Bachelor Degree in Computer Engineering - **SECOND PARTIAL** **31st-may-2013**

Group E Surname, name: ______________________________ Signature:

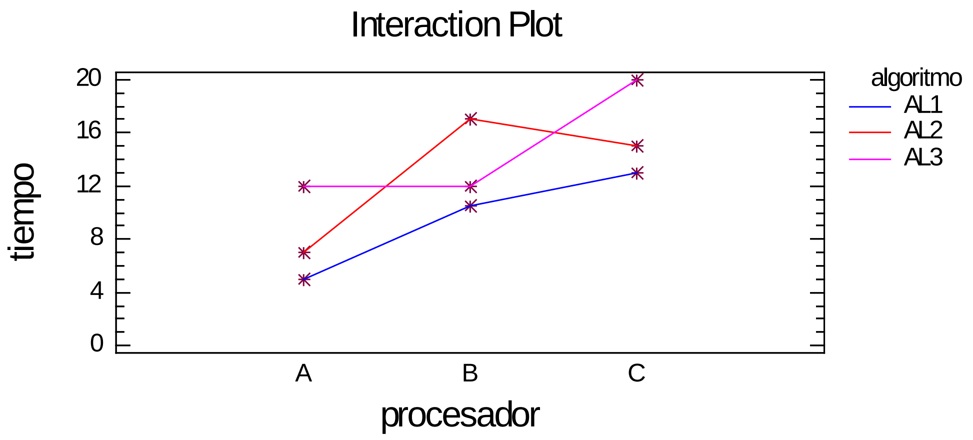
1) **[4 points]** One experiment has been conducted to study the effect of type of processor and algorithm on the time required to inverse big data matrixes. For this purpose, two matrixes of similar characteristics (matrix_1 and matrix_2) have been inversed with 3 different algorithms (AL1, AL2, AL3) and with 3 types of processors (A, B, C), resulting 18 values of time (in milliseconds). Processor A has a RAM memory of 10 MB, processor B has 30 MB of RAM, and C has 20 MB of RAM. Data were analyzed with ANOVA, resulting the following table. Factor "matrix" was not statistically significant and it was not considered in this study.

```
Analysis of Variance for TIME - Type III Sums of Squares
----------------------------------------------------------------------------
Source                     Sum of Squares      Df      Mean Square         F-Ratio
----------------------------------------------------------------------------
MAIN EFFECTS
 A:processor                      197,444
 B:algorithm                                                 41,7222

INTERACTIONS
 AB                               66,8889

RESIDUAL
----------------------------------------------------------------------------
TOTAL (CORRECTED)                 416,278
----------------------------------------------------------------------------
```

1a) Complete the summary table of ANOVA. **[1 point]**

1b) What effects are statistically significant, considering α=0.05? Justify conveniently the answer. **[1 point]**

1c) The interaction plot corresponding to both factors is shown below. What information is deduced from this plot, taking into account the results in the ANOVA table? **[1 point]**



1d) Based on the plot shown above, obtain the Means Plot for the factor "processor". Draw also the LSD intervals, taking into account that the total width of these LSD intervals (for a confidence level of 95%) is 3.6 units (that is, $\overline{x_i} \pm 1.8$). **[0.5 points]**

1e) Based on the plot obtained in question d), can we affirm that the RAM memory of processors used in this experiment has a linear or quadratic effect on the time for matrix inversion? Justify your answer. **[0.5 points]**

2) One researcher intends to design an experiment with all factors either at 2 levels, or at 3 levels. What is the main advantage of using all factors at 3 levels? What is the main disadvantage? **[1 point]**

3) **[2.5 points]** The computer department of a mobile phone company wants to obtain a statistical model to predict the monthly expenses (€) of customers who hire a certain telephone rate (restricted to people over 20 years). A representative sample of customers is randomly selected, and two variables are considered: age and time spent in the company (years). Data were analyzed with Statgraphics, and the following table was obtained:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: EXPENSES
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 45,0561        3,73421       12,0658          0,0000
Age                     -0,16612      0,0708991      [redacted]      [redacted]
Time                     3,22028       0,354496       9,08411          0,0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     9315,25      2      4657,63      43,24       0,0000
Residual                  10556,5     98       107,72
-----------------------------------------------------------------------------
```

Answer the following questions, justifying conveniently your responses.

3a) One technician considers that age does not affect the expenses at the population level because the estimated regression coefficient of this variable is small (value of -0.16612). Considering α=0.05, is it possible to admit the opinion of the technician? **[1 punto]**

3b) Calculate the monthly expenses expected for a customer of age 40 years who has remained in the company for 3 years. **[0.5 puntos]**

3c) If a customer of age 40 years has remained 3 years in the company, what is the probability to spend more than 45 euros in one month? **[1 point]**

4) **[2.5 p.]** Given these pairs of values: (X=1, Y=5), (X=2, Y=7), (X=4, Y=9), (X=5; Y=11),

a) Calculate the covariance: cov (X;Y).

b) The variance of X is 10/3. If a simple linear regression is fitted with these pairs of values, obtain the mathematical equation of the regression model.

c) Calculate the residual corresponding to X=1.

SOLUTION

1a) Total degrees of freedom (DF) = nº data - 1 = 17. DF_A=DF_B = nº variants - 1 = 3-1 = 2.
DF of the interacction = 2·2 = 4. DF residual = 17 - 2 - 2 - 4 = 9.
SS_B = 41.722 · 2 = 83.444 ; SS_{res} = 416.28 - 197.44 - 83.44 - 66.89 = 68.5
MS = SS / df; F-ratio = MS / MS_{res} The complete ANOVA table is the following:

```
Analysis of Variance for TIME - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                  Sum of Squares     Df    Mean Square    F-Ratio     P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:processor                   197,444      2        98,7222      12,97      0,0022
 B:algorithm                   83,4444      2        41,7222       5,48      0,0277

INTERACTIONS
 AB                            66,8889      4        16,7222       2,20      0,1502

RESIDUAL                          68,5      9        7,61111
--------------------------------------------------------------------------------
TOTAL (CORRECTED)              416,278     17
--------------------------------------------------------------------------------
```

1b) $F_{ratio_A} \approx F_{2;9}$; $F_{ratio_B} \approx F_{2;9}$; $F_{ratio_{A \cdot B}} \approx F_{4;9}$ Considering α=0.05, the null hypothesis is rejected if the F-ratio obtained is higher than the critical value from the table: $F_{2;9}^{0.05} = 4.26$; $F_{4;9}^{0.05} = 3.63$

As the F-ratios for factor processor (12.97) and for factor algorithm (5.48) are higher than 4.26, the null hypothesis is rejected: the simple effect of both factors is statistically significant.
As the F-ratio of the interaction: 2.2 < 3.63, there is not enough evidence to affirm that the double interaction between both factors is statistically significant.
Note: p-values cannot be computed "by hand", but they have also been included in the table.

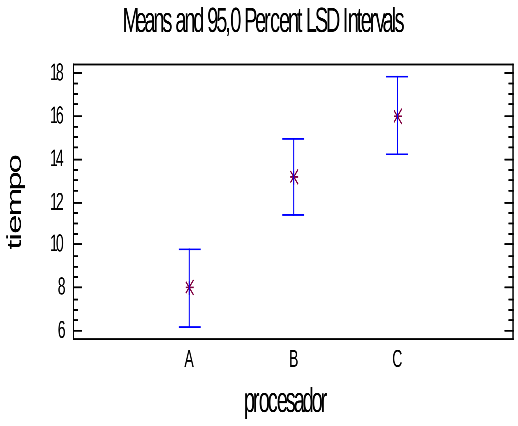
1c) The null hypothesis associated to factor processor is H_0: m_{procA} = m_{procB} = m_{procC}. According to the results from the ANOVA table, the null hypothesis is rejected. Based on the interaction plot, the average time of A is lower than the time for B and lower than C. At the population level, the average time of A will be lower than in the case of C, but with the information available it is not possible to know if the average time of F differs significantly from the others.

Similarly, the null hypothesis associated to factor algorithm is that the average time for AL1, AL2 and AL3 is the same. According to question 1b), this null hypothesis is rejected. Based on the interaction plot, it can be deduced that the average time for AL1 [mean=(5+10+13)/3] is lower than for AL2 [mean=(7+17+15)/3] and also lower than in the case of AL3 [mean=(12+12+20)/3)]. Thus, it can be concluded that algorithm 1 will require less time in average, at the population level, than algorithm 3, but it is not clear if the average time of AL2 will differ significantly from the others.

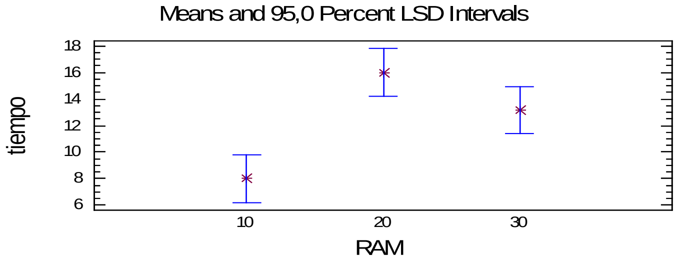
The interaction plot also shows that the three lines (AL1, AL2, AL3) are not parallel. However, the double interaction is not statistically significant. Therefore, there is not enough evidence to conclude that the effect of processor depends on the type of algorithm, or vice versa.

1d) The values deduced from the interaction plot are indicated in the following table. From these values it is possible to calculate the average time for each type of processor. LSD interval for processor A: 8 ± 1.8 ; processor B: 13 ± 1.8 ; processor C: 16 ± 1.8. These intervals are drawn in the following figure:

processor	algorithm	time	mean
A	AL1	5	
A	AL2	7	8
A	AL3	12	
B	AL1	10	
B	AL2	12	13
B	AL3	17	
C	AL1	13	
C	AL2	15	16
C	AL3	20	



1e) The figure obtained in the previous section seems to indicate that the effect of processor is linear. But this is not the case, because processor A has a RAM memory of 10 MB, B has 30 MB and C has 20 MB. If the order of the variants (processors) is changed according to the RAM memory, the following figure is obtained. Taking into account that the LSD interval for 10 RAM does not overlap with the other two intervals, and given that the relationship is not linear, it can be deduced that there is a quadratic relationship between time and RAM memory. A relative maximum value of time is reached corresponding to a RAM memory slightly lower than 20 MB.



2) The advantage of using an experimental design with all factors at 3 levels is that it is possible to study if the effect is linear or quadratic, in case of being statistically significant. The disadvantage is that the total number of experimental trials required is obviously higher than in the case of factors at two levels.

3a) Total degrees of freedom (DF) appearing in the global significance test of the model = N-1:
$DF_{total} = DF_{model} + DF_{residual}$ = 98 + 2 = 100 = N-1. Thus, N=101 (number of observations in the model).

In the model: $Expenses = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot time$, the null hypothesis $H_0 : \beta_1 = 0$ will be accepted if the following condition is satisfied: $\left| b_1 / S_{b_1} \right| < t^{\alpha/2}_{N-1-I}$ In this case, I=2 (two explicative variables in the model), N=101, b_1 = -0.16612 (estimated value of the coefficient associated to age), s_{b1} = 0.0708991 (standard error)
Crictical value from the t-table: $t^{\alpha}_{N-1-I} = t^{0.025}_{101-1-2} = t^{0.025}_{98} = 1.99$
The ratio $\left| b_1 / S_{b_1} \right| = \left| -0.16612/0.0708991 \right| = 2.343$ is higher than the critical value 1.99 and, hence, the null hypothesis is rejected. Thus, there is enough evidence to conclude that the coefficient β_1 is different from zero at the population. Therefore, the opinion of the technician is **NOT** admissible.

3b) Regression equation: Expenses = 45.0561 - 0.16612·age + 3.22028 · time
If age=40 and time=3: E(expenses)= 45.0561 - 0.16612 · 40 + 3.22028 · 3 = **48.07** eur

3c) Residual variance = residual square mean = 107.72
If age=40 and time=3: P(expenses>45) = P [N(m=48.07; s^2=107.72) >45] =
$= P\left[N(0;1) > (45-48.07)/\sqrt{107.72}\right] = P\left[N(0;1) > -0.296\right] = 1 - 0.384 =$ **0.616**

4a) $\bar{x} = 3$; $\bar{y} = 8$; $\text{cov}(x,y) = \sum\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)/(n-1)$
$\text{cov} = \left[(1-3)\cdot(5-8)+(2-3)\cdot(7-8)+(4-3)\cdot 9-8)+(5-3)\cdot(11-8)\right]/3 = (6+1+1+6)/3 =$ 14/3= **4.67**

4b) $b = \text{cov}/s_x^2 = (14/3)/(10/3) = 1.4$; $a = \bar{y} - b\cdot\bar{x} = 8 - 1.4\cdot 3 = 3.8$
Regression line: $y = 3.8 + 1.4\cdot x$

4c) $residual_{x=1} = y_{observed} - y_{predicted} = 5 - (3.8 + 1.4\cdot 1) = 5 - 5.2 =$ **-0.2**