University of Padova

Master Degree in *Computer Science*

a.y. 2017/2018

Data Mining

Teacher: Annamaria Guolo

Written assessment: June, 6, 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 comparison between two linear models can be performed through the F statistic

(a) always (b) never (c) only for large sample size
(d) only if the models are nested

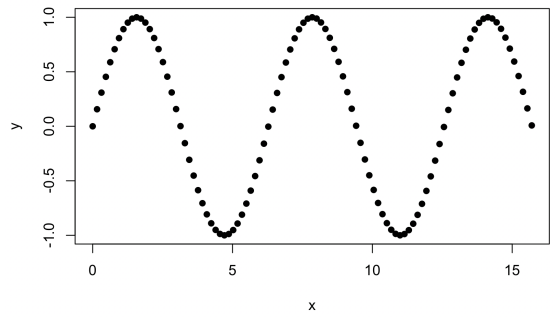
2) Type I error is

(a) the rejection of H_0 when it is true (b) the rejection of H_0 when it is false
(c) always smaller than 0.05 (d) non of the previous responses is true

3) A confidence interval for the coefficient associated to a covariate in the linear regression model based on n observations

(a) has a smaller width as n decreases (b) has a larger width as n increases
(c) does not depend on n (d) none of the previous responses is true

4) The following plot suggests that the correlation between X and Y is



(a) 0 (b) 1 (c) -1 (d) 0.7

5) Incorrelation between X and Y suggests that in the linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$

(a) the estimate of β_0 is 0 (b) the p-value associated to X is close to 1
(c) $R^2 = 0.3$ (d) the residual standard error is 0.5

Exercise.
Consider the dataset about the information of 120 houses on sale. Data include the following information:

- `Lprice`: Natural logarithm of the asking price (in thousands of dollars)
- `State`: Location of the house in US (CA, NJ, NY, PA)
- `Size`: Area of all rooms (in thousand square feet)
- `Beds`: Number of bedrooms
- `Baths`: Number of bathrooms

a) We estimate a linear regression model to explain the relationship between the logarithm of the price and some characteristics of the house. This is the output from `R`

```
Call:
lm(formula = Lprice ~ Size + State)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77915 -0.27947 -0.03962  0.29892  1.69701

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.84306    0.13469  35.957  < 2e-16 ***
Size         0.52416    0.04409  11.888  < 2e-16 ***
StateNJ     -0.13825    0.13835  -0.999 0.319735
StateNY     -0.01309    0.13842  -0.095 0.924802
StatePA     -0.47680    0.13860  -3.440 0.000811 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5358 on 115 degrees of freedom
Multiple R-squared:  0.5864,    Adjusted R-squared:  0.572
F-statistic: 40.76 on 4 and 115 DF,  p-value: < 2.2e-16
```

a.1) Write the expression of the estimated model. Describe how `R` handles the qualitative variable `State` 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) Which kind of information about the residuals is provided in the output? Why is this information useful?

a.4) Predict the price (on the original scale) for a house of 2 thousand square feet located in CA and compare it to that of an equal-size house located in NJ.

a.5) Compute a confidence interval at nominal level 0.95 for the coefficients associated to `Size`. Explain assumptions if any.

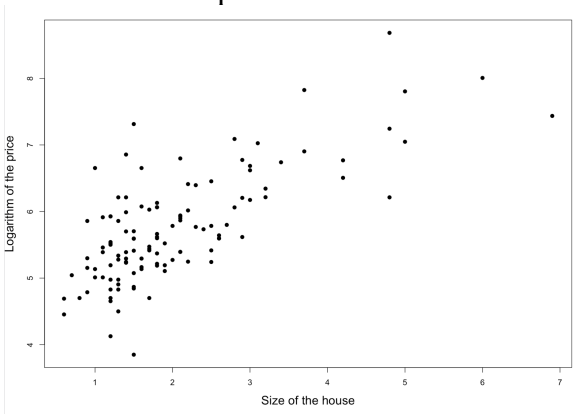
b) The extension of the model including the number of beds and bathrooms provides the output in the following page

b.1) Does it make sense to maintain the new variables in the model? Can be the model simplified? Why and how?

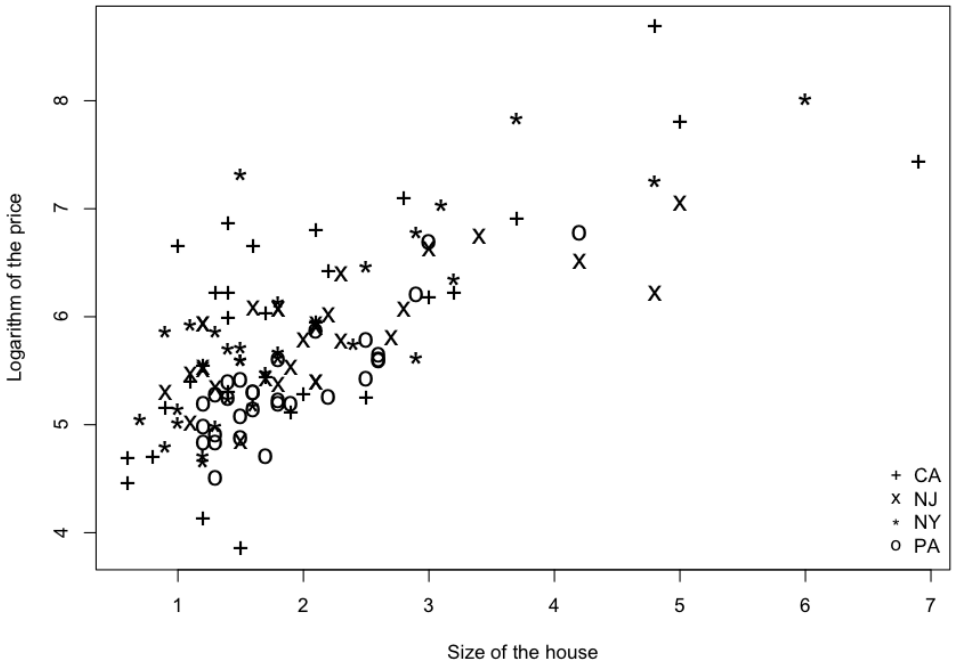
```
Call:
lm(formula = Lprice ~ Size + State + Beds + Baths)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76421 -0.27840 -0.06076  0.28479  1.70864

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.76738    0.17534  27.190  < 2e-16 ***
Size         0.39523    0.07576   5.217 8.33e-07 ***
StateNJ     -0.06764    0.14116  -0.479 0.632752
StateNY      0.03983    0.13761   0.289 0.772750
StatePA     -0.47654    0.13629  -3.497 0.000675 ***
Beds        -0.05569    0.06198  -0.899 0.370801
Baths        0.21060    0.08456   2.491 0.014204 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5262 on 113 degrees of freedom
Multiple R-squared:  0.6079,    Adjusted R-squared:  0.5871
F-statistic:  29.2 on 6 and 113 DF,  p-value: < 2.2e-16
```

b.2) Compare the two models using statistic F, explaining the hypothesis test and discussing the result. Consider the significance level equal to 0.05.

c) The following plot shows the distribution of the logarithm of the price versus that of `Size`. Does the plot provide any information useful to improve the model?



d) The following plot shows the distribution of the logarithm of the price versus that of `Size` by distinguishing the `State`. Does the plot suggest an interaction between the two covariates useful to improve on the fitting of 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

$$F_{0.025;2,113} = 0.0025 \;\; F_{0.025;113,2} = 0.262 \;\; F_{0.05;2,113} = 0.051$$

$$F_{0.05;1,113} = 0.0039 \;\; F_{0.95;2,113} = 3.077 \;\; F_{0.95;1,113} = 3.925$$