4,45

Trabajo 1

Estudiantes

Juan Camilo Berrío López
Maria José Herrera Arango
Juan Pablo Perez Arrubla
Susana Villa Vasquez

Equipo #26

Docente

Francisco Javier Rodríguez Cortés

Asignatura

Estadística II

Sede Medellín

30 de marzo de 2023

Índice

1. Pregunta 1 **3**

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 5

1.5. Coeficiente de determinación múltiple R^2 5

2. Pregunta 2 **5**

2.1. Planteamiento pruebas de hipótesis y modelo reducido 5

2.2. Estadístico de prueba y conclusión 6

3. Pregunta 3 **6**

3.1. Prueba de hipótesis y prueba de hipótesis matricial 6

3.2. Estadístico de prueba 7

4. Pregunta 4 **7**

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 influenciales 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 5
5. Valores h_{ii} para las observaciones 11
6. Valores DFFITS para las observaciones 13

1. Pregunta 1

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

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

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
β_0	-0.2579
β_1	0.1672
β_2	-0.0032
β_3	0.0610
β_4	0.0204
β_5	0.0018

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.2579 + 0.1672 X_{1i} - 0.0032 X_{2i} + 0.061 X_{3i} + 0.0204 X_{4i} + 0.0018 X_{5i};\ 1 \leqslant i \leqslant 45$$

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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,39} \tag{1}$$

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	70.719	5	14.143803	14.1915	6.3928e-08
Error	38.869	39	0.996641		

De la tabla Anova, se observa un valor P cercano a 0, esto hace que no haya la evidencia suficiente para aceptar 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 a la luz de los datos presentados se puede concluir que 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
β_0	-0.2579	1.8953	-0.1361	0.8924
β_1	0.1672	0.1078	1.5503	0.1292
β_2	-0.0032	0.0364	-0.0866	0.9314
β_3	0.0610	0.0202	3.0221	0.0044
β_4	0.0204	0.0087	2.3351	0.0248
β_5	0.0018	0.0010	1.6950	0.0980

Respecto a los valores P presentados en la tabla ANOVA se puede concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_3 y β_4 son significativos, pues sus P-valores son menores a α.

A continuación, se presenta la prueba de hipótesis utilizada para el parametro β_3.

$$\begin{cases} \mathrm{H}_0 : \beta_3 = 0 \\ \mathrm{H}_1 : \beta_3 \neq 0 \end{cases}$$

De la misma forma, para el parámetro β_4.

$$\begin{cases} \mathrm{H}_0 : \beta_4 = 0 \\ \mathrm{H}_1 : \beta_4 \neq 0 \end{cases}$$

1.4. Interpretación de los parámetros

$\hat{\beta}_3$**:** El número promedio de camas en el hospital aumenta la probabilidad de adquirir una infección durante la estadía en un 6,1 %, en cuanto las demás variables están fijas, ya que al haber más camas (disponibles u ocupadas), incrementa el riesgo de contagio entre las personas que se encuentren dentro del hospital.

$\hat{\beta}_4$**:** El número promedio de pacientes en el hospital por día, incrementa la probabilidad de contraer una infección intrahospitalaria en un 2,04 %, ya que al momento de haber más personas dentro de él se extenderá la infección de una manera más rápida. El aumento se da mientras las demás variables están fijas.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6453$, lo que significa que aproximadamente el 64.53 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe. Hay que tener cuidado a la hora de utilizar el R^2 no ajustado dado que puede llegar a inflar el coeficiente, es decir, podría esbozar una interpretación incorrecta respecto a la variabilidad total observada en el modelo.

¿cómo se calcula?

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_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	38.869	X1 X2 X3 X4 X5
Modelo reducido	50.767	X3 X4

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i \ ; \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 45$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_1, \beta_3, \beta_4) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,39} \\ &= \frac{(50.767 - 38.869)/3}{38.869/39} \\ &= 3.9793 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,39} = 2.8451$, se puede ver que $F_0 > f_{0.95,3,39}$. Por lo tanto, teniendo en cuenta los datos que tenemos, y a la luz de la prueba de hipótesis, poseemos información suficiente para rechazar la hipótesis nula, en consecuencia, se acepta la hipótesis alternativa, concluyendo entonces que el subconjunto es significativo.

En consecuencia, no se pueden descartar las variables implicadas dado que el subconjunto es distinto de cero, al menos una de ellas no es nula.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta de si el parámetro β_2 podría ser igual a cinco veces el parámetro β_5, así como también la comprobación de si el doble del parámetro β_1 podría ser igual a triple del parámetro β_4. Por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = 5\beta_5;\ 2\beta_1 = 3\beta_4 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

$$L = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & -5 \\ 0 & 2 & 0 & 0 & -3 & 0 \end{bmatrix}$$

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_3 X_{3i} + \beta_4 X^*_{4i} + \beta_5 X^*_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 38.869/2)}{38.869/39} \overset{H_0}{\sim} f_{2,39} \tag{3}$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para comprobar este supuesto, se deberá realizar una prueba de hipótesis ~~Shapiro-Wilk~~, en donde lograremos determinar si el conjunto de datos proviene de una distribución normal, además, se acompañará de un gráfico cuantil-cuantil y así afirmar o negar esta suposición:

$$\begin{cases} \mathrm{H}_0 : \varepsilon_i \sim \text{Normal} \\ \mathrm{H}_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$

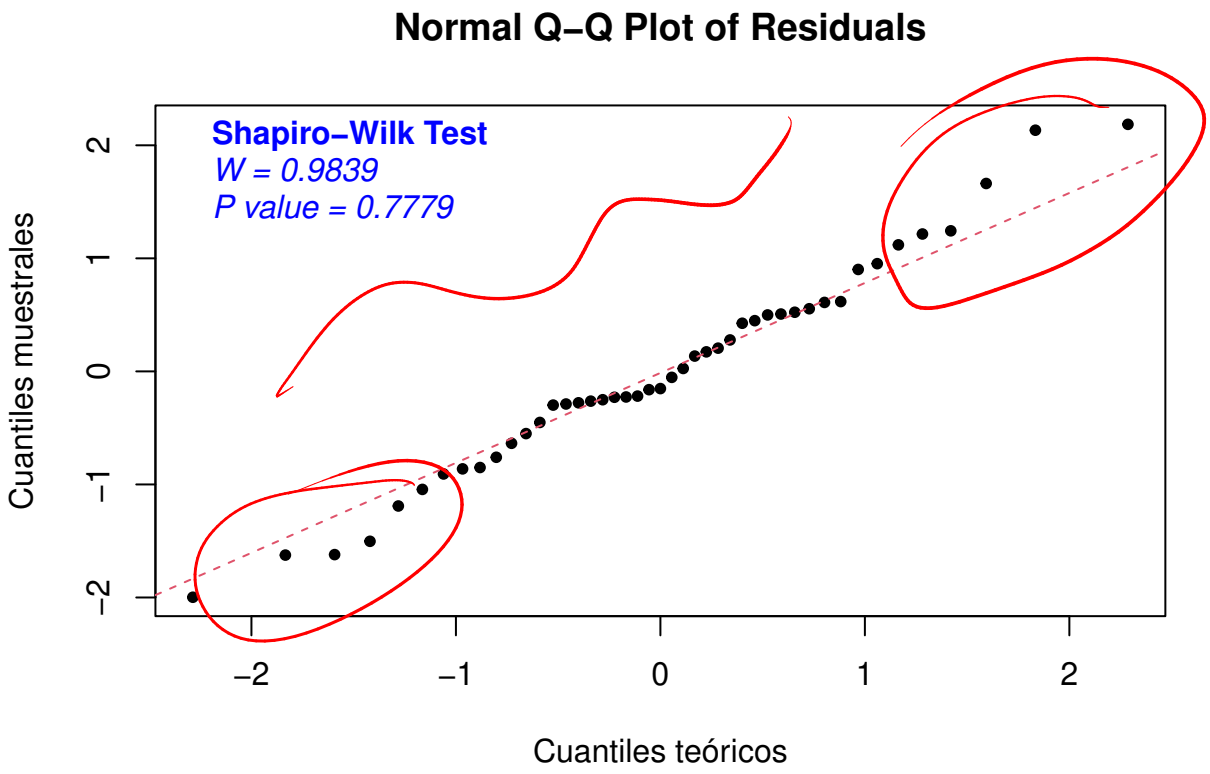

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

El valor P al ser aproximadamente igual a 0.9839 y teniendo en cuenta el nivel de significancia $\alpha = 0.05$, el valor P es mucho mayor y por lo tanto, a la luz de los resultados y con los datos que tenemos, no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal.

No obstante, tenemos que tener cuidado, porque observando la gráfica, el patrón de los residuales no sigue la línea roja que representa el ajuste de la distribución de los residuales a una distribución normal, notamos unos movimientos bruscos en las colas superior e inferior así como patrones irregulares en los datos. Por lo anterior, es plausible declarar que el supuesto de normalidad NO se cumple a pesar de que la prueba de hipótesis era apropiada.

4.1.2. Varianza constante

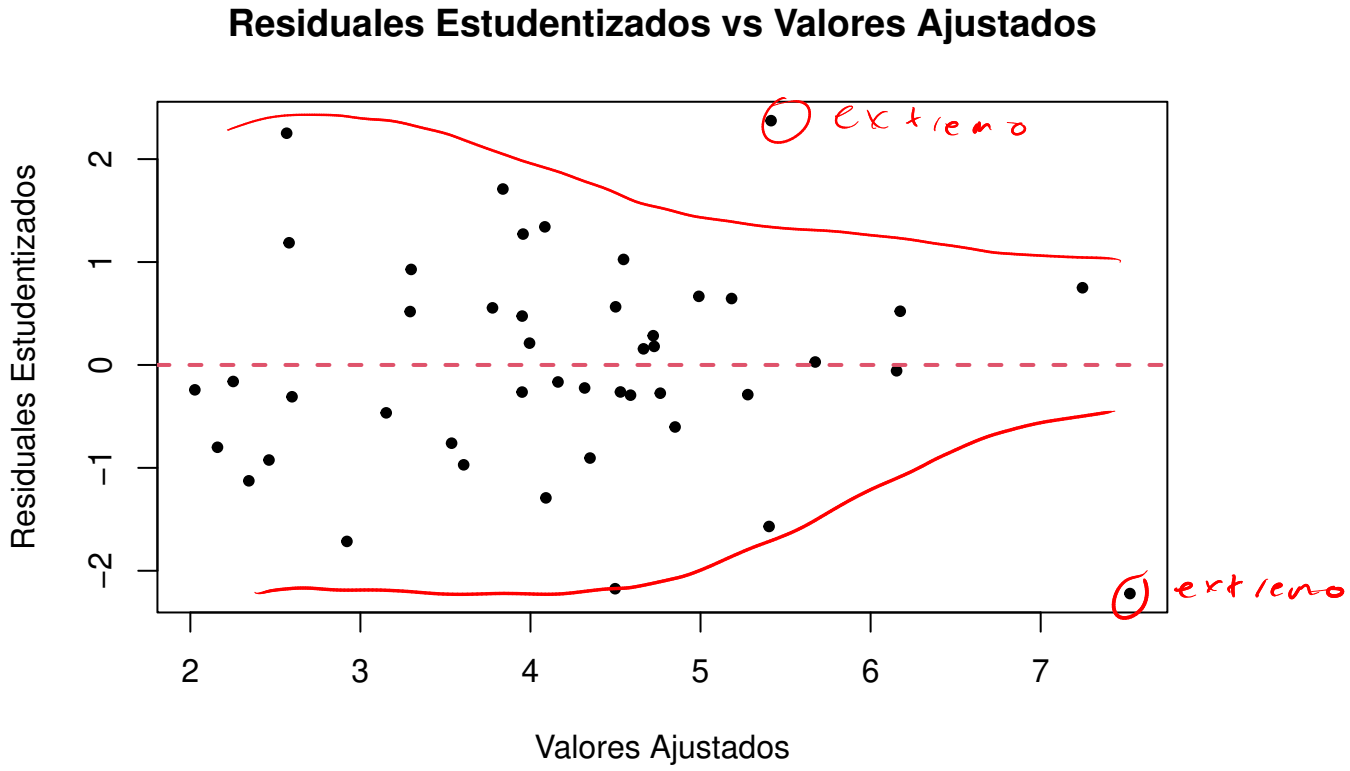

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se observa que no existe un comportamiento lineal entre los puntos. Esto se debe a que existe una variabilidad considerable de los datos, generando que los residuales se alejen de un comportamiento ideal, es decir, no hay linealidad en el modelo. En conclusión, el modelo no cumple el supuesto de varianza constante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

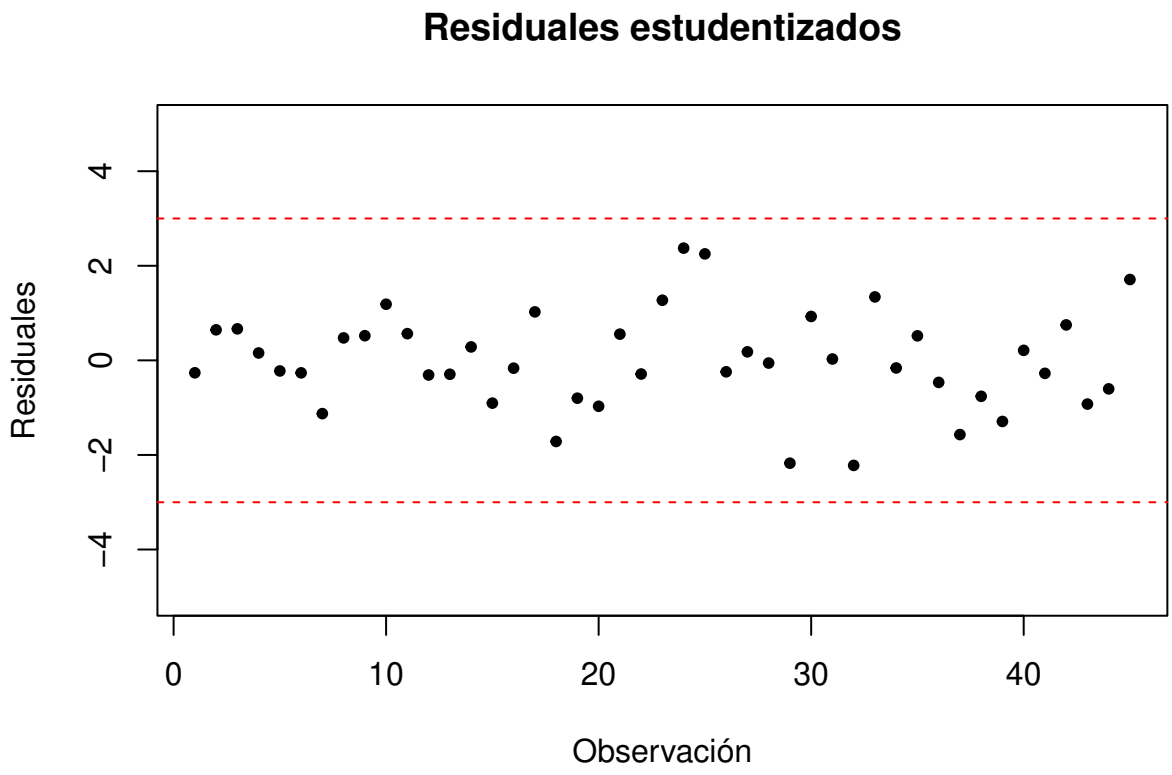

Figura 3: Identificación de datos atípicos

De la gráfica anterior, es posible deducir que no existen datos atípicos en el conjunto de datos ya que ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

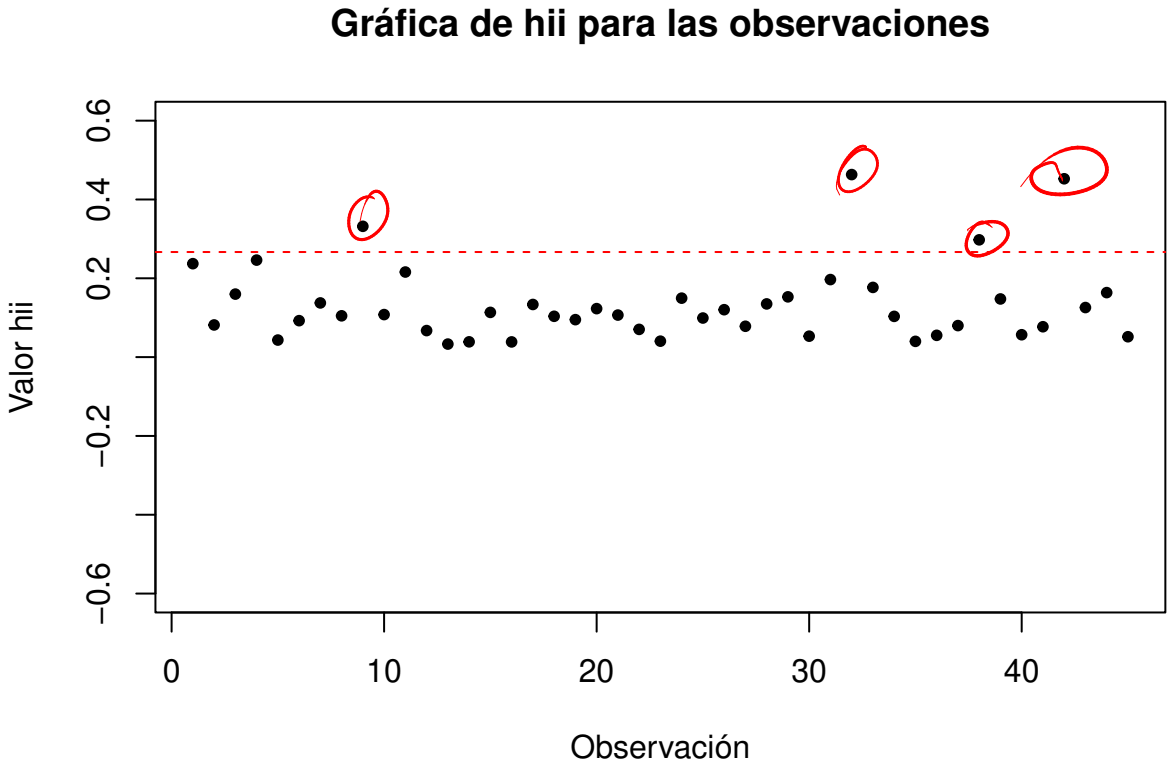

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Valores h_{ii} para las observaciones

	h_{ii}
9	0.3319
32	0.4632
38	0.2976
42	0.4526

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $2 \times \frac{p}{n} = 2 \times \frac{6}{45} = 0.26667$, se puede apreciar que existen 4 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n} = 0.26667$, los cuales son los presentados en la tabla. Específicamente estamos hablando de las observaciones 9, 32, 38 y 42 las cuales cumplen el criterio que los vuelve puntos de balanceo.

4.2.3. Puntos influenciales

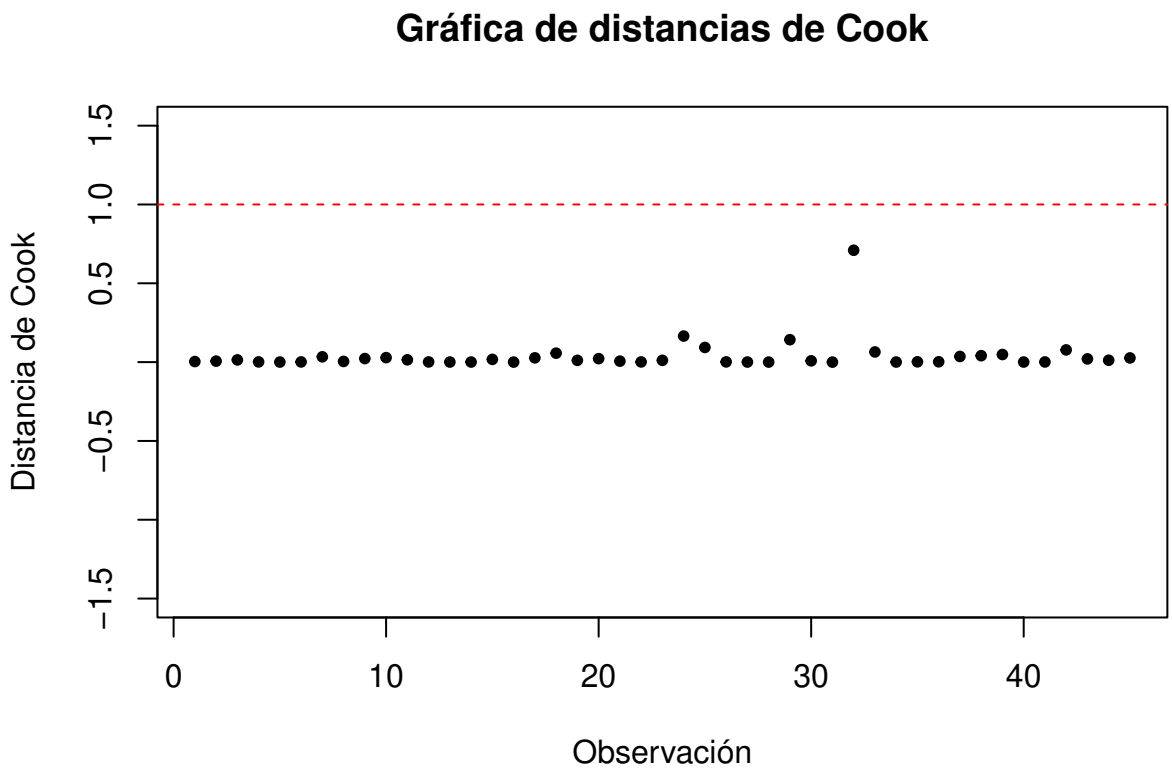

Figura 5: Criterio distancias de Cook para puntos influenciales

Respecto al criterio de distancias de Cook, en el cual para cualquier punto que cumpla la condición $D_i > 1$, será una observación influencial, notamos que ninguno de los datos cumple con la condición presentada.

✓

2 pt

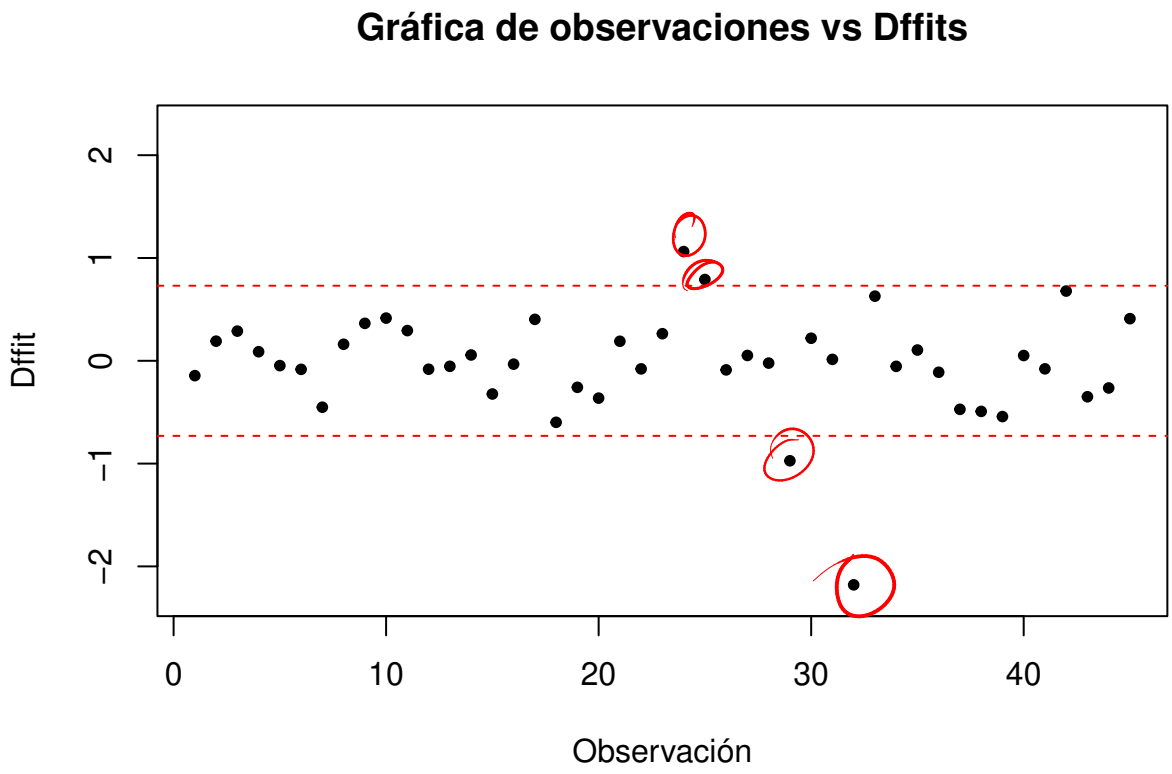

Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Valores DFFITS para las observaciones

	DFFITS
24	1.0626
25	0.7907
29	-0.9725
32	-2.1795

Como se puede ver, las observaciones 24, 25, 29 y 32 son observaciones influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffits}| > 2\sqrt{\frac{6}{39}} = 0.7845$, es un punto influencial.

4.3. Conclusión

Finalmente, podemos observar que el modelo de regresión NO cumple con el supuesto de normalidad ya que en la gráfica vemos que el patrón de los residuales no sigue la línea roja

que representa el ajuste de la distribución de los residuales a una distribución normal, de igual manera tampoco cumple con el supuesto de la varianza, pues no hay un comportamiento lineal de los residuales. En consecuencia, el modelo NO es válido.

Por otro lado, su R^2 no es ideal, pero sobrepasa el 60 %, los parámetros β_3 y β_4 son significativos (a un nivel $\alpha = 0.05$) mientras que todos los demás no lo son. Las observaciones 24, 25, 29 y 32 son observaciones influenciales según el criterio DFFITS cuya condición satisface $|D_{ffits}| > 2\sqrt{\frac{6}{39}} = 0.7845$. Por ello, consideramos que es importante darle especial estudio a estas observaciones con el fin de hacer el modelo lo más apropiado posible.

¿Qué es ideal?