Trabajo 1

3,8

Estudiantes

Cristian Camilo Rojas Hurtado
Juan Pablo Chalarca Jaramillo
Jhon Fernando Castellar Florez
Mario Javier Mosquera Vidales

Equipo 62

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

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

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

Dónde tenemos como variables en el modelo:

- 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	-0.7836
β_1	0.2133
β_2	0.0158
β_3	0.0634
β_4	0.0101
β_5	0.0018

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.7836 + 0.2133 X_{1i} + 0.0158 X_{2i} + 0.0634 X_{3i} + 0.0101 X_{4i} + 0.0018 X_{5i};\ 1 \leqslant i \leqslant 54$$

1.2. Significancia de la regresión

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

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

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	62.1975	5	12.439494	12.7229	6.87795e-08
Error	46.9309	48	0.977726		

De la tabla Anova, se observa un valor P aproximadamente muy cercano a cero, siendo Vp $< \alpha$, por lo que se rechaza la hipótesis nula H_0 en la que $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto la regresión si 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	-0.7836	1.7701	-0.4427	0.6600
β_1	0.2133	0.0830	2.5715	0.0133
β_2	0.0158	0.0344	0.4600	0.6476
β_3	0.0634	0.0167	3.8024	0.0004
β_4	0.0101	0.0085	1.1973	0.2371
β_5	0.0018	0.0008	2.1454	0.0370

Los P-valores presentes en la anterior tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$ y haciendo el siguiente juego de hipótesis para cada parámetro:

6pt

$$\begin{cases} \mathrm{H}_0 : \beta_i = 0; \ \mathrm{i} = 1,2,3,4,5 \\ \mathrm{H}_1 : \beta_i \neq 0 \end{cases}$$

Los parámetros β_1, β_3 y β_5 son significativos rechazando H_0 y aceptando la hpótesis alternativa, pues sus P-valores son menores a α, en caso contrario, no rechazamos H_0 en β_2 y β_4 dado que su Valor-P es mayor que α.

1.4. Interpretación de los parámetros

1pt

En qué cantidad? Y cuand pas demás constantes

$\hat{\beta}_1$**:** Es la cantidad de días que está cada paciente en el hospital, lo cual puede afectar de forma importante en qué tanto tiempo están directamente relacionados con infección a las posibles enfermedades que haya ente los pacientes del hospital.

$\hat{\beta}_3$**:** La cantidad de camas aumenta la cantidad de personas que pueden estar albergadas dentro del hopital, pero también que estén prestos a una posible infección dentro de las instalaciones.

$\hat{\beta}_5$**:** El número de enfermeras dentro del hospital es relevante dado que estas por una parte pueden ser mismas transmisoras de enfermedades por el contacto con los diferentes pacientes y así llegar a infectar algunos, como también con los cuidados se interfiere en la mejora o no de cada uno, su timepo estadía y demás.

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5699$, lo que significa que aproximadamente el 57 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

2pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

más bajo, y eran 3

Las covariable con el P-valor más alto en el modelo fueron 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:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	46.931	X1 X2 X3 X4 X5
Modelo reducido	48.809	X1 X3 X5

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

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

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_1, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,48} \\ &= \frac{0.626}{0.977} \\ &= 0.640 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,48} = 2.7981$, se puede ver que $F_0 < f_{0.95,3,48}$ y por ello decimos que no se rechaza la hipótesis nula, dando a β_2, β_4 como no significativos.

Finalmente, concluimos que si se pueden descartar estas variables del modelo basados en la anterior prueba de hipótesis , quedando así con un modelo reducido.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea que el Número de camas con respecto al Riesgo de infección es igual a Número de enfermeras, como también que la Duración de la estadía con respecto al Riesgo de infección es dos veces la Rutina de cultivos, por lo que se plantea la siguiente prueba de hipótesis:

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

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

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

El modelo reducido está dado por:

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

Donde $X_{1i}^* = X_{1i} + \frac{X_{2i}}{2}$ y $X_{3i}^* = 3X_{1i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 46.9309))/2}{0.9777} \overset{H_0}{\sim} f_{2,48} \tag{3}$$

Dado que no se conoce el SSE del modelo reducido (MR) tenemos la anterior ecuación expresada como resultante en el estadístico de prueba F_0, además de que utilizamos 2 que es el número de ecuaciones con las que se trabajó la prueba de hipótesis.

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:

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

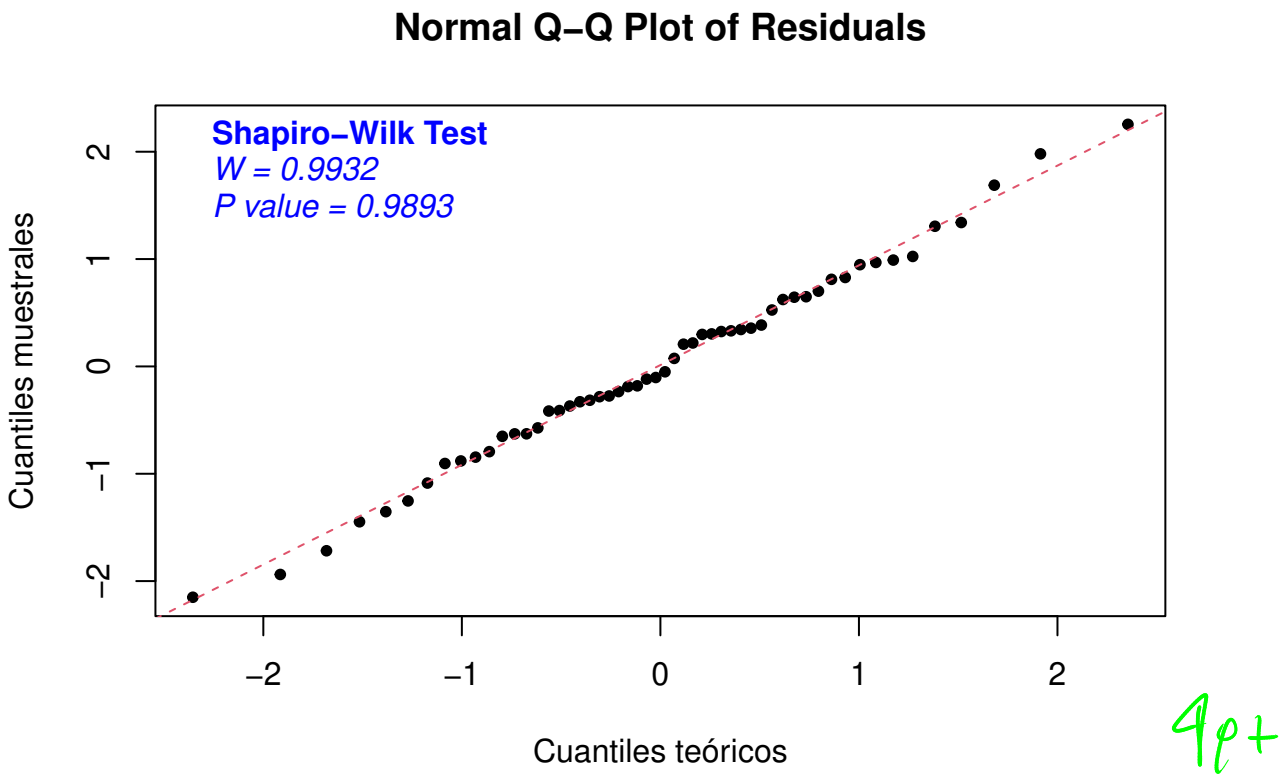

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.9893 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, de acuerdo a la gráfica de comparación de cuantiles podemos ver ciertas irregularidades, pero concluimos que no son suficientes para rechazar el cumplimiento del supuesto de normalidad. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

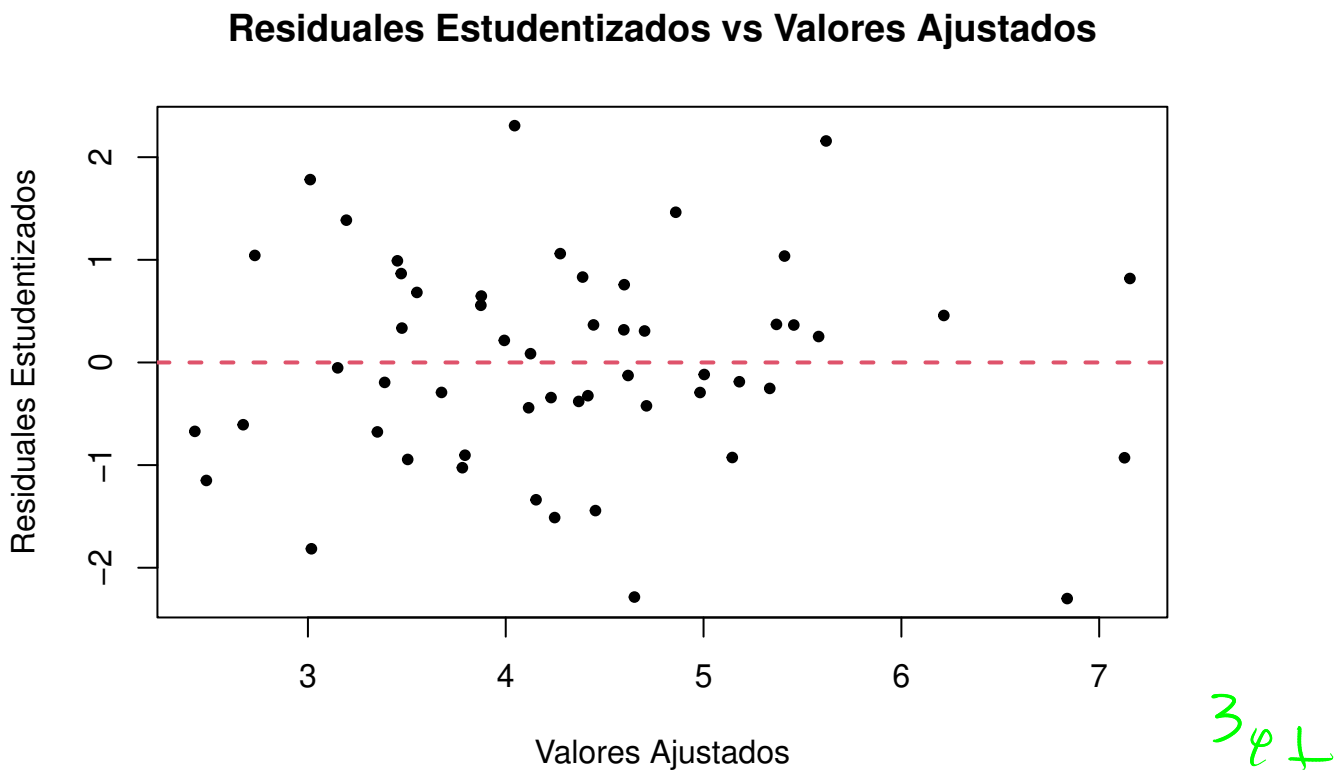

Figura 2: Gráfico residuales estudentizados vs valores ajustados

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 observar un patrón concluimos que no hay falta de ajuste y es posible observar media 0.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

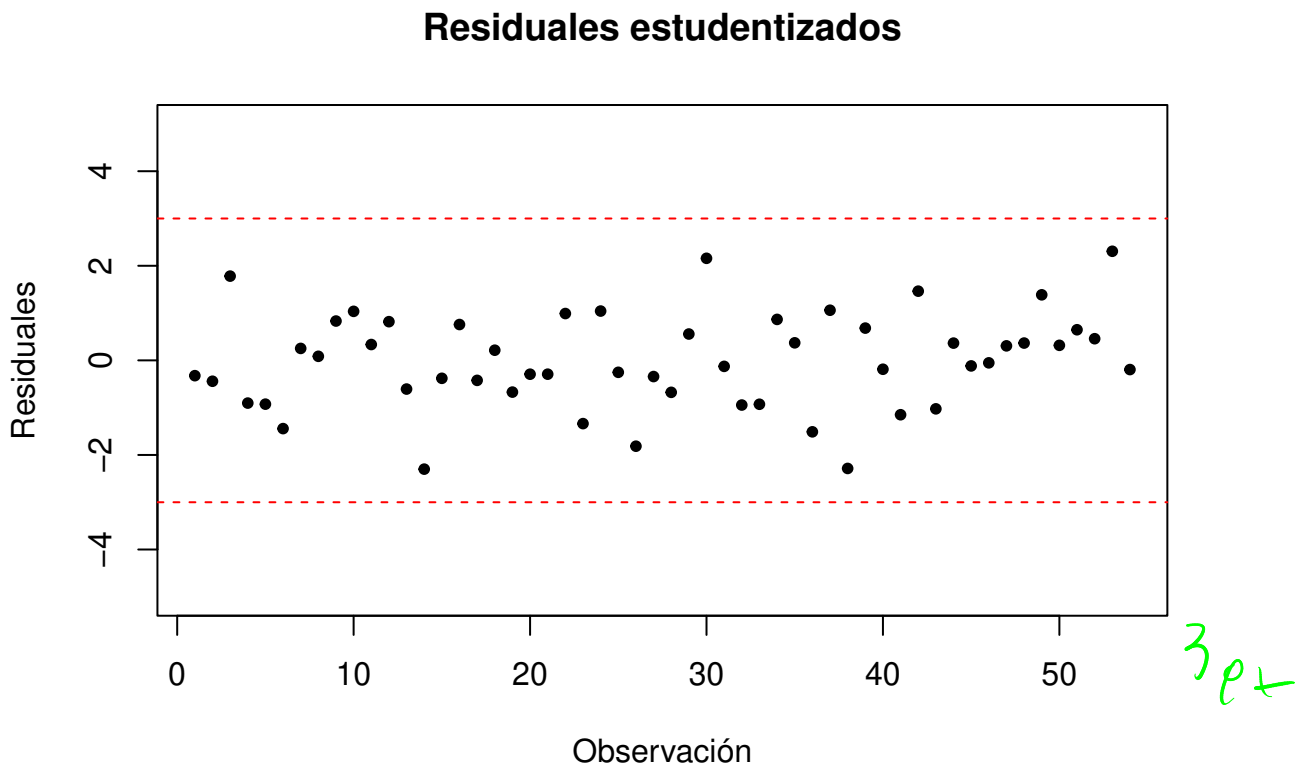

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$. es importante recordar que la falta de datos atípicos no significa necesariamente que los datos sean perfectos, y aún es fundamental realizar un análisis exploratorio y verificar que se cumplan las suposiciones del modelo para garantizar la validez de los resultados.

4.2.2. Puntos de balanceo

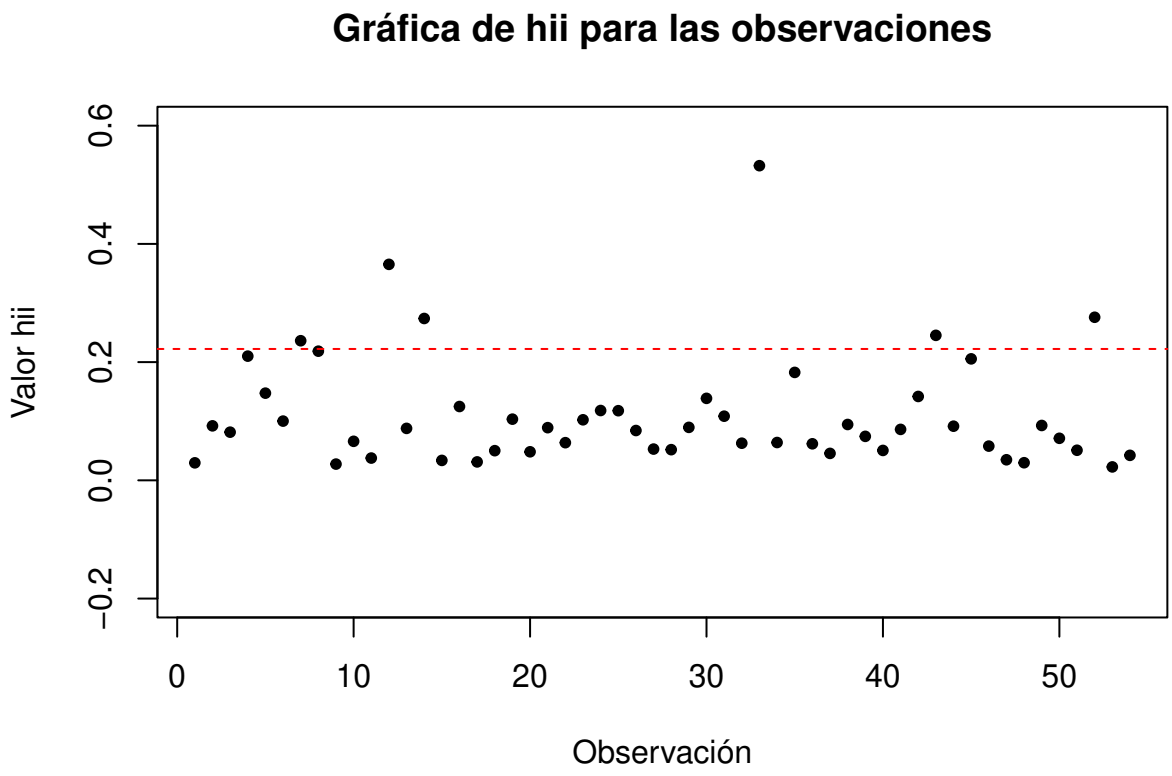

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 7    0.2525  0.0033    0.2362  0.1390
## 12   0.8182  0.0642    0.3654  0.6187
## 14  -2.3002  0.3325    0.2738 -1.4816
## 33  -0.9284  0.1635    0.5323 -0.9889
## 43  -1.0254  0.0569    0.2453 -0.5849
## 52   0.4579  0.0133    0.2759  0.2803
```

causan...?

2p+

Al examinar el gráfico que muestra las observaciones frente a los valores h_{ii}, notamos que la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$. En este gráfico, se identifican claramente 6 puntos de datos del conjunto que cumplen con el criterio en el cual $h_{ii} > 2\frac{p}{n}$. Estos puntos específicos se encuentran detallados en la tabla adjunta

4.2.3. Puntos influenciales

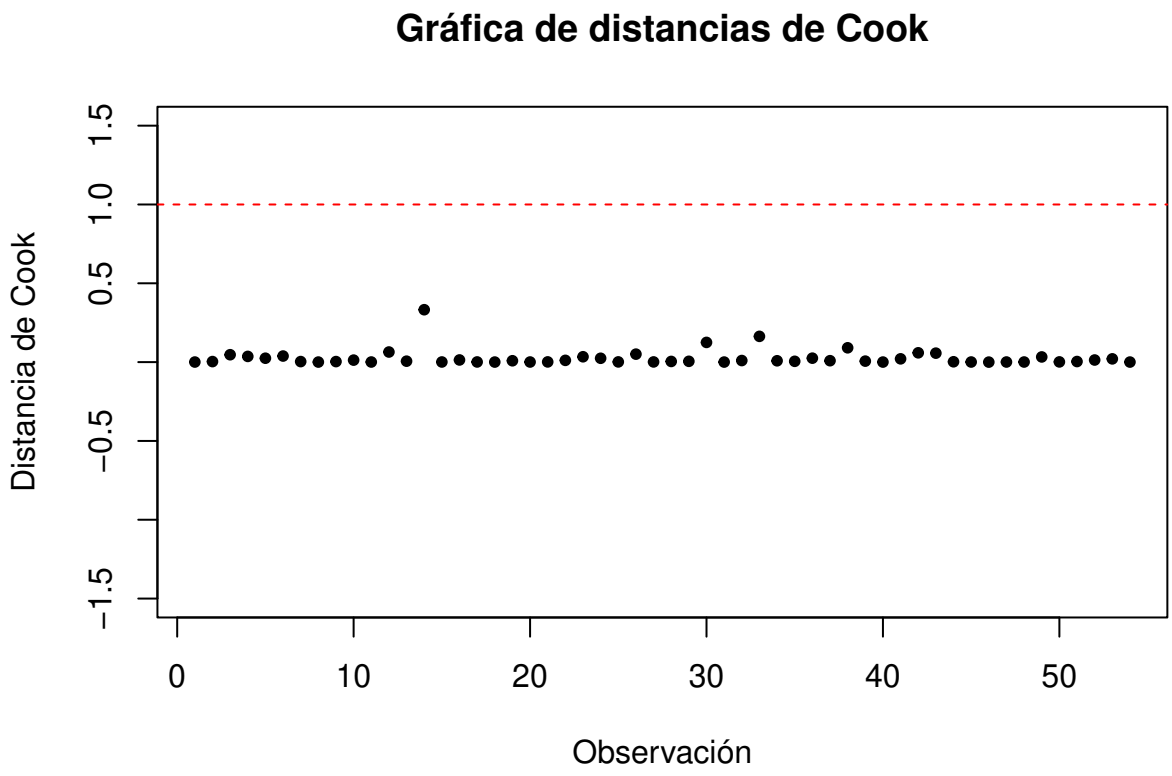

Figura 5: Criterio distancias de Cook para puntos influenciales

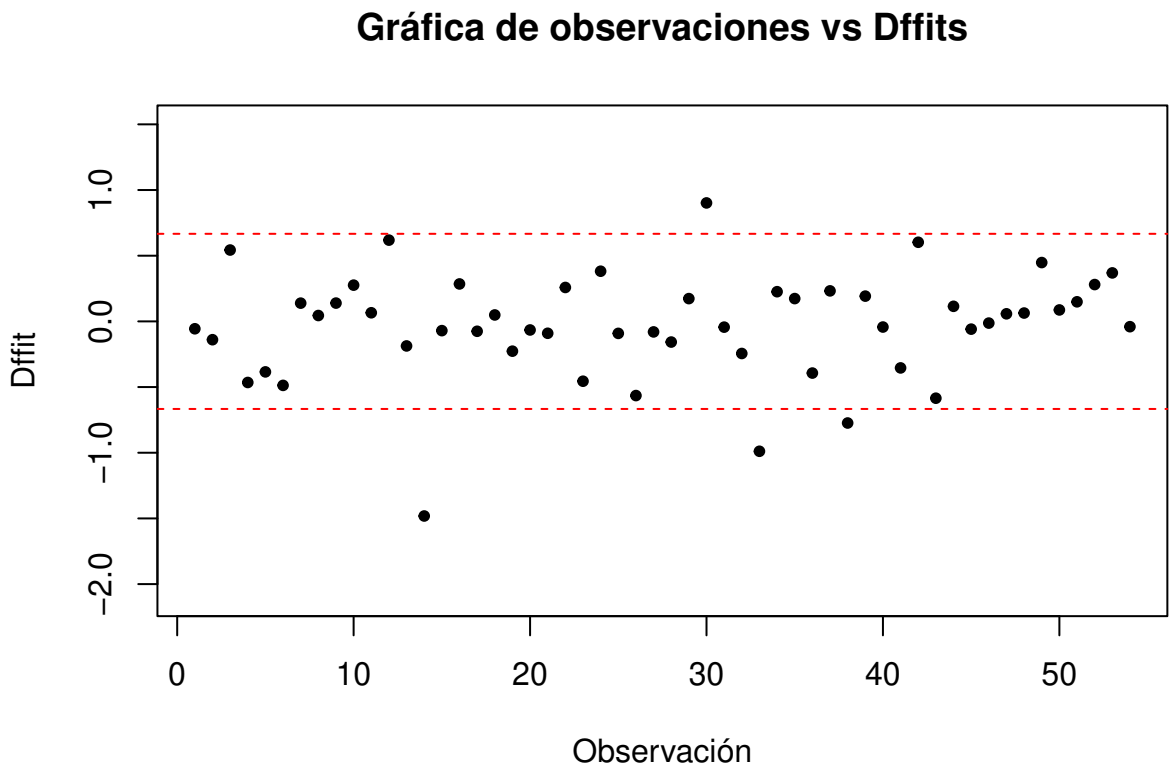

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 14  -2.3002  0.3325    0.2738 -1.4816
## 30   2.1580  0.1250    0.1387  0.9019
## 33  -0.9284  0.1635    0.5323 -0.9889
## 38  -2.2854  0.0908    0.0945 -0.7737
```

Como se puede ver,las observaciones ... 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.

Como se puede ver,las observaciones 14, 30, 33 y 38 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. esto sugiere que la observación puede ser influyente en términos de predicción, pero no necesariamente en la estructura general del modelo o en la estimación de los coeficientes de regresión.

4.3. Conclusión

Asumimos la validez del modelo por el cumplimiento de los supuestos, los puntos extremos pueden tener varias causas de acuerdo al contexto del problema, y no significan necesariamente errores o problemas en los datos, estos puntos extremos pueden señalarnos eventos inusuales y contener información importante teniendo en cuenta la naturaleza de la influencia de la observación del riesgo de infección frente a las variables predictoras