Trabajo 1

Estudiantes

3,8

David Leon Ruiz Herrera
Luisa Camila Rios Ramirez
Cristhian Gallego Avila

Equipo 63

Docente

Javier Armando Lozano

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
05 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

19pt

Teniendo en cuenta la base de datos del Equipo63, en la cual hay 5 variables regresoras definidas como:

Y: Riesgo de infección [%] X_1: Duración de la estadía [días] X_2: Rutina de cultivos [por cada 100] X_3: Número de camas X_4: Censo promedio diario X_5: Número de enfermeras

Entonces, se plantea el siguiente modelo de regresión lineal múltiple(RLM):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i;$$

donde

$$\varepsilon_i \overset{idd}{\sim} N(0, \sigma^2) \quad 1 \leq i \leq 54$$

1.1. Modelo de regresión

Al ajustar el modelo planteado según los datos, se obtiene la siguiente tabla de coeficientes

3pt

Cuadro 1: Valores de los coeficientes

	Valor del parametro
β_0	1.8165
β_1	0.1914
β_2	-0.0185
β_3	0.0596
β_4	0.0022
β_5	0.0018

Por ende, el modelo de regresión ajustado es:

$$\hat{Y}_i = 1.8165 + 0.1914 X_{1i} - 0.0185 X_{2i} + 0.0596 X_{3i} + 0.0022 X_{4i} + 0.0018 X_{5i}$$

donde:

$$1 \leq i \leq 54$$

1.2. Significancia de la regresión

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:

5pt

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,48} \tag{1}$$

Además, sea esta la tabla ANOVA:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	grados de libertad	Cuadrado medio	F_0	P-valor
Regresión	47.9174	5	9.58347	12.5994	7.81137e-08
Error	36.5102	48	0.76063		

De la tabla Anova, se observa que bajo un nivel de significacia del 5 %, valor p $< \alpha$, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, entonces al menos un parametro del modelo de regresión mútiple es diferente de 0, es decir, la regresión es estadísticamente significativa.

1.3. Significancia de los parámetros

Primero observemos el juego de hipotesis para la prueba individual de la significancia de los parametros.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \text{ con } 0 \leqslant j \leqslant 5 \end{cases}$$

En el siguiente cuadro se presenta la Tabla de coeficientes, la cual permitirá, entre otras cosas, determinar cuáles de los parametros son significativos en nuestro modelo:

6pt

Cuadro 3: Tabla de los coeficientes

	Estimate	Std.error	T_{0j}	Valor P
β_0	1.8165	1.8257	0.9949	0.3248
β_1	0.1914	0.0788	2.4283	0.0190
β_2	-0.0185	0.0340	-0.5437	0.5892
β_3	0.0596	0.0148	4.0278	0.0002
β_4	0.0022	0.0077	0.2815	0.7795
β_5	0.0018	0.0008	2.2952	0.0261

Los respectivos valores P nos permiten concluir que con un nivel de significancia de $\alpha = 0.05$, los parámetros β_1 β_3 y β_5 son significativos, pues sus P-valores son menores a α, por lo que se rechaza H_0.

1.4. Interpretación de los parámetros

3pt

Importante mencionar que β_0 no tiene interpretación pues no hay una coordenada

$$(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}) = (0, 0, 0, 0, 0)$$

$\hat{\beta}_1$:=0.1914 indica que por cada cantidad de aumento de la duración de la estadía [días], el promedio del resultado en la prueba de riesgo de infección aumenta en 0.1914 unidades, cuando las demas variables predictoras se mantienen fijas.

$\hat{\beta}_3$:=0.0596 indica que por cada cantidad de aumento en el número de camas, el promedio del resultado en la prueba de riesgo de infección aumenta en 0.0596 unidades, cuando las demas variables predictoras se mantienen fijas.

$\hat{\beta}_5$:=0.0018 indica que por cada cantidad de aumento en el número de enfermeras, el promedio del resultado en la prueba de riesgo de infección aumenta en 0.0018 unidades, cuando las demas variables predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2 2 pt

El modelo tiene un $R^2 = 0,5675$, se interpreta que el 56.75 % de la variabilidad total en los resultados del porcentaje de riesgo de infección es explicada por el modelo de regresión múltiple propuesto, es decir, nos indica una poca asociación lineal, pero esto no quiere decir que no se garantice los supuestos básicos del modelo, que se comprabarán más adelante.

¿cómo se calcula?

2. Pregunta 2 0 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

→ bajo...

Las variables regresoras con los valores P más alto en el modelo fueron X_1, X_3, X_5 (Para este juego de hipótesis B_0 no es tomado en cuenta), por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

?

Nada coincide

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	36.510	X1 X2 X3 X4 X5
Modelo reducido	36.936	X1 X3 X5

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$ ✗

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_0,\beta_1,\beta_3,\beta_5) - SSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5))/3}{MSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)} \overset{H_0}{\sim} F_{2,48} \\
&= \frac{(36.936 - 36.510)/2}{0.76063} \qquad (2)\\
&= 0,2800
\end{aligned}
$$

0 pt

Si $F_0 < f_{\alpha,k,n-p}$, se rechaza

Ahora, comparando el F_0 con $f_{0.95,2,48} = 3.1907$, se puede ver que $F_0 < f_{0.95,2,48}$

0 pt

Por lo tanto, no se rechaza H_0, es decir que las variables: Rutina de cultivos [por cada 100] y Censo promedio diario.

3. Pregunta 3

3 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si las variables predictoras X_2 y X_4 son colineales y las variables predictoras X_1 y X_3 presentan colinealidad en el modelo. por consiguiente se plantea la siguiente prueba de hipótesis:

$$
\begin{cases}
\mathrm{H}_0 : \beta_2 = \beta_4 \ , \ \beta_1 = \beta_3 \\
\mathrm{H}_1 : \text{Al menos una de las igualdades no se cumple}
\end{cases}
$$

lo que es equivalente a lo siguiente:

$$
\begin{cases}
\mathrm{H}_0 : \beta_2 - \beta_4 = 0, \ \beta_1 - \beta_3 = 0 \\
\mathrm{H}_1 : \text{Al menos una de las igualdades no se cumple}
\end{cases}
$$

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

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

2 pt

El modelo reducido está dado por:

$$
Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_2 X^*_{2,i} + \beta_5 X_{5,i} + \varepsilon_i \ , \ \varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2) \ \ 1 \leqslant i \leqslant 54
$$

0 pt

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

3.2. Estadístico de prueba

→ Reemplazar

El estadístico de prueba F_0 está dado por:

1 pt

$$
F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,48} \tag{3}
$$

4. Pregunta 4

16pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

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

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

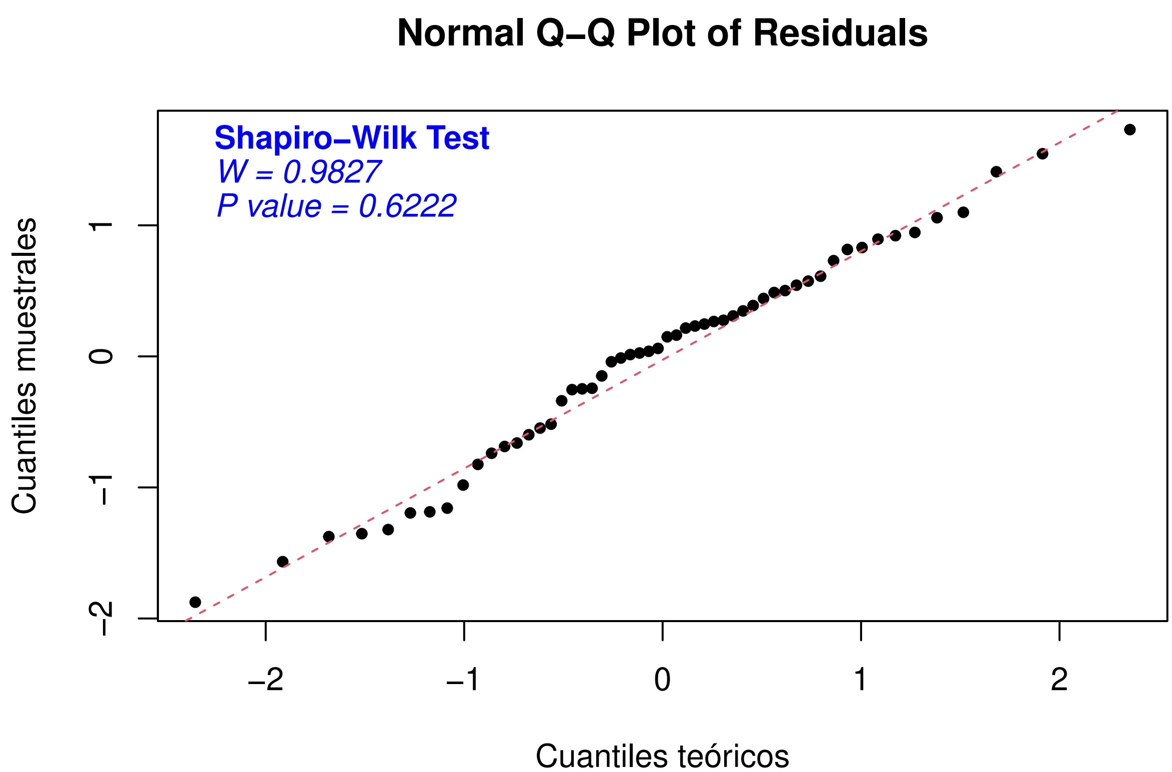


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

4pt

Al ser el P-valor aproximadamente igual a 0.6222 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mayor y por lo tanto, no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal, aunque hay desviaciones en los dos extremos, pues hay varios datos alejados de la linea roja, por lo que se identifican como posibles datos atípicos, de balanceo o de influencia.

Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

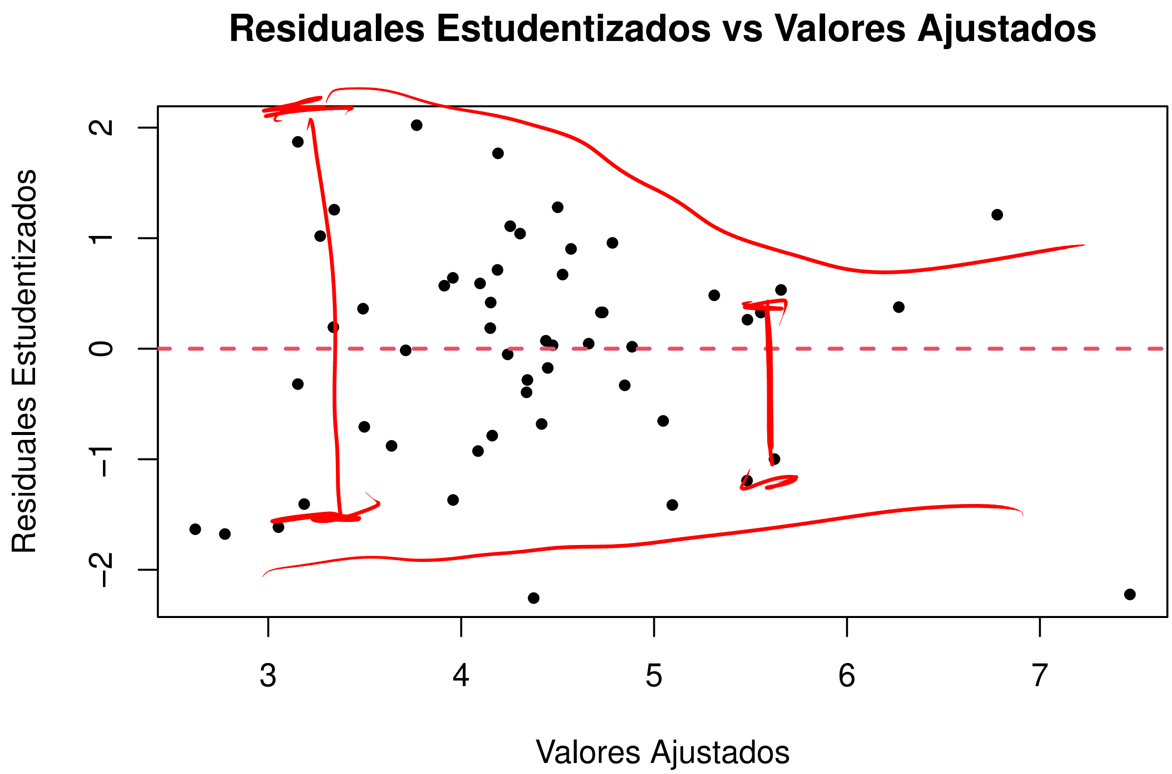


Figura 2: Gráfico residuales estudentizados vs valores ajustados

Siempre pasa con r. estud. ?

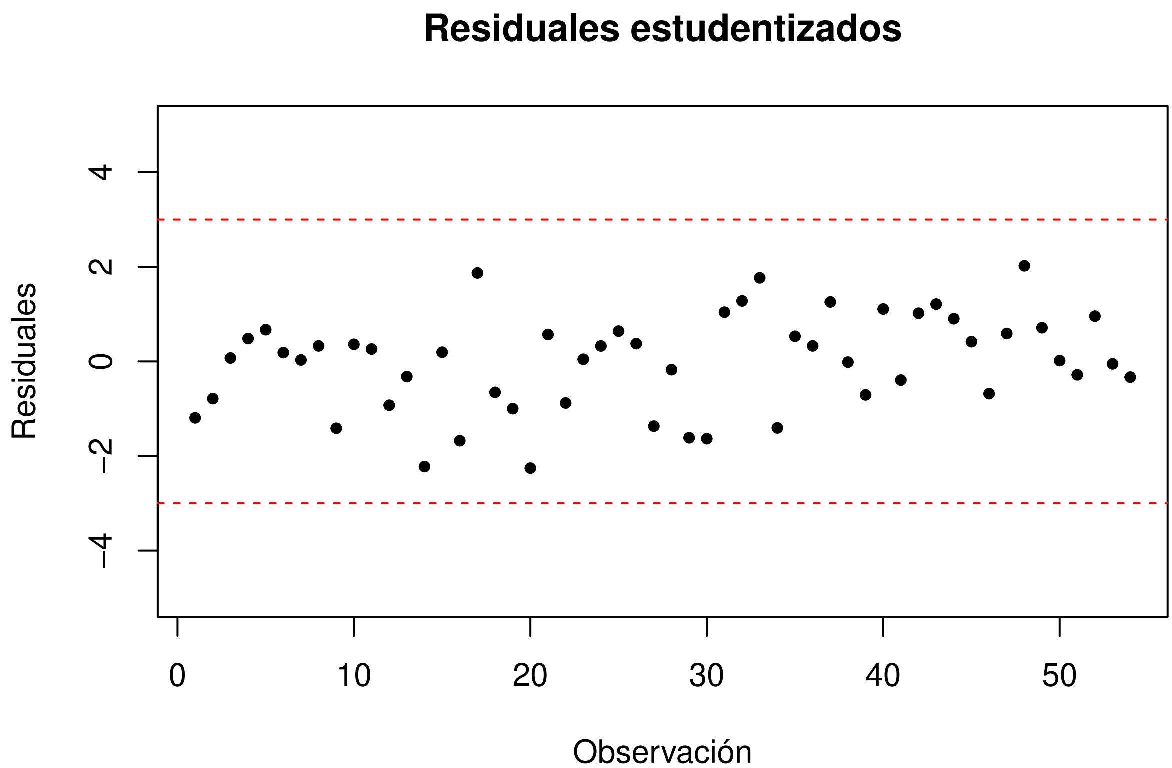
En el gráfico de residuales estudentizados vs valores ajustados se puede observarla mayoría de los datos entre los dos primeros tercios, especialmente en la primera mitad del gráfico, que el supuesto de varianza constante se cumple y que además se organizan alrededor del 0. En general, el supuesto se cumple pero vale mencionar de que en último tercio hay presencia de algunos datos que puedan afectar este supuesto dado que se encuentran alejados de los demás, y como mencionamos en la prueba de normalidad anterior, son posibles valores extremos.

había patrón de decrecimiento

2pt

4.2. Verificación de las observaciones

4.2.1. Datos atípicos



3e+

Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica de dispersión anterior, no hay datos atípicos en el modelo, ya que ninguno de los datos se encuentra por fuera del rango (-3,3), es decir que no cumplen que $|r_i| > 3$.

4.2.2. Puntos de balanceo

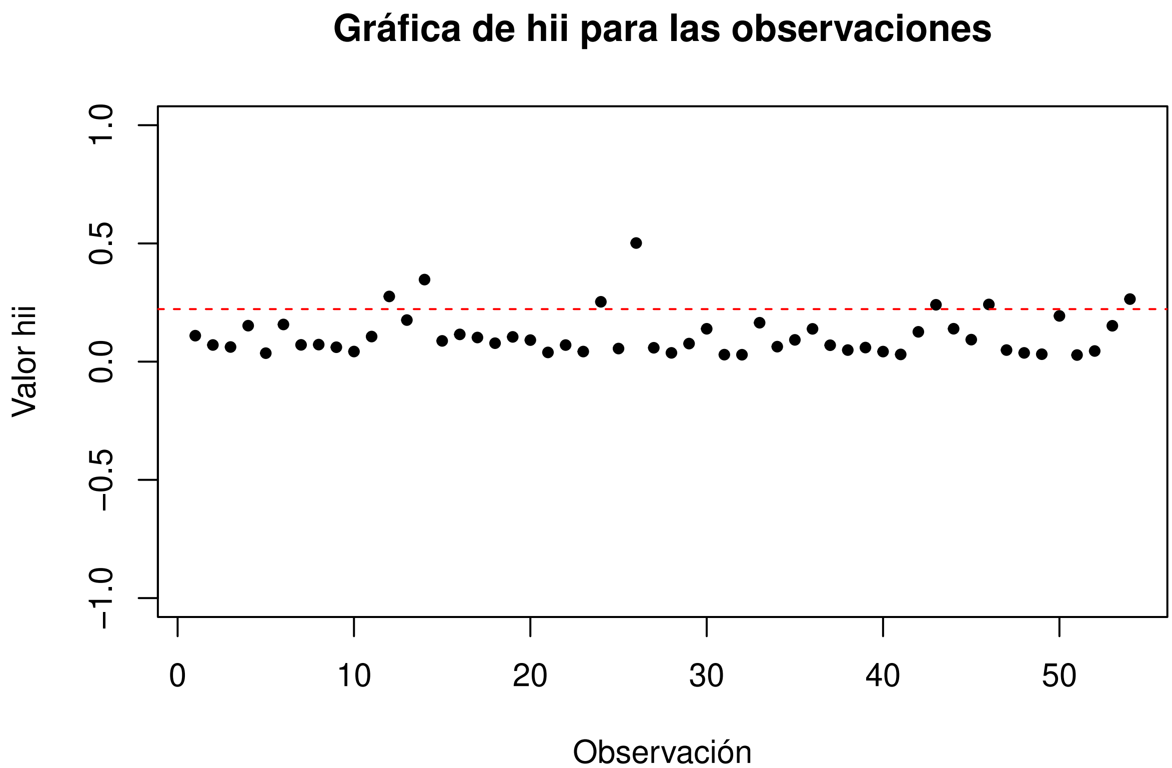


Figura 4: Identificación de puntos de balanceo

	hii
12	0.2757
14	0.3470
24	0.2531
26	0.5017
43	0.2405
46	0.2419
54	0.2646

2pt

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, es decir $h_{ii} = 0.222$, se reconocen 7 puntos de balanceo bajo el criterio que su respectivo $h_{ii} > 0.222$, los cuales están presentados en la tabla.

¿Cuáles son...?

4.2.3. Puntos influenciales

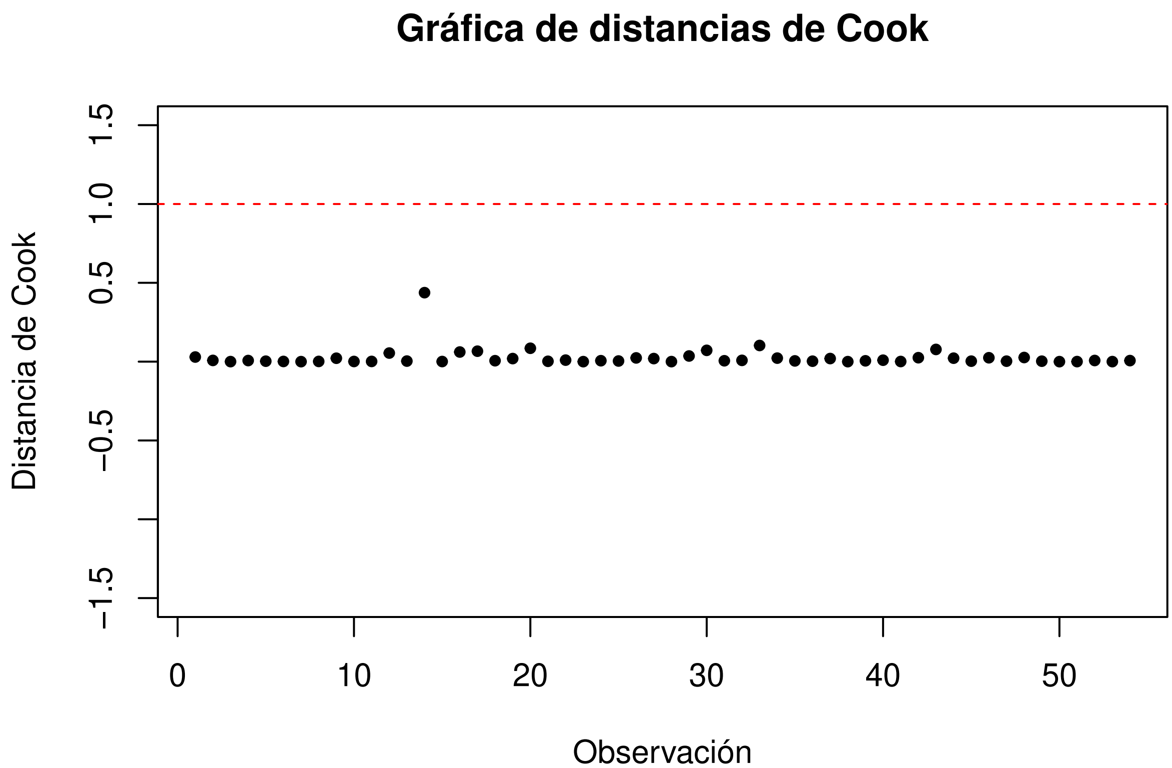


Figura 5: Criterio distancias de Cook para puntos influenciales

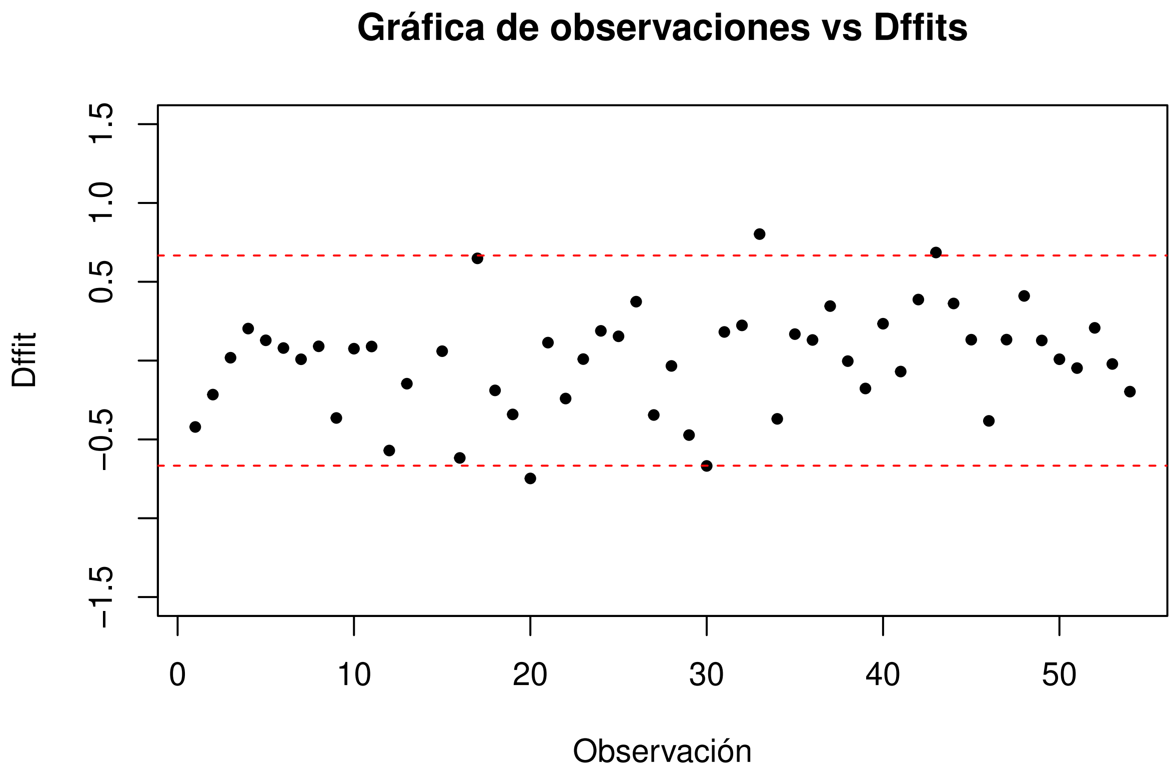


Figura 6: Criterio Dffits para puntos influenciales

	Dffits
14	-1.6929
20	-0.7475
30	-0.6682
33	0.8025
43	0.6856

→ No se ve

causan...?

2pt

Como se puede ver,las observaciones {14, 20, 30, 33, 43} son puntos influenciales según el criterio de Dffits, pues si $|D_{ffit}| > 2\sqrt{\frac{6}{54}}$, es un punto influencial, lo cual puede verse representado en la tabla. 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.

4.3. Conclusión

3pt

En conclusión, respecto a la validez del modelo podemos decir que es valido dado que se cumplen los supuestos los cuales se les planteó una prueba de hipotesis, como lo son el normalidad, y varianza constante igual a σ^2