4,1

Trabajo 1

Estudiantes

Mateo Patiño Gómez
Jhon Alexander Valenzuela Benavides
Linder Yolian Rodriguez Cortes
Kevin Jair Quinones Sierra

Equipo 7

Docente

L **Julieth Veronica Guarin 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

17 pt

Teniendo en cuenta la base de datos del Equipo007, en la cual hay 5 variables regresoras definidas como:

Y: Riesgo de infección [%] X_1: Duración de la estadía [días] X_2: Rutina de cultivos [por cada 100] X_3: Número de camas X_4: Censo promedio diario X_5: Número de enfermeras

esta es regresora?

Entonces, se plantea el siguiente modelo de regresión lineal múltiple(RLM):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i;$$

donde

$$\varepsilon_i \overset{idd}{\sim} N(0, \sigma^2) \;\; 1 \leq i \leq 65$$

1.1. Modelo de regresión

3 pt

Al ajustar el modelo planteado según los datos, se obtiene la siguiente tabla de coeficientes

Cuadro 1: Valores de los coeficientes

	Valor del parametro
β_0	0.6103
β_1	0.1825
β_2	-0.0040
β_3	0.0312
β_4	0.0199
β_5	0.0008

Por ende, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.6103 + 0.1825X_{1i} - 0.004X_{2i} + 0.0312X_{3i} + 0.0199X_{4i} + 8 \times 10^{-4} X_{5i}$$

donde:

$$1 \leq i \leq 65$$

1.2. Significancia de la regresión 5 pt

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

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

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

Además, sea esta la tabla ANOVA:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	grados de libertad	Cuadrado medio	F_0	P-valor
Regresión	54.3368	5	10.867365	11.8384	5.99114e-08
Error	54.1607	59	0.917978		

De la tabla Anova, se observa que bajo un nivel de significacia del 5 %, valor p $< \alpha$, por lo que se rechaza la hipótesis nula en la que $\cancel{\beta_j = 0 \text{ con } 1 \leqslant j \leqslant 5}$, entonces al menos un parametro del modelo de regresión mútiple es diferente de 0, es decir, la regresión es estadísticamente significativa.

→ esto es individual

1.3. Significancia de los parámetros 4 pt

Primero observemos el juego de hipotesis para la prueba individual de la significancia de los parametros.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \text{ con } 0 \leqslant j \leqslant 5 \end{cases}$$

En el siguiente cuadro se presenta la Tabla de coeficientes, la cual permitirá, entre otras cosas, determinar cuáles de los parametros son significativos en nuestro modelo:

Cuadro 3: Tabla de los coeficientes

	Estimate	Std.error	T_{0j}	Valor P
β_0	0.6103	1.8419	0.3313	0.7416
β_1	0.1825	0.0818	2.2310	0.0295
β_2	-0.0040	0.0359	-0.1112	0.9118
β_3	0.0312	0.0150	2.0832	0.0416
β_4	0.0199	0.0079	2.5221	0.0144
β_5	0.0008	0.0007	1.0838	0.2829

Los respectivos valores P nos permiten concluir que con un nivel de significancia de $\alpha = 0.05$, los parámetros β_0, β_1 β_3 y β_4 son significativos, pues sus P-valores son menores a α, por lo que se rechaza H_0. × para β_0, 0,1416 les parece < que 0,05?

1.4. Interpretación de los parámetros 2 pt

Importante mencionar que β_0 no tiene interpretación pues no hay una coordenada

$$(X_{1i},\, X_{2i},\, X_{3i},\, X_{4i},\, X_{5i}) = (0,\, 0,\, 0,\, 0,\, 0)$$

probabilidad promedio

$\hat{\beta}_1$**:** Indica que por cada unidad que se aumente en la variable duración de estadía (X_1), el promedio de Riesgo de infección aumenta en 0.1825 unidades mientras las demás predictoras permanezcan fijas.

$\hat{\beta}_3$**:** Indica que por cada unidad que se aumente el número de camas (X_3), el promedio de Riesgo de infección aumenta en 0.0312 unidades mientras las demás predictoras permanecen fijas.

$\hat{\beta}_4$**:** Indica que por cada unidad que se aumente del censo promedio diario (X_4), el promedio de Riesgo de infección aumenta en 0.0199 unidades mientras las demás predictoras sean constantes.

1.5. Coeficiente de determinación múltiple R^2 3 pt

El modelo tiene un $R^2 = 0.5008$, se interpreta que el 50.08 % de la variabilidad total en ~~los resultados del~~ porcentaje de riesgo de infección es explicada por el modelo de regresión múltiple propuesto, es decir, nos indica una poca asociación lineal, pero esto no quiere decir que no se garantice los supuestos básicos del modelo, que se comprabarán más adelante.

2. Pregunta 2 3 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las variables regresoras con los valores P más alto en el modelo fueron $\boxed{X_1, X_2, X_4}$ (Para este juego de hipótesis B_0 no es tomado en cuenta), por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 3, 5 \end{cases}$$

No es cierto, además de que no es congruente con H0

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	54.161	X1 X2 X3 X4 X5
Modelo reducido	61.213	X1 X4

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

2.2. Estadístico de prueba y conclusión

2 pt

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_4) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \overset{H_0}{\sim} F_{3,59} \\ &= \frac{(61.213 - 54.161)/3}{0.917983} \\ &= 2.5607 \end{aligned} \tag{2}$$

1 pt

Ahora, comparando el F_0 con $f_{0.95,3,59} = 2.7608$, se puede ver que $F_0 < f_{0.95,3,59}$

Por lo tanto, no se rechaza H_0, es decir que las variables: Rutina de cultivos [por cada 100], número de camas y número de enfermeras pueden ser descartadas del modelo.

Conclusión directa es que no son signific. y por tanto

3. Pregunta 3

4 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si las variables predictoras X_2 y X_4 son colineales y las variables predictoras X_1 y X_3 presentan colinealidad en el modelo. por consiguiente se plantea la siguiente prueba de hipótesis:

No es una pregunta de colinealidad => $X_2 = \beta_0 + \beta_4 X_4$

$$\begin{cases} \text{H}_0 : \beta_2 = \beta_4\ ,\ \ \beta_1 = \beta_3 \\ \text{H}_1 : \text{Al menos una de las igualdades no se cumple} \end{cases}$$

lo que es equivalente a lo siguiente:

$$\begin{cases} \text{H}_0 : \beta_2 - \beta_4 = 0,\ \ \beta_1 - \beta_3 = 0 \\ \text{H}_1 : \text{Al menos una de las igualdades no se cumple} \end{cases}$$

2 pt

reescribiendo matricialmente:

$$\begin{cases} \mathrm{H}_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ \mathrm{H}_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Con **L** dada por

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

0,5 pt

El modelo reducido está dado por:

n=65

$$Y_i = \beta_o + \beta_1 X_{1i}^* + \beta_2 X_{2,i}^* + \beta_5 X_{5,i} + \varepsilon_i \ , \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \ 1 \leqslant i \leqslant 50$$

Donde $X_{1i}^* = X_{1i} + X_{3i}$ y $X_{2i}^* = 3X_{2i} + X_{4i}$

3.2. Estadístico de prueba

1,5 pt

reemplazar lo que conocen

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,59} \tag{3}$$

4. Pregunta 4

17 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

3 pt

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis ~~shapiro-wilk,~~ acompañada de un gráfico cuantil-cuantil:

$$\begin{cases} \mathrm{H}_0 : \varepsilon_i \sim \text{Normal} \\ \mathrm{H}_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$

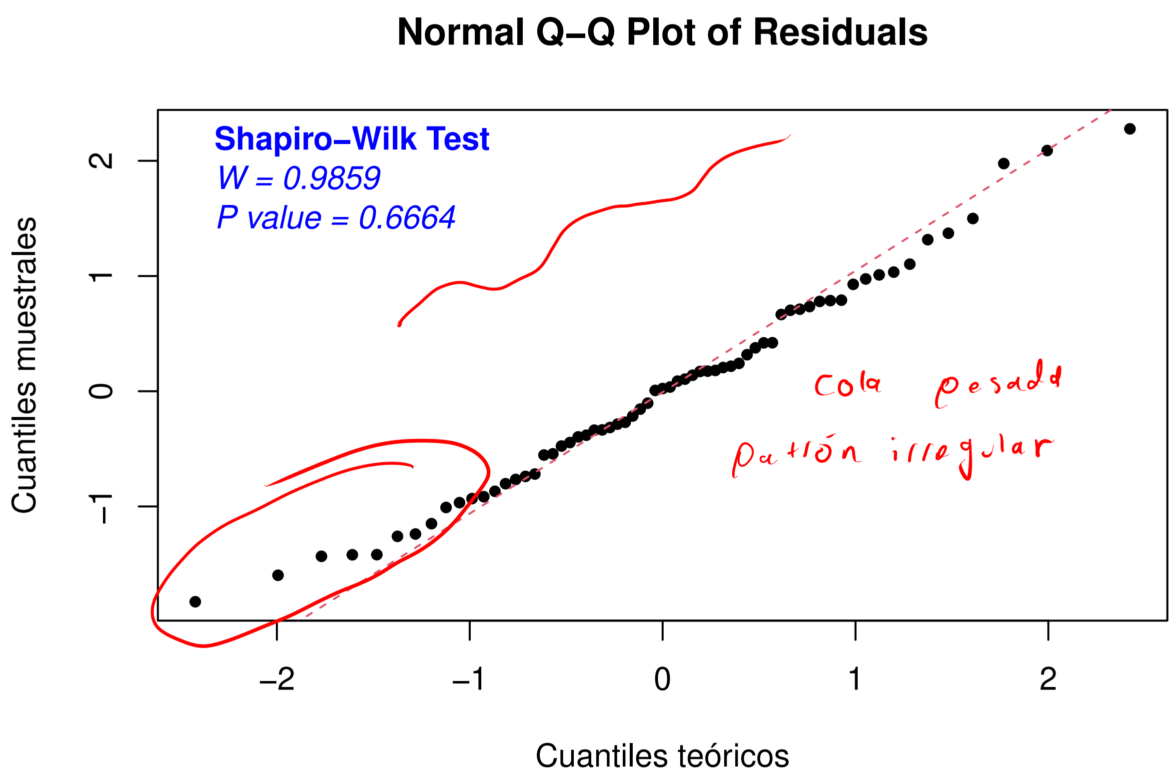


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

¿Lo hacen o no?

Al ser el P-valor aproximadamente igual a 0.6664 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal, aunque hay desviaciones considerables en los dos extremos, pues hay varios datos alejados de la línea roja, por lo que se identifican como posibles datos atípicos, de balanceo o de influencia.

¿No es cualquier línea

Ahora se validará si la varianza cumple con el supuesto de ser constante.

No terminan de decir gráficamente si distribuye normal, criterio gráfico es más importante.

4.1.2. Varianza constante

3pt

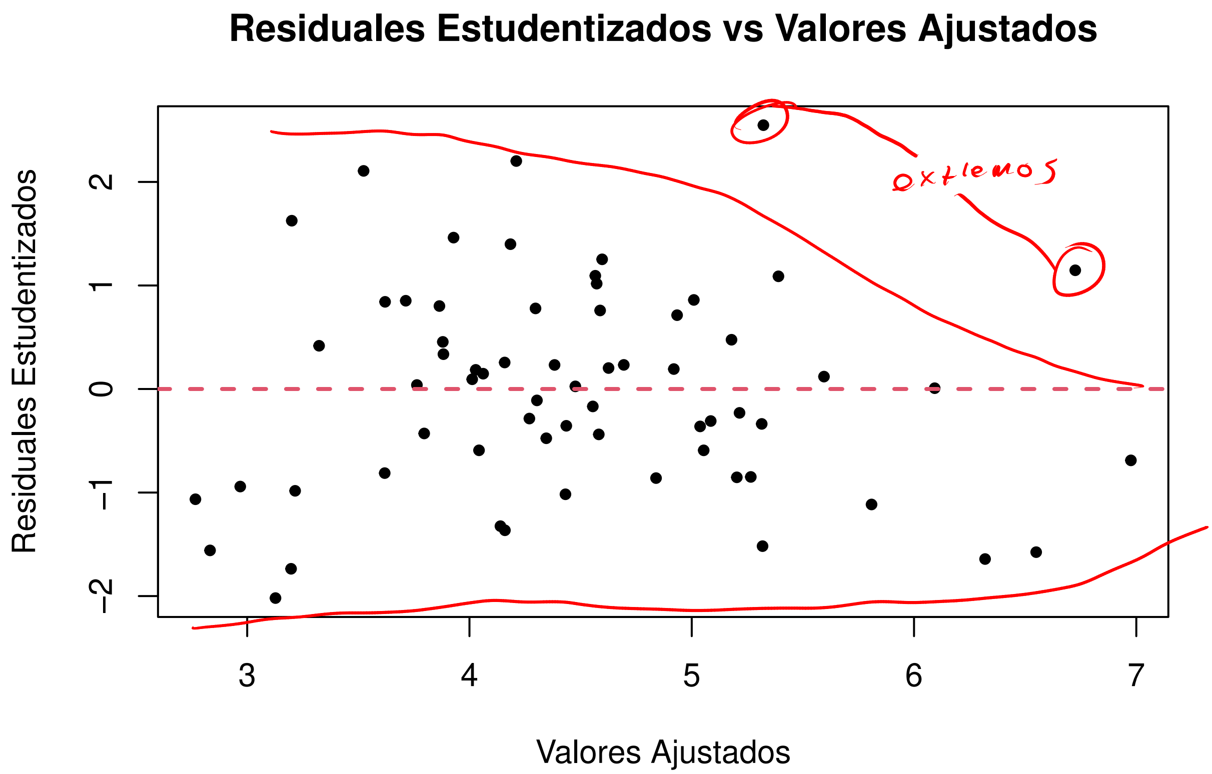


Figura 2: Gráfico residuales estudentizados vs valores ajustados

En la gráfica anterior se observa que la varianza no tiene una tendencia constante, es decir el H_0 no se cumple, pues se observa un aumento de la dispersión hasta el centro de la gráfica y luego se presenta un leve decrecimiento en los puntos. Esto podría ser debido a la posible presencia de datos extremos. ✓

¿cuál Ho que nunca lo plantearon?

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3.07

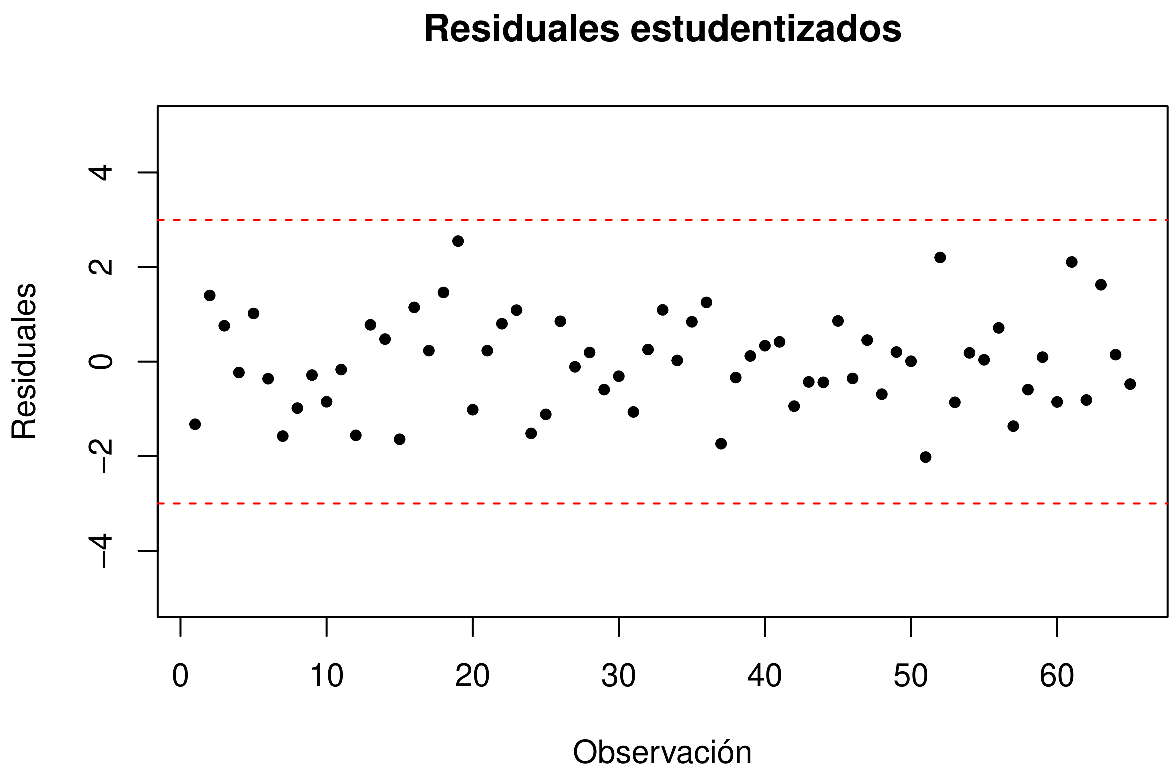


Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica de dispersión anterior, no hay datos atípicos en el modelo, ya que ninguno de los datos se encuentra por fuera del rango (-3,3), es decir que no cumplen que -3 $< r_i >$ 3.

4.2.2. Puntos de balanceo

2,5 pt

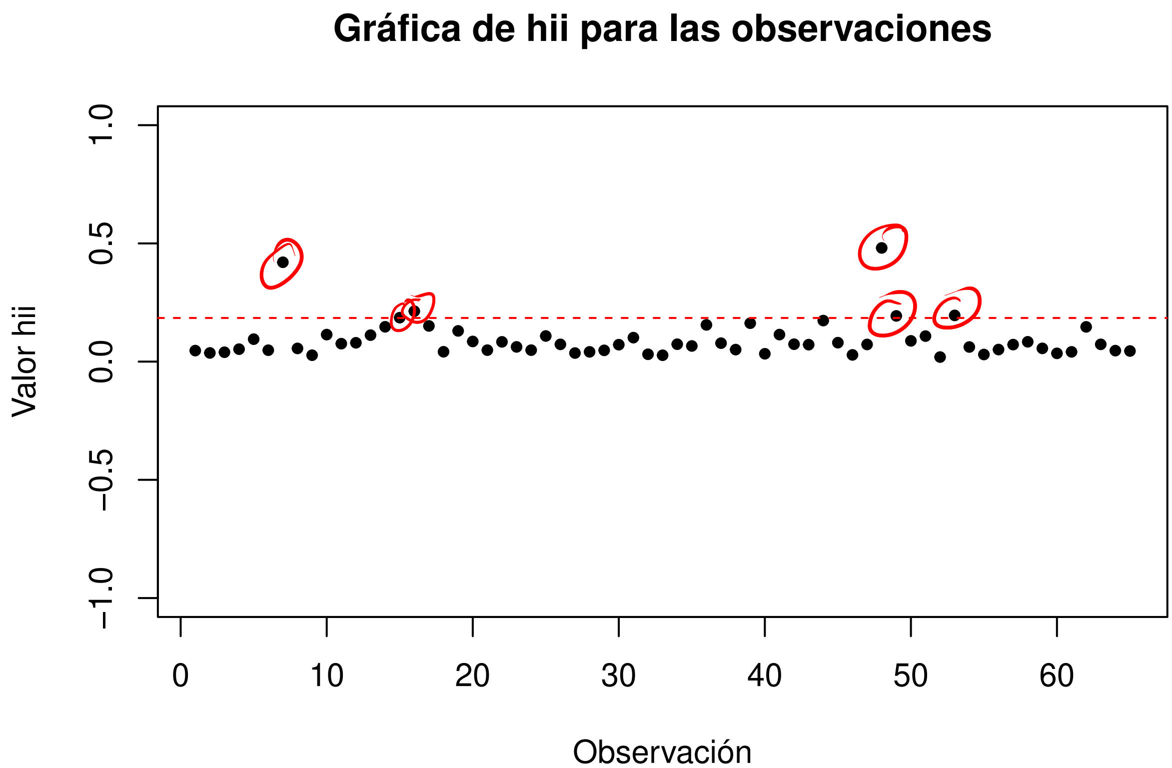


Figura 4: Identificación de puntos de balanceo

	hii
7	0.4202
15	0.1859
16	0.2133
48	0.4808
49	0.1933
53	0.1955

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, es decir $h_{ii} = 0.1846$, se reconocen 6 puntos de balanceo bajo el criterio que su respectivo $h_{ii} > 0.1846$, los cuales están presentados en la tabla

¿Qué causan estos puntos en el modelo?

4.2.3. Puntos influenciales

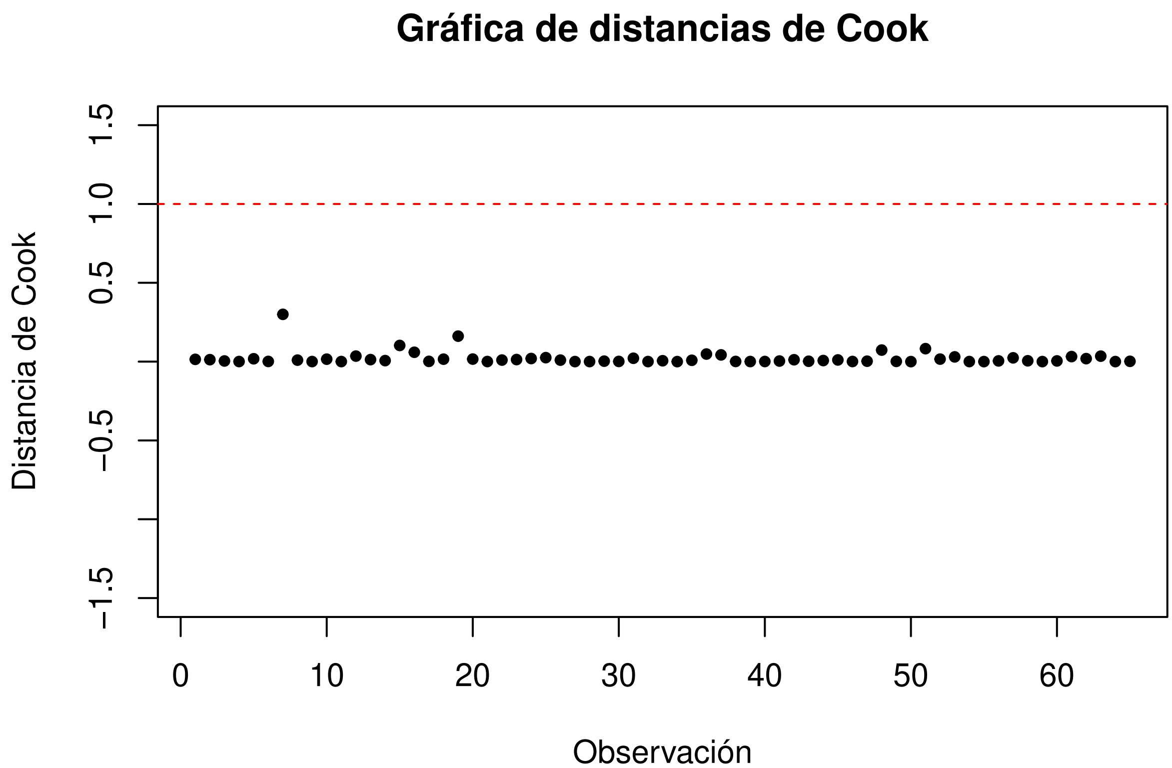
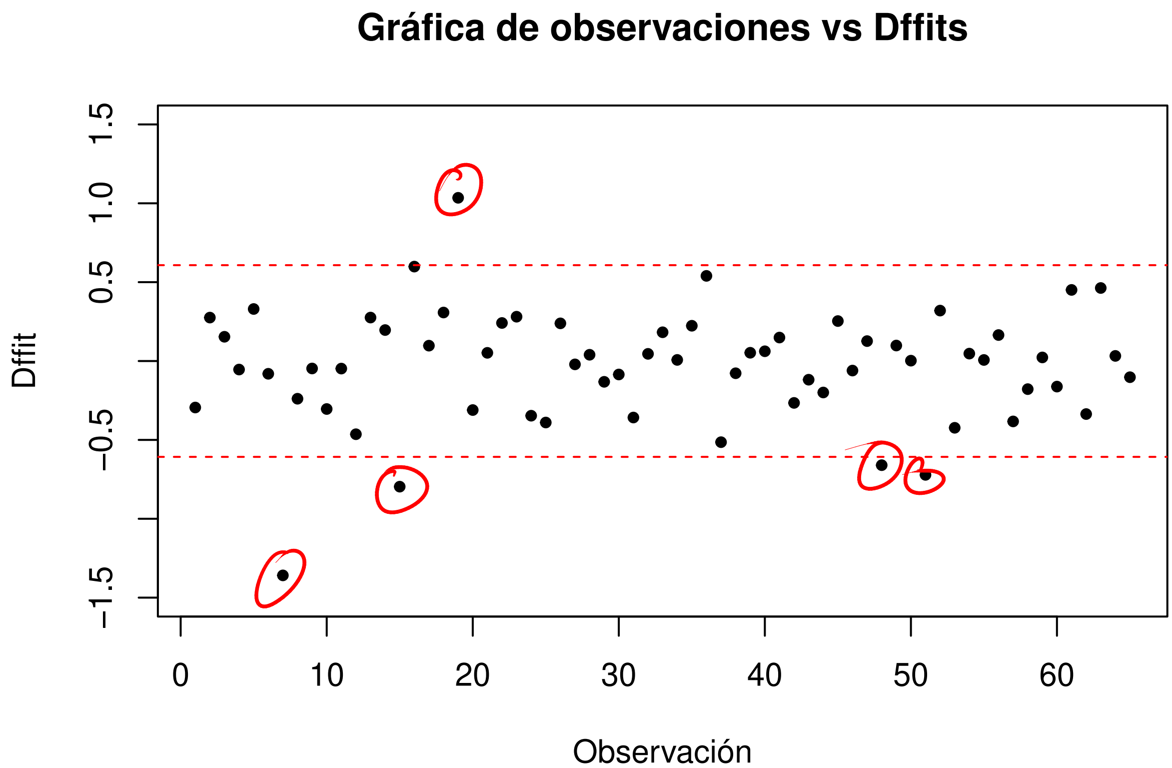


Figura 5: Criterio distancias de Cook para puntos influenciales

No dejen esto espacio así



3 pt

Figura 6: Criterio Dffits para puntos influenciales

	Dffits
7	-1.3589
15	-0.7964
19	1.0348
48	-0.6603
51	-0.7217

¿cuánto da?

Como se puede ver,las observaciones $\{7, 15, 19, 48, 51\}$ son puntos influenciales según el criterio de Dffits, pues si $|D_{ffit}| > 2\sqrt{\frac{6}{65}}$, es un punto influencial, lo cual puede verse representado en la tabla. Cabe destacar también que con el criterio de distancias de Cook, en el cual para cualquier punto cuya $D_i > 1$, es un punto influencial, ninguno de los datos cumple con serlo.

¿Qué causan estos puntos

4.3. Conclusión

2,5 pt

En conclusión, respecto a la validez del modelo podemos decir que este no es válido debido a que no se cumple uno de los dos principales supuestos del modelo, como lo es el de varianza constante igual a σ^2debido a la presencia de distintas observaciones extremas, para este caso lo son puntos influenciales y de balanceo, donde algunos de estos cumplen el criterio de ser tanto de balanceo como influenciales y esto afecta el ajuste en el modelo (como se explicó con la gráfica para el supuesto de normalidad aunque el Pvalue $> \alpha$ de la prueba Shapiro-Wilk) y a su vez también al supuesto de la varianza. → No afectan sólo eso, es más, los influenciales no necesariamente lo hacen.

→ ¿cuándo demostraron esta aseveración?