vrije Universiteit *amsterdam*

School of Business and Economics

Exam:	Data Analysis 1
Code:	E_EOR1_DA1
Examinator:	dr. Paolo Gorgi
Co-reader:	dr. Hande Karabiyik
Date:	February 2, 2018
Time:	15:15
Duration:	2 hours
Calculator allowed:	Yes
Graphical calculator allowed:	No
Number of questions:	3
Type of questions:	Open
Answer in:	English
Credit score:	100 credits counts for a 10
Grades:	The grades will be made public within 10 working days
Inspection:	TBA
Number of pages:	4 (including front page)

Good luck!

(This page is intentionally left blank.)

Question 1 (30/100 points)

(a) Find the sample variance of the following 3 data points:

$$1.0\,; \quad 3.3\,; \quad 2.6$$

(b) For a certain variable you have obtained that the skewness is -2.2 and the kurtosis is 12.5. What can you say about the distribution of the observations? Would you expect to have some outliers?

(c) Consider the following R code

```
> x <- c(1, 5, 7, 3, 2)
> z <- x[x>=3]
```

What is in the R object `z`? Explain briefly what the R code is doing.

(d) The following R code with a `for` loop is given

```
> v <- 1:5
>
> for(i in 1:5){
+   if(v[i]==4) {break}
+   v[i] <- v[i]-1
+ }
```

What is in the R object `v` after running the `for` loop? Explain briefly what the R code is doing.

Question 2 (40/100 points)

(a) Available is a dataset with 2 variables and $n = 100$ 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 ($\hat{\beta}_0$ and $\hat{\beta}_1$), the RSS and the TSS are obtained:

$$\hat{\beta}_0 = 2.6, \quad \hat{\beta}_1 = 1.5, \quad RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 308.6, \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 558.6.$$

(i) Interpret the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(ii) Obtain a prediction for the variable y when the observed x is equal to 3.5.

(iii) Obtain the R^2 and the standard error of the regression (SRE).

(b) A colleague of yours has estimated the linear regression model $y_i = \beta_0 + \beta_1 x_i + u_i$ using a certain dataset. She claims that the adjusted-R^2 (R^2_{Adj}) obtained from the regression is negative. Is this possible? Why? What can you say about the relationship between the variable y_i and x_i?

(c) Consider the regression model without intercept given by $y_i = \beta_1 x_i + u_i$.

(i) Show that the OLS estimate of β_1 is

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

by setting the derivative of the sum of squares to zero.

(ii) Show that, in general, the equality $TSS = ESS + RSS$ is no longer true in the regression model without the intercept.

Question 3 (30/100 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	$n_{11} = 125$	$n_{12} = 21$
	Π_2	$n_{21} = 13$	$n_{22} = 174$

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 either population Π_1 or Π_2. We know that the populations Π_1 and Π_2 have an exponential distribution with rates $\lambda_1 = 1$ and $\lambda_2 = 2$, respectively. Note that the density function of an exponential distribution with rate $\lambda > 0$ is

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

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

(ii) Obtain the misclassification probabilities p_{12} and p_{12}.

(iii) Assume that $C(1|2) = 2C(2|1)$, that is, the misclassification cost $C(1|2)$ is 2 times the misclassification cost $C(2|1)$. How would you expect the regions R_1 and R_2 obtained from the ECM discriminant rule to differ from the ones obtained from the ML rule? Justify your answer. *Do not calculate the ECM discriminant regions!*

End of the exam!