vrije Universiteit amsterdam

Studentnumber:
Name:

School of Business and Economics

Exam:	Data Analysis 1
Code:	E_EOR1_DA1
Examinator:	Paolo Gorgi
Co-reader:	Hande Karabiyik
Date:	February 4, 2022
Time:	15:30
Duration:	2 hours
Calculator allowed:	**Yes**
Graphical calculator allowed:	**No**
Scrap paper	**Yes**
Number of questions:	3
Type of questions:	Open
Answer in:	English

Remarks:

Credit score:	100 credits counts for a 10
Grades:	The grades will be made public within 10 working days
Inspection:	TBA
Number of pages:	5

Good luck!

(This page is intentionally left blank.)

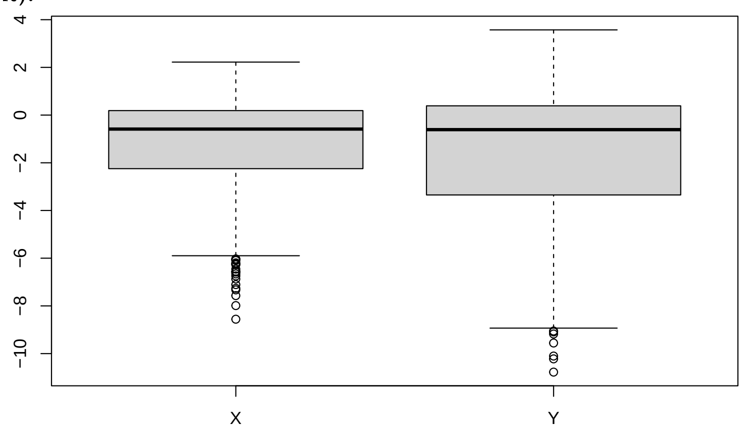
Question 1 (33/100 points)

(a) Consider the following data points

$$-6.7; \quad -2.5; \quad 3.2; \quad -2.1;$$

Obtain the sample mean and the sample variance.

(b) You have available a dataset that contains two variables x and y. For each variable, you have obtained the boxplot given below (boxplot of x is on the left and boxplot of y is on the right).



A colleague of yours makes the following statements:

(i) *"I expect both variables to have a negative skewness"*.

(ii) *"I expect the sample variance of x to be smaller than the sample variance of y"*.

(iii) *"I expect the kurtosis of y to be larger than the kurtosis of x"*.

(iv) *"I expect the variables to have a positive correlation r_{xy}"*.

For each statement, say whether you agree or not. Justify your answers.

(c) The R vectors "`age`" and "`salary`" contain the age and the monthly salary of 1000 individuals. The following R code is given:

```
n <- length(age)
out <- rep(0,n)
k <- 1
while(k <= n){
    if(age[k]>45){
        out[k] <- income[k]
        if(income[k]<=mean(income)){out[k] <- 0}
    }
  k <- k+1
}
```

Explain briefly what the R code is doing. What is contained in "`out`" after the *while loop*?

How would you write some R code that produces the same result but without using a loop? Sketch the code and explain what it does.

Question 2 (34/100 points)

(a) You have available a dataset that contains the variables `math_score` and `country` for some high school students. The variable `math_score` reports the result of an international math test and the variable `country` indicates the country of residence of the student. The variable `country` takes 3 possible values: 0 if the student is a resident of Belgium, 1 if the student is a resident of The Netherlands, and 2 if the student is a resident of Germany. You are interested in regressing `math_score` on `country`. Write down the regression model you would consider. Justify your choice. Discuss the interpretation of the regression coefficients of the model you have proposed.

(b) Available is a dataset with 2 variables and $n = 12$ observations for each of the 2 variables. Consider a linear regression model of the form $y_i = \beta_0 + \beta_1 x_i + u_i$. The OLS estimates of β_0 and β_1, the R^2 and the explained sum of squares (ESS) are obtained:

$$\hat{\beta}_0 = -6.5, \quad \hat{\beta}_1 = -3.1, \quad R^2 = 0.90, \quad ESS = 122.5.$$

(i) Obtain a prediction of y given $x = 3.5$.

(ii) Obtain the standard error of the regression (SER).

(iii) Obtain the sample variances of the variables s_x^2 and s_y^2.

(c) A colleague of yours has estimated the following regression models using a variable of interest y_i and 2 regressors, $x_{1,i}$ and $x_{2,i}$, $i = 1, \ldots, n$.

(1) $y_i = \beta_0 + \beta_1 x_{1,i} + u_i$.

(2) $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{1,i}^2 + u_i$.

(3) $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$.

Your colleague makes the following 2 statements:

(i) *"If the adjusted R^2 (adj-R^2) of model (1) is larger than the adj-R^2 of model (2), we can conclude that the relationship between y and x_1 is linear."*

(ii) *"I have obtained that the R^2 of model (3) is larger than the R^2 of model (2). Instead, the adj-R^2 of model (2) is larger than the adj-R^2 of model (3). There must be an error since R^2 and adj-R^2 provide the same information."*

Comment on each statement and say whether you agree or not. Justify your answers.

Question 3 (33/100 points)

(a) We have an observation x that we want to classify as a member of any of the three populations Π_1, Π_2 and Π_3. We know that population Π_1 has an exponential distribution[1] with rate $\lambda = 1$, population Π_2 has an exponential distribution with rate $\lambda = 2$ and population Π_3 has an exponential distribution with rate $\lambda = 4$.

(i) Obtain the discriminant regions R_1, R_2 and R_3 based on the Maximum Likelihood (ML) discriminant rule.

(ii) Obtain the probabilities of correct classification p_{11}, p_{22} and p_{33} of the ML rule.

(b) Consider the ML discriminant rule with two normal populations with means μ_1 and μ_2, $\mu_1 > \mu_2$, and the same variance σ^2. The discriminant regions are $R_1 = \{x : x > \frac{\mu_2+\mu_1}{2}\}$ and $R_2 = \{x : x \leq \frac{\mu_2+\mu_1}{2}\}$.

Show that the misclassification probabilities are given by

$$p_{12} = p_{21} = \Phi\left(-\frac{\mu_1 - \mu_2}{2\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

(c) You have implemented for a given dateset the ML discriminant rule based on two normal populations with means μ_1 and μ_2, $\mu_1 > \mu_2$, and the same variance σ^2. A colleague of yours suggest to estimate the misclassification probabilities p_{12} and p_{21} as follows

$$\hat{p}_{12} = \hat{p}_{21} = \Phi\left(-\frac{\bar{x}_1 - \bar{x}_2}{2s}\right),$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample mean of the observations from populations 1 and 2, and s is the sample standard deviation. Discuss potential advantages (if any) and disadvantages (if any) of the method proposed by your colleague. Could you present an alternative approach to estimate p_{12} and p_{21}?

End of the exam!

[1] The probability density function of an exponential distribution with rate $\lambda > 0$ is

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$