3,6

Trabajo 1

Estudiantes

Juan Pablo Muñoz Jimenez

Los otros estudiantes asignados a mi equipo no respondieron.

Ay :C

Equipo #47

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:

17 pt

Guiándonos según la base de datos brindada (la cual es la del equipo #47) existen 5 variables regresoras denominadas por:

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

Sabiendo esto, se plantea el modelo de regresión lineal multiple:

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

1.1. Modelo de regresión

Realizando el ajuste del modelo, se generaron los siguientes coeficientes como resultado:

Cuadro 1: Tabla de valores de los coeficientes del modelo.

	Valor del parámetro
β_0	-0.1074
β_1	0.1925
β_2	0.0200
β_3	0.0910
β_4	-0.0023
β_5	0.0017

2 pt

No va en c. ajustada

Entonces, el modelo de regresión ajustado quedaría así:

$$\hat{Y}_i = -0.1074 + 0.1925 X_{1i} + 0.02 X_{2i} + 0.091 X_{3i} - 0.0023 X_{4i} + 0.0017 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 50$$

Y por consiguiente para realizar un analisis de la regresión, entonces se formula el siguiente juego de hipotesis:

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

Que tiene como estadistico de prueba:

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

$F_0 = \frac{MSR}{MSE} \neq \frac{MST}{MSE}$ 9 pt

Entonces con base a la siguiente tabla ANOVA:

Cuadro 2: Tabla ANOVA del modelo

	Suma Cuadratica	g.l.	Media Cuadratica	F_0	P-valor
Regresión	47.3921	5	9.478411	12.0552	2.21457e-07
Error	34.5951	44	0.786253		

Se sabe que se indica un valor P aproximado a cero, lo que lleva al rechazo de la hipótesis nula. En su lugar, se acepta la hipótesis alternativa que señala que al menos uno de los β_j es distinto de cero. Por lo tanto, se concluye que se rechaza la hipótesis nula y se acepta la hipótesis alternativa debido a la significancia de al menos uno de los parámetros. -) un poco redundante pero bien.

1.2. Significancia de los parámetros

6 pt

Según la siguiente tabla, podemos determinar los valores significativos de los mismos:

Cuadro 3: Resumen de los coeficientes

	Valor del parametro	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.1074	2.1125	-0.0508	0.9597
β_1	0.1925	0.0785	2.4514	0.0183
β_2	0.0200	0.0384	0.5207	0.6052
β_3	0.0910	0.0172	5.2854	0.0000
β_4	-0.0023	0.0087	-0.2593	0.7966
β_5	0.0017	0.0009	2.0189	0.0496

Para saber cuales parametros son significativos, se observa la tabla y se toma un nivel de significancia de $\alpha = 0.05$ que como no lo indica el trabajo, se toma este normalmente como referencia.

Se concluye entonces que los parametros β_1, β_3 y β_5 son significativos porque sus P-valores son menores que alfa ($\alpha = 0.05$) los cuales son 0.0183, 0.0000 y 0.0496 respectivamente y se concluye que hay diferencias significativas. Tomando como ejemplo el parametro β_3 el cual su P-valor es 0.0000(o 0.0000037), significa que es muy poco probable, practicamente casi nulo, que los resultados que obtuvieron estos parametros en el estudio sean el resultado del azar o de alguna fluctuación aleatoria y además es una evidencia muy fuerte para rechazar una hipotesis nula.

1.3. Interpretación de los parámetros 2 pt

$\hat{\beta}_1$**: Este parámetro está asociado la duración de la estadía (en días), al aumentar sus valores(que aumentarían de 0.1925), la probabilidad de riesgo de infección sería mayor. A mayor duración de estadía (días) en el hospital, mayor riesgo de infección se presenta dado a que su parámetro es positivo.**

x X_1 aumenta en 1, la pue aumenta 0,1925 es Y

$\hat{\beta}_3$**: Este parámetro está asociado al número promedio de camas en el hospital, nuevamente este parámetro es positivo lo cual significa un aumento de camas, al aumentar estas, la probabilidad de riesgo de infección también aumentaría en el hospital dado a la cantidad de personas.**

$\hat{\beta}_5$**: Este parámetro está asociado al numero promedio de enfermeras en el hospital, según la tabla y el parámetro (que es positivo), a mayor enfermeras, mayor sería la probabilidad de riesgo de infección.**

1.4. Coeficiente de determinación múltiple R^2 3 pt

El modelo de regresión presentado en este trabajo explica alrededor del 57.8 % de la variabilidad total observada en la respuesta, como se indica por su coeficiente de determinación múltiple R^2 de 0.578.

¿Cómo se calcula?

2. Pregunta 2 4 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Analizando la tabla anterior "Cuadro 3: Resumen de los coeficientes" se puede observar que X_2, X_4 y X_5 son las tres variables con los valores-p más grandes, haciendo excepción en X_0 la cual no se toma debido a que se debe tomar valores que acompañan variables. Entonces, se busca realizar una prueba de hipótesis utilizando una tabla que contenga todas las regresiones posibles.

Prueba de hipotesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones posibles

	Suma cuadrática de error	Covariables en el modelo
Modelo completo:	34.595	X1 X2 X3 X4 X5
Modelo reducido:	38.390	X1 X3

Entonces 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 y se pone 2 en el cuantil, al igual que se divide, dado a que son dos hipotesis, entonces: X son 3 parámetros en H_0

2 pt

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_1, \beta_3, \beta_5) - SSE(\beta_0, \ \cdots, \ \beta_5))/2 \ 3}{MSE(\beta_1, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{2\ 3,44} \\
&= \frac{(38.390 - 34.595)/2 \ 3}{38.390/44} \\
&= 2.1747851 \quad \times
\end{aligned}
\tag{2}
$$

2 pt

Ahora, comparando el $F_0 = 2.1747851$ con $f_{0.95,3,44} = 2.8165$, se puede ver que $F_0 < f_{0.95,1,45}$ Entonces como F_0 es menor a $f_{0.95,1,45}$, el conjunto no es significativo y no se rechaza la hipotesis nula y por consiguiente es posible descartar las variables del subconjunto del modelo. ✓

3. Pregunta 3

0,5 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Si planteamos la pregunta: ¿Existe una relación significativa entre el riesgo de infección y las variables Duración de la estadía, Rutina de cultivos R, Número de camas, Censo promedio diario y Número de enfermeras en los hospitales de EE.UU? X Es el efecto de las var's

La hipótesis nula sería:

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

esto no es de exclusivo de PH lineal general es signif. de regresión

lo que significa que ninguna de las variables explicativas (X1, X2, X3, X4, X5) tiene un efecto significativo sobre el riesgo de infección Y.

Con **L** dada por

0 pt

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

0 pt

El modelo reducido se obtiene al igualar todas las pendientes de regresión a cero, es decir, $Y = \beta_0$. $+ \varepsilon_i$

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

3.2. Estadístico de prueba

este lo conocías

El estadístico de prueba F_0 se calcula como:

0,5 pt

$$
F_0 = \frac{(SSE(MR) - SSE(MF))/2 \ 5}{MSE(MF)} \overset{H_0}{\sim} f_{2,44} = \frac{(SSE(MR) - 34.595)/2 \ 5}{38.390/44}
\tag{3}
$$

Si el valor p del estadístico de prueba es menor que el nivel de significancia (por ejemplo, 0.05), se rechaza la hipótesis nula y se concluye que al menos una de las variables explicativas tiene un efecto significativo sobre el riesgo de infección.

19,5 pt

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

4pt

Para comprobar si este supuesto es válido, se llevará a cabo una prueba de hipótesis de ~~Shapiro-Wilk~~ y se complementará con un gráfico de cuantil-cuantil.

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

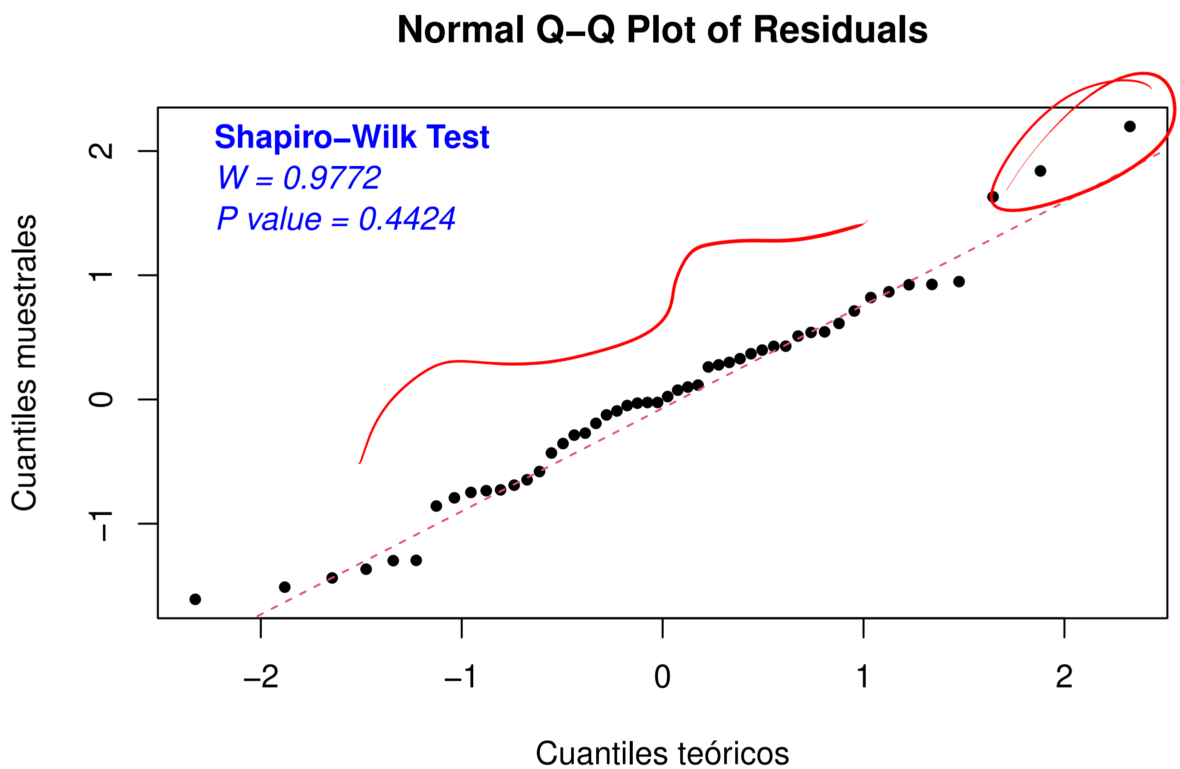

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

En el gráfico se observa alta dispersion de puntos al principio y al final de la linea formada, lo esperado para que el grafico fuese normal sería una distribución más concentrada por la media de la distribución de estos. Si se toma un nivel estandar de significancia $\alpha = 0.05$, el p-valor(0.4424) en este caso se distancia mucho de este alfa y es mayor que este, entonces la hipótesis nula, que establece que los datos se distribuyen normalmente, no sería tan probable en este caso. En otras palabras, se rechazaría la normalidad.

A continuación, se verificará si la varianza mantiene su estabilidad, lo cual es un supuesto importante que debe ser comprobado.

4.1.2. Varianza constante

0pt

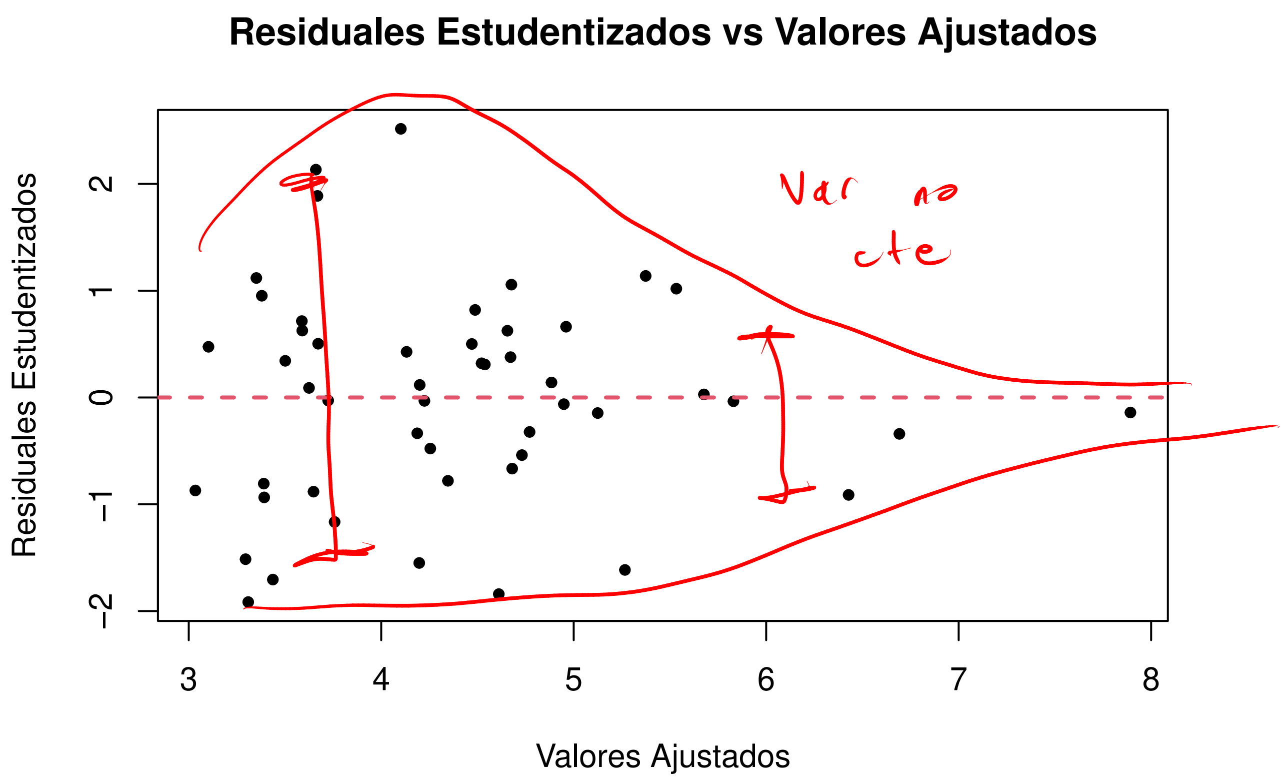

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de Residuales "Estudentizados vs Valores Ajustados" no se observa muy uniformemente los puntos sobre la linea que pasa por 0, pocos de estos se situan sobre esta. La dispersión de los puntos no es uniforme en toda la gama de los valores, entonces es probable que el gráfico tenga una varianza no constante pero no se afirma aún esto dado a que no existe evidencia suficiente como para dar esta conclusión. → si hay y mucha

huh? no tiene nada que ver.

4.1.3. Datos atípicos

3pt

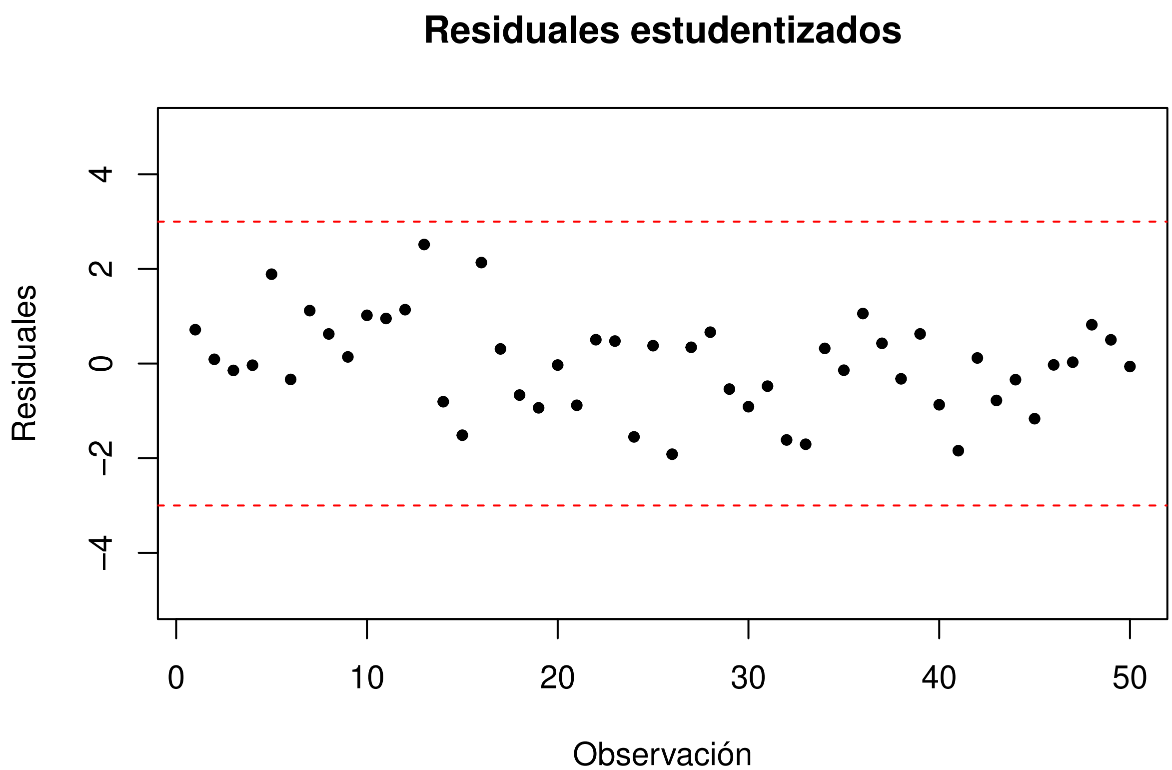

Figura 3: Identificación de datos atípicos

En otras palabras, la gráfica muestra que no hay valores inusuales o extremos en el conjunto de datos, ya que ninguno de los residuos estandarizados supera el criterio establecido de $|r_{estud}| > 3$.

4.1.4. Puntos de balanceo

2 pt

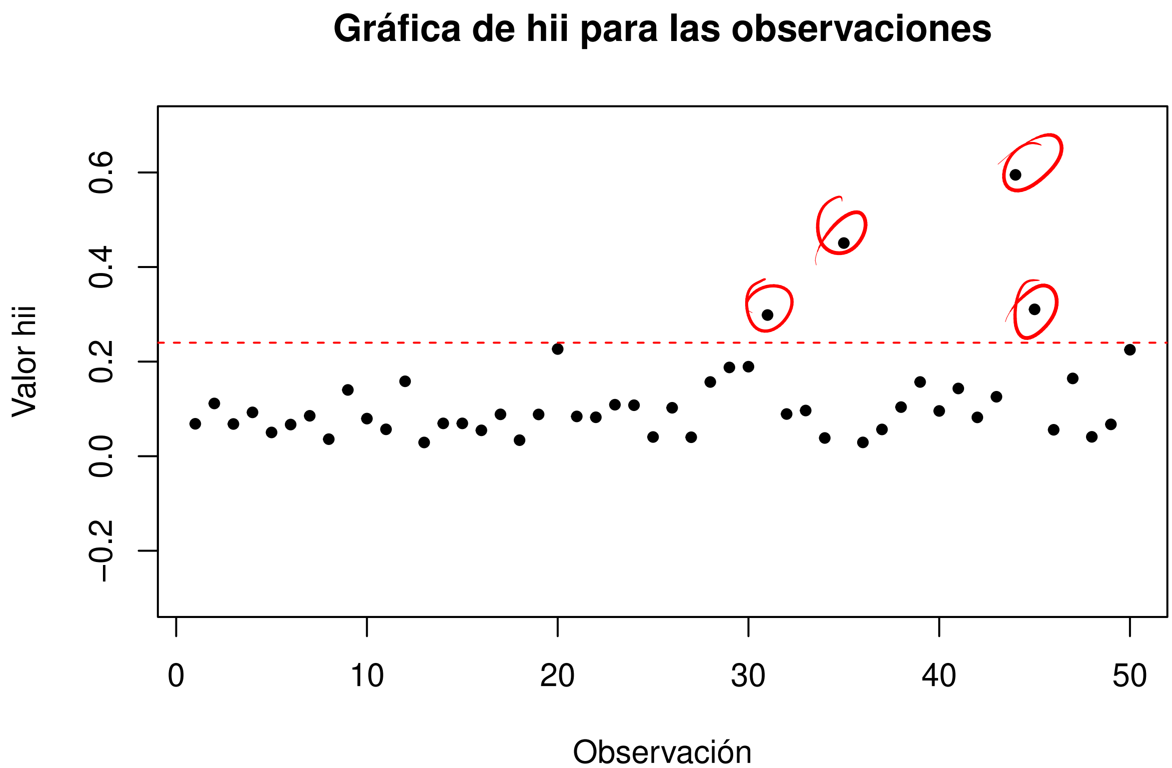

Figura 4: Identificación de puntos de balanceo

```
##    hii.value hii.value.1
## 31    0.2985      0.2985
## 35    0.4507      0.4507
## 44    0.5949      0.5949
## 45    0.3105      0.3105
```

→ tabla, aunque igual no baja nota.

Observando la gráfica “Gráfica de hii h_{ii} para las observaciones” se evidencian 4 puntos de balanceo por encima de la linea punteada roja quien representa el valor $h_{ii} = 2\frac{6}{50} = 0.24$ y así también como se puede ver en la tabla de abajo del gráfico confirmando la existencia de estos mostrando los puntos de balanceo.

¿Qué causan?

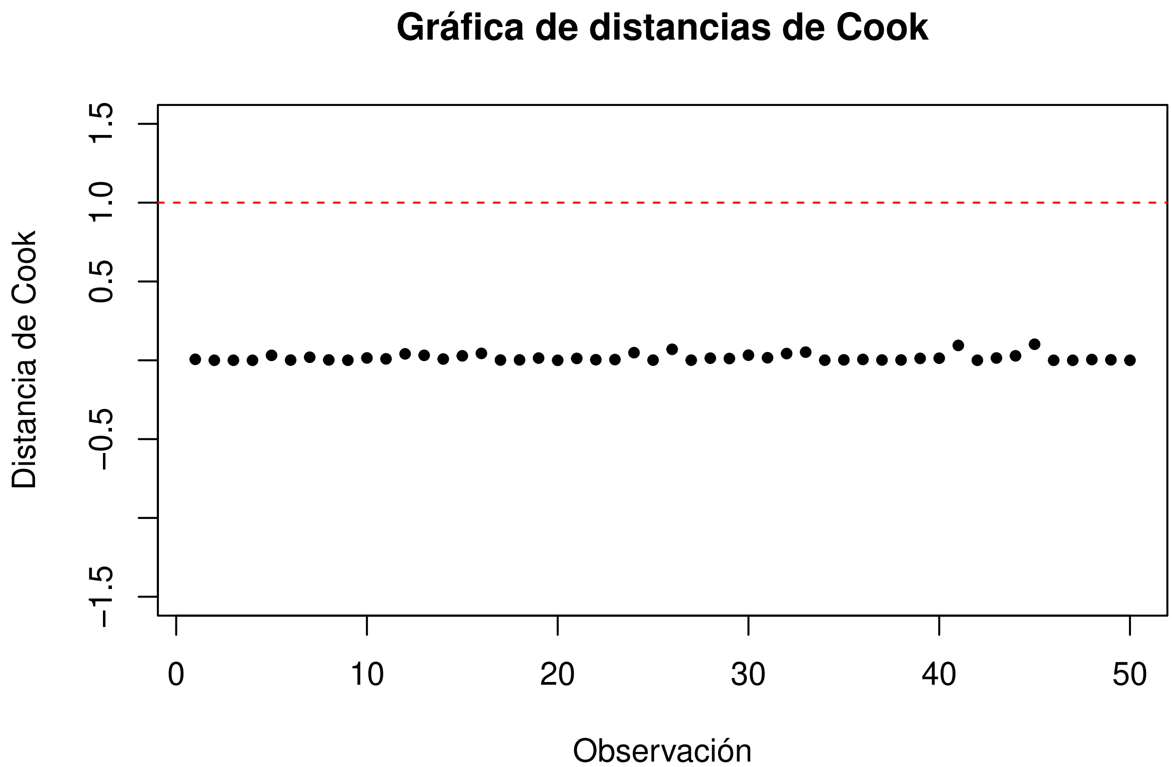

Figura 5: Criterio distancias de Cook para puntos influenciales

~~No hay puntos influenciales mayores a 1~~ (linea roja punteada) en esta Gráfica de distancias de Cook.

No hay distancias de cook mayor a 1

1pt

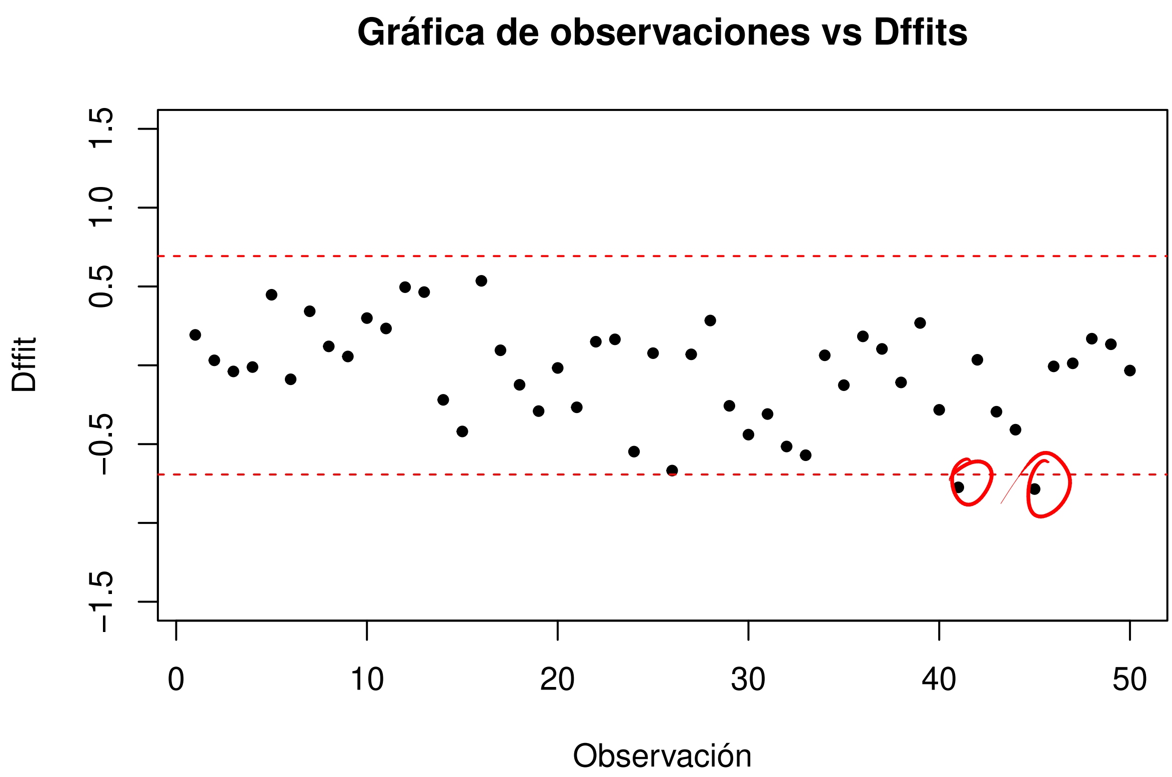

Figura 6: Criterio Dffits para puntos influenciales

1,5 pt

```
##     Dffits Dffits.1
## 41 -0.7743  -0.7743
## 45 -0.7852  -0.7852
```

Analizando la gráfica y comprobando con la tabla donde se muestran los puntos influenciales en la misma, se hace evidencia de 2 puntos influencias bajo el criterio de Dffits el cual está dado por $|D_{ffit}| > 2\sqrt{\frac{6}{50}} = 0.6928$ (quien representa las dos lineas punteadas tanto positiva como negativas) estos dos puntos siendo menores a -0.6928, siendo -0.7743 y -0.7852.

¿Qué causan?

4.2. Conclusión

3 pt

Al no ser cumplido el supuesto de normalidad debido a la dispersion de sus puntos al principio y al final de la media punteada, como también el supuesto de varianza y su estabilidad debido a que la dispersión de los puntos no es uniforme en toda la gama de los valores y además la existencia de varios puntos influenciales, entonces se determina la invalidez del modelo.