3,5

Trabajo 1

Estudiantes

Castro Parra Alejandro
Diaz Jaramillo Julian David
Mejia Arango Maria Camila
Restrepo Duran Mariana

Equipo 34

Docente

Mateo Ochoa Medina

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

16,5 pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

Donde:

- Y: Riesgo de infección
- X_1: Duración de la estadia (en días)
- X_2: Rutina de cultivos (por cada 100)
- X_3: Número de camas (durante el periodo de estudio)
- X_4: Censo promedio diario (durante el periodo de estudio)
- X_5: Número de enfermeras (durante el periodo de estudio)

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	-1.5846
β_1	0.1032
β_2	0.0495
β_3	0.0772
β_4	0.0074
β_5	0.0024

3,5 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.5846 + 0.1032 X_{1i} + 0.0495 X_{2i} + 0.0772 X_{3i} + 0.0074 X_{4i} + 0.0024 X_{5i}$$

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{12.460902}{0.751937} \overset{H_0}{\sim} f_{5,63} \qquad (1)$$

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	62.3045	5	12.460902	16.5717	2.04767e-10
Error	47.3720	63	0.751937		

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 $0 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto la regresión es significativa.

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	-1.5846	1.4558	-1.0885	0.2805
β_1	0.1032	0.1028	1.0042	0.3191
β_2	0.0495	0.0271	1.8290	0.0721
β_3	0.0772	0.0142	5.4328	0.0000
β_4	0.0074	0.0066	1.1285	0.2634
β_5	0.0024	0.0007	3.6134	0.0006

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_3 y β_5 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

Interpreten sólo los parámetros significativos, respecto a β_0 ya saben que se debe cumplir que el 0 esté en el intervalo

$\hat{\beta}_3 = 0.0772$**: indica que por cada unidad que aumenten las camas en el hospital el promedio de riesgo de infección aumenta en 0.0772 unidades porcentuales, mientras los demás factores se mantienen constantes**

$\hat{\beta}_5 = 0.0024$**: indica que por cada enfermera equivalente a tiempo completo en el hospital aumenta el riesgo de infección en 0.0024 unidades porcentuales, mientras los demás factores se mantienen constantes**

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6068$, lo que significa que aproximadamente el 60.68 % 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 covariable con el P-valor más bajo en el modelo fueron X_2, 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:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	47.372	X1 X2 X3 X4 X5
Modelo reducido	75.582	X1 X4

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSE(\beta_0, \beta_1, \beta_4) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,63}$$
$$= \frac{75.582}{0.751937}$$
$$= 100.516 \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 > f_{0.95,1,45}$ y por tanto se rechaza la hipotesis nula, indicando que algunos de los

β_j

es diferente de 0

¿Es posible o no descartar las variables del subconjunto? No, ya que al menos una de las variables de la regresión es significativa

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hacen las preguntas: ¿El número promedio de enfermeras es igual al doble del número de promedio de pacientes? y, ¿La duración promedio de la estadia de los pacientes es igual a tres veces el número promedio de camas en el hospital?. Consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_5 = 2\beta_4;\ \beta_1 = 3\beta_3 \\ \mathrm{H}_1 : \beta_5 \neq 2\beta_4, \beta_1 \neq 3\beta_3 \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X_{1i}^* + \beta_2 X_{2i} + \beta_5 X_{5i}^* + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

Donde $X_{1i}^* = X_{1i} + 3X_{3i}$ $X_{5i}^* = X_{5i} + 2X_{4i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por: 2 pt

$$F_0 = \frac{(SSE(MR) - 47.3720)/2}{0.751937} \overset{H_0}{\sim} f_{2,63} \tag{3}$$

4. Pregunta 4 13 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:

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

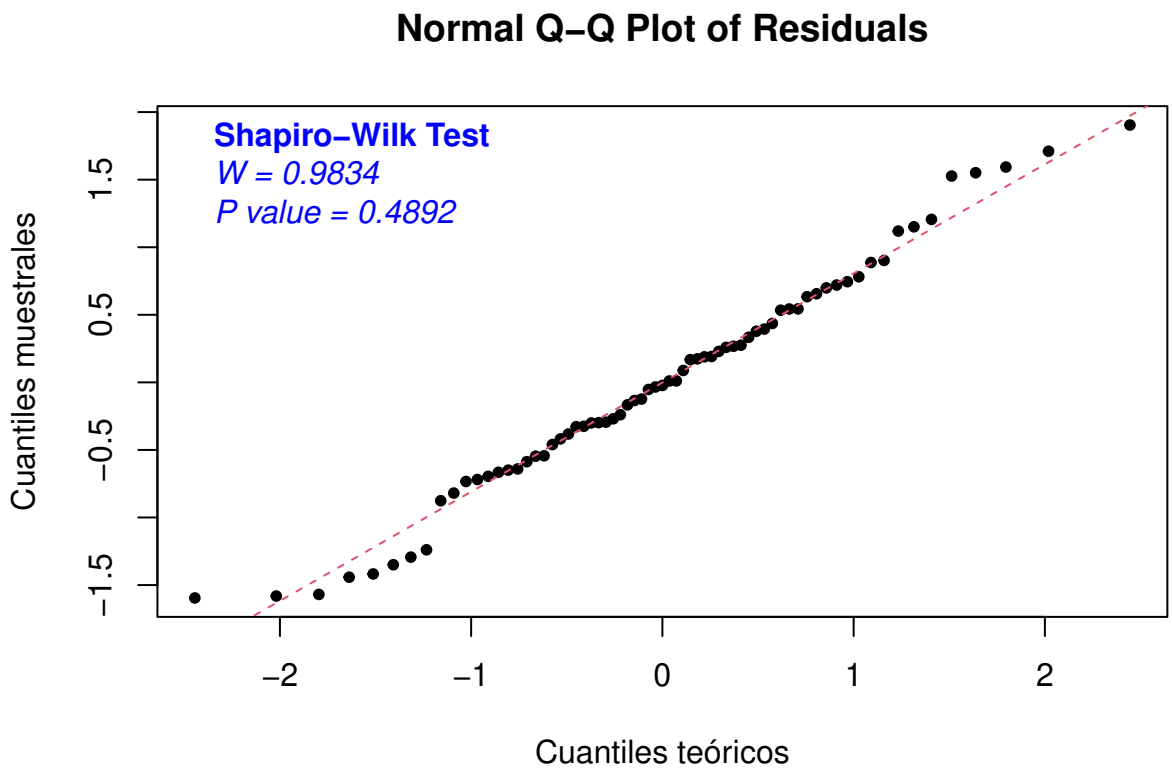

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.4892 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. Este resultado se ve respaldado por la gráfica de comparación de cuantiles, donde la mayoría de los puntos se encuentran cercanos a la línea del Q-Q Plot, mostrando colas estables y ausencia de patrones irregulares.

4.1.2. Varianza constante

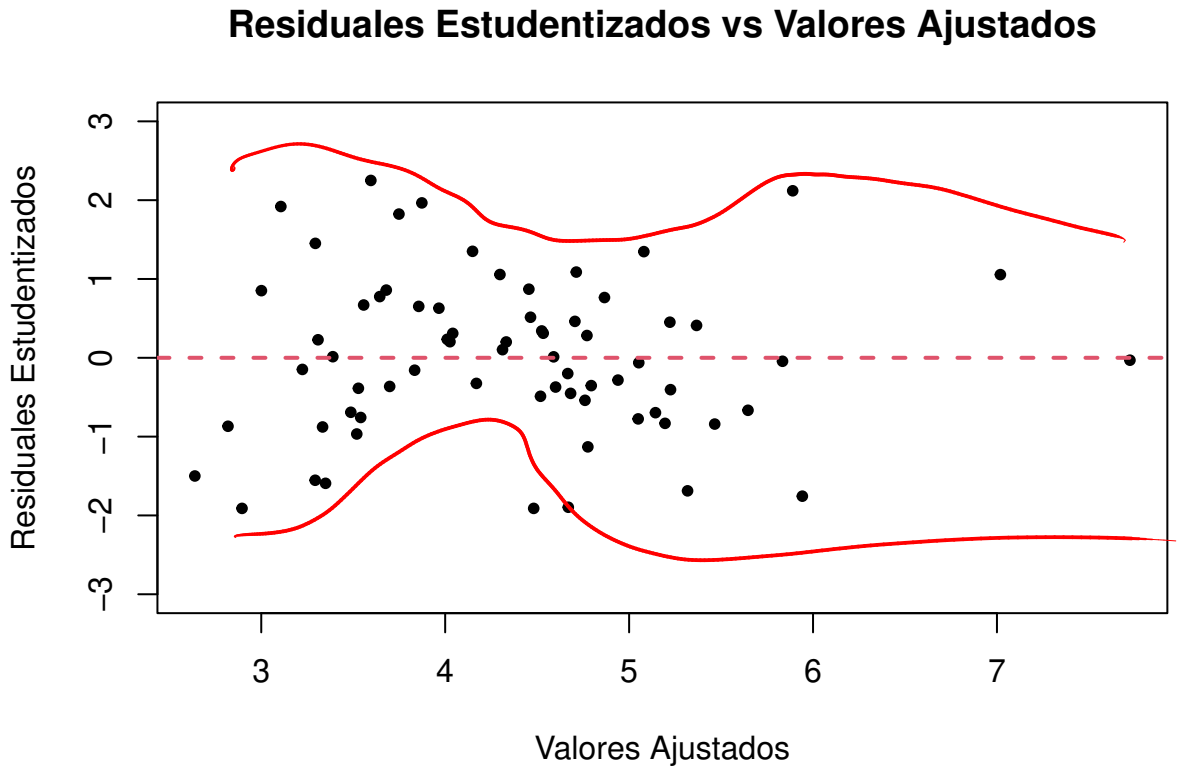

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuos estandarizados vs valores ajustados, se puede observar una curvatura en la nube de puntos, en la que la dispersión de los residuos primero aumenta, luego disminuye y vuelve a aumentar. Esto sugiere que del modelo se puede diferir dos condiciones: varianza constante y falta de ajuste. Aunque podamos aceptar que el supuesto de varianza constante es válido, debemos tener en cuenta que el modelo no se ajusta adecuadamente a los datos y, por lo tanto, podría no ser un modelo válido.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

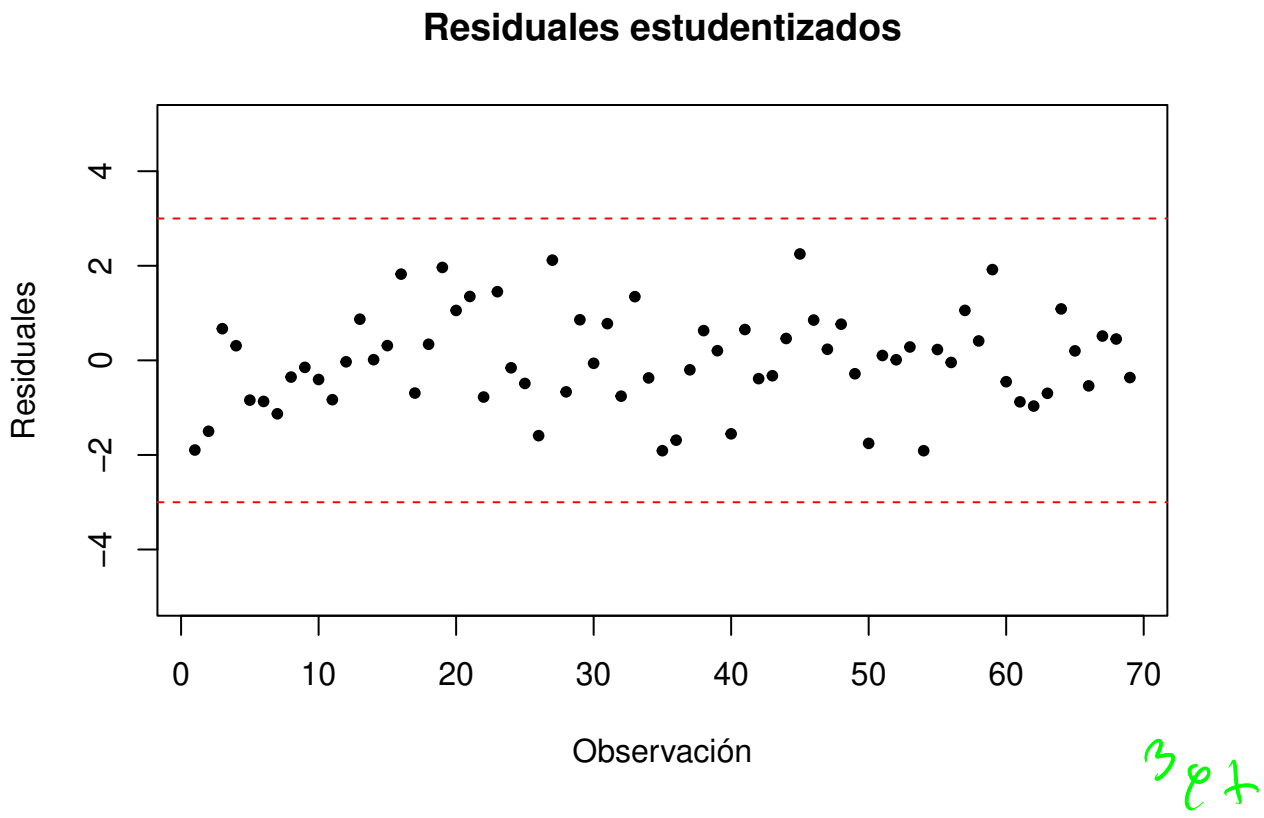

Figura 3: Identificación de datos atípicos

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

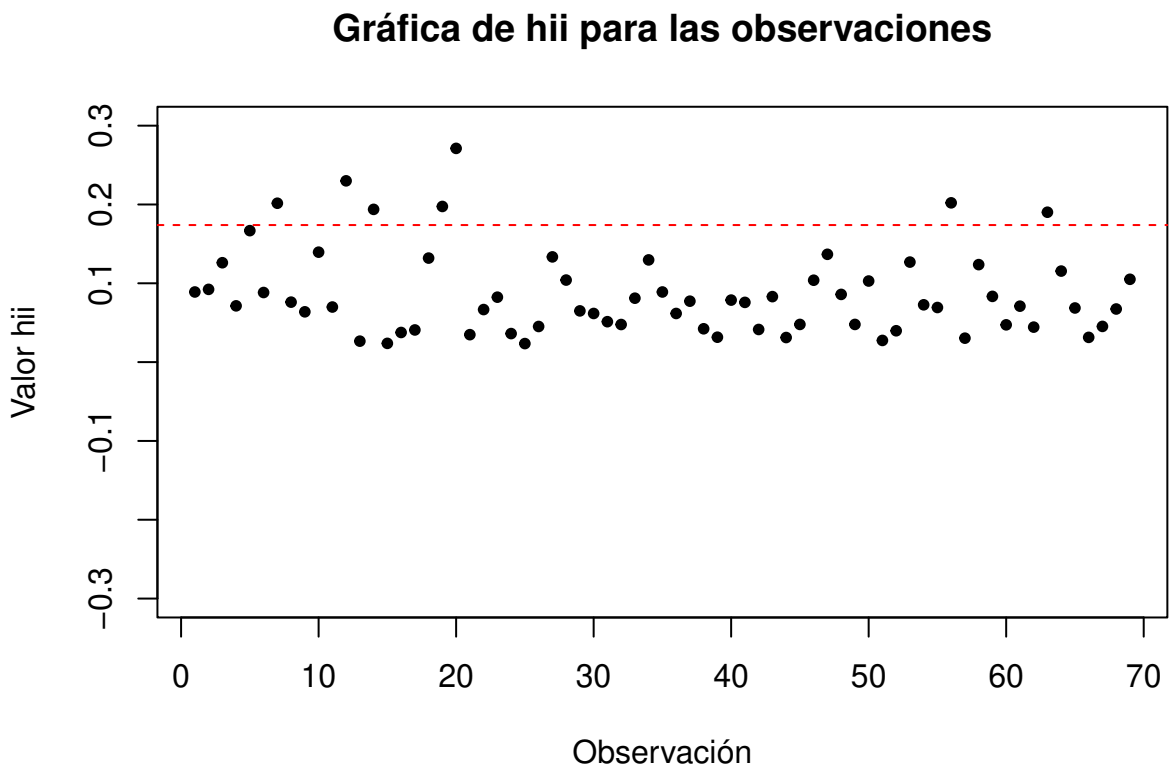

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 7   -1.1302  0.0538    0.2016 -0.5692
## 12  -0.0303  0.0000    0.2300 -0.0164
## 14   0.0140  0.0000    0.1938  0.0068
## 19   1.9654  0.1584    0.1975  0.9982
## 20   1.0554  0.0691    0.2712  0.6444
## 56  -0.0440  0.0001    0.2021 -0.0220
## 63  -0.6963  0.0190    0.1902 -0.3361
```

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{6}{69}$, se puede apreciar que existen 7 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{6}{69}$, los cuales son los presentados en la tabla.

4.2.3. Puntos influenciales

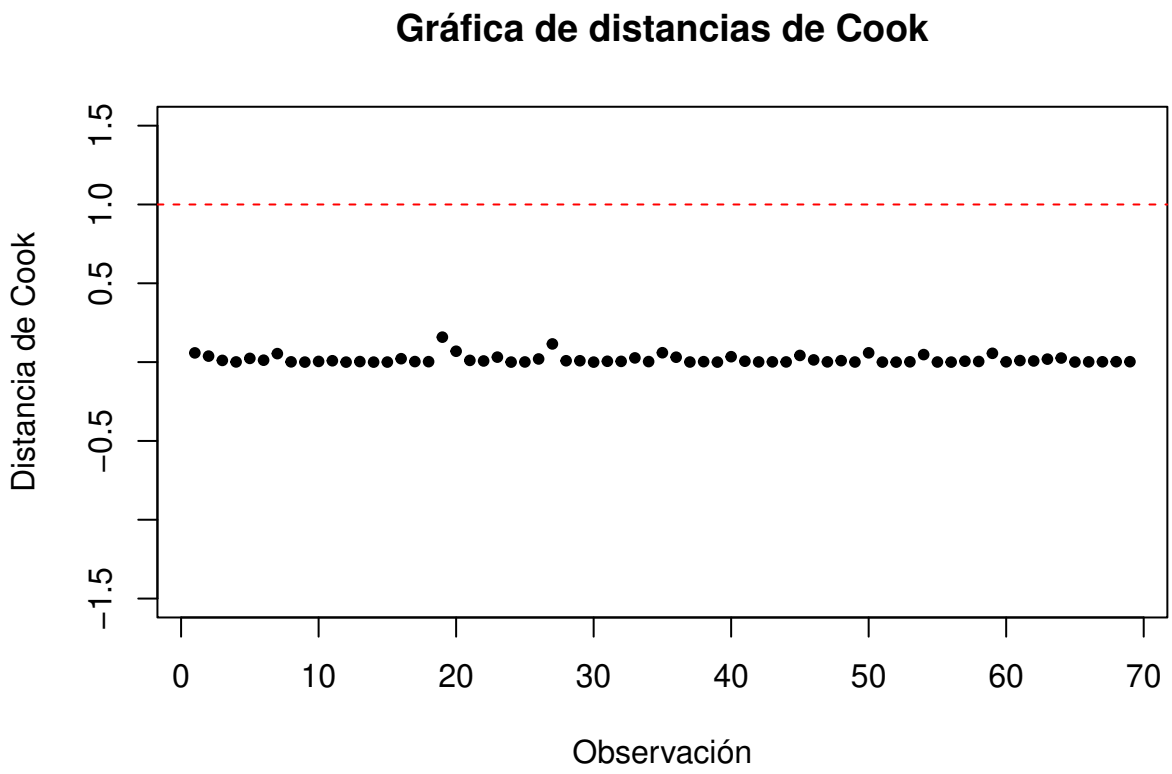

Figura 5: Criterio distancias de Cook para puntos influenciales

Según 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.

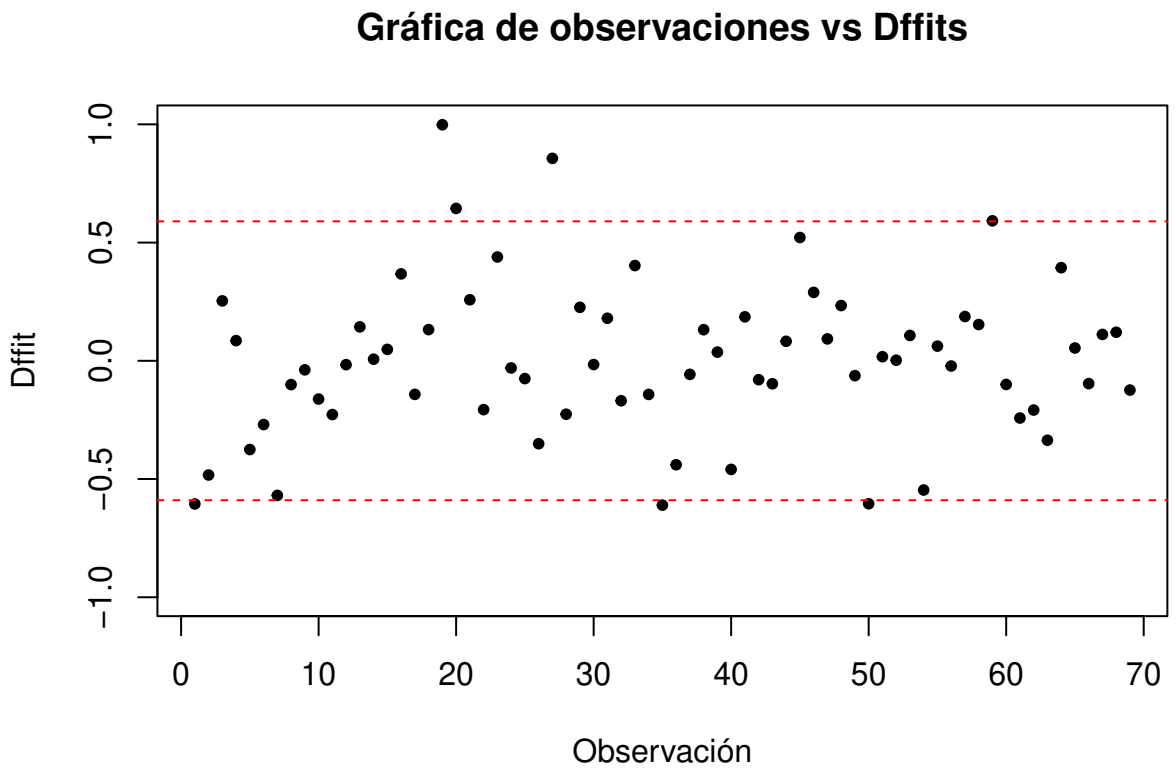

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 1   -1.8961  0.0586    0.0890 -0.6055
## 19   1.9654  0.1584    0.1975  0.9982
## 20   1.0554  0.0691    0.2712  0.6444
## 27   2.1189  0.1153    0.1335  0.8561
## 35  -1.9107  0.0595    0.0890 -0.6105
## 50  -1.7556  0.0589    0.1029 -0.6047
## 59   1.9195  0.0559    0.0834  0.5920
```

Como se puede ver, las observaciones 1, 19, 20, 27, 35, 50 y 59 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{6}{69}}$, es un punto influencial.

4.3. Conclusión

En conclusión, nuestros análisis han revelado que, en general, el modelo cumple con los supuestos de normalidad y varianza constante. Según las pruebas de hipótesis realizadas existen parámetros que ejercen un impacto significativo en la respuesta del modelo. Sin

Validez es dada sólo por supuestos.

embargo, el valor del R^2 queda muy abierto a la interpretabilidad sobre si es lo suficientemente alto para que el modelo sea válido. Además, hemos observado un patrón en la gráfica de residuales estudentizados vs valores ajustados que sugiere una falta de ajuste en el modelo. Se ha planteado la posibilidad de mejorar esta situación mediante una transformación, aunque esto podría afectar la interpretabilidad de los resultados.

Dada esta información, es crucial subrayar que la interpretación final de la aplicabilidad del modelo debe ser realizada por un experto en el campo, teniendo en cuenta la naturaleza específica del problema y la relevancia de los resultados en el contexto de la investigación o aplicación práctica.