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:	January 29, 2021
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

$$7.7; \quad -3.5; \quad 8.6; \quad 2.8; \quad -1.7; \quad -1.2; \quad 0.5$$

Obtain the 1st, 2nd and 3rd quartile of this dataset.

(b) You have obtained the skewness and kurtosis of a certain variable. The skewness is $\gamma_1 = -2.35$ and the kurtosis is $\gamma_2 = 7.33$. A colleague of yours makes the following statements:

(i) *"The variable is leptokurtic and the left tail is heavier than the right tail"*.

(ii) *"There are outliers on the left tail of the distribution but not on the right tail"*.

(iii) *"I expect the mean to be larger than the median because the median is robust to outliers"*.

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

(c) The data frame `grades` contains the variables `name`, `math1`, `math2` and `math_final`. The variable `name` contains the names of the students, the variable `math1` contains the grades of the 1st part of a Math exam, the variable `math2` contains the grades of the 2nd part of the Math exam, and the variable `math_final` contains the final grade, which is missing. The dataset is given below:

```
> grades
  names math1 math2 math_final
1   Bob   6.5   4.5         NA
2  Lucy   5.5   8.5         NA
3   Eve   5.0   7.5         NA
4  Mark   7.0   5.0         NA
```

The following R code is given:

```
n <- length(grades$names)
k <- 1
repeat{
  grades$math_final[k] <- 0.3*grades$math1[k]+0.7*grades$math2[k]
  k <- k+1
  if(k>n){break}
}
```

Explain briefly what the R code is doing. What will be contained in `math_final` after the *repeat loop*?

The teacher of the course wants to set a minimum grade of 5.5 for each part of the exam in order to pass the course. She wants the final grade to be equal to the minimum between the grades of the two parts if at least one of the two parts has a grade that is lower than 5.5. How would you adjust the code to account for this? Sketch the code and explain what it does.

Question 2 (34/100 points)

(a) You have a dataset with three variables: x, y and z. The sample means, variances and covariances are obtained:

$$\bar{x} = -6.9, \quad \bar{y} = 0.7, \quad \bar{z} = 1.2; \qquad s_x^2 = 6.3, \quad s_y^2 = 1.0, \quad s_z^2 = 9.1;$$

$$s_{xy} = 1.3, \quad s_{xz} = 3.5, \quad s_{yz} = -2.6.$$

Say whether you agree or not with the following statements. Justify your answers.

(i) *"The relationship between x and z is stronger than the relationship between x and y."*

(ii) *"The simple linear regression model $y = \beta_0 + \beta_1 x + u$ will produce better predictions of y than the simple linear regression model $y = \beta_0 + \beta_1 z + u$."*

(b) Available is a dataset with 2 variables and $n = 16$ 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 standard error of the regression (SER) are obtained:

$$\hat{\beta}_0 = 2.3, \quad \hat{\beta}_1 = -1.7, \quad R^2 = 0.85, \quad SER = 3.5.$$

(i) Obtain a prediction of y given $x = -2.0$.

(ii) Obtain the total sum of squares (TSS), the residuals sum of squares (RSS) and the explained sum of squares (ESS) of the regression.

(iii) Obtain the sample correlation r_{xy} between x and y.

(c) A colleague of yours has estimated the following regression models using a variable of interest y_i and 3 regressors, $x_{1,i}$ $x_{2,i}$ and $x_{3,i}$, $i = 1, \ldots, n$.

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

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

(3) $y_i = \beta_0 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + u_i$.

Your colleague makes the following 2 statements:

(i) *"The adjusted R^2 (adj-R^2) of model (1) is larger than the adj-R^2 of model (2). Therefore, model (1) is better than model (2). This also means that there is no relationship between y and x_2."*

(ii) *"I have obtained that the R^2 of model (1) is larger than the R^2 of model (3). There must be an error since model (3) has more variables than model (1) and therefore its R^2 must be larger."*

For each statement, 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 a uniform distribution between 0 and 2 (i.e. $f_2(x) \sim U(0,2)$) and population Π_3 has a uniform distribution between -1 and 3 (i.e. $f_3(x) \sim U(-1,3)$).

(i) Obtain the discriminant regions R_1, R_2 and R_3 based on the Maximum Likelihood (ML) discriminant rule. Draw a graph of the densities $f_1(x)$, $f_2(x)$ and $f_3(x)$ of the three populations.

(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 missclassification 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. Discuss how the means μ_1 and μ_2 and the variance σ^2 of the normal distributions affect the missclassification probabilities p_{12} and p_{21}.

(c) Assume we have two normal[2] populations Π_1 and Π_2 with means equal to zero and different variances σ_1^2 and σ_2^2, $\sigma_1^2 > \sigma_2^2$. More specifically, we have $f_1(x) \sim N(0, \sigma_1^2)$ and $f_2(x) \sim N(0, \sigma_2^2)$. Derive the discriminant regions R_1 and R_2 of the ML discriminant rule.

End of the exam!

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

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

[2] The probability density function of a normal $N(\mu, \sigma^2)$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$