

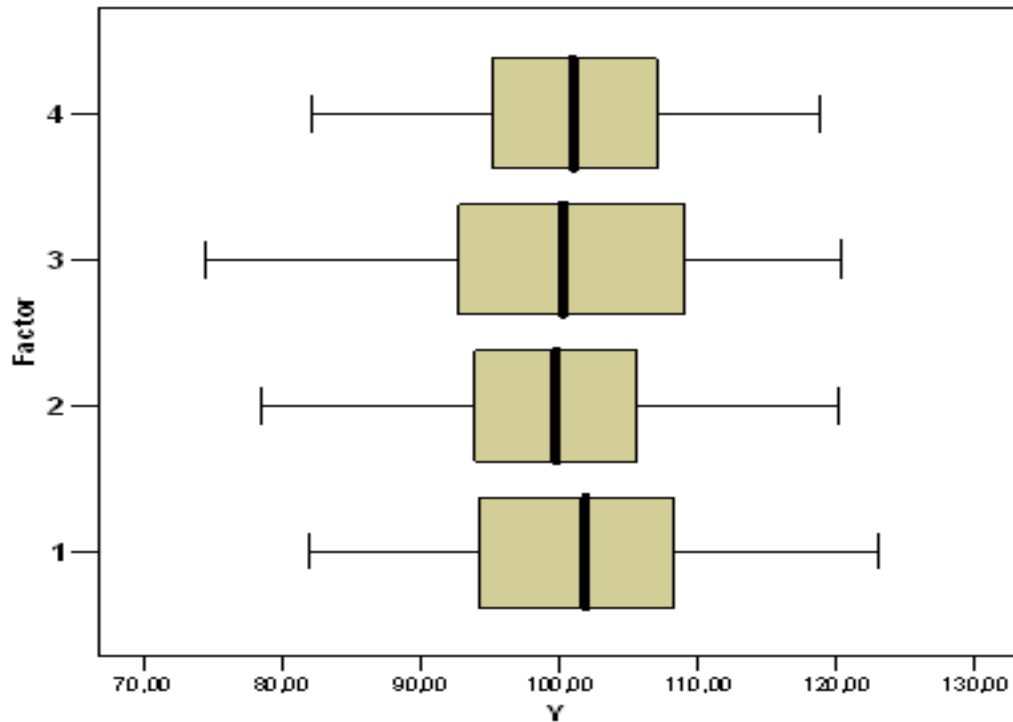
ESTADISTICA II
INGENIERIA INFORMATICA, 3^{ER} Curso
13 - Septiembre - 2.006 Primera Parte - Test

Las respuestas del TEST son las siguientes:

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

CUESTIONES

1. La eficacia de un estimador insesgado $\hat{\theta}_n$ de θ viene dada por
 - A. $(\text{Sesgo}(\hat{\theta}_n))^2 + \text{var}(\hat{\theta}_n)$
 - B. $\text{var}(\hat{\theta}_n)$
 - C. $1/\text{var}(\hat{\theta}_n)$ **Solución**
 - D. $ECM(\hat{\theta}_n)$
2. En el estudio de un diseño de experimentos con un factor se obtiene que el contraste de Levene es significativo. Por tanto
 - A. Las varianzas de los residuos en cada nivel no son iguales. **Solución**
 - B. Existe dependencia positiva en la serie de residuos.
 - C. Se verifica la hipótesis de homocedasticidad.
 - D. Existe una relación lineal entre la variable respuesta y el tiempo. Debe introducirse el tiempo como regresora en el modelo.
3. Un diseño de experimentos equilibrado
 - A. Es un diseño con factores sin interacción.
 - B. Es un diseño sin factores bloque.
 - C. Es un diseño con factores con interacción y replicados.
 - D. Es un diseño en el que todos los tratamientos se han asignado a un número igual de unidades experimentales. **Solución**
4. El gráfico de cajas de la variable respuesta de un diseño de experimentos con un factor frente a los niveles del factor es el siguiente. En cada nivel se tienen 100 observaciones y de este gráfico se deduce



- A. El factor es significativo y se verifican las hipótesis.
 B. Se verifica la hipótesis de normalidad pero falla la de homocedasticidad.
 C. Que el factor probablemente no sea significativo, no se observa el incumplimiento de alguna hipótesis. **Solución**
 D. Se verifica la hipótesis de homocedasticidad pero falla la de normalidad.
5. Los residuos (e_t) de un diseño de experimentos según el orden de obtención (t) son los siguientes

t	1	2	3	4	5	6	7	8	9	10
e_t	0'93	1'07	1'27	1'75	1'82	-0'70	-0'80	-1'11	-1'91	-1'95
t	11	12	13	14	15	16	17	18	19	20
e_t	-2'58	0'19	0'31	0'49	0'54	-0'01	-0'23	-0'41	0'71	0'62

Haciendo los contrastes de rachas: número total de rachas (R) y número de rachas ascendentes y descendentes (T) se obtiene

- A. $T = 5$, $R = 7$, existe dependencia negativa.
 B. $T = 6$, $R = 5$, existe dependencia positiva. **Solución** (Mediana de los residuos 0'25)
 C. $T = 5$, $R = 7$, existe independencia.
 D. $T = 6$, $R = 5$, existe independencia.
6. Utilizando los datos de la cuestión anterior y teniendo en cuenta que

$$\sum_{t=1}^{20} e_t = 0 \quad \sum_{t=1}^{20} e_t^2 = 28'24 \quad \sum_{t=1}^{20} e_t^3 = -17'48$$

y que el coeficiente de asimetría muestral (CA) tiene varianza $6/n$. Se deduce que

- A. $CA = -0'52$ y por tanto la distribución de los residuos es simétrica. **Solución**
- B. $CA = -3'56$ y por tanto la distribución de los residuos no es normal.
- C. $CA = -0'52$ y por tanto la distribución de los residuos es normal.
- D. $CA = 1'23$ y no se obtienen conclusiones acerca de la distribución de los residuos.

$$CA = \frac{\frac{-17.48}{20}}{\left(\sqrt{\frac{28.24}{20}}\right)^3} = -0.52 \Rightarrow CAS = \frac{-0.52}{\sqrt{\frac{6}{20}}} = -0.949 \Rightarrow \text{simetría}$$

7. Se quiere ajustar un diseño de experimentos con un factor fijo y un factor aleatorio. Por tanto, se supone que

- A. Los efectos del factor aleatorio (τ_i) son parámetros desconocidos.
- B. No existen diseños de experimentos con factores fijos y aleatorios.
- C. Los efectos del factor aleatorio (τ_i) son variables aleatorias de media cero y varianza desconocida. **Solución**
- D. Ninguna de las otras tres respuestas.

8. En un cuadrado latino con seis niveles en cada factor se ha obtenido que $scR = \sum_t e_t^2 = 20$. Un intervalo de confianza al 95% para la varianza del modelo es

- A. (0.52; 2.33)
- B. (0.16; 1.35)
- C. (1.12; 3.37)
- D. Ninguna de las otras tres respuestas. **Solución**

$$\begin{aligned} \frac{scR}{\sigma^2} &\sim \chi_{gl}^2 = \chi_{20}^2 \Rightarrow \chi_{20}^2(0.025) < \frac{20}{\sigma^2} < \chi_{20}^2(0.975) \\ 9.5908 &< \frac{20}{\sigma^2} < 34.170 \Rightarrow \frac{20}{34.170} < \sigma^2 < \frac{20}{9.5908} \\ \mathbf{0.5853} &< \sigma^2 < \mathbf{2.0853} \end{aligned}$$

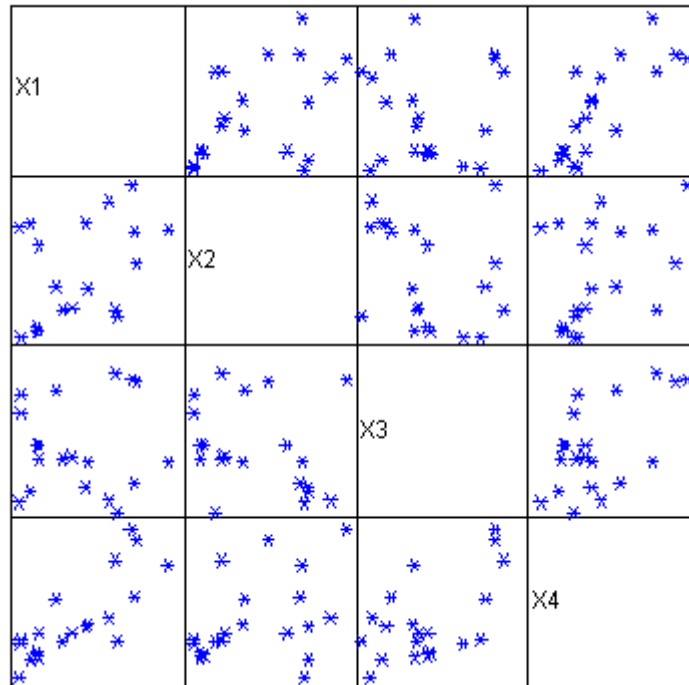
9. El Teorema de Gauss Markov afirma que bajo determinadas hipótesis en un modelo de regresión lineal múltiple

- A. Los estimadores mínimo cuadráticos son los más eficientes dentro de la clase de los estimadores insesgados y lineales. **Solución**
- B. Los estimadores mínimo cuadráticos son los de menor *ECM* dentro de la clase de los estimadores de mínima varianza.
- C. Los estimadores mínimo cuadráticos son asintóticamente más eficientes que los de máxima verosimilitud.
- D. Los estimadores mínimo cuadráticos son los de menor varianza pero no necesariamente sesgados.

10. En un modelo de regresión lineal simple la varianza del estimador $\hat{\alpha}_0$

- A. Disminuye al disminuir la ($Var(\hat{\alpha}_1)$). **Solución**
- B. Disminuye al aumentar la media de la regresora.

- C. Disminuye al aumentar la varianza del modelo.
 D. Disminuye al disminuir el tamaño muestral.
11. Al ajustar un modelo de regresión lineal múltiple, se obtiene que el $FIV(1) = 20$ (Factor de Incremento de la Varianza de la regresora x_1). Por tanto
- A. Existe una fuerte relación lineal entre la regresora x_1 y la respuesta Y .
 B. La regresora x_1 es significativa y debe estar en el modelo.
 C. Existe una fuerte relación lineal entre la regresora x_1 y las otras regresoras. **Solución**
 D. La regresora x_1 es ortogonal a las otras regresoras.
12. Al ajustar un modelo de regresión lineal múltiple con cuatro regresoras se obtiene el siguiente gráfico matricial de dispersión de las regresoras. De este gráfico se deduce:



- A. La existencia de observaciones influyentes a priori.
 B. Que el modelo de regresión lineal es adecuado.
 C. La existencia de multicolinealidad. **Solución**
 D. No se observa ningún hecho destacable.

ESTADISTICA II, Ingeniería Informática,

Problemas, 13 - Septiembre - 2.006

Problema 1. Se ha llevado a cabo un diseño de experimentos para estudiar la posible influencia de la variable "Tipo de examen" en la "Nota final" de una asignatura, controlando la influencia del "Profesor" y del "Tiempo (en horas) de estudio semanales". Se consideran 3 profesores distintos (P1, P2 y P3), 3 tipos de examen (E1=Test, E2=Problemas y E3=Mixto) y 3 tiempos de dedicación semanales (T1="Menos de 2 horas", T2="Entre 2 y 5 horas" y T3="Más de 5 horas"). Los resultados del experimento sobre 9 alumnos son los de la tabla adjunta (en cada casilla aparece el nivel del factor "Tiempo" y la nota final correspondiente que sacó cada alumno para los niveles de los factores "Tipo de examen" y "Profesor" en su fila y columna):

		Profesor (factor B)		
		P1	P2	P3
Tipo de Examen (factor A)	E1	T1	T2	T3
		4	7	5.5
	E2	T2	T3	T1
		7.5	8	4.5
	E3	T3	T1	T2
		7	4.25	6.5

P.1. Formular matemáticamente este diseño de experimentos e indicar las hipótesis que se hacen. Calcular las estimaciones de los efectos de los factores E, P y T.

Solución:

El modelo matemático es

$$Y_{ij(k)} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij(k)},$$

con $\varepsilon_{ij(k)}$ v. a. independientes $N(0, \sigma^2)$.

Para estimar los efectos de los factores E y P, calculamos:

	P1	P2	P3	$\bar{y}_{i.}$	$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$
E1	T1	T2	T3	5.5	-0.528
	4	7	5.5		
E2	T2	T3	T1	6.67	0.642
	7.5	8	4.5		
E3	T3	T1	T2	5.92	-0.108
	7	4.25	6.5		
$\bar{y}_{..j}$	6.167	6.417	5.5	$\bar{y}_{..} = 6.028$	
$\hat{\beta}_j = \bar{y}_{..j} - \bar{y}_{..}$	0.139	0.389	-0.528		

Para el factor T, tenemos:

$\bar{y}_{..(k)}$	$\hat{\gamma}_k = \bar{y}_{..(k)} - \bar{y}_{..}$
4.25	-1.778
7	0.972
6.83	0.802

P.2. Completar la tabla ANOVA e indicar qué efectos son significativos (nivel de significación 5%). Obtener los coeficientes de determinación.

Solución:

F. var.	sc	gl	SCM	F	p-valor	¿significativo?
Factor E	2.0972	2	1.0486	6.04	0.1420	NO
Factor P	1.3472	2	0.6736	3.88	0.2049	NO
Factor T	14.264	2	7.1319	41.08	0.0238	SÍ
Residual	0.3472	2	0.1726			
Global	18.055	8	2.2568			

Las sumas de cuadrados se han calculado de la siguiente manera:

$$scE = K \sum_{i=1}^K \hat{\alpha}_i^2 = 3 (0.528^2 + 0.642^2 + 0.108^2) = 2.0972$$

$$scP = K \sum_{j=1}^K \hat{\beta}_j^2 = 3 (0.139^2 + 0.389^2 + 0.528^2) = 1.3472$$

$$scT = K \sum_{k=1}^K \hat{\gamma}_k^2 = 3 (1.778^2 + 0.972^2 + 0.802^2) = 14.264$$

$$scR = scG - (scE + scP + scT) = 0.3472$$

Los coeficientes de determinación son:

$$R^2 (\text{"Examen"}) = \frac{2.0972}{18.055} = 0.11616$$

$$R^2 (\text{"Profesor"}) = \frac{1.3472}{18.055} = 0.0746$$

$$R^2 (\text{"Tiempo"}) = \frac{14.264}{18.055} = 0.79$$

$$R^2 (\text{"TOTAL"}) = \frac{18.055 - 0.3472}{18.055} = 0.98077 = 0.116 + 0.0746 + 0.79$$

P.3. Teniendo en cuenta los resultados del apartado anterior, indicar qué modelo más simple sirve para explicar los datos. Completar la tabla ANOVA, indicando si los nuevos factores son significativos al 5%. Calcular los coeficientes de determinación.

Solución:

Como solamente es significativo el factor "Tiempo de estudio", se puede ajustar un diseño de una vía

F. var.	sc	gl	SCM	F	p-valor	¿significativo?
Factor T	14.264	2	7.1319	11.29	0.0093	SÍ
Residual	3.7917	6	0.6319			
Global	18.055	8	2.2568			

El coeficiente de determinación es

$$R^2 = \frac{14.264}{18.055} = 0.79$$

P.4. Ajustar los datos a un diseño cuya única vía sea el factor T, calcular las estimaciones de los efectos y calcular los grupos homogéneos, al 95%, según el método LSD.

Solución:

Reordenando los datos según el diseño de una vía tenemos

	T1	T2	T3	
E1	4	7	5.5	
E2	4.5	7.5	8	
E3	4.25	6.5	7	
\bar{y}_i	4.25	7	6.83	$\bar{y}_{..} = 6.028$
$\hat{\gamma}_i = \bar{y}_i - \bar{y}_{..}$	-1.778	0.972	0.802	

El modelo matemático es

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \gamma_i + \varepsilon_{ij}$$

Para estudiar los grupos homogéneos, realizamos los siguientes contrastes:

$$\text{Contraste 1} \begin{cases} H_0^{12} : \alpha_1 = \alpha_2 \\ H_1^{12} : \alpha_1 \neq \alpha_2 \end{cases} \quad \text{Contraste 2} \begin{cases} H_0^{13} : \alpha_1 = \alpha_3 \\ H_1^{13} : \alpha_1 \neq \alpha_3 \end{cases} \quad \text{Contraste 3} \begin{cases} H_0^{23} : \alpha_2 = \alpha_3 \\ H_1^{23} : \alpha_2 \neq \alpha_3 \end{cases}$$

El valor crítico es

$$\frac{\bar{y}_i - \bar{y}_j}{\hat{s}_R \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{n-I} \Rightarrow |\bar{y}_i - \bar{y}_j| < \hat{s}_R \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \times t_6(0.975)$$

$$\Rightarrow |\bar{y}_i - \bar{y}_j| < 0.795 \times \sqrt{\frac{1}{3} + \frac{1}{3}} \times 2.4469 = \mathbf{1.5883}$$

Constraste 1: $|\bar{y}_1 - \bar{y}_2| = |4.25 - 7| = 2.75 \geq 1.5883 \Rightarrow$ Se rechaza H_0^{12}

Constraste 2: $|\bar{y}_1 - \bar{y}_3| = |4.25 - 6.83| = 2.58 \geq 1.5883 \Rightarrow$ Se rechaza H_0^{13}

Constraste 3: $|\bar{y}_2 - \bar{y}_3| = |7 - 6.83| = 0.17 < 1.5883 \Rightarrow$ Se acepta H_0^{23}

Los grupos homogéneos son: T1 y (T2,T3).

P.5. *Calcular los residuos del modelo del apartado anterior. ¿Hay alguno atípico?*

Solución:

Teniendo en cuenta que

$$\hat{y}_{ij} = \hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$$

y que la estimación de la varianza residual es $\hat{s}_R^2 = 0.6319$ (y por tanto $\hat{s}_R = 0.795$) entonces los residuos son:

Factor T	y_{ij}	\hat{y}_{ij}	$e_{ij} = y_{ij} - \hat{y}_{ij}$	$r_{ij} = \frac{e_{ij}}{\hat{s}_R}$
T1	4	4.25	-0.25	-0.3144
T1	4.5	4.25	0.25	0.3144
T1	4.25	4.25	0	0
T2	7	7	0	0
T2	7.5	7	0.5	0.6289
T2	6.5	7	-0.5	-0.6289
T3	5.5	6.83	-1.33	1.673
T3	8	6.83	1.167	1.468
T3	7	6.83	0.167	0.21

No hay datos atípicos.

ESTADISTICA II, Ingeniería Informática,

Problemas, 13 - Septiembre - 2.006

Problema 2:

En una red de comunicaciones se quiere estudiar la relación entre la variable respuesta $Y =$ "velocidad de transmisión" y dos regresoras $X_1 =$ "grosor del cable" y $X_2 =$ "temperatura ambiente". Para ello se han hecho diferentes pruebas y se ha tomado una muestra de 50 observaciones. A partir de esta muestra se han obtenido los siguientes datos

$$X^t X = \begin{pmatrix} 50 & 90.2 & -12 \\ 90.2 & 178.26 & -27.5 \\ -12 & -27.5 & 4067 \end{pmatrix}$$
$$X^t Y = \begin{pmatrix} 514.2 \\ 983.59 \\ 1861.1 \end{pmatrix} \quad Y^t Y = \sum_{i=1}^{50} y_i^2 = 6630.12$$

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

P.6. En un primer análisis se estudia la relación lineal simple entre la respuesta $Y =$ "velocidad de transmisión" y la regresora $X_1 =$ "grosor del cable". Calcular un intervalo de confianza al 95% para el coeficiente de regresión del modelo (α_1)

Solución:

$$\hat{\alpha}_1 = \frac{s_{1Y}}{s_1^2} = \frac{\frac{983.59}{50} - \frac{90.2}{50} * \frac{514.2}{50}}{\frac{178.26}{50} - \left(\frac{90.2}{50}\right)^2}$$
$$= \frac{19.672 - 1.804 * 10.284}{3.5652 - (1.804)^2} = \frac{1.1197}{0.31078} = 3.6029$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x}_1 = 10.284 - 3.6029 * 1.804 = 3.7844$$

La recta de regresión estimada es

$$\hat{y} = 3.7844 + 3.6029 \cdot x_1$$

Cálculo del intervalo de confianza

$$Var(\hat{\alpha}_1) = \frac{\hat{s}_R^2}{n s_1^2} = \frac{23.758}{50 \cdot 0.31078} = 1.5289$$
$$\sigma(\hat{\alpha}_1) = \sqrt{1.5289} = 1.2365$$

Cálculo de la varianza residual del modelo

$$\sum e_i^2 = 6630.12 - (3.7844 \cdot 514.2 + 3.6029 \cdot 983.59) = 1140.4$$
$$\hat{s}_R^2 = \frac{1140.4}{48} = 23.758$$

El intervalo de confianza al 95% es

$$\begin{aligned}\alpha_1 &\in 3.6029 \mp 1.2365 \cdot t_{48}(0.975) = 3.6029 \mp 1.2365 \cdot 1.96 \\ \alpha_1 &\in (1.1794; 6.0264)\end{aligned}$$

El intervalo de confianza al 90% es

$$\begin{aligned}\alpha_1 &\in 3.6029 \mp 1.2365 \cdot t_{48}(0.95) = 3.6029 \mp 1.2365 \cdot 1.66 \\ \alpha_1 &\in (1.5503; 5.6555)\end{aligned}$$

P.7. Utilizando el modelo anterior, calcular un intervalo de predicción al 95% para una observación cuyo grosor es $x_{1,t} = 1$

Solución:

La predicción es

$$\hat{y}(x_{1,t} = 1) = 3.7844 + 3.6029 \cdot 1 = 7.3873$$

El número de observaciones equivalentes es

$$n_t = \frac{n}{1 + \frac{(x_{1t} - \bar{x}_1)^2}{s_1^2}} = \frac{50}{1 + \frac{(1 - 1.804)^2}{0.31078}} = 16.234$$

$$\begin{aligned}Var(\hat{y}(1)) &= \hat{s}_R^2 \cdot \left(1 + \frac{1}{n_t}\right) = 23.758 \cdot \left(1 + \frac{1}{16.234}\right) = 25.221 \\ \sigma(\hat{y}(1)) &= \sqrt{25.221} = 5.0221\end{aligned}$$

El intervalo de predicción es

$$\begin{aligned}\hat{y}(1) &\in 7.3873 \mp 5.0221 \cdot t_{48}(0.975) \\ &= 7.3873 \mp 5.0221 \cdot 1.96 \\ \hat{y}(1) &\in (-2.456; 17.231)\end{aligned}$$

P.8. A continuación se ajusta un modelo con las dos regresoras se obtiene

$$\hat{y}_t = 3.68 + 3.73 \cdot \text{grosor} + 0.495 \cdot \text{temp}$$

Contrastar la hipótesis $H_0 \equiv \alpha_1 = \alpha_2 = 0$

Solución:

Es el denominado contraste conjunto de regresión de la F que se resuelve construyendo la siguiente tabla ANOVA

$$\begin{aligned}scR &= \sum e_t^2 \\ &= 6630.12 - (3.68 \cdot 514.2 + 3.73 \cdot 983.59 + 0.495 \cdot 1861.1) \\ &= 147.83\end{aligned}$$

$$scG = ns_Y^2 = 50 \cdot \left(\frac{6630.12}{50} - \left(\frac{514.2}{50} \right)^2 \right) = 1342.1$$

:

F. var.	sc	gl	SCM	F	p-valor	¿significativo?
Modelo	1194.3	2	597.15	189.85	0.000	SÍ
Residual	147.83	47	3.1453	$\hat{s}_R = 1.7735$		
Global	1342.1	49	27.390	$\hat{s}_Y = 5.2335$		

Se **RECHAZA** claramente la hipótesis de NO influencia del modelo. El contraste de regresión es **significativo**.

P.9. ¿cuánto mejora el coeficiente de determinación corregido por grados de libertad al pasar del modelo con una regresora (grosor) al modelo con dos regresoras (grosor y temperatura)? Interpretar este coeficiente en ambos casos.

Solución:

Con una regresora

$$\begin{aligned}\bar{R}_1^2 &= 1 - \frac{\hat{s}_{R,1}^2}{\hat{s}_Y^2} = 1 - \frac{23.758}{27.390} = 0.1326 \\ \Rightarrow \bar{R}_1 &= \sqrt{0.1326} = 0.36414\end{aligned}$$

Por tanto, la influencia de la variable grosor es pequeña

Con dos regresoras

$$\begin{aligned}\bar{R}_2^2 &= 1 - \frac{\hat{s}_{R,2}^2}{\hat{s}_Y^2} = 1 - \frac{3.1453}{27.390} = 0.88517 \\ \Rightarrow \bar{R}_2 &= \sqrt{0.88517} = 0.94083\end{aligned}$$

El modelo con las dos variables explica claramente el comportamiento de la respuesta.

El incremento en el \bar{R}^2 al cambiar de modelo es muy fuerte

$$\Delta \bar{R}^2 = \bar{R}_2^2 - \bar{R}_1^2 = 0.88517 - 0.1326 = 0.75257$$

P.10. De la primera observación muestral se tiene la siguiente información

$$y_1 = 10.5 \quad x_{11} = 2 \quad x_{12} = -4 \quad DFITS_1 = 0.20$$

Estudiar si la observación es influyente y/o atípico.

Será de utilidad el siguiente dato

$$(X^t X)^{-1} = 10^{-3} \begin{pmatrix} 229.5 & -116.1 & -0.108 \\ -116.1 & 64.39 & 0.0927 \\ -0.108 & 0.0927 & 0.0246 \end{pmatrix}$$

Solución:

Residuo ordinario

$$\begin{aligned}e_1 &= y_1 - \hat{y}_1 = 10.5 - (3.68 + 3.73 \cdot 2 + 0.495 \cdot (-4)) \\ &= 10.5 - 9.16 = \mathbf{1.34}\end{aligned}$$

Influencia

$$\begin{aligned} h_1 &= 10^{-3} \cdot \begin{pmatrix} 1 & 2 & -4 \end{pmatrix} \cdot \begin{pmatrix} 229.5 & -116.1 & -0.108 \\ -116.1 & 64.39 & 0.0927 \\ -0.108 & 0.0927 & 0.0246 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \\ -4 \end{pmatrix} \\ &= \mathbf{0.022434} \end{aligned}$$

Residuo tipificado

$$\begin{aligned} \text{Var}(e_1) &= \hat{s}_R^2 (1 - h_1) = 3.1453(1 - 0.022434) = 3.0747 \\ \Rightarrow \sigma(e_1) &= \sqrt{3.0747} = 1.7535 \end{aligned}$$

$$r_1 = \frac{e_1}{\sigma(e_1)} = \frac{1.34}{1.7535} = \mathbf{0.76419}$$

La observación 1 no es atípica

"Influencia a priori"

$$E(h_i) = \frac{k+1}{n} = \frac{3}{50} = 0.06$$

$$h_1 = 0.022434 < 2 \cdot E(h_i) = 0.12$$

La observación 1 no es influyente a priori

"Influencia a posteriori"

$$DFITS_1 = 0.20 < 2 \cdot \sqrt{\frac{k}{n}} = 2 \cdot \sqrt{\frac{2}{50}} = 0.40$$

La observación 1 no es influyente a posteriori