Trabajo 1

4,1

Estudiantes

Camilo Andrés Espíndola Aldana

Equipo 64

Docente

Javier Armando Lozano Rodriguez

Asignatura

Estadística II

Sede Medellín

5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

17 pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

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	-1.0857
β_1	0.2179
β_2	0.0293
β_3	0.0509
β_4	0.0074
β_5	0.0012

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.0857 + 0.2179 X_{1i} + 0.0293 X_{2i} + 0.0509 X_{3i} + 0.0074 X_{4i} + 0.0012 X_{5i}$$

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,63} \tag{1}$$

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	45.8937	5	9.178743	10.5334	2.23107e-07
Error	54.8979	63	0.871395		

De la tabla Anova, se observa un valor P muy bajo, por lo que se rechaza 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	-1.0857	1.4715	-0.7378	0.4634
β_1	0.2179	0.0713	3.0568	0.0033
β_2	0.0293	0.0275	1.0643	0.2912
β_3	0.0509	0.0146	3.4917	0.0009
β_4	0.0074	0.0066	1.1325	0.2617
β_5	0.0012	0.0007	1.8039	0.0760

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.

1.4. Interpretación de los parámetros

3 pt

Se tiene que los siguientes parámetros son significativos:

$\hat{\beta}_1$**: En promedio, por cada unidad de aumento en la duración de la estadía, el riesgo de infección aumenta en 0.2179 unidades, cuando las demás variables permanecen constantes**

$\hat{\beta}_3$**: En promedio, por cada unidad de aumento en el número de camas, el riesgo de infección aumenta en 0.0509 unidades, cuando las demás variables permanecen constantes**

1.5. Coeficiente de determinación múltiple R^2

2 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.46$, lo que significa que aproximadamente el 46 % 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

3 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más bajo en el modelo fueron X_1, X_3, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	54.898	X1 X2 X3 X4 X5
Modelo reducido	83.537	X2 X4

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_2, \beta_4) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{9.54633}{0.871395} \\ &= 10.955 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 > f_{0.95,1,45}$ y por tanto se rechaza H_0, de donde se descartan las variables del subconjunto.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hacen las preguntas. 1.Existe alguna relación entre el numero de camas y el numero de enfermeras? 2.El efecto de la duración de la estadía, sobre el censo promedio diario, es igual a 2 veces el efecto del censo promedio diario sobre le riesgo de infección?

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

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + 2\beta_4 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_3 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por: 2 pt

$$F_0 = \frac{(SSE(MR) - 54.8979)/2}{0.871395} \overset{H_0}{\sim} f_{2,63} \tag{3}$$

4. Pregunta 4 16 pt

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:

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

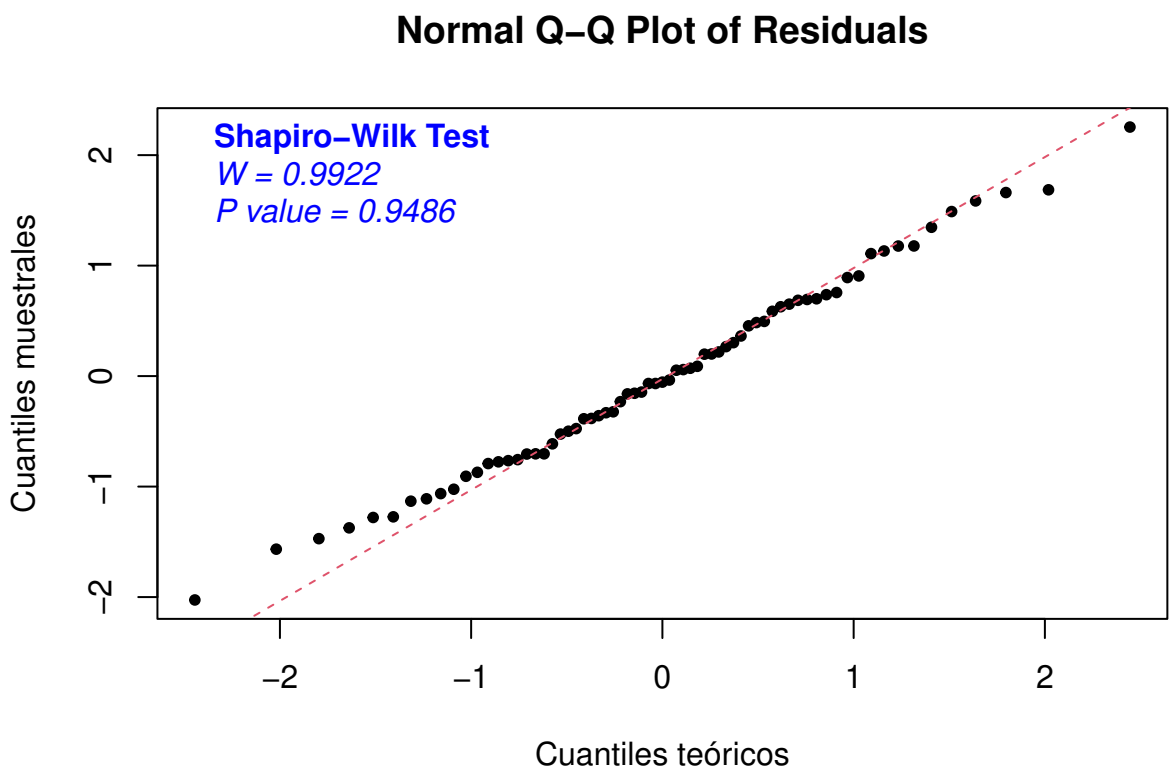

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

3 pt

Al ser el P-valor aproximadamente igual a 0.9922 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, se podría pensar que se rechaza la hipotesis nula y los datos distribuyen normal con media μ y varianza σ^2, sin embargo al observar la gráfica vemos que en realidad los datos no distribuyen normal, porque hay muchos datos alejados de la recta roja punteada, además por los extremos de la recta hay colas bien diferenciadas. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

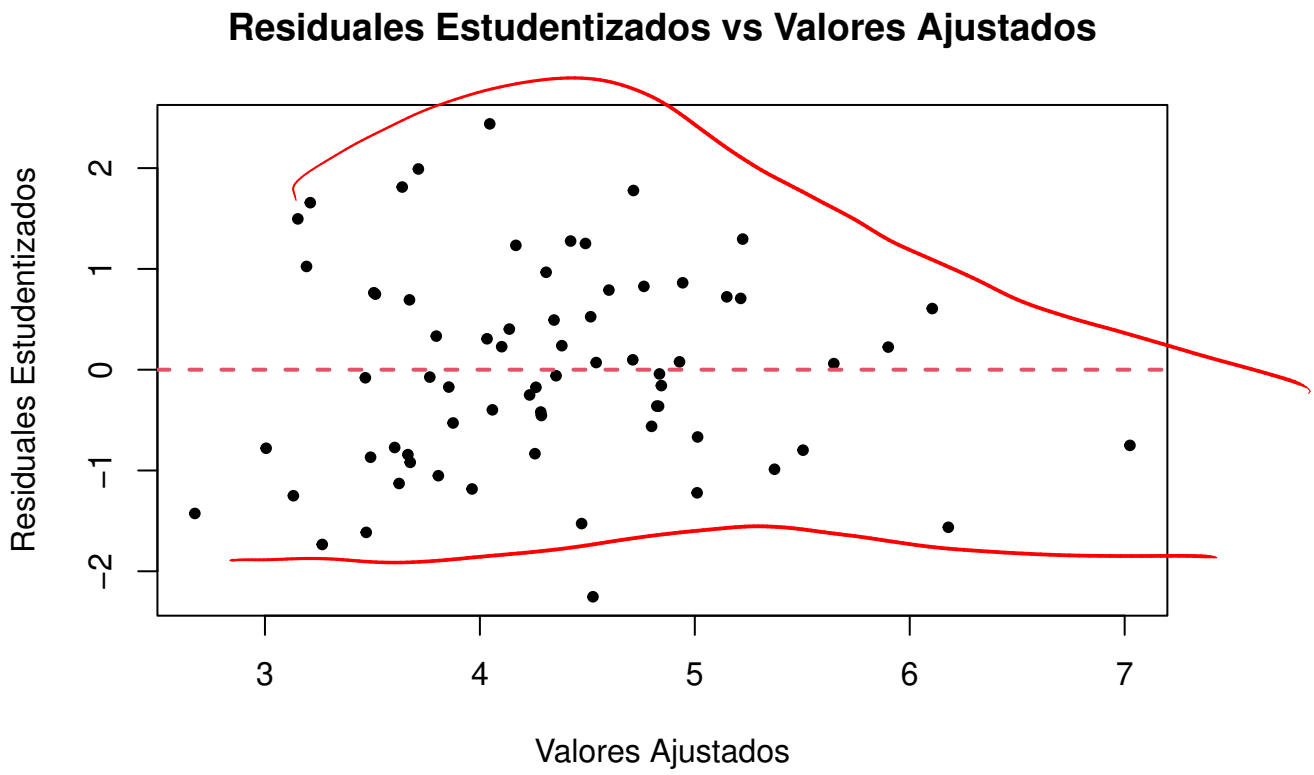

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se pueden observar patrones por ejemplo entre un valor de $x = 5$ y $x = 7$ se ve una flecha, por lo que se puede afirmar que la varianza no es constante, aunque la media si es aproximadamente 0.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

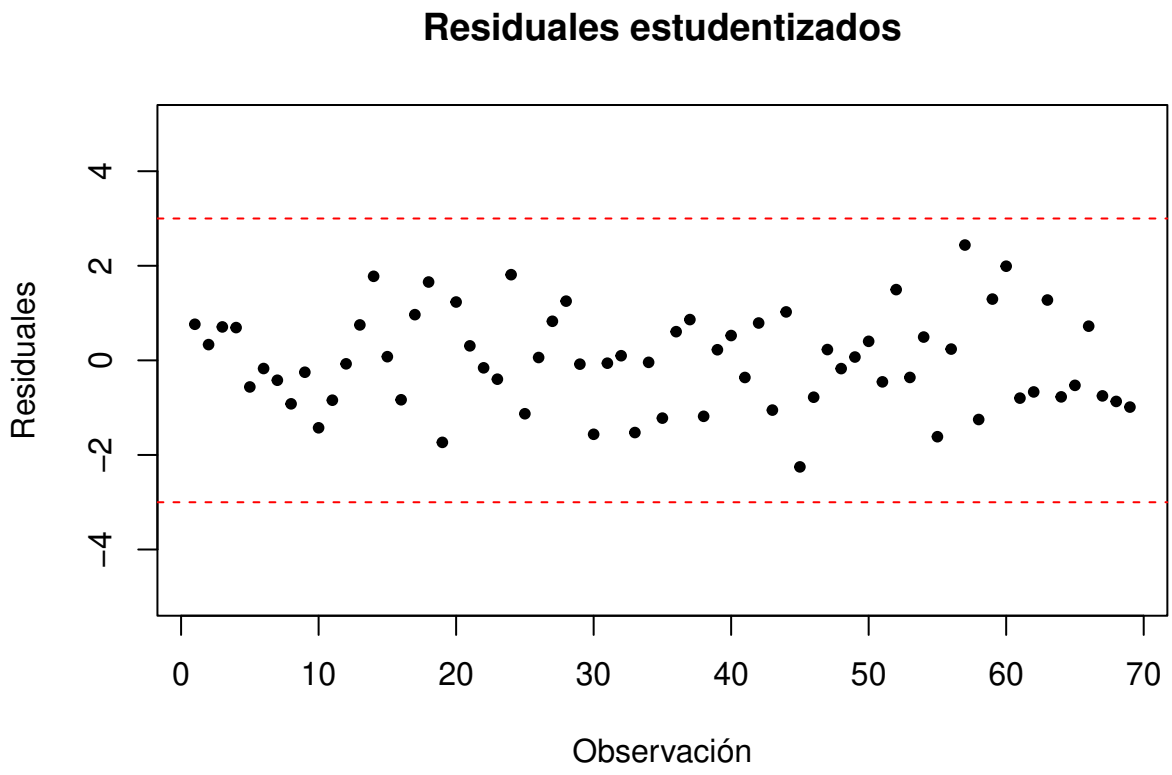

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

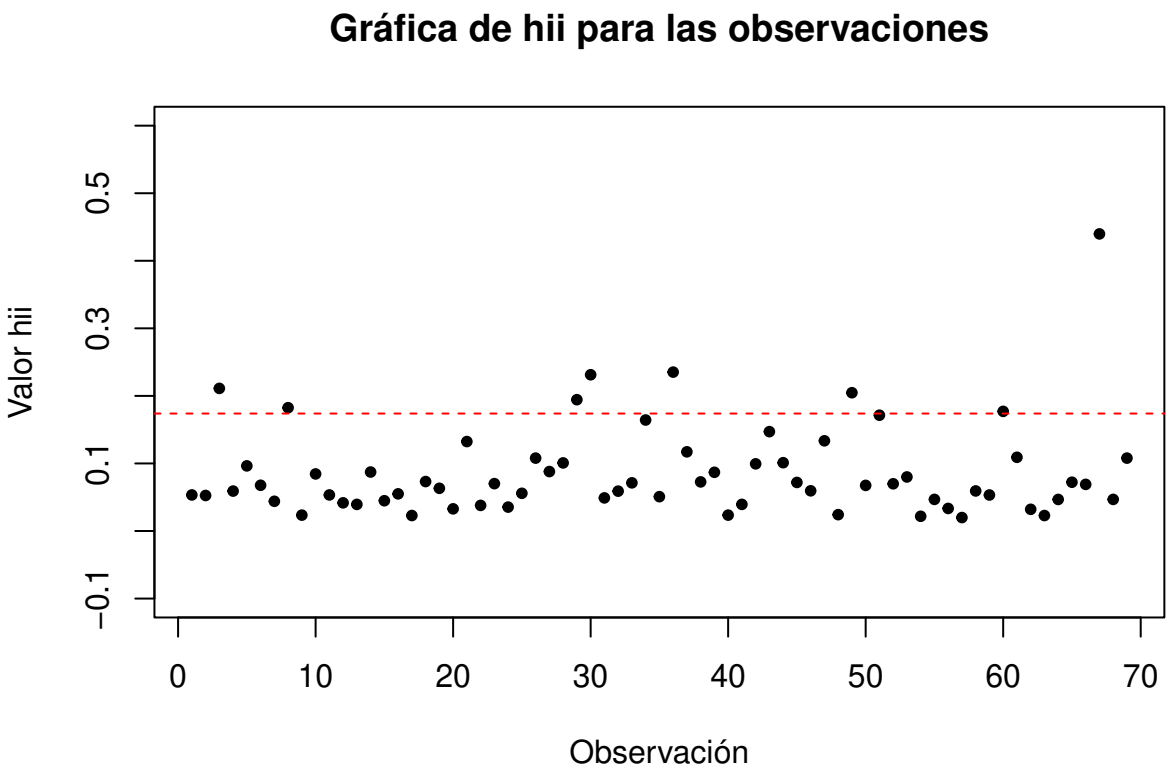

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 3    0.7073  0.0223    0.2111  0.3645
## 8   -0.9194  0.0314    0.1825 -0.4338
## 29  -0.0801  0.0003    0.1944 -0.0390
## 30  -1.5634  0.1225    0.2313 -0.8676
## 36   0.6066  0.0189    0.2352  0.3347
## 49   0.0707  0.0002    0.2047  0.0356
## 60   1.9912  0.1423    0.1771  0.9468
## 67  -0.7509  0.0738    0.4398 -0.6630
```

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 8 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.

4.2.3. Puntos influenciales

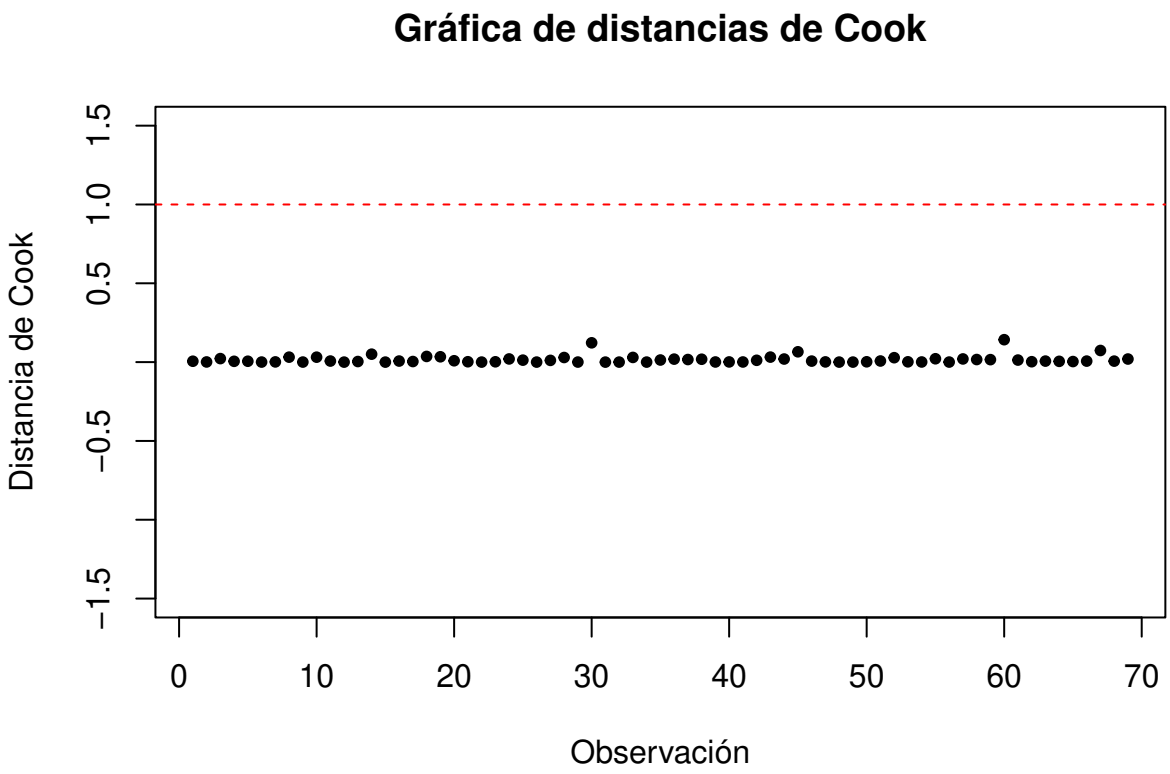

Figura 5: Criterio distancias de Cook para puntos influenciales

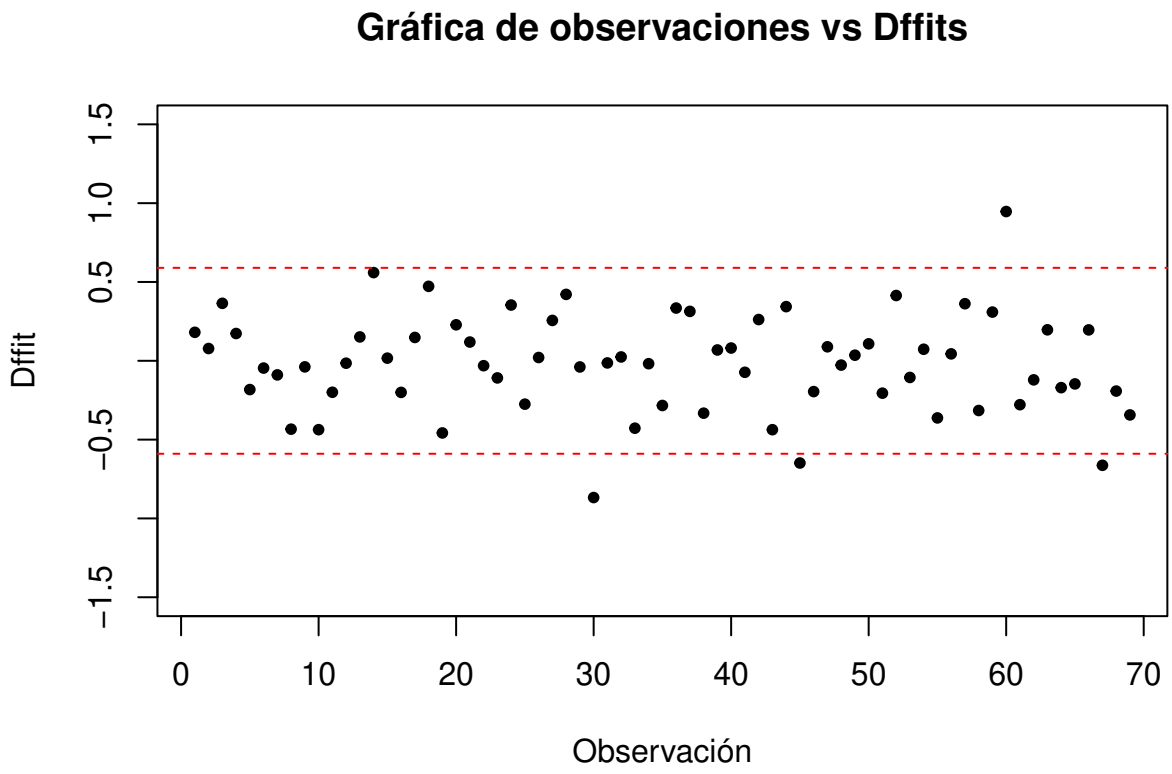

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 30  -1.5634  0.1225    0.2313 -0.8676
## 45  -2.2529  0.0655    0.0719 -0.6485
## 60   1.9912  0.1423    0.1771  0.9468
## 67  -0.7509  0.0738    0.4398 -0.6630
```

Ni hablan de estos 2 pt

Como se puede ver,las observaciones de la tabla 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, ya que todos los puntos están por debajo de 1.

3 pt

4.3. Conclusión

El modelo no es válido por las siguientes razones: 1. Por el criterio gráfico se observa que no se cumple el supuesto de que los datos se distribuyan normal. 2. La varianza no es constante, directamente no se está cumpliendo es supuesto del modelo de regresión lineal, lo que puede provocar que las estimaciones de los coeficientes de regresión sean ineficientes y

sesgadas, afectando la presición de las pruebas de hipótesis. 3. Los puntos de balanceo y los puntos extremos tamién pueden afectar el supuesto de linealidad, el supuesto de normalidad de los errores, el supuesto de independencia de los errores y el supuesto de homocedasticidad.

la 3 no es tan dificil