4,3

Trabajo 1

Estudiantes

Brayan Alberto Patiño Alzate
Jenifer Tatiana Atehortua Duque
Tomas Villa Machado
Equipo # 23

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

Sede Medellín
30 de Marzo de 2023

Índice

Índice de figuras

Índice de tablas

1. Pregunta 1

Teniendo en cuenta la base de datos asignada, se realiza el ajuste del modelo de regresión lineal múltiple (RLM), explicando la eficacia sobre el control de infecciones hospitalarias 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:

Tabla 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-1.3940
β_1	0.1174
β_2	0.0298
β_3	0.0768
β_4	0.0158
β_5	0.0022

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.394 + 0.1174 X_{1i} 0.0298 X_{2i} + 0.0768 X_{3i} + 0.0158 X_{4i} + 0.0022 X_{5i}$$

1.2. Significancia de la regresión

Para analizar la significancia de la regresión, se plantea la siguiente prueba 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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,44} \tag{1}$$

$$F_0 = \frac{13.471360}{0.975073} = 13.8157 \tag{2}$$

Ahora, se presenta la tabla Anova:

Tabla 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	67.3568	5	13.471360	13.8157	4.02125e-08
Error	42.9032	44	0.975073		

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 $\beta_j \neq 0$, en este caso el P-valor nos permite concluir que sí, el modelo de regresión es significativo, rechazando la hipótesis nula planteada anteriormente, esto quiere decir que alguna de las variables predicctorias es significativa en el riesgo de adquirir infección en hospitales.

1.3. Significancia de los parámetros

6 pt

En el siguiente cuadro respecto a los parametros individuales se presenta información de los los mismos, lo cual permitirá determinar cuáles de ellos son significativos.

Tabla 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-1.3940	1.6333	-0.8535	0.3980
β_1	0.1174	0.0828	1.4184	0.1631
β_2	0.0298	0.0321	0.9288	0.3580
β_3	0.0768	0.0169	4.5399	0.0000
β_4	0.0158	0.0078	2.0150	0.0500
β_5	0.0022	0.0007	2.9178	0.0055

→ Brutal :O

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_3 y β_5 son significativos, Los P-valores mostrados en la tabla me indican cuales de los parámetros está aportando significativamente al modelo de regresión, ya que sus P-valores 0.0000 y 0.0055 respectivamente son menores al valor alfa.

1.4. Interpretación de los parámetros

2 pt

Con lo anterior se puede determinar que $\hat{\beta}_3$: 0.0768, indica que por cada unidad que aumente el promedio de camas en el hospital X_3, el riesgo de infección aumenta 0.0768 unidades, cuando las demás variables predictoras se mantienen fijas.

→ probabilidad promedio

Asi mismo $\hat{\beta}_5$: 0.0022 indica que por cada unidad que aumente el promedio de enfermeras en el hospital X_5, el riesgo de infección aumenta 0.0022 unidades, cuando las demás variables predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

3 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6108906$, lo que significa que aproximadamente el 61.08 % de la variabilidad total observada en los resultados de la prueba de riesgo es explicada por el modelo de regresión propuesto en el presente informe.

¿Cómo se calcula?

2. Pregunta 2

9 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más alto en el modelo fueron, la duracion de la estadia X_1 (Valor P= 0.1631),rutina de cultivos X_2 (Valor P= 0.3580) y censo promedio diario X_4 (Valor P= 0.0500), ahora se

procede a plantear la prueba de hipótesis para verificar la significancia en simultaneo del subconjunto con valores p más altos del modelo de regresión.

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_4 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 4 \end{cases}$$

Tabla 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	42.903	X1 X2 X3 X4 X5
Modelo reducido	55.477	X3 X5

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,44} \\ &= \frac{(55.477 - 42.9032)/3}{0.975073} \\ &= 4.298413 \end{aligned} \tag{3}$$

Ahora, comparando el F_0 con $f_{0.95,3,44} = 2.8165$, se puede ver que $F_0 > f_{0.95,3,44}$

Por consiguiente, con un nivel de significancia de 0.05 comparando F_0 con $f_{0.95,3,44=4.0085}$, se puede ver que $F_0 > f_{0.95,3,44}$, como F_0 esta dentro de la region de rechazo, se rechaza H_0 y se concluye que las variables X_1, X_2 y X_4 afectan de forma conjunta a la probabilidad de riesgo de infección.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

¿Será que las variables predictoras X1 y X3; X2 y X4 presentan colinealidad en el modelo establecido? Por consiguiente se plantea la siguiente prueba de hipótesis:

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

reescribiendo matricialmente:

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

1,5 pt

Con **L** dada por

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

El modelo reducido está dado por:

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

1 pt

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

2 pt

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

$$F_0 = \frac{(SSE(MR) - 42.9032/2}{0.975073} \tag{5}$$

4. Pregunta 4

15,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

4 pt

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

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

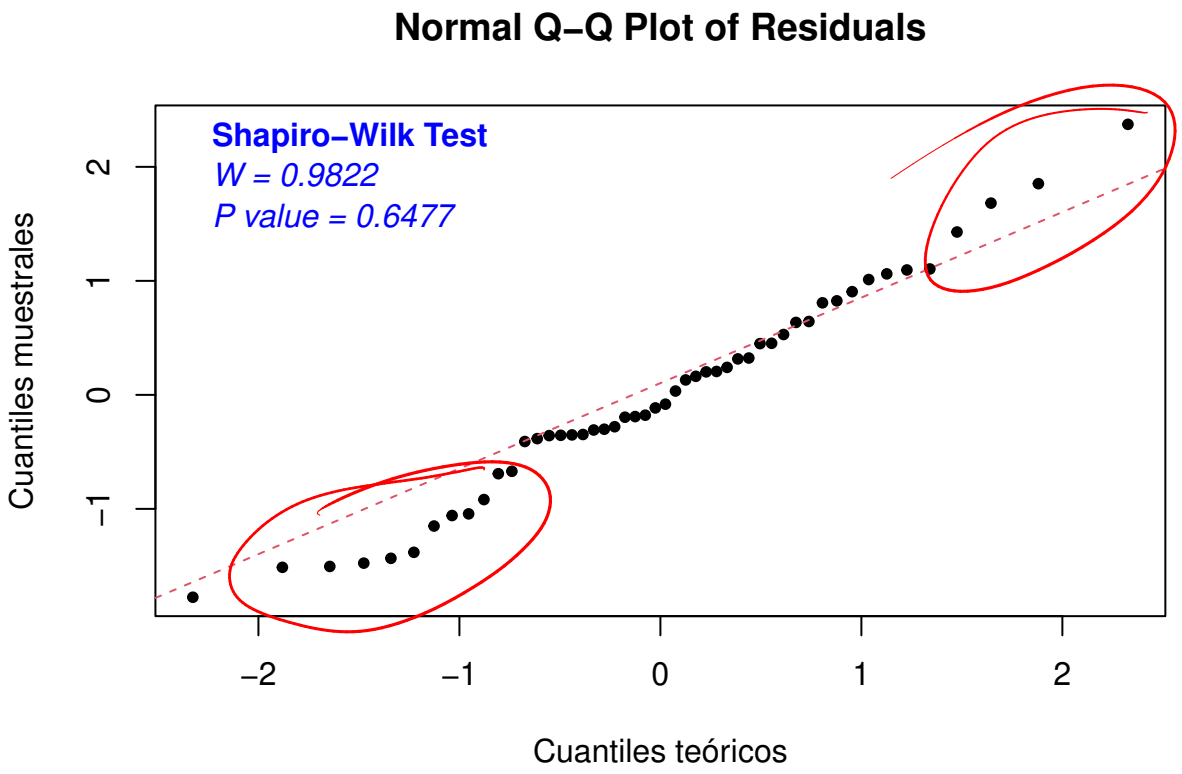

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Normalidad

Tomando en cuenta el valor p no se rechaza H_0, se podría decir que hay normalidad, sin embargo la gráfica nos muestra patrones irregulares, además las colas están alejadas de la línea de tendencia marcada en la gráfica lo que supone que el supuesto de normalidad no se este cumpliendo como lo indica el valor p, por consiguiente al tener mas peso el método grafico frente al valor se rechaza el supuesto de normalidad por medio del grafico de cuanti – cuantil. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

1,5 pt

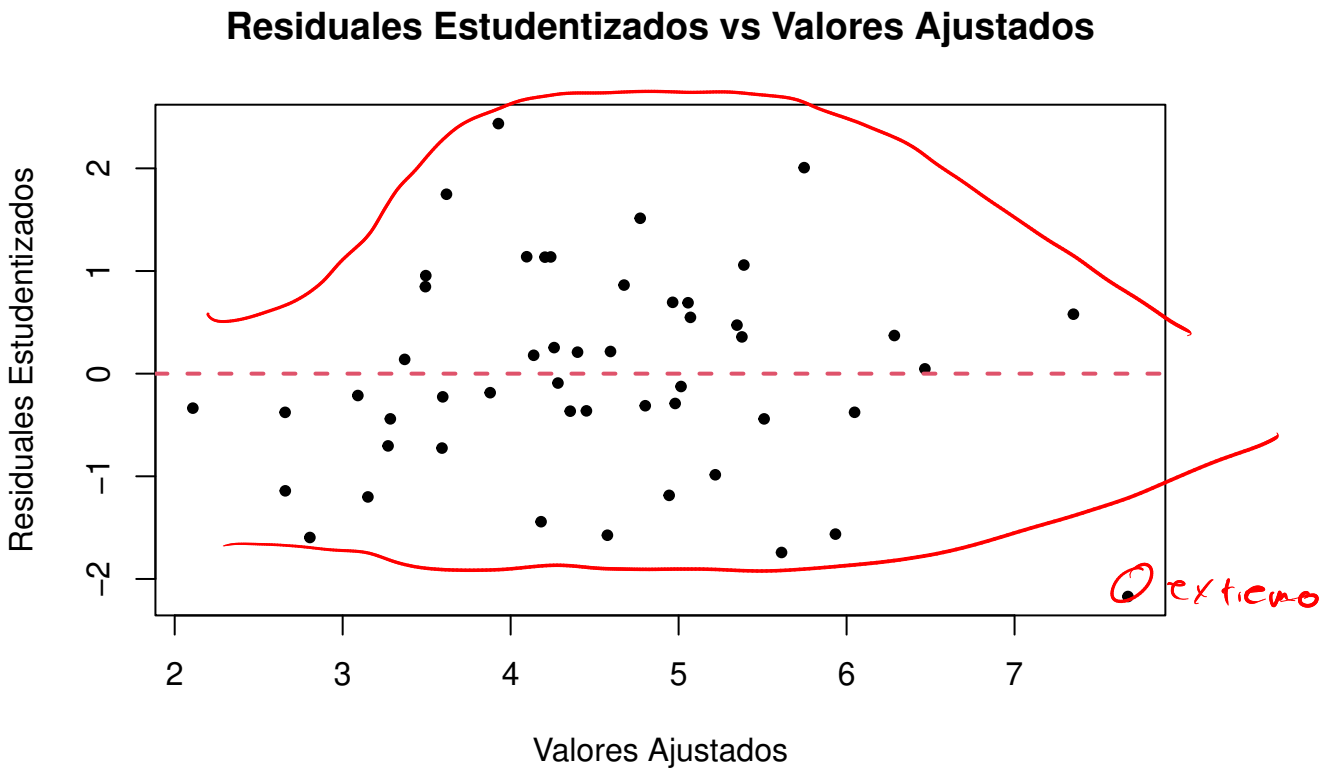

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Varianza En la gráfica observamos que la varianza no tiene una tendencia marcada hacia la linealidad, o hacia valores constantes, ya que los datos se muestran dispersos a lo largo de la gráfica, se puede decir que no se cumple el supuesto de varianza constante.

X

La varianza efectivamente no se cumple pero el análisis hecho no explica este hecho lo suficientemente claro

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

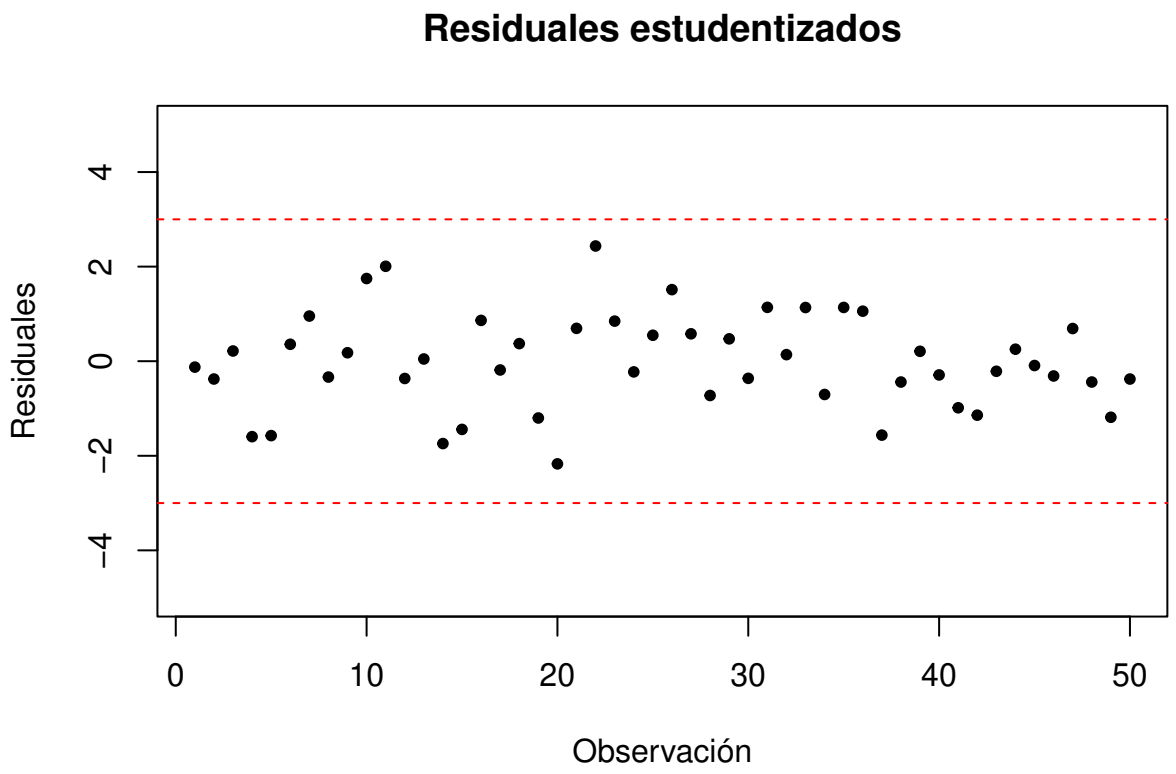

Figura 3: Identificación de datos atípicos

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

2 pt

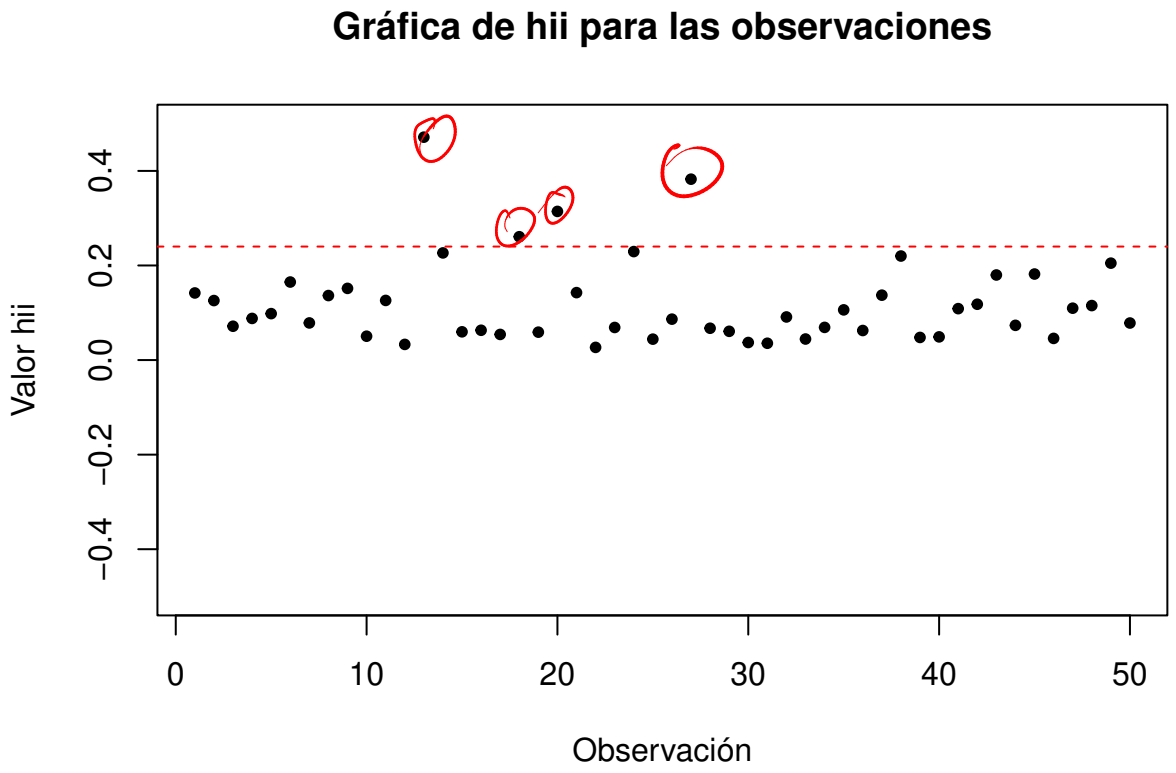

Figura 4: Identificación de puntos de balanceo

Tabla 5: Resumen de diagnostico

	$Res.stud$	$Cook.D_i$	$h_{ii} value$	$Dffits$
13	0.0469	0.0003	0.4713	0.0438
18	0.3716	0.0081	0.2608	0.2186
20	-2.1709	0.3600	0.3143	-1.5375
27	0.5793	0.0346	0.3825	0.4525

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.24$, se puede apreciar que existen 4 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$, son puntos de balanceo los cuales son las observaciones 13, 18, 20, 27 como se muestra tabla.

¿Qué causan estos puntos?

4.2.3. Puntos influenciales

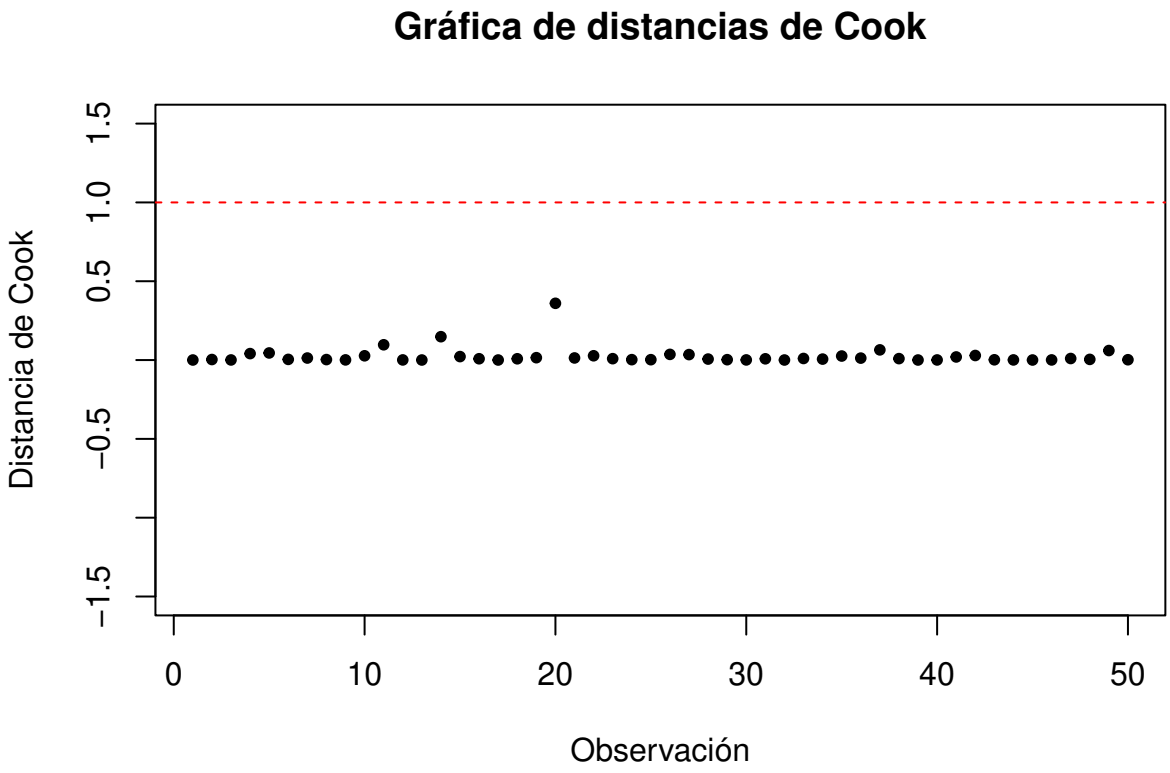

Figura 5: Criterio distancias de Cook para puntos influenciales

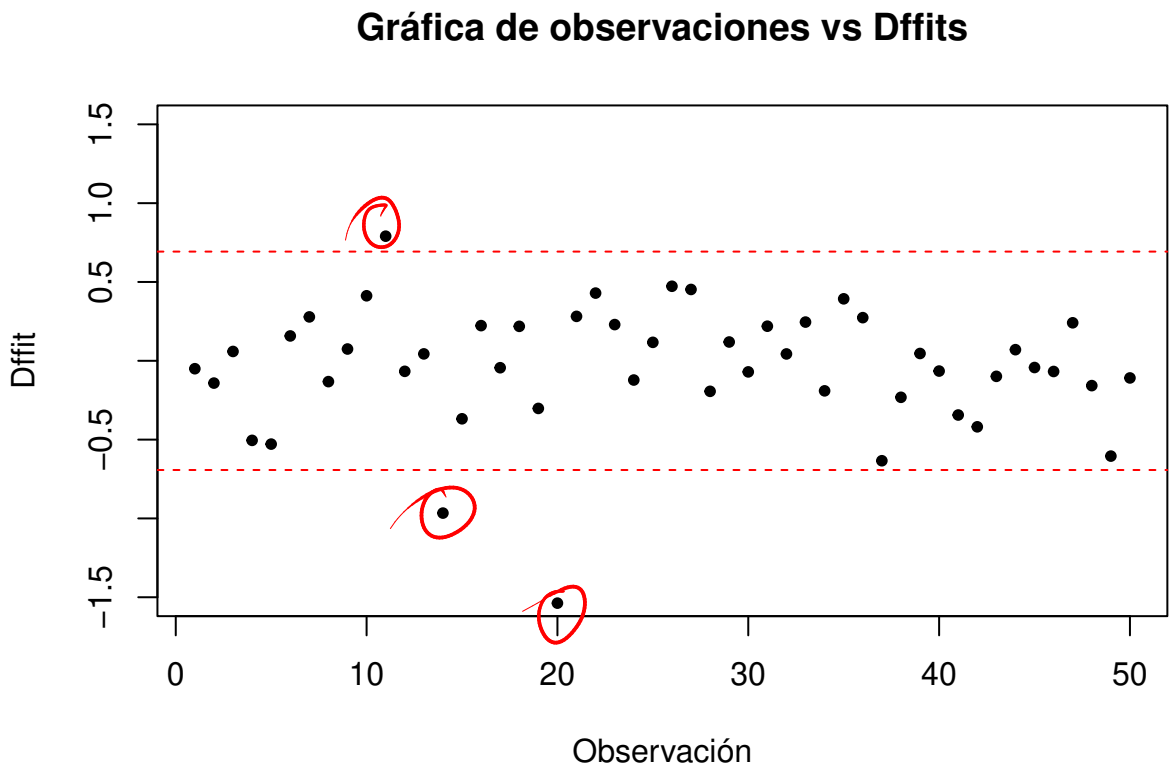

Figura 6: Criterio Dffits para puntos influenciales

Tabla 6: Resumen de diagnostico

	$Res.stud$	$Cook.D_i$	$h_{ii}value$	$Dffits$
11	2.0069	0.0969	0.1261	0.7906
14	-1.7421	0.1482	0.2266	-0.9661
20	-2.1709	0.3600	0.3143	-1.5375

4pt

Como se puede ver,las observaciones 11, 14, 20 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.6928203$, 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.

¿qué causan estos puntos dependiendo del criterio?

4.3. Conclusión

1pt

Para la validez del modelo se puede decir que se hallaron varios puntos de balanceo (4 en total) 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 me 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 riego de infección. El modelo podría generar inconsistencias por las razones mencionadas anteriormente.

X

La validez la da únicamente el cumplimiento de supuestos, los puntos extremos se analizan es porque pueden afectar en cierta medida a los mismos.