4,6

Trabajo 1

Estudiantes

María Fernanda Álvarez Restrepo
Manuel Alejandro Arriola Grisales
Lenny Alejandro Carvajal Taborda
Laura Lopera Zuleta

Equipo 24

Docente

Mateo Ochoa Medina

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

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

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

Donde:

- Y: Riesgo de infección
- X_1:Duración de la estadía en el hospital
- 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:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-2.0519
β_1	0.1035
β_2	0.0486
β_3	0.0372
β_4	0.0204
β_5	0.0020

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -2.0519 + 0.1035 X_{1i} + 0.0486 X_{2i} + 0.0372 X_{3i} + 0.0204 X_{4i} + 0.002 X_{5i} \quad 1 \leqslant i \leqslant 64$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

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

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	49.6816	5	9.936329	11.6965	7.72289e-08
Error	49.2719	58	0.849516		

De la tabla Anova, se observa un valor P muy cercano a 0, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, por ende, se acepta la hipótesis alternativa en la que algún $\beta_j \neq 0$, de allí se afirma que la regresión es significativa, lo cual quiere decir que el riesgo de infección depende significativamente de al menos una de las predictoras del modelo.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar cuáles de ellos son significativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-2.0519	1.7304	-1.1858	0.2406
β_1	0.1035	0.1003	1.0317	0.3065
β_2	0.0486	0.0291	1.6693	0.1005
β_3	0.0372	0.0128	2.9002	0.0053
β_4	0.0204	0.0067	3.0302	0.0036
β_5	0.0020	0.0007	2.7641	0.0076

Si tomamos en cuenta un valor de $\alpha = 0.05$, los P-valores presentes en la tabla permiten concluir que los parámetros β_3, β_4 y β_4 son significativos, pues su valor-P es menor a 0.05. Por otro lado, se halla que β_0, β_1 y β_2 son individualmente no significativos en presencia de los demás parámetros.

1.4. Interpretación de los parámetros

De acuerdo con el punto anterior, se interpretan β_3, β_4 y β_5.

$\hat{\beta}_3$**:** 0.0372 indica que por cada unidad que se aumente el número de camas, el promedio del riesgo de infección aumenta en 0.0372 unidades, cuando las demás predictoras se mantienen fijas

$\hat{\beta}_4$**:** 0.0204 indica que por cada unidad que se aumente el censo promedio diario, el promedio del riesgo de infección aumenta en 0.0204 unidades, cuando las demás predictoras se mantienen fijas

$\hat{\beta}_5$**:** 0.0020 indica que por cada unidad que se aumente el número de enfermeras, el promedio del riesgo de infección aumenta en 0.0020 unidades, cuando las demás predictoras se mantienen fijas

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5021$, lo que significa que aproximadamente el 50 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más pequeño en el modelo fueron X_4, X_3, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 3, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	49.272	X1 X2 X3 X4 X5
Modelo reducido	74.328	X1 X2

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

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

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,58} \\
&= \frac{8.352}{0.8495} \\
&= 9.8317
\end{aligned} \tag{2}
$$

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 se rechaza H0, de ahí que no es posible descartar las variables del subconjunto. Entonces, el riesgo promedio de infección depende de al menos una de las variables mencionadas.(número de camas, censo promedio diario y número de enfermeras)

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si ¿El efecto del censo promedio diario es el doble del efecto de la rutina de cultivos sobre el riesgo de infección? y ¿El efecto del número de camas es igual al efecto del número de enfermeras sobre el riesgo de infección?. Por consiguiente se plantea la siguiente prueba de hipótesis:

$$
\begin{cases}
\mathrm{H}_0 : \beta_4 = 2\beta_2;\ \beta_3 = \beta_5 \\
\mathrm{H}_1 : \text{Alguna de las igualdades no se cumple}
\end{cases}
$$

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

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

El modelo reducido está dado por:

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

Donde $X_{2i}^* = X_{2i} + 2X_{4i}$ y $X_{3i}^* = X_{3i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

2 pt

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

$$F_0 = \frac{(SSE(MR) - 49.2719)/2}{0.849516} \overset{H_0}{\sim} f_{2,58} \tag{4}$$

4. Pregunta 4

18 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

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

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

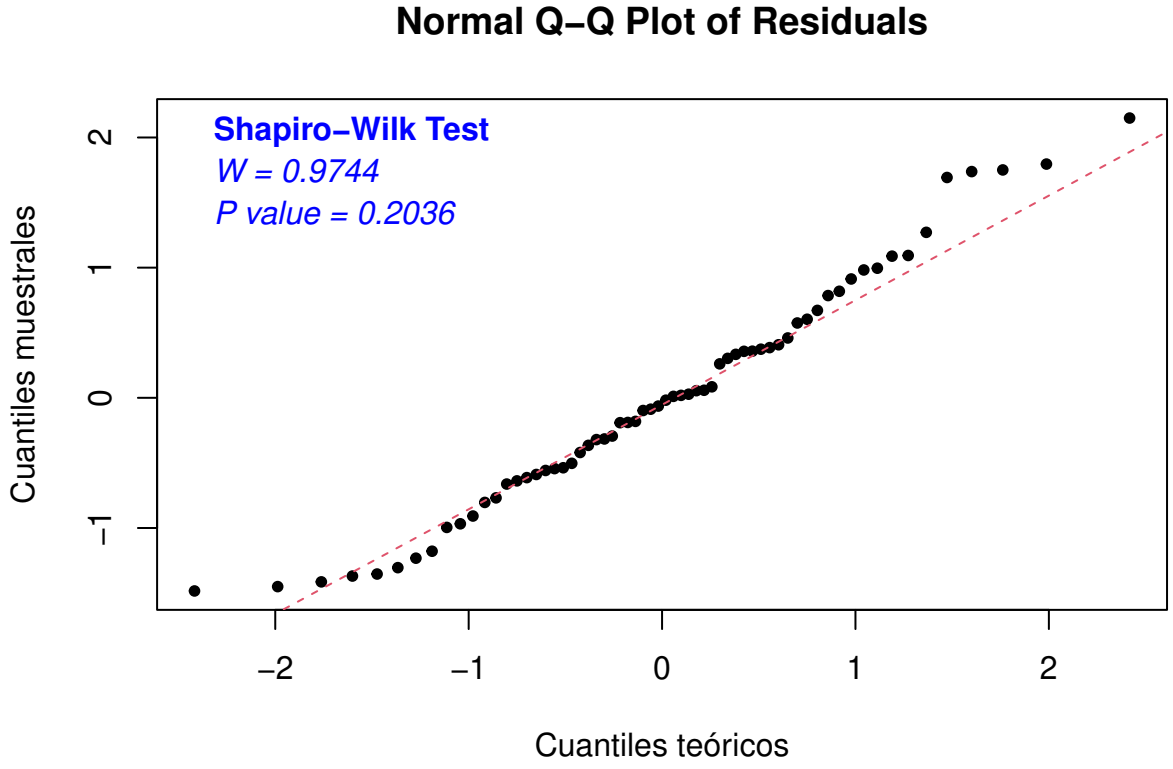

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

En base a la prueba de Shapiro-Wilk se obtiene un P-valor aproximadamente igual a 0.2036 y tomando en cuenta un $\alpha = 0.05$ y ya que el P-valor es mucho mayor se podria concluir que no se rechaza la hipotesis nula, es decir que los residuales se distribuyen de forma normal con media μ y varianza σ^2 pero a pesar de esto en el grafico se puede observar un sesgo positivo en la distribucion de los residuales junto con patrones irregulares en el centro y las colas indicando que los errores no se distribuyen de forma normal, debido a esto se rechaza la hipotesis nula y se concluye que no se distribuyen de forma normal.

4.1.2. Varianza constante

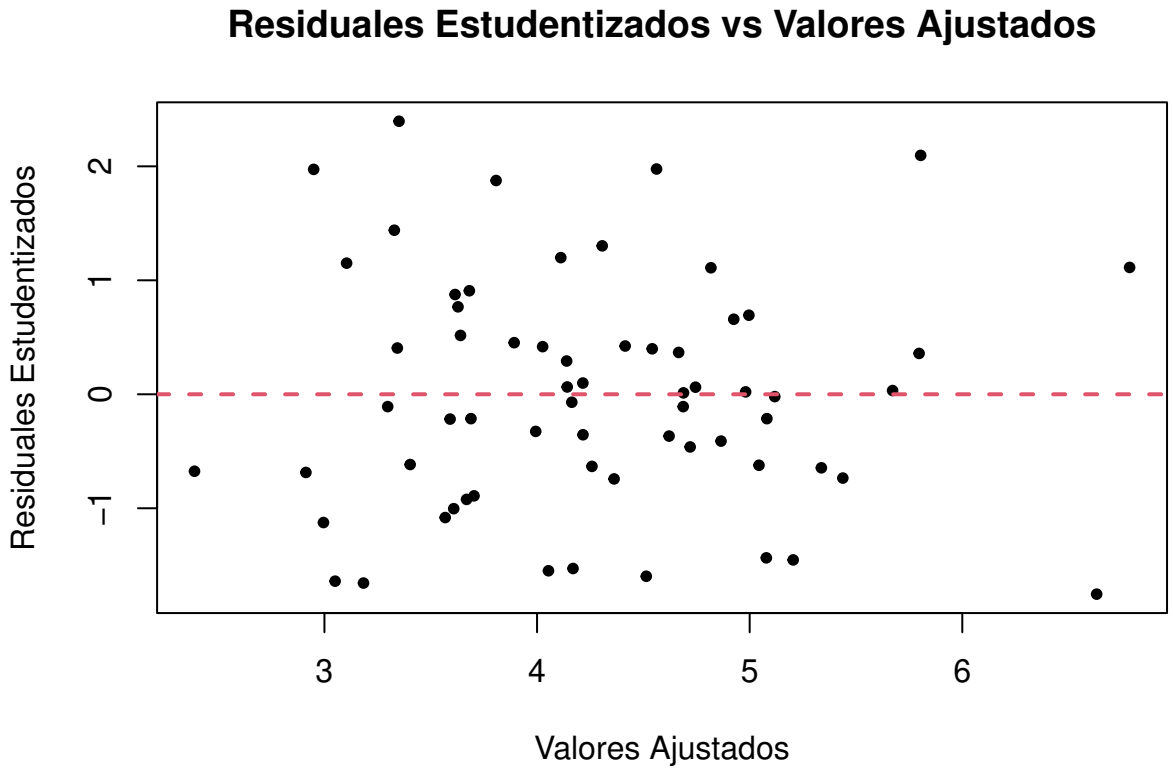

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados no se puede observar ningun patron que indique que la varianza aumente o disminuzca ni tampoco un comportamiento o dependencia que permita descartar el supuesto de varianza constante, ya que la evidencia para rechazar este supuesto es insuficiente se acepta el supuesto de varianza constante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

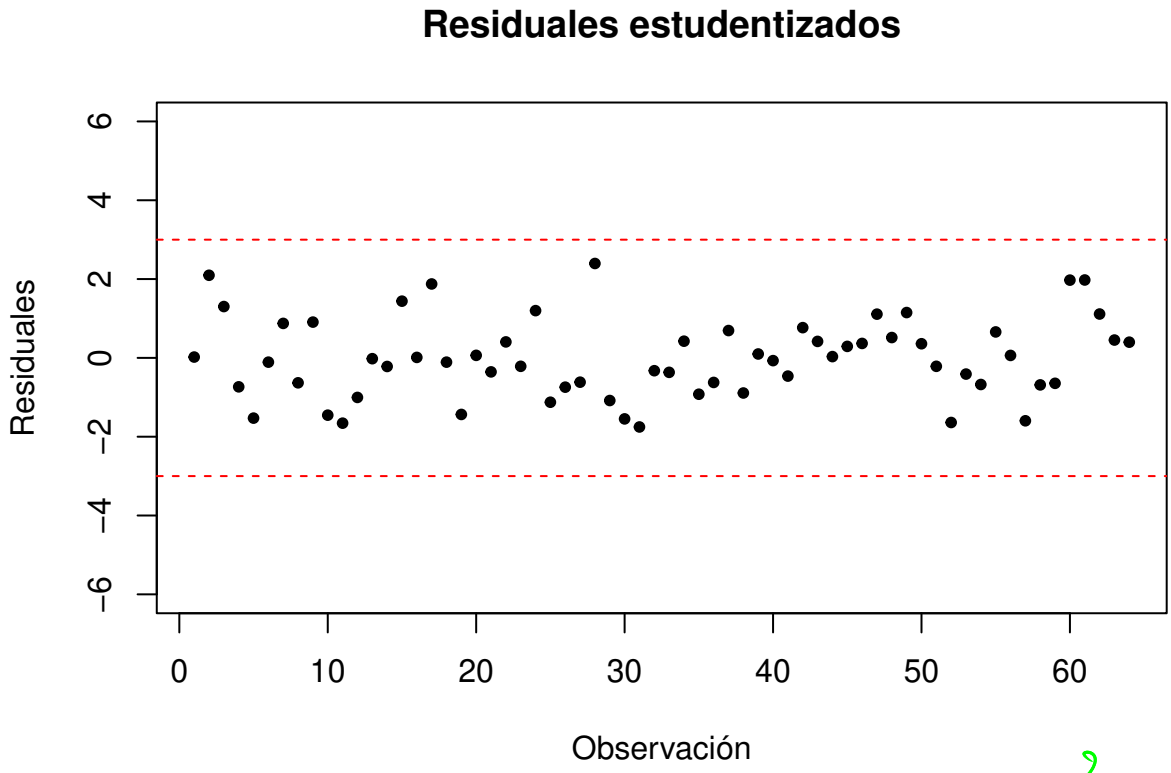

Figura 3: Identificación de datos atípicos

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

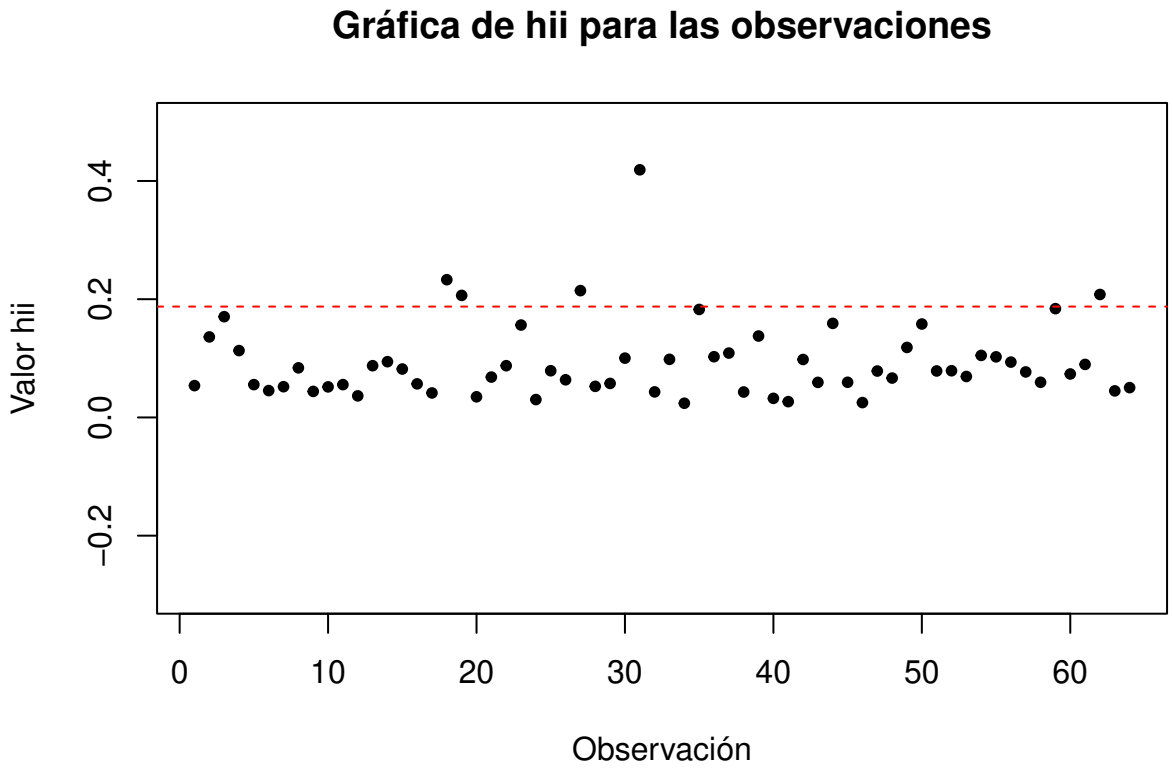

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 18  -0.1083  0.0006    0.2330 -0.0592
## 19  -1.4352  0.0893    0.2064 -0.7388
## 27  -0.6162  0.0173    0.2145 -0.3203
## 31  -1.7536  0.3695    0.4189 -1.5168
## 62   1.1130  0.0542    0.2080  0.5717
```

En la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja es el valor $h_{ii} = 2\frac{p}{n}$, se pueden observar 5 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.

4.2.3. Puntos influenciales

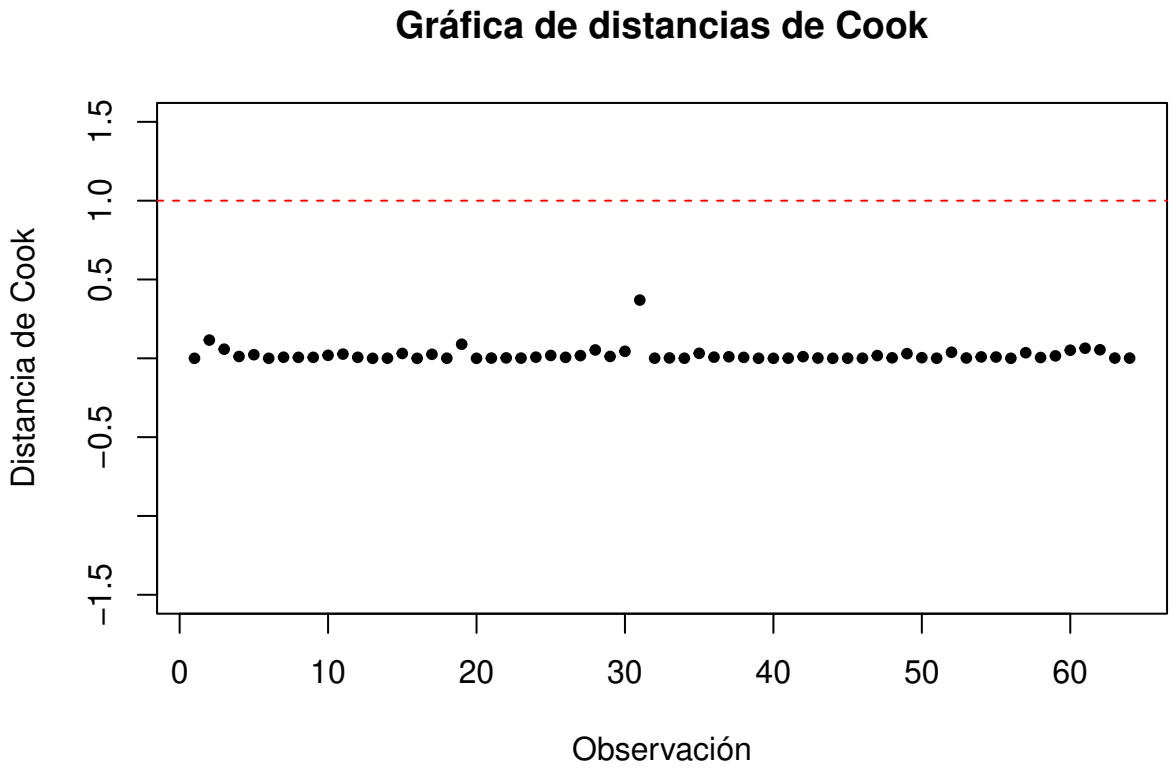

Figura 5: Criterio distancias de Cook para puntos influenciales

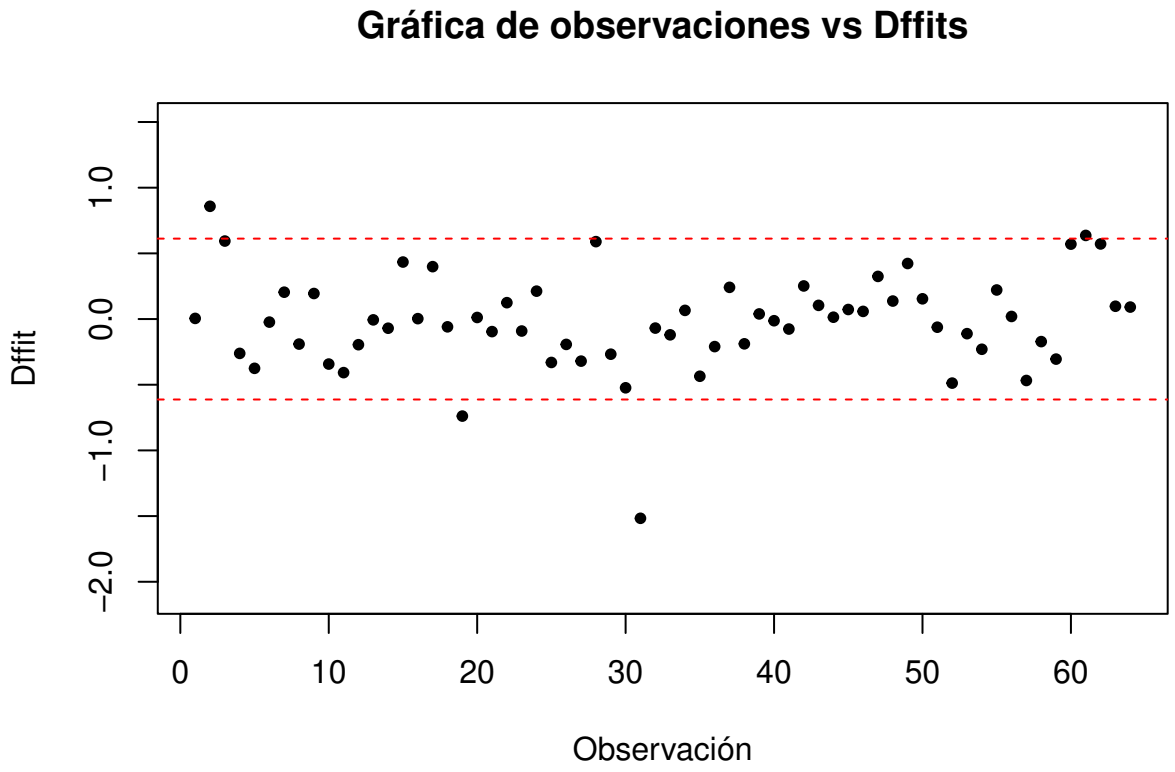

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 2    2.0962  0.1155    0.1362  0.8584
## 19  -1.4352  0.0893    0.2064 -0.7388
## 31  -1.7536  0.3695    0.4189 -1.5168
## 61   1.9763  0.0641    0.0897  0.6367
```

3 pt

¿ Qué causan ?

Como se puede ver, las observaciones 2, 19, 31 y 61 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial y también se ve 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 este.

4.3. Conclusión

3 pt

Todos los supuestos a excepción del supuesto de normalidad se cumplen y ya que los puntos influenciales por si solos son incapaces de explicar estas irregularidades, junto a la distorsion de los parametros debido a los puntos influenciales se concluye que el modelo no es valido ya que puede haber sesgo en los coeficientes y sus predicciones, intervalos, estimaciones e hipotesis pueden ser inexactas.