# Trabajo 1

4,0

Estudiantes

**Maria Camila Monsalve Perez**
**Ricardo Gil Alzate**
**Santiago Barrada Gonzalez**
**Yanith Alejandra Salazar Ossa**

Equipo 8

Docente

**Julieth Veronica Guarin Escudero**

Asignatura

**Estadistica II**

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medell??n
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:

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

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 |
|---|---|
| $\beta_0$ | -0.6544 |
| $\beta_1$ | 0.2266 |
| $\beta_2$ | 0.0171 |
| $\beta_3$ | 0.0511 |
| $\beta_4$ | 0.0089 |
| $\beta_5$ | 0.0013 |

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.6544 + 0.2266 X_{1i} + 0.0171 X_{2i} + 0.0511 X_{3i} + 0.0089 X_{4i} + 0.0013 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

## 1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

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 | 66.7160 | 5 | 13.343199 | 16.5359 | 4.0459e-10 |
| Error | 46.8015 | 58 | 0.806922 | | |

De la tabla Anova, se observa un valor P aproximado a cero, con un nivel de significancia $\alpha = 0.05$, por lo tanto 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 |
|---|---|---|---|---|
| $\beta_0$ | -0.6544 | 1.3291 | -0.4924 | 0.6243 |
| $\beta_1$ | 0.2266 | 0.0689 | 3.2879 | 0.0017 |
| $\beta_2$ | 0.0171 | 0.0245 | 0.6971 | 0.4885 |
| $\beta_3$ | 0.0511 | 0.0124 | 4.1193 | 0.0001 |
| $\beta_4$ | 0.0089 | 0.0064 | 1.3922 | 0.1692 |
| $\beta_5$ | 0.0013 | 0.0006 | 1.9781 | 0.0527 |

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 $\alpha$.

### 1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** 0.2266 indica que por cada dia de estadia promedio de los pacientes en el hospital ($X_1$) el promedio de riesgo de infección aumenta en un 0.2266, cuando las demás variables predictoras permanecen constantes.

$\hat{\beta}_3$**:** 0.0511 indica que por cada cama adicional promedio ($X_3$) el promedio de riesgo de infección aumenta en un 0.0511, cuando las demás variables predictoras permanecen constantes.

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

para calcular el coeficiente de determinación se usa la siguiente fórmula:

$$R^2 = \frac{SSR}{SST} \tag{2}$$

Según los datos de la tabla ANOVA el valor de $R^2$ es 0.5877, lo que significa que aproximadamente el 58.77 % 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 tres covariables con el P-valor más bajo 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 | 46.802 | X1 X2 X3 X4 X5 |
| Modelo reducido | 89.677 | X2 X4 |

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

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

### 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \ \cdots, \ \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{(89.677 - 46.802)/3}{0.806922} \\ &= 17.7113 \end{aligned} \tag{3}$$

Ahora, comparando el $F_0$ con $f_{0.05,3,58} = 0.1166$, se puede ver que $F_0 > f_{0.05,3,58}$ y por tanto se rechaza la hipotesis nula $\beta_1 = \beta_3 = \beta_5 = 0$ y se concluye que el riesgo de infeccion promedio depende de al menos una de la variables asociadas a los valores p mas pequeños

## 3. Pregunta 3

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

Se quiere probar que $H_0 : \beta_1 = 2\beta_5 3\beta_2 = \beta_4$ por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 2\beta_5; \ \beta_2 = 3\beta_4 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

Reescribiendo matricialmente tenemos que:

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

Con $\mathbf{L}$ dada por

$$L = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & -3 & 0 \end{bmatrix}$$

El modelo reducido está dado por:

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

Donde $X_{1i}^* = X_{1i} + 2X_{5i}$ y $X_{2i}^* = X_{2i} + 3X_{4i}$

2 pt

### 3.2. Estadístico de prueba

El estadístico de prueba $F_0$ está dado por:

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

Entonces tenemos que:

$$F_0 = \frac{(SSE(MR)^* - 46.802))/2}{0.806922} \overset{H_0}{\sim} f_{2,58} \tag{5}$$

## 4. Pregunta 4

16,5 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:

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

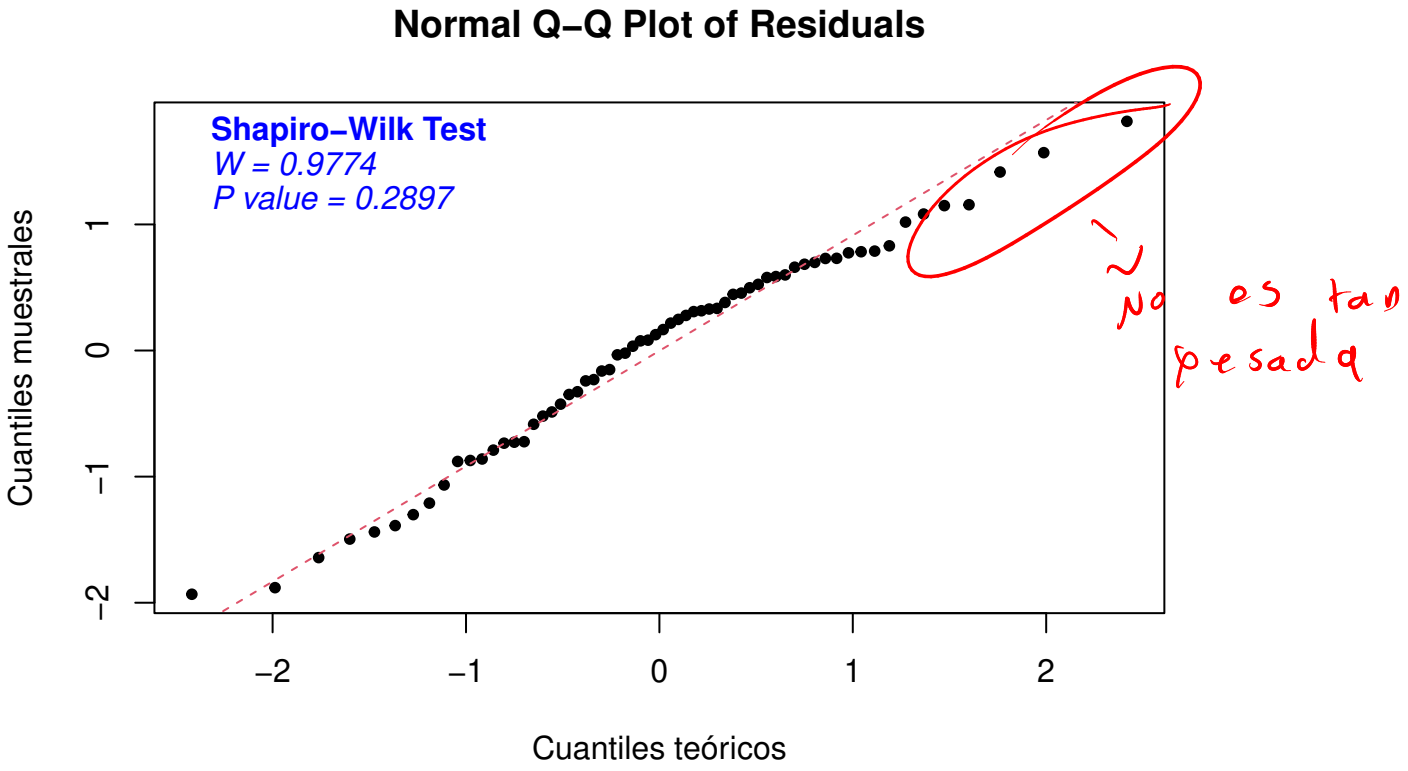



Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.2897 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 la cola superior más pesada y con un patron irregular, 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

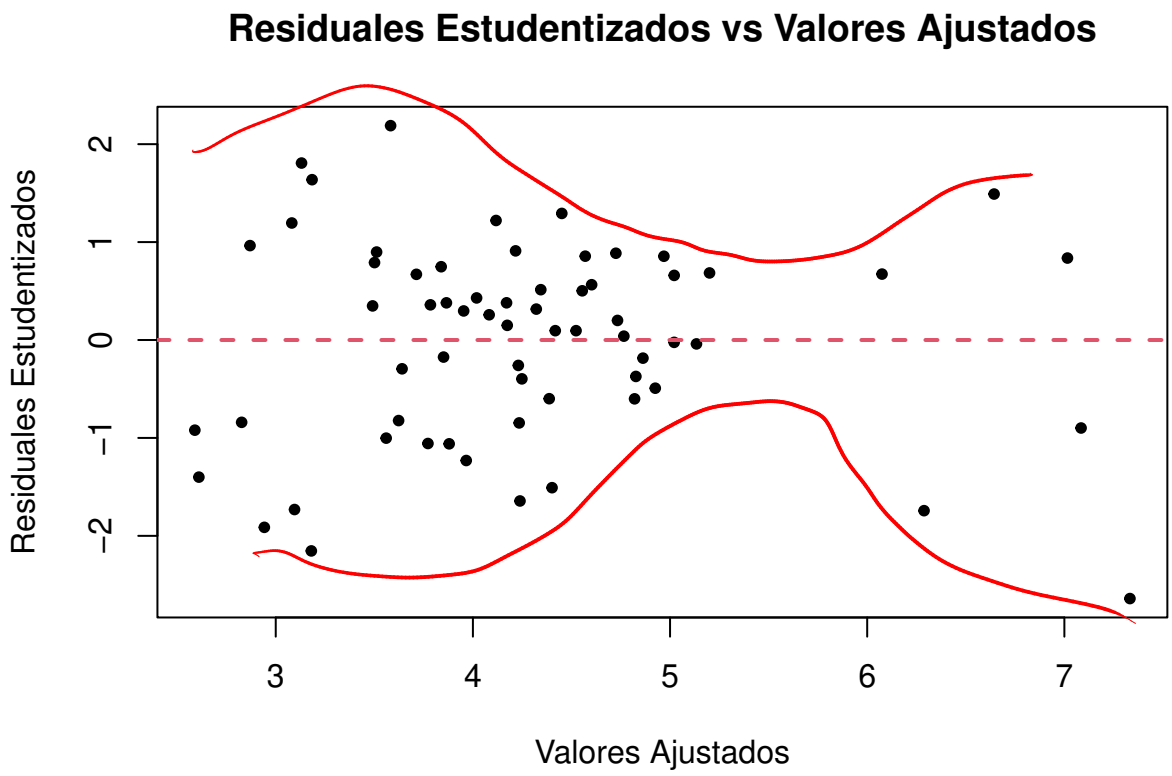



Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que hay patrones con varianza irregular y una disminución de la dispersión de algunos datos en el centro de la grafica, mientras que al inicio y al final aumenta este comportamiento, lo que nos permite descartar la existencia de varianza constante en este modelo. Es posible que algunas observaciones extremas estén afectando este análisis.

## 4.2. Verificación de las observaciones

### 4.2.1. Datos atípicos

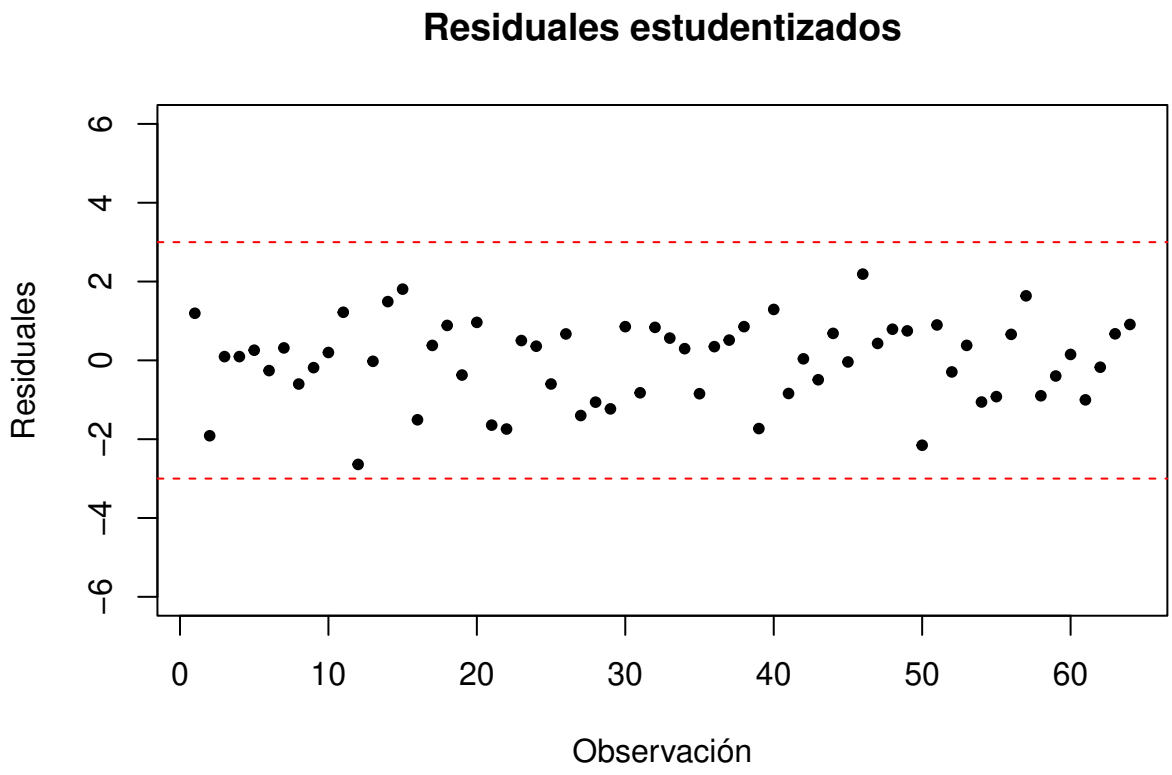



Figura 3: Identificación de datos atípicos

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

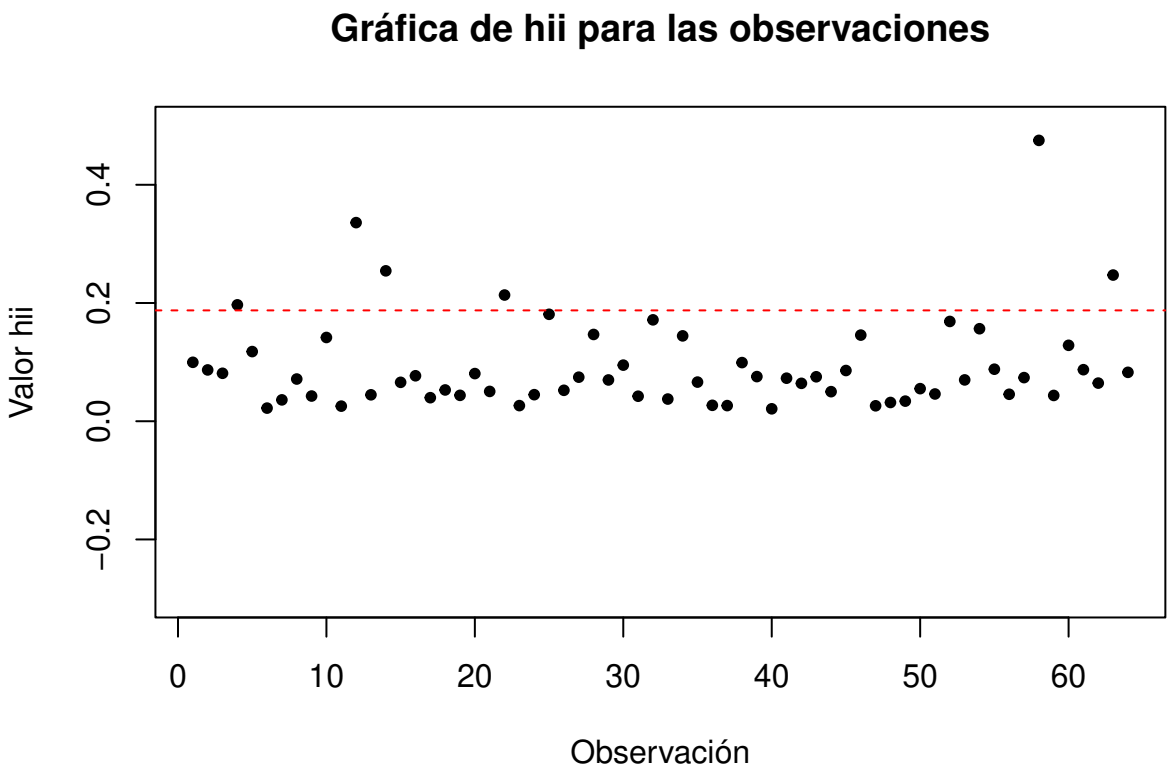



Figura 4: Identificación de puntos de balanceo

Al observar la gráfica de observaciones vs valores $h_{ii}$, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.1875$, se puede apreciar que existen 6 datos del conjunto que son puntos de balanceo (4,12,14,22,58,63) según el criterio bajo el cual $h_{ii} > 2\frac{p}{n} = 0.1875$.

¿Qué causan? 2p+

### 4.2.3. Puntos influenciales

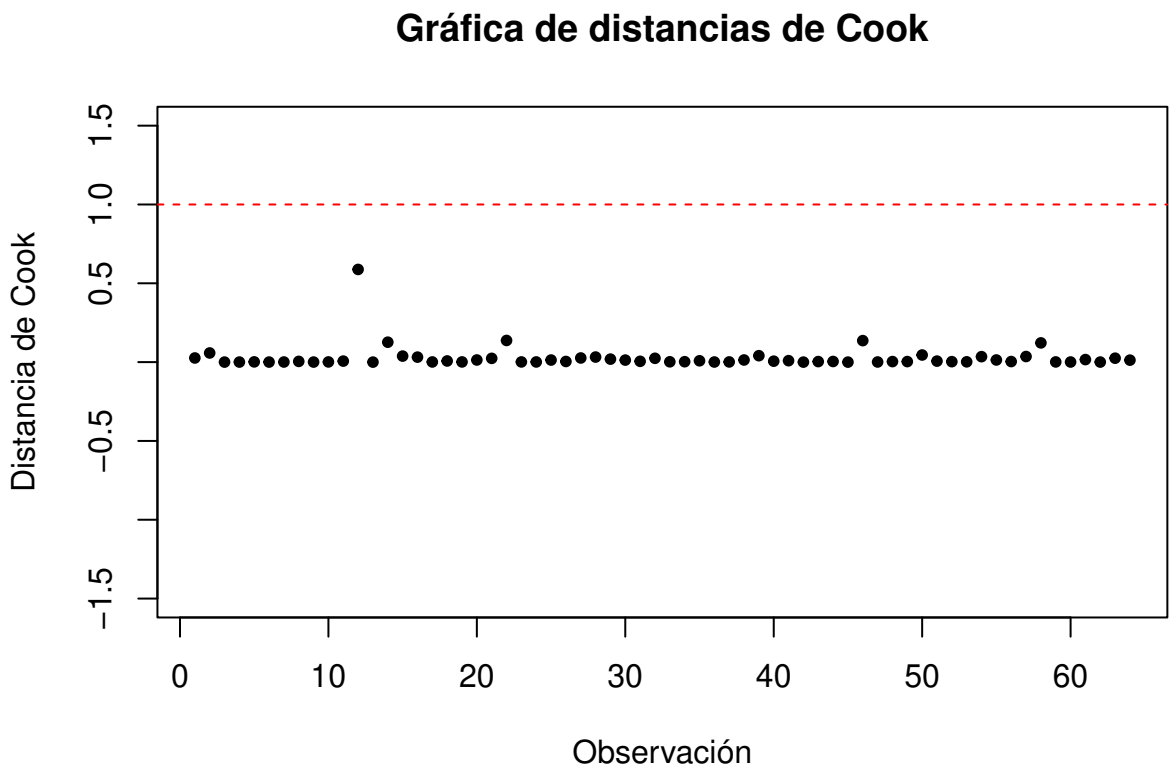
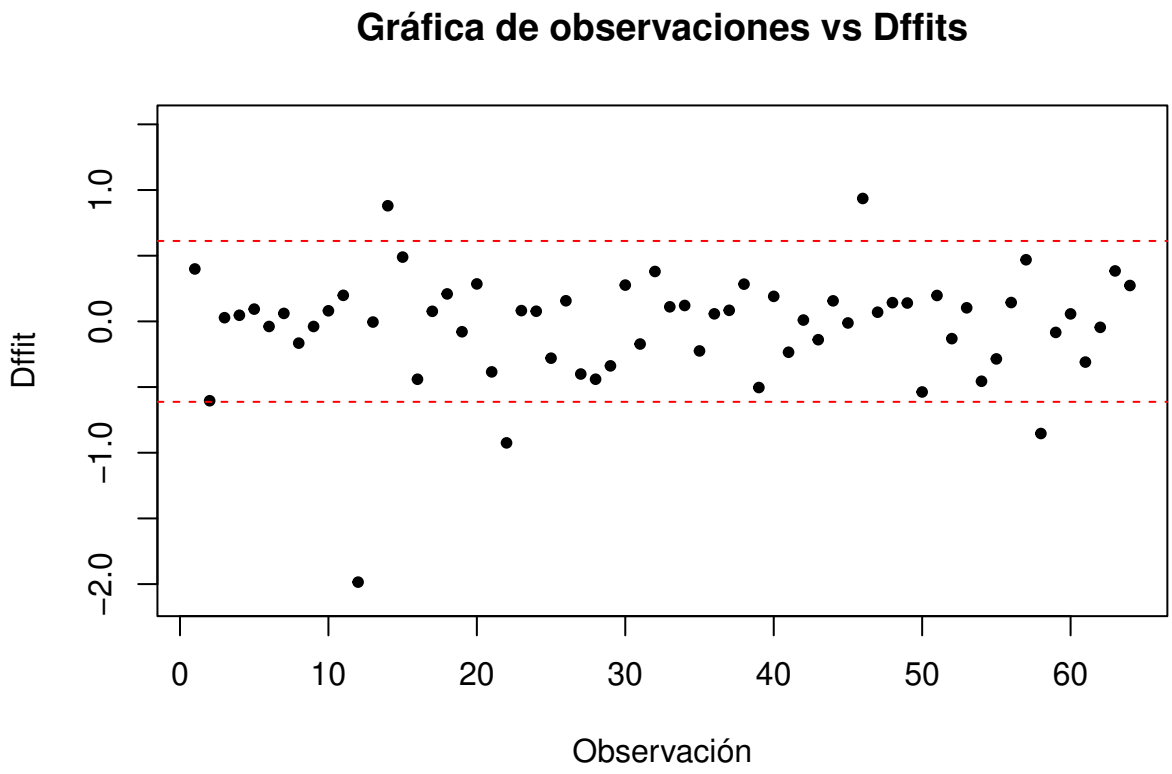



Figura 5: Criterio distancias de Cook para puntos influenciales

podemos apreciar en la grafica que, según el criteria de Cook donde los $D_i > 1$, no hay datos influénciales que afecten las estimaciones de los parametros




3pt

Figura 6: Criterio Dffits para puntos influenciales

Como se puede ver,las observaciones 12,14,22,46,58 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.612$, es un punto influencial.

¿Qué causan?

## 4.3. Conclusión

2,5pt

Al analizar la gráfica de cuantiles podemos evidenciar que este modelo no distribuye normal a pesar de tener un valor-P superior a $\alpha = 0.05$, también se pudo comprobar mediante el gráfico de residuales estudentizados vs valores ajustados que no tiene varianza constante debido a que algunos datos se dispersan en los extremos. Por otro lado, al realizar la identificación de observaciones extremas podemos encontrar que en nuestro modelo no existen datos atípicos que pueda afectar el resultado de ajuste de la regresión, ádemas en la gráfica de los puntos de balanceo se pudieron observar 6 datos que pueden justificar la NO normalidad de los errores estándar y la poca variabilidad explicada en el $R^2 = 0.5877$. Por último, los puntos influenciales que encontramos fueron 5, los cuales tiene un impacto notable en coeficientes de la regresión ajustada.

En conclusión, por lo mencionado anteriormente podemos afirmar que el modelo de regresión presentado en este trabajo NO es válido. → ojo, deben especificar que es por el no cumplimiento de supuestos unicamente