Trabajo 1

4,4

Estudiantes

Andrea Sepúlveda Gaviria
Juliana Ramírez Tamayo
Angela Maria Quinche Enríquez
Carlos Julio Nieto Morales

Equipo 03

Docente

Julieth Verónica Guarín Escudero

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:

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

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	0.0224
β_1	0.2205
β_2	-0.0006
β_3	0.0564
β_4	0.0111
β_5	0.0015

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.0224 + 0.2205 X_{1i} - 0.0006 X_{2i} + 0.0564 X_{3i} + 0.0111 X_{4i} + 0.0015 X_{5i};\ 1 \leqslant i \leqslant 69$$

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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,63} \tag{1}$$

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	55.3615	5	11.072298	13.5717	5.49251e-09
Error	51.3976	63	0.815836		

Tomando $\alpha = 0.05$ se observa en la tabla Anova un valor P muy pequeño aproximadamente igual a 0, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, es decir, al menos un parámetro β_j con $1 \leqslant j \leqslant 5$ es significativo.

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	0.0224	1.3870	0.0162	0.9872
β_1	0.2205	0.0932	2.3654	0.0211
β_2	-0.0006	0.0263	-0.0244	0.9806
β_3	0.0564	0.0148	3.7986	0.0003
β_4	0.0111	0.0069	1.6127	0.1118
β_5	0.0015	0.0007	2.2588	0.0274

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3 y β_5 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

3pt

La interpretración de los parámetros β_1, β_3 y β_5 es la siguiente (No se tienen en cuenta los demás parámetros ya que no son significativos):

$\hat{\beta}_1$**:** Por cada día que aumenta la duración de la estadía de los pacientes en el hospital el riesgo de infección se incrementa en promedio 0.2205 % cuando las demás variables se mantienen constantes.

$\hat{\beta}_3$**:** A medida que aumenta el número de camas en el hospital el riesgo de infección se incrementa en promedio 0.0564 % cuando las demás variables se mantienen constantes.

$\hat{\beta}_5$**:** A medida que aumenta el número de enfermeras en el hospital el riesgo de infección se incrementa en promedio 0.0015 % cuando las demás variables se mantienen constantes.

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5186$, lo que significa que aproximadamente el 51.86 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

¿Cómo se calculó?

2. Pregunta 2

5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más pequeño en el modelo fueron X_1, 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:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	51.398	X1 X2 X3 X4 X5
Modelo reducido	80.962	X2 X4

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

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

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0,\beta_2,\beta_4) - SSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5))/3}{MSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{9.8547}{0.8158} \\ &= 12.0798 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 > f_{0.95,3,63}$, por lo que se rechaza la hipótesis nula, esto quiere decir que al menos una de las variables del subconjunto es significativa y por lo tanto no es posible descartarlas.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si β_1 es equivalente a β_4 y si β_3 es equivalente a β_5 por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_4;\ \beta_3 = \beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_2 X_{2i} + \beta_3 X^*_{3i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2);\ 1 \leqslant i \leqslant 69$$

Donde $X^*_{1i} = X_{1i} + X_{4i}$ y $X^*_{3i} = X_{3i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 51.398/2}{0.8158} \overset{H_0}{\sim} f_{2,63} \tag{3}$$

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:

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

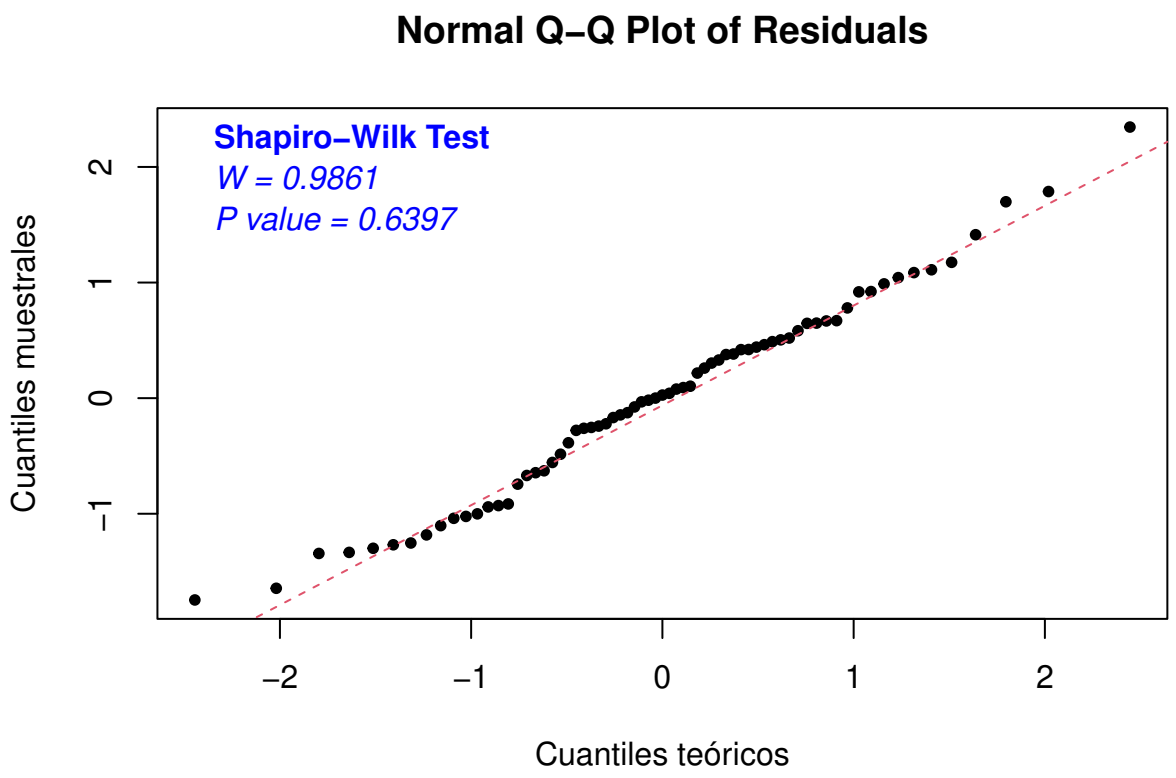

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.6397 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, además la gráfica de comparación de cuantiles permite ver que la mayoría de datos están muy cerca de la línea roja, presentando patrones irregulares. Cabe destacar que en las colas hay una leve dispersión de unos pocos datos, pero se termina aceptando la hipótesis nula, es decir, el supuesto se cumple.

4.1.2. Varianza constante

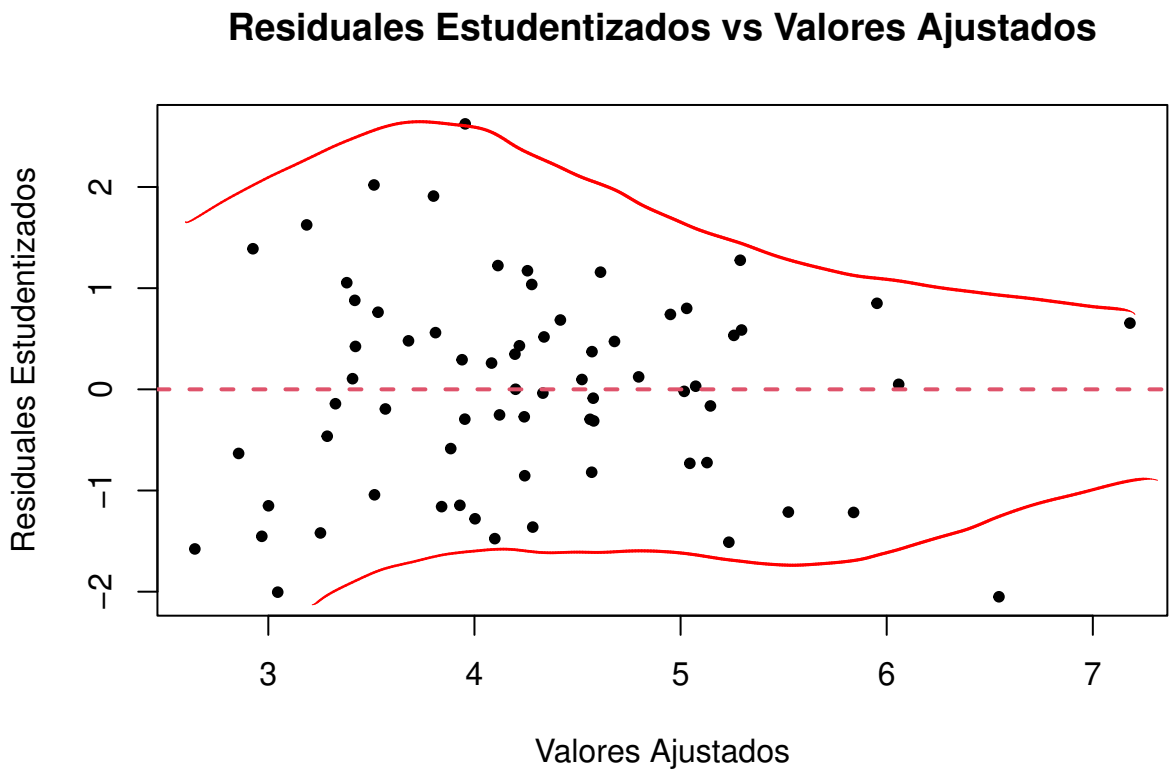

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que hay patrones en los que la varianza aumenta y decrece, es decir, no hay ningún comportamiento que permita aceptar una varianza constante, por lo que al no haber evidencia suficiente a favor de este supuesto se rechaza la hipótesis nula, es decir, no se cumple el supuesto de varianza constante. Además es posible observar media 0.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

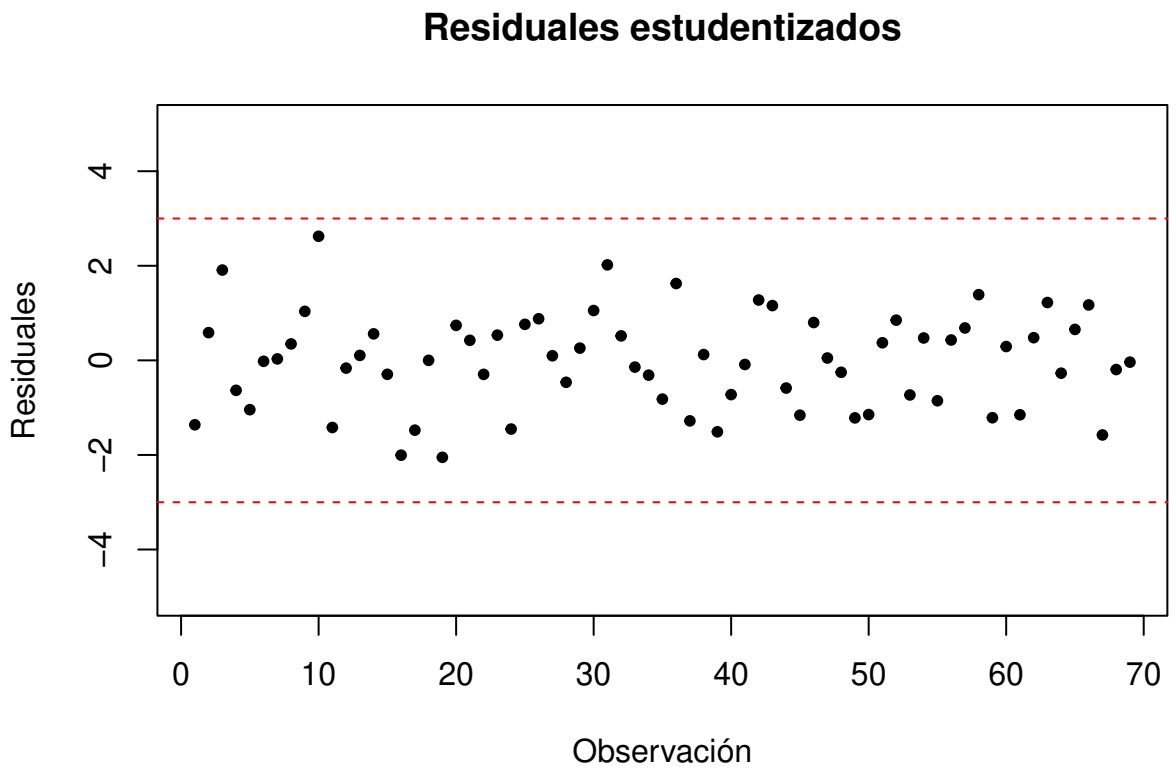

Figura 3: Identificación de datos atípicos

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

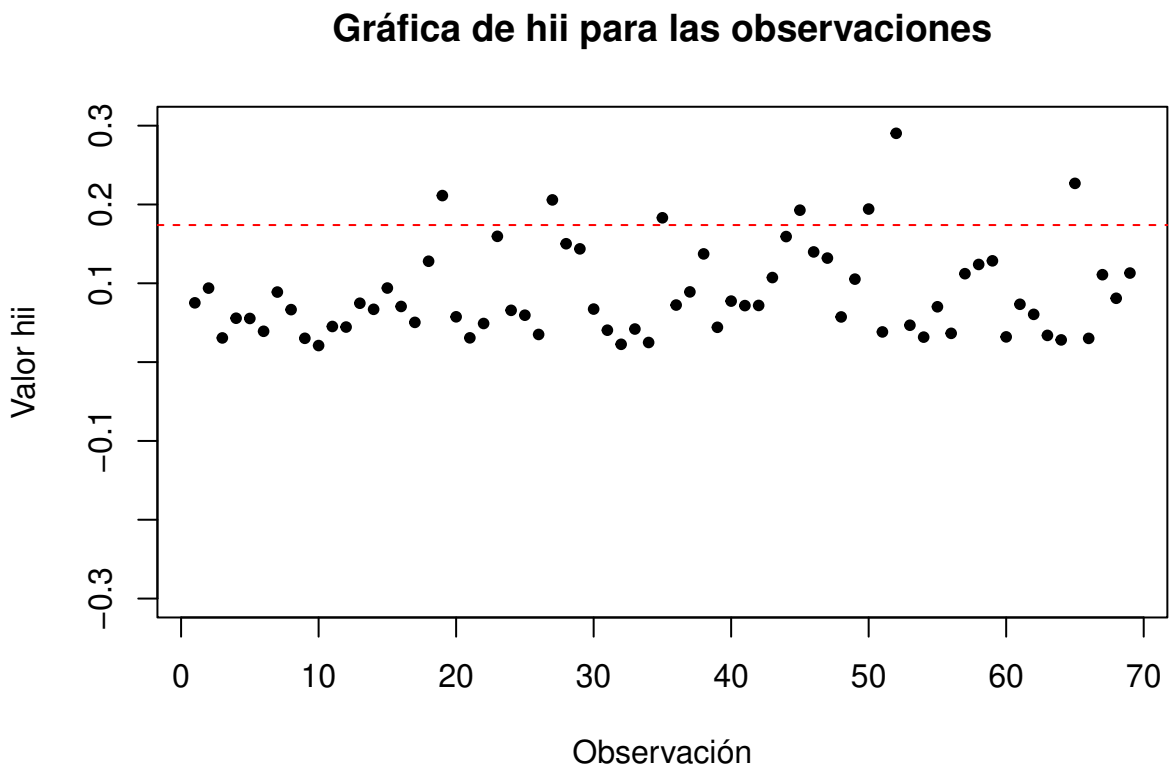

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 19  -2.0502  0.1876    0.2113 -1.0895
## 27   0.0973  0.0004    0.2058  0.0492
## 35  -0.8193  0.0251    0.1831 -0.3869
## 45  -1.1593  0.0535    0.1928 -0.5681
## 50  -1.1462  0.0528    0.1942 -0.5642
## 52   0.8504  0.0493    0.2903  0.5427
## 65   0.6546  0.0209    0.2268  0.3529
```

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 7 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.

¿Qué causan? 2 pt

4.2.3. Puntos influenciales

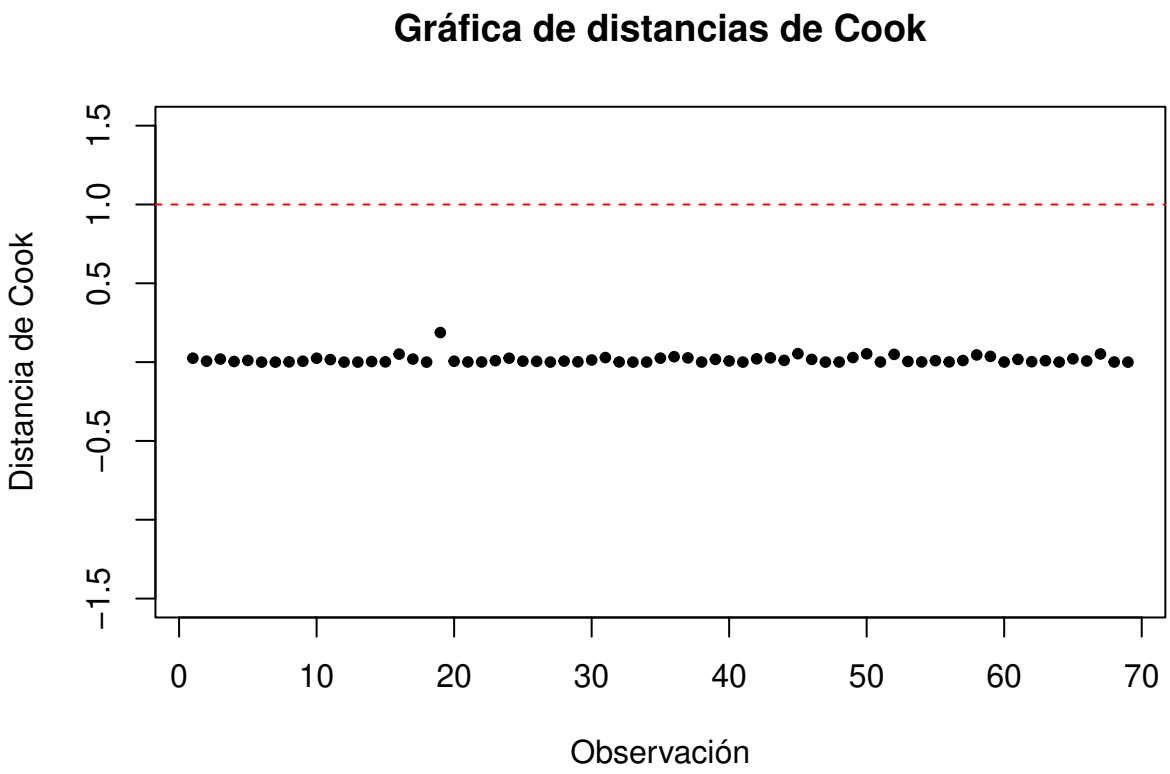

Figura 5: Criterio distancias de Cook para puntos influenciales

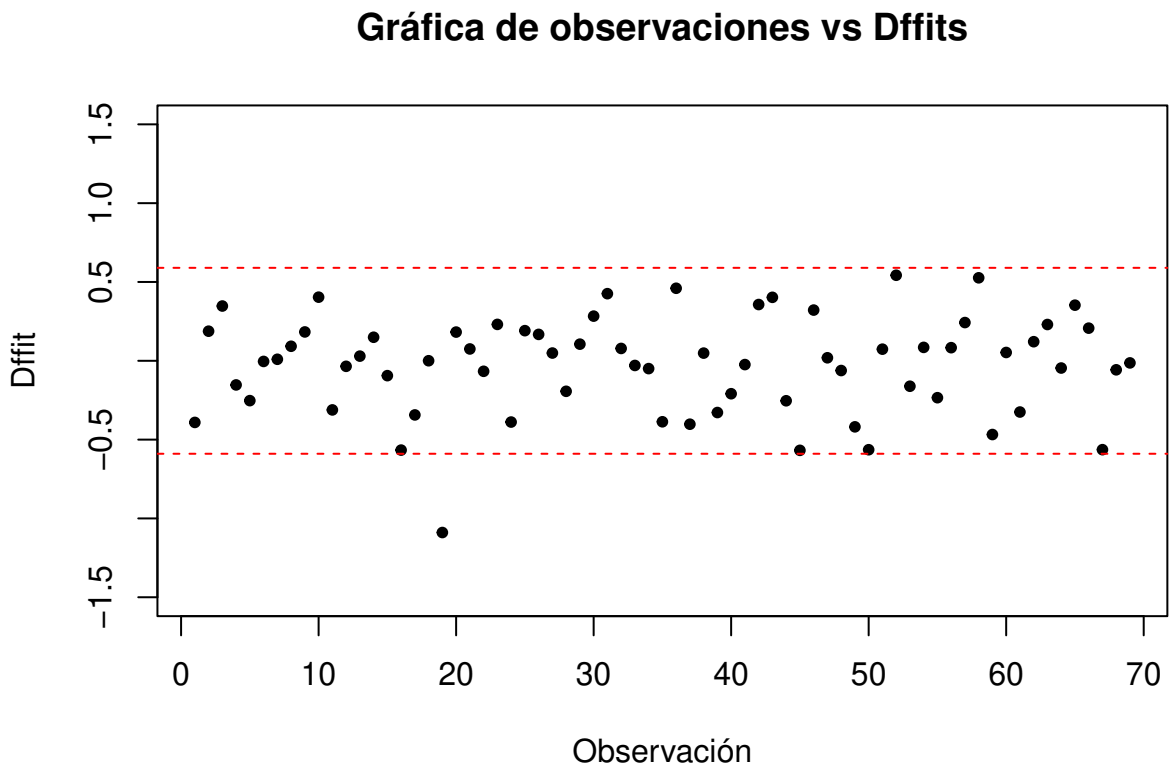

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 19  -2.0502  0.1876    0.2113 -1.0895
```

Como se puede ver,la observación número 19 es un punto influencial 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

Para verificar la validez de un modelo se deben tener en cuenta varios aspectos. Aunque el modelo trabajado no cuenta con varianza constante, tiene puntos de balanceo y un punto influencial, lo que puede afectar su validez, este si cumple con el supuesto de normalidad y presenta una ausencia de valores atípicos. Ahora bien, considerando todos los supuestos en conjunto el modelo de regresión parece válido.