Trabajo 1

4,3

Estudiantes

Felipe Vélez Fernández
Felipe Taborda Medina
Alvaro Jesus Sanchez Zarama

Equipo 54

Docente

Carlos Mario Lopera

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

5 de octubre de 2023

Índice

1. Pregunta 1 **3**

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

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

1. Pregunta 1

19pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ \ 1 \leqslant i \leqslant 54$$

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
β_0	0.7036
β_1	0.2167
β_2	-0.0075
β_3	0.0421
β_4	0.0122
β_5	0.0010

3pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.7036 + 0.2167 X_{1i} - 0.0075 X_{2i} + 0.0421 X_{3i} + 0.0122 X_{4i} + 0.001 X_{5i};\ \ 1 \leqslant i \leqslant 54$$

1.2. Significancia de la regresión

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

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

5pt

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,48} \qquad (1)$$

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	49.0289	5	9.805788	10.1723	1.09697e-06
Error	46.2703	48	0.963965		

De la tabla Anova, se obtienen los valores del estadístico de prueba $F_0 = 10.1723$ y su correspondiente valor-P, $V_p = 1.09697e - 06$.

Como $V_p < 0.05 = \alpha$,se rechaza 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 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	0.7036	1.9519	0.3605	0.7201
β_1	0.2167	0.0978	2.2161	0.0315
β_2	-0.0075	0.0374	-0.1993	0.8429
β_3	0.0421	0.0150	2.8080	0.0072
β_4	0.0122	0.0078	1.5728	0.1223
β_5	0.0010	0.0009	1.2092	0.2325

6pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1 y β_3 son significativos, pues sus P-valores son menores a α.

Por otro lado, se encuentra que β_0, β_2, β_4, β_5 son individualmente no significativos en presencia de los demás parámetros.

1.4. Interpretación de los parámetros

3pt

Interpreten sólo los parámetros significativos, respecto a β_0 ya saben que se debe cumplir que el 0 esté en el intervalo

$\hat{\beta}_1$= **0.2167** indica que por cada unidad de aumento en la duración de la estadía el promedio del resultado en la eficacia en el control de infecciones hospitalarias aumenta en 0.2167 unidades, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_3$= **0.0421** indica que por cada unidad de aumento en el número de camas el promedio del resultado en la eficacia en el control de infecciones hospitalarias aumenta en 0.0421 unidades, cuando las demás variables predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.51447$, lo que significa que aproximadamente el 51.447 de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

¿cómo se calcula?

2. Pregunta 2

4pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

era el más bajo

Las covariable con el P-valor más alto en el modelo fueron X_2, X_4, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	46.270	X1 X2 X3 X4 X5
Modelo reducido	50.313	X1 X3

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,48} \\ &= \frac{1.34767}{0.963965} \\ &= 1.39804 \end{aligned} \tag{2}$$

2pt

2pt

Ahora, comparando el F_0 con $f_{0.95,3,48} = 2.7981$, se puede ver que $F_0 < f_{0.95,3,48}$,entonces no se rechaza H_0 y se concluye que el conjunto de predictoras individualmente no significativas, en presencia de los demás parámetros, se pueden descartar del modelo.

3. Pregunta 3

5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Preguntas(Compruebe Si estas suceden a la vez):

- 1. El efecto de la duración de la estadia (X1) sobre la infeccion, es igual a 3 veces al efecto del censo promedio diario (X4) sobre la infeccion.
- 2. El efecto de la rutina de cultivos (X2) sobre la infeccion, es igual al efecto del numero de camas (X3) sobre la infeccion.
- 3. El efecto del numero de camas (X3) sobre la infeccion, es igual a 2 veces al efecto del numero de enfermeras (X5) sobre la infeccion.

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 3\beta_4;\ \beta_2 = \beta_3;\ \beta_3 = 2\beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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 **L** dada por:

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

El modelo reducido está dado por:

3pt

$$Y_i = \beta_o + \beta_3 X_{3i}^* + \beta_4 X_{4i}^* + \beta_5 X_{5i}^* + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

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

3.2. Estadístico de prueba

2pt

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/3}{MSE(MF)} \overset{H_0}{\sim} f_{3,48} = \frac{(SSE(MR) - 46.2703)/3}{0.963965} \overset{H_0}{\sim} f_{3,48} \qquad (3)$$

4. Pregunta 4

15pt

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:

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

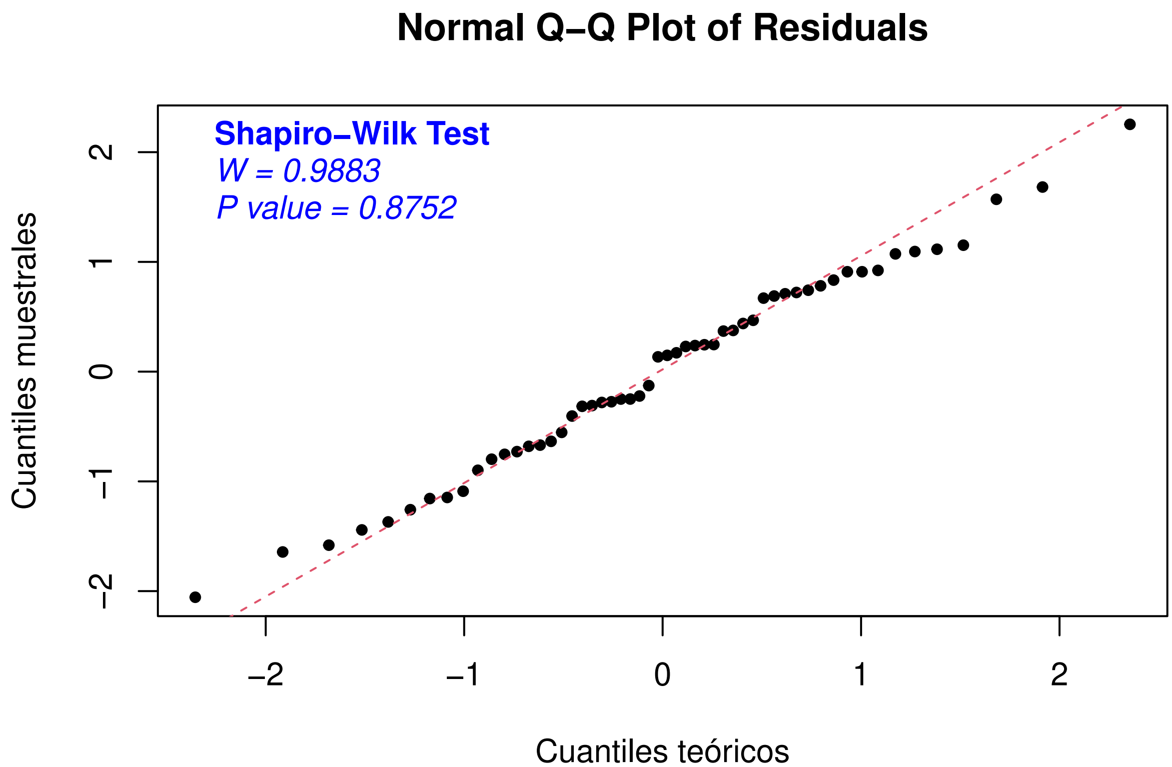


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

3 pt

Al ser el $V_p = 0.8752$ y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto,no se rechaza la hipótesis nula y se puede concluir que el supuesto de normalidad se cumple, es decir que los datos distribuyen normal con media μ y varianza σ^2, ademas la gráfica de comparación de cuantiles permite ver que los residuales se ajustan en su mayoría a la recta normal. al esto ser asi se puede sustentar con ayuda del gráfico que los datos distribuyen normal.

× Patrón muy irregular

4.1.2. Varianza constante

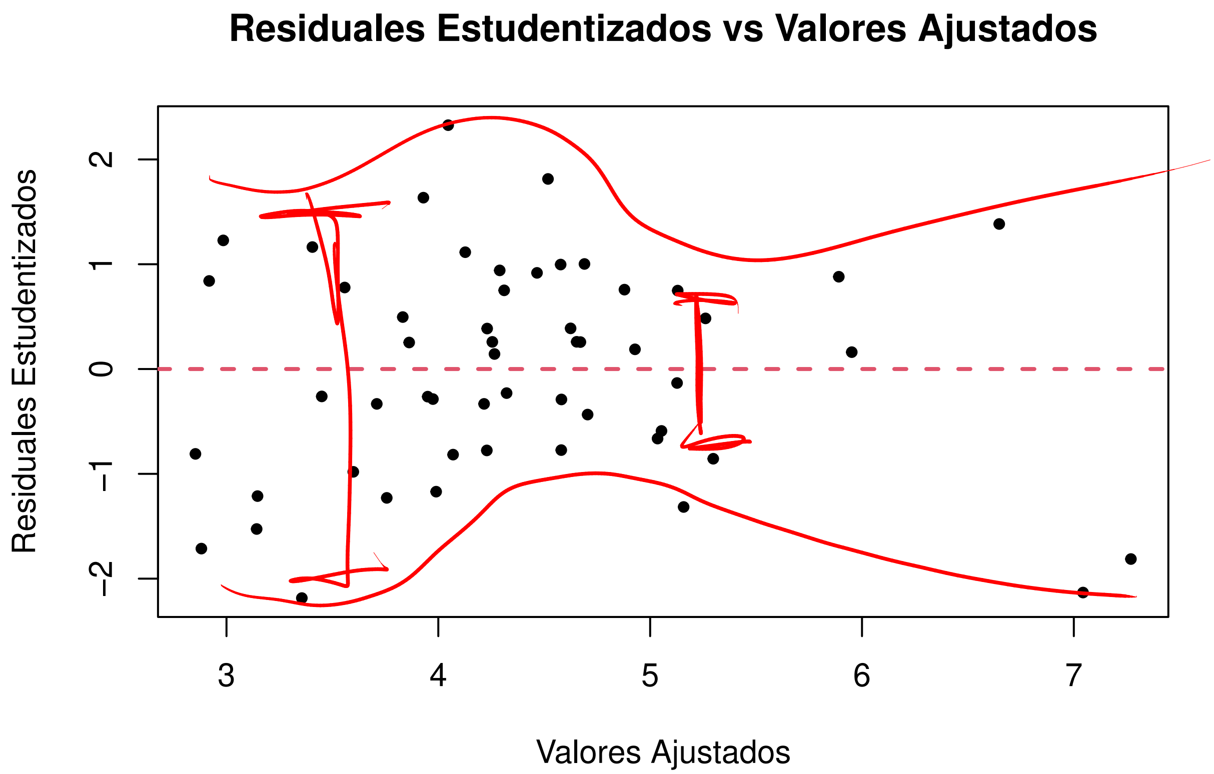
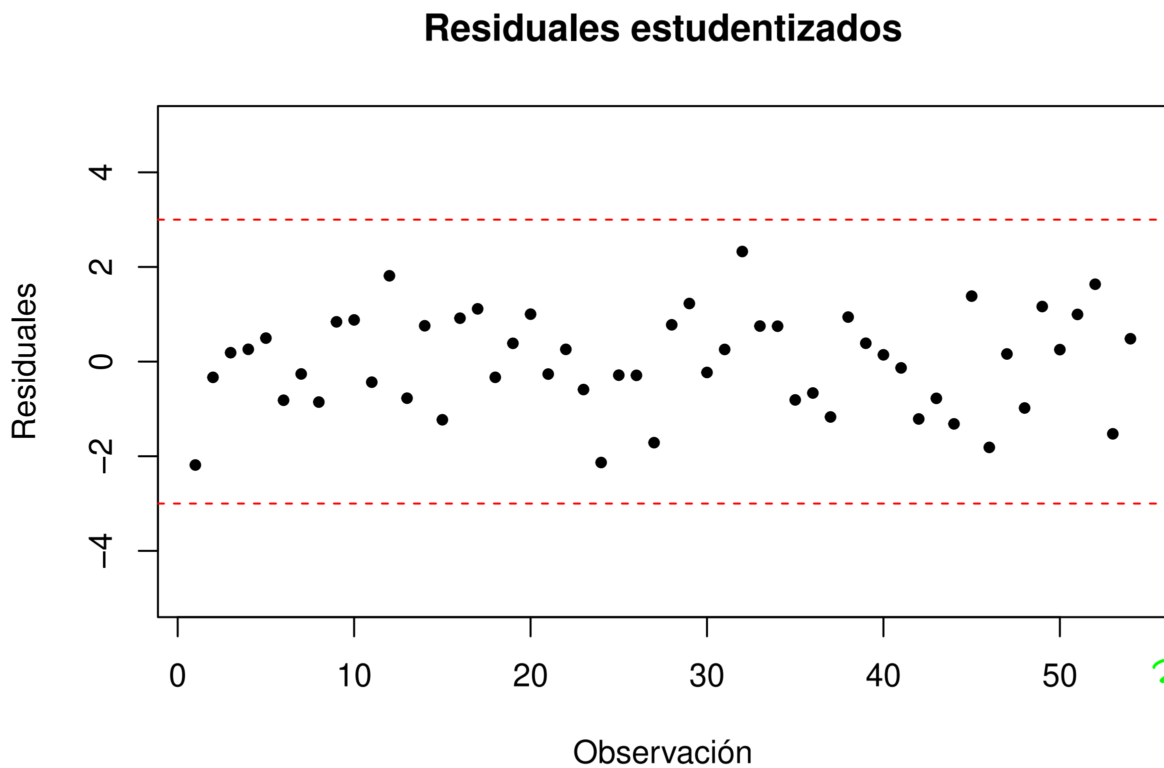


Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados Es evidente, que no se observa un patrón marcado en la distribución de los residuales, ni agrupaciones notables de los mismos, por lo tanto podríamos indicar que el supuesto de varianza constante sobre los residuales se cumple. Por lo tanto como a la luz de los residuales el supuesto de normalidad y varianza constante se cumplen, podemos concluir que el modelo es apto para hacer estimaciones y prediciones sobre el riesgo de infección. Sí lo hay

4.2. Verificación de las observaciones

4.2.1. Datos atípicos



3pt

Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica anterior, ningún residual estudentizado sobrepasa el criterio $|r_{estud}| > 3$ por lo tanto se concluye que no hay valores atípicos en el conjunto de datos.

4.2.2. Puntos de balanceo

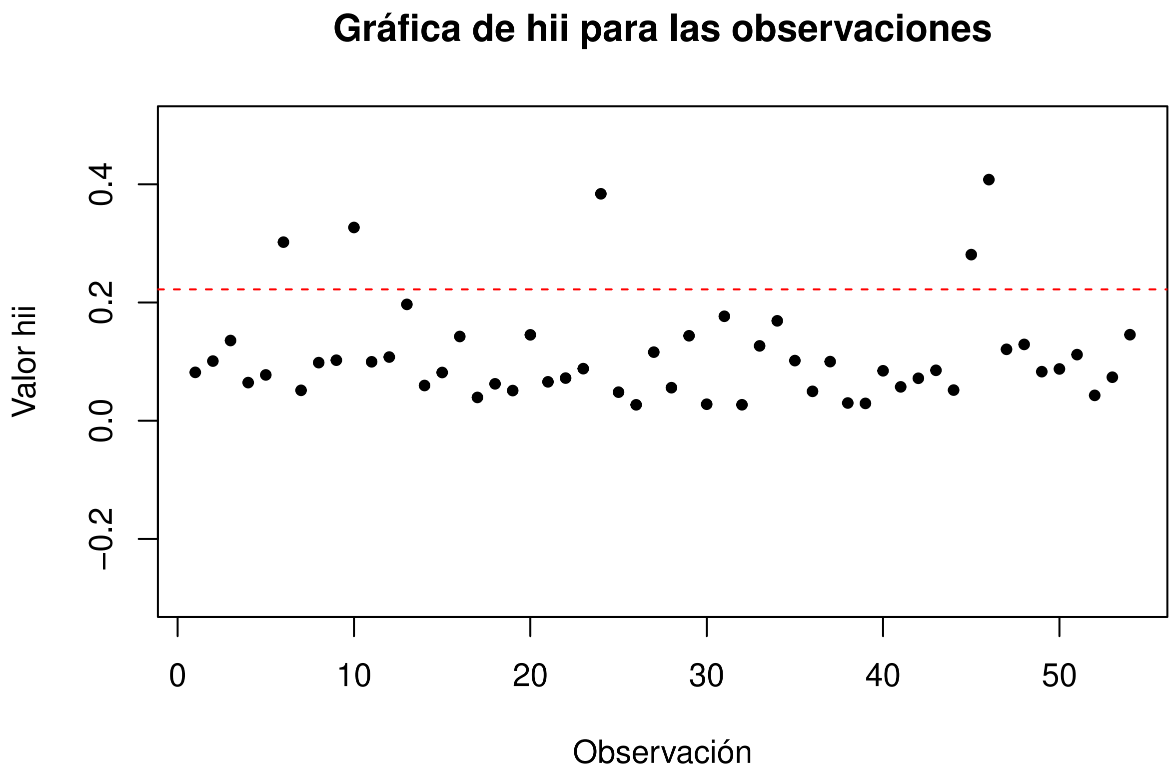


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 6   -0.8165  0.0481    0.3021 -0.5354
## 10   0.8807  0.0628    0.3269  0.6123
## 24  -2.1325  0.4723    0.3839 -1.7508
## 45   1.3840  0.1248    0.2810  0.8739
## 46  -1.8127  0.3773    0.4079 -1.5426
```

Causan ... ?

2pt

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} > 0.2222222$, se puede apreciar que existen 5 puntos de balanceo que son las observaciones 6, 10, 24, 45 y 46. según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$, los cuales son los presentados en la tabla.

4.2.3. Puntos influenciales

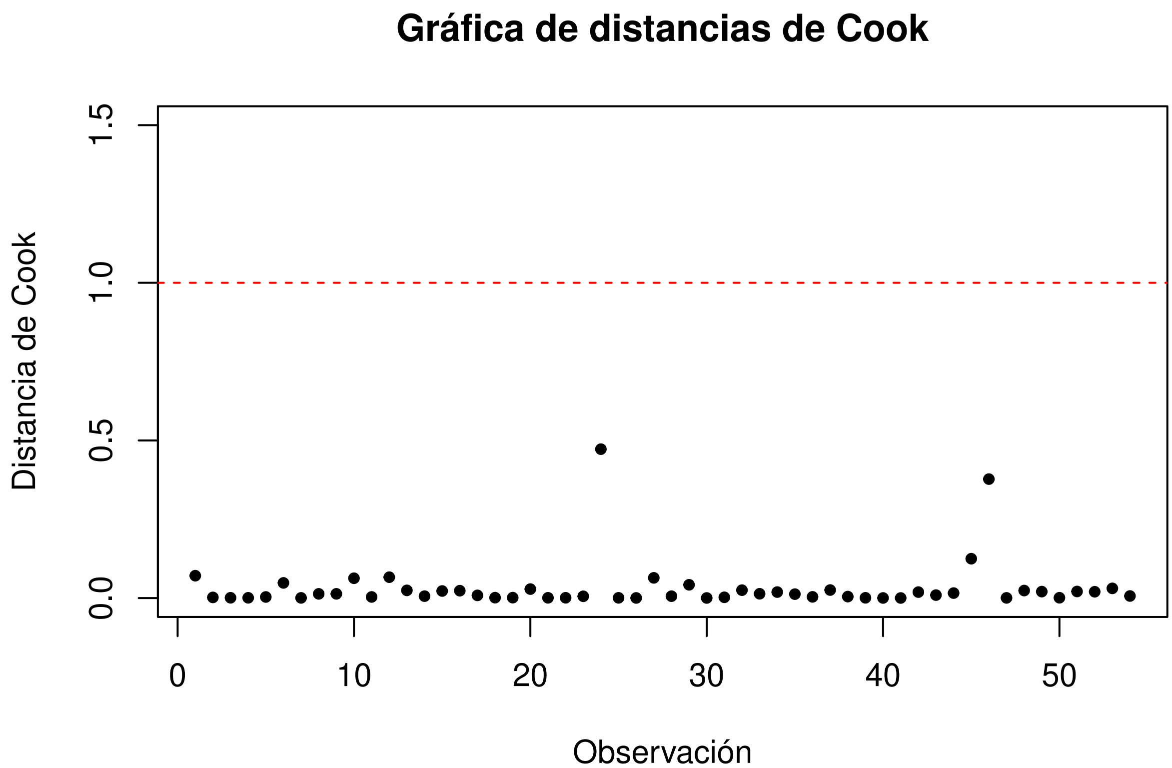


Figura 5: Criterio distancias de Cook para puntos influenciales

Mediante la gráfica de este criterio es posible ver que no existe ningún punto influencial que esté dado por $Di > 1$.

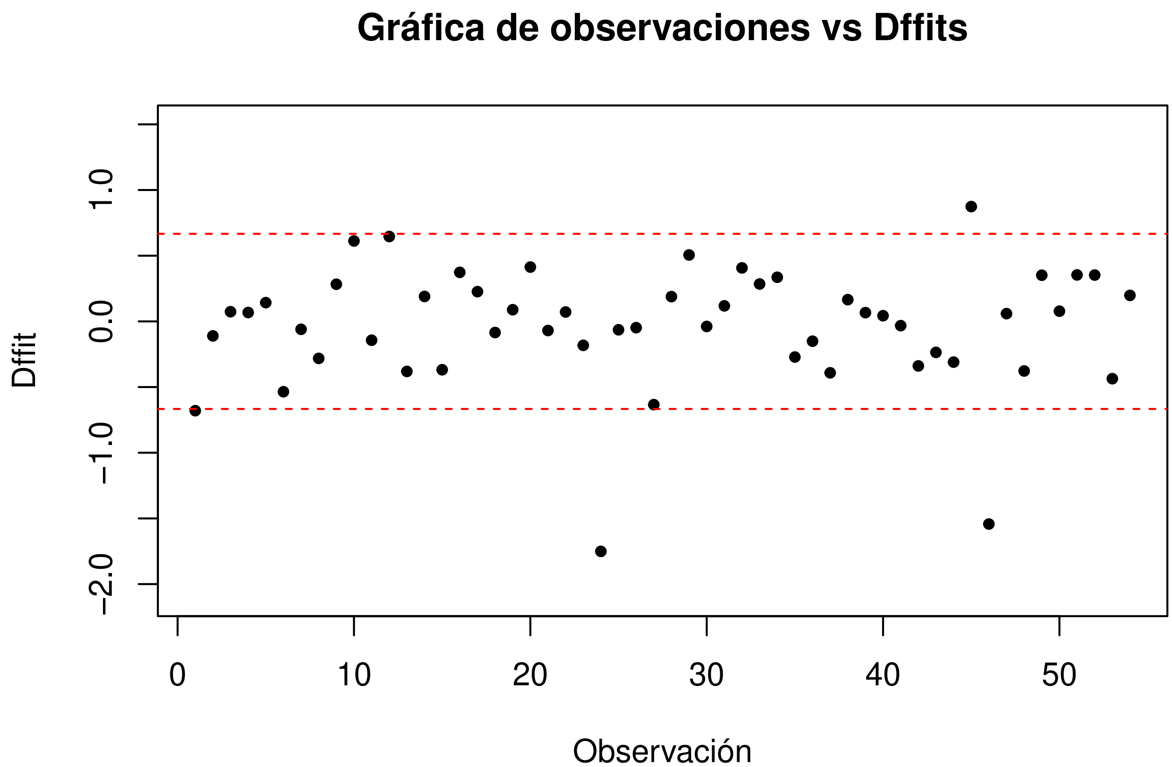


Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 1   -2.1851  0.0709    0.0818 -0.6799
## 24  -2.1325  0.4723    0.3839 -1.7508
## 45   1.3840  0.1248    0.2810  0.8739
## 46  -1.8127  0.3773    0.4079 -1.5426
```

3pt

influencial

Como se puede ver,las observaciones 1, 24, 45, 46 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{6}{54}} = 0.666666$, es un punto ~~influenciable.~~ Cabe destacar también que con el criterio de distancias de Cook, en el cual para cualquier punto cuya $D_i > 1$, es un punto influenciable, ninguno de los datos cumple con serlo.

Dado que se cumple el criterio de Dffits las observaciones 1, 24, 45, 46 son puntos influenciales.

Causan...?

4.3. Conclusión

2pt

D Bajo qué criterio. Es o no es?

~~Para nosotros,~~ el modelo que hemos empleado no se considera completamente válido. A pesar de que cumple con los supuestos de normalidad y varianza constante, la presencia de valores influenciables puede alterar el comportamiento del modelo de manera significativa. Es fundamental resaltar que, aunque los supuestos puedan estar satisfechos en términos generales, es crucial realizar un análisis exhaustivo de los valores influenciables detectados. Estos valores influenciables podrían estar ejerciendo una influencia significativa en la validez de los supuestos del modelo y en la interpretación de los resultados.

Además, el modelo de regresión utilizado en este análisis muestra un coeficiente de determinación múltiple(R^2) de aproximadamente 0.51447. Este valor indica que alrededor del 51.447 % de la variabilidad total observada en la variable dependiente es explicada por el modelo. En otras palabras, el modelo es capaz de explicar una parte significativa de la variabilidad en los datos.

Sin embargo, es importante destacar que, aunque el R^2 es una medida útil de la capacidad de ajuste del modelo, no es la única consideración al evaluar la validez del modelo. La validez del modelo debe ser evaluada considerando otros factores como la significancia de los coeficientes, la evaluación de supuestos y la presencia de valores extremos o influenciables. En este caso, al analizar todas estas consideraciones de validez, se observa que para un modelo con cinco coeficientes, solo dos de ellos resultan ser significativos, lo cual plantea interrogantes sobre la validez del modelo, especialmente debido a la influencia potencial de los valores extremos en la significancia de los coeficientes. Es importante tener en cuenta que, incluso si los supuestos se cumplen, la presencia de puntos influenciables o extremos puede impactar significativamente en la validez del modelo y en la interpretación de los resultados.