# Trabajo 1-equipo 43

3,4


Estudiantes

**Brayan Estiven Arias Davila**
**Juan Paulo Lemus Cano**
**Mariana Pitalua Martinez**
**Daniel Gonzalez Henao**

Equipo

Docente

**Carlos Mario Lopera**


Asignatura

**Estadadistica II**


Sede Medellin
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... acá dicen el nombre de las variables

- Y: Riesgo de infección
- $X_1$:Duracción de la estadia
- $X_2$: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 |
|---|---|
| $\beta_0$ | -0.7016 |
| $\beta_1$ | 0.1575 |
| $\beta_2$ | 0.0210 |
| $\beta_3$ | 0.0497 |
| $\beta_4$ | 0.0164 |
| $\beta_5$ | 0.0007 |

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.7016 + 0.1575 X_{1i} + 0.021 X_{2i} + 0.0497 X_{3i} + 0.0164 X_{4i} + 7\times10^{-4} X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 113$$

## 1.2. Significancia de la regresión

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

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

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 | 58.1899 | 5 | 11.637974 | 14.2556 | 4.28863e-09 |
| Error | 47.3501 | 58 | 0.816382 | | |

De la tabla Anova, se observa un valor P es muy pequeño (4.28863e-09), 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 |
|---|---|---|---|---|
| $\beta_0$ | -0.7016 | 1.4295 | -0.4908 | 0.6254 |
| $\beta_1$ | 0.1575 | 0.0674 | 2.3367 | 0.0229 |
| $\beta_2$ | 0.0210 | 0.0261 | 0.8058 | 0.4236 |
| $\beta_3$ | 0.0497 | 0.0127 | 3.9265 | 0.0002 |
| $\beta_4$ | 0.0164 | 0.0066 | 2.5027 | 0.0152 |
| $\beta_5$ | 0.0007 | 0.0007 | 0.9401 | 0.3511 |

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

$\hat{\beta}_1$**:Indica que cada unidad que se aumente en la duración de la estadía el promedio del riesgo de infección aumenta en 0,1575 unidades, cuando las demás predictoras se mantienen fijas.**

$\hat{\beta}_3$**:Indica que por cada unidad que aumente el numero de camas el promedio de riesgo de infección aumenta en 0,0497 unidades, cuando las demás se mantiene fijas.**

$\hat{\beta}_4$**:Indica que por cada unidad que aumente el censo promedio diario, la respuesta media de riesgo de infección aumenta en 0,0164 unidades, cuando las demás predictoras se mantienen fijas.**

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

Segun la información obtenidad de la tabla ANOVA del modelo ajustado y utilizando la formula para hallar $R^2 = SSR/SST$, se toma el valor del $SST = SSE + SSR = (58.1899 + 47.3502)$ se obtiene finalmente un coeficiente de determinación de $R^2 = 0.5514$.Lo cual,significa que dicho modelo explica aproximadamente el 55.14 % de la variabilidad total de la respuesta.

## 2. Pregunta 2

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

Las covariables con el valor-P más bajo en el modelo fueron $X_1, X_3, X_4$, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_4 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 4 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

| | $SSE$ | Covariables en el modelo |
|---|---|---|
| Modelo completo | 47.350 | X1 X2 X3 X4 X5 |
| Modelo reducido | 86.603 | X2 X5 |

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

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

### 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_0,\beta_3,\beta_5) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,58} \\
&= \frac{(86.603 - 47.350)/3}{47.350/58} \\
&= 16.02727
\end{aligned} \tag{2}
$$

Ahora con un nivel de significancia de $\alpha = 0.95$ y con un cuantil $f_{0.95,3,58} = 2.7636$, se puede ver que $F_0 > f_{0.95,3,58}$ y por tanto qué se rechaza $H_0$, entonces se concluye que las variables del subconjunto no se pueden retirar del modelo.

# 3. Pregunta 3

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

El efecto de la duración de la estadia $X_1$ sobre el riesgo de infección $Y_i$ es igual a dos veces el número de enfermeras $X_5$;por consiguiente se plantea la siguiente prueba de hipótesis: Pregunta 2: El efecto de 3 veces el número de camas sobre el riesgo de infección es igual al censo promedio diario; por lo tanto se pkante la siguiente prueba de hipotesis

$$
\begin{cases}
\mathrm{H}_0 : \beta_1 = 2\beta_5;\ \beta_4 = 3\beta_3 \\
\mathrm{H}_1 : \text{Alguna de las igualdades no se cumple}
\end{cases}
$$

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

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

El modelo reducido está dado por:

$$
Y_o = \beta_o + \beta_2 X_{2i} + \beta_3 X^*_{3i} + \beta_5 X^*_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2);\ 1 \leqslant i \leqslant 64
$$

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

## 3.2. Estadístico de prueba

El estadístico de prueba $F_0$ está dado por:

$$
F_0 = \frac{(SSE(MR) - 47.350)/2}{0.8163} \overset{H_0}{\sim} f_{2,58} \tag{3}
$$

# 4. Pregunta 4

13,5 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:

$$\begin{cases} H_0 : \varepsilon_i \sim \text{Normal} \\ H_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$

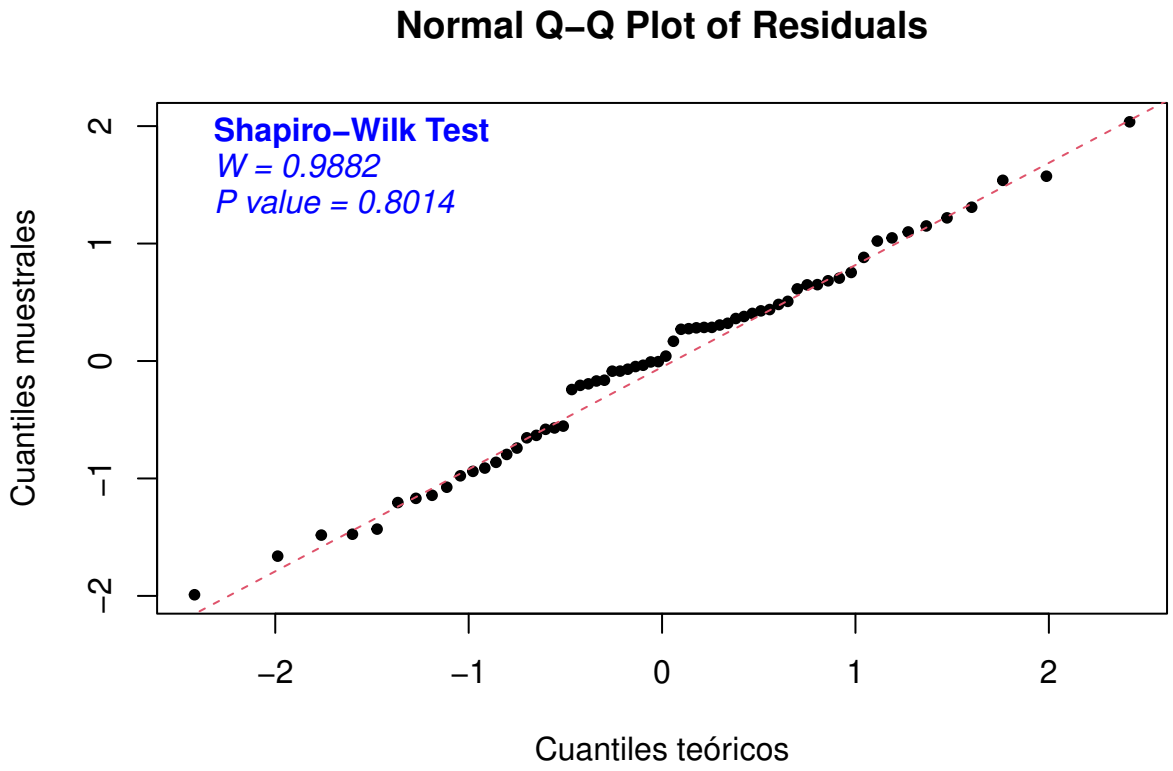


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

3,5 pt

Al ser el P-valor aproximadamente igual a 0.8014 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 0 y varianza $\sigma^2$,.El modelo parece ser válido en términos de normalidad de los residuales y homocedasticidad de los mismos, ya que no se encontraron pruebas suficientes para rechazar estos supuestos; tambien se puede ver graficamente que los valores estan cercanos a la linea recta distribuyendo de forma normal.

Faltó más análisis gráfico

### 4.1.2. Varianza constante

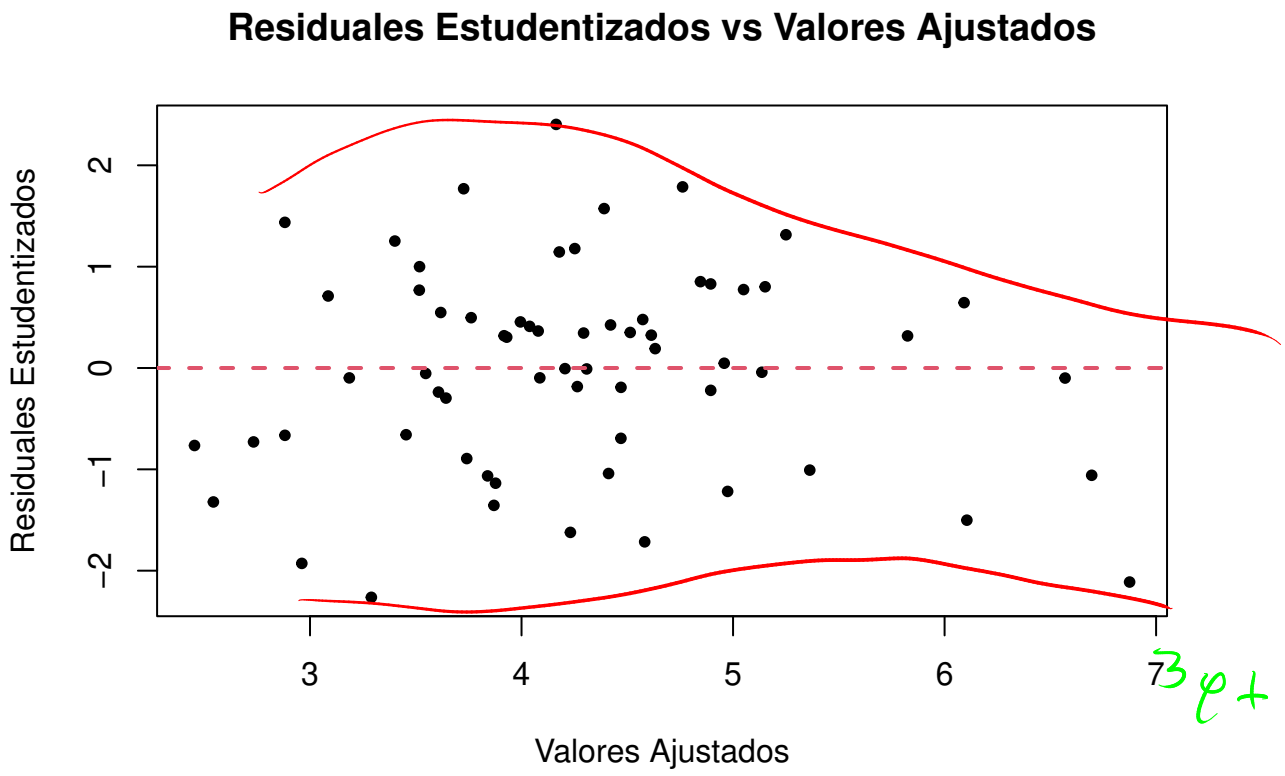


Figura 2: Gráfico residuales estudentizados vs valores ajustados

La ausencia de patrones visibles en el gráfico de "Residuales Estudentizados vs Valores Ajustados" sugiere que la varianza de los errores es constante a lo largo de los valores ajustados, lo que es consistente con el supuesto de varianza constante y fortalece la validez del modelo en términos de homocedasticidad.

## 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

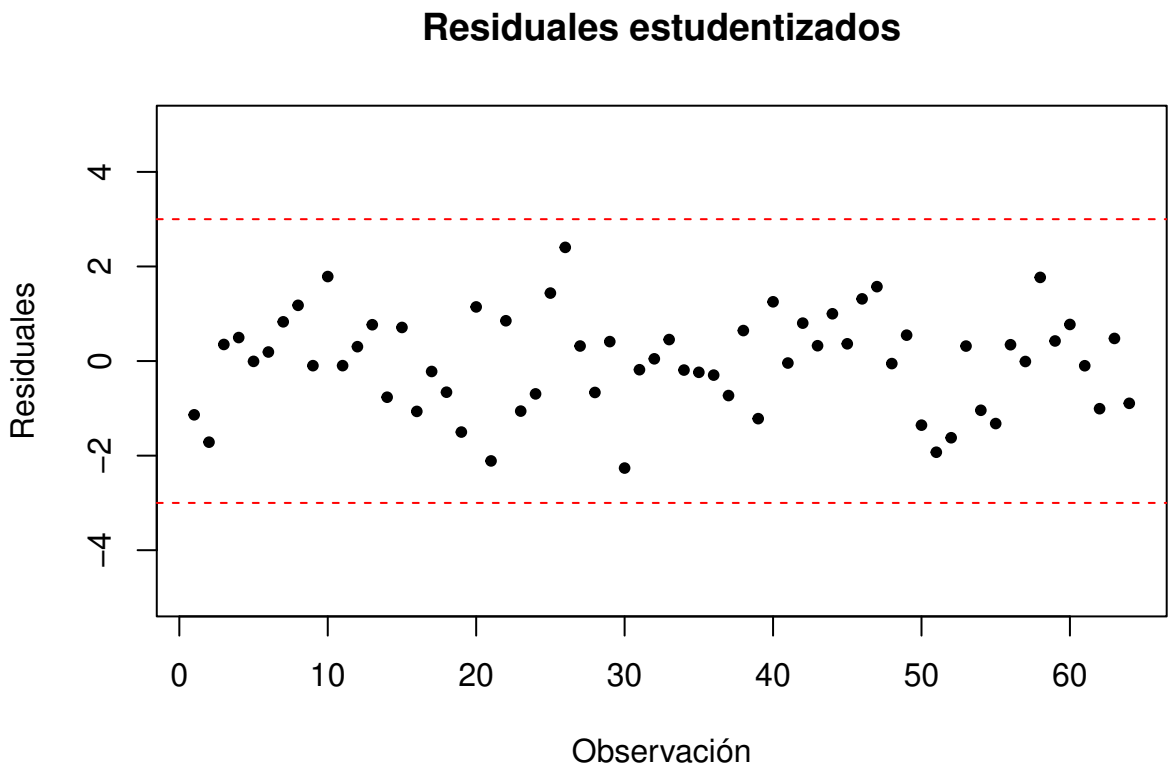


Figura 3: Identificación de datos atípicos

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

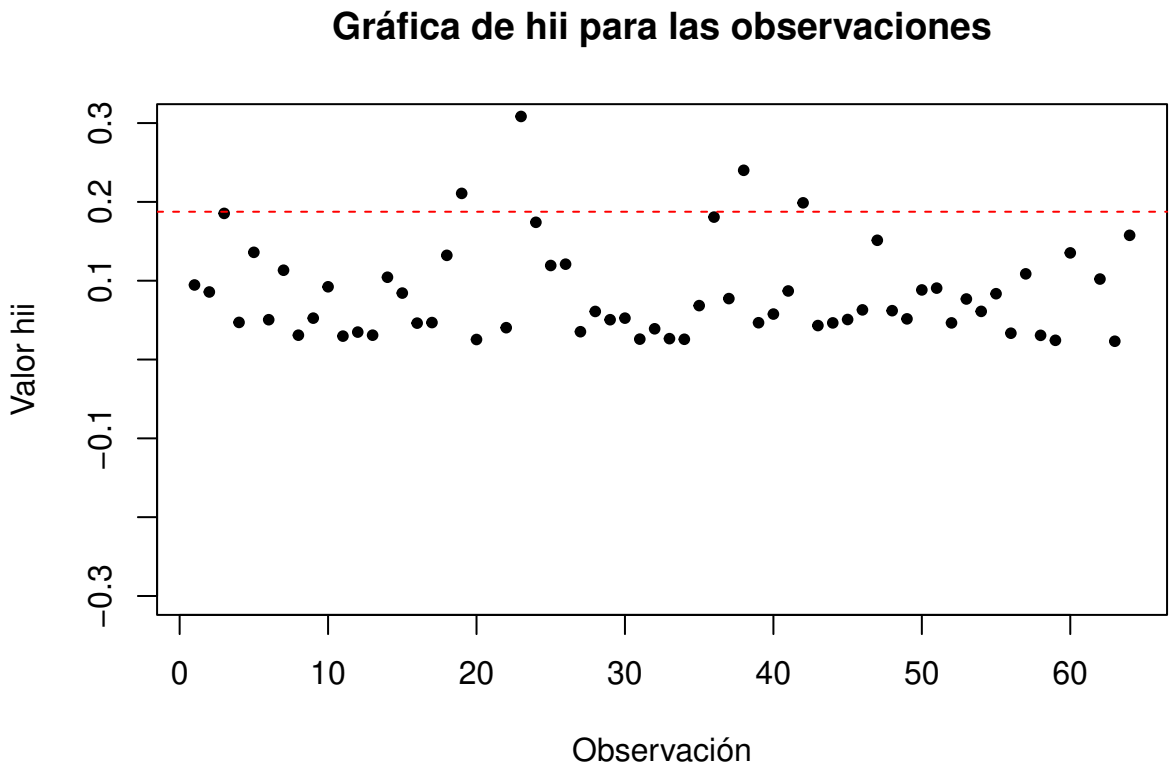


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 19  -1.5015  0.1004    0.2108 -0.7847
## 21  -2.1121  0.5014    0.4028 -1.7897
## 23  -1.0586  0.0833    0.3084 -0.7076
## 38   0.6446  0.0219    0.2401  0.3605
## 42   0.8016  0.0266    0.1988  0.3980
## 61  -0.0993  0.0011    0.3969 -0.0799
```

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 que hacen referencia a las observaciones 19,21,23,38,42 y 61,según el criterio bajo el cual $h_{ii} > 0.1875$, los cuales son los presentados en la tabla.

#### 4.2.3. Puntos influenciales

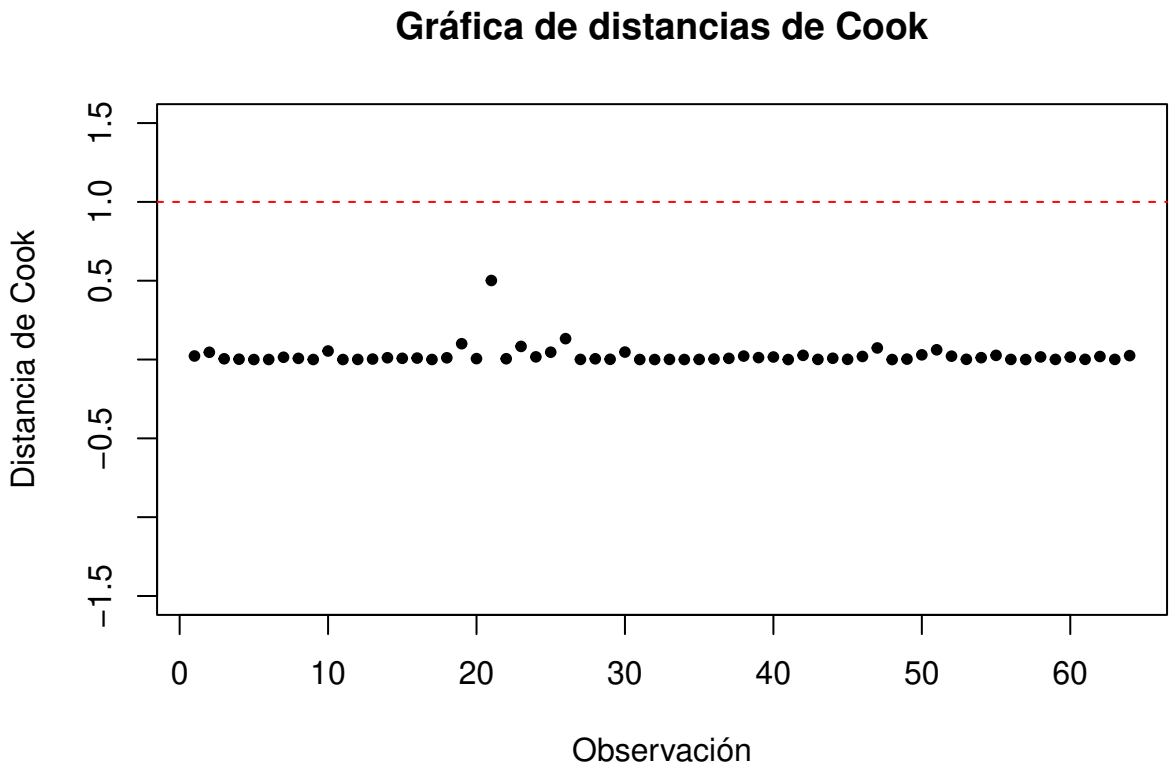


Figura 5: Criterio distancias de Cook para puntos influenciales

Como arrojo el analisis de distancias de Cook,no hay datos influenciales segun este criterio; sin embargo, se procede a analizar el criterio de los DFFITS para verificar o descartar la existencia de puntos influenciales.

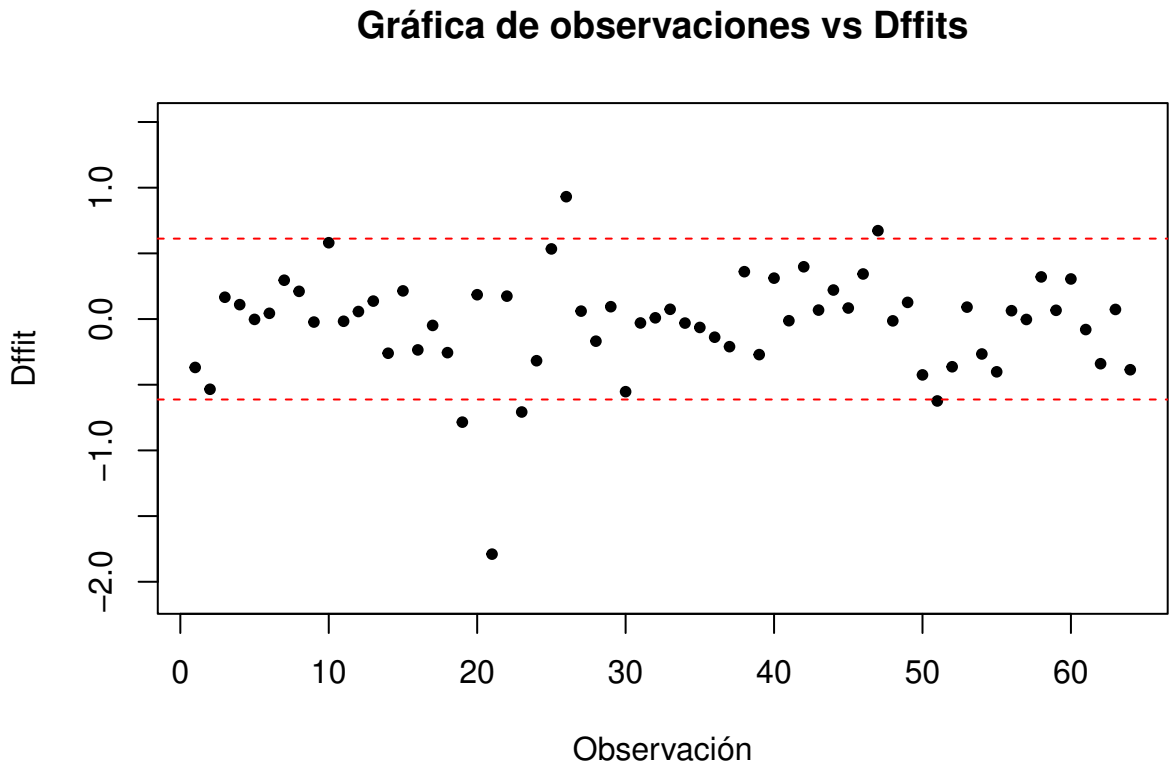


Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 19  -1.5015  0.1004    0.2108 -0.7847
## 21  -2.1121  0.5014    0.4028 -1.7897
## 23  -1.0586  0.0833    0.3084 -0.7076
## 26   2.4037  0.1325    0.1210  0.9315
## 47   1.5731  0.0735    0.1513  0.6729
## 51  -1.9282  0.0617    0.0906 -0.6235
```

Como se puede ver,las observaciones 19,21,23,26,47,51 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.6124$. En sintesis, unificando los analisis de ambos criterios solo el criterio de los Dffits nos proporciona puntos influenciales, los cuales hacen referencia a las observaciones ya mencionadas.

## 4.3. Conclusión

En conclusión el modelo es valido ya que es estadísticamente significativo y los errores cumplen con los supuestos de normalidad y homogenidad de varianza, teniendo en cuenta la presencia de valores extremos al interpretar los resultados del modelo, ya que pueden tener un efecto significativo en las estimaciones y conclusiones derivadas del mismo..El analisis de valores extremos arroja que en las observaciones [19,21,23] existen puntos de balanceo e influenciales al mismo tiempo,mientras que en las observaciones [38,42,61] son puntos de balanceo y las observaciones [26,47,51] son datos influenciales individualmente