

# **Trabajo 1**

Estudiantes

**Juan Camilo Gutierrez Martinez**  
**Maria Fernanda Calle Agudelo**  
**Jaider Castañeda Villa**

Equipo 8

Docente

**Julieth Veronica Guarin Escudero**

Asignatura

**Estadística II**



Sede Medellín  
30 de marzo de 2023

# Índice

<b>1. Pregunta 1</b>	<b>3</b>
1.1. Modelo de regresion . . . . .	3
1.2. Significancia de la regresión . . . . .	3
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>6</b>
2.1. Planteamiento pruebas de hipótesis y modelo reducido . . . . .	6
2.2. Estadístico de prueba y conclusión . . . . .	6
<b>3. Pregunta 3</b>	<b>7</b>
3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . . .	7
3.2. Estadístico de prueba . . . . .	7
<b>4. Pregunta 4</b>	<b>8</b>
4.1. Supuestos del modelo . . . . .	8
4.1.1. Normalidad de los residuales . . . . .	8
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 influenciales . . . . .	12
4.3. Conclusión . . . . .	14

## 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.	Valores coeficientes . . . . .	3
2.	Tabla ANOVA para el modelo . . . . .	4
3.	Resumen de los coeficientes . . . . .	5
4.	Resumen tabla de todas las regresiones posibles . . . . .	6
5.	Puntos de balanceo . . . . .	11
6.	Puntos influenciales . . . . .	13

# 1. Pregunta 1

Se toma la base de datos 8, en la cual hay 5 variables regresoras, denominadas como:

$Y$ : 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

A partir de ello se plantea el siguiente modelo inicial:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2); 1 \leq i \leq 65$$

## 1.1. Modelo de regresion

Al realizar el ajuste al modelo con el fin de obtener la relacion de la variable respuesta con las variables regresoras obtenemos los siguientes coeficientes respectivamente:

Cuadro 1: Valores coeficientes

	Valor del parametro
$\beta_0$	-0.7194
$\beta_1$	0.0986
$\beta_2$	0.0274
$\beta_3$	0.0628
$\beta_4$	0.0146
$\beta_5$	0.0024

La ecuacion de la regresion ajustada es:

$$\hat{Y}_i = -0.7194 + 0.0986X_{1i} + 0.0274X_{2i} + 0.0628X_{3i} + 0.0146X_{4i} + 0.0024X_{5i}; 1 \leq i \leq 65$$

## 1.2. Significancia de la regresión

Se realizara un análisis de varianza para probar la significancia de los parametros, el cual se establece con el siguiente juego de hipotesis:

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \stackrel{H_0}{\sim} f_{5,59} \quad (1)$$

Se hace la comparacion con la distribucion  $f$  debido a que  $F_0$  es un analisis de significancia global del modelo, o sea, que tanto cambia el modelo segun las variables regresoras, la distribucion  $f$  es un analisis de varianza segun la cantidad de parametros y datos, si  $f \leq F_0$  significa que el modelo tiene una varianza mayor, por lo que, tiene alguna relacion con al menos una de las variables regresoras.

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	Grados de libertad	Cuadrado medio	$F_0$	P-valor
Regresión	57.1974	5	11.439480	11.553	8.42693e-08
Error	58.4204	59	0.990177		

De la tabla Anova, se observa un valor P casi igual a 0, lo que permite rechazar la hipótesis nula en la que  $\beta_j = 0$  con  $1 \leq j \leq 5$ , aceptando la hipótesis alternativa en la que algún  $\beta_j \neq 0$ , esto nos dice que hay al menos una relacion entre las variable respuesta y las regresoras permitiendonos asi concluir la significancia del modelo.

### 1.3. Significancia de los parámetros

Antes se realizo una prueba general del modelo con el fin de saber si nos proporcionaba alguna informacion, ahora se realizara una prueba de hipotesis sobre los coeficientes individuales del modelo con el fin de saber cuales son significativos o no, se establece primero el juego de hipotesis:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \quad j = 1, 2, \dots, 5 \end{cases}$$

El estadistico es el siguiente:

$$T_{j,0} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \stackrel{H_0}{\sim} t_{59} \quad (2)$$

En el siguiente cuadro se presenta información de los parámetros:

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	$T_{0j}$	P-valor
$\beta_0$	-0.7194	1.5430	-0.4662	0.6428
$\beta_1$	0.0986	0.0791	1.2462	0.2176
$\beta_2$	0.0274	0.0286	0.9580	0.3419
$\beta_3$	0.0628	0.0153	4.1097	0.0001
$\beta_4$	0.0146	0.0075	1.9512	0.0558
$\beta_5$	0.0024	0.0008	2.9449	0.0046

Usando el criterio de rechazo del valor-P  $\alpha > Val - P$  determinaremos que valores son significativos y cuales no, dado que no nos dan un valor especifico para  $\alpha$  diremos que  $\alpha = 0.05$  viendo la tabla solo hay dos valores significativos, o sea, que rechazamos su hipotesis nula que son  $\beta_3$  y  $\beta_5$  ya que sus P-valores son menores que  $\alpha$ ,  $\beta_0$  no hubiera sido interpretable en caso de que fuera significativa debido a que ninguna de las  $X_{j,i}$  contiene al 0 en sus datos.

#### 1.4. Interpretación de los parámetros

$\hat{\beta}_3$ : Significa que por cada unidad que aumente  $X_3$  el promedio en el riesgo de infección aumenta en 0.0628 unidades cuando las demas variables se mantienen constantes, esto en otras palabras es que a medida que hayan mas camas mas aumenta la media del riesgo de infeccion.

$\hat{\beta}_5$ : Significa que por cada unidad que aumente  $X_5$  el promedio en el riesgo de infección aumenta en 0.0024 unidades cuando las demas variables se mantienen constantes, aqui nos dice que segun el aumento de la cantidad de enfermeras aumenta la media del riesgo de infeccion.

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

El modelo tiene un coeficiente de determinación múltiple  $R^2 = 0.4947$ , lo que significa que aproximadamente el 49.47% de la variabilidad de  $Y$  es explicada por el modelo de regresion ajustado debido a las variables independientes, el resto de la variabilidad es explicada por la variabilidad residual, o sea,  $1 - R^2$ .

## 2. Pregunta 2

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

Según el Cuadro 3, las covariables que tienen el valor-P más alto en el modelo son  $X_1, X_2, X_4$ . Por medio de la tabla de todas las regresiones posibles se quiere hacer la siguiente prueba de hipótesis que permita concluir si el subconjunto de variables es significativo:

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

Cuadro 4: Resumen tabla de todas las regresiones posibles

	Suma cuadratica de error	Covariables en el modelo
Modelo completo	58.420	X1 X2 X3 X4 X5
Modelo reducido	69.028	X3 X5

Un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon; \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); 1 \leq i \leq 65$$

### 2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \dots, \beta_5))/3}{MSE(\beta_0, \dots, \beta_5)} \stackrel{H_0}{\sim} f_{3,59} \\ &= \frac{[69.028 - 58.420]/3}{0.990177} \\ &= 3.5711 \end{aligned} \tag{3}$$

Usando una significancia de  $\alpha = 0.05$ :

Si se compara el  $F_0$  con  $f_{0.95,3,59} = 2.7608$ , se puede ver que  $F_0 > f_{0.95,3,59}$ .

Como  $F_0 > f_{0.95,3,59}$ , entonces se rechaza  $H_0$ , por lo tanto, el subconjunto es significativo, en presencia de los demás parámetros.

Por lo anterior, llegamos a la conclusión que las variables no se pueden descartar del modelo porque el riesgo promedio de infección depende de al menos una de las variables presentes en el subconjunto.

### 3. Pregunta 3

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

Queremos probar si:

$$2\beta_1 = \beta_2; \quad 5\beta_3 = \beta_4; \quad \beta_4 = \beta_5$$

Para esto tenemos la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : 2\beta_1 = \beta_2; \quad 5\beta_3 = \beta_4; \quad \beta_4 = \beta_5 \\ H_1 : 2\beta_1 \neq \beta_2 \text{ ó } 5\beta_3 \neq \beta_4 \text{ ó } \beta_4 \neq \beta_5 \end{cases}$$

Podemos reescribirlas de la siguiente manera:

$$\begin{cases} H_0 : 2\beta_1 - \beta_2 = 0; \quad 5\beta_3 - \beta_4 = 0; \quad \beta_4 - \beta_5 = 0 \\ H_1 : 2\beta_1 - \beta_2 \neq 0; \quad 5\beta_3 - \beta_4 \neq 0; \quad \beta_4 - \beta_5 \neq 0 \end{cases}$$

Y ahora en términos matriciales:

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

Donde la matriz  $\mathbf{L}$  está dada por:

$$L = \begin{bmatrix} 0 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Para obtener el modelo reducido operamos:

$$Y_i = \beta_0 + \beta_1 X_{1i} + 2\beta_1 X_{2i} + \beta_3 X_{3i} + 5\beta_3 X_{4i} + 5\beta_3 X_{5i} + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); \quad 1 \leq i \leq 65$$

Agrupando, el MR estará dado por:

$$Y_i = \beta_0 + \beta_1 X_{1,2i}^* + \beta_3 X_{3,4,5i}^* + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2); \quad 1 \leq i \leq 65$$

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

#### 3.2. Estadístico de prueba

El estadístico de prueba  $F_0$  es el siguiente:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/3}{MSE(MF)} \stackrel{H_0}{\sim} f_{3,59} \quad (4)$$

Reemplazando el  $SSE(MF)$  y el  $MSE(MF)$  conocidos:

$$F_0 = \frac{(SSE(MR) - 58.420)/3}{0.990177} \stackrel{H_0}{\sim} f_{3,59} \quad (5)$$



## 4. Pregunta 4

### 4.1. Supuestos del modelo

#### 4.1.1. Normalidad de los residuales

Para validar el supuesto de normalidad, se plante la prueba de hipótesis Shapiro-Wilk, seguida de un gráfico cuantil-cuantil:

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

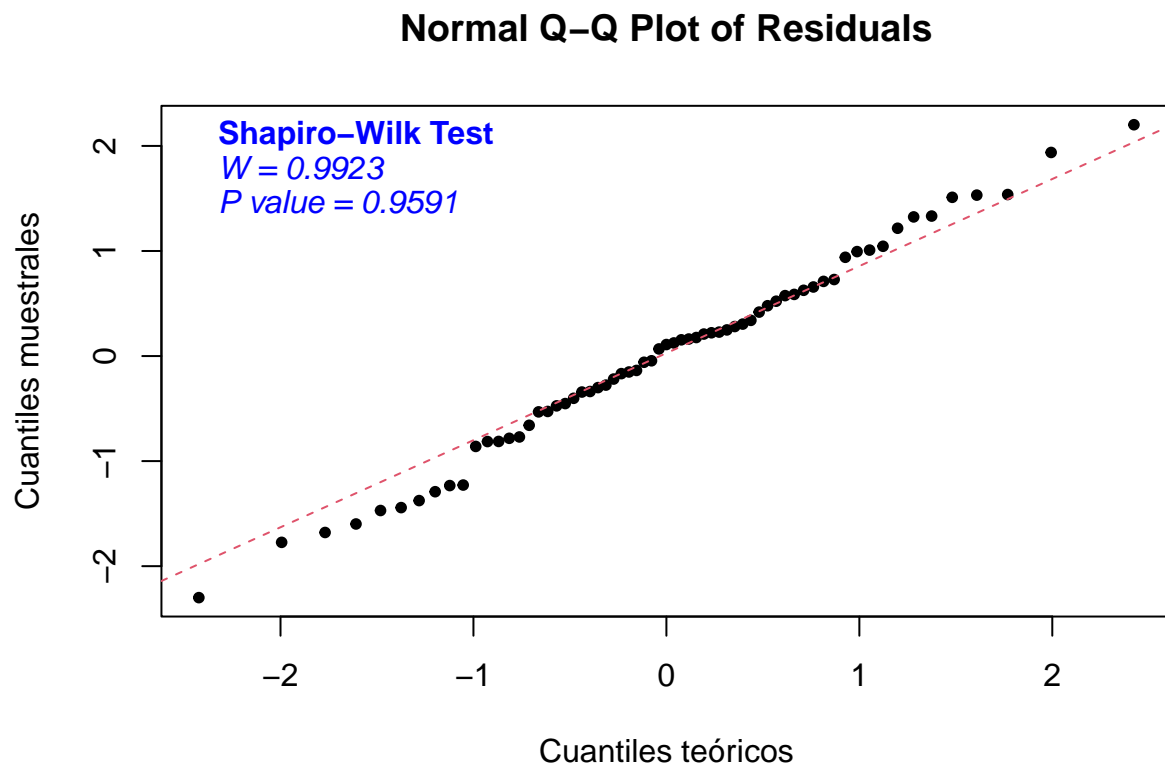


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Analizando en un principio el valor-P de la prueba Shapiro-Wilk utilizando un nivel de significancia  $\alpha = 0.05$ , podemos observar que este valor-P es superior pues  $0.9591 > 0.05$ , por lo que no se cumple el criterio de rechazo, no rechazaríamos la hipótesis nula y se cumpliría el supuesto de normalidad. No obstante, en la gráfica de cuantiles podemos ver que en las colas, hay bastantes datos que no se alinean a la línea punteada roja, por lo que, dándole más peso a la prueba gráfica se rechaza el cumplimiento del supuesto.

#### 4.1.2. Varianza constante

Para validar el supuesto de varianza constante analizaremos el gráfico de los residuales estudentizados vs los valores ajustados:

$$\begin{cases} H_0 : V[\varepsilon_i] = \sigma^2 \\ H_1 : V[\varepsilon_i] \neq \sigma^2 \end{cases}$$

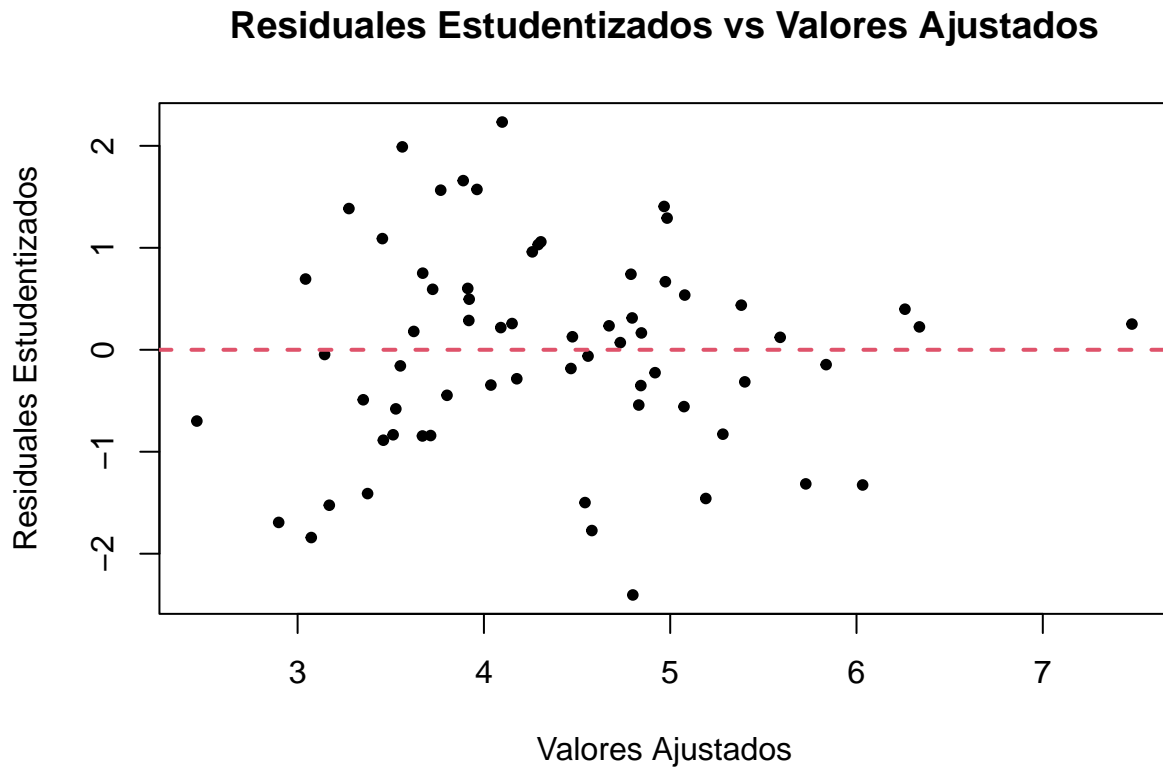


Figura 2: Gráfico residuales estudentizados vs valores ajustados

Del análisis de la gráfica encontramos que el patrón de los puntos indica un aumento de la dispersión hasta poco antes del centro de la gráfica y después hay un decrecimiento de la dispersión. Si nos paramos en puntos de las x vemos que en algunos hay diferentes amplitudes en la nube de puntos. En base a esto podemos concluir que el supuesto de varianza constante no se cumple. Es posible que algunas observaciones extremas estén afectando nuestro análisis.

## 4.2. Verificación de las observaciones

### 4.2.1. Datos atípicos

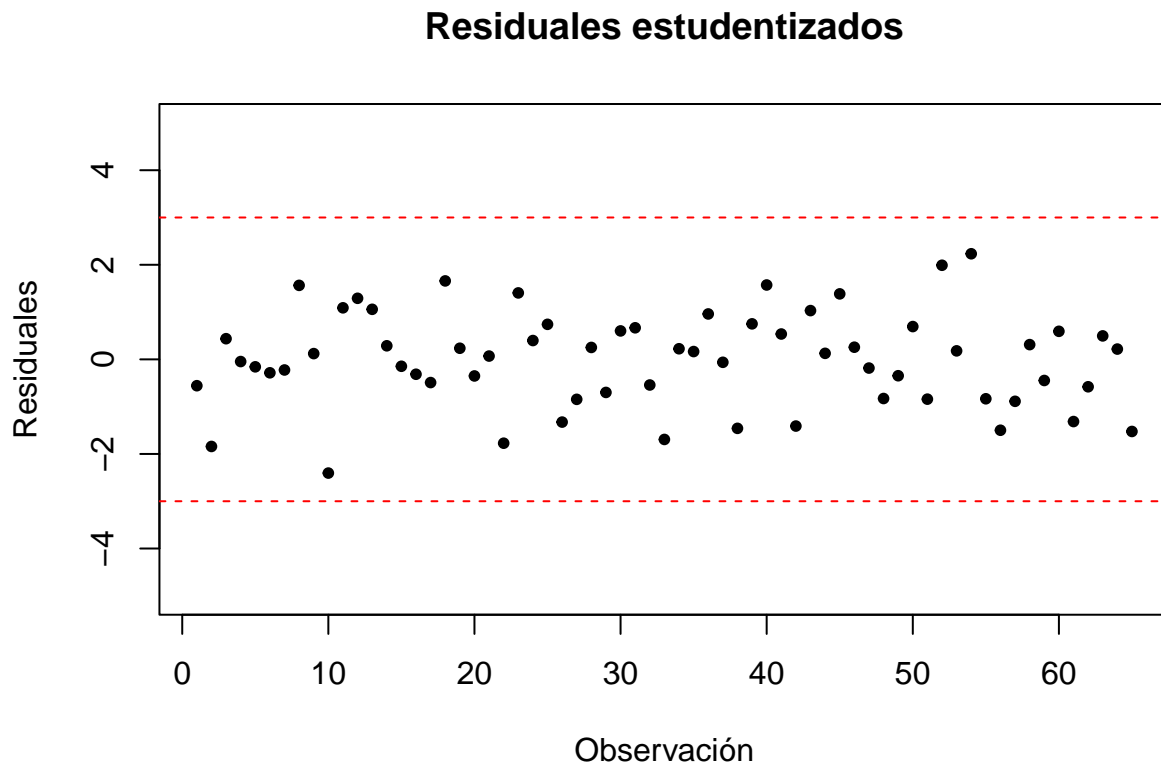


Figura 3: Identificación de datos atípicos

Las observaciones atípicas son aquellas que se encuentran separadas del resto de las observaciones (En el espacio de las  $Y$ ) y por tanto puede afectar los resultados del ajuste del modelo de regresión. Tal como nos los muestra la gráfica anterior, no se observa ningún dato atípico en el conjunto de datos que tenemos, esto porque ningún residual estudentizado sobrepasa el criterio correspondiente  $|r_{estudentizados}| > 3$ .

#### 4.2.2. Puntos de balanceo

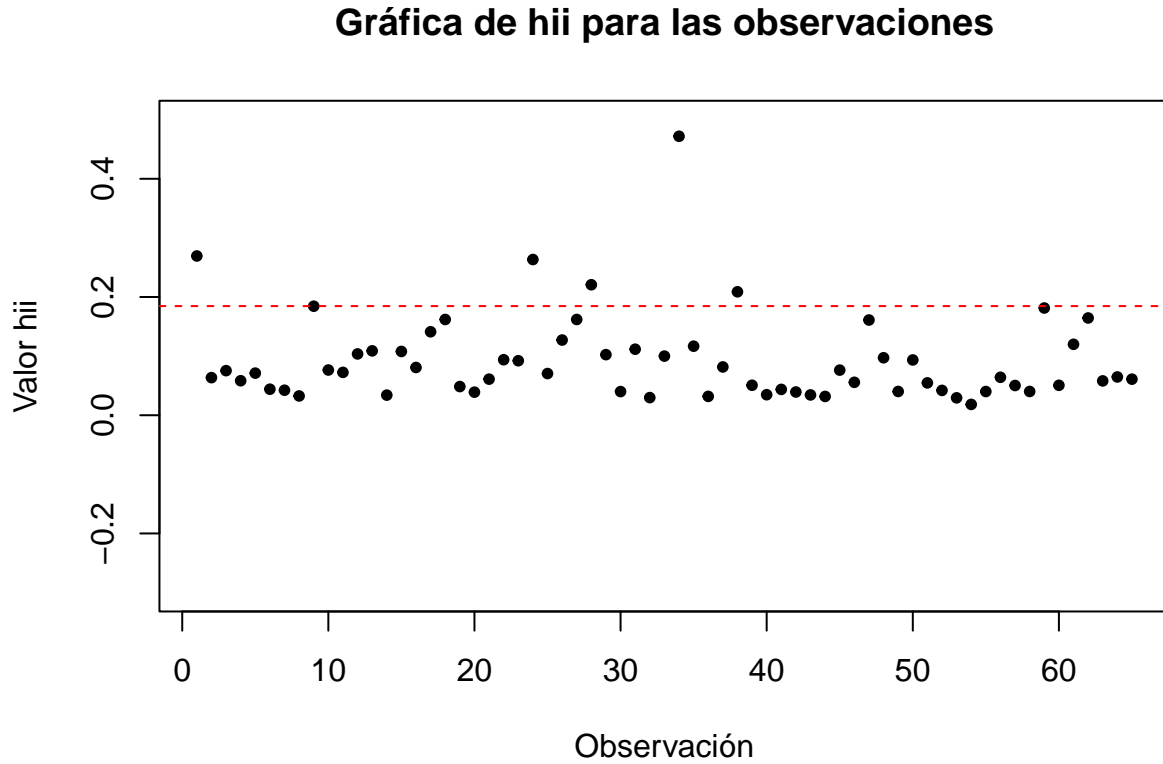


Figura 4: Identificación de puntos de balanceo

Cuadro 5: Puntos de balanceo

	Valores $h_{ii}$
1	0.2694
24	0.2634
28	0.2207
34	0.4719
38	0.2088

Analizando la gráfica, concluimos que tenemos 5 puntos de balanceo (los dos que están exactamente en la línea son menores al valor  $h_{ii} = 2\frac{p}{n} = 2\frac{6}{65} = 0.1846$  que representa la línea punteada roja). Las 5 observaciones 1, 24, 28, 34 y 38 ,como se muestra en la tabla, cumplen con el criterio correspondiente  $h_{ii} > 2\frac{p}{n}$ . Dichos puntos representan observaciones en el espacio de las predictoras, alejadas del resto de la muestra, por lo que pueden controlar

ciertas propiedades de nuestro modelo ajustado. Posiblemente puedan afectar el  $R^2$  y los errores estándar de los coeficientes estimados.

#### 4.2.3. Puntos influenciales

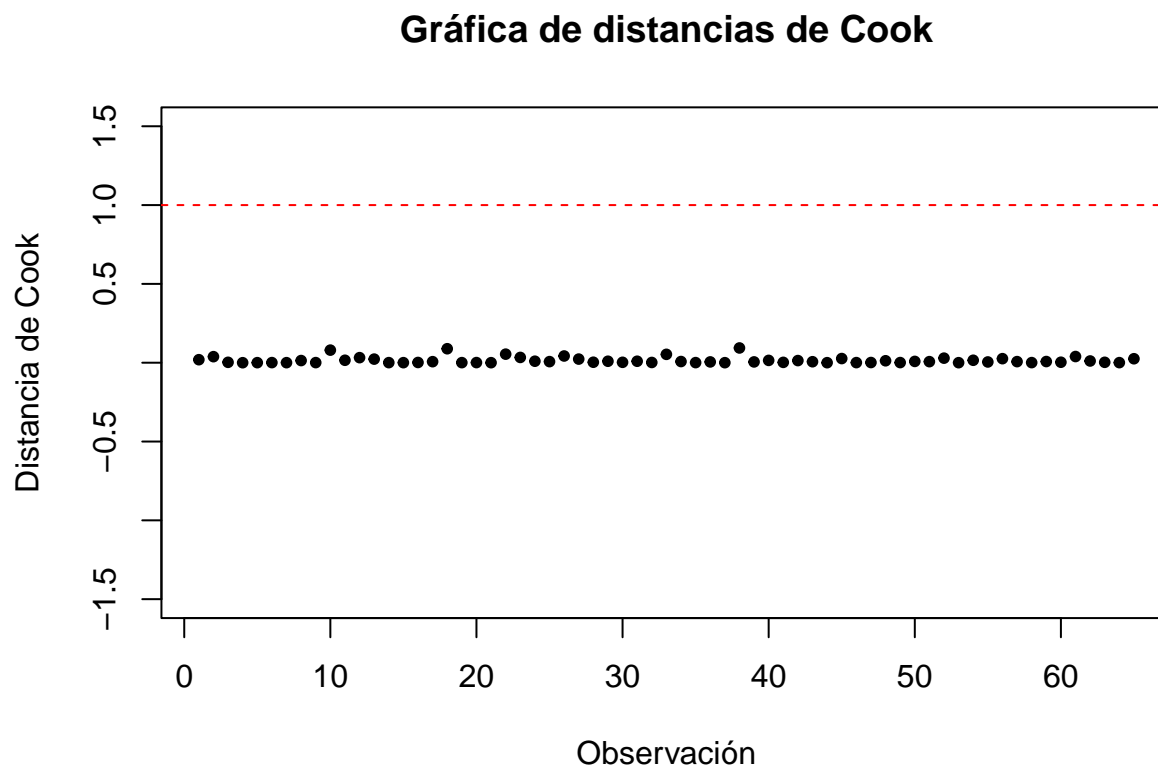


Figura 5: Criterio distancias de Cook para puntos influenciales

Las observaciones influenciales son aquellas que tienen un impacto notable sobre los coeficientes de regresión ajustados. Con el criterio Cook que nos indica que para que sea un punto influyente  $D_i > 1$ , observamos que ningún punto cumple, tal cual como se ve en la gráfica ya que ninguno sobrepasa la línea roja punteada.

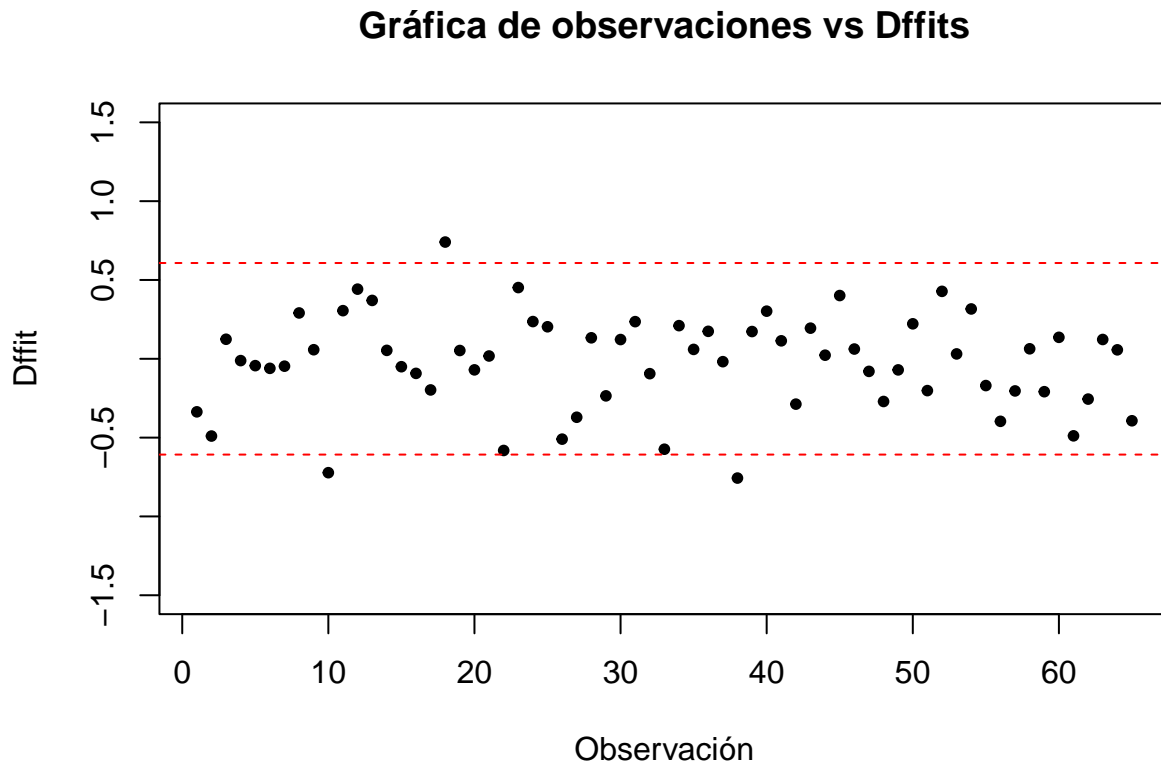


Figura 6: Criterio Dffits para puntos influyentes

Cuadro 6: Puntos influyentes

	Dffits
10	-0.7226
18	0.7406
38	-0.7567

Tal como nos lo muestra la gráfica y la tabla, tenemos 3 puntos influyentes, provenientes de la observación 10, 18 y 38, que cumplen el criterio del diagnóstico DFFITS el cual establece que es un punto influyente si  $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$ , para este conjunto de datos en particular  $|D_{ffit}| > 2\sqrt{\frac{6}{56}} = 0.6076$ . Teniendo en cuenta la definición de punto influyente podemos decir que la exclusión de alguno de dichos puntos del modelo puede causar cambios importantes en la ecuación de regresión ajustada.

### 4.3. Conclusión

Se llega a la conclusión que el modelo propuesto no es válido, pues no se cumple el supuesto que indica que los errores se distribuyen normal, y tampoco se cumple el supuesto de los errores con varianza constante.

Por esta razón, es inviable hablar de linealidad tal y como se presentaron los datos, a pesar de que los demás supuestos si tuvieron validez (errores con media cero y errores mutuamente independientes).

Es importante recalcar la posibilidad de que los puntos de balanceo e influencias mostrados anteriormente estén afectando en la validación de los supuestos. Por consiguiente, es necesario analizar estas observaciones individualmente, y realizar el análisis respectivo que permita explicar el por qué ocurrieron.