Trabajo 1

3,35

Estudiantes

María Camila Gutiérrez Ruiz

Andrea Carolina Vergara Baquero

Jhonatan Efrén Ortiz Rodríguez

Bryan Andres Garcia Villa

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

h

Sede Medellín

30 de Marzo 2023

Índice

Índice de figuras

Índice de cuadros

Pregunta 1

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

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

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	0.5504
β_1	0.2601
β_2	-0.0124
β_3	0.0503
β_4	0.0106
β_5	0.0014

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.5504 + 0.2601 X_{1i} - 0.0124 X_{2i} + 0.0503 X_{3i} + 0.0106 X_{4i} + 0.0014 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

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:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,44} \qquad (1)$$

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	63.7643	5	12.752866	13.1302	7.69677e-08
Error	42.7357	44	0.971265		

si lo ponen así es individual

De la tabla Anova, se observa un valor P aproximadamente igual a 0, por lo que 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

5 pt

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	0.5504	1.9352	0.2844	0.7774
β_1	0.2601	0.0857	3.0363	0.0040
β_2	-0.0124	0.0364	-0.3400	0.7355
β_3	0.0503	0.0177	2.8469	0.0067
β_4	0.0106	0.0092	1.1453	0.2583
β_5	0.0014	0.0008	1.8086	0.0773

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 lo tanto, se concluye que los β_2, β_4 y β_5 no son significativos pues sus P-valores son mayores al $\alpha = 0.05$.

¿y p0?

1.4 Interpretación de los parámetros

2 pt

$\hat{\beta}_1$: Con un valor- P pequeño se rechaza la hipótesis nula que indica que este parámetro es igual a cero por lo tanto es significativo. La estimación de $\hat{\beta}_1 = 0.2601$ lo que quiere decir que por un aumento en el promedio de la duración de la estadía se espera que se incremente en un 0.2601 el riesgo de infección, siempre que las otras covariables permanezcan constantes.

$\hat{\beta}_3$: Con un valor-P pequeño se rechaza la hipótesis nula que indica que este parámetro es igual a cero por lo tanto es significativo. La estimación de $\hat{\beta}_3 = 0.0503$ quiere decir que por un aumento en el promedio de el numero de camas se espera que se incremente en un 0.0503 el riesgo de infección, siempre que las otras covariables permanezcan constantes.

↳ porcentaje promedio

1.5 Coeficiente de determinación múltiple R^2

3 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5987$, lo que significa que aproximadamente el 59.87 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el este informe.

¿cómo se calcula?

Pregunta 2

2.1 Planteamiento pruebas de hipótesis y modelo reducido

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	42.736	X1 X2 X3 X4 X5
Modelo reducido	48.866	X1 X3

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 55$$

2.2 Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} F_{3,44} \\ &= \frac{(48.86 - 42.73)/3}{0.97} \\ &= 2.104 \end{aligned} \tag{2}$$

Ahora, comparando el $F_0 = 2.104$ con $f_{0.95,3,44} = 2.790$ se puede concluir que el F_0 es menor que f_0 por lo tanto la hipótesis nula que es $\beta_2{=}\beta_4{=}\beta_5{=}0$ no se rechaza, lo que quiere decir que bajo esta prueba para subconjuntos de variables ninguno de los parámetros es significativo y pueden ser sacados del modelo.

Pregunta 3

3.1 Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

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

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

Con $\mathbf{L}$ dada por

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_3 X^*_{3i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

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

3.2 Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,44} \tag{3}$$

$$F_0 = \frac{(SSE(MR) - (42.73))/2}{(0.97)} \overset{H_0}{\sim} f_{2,44} \tag{4}$$

El SSE (MR) no se puede calcular debido a que es derivado a las variables dependientes X_n

Pregunta 4

Supuestos del modelo

Normalidad de los residuales

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

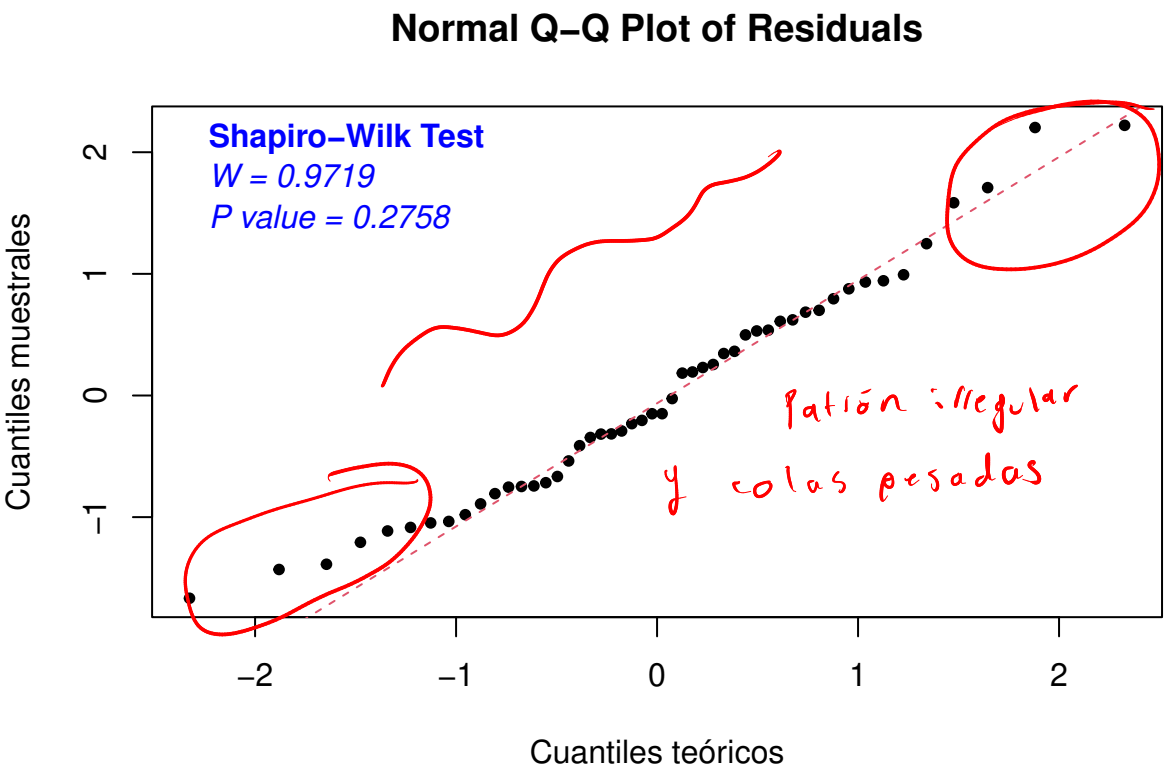

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

En este caso, el valor P obtenido es 0.2758, lo que indica que no hay suficiente evidencia para rechazar la hipótesis nula de que los datos siguen una distribución normal. Por lo tanto, se puede concluir que los datos podrían seguir una distribución normal. X

No hicieron análisis gráfico que es aún más importante y el vp ya es un poco bajo y les da indicio de rechazar

Varianza constante 2,5 pt

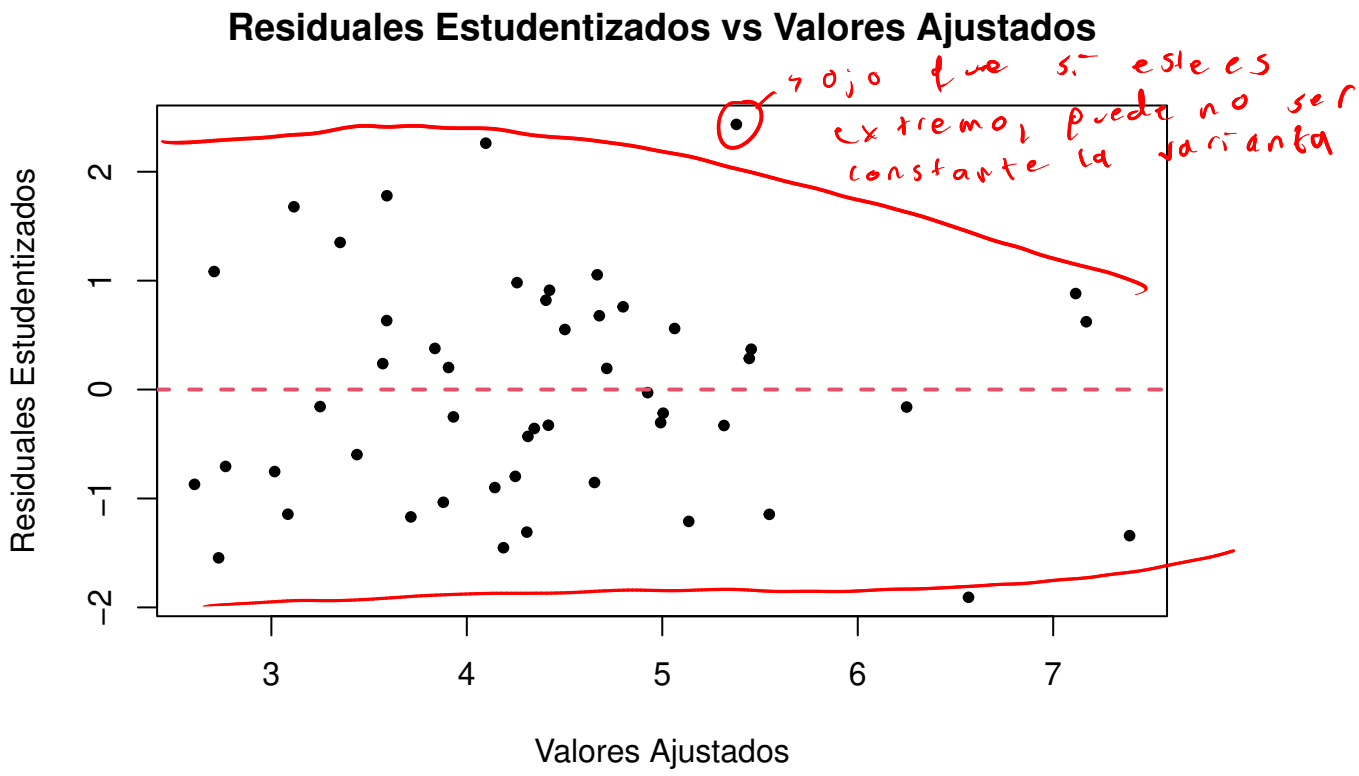

Figura 2: Gráfico residuales estudentizados vs valores ajustados

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. ✓

Verificación de las observaciones

Datos atípicos

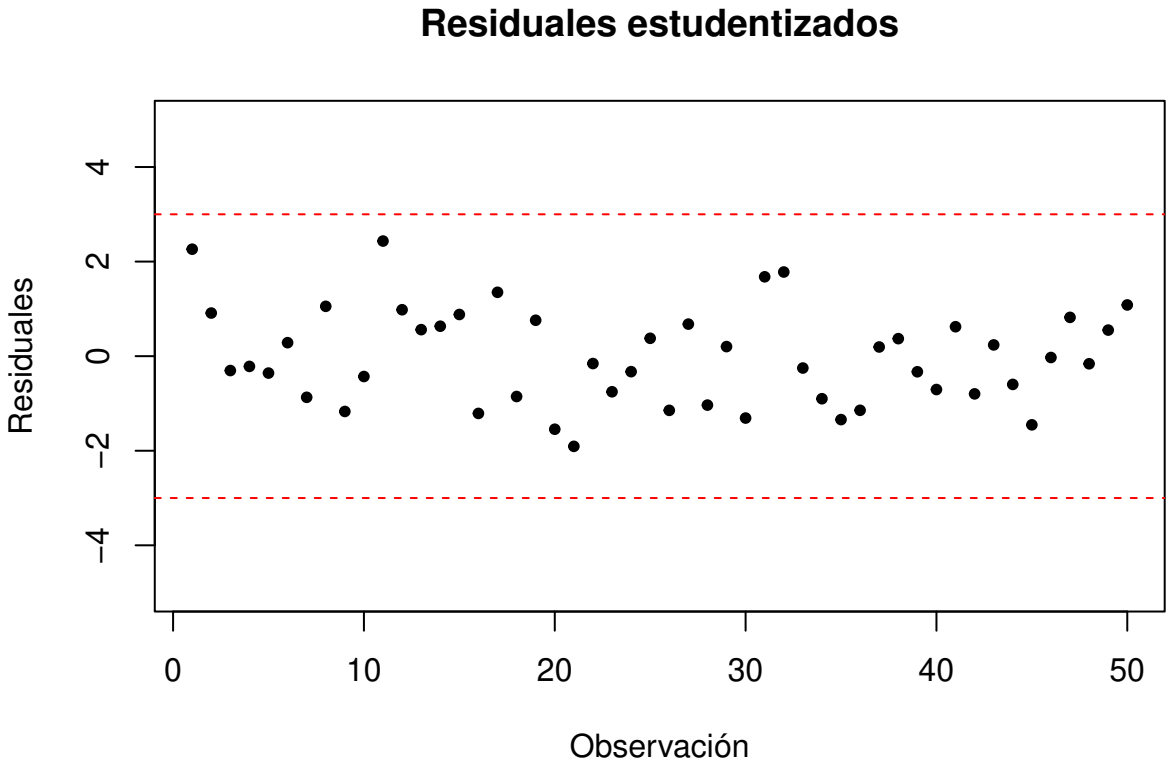

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

Puntos de balanceo

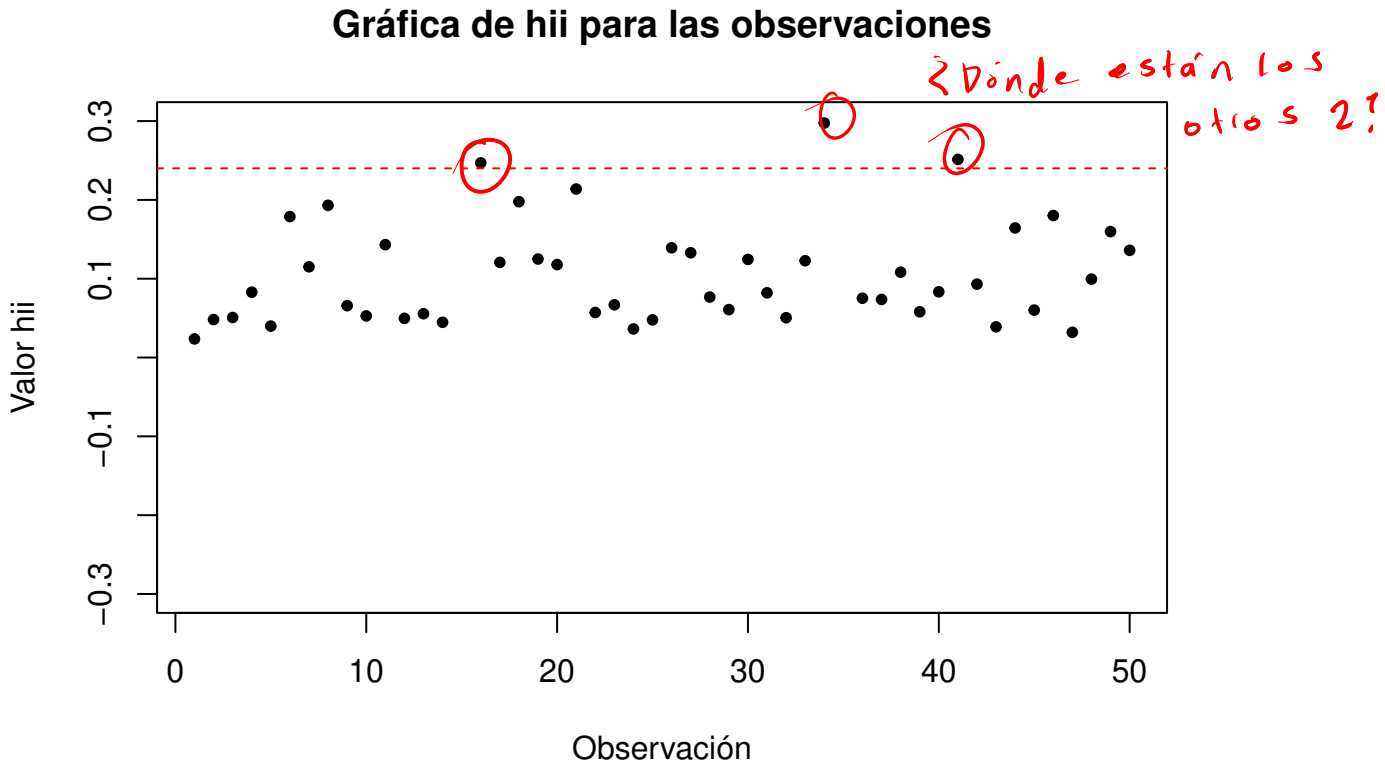

Figura 4: Identificación de puntos de balanceo

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 5 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.

Puntos influenciales

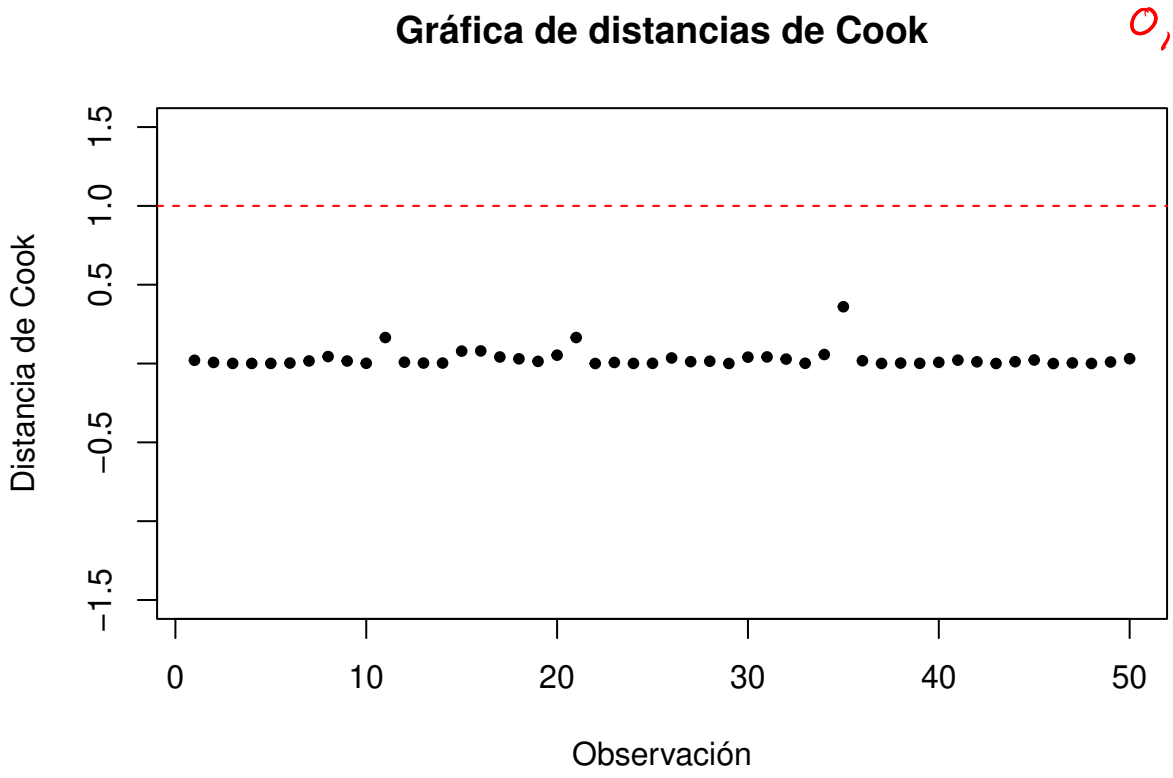

Figura 5: Criterio distancias de Cook para puntos influenciales

Para la validez del modelo se puede decir que se hallaron varios puntos de balanceo, lo que hace variar el coeficiente R^2 ocasionando un aumento de su valor por parte de las variables predictoras al riesgo de infección. Sin embargo, los puntos influénciales afectan el modelo haciendo un efecto mayor sobre la recta de regresión jalando el modelo en su dirección, lo que puede conllevar a generar errores en las variables de predicción sobre el riesgo de infección. El modelo podría generar inconsistencias por las razones mencionadas anteriormente.

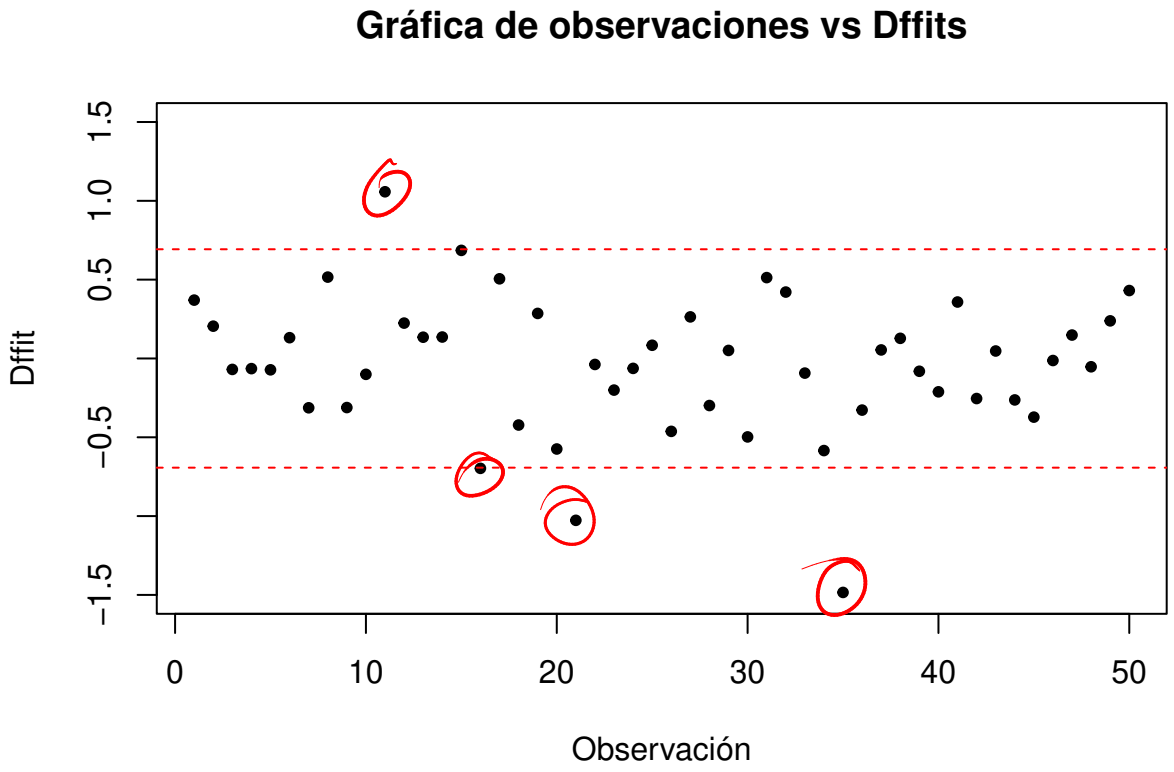

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 11   2.4344  0.1650     0.1432  1.0575
## 16  -1.2107  0.0801     0.2470 -0.6972
## 21  -1.9074  0.1649     0.2138 -1.0268
## 35  -1.3415  0.3604     0.5458 -1.4844
```

El criterio D_{ffits} es una medida que se utiliza para evaluar la influencia de cada observación en el modelo de regresión. En general, un valor de D_{ffits} mayor que 1 o menor que -1 indica que una observación puede tener un impacto significativo en los resultados del modelo y puede ser considerada como un punto influyente.

En este caso, los valores de D_{ffits} para las observaciones 11, 16, 21 y 35 son 1.0575, -0.6972, -1.0268 y -1.4844 respectivamente, lo que indica que la observación 11 puede tener un impacto significativo en el modelo, mientras que las observaciones 16, 21 y 35 podrían ser consideradas como puntos influyentes

Conclusión

1. El coeficiente de determinación múltiple R^2 es una medida que indica la proporción de la variabilidad total en la variable de respuesta que puede ser explicada por el modelo de regresión. Cuanto mayor sea el valor de R^2 , mayor será la capacidad del modelo para explicar la variabilidad en la variable de respuesta. Por lo tanto, al tener un alto valor de R^2 , podemos concluir que el modelo de regresión propuesto es capaz de explicar una

gran parte de la variabilidad en la variable de respuesta. Sin embargo, esto no significa que el modelo sea perfecto o que explique la totalidad de la variabilidad en la respuesta, ya que siempre puede haber otros factores no considerados en el modelo que también estén afectando la variable de respuesta. ✓

2. Con base al estadístico de prueba para subconjuntos de variables, se podría concluir que es posible sacar del modelo los parámetros β_2 , β_4 y β_5 sin que esto afecte significativamente la capacidad del modelo para explicar la variabilidad en la variable de respuesta. ✓

No dan respuesta a si el modelo es válido o no