Trabajo 1

3,1

Estudiantes

Ana Maria Barragan Ariza
Sofia Vanegas Arango
Esteban Garcia
Maria Hernandez

Equipo 16

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

15 pt

Considerando la información proporcionada en la base de datos, la cual incluye 5 variables predictoras identificadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Donde:

- Y: Riesgo de infección
- X_1: Duracion de la estadia
- 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 realizar el ajuste del modelo, se obtienen los siguientes valores para los coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.0465
β_1	0.1501
β_2	0.0252
β_3	0.0668
β_4	0.0044
β_5	0.0009

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.0465 + 0.1501 X_{1i} + 0.0252 X_{2i} + 0.0668 X_{3i} + 0.0044 X_{4i} + 9 \times 10^{-4} X_{5i};\ 1 \leqslant i \leqslant 64$$

1.2. Significancia de la regresión

Para evaluar la importancia de la regresión, se presenta el siguiente conjunto 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:

$$F_0 = \frac{SSR/K}{MSE} \overset{H_0}{\sim} f_{5,58} \tag{1}$$

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresion	57.0641	5	11.412821	13.3333	1.17778e-08
Error	49.6459	58	0.855964		

A partir de los resultados de la tabla Anova, se aprecia un valor P cercano a 0.05. Esto lleva al rechazo de la hipótesis nula, donde $\beta_j = 0$ para $0 \leqslant j \leqslant 5$, y a la aceptación de la hipótesis alternativa, indicando que al menos un $\beta_j \neq 0$. En consecuencia, se concluye que la regresión tiene relevancia estadística.

1.3. Significancia de los parámetros

En la tabla siguiente se proporciona informacion acerca de los parámetros, lo cual posibilitará identificar cuáles de ellos poseen importancia estadística.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.0465	1.4076	-0.0330	0.9738
β_1	0.1501	0.0657	2.2840	0.0261
β_2	0.0252	0.0257	0.9789	0.3317
β_3	0.0668	0.0132	5.0672	0.0000
β_4	0.0044	0.0070	0.6281	0.5324
β_5	0.0009	0.0006	1.4178	0.1616

La información en la tabla revela que, con un nivel de significancia $\alpha = 0.05$, los parámetros β_1 y β_3 son estadísticamente significativos, ya que sus P-valores son inferiores a α.

1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** Duracion de la estadia

Por cada unidad adicional en la Variable 1, se espera un aumento promedio de 0.0261 unidades en la variable de respuesta, manteniendo constantes las demás variables.

$\hat{\beta}_3$**:** Numero de camas

La Variable 3 no tiene un efecto lineal significativo en la variable de respuesta después de considerar las otras variables en el modelo.

1.5. Coeficiente de determinación múltiple R^2

El modelo exhibe un coeficiente de determinación múltiple $R^2 = 0.87000233071$, indicando que alrededor del 87 % de la variabilidad total en la respuesta se explica mediante el modelo de regresión descrito en este informe.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con los P-valores más elevados en el modelo fueron X_2, X_4, X_5 Por ende, mediante el análisis de todas las posibles regresiones en la tabla, se busca realizar la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	49.646	X1 X2 X3 X4 X5
Modelo reducido	52.209	X1 X3

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{0.854333333}{0.129286458} \\ &= 6.60806509 \end{aligned} \tag{2}$$

En la comparación entre el valor F_0 y $f_{0.95,3,58} = 2.7636$, se observa que $F_0 > f_{5,16,33}$. Dado que F_0 es mayor, se rechaza la hipótesis nula.

Esto implica que es factible descartar las variables X_2, X_4 y X_5 del modelo, ya que la evidencia sugiere que al menos una de estas variables tiene un efecto significativo. Por lo tanto, se puede contemplar un modelo simplificado que no incluya estas variables.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se dice que el hospital desea comparar si la duración de la estadia y el número de enfermeras tienen el mismo efecto sobre la respuesta, además con el fin de estudiar su capacidad de atención también les interesa saber si el censo promedio es igual al numero de camas, si vemos este problema como una prueba de hipotesis tenemos:

$$H_o : \begin{cases} \beta_1 - \beta_5 = 0 \\ \beta_3 - \beta_4 = 0 \end{cases}$$

si se ve de la forma $H_o : \mathbf{L}\underline{\beta} = 0$ se tiene

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

Con $\mathbf{L}$ dada por

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

Cuadro 5: Todas las regresiones posible modelo reducido

k	R_sq	adj_R_sq	SSE	Cp	Variables_in_model
1	0.309	0.298	73.746	6.422	X34
1	0.065	0.050	99.737	29.831	X15
2	0.365	0.345	67.727	3.000	X15 X34

Así tenemos que el modelo reducido (RM) es

$$Y = \beta_0 + \beta_1(X_1 + X_5) + \beta_3(X_3 + X_4) + \varepsilon$$

Ahora bien, si configuramos las nuevas variables: $X_{15} = X_1 + X_5$ y $X_{34} = X_3 + X_4$ podemos construir la tabla de todas las regresiones con el modelo reducido

Luego, se tiene que el estadístico de prueba está dado por:

$$F_0 = \frac{MSR}{MSE} = \frac{6.027033}{0.855964} = 7.041223$$

$$MSR = \frac{SSE(RM) - SSE(FM)}{Gl_{RM} - Gl_{FM}} = \frac{67.727 - 49.6459}{61 - 58} = 6.027033$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Con el fin de corroborar la suposición de normalidad en nuestros datos, proponemos llevar a cabo la siguiente prueba de hipótesis. Utilizaremos la prueba de Shapiro-Wilk, una herramienta estadística ampliamente reconocida para evaluar la normalidad de una muestra. Nuestra hipótesis nula (H0) plantea que la muestra sigue una distribución normal, mientras que la hipótesis alternativa (H1) sugiere que la distribución difiere de la normal. Como complemento a esta prueba, generaremos un gráfico cuantil-cuantil para visualizar cualquier desviación de la normalidad. Este enfoque nos permitirá realizar una evaluación completa respaldada estadísticamente sobre la normalidad en nuestros datos.

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

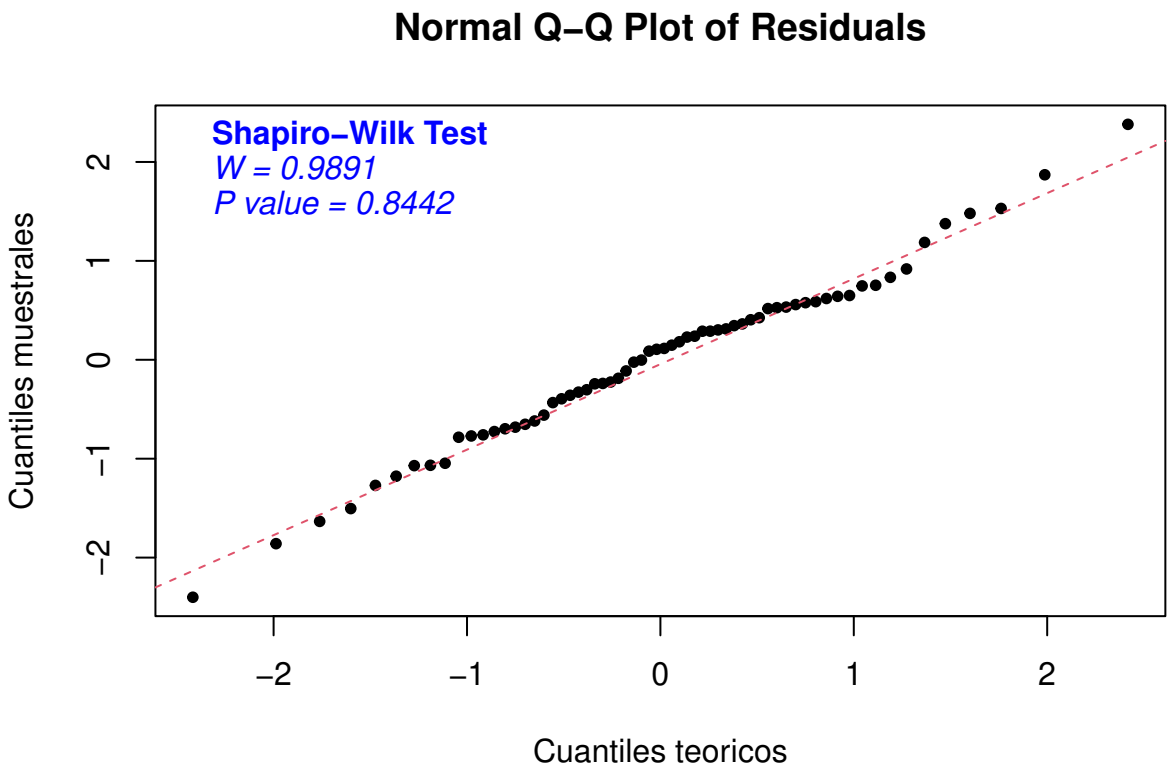

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

2pt

Dado que el P-valor es cercano a 0.8442 y considerando el nivel de significancia $\alpha = 0.05$, el valor p es significativamente mayor. Por lo tanto, no se rechaza la hipótesis nula, indicando que los datos siguen una distribución normal con media μ y varianza σ^2. A continuación, procederemos a verificar si la varianza cumple con la suposición de ser constante.

Análisis gráfico es aún más importante.

4.1.2. Varianza constante

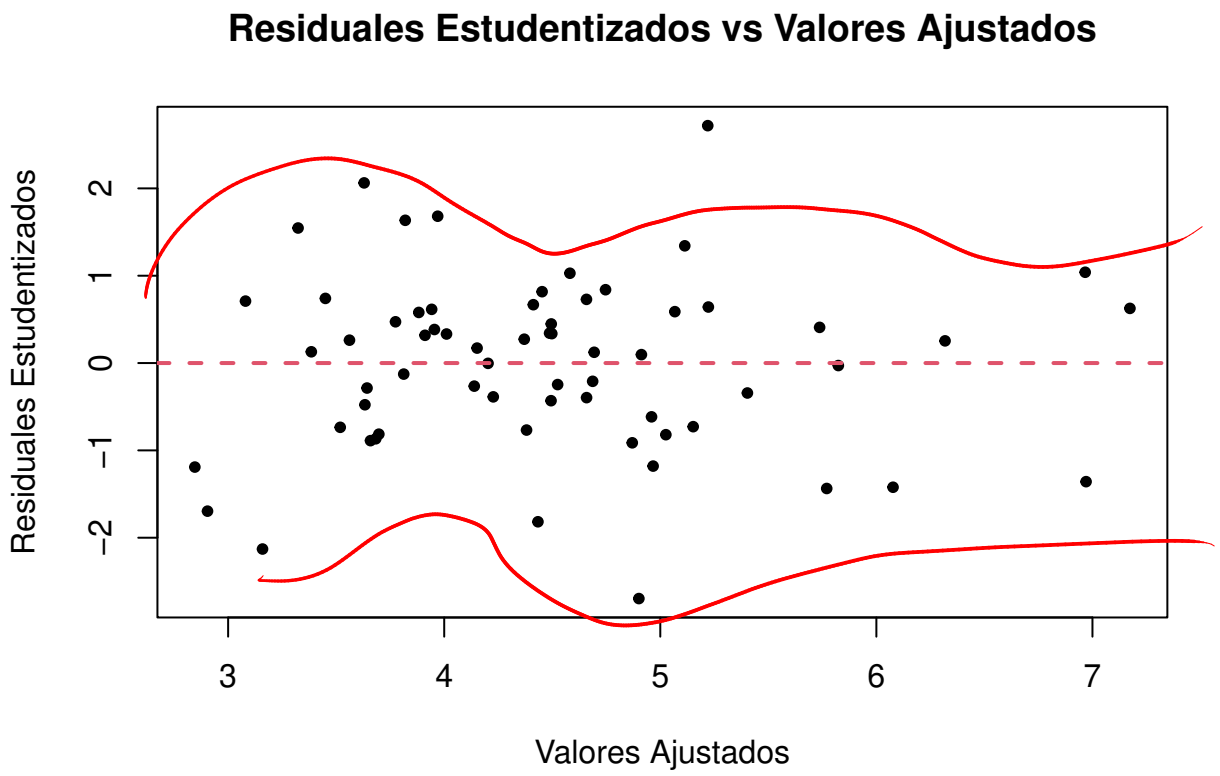

Figura 2: Gráfico residuales estudentizados vs valores ajustados

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

Tengan cuidado acá, modifiquen los límites de las gráficas para que tenga sentido con lo que observan en la tabla diagnóstica. También, consideren que en aquellos puntos extremos que identifiquen deben explicar el qué causan los mismos en el modelo.

4.2.1. Datos atípicos

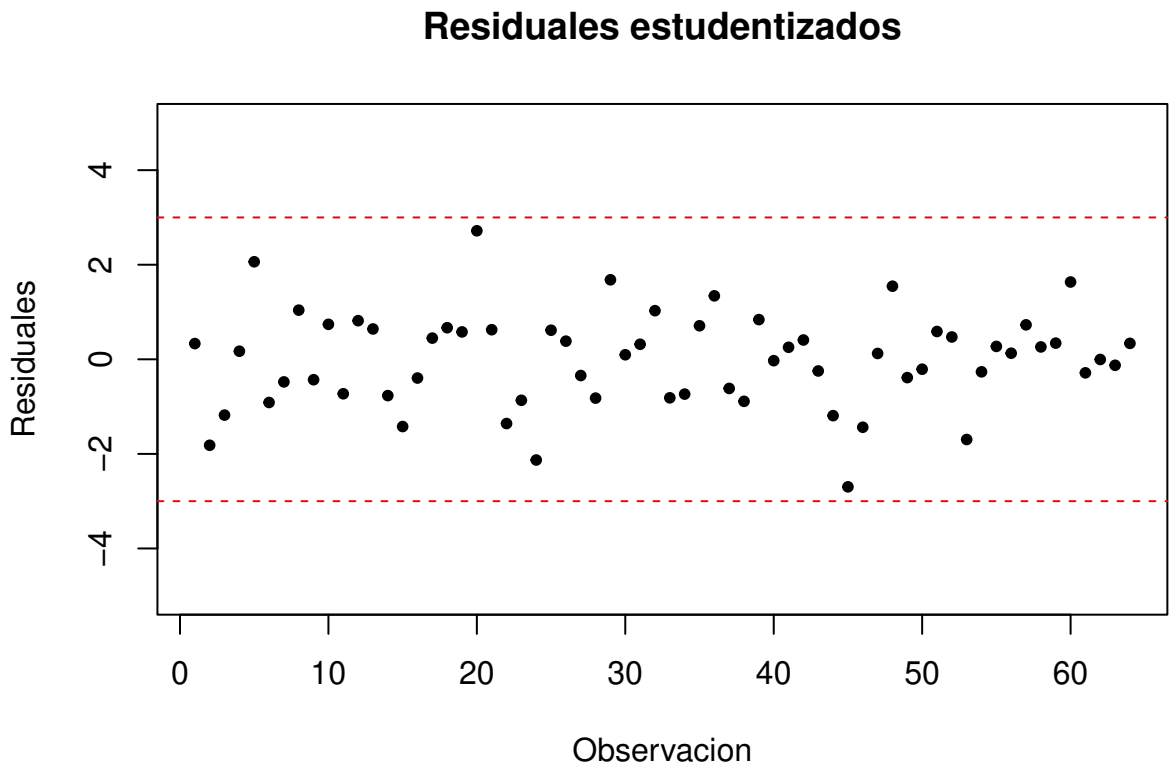

Figura 3: Identificación de datos atípicos

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

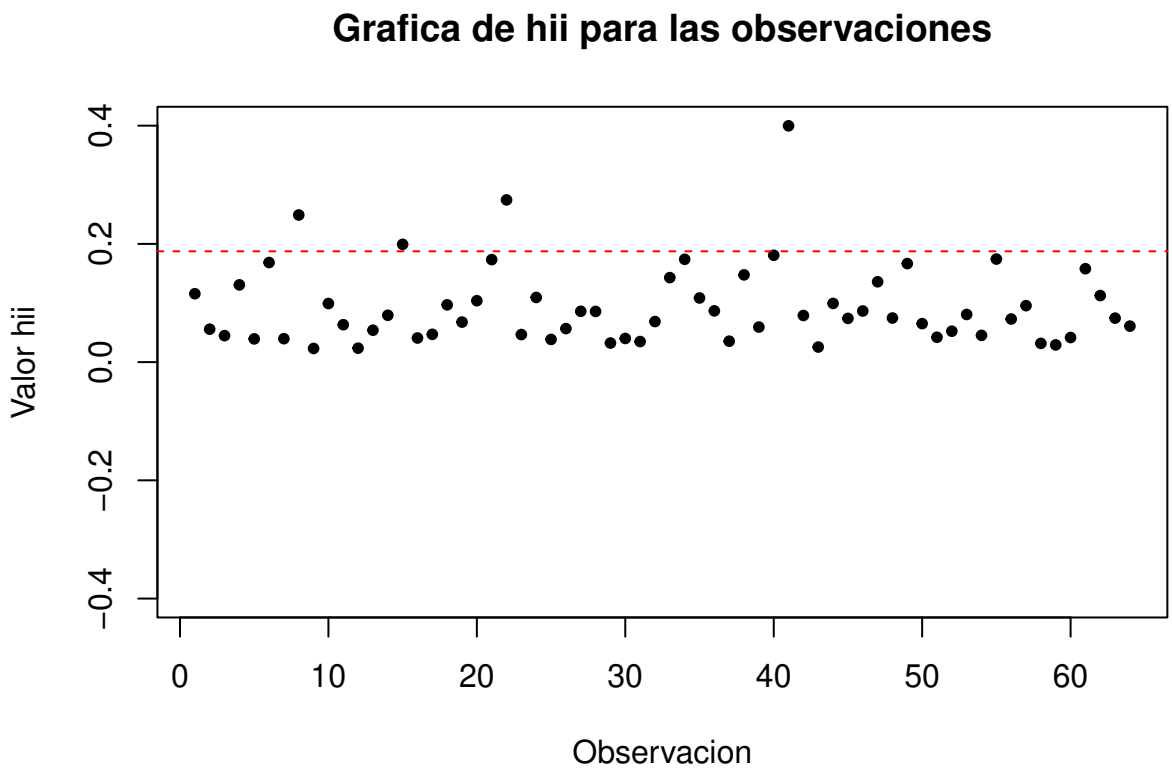

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 8    1.0397  0.0597    0.2489  0.5989
## 15  -1.4218  0.0838    0.1992 -0.7156
## 22  -1.3582  0.1163    0.2744 -0.8415
## 41   0.2541  0.0072    0.3998  0.2057
```

Al examinar el gráfico de observaciones frente a los valores h_{ii}, donde la línea punteada roja indica el valor $h_{ii} = 2\frac{p}{n}$, se nota la presencia de 4 puntos en el conjunto de datos que cumplen con el criterio de $h_{ii} > 2\frac{p}{n}$, como se detalla en la tabla adjunta.

4.2.3. Puntos influenciales

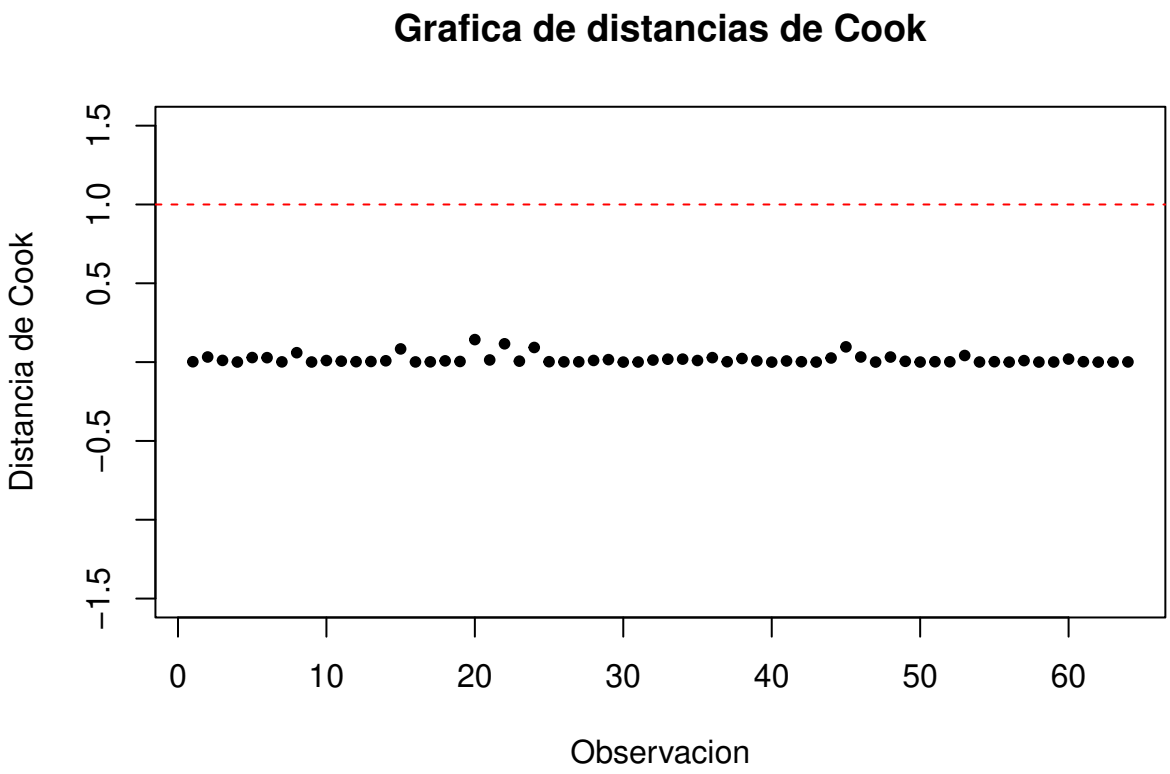

Figura 5: Criterio distancias de Cook para puntos influenciales

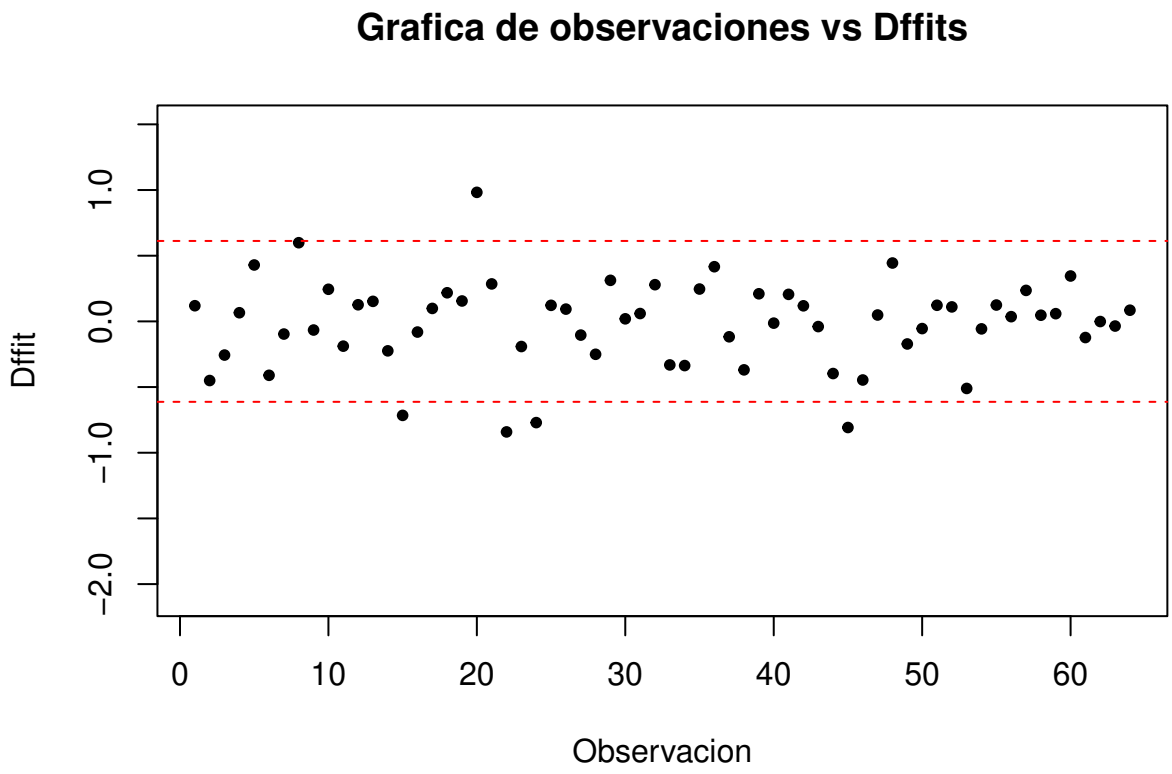

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 15  -1.4218  0.0838    0.1992 -0.7156
## 20   2.7181  0.1428    0.1039  0.9824
## 22  -1.3582  0.1163    0.2744 -0.8415
## 24  -2.1293  0.0928    0.1094 -0.7706
## 45  -2.6966  0.0968    0.0740 -0.8079
```

Como se puede ver,las observaciones los puntos 15, 20, 22, 24 y 45 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial. Cabe destacar también que con el criterio de distancias de Cook, en el cual para cualquier punto cuya $D_i > 1$, es un punto influencial, ninguno de los datos cumple con serlo.

4.3. Conclusión

Con un valor de p de 0.8442 en la prueba de normalidad y una gráfica de Shapiro-Wilk que indica linealidad, no hay suficiente evidencia para descartar la hipótesis nula de

que los datos siguen una distribución normal. En consecuencia, podemos inferir con un nivel de confianza del 95 % que los datos exhiben una distribución normal.

Dado que los residuales estandarizados se encuentran en un rango aceptable de -3 a 3 y la gráfica de distancias de Cook no indica la presencia de observaciones influyentes, la mayoría de las observaciones no tienen un impacto sustancial en el modelo. Sin embargo, al examinar la gráfica de Hii, se identificaron 4 puntos con valores ligeramente mayores, indicando que estas observaciones podrían tener una influencia moderada en los coeficientes de regresión.

válido o no ?