Estadística II - Taller 06 **Semestre: 2021-02**

Profesores: Carlos M. Lopera-Gómez y Raúl Alberto Pérez

Monitor: Simon Pedro Galeano

---

1. Responda las siguientes preguntas.

a) En un modelo de regresión es posible que $R^2 = R^2_{adj}$, ¿bajo qué condiciones? ¿Es esto factible en modelos de regresión?

b) Se tiene un modelo de regresión con k covariables, ¿cuál es el número de observaciones para obtener una estimación de la varianza?.

c) Suponga que se tiene $\underline{\mathbf{x}}_0 = [1, x_{01}, \cdots, x_{0k}]$ adecuado para hacer inferencias respecto a la respuesta media de $\underline{\mathbf{x}}_0$, ¿la siguiente ecuación es correcta?

$$P\left(0 < \frac{\hat{Y}_0 - \mathbb{E}[Y|\underline{\mathbf{x}}_0|]}{se(\hat{Y}_0)} < t_{\alpha/2,n-p}\right) = 0.5 - \frac{\alpha}{2}$$

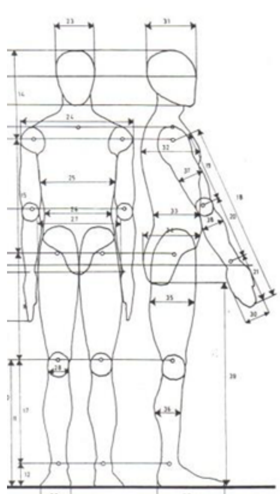
donde $\hat{Y}_0 = \underline{\mathbf{x}}_0\underline{\beta}$.

d) Suponga que $\underline{\mathbf{x}}_0 = [1, x_{01}, \cdots, x_{0k}]^T$ es un punto en el que no se comete extrapolación, luego $\underline{\mathbf{x}}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\underline{\mathbf{x}}_0 < 1$.

e) Considere a la entrada h_{ii} de la matriz $\mathbf{H}$, se tiene que $\sum_{i=1}^{n} h_{ii}$ es igual al número de covariables en el modelo.

f) En un modelo de regresión suponga que $2p > n$ y que para i $= 3$ $h_{ii} > \frac{2p}{n}$, ¿dicha observación es un punto de balanceo?

2. Considere la siguiente base de datos



Obs	Peso (Kg)	Sexo	Estatura (m)	Circunferencia cuello (cm)	Circunferencia muñeca (cm)
1	47.6	F	1.57	29.5	13.9
2	68.1	M	1.66	38.4	16
3	68	M	1.9	36.5	16.6
4	80	M	1.76	38	17.1
5	68.1	M	1.83	38	17.1
6	56.1	F	1.66	33	14.7
7	54.2	F	1.65	32.5	15.4
8	69.2	M	1.78	40.5	16.5
9	74.3	M	1.68	38	16.1
10	73.3	M	1.69	37.5	16.3
11	102.2	M	1.79	41.5	17.1
12	46.7	F	1.49	31.5	13.8
13	63.8	M	1.74	38	16.4
14	76.9	M	1.73	39.5	17.6
15	52.5	F	1.52	32.5	14.4
16	67.3	M	1.76	36.5	16.1
17	79.1	M	1.82	38	18
18	58.4	F	1.62	33	14.3
19	59.3	F	1.68	32	14.2
20	57.3	F	1.61	32	14.7
21	67.6	F	1.64	34.5	15.3
22	62.7	F	1.67	33	15.3
23	71.9	M	1.64	38.5	16.8
24	74.9	M	1.75	40	16.8
25	73	M	1.85	37.2	16.4
26	63.8	M	1.71	35	15.6

a) Ajuste un modelo de regresión usando la estatura como respuesta y al resto como covariables (excepto al sexo).

b) Plantee un contraste usando la hipótesis general lineal para comparar efectos de covariables sobre la respuesta.

c) Valide los supuestos del modelo, encuentre puntos de balanceo e influencia, también identifique outliers.

Ejercicio 2

Solución

a)

Vamos a ajustar el siguiente modelo de regresión

$$\text{Estatura}_i = \beta_0 + \beta_1 \text{Peso}_i + \beta_2 \text{circuncuello}_i + \beta_3 \text{circunmuneca}_i + \varepsilon_i,\ 1 \leq i \leq 26;\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Una vez ajustado el modelo, se obtuvo la siguiente estimación

$$\widehat{\text{Estatura}}_i = 0.7123 + 0.0011\text{Peso}_i - 0.0056\text{circuncuello}_i + 0.0705\text{circunmuneca}_i\ 1 \leq i \leq 26$$

b)

Se propone contrastar la hipótesis de que el efecto de las muñecas es igual a la mitad de veces el del cuello, además, se quiere observar si el efecto del peso es no significativo.

$$\begin{cases} H_0 : 2\beta_3 = \beta_2,\ \beta_1 = 0 \\ H_1 : 2\beta_3 \neq \beta_2,\ \beta_1 \neq 0 \end{cases}$$

Expresando la hipótesis anterior de manera matricial se obtiene que

$$\begin{cases} H_0 : \mathbf{L}\underline{\beta} = \underline{0} \\ H_1 : \mathbf{L}\underline{\beta} \neq \underline{0} \end{cases}$$

donde $L = \begin{bmatrix} 0 & 0 & -1 & 2 \\ 0 & 1 & 0 & 0 \end{bmatrix}$.

Ahora, se especifica el modelo reducido

$$\text{Estatura}_i = \beta_0 + 2\beta_3 \text{circuncuello}_i + \beta_3 \text{circunmuneca}_i + \varepsilon_i,\ 1 \leq i \leq 26;\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

lo cual es equivalente a

$$\text{Estatura}_i = \beta_0 + \beta_3(2\text{circuncuello}_i + \text{circunmuneca}_i) + \varepsilon_i,\ 1 \leq i \leq 26;\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

o

$$\text{Estatura}_i = \beta_0 + \beta_3 X_{3i} + \varepsilon_i,\ 1 \leq i \leq 26;\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

con $X_{3i} = 2\text{circuncuello}_i + \text{circunmuneca}_i$.

Se tiene que

$$\begin{aligned} F_{0obs} &= \frac{\frac{SSE(RM) - SSE(FM)}{\text{g.l}_{SSE(RM)} - \text{g.l}_{SSE(FM)}}}{MSE} \\ &= \frac{\frac{0.122730 - 0.092836}{24 - 22}}{0.004220} \\ &= 3.541943 \end{aligned}$$

Así, $P(F_{2,24} > F_{0obs}) = 0.0449$ y por tanto, a un nivel de significancia del 5 % se rechaza la hipótesis nula.

d)

Se procede a validar los supuestos del modelo

Se analiza la homocedasticidad y los datos atípicos

del gráfico anterior se puede observar claramente una violación del supuesto de varianza constante y que no se presetan datos atípicos.

Posteriormente se analiza la influencia y el balanceo

Respecto a la influencia solo se encuentra una observación, lo cual no es grave ni representativa, mientras que para el balanceo se tienen 3 observaciones que se encuentran muy alejadas del resto pero no afectan las estimaciones de los parámetros.

Por último se verifica la normalidad

No se ven desviaciones graves del supuesto de normalidad y adicionalmente el test de Shapiro-Wilk no rechaza la hipótesis de que los residuales vienen de una población normal.

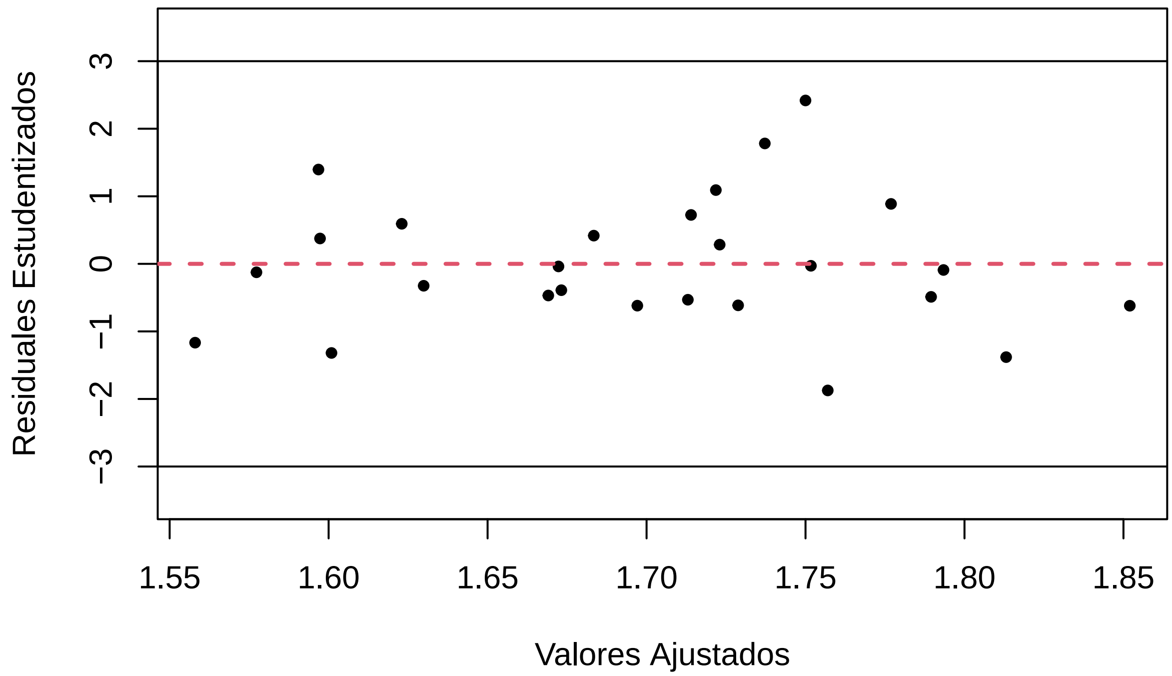


Figura 1: Estudentizados vs ajustados

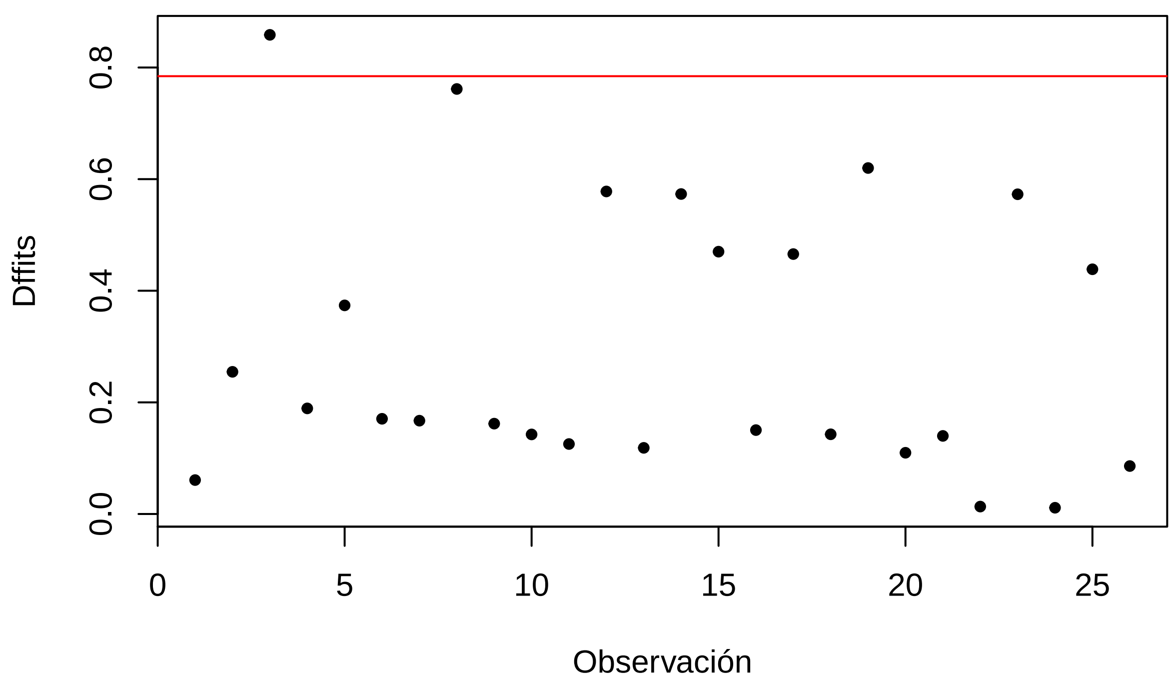


Figura 2: Análisis de influencia

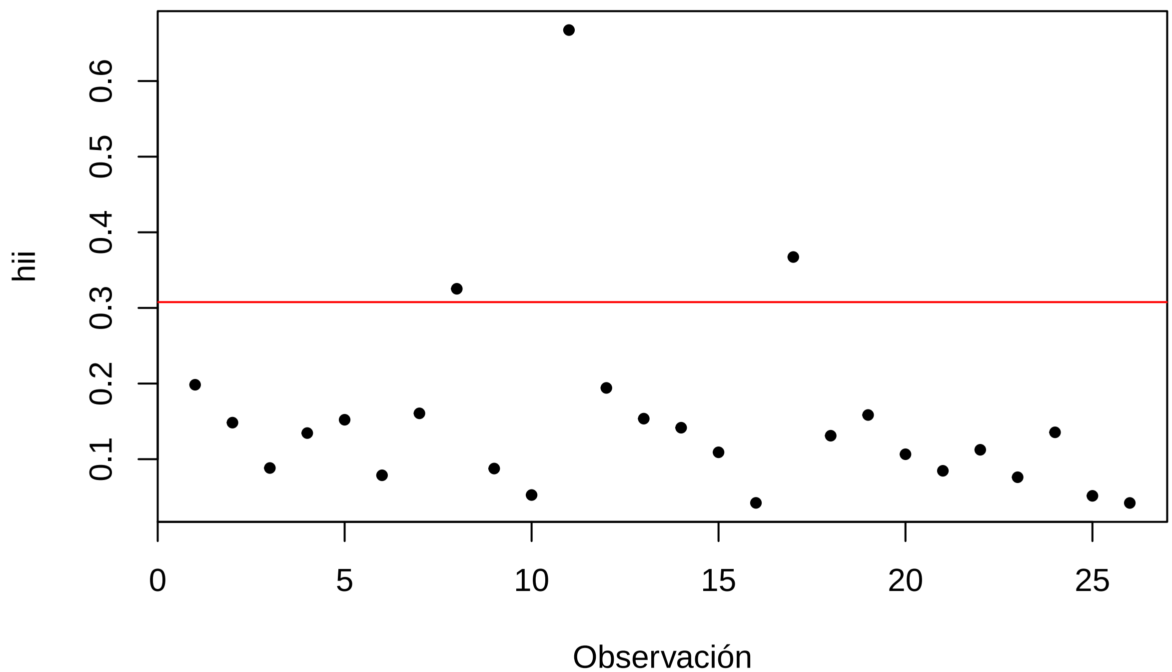


Figura 3: Análisis de balanceo

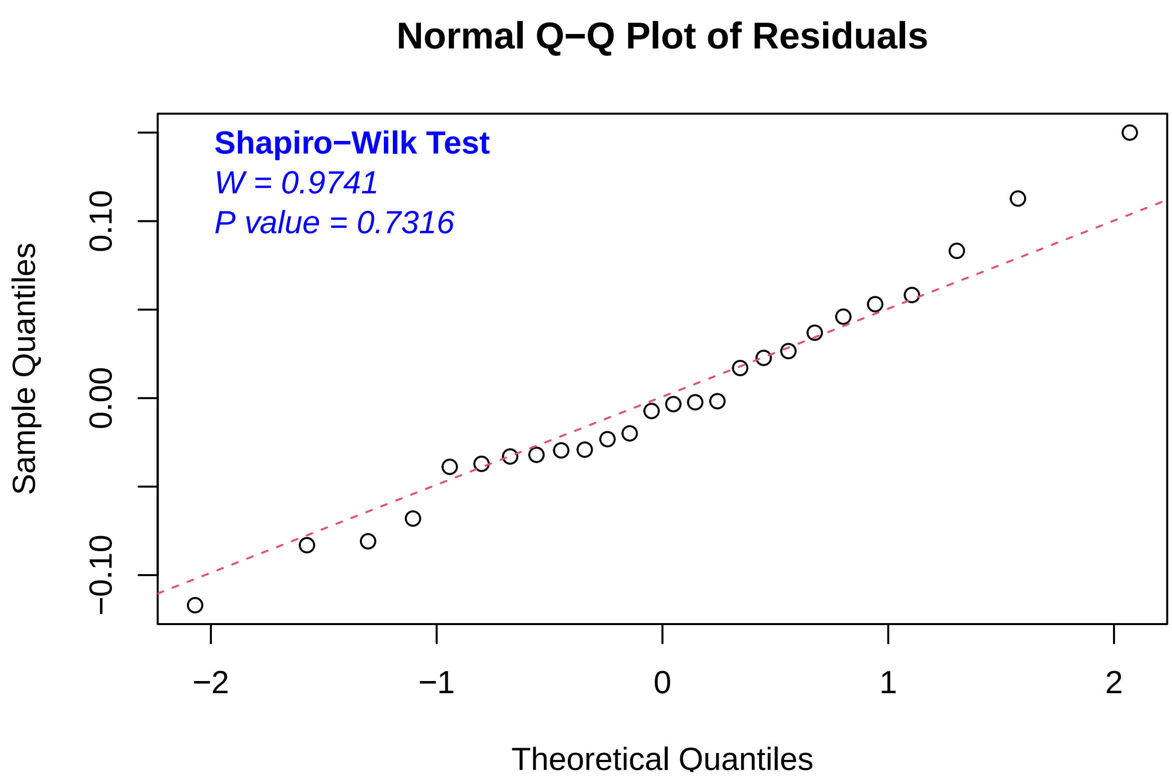


Figura 4: Análisis de normalidad