Trabajo 1

3,8

Estudiantes

Luis Miguel Alzate Cañas
Manuela Ferrer Cuervo
Esteban Gómez Benítez
Jorge Humberto Gaviria Botero

Equipo 41

Docente

Carlos Mario Lopera

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

17,5 pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

Donde ... acá dicen el nombre de las variables

- Y: Riesgo de infección
- X_1: Duracion 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
β_0	-0.5798
β_1	0.1456
β_2	0.0032
β_3	0.0441
β_4	0.0250
β_5	0.0021

2pt

No va en ec. ajustada

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.5798 + 0.1456 X_{1i} + 0.0032 X_{2i} + 0.0441 X_{3i} + 0.025 X_{4i} + 0.0021 X_{5i} + \varepsilon_i,\ 1 \leqslant i \leqslant 59$$

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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,53} \tag{1}$$

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	83.2914	5	16.658283	16.8377	6.19931e-10
Error	52.4354	53	0.989346		

De la tabla Anova, se observa un valor P aproximadamente igual a 0, por lo que no se haya la evidencia suficiente para rechazar la hipótesis nula en la que $\beta_j = 0$ con $0 \leqslant j \leqslant 5$, 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
β_0	-0.5798	1.6352	-0.3546	0.7243
β_1	0.1456	0.0765	1.9021	0.0626
β_2	0.0032	0.0305	0.1042	0.9174
β_3	0.0441	0.0146	3.0226	0.0039
β_4	0.0250	0.0079	3.1588	0.0026
β_5	0.0021	0.0007	2.9368	0.0049

Las pruebas de significancia para los parámetros establecen el siguiente juego de hipótesis.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \text{ para } j = 0, 1, ..., 5 \end{cases}$$

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros individuales β_3, β_4 y β_5 son significativos cada uno en presencia de los demás parámetros, pues sus P-valores son menores a α.

Por otro lado, se encuentra que β_0, β_1 y β_2 son individualmente no significativos en presencia de los demás parámetros

1.4. Interpretación de los parámetros

3pt

Para poder realizar una interpretación de los parámetros presentes en el modelo se debe identificar primero aquellos parámetros susceptibles de interpretación, es decir, unicamente se pueden interpretar los parametros que son significativos individualmente, en este caso son: β_3, β_4, β_5.

$\hat{\beta}_3$= 0.0441 indica que por cada dia que se aumente la estadia de un paciente (X_3) el promedio de la probabilidad de que el paciente adquiera una infección en el hospital aumenta en un 0.0441 %, cuando las demás predictoras se mantienen fijas

$\hat{\beta}_4$= 0.0250 indica que por cada paciente que aumente en el hospital (X_4) el promedio de la probabilidad de que el paciente adquiera una infección en el hospital aumenta en un 0.025 %, cuando las demás predictoras se mantiene fijas

$\hat{\beta}_5$= 0.0021 indica que por cada enfermera a tiempo completo que aumente en el hospital (X_4) el promedio de la probabilidad de que el paciente adquiera una infección en el hospital aumenta en un 0.0021 %, cuando las demás predictoras se mantiene fijas

1.5. Coeficiente de determinación múltiple R^2

2,5 pt
→ ¿cómo se calcula?

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6137$, lo que significa que aproximadamente el 61.37 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

Aunque el R^2 es usado com una medida de bonda del ajuste de la función de regresión, es necesario tener presente que valores grandes de R^2 no implican necesariamente que la superficie ajustada sea útil, ademas, cuando se agregan más variables predictorias al modelo, el R^2 tiende a no decrecer, aún cuando existan dentro del grupo de variables, un subconjunto de ellas que no aportan significativamente.

Como medida de bondad de ajuste se prefiere usar otros estadísticos que penalicen al modelo por el número de variables incluidas, entre ellos se tienen el MSE, y el R^2 ajustado, estas dos medidas son equivalentes, dado que éste último se define:

$$R^2_{adj} = 1 - \frac{(n-1)MSE}{SST} = 0.577$$

En nuestro caso el $R^2_{adj} = 0.577$ es menor que el $R^2 = 0.6137$, lo que indica que en el modelo pueden haber variables que no aporten significativamente.Es decir, que se pueden quitar variables que no aporten del modelo

2. Pregunta 2

5pt

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 para probar la significancia:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_4 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 4 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	52.435	X1 X2 X3 X4 X5
Modelo reducido	74.387	X3 X5

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,53} \\ &= \frac{(74.387 - 52.435)/3}{0.989346} \\ &= 7.2393 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,53} = 2.7791$, se puede ver que $F_0 > f_{0.95,1,45}$ entonces el subconjunto es significativo.

Basándonos en esto, podemos rechazar la hipótesis nula H_0 y afirmar la validez de la hipótesis alternativa. Por lo tanto, podemos concluir que la variable y está condicionada por al menos una de las variables vinculadas a los parámetros del subconjunto examinado, y no es posible excluir su influencia.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se desea saber si la variable independiente X_2 (Rutina de cultivos) tiene un efecto significativo en la variable dependiente Y (Riesgo de infección). Determinamos la hipótesis nula:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = 0 \\ \mathrm{H}_1 : \text{La igualdad no se cumple} \end{cases}$$

Esto significa que en caso de que la Hipótesis Nula se cumpla, no hay relación lineal significativa entre la Rutina de cultivos y el Riesgo de infección en los hospitales.

La matriz L para esta prueba de hipótesis es:

$$\begin{cases} \mathrm{H}_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ \mathrm{H}_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Con **L** dada por

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 x_5 + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/1}{MSE(MF)} \overset{H_0}{\sim} f_{1,53} = \frac{(SSE(MR) - 52.435)/1}{0.989346} \overset{H_0}{\sim} f_{1,53} \qquad (3)$$

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

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

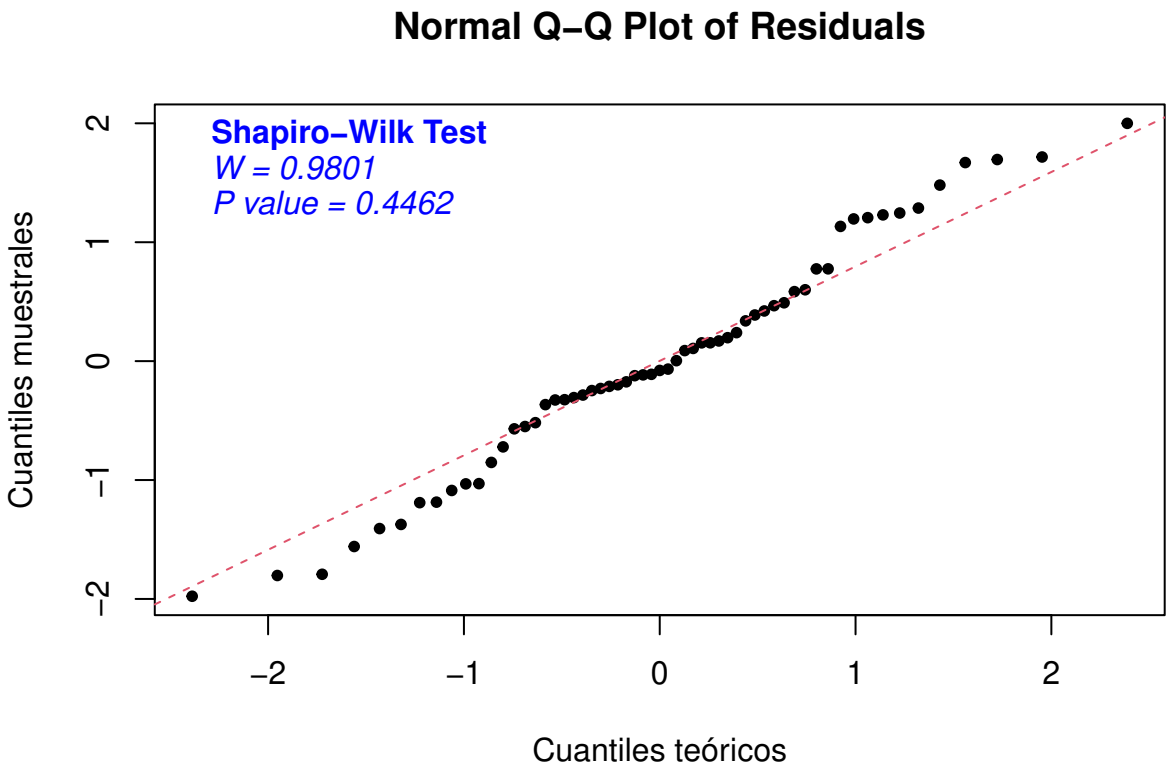

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.4462 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 μ y varianza σ^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

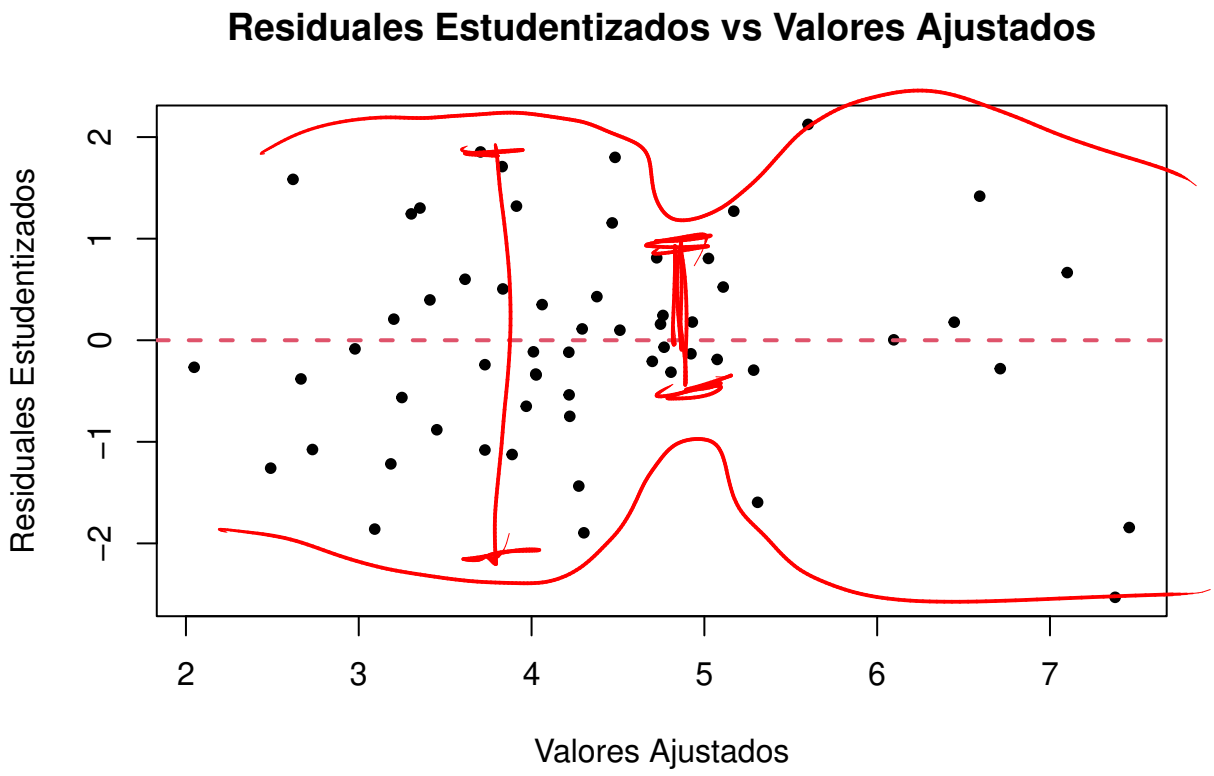

Figura 2: Gráfico residuales estudentizados vs valores ajustados

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

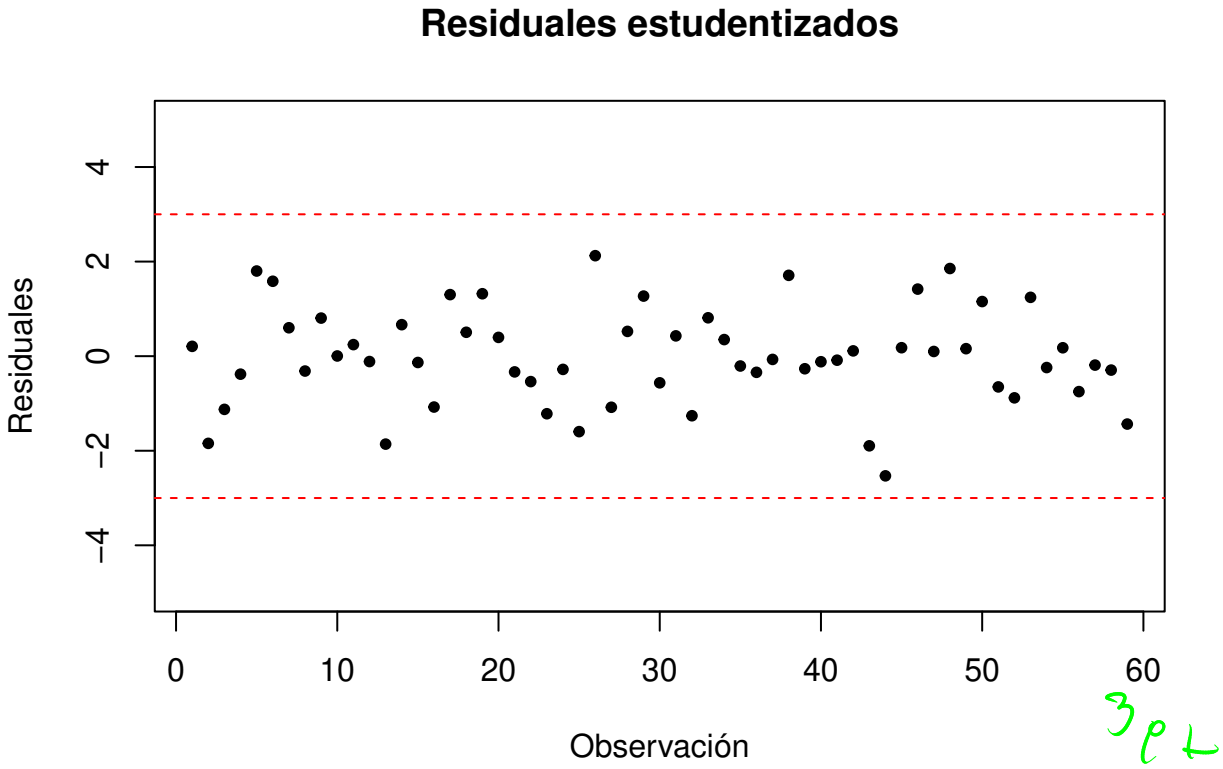

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$. Al no haber datos atípico, estos no afectan los resultados del ajuste del modelo de regresión.

4.2.2. Puntos de balanceo

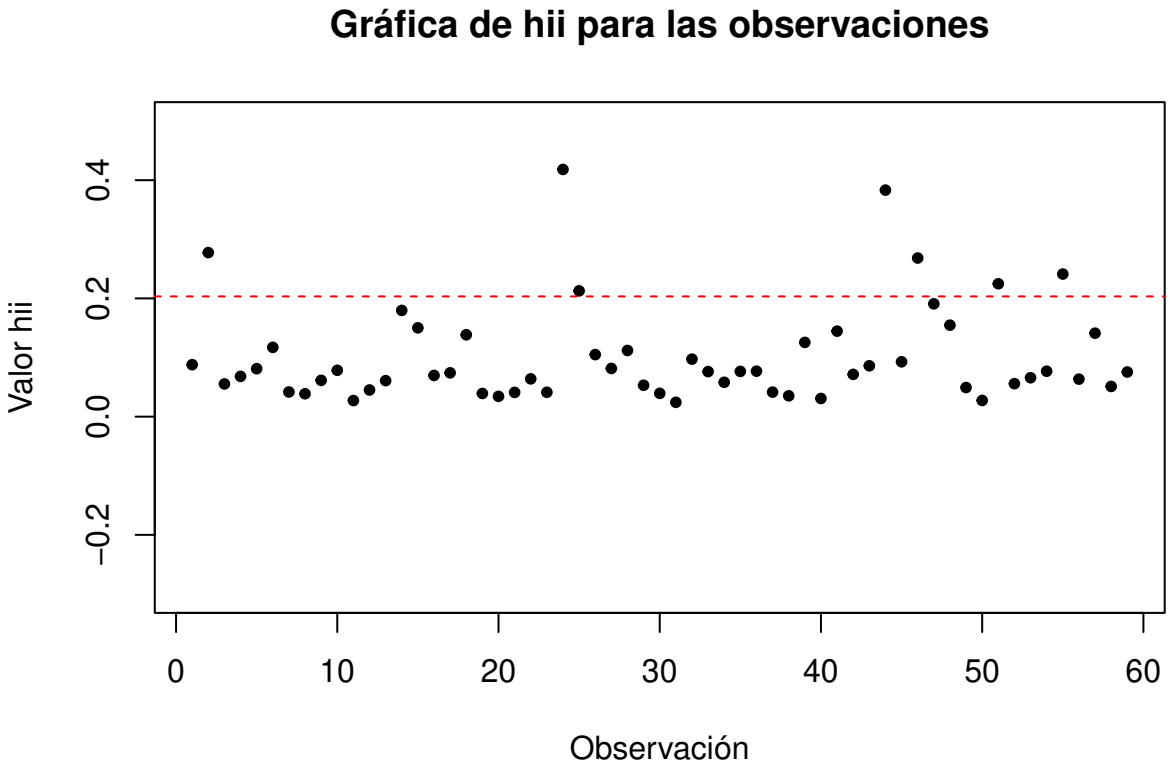

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 2   -1.8439  0.2176    0.2775 -1.1699
## 24  -0.2798  0.0094    0.4182 -0.2351
## 25  -1.5957  0.1147    0.2128 -0.8422
## 44  -2.5309  0.6631    0.3832 -2.1072
## 46   1.4181  0.1229    0.2683  0.8671
## 51  -0.6502  0.0204    0.2248 -0.3482
## 55   0.1777  0.0017    0.2412  0.0993
```

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 7 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.Que hayan puntos de balanceo quiere decir que afecta la normalidad de la regresión.

4.2.3. Puntos influenciales

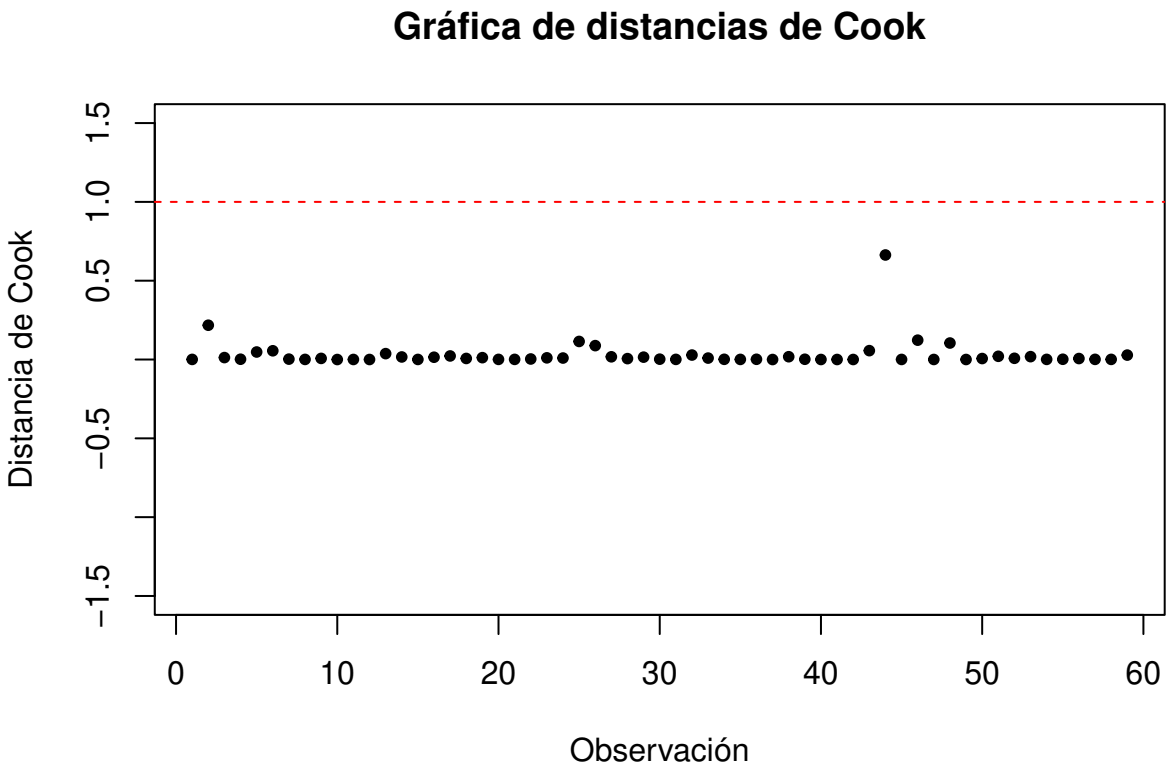

Figura 5: Criterio distancias de Cook para puntos influenciales

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.

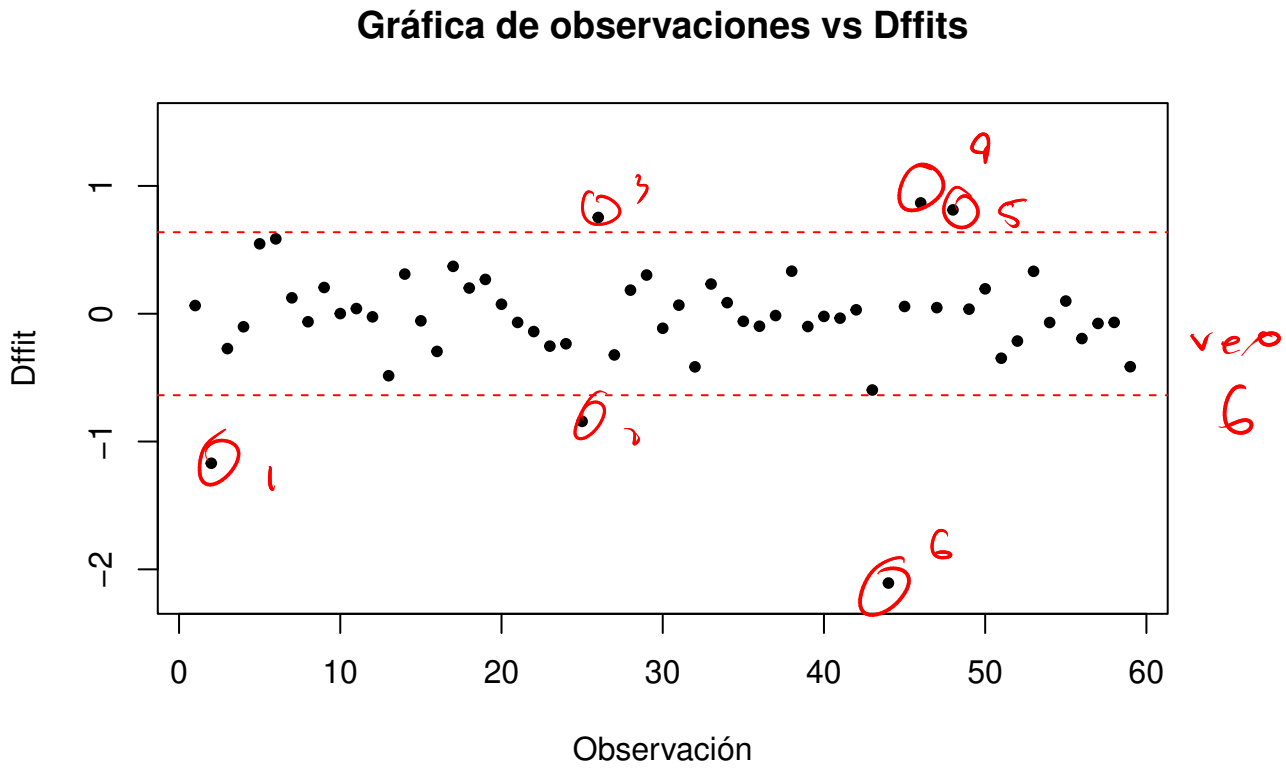

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 2   -1.8439  0.2176    0.2775 -1.1699
## 25  -1.5957  0.1147    0.2128 -0.8422
## 26   2.1249  0.0882    0.1049  0.7533
## 44  -2.5309  0.6631    0.3832 -2.1072
## 46   1.4181  0.1229    0.2683  0.8671
## 48   1.8538  0.1048    0.1547  0.8123
```

Como se puede ver, en la observación anterior, son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.637$, es un punto influencial, se puede apreciar en la tabla y en el gráfico, que existen 6 datos del conjunto que son influenciales por el criterio de Dffits. Al haber puntos influenciables, se ven afectos los coeficientes de regresión ajustados.

4.3. Conclusión

la invalidez del modelo queda claramente demostrada al no cumplir con el supuesto fundamental de normalidad. Este hecho se ve claramente reflejado en el análisis del gráfico "cuantil-cuantil y normalidad de residuales", donde se observan colas más pesadas y patrones irregulares, lo cual es una clara indicación de la falta de normalidad en los datos de los residuales. Este resultado compromete la validez de las inferencias realizadas a partir del modelo, ya que las pruebas y conclusiones basadas en la normalidad de los residuales no pueden ser consideradas confiables.

Además, es importante destacar que este problema se ve agravado por la presencia de puntos de balanceo que ejercen una influencia significativa en la normalidad de los residuales. Estos puntos distorsionan la distribución de los errores y afectan negativamente la validez del modelo.