# Trabajo 1

4,7


Estudiantes

**Diego Fernando Gallego Romero**
**Juan José Munera Pulgarín**
**Juan Esteban Rivera Henao**
**Juan José Zapata Moreno**

Equipo 47

Docente

**Carlos Mario Lopera**


Asignatura

**Estadística II**


Sede Medellín
5 de octubre de 2023

# Índice

# Índice de figuras



# Índice de cuadros

# 1. Pregunta 1

1 pt

Teniendo en cuenta la base de datos brindada, en la cual hay 5 variables regresoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

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 |
|---|---|
| $\beta_0$ | -0.1308 |
| $\beta_1$ | 0.2361 |
| $\beta_2$ | 0.0110 |
| $\beta_3$ | 0.0618 |
| $\beta_4$ | 0.0054 |
| $\beta_5$ | 0.0012 |

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.1308 + 0.2361 X_{1i} + 0.011 X_{2i} + 0.0618 X_{3i} + 0.0054 X_{4i} + 0.0012 X_{5i}$$

## 1.2. Significancia de la regresión

Para analizar la significancia de la regresión, se plantea el siguiente juego de hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,58} \tag{1}$$

Ahora, se presenta la tabla Anova:

5 pt

Cuadro 2: Tabla ANOVA para el modelo

| | Sumas de cuadrados | g.l. | Cuadrado medio | $F_0$ | P-valor |
|---|---|---|---|---|---|
| Regresión | 56.7595 | 5 | 11.351908 | 12.2659 | 3.96152e-08 |
| Error | 53.6780 | 58 | 0.925482 | | |

De la tabla Anova, se observa un valor P 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 $\beta_j \neq 0$, 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 |
|---|---|---|---|---|
| $\beta_0$ | -0.1308 | 1.6324 | -0.0802 | 0.9364 |
| $\beta_1$ | 0.2361 | 0.0782 | 3.0214 | 0.0037 |
| $\beta_2$ | 0.0110 | 0.0305 | 0.3601 | 0.7201 |
| $\beta_3$ | 0.0618 | 0.0158 | 3.9209 | 0.0002 |
| $\beta_4$ | 0.0054 | 0.0074 | 0.7285 | 0.4692 |
| $\beta_5$ | 0.0012 | 0.0007 | 1.7434 | 0.0866 |

6 pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros $\beta_1$ y $\beta_3$ son significativos, pues sus P-valores son menores a $\alpha$.

3 pt

### 1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** por cada día que incrementa la duración de la estadía de los pacientes en el hospital, la probabilidad promedio estimada de adquirir infección allí aumenta significativamente en 0.2361 cuando los valores en las demás variables predictoras permanecen constantes.

$\hat{\beta}_3$**:** por cada unidad de incremento del número de camas promedio en el hospital durante el periodo de estudio, la probabilidad promedio estimada de adquirir infección allí aumenta significativamente en 0.0618 cuando los valores en las demás variables predictoras permanecen constantes.

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

2 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.513951$, lo que significa que aproximadamente el 51.3951 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en este informe.

¿Cómo se calcula?

## 2. Pregunta 2

5 pt

### 2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables 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 | 53.678 | X1 X2 X3 X4 X5 |
| Modelo reducido | 90.914 | X2 X4 |

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

### 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \ \cdots, \ \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{12.4120}{0.9255} \\ &= 13.4111 \end{aligned} \tag{2}$$

Comparando este valor con el cuantil crítico $f_{0.95,3,58} = 2.7636$ (que corresponde al nivel de significancia de $\alpha = 0.05$ con 3 y 58 grados de libertad), se observa que $F_0$ es significativamente mayor que $f_{0.95,3,58}$. Dado que el valor calculado del estadístico de prueba supera el valor crítico, se rechaza la hipótesis nula. Esto sugiere que al menos uno de los parámetros no es igual a cero y tiene un efecto significativo en el modelo analizado, por lo tanto, no pueden ser descartados.

## 3. Pregunta 3

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

Se hace la pregunta si $\beta_1 = \beta_2$ y $2\beta_3 = \beta_5$; por consiguiente, se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2; \ \beta_5 = 2\beta_3 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i}^* + \beta_3 X_{3i}^* + \beta_4 X_{4i} + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 64$$

Donde $X_{1i}^* = X_{1i} + X_{2i}$ y $X_{3i}^* = X_{3i} + 2X_{5i}$

### 3.2. Estadístico de prueba

El estadístico de prueba $F_0$ está dado por:

$$F_0 = \frac{(SSE(MR) - 53.678)/2}{0.9255} \overset{H_0}{\sim} f_{2,58} \tag{3}$$

2 pt

# 4. Pregunta 4

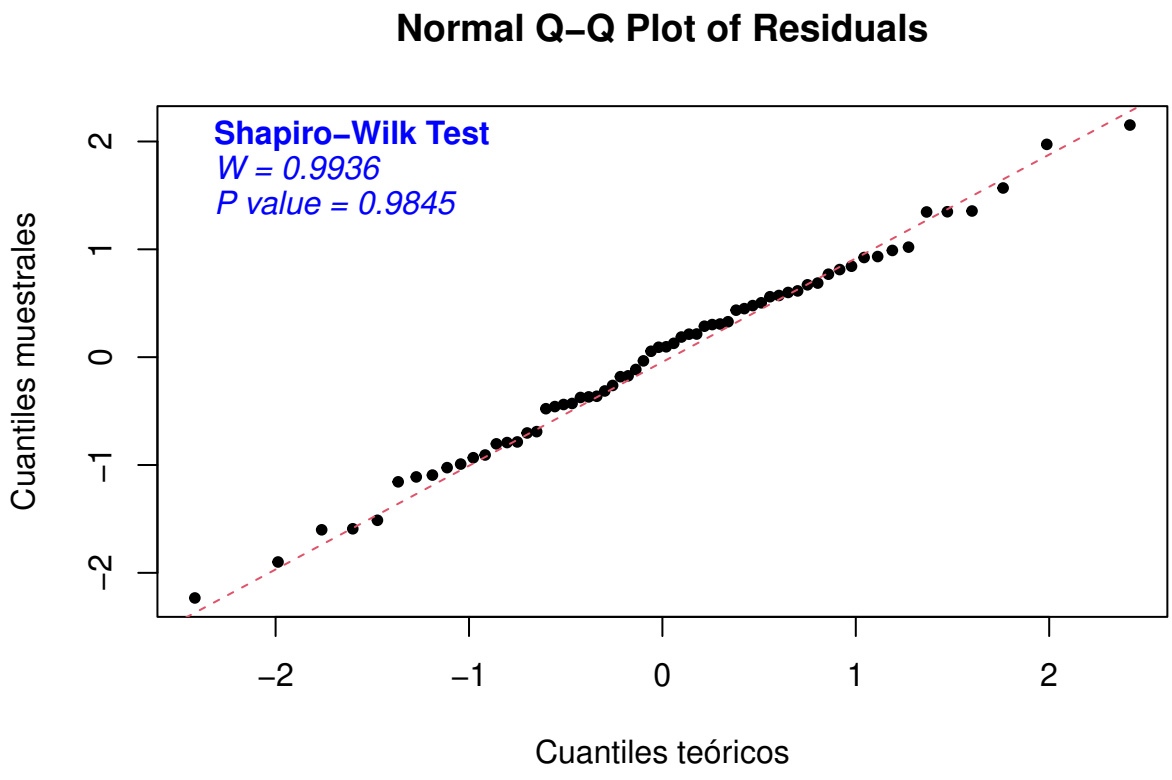
18 pt

## 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:

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




4 pt

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Utilizando la prueba de Shapiro-Wilk para evaluar la normalidad, se obtuvo un P-valor de 0.9845, considerablemente mayor que el nivel de significancia $\alpha = 0.05$, inclusive muy cercano a 1, sugiriendo que los datos podrían seguir una distribución normal. Además el análisis de la gráfica de comparación de cuantiles devela que los puntos están muy cercanos a la línea y no siguen un patrón claro, reafirmando la evidencia de normalidad en los datos.

### 4.1.2. Varianza constante

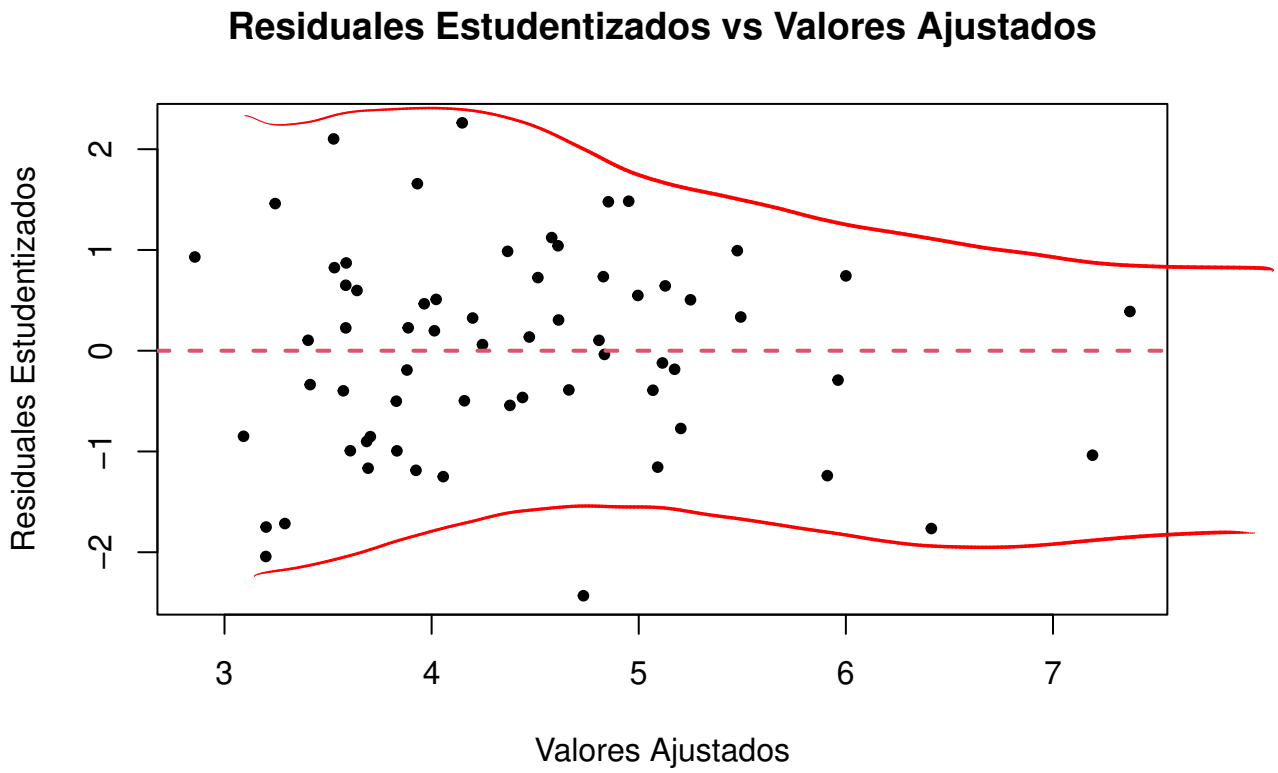



Figura 2: Gráfico residuales estudentizados vs valores ajustados

En este gráfico los datos se distribuyen de manera uniforme sin mostrar patrones que sugieran un aumento o disminución en la varianza. Esta distribución uniforme respalda la suposición de varianza constante. Además, la distribución de los residuales se centra alrededor de una media de 0.

## 4.2. Verificación de las observaciones

### 4.2.1. Datos atípicos

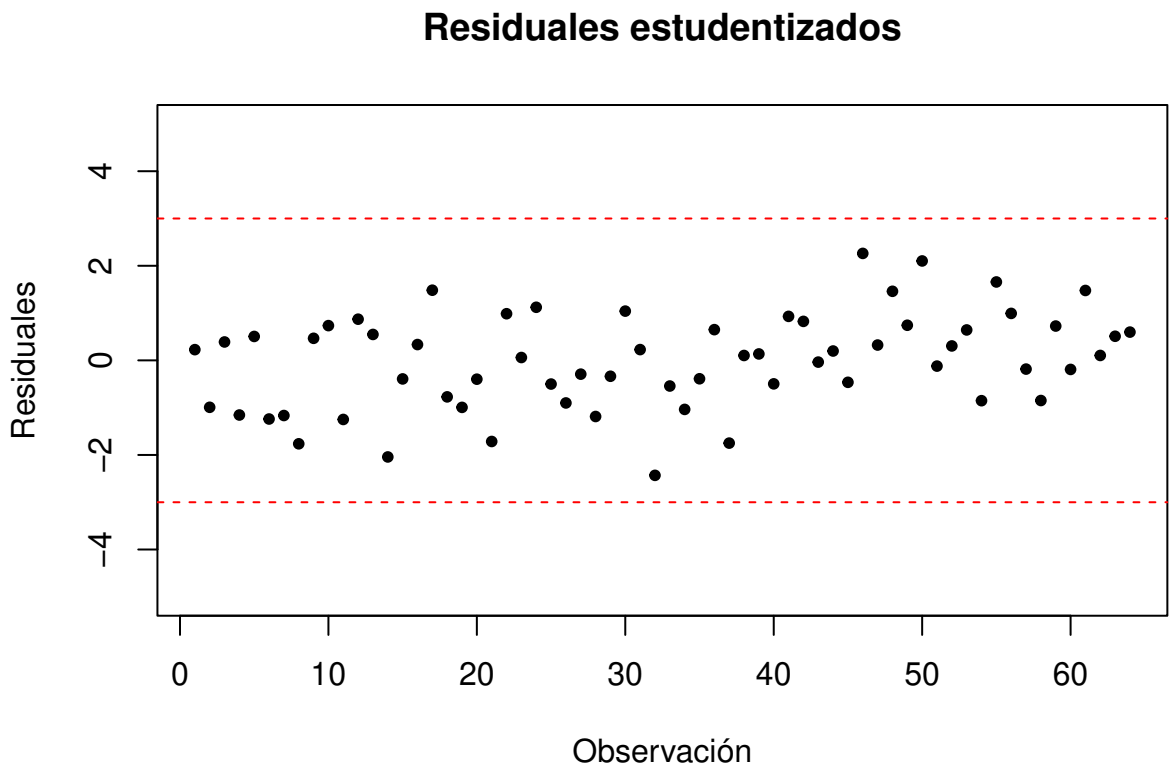



Figura 3: Identificación de datos atípicos

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

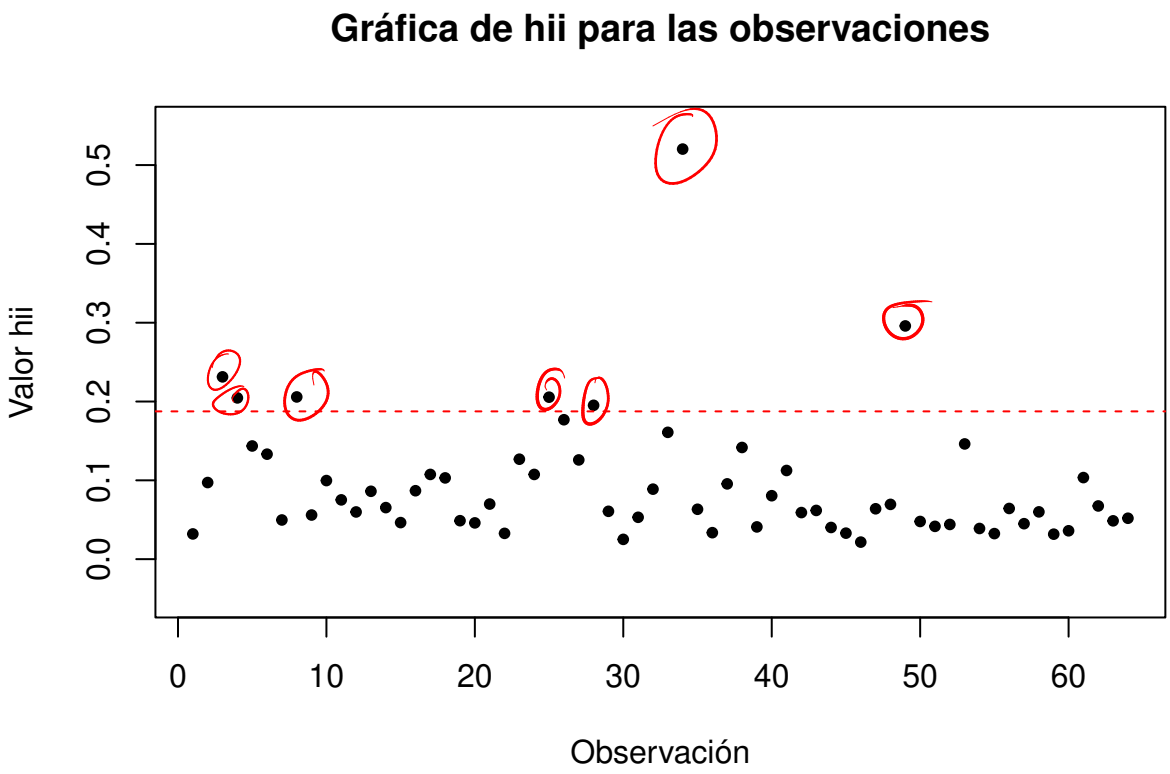



Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 3    0.3893  0.0076    0.2313  0.2120
## 4   -1.1555  0.0572    0.2046 -0.5878
## 8   -1.7646  0.1345    0.2058 -0.9154
## 25  -0.5009  0.0108    0.2055 -0.2531
## 28  -1.1871  0.0570    0.1952 -0.5868
## 34  -1.0369  0.1944    0.5203 -1.0807
## 49   0.7429  0.0387    0.2959  0.4797
```

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.

### 4.2.3. Puntos influenciales

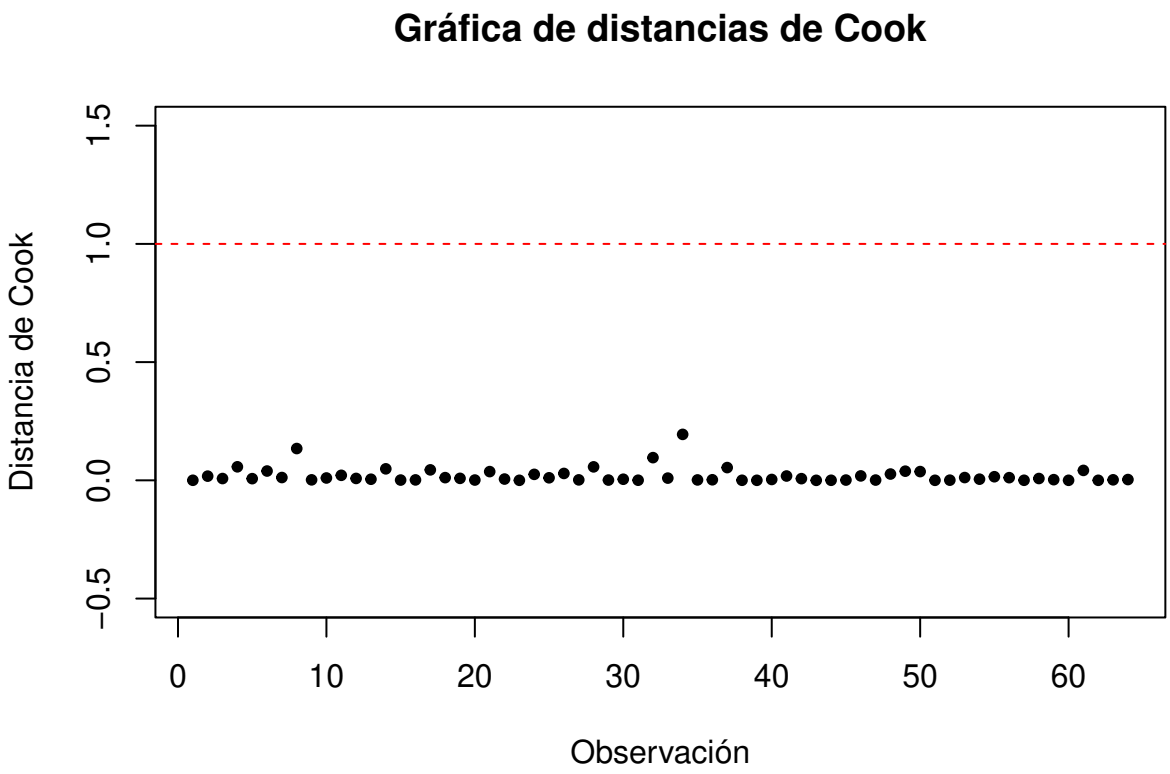



Figura 5: Criterio distancias de Cook para puntos influenciales

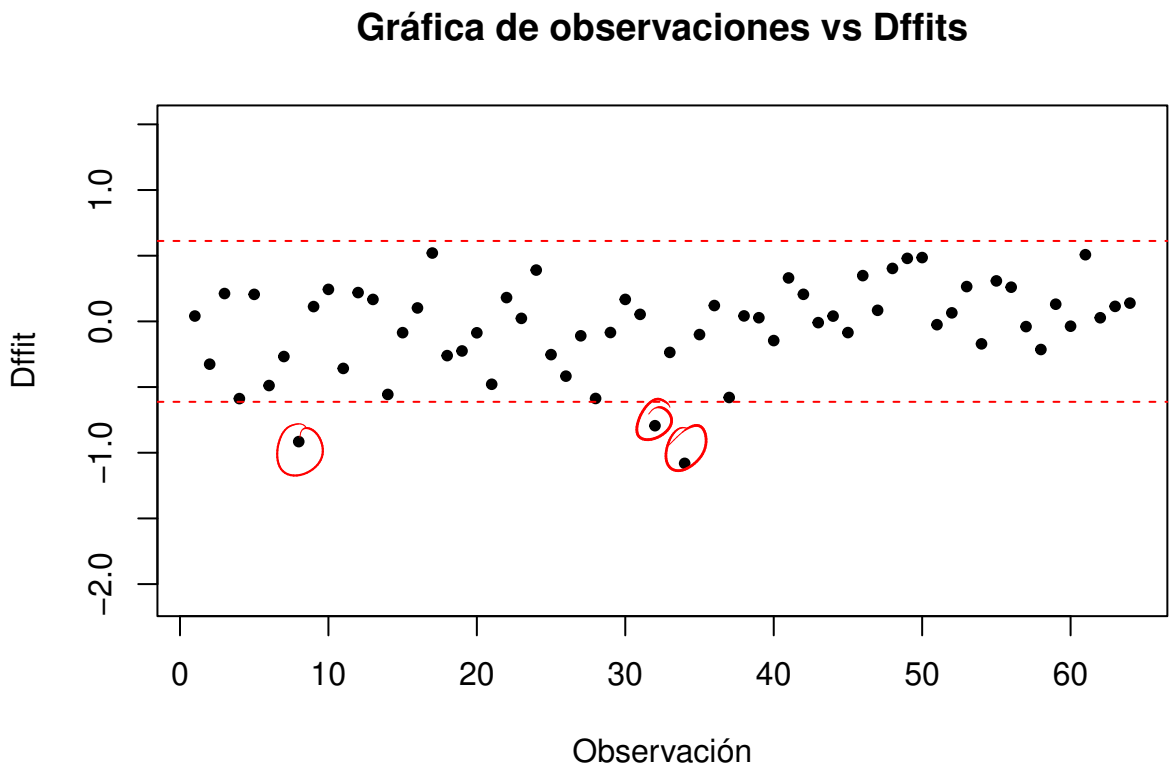



Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 8   -1.7646  0.1345    0.2058 -0.9154
## 32  -2.4316  0.0960    0.0888 -0.7941
## 34  -1.0369  0.1944    0.5203 -1.0807
```

Como se puede ver,las observaciones 8, 32 y 34 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. Un punto influencial puede inclinar o desplazar la línea de regresión, lo que lleva a una interpretación errónea de la relación entre las variables: hay que tener especial precaución con la observación 34, ya que su valor Dffits es -1.0807, muy superior a las demás observaciones realizadas. También es de anotar que su hii es el más alto por lejos, siendo de 0.5203. Ambos indicadores señalan un alto grado de influencia.

## 4.3. Conclusión

Tras un análisis detallado de los supuestos y las observaciones del modelo, concluimos que, aunque nuestro modelo muestra una robustez general, la presencia de puntos de balanceo

y, en particular, puntos influyentes, plantea serias preocupaciones. Estos puntos extremos tienen el potencial de sesgar las estimaciones y comprometer la validez del modelo. En particular, las observaciones influyentes, como las identificadas en los análisis de distancias de Cook y Dffits, pueden distorsionar significativamente tanto las estimaciones de los parámetros como las predicciones resultantes.

Dada la influencia desproporcionada de estos puntos (en especial de la observación 34), es esencial reconsiderar la validez del modelo en su forma actual. Aunque puede ser tentador excluir estos puntos y reajustar el modelo, es crucial entender por qué estos puntos son extremos en primer lugar. Puede ser más informativo y útil identificar y comprender la naturaleza de estos puntos antes de tomar decisiones sobre su exclusión.

Por lo tanto, antes de considerar este modelo como definitivo o basar decisiones importantes en él, se recomienda un examen más profundo de estos puntos extremos y, posiblemente, la realización de análisis adicionales que los excluyan o consideren técnicas de modelado más robustas.

Válido o no? No dieron una conclusión concisa al respecto.