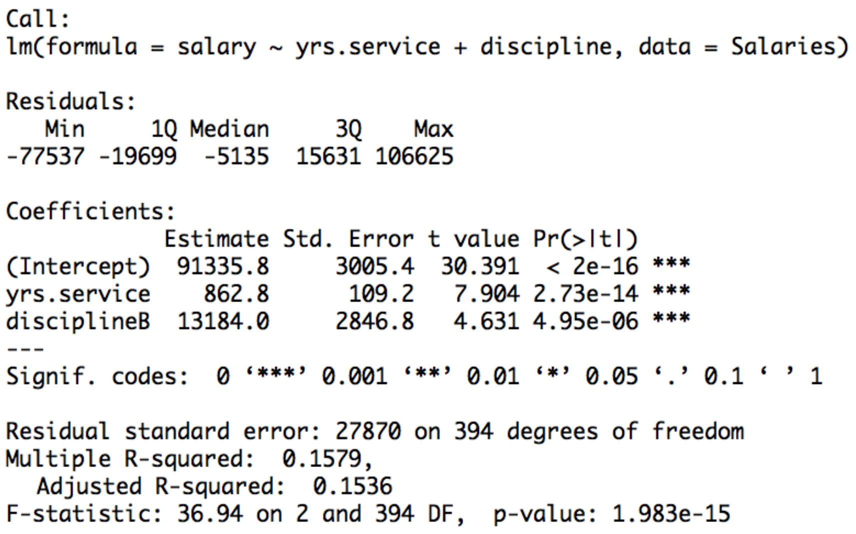
(a) decreases (b) increases (c) decreases until a certain n and then it increases
(d) it does not change

5) Errors ε in the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \ldots, n$, are assumed to be

(a) with mean equal to 1 (b) with variance equal to 1
(c) incorrelated with the covariates (d) incorrelated with Y

Exercise.
Consider the data about 397 teachers in a US college in the academic year 2008-2009. Data refer to years of service, discipline (A= theoretical, B= applied) and salary for 9 months in dollars.

a) We estimate a linear regression model to explain the relationship between the salary and the years of service and the discipline. This is the output from `R`



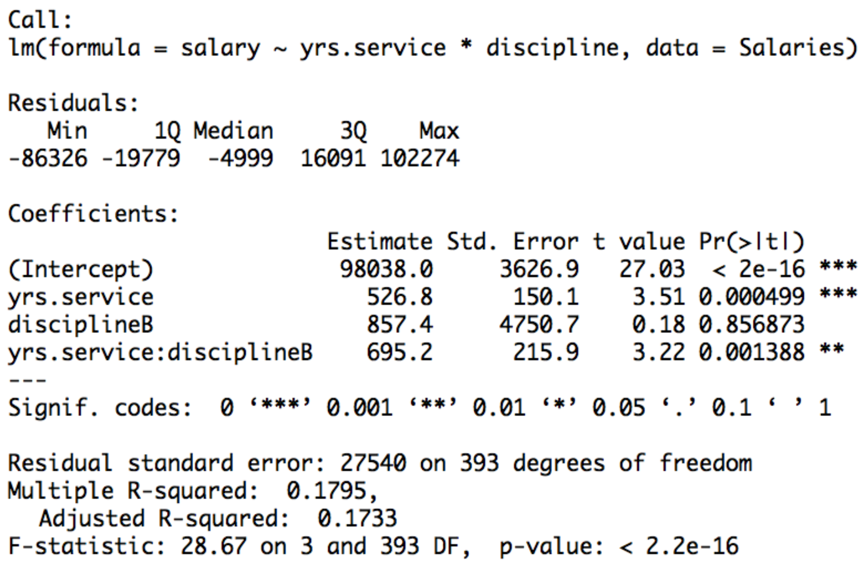
```
Call:
lm(formula = salary ~ yrs.service + discipline, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max 
-77537 -19699  -5135  15631 106625 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  91335.8     3005.4  30.391  < 2e-16 ***
yrs.service    862.8      109.2   7.904 2.73e-14 ***
disciplineB  13184.0     2846.8   4.631 4.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27870 on 394 degrees of freedom
Multiple R-squared:  0.1579,
  Adjusted R-squared:  0.1536 
F-statistic: 36.94 on 2 and 394 DF,  p-value: 1.983e-15
```

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) Provide a 95% confidence interval for the parameter associated to `yrs.service`, explaining possible assumptions, if any.

b) The extension of the model including the interaction between `yrs.service` and `discipline` provides the following output



```
Call:
lm(formula = salary ~ yrs.service * discipline, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max 
-86326 -19779  -4999  16091 102274 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              98038.0     3626.9   27.03  < 2e-16 ***
yrs.service                526.8      150.1    3.51 0.000499 ***
disciplineB                857.4     4750.7    0.18 0.856873    
yrs.service:disciplineB    695.2      215.9    3.22 0.001388 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27540 on 393 degrees of freedom
Multiple R-squared:  0.1795,
  Adjusted R-squared:  0.1733 
F-statistic: 28.67 on 3 and 393 DF,  p-value: < 2.2e-16
```

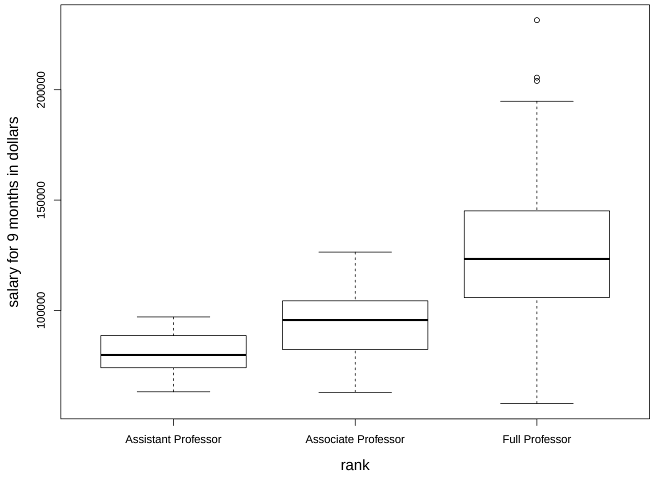
b.1) Does it make sense to maintain the interaction in the model? Can we simplify the model? Why?

b.2) Compare the two models using R^2 and discuss.

b.3) Compare the two models using statistic F, explaining the hypothesis test and discussing the result. Consider the significance level equal to 0.05.

b.4) Predict the salary for a teacher of a theoretical discipline with 20 years of service. Predict the salary for the teacher with the same years of service in case he/she teaches an applied discipline.

c) The following plot shows the distribution of the salary by distinguishing the rank of the teacher



c.1) Suppose to insert variable `rank` as a covariate (with no interactions) in the linear regression model with salary as response. Which level should be the baseline level? How many and which dummy variable would be constructed?

c.2) Discuss the plot. What could we expect in terms of significance of the parameters associated to variable `rank` in case `rank` would be inserted in the model?

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

$$F_{0.025;1,393} = 0.00098 \;\; F_{0.025;393,1} = 0.1975 \;\; F_{0.975;1,393} = 5.063 \;\; F_{0.975;393,1} = 1016.962$$

$$F_{0.05;1,393} = 0.0039 \;\; F_{0.025;393,1} = 0.2587 \;\; F_{0.95;1,393} = 3.865 \;\; F_{0.95;393,1} = 253.9898$$