Trabajo 1

Estudiantes

Alejandro Diaz López
Juan José Flórez Ospina
Juan Diego Giraldo Jaramillo
Mariam Saavedra Navaja

Equipo 6

Docente:

Julieth Veronica Guarin Escudero

Asignatura:

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín

30 de marzo 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 : Riesgo de infección

X_1: Duración de la estadía

X_1: Rutina de vultivos

X_1: Número de camas

X_1: Censo promdeio diario

X_1: Número de enfermeras

$$Y_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 60$$

1.1. Modelo de regresión

Al cargar y ajustar el modelo lineal, se obtienen los siguientes coeficientes estimados del MRLM:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	0.0882
β_1	0.1873
β_2	-0.0100
β_3	0.0605
β_4	0.0203
β_5	0.0008

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.0882 + 0.1873 X_{1i} - 0.01 X_{2i} + 0.0605 X_{3i} + 0.0203 X_{4i} + 0.0008 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 60$$

1.2. Significancia de la regresión

Para realizar la prueba de significancia del modelo de regresión planteamos el siguiente juego de hipótesis

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \text{algún } \beta_j \neq 0,\ \text{j=1, 2, 3, 4, 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,54} \tag{1}$$

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	Grados de libertad.	Cuadrado medio	F_0	valor P
Regresión	91.8115	5	18.362302	21.6734	7.80158e-12
Error	45.7503	54	0.847228		

De la tabla Anova, se compara el valor P que es de 7.80158e-12, con un nivel de significancia $\alpha = 0.05$ permitiendo que se rechaze la hipótesis nula en la que $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, probando que existe una relacion en la regresión.

1.3. Significancia de los parámetros

En la siguiente tabla se aprecia la información de los parámetros que permite determinar la significancia de cada uno de estos.

Cuadro 3: Resumen de los coeficientes

	$Valorestimado$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	0.0882	1.5010	0.0588	0.9533
β_1	0.1873	0.0690	2.7145	0.0089
β_2	-0.0100	0.0285	-0.3514	0.7267
β_3	0.0605	0.0154	3.9187	0.0003
β_4	0.0203	0.0072	2.8405	0.0063
β_5	0.0008	0.0008	1.0852	0.2826

Para analizar la significancia de los coeficientes en el modelo lineal emplearemos un nivel de significancia = 0.05 para comparar con el valor P arrojado por el resumen de los coeficientes.

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

$\hat{\beta}_3$:

$\hat{\beta}_5$:

1.5. Coeficiente de determinación múltiple R^2

Para hallar el coeficiente de determinación múltiple R^2 empleamos el SSR y SSE dados por la tabla ANOVA

$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE} = \frac{91.8115}{137.5618} = 0.667420$

Este coeficiente de determinación nos dice que aproximadamente el 66.74 % de la variabilidad 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

Las variables con el valor P más alto en el modelo fueron X_1, X_2, X_5

para probar la significancia simultánea de estos tres coeficientes de la regresión planteamos las hipótesis:

$$\begin{cases} \text{H}_0 : \beta_1 = \beta_2 = \beta_5 = 0 \\ \text{H}_1 : \text{algún } \beta_j \neq 0, \text{ j=1, 2, 5} \end{cases}$$

Cuadro 4: tabla de todas las regresiones resumida

	SSE	Covariables en el modelo
Modelo completo	45.750	X1 X2 X3 X4 X5
Modelo reducido	57.368	X3 X4

El modelo completo es el visto en el inicio de la pregunta 1.

El modelo reducido es de la forma:

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 60$$

2.2. Estadístico de prueba y conclusión

Se calcula el estadistico de la prueba de la forma:

$$
\begin{aligned}
F_0 &= \frac{SSR(\beta_1, \beta_2, \beta_5 | \beta_0, \beta_3, \beta_4)/3}{SSE(\beta_0, \beta_1, \beta_3, \beta_4, \beta_5)/54} \\
&= \frac{(SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0, \beta_1, \beta_3, \beta_4, \beta_5))/3}{SSE(\beta_0, \beta_1, \beta_3, \beta_4, \beta_5)/54} \\
&= \frac{11.618/3}{47.750/54} = 4.349557 \overset{H_0}{\sim} f_{3,54}
\end{aligned} \tag{2}
$$

Para el criterio de decisión calculamos el valor crítico a un nivel de significancia $\alpha = 0.05$ de una distribución $f_{0.05,3,54} = 4.349557$

Como $F_0 > f_{0.05,3,54}$, se rechaza la hipotesis nula, lo que quiere decir que al menos una de las variables regresoras asociadas a la Duración de la estadía, Rutina de culitivos y Número de enfermeras (X_1, X_2, X_5), es significativa en presencia del resto de variables y por lo tanto hace a este conjunto un conjunto significativo y no podemos descartarlo. subconjunto.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se quiere estudiar si el efecto de la duración de estadía de los pacientes en el hospital es igual a la rutina de cultivos realizados en los pacientes sin síntoma de infección hospitalaria, por cada 100 pacientes. Además deseamos estudiar si el efecto promedio de camas en el hospital durante el periodo del estudio es igual al efecto del número promedio de pacientes en el hospital por día durante el periodo de estudio.

Para responder a la pregunta se plantea la siguiente prueba de hipótesis:

$$
\begin{cases}
\mathrm{H}_0 : \beta_1 = \beta_2; \ \beta_3 = \beta_4 \\
\mathrm{H}_1 : \beta_1 \neq \beta_2 \ \ \text{ó} \ \ \beta_3 \neq \beta_4
\end{cases}
$$

O equivalentemente,

$$
\mathrm{H}_0 : \beta_1 - \beta_2 = 0 \ ; \ \beta_3 - \beta_4 = 0
$$

Además, se puede representar matricialmente de la siguiente forma:

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

Con $\mathbf{L}$ dada por

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

El modelo reducido es:

$$\begin{aligned} Y &= \beta_o + \beta_1(X_1 + X_2) + \beta_3(X_3 + X_4) + \beta_5 X_5 + \varepsilon \\ &= \beta_o + \beta_1 X_{1,2} + \beta_3 X_{3,4} + \beta_5 X_5 \end{aligned} \tag{3}$$

Donde $X_{1,2} = X_1 + X_2$ y $X_{3,4} = X_3 + X_4$

3.2. Estadístico de prueba

Se tiene que el estadístico de prueba F_0 está dado por:

$$F_0 = \frac{SSH/gl.ssh}{MSE(MF)} = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,54} \tag{4}$$

donde SSE(RM) corresponde al error estándar del modelo reducido, SSE(RM) corresponde al error estándar del modelo full, r corresponde al número de filas linealmente independientes en la matriz L que son 2, y el MSE(FM) a la media de errores estándar del modelo full.

Si $F_0 > f_{0.05,2,54}$ entonces se rechaza la hipotesis nula y al menos una de los dos supuestos que queremos probar no se cumple con una significancia del 95 %.

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:

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

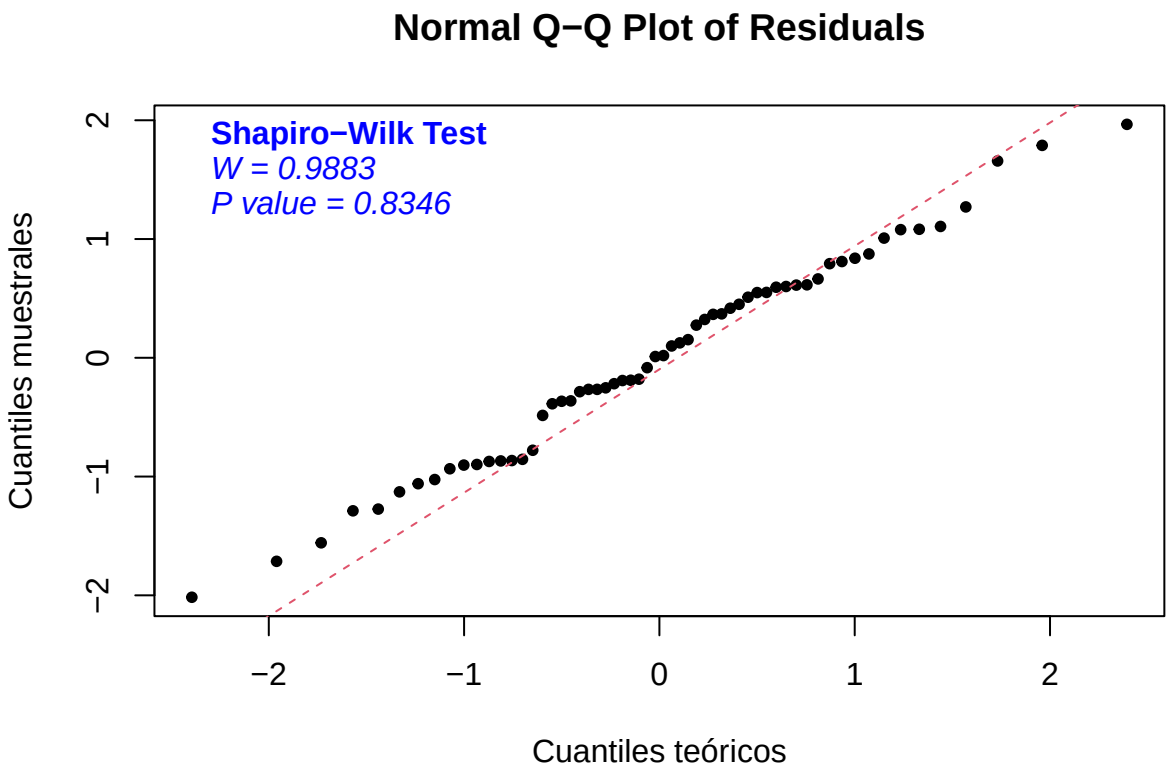

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.5951 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

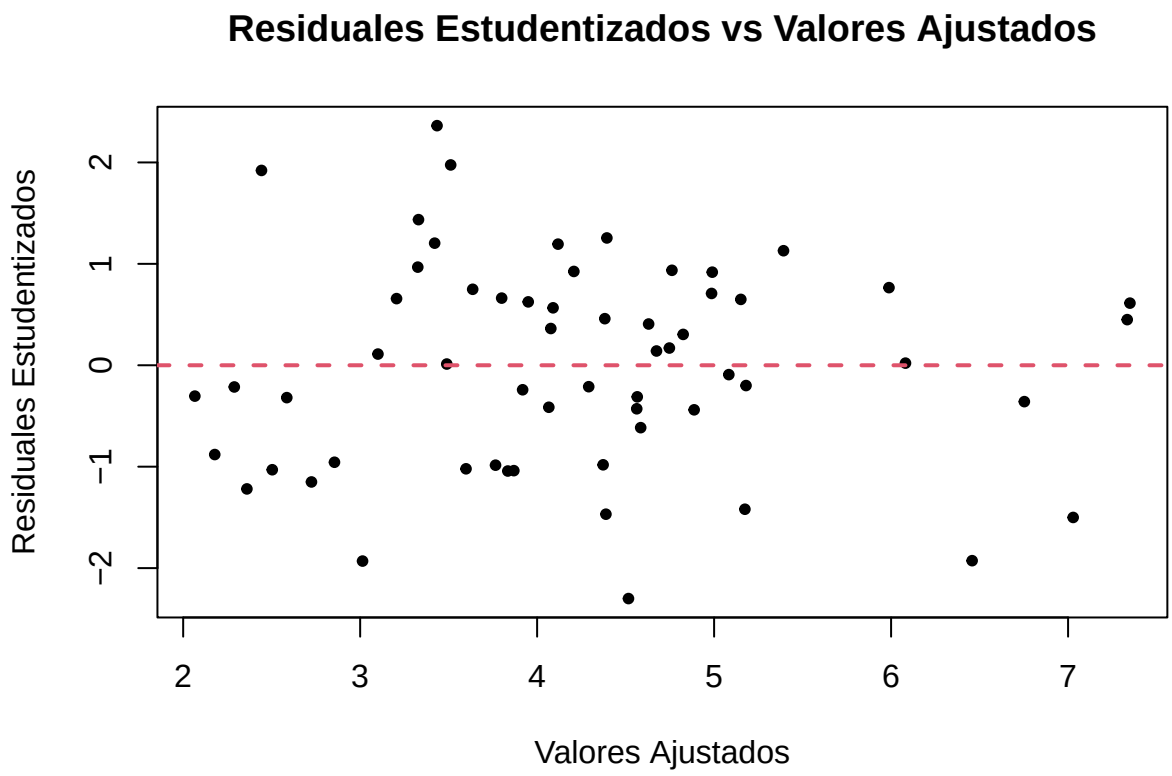

Figura 2: Gráfico residuales estudentizados vs valores ajustados

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

4.2.1. Datos atípicos

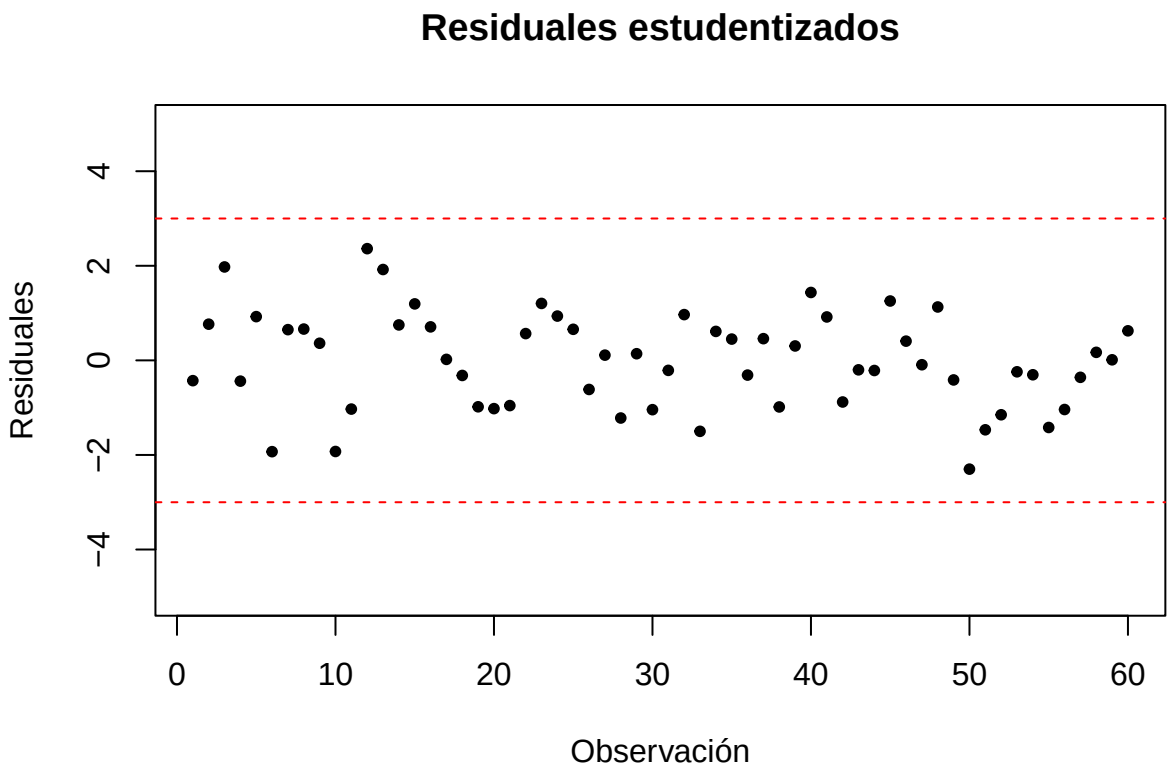

Figura 3: Identificación de datos atípicos

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

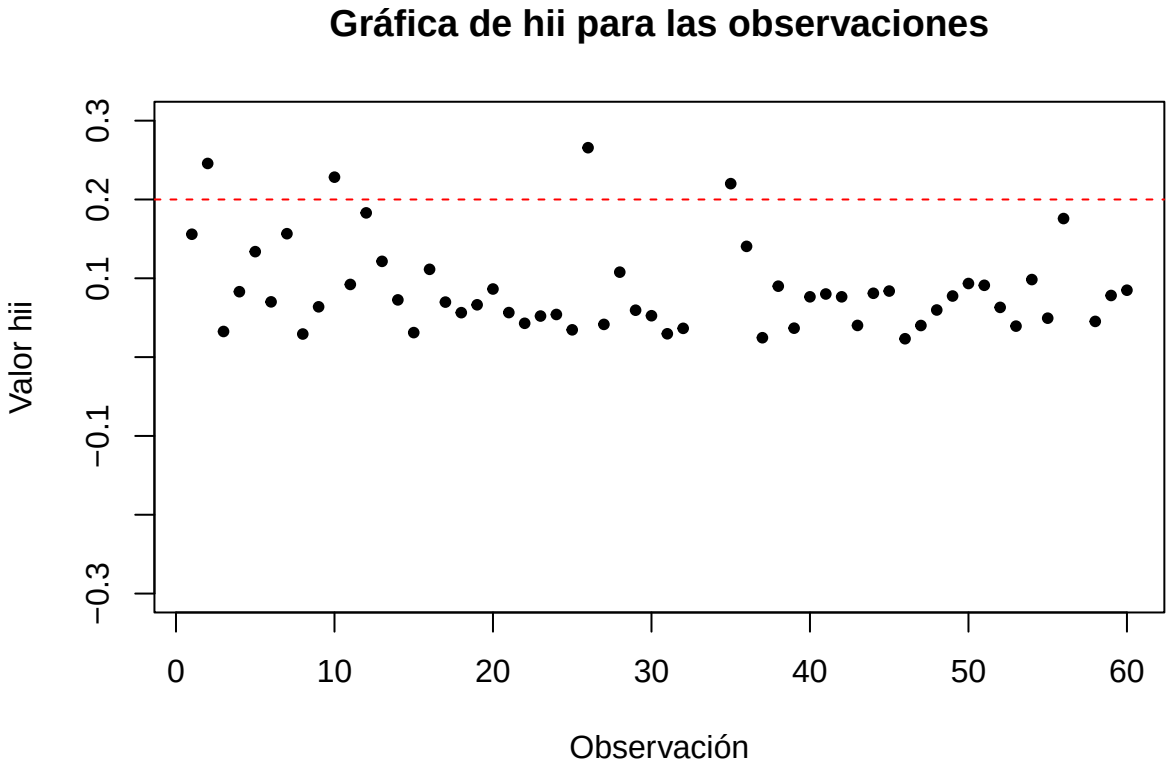

Figura 4: Identificación de puntos de balanceo

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

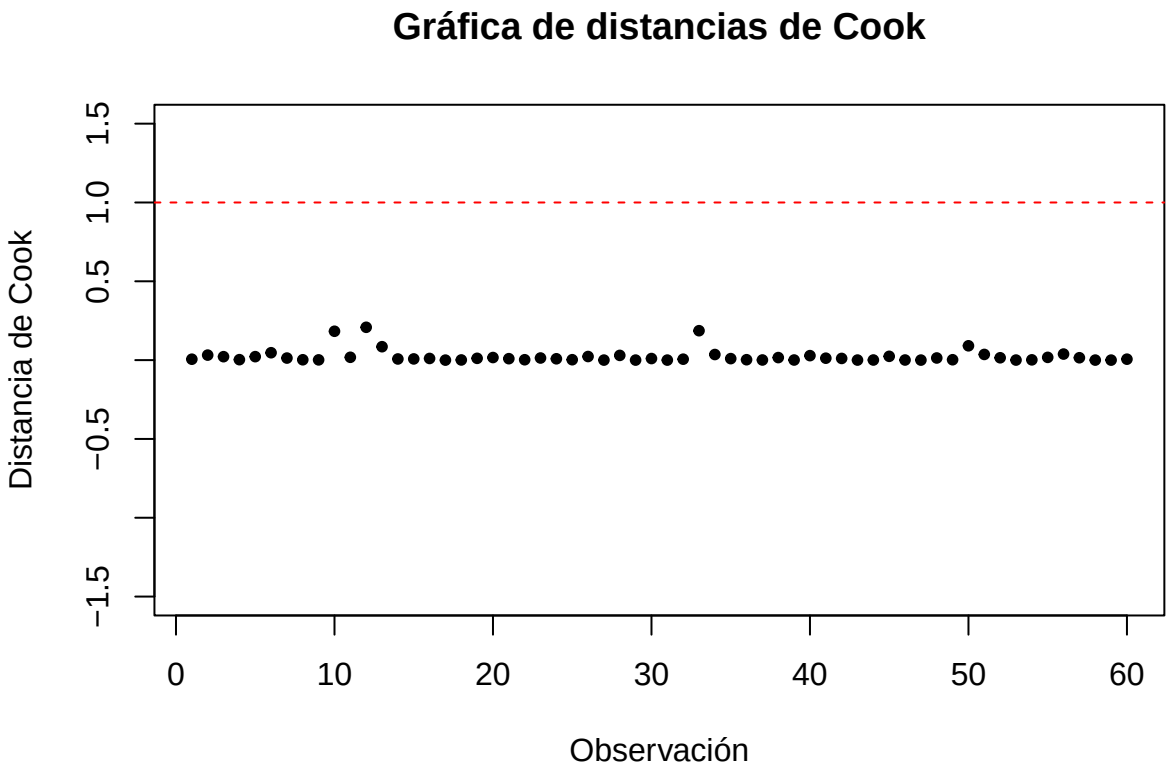

Figura 5: Criterio distancias de Cook para puntos influenciales

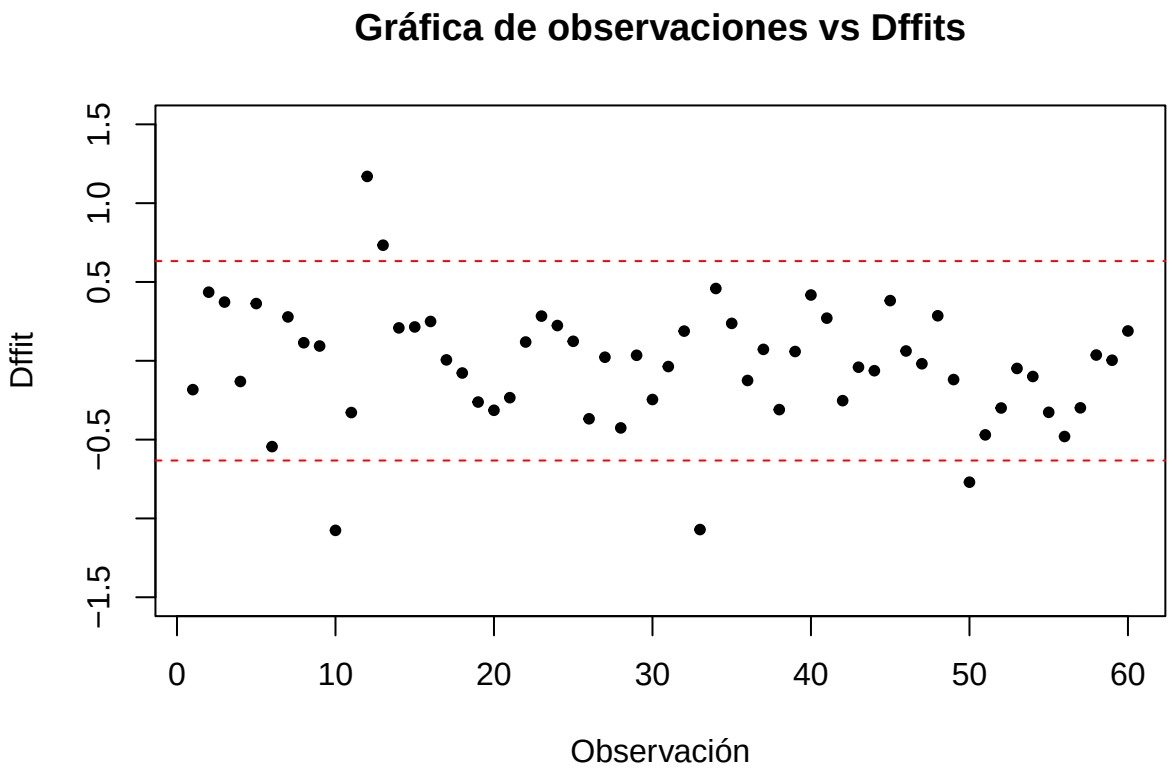

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 10  -1.9267  0.1830    0.2283 -1.0758
## 12   2.3624  0.2083    0.1830  1.1697
## 13   1.9211  0.0851    0.1215  0.7334
## 33  -1.5005  0.1866    0.3321 -1.0707
## 50  -2.3003  0.0908    0.0934 -0.7700
```

Como se puede ver,las observaciones ... 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