ESTADISTICA II

INGENIERIA INFORMATICA, 3^{ER} Curso

15 - septiembre - 2.008 **Primera Parte - Test**

Apellidos y Nombre: .....

D.N.I. : .....

Nota : En la realización de este examen sólo esta permitido utilizar calculadoras que, a lo sumo, tengan funciones estadísticas básicas. **No se pueden utilizar calculadoras programables.**

Existe una sóla respuesta correcta por pregunta.

Cada respuesta correcta se valorará con 1 punto y cada incorrecta con -1/3. Las preguntas no contestadas no se valoran. Si se marcan varias respuestas a la vez se considerará la pregunta no contestada.

El valor de esta primera parte del examen es de CINCO PUNTOS sobre diez.

Responder con letras mayúsculas y bolígrafo.

Las respuestas elegidas que se considerarán válidas son las que se consignen en el cuadro que se adjunta a continuación.

Pregunta	**1**	**2**	**3**	**4**	**5**	**6**
Respuesta	**A**	**D**	**D**	**C**	**D**	**A**
Pregunta	**7**	**8**	**9**	**10**	**11**	**12**
Respuesta	**B**	**B**	**A**	**D**	**C**	**D**

CUESTIONES

1. Si la probabilidad de error de tipo II en un contraste bilateral es $0'09$, entonces

A. La probabilidad de aceptar H_0 cuando es falso es de $0'09$. **Solución**

B. La probabilidad de aceptar H_0 cuando es falso es de $0'91$.

C. Se rechazará la hipótesis nula el $4'5\%$ de las veces

D. Ninguna de las otras respuestas.

2. Sea $\hat{F}_n$ la distribución empírica asociada a una muestra de tamaño n. Entonces $\hat{F}_n(x)$

A. es una función escalonada con saltos iguales a x_i en los puntos $\frac{i}{n}$.

B. es igual a la distribución teórica a la muestra, $F(x)$, bajo la hipótesis de normalidad.

C. es igual a la recta $y = x$ en el gráfico de normalidad, bajo la hipótesis de normalidad.

D. es igual a $\dfrac{n_i(x)}{n}$, siendo $n_i(x)$ el número de observaciones menores o iguales que x. **Solución**

3. En un diseño completamentamente aleatorizado con un sólo factor

A. Bajo la hipótesis de normalidad el contraste de la F tiene distribución normal.

B. El estimador de la varianza del modelo por máxima verosimilitud es la varianza residual ($\hat{s}^2_R$).

C. Un estimador insesgado de la varianza de la respuesta es la varianza residual ($\hat{s}^2_R$).

D. Bajo la hipótesis de normalidad los estimadores de máxima verosimilitud y de mínimos cuadrados de los efectos coinciden. **Solución**

4. Se ha observado la duración de diez llamadas telefónicas y se ha obtenido

$2'1$	$7'5$	$6'2$	$7'0$	$4'4$	$3'6$	$5'9$	$8'8$	$4'0$	$1'0$

Para contrastar si la duración de una llamada sigue una distribución uniforme $(0; 10)$, se ha aplicando el contraste de Kolmogorov-Smirnov. Se obtiene

A. Se rechaza la hipótesis para $0'10 < \alpha < 0'20$.

B. Se rechaza la hipótesis para $\alpha < 0'10$.

C. Se acepta la hipótesis para $\alpha < 0'20$. **Solución**

D. Ninguna de las otras tres respuestas.
Solución

x	$F(x)$	$F_n(x_{i-1})$	$F_n(x_i)$	
$1'0$	0'10	0'00	0'10	0'10
$2'1$	0'21	0'10	0'20	0'11
$3'6$	0'36	0'20	0'30	0'16
$4'0$	0'40	0'30	0'40	0'10
$4'4$	0'44	0'40	0'50	0'06
$5'9$	0'59	0'50	0'60	0'09
$6'2$	0'62	0'60	0'70	0'08
$7'0$	0'70	0'70	0'80	0'10
$7'5$	0'75	0'80	0'90	0'15
$8'8$	0'88	0'90	1'00	0'12
				KS=0'16

$$KS = 0'16 \Rightarrow p - valor > 0'20$$

5. En un diseño de experimentos hay un factor que se supone que tiene poca influencia en la respuesta y no se ha introducido en el modelo, entonces

A. Los resultados de este diseño de experimentos serán malos.

B. El factor no influye en los resultados si es fijo pero su influencia puede ser fuerte si es un factor aleatorio.

C. El factor no influye o influye poco.

D. Se debe aleatorizar la aplicación de los niveles de este factor. **Solución**

6. Sea C_s el coeficiente de asimetría estandarizado y C_p el coeficiente de apuntamiento estandarizado de los residuos de un modelo de diseño de experimentos. Entonces el estadístico $d = C_S^2 + C_P^2$

A. Se utiliza para contrastar la normalidad del modelo, sigue una distibución chi-cuadrado. **Solución**

B. Se utiliza para contrastar la normalidad del modelo, sigue una distibución normal.

C. Se utiliza para contrastar la independencia del modelo, sigue una distibución chi-cuadrado.

D. Se utiliza para contrastar la independencia del modelo, sigue una distibución normal.

7. En un modelo de diseño de experimentos con un factor aleatorio (α), se verifica que

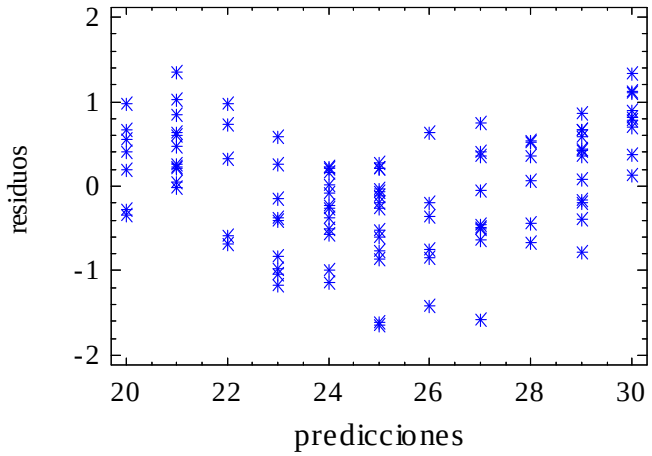
A. El factor aleatorio (α) influye en la respuesta media.

B. Los efectos del factor (α) son variables aleatorias de las que deseamos contrastar si su varianza es nula. **Solución**

C. Los efectos del factor (α) son parámetros desconocidos que queremos estimar.

D. Las respuestas de un mismo nivel son independientes.

8. Se estudia un modelo de diseño de experimentos con dos factores sin interacción. En la figura adjunta se representa el gráfico de los residuos frente a las predicciones del modelo ajustado. Se observa



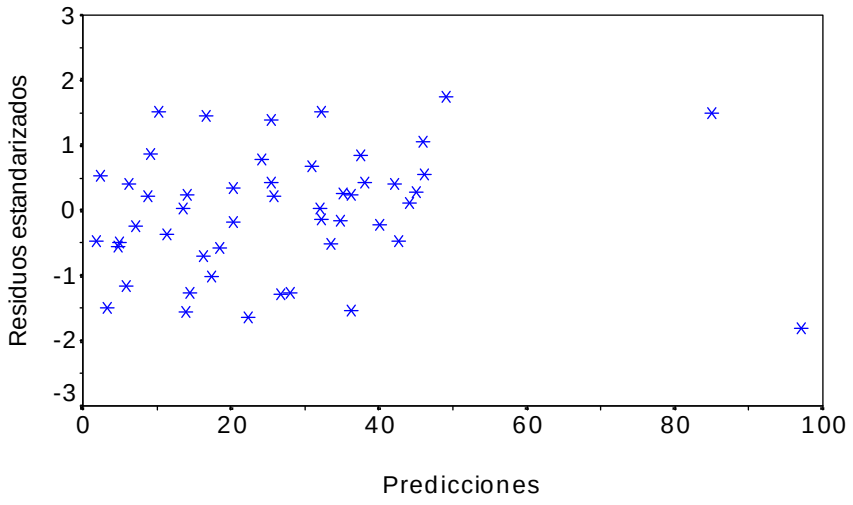
A. Homocedasticidad y falta de normalidad.

B. Homocedasticidad y la existencia de interacción entre los factores. **Solución**

C. Existe normalidad pero hay dependencia positiva.

D. Ninguna de las otras respuestas.

9. En la figura adjunta se representa el gráfico de residuos estandarizados frente a las predicciones en un modelo de regresión lineal múltiple con cinco variables regresoras. Se observa



A. Que el ajuste es adecuado, no se observa ningún problema. **Solución**

B. La existencia de observaciones influyentes a posteriori.

C. Falta de normalidad por tener una curtosis alta.

D. La existencia de datos atípicos.

10. En un modelo de regresión lineal múltiple se verifica que $X^t\hat{Y}$ es igual a

A. $\left(X^tX\right)^{-1}X^tY$

B. $X\left(Y-\hat{Y}\right)$

C. $Y^t{}_{\varepsilon}$

D. X^tY **Solución.**

Solución.

$$X^t\hat{Y} = X^t \cdot X\hat{\alpha} = X^t \cdot X \cdot \left(X^tX\right)^{-1}X^tY = \left[\left(X^tX\right)\left(X^tX\right)^{-1}\right]X^tY = X^tY$$

11. Al ajustar un modelo de regresión lineal múltiple con cuatro regresoras a una muestra de 101 observaciones, se ha obtenido que el estadístico del contraste conjunto de regresión es $\hat{F}_M = 20$. Entonces se obtiene que

A. $\bar{R}^2 = 0,835$

B. $\bar{R}^2 = 0,151$

C. $\bar{R}^2 = 0,432$ **Solución**

D. Ninguna de las otras tres respuestas.

Solución:

$$\begin{aligned}
\hat{F}_M &= 20 = \frac{\hat{s}^2_{\text{exp}}}{\hat{s}^2_R} = \frac{scE/4}{scR/96} = \frac{scE}{scR} \cdot \frac{96}{4} \Longrightarrow \frac{scE}{scR} = 20 \cdot \frac{4}{96} \\
\bar{R}^2 &= 1 - \frac{\hat{s}^2_R}{\hat{s}^2_Y} = 1 - \frac{scR/96}{scT/100} = 1 - \frac{scR}{scT} \cdot \frac{100}{96} \\
\frac{scT}{scR} &= 1 + \frac{scE}{scR} = 1 + 20 \cdot \frac{4}{96} = \frac{176}{96} \\
\bar{R}^2 &= 1 - \frac{96}{176} \cdot \frac{100}{96} = 0.431\,82
\end{aligned}$$

12. En un modelo de regresión lineal simple, la varianza de la predicción en un punto x_t

A. Ninguna de las otras tres respuestas.

B. Aumenta al aumentar el número de observaciones equivalentes (n_t).

C. Es inversamente proporcional a la varianza del modelo.

D. Toma el valor mínimo en el punto $x_t = \bar{x}$. **Solución**

ESTADISTICA II, Ingeniería Informática,
Problemas, 15 - septiembre - 2.008

Cada una de los dos problemas tiene una valoración de 2.5 puntos sobre diez.

Para aprobar el examen es necesario obtener una puntuación igual o superior a 1 punto en cada uno de los dos problemas.

Problema 1.

En un estudio (e.g. Hix y Hartson, 1986) para evaluar la utilidad de un entorno interactivo para el diseño de cuadros de diálogo (user-interface management system) frente al método tradicional de escribir el código fuente, se midió el tiempo (en horas) de creación y modificación de interfaces empleando distintos métodos (UIMS, Programación, UIMS+Programación). Los tiempos (en horas) obtenidos, correspondientes a K=2 programadores expertos en cada combinación de niveles, se muestran a continuación:

	Método		
Tarea	Programación	UIMS	UIMS+Programación
Creación	1.9	0.8	0.7
	2.4	0.9	0.8
Modificación	0.8	0.4	0.2
	0.9	0.6	0.6

$\Sigma y_{ij} = 11$
$\Sigma y_{ij}^2 = 14.320$

P.1.Formular el modelo de diseño de experimentos asociado a este problema y obtener las estimaciones puntuales de los efectos de los factores. *(0.6 puntos)*

Solución:

Se trata de un diseño completamente aleatorizado con dos factores tratamiento con $I = 3$ y $J = 2$ niveles, respectivamente, y se supone que hay interacción entre ambos factores.

La formulación matemática del modelo es la siguiente:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

con ε_{ijk} v.a. independientes con distribución $N\left(0, \sigma^2\right)$.

Para estimar los parámetros, calculamos las medias en cada casilla, por filas y por columnas:

Medias				
$\overline{y}_{ij.}$	M_1	M_2	M_3	$\overline{y}_{.j.}$
$T1$	2.15	0.85	0.75	1.25
$T2$	0.85	0.5	0.4	0.583
$\overline{y}_{i..}$	1.5	0.675	0.575	$\overline{y}_{...}$ =0.917

Calculamos ahora las estimaciones de los efectos:

$$\begin{aligned} \widehat{\alpha}_i &= \overline{y}_{i..} - \overline{y}_{...}, \\ \widehat{\beta}_j &= \overline{y}_{.j.} - \overline{y}_{...}, \\ \left(\widehat{\alpha\beta}\right)_{ij} &= \overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...} \end{aligned}$$

Parámetros				
$\left(\widehat{\alpha\beta}\right)_{ij}$	M_1	M_2	M_3	$\widehat{\beta}_j$
$T1$	0.317	-0.159	-0.159	0.334
$T2$	-0.317	0.159	0.159	-0.334
$\widehat{\alpha}_i$	0.584	-0.242	-0.342	

P.2. Obtener la tabla ANOVA, realizando los contrastes (con un nivel de significación del 5%) y calculando los coeficientes de determinación correspondientes. *(0.8 puntos)*

Solución:
Las sumas de cuadrados son:

$$\begin{aligned}
scT_\alpha &= 4\sum_{i=1}^{3}\widehat{\alpha}_i^2 = 4\times\left(0.584^2+0.242^2+0.342\right) = 2.066\\
scT_\beta &= 6\sum_{j=1}^{2}\widehat{\beta}_j^2 = 6\times 2\times 0.334^2 = 1.339\\
scT_{\alpha\beta} &= 2\sum_{i=1}^{3}\sum_{j=1}^{2}\left(\widehat{\alpha\beta}\right)_{ij}^2 = 2\times 2\times\left(0.317^2+0.159^2+0.159^2\right) = 0.604\\
scG &= \sum_{ijk}(y_{ijk}-\overline{y}_{...})^2 = \sum_{ijk}y_{ijk}^2 - n\overline{y}_{...}^2 =\\
&= 14.320 - 12\times 0.917^2 = 4.229\\
scR &= \sum_{ijk}e_{ijk}^2 = scG - scT_\alpha - scT_\beta - scT_{\alpha\beta}\\
&= 4.229 - 2.066 - 1.339 - 0.604 = 0.22
\end{aligned}$$

y la tabla ANOVA:

F. var.	**sc**	**gl**	**SCM**	**F**	**p-valor**
Método	2.066	2	1.033	28.17	$p < P\left(F_{2,6}\geq 14.54\right) = 0.005$
Tarea	1.339	1	1.339	36.52	$p < P\left(F_{1,6} > 18.63\right) = 0.005$
Interacción	0.604	2	0.302	8.24	$0.01 < p < 0.025$
Residual	0.22	6	0.037		
Global	4.229	11	0.384		

Se acepta que hay interacción, el método y el tipo de tarea influyen significativamente en el tiempo (medio), el efecto de uno de los factores depende del nivel del otro factor (el efecto del método depende del tipo de tarea).

Los coeficientes de determinación son:

$$
\begin{aligned}
R^2\,(\text{"Método"}) &= \frac{2.066}{4.229} = 0.489 \\
R^2\,(\text{"Tarea"}) &= \frac{1.339}{4.229} = 0.317 \\
R^2\,(\text{"Interacción"}) &= \frac{0.604}{4.229} = 0.143 \\
R^2\,(\text{"Total"}) &= 1 - \frac{0.22}{4.229} = 0.948
\end{aligned}
$$

El modelo explica un 94.8% de la variabilidad de la respuesta.

P.3. Obtener una estimación puntual y por intervalo de confianza al 95% de la varianza del modelo. *(0.4 puntos)*

Solución:
Estimación puntual:

$$\hat{\sigma}^2 = \hat{S}_R^2 = 0.037$$

Estimación por intervalo de confianza:

$$I_{0.95}(\sigma^2) = \left(\frac{scR}{\chi^2_{6,0.975}}, \frac{scR}{\chi^2_{6,0.025}} \right) = \left(\frac{0.22}{14.449}, \frac{0.22}{1.2373} \right) = (0.015, 0.178)$$

P.4. Obtener estimaciones puntuales y por intervalo de confianza al 95% para las diferencias de tiempos medios con los distintos métodos. *(0.7 puntos)*

Solución:
Nos piden estimaciones puntuales y por intervalos de confianza de los estadísticos $\theta_{ik} = \mu_{i.} - \mu_{k.}$.La estimación puntual será:

$$\hat{\theta}_{ik} = \hat{\mu}_{i.} - \hat{\mu}_{k.} = \overline{y}_{i..} - \overline{y}_{k..}.$$

En el calculo de los intervalos de confianza, si se utiliza el método de Scheffé el valor crítico sería:

$$\omega_S = \sqrt{(I-1)\, F_{I-1,IJ(K-1),(1-\alpha)}} = \sqrt{2 \times F_{2,6,0.95}} = \sqrt{2 \times 5.14} = 3.206.$$

Si se utiliza el método DMS (LSD) el valor crítico sería:

$$t_{6,0.975} = 2.4469.$$

Como se trata de un diseño equilibrado en todos los casos el error estandar es el mismo:

$$\sigma\left(\hat{\theta}_{ik}\right) = \hat{S}_R \sqrt{\frac{2}{4}} = \sqrt{\frac{0.037}{2}} = 0.136.$$

- $\mathbf{M_1 - M_2}$:

$$
\begin{aligned}
\overline{y}_{1..} - \overline{y}_{2..} &= 1.5 - 0.675 = 0.825 \\
I_{0.95}^{DMS}(\mu_{1.} - \mu_{2.}) &= (0.825 \pm 2.4469 \times 0.136) = (0.492, 1.158) \\
I_{0.95}^{Scheffé}(\mu_{1.} - \mu_{2.}) &= (0.825 \pm 3.206 \times 0.136) = (0.389, 1.261)
\end{aligned}
$$

- $\mathbf{M_1 - M_3}$:

$$\begin{aligned} \overline{y}_{1..} - \overline{y}_{3..} &= 1.5 - 0.575 = 0.925 \\ I_{0.95}^{DMS}(\mu_{1.} - \mu_{3.}) &= (0.925 \pm 2.4469 \times 0.136) = (0.592, 1.258) \\ I_{0.95}^{Scheffé}(\mu_{1.} - \mu_{3.}) &= (0.925 \pm 3.206 \times 0.136) = (0.48\,9, 1.\,361) \end{aligned}$$

- $\mathbf{M_2 - M_3}$:

$$\begin{aligned} \overline{y}_{2..} - \overline{y}_{3..} &= 0.675 - 0.575 = 0.1 \\ I_{0.95}^{DMS}(\mu_{2.} - \mu_{3.}) &= (0.1 \pm 2.4469 \times 0.136) = (-0.233\,, 0.433) \\ I_{0.95}^{Scheffé}(\mu_{2.} - \mu_{3.}) &= (0.1 \pm 3.206 \times 0.136) = (-0.336, 0.536) \end{aligned}$$

ESTADISTICA II, Ingeniería Informática,
Problemas, 15 - septiembre - 2.008

Problema 2:

En una clase de 30 estudiantes se realiza un estudio para determinar la relación existente entre la variable $Y =$ *"Nota obtenida por el estudiante en la asignatura de Estadística"* y las variables $X_1 =$ *"Nota esperada por el estudiante"*, $X_2 =$ *"Horas semanales de estudio dedicadas a la asignatura"* y $X_3 =$ *"Nota media del estudiante en las demás asignaturas"*.

En una primera etapa se estudia la relación lineal existente entre la variable Y y la variable X_2. Para ello se tienen en cuenta los siguientes datos:

$$\sum_{i=1}^{30} x_{i2} = 196 \qquad \sum_{i=1}^{30} x_{i2}^2 = 1600$$

$$\sum_{i=1}^{30} y_i = 167 \qquad \sum_{i=1}^{30} y_i^2 = 1161.25$$

$$\sum_{i=1}^{30} y_i x_{i2} = 1355.5$$

P.5. Estima los coeficientes y obtén el modelo de regresión lineal correspondiente a esta primera etapa. *(Valor 0.50)*

Solución:

El modelo ajustado es:

$$\hat{Y} = \hat{\alpha_0} + \hat{\alpha_1} X_2$$

donde

$$\begin{aligned} \hat{\alpha_1} &= \frac{S_{X_2Y}}{S_{X_2}^2} = \frac{\frac{1355.5}{30} - \frac{167}{30}\frac{196}{30}}{\frac{1600}{30} - \left(\frac{196}{30}\right)^2} = 0.827733722 \\ \text{y} & \\ \hat{\alpha_0} &= \bar{Y} - \hat{\alpha_1}\bar{X_2} = \frac{167}{30} - \hat{\alpha_1}\frac{196}{30} = 0.158806349. \end{aligned}$$

P.6. Haz el contraste de regresión (tabla ANOVA) del modelo obtenido en esta primera etapa e interprétalo. *(Valor 0.50)*

Solución:

$$\begin{aligned} scR &= \sum_{i=1}^{30} e_i^2 = \sum_{i=1}^{30} y_i^2 - \hat{\alpha_0}\sum_{i=1}^{30} y_i - \hat{\alpha_1}\sum_{i=1}^{30} y_i x_{i2} \\ &= 1161.25 - \hat{\alpha_0}167 - \hat{\alpha_1}1355.5 = 12.73627955. \end{aligned}$$

$$scG = \sum_{i=1}^{30} \left(y_i - \bar{Y}\right)^2 = \sum_{i=1}^{30} {y_i}^2 - 30\bar{Y}^2$$
$$= 1161.25 - \frac{(167)^2}{30} = 231.6166667.$$

Fuentes de variación	Suma de cuadrados	Grados de libertad	Varianzas
Modelo	218.8803872	1	$\widehat{S}_E^2 = \frac{scE}{1} = 218.8803872$
Residual	12.73627955	28	$\widehat{S}_R^2 = \frac{scR}{28} = 0.454867126$
Global	231.6166667	29	$\widehat{S}_G^2 = \frac{scG}{29} = 7.98678161$

Contraste de regresión:

$$\begin{cases} H_0: \text{ El modelo no es significativo, i.e, } \alpha_1 = \alpha_2 = 0 \\ H_1: \text{ El modelo es significativo} \end{cases}$$

Calculamos el estadístico asociado al contraste de regresión:

$$\widehat{F}_R = \frac{\widehat{S}_E^2}{\widehat{S}_R^2} = 481.1963202 \sim F_{1,28} \text{ (si } H_0 \text{ es cierta)}$$

$p - valor = p(F_{1,28} \geq 481.1963202) \simeq 0 \Rightarrow$ Rechazamos H_0, i.e., el modelo es significativo.

P.7. Obtén el intervalo de predicción al 95% **para la nota de un alumno que dedica** 8 **horas semanales al estudio de esta asignatura.** *(Valor 0.50)*

Solución:
El intervalo de predicción para y_i viene dado por

$$\left[\widehat{y}_i - \widehat{S}_R\sqrt{1+h_{ii}}t_{28}(0.975), \widehat{y}_i + \widehat{S}_R\sqrt{1+h_{ii}}t_{28}(0.975)\right].$$

$$\widehat{y}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 8 = 0.159 + 80.828 = 6.780676125 \text{ y } t_{28}(0.975) = 2.0484$$
$$\widehat{S}_R^2 = 0.454867126 \Rightarrow \widehat{S}_R = 0.6744$$
$$h_{ii} = \frac{1 + \left(\frac{8 - \bar{X}}{S_X}\right)^2}{n} = 0.0483.$$

De esta forma, sustituyendo obtenemos que el intervalo de predicción al 95% para la nota de ese alumno viene dado por

$$[5.366, \quad 8.195].$$

P.8. En una segunda etapa añadimos las variables regresoras X_1 y X_3.
En este caso obtenemos el siguiente modelo ajustado, donde los errores estándar aparecen entre paréntesis:

$$\begin{array}{rcccc} \hat{Y} = & -0.58 & +0.2X_1 & +0.6X_2 & +0.23X_3 \\ & (0.325) & (0.07) & (0.074) & (0.092) \end{array}$$

Estudia qué variables regresoras de las utilizadas en esta segunda etapa son significativas al 95%**.** *(Valor 0.50)*

Solución:
Se obtiene

$\widehat{\alpha}_i$	-0.58	0.2	0.6	0.23
$\sigma(\widehat{\alpha}_i)$	0.325	0.07	0.074	0.092
$\widehat{t}_i = \dfrac{\widehat{\alpha}_i}{\sigma(\widehat{\alpha}_i)}$	-1.784	2.857	8.108	2.5

Para cada coeficiente α_i consideramos el siguiente contraste de hipótesis: $\begin{cases} H_0 : \alpha_i = 0 \\ H_1 : \alpha_i \neq 0 \end{cases}$
Sabemos que $\widehat{t}_i \sim t_{26}$ (si H_0 es cierta). Sabemos además que $t_{26}(0.975) = 2.0555$.
Por tanto, para $i = 1, 2, 3$

$$|\widehat{t}_i| > 2.0555 \Rightarrow p - valor = 2p(t_{26} > |\widehat{t}_i|) < 0.05 \Rightarrow$$

Las tres variables regresoras son significativas.

P.9. Calcula los coeficientes de correlación parcial y simple entre las variables Y y X_2. *(Valor 0.50)*

Solución:
Coeficiente de correlación parcial:

$$r^2_{(Y,X_2).(X_1,X_3)} = \frac{\hat{t_2}^2}{\hat{t_2}^2 + (n-4)} = \frac{0.8108^2}{0.8108^2 + 26} = 0.716594741 \Rightarrow r_{(Y,X_2).(X_1,X_3)} = 0.846519191.$$

Coeficiente de correlación simple:

$$r_{X_2Y} = \frac{S_{X_2Y}}{S_{X_2} S_Y} = 0.972116961.$$