Trabajo 1

3,1

Estudiantes

Juan Carlos Múnera Arango
Nicolás Pérez Vásquez
Jorge Andrés Higuita Monsalve
Emmanuel Alberto Mejia Arango

Equipo #43

Docente

Francisco Javier Rodriguez Cortes

Asignatura

Estadística II

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

Teniendo en cuenta la base de datos asignado, la cual es la **Equipo43.txt**, las covariables son: Duración de la estadía(DE), Rutina de cultivos(RU), Número de camas(NC), Censo promedio diario(CPD), Número de enfermeras(NE).

El modelo que se propone es:

$$Riesgo\ de\ Infeccion_i = \beta_0 + \beta_1 DE_i + \beta_2 RU_i + \beta_3 NC_i + \beta_4 CPD_i + \beta_5 NE_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

1.1. Modelo de regresión

Al ajustar el modelo anterior se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	-0.4056
$\hat{\beta}_1$	0.1431
$\hat{\beta}_2$	0.0179
$\hat{\beta}_3$	0.0494
$\hat{\beta}_4$	0.0151
$\hat{\beta}_5$	0.0019

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0,4056 + 0,1431`X_{1i}0,0179X_{2i} + 0.0494X_{3i} + 0,0151X_{4i} + 0,0019X_{5i},; \ 1 \leqslant i \leqslant 65$$

donde $1 \leq i \leq 65$

1.2. Significacia de la regresión

Para la significancia de la regresión se hará uso de la siguiente tabla anova, usando un estadístico de prueba **F**:

Cuadro 2: Tabla anova significancia de la regresión

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	Valor-P
Modelo de regresión	71.4589	5	14.2918	13.6816	7.21218e-09
Error	61.6313	59	1.0446		

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 3, 4, 5 \end{cases}$$

De la tabla anova, se concluye que se rechaza la hipótesis nula para la no significancia de los parámetros, por lo tanto, la regresión es significativa y algún parámetro por consiguiente es significativo.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar cual de estos es significativo para el modelo de regresión:

Cuadro 3: Resumen de los coeficientes

	Estimación β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-0.4056	1.8155	-0.2234	0.8240
β_1	0.1431	0.1018	1.4054	0.1651
β_2	0.0179	0.0340	0.5269	0.6003
β_3	0.0494	0.0136	3.6233	0.0006
β_4	0.0151	0.0078	1.9287	0.0586
β_5	0.0019	0.0008	2.2862	0.0259

1.4. Significancia de los parámetros

- Ya que no se especifica el valor de α se asume el $\alpha = 0.05$.
- Un estadístico de prueba para los parámetros del modelo sería $T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{59}$ *bajo* H_0
- $\hat{\beta}_3$: Su valor-P es menor que un $\alpha = 0.05$ entonces rechazamos la hipótesis nula, luego, el parámetro es significativo para el modelo.
- $\hat{\beta}_5$: Su valor-P es menor que un $\alpha = 0.05$ entonces rechazamos la hipótesis nula, luego, el parámetro es significativo para el modelo.
- $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_4$: Para estos parámetros, se acepta la hipótesis nula y determinamos que no son significativos para el modelo de regresión.
- 6: Del modelo anterior podemos concluir que el Promedio de Riesgo de Infeccion se puede predecir principalmente con las variables Numero de camas(NC) y Numero de Enfermeras(NE).

1.5. Interpretación de los parámetros estimados

- $\hat{\beta}_0$: El parámetro no tiene interpretación ya que el valor (0,0,0,0,0,0) no está en las observaciones, es decir, no está en las covariables.

- $\hat{\beta}_3$: El cambio promedio en el Riesgo de Infeccion aumenta 0.0494 por cada unidad de cambio en el Numero de camas(NC), cuando las demás variables predictoras se mantienen constantes.
- $\hat{\beta}_5$: El cambio promedio en el Riesgo de Infeccion aumenta 0.0019 por cada unidad de cambio en el Numero de Enfermeras(NE), cuando las demás variables predictoras se mantienen constantes.

1.6. Coeficiente de determinación múltiple R^2

Para el cálculo del $R^2 = \dfrac{SSR}{SST}$, que se puede calcular de la tabla anova, el modelo tiene un $R^2 = 0.5369$, es decir, el modelo de regresión múltiple lineal explica el 53.7 % de la variabilidad total del porcentaje de grasa corporal.

2. Pregunta 2

2.1. Planteamiento prueba de hipótesis y modelo reducido

Los parámetros cuyos valores-P fueron los más altos corresponden a $\beta_0, \beta_1, \beta_2$, por lo tanto, se plantea la siguiente prueba de hipótesis:

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

El modelo completo es el definido en la sección 1.1, y el modelo reducido es:

$$\text{MR: } Y_i = \beta_0 + \beta_1 DE_i + \beta_2 RU_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Se presenta la siguiente tabla con el resumen de todas las regresiones para plantear el estadístico de prueba:

Cuadro 4: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo Completo	61.631	X1 X2 X3 X4 X5
Modelo Reducido	65.090	X3 X4 X5

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSE(\beta_0, \beta_3, \beta_4, \beta_5) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/2}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)} \overset{H_0}{\sim} f_{2,59}$$

$$= \frac{(65.090 - 61.631)/2}{61.631/59} = 1.6556684$$

Ahora, comparando a un nivel de significancia de $\alpha = 0.05$, F_0 con $f_{2,59} = 3.1531233$

Entonces se concluye que aceptamos la hipostesis nula,pues su estadístico de prueba es menor que el estadístico $f_{(0.05,2,59)}$, pero esto no necesariamente significa que podamos descartar las variables del modelo.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = -\beta_3, \beta_2 = 0.1\beta_4 \\ \mathrm{H}_1 : \beta_1 \neq -\beta_3 \vee \beta_2 \neq 0.1\beta_4 \end{cases}$$

Luego igualando a 0 las hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 + \beta_3 = 0 \wedge \beta_2 - 0.1\beta_4 = 0 \\ \mathrm{H}_1 : \beta_1 + \beta_3 \neq 0, \beta_2 - 0.1\beta_4 \neq 0 \end{cases}$$

Donde la hipótesis nula H_0 contiene $m = 2$ ecuaciones, donde H_0 esta dado como:

$$\mathrm{H}_0 : \begin{cases} \beta_1 + \beta_3 = 0 \\ \beta_2 - 0.1\beta_4 = 0 \end{cases}$$

de la anterior H_0 se puede construir la matriz $\mathbf{L}$ y asi formar el sistema de ecuaciones de la forma $\mathbf{L}\beta = 0$:

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -0.1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

donde, la matriz $\mathbf{L}$ tiene un $r = 2$ filas linealmente independientes, con un modelo reducido dado por:

MR: $Riesgo\,de\,Infeccion_i = \beta_0 + \beta_1(DE_i - NC_i) + \beta_2(RU_i + 0.1CPD_i) + \beta_5 NE_i + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$

3.2. Estadístico de prueba

El estadistico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} = \frac{(SSE(MR) - SSE(MF))/2}{SSE(MF)/59} \overset{F_0}{\sim} f_{2,59} \qquad (1)$$

donde se rechaza H_0, si $F_0 > f_{\alpha,2,59}$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis (Shapiro Wilk):

$$\begin{cases} H_0 : \varepsilon_i \sim \mathcal{N} \\ H_1 : \varepsilon_i \nsim \mathcal{N} \end{cases}$$

Acompañado de un gráfico cuantil-cuantil:

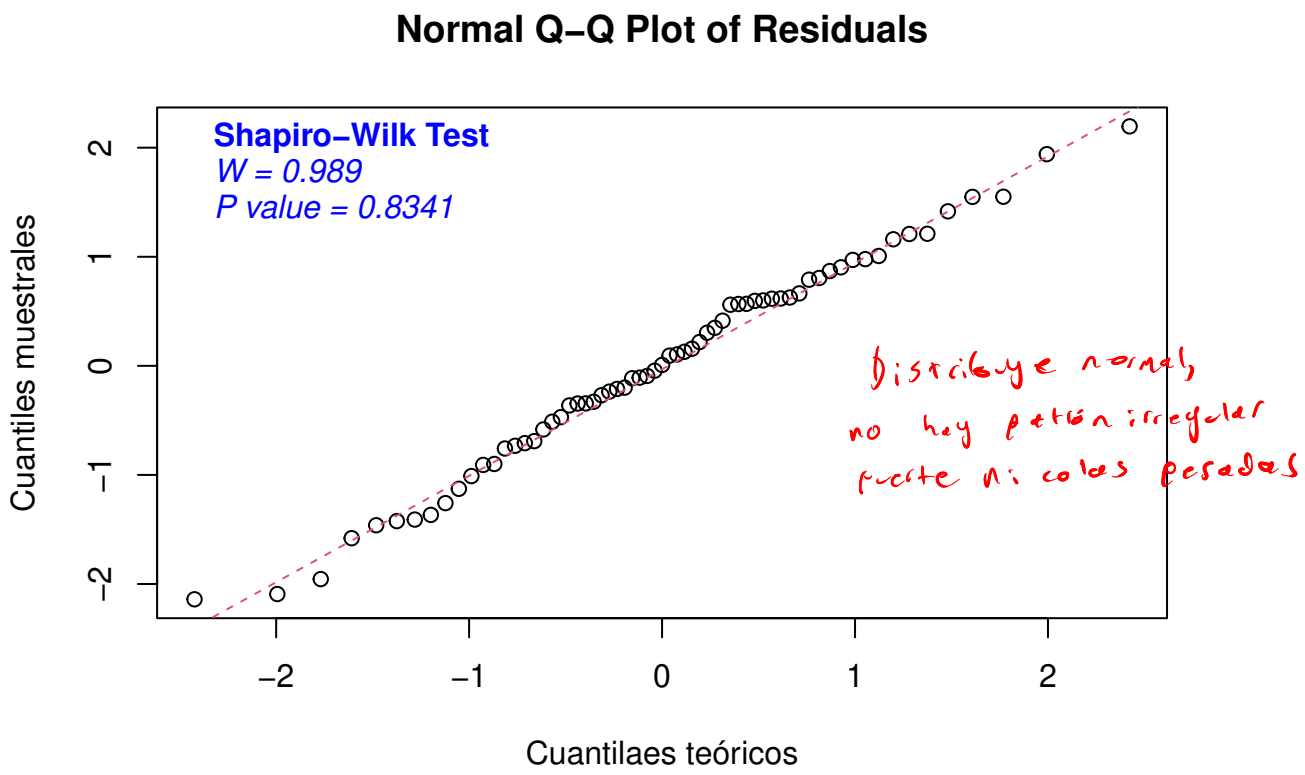

Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Al ser el P-valor aproximadamente igual a 0.8341 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 μ y varianza σ^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.

4.2. Supuesto de media 0 y varianza constante

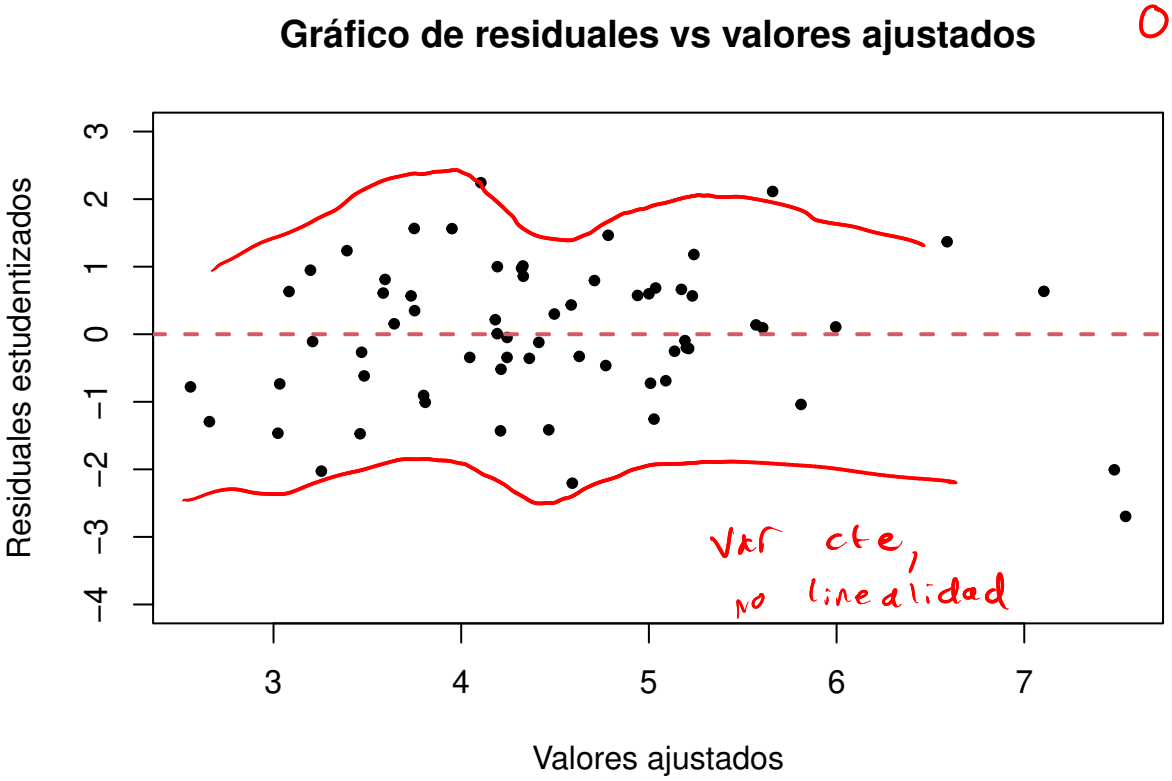

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Según el gráfico de los residuales estudentizados vs valores ajustados, se puede observar un patron de arco teniendo en cuenta los datos superiores, por lo que concluimos que el modelo no cumple con el supuesto de varianza constante.

4.3. Observaciones extremas

4.3.1. Datos atípicos

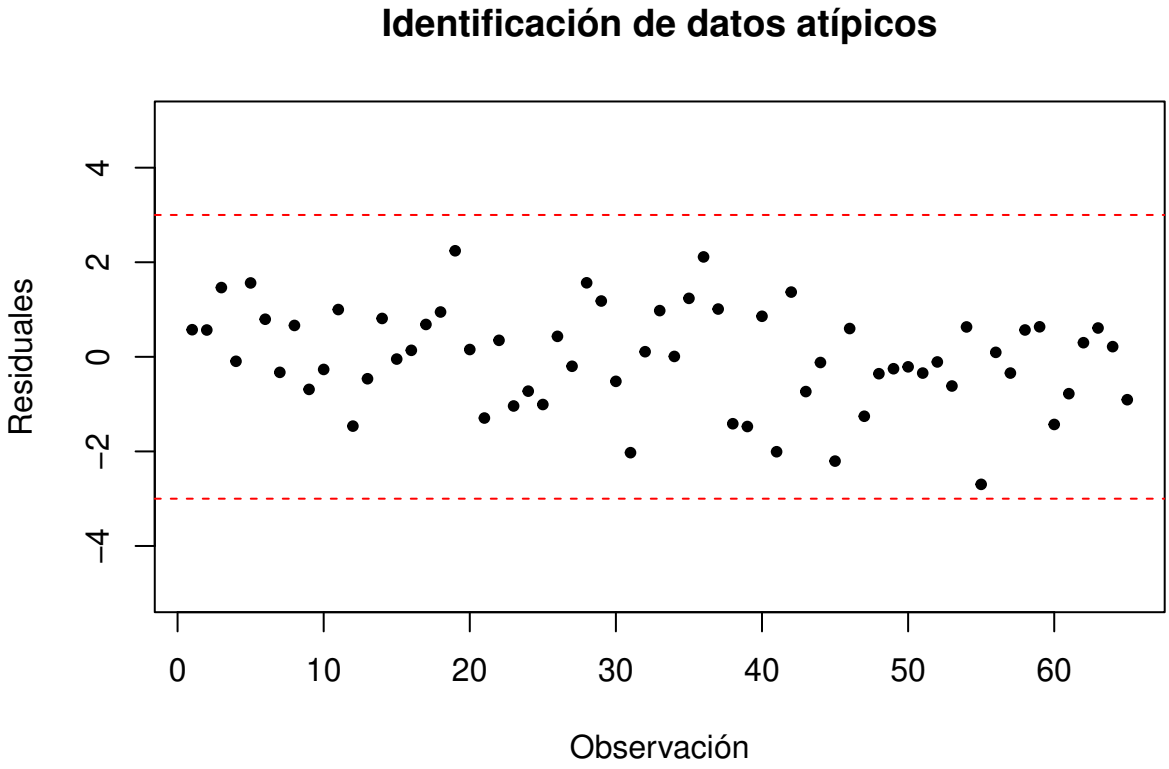

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

4.4. Puntos de balanceo

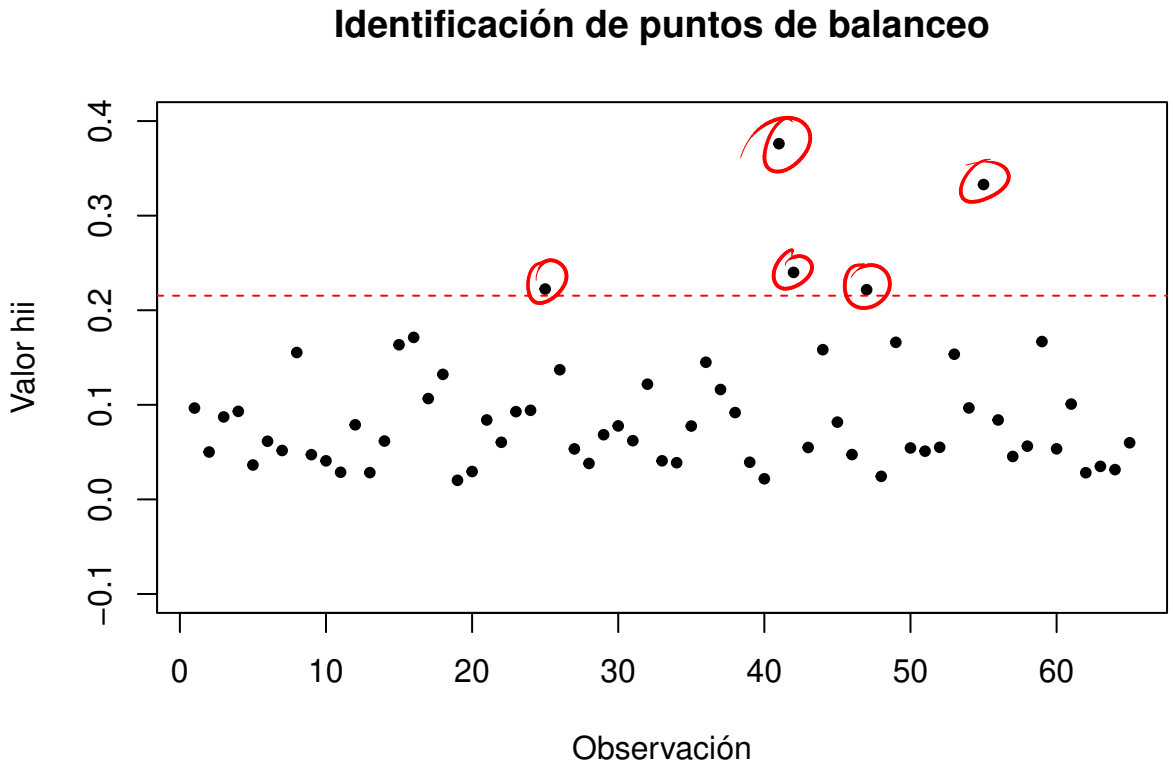

Figura 4: Identificación de puntos de balanceo

Según el criterio de h_{ii} existen 5 puntos con un valor $h_{ii} > 0.2153846$, las cuales son las observaciones 25, $41, 42, 47, y$ 55 . Estos son los puntos de balanceo del modelo.

4.4.1. Puntos influenciales

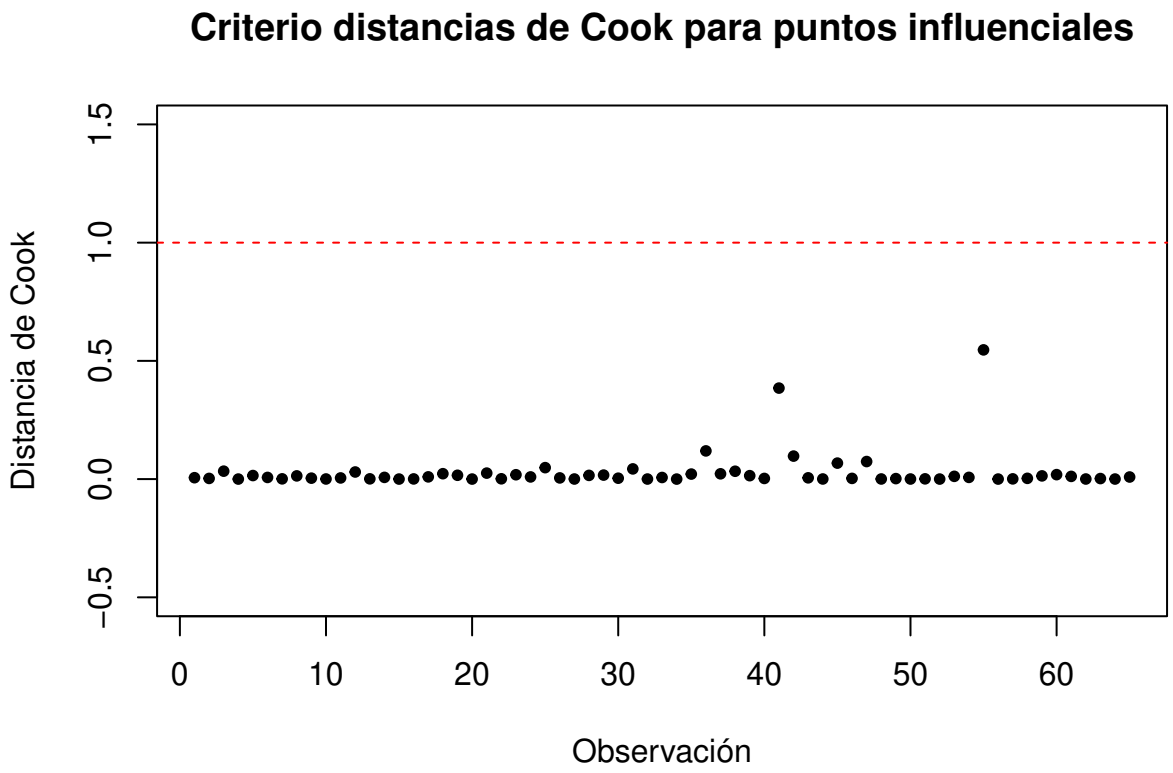

Figura 5: Criterio distancias de Cook para puntos influenciales

Según el criterio de Cook no hay valores influenciales que afecten la estimación de los parámetros.

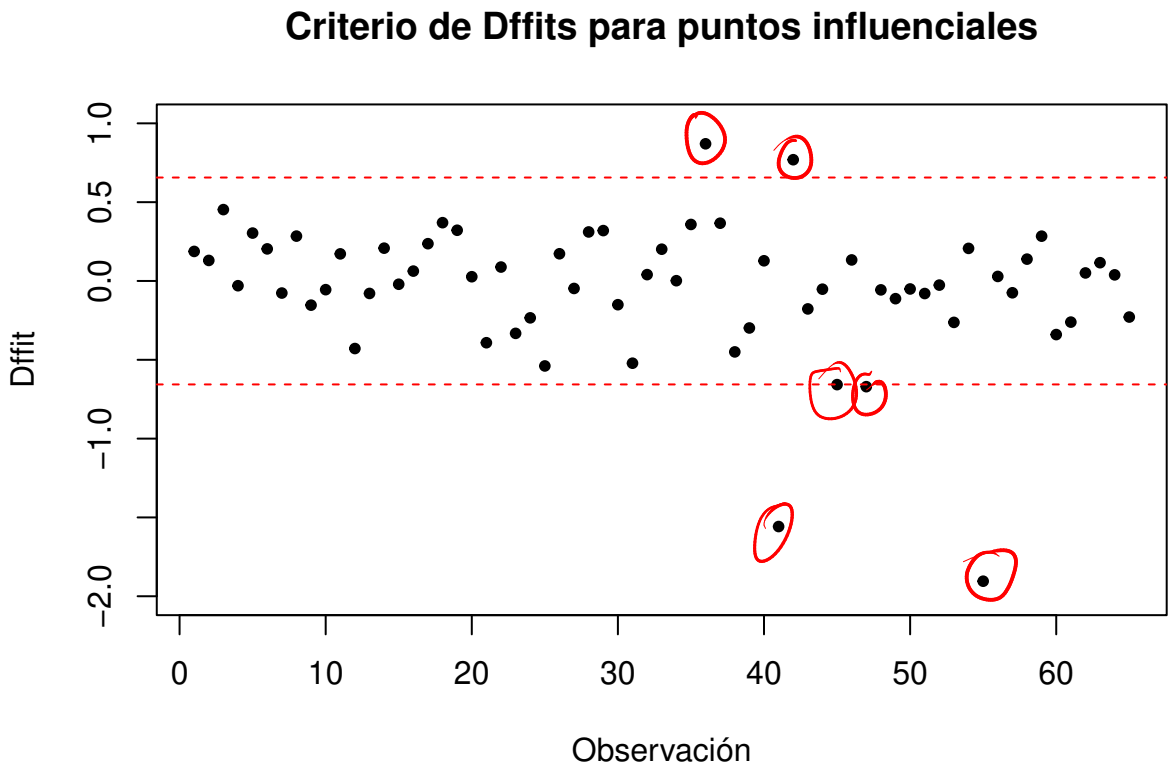

Figura 6: Criterio de Dffits para puntos influenciales

```
##    res.stud Cook.D hii.value   Dffts
## 36   2.1131 0.1192    0.1450  0.8703
## 41  -2.0067 0.3848    0.3761 -1.5579
## 42   1.3691 0.0972    0.2400  0.7693
## 45  -2.2049 0.0676    0.0817 -0.6575
## 47  -1.2565 0.0742    0.2217 -0.6706
## 55  -2.6964 0.5465    0.3329 -1.9046
```

Según el criterio de Dffits las observaciones $36, 41, 42, 45, 47$ y 55 son puntos influenciales que afectan las estimaciones de $\hat{y}$.

4.5. Conclusiones

- 1: El modelo no es válido, ya que no cumple con los supuestos de normalidad de los errores.
- 2: Del punto dos se concluye que aceptamos la hipotesis nula donde B1=B2, pero que esto no necesariamente significa que los parametros se pueden descartar del modelo.
- 3: Se podría usar alguna transformación sobre el modelo para buscar cumplir con el supuesto de normalidad de los errores.
- 4: Las observaciones $41, 42, 47, 55$ son puntos de balanceo y puntos influencial al mismo tiempo.