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:	March 16, 2021
Time:	16:00
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 dataset

$$5.2; \quad -3.2; \quad -2.6; \quad 1.9;$$

Obtain the sample mean and sample variance.

(b) You have available a dataset that contains two variables x and y. For each variable, you have obtained the first, second and third quartile, which are given by
-Variable x: $Q_1 = 1.4$, $Q_2 = 2.8$ and $Q_3 = 7.4$
-Variable y: $Q_1 = 2.2$, $Q_2 = 4.3$ and $Q_3 = 6.4$
A colleague of yours makes the following statements:
(i) *"About 75% of the observations of x are contained in the interval* $[1.4, 7.4]$ *and about 75% of the observations of y are contained in the interval* $[2.2, 6.4]$*"*.
(ii) *"I expect the sample variance of x to be larger than the sample variance of y"*.
(iii) *"I expect x to have skewness close to zero and instead y to have a strong negative skewness"*.
For each statement, say whether you agree or not. Justify your answers.

(c) The R vector "`income`" contains the annual income of 350 citizens. The following R code is given:

```
n <- length(income)
x <- rep(0,n)

for(i in 1:n){
   if(income[i]<median(income)){next}
   x[i] <- income[i]
}
```

What is contained in the R object `x` after running the for loop given above? Explain briefly what the R code is doing. 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) 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 sample means $\bar{x}$ and $\bar{y}$, the sample variances s_x^2 and s_y^2, and the sample correlation r_{xy} between x and y are given:

$$\bar{x} = -1.5, \quad \bar{y} = 0.9, \quad s_x^2 = 2.1, \quad s_y^2 = 3.5, \quad r_{xy} = -0.95.$$

(i) Obtain the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(ii) Obtain the R^2 of the regression.

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

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

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

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

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

The colleague makes the following 2 statements:

(i) "*I have obtained that the R^2 of model (1) is larger than the R^2 of model (2). This means that the relationship between y and x_1 is linear.*"

(ii) "*The OLS estimate of β_1 is positive in model (1) and negative in model (2). This means that model (1) suggests a positive relationship between x_1 and y and model (2) suggests a negative relationship between x_1 and y. There must be an error.*"

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

(c) Consider the regression model without the intercept $y_i = \beta_1 x_i + u_i$ with OLS estimate of β_1 given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

Does the estimated regression line of this model go through the point ($\bar{x}$,$\bar{y}$), where $\bar{x}$ and $\bar{y}$ are the sample means of x and y? Justify your answer.

Question 3 (33/100 points)

(a) Consider the following *confusion matrix* containing the number of misclassified and correctly classified observations for the populations Π_1 and Π_2.

		True membership	
		Π_1	Π_2
Predicted	Π_1	$n_{11} = 97$	$n_{12} = 37$
	Π_2	$n_{21} = 4$	$n_{22} = 215$

Obtain the estimated probabilities of misclassification $\hat{p}_{12}$ and $\hat{p}_{21}$ and the apparent error rate (APER).

(b) 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 $\lambda = 4$ and population Π_3 has a uniform distribution between -1 and 0 (i.e. $f_3(x) \sim U(-1, 0)$).

(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.

(c) Assume we have two normal 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 $f_2(x) \sim N(0, \sigma_2^2)$. The discriminant regions R_1 and R_2 of the ML discriminant rule are

$$R_1 = \Big(-\infty, -g(\sigma_1^2, \sigma_2^2) \Big] \cup \Big[g(\sigma_1^2, \sigma_2^2), +\infty \Big), \quad R_2 = \Big(-g(\sigma_1^2, \sigma_2^2), g(\sigma_1^2, \sigma_2^2) \Big).$$

where

$$g(\sigma_1^2, \sigma_2^2) = \sqrt{\log(\sigma_1^2/\sigma_2^2) \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 - \sigma_2^2}}$$

Derive the expressions of the probabilities of misclassification p_{12} and p_{21} (as functions of σ_1^2 and σ_2^2). Discuss how the variances σ_1^2 and σ_2^2 affect the misclassification probabilities p_{12} and p_{21}.

End of the exam!

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

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