Trabajo 1

3,9

Estudiantes

Alejandro Torrado Calderón
Laura Valentina Cárdenas Luna
Sebastián Lavergne Vélez
Esteban Espinosa Parra

Equipo 14

Docente

Julieth Verónica Guarín 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: 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: 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	-2.1865
β_1	0.2755
β_2	0.0464
β_3	0.0412
β_4	0.0070
β_5	0.0008

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -2.1865 + 0.2755 X_{1i} + 0.0464 X_{2i} + 0.0412 X_{3i} + 0.007 X_{4i} + 8 \times 10^{-4} X_{5i}$$

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:

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

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	45.5096	5	9.101912	9.64814	2.01654e-06
Error	45.2825	48	0.943385		

De la tabla Anova, se observa un valor P cercano 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	-2.1865	1.7133	-1.2762	0.2080
β_1	0.2755	0.1077	2.5585	0.0137
β_2	0.0464	0.0313	1.4828	0.1447
β_3	0.0412	0.0139	2.9690	0.0047
β_4	0.0070	0.0079	0.8822	0.3820
β_5	0.0008	0.0008	1.1116	0.2719

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1 y β_3 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

$\hat{\beta}_1$: La duración promedio de la estadía de los pacientes del hospital aumenta la probabilidad de contraer una infección en un 27.55 % fijando las demás variables, esto se debe a que entre más tiempo se pasa en un ambiente contaminado, es más posible el riesgo de padecimiento.

$\hat{\beta}_3$: El aumento en la cantidad promedio de camas en un hospital incrementa la probabilidad de contraer una infección durante la estadía en un 6.1 % en comparación con cuando todas las demás variables se mantienen constantes. Esto se debe a que la presencia de un mayor número de camas, ya sea ocupadas o disponibles, aumenta el riesgo de transmisión de infecciones entre las personas dentro del hospital.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.50125$, lo que significa que aproximadamente el 50.12 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe. Hay que tener cuidado a la hora de utilizar el R^2 no ajustado dado que puede llegar a inflar el coeficiente, es decir, podría esbozar una interpretación incorrecta respecto a la variabilidad total observada en el modelo.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más alto en el modelo fueron X_2, 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:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	45.282	X1 X2 X3 X4 X5
Modelo reducido	49.302	X1 X3

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

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

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \ \cdots, \ \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,48} \\ &= \frac{(49.302 - 45.282)/3}{0.943385} \\ &= 4.26125 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,48} = 2.7981$, se puede ver que $F_0 > f_{0.95,1,45}$ y En consecuencia, con los datos disponibles y considerando la prueba de hipótesis realizada, se tiene suficiente evidencia para rechazar la hipótesis nula y, por lo tanto, aceptar la hipótesis alternativa. Esto lleva a concluir que el conjunto de variables es significativo. En otras palabras, no se puede descartar las variables involucradas, ya que al menos una de ellas no es nula, dado que el conjunto de variables no es igual a cero.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si la duración promedio de la estadía tiene alguna relación con el número promedio de camas en el hospital. Así como también si la razón del número de cultivos realizados en pacientes sin síntomas está en proporción del número promedio de pacientes. Por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 2\beta_3;\ \beta_2 = \dfrac{2}{3}\beta_4 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

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

Donde $X^*_{3i} = 2X_{1i} + X_{3i}$ y $X^*_{4i} = \frac{2}{3}X_{2i} + X_{4i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(49.302 - 45.282)/2}{0.943385} \overset{H_0}{\sim} f_{2,48} \tag{3}$$

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 realizará por medio de shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

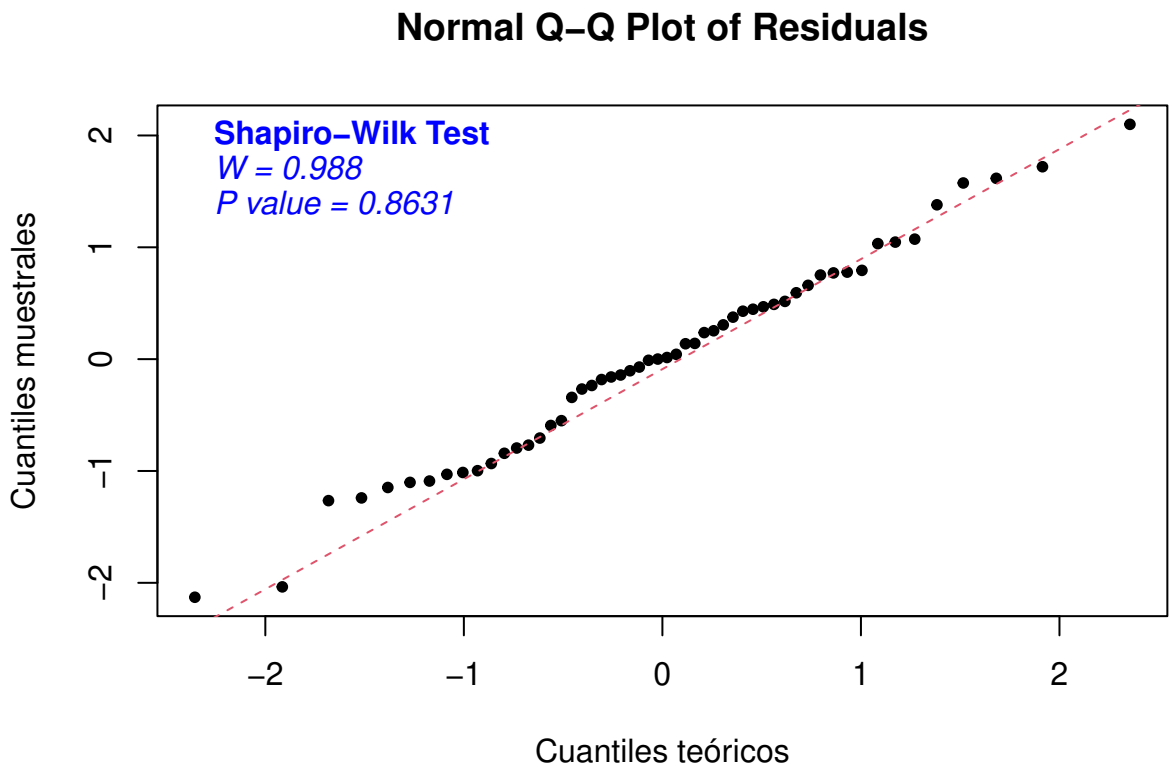

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

En primer lugar, al analizar la gráfica de comparación de cuantiles, se puede notar que en el intervalo de valores entre -1 y 1, los datos parecen seguir una distribución normal. Sin embargo, en los extremos de la gráfica, se observan patrones irregulares que sugieren que la distribución no es completamente normal.

En cuanto al Valor-P, que es aproximadamente 0.8631, y considerando un nivel de significancia de $\alpha = 0.05$, dado que el Valor-P es significativamente mayor que el valor de alpha, no se tendría suficiente evidencia para rechazar la hipótesis nula. Esto implica que los datos podrían seguir una distribución normal con una media $\mu = 0$ y una varianza constante σ^2.

Sin embargo, al tener en cuenta el análisis gráfico más detallado, se llega a la conclusión de que la hipótesis nula debe ser rechazada. En otras palabras, los datos no siguen estrictamente una distribución normal, ya que se observan patrones irregulares en la gráfica, lo que lleva a deducir que los valores ϵ_i no se distribuyen de manera normal.

4.1.2. Varianza constante

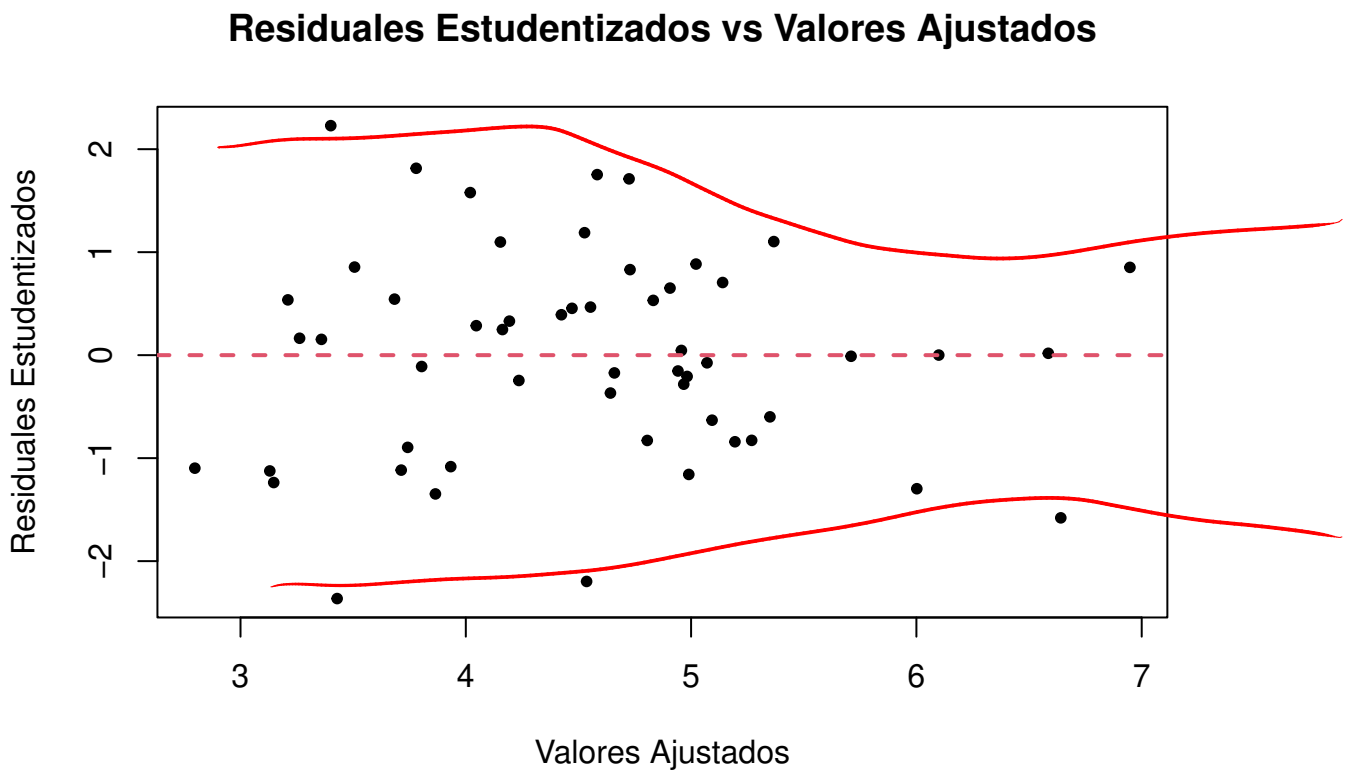

Figura 2: Gráfico residuales estudentizados vs valores ajustados

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

4.2.1. Datos atípicos

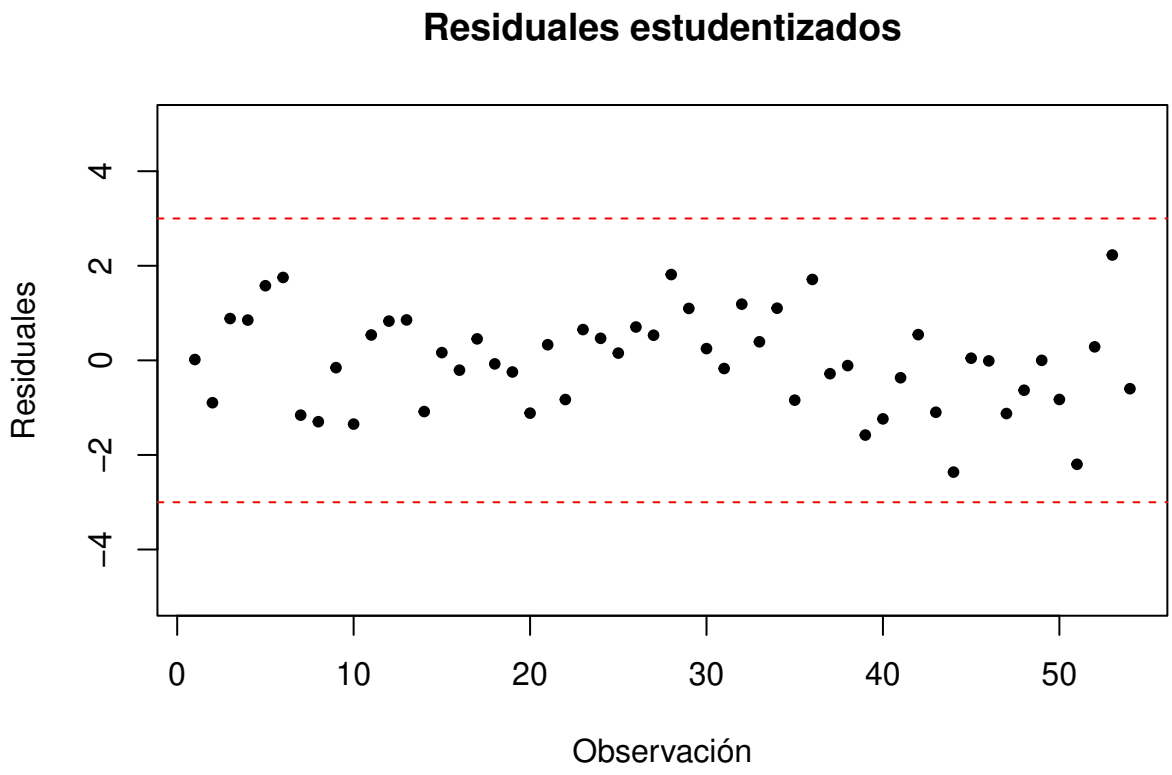

Figura 3: Identificación de datos atípicos

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

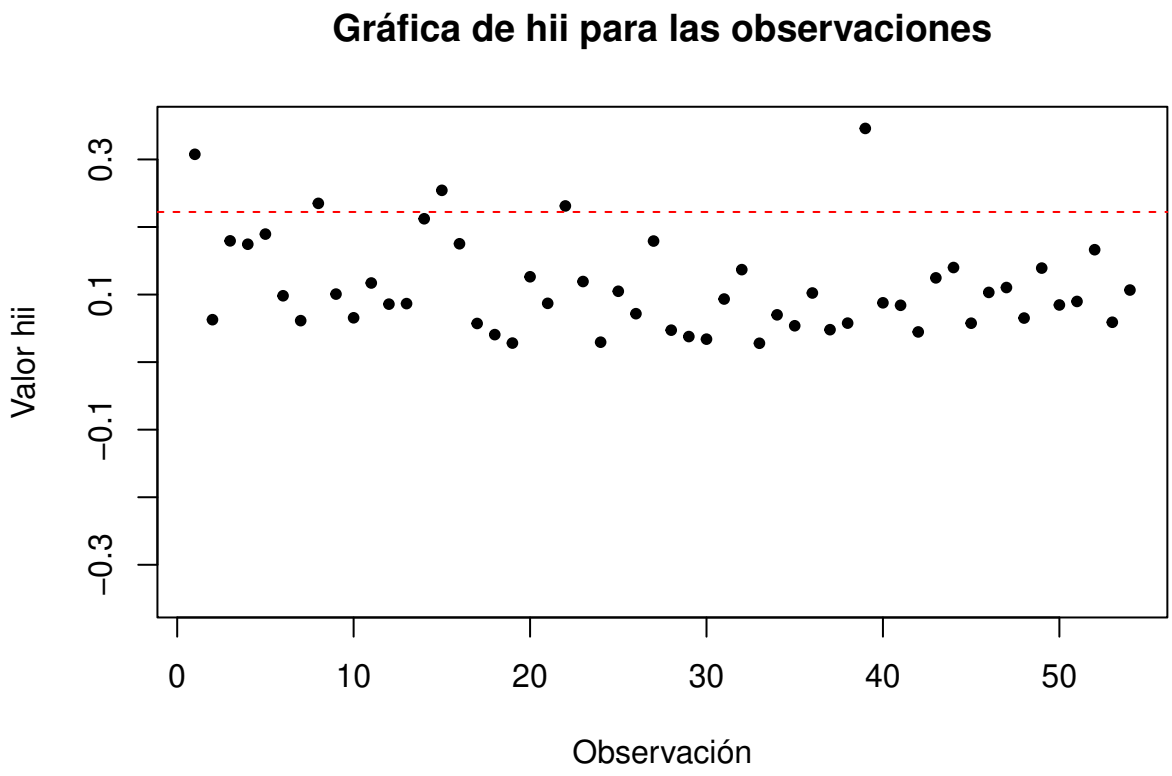

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 1    0.0181  0.0000    0.3076  0.0119
## 8   -1.2971  0.0862    0.2350 -0.7243
## 15   0.1641  0.0015    0.2543  0.0949
## 22  -0.8284  0.0344    0.2312 -0.4528
## 39  -1.5797  0.2200    0.3459 -1.1676
```

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 2 * \frac{6}{54} = 0.22222$, se puede apreciar que existen 5 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 0.22222$, los cuales son los presentados en la tabla.

¿Qué causan?

2p+

4.2.3. Puntos influenciales

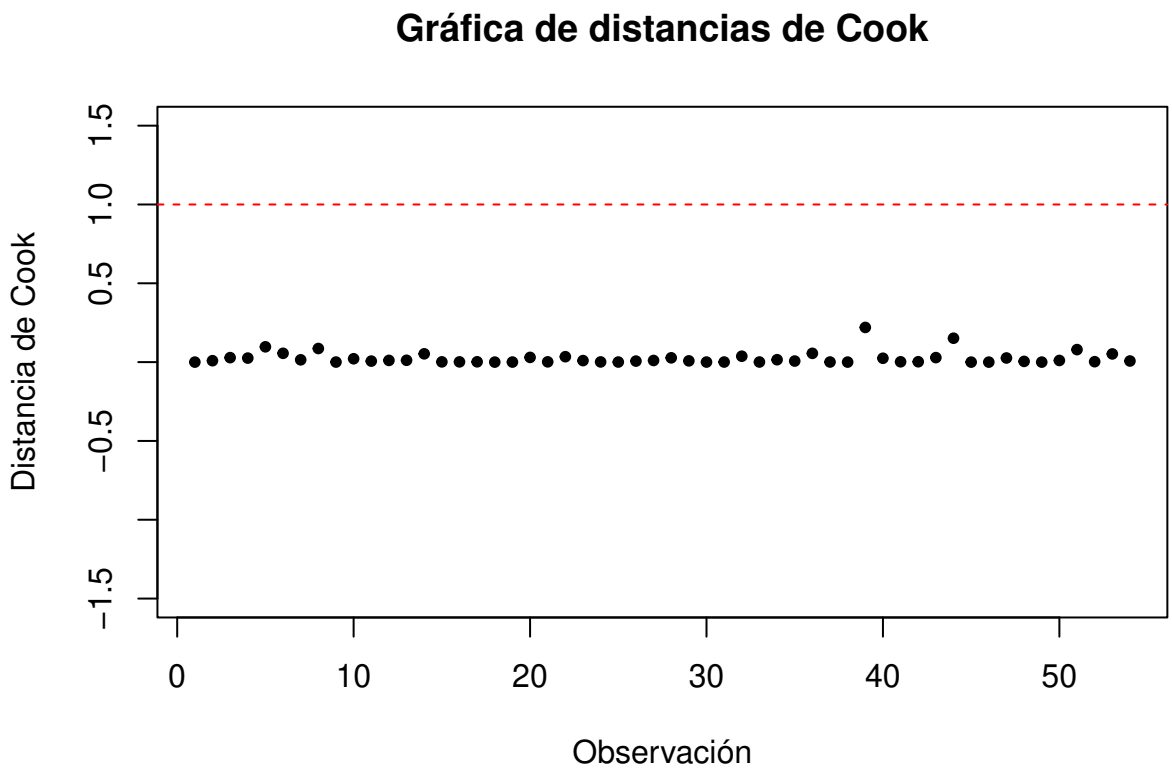

Figura 5: Criterio distancias de Cook para puntos influenciales

```
## [1] res.stud  Cooks.D   hii.value Dffits
## <0 rows> (or 0-length row.names)
```

Al analizar los datos observados, se puede notar que todos se encuentran por debajo de la línea roja punteada que representa el criterio $D_i > 1$. No hay ningún punto que se destaque de manera significativa del resto en el gráfico, lo que significa que no hay observaciones atípicas que tengan un impacto notable en los coeficientes de regresión ajustados. En otras palabras, no existe ningún dato individual que influya de manera sustancial en los resultados de la regresión.

Ojo, una cosa es atípico y otra influencial

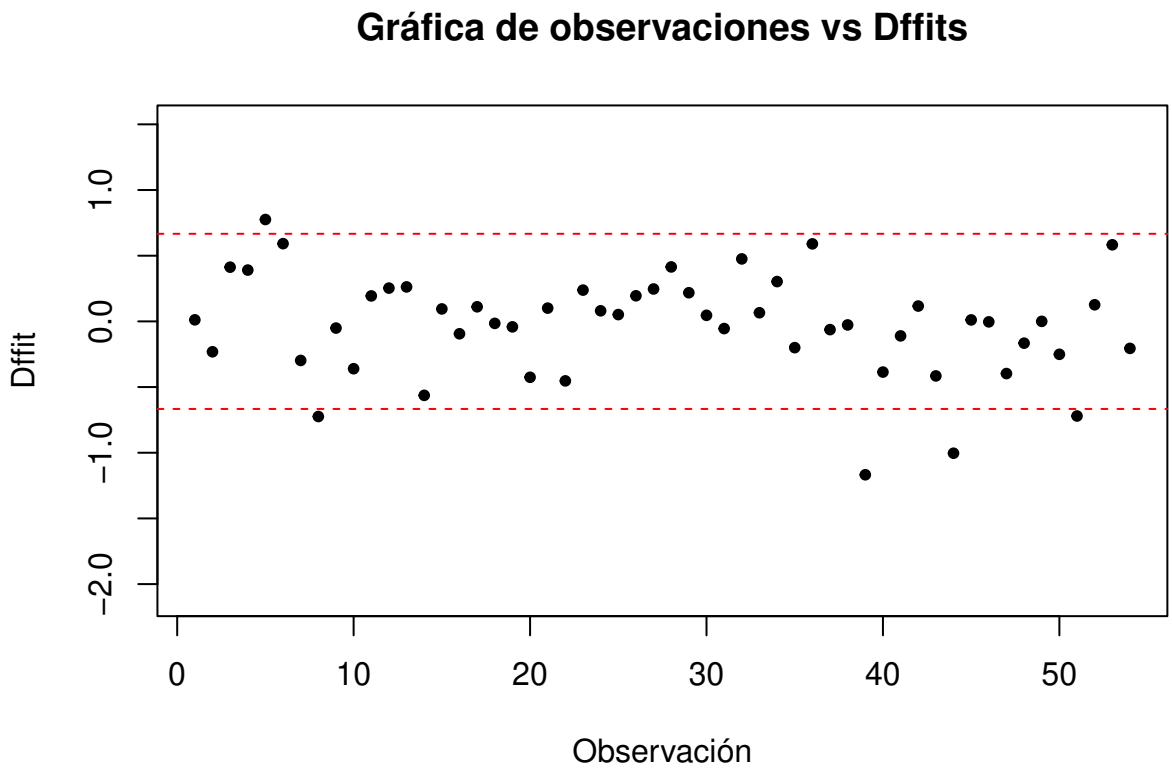

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 5    1.5783  0.0971    0.1895  0.7756
## 8   -1.2971  0.0862    0.2350 -0.7243
## 39  -1.5797  0.2200    0.3459 -1.1676
## 44  -2.3635  0.1515    0.1400 -1.0037
## 51  -2.1974  0.0794    0.0898 -0.7203
```

Como se puede ver hay 5 observaciones que no se encuentran entre las líneas rojas, estos son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{6}{54}} = 0.66667$, es un punto influencial

4.3. Conclusión

Los resultados de la regresión muestran una asociación estadísticamente significativa entre las variables analizadas. No obstante, es importante señalar que los residuos no siguen una distribución normal y que el supuesto de varianza constante no se cumple, lo que plantea dudas sobre la adecuación del modelo para representar los datos de manera precisa. Es posible que se requieran enfoques adicionales, como la transformación de variables o la exploración

de modelos alternativos, para mejorar la idoneidad del modelo y asegurar la confiabilidad de los resultados.

Finalmente se da el modelo como no válido debido a las explicaciones expuestas anteriormente.