Grado en Ingeniería Informática - **SEGUNDO PARCIAL** **31-mayo-2013**

Grupo: ____ Apellidos, nombre: _____ Firma:

1) **[4 puntos]** Se ha realizado un ensayo para estudiar el efecto del tipo de procesador y algoritmo en el tiempo que se tarda para invertir grandes matrices de datos. Para ello, dos matrices de características similares (matriz_1 y matriz_2) se han invertido con tres algoritmos distintos (AL1, AL2, AL3) y con tres tipos de procesador (A, B, C), obteniéndose en total 18 valores de tiempo (en milisegundos). El procesador A tiene una memoria RAM de 10 MB, el procesador B tiene una RAM de 30 MB y el C tiene una RAM de 20 MB. Los datos de tiempo se han analizado con ANOVA, obteniéndose la siguiente tabla. El factor "matriz" no ha resultado ser estadísticamente significativo y no se ha considerado en el estudio.

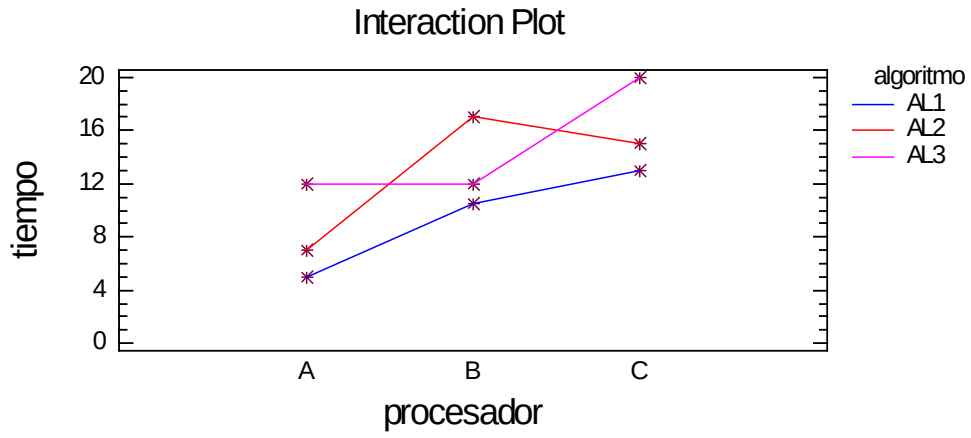
Analysis of Variance for TIEMPO - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:procesador	197,444			
B:algoritmo			41,7222	
INTERACTIONS				
AB	66,8889			
RESIDUAL				
TOTAL (CORRECTED)	416,278			

1a) Completar la tabla resumen del ANOVA. **[1 punto]**

1b) ¿Qué efectos son estadísticamente significativos, considerando α=0.05? Justificar convenientemente la respuesta. **[1 punto]**

1c) El gráfico de la interacción entre los dos factores se muestra a continuación. ¿Qué información se deduce a la vista de este gráfico, teniendo en cuenta los resultados del ANOVA? **[1 punto]**



1d) A la vista de la figura anterior, dibujar el gráfico de medias del factor procesador. Dibujar también los intervalos LSD, teniendo en cuenta que su anchura total para un nivel de confianza del 95% es de 3.6 unidades (es decir, $\overline{x_i} \pm 1.8$). **[0.5 puntos]**

1e) A la vista del gráfico del apartado d), ¿puede afirmarse que la memoria RAM del procesador tiene un efecto lineal o cuadrático en el tiempo de inversión de estas matrices? Justifica la respuesta. **[0.5 puntos]**

2) Un investigador se plantea diseñar un experimento con todos los factores a 2 niveles, o bien a 3 niveles. ¿Qué ventajas tiene emplear todos los factores a 3 niveles? ¿Qué inconvenientes tiene? **[1 punto]**

3) **[2.5 puntos]** El departamento informático de una compañía de teléfonos móviles quiere obtener un modelo estadístico para predecir la facturación mensual (€) de los clientes que contratan una determinada tarifa (restringida a clientes mayores de 20 años). Se selecciona al azar una muestra representativa de clientes y se obtienen dos variables: edad y tiempo de permanencia en la compañía (años). Los datos se han analizado con Statgraphics, resultando la siguiente tabla:

Multiple Regression Analysis

Dependent variable: FACTURACION

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	45,0561	3,73421	12,0658	0,0000
edad	-0,16612	0,0708991	[illegible]	[illegible]
tiempo_perm	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		

Responde a las siguientes preguntas justificando convenientemente las respuestas:

3a) Un técnico opina que la edad no afecta a la facturación a nivel poblacional, ya que su coeficiente estimado (de valor -0,16612) es pequeño. Considerando α=0.05, ¿es admisible la opinión del técnico? **[1 punto]**

3b) Calcular la facturación media esperada de un cliente de 40 años de edad cuyo tiempo de permanencia en la compañía sea de 3 años. **[0.5 puntos]**

3c) Si un cliente de 40 años de edad lleva 3 años en la compañía, ¿cuál es la probabilidad de que su facturación mensual supere los 45 euros. **[1 punto]**

4) **[2.5 p.]** Dados los siguientes pares de valores: (X=1, Y=5), (X=2, Y=7), (X=4, Y=9), (X=5; Y=11),

a) Calcular la covarianza entre X e Y.

b) La varianza de X vale 10/3. Si se realiza una regresión lineal simple con estos valores, obtener la ecuación matemática del modelo resultante.

c) Calcular el residuo correspondiente a X=1.

SOLUCIÓN

1a) Grados de libertad (GL) totales = nº datos - 1 = 17. GL_A=GL_B = nº variantes - 1 = 3-1 = 2.
GL de la interacción = 2·2 = 4. GL residuales = 17 - 2 - 2 - 4 = 9.
SC_B = 41.722 · 2 = 83.444 ; SC_{res} = 416.28 - 197.44 - 83.44 - 66.89 = 68.5
CM = SC / GL; F-ratio = CM / CM_{res} La tabla completa del ANOVA es la siguiente:

Analysis of Variance for TIEMPO - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:procesador	197,444	2	98,7222	12,97	0,0022
B:algoritmo	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) $F_{ratio_A} \approx F_{2;9}$; $F_{ratio_B} \approx F_{2;9}$; $F_{ratio_{A\cdot B}} \approx F_{4;9}$ Considerando α=0.05, se rechazará la hipótesis nula si la F-ratio obtenida es mayor al valor crítico de tablas: $F_{2;9}^{0.05} = 4.26$; $F_{4;9}^{0.05} = 3.63$

Como la F-ratio del factor procesador (12.97) y la del factor algoritmo (5.48) son mayores a 4.26, se rechaza la hipótesis nula: el efecto simple de ambos factores es estadísticamente significativo.
Como la F-ratio de la interacción: 2.2 < 3.63, no hay suficiente evidencia para afirmar que la interacción doble entre los factores sea estadísticamente significativa.
Nota: los p-valores no se pueden calcular a mano, pero se han incluido también en la tabla.

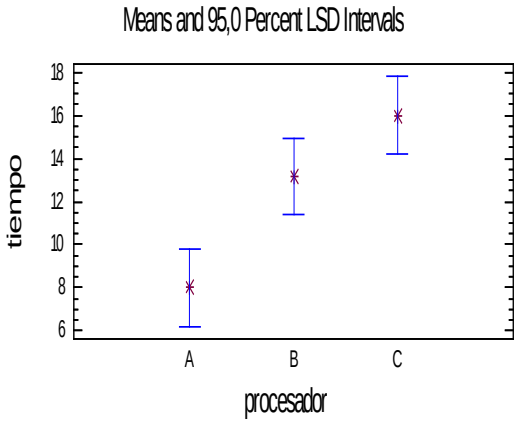
1c) La hipótesis nula asociada al factor procesador es H_0: m_{procA} = m_{procB} = m_{procC}. Según el apartado anterior se rechaza esta hipótesis nula. A la vista de la figura, el tiempo medio de A es menor al de B y éste a su vez inferior al de C. A nivel poblacional, el tiempo medio de A será inferior al de C, pero con la información disponible no está claro si el tiempo medio de B difiere significativamente de los otros dos.

De modo análogo, la hipótesis nula asociada al factor algoritmo es que el tiempo medio de los tres tipos de algoritmo es el mismo. Según el apartado 1b) se rechaza esta hipótesis nula. A la vista de la figura, se deduce que el tiempo medio de AL1 [media=(5+10+13)/3] es inferior al de AL2 [media=(7+17+15)/3] y a su vez inferior al de AL3 [media=(12+12+20)/3)]. Por tanto, puede afirmarse que a nivel poblacional el algoritmo 1 tardará un tiempo significativamente menor al algoritmo 3, pero no está claro si el tiempo medio del algoritmo 2 difiere significativamente de los otros dos.

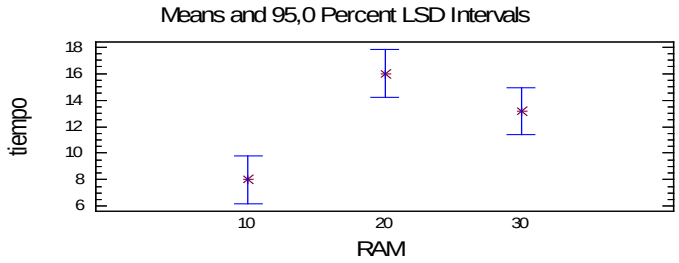
La figura muestra también que las tres líneas (AL1, AL2, AL3) no son paralelas. Sin embargo, la interacción doble entre los dos factores no es estadísticamente significativa. Por tanto, no hay evidencia suficiente para afirmar que el efecto del procesador dependa del tipo de algoritmo, o viceversa.

1d) Los valores que se deducen del gráfico de la interacción se indican en la siguiente tabla, a partir de los cuales se calcula el tiempo medio para cada tipo de procesador. Intervalo LSD para procesador A: 8 ± 1.8 ; procesador B: 13 ± 1.8 ; procesador C: 16 ± 1.8. Los intervalos se dibujan en el siguiente gráfico:

procesador	algoritmo	tiempo	media
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) A la vista del gráfico anterior parece que el efecto de procesador es de tipo lineal pero no es así, pues el procesador A tiene una memoria RAM de 10 MB, el procesador B tiene una RAM de 30 MB y el C tiene una RAM de 20 MB. Si reordenamos el gráfico en función de la memoria RAM queda la figura indicada a continuación. Teniendo en cuenta que el intervalo LSD de 10 RAM no se solapa con los otros dos, y dado que claramente NO es una relación lineal, se deduce que la relación es de tipo cuadrático, alcanzándose un máximo relativo de tiempo con una memoria RAM algo inferior a 20 MB.



2) La ventaja de diseñar un experimento con todos los factores a 3 niveles es que puede estudiarse si el efecto es lineal o cuadrático (en caso de resultar estadísticamente significativo). El inconveniente es que el número total de pruebas que requiere es lógicamente superior que en el caso de factores a dos niveles.

3a) Los grados de libertad (GL) totales que aparecen en el test de significación global del modelo son N-1 (ver formulario) y se calculan como: $GL_{totales}$ = GL del modelo + GL residuales = 98 + 2 = 100 = N-1.
Por tanto, N=101, que es el número de observaciones del modelo.

Dado el modelo $facturación = \beta_0 + \beta_1 \cdot edad + \beta_2 \cdot tiempo$, se aceptará la hipótesis nula $H_0 : \beta_1 = 0$ si se cumple: $\left| b_1 / S_{b_1} \right| < t_{N-1-I}^{\alpha/2}$ En este caso, I=2 (hay dos variables explicativas), N=101,

b_1 = -0.16612 (valor estimado del coeficiente asociado a edad), s_{b1} = 0.0708991 (standard error)

Valor crítico: $t_{N-1-I}^{\alpha} = t_{101-1-2}^{0.025} = t_{98}^{0.025} = 1.99$

$\left| b_1 / S_{b_1} \right| = \left| -0.16612 / 0.0708991 \right| = 2.343$ que es superior al valor crítico 1.99, de modo que se rechaza la hipótesis nula. Por tanto, hay evidencia suficiente para afirmar que el coeficiente β_1 es distinto de cero a nivel poblacional. Así pues, **NO** es admisible la opinión del técnico.

3b) Ecuación de regresión: facturación = 45.0561 - 0.16612·edad + 3.22028 · tiempo
Si edad=40 y tiempo=3: E(facturación)= 45.0561 - 0.16612 · 40 + 3.22028 · 3 = **48.07** eur

3c) Varianza de los residuos = cuadrado medio residual = 107.72
Si edad=40 y tiempo=3: P(facturación>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) $b = \text{cov} / s_x^2 = (14/3)/(10/3) = 1.4$; $a = \bar{y} - b \cdot \bar{x} = 8 - 1.4 \cdot 3 = 3.8$

recta regresión: $y = 3.8 + 1.4 \cdot x$

4c) $residuo_{x=1} = y_{observado} - y_{predicho} = 5 - (3.8 + 1.4 \cdot 1) = 5 - 5.2 =$ **-0.2**