

# **Trabajo 1**

Estudiantes

**Alexander Bustos Betancur**  
**David Felipe Ruiz Figueroa**  
**Yuliana Vanessa Rodríguez Posada**

Equipo 61

Docente

**Javier Armando Lozano Rodriguez**

Asignatura

**Estadística II**



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

Sede Medellín  
5 de octubre de 2023

# Índice

<b>1. Pregunta 1</b>	<b>3</b>
1.1. Modelo de regresión . . . . .	3
1.2. Significancia de la regresión . . . . .	4
1.3. Significancia de los parámetros . . . . .	4
1.4. Interpretación de los parámetros . . . . .	5
1.5. Coeficiente de determinación múltiple $R^2$ . . . . .	5
<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 . . . . .	6
<b>3. Pregunta 3</b>	<b>6</b>
3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . . .	6
3.2. Estadístico de prueba . . . . .	7
<b>4. Pregunta 4</b>	<b>7</b>
4.1. Supuestos del modelo . . . . .	7
4.1.1. Normalidad de los residuales . . . . .	7
4.1.2. Varianza constante . . . . .	9
4.2. Verificación de las observaciones . . . . .	10
4.2.1. Datos atípicos . . . . .	10
4.2.2. Puntos de balanceo . . . . .	11
4.2.3. Puntos influyentes . . . . .	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

# 1. Pregunta 1

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i, \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); 1 \leq i \leq 69$$

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
$\beta_0$	-0.2449
$\beta_1$	0.1836
$\beta_2$	0.0208
$\beta_3$	0.0689
$\beta_4$	0.0044
$\beta_5$	0.0016

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.2449 + 0.1836X_{1i} + 0.0208X_{2i} + 0.0689X_{3i} + 0.0044X_{4i} + 0.0016X_{5i}; 1 \leq i \leq 69$$

## 1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \stackrel{H_0}{\sim} f_{5,63} \quad (1)$$

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	63.7619	5	12.752383	20.867	3.08815e-12
Error	38.5010	63	0.611127		

De la tabla Anova se obtienen los valores del estadístico de prueba  $F_0 = 20.867$  y su correspondiente valor-P,  $Vp = 3.08815e - 12$ . Como  $Vp < 0.05$  se rechaza  $H_0$  en la que  $\beta_j = 0$  con  $0 \leq j \leq 5$  y se concluye que el modelo de regresión múltiple es significativa, aceptamos la hipótesis alternativa en la que algún  $\beta_j \neq 0$ . Por lo cual la probabilidad promedio estimada de adquirir una infección en el hospital depende significativamente de al menos una de las predictoras del modelo.

## 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
$\beta_0$	-0.2449	1.2309	-0.1990	0.8429
$\beta_1$	0.1836	0.0624	2.9414	0.0046
$\beta_2$	0.0208	0.0239	0.8696	0.3878
$\beta_3$	0.0689	0.0113	6.0984	0.0000
$\beta_4$	0.0044	0.0057	0.7697	0.4444
$\beta_5$	0.0016	0.0006	2.7616	0.0075

Los P-valores presentes en el Cuadro 3 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  $\alpha$ . Mientras que los parámetros  $\beta_0$ ,  $\beta_2$  y  $\beta_4$  son individualmente no significativos en presencia de los demás parámetros.

## 1.4. Interpretación de los parámetros

A continuación se hará una interpretación de los parámetros significativos teniendo en cuenta que se debe cumplir que el 0 esté en el intervalo.

$\hat{\beta}_1 = 0.1836$ : Este parámetro indica que por cada día que aumente la estadía del paciente en el hospital el promedio de adquirir una infección aumenta en 0.1836, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_3 = 0.0208$ : Este parámetro indica que por cada cama se instale de más en el hospital el promedio de infección aumenta un 0.0205.

$\hat{\beta}_5 = 0.0016$ : Esto indica que por cada enfermera adicional en el hospital, el promedio de infección aumenta en 0.0016.

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

El modelo tiene un coeficiente de determinación múltiple  $R^2 = SSR/SST = 63.7619/(63.7619 + 38.5010) = 0.6235$ , lo que significa que aproximadamente el 62.35 % de la variabilidad total observada en el riesgo de infección es explicado por el modelo de regresión propuesto en el presente informe.

El modelo tiene un coeficiente de determinación múltiple ajustado  $R_a^2 = 1 - ((n - 1)MSE)/SST = 1 - ((69 - 1)(0.611127)/(63.7619 + 38.5010)) = 0.59363$ , como el valor de  $R_a^2 = 0.59363$  es menor que  $R^2 = 0.6235$  podemos concluir que en el modelo de regresión lineal múltiple propuesto pueden haber variables predictorias que no aporten significativamente.

# 2. Pregunta 2

## 2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las 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:

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

Cuadro 4: Resumen tabla de todas las regresiones

	$SSE$	Covariables en el modelo				
Modelo completo	38.501	X1	X2	X3	X4	X5
Modelo reducido	80.626		X2	X4		

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon_i; \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); 1 \leq i \leq 69$$

## 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned}
 F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \dots, \beta_5))/3}{MSE(\beta_0, \dots, \beta_5)} \stackrel{H_0}{\sim} f_{3,63} \\
 &= \frac{14.0416}{0.61113} \\
 &= 22.9764
 \end{aligned} \tag{2}$$

Ahora, comparando el  $F_0$  con  $f_{0.95,3,63} = 2.7505$ , se puede ver que  $F_0 > f_{0.95,3,63}$  y por tanto se rechaza  $H_0$ .

Es posible o no descartar las variables del subconjunto? No, debido a que hay evidencia suficiente para que las variables  $X_1, X_3, X_5$  son significativamente diferentes de cero en el modelo reducido y no se pueden descartar del subconjunto.

## 3. Pregunta 3

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

Se quiere observar si el efecto de la duración promedio de la estadía de todos los pacientes dentro del hospital es igual al número promedio de camas que hay en el hospital durante el periodo de prueba, además, si el efecto del número promedio de pacientes en el hospital por día durante el estudio es igual al número promedio de enfermeras contratadas dentro del hospital. Para esto se hace la prueba de hipótesis que se muestra a continuación.

$$\begin{cases} H_0 : \beta_1 = \beta_3; \beta_4 = \beta_5 \\ H_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i}^* + \beta_2 X_{2i} + \beta_4 X_{4i}^* + \varepsilon$$

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

### 3.2. Estadístico de prueba

El estadístico de prueba  $F_0$  está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \stackrel{H_0}{\sim} f_{2,63} \quad (3)$$

$$F_0 = \frac{MSH}{MSE} = \frac{SSH/2}{MSE} = \frac{(SSE(MR) - 38.5010)/2}{0.611127} \stackrel{H_0}{\sim} f_{2,63} \quad (4)$$

Para poder realizar el cálculo completo es necesario conocer el valor de SSE(MR) el cual no se puede obtener de la tabla de todas las regresiones posibles, siendo así que se deja indicada la expresión con los valores conocidos.

## 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 que se realizará por medio de la prueba de shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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



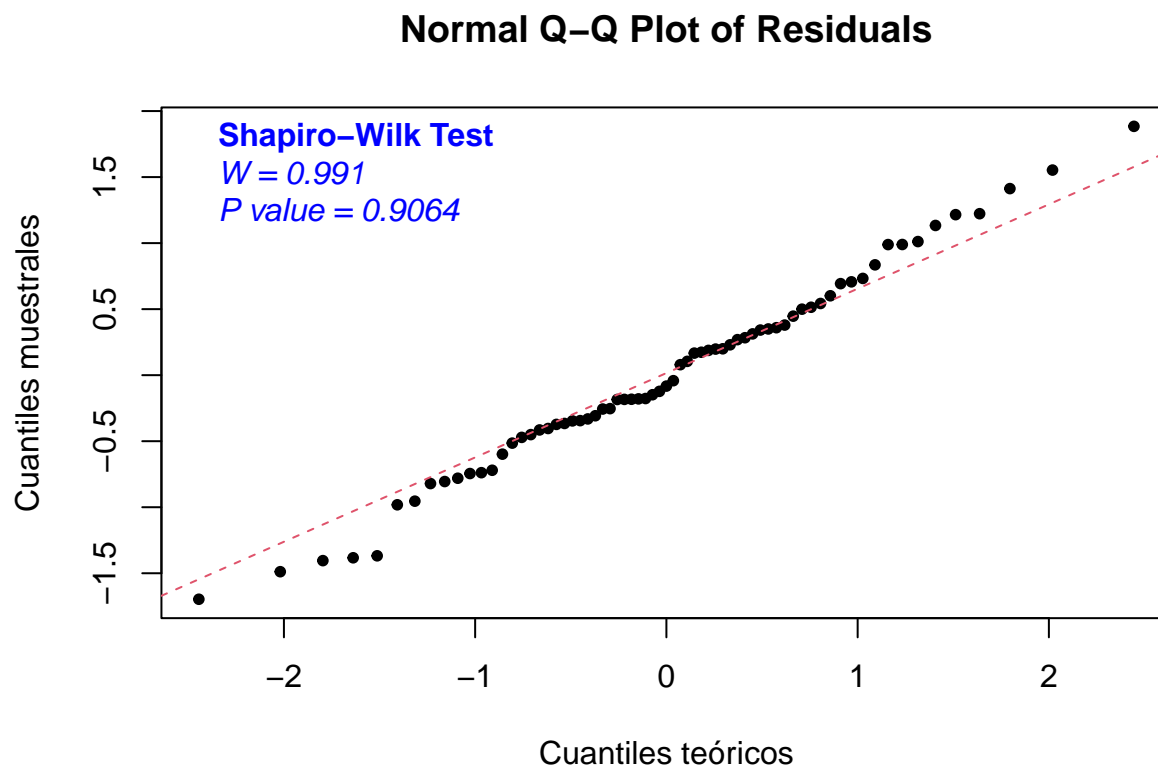


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Con la prueba de shapiro-wilk podemos observar que el P-valor es aproximadamente igual a 0.9064 y teniendo en cuenta que el nivel de significancia  $\alpha = 0.05$ , el P-valor es mucho mayor y por lo que podríamos llegar a decir no se rechaza la hipótesis nula, es decir que los datos distribuyen normal con media  $\mu$  y varianza  $\sigma^2$ . Pero si observamos detenidamente la gráfica, permite ver colas más pesadas y patrones irregulares, y el patrón no sigue la línea de ajuste siendo así que el supuesto de normalidad no se cumple. Al tener más poder el análisis gráfico, se termina por rechazar el cumplimiento de este supuesto.

#### 4.1.2. Varianza constante

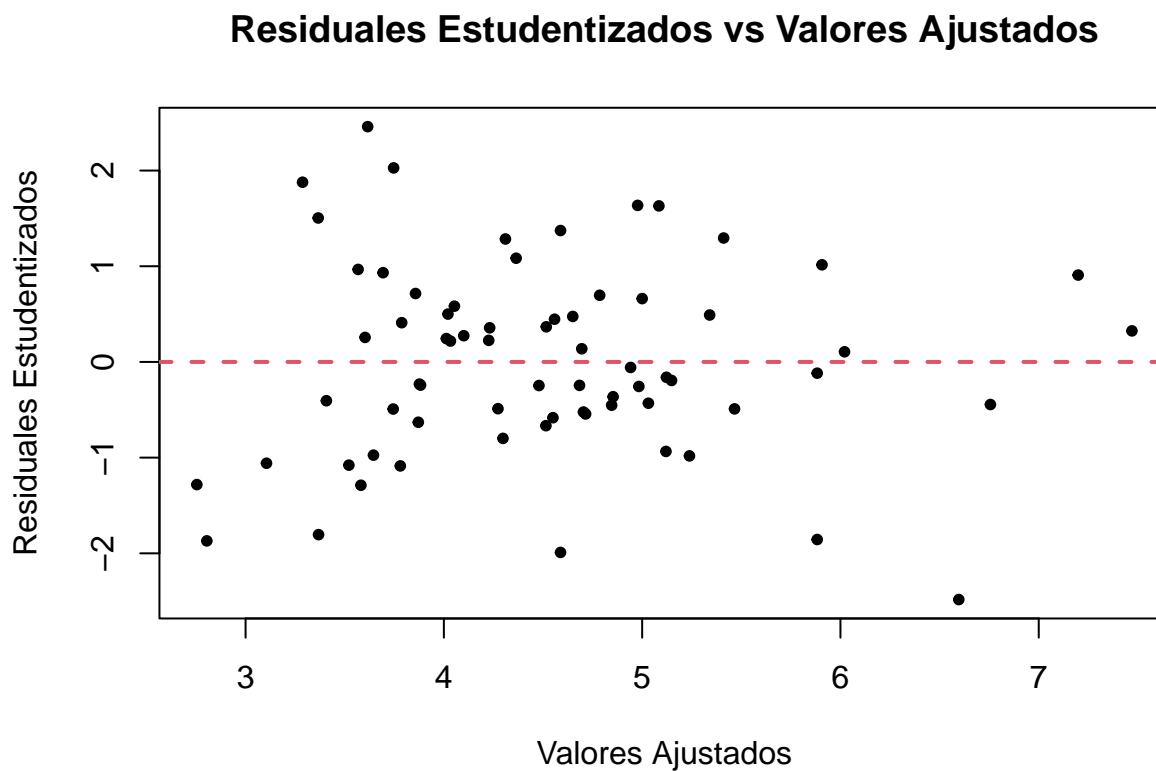


Figura 2: Gráfico residuales estudentizados vs valores ajustados

Por medio del 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. Por lo tanto podríamos decir que el modelo presenta homocedastidad.

## 4.2. Verificación de las observaciones

### 4.2.1. Datos atípicos

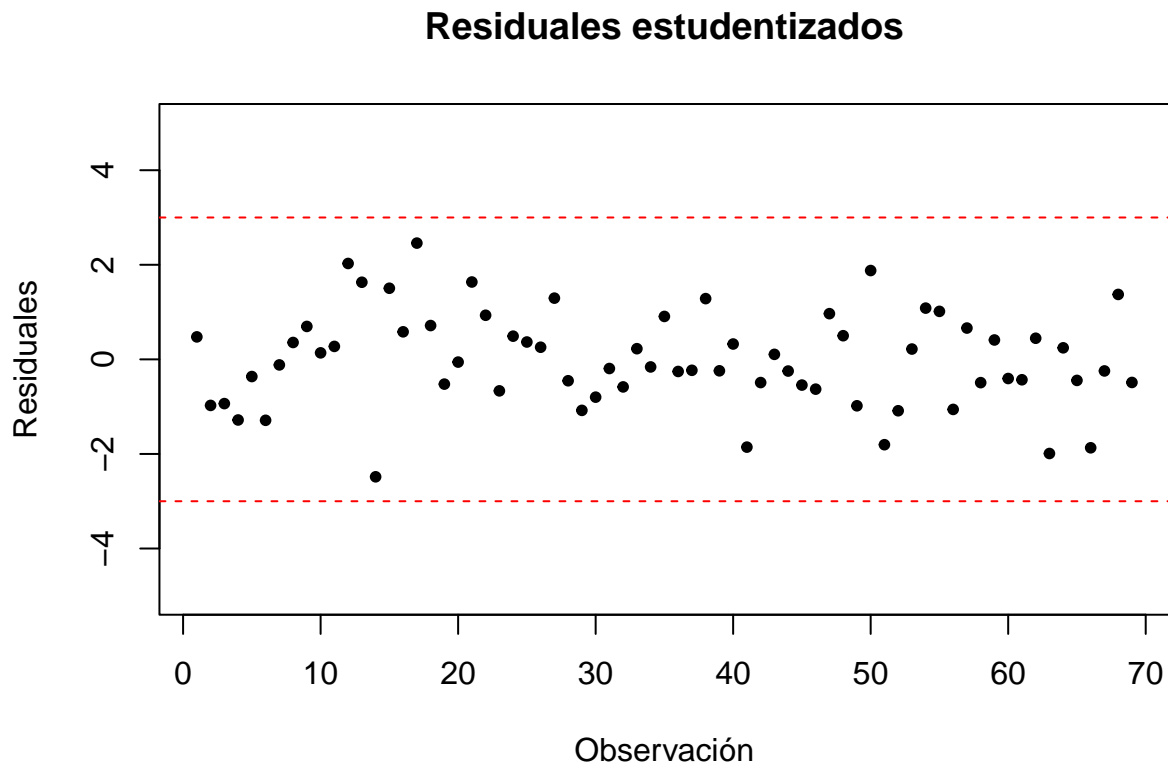


Figura 3: Identificación de datos atípicos

```
## [1] res.stud  Cooks.D   hii.value Dffits
## <0 rows> (or 0-length row.names)
```

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

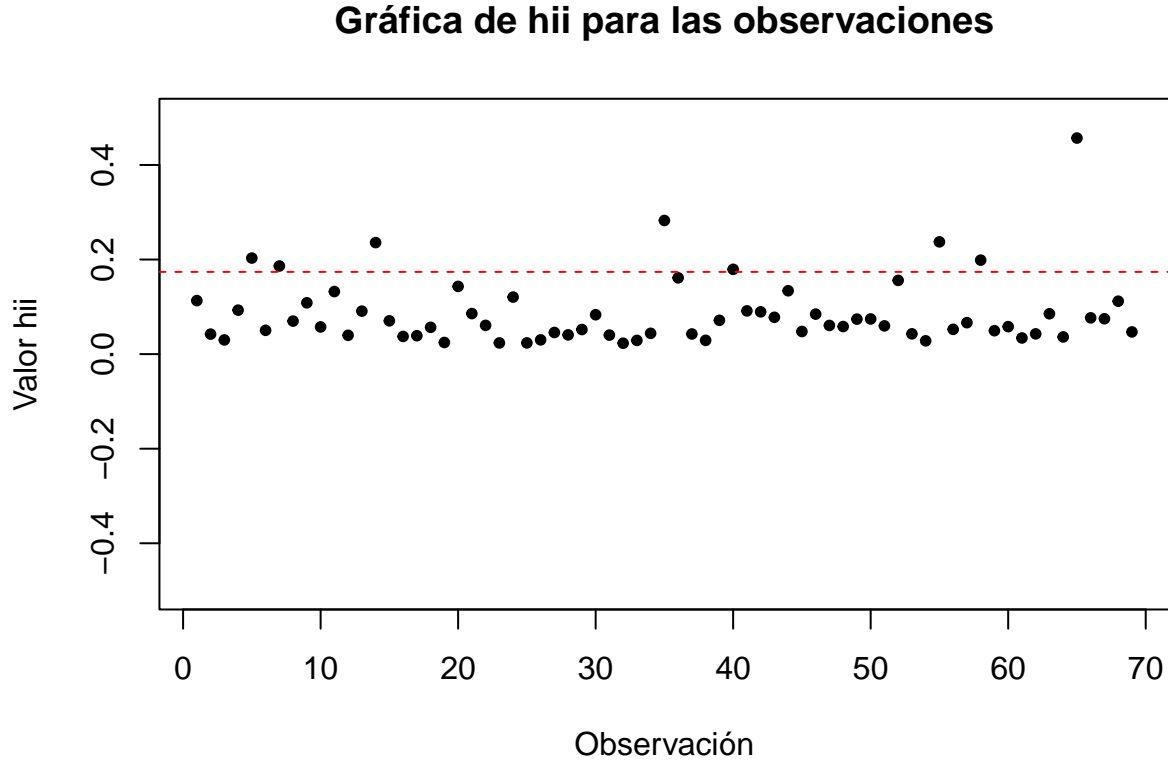


Figura 4: Identificación de puntos de balanceo

	res.stud	Cooks.D	hii.value	Dffits
5	-0.3633	0.0056	0.2031	-0.1821
7	-0.1175	0.0005	0.1863	-0.0558
14	-2.4832	0.3173	0.2359	-1.4412
35	0.9080	0.0541	0.2826	0.5690
40	0.3246	0.0038	0.1793	0.1506
55	1.0153	0.0535	0.2374	0.5666
58	-0.4922	0.0100	0.1987	-0.2436
65	-0.4452	0.0278	0.4569	-0.4057

Al observar la gráfica de observaciones vs valores  $h_{ii}$ , la línea punteada roja representa el valor  $h_{ii} = 2\frac{6}{69} = 0.17391$ , en ella se puede apreciar que existen 8 observaciones del conjunto que son puntos de balanceo según el criterio bajo el cual  $h_{ii} > 0.17391$ . Estos puntos de balanceo son los presentados en la tabla.

### 4.2.3. Puntos influyentes

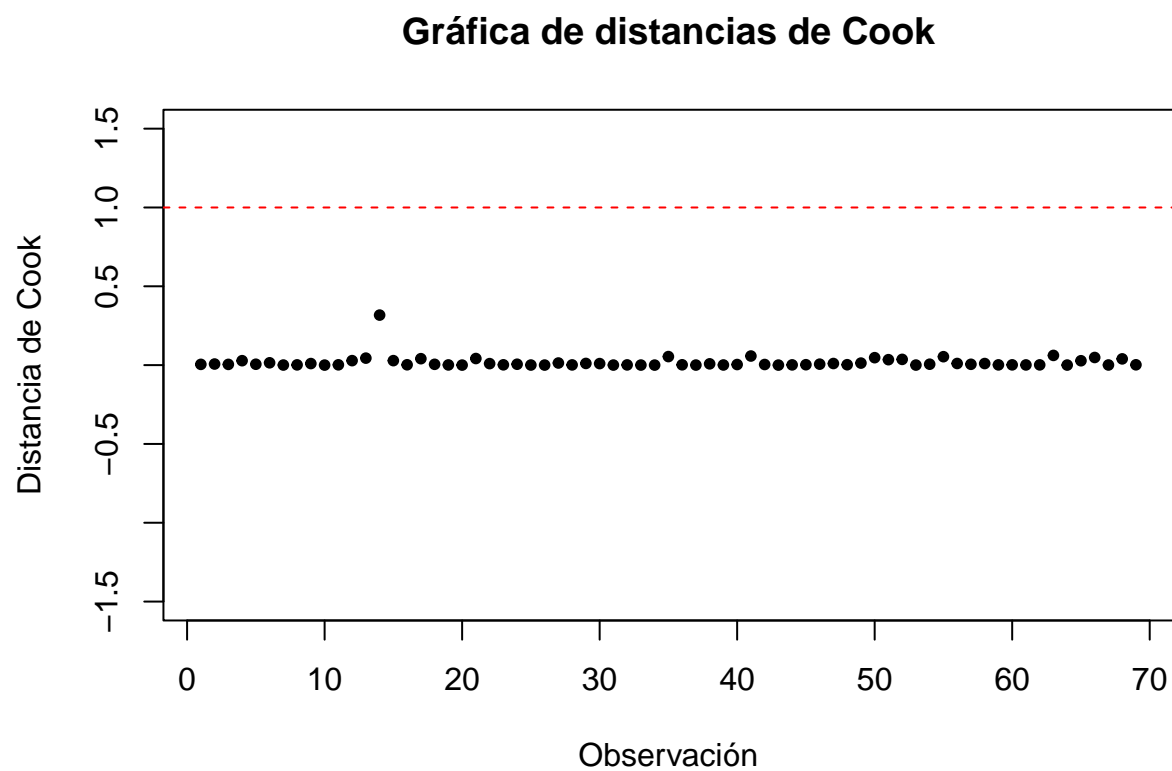


Figura 5: Criterio distancias de Cook para puntos influyentes

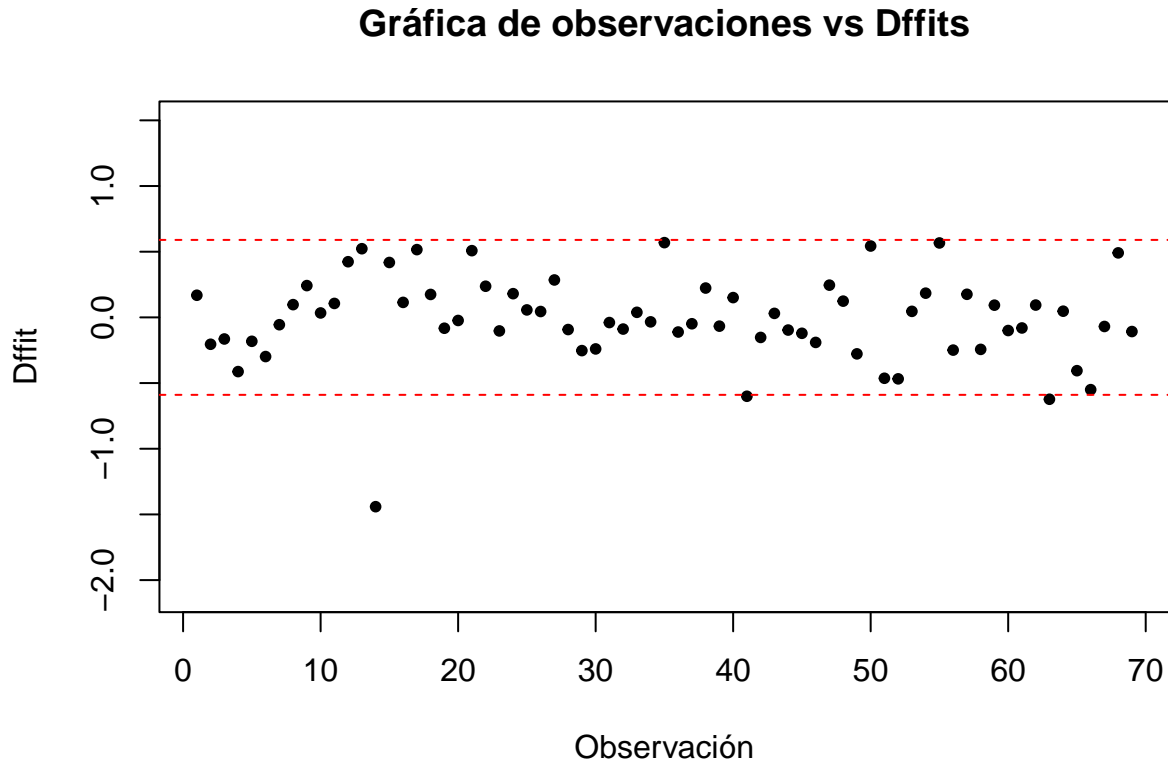


Figura 6: Criterio Dffits para puntos influyentes

##	res.stud	Cooks.D	hii.value	Dffits
## 14	-2.4832	0.3173	0.2359	-1.4412
## 41	-1.8558	0.0577	0.0914	-0.6005
## 63	-1.9907	0.0616	0.0853	-0.6231

Segun las dos graficas y la tabla de datos que tenemos podemos hacer el siguiente analisis de las observaciones:

Con el criterio de Cook, un punto es influyente cuando se cumple que  $D_i > 1$ , pero segun el análisis grafico y observando la tabla se puede concluir que, con este criterio, no hay puntos influyentes en el modelo.

Ahora, según el criterio de Dffits, si se cumple para cualquier punto que  $|D_{ffit}| > 2\sqrt{\frac{6}{69}} = 0.58977$ . Usando este criterio, y viendo la grafica junto con la tabla se puede observar que el modelo tiene tres puntos influyentes (14, 41, 63).

### 4.3. Conclusión

-Usando el coeficiente de determinación: El  $R^2$  nos indica la proporción de la muestra explicada por la regresión. Cuando un  $R^2$  está cerca de 1, se garantiza que el modelo se ajuste

bien a los datos. En este caso, se obtuvo un  $R$  cuadrado bastante alejado del 1  $R^2 = 0.6235$ , por lo cual se puede concluir que no es el mejor para ajustarse bien a la muestra de datos obtenida, aunque, sí es un modelo en donde la regresión explica en mayor parte los resultados.

-Usando los residuales: Usando la prueba de Shapiro-Wilk, el modelo cumple con el supuesto de normalidad, pero apreciando de manera gráfica, el patrón de residuales no sigue la línea de ajuste.

- En el modelo no se presentaron datos atípicos.
- Las observaciones 5, 7, 14, 35, 40, 55, 58, 65 son puntos de balanceo bajo el criterio de  $h_{ii} = 2 \frac{6}{69} = 0.17391$ .
- Los puntos 14, 41, 63 son puntos influyentes bajo el criterio de Diffs. Esto podría generar problemas de correlación entre las variables del modelo.

En conclusión, es un modelo que no se ajusta de la mejor manera a la muestra estudiada, es probable que los datos obtenidos con el, no sean los más precisos.