University of Padova

Master Degree in *Computer Science*

a.y. 2017/2018

Data Mining

Teacher: Annamaria Guolo

Intermediate assessment: April 12, 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) If in the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ the p-value associated to the null hypothesis of significance for β_1 is equal to 0.25, we can conclude that

(a) X is not associated to Y (b) $\beta_1 = 1$ (c) $R^2 = -1$ (d) $\rho_{XY} = 1$

2) Consider the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$. The test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ based on a sample of size 100 gives a value of the appropriate test statistic equal to -1.84. Then, the p-value is

(a) $P(t_{98} < -1.84)$ (b) $P(t_{98} > -1.84)$
(c) $2\min\{P(t_{98} < -1.84), P(t_{98} > -1.84)\}$ (d) none of the above

3) In the model $Y = \beta_0 + \beta X_1 + \cdots + \beta_p X_p + \varepsilon$ the coefficient β_j, $j = 1, \ldots, p$, indicates

(a) the average effect of X_j on Y while keeping the other covariates fixed
(b) the average effect of X_j on the other covariates (c) the average effect of Y on X_j
(d) the average effect of X_j on Y and on the other covariates

4) The explained deviance in a linear regression model

(a) is always larger than the residual deviance (b) decreases as the residual deviance increases
(b) increases as the residual deviance increases (d) decreases as the total deviance increases

5) The correlation coefficient ρ_{XY} is

(a) $0 \leq \rho_{XY} \leq 1$ (b) $\rho_{XY} \geq 0$ (c) $-\infty \leq \rho_{XY} \leq +\infty$ (d) $-1 \leq \rho_{XY} \leq 1$

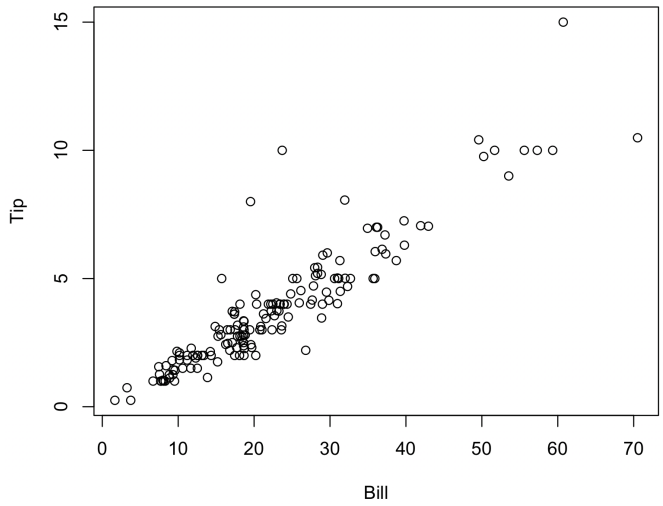
Exercise
Please provide clear and concise replies to the questions. When a computation is required, remember to report all the steps and not only the final result.

We analyze data about tips to the waiters given by 157 customers in a New York bistro. The variables are

- `tip`: Size of the tip to the waiter (in dollars)
- `bill`: Size of the bill (in dollars)
- `card`: Paid with a credit card? y/n
- `guests`: Number of people in the group
- `day`: Day of the week (Mon, Tue, Wed, Thu, Fri)

a) We want to estimate a model relating the size of the tip to the size of the bill and the other variables in the dataset.

a.1) A preliminary analysis of tip and bill provides the following graph.



On the basis of the graph, do you think it is reasonable to fit a linear regression model? Why?

a.2) The following output refers to a linear regression model relating the tip to the bill, the use of the credit card and the number of guests.

```
Call:
lm(formula = tip ~ bill + card + guests, data = ristorante)

Residuals:
      Min        1Q    Median        3Q       Max
-2.383831 -0.478400 -0.107596  0.272314  5.983526

Coefficients:
              Estimate  Std. Error  t value Pr(>|t|)
(Intercept) -0.25468361  0.20272531 -1.25630  0.21092
bill         0.18301835  0.00845809 21.63826  < 2e-16 ***
cardy        0.04217011  0.18281511  0.23067  0.81788
guests      -0.03318844  0.10281504 -0.32280  0.74729
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.98534 on 153 degrees of freedom
Multiple R-squared:  0.837519,
  Adjusted R-squared:  0.834334
F-statistic: 262.884 on 3 and 153 DF,  p-value: < 2.22e-16
```

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 90% confidence interval for the parameter associated to `bill`, explaining possible assumptions, if any.

a.4) Predict the tip in case of a dinner for 4 guests, when the bill is equal to 28 dollars and it is paid by cash. How does the tip change in case the customer pays by credit card? Comment.

a.5) What it the residual standard error in the output of the model? How is it computed?

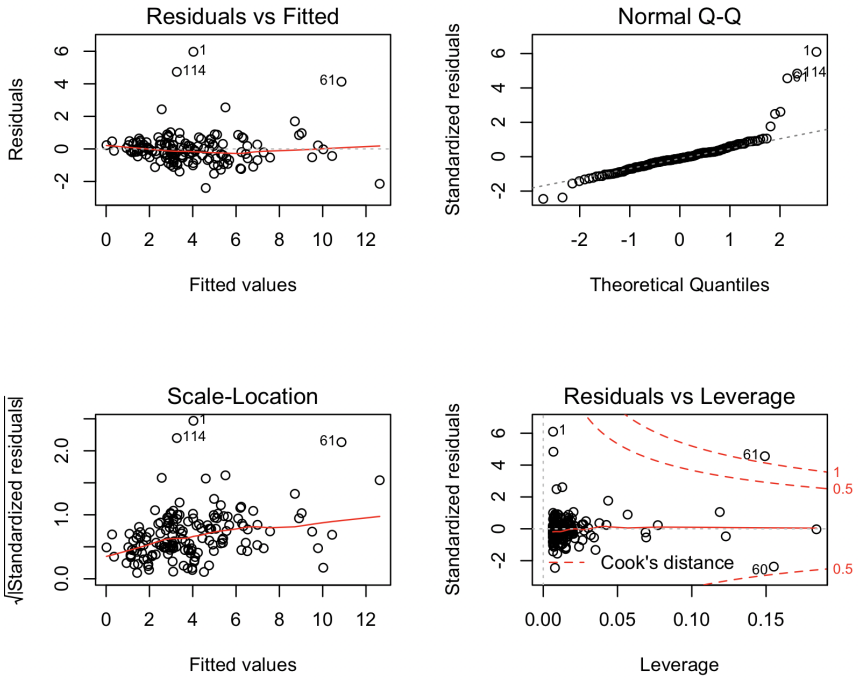
b) We reduce the model by eliminating variable `card`. The new model is

```
Call:
lm(formula = tip ~ bill + guests, data = ristorante)

Residuals:
      Min        1Q    Median        3Q       Max
-2.400753 -0.490897 -0.110722  0.271571  5.968875

Coefficients:
             Estimate  Std. Error  t value Pr(>|t|)
(Intercept) -0.25235366  0.20185016 -1.25020  0.21312
bill         0.18375096  0.00781501 23.51257  < 2e-16 ***
guests      -0.03570951  0.10191775 -0.35038  0.72654
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.982307 on 154 degrees of freedom
Multiple R-squared:  0.837463,
  Adjusted R-squared:  0.835352
F-statistic: 396.738 on 2 and 154 DF,  p-value: < 2.22e-16
```

b.1) Do you think it has been a good idea to eliminate variable `card`? Why?

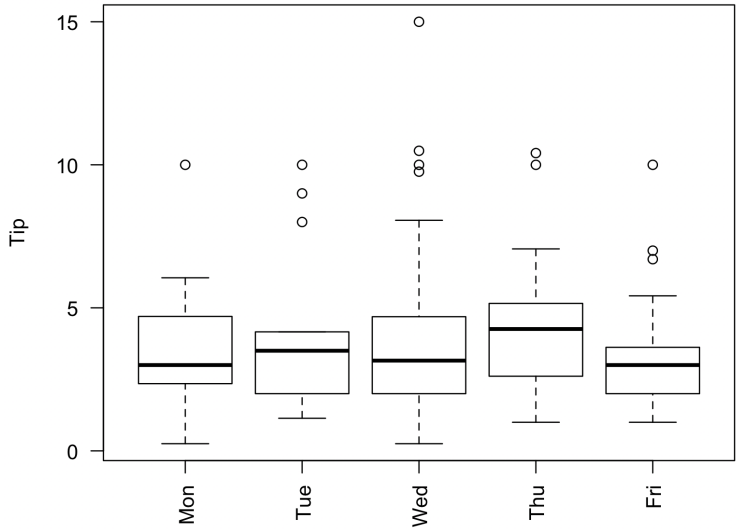
b.2) Compare the two models using the F statistic, explaining the hypothesis test and discussing the result. Consider the significance level equal to 0.02.

c) The following graph shows the residual analysis of the model without variable `card`



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.

d) The following plot is the distribution of `tip` as a function of the variable `day`



d.1) On the basis of the graph, suggest whether inserting variable `day` as a covariate with no interaction in the linear regression model without `card` can be useful to improve the accuracy of the model.

d.2) If we insert variable `day` in the model, which level should be the baseline level? How many and which dummy variables would be constructed?

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

$F_{0.95;1,153} = 3.90$ $F_{0.95;153,1} = 253.48$ $F_{0.02;1,153} = 0.0006$ $F_{0.98;1,153} = 5.527$ $F_{0.98;153,1} = 1586.021$ $F_{0.02;153,1} = 0.1809$