# Trabajo 1


Estudiantes

**Matheo Muñoz Betancur**

Equipo

Docente

**Carlos Mario Lopera**


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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

Donde ... acá dicen el nombre de las variables

- Y: Hospitales
- $X_1$:

## 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 |
|---|---|
| $\beta_0$ | -0.5798 |
| $\beta_1$ | 0.1456 |
| $\beta_2$ | 0.0032 |
| $\beta_3$ | 0.0441 |
| $\beta_4$ | 0.0250 |
| $\beta_5$ | 0.0021 |

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.5798 + 0.1456 X_{1i} + 0.0032 X_{2i} + 0.0441 X_{3i} + 0.025 X_{4i} + 0.0021 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

## 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:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,53} \tag{1}$$

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 | 83.2914 | 5 | 16.658283 | 16.8377 | 6.19931e-10 |
| Error | 52.4354 | 53 | 0.989346 | | |

De la tabla Anova, se observa un valor P aproximadamente igual a 1, 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 |
|---|---|---|---|---|
| $\beta_0$ | -0.5798 | 1.6352 | -0.3546 | 0.7243 |
| $\beta_1$ | 0.1456 | 0.0765 | 1.9021 | 0.0626 |
| $\beta_2$ | 0.0032 | 0.0305 | 0.1042 | 0.9174 |
| $\beta_3$ | 0.0441 | 0.0146 | 3.0226 | 0.0039 |
| $\beta_4$ | 0.0250 | 0.0079 | 3.1588 | 0.0026 |
| $\beta_5$ | 0.0021 | 0.0007 | 2.9368 | 0.0049 |

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

## 1.4. Interpretación de los parámetros

Interpreten sólo los parámetros significativos, respecto a $\beta_0$ ya saben que se debe cumplir que el 0 esté en el intervalo

**$\hat{\beta}_i$:**

**$\hat{\beta}_j$:**

**$\hat{\beta}_2$:**

### 1.5. Coeficiente de determinación múltiple $R^2$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.hey$, lo que significa que aproximadamente el $hey\,\%$ de la variabilidad total 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 y modelo reducido

Las covariable con el P-valor más alto en el modelo fueron $X_1, X_2, X_4$, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

Cuadro 4: Resumen tabla de todas las regresiones

| | $SSE$ | Covariables en el modelo |
|---|---|---|
| Modelo completo | 52.435 | X1 X2 X3 X4 X5 |
| Modelo reducido | 74.387 | X3 X5 |

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

### 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,53} \\ &= \frac{numerador}{denominador} \\ &= final \end{aligned} \tag{2}$$

Ahora, comparando el $F_0$ con $f_{0.95,3,53} = 2.7791$, se puede ver que $F_0 ? f_{0.95,1,45}$ y por tanto qué se rechaza ...

Es posible o no descartar las variables del subconjunto?

# 3. Pregunta 3

## 3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si ¿...? por consiguiente se plantea la siguiente prueba de hipótesis:

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

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_2 X^*_{2i} + \beta_3 X^*_{3i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

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

**Nota:** NO usen esta misma prueba

## 3.2. Estadístico de prueba

El estadístico de prueba $F_0$ está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{100,53} \tag{3}$$

Aquí deben igualar después esa ecuación a sí misma con los valores conocidos reemplazados, es decir, el SSE(MF) y el MSE(MF).

# 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:

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

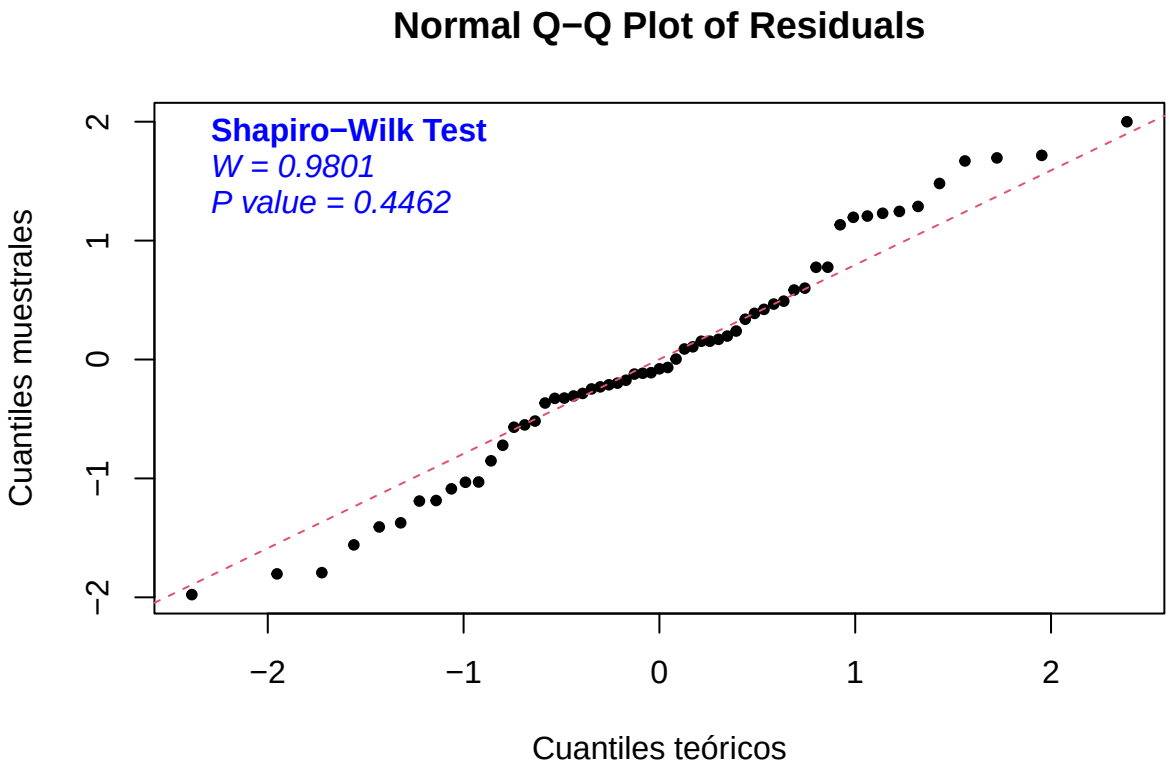


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.4462 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 $\mu$ y varianza $\sigma^2$, 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

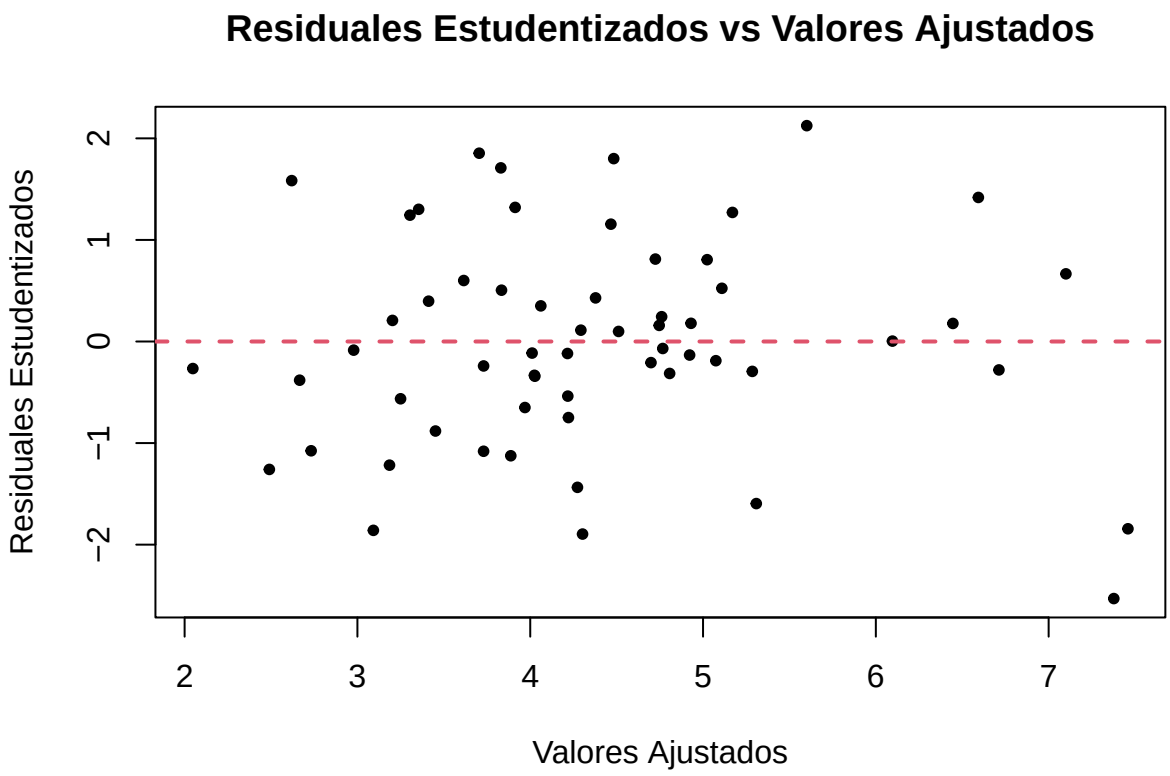


Figura 2: Gráfico residuales estudentizados vs valores ajustados

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

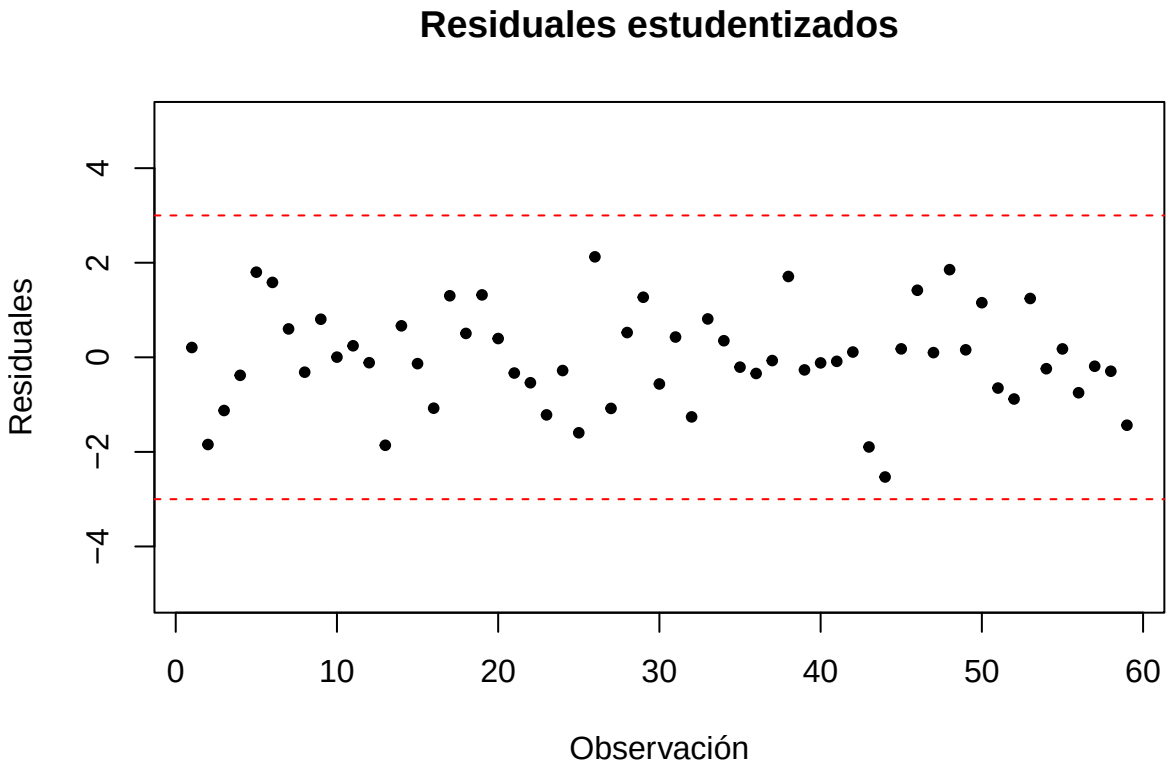


Figura 3: Identificación de datos atípicos

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

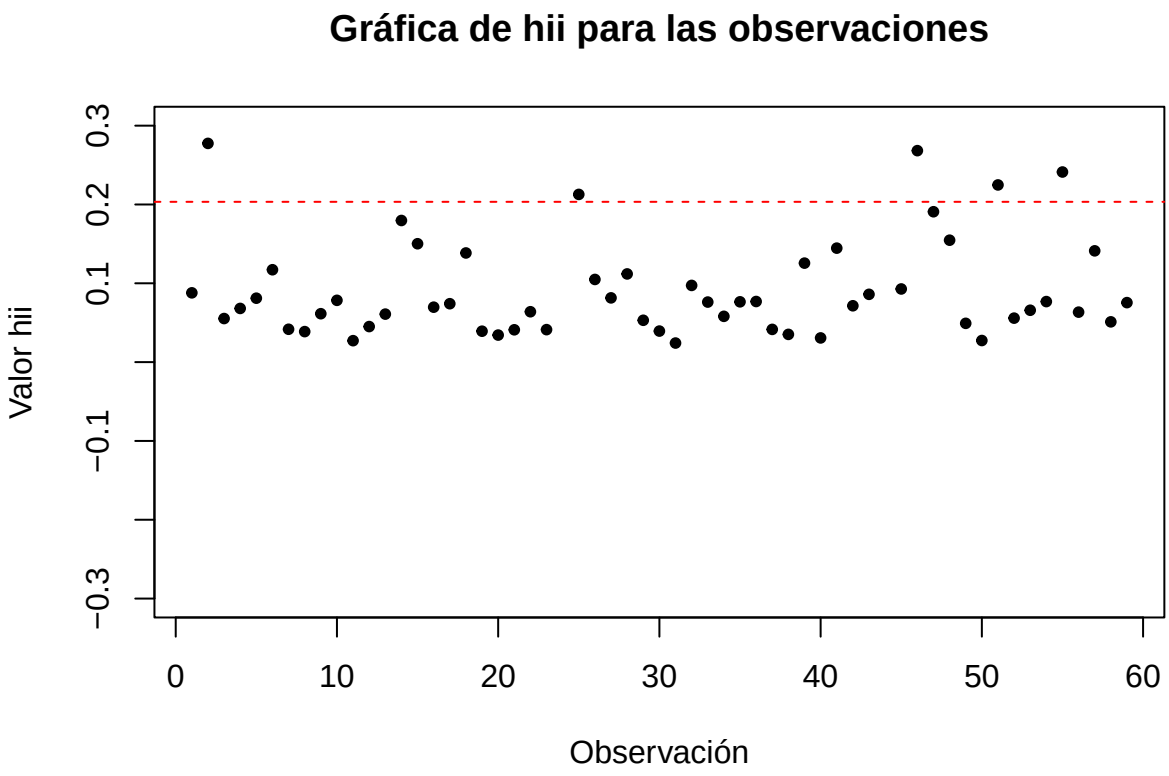


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 2   -1.8439  0.2176    0.2775 -1.1699
## 24  -0.2798  0.0094    0.4182 -0.2351
## 25  -1.5957  0.1147    0.2128 -0.8422
## 44  -2.5309  0.6631    0.3832 -2.1072
## 46   1.4181  0.1229    0.2683  0.8671
## 51  -0.6502  0.0204    0.2248 -0.3482
## 55   0.1777  0.0017    0.2412  0.0993
```

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

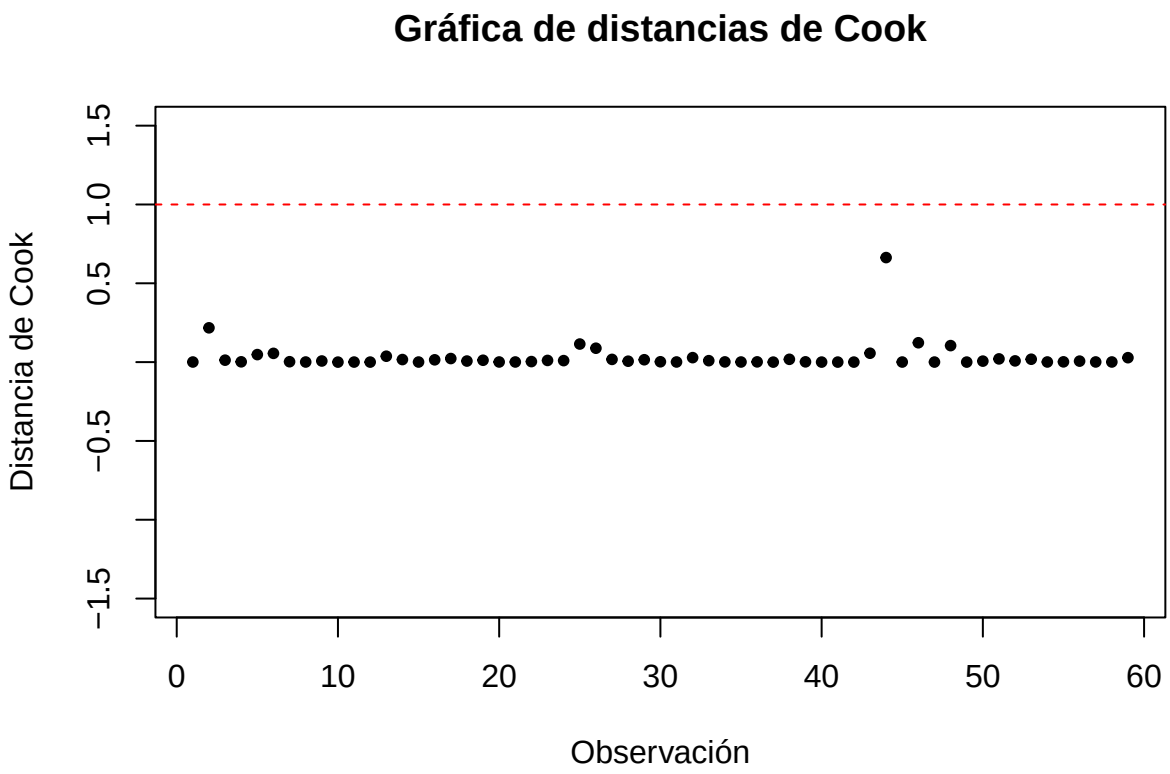


Figura 5: Criterio distancias de Cook para puntos influenciales

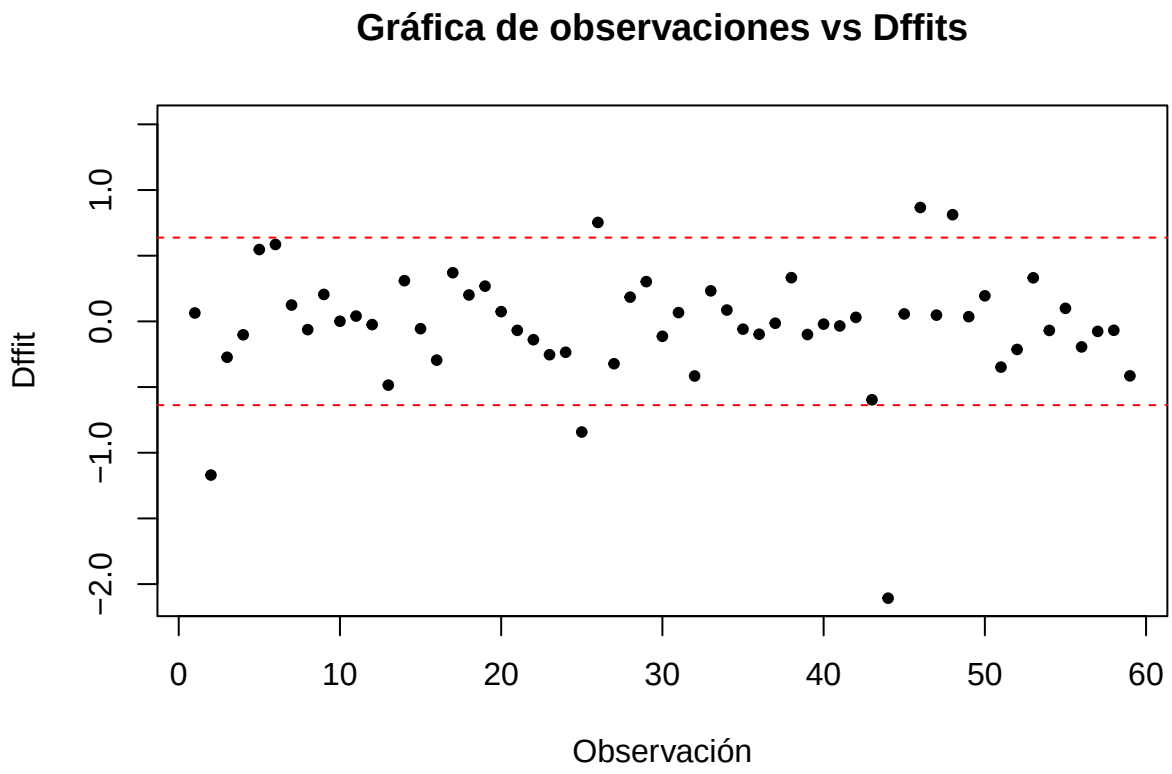


Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 2   -1.8439  0.2176    0.2775 -1.1699
## 25  -1.5957  0.1147    0.2128 -0.8422
## 26   2.1249  0.0882    0.1049  0.7533
## 44  -2.5309  0.6631    0.3832 -2.1072
## 46   1.4181  0.1229    0.2683  0.8671
## 48   1.8538  0.1048    0.1547  0.8123
```

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

Acá como mínimo deben decir si el modelo es válido o no, argumentar por qué y cómo esto se ve afectado por estos puntos extremos.