Trabajo 1

4,3

Estudiantes

Astrid Isbed Chapid Lucero
Julian Sebastian Salazar Marin
Doris Alejandra Ramos Figueroa
Juan Gabriel Carvajal Negrete

Equipo 12

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

Donde:

- Y_i: Riesgo de infección
- X_1: Duración de la estadía
- X_2: Rutina de cultivos
- X_3: Número de camas
- X_4: Censo promedio diario
- X_5: Numero 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	-1.0485
β_1	-0.0169
β_2	0.0463
β_3	0.0515
β_4	0.0192
β_5	0.0036

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.0485 - 0.0169 X_{1i} + 0.0463 X_{2i} + 0.0515 X_{3i} + 0.0192 X_{4i} + 0.0036 X_{5i};\ 1 \leqslant i \leqslant 54$$

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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,48} \tag{1}$$

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	49.3214	5	9.86428	8.31072	1.02094e-05
Error	56.9729	48	1.18693		

De la tabla Anova, se observa un valor P aproximadamente igual a 1.02094e-05, 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	-1.0485	1.9479	-0.5383	0.5929
β_1	-0.0169	0.0927	-0.1827	0.8558
β_2	0.0463	0.0343	1.3511	0.1830
β_3	0.0515	0.0218	2.3613	0.0223
β_4	0.0192	0.0111	1.7289	0.0903
β_5	0.0036	0.0011	3.2147	0.0023

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 $\alpha = 0.05$.

1.4. Interpretación de los parámetros

- β_3 : Indica que por cada unidad de incremento en el numero de camas en promedio el riesgo de infección aumenta en 0.051514060 cuando las demas predictoras permanecen fijas.
- β_5 : Indica que por cada enfermera adicional en promedio el riesgo de infección aumenta en 0.003579507 cuando las demas predictoras permanecen fijas.

1.5. Coeficiente de determinación múltiple R^2

2 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.4640079$, lo que significa que aproximadamente el 46 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.Esta medida nos proporciona una idea de cuánta variabilidad en la variable de respuesta se captura y se explica adecuadamente mediante las variables independientes incluidas en el modelo. Es importante destacar que, aunque el modelo explica una parte significativa de la variabilidad, aún existe un porcentaje considerable de variabilidad que no se ha tenido en cuenta o no se puede explicar mediante las variables consideradas.

¿Cómo se calcula?

2. Pregunta 2

4 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más bajo en el modelo fueron X_3, X_4, 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	56.973	X1 X2 X3 X4 X5
Modelo reducido	87.645	X1 X2

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i; \quad 1 \leqslant i \leqslant 54$$

¿supuestos

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

2 pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,48} \\ &= \frac{10.224}{1.186937} \\ &= 8.613768 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,48} = 2.7981$, se puede ver que $F_0 > f_{0.95,3,48}$ y por tanto se rechaza la hipótesis nula y concluimos con un nivel de significancia de $\alpha = 0.05$ que hay al menos una de las variables tiene un impacto estadisticamente significativo en la variable respuesta.

Las variables en el subconjunto no pueden ser descartadas, ya que al menos una covariable muestra una relación significativa con la variable de respuesta.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea las siguientes preguntas: ¿Será que el efecto de la variable "Duración de la estadía" es igual al efecto de la variable "Rutina de cultivos" sobre la variable respuesta? Además, ¿El efecto de la variable "Número de camas" es tres veces el efecto de la variable "Número de enfermeras" sobre la variable respuesta?. Para responder a estas preguntas , se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2;\ \beta_3 = 3\beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{2i} + \beta_5 X^*_{3i} + \beta_4 X_{5i} + \varepsilon_i;\ 1 \leqslant i \leqslant 54$$

Donde:

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$\begin{aligned} F_0 &= \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,48} \\ &= \frac{(SSE(MR) - 56.9729)/2}{1.18693} \overset{H_0}{\sim} f_{2,48} \end{aligned} \tag{3}$$

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 que se evaluará por medio del test de shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

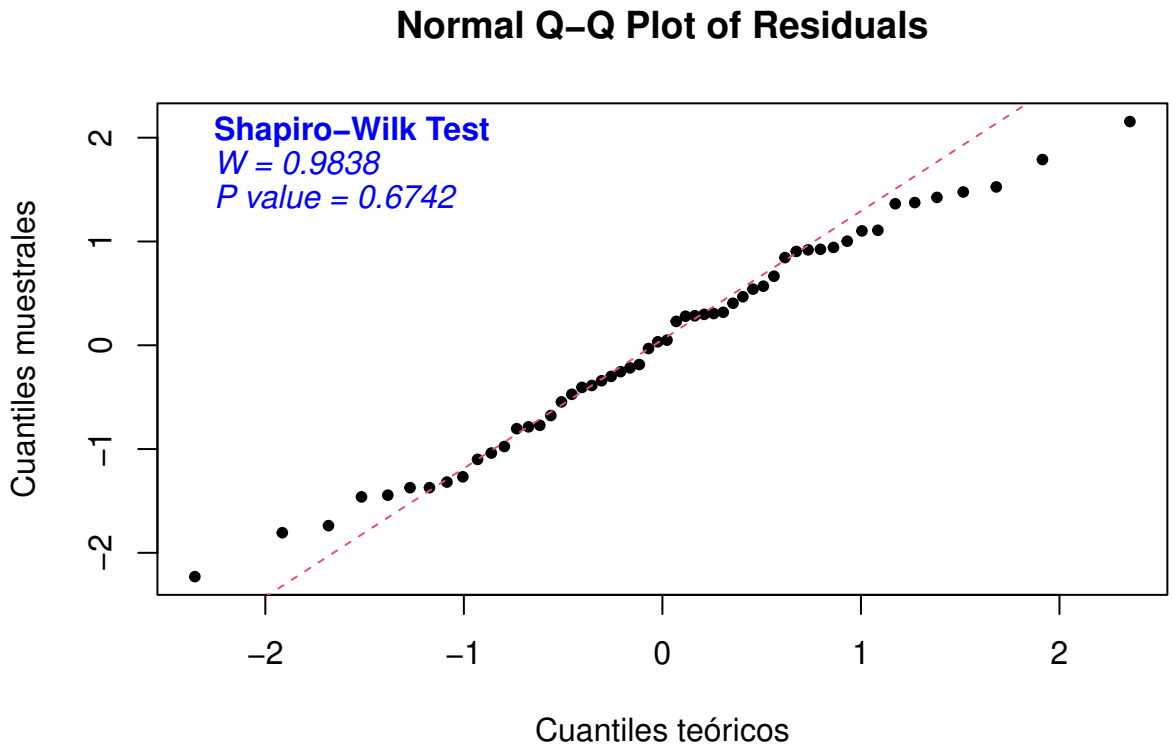

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.6742 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

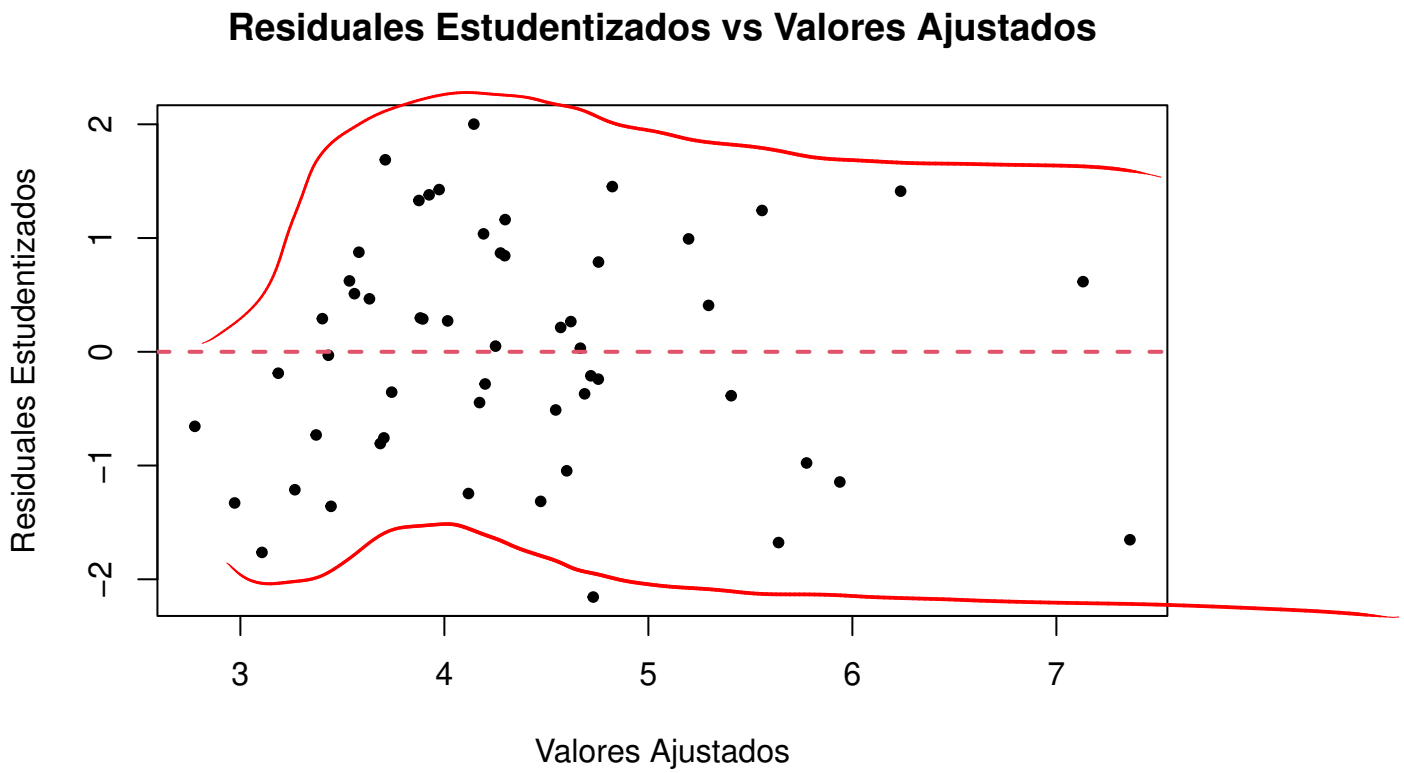

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente, decrezca ni un comportamiento que permita descartar una varianza constante, al no haber evidencia suficiente en contra de este supuesto se acepta como cierto. Además es posible observar media 0.

4.2. Verificación de las observaciones

Tengan cuidado acá, modifiquen los límites de las gráficas para que tenga sentido con lo que observan en la tabla diagnóstica. También, consideren que en aquellos puntos extremos que identifiquen deben explicar el qué causan los mismos en el modelo.

4.2.1. Datos atípicos

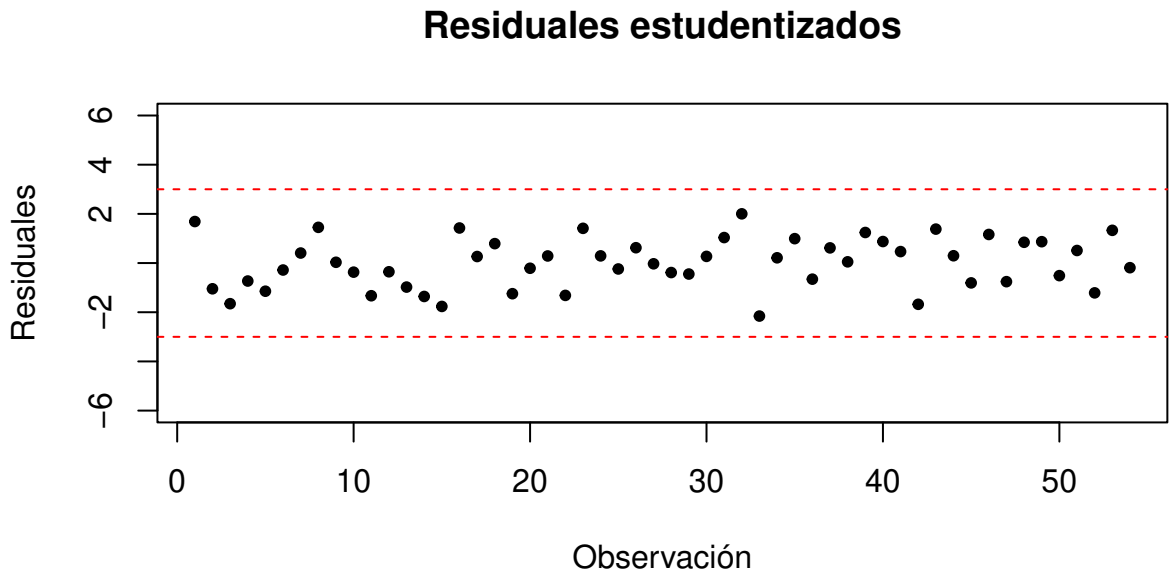

Figura 3: Identificación de datos atípicos

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

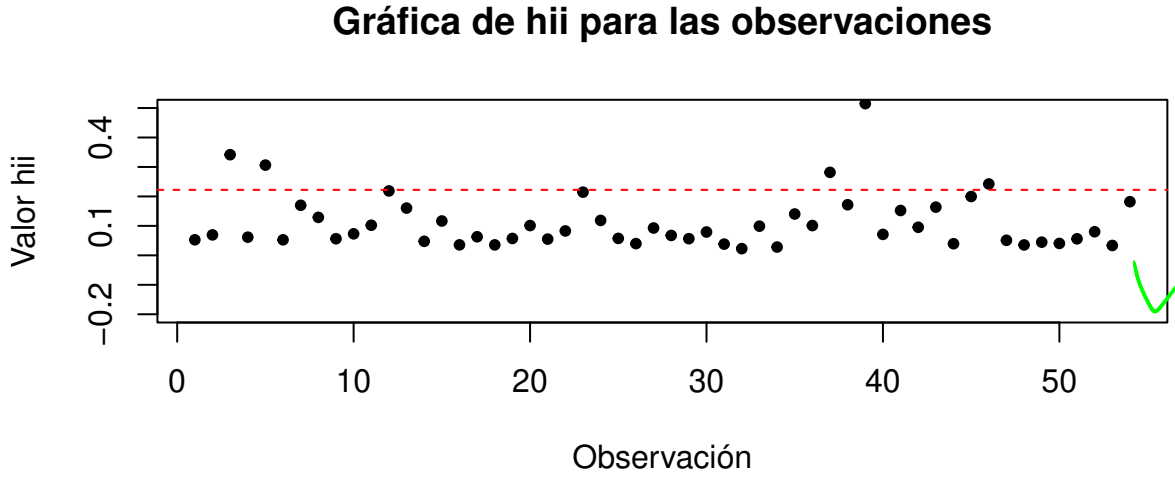

Figura 4: Identificación de puntos de balanceo

	res.stud	Cooks.D	hii.value	Dffits
3	-1.6524	0.2362	0.3416	-1.2129
5	-1.1446	0.0964	0.3063	-0.7631
37	0.6165	0.0248	0.2817	0.3836
39	1.2425	0.2732	0.5150	1.2878
46	1.1621	0.0719	0.2422	0.6594

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, se puede apreciar que existen 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

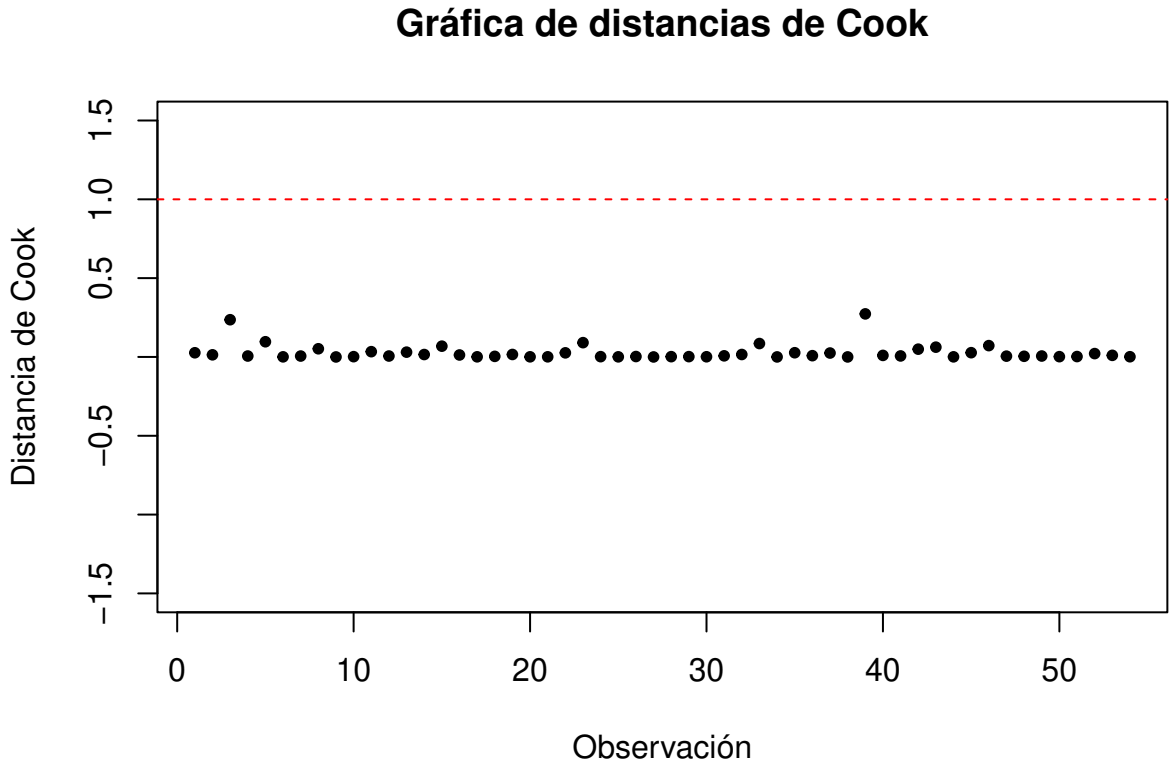

Figura 5: Criterio distancias de Cook para puntos influenciales

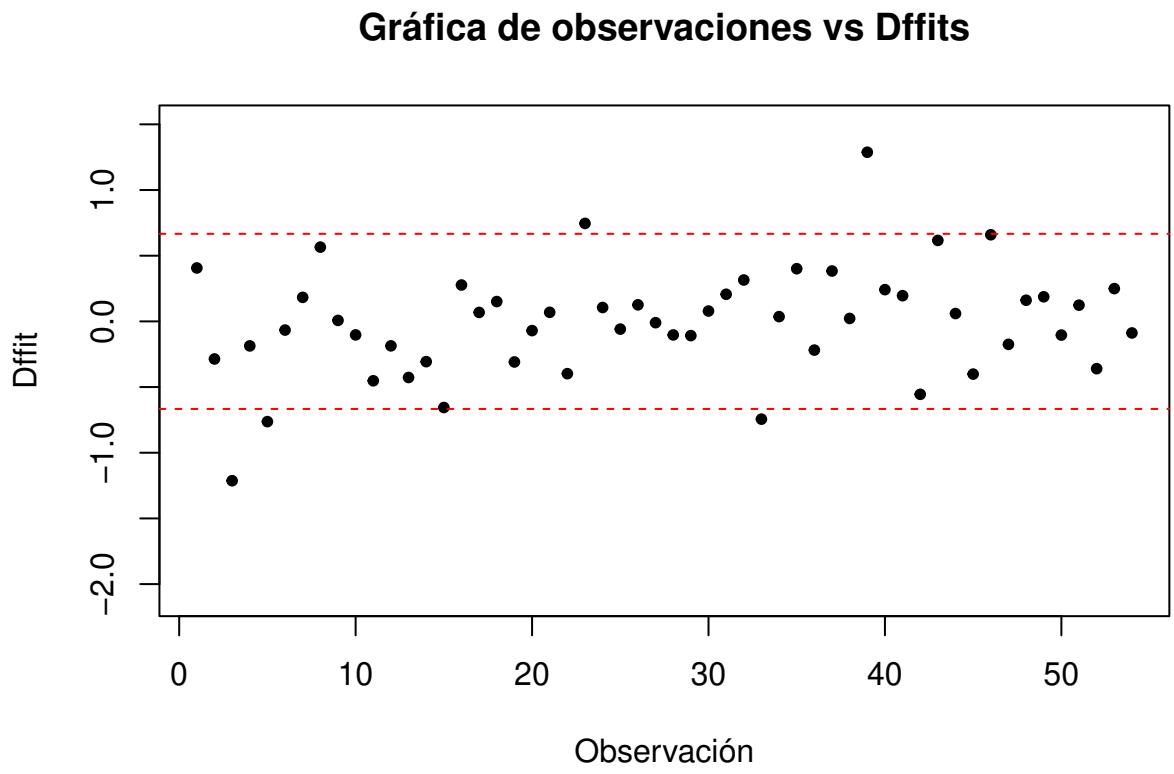

Figura 6: Criterio Dffits para puntos influenciales

	res.stud	Cooks.D	hii.value	Dffits
3	-1.6524	0.2362	0.3416	-1.2129
5	-1.1446	0.0964	0.3063	-0.7631
23	1.4120	0.0907	0.2145	0.7457
33	-2.1561	0.0851	0.0990	-0.7442
39	1.2425	0.2732	0.5150	1.2878

Como se puede ver,las observaciones 3,5,23,33,39 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. 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.

4.3. Conclusión

El modelo no es válido porque no se cumplen todos los supuestos, ya que en el supuesto de normalidad se puede observar que se rechaza el cumplimiento del supuesto debido a que en los extremos podemos ver qué las colas son más pesadas, además de tener patrones de irregularidades.