# Trabajo 1

4,4

Estudiantes

-1pt

Equipo 50

Docente

Asignatura

**Estadística II**

Sede Medellín

5 de octubre de 2023

# Índice


**1. Pregunta 1** **3**
- 1.1. Modelo de regresión . . . 3
- 1.2. Significancia de la regresión . . . 3
- 1.3. Significancia de los parámetros . . . 4
- 1.4. Interpretación de los parámetros . . . 5
- 1.5. Coeficiente de determinación múltiple $R^2$ . . . 5

**2. Pregunta 2** **5**
- 2.1. Planteamiento pruebas de hipótesis y modelo reducido . . . 5
- 2.2. Estadístico de prueba y conclusión . . . 6

**3. Pregunta 3** **6**
- 3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . 6
- 3.2. Estadístico de prueba . . . 7

**4. Pregunta 4** **7**
- 4.1. Supuestos del modelo . . . 7
  - 4.1.1. Normalidad de los residuales . . . 7
  - 4.1.2. Varianza constante . . . 8
- 4.2. Verificación de las observaciones . . . 10
  - 4.2.1. Datos atípicos . . . 10
  - 4.2.2. Puntos de balanceo . . . 10
  - 4.2.3. Puntos influenciales . . . 12
- 4.3. Conclusión . . . 13

# Índice de figuras


1. Gráfico cuantil-cuantil y normalidad de residuales . . . . . 8
2. Gráfico residuales estudentizados vs valores ajustados . . . . . 9
3. Identificación de datos atípicos . . . . . 10
4. Identificación de puntos de balanceo . . . . . 11
5. Criterio distancias de Cook para puntos influenciales . . . . . 12
6. Criterio Dffits para puntos influenciales . . . . . 13


# Índice de cuadros


1. Tabla de valores coeficientes del modelo . . . . . 3
2. Tabla ANOVA para el modelo . . . . . 4
3. Resumen de los coeficientes . . . . . 4
4. Resumen tabla de todas las regresiones . . . . . 6
5. Tabla de puntos de balanceo . . . . . 11
6. Tabla de puntos influenciables . . . . . 12

# 1. Pregunta 1

Teniendo en cuenta la base de datos brindada, donde hay 5 variables regresoras dadas por el siguiente modelo de regresión lineal multiple:

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

Donde:

- Y: Riesgo de infección
- $X_1$:Duración de la estadía [días]
- $X_2$:Rutina de cultivos [por cada 100]
- $X_3$:Numero de camas
- $X_4$:Censo promedio diario
- $X_5$:Numero 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 |
|---|---|
| $\beta_0$ | -1.2506 |
| $\beta_1$ | 0.1842 |
| $\beta_2$ | 0.0250 |
| $\beta_3$ | 0.0646 |
| $\beta_4$ | 0.0124 |
| $\beta_5$ | 0.0017 |

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.2506 + 0.1842 X_{1i} + 0.025 X_{2i} + 0.0646 X_{3i} + 0.0124 X_{4i} + 0.0017 X_{5i};\ 1 \leqslant i \leqslant 64$$

## 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:

Cuadro 2: Tabla ANOVA para el modelo

| | Sumas de cuadrados | g.l. | Cuadrado medio | $F_0$ | P-valor |
|---|---|---|---|---|---|
| Regresión | 63.4148 | 5 | 12.68296 | 12.4881 | 3.0654e-08 |
| Error | 58.9052 | 58 | 1.01561 | | |

De la tabla Anova, se observa que bajo un nivel de significacia del 5 %, valor p =3.0654e-08 $< \alpha = 0.05$, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, lo que signifca que entonces al menos un parametro del modelo de regresión mútiple es diferente de 0, es decir, la regresión es globalmente significativa.

## 1.3. Significancia de los parámetros

Primero observemos el juego de hipotesis para la prueba individual de la significancia de los parametros.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \text{ con } 0 \leqslant j \leqslant 5 \end{cases}$$

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 |
|---|---|---|---|---|
| $\beta_0$ | -1.2506 | 1.6154 | -0.7742 | 0.4420 |
| $\beta_1$ | 0.1842 | 0.1105 | 1.6679 | 0.1007 |
| $\beta_2$ | 0.0250 | 0.0298 | 0.8407 | 0.4040 |
| $\beta_3$ | 0.0646 | 0.0168 | 3.8418 | 0.0003 |
| $\beta_4$ | 0.0124 | 0.0077 | 1.6059 | 0.1137 |
| $\beta_5$ | 0.0017 | 0.0008 | 2.2145 | 0.0307 |

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros $\beta_1$ $\beta_3$y $\beta_5$ son significativos, pues sus P-valores son menores a $\alpha$.

### 1.4. Interpretación de los parámetros

Lo primero es identificar aquellos parámetros susceptibles de interpretación, esto es, solo se podrán interpretar los que vimos significativos, en esto caso son:

2 pt

$\hat{\beta}_1$:

Indica que por cada unidad de aumento en la duración de la estadía, el promedio de riesgo de infección en el hospital aumenta en 0.1842, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_3$:

Indica que por cada unidad de aumento en el número de camas, el promedio de riesgo de infección en el hospital aumenta en 0.0646, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_5$:

Indica que por cada unidad de aumento en el número de enfermeras, el promedio de riesgo de infección en el hospital aumenta en 0.0017, cuando las demás variables predictoras se mantienen fijas.

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

3 pt

Extrayendo valores de la tabla ANOVA, tenemos que:

$$R^2 = \frac{SSR}{SST} = \frac{63.4148}{(63.4148 + 58.9052)} = 0.5184336$$

- Es decir, el 51.843 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital es explicado por el modelo propuesto.
- Simultaneamente, el 48.15 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital es explicado por el error del modelo.

## 2. Pregunta 2

5 pt

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

Las covariable con el P-valor más bajos en el modelo fueron $X_1$, $X_3$ y $X_5$ , por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

Cuadro 4: Resumen tabla de todas las regresiones

| | $SSE$ | Covariables en el modelo |
|---|---|---|
| Modelo completo | 58.905 | X1 X2 X3 X4 X5 |
| Modelo reducido | 94.581 | 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 se muestra a continuación:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{94.581 - 58.905/3}{1.01561} \\ &= 11.70922 \end{aligned} \tag{2}$$

Ahora, comparando el $F_0$ con $f_{0.95,3,58} = 2.7636$, se puede ver que $F_0 > f_{0.95,3,58}$ entonces se rechaza la hipotesis nula, y concluimos que al menos un $\beta_j$ es distinto de cero con j=1,3,5.

Por lo que no podemos descartar el subconjunto dado que al menos uno de los parametros aportan significativamente al modelo.

## 3. Pregunta 3

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

Se hace la pregunta si $\beta_3 = 6\beta_1$, $\beta_2 = 2\beta_5$ por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \text{H}_0 : \beta_1 = 6\beta_3;\ \beta_2 = 2\beta_5 \\ \text{H}_1 : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

Lo que es equivalente a lo siguiente:

$$\begin{cases} \text{H}_0 : \beta_1 - 6\beta_3 = 0;\ \beta_2 - 2\beta_5 = 0 \\ \text{H}_1 : \text{Al menos una de las desigualdades no se cumple} \end{cases}$$

Reescribiendo matricialmente:

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

Con **L** dada por:

$$L = \begin{bmatrix} 0 & 1 & 0 & -6 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -2 \end{bmatrix}$$

El modelo completo esta dado por:

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

El modelo reducido está dado por:

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

Donde:

$X^*_{3i} = 6X_{1i} + X_{3i}$ y $X^*_{5i} = 2X_{2i} + X_{5i}$

## 3.2. Estadístico de prueba

El estadístico de prueba $F_0$ está dado por:

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

Al reemplazar con los valores conocidos, se encuentra lo siguiente:

$$F_0 = \frac{(SSE(MR) - 58.905)/2}{1.01561} \overset{H_0}{\sim} f_{2,58} \tag{4}$$

# 4. Pregunta 4

## 4.1. Supuestos del modelo

### 4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se planteará el siguiente test de Shapiro-Wilk que se utiliza para determinar si un conjunto de datos puede distribuirse mediante la distribucción normal , acompañada de un gráfico cuantil-cuantil:

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

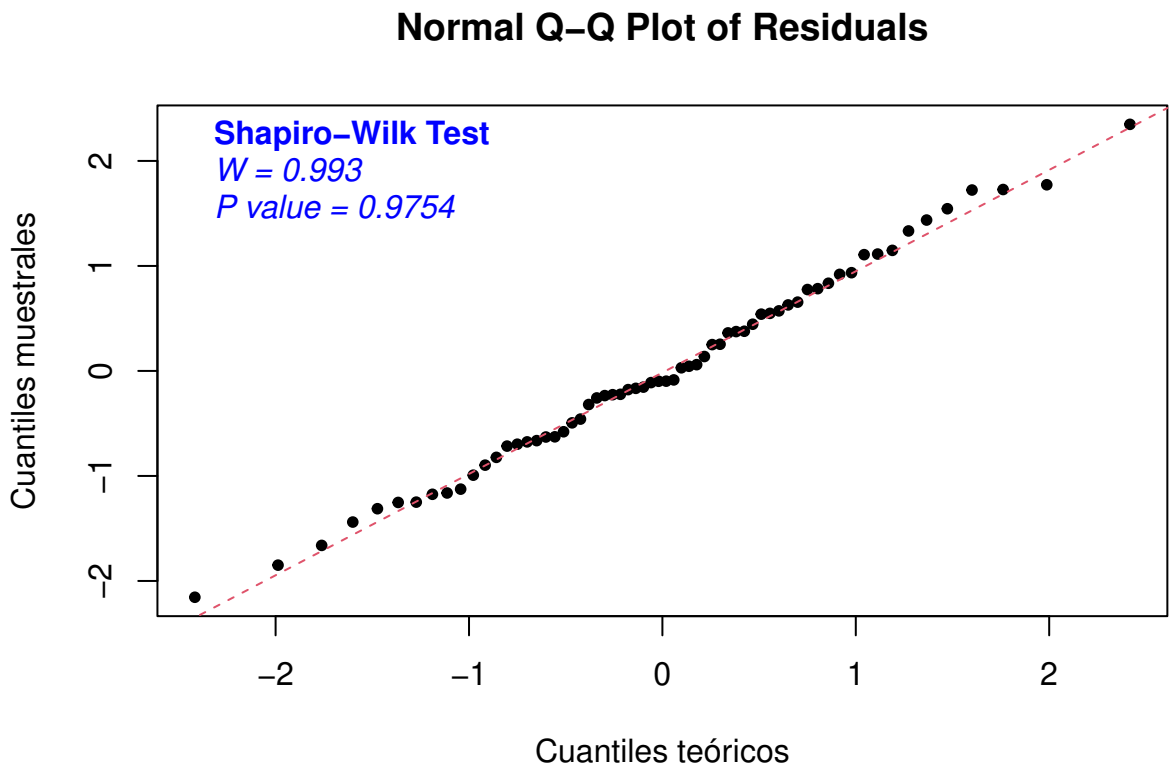


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.9754 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 $\sigma^2$, lo cual es rectificadopor el analisis del grafico de comparacion de cuantiles donde se observa claramente, que un buen ajuste de los datos al rededor de la recta, de lo cual podemos deducir que los residuales se distribuyen normal

### 4.1.2. Varianza constante

$$H_0 : \mathbf{V}[\varepsilon_i] = \sigma^2 \qquad \textbf{vs} \qquad H_a : \mathbf{V}[\varepsilon_i] \neq \sigma^2$$

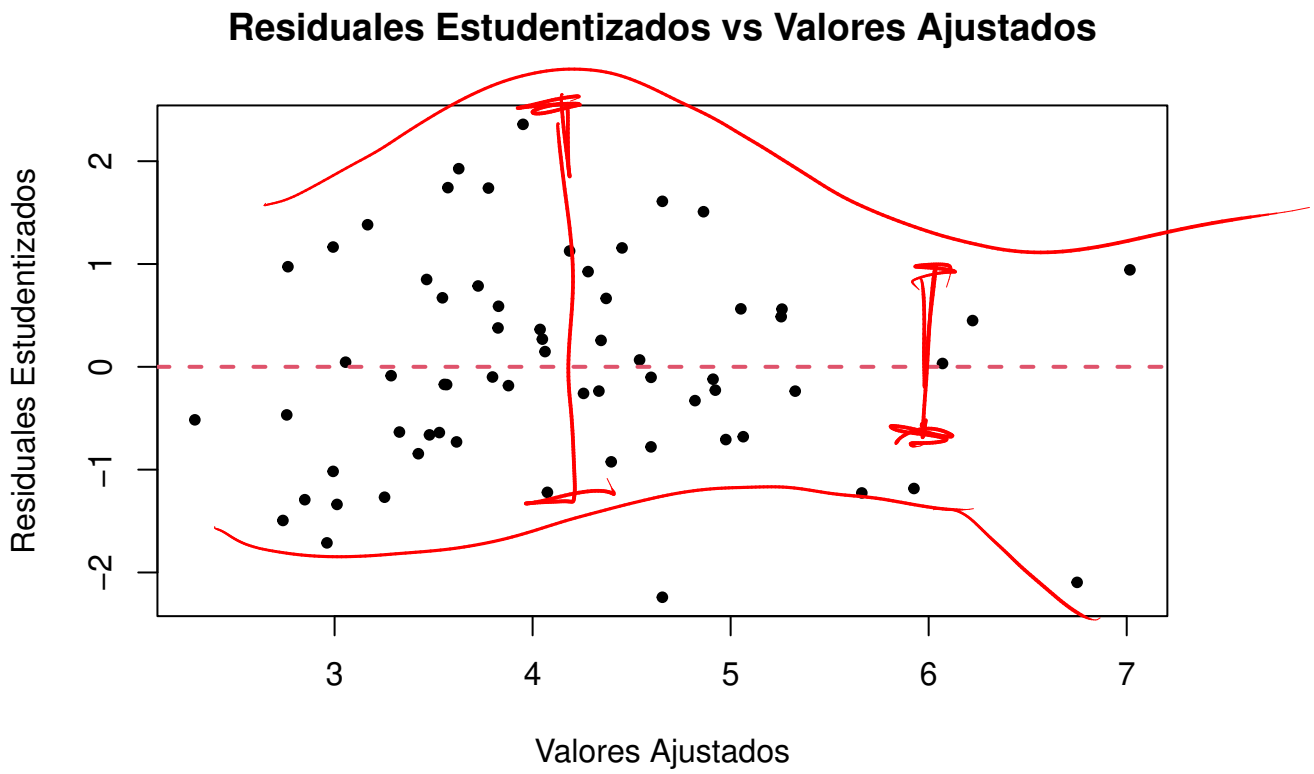


Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones muy marcados 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

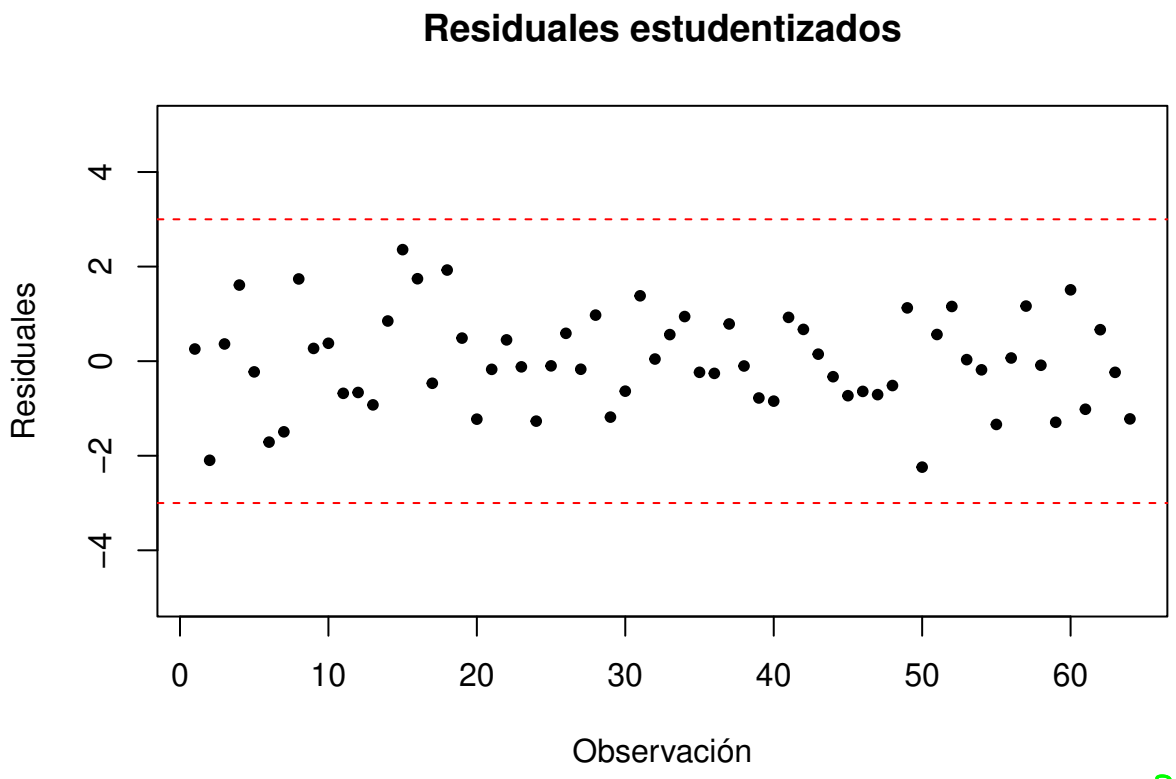


Figura 3: Identificación de datos atípicos

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

Se puede apreciar en la siguiente tabla que hay 7 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n} = 0.1875$, los cuales son los presentados en la tabla.

Cuadro 5: Tabla de puntos de balanceo

| | res.stud | Cooks.D | hii.value | Dffits |
|---|---|---|---|---|
| 2 | -2.0963 | 0.2231 | 0.2335 | -1.1931 |
| 12 | -0.6619 | 0.0238 | 0.2462 | -0.3764 |
| 22 | 0.4499 | 0.0148 | 0.3047 | 0.2958 |
| 27 | -0.1710 | 0.0012 | 0.2006 | -0.0849 |
| 34 | 0.9432 | 0.0699 | 0.3204 | 0.6469 |
| 39 | -0.7786 | 0.0267 | 0.2092 | -0.3991 |
| 56 | 0.0671 | 0.0002 | 0.2176 | 0.0351 |

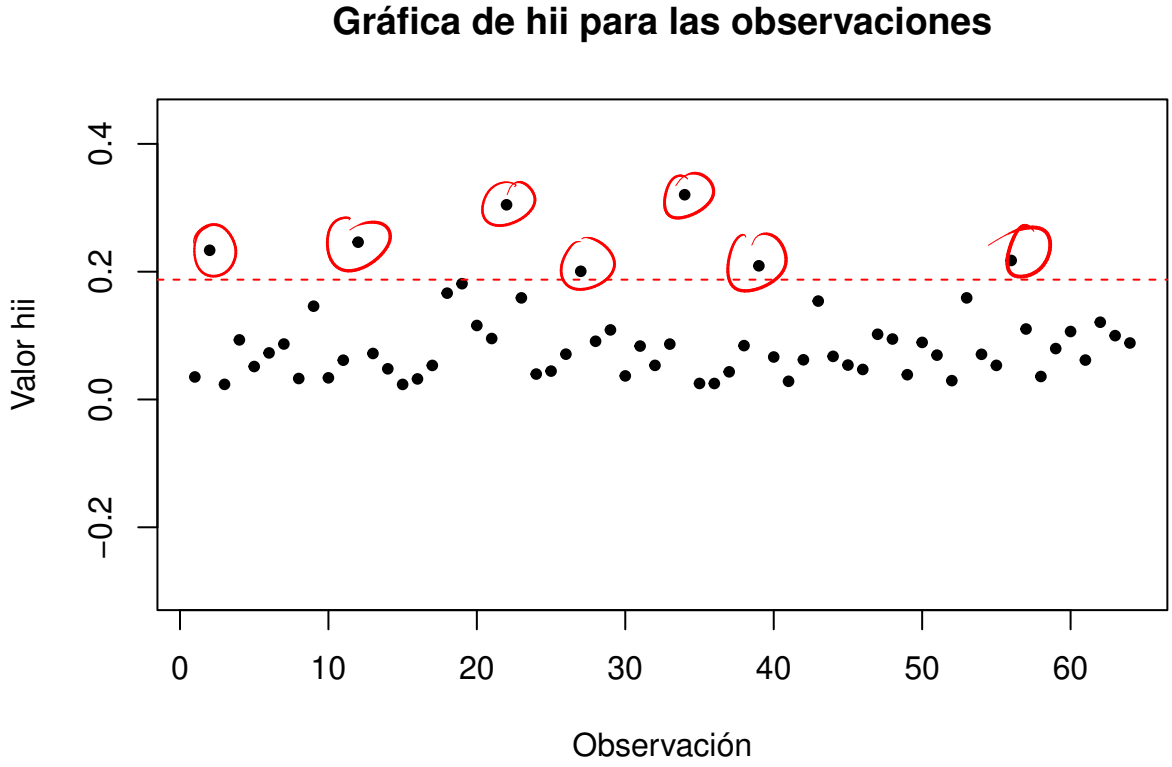


Figura 4: Identificación de puntos de balanceo

Al observar la gráfica de observaciones vs valores $h_{ii}$, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.1875$, se puede apreciar que existen 7 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p=6}{n=64}$, los cuales son los presentados en la tabla.

3pt

Cuadro 6: Tabla de puntos influenciables

| | res.stud | Cooks.D | hii.value | Dffits |
|---|---|---|---|---|
| 2 | -2.0963 | 0.2231 | 0.2335 | -1.1931 |
| 18 | 1.9270 | 0.1235 | 0.1664 | 0.8821 |
| 34 | 0.9432 | 0.0699 | 0.3204 | 0.6469 |
| 50 | -2.2409 | 0.0820 | 0.0893 | -0.7278 |

### 4.2.3. Puntos influenciales

Para los puntos influenciales tenemos dos criterios. El primero es la distacia de Cook que dice que la observación $i$ será influencial si su $D_i > 1$

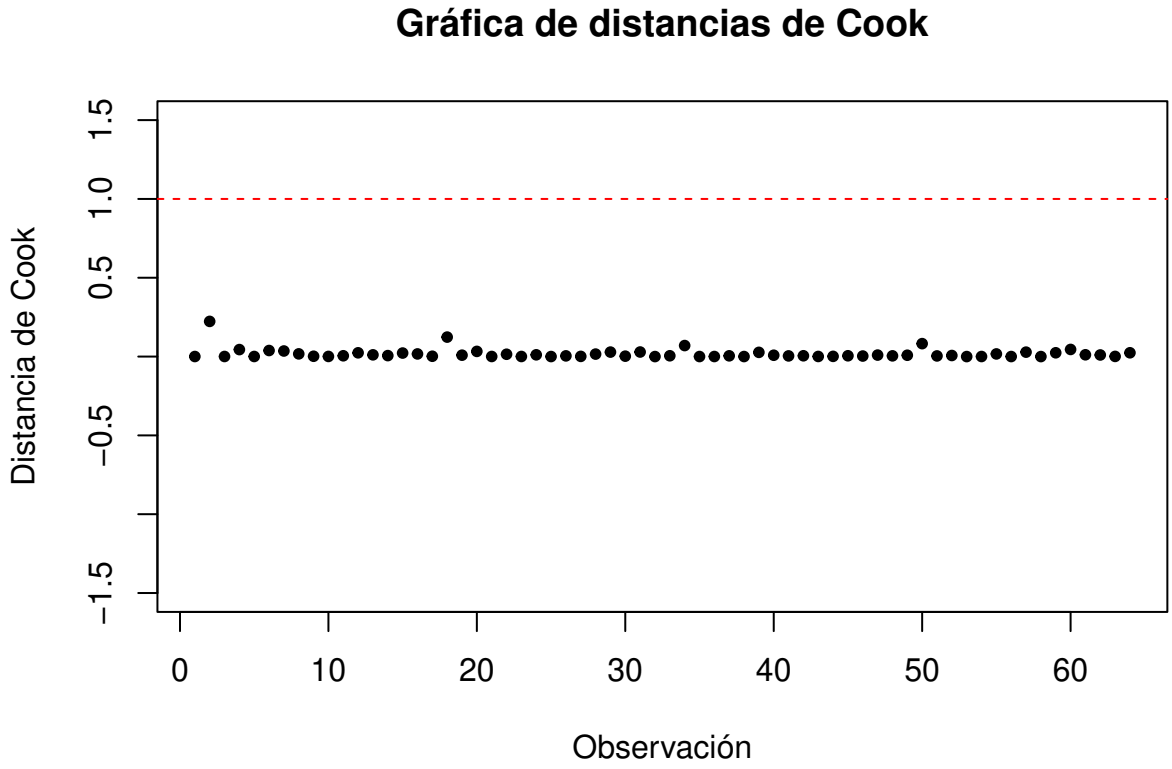


Figura 5: Criterio distancias de Cook para puntos influenciales

Podemos observar graficamente que por el criterio distancia de Cook no hay ninguna observación influencial.

También queda verificado con la tabla que hicimos para los datos atípicos.

En la tabla podemos ver claramente que las observaciones 2, 18, 34 y 50 son influenciales con por el criterio de Dffits.

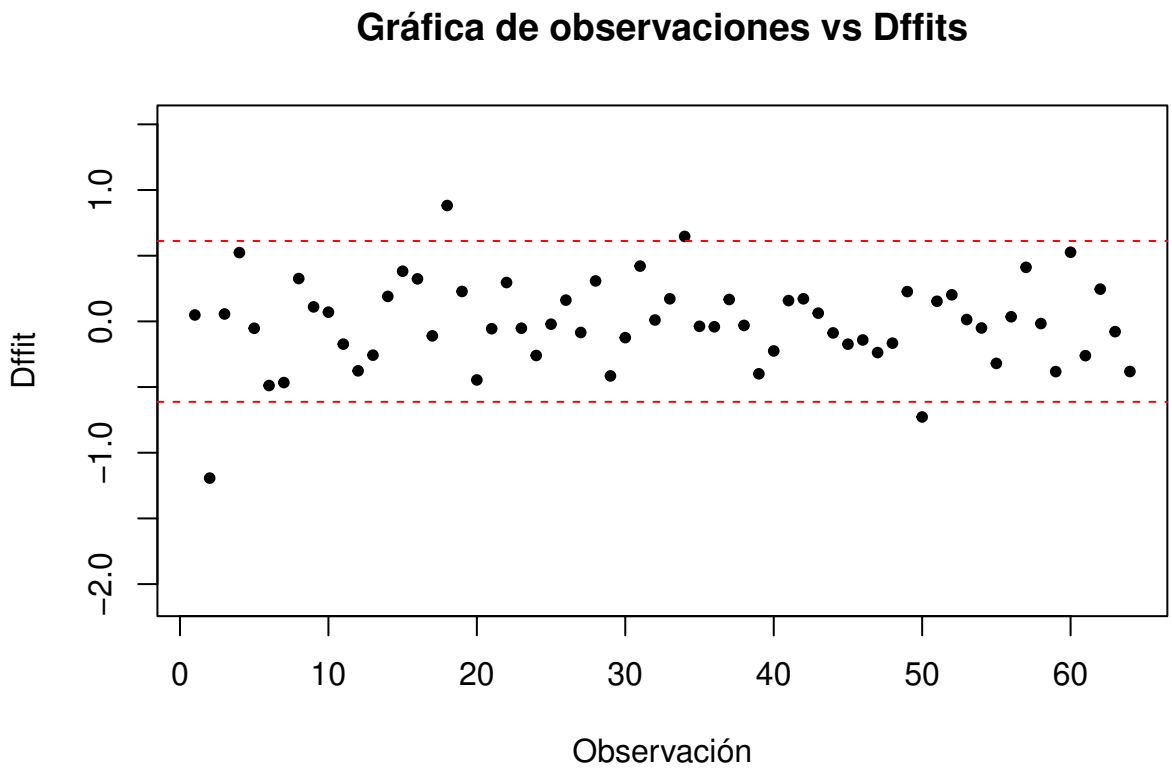


Figura 6: Criterio Dffits para puntos influenciales

Graficamente confirmamos lo visto en la anterior tabla donde hay 4 puntos influenciales según el criterio 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

Teniendo en cuenta que el supuesto de varianza se cumple y el de normalidad no, concluimos que el modelo no es valido, cabe resaltar que los supuestos cumplidos y no pueden estar siendo afectados por los puntos de balanceo o influenciales, también como se puede ver afectados los resumenes estadísticos como el $R^2$ entre otros a pesar de que no encontramos datos atípicos. Se recomienda investigar los puntos de balanceo a ver si se pueden descartar o están en una escala diferente, cualquier problema con esos puntos se debe intentar resolver antes de utilizar el modelo.