Trabajo 1

4,7

Estudiantes

Santiago Blandón Gutiérrez
Melida Maribel Vásquez Ordoñez
Johari Daniela Ardila Preciado
Victor David Usuga Duque

Equipo 48

Docente

Carlos Mario Lopera

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

18 pt

Teniendo en cuenta la base de datos brindada, en la cual hay 64 muestras y 5 variables regresoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Las variables son:

- Y: Riesgo de infeccion
- X_1: Duracion de la estadia (días)
- X_2: Rutina de cultivos
- X_3: Número de camas
- X_4: Censo promedio diario
- X_5: Número de enfermeras

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes coeficientes $\hat{\beta}$:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.2875
β_1	0.2735
β_2	0.0016
β_3	0.0538
β_4	0.0079
β_5	0.0017

Por lo tanto, el modelo de regresión ajustado es:

2 pt

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i}$$

Reemplacen

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2, 3, 4, 5} \end{cases}$$

Con un estadístico de prueba de la forma:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,58} \tag{1}$$

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	59.4077	5	11.881542	15.1683	1.62985e-09
Error	45.4321	58	0.783313		

De la tabla Anova, se observa un valor P aproximadamente igual a 0, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$. Se acepta la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto existe relación de regresión y la RLM es significativa.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros $\hat{\beta}$, este cuadro se uso para determinar cuáles parámetros son significativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.2875	1.6570	-0.1735	0.8628
β_1	0.2735	0.0868	3.1524	0.0026
β_2	0.0016	0.0306	0.0527	0.9582
β_3	0.0538	0.0138	3.8931	0.0003
β_4	0.0079	0.0066	1.1817	0.2421
β_5	0.0017	0.0007	2.3944	0.0199

De los P-valores presentes en la tabla se concluye que para un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3 y β_5 son significativos individualmente, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

El primer paso es identificar aquellos parámetros susceptibles de interpretación, es decir, solo se podrán interpretar parámetros que resultaron significativos individualmente, como ya lo vimos, en este caso son: β_1, β_3, β_5.

3pt

- β_1 : 0,2735 indica que ante un aumento en una unidad de duración promedio en la estadía de todos los pacientes en el hospital (en días), la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) aumenta en 0,2735 %, cuando las demás predictoras se mantienen fijas.
- β_3 : 0,0538 indica que ante un aumento en una unidad promedio de camas en el hospital durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) aumenta 0,0538 %, cuando las demas predictoras se mantienen fijas.
- β_5 :0,0017 indica que ante un aumento en una unidad del número promedio de enfermeras presentes a tiempo completo en el hospital durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) aumenta en 0,0017 %, cuando las demas predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5666$, lo que significa que aproximadamente el 56.66 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto.

¿cómo se calcula?

2. Pregunta 2

3,5 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más pequeño en el modelo fueron X_1, X_3 y X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	45.432	X1 X2 X3 X4 X5
Modelo reducido	82.645	X2 X4

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_{2i} + \hat{\beta}_4 X_{4i}$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{12.4043}{0.7833} \\ &= 15.8357 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,58} = 2.7636$, se puede ver que $F_0 > f_{0.95,3,58}$ y por tanto qué se rechaza H_0. De esto se concluye que estas variables tienen significancia simultánea y no se pueden descartar del subconjunto.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si ¿Se quiere saber si X_1, X_3 y X_5 contribuyen a Y cuando β_1 es 3 veces β_3 y β_5 sea 2 veces β_1? por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 3\beta_3;\ 2\beta_1 = \beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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 $\mathbf{L}$ dada por

$$L = \begin{bmatrix} 0 & 1 & 0 & -3 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & -1 \end{bmatrix}$$

2 pt

El modelo reducido está dado por:

$$Y_i = \beta_o + 3\beta_3 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + 6\beta_3 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

1 pt

$$Y_i = \beta_o + \beta_2 X_{2i} + \beta_3 [3X_{1i} + X_{3i} + 6X_{5i}] + \beta_4 X_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

3.2. Estadístico de prueba

2 pt

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 45.4321)/2}{0.783313)} \overset{H_0}{\sim} f_{2,58} \tag{3}$$

4. Pregunta 4

20 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

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

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

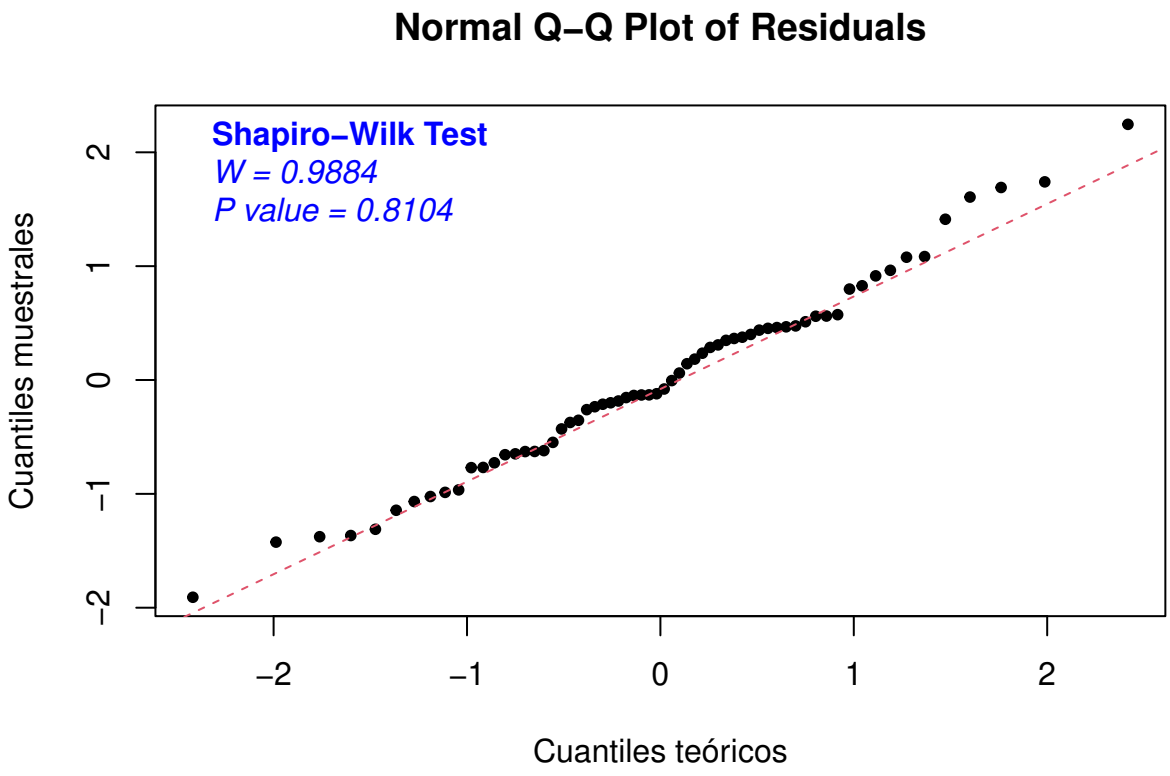

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.8104 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 con media μ y varianza σ^2, sin embargo la gráfica de comparación de cuantiles permite ver colas más pesadas y patrones irregulares, al tener más poder el análisis gráfico, se termina por rechazar el cumplimiento de este supuesto. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

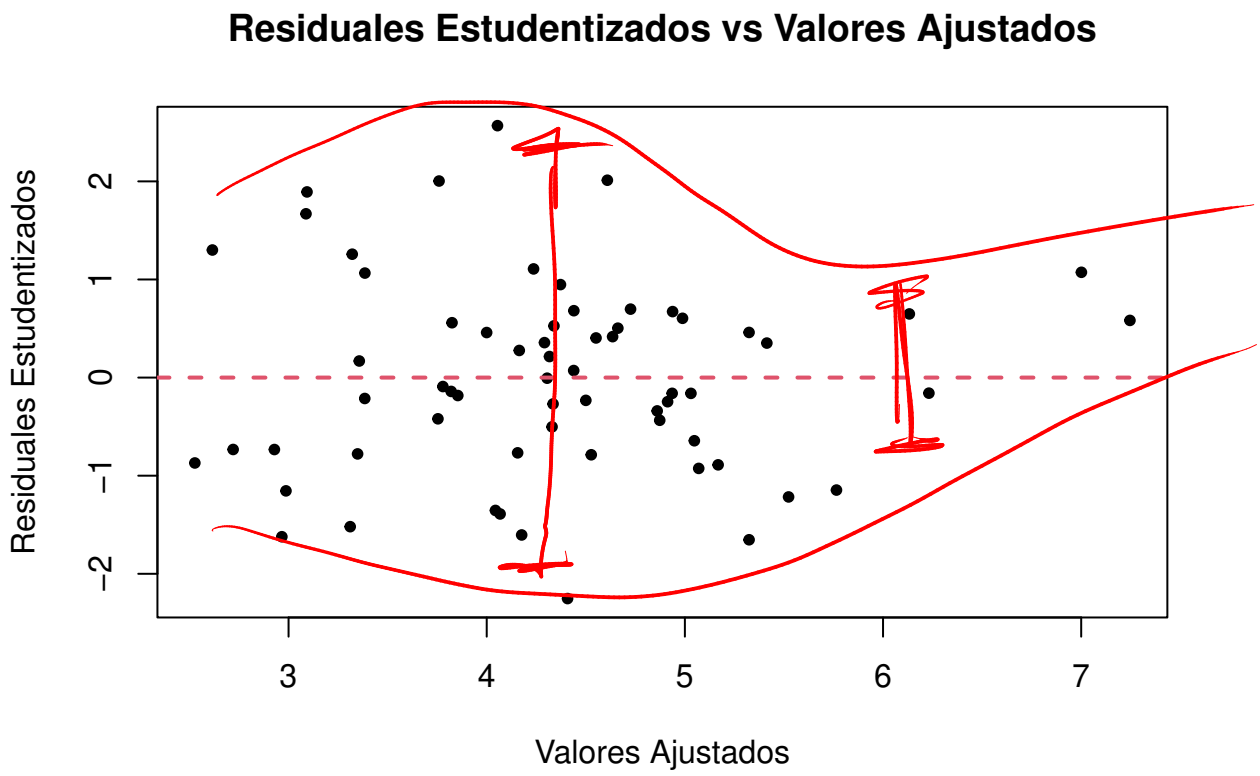

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que la dispersión de los puntos inicia en una concentración a la izquierda cercana al menos 1, que se transforma en un aumento de la dispersion hacía los extremos y se mantiene algo constante hasta la mitad del gráfico, luego se nota decrecimiento en dirección al centro de la figura, es decir, hay patrones en los que la varianza aumenta o disminuye que permiten rechazar el supuesto de varianza contanstante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

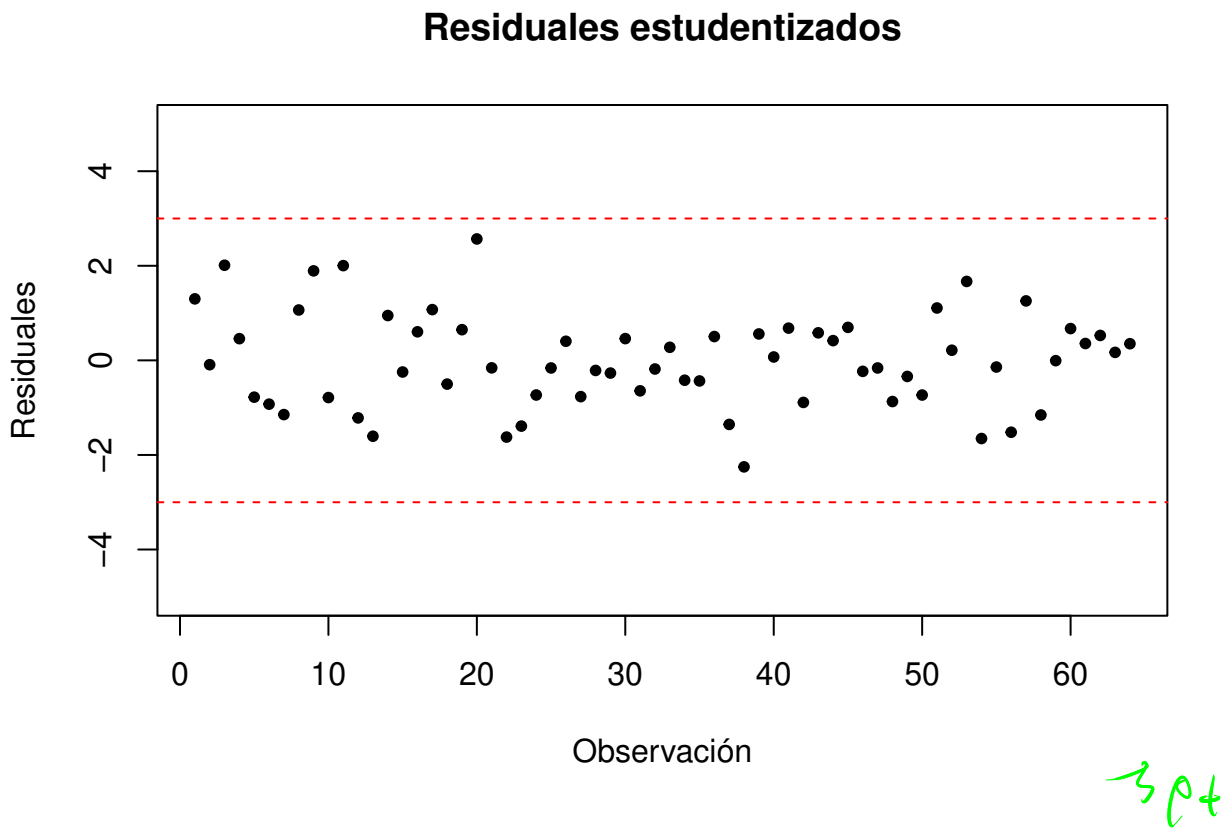

Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica anterior, no hay datos atípicos en el conjunto de datos pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

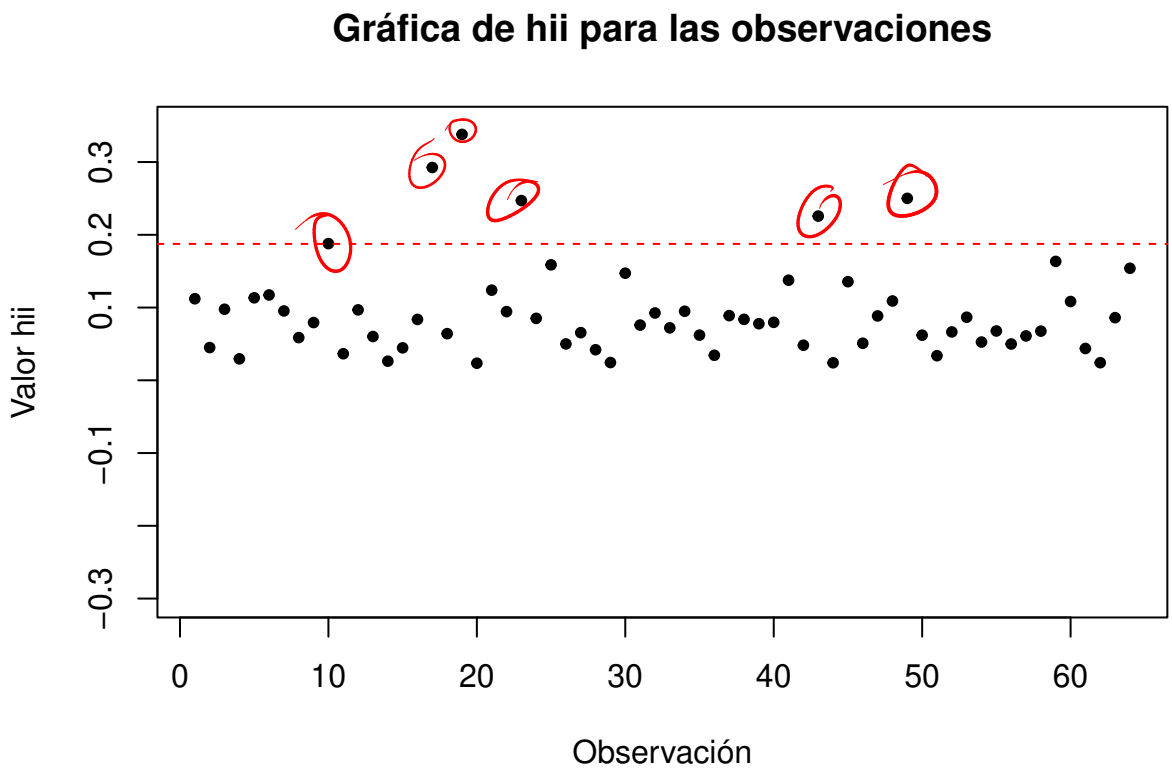

Figura 4: Identificación de puntos de balanceo

	res.stud	Cooks.D	hii.value	Dffits
10	-0.7868	0.0239	0.1879	-0.3772
17	1.0730	0.0793	0.2925	0.6908
19	0.6488	0.0358	0.3380	0.4613
23	-1.3900	0.1056	0.2470	-0.8026
43	0.5828	0.0165	0.2257	0.3129
49	-0.3398	0.0064	0.2501	-0.1947

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$ con $h_{ii} = 0.1874$, se puede apreciar que existen 6 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$, los cuales son los presentados en la tabla de arriba.

4.2.3. Puntos influenciales

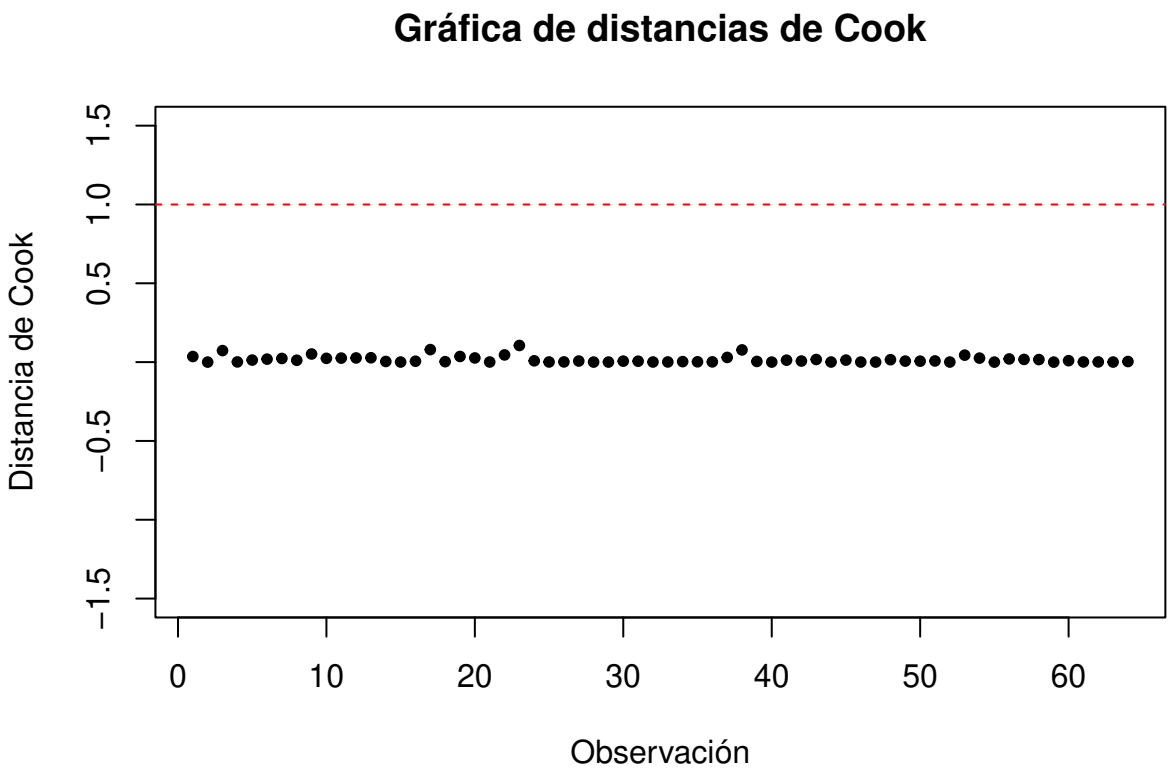

Figura 5: Criterio distancias de Cook para puntos influenciales

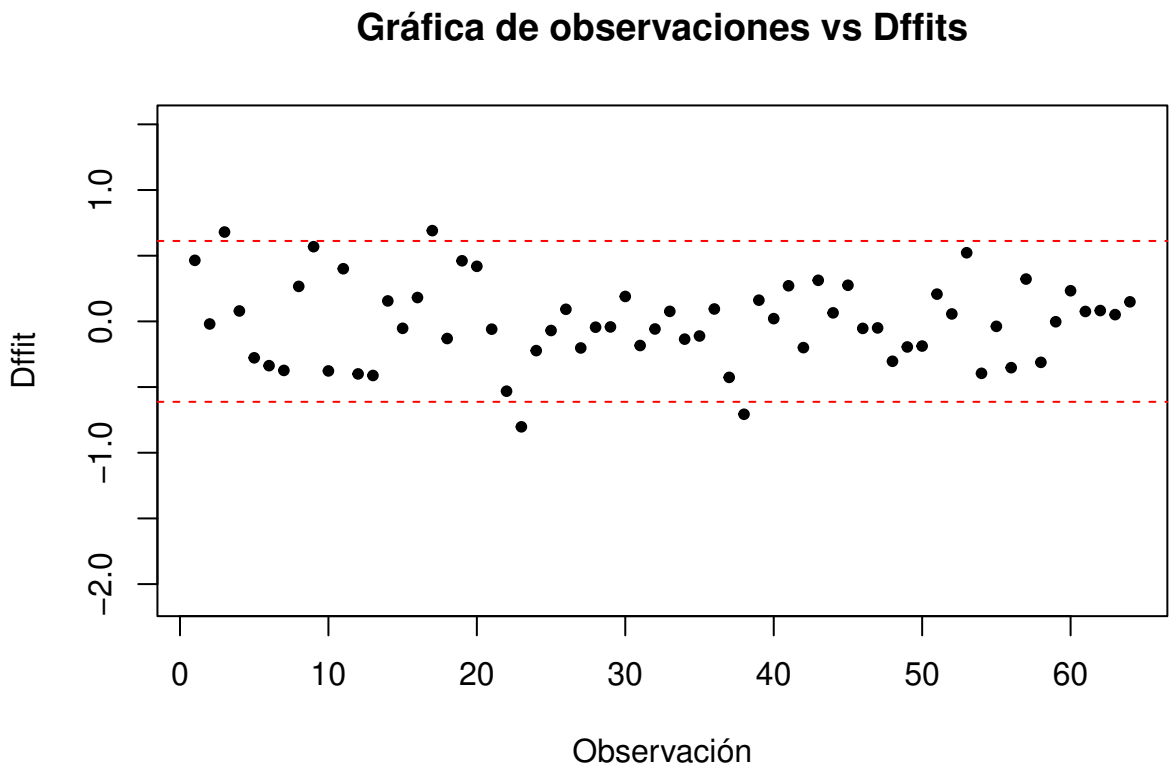

Figura 6: Criterio Dffits para puntos influenciales

	res.stud	Cooks.D	hii.value	Dffits
3	2.0115	0.0730	0.0977	0.6804
17	1.0730	0.0793	0.2925	0.6908
23	-1.3900	0.1056	0.2470	-0.8026
38	-2.2524	0.0774	0.0838	-0.7070

9 pt

Como se puede ver, las observaciones 3, 17, 23, 38 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$ y calculando $2\sqrt{\frac{p}{n}} = 0.61237$ se corrobora que son puntos influenciales. 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, ninguna observación cumple con este criterio.

4.3. Conclusión

3 pt

Como vimos anteriormente, el supuesto de normalidad y varianza constante de los errores del modelo no se cumple, por lo tanto, el modelo no es valido para hacer estimaciones y predicciones. Es importante resaltar que estos procesos de prueba se hicieron aún considerando los posibles valores extremos que pudiese tener el modelo. Para identificar

justamente estos valores que pueden alterar el modelo, se parte del análisis de observaciones extremas que acabamos de desarrollar, donde se obtuvo:

- Ninguna de las observaciones es atípica.
- Las observaciones 10, 17, 19, 23, 43, 49 son puntos de balanceo, por lo tanto estas son observaciones en el espacio de las variables predictoras que están alejadas del resto de la muestra y posiblemente afectan al R^2 y los Se de los coeficientes estimados.
- Las observaciones 3, 17, 23, 38 son influenciales, en consecuecia, son observaciones que tienen un impacto notable sobre los coeficientes de regresión ajustados. Es así, una observación que hala al modelo en su dirección.

Se apreció que en efecto se detectó la presencia de observaciones extremas que tendrán que ser estudiadas antes de usar el modelo y así evaluar de nuevo su validez como predictor o estimador de valores de la respuesta.