3,7

Trabajo 1

Estudiantes

Arturo De Jesus Rangel Julio

Equipo #48

Docente

Francisco Javier Rodriguez Cortes

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

14,5 pt

Teniendo en cuenta la base de datos 48, encontamos que tiene 5 variables regresoras las cualoes son:

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

Así, al palntear el modelo de regresión lineal múltiple tenemos:

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

1.1. Modelo de regresión

1,5 pt

De esta manera al traer el modelo ajusuastado, obtenemos los siguientes coeficientes para los valores de beta:

Cuadro 1: Tabla de valores para los beta

	V. parámetros
β_0	-1.8188
β_1	0.1857
β_2	0.0357
β_3	0.0511
β_4	0.0165
β_5	0.0012

No va en ec. ajustada

entonces, el resultado del modelo de regresión ajustado es:

$$\hat{Y}_i = -1.8188 + 0.1857X_{1i} + 0.0357X_{2i} + 0.0511X_{3i} + 0.0165X_{4i} + 0.0012X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

50

1.2. Significancia de la regresión

4 pt

Así si queremos analizar la significancia de la regresión, se debe plantear un juego de hipótesis que ayuden a interpretar su validez:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

$\frac{MST}{MSE}$

SE utiliza un estadístico de prueba que es:

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

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

Presentando la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Suma Cuadratica	G.libertad.	media cuadratica	F_0	P-valor
Regresión	47.573	5	9.51460	10.6996	9.03818e-07
Error	39.127	44	0.88925		

Viendo los resultadod obtenido en la tabla Anova, se observa un valor P es un valor muy pequeño que lo hace muy cercano a 0, por lo que para cualquier nivel de significancia (α) se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, por tanto se aceptan la hipótesis alternativa que nos dice que algún $\beta_j \neq 0$, por esta razon la regresión general es significativa.

1.3. Significancia de los parámetros

6 pt

En el siguiente cuadro podemos encontrar detalladamente la información de los parámetros estudiados, de esta manera podrmos determina con mayor precision cuáles de ellos son significativos y cuales no.

Pureba de hipotesis para los parametros:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j ! = 0 \end{cases}$$

$j = 0, 1, \ldots, 5$

Cuadro 3: Resumen de los coeficientes beta

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-1.8188	1.8419	-0.9875	0.3288
β_1	0.1857	0.0812	2.2857	0.0271
β_2	0.0357	0.0335	1.0665	0.2920
β_3	0.0511	0.0150	3.4040	0.0014
β_4	0.0165	0.0074	2.2220	0.0315
β_5	0.0012	0.0007	1.6785	0.1003

De la tabla anterior al nalizar los P-valores resultantes, permiten llegar a la concluición de que con un nivel de significancia $\alpha = 0.05$, para los parámetros β_2 y β_5 se acepta la hipotesis nula(H_0) y para los Parametros β_1, β_3 y β_4 se acepta la hipotesis alternativa(H_1), por lo que son significativos. Ya que sabemos que para que los P-valores cumplan la prueba de significancia deben ser menores a α.

1.4. Interpretación de los parámetros

0 pt

$\hat{\beta}_1$**:** Por cada unidad de aumento en la Probabilidad promedio estimada de adquirir infección en el hospital (Y), la Duración promedio de la estadía de todos los pacientes en el hospital(X_1) aunmenta significativamente en 0.1857 unidades, cuando los valores en las demas predictoras se mantiene fijo.

$\hat{\beta}_3$**:** Por cada unidad de aumento en la Probabilidad promedio estimada de adquirir infección en el hospital (Y), el Número promedio de camas en el hospital durante el periodo del estudio(X_3) aunmenta significativamente en 0.0511 unidades, cuando los valores en las demas predictoras se mantiene fijo.

$\hat{\beta}_4$**:** Por cada unidad de aumento en la Probabilidad promedio estimada de adquirir infección en el hospital (Y), el Número promedio de pacientes en el hospital por día durante el periodo del estudio(X_4) aunmenta significativamente en 0.0165 unidades, cuando los valores en las demas predictoras se mantiene fijo.

Unidad de aumento en X_j, Y aumenta β_j unidades

3 pt

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5487$, de lo cual identificamos que el porcentage de la variabilidad de el Riesgo de infección explicado por el modelo propuesto en el presente es 54.77 %.

¿cómo se calcula?

2. Pregunta 2

3 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Encontramos que las 3 covariable con el P-valor más alto en el modelo fueron X_2, X_4, X_5, así vemos que a partir 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	39.127	X1 X2 X3 X4 X5
Modelo reducido	45.921	X1 X3

asi encontramos que un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba de la siguiente forma:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} F_{3,44} \\ &= \frac{45.921 - 39.127}{0.88925} \\ &= 1.921 \end{aligned} \tag{2}$$

$F_0 < F_{0{,}95,3,44}$

Ahora, al realizar una comparacion entre F_0 con $f_{0.95,3,44} = 2.8165$, se puede encontrar que $\cancel{F_0 > f_{0.95,1,45}}$. Debido a esto, el subconjunto es significactivo. por tal razon no es posible descartar las variables del modelo ya que se rechasa la hipotesis nula (H_0) y almenos unas de sus variables es distinto de 0, por lo que se acepta la hipotesis alternativa (H_1). 0pt

3. Pregunta 3

5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

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

→ ojo con eso.

$$\begin{cases} \text{H}_0 : \beta_3 = 5\beta_4;\ \beta_2 = 2\beta_5 \\ \text{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

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

donde la **L** dada por

$$L = \begin{bmatrix} 0 & 0 & 0 & 1 & -5 & 0 \\ 0 & 0 & 1 & 0 & 0 & -2 \end{bmatrix}$$

3pt

El modelo reducido está dado por:

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

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

2pt

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

&=

$$F_0 = \frac{(SSE(MR) - 39.127)/2}{0.88925} \overset{H_0}{\sim} f_{2,44} \tag{4}$$

4. Pregunta 4

14,5 pt

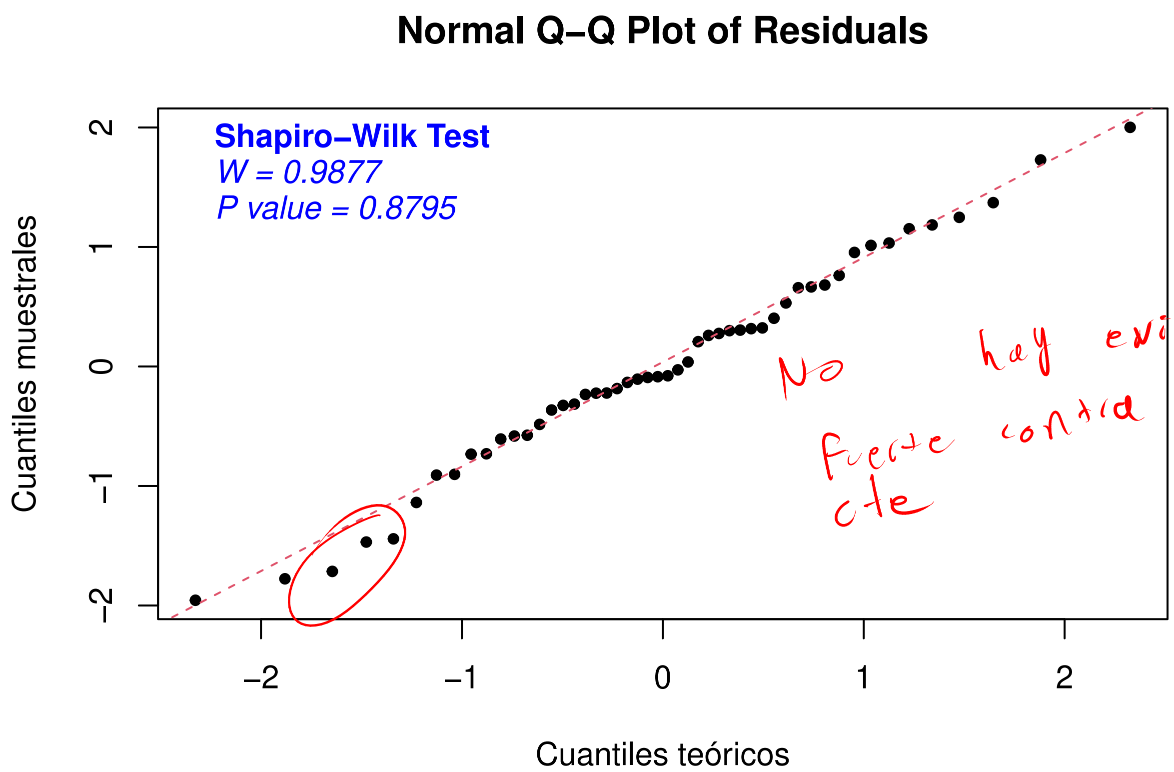
4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

3,5 pt

Para la verificación de este supuesto, se ara el planteamiento de la siguiente prueba de hipótesis ~~shapiro~~ wilk, de igual manera se agregara un gráfico cuantil-cuantil:

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



No hay evidencia fuerte contra ver cte

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.8795 y al ser $\alpha = 0.05$, el p-valor es mucha mas alto que el α por lo que no se rechasa la hipotesis nula (h_0)y al ser tan cercano a 1 esto muestra que cumple con la prueva de normalida, por lo que los dato se distruibuyen normal con ~~media~~ μ y ~~varianza~~ σ^2. Pero en este caso hay que tener el cuenta que la grafica aunque no se logran identificar colas muy pesadas, claramente se pueden ver patrones marcados e irregulares, por lo que se toma la decision de rechasar el cumplimiento del supuesto ya que la grafica tiene mas importancia a la hora de decidir. ahora vamos a proceder a ver si la varianza es constante. X

Te valgo el análisis pero en esta base no se rechaza.

No se está probando media y var ctes.

4.1.2. Varianza constante

1 pt

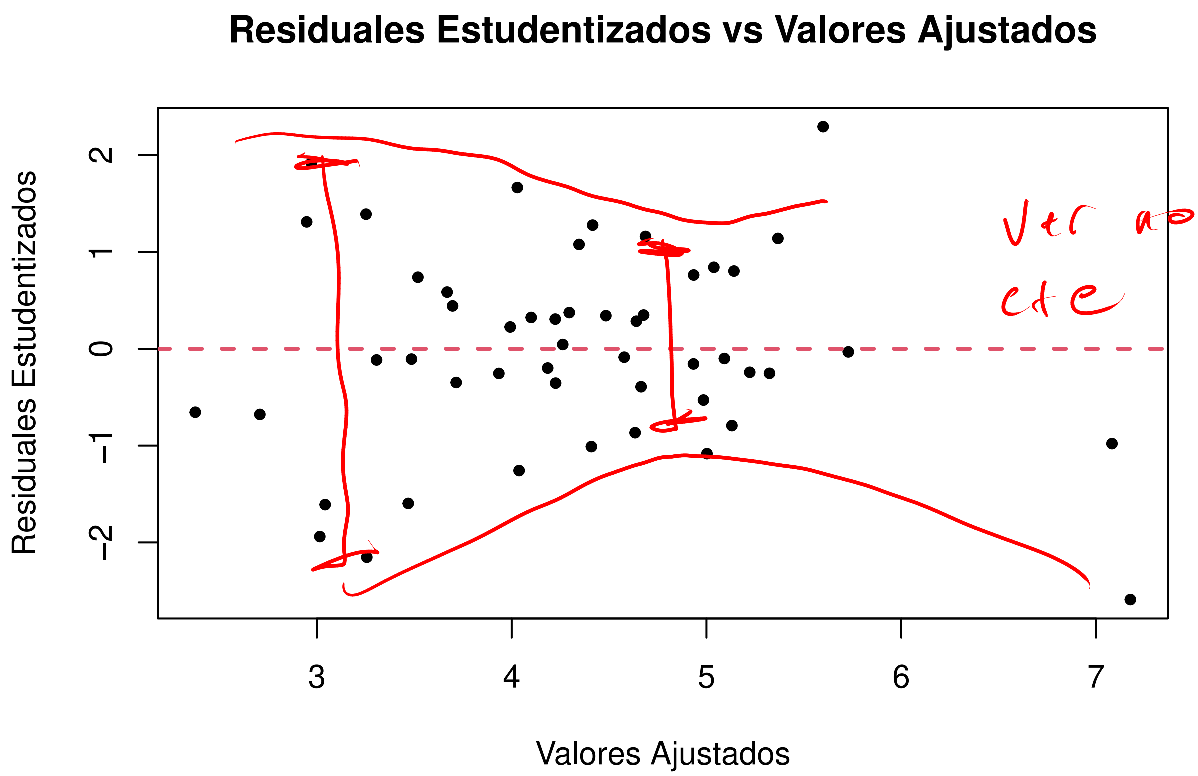


Figura 2: Gráfico residuales estudentizados vs valores ajustados

Al trazar las dos lineas una por debajo y una arriba de los dados, se puede identificar que no existen comportamientos que me hagan rechazar una varianza constante.De igual forma no vemos patrones de aumento o disminucion de la varianza. tambien podemos observa que se tiene media 0.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3 e t

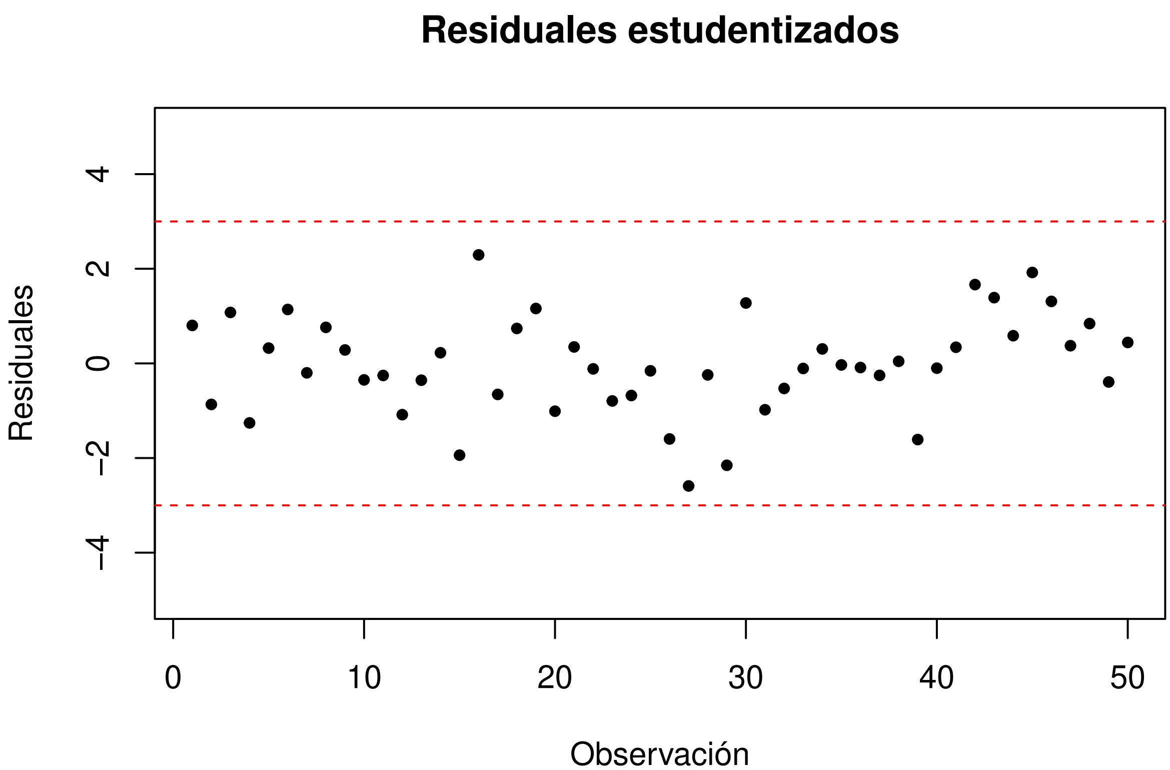


Figura 3: Identificación de datos atípicos

podemos ver en la grafica anterio que no se observan datos atipicos en le conjunto de datos pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

1pt

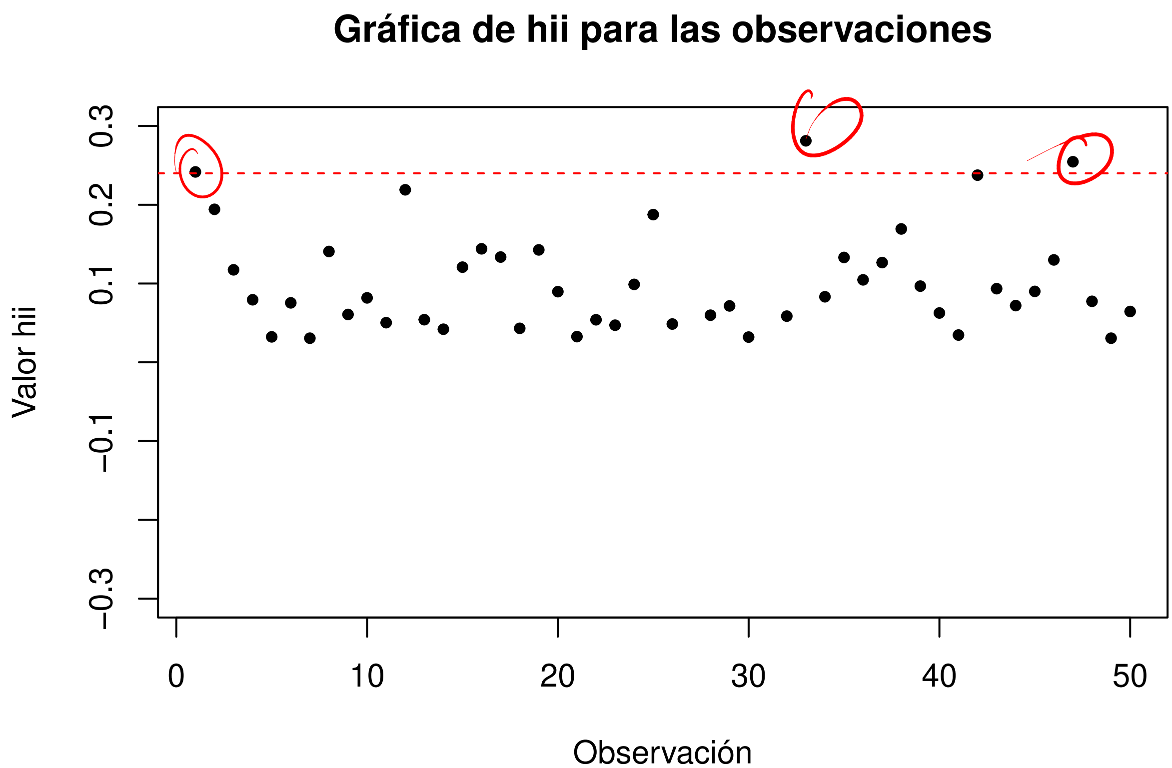


Figura 4: Identificación de puntos de balanceo

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

Tienes de hecho 5 puntos de balanceo.

¿cuál? No está.

	hii.value	Dffits
1	0.2417	0.4512
27	0.4710	-2.6243
31	0.6027	-1.2056
33	0.2812	-0.0662
47	0.2546	0.2162

4.2.3. Puntos influenciales

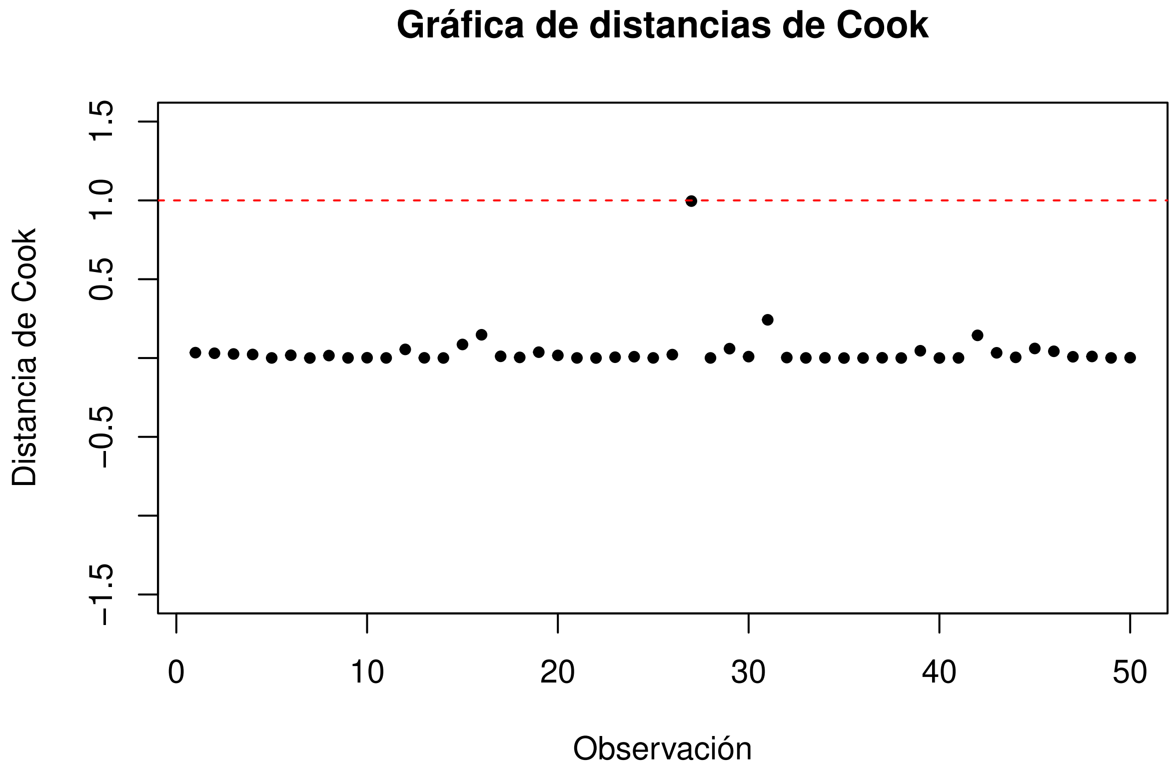
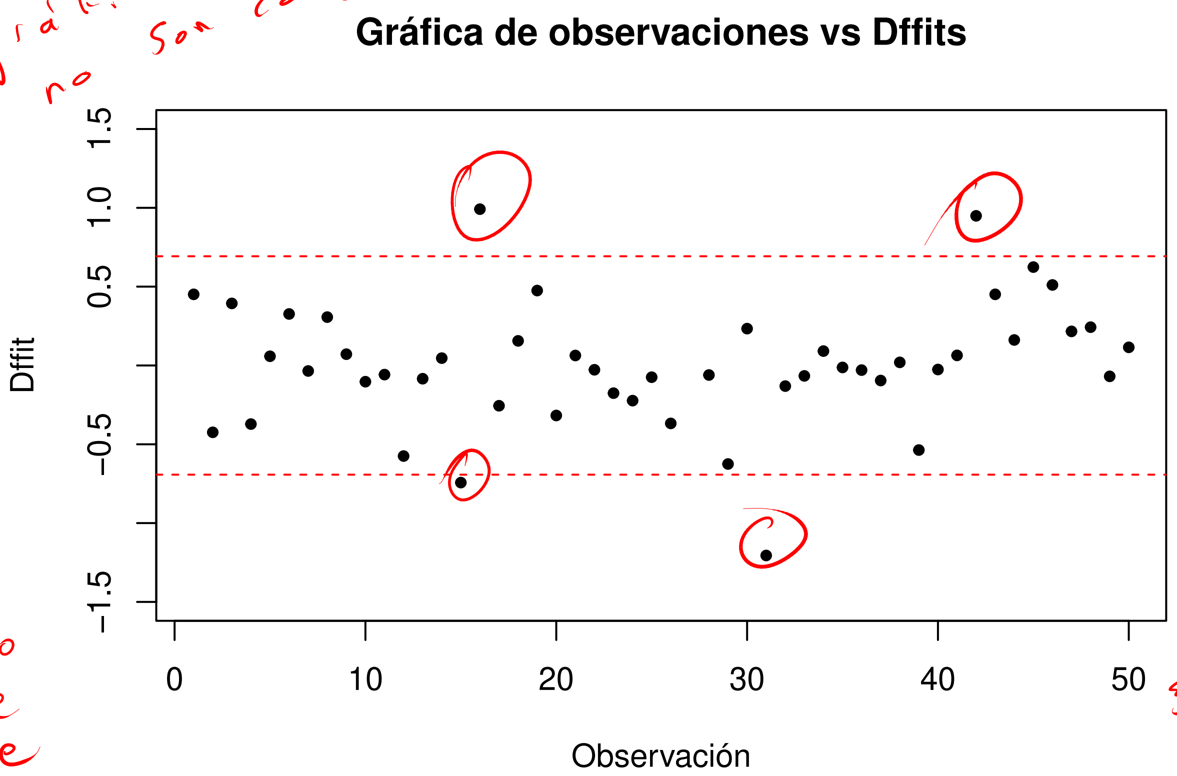


Figura 5: Criterio distancias de Cook para puntos influenciales

Se ven 4 en gráfica y tienen 5, no son congruentes.



3 pt

Figura 6: Criterio Dffits para puntos influenciales

no se ve

```
##    res.stud Cooks.D hii.value  Dffits
## 15  -1.9393  0.0861    0.1208 -0.7431
## 16   2.2936  0.1476    0.1441  0.9916
## 27  -2.5901  0.9955    0.4710 -2.6243
## 31  -0.9793  0.2425    0.6027 -1.2056
## 42   1.6650  0.1440    0.2376  0.9493
```

tabla, no salida de R

¿cuáles? ¿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, ninguno de los datos cumple con serlo.

¿Qué causan?

4.3. Conclusión

3 pt

El modelo no es valido devido a que no se cumple el supuesto de normalidad ya que atravez de grafica llegamos a dicha comclucion, esto se puede dar devido a los puntos de balance que pueden ver por fuera de la condicion propuesta en grafica de que observa los Valores de hii.