Trabajo 1

9,2

Estudiantes

Maria Clara Gonzalez Arismend
Manuel Jose Gomez Echeverri
Juan Pablo Campo Dorian
David Esteban Munera Agudelo

Equipo 42

Docente

Carlos Mario Lopera

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

19 pt

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

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

Donde: Y: Riesgo de infección: Probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).

X1: Duración de la estadía: Duración promedio de la estadía de todos los pacientes en el hospital (en días).

X2: Rutina de cultivos: Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100.

X3: Número de camas: Número promedio de camas en el hospital durante el periodo del estudio.

X4: Censo promedio diario: Número promedio de pacientes en el hospital por día durante el periodo del estudio.

X5: Número de enfermeras: Número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio

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	2.8494
β_1	0.3432
β_2	-0.0567
β_3	0.0682
β_4	-0.0032
β_5	0.0013

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 2.8494 + 0.3432 X_{1i} - 0.0567 X_{2i} + 0.0682 X_{3i} - 0.0032 X_{4i} + 0.0013 X_{5i}$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

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

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	69.7973	5	13.959454	26.8852	7.47758e-13
Error	24.9227	48	0.519224		

Tomando un Valor de significancia de $\alpha = 0.05$ y basándonos en la tabla ANOVA, se encuentra que el valor P es 0, lo que es menor a α con lo que concluimos que se rechaza la hipótesis nula, por lo tanto el modelo es **significativo**. Lo anterior quiere decir que el riesgo de infección depende significativamente de una o varias variables predictoras.

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	2.8494	1.6044	1.7760	0.0821
β_1	0.3432	0.0895	3.8353	0.0004
β_2	-0.0567	0.0290	-1.9586	0.0560
β_3	0.0682	0.0131	5.2154	0.0000
β_4	-0.0032	0.0066	-0.4856	0.6295
β_5	0.0013	0.0007	1.7979	0.0785

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

1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** Indica que por cada unidad que aumenta el **Tiempo de duracion en la estadía**, aumenta el promedio del riesgo de infección en **0.3432** cuando las demás variables se mantienen fijas

$\hat{\beta}_3$**:** Indica que por cada unidad que aumenta el **Número de camas**, aumenta el promedio de riesgo de riesgo de infección en **0.0682** cuando las demás variables se mantienen fijas

1.5. Coeficiente de determinación múltiple R^2

Calculamos que el modelo tiene un coeficiente de determinación múltiple $R^2 = 0.73688$, lo que significa que aproximadamente el 73.688 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Los parámetros con el P-valor más alto en el modelo fueron β_2,β_4 y β_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	24.923	X1 X2 X3 X4 X5
Modelo reducido	65.407	X2 X4 X5

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:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,48} \\ &= \frac{(65.407 - 24.923)/3}{0.519224} \\ &= 25.990067 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,48} = 2.7981$, se puede ver que $F_0 > f_{0.95,1,45}$ y por tanto se rechaza la hipotesis nula H_0 lo cual indica que los parámetros **sí** son significativos y **no es posible** descartar las Covariables del subconjunto.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si ¿El valor de la variable β_2 es equivalente al valor de β_1 y el valor de β_4 es equivalente al β_3? por consiguiente se plantea la siguiente prueba de hipótesis:

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

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

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

El modelo reducido está dado por:

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

$$Y_i = \beta_o + \beta_2 X^*_{2i} + \beta_4 X^*_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,48} \tag{3}$$

$$F_0 = \frac{(SSE(MR) - 24.9227)/2}{0.519224)} \overset{H_0}{\sim} f_{2,48} \tag{4}$$

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

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

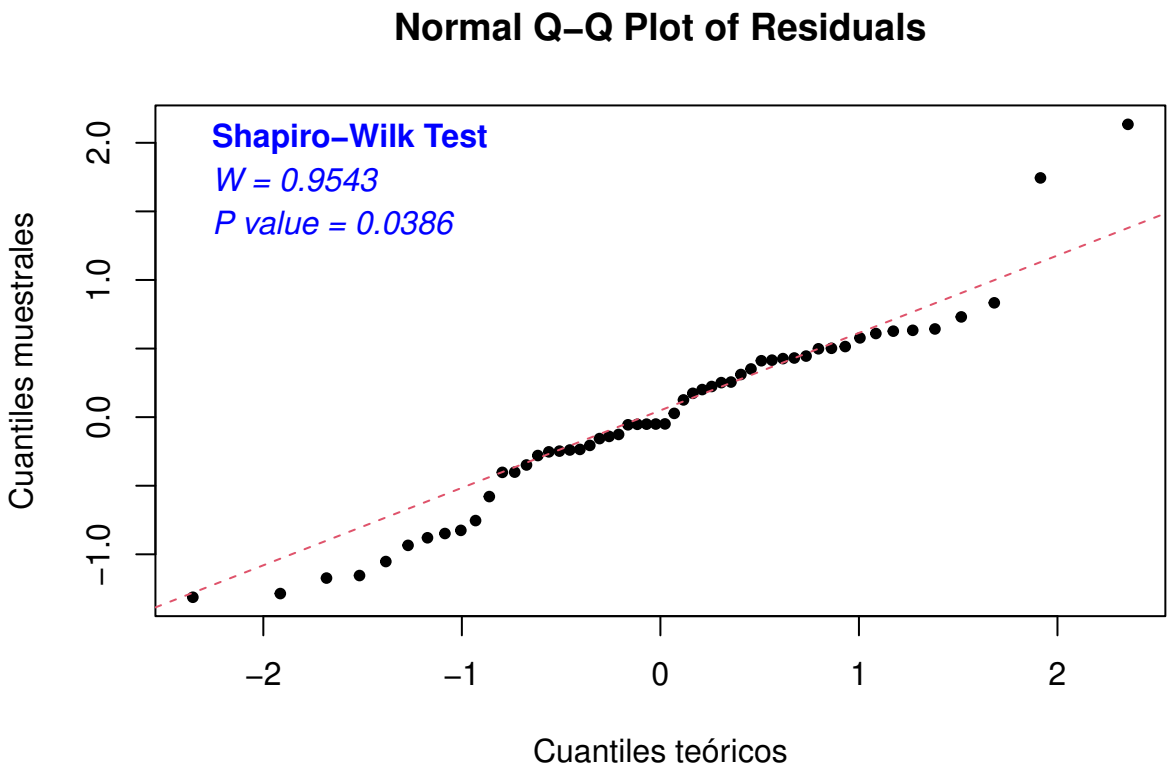

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

El valor p es de 0.0386, que en comparación con el $\alpha = 0.05$, es menor. Esto podría indicarnos que la prueba de hipótesis con la que intentamos comprobar la normalidad de los errores, debe ser rechazada. Además, es posible observar desde el criterio gráfico, que el patrón es irregular y las colas son pesadas. Tanto con la gráfica como con el valor p, es posible concluir que los errores **no distribuyen de forma normal**

Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

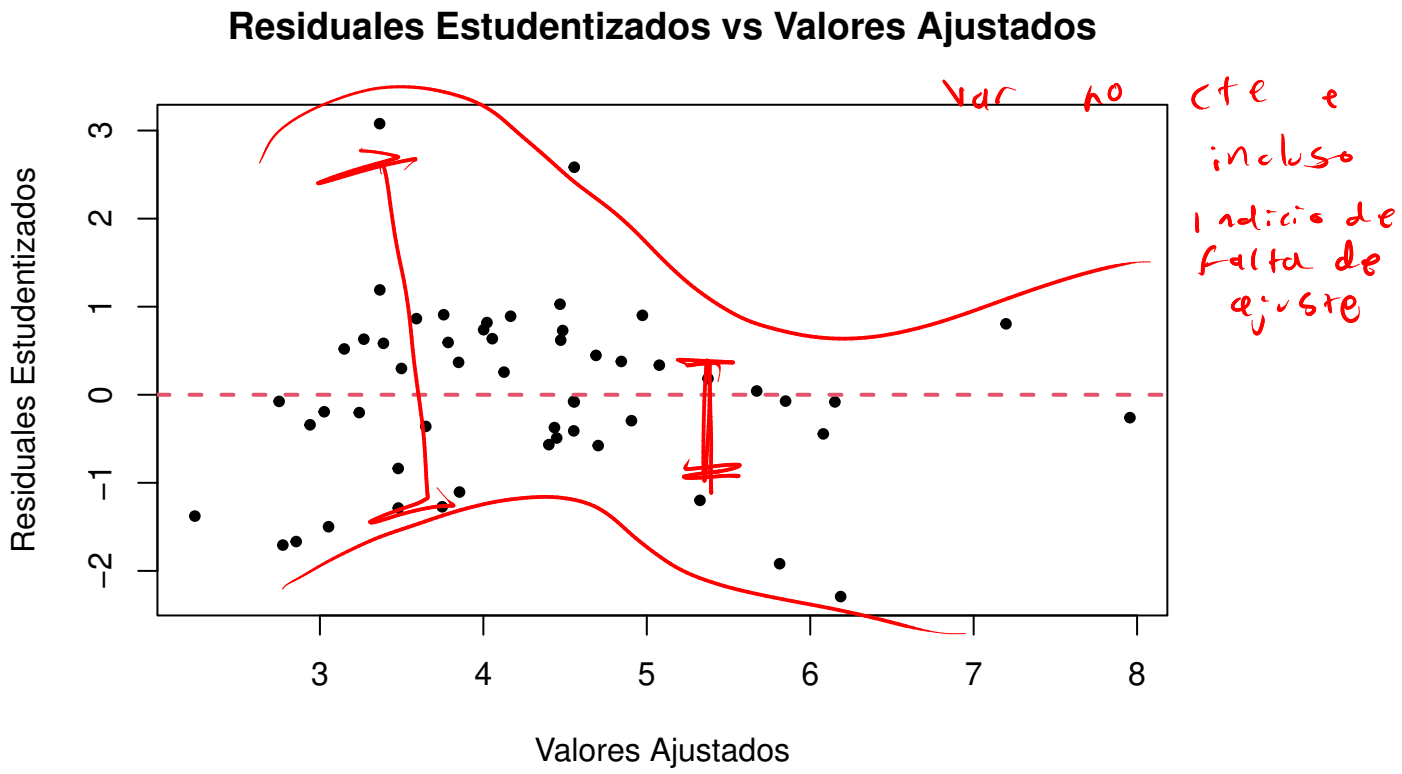

Figura 2: Gráfico residuales estudentizados vs valores ajustados

De la gráfica anterior, podemos determinar que existe un patrón de dispersión y se puede notar también que existen observaciones atípicas en el gráfico.

Por tanto, esto lleva determinar una varianza **no constante**. Por lo tanto este supuesto no se cumple.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

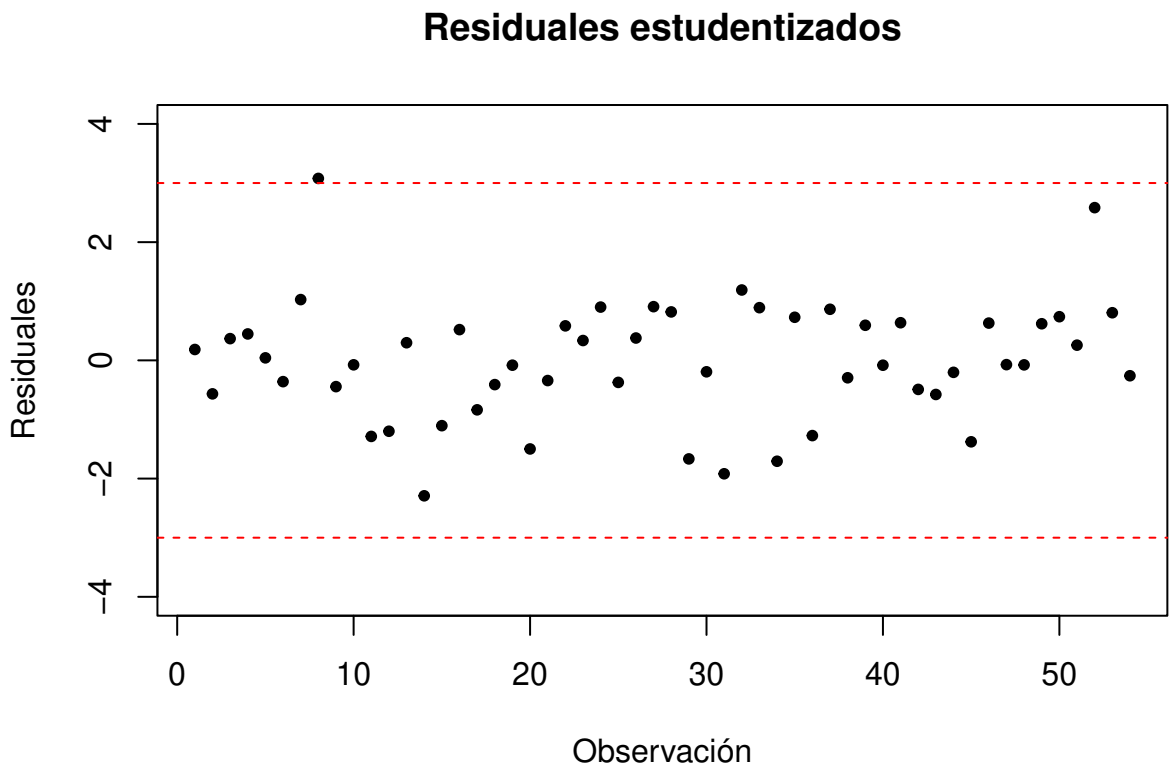

Figura 3: Identificación de datos atípicos

```
##   res.stud Cooks.D hii.value Dffits
## 8    3.078  0.1258    0.0738 0.9597
```

Analizando la gráfica Residuales estudentizados, podemos observar que un dato se sale del límite de la gráfica, pero además según el criterio que nos dice $|r_{estud}| > 3$ no permite confirmar que en efecto sólo hay un dato atípico en el modelo.

4.2.2. Puntos de balanceo

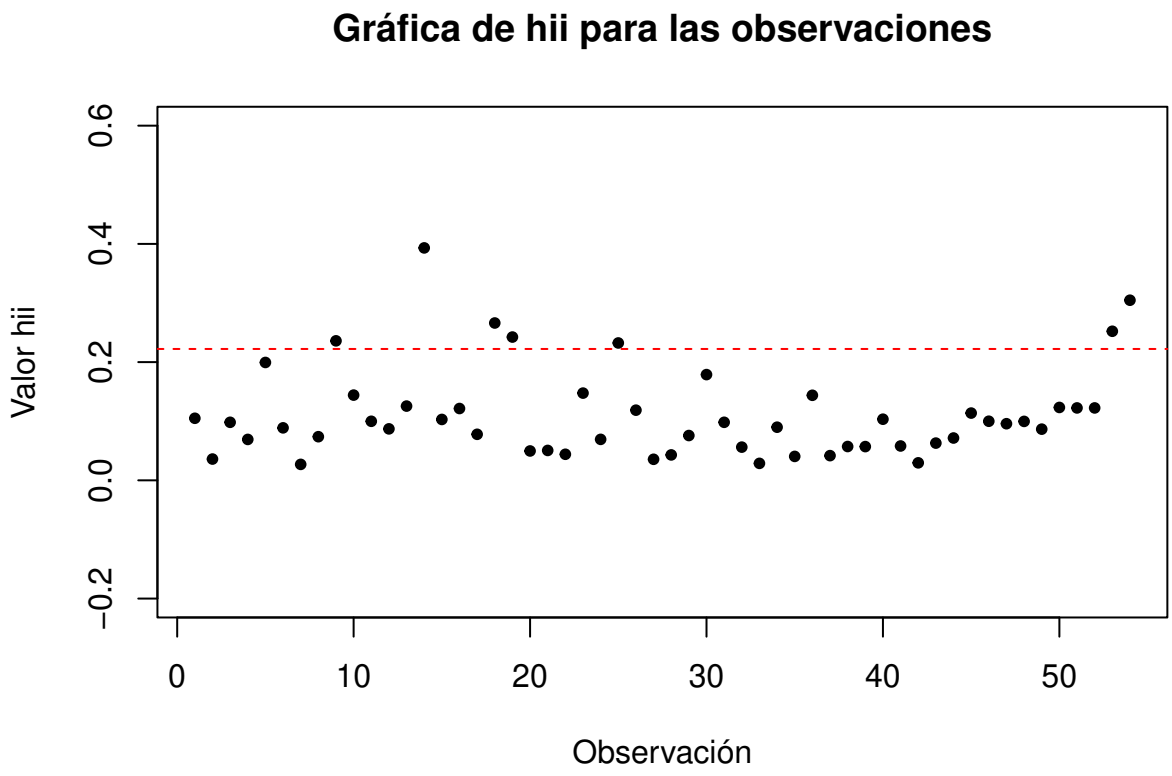

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 9   -0.4445  0.0102    0.2360 -0.2450
## 14  -2.2915  0.5670    0.3932 -1.9341
## 18  -0.4100  0.0102    0.2662 -0.2448
## 19  -0.0815  0.0004    0.2424 -0.0456
## 25  -0.3726  0.0070    0.2324 -0.2031
## 53   0.8055  0.0365    0.2522  0.4660
## 54  -0.2601  0.0049    0.3047 -0.1705
```

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.

4.2.3. Puntos influenciales

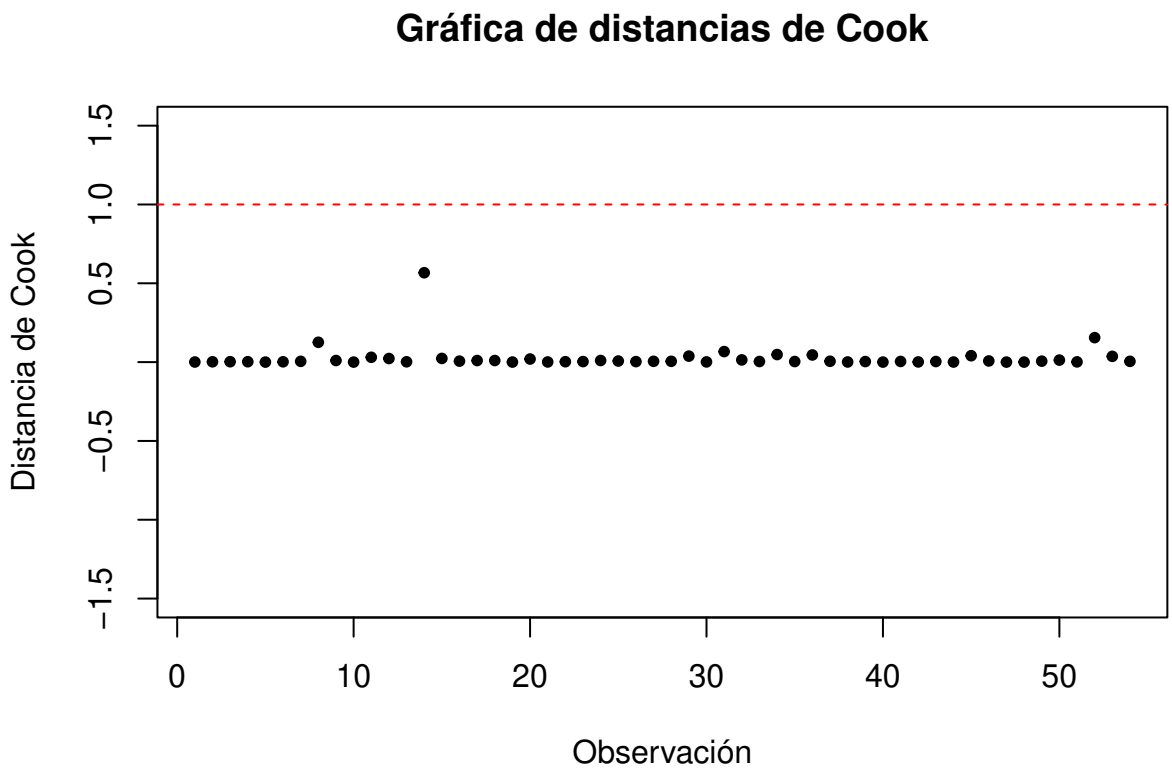

Figura 5: Criterio distancias de Cook para puntos influenciales

El criterio de cook dice que la observación i será influencial si $D_i > 1$, por lo que podemos concluir que no hay puntos influenciales.

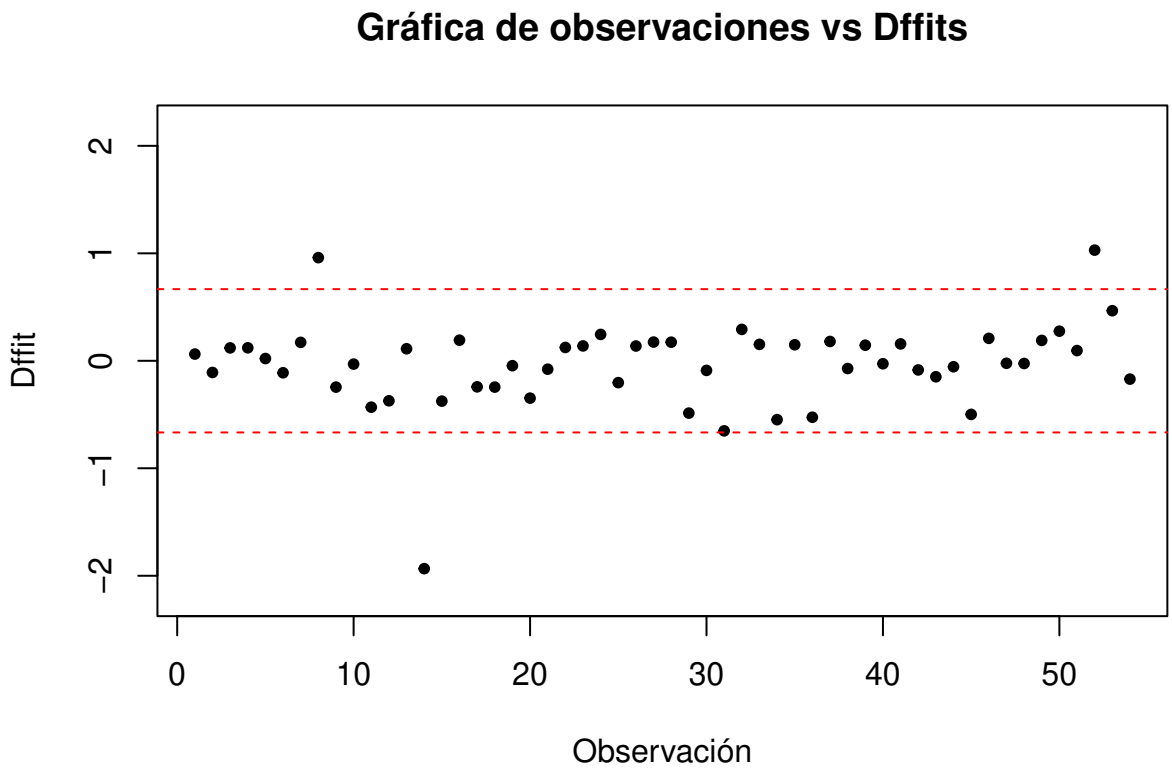

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 8    3.0780  0.1258    0.0738  0.9597
## 14  -2.2915  0.5670    0.3932 -1.9341
## 52   2.5837  0.1553    0.1225  1.0296
```

causan...! 3pt

A partir de la gráfica anterior y del criterio que indica que cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial, es posible afirmar que los puntos 8, 14 y 52 cumplen.

4.3. Conclusión

2,5pt redacción ...

Al realizar los analisis, se encuentra que los supuestos del modelo, tales como el supuesto de normalidad de los residuales, hayando que **no se cumple.**

Se encuentra tambien, que la varianza **no es contante** Además, se detectaron gran cantidad de puntos de balanceo, lo que sugiere que el modelo no es el más adecuado para describir el comportamiento de la variable respuesta Y : *Riesgo de infección*

Por tanto **no se acepta** el modelo enontrado de regresion lineal multiple.

5 No se pide aceptar o rechazar el modelo