Trabajo 1

Estudiantes

Verónica Pérez Zea
Tomás Gutiérrez Orrego
Guillermo Toloza Guzmán
Juan Fernando Misas Marín

Equipo #02

Docente

Julieth Verónica Guarín Escudero

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

19 pt

Teniendo en cuenta la base de datos asignada, la cual es **Equipo02.txt**, las 5 variables regresoras son:

Y: Riesgo de infección

X_1: Duración de la estadía

X_2: Rutina de cultivos

X_3: Número de camas

X_4: Censo promedio diario

X_5: Número de enfermeras

Entonces, se plantea un modelo de RLM para el problema:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,\ i = 1, 2, ..., 65$$

que tiene como supuestos lo siguiente:

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ \forall_i = 1, 2, ..., 65$$

1.1. Modelo de regresión

3 pt

Al ajustar el modelo anterior se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	1.3742
$\hat{\beta}_1$	0.2122
$\hat{\beta}_2$	-0.0134
$\hat{\beta}_3$	0.0545
$\hat{\beta}_4$	0.0055
$\hat{\beta}_5$	0.0017

Por lo tanto, el modelo de regresión ajustado es:

$$\widehat{Y} = 1.3742 + 0.2122X_{i1} - 0.0134X_{i2} + 0.0545X_{i3} + 0.0055X_{i4} + 0.0017X_{i5},\ i = 1, 2, ..., 65$$

1.2. Significancia de la regresión 5 pt

Para analizar la significancia de la regresión, se plantea el siguiente juego de hipótesis:

$$\begin{aligned} &\mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0, \ \text{vs.} \\ &\mathrm{H}_1 : \text{Algún } \beta_j \neq 0, \ j = 1, 2, 3, 4, 5 \end{aligned}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{SSR/5}{SSE/(n-p)}$$

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,59} \tag{1}$$

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	Valor-P
Modelo	55.3626	5	11.072524	12.8712	1.79699e-08
Error	50.7549	59	0.860253		

De la tabla ANOVA anterior se observa que, como Valor-P $< 0.05 = \alpha$, se rechaza H_0 concluyendo que el modelo de RLM propuesto es significativo. Esto quiere decir, que el riesgo de infección depende significativamente de al menos una de las predictoras del modelo.

1.3. Significancia de los parámetros 6 pt

Cuadro 3: Resumen de los coeficientes

sobran los ^

	Estimación $\hat{\beta}_j$	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
$\hat{\beta}_0$	1.3742	1.6491	0.8333	0.4080
$\hat{\beta}_1$	0.2122	0.0774	2.7414	0.0081
$\hat{\beta}_2$	-0.0134	0.0307	-0.4359	0.6645
$\hat{\beta}_3$	0.0545	0.0161	3.3943	0.0012
$\hat{\beta}_4$	0.0055	0.0074	0.7417	0.4612
$\hat{\beta}_5$	0.0017	0.0008	2.1930	0.0323

Se establece el siguiente juego de hipótesis:

$$\left.\begin{aligned} &\mathrm{H}_0 : \beta_j = 0 \\ &\mathrm{H}_1 : \beta_j \neq 0 \end{aligned}\right. \text{para } j = 1, 2, 3, 4, 5$$

Cuyo estadístico de prueba es:

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \overset{H_0}{\sim} t_{59} \tag{2}$$

Los resultados de las pruebas: valor del estadístico de prueba y el valor-P para la prueba, se obtienen en las dos últimas columnas de la tabla de resumen de los coeficientes.

A un nivel de significancia $\alpha = 0.05$ se concluye que los parámetros individuales $\beta_1, \beta_3, \beta_5$ son significativos cada uno en presencia de los demás parámetros.

Por otro lado, se encuentra que $\beta_0, \beta_2, \beta_4$ son individualmente no significativos en presencia de los demás parámetros.

1.4. Interpretación de los parámetros estimados 2 pt

Solo se podrán interpretar parámetros que resultaron significativos individualmente, en este caso son: $\beta_1, \beta_3, \beta_5$.

porcentaje promedio

- $\hat{\beta}_1 = 0.2122$ indica que por cada unidad que se aumente la duración de la estadía (X_1) el promedio de Riesgo de Infección aumenta significativamente en 0.2122 unidades, cuando las demás predictoras se mantienen fijas.
- $\hat{\beta}_3 = 0.0545$ indica que por cada unidad que se aumente el número de camas (X_3) el promedio de Riesgo de Infección aumenta significativamente en 0.0545 unidades, cuando las demás predictoras se mantienen fijas.
- $\hat{\beta}_5 = 0.0017$ indica que por cada unidad que se aumente el número de enfermeras (X_5) el promedio de Riesgo de Infección aumenta significativamente en 0.0017 unidades, cuando las demás predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2 3 pt

Para el cálculo del R^2 se usa la tabla Anova:

$$R^2 = \frac{SSR}{SST} = \frac{55.3626}{55.3626 + 50.7549} = 0.5217$$

El modelo tiene un $R^2 = 0.5217$, es decir, que aproximadamente el 52.17 % de la variabilidad total en el Riesgo de Infección es explicado por el modelo de RLM propuesto.

2. Pregunta 2

3p+

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el Valor-P más alto en el modelo fueron X_2, X_4, X_5, por lo tanto se muestra la siguiente tabla de todas las regresiones posibles:

Cuadro 4: Resumen tabla de todas las regresiones

2p+

	SSE	Covariables en el modelo
Modelo completo	50.755	X1 X2 X3 X4 X5
Modelo reducido	55.916	X1 X3

Luego, un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i,\ i = 1, 2, ..., 65$$
$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),\ \forall_i = 1, 2, ..., 65$$

Probar la significancia simultánea de las variables con el Valor-P más alto, equivale a la siguiente prueba de hipótesis:

$$\text{H}_0 : \beta_2 = \beta_4 = \beta_5 = 0 \text{ vs.}$$
$$\text{H}_1 : \text{Algún } \beta_j \neq 0,\ j = 2, 4, 5$$

Revisen bien el uso de esa notación

Para esta prueba de hipótesis se tiene como estadístico de prueba a:

$$F_0 = \frac{MSExtra}{MSE} = \frac{\cancel{MSR(\beta_0, \beta_1, \beta_3 \mid \beta_2, \beta_4, \beta_5)}}{MSE} = \frac{[SSR(\beta_0, \beta_1, \beta_3 \mid \beta_2, \beta_4, \beta_5)]/3}{MSE}$$

MSR$(\beta_2, \beta_4 / \beta_0, \beta_1, \beta_3, \beta_5)$

$$= \frac{[SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)]/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)}$$
$$= \frac{[55.916 - 50.755]/3}{55.916/59} = 1.815$$

Para el criterio de decisión se requiere obtener el valor crítico de una distribución $f_{3,65-6} = f_{3,59}$ a un nivel de significancia $\alpha = 0.05$, esto es, $f_{0.05,3,59} = 2.7608$.

Como $F_0 = 1.815 < f_{0.05,3,59} = 2.7608$, entonces no se rechaza H_0 y se concluye que el subconjunto no es significativo para el promedio de Riesgo de Infección, el cual no depende de al menos una variable asociada al Valor-P más alto.

1p+

¿Se pueden descartar o no?

3. Pregunta 3 3,5 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si el número promedio de camas en el hospital es igual al número promedio de pacientes en el hospital o si la duración promedio de la estadía de los pacientes es igual a la mitad del número promedio de enfermeras en el hospital.

Por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{aligned} &\mathrm{H}_0 : \beta_3 = \beta_4,\ \ \beta_1 = 0.5\beta_5 \text{ vs.} \\ &\mathrm{H}_1 : \beta_3 \neq \beta_4 \ \vee\ \beta_1 \neq 0.5\beta_5 \end{aligned}$$

Veamos H_0 como un sistema de dos ecuaciones:

$$H_0 : \begin{cases} \beta_3 - \beta_4 = 0 \\ \beta_1 - 0.5\beta_5 = 0 \end{cases}$$

que en forma matricial se puede expresar como:

sólo esto es $\mathbf{L}$

1,5 pt

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -0.5 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Por tanto, se tiene una prueba de hipótesis lineal general, con:

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -0.5 \end{bmatrix}$$

→ Ahora sí la definen bien?

que tiene $r = 2$ filas linealmente independientes.

Entonces, el modelo reducido en este caso es:

$\beta_1(X_{1i} + 2X_{5i})$

0pt

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_3 X_{i4} + \beta_1 X_{i5} + \varepsilon_i,\ \ i = 1, 2, ..., 65 \\ &= \beta_0 + \boxed{\beta_1(X_{i1} + X_{i5})} + \beta_2 X_{i2} + \beta_3(X_{i3} + X_{i4}) + \varepsilon_i,\ \ i = 1, 2, ..., 65 \\ &= \beta_0 + \beta_1 X_{i1,5} + \beta_2 X_{i2} + \beta_3 X_{i3,4} + \varepsilon_i,\ \ i = 1, 2, ..., 65 \end{aligned}$$

→ ¿De dónde sale 5? nunca lo dijeron

donde $X_{i1,5} = X_{i1} + 0.5X_{i5}$, y $X_{i3,4} = X_{i3} + X_{i4}$.

2pt

Finalmente, la expresión para el estadístico de prueba es:

$$F_0 = \frac{MSH}{MSE} = \frac{SSH/2}{MSE} = \frac{(SSE(MR)^* - SSE(MF))/2}{MSE(MF)} = \frac{(SSE(MR)^* - 50.755]/2}{55.916/59} \overset{H_0}{\sim} f_{2,59}$$

Solo resta establecer el valor en (*) que es SSE(MR), el cual no se puede obtener de la tabla de todas las regresiones posibles, ya que ésta no admite sumas de variables entre sus opciones.

4. Pregunta 4

13,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

1 pt

Shapiro wilkes un método, no una P.H.

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis ~~Shapiro-Wilk:~~

$$H_0 : \varepsilon_i \sim \text{Normal vs.}$$
$$H_1 : \varepsilon_i \nsim \text{Normal}$$

Acompañado de un gráfico cuantil-cuantil:

colas muy pesadas

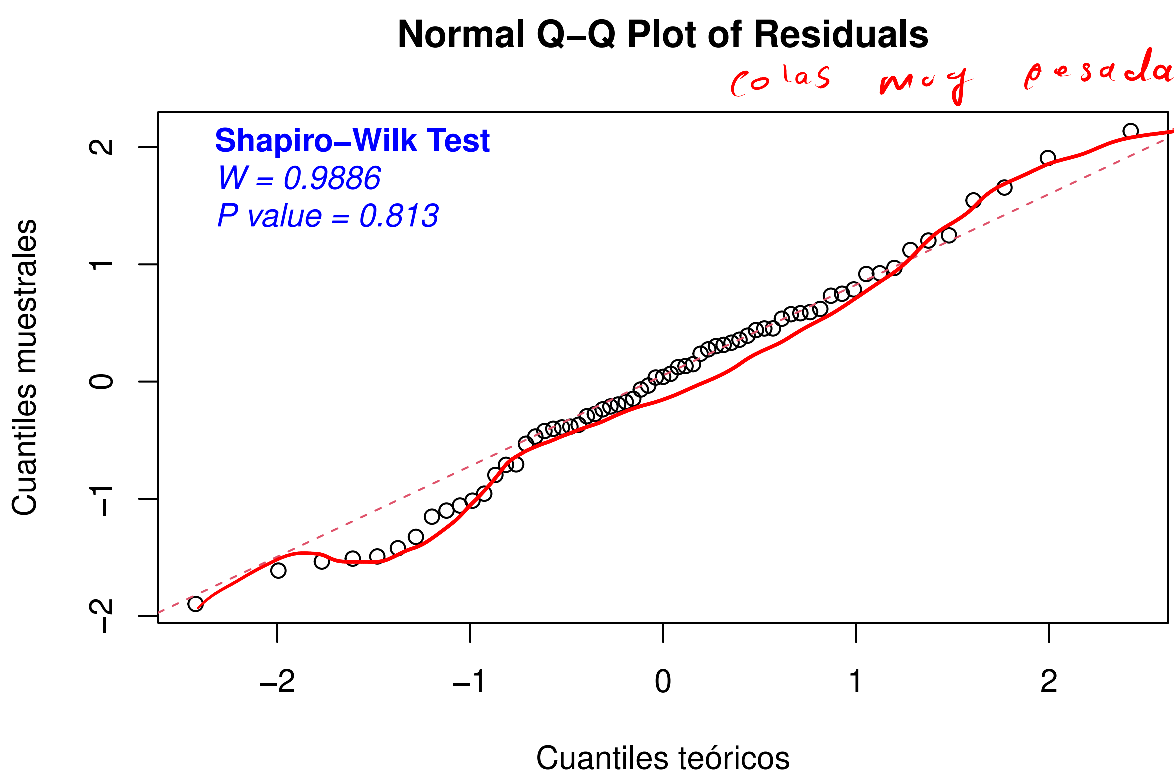

Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Como el patrón de los residuales (aproximadamente el ~~85 %~~ de los valores del centro) sigue la línea roja que representa el ajuste de la distribución de los residuales a una distribución normal, se concluye que el supuesto de normalidad se cumple. Lo cual se ratifica en el resultado de la prueba de normalidad Shapiro-Wilk con un Valor-P = $0.813 > \alpha = 0.05$. X

Nada que ver, todos los datos deben seguir la línea sin patrones y son mucho menos del 85% que lo hacen.

Gráfico es más importante que val-P

4.1.2. Varianza constante

1,5 pt

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis:

$$\begin{aligned} &H_0 : V[\varepsilon_i] = \sigma^2 \text{ vs.} \\ &H_1 : V[\varepsilon_i] \neq \sigma^2 \end{aligned}$$

Para ello se usa la siguiente gráfica:

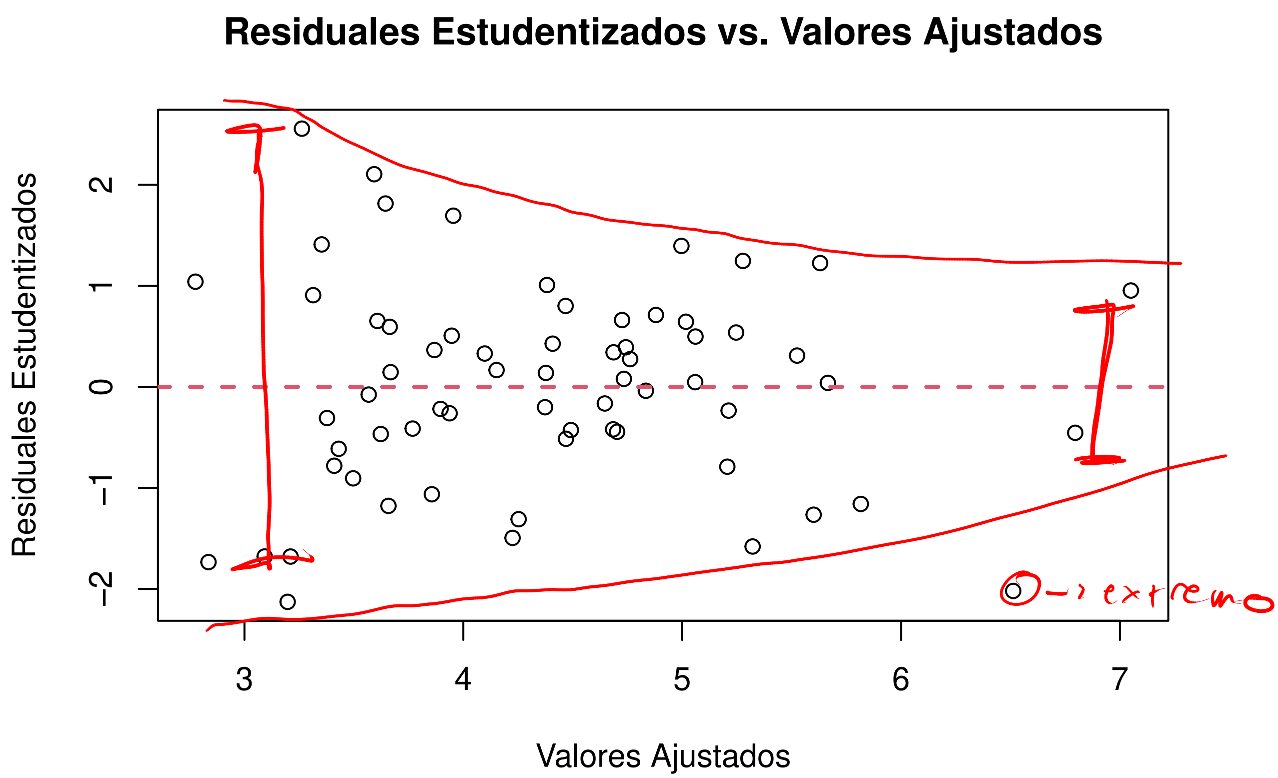

Figura 2: Gráfico residuales estudentizados vs. valores ajustados

Se observa una nube de puntos hasta el segundo tercio de la gráfica que puede indicar que el supuesto se cumple, pero en el resto de la gráfica se observan valores extremos alejados de la nube, por lo tanto afectan este supuesto.

En conclusión, se dice que el supuesto se cumple pero se advierte de la existencia de valores extremos alejados de la nube principal de datos. Además es posible observar media 0.

4.2. Verificación de las observaciones

Para identificar si en el modelo hay observaciones extremas, se deben calcular los estadísticos que nos permiten aplicar criterios en ese sentido, los cuales incluyen: residuales estudentizados, los valores de la diagonal de la matriz H (los h_{ii}), la distancia de Cook (D_i) y los DFFITS.

4.2.1. Datos atípicos

3e+

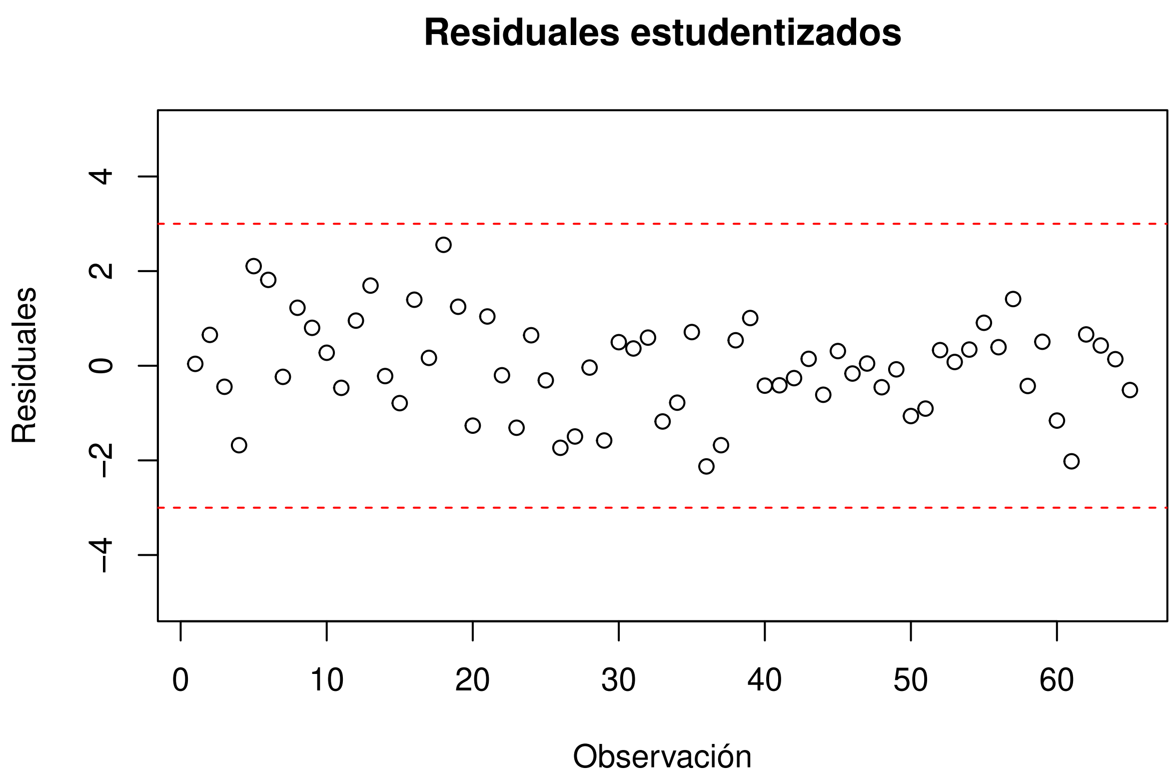

Figura 3: Identificación de datos atípicos

Se considera que una observación es atípica cuando su residual estudentizado r_i, es tal que: $|r_i| > 3$.

De acuerdo a la gráfica de residuales estudentizados se tiene que no hay observaciones atípicas.

4.2.2. Puntos de balanceo

2pt

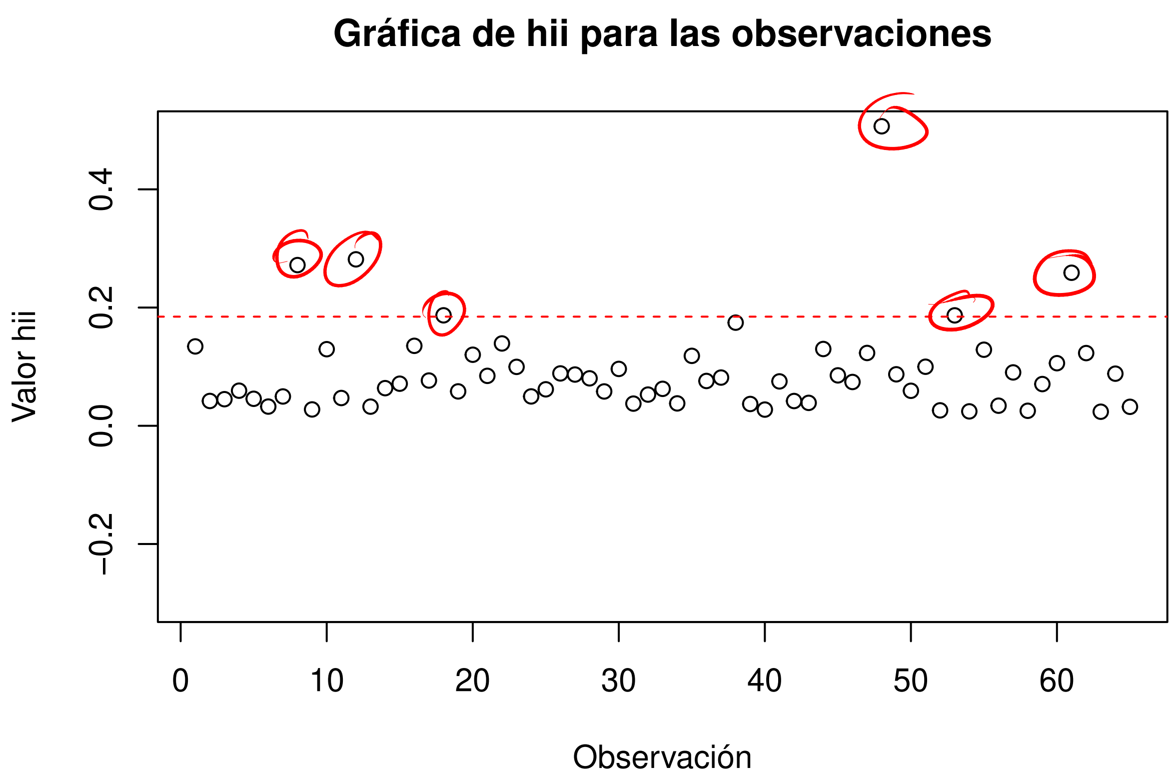

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Tabla de valores para el diagnóstico de puntos de balanceo

	Residuales Estudentizados	Distancia de Cook	Valores hii	Diagnóstico DFFITS
8	1.2251	0.0934	0.2719	0.7520
12	0.9536	0.0594	0.2816	0.5966
18	2.5557	0.2502	0.1869	1.2881
48	-0.4545	0.0354	0.5067	-0.4575
53	0.0793	0.0002	0.1868	0.0377
61	-2.0199	0.2375	0.2589	-1.2268

Se asume que la observación i es un punto de balanceo si $h_{ii} > 2\frac{p}{n}$.

Tenemos que: $h_{ii} > 2\frac{p}{n} = 2(\frac{6}{65}) = 0.1846$.

De acuerdo a la columna valores hii de la diagonal de la matriz H se tiene que las observaciones 8, 12, 18, 48, 53 y 61 son puntos de balanceo.

¿Qué causan los puntos de balanceo en el modelo?

4.2.3. Puntos influenciales

3 pt

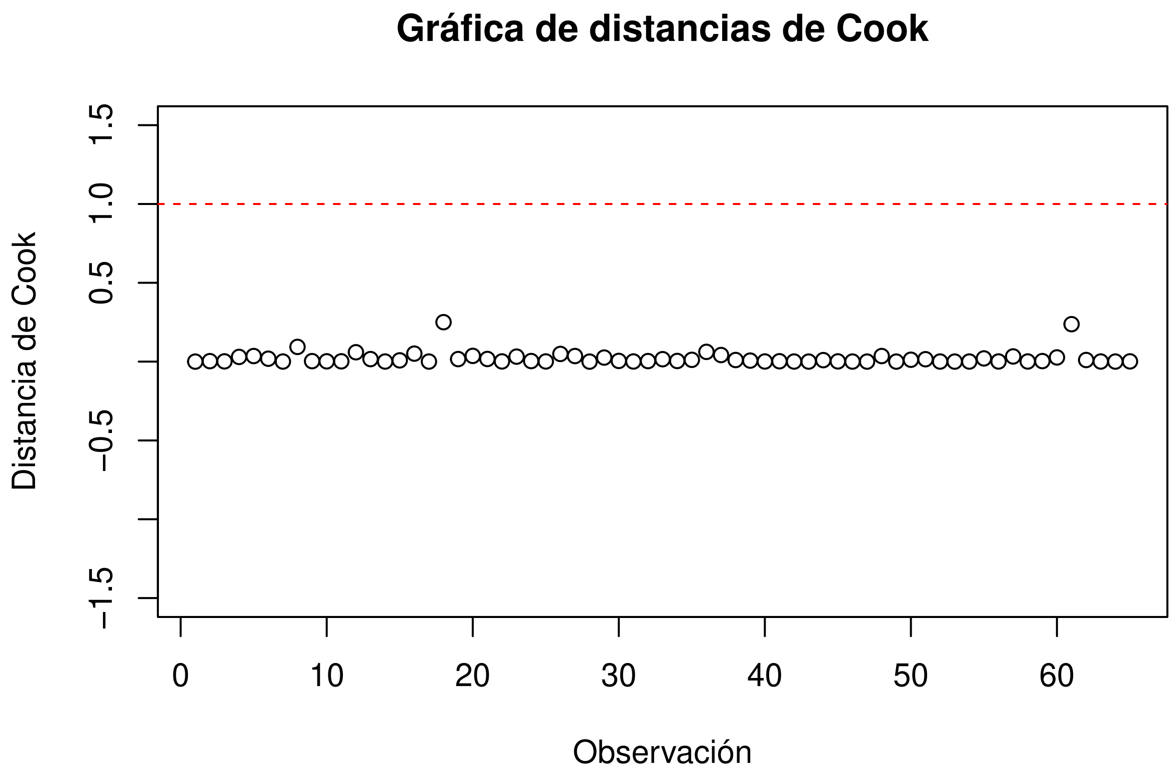

Figura 5: Criterio distancias de Cook para puntos influenciales

Se dice que la observación i será influencial si $D_i > 1$.

Según el criterio de Distancia de Cook no hay valores influenciales que afecten la estimación de los parámetros. → muy bien

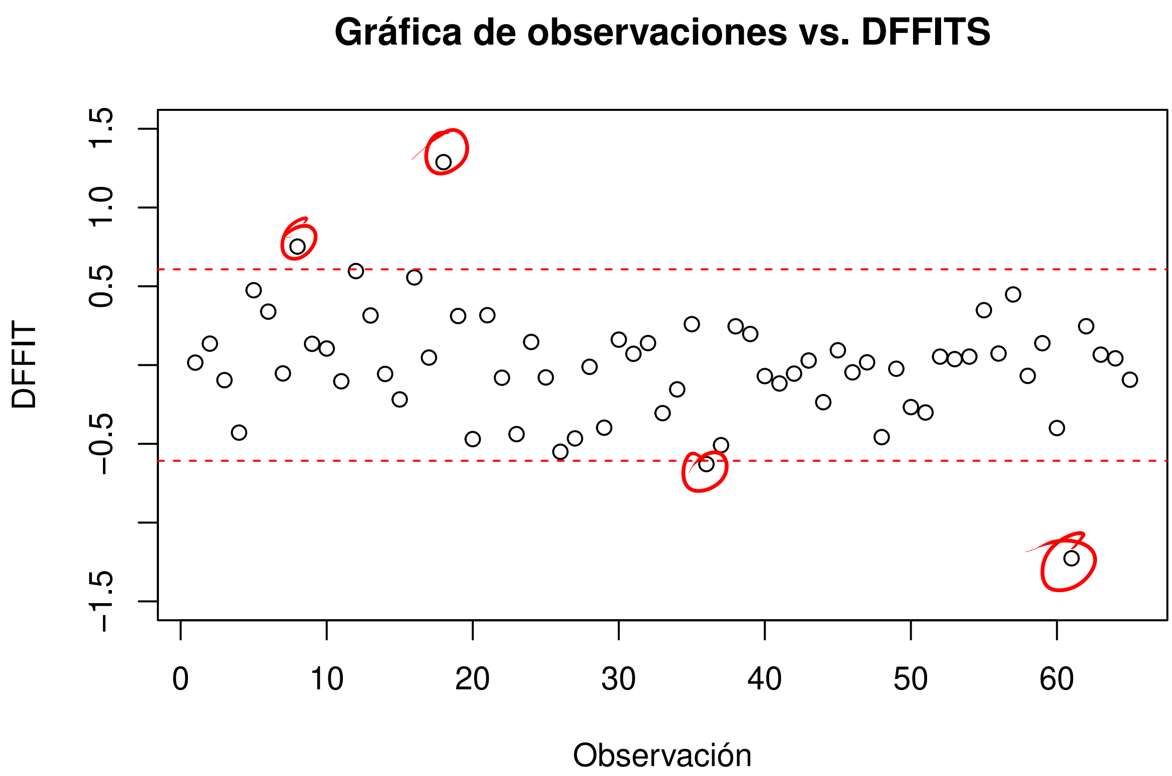

Figura 6: Criterio DFFITS para puntos influenciales

Cuadro 6: Tabla de valores para el diagnóstico de DFFITS

	Residuales Estudentizados	Distancia de Cook	Valores hii	Diagnóstico DFFITS
8	1.2251	0.0934	0.2719	0.7520
18	2.5557	0.2502	0.1869	1.2881
36	-2.1278	0.0619	0.0758	-0.6286
61	-2.0199	0.2375	0.2589	-1.2268

Se dice que la observación i será influencial si $[DFFITS_i] > 2\sqrt{\frac{p}{n}}$.

Tenemos que el criterio basado en DFFITS debe superar en valor absoluto a $2\sqrt{\frac{6}{65}} = 0.6076$.

De acuerdo a la columna Diagnóstico DFFITS tenemos que las observaciones 8, 18, 36 y 61 son influenciales.

¿Qué causan según este criterio?

5. Conclusiones

3pt

- El modelo es válido, ya que cumple con los supuestos de normalidad de los residuales y varianza constante. → Ninguno, pero al menos son congruentes
- En el punto 2, no se esperaba descartar las variables del subconjunto (X_2, X_4, X_5), ya que en el punto 1, el parámetro para (X_5) resultó ser significativos para el modelo de regresión lineal múltiple, lo que nos lleva a pensar que las otras dos variables (X_2, X_4) tuvieron mucho más peso al no ser significativas y así descatar por completo el subconjunto de variables. ✓
- Las observaciones 8, 18 y 61 son puntos de balanceo y puntos influenciales al mismo tiempo. Además, las observaciones 12, 48, y 53 solo son puntos de balanceo y la observación 36 sólo es un punto influencial. ✓