DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Multivariate Statistical Methods for Engineering and Management

Master in Industrial Engineering and Management

1st Semester – 2021/2022

2nd Exam 25/02/2022 – 10:30h – Room: A3 **Duration: 2h**

Justify your answers

Group I 10 points

1. The **gala** is a data set on the species diversity on the Galapagos Islands. The data were collected to study the relationship between the number of plant species (**Species**) and various geographic variables on the 30 Galapagos islands. A regression model was fitted to this data, with **Species** as the response and the three predictors: **Area** (area of the island, km^2), **Elevation** (highest elevation of the island, m) and **Adjacent** (area of the adjacent island, km^2). The results obtained using R (with some missing values) are:

```
Call:
lm(formula = Species ~ Area + Elevation + Adjacent, data = gala)

Coefficients:
           Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -5.71893   16.90706  -0.338  0.73789
Area        -0.02031    0.02181    ??      ??
Elevation    0.31498    0.05211   6.044  2.2e-06 ***
Adjacent    -0.07528    0.01698  -4.434  0.00015 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61.01 on 26 degrees of freedom
Multiple R-squared:  0.746,Adjusted R-squared:  0.7167
```

For the **Baltra** island the observed value was:

Species=58 and $\mathbf{x}_{\text{Baltra}}^{\text{T}} = (1, \textbf{Area}, \textbf{Elevation}, \textbf{Adjacent})^{\text{T}} = (1, 25.09, 346, 1.84)$. Using the R output:

(a) Write the fitted regression function. Obtain the residual for the **Baltra** observation. (1.0)

(b) Test, with $\alpha = 0.01$, whether there is a linear association between the expected value of **Species** and the three predictors variables under study. State the hypotheses, test statistic, decision rule and the conclusion. What are the assumptions that you have to admit in order to solve this question? (2.5)

(c) Test if the predictor variable **Area** is statistically helpful. Decide based on the p-value. (1.5)

(d) Derive a 95% confidence interval to the expected value of **Species** variable for the **Baltra** island, knowing that $\mathbf{x}_{\text{Baltra}}^{\text{T}}\mathbf{C}\,\mathbf{x}_{\text{Baltra}} = 0.0468$. (2.0)

2. For a study with $a = 3$ and $N = 83$, complete the one-way Anova table (3.0)

Source	SS	df	MS	F
Treatments	—	—	12.50	—
Error	190	—	—	
Total	—	—		

and check out if the mean of the response variable is the same for all treatments. State the hypothesis under test and the decision rule for $\alpha = 0.05$.

Group II 10 points

1. Consider the variables X_1=**Area**, X_2=**Elevation** and X_3=**Adjacent** from the **gala** data set presented in **Group I**. The sample covariance matrix of the **gala** standardized data set, ($z_i = \frac{x_i - \bar{x_i}}{s_i}$), have the following orthonormal eigenvectors ($\hat{\gamma}_i$) and eigenvalues, $i = 1, 2, 3$:

$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
-0.5806	-0.5683	0.5830
-0.6729	-0.0682	-0.7366
-0.4583	**a**	0.3428
$\hat{\lambda}_1$=2.0157	$\hat{\lambda}_2$= 0.8306	**b**

(a) Compute the missing values **a** and **b**. (1.5)

(b) Write the first two sample principal components and interpret them. (1.5)

(c) Decide, based on the percentage of total sample variance explained, how many sample principal components should be retained. (1.5)

(d) Obtain the sample correlation between the first sample principal component and each standardized variable and interpret them. Compare with (b) result. (1.5)

2. The values of four binary variables are measured for each of five individuals as follows:

	variables			
	x_1	x_2	x_3	x_4
o_1	1	0	1	0
o_2	1	1	1	1
o_3	0	1	1	0
o_4	0	1	1	1
o_5	0	1	0	0

Given two individuals, o_i and o_j, the Jaccard similarity coefficient is defined as $J_{ij} = \frac{a}{a+b+c}$, where a, b and c are the counts specified in contingency table such as: $\#\{(1,1)\} = a$, $\#\{(1,0)\} = b$ and $\#\{(0,1)\} = c$.

(a) Compute the dissimilarity matrix **D**, using $d_{ij} = 1 - J_{ij}$. (2.0)

(b) Taking **D**, use the average linkage method to group the previous individuals and draw the corresponding dendrogram. How many clusters would you recommend? (2.0)