University of Padova

Master Degree in *Computer Science*

a.y. 2017/2018

Data Mining

Teacher: Annamaria Guolo

Example of intermediate assessment: Solution

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

The right answer is (a).

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

The right answer is (d).

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

The right answer is (a).

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

The right answer is (a).

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

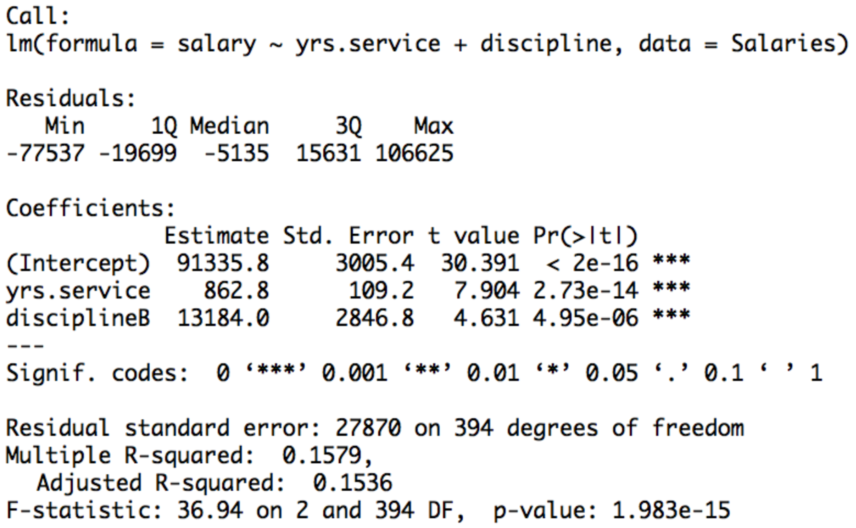
(c) incorrelated with the covariates (d) incorrelated with Y

The right answer is (c).

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. The estimated model is

$$\widehat{\texttt{Salary}} = 91335.8 + 862.8 \times \texttt{yrs.service} + 13184 \times I_{discipline=B}$$

where $I_{discipline=B}$ is the indicator function equal to 1 if `discipline=B` and equal to 0 if `discipline=A`. This is a consequence of variable `discipline` being qualitative with two levels, A and B, of which level A is the baseline level in `R`, on the basis of the alphabetical order. The qualitative variable with two levels is transformed into a numerical variable (quantitative) that assumes values 0 and 1 as described above. As a consequence, coefficient 13184 indicates the increase of salary for a subject teaching discipline B with respect to a subject teaching discipline A.

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.

The output of the model indicates that both the coefficients associated to the covariates are significantly far from zero, as the p-values associated to the hypothesis test for the coefficients being equal to zero are very small (almost zero). For this reason the model cannot be simplified. The small value of R^2 suggests that the model does not have a good fit (accuracy): it could be improved by inserting new covariates, interactions between covariates or polynomials for quantitative covariates.

a.3) Provide a 95% confidence interval for the parameter associated to `yrs.service`, explaining possible assumptions, if any.

Let β_1 be the coefficient associated to `yrs.service`, let $\hat{\beta}_1$ be its estimate and let $SE(\hat{\beta}_1)$ be the standard error. Thus

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-3}$$

and the required confidence interval is

$$\hat{\beta}_1 - t_{0.975,n-3} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{0.975,n-3} SE(\hat{\beta}_1)$$

Since $n - 3 = 394$ is very large, Student t distribution can be approximated to the $N(0,1)$ distribution (standard Normal): the interval can then be approximated with that based on the quantiles of a $N(0,1)$ distribution,

$$\hat{\beta}_1 - z_{0.975} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + z_{0.975,n-3} SE(\hat{\beta}_1),$$

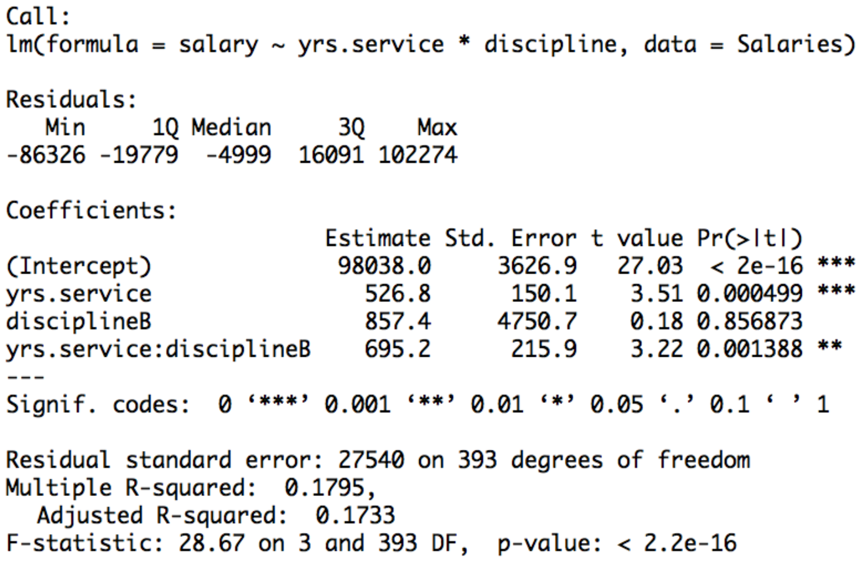
that is,

$$862.8 - 1.96 \times 109.2 \leq \beta_1 \leq 862.8 + 1.96 \times 109.2$$

giving rise to

$$648.768 \leq \beta_1 \leq 1076.832$$

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?

Yes, it makes sense as the p-values associated to the interaction term is very small. The associated hypothesis test for the coefficient of the interaction term being null is thus rejected. Although variable `discipline` is not significant, the model cannot be simplified given the principle of hierarchy, as the variable is included in the interaction and the interaction is significant.

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

Moving from the first model to the second model implies R^2 increasing from 0.1579 to 0.1795. Given that the new covariate (interaction term) is significant, the increase in R^2 is really an indication of an amelioration of the model accuracy.

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.

Consider the hypothesis test $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$, where β_3 is the coefficient associate to the interaction between `yrs.service` and `discipline`. Since the models are nested, we can compute the F statistic

$$F = \frac{RSS_0 - RSS}{RSS} \frac{n-p-1}{q},$$

where $q = 1$, $n - p - 1 = 397 - 3 - 1 = 393$, RSS_0 is the residuals sum of squares in the reduced model (the model with less variables) and RSS is the residuals sum of squares in the larger model. Given that $RSS_0 = RSE^2(n-p-1) = 27870^2 \times 394 = 306034338600$ and that $RSS = 27540^2 \times 393 = 298071478800$, the values of F is

$$\frac{306034338600 - 298071478800}{298071478800}\frac{393}{1} = 10.49884$$

Compare this value with the quantile of order 0.95 of a $F_{1,393}$ distribution, that si equal to 3.865. The observed value of F is larger than the quantile: thus we reject the null hypothesis at significance level 0.05. The variable remains inside the model, that is, we are not allowed to move to the reduced model without the interaction term. In this case, as the two models differ for one covariate only, the observed value of F coincides with the square of the observed value of t statistic for the significance of the interaction term, that is, the square of 3.22 (rounded).

As the requirement specifies the significance level 0.05, it is expected to carry out the hypothesis test in the way described above. If we would like to compute the p-value, instead, we would proceed as follows

$$P(F_{1,393} \geq 10.49884) = 1 - P(F_{1,393} < 10.49884) = 0.0013,$$

the small value confirms that we can reject the null hypothesis.

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.

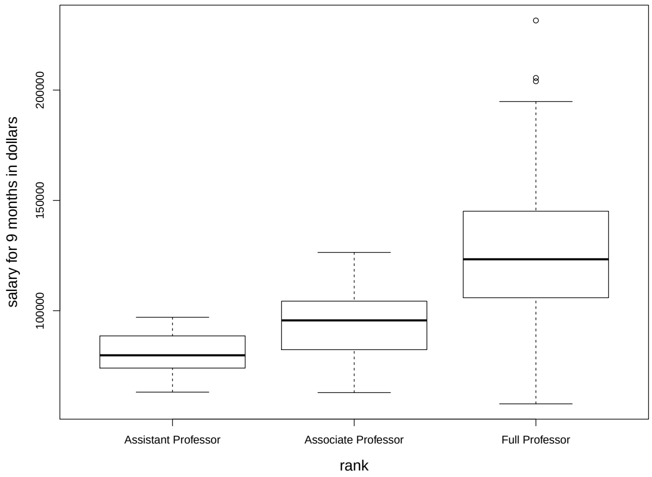
Teacher of a theoretical discipline:

$$\widehat{\texttt{Salary}} = 98038 + 526.8 \times 20 + 857.4 + 695.2 \times 20 = 123335.4$$

teacher of an applied discipline:

$$\widehat{\texttt{Salary}} = 98038 + 526.8 \times 20 = 108574$$

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?

Baseline level is Assistant Professor, given the alphabetical order. Since the variable has three levels, R constructs 2 dummies or 2 indicators: the first dummy assumes value 0 if the subject is not Associate professor and value 1 if the subject is Associate Professor; the second dummy assumes value 0 if the subject is not Full professor and value 1 if the subject is Full Professor.

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?

The plot shows that the distribution of variable Salary varies according to the levels of rank. In particular, the median of salary for an Assistant Professor is smaller than that for an Associate Professor, which is smaller of that for a Full professor. The variability of salary for a Full Professor is the largest.
After inserting the variabile rank in the model, we expect the coefficients associated to the two dummies being significantly far from zero. Moreover, given the medians in the boxplots, we expect that the coefficient associated to the Associate Professor will be positive (and significantly far from zero), indicating an increase of salary with respect to an Assistant Professor (baseline). We expect an even larger increase for the coefficient associated to the dummy for a Full Professor (and significantly far from zero). Given the width of the whiskers of the boxplots, we also expect a larger standard error for the estimate of the dummy associated to Full Professor with respect to that for Associate Professor.
In addition, the boxplot of the salary for a Full Professor shows the presence of three possibile outliers, that is, three teachers with a salary higher than the remaining. These subjects might be associated to larger residuals in the model fit: a residual analysis based on leverages and Cook's distance could clarify whether they are outliers or not.

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

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

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