ESTADISTICA II

INGENIERIA INFORMATICA, 3ER Curso

27 - Junio - 2.006 **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	**C**	**A**	**D**	**B**	**B**	**D**
Pregunta	**7**	**8**	**9**	**10**	**11**	**12**
Respuesta	**C**	**B**	**C**	**D**	**C**	**A**

CUESTIONES

1. En el contraste de $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$ se obtiene que $Potencia\,(\theta_0) = 0.06$. Entonces

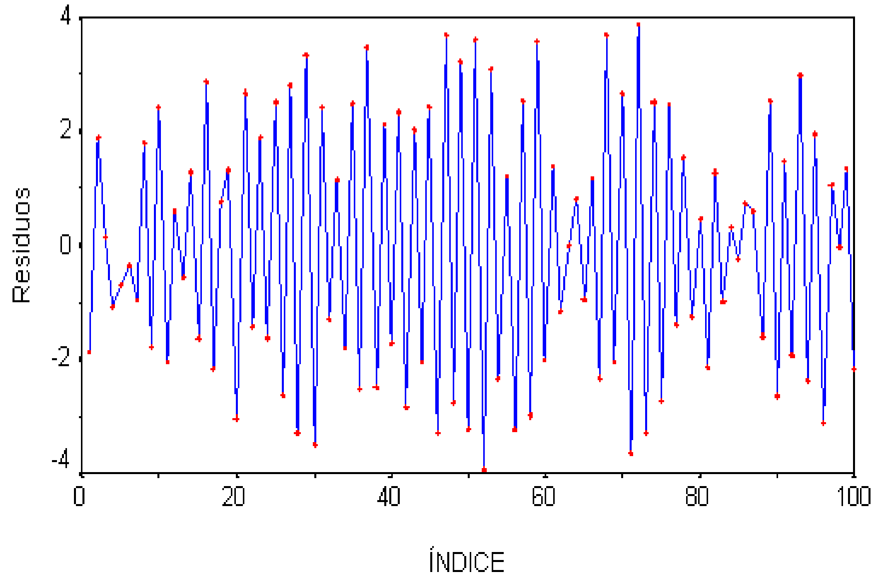
A. La probabilidad de error de tipo 2 es 0.06

B. $P(aceptar\ H_0/\theta = \theta_1) = 0.94$

C. La probabilidad de error de tipo 1 es 0.06 **Solución**

D. $P(aceptar\ H_0/\theta = \theta_0) = 0.06$

2. En el ajuste de un modelo estadístico el gráfico de residuos frente al índice (orden de entrada de los datos) es el siguiente. En este gráfico se observa



A. Que existe dependencia negativa. **Solución**

B. Que existe dependencia positiva.

C. Que existe independencia y homocedasticidad.

D. Que existe independencia pero hay heterocedasticidad.

3. En el ajuste de un modelo estadístico, al hacer el contraste de Box-Ljung para las seis primeras autocorrelaciones ($m = 6$) de los residuos del modelo se obtiene un estadístico $Q_{BL} = 10$. Por tanto

A. Se acepta que los residuos son independientes.

B. Se acepta con $\alpha = 0.10$ que las seis primeras autocorrelaciones de los residuos son cero.

C. Se rechaza que los residuos son independientes.

D. Se acepta con $\alpha = 0.05$ que las seis primeras autocorrelaciones de los residuos son cero. **Solución**

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

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

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

5. Para contrastar la igualdad de las medias de dos muestras independientes de igual varianza se utiliza el test de la t. El $p - valor$ que se obtiene es

A. Igual al que proporciona el test de la F en un diseño de experimentos con dos factores, uno fijo y uno aleatorio.

B. Igual al que proporciona el test de la F en un diseño de experimentos con un factor completamente aleatorizado. **Solución**

C. Igual al que proporciona el test de la F en un diseño de experimentos en bloques completamente aleatorizados.

D. Igual al que proporciona el test de la F en un diseño de experimentos con dos factores e interacción.

6. El contraste de Kruskal-Wallis se utiliza para contrastar

A. Que los residuos de un diseño de experimentos con un factor son homocedásticos.

B. Que los residuos de un diseño de experimentos con un factor son heterocedásticos.

C. Que los residuos de un diseño de experimentos con un factor son independientes.

D. Que en un diseño de experimentos con un factor las medias de la respuesta condicionada a los niveles son iguales. **Solución**

7. Al ajustar un modelo de diseño de experimentos con un factor aleatorio: $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, 2, 3, 4, \quad j = 1, 2, \ldots, 10$. Se ha obtenido que $scT = 252$ (Total) y $scR = 72$ (residual). Entonces la estimación de $Var(\tau_i) = \sigma_\tau^2$ es

A. 4.92

B. 2.0

C. 5.8 **Solución**

D. Ninguna de las otras tres respuestas.

8. En el diseño de experimentos cuadrado latino, donde los factores tienen K niveles, el número de restricciones que verifican los residuos es

A. $K^2 - 1$

B. $3K - 2$ **Solución**

C. $K^2 - 3K + 2$

D. Ninguna de las otras tres respuestas.

9. En el modelo de regresión lineal múltiple $\vec{Y} = X^t\vec{\alpha} + \vec{\varepsilon}$. Se verifica que

A. el vector $H\vec{Y}$ es ortogonal a $\vec{\varepsilon}$

B. el vector $H\vec{Y}$ es ortogonal a $H\hat{Y}$

C. el vector $H\vec{Y}$ es ortogonal a $\vec{Y} - \hat{Y}$ **Solución**

D. el vector $H\vec{Y}$ es ortogonal a $\hat{Y}$

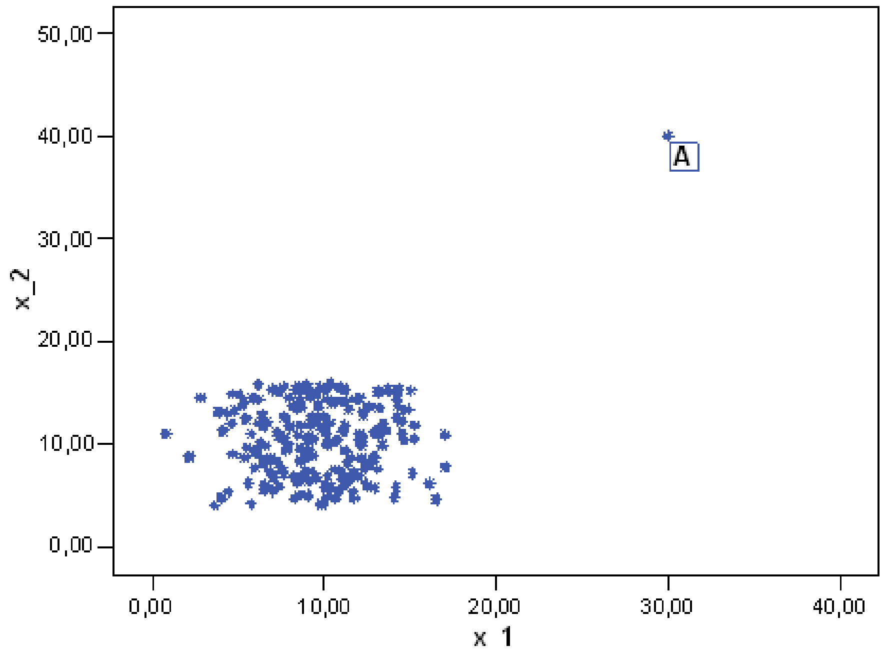
10. En el ajuste de un modelo de regresión lineal múltiple con dos regresoras, el gráfico de dispersión de las dos regresoras (x_1 y x_2) es el de la figura adjunta. De este gráfico se deduce

A. El punto A es influyente y atípico.

B. El punto A es influyente a priori pero no a posteriori.

C. El punto A origina problemas de heterocedasticidad y falta de normalidad.

D. El punto A origina multicolinealidad entre las dos regresoras. **Solución**



11. En el modelo de regresión lineal simple, la hipótesis de linealidad afirma que

A. $\hat{Y}_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i$

B. $E(Y_i/x_i) = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$

C. $E(Y_i/x_i) = \alpha_0 + \alpha_1 x_i$ **Solución**

D. $Y_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$

12. Al ajustar un modelo de regresión lineal múltiple con tres regresoras a una muestra de 40 observaciones, se ha obtenido que el estadístico del contraste conjunto de regresión es $\hat{F}_M = 10$. Entonces se obtiene que

A. $R^2 = \dfrac{1}{2.2}$ **Solución**

B. $R^2 = \dfrac{12}{23}$

C. $R^2 = \dfrac{11}{36}$

D. Ninguna de las otras tres respuestas.

ESTADISTICA II, Ingeniería Informática,

Problemas, 27 - Junio - 2.006

Apellidos, Nombre: ....

D.N.I.: ....

Responder de forma concisa y justificada a las siguientes cuestiones.
Las respuestas se escribirán con bolígrafo a continuación de las preguntas.
Cada una de las preguntas tiene una valoración máxima de 0.5 puntos sobre cinco.
Para aprobar el examen es necesario obtener una puntuación igual o superior a 1 punto en cada uno de los dos problemas.

Problema 1. *Para mejorar la velocidad de un procesador, se quiere saber el efecto, en el tiempo de realización de un determinado código, de la temperatura de funcionamiento (normal: 50º, alta:70º o muy alta: 100º) y del tipo de circuito (C1, C2 y C3). Los resultados del experimento son los siguientes:*

		Factor C		
		$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_3$
Factor T	$\mathbf{T}_1$	0.62 0.77	0.34 0.30	0.35 0.38
	$\mathbf{T}_2$	0.55 0.81	0.50 0.48	0.55 0.61
	$\mathbf{T}_3$	0.6 0.82	0.52 0.54	0.62 0.68

$\Sigma y_i = 10.04$

P.1. Formular el modelo de un diseño de experimentos con dos factores e interacción, y detallar todas las suposiciones que se hacen en él. Calcular las estimaciones de los efectos T, C e interacción.

Solución:
El modelo es $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, 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				
$\overline{y}_{ij\cdot}$	C_1	C_2	C_3	$\overline{y}_{i\cdot\cdot}$
T_1	0.695	0.32	0.365	0.46
T_2	0.68	0.49	0.58	0.583
T_3	0.71	0.53	0.65	0.63
$\overline{y}_{\cdot j\cdot}$	0.695	0.447	0.528	$\overline{y}_{\cdots} = \frac{10.04}{18} = 0.556$

Calculamos ahora los parámetros:

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

Parámetros				
$\left(\widehat{\alpha\beta}\right)_{ij}$	C_1	C_2	C_3	$\widehat{\alpha}_i$
T_1	0.096	-0.031	-0.067	-0.096
T_2	-0.042	0.016	0.025	0.027
T_3	-0.059	0.009	0.048	0.074
$\widehat{\beta}_j$	0.139	-0.109	-0.028	

P.2. Completar la tabla ANOVA e indicar qué efectos que son significativos y cuáles no (nivel de significación 5%).

Solución:
Las sumas de cuadrados son:

$$\begin{aligned} VE(C) &= \Sigma\Sigma\Sigma\widehat{\beta}_j^2 = 2 \times 3 \times \left(0.139^2 + 0.109^2 + 0.028^2\right) = 0.1911 \\ VE(T) &= \Sigma\Sigma\Sigma\widehat{\alpha}_i^2 = 2 \times 3 \times \left(0.096^2 + 0.027^2 + 0.074^2\right) = 0.0926 \\ VR &= \Sigma e_{ijk}^2 = VG - VE(CT) - VE(C) - VE(T) \\ &= 0.4045 - 0.04629 - 0.1911 - 0.0926 = 0.0745 \end{aligned}$$

F. var.	sc	gl	SCM	F	p-valor	¿significativo?
Factor C	0.1911	2	0.0956	11.5	0.0033	SÍ
Factor T	0.0926	2	0.04629	5.59	0.0264	SÍ
Interacción	**0.04629**	4	0.01157	1.40	0.3096	NO
Residual	0.0745	9	0.00828			
Global	**0.4045**	17	0.02379			

P.3. Contrastar, con $\alpha = 0.05$, si el promedio de los efectos de los niveles C1 y C2 coincide con el efecto del nivel C3 del factor C.

Solución:
Las hipótesis a contrastar son

$$\begin{array}{l} H_0 : \frac{1}{2}\beta_1 + \frac{1}{2}\beta_2 = \beta_3 \\ H_1 : \frac{1}{2}\beta_1 + \frac{1}{2}\beta_2 \neq \beta_3 \end{array}, \quad \text{es decir,} \quad \begin{array}{l} H_0 : \theta = \frac{1}{2}\beta_1 + \frac{1}{2}\beta_2 - \beta_3 = \sum_{j=1}^{J=3} b_j\beta_j = 0 \\ H_1 : \theta = \frac{1}{2}\beta_1 + \frac{1}{2}\beta_2 - \beta_3 = \sum_{j=1}^{J=3} b_j\beta_j \neq 0 \end{array}$$

El estimador mínimo-cuadrático de θ es

$$\widehat{\theta} = \sum_{j=1}^{J} b_j \widehat{\beta}_j = \frac{1}{2}\widehat{\beta}_1 + \frac{1}{2}\widehat{\beta}_2 - \widehat{\beta}_3 = \frac{1}{2}0.139 - \frac{1}{2}0.109 + 0.028 = 0.043 = \sum_{j=1}^{J} b_j \overline{y}_{\cdot j \cdot}$$

Sabiendo que

$$\widehat{\theta} = \sum_{j=1}^{J} b_j \widehat{\beta}_j \in N\left(\theta = \sum_{j=1}^{J} b_j \beta_j, \sigma^2 \sum_{j=1}^{J} \frac{b_j^2}{n_j}\right) \text{ entonces } \frac{\sum_{j=1}^{J} b_j \widehat{\beta}_j - \sum_{j=1}^{J} b_j \beta_j}{\widehat{s}_R \sqrt{\sum_{j=1}^{J} \frac{b_j^2}{n_j}}} \in t_{IJ(K-1)}$$

El valor del estadístico de contraste es

$$\frac{0.043}{\sqrt{0.00828}\sqrt{\frac{\frac{1}{4}}{6} + \frac{\frac{1}{4}}{6} + \frac{1}{6}}} = 0.945\,11$$

y por tanto el p-valor es

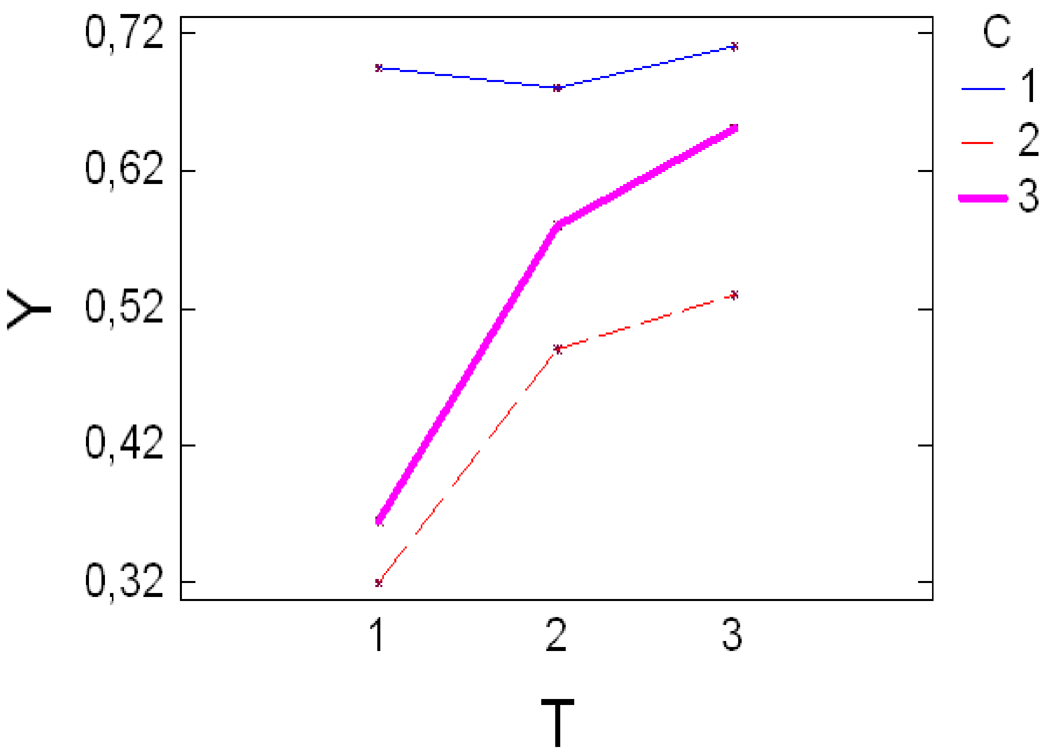
$$p - valor = 2P\left(t_9 \geq 0.94511\right) = 2 \times 0.184642 = 0.369\,3 \Rightarrow \text{ Acepto } \mathbf{H}_0.$$

P.4. Dibujar el gráfico de interacciones poniendo en el eje de abscisas el factor T. Interpretar el resultado.

Solución:

Teniendo en cuenta la tabla de medias

Medias			
$\overline{y}_{ij\cdot}$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_3$
$\mathbf{T}_1$	0.695	0.32	0.365
$\mathbf{T}_2$	0.68	0.49	0.58
$\mathbf{T}_3$	0.71	0.53	0.65



P.5. Calcular la tabla ANOVA del modelo de dos factores sin interacción. En este caso ¿qué factores son significativos? Calcular el coeficiente de determinación de este modelo.

Solución:

F. var.	**sc**	**gl**	**SCM**	**F**	**p-valor**	**¿significativo?**
Factor C	0.1911	2	0.0956	10.29	0.0021	SÍ
Factor T	0.0926	2	0.04629	4.983	0.0248	SÍ
Residual	0.1208	13	0.00929			
Global	**0.4045**	17	0.02379			

No se puede eliminar ningún factor del análisis.
El coeficiente de determinación es

$$\begin{aligned} R^2 &= \frac{scE}{scG} = \frac{scE\,(C)}{scG} + \frac{scE\,(T)}{scG} \\ &= \frac{0.1911}{0.4045} + \frac{0.0926}{0.4045} = 0.472\,4 + 0.228\,9 = \mathbf{0.701\,36} \end{aligned}$$

ESTADISTICA II, Ingeniería Informática,

Problemas, 27 - Junio - 2.006

Apellidos, Nombre: ....

D.N.I.: ....

Problema 2:

En base a una muestra de 22 observaciones se quiere estudiar la relación lineal que existe entre una variable respuesta (Y) *y tres variables regresoras* $(X_1, X_2$ y $X_3)$. *En una primera etapa se estudia la relación lineal simple entre la respuesta* (Y) *y una regresora* (X_1).*Se obtienen los siguientes datos*

$$\sum_{i=1}^{22} x_{i1} = 91 \qquad \sum_{i=1}^{22} x_{i1}^2 = 521$$
$$\sum_{i=1}^{22} y_i = -69 \qquad \sum_{i=1}^{22} y_i^2 = 1091$$
$$\sum_{i=1}^{22} y_i \cdot x_{i1} = 9'84$$

En todos los casos en que se pida estudiar la significatividad de un contraste utilizar $\alpha = 0'05$

P.6. Calcular la predicción para la observación muestral 22, cuyos datos son: $(x_{22,1} = 10; \quad y_{22} = 18)$.

Solución:

$$\hat{\alpha}_1 = \frac{s_{1y}}{s_1^2} = \frac{13.42}{6.572} = 2.042$$

$$\begin{aligned} s_1^2 &= \overline{x_1^2} - (\bar{x}_1)^2 = \frac{521}{22} - \left(\frac{91}{22}\right)^2 = 6.5723 \\ s_1 &= \sqrt{6.5723} = 2.5636 \end{aligned}$$

$$s_{1y} = \overline{x_1 \cdot y} - \bar{x}_1 \cdot \bar{y} = \frac{9.84}{22} - \frac{91}{22} \cdot \frac{-69}{22} = 13.42$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \cdot \bar{x}_1 = \frac{-69}{22} - 2.042 \cdot \frac{91}{22} = -11.583$$

Por tanto la recta de regresión es:

$$\hat{\mathbf{Y}} = -\mathbf{11.583} + \mathbf{2.042} \cdot \mathbf{X}_1$$

Y la predicción para la observación 22 es

$$\begin{aligned}\hat{y}_{22} &= -11.583 + 2.042 \cdot x_{22,1} = \\ &= -11.583 + 2.042 \cdot 10 = \mathbf{8.837}\end{aligned}$$

P.7. Utilizando el modelo de la cuestión anterior, estudiar si la observación muestral 22 es atípica y si es influyente a priori.

Solución:
El residuo ordinario es

$$e_{22} = y_{22} - \hat{y}_{22} = 18 - 8.837 = \mathbf{9.163}$$

El residuo estandarizado es

$$r_{22} = \frac{e_{22}}{\hat{s}_R\sqrt{1-h_{22}}} = \frac{9.163}{3.80\sqrt{1-0.283}} = \mathbf{2.8477}$$

Por tanto la observación 22 es atípica aunque no de forma exagerada.
Para el cálculo de $\hat{s}_R$ se utliza lo siguiente

$$\begin{aligned}\sum_{i=1}^{22} e_i^2 &= \sum_{i=1}^{22} y_i^2 - \left(\hat{\alpha}_0 \cdot \sum_{i=1}^{22} y_i + \hat{\alpha}_1 \cdot \sum_{i=1}^{22} y_i \cdot x_{i1}\right) \\ &= 1091 - \left(-11.583 \cdot (-69) + 2.042 \cdot 9'84\right) \\ &= \mathbf{288.77}\end{aligned}$$

$$\begin{aligned}\hat{s}_R^2 &= \frac{\sum_{i=1}^{22} e_i^2}{g.l.} = \frac{288.77}{20} = \mathbf{14.439} \\ \hat{s}_R &= \sqrt{14.439} = \mathbf{3.80}\end{aligned}$$

Se calcula el valor de influencia

$$\begin{aligned}h_{22} &= \frac{1}{n}\left(1 + \frac{(x_{22,1} - \bar{x}_1)^2}{s_1^2}\right) \\ &= \frac{1}{22}\left(1 + \frac{\left(10 - \frac{91}{22}\right)^2}{6.5723}\right) = \mathbf{0.28324}\end{aligned}$$

$$n_{22} = \frac{1}{h_{22}} = \frac{1}{0.28324} = 3.5306$$

La media de los valores de influencia es

$$E(h) = \frac{k+1}{n} = \frac{2}{22} = 0.0909$$

Como $h_{22} = 0.28324 > 2 \cdot 0.0909 = 0.1818$, **la observación 22 es influyente a priori** (su $x_{22,1} = 10$ se separa claramente de la media $\bar{x}_1 = \frac{91}{22} = 4.1364$).

P.8. Utilizando el modelo de la cuestión anterior, estudiar el contraste de regresión e interpretarlo.

Solución:
El contraste de regresión es

$$\begin{aligned} H_{0,reg}: &\quad \text{“el modelo de regresión lineal ajustado } \textbf{no} \text{ es influyente”} \approx \ \alpha_1 = 0 \\ H_1: &\quad \text{“el modelo ajustado es influyente”} \approx \ \alpha_1 \neq 0 \end{aligned}$$

El estadístico del contraste es $\hat{d}_{reg}$

$$\begin{aligned} \hat{d}_{reg} &= \frac{\hat{s}_e^2}{\hat{s}_R^2} = \frac{585.82}{14.439} = 40.572 \sim F_{1,20}? \\ &\Rightarrow \ \hat{d}_{reg} = 40.572 > F_{1,20}(0.95) = 4.35 \end{aligned}$$

:

Se rechaza $H_{0,reg}$, por tanto **el modelo es significativo.**

Se calcula $\hat{s}_e^2$

$$\hat{s}_e^2 = \frac{scE}{1} = \mathbf{585.82}$$

$$\begin{aligned} scE &= scG - scR = n \cdot s_Y^2 - scR \\ &= 22 \cdot 39.754 - 288.77 \\ &= 874.59 - 288.77 = \mathbf{585.82} \end{aligned}$$

$$s_Y^2 = \overline{y^2} - (\bar{y})^2 = \frac{1091}{22} - \left(\frac{-69}{22}\right)^2 = \mathbf{39.754}$$

Se **rechaza** $H_{0,reg}$ a un nivel del m0.05y se asume que el modelo ajustado es significativo.

P.9. *En una segunda etapa se estudia el modelo de regresión lineal múltiple con las tres variables regresoras, obteniendo el siguiente modelo ajustado*

$$\hat{Y} = -19.8 + 1.7 \cdot x_1 + 2.54 \cdot x_2 - 11.6 \cdot x_3$$

siendo para este modelo la suma de los residuos al cuadrado

$$\sum_{i=1}^{22} e_i^2 = 192$$

y la diagonal de la matriz $\left(X^t X\right)^{-1}$

$$diag\left(X^t X\right)^{-1} = \begin{pmatrix} 2.57 & & & \\ & 0.012 & & \\ & & 0.129 & \\ & & & 7.584 \end{pmatrix}$$

En base a estos datos estudiar qué regresoras son significativas.

Solución:
La varianza residual en este caso es

$$\hat{s}_R^2 = \frac{\sum_{i=1}^{22} e_i^2}{g.l.} = \frac{192}{18} = \mathbf{10.667}$$

$$\begin{aligned}
Var(\hat{\boldsymbol{\alpha}}) &= \hat{s}_R^2 \cdot \left(X^t X\right)^{-1} \Rightarrow \\
Var(\hat{\alpha}_1) &= \hat{s}_R^2 \cdot q_{11} = 10.667 \cdot 0.012 = 0.128 \Rightarrow \sigma(\hat{\alpha}_1) = \sqrt{0.128} = 0.3578 \\
Var(\hat{\alpha}_2) &= \hat{s}_R^2 \cdot q_{22} = 10.667 \cdot 0.129 = 1.376 \Rightarrow \sigma(\hat{\alpha}_1) = \sqrt{1.376} = 1.173 \\
Var(\hat{\alpha}_3) &= \hat{s}_R^2 \cdot q_{33} = 10.667 \cdot 7.584 = 80.899 \Rightarrow \sigma(\hat{\alpha}_1) = \sqrt{80.899} = 8.9944
\end{aligned}$$

Dado que $P(t_{18} > 2.10) = 0.025$ se obtiene finalmente

Parám (α_i)	Est ($\hat{\alpha}_i$)	$\sigma(\hat{\alpha}_i)$	$\hat{t}_i = \frac{\hat{\alpha}_i}{\sigma(\hat{\alpha}_i)}$	**¿es significativa** X_i**?**
α_1	1.7	0.3578	$\frac{1.7}{0.3578} = 4.7513$	**SI**
α_2	2.54	1.173	$\frac{2.54}{1.173} = 2.1654$	**SI**
α_3	-11.6	2.7539	$\frac{-11.6}{8.9944} = :-1.2897$	**NO**

P.10. Utilizando el modelo de la cuestión anterior, calcular la correlación parcial y simple entre las variables Y y X_1.

Solución:
Coeficiente de correlación simple entre las variables $\mathbf{Y}$ y $\mathbf{X}_1$

$$r_{Y1} = \frac{s_{Y1}}{s_1 \cdot s_Y} = \frac{13.42}{2.5636 \cdot \sqrt{39.754}} = \mathbf{0.83026}$$

Coeficiente de correlación parcial entre las variables $\mathbf{Y}$ y $\mathbf{X}_1$

$$\begin{aligned}
r_{y1.23}^2 &= \frac{\hat{t}_1^2}{\hat{t}_1^2 + n - (k+1)} \\
&= \frac{4.7513^2}{4.7513^2 + 18} = 0.55638 \\
&\Rightarrow \; r_{y1.23} = \sqrt{0.55638} = \mathbf{0.74591}
\end{aligned}$$