2,15

Trabajo 1

Estudiantes

David Leon Ruiz Herrera
Carlos Mario Chavez Aguilera
Sebastian Jaramillo Correa

Equipo 29#

Docente

Francisco Javier Rodriguez

Asignatura

Estadística II

Sede Medellín

30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1 12 pt

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

¿Quiénes son esas var's? ¿Qué representan?

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2);\ 1 \leqslant i \leqslant 45$$

1.1. Modelo de regresión 1,5 pt

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.4268
β_1	0.1468
β_2	0.0217
β_3	0.0665
β_4	0.0088
β_5	0.0019

No va en ec. ajustada

su n = 45

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.4268 + 0.1468 X_{1i} 0.0217 X_{2i} + 0.0665 X_{3i} + 0.0088 X_{4i} + 0.0019 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2);\ 1 \leqslant i \leqslant 50$$

1.2. Significancia de la regresión 4 pt

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:

$$F_0 = \frac{MSR}{MSE} \sim F_{5,39}$$

$$F_0 = \frac{\cancel{MST}}{MSE} \overset{H_0}{\sim} f_{5,39} \tag{1}$$

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	54.4339	5	10.886776	12.8548	2.06507e-07
Error	33.0292	39	0.846903		

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, Note que $f_0 >$ f0,05,5,5 por lo tanto con un 95 % de significancia al menos un β_j es diferente de 0, Lo que significa que por lo menos uno de los parámetros si es significativo en presencia de los otros a la hora de explicar el riesgo de infección.

¿y dónde veo ese valor?

1.3. Significancia de los parámetros

4 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	-0.4268	1.7623	-0.2422	0.8099
β_1	0.1468	0.0760	1.9311	0.0608
β_2	0.0217	0.0318	0.6831	0.4986
β_3	0.0665	0.0147	4.5218	0.0001
β_4	0.0088	0.0086	1.0294	0.3096
β_5	0.0019	0.0008	2.4198	0.0203

Los P-valores presentes en la tabla permiten concluir con un $\alpha = 0.05$, que Si P-value < nivel confianza 5 % es significativo el ~~beta~~, Si P-value > nivel ~~confianza~~ 5 % No es significativo, por lo tanto los parámetros β_3 y son significativos, pues sus P-valores son menores a α.

s β_j

¿β_3 y ...?

significancia

1.4. Interpretación de los parámetros

1 pt

$\hat{\beta}_0$**:** t value = -0.2422038 < t(0.025,39) = 2,023. No se puede determinar significancia de X0. No es interpretable.

$\hat{\beta}_1$**:** t value = 1.9310960 < t(0.025,39) = 2,023. No se puede determinar significancia de X1. No es interpretable.

¿cómo que no? sólo no lo son

$\hat{\beta}_2$: t value = 0.6831252 < t(0.025,39) = 2,023. No se puede determinar significancia de X2. No es interpretable.

probabilidad promedio de considerar la infección ¡lean bien de qué tratan sus datos

$\hat{\beta}_3$: t value = 4.5218340 > t(0.025,39) = 2,023. Se puede determinar que el parámetro es significativo, con el aumento de 1 cama también aumentaría la cantidad de infecciones explicado a una tasa de ~~0.0665003~~ %. Considerando que los demás parámetros estén constantes.

6,65003%

¿? $\hat{\beta}_4$: t value = 1.0294096 < t(0.025,39) = 2,023. ~~No se puede determinar significancia~~ de $\boxed{\text{X1}}$. No es interpretable.

$\hat{\beta}_5$: t value = 2.4198102 > t(0.025,39) = 2,023. Se puede determinar que el parámetro es significativo, con el aumento de 1 enfermera también aumentaría la cantidad de infecciones explicado a una tasa de ~~0.0018606~~ %. Considerando que los demás parámetros estén constantes.

0,18606%

1.5. Coeficiente de determinación múltiple R^2

1,5 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6224$ Esto nos deja saber que con los datos que se tienen, el \$ 62.24 % \$ del modelo es explicado por las variables predictoras. Este ~~parámetro~~ podría ser mayor lo que daría a entender que no necesariamente están bien explicados los valores de la regresión por las variables predictoras dadas. Son poco significativas.

→ Esto no habla de significancia

→ No! La respuesta es explicada por el modelo. ¿cómo se calcula?

0,5 pt

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable 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:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_4 = 0 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 4 \end{cases}$$

Si $\beta_1 = \beta_2 = \beta_4 = 0$

MR: $Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon_i$; $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

var's en MR: X_3, X_5

Nada que ver

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	33.029	X1 X2 X3 X4 X5
Modelo reducido	38.839	X1 X3

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

0,5 pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,39} \\ &= \frac{33.029 - 45.750}{38.839 - 39} \\ &= \frac{-12.721}{-0.161} \\ &= -79.01242 \end{aligned} \tag{2}$$

esto es MF

esto es MR y ni siquiera es el valor que corresponde a su ejercicio

¿?

menos * menos = +

Es posible o no descartar las variables del subconjunto

¿? 0pt

3. Pregunta 3

0,5 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si … por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \text{H}_0 : \beta_1 = \beta_4;\ \beta_5 = \beta_2;\ \beta_3 = 0 \\ \text{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

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

Con $\mathbf{L}$ dada por

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

0pt

El modelo reducido está dado por:

esto debería ser 0

$$Y_i = \beta_o + \beta_1 X_{1i}^* + \beta_3 X_{2i}^* + \boxed{\beta_3 X_{3i}} + \beta_1 X_{4i}^* + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

0pt

$$Y_i = \beta_o + \beta_1 (X_{1i}^* + X_{4i}^*) + \beta_3 (X_{2i}^* + X_{3i}^*) + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

$Y_i = \beta_0 + \beta_1 (x_{1i} + x_{4i}) + \beta_2 (x_{2i} + x_{5i}) + \varepsilon_i\ ;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)\ ;\ 1 \leq i \leq 45$

3.2. Estadístico de prueba

Reemplazar lo que conocen

0,5 pt

El estadístico de prueba F_0 está dado por:

$\to 3$

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,39'} \tag{3}$$

$\to 3$

4. Pregunta 4 8,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales 2 pt

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

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

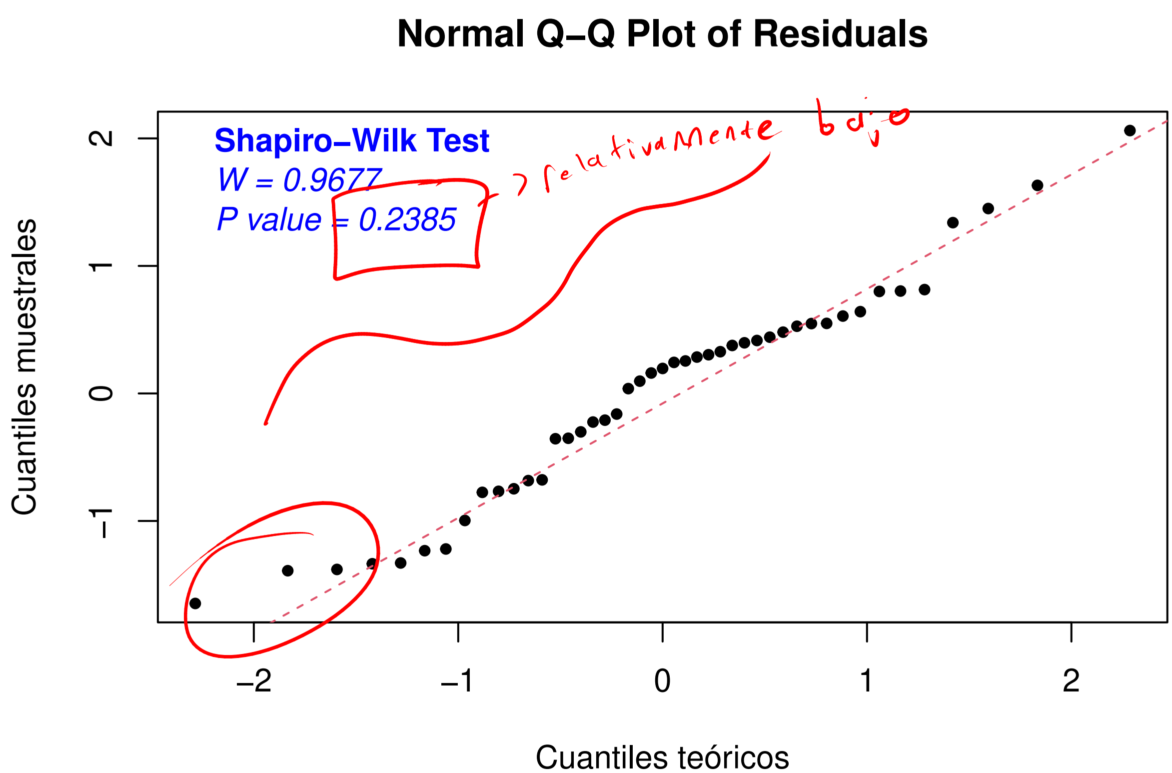


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Hasta acá tiene sentido lo que escriben

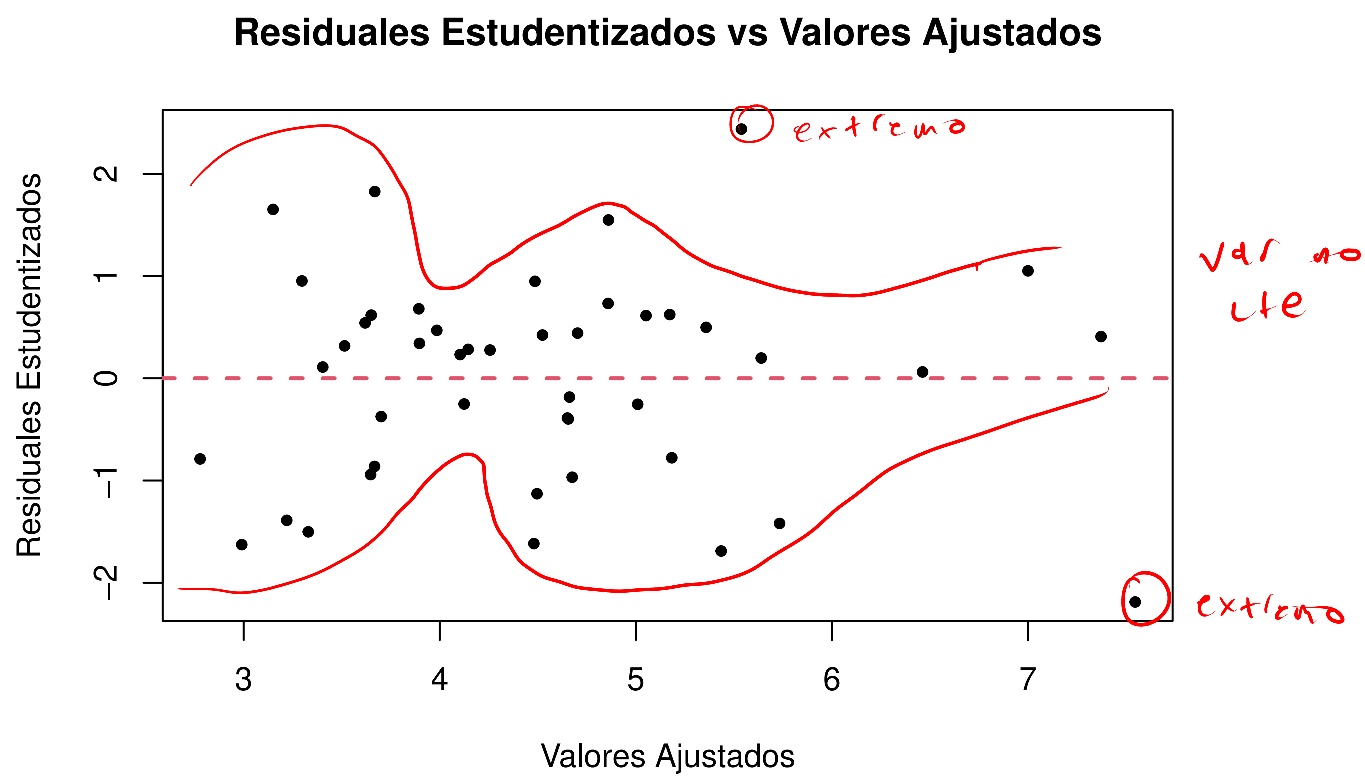
Al ser el P-valor aproximadamente igual a 0.2385 Con una confianza del 0.95 se puede decir que los residuales del modelo se distribuyen de forma normal ya que p-value es mayor a 0.05, es decir que los datos distribuyen normal con media μ y varianza σ^2, y patrones irregulares, al tener más poder el análisis gráfico, se termina por rechazar el cumplimiento de este supuesto. Ahora se validará si la varianza cumple con el supuesto de ser constante. ¿?

Ni siquiera lo han hecho

huh? ¿por qué?

lpt

4.1.2. Varianza constante



var no cte

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados, se puede concluir que estos no son constantes, ya que en la grafica se ve que estos no tienen un comportamiento ordenado.

No se trata de que sea ordenado, análisis deficiente

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3p +

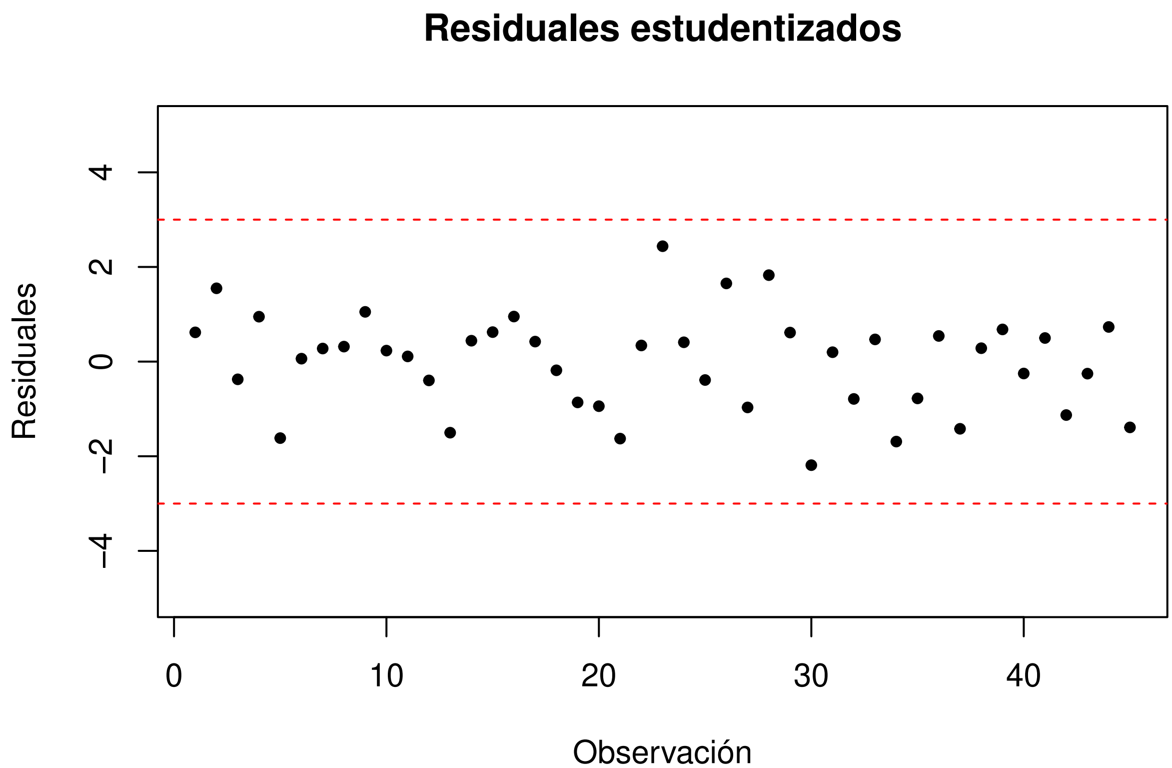


Figura 3: Identificación de datos atípicos

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

1 pt

Ampliar límites

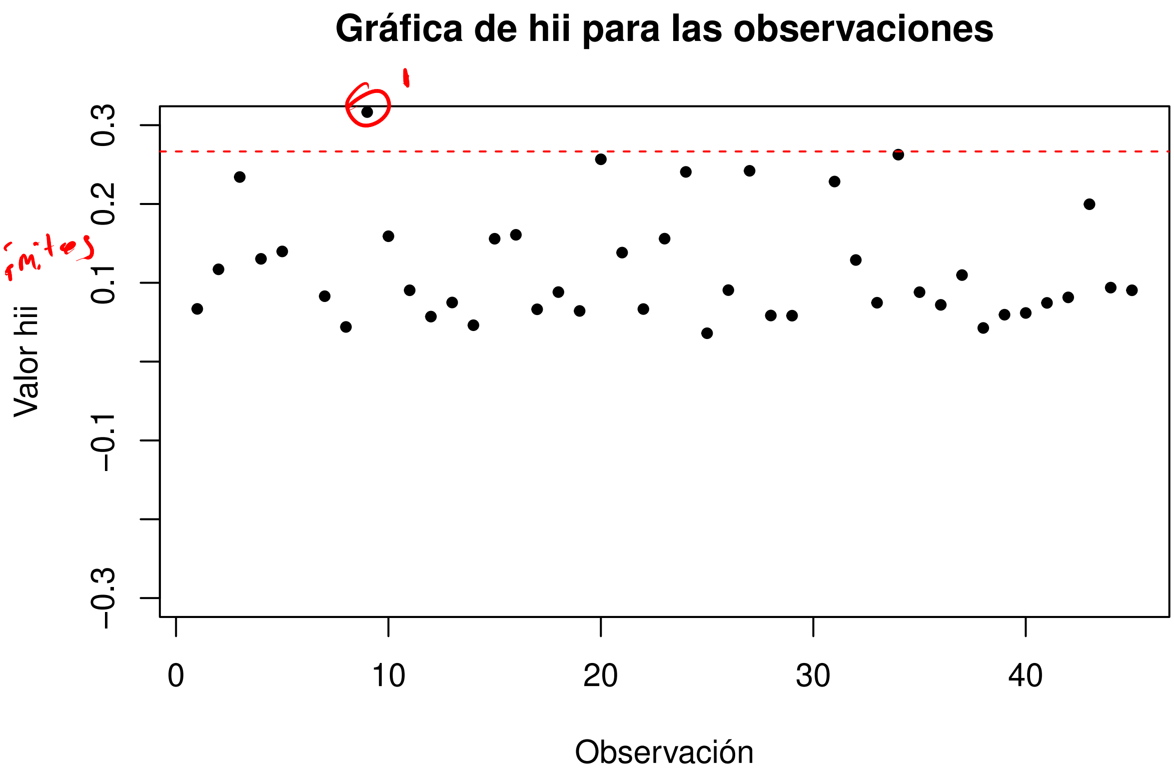


Figura 4: Identificación de puntos de balanceo

Sólo se ve 1.

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 3 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 6, 9 y 30.

→ 3 tablas? o esas son los datos? cuál es su hii? hay una tabla?

Qué causan en el modelo?

4.2.3. Puntos influenciales

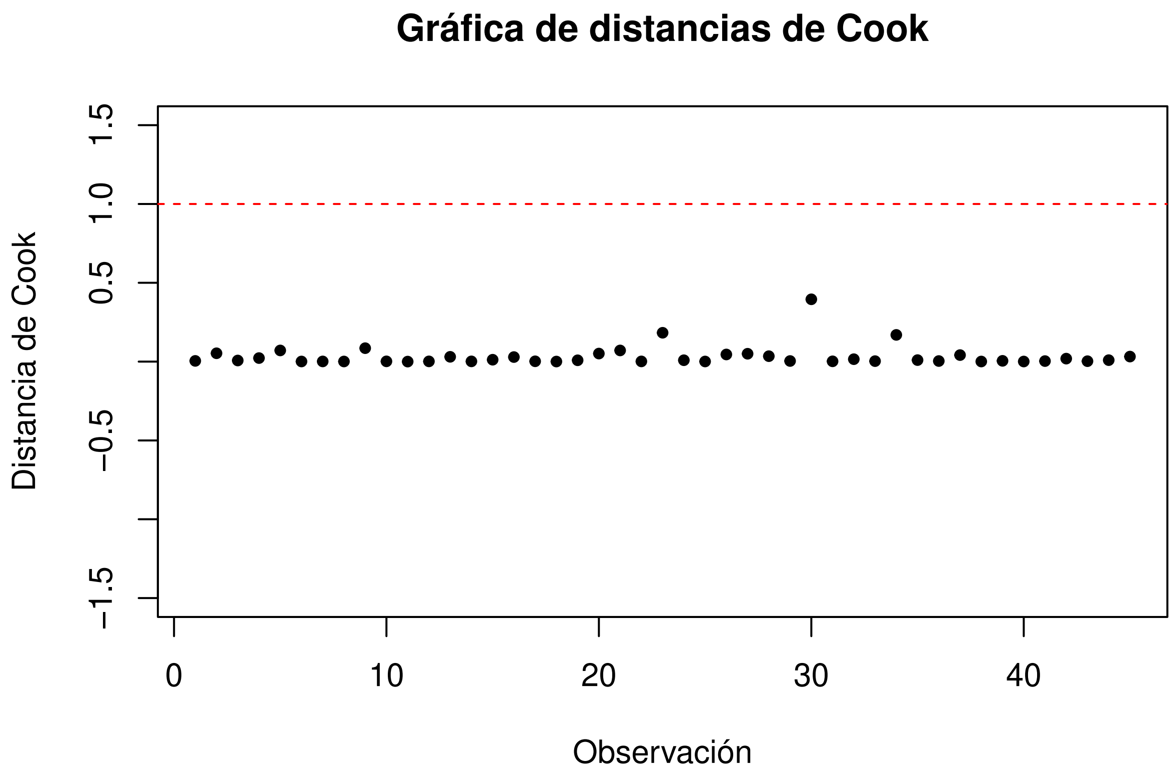
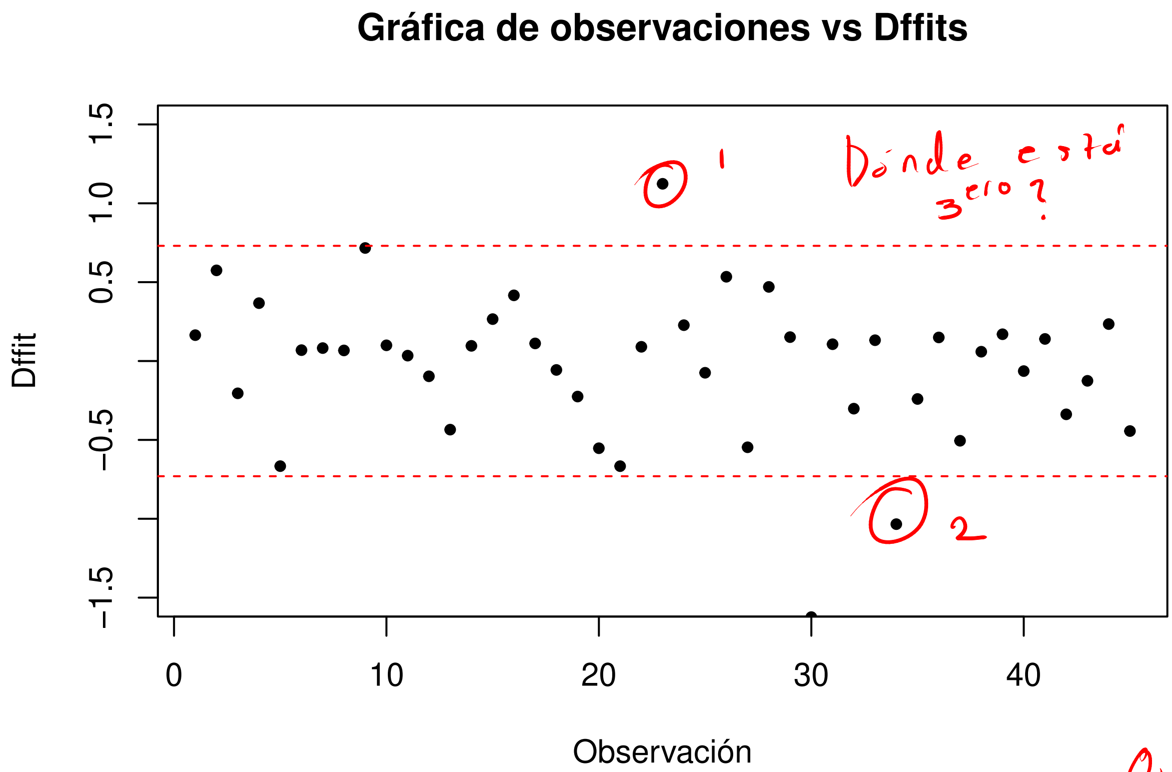


Figura 5: Criterio distancias de Cook para puntos influenciales



Dónde está el 3ero?

Ampliar límites

0,5 pt

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 23   2.4382  0.1831    0.1560  1.1237
## 30  -2.1884  0.3954    0.3313 -1.6233
## 34  -1.6900  0.1695    0.2626 -1.0341
```

→ tabla, no salida

¿cuáles? 23, 30, 34 ¿cuánto da?

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, 23 datos cumplen. → Ni uno cumple!!

¿Qué causan según cada criterio?

1pt

4.3. Conclusión

Desde los análisis anteriores podemos encontrar varias cosas, entre ellas que aunque el modelo es significativo en su generalidad no es un buen modelo en cuanto a predicción porque incluye 3 variables no significativas dentro del mismo además que el coeficiente múltiple R2 el cual no penaliza por adherir tales variables es aún muy bajo, por lo tanto esto puede darse debido a que las variables significativas del modelo no tienen una relación lineal muy adecuada con la variable de respuesta "Riesgo de infección" lo cual se puede evidenciar en

→ ¿qué es adecuado? ¿Encontraron no linealidad?

el poco aumento marginal que aportan estas variables cuando el resto está constante, esto debido posiblemente a que "Riesgo de infección" depende de otros factores.

No responden si el modelo es válido o no.