

# Trabajo 1

Estudiantes

Docente

Asignatura

## Estadística II



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

Sede Medellín  
30 de marzo de 2023

# Índice

<b>1. Pregunta 1</b>	<b>3</b>
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 . . . . .	4
1.5. Coeficiente de determinación múltiple $R^2$ . . . . .	4
<b>2. Pregunta 2</b>	<b>5</b>
2.1. Planteamiento pruebas de hipótesis y modelo reducido . . . . .	5
2.2. Estadístico de prueba y conclusión . . . . .	5
<b>3. Pregunta 3</b>	<b>5</b>
3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . . .	5
3.2. Estadístico de prueba . . . . .	6
<b>4. Pregunta 4</b>	<b>6</b>
4.1. Supuestos del modelo . . . . .	6
4.1.1. Normalidad de los residuales . . . . .	6
4.1.2. Varianza constante . . . . .	8
4.2. Verificación de las observaciones . . . . .	9
4.2.1. Datos atípicos . . . . .	9
4.2.2. Puntos de balanceo . . . . .	10
4.2.3. Puntos influyentes . . . . .	11
4.3. Conclusión . . . . .	12

## Índice de figuras

1.	Gráfico cuantil-cuantil y normalidad de residuales . . . . .	7
2.	Gráfico residuales estudentizados vs valores ajustados . . . . .	8
3.	Identificación de datos atípicos . . . . .	9
4.	Identificación de puntos de balanceo . . . . .	10
5.	Criterio distancias de Cook para puntos influenciales . . . . .	11
6.	Criterio Dffits para puntos influenciales . . . . .	12

## 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 . . . . .	5

# 1. Pregunta 1

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); \quad 1 \leq i \leq 55$$

## 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	
$\beta_0$	2.9148
$\beta_1$	0.0875
$\beta_2$	-0.0193
$\beta_3$	0.0628
$\beta_4$	0.0026
$\beta_5$	0.0022

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 2.9148 + 0.0875X_{1i} - 0.0193X_{2i} + 0.0628X_{3i} + 0.0026X_{4i} + 0.0022X_{5i} + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); \quad 1 \leq i \leq 50$$

## 1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MST}{MSE} \stackrel{H_0}{\sim} f_{5,49} \quad (1)$$

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	32.9947	5	6.59895	7.93706	1.53888e-05
Error	40.7391	49	0.83141		

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  $1 \leq j \leq 5$ , 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
$\beta_0$	2.9148	1.8579	1.5688	0.1231
$\beta_1$	0.0875	0.1123	0.7790	0.4397
$\beta_2$	-0.0193	0.0306	-0.6292	0.5321
$\beta_3$	0.0628	0.0169	3.7108	0.0005
$\beta_4$	0.0026	0.0083	0.3135	0.7553
$\beta_5$	0.0022	0.0007	2.8909	0.0057

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

### 1.4. Interpretación de los parámetros

$\hat{\beta}_3$ :

$\hat{\beta}_5$ :

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

El modelo tiene un coeficiente de determinación múltiple  $R^2 = 0.4475$ , lo que significa que aproximadamente el 44.75 % 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 covariable con el P-valor más alto en el modelo fueron  $X_1, X_2, X_4$ , por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

Cuadro 4: Resumen tabla de todas las regresiones

	$SSE$	Covariables en el modelo				
Modelo completo	40.739	X1	X2	X3	X4	X5
Modelo reducido	41.663			X3	X5	

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon_i; \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); 1 \leq i \leq 55$$

### 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \dots, \beta_5))/3}{MSE(\beta_0, \dots, \beta_5)} \stackrel{H_0}{\sim} F_{3,49} \\ &= \\ &= \end{aligned} \tag{2}$$

Ahora, comparando el  $F_0$  con  $f_{0.95,3,49} = 2.7939$ , se puede ver que  $F_0 > f_{0.95,3,49}$

Es posible o no descartar las variables del subconjunto

## 3. Pregunta 3

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

Se hace la pregunta si ... por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \beta_1 = 3\beta_3; \beta_2 = \beta_4 \\ H_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

$$\begin{cases} H_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ H_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Con  $\mathbf{L}$  dada por

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + \beta_5 X_{5i} + \varepsilon_i, \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); 1 \leq i \leq 50$$

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

### 3.2. Estadístico de prueba

El estadístico de prueba  $F_0$  está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \stackrel{H_0}{\sim} f_{2,49} \quad (3)$$

## 4. Pregunta 4

### 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 shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

$$\begin{cases} H_0 : \varepsilon_i \sim \text{Normal} \\ H_1 : \varepsilon_i \not\sim \text{Normal} \end{cases}$$

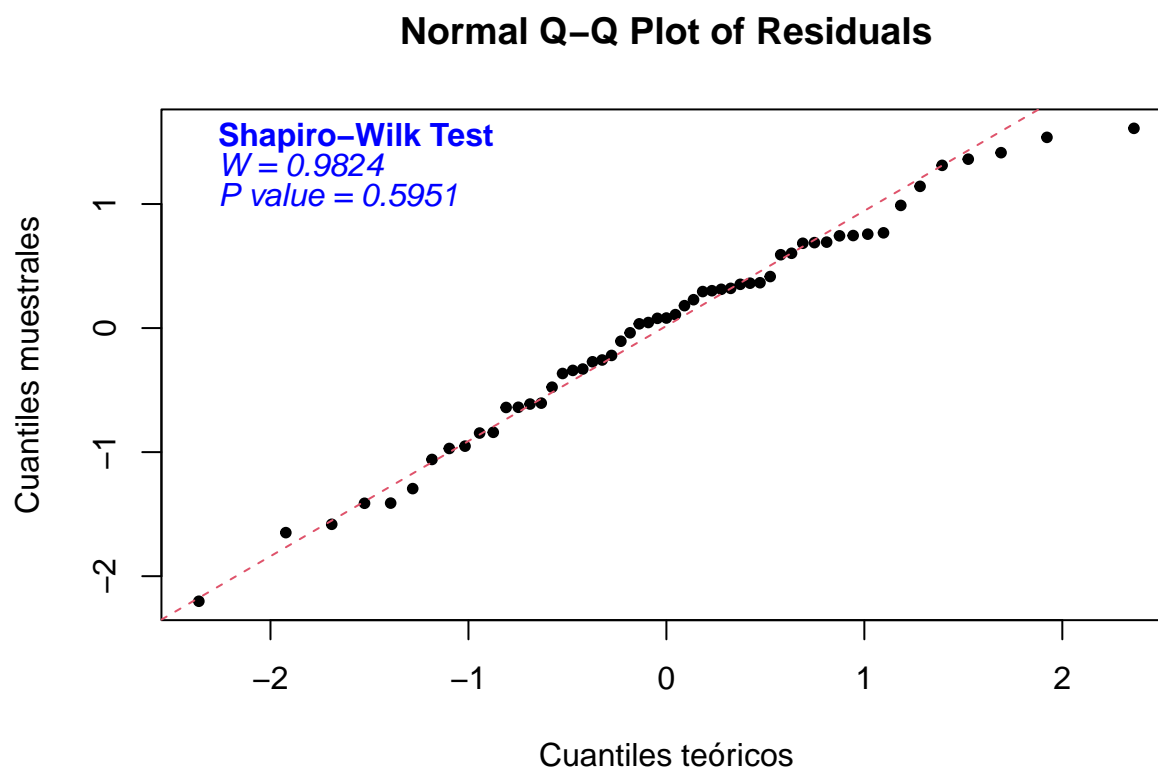


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.5951 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 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.



#### 4.1.2. Varianza constante

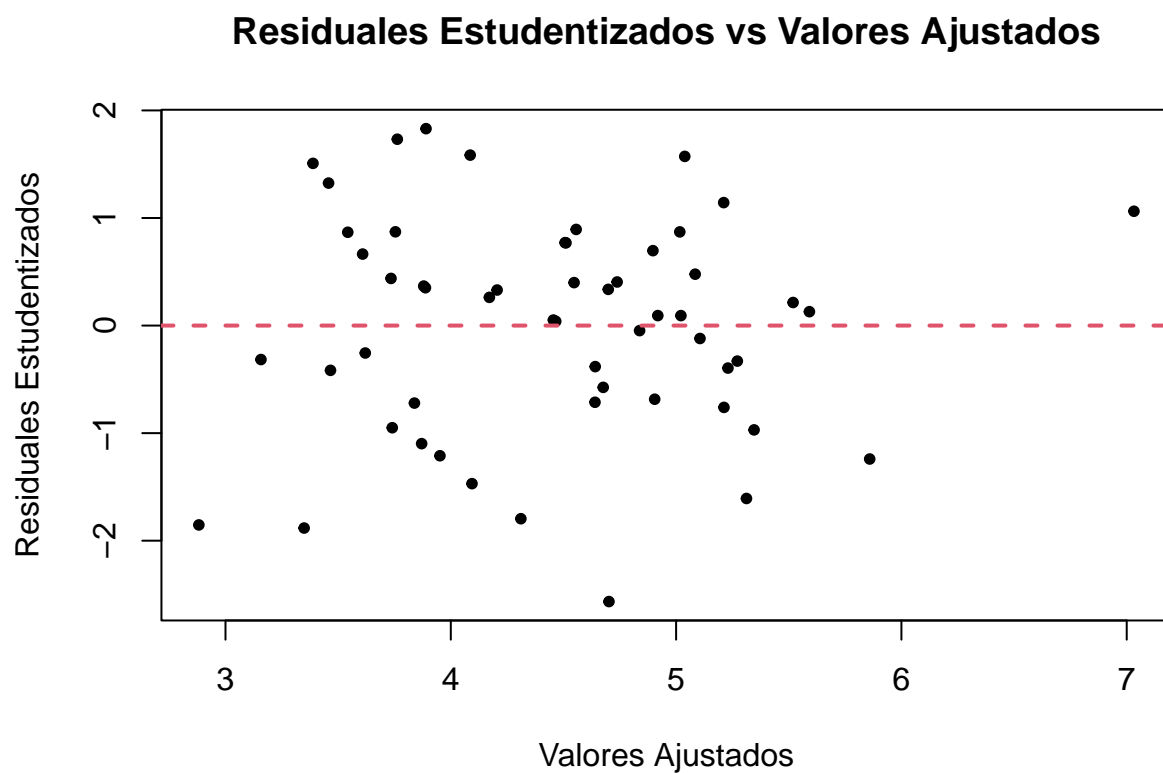


Figura 2: Gráfico residuales estudentizados vs valores ajustados

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

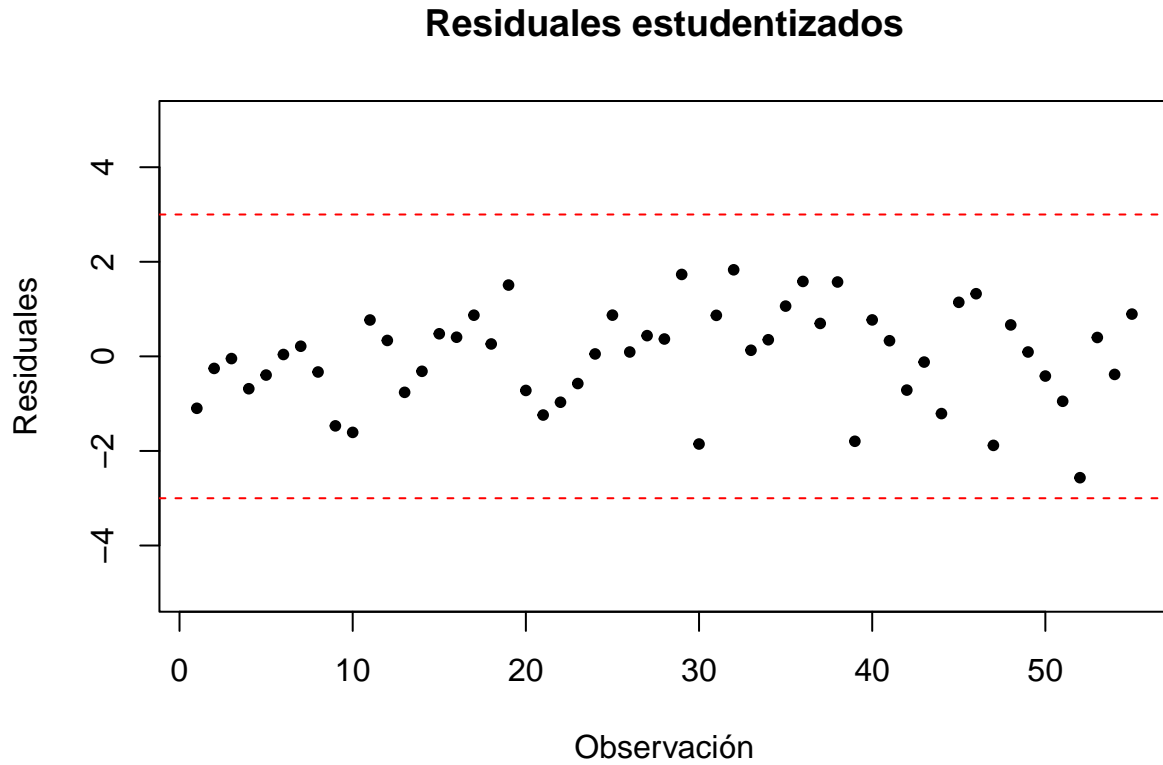


Figura 3: Identificación de datos atípicos

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

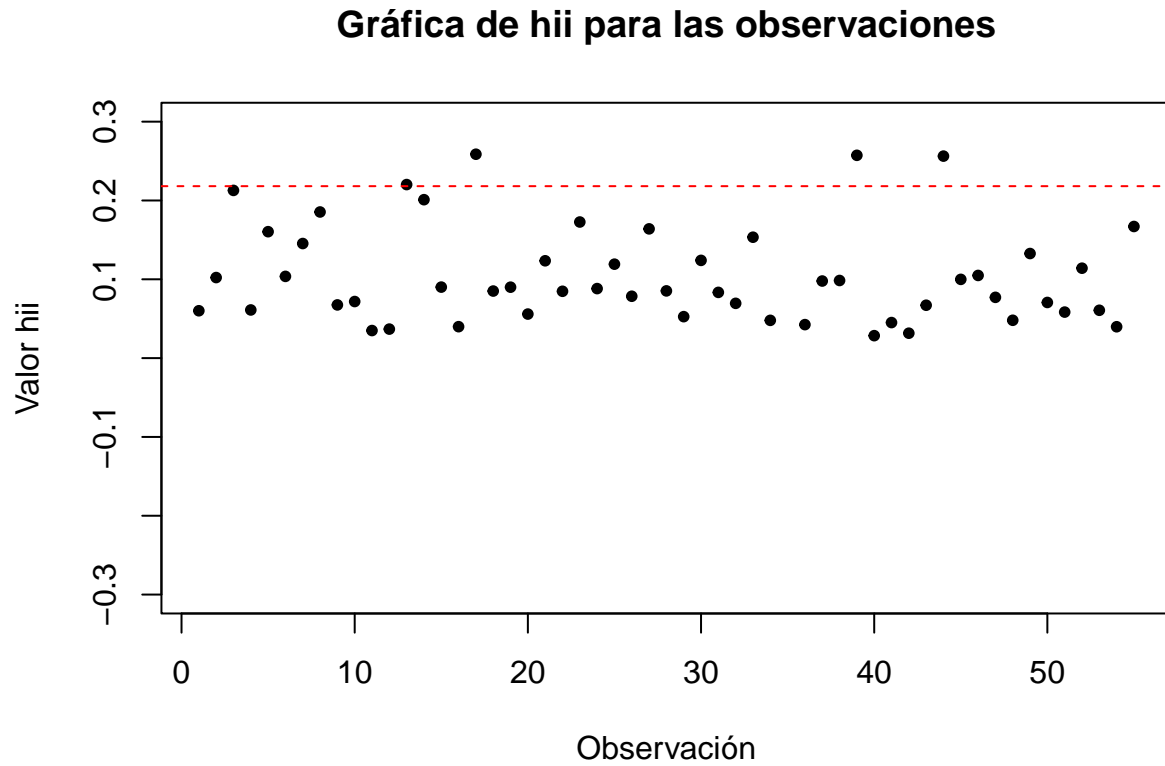


Figura 4: Identificación de puntos de balanceo

Al observar la gráfica de observaciones vs valores  $h_{ii}$ , donde la línea punteada roja representa el valor  $h_{ii} = 2\frac{p}{n}$ , se puede apreciar que existen 5 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 influyentes

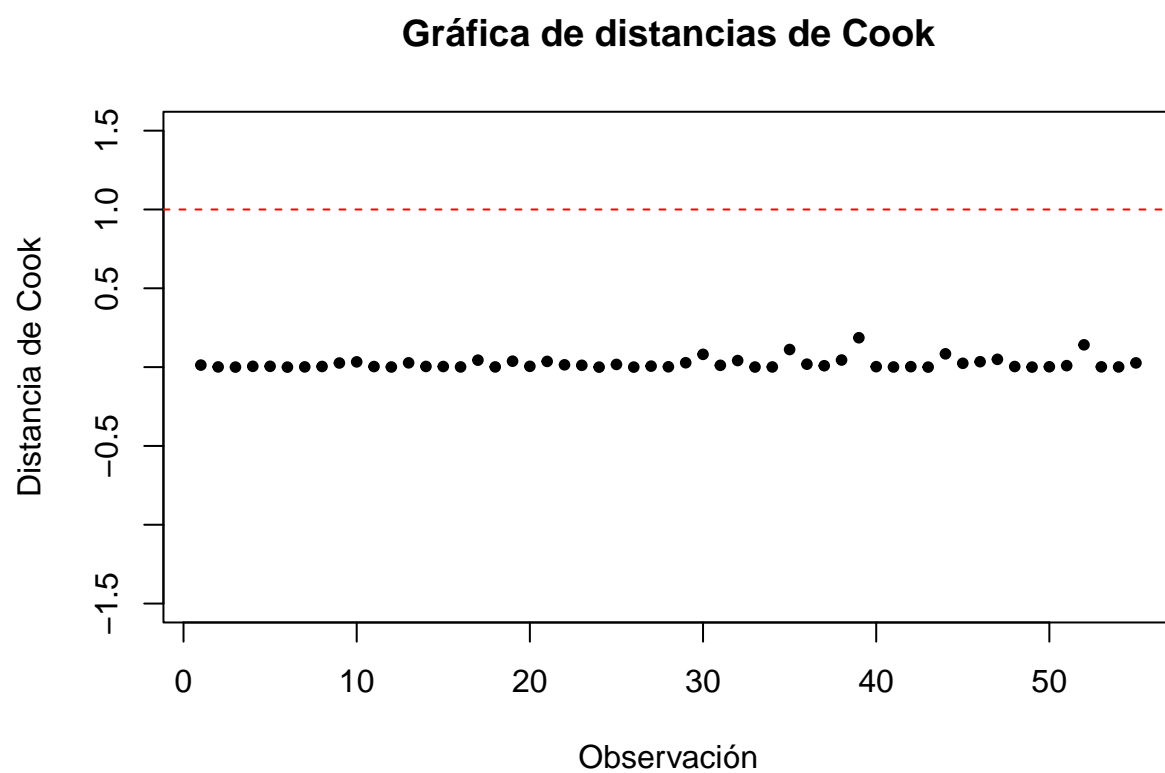


Figura 5: Criterio distancias de Cook para puntos influyentes

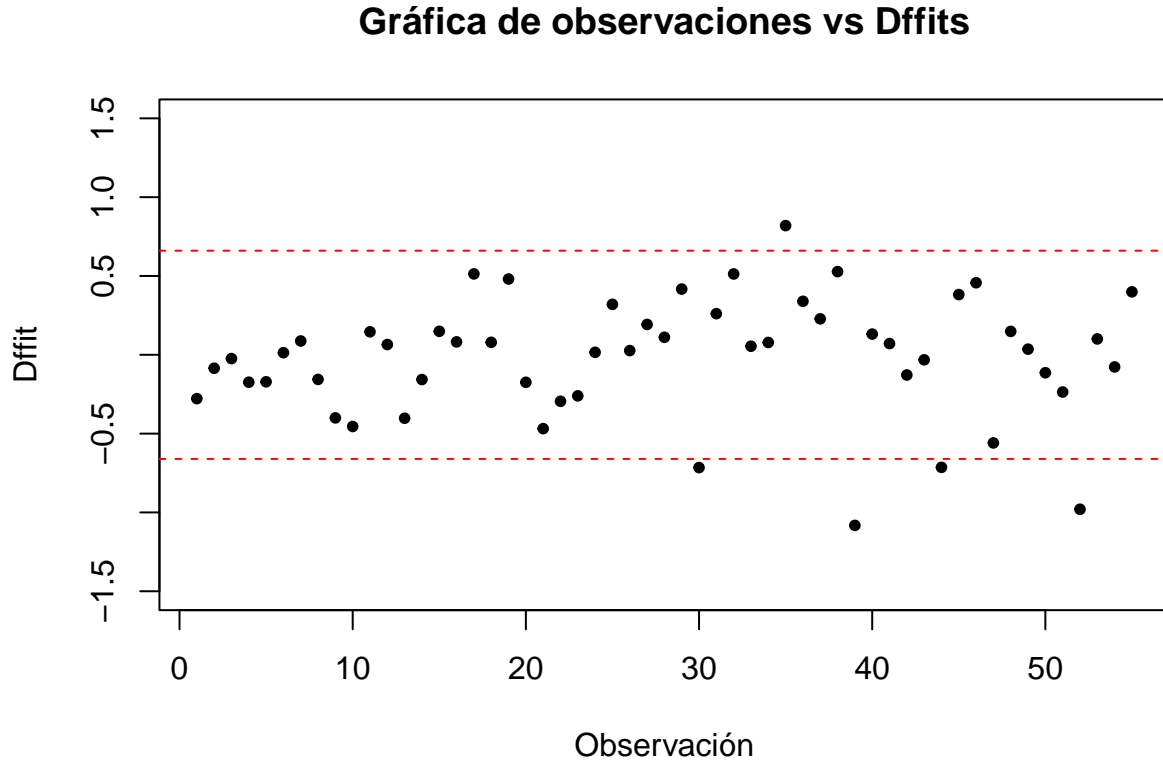


Figura 6: Criterio Dffits para puntos influyentes

##	res.stud	Cooks.D	hii.value	Dffits
## 30	-1.8533	0.0810	0.1240	-0.7155
## 35	1.0628	0.1116	0.3722	0.8195
## 39	-1.7957	0.1862	0.2573	-1.0824
## 44	-1.2104	0.0842	0.2563	-0.7141
## 52	-2.5658	0.1414	0.1141	-0.9798

Como se puede ver, las observaciones ... son puntos influyentes según el criterio de Dffits, el cual dice que para cualquier punto cuyo  $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$ , es un punto influyente. Cabe destacar también que con el criterio de distancias de Cook, en el cual para cualquier punto cuya  $D_i > 1$ , es un punto influyente, ninguno de los datos cumple con serlo.

### 4.3. Conclusión