University of Padova

Master Degree in *Computer Science*

a.y. 2017/2018

Data Mining

Teacher: Annamaria Guolo

Written assessment: September, 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.

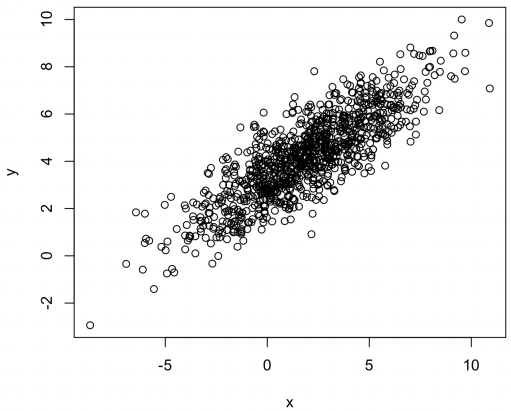
1) The classical hypotheses on the errors in the linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ include

(a) Student t distribution
(b) mean equal to zero
(c) increasing variance
(d) correlation between the errors

2) The predicted values of Y from a linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ are close to the observed values of Y

(a) when the residual deviance is close to zero
(b) when the total deviance is equal to 1
(c) as R^2 is close to zero
(d) when the sample size is small

3) The following plot suggests that the estimate of β_1 in the linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ could be equal to



(a) 0.5 (b) 0 (c) -0.5 (d) 2

4) How can we avoid the effects of multicollinearity between covariates X_1 and X_2 in the linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$?

(a) by eliminating both the covariates X_1 and X_2
(b) by eliminating X_1
(c) increasing the sample size
(d) by eliminating X_3

5) In a fitted linear regression model, the residual standard error is an estimate of

(a) the square root of the variance of the errors
(b) the variance of the covariate
(c) the coefficient relating the covariate and the response
(d) R^2

Exercise.
Consider the data about cigarette consumption for the 48 continental US States in two years, for an amount of 96 observations. Data include the following information:

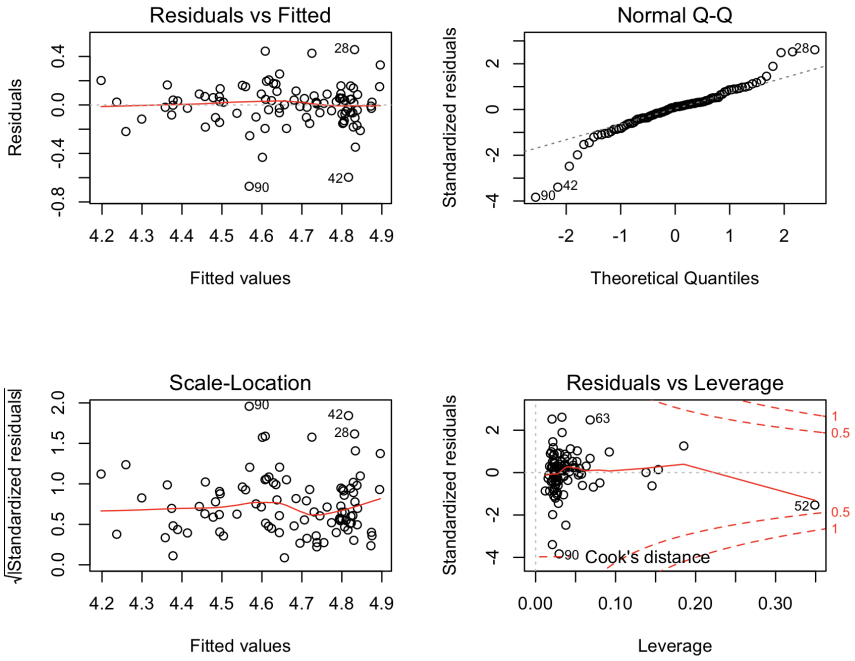
- `log.packs`: the natural logarithm of the number of packs of cigarettes per capita
- `income`: State personal income
- `tax`: Average state, federal and average local taxes
- `price`: Average price of cigarettes
- `year`: year of survey (1985 or 1995)

a) We estimate a linear regression model to explain the relationship between the logarithm of the number of packs and income, tax and price. This is the output from `R` (for clarity, consider that, for example, -2.244e-03 is equal to -0.002244)

```
Call:
lm(formula = log.packs ~ income + price + tax)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67086 -0.07389  0.01917  0.08624  0.45641

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.175e+00  6.359e-02  81.378   <2e-16 ***
income      -2.761e-10  1.611e-10  -1.714   0.0899 .
price       -2.244e-03  9.528e-04  -2.355   0.0206 *
tax         -3.764e-03  2.591e-03  -1.453   0.1497
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1776 on 92 degrees of freedom
Multiple R-squared:  0.4857,
  Adjusted R-squared:  0.4689
F-statistic: 28.96 on 3 and 92 DF,  p-value: 2.834e-13
```

a.1) Write the expression of the estimated model. 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.2) 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.3) Compute a confidence interval at nominal level 0.95 for the coefficients associated to `tax`. Explain assumptions if any. How can you comment on the result?

b) The model fitted excluding some variables provides the following output

```
Call:
lm(formula = log.packs ~ price)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62667 -0.08771  0.00589  0.08516  0.49617

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.2017203  0.0633872  82.063  < 2e-16 ***
price       -0.0037446  0.0004227  -8.858 4.92e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1808 on 94 degrees of freedom
Multiple R-squared:  0.455,
  Adjusted R-squared:  0.4492
F-statistic: 78.46 on 1 and 94 DF,  p-value: 4.921e-14
```

b.1) Write the expression of the estimated model. Comment on the model.

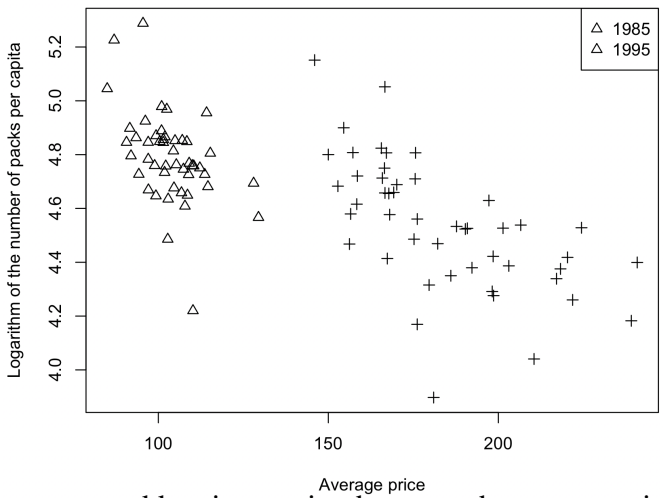
b.2) Compare the two models using statistic F, explaining the hypothesis test and discussing the result. Consider the significance level equal to 0.05.

c) The model in b) has been extended including covariate `year`, as shown in the following output

```
Call:
lm(formula = log.packs ~ price + year)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65654 -0.07330  0.01388  0.08494  0.44396

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.4772147  0.1046608  52.333  < 2e-16 ***
price       -0.0066276  0.0009808  -6.757 1.21e-09 ***
year1995     0.2761259  0.0856420   3.224  0.00174 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1724 on 93 degrees of freedom
Multiple R-squared:  0.5098,
  Adjusted R-squared:  0.4992
F-statistic: 48.35 on 2 and 93 DF,  p-value: 4.018e-15
```

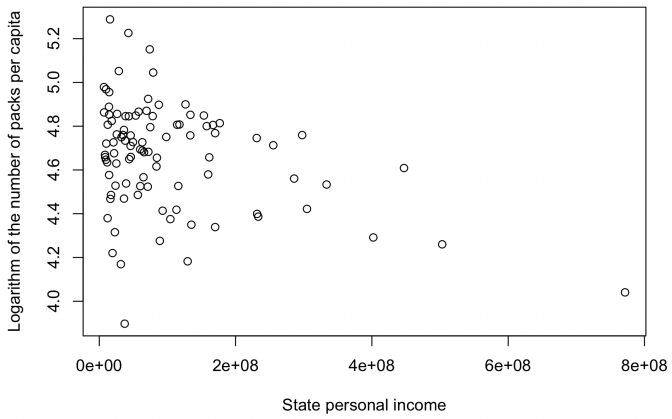
c.1) Write the expression of the estimated model. What kind of variable `year` is? How can the coefficient `year1995` in the output be interpreted? Comment on the model.

c.2) The following plot shows the distribution of the logarithm of the number of packs with varying `price` for different levels of `year` (triangles for 1985 and crosses for 1995)



Does the plot suggest to add an interaction between the two covariates in the model? Why?

d.1) The following graph is the dispersion plot of the logarithm of the number of packs and `income`



Does the graph suggest the possibility to improve the previous models with some interesting covariate/covariates? How? 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

$$F_{0.95;1,96} = 3.9401 \;\; F_{0.05;2,92} = 0.0513 \;\; F_{0.95;2,96} = 3.091 \;\; F_{0.95;2,92} = 3.095$$