3,45

Trabajo 1

Estudiantes

Cristian Javier Rios Arrieta
Juan Camilo Martinez Oviedo
Jainy Meg Montes Ortíz
Valentina Valencia Quiceno

Docente

Francisco Javier Rodriguez Cortes

Asignatura

Estadística II

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

18,5 pt

Considerando la información contenida en el archivo "Equipo36.txt" que incluye 5 variables predictoras con nombres específicos, se puede afirmar que:

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

Entonces, se plantea el modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i;\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 55$$

1.1. Modelo de regresión

Cuando se realiza un ajuste al modelo, se generan los siguientes valores de coeficientes

Cuadro 1: Valores de coeficientes

	Valor del parámetro
β_0	1.8435
β_1	0.1573
β_2	-0.0215
β_3	0.0570
β_4	0.0093
β_5	0.0021

3pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 1.8435 + 0.1573X_{1i} - 0.0215X_{2i} + 0.057X_{3i} + 0.0093X_{4i} + 0.0021X_{5i}$$

1.2. Significancia de la regresión

4pt

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

$F_0 = \frac{MSR}{MSE} \neq \frac{MST}{MSE}$

$$F_0 = \frac{\cancel{MST}}{MSE} \overset{H_0}{\sim} f_{5,49} \tag{1}$$

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	Grados de libertad	Cuadrado medio	F_0	P-valor
Regresión	57.7523	5	11.550463	13.4269	2.96616e-08
Error	42.1520	49	0.860246		

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

6 pt

Este cuadro contiene información sobre los parámetros, que ayudará a identificar cuáles de ellos son sognificativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	1.8435	1.8840	0.9785	0.3326
β_1	0.1573	0.0766	2.0533	0.0454
β_2	-0.0215	0.0336	-0.6408	0.5246
β_3	0.0570	0.0143	3.9857	0.0002
β_4	0.0093	0.0087	1.0724	0.2888
β_5	0.0021	0.0008	2.5208	0.0150

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3 y β_5 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

2,5 pt

$\hat{\beta}_1$**:** Indica que cuando aumente la duración de la estadía de los pacientes en el hospital (en días), el promedio del riesgo de infección también aumentará sigificativamente en 0.1573 unidades, mientras las demás predictoras se mantienen fijas.

probabilidad promedio...

$\hat{\beta}_3$**:** Indica que por cada unidad que aumente el número promedio de camas en el hospital, el promedio del riesgo de infección también aumentará sigificativamente en 0.0570 unidades, mientras las demás predictoras se mantienen fijas.

probabilidad promedio

$\hat{\beta}_5$**:** Indica que por cada unidad que aumente el número promedio de enfermeras durante el estudio, el promedio del riesgo de infección también aumentará sigificativamente en 0.0021 unidades, mientras las demás predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

3pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5781$, lo que significa que aproximadamente el 57.81 % 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

3pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más alto en el modelo fueron X_1, 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:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	42.152	X1 X2 X3 X4 X5
Modelo reducido	43.624	X1 X3 X5

$\beta_1 = 0 \Rightarrow \beta_1 \cdot X_1 = 0$

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 55$$

β_1 vale 0 según Ho.

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_1, \beta_3, \beta_5) - SSE(\beta_0, \ \cdots\ , \ \beta_5))/3}{MSE(\beta_0, \ \cdots\ , \ \beta_5)} \overset{H_0}{\sim} f_{3,49} \\ &= \frac{(43.624 - 42.152)/3}{0.860244898} \\ &= 0.5703802113 \end{aligned} \tag{2}$$

1pt

Por lo menos son congruentes con el error...

Ahora, comparando a un nivel de significancia $\alpha = 0.05$, F_0 con $f_{0.95,3,49} = 2.7939$.

2pt

Entonces se concluye que las variables β_1, β_2, β_4 no son significativas al tener un valor p mayor al valor alpha asignado, por lo que se pueden retirar dichas variables del modelo.

3. Pregunta 3

4 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 2\beta_4;\ \ \beta_2 = \beta_3 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

3pt

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_2 X_{2i}^* + \beta_4 X_{4i}^* + \beta_5 X_{5i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ \ 1 \leqslant i \leqslant 55$$

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

1pt

$$F_0 = \frac{(SSE(MR) - 42.152/2}{0.860244898} \overset{H_0}{\sim} f_{2,49} \tag{3}$$

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF} \overset{H_0}{\sim} f_{2,49}$$

$$= \frac{(SSE(MR) - 42{,}152)/2}{0{,}860244898}$$

4. Pregunta 4

19 pt

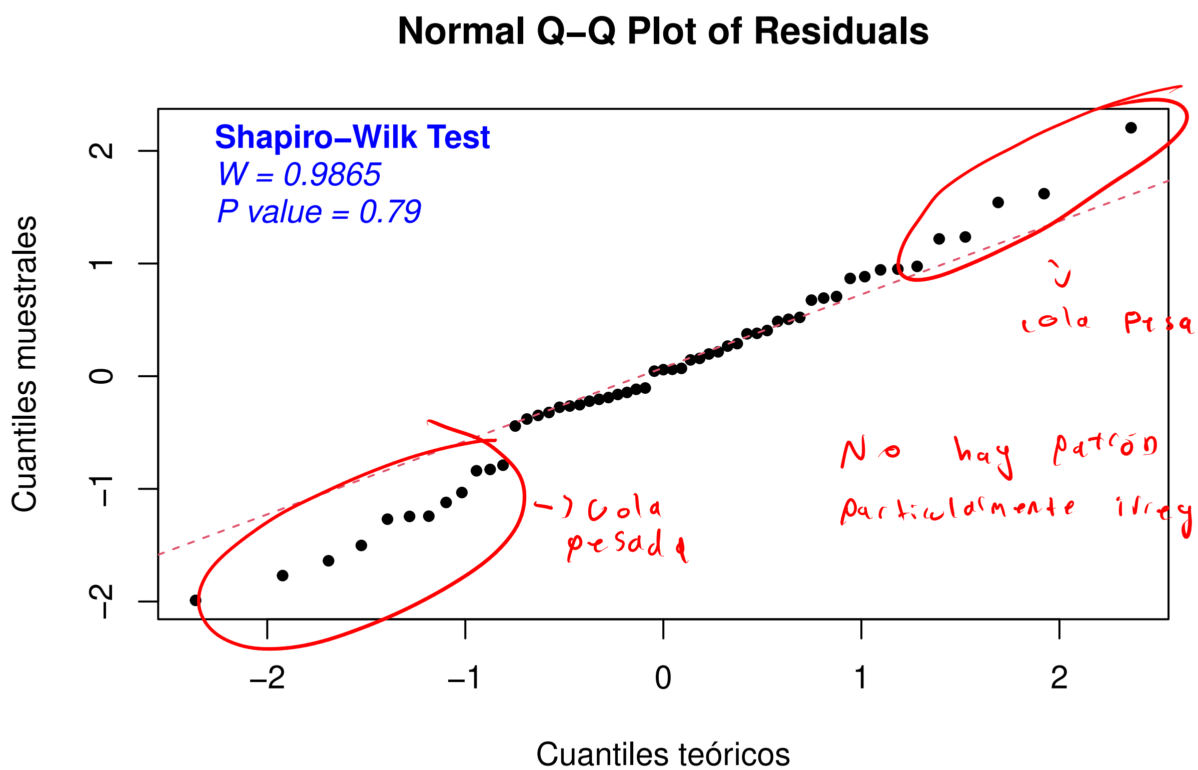
4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

3,5 pt

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis sh~~apiro-wi~~lk, acompañada de un gráfico cuantil-cuantil:

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



cola pesada

cola pesada

No hay patrón particularmente irregular

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Dado que el valor del P-valor es cercano a 0.79 y considerando que el nivel de significancia es de $\alpha = 0.05$, se concluye que el P-valor es significativamente mayor, lo que indica que no se puede rechazar la hipótesis nula. Esto significa que los datos se distribuyen normalmente con una media de 0 y una varianza de σ^2, sin embargo la gráfica de comparación de cuantiles permite ver colas más pesadas y patrones irregulares, al tener más poder el análisis gráfico, se termina por rechazar el cumplimiento de este supuesto. Ahora se procederá a validar si la varianza se mantiene constante, como se supone.

No están probando eso. NO los hay

4.1.2. Varianza constante

2,5 pt

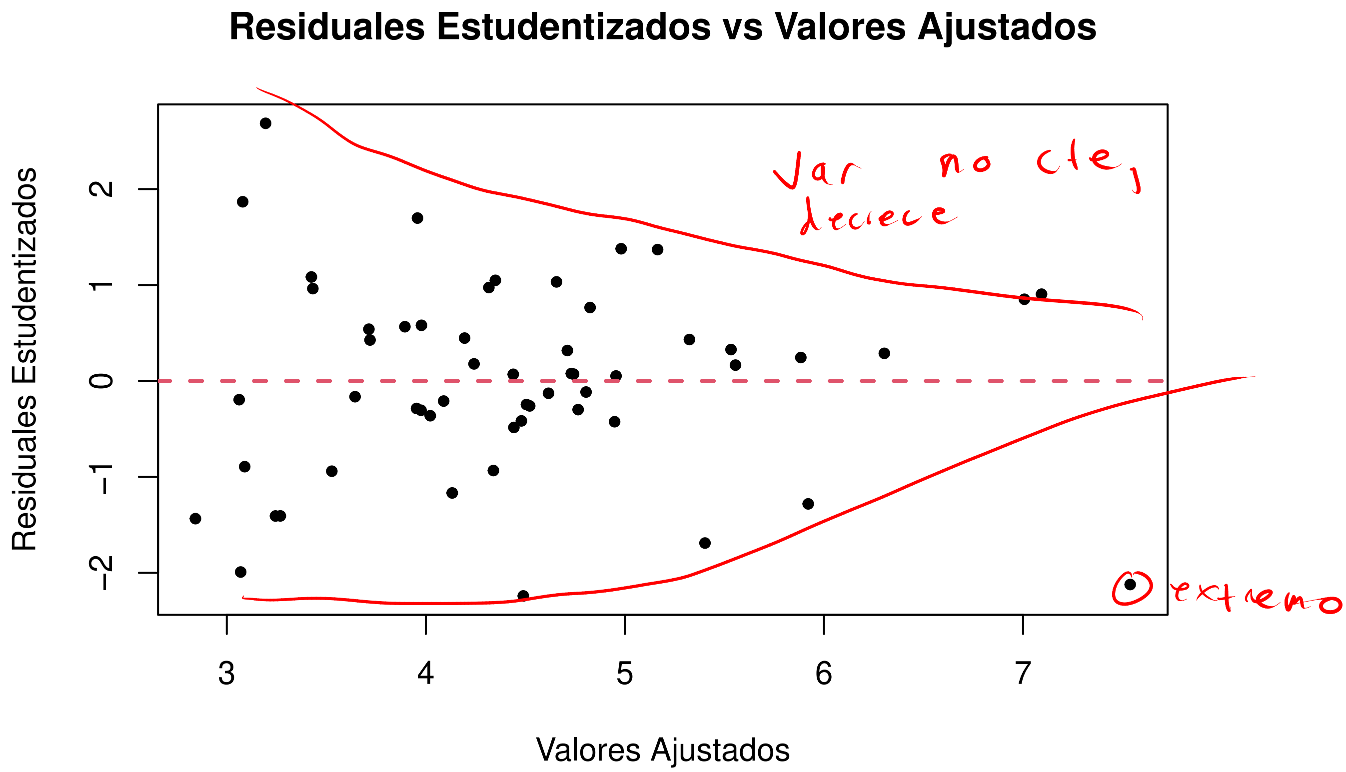


Figura 2: Gráfico residuales estudentizados vs valores ajustados

Se puede afirmar que no tiene varianza constante, ya que ~~se puede apreciar que la distancia~~ de los puntos alrededor de su tendencia no es igual. Hay variabilidad en vez de dispersión.

Hasta acá iban bien.

huh?

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3pt

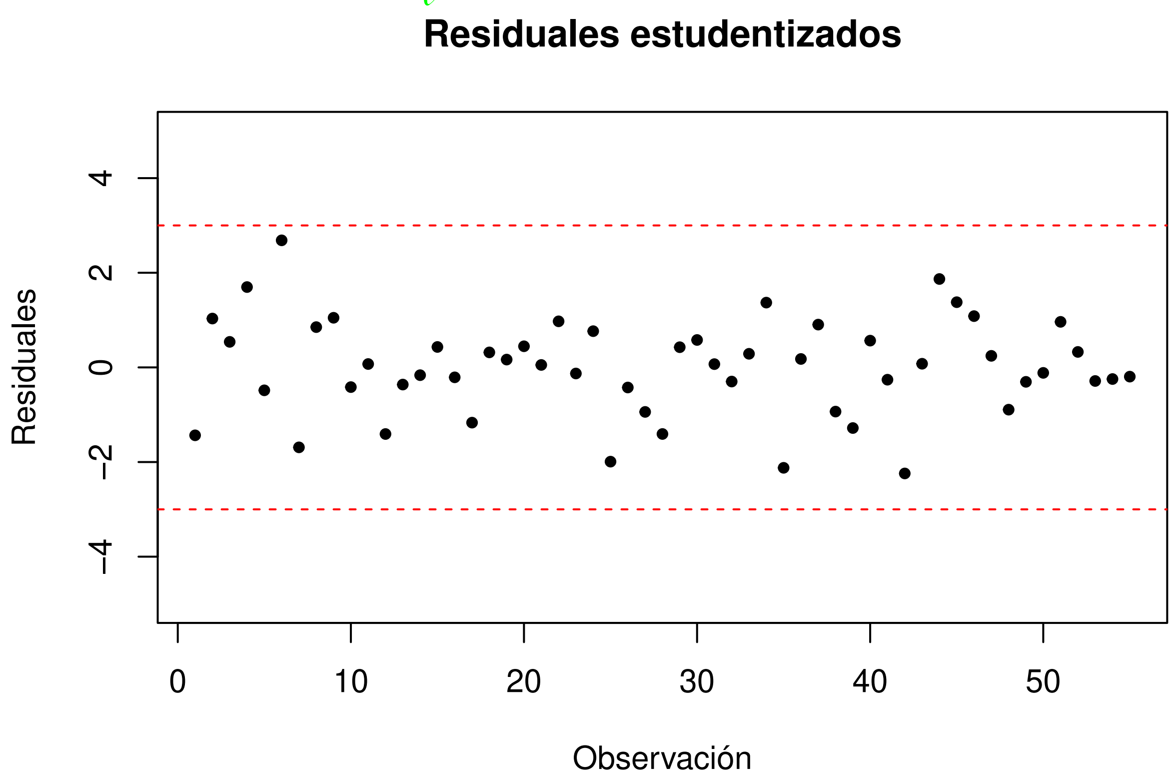


Figura 3: Identificación de datos atípicos

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

lpt

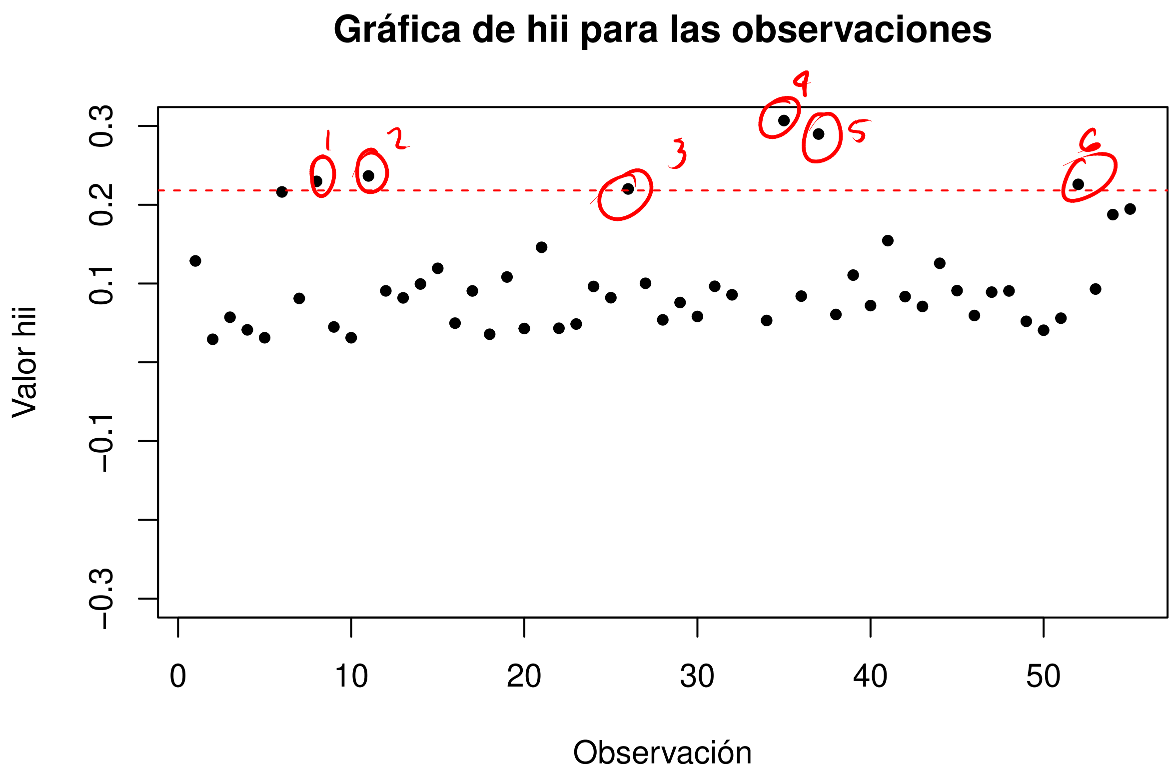


Figura 4: Identificación de puntos de balanceo

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.2181$, 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.

¿De dónde sacan eso?
En la gráfica veo 6
¿Cuáles son? ¿Qué causan?

tabla? Dónde?

copy + paste de la plantilla.

4.2.3. Puntos influenciales

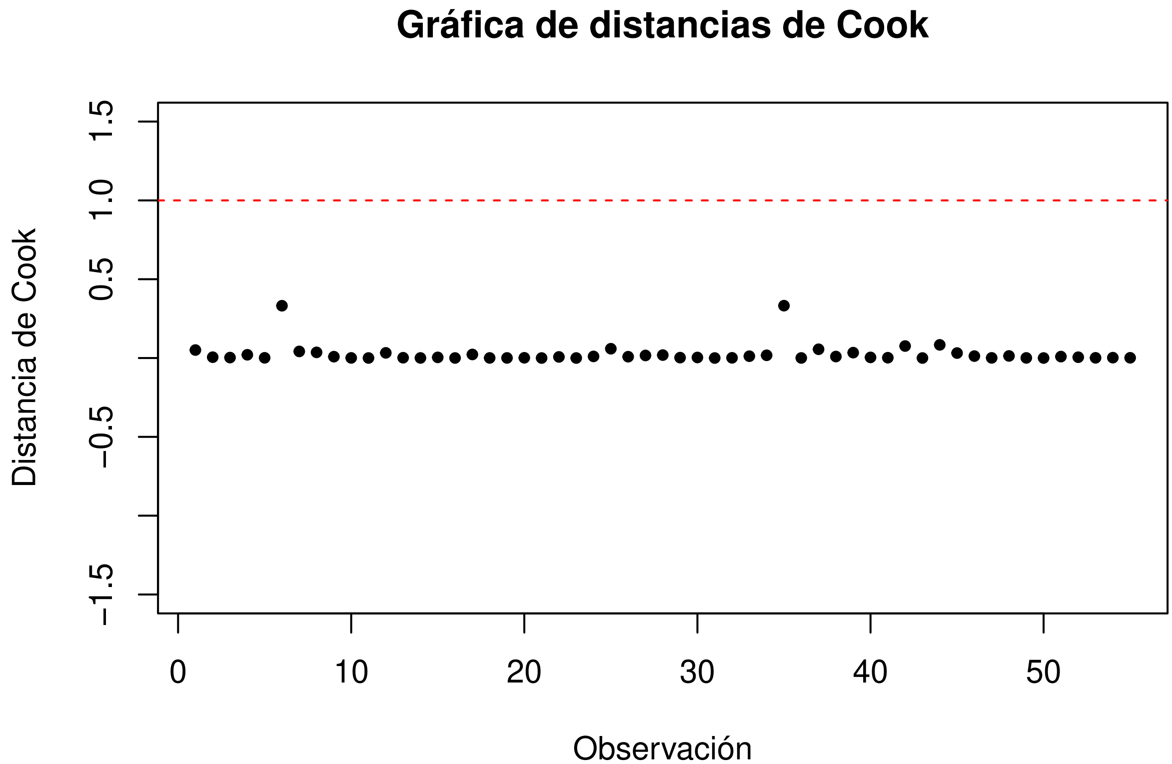


Figura 5: Criterio distancias de Cook para puntos influenciales

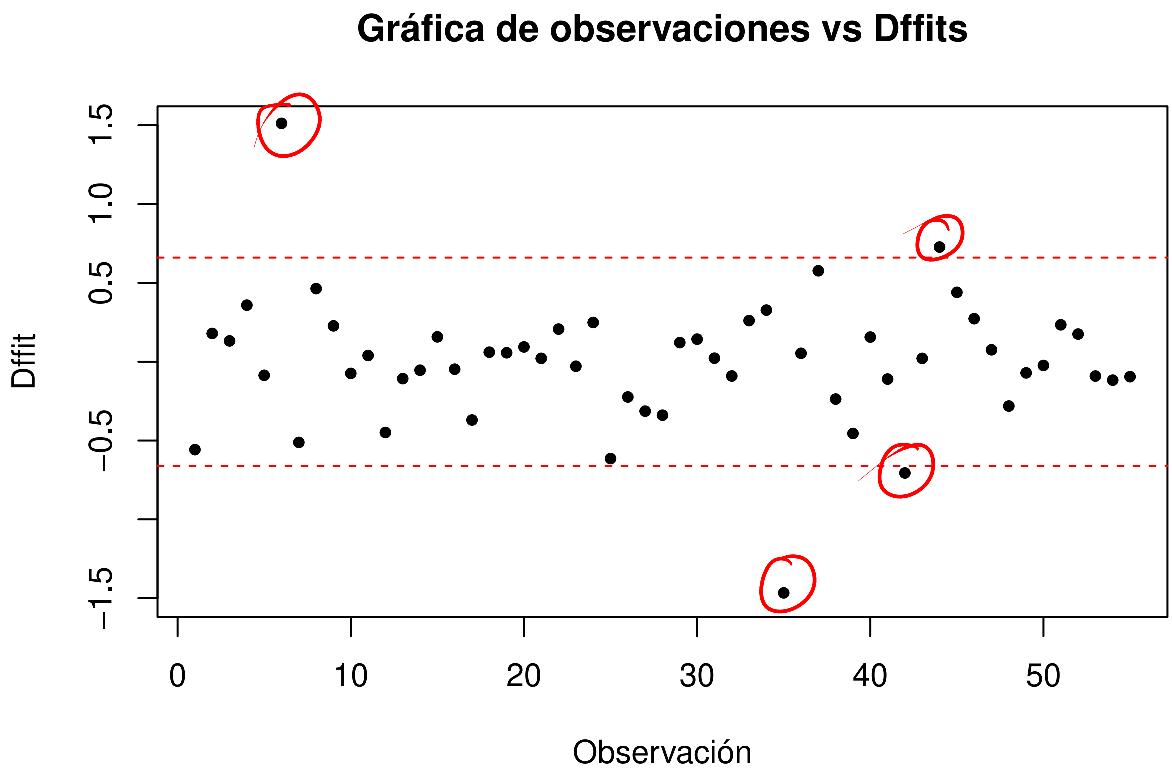


Figura 6: Criterio Dffits para puntos influenciales

3pt

```
##     res.stud Cooks.D hii.value  Dffits
## 6     2.6855  0.3317    0.2163  1.5120
## 35   -2.1218  0.3323    0.3069 -1.4665
## 42   -2.2407  0.0761    0.0834 -0.7060
## 44    1.8680  0.0836    0.1257  0.7275
```

→ Volver tabla, aunque no los baje es mal presentado.

Como se puede ver, 4 observaciones: 6, 35, 42 y 44 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.6606$, 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?

seguros?

4.3. Conclusión

1pt

Al hacer un análisis del modelo y de la significancia de la regresión se comprobó que no es el mejor modelo que se pudo haber planteado, puesto que no todos los parámetros que acompañan a las variables eran significativos, pues β_2, β_4 aparte del intercepto β_0, no resultan útiles en el intento por explicar la variable respuesta (Riesgo de infección), por lo que se puede prescindir de ellos y contar aún con un buen modelo. Por otro lado, en cuanto a los supuestos distribucionales de los residuales se determinó normalidad en los errores y varianza no constante. También, en cuanto a las observaciones no se encontró datos atípicos en la gráfica de residuales estudentizados, sin embargo en la gráfica de valores hii se evidencian alrededor de 7 puntos de balanceos que cumplen $h_{ii} > 2\frac{p}{n}$. Por último, existen puntos influenciales

Se contradicen, habían rechazado normalidad.

según el criterio de Dffits, pero para el criterio de distancias de cook, según la gráfica, no hay datos que cumplan $D_i > 1$.

No dicen si el modelo es válido.