3,25

Trabajo 1

Estudiantes

Emanuel Cardona López
Ivanna Lucía Montes Otero
Juan Manuel Ortiz Echeverri
Sebastián Rengifo Jaramillo
Equipo # 04

Docente

Julieth Verónica Guarín Escudero

Asignatura

Estadística II

Sede Medellín
30 de Marzo de 2023

Índice

Índice de figuras

Índice de tablas

1. Pregunta 1

Teniendo en cuenta la base de datos asignada, la cual es la **Equipo04.txt**, las covariables son: Duración de la estadía (X1), Rutina de cultivos (X2), Número de camas (X3), Censo promedio diario (X4) y Número de enfermeras(X5).

El modelo que se propone es:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

1.1. Modelo de regresión

Al ajustar el modelo anterior se obtienen los siguientes coeficientes:

Tabla 1: Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	2.3243
$\hat{\beta}_1$	0.3016
$\hat{\beta}_2$	-0.0339
$\hat{\beta}_3$	0.0506
$\hat{\beta}_4$	-0.0043
$\hat{\beta}_5$	0.0021

Por lo tanto, el modelo de regresión ajustado es:

$$\widehat{Y}_i = 2.3243 + 0.3016 X_{1i} - 0.0339 X_{2i} - 0.0506 X_{3i} - 0.0043 X_{4i} + 0.0021 X_{5i}$$

donde $1 \leq i \leq 45$

1.2. Significancia de la regresión

Para la significancia de la regresión se hará uso de la siguiente tabla anova, usando un estadístico de prueba **F**: Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,39} \tag{1}$$

Tabla 2: Tabla anova significancia de la regresión

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	Valor-P
Modelo de regresión	23.8469	5	4.769381	5.46737	0.000661129
Error	34.0211	39	0.872336		

$$\begin{cases} \text{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ \text{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 3, 4, 5 \end{cases}$$

De la tabla anova, se concluye que se rechaza la hipótesis nula para la no significancia de los parámetros, por lo tanto, la regresión es significativa y algún parámetro por consiguiente es significativo.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar cuál de estos es significativo para el modelo de regresión:

Tabla 3: Resumen de los coeficientes

	Estimación β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	2.3243	1.7631	1.3183	0.1951
β_1	0.3016	0.1233	2.4454	0.0191
β_2	-0.0339	0.0315	-1.0778	0.2877
β_3	0.0506	0.0247	2.0506	0.0471
β_4	-0.0043	0.0087	-0.4885	0.6279
β_5	0.0021	0.0009	2.2430	0.0306

1.4. Significancia de los parámetros

- $\hat{\beta}_1$: Su valor-P es menor que un $\alpha = 0.05$ entonces rechazamos la hipótesis nula, luego, la variable predictora Duración de la estadía es significativa para el modelo.
- $\hat{\beta}_3$: Su valor-P es menor que un $\alpha = 0.05$ entonces rechazamos la hipótesis nula, luego, la variable predictora Número de camas es significativa para el modelo.
- $\hat{\beta}_5$: Su valor-P es menor que un $\alpha = 0.05$ entonces rechazamos la hipótesis nula, luego, la variable predictora Número de enfermeras es significativa para el modelo.
- $\hat{\beta}_0, \hat{\beta}_2, \hat{\beta}_4$: Para estos parámetros, se acepta la hipótesis nula y determinamos que las variables predictoras asociadas a estos parámetros no son significativas para el modelo de regresión.

1.5. Interpretación de los parámetros estimados

- $\hat{\beta}_1$: La probabilidad promedio de riesgo de infección aumenta 0.3016 cuando el paciente permanece más días en el hospital. Esto se concluye cuando los demás parámetros ($\hat{\beta}_3, \hat{\beta}_5$) permanecen constantes.
- $\hat{\beta}_3$: La probabilidad promedio de riesgo de infección aumenta 0.0506 cuando hay un mayor número de pacientes en el hospital.Esto se concluye cuando los demás parámetros ($\hat{\beta}_1, \hat{\beta}_5$) permanecen constantes.
- $\hat{\beta}_5$: La probabilidad promedio de riesgo de infección aumenta 0.0021 si hay un mayor número de enfermeras a tiempo completo en el hospital.Esto se concluye cuando los demás parámetros ($\hat{\beta}_1, \hat{\beta}_3$) permanecen constantes.

1.6. Coeficiente de determinación múltiple R^2

Para el cálculo del $R^2 = \dfrac{SSR}{SST}$, que se puede calcular de la tabla anova, el modelo tiene un $R^2 = 0.4121$, es decir, el modelo de regresión múltiple lineal explica el 41.1 % de la variabilidad total del porcentaje de infección.

2. Pregunta 2

3,5 pt

2.1. Planteamiento prueba de hipótesis y modelo reducido

Los parámetros cuyos valores-P fueron los más altos corresponden a $\beta_2, \beta_3, \beta_4$, por lo tanto, se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0, para } j = 2, 4 \end{cases}$$

El modelo completo es el definido en la sección 1.1, y el modelo reducido es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

Se presenta la siguiente tabla con el resumen de todas las regresiones para plantear el estadístico de prueba:

Tabla 4: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo Completo	34.021	X1 X2 X3 X4 X5
Modelo Reducido	39.730	X1 X5

3 pt

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_5) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \overset{H_0}{\sim} f_{3,39} \\ &= \frac{(39.730 - 34.021)/3}{34.021/39} = 2.1815055 \end{aligned}$$

es un cuantil, no un estadístico

Ahora, comparando a un nivel de significancia de $\alpha = 0.05$, F_0 con $f_{3,39} = 2.8450678$

Entonces se concluye que las variables del subconjunto se pueden descartar del modelo, es decir, aceptamos la hipótesis nula. Pues su estadístico de prueba es menor que el estadístico $f_{(0.05,3,39)}$.

se concluye primero que el subconjunto es signific. y ahí sí que se pueden descartar

3. Pregunta 3

2 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2, \beta_3 = \beta_5 \\ \mathrm{H}_1 : \beta_1 \neq \beta_2 \vee \beta_3 \neq \beta_5 \end{cases}$$

Reescribimos matricialmente:

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

Con una matriz $\mathbf{L}$:

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

Luego, la matriz $\mathbf{L}$ tiene un rango de 2, con un modelo reducido dado por:

$$Y_i = \beta_o + \beta_1(X_{1i} + X_{2i}) + \beta_3(X_{3i} + X_{5i}) + \beta_4 X_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

Reescrito como:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_3 X^*_{3i} + \beta_4 X_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

Donde $X^*_{1i} = X_{1i} + X_{2i}$ y $X^*_{3i} = 3X_{3i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,39} \tag{2}$$

4. Pregunta 4

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 ~~shapiro-wilk~~, acompañada de un gráfico cuantil-cuantil:

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

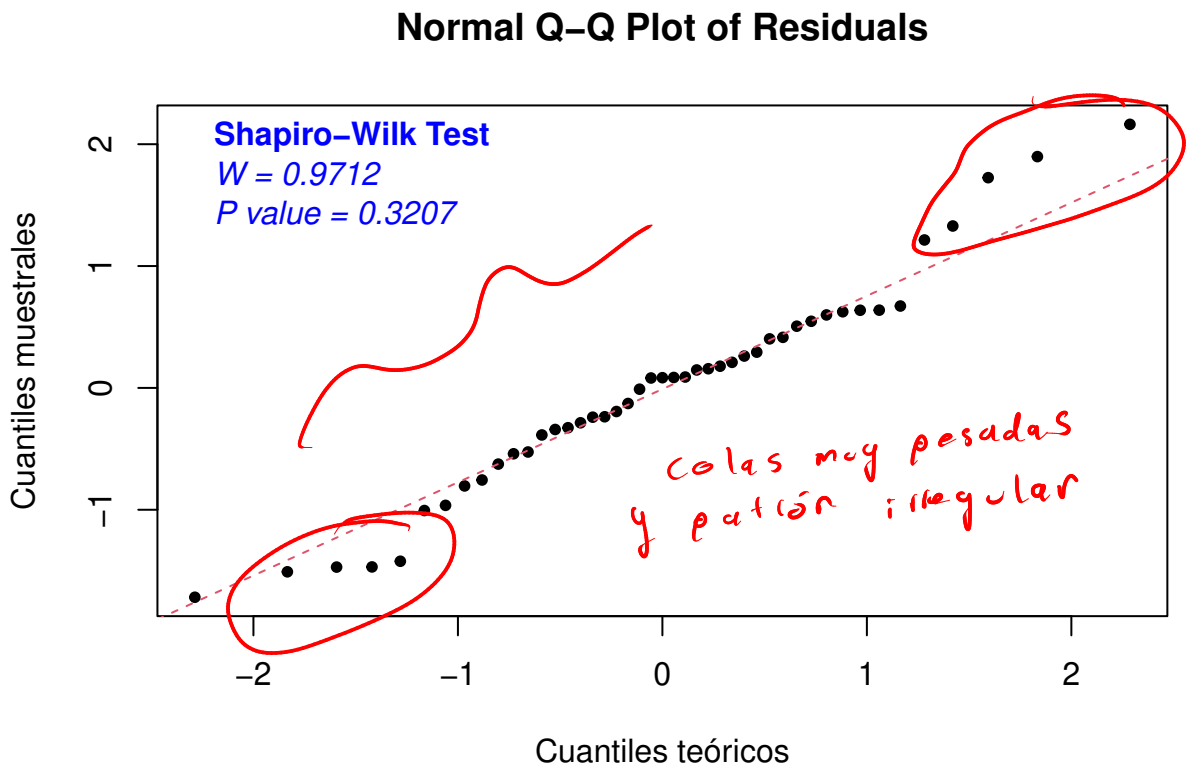

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Como el valor P es mayor que $\alpha = 0.05$, no rechazaríamos la hipótesis nula, y se podría concluir que los errores tienen una distribución normal con media μ y varianza σ^2, pero si analizamos el gráfico no se evidencia un comportamiento lineal en las colas y el patrón de los errores es asimétrico. Por lo tanto rechazamos la hipótesis nula y por ende los errores no se distribuyen normal.

El análisis del gráfico pudo ser más exhaustivo y claro, sin embargo está muy bien

4.1.2. Varianza constante

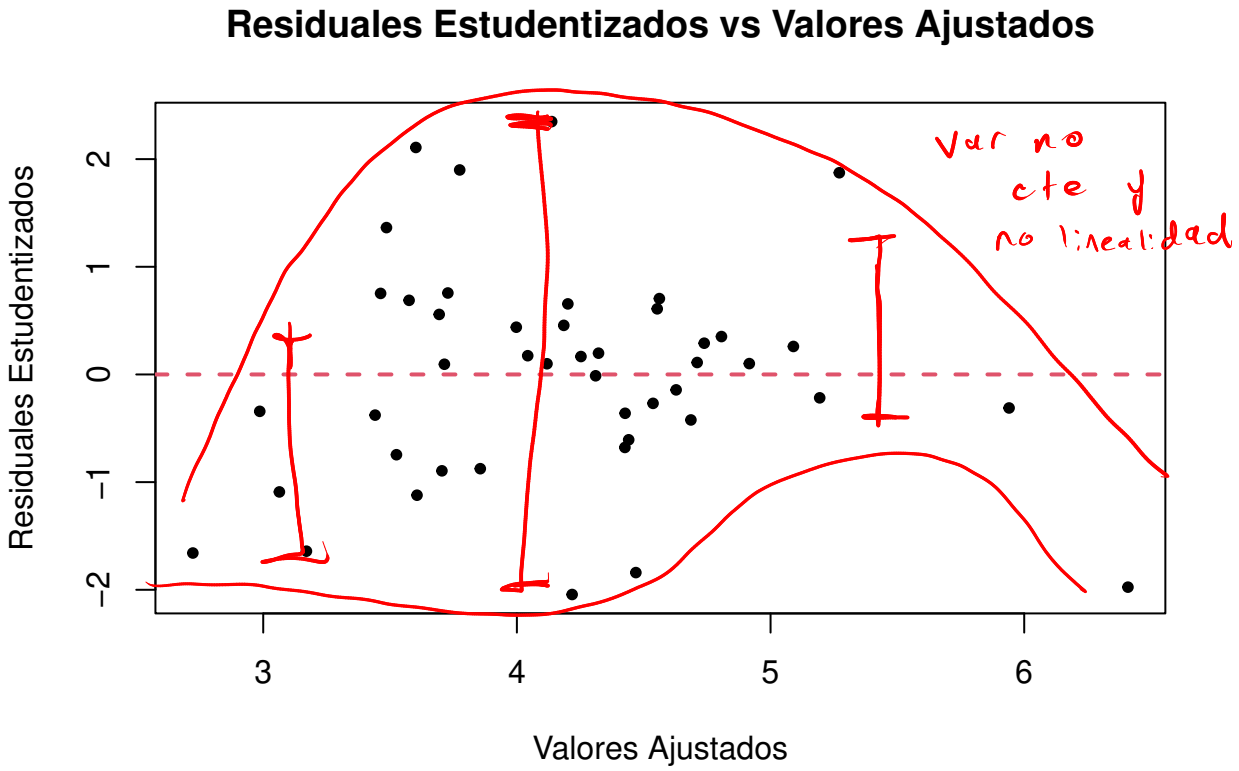

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Del grafico de Valores Ajustados vs Residuales Estudentizados no se detectan comportamientos extremos que puedan demostrar que la varianza no es constante, por lo tanto se asume que la varianza es constante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

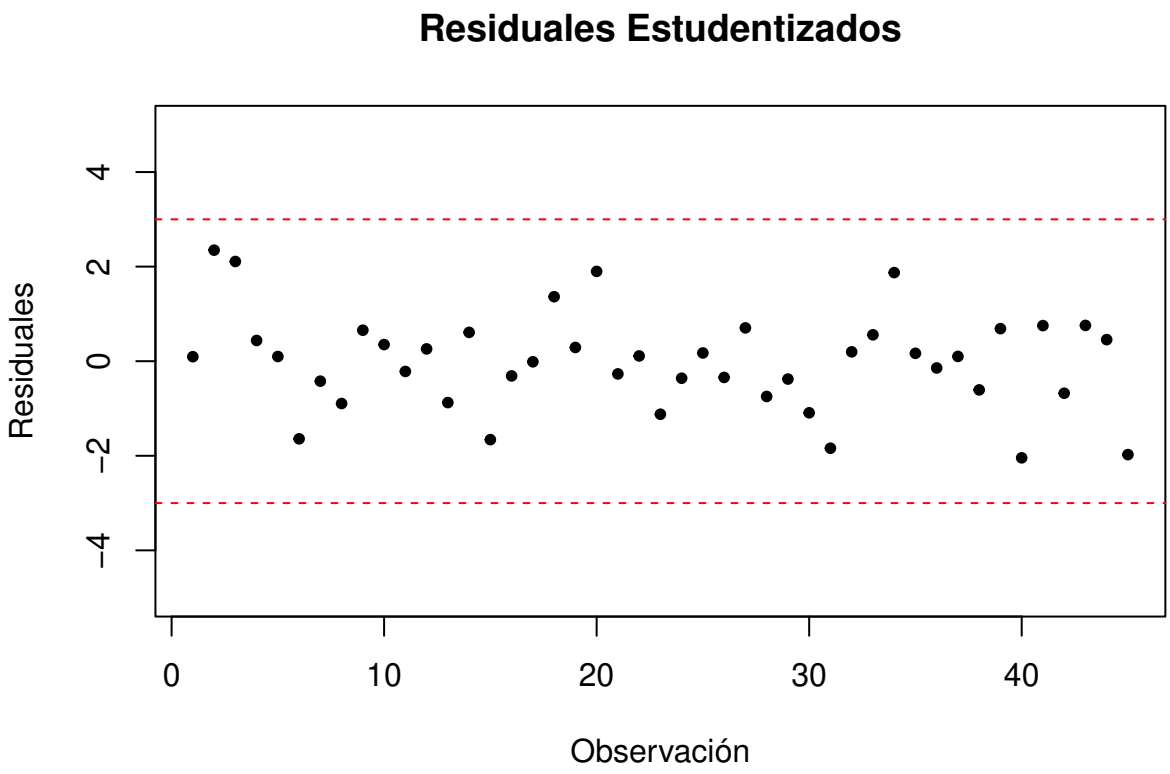

Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica anterior, los datos no sobrepasan los valores de 3 y -3, se puede concluir que no hay valores atípicos bajo el criterio $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

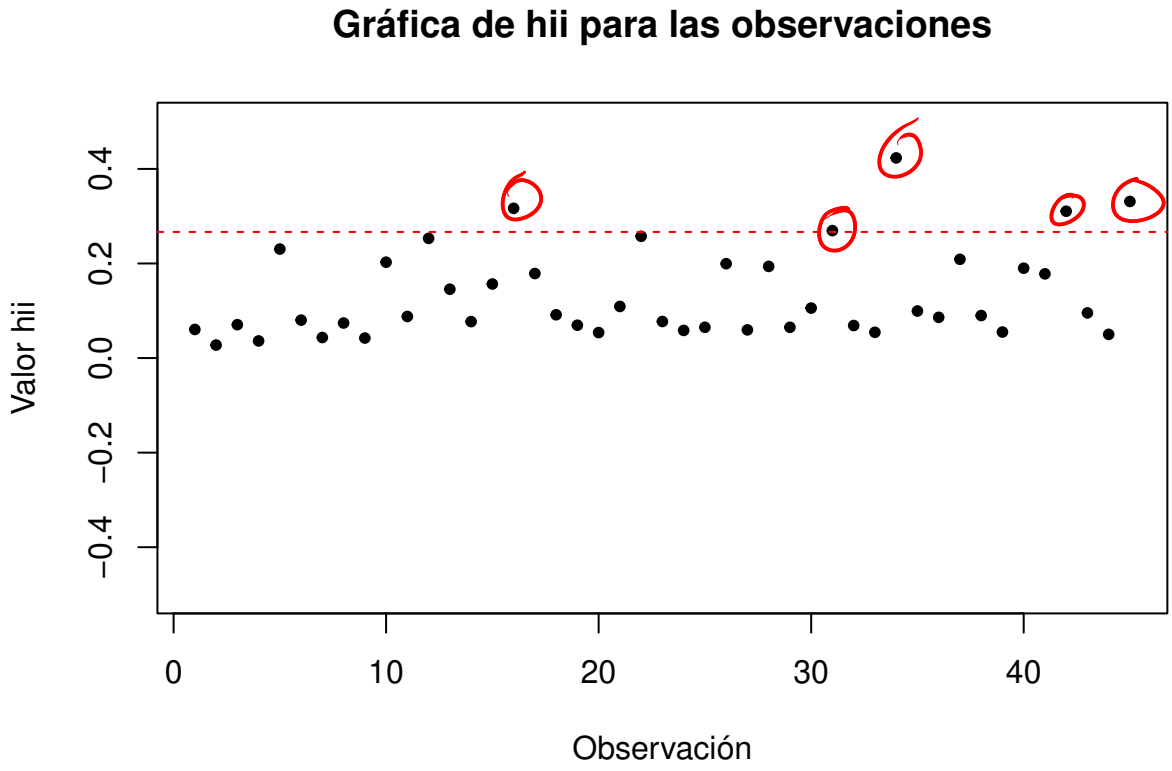

Figura 4: Identificación de puntos de balanceo

Al analizar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor$h_{ii} > 0.2666667$, podemos darnos cuenta que existen 5 datos que exceden dicho valor, y por ende se pueden catalogar como puntos de balanceo. Estos datos son los presentados en cuales son los presentados en la tabla.

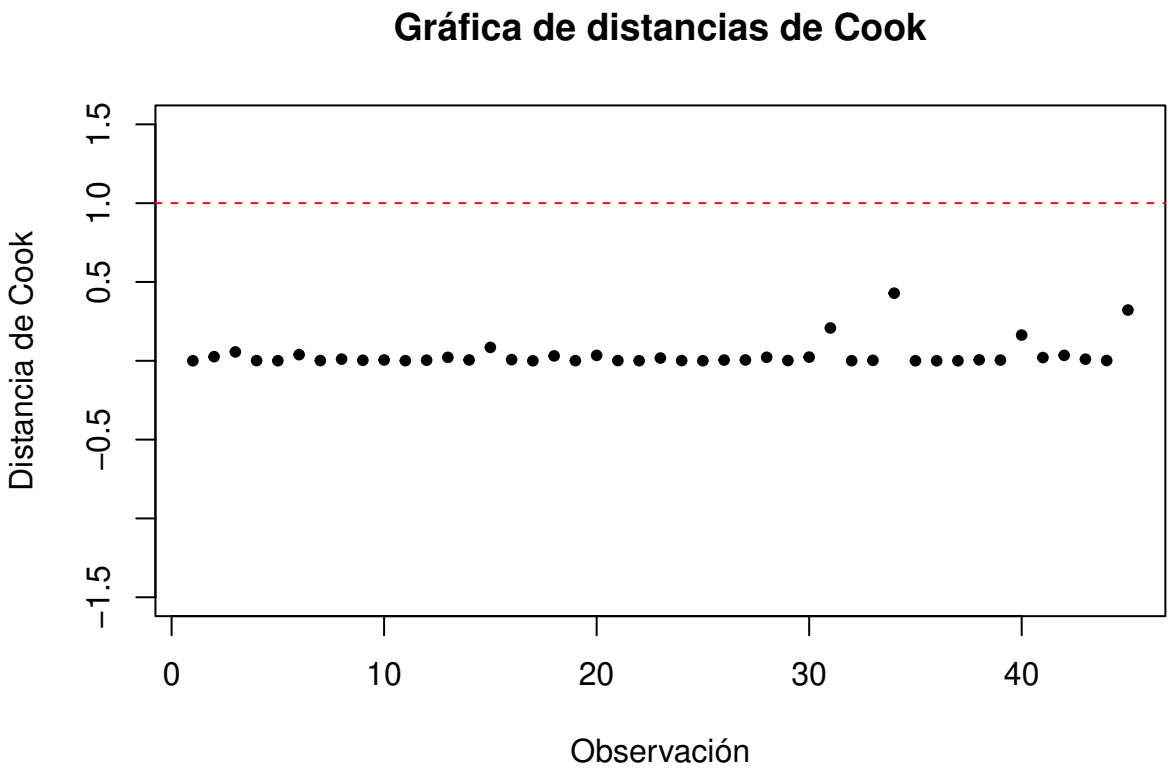

Figura 5: Criterio distancias de Cook para puntos influenciales

Al observar la gráfica podemos darnos cuenta que ninguno de los datos supera el valor de 1, por lo tanto no se puede asegurar que alguno de los valores sea de carácter influencial. según este criterio

Tabla 5: Observaciones con valores Hii, Distancia de Cooks y Dffits

	res.stud	Cooks.D	hii.value	Dffits
15	-1.6585	0.0851	0.1566	-0.7318
31	-1.8402	0.2082	0.2695	-1.1546
34	1.8733	0.4291	0.4232	1.6604
40	-2.0437	0.1633	0.1900	-1.0341
45	-1.9755	0.3220	0.3311	-1.4462

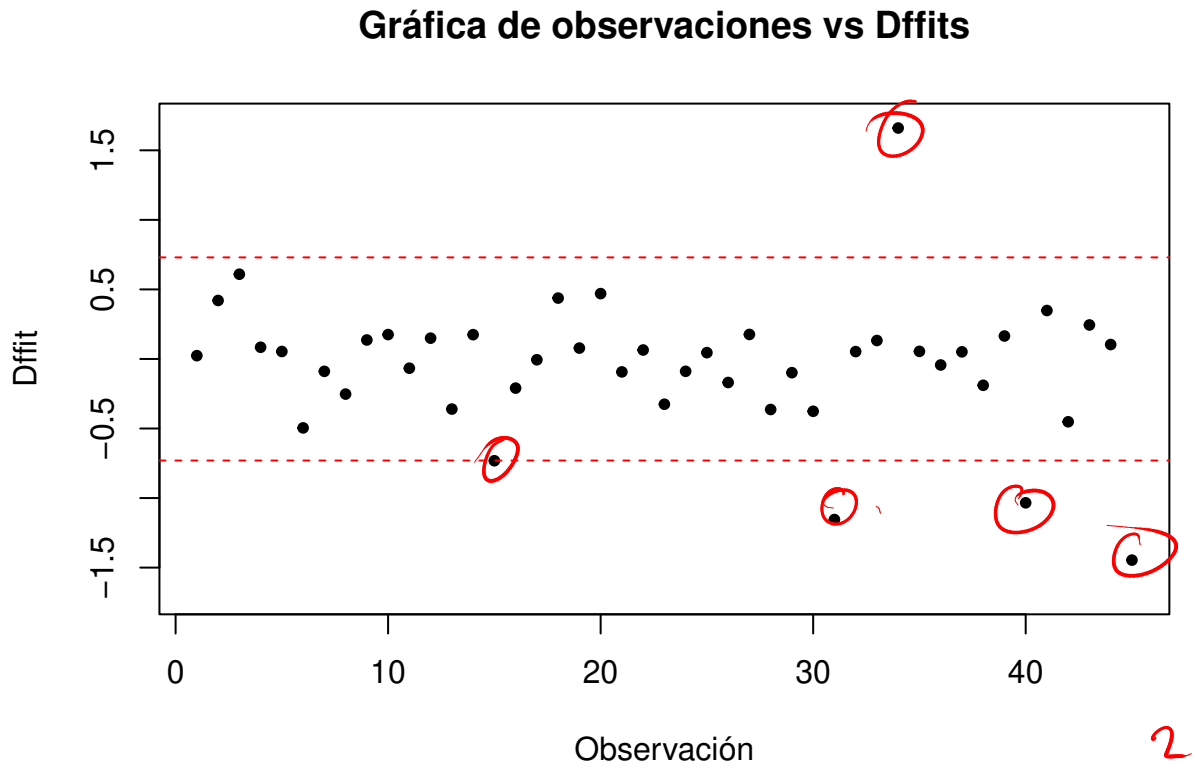

Figura 6: Criterio Dffits para puntos influenciales

Según el criterio de Dffits el cual dice $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$ es punto influencial. Si analizamos la gráfica podemos darnos cuenta que 5 valores cumplen dicho criterio y por ende pueden catalogarse como influenciales

4.3. Conclusión

- 1: El modelo no es válido, ya que no cumple con los supuestos de normalidad de los errores.
- 2: No se encuentran puntos atípicos, pero sí puntos de balanceo, los cuales son: 15, 31, 34, 40, 45.
- 3: por el criterio de distancias de Cook no se puede asegurar que los puntos atípicos o de balanceo sean influenciales
- 4: por el criterio de Dffits se puede asegurar que hay 5 datos influenciales los cuales son: 15, 31, 34, 40, 45.