Trabajo 1

4,6

Estudiantes

Luis Fernando Henao Henao
Joan Sebastian Lopez Salas
Juan Manuel Roldán Zalazar
José Manuel Rueda Jaramillo

Equipo 13

Docente

Julieth Verónica Guarín Escudero

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

18pt

Teniendo en cuenta la base de datos asignada a nuestro grupo, se plantea el siguiente de RLM en el cual hay 5 variables regresoras:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

Donde:

- 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
- ε: Errores aleatorios

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores de coeficientes del modelo ajustado

	Valor del parámetro
$\hat{\beta}_0$	-2.1975
$\hat{\beta}_1$	0.2492
$\hat{\beta}_2$	0.0442
$\hat{\beta}_3$	0.0495
$\hat{\beta}_4$	0.0120
$\hat{\beta}_5$	0.0010

3pt

Con base a la tabla de valores de coeficientes del modelo ajustado, se obtiene la ecuación de regresión ajustada:

$$\hat{Y}_i = -2.1975 + 0.2492 X_{1i} + 0.0442 X_{2i} + 0.0495 X_{3i} + 0.012 X_{4i} + 0.001 X_{5i}\ ;\ 1 \leqslant i \leqslant 59$$

1.2. Significancia de la regresión

β_0 no va acá

Se plantea el siguiente juego de hipótesis para analizar la significancia de la regresión:

$$\begin{cases} H_0 : \cancel{\beta_0} = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \text{Algún } \beta_j \neq 0 \text{ para j=1, 2,..., 5.} \end{cases}$$

y β_0 ?

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,53} \tag{1}$$

Para el análisis del juego de hipótesis se presenta la tabla Anova relacionada:

3pt

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresion	49.2885	5	9.857709	10.7054	3.81717e-07
Error	48.8033	53	0.920817		

De la tabla Anova, se observa un valor del estadístico de prueba $F_0 = 10.7054$ y su correspondiente valor-P aproximadamente igual a 0, por lo que, con un nivel de significancia de $\alpha = 0.05$, se rechaza la hipótesis nula en la que $\beta_j = 0$ con $\cancel{0} \leqslant j \leqslant 5$, 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

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar cuáles de ellos son significativos.

Cuadro 3: Tabla resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-2.1975	1.7061	-1.2880	0.2033
β_1	0.2492	0.1157	2.1537	0.0358
β_2	0.0442	0.0281	1.5732	0.1216
β_3	0.0495	0.0132	3.7614	0.0004
β_4	0.0120	0.0088	1.3636	0.1785
β_5	0.0010	0.0007	1.4474	0.1537

6pt

Los P-valores presentes en la tabla permiten concluir que, con un nivel de significancia $\alpha = 0.05$, los parámetros individuales $\hat{\beta}_1$ y $\hat{\beta}_3$ son significativos en presencia de los demás parámetros, pues sus P-valores son menores a α.

Por otro lado, los parametros $\hat{\beta}_0$, $\hat{\beta}_2$, $\hat{\beta}_4$ y $\hat{\beta}_5$ no son individualmente significativos en presencia de los demás parámetros.

1.4. Interpretación de los parámetros estimados

Los parámetros suceptibles de interpretación son aquellos que resultaron significativos individualmete, los cuales son $\hat{\beta}_1$ y $\hat{\beta}_3$.

3pt

- $\hat{\beta}_1$ indica que por cada día más de estadía en el hospital ($\mathbf{X_1}$), la probabilidad promedio estimada de adquirir infección en el hospital aumenta a razón de 0.2492 mientras las otras variables predictoras permanezcan constantes.
- $\hat{\beta}_3$ indica que por cada cama que se agregue durante el perido de estudio ($\mathbf{X_3}$), la probabilidad promedio estimada de adquirir infección en el hospital aumenta a razón de 0.0495 mientras las otras variables predictoras permanezcan constantes.

1.5. Coeficiente de determinación múltiple R^2

3pt

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} = \frac{49.2885}{49.2885 + 48.8033} = 0.5024732 \tag{2}$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.50247$, lo que significa que aproximadamente el 50 % de la variabilidad total observada en la respuesta es explicada por el modelo de RLM propuesto en el presente informe.

2. Pregunta 2

4,5 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más pequeño en el modelo fueron X_1, X_2 y X_3, por lo tanto, a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis para probar la significancia simultánea de las covariables mencionadas:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \neq 0 \text{ para j=1, 2, 3.} \end{cases}$$

Para probar este juego de hipótesis se usa la suma de cuadrados extra. Para el cálculo de esta SSextra se deben definir los modelos completo y reducido (bajo H_0); del siguiente cuadro se identifican las SSE de los modelos mencionados.

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	48.803	X1 X2 X3 X4 X5
Modelo reducido	73.137	X4 X5

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

$$Y_i = \beta_0 + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

2.2. Estadístico de prueba y conclusión

2,5pt

Para esta prueba de hipótesis se construye el estadístico de prueba F_0 como:

$$F_0 = \frac{(SSR(\beta_1, \beta_2, \beta_3|\beta_0, \beta_4, \beta_5))/3}{MSE(\beta_0,\ \cdots\ ,\ \beta_5)} = \frac{(SSE(\beta_0, \beta_4, \beta_5) - SSE(\beta_0,\ \cdots\ ,\ \beta_5))/3}{MSE(\beta_0,\ \cdots\ ,\ \beta_5)} \overset{H_0}{\sim} f_{3,53} \tag{3}$$

$F_0 \sim f_{3,53}$

(s No

$F_0 = \# \sim ?$

$$F_0 = \frac{5.05233}{0.92081} = \cancel{5.48679} \overset{H_0}{\sim} f_{3,53} \tag{4}$$

la distri. se coloca antes de calcular F_0

Ahora, comparando el F_0 con $f_{0.95,3,53} = 2.7791$, se puede ver que $F_0 > f_{0.95,3,53}$, y por lo tanto, se rechaza H_0.

Se concluye que no es posible descartar las variables del subconjunto ya que con un $\alpha = 0.05$, la probabilidad promedio estimada de adquirir infección en el hospital (Y) depende de al menos una de las variables predictoras X_1, X_2 y X_3.

2pt

3. Pregunta 3

5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hacen las siguientes preguntas:

- ¿ El efecto de la razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria por cada 100 pacientes ($\mathbf{X_2}$) sobre la probabilidad promedio estimada de adquirir infección en el hospital ($\mathbf{Y}$) es igual a 2 veces el efecto del número promedio de camas en el hospital durante el periodo del estudio ($\mathbf{X_3}$) sobre la probabilidad promedio estimada de adquirir infección en el hospital ($\mathbf{Y}$)?

- ¿ El efecto de la duración promedio de la estadía de todos los pacientes en el hospital en días ($\mathbf{X_1}$) sobre la probabilidad promedio estimada de adquirir infección en el hospital ($\mathbf{Y}$) es igual al efecto del número promedio de pacientes en el hospital por día ($\mathbf{X_4}$) más 2 veces el efecto del número promedio de enfermeras, equivalentes a tiempo completo, ($\mathbf{X_5}$) sobre la probabilidad promedio estimada de adquirir infección en el hospital ($\mathbf{Y}$) durante el periodo del estudio?

Por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = 2\beta_3;\; \beta_1 = \beta_4 + 2\beta_5 \\ \mathrm{H}_1 : \beta_2 \neq 2\beta_3 \text{ o } \beta_1 \neq 2\beta_5 + \beta_4 \end{cases}$$

Reescribiendo matricialmente:

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

2pt

Con $\mathbf{L}$ dada por

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

La matriz $\mathbf{L}$ tiene $r = 2$ filas linealmente independientes.

El modelo reducido está dado por:

1pt

$$MR : Y_i = \beta_0 + \beta_3 X_{3i}^* + \beta_4 X_{4i}^* + \beta_5 X_{5i}^* + \varepsilon_i,\; \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\; 1 \leqslant i \leqslant 59$$

Donde $X_{3i}^* = 2X_{2i} + X_{3i}$, $X_{4i}^* = X_{1i} + X_{4i}$, y $X_{5i}^* = 2X_{1i} + X_{5i}$.

3.2. Estadístico de prueba

2pt

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{SSH/2}{MSE(MF)} = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} = \frac{(SSE(MR) - 48.803)/2}{0.9208} \overset{H_0}{\sim} f_{2,53} \quad (5)$$

Solo falta determinar el valor de $SSE(MR)$, el cual no se puede obtener de la tabla de todas las regresiones posibles, ya que esta no admite sumas de variables entre sus opciones.

4. Pregunta 4 18,5 pt

4.1. Supuestos del modelo

Procedemos a validar los supuestos de normalidad y varianza constante de los errores del modelo.

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis que se realizará por medio de Shapiro-Wilk, acompañada de un gráfico cuantil-cuantil:

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

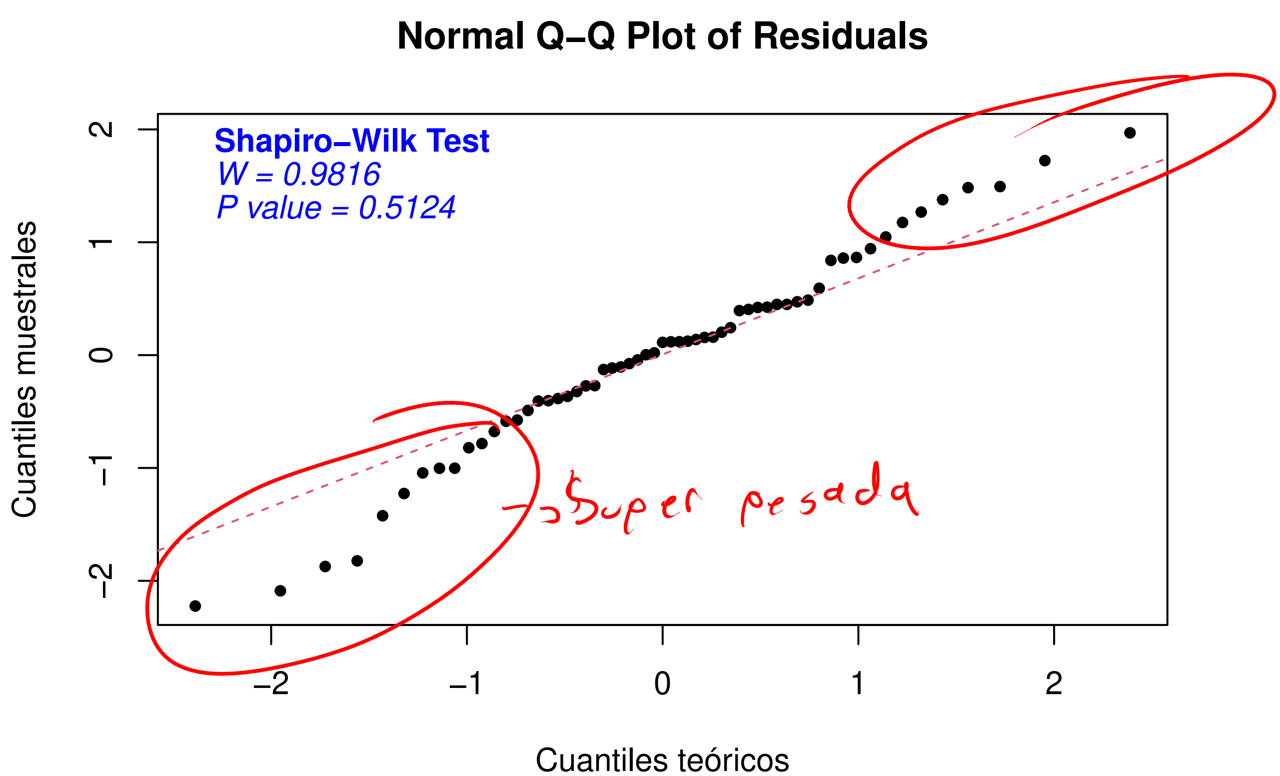


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

A pesar de que la pruebra de normalidad de Shapiro-Wilk indica que, con una significancia de $\alpha = 0.05$, lo errores se distribuyen normal($VP = 0.6692 > \alpha$), al revisar el

> Análisis muy breve, no hacen notoriedad de la pesada que es la cola inferior.

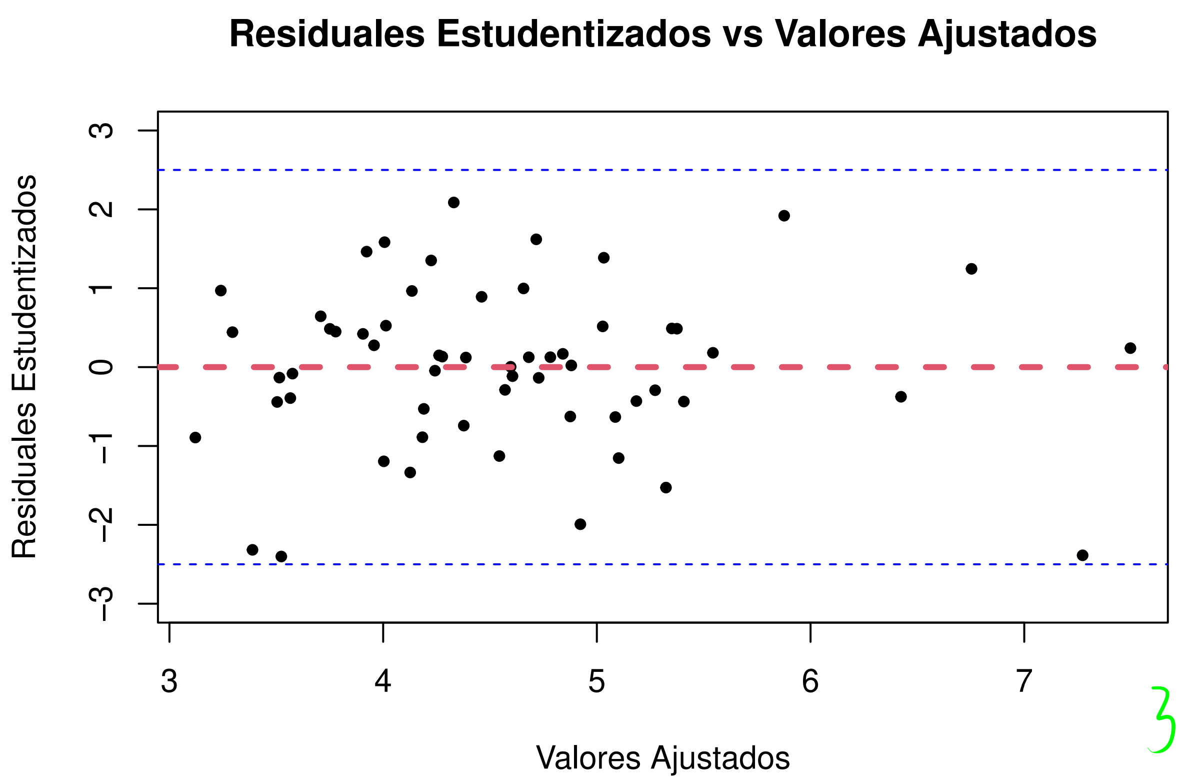
3,5pt

gráfico de cuantiles se evidencia que el patrón no sigue la línea roja que representa el ajuste de la distribución de los residuales a una distribución normal, mostrando colas pesadas en los extremos. Por este motivo, se rechaza H_0 y se concluye que el supuesto de normalidad no se cumple.

4.1.2. Varianza constante

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis acompañada de un gráfico de residuales estudentizados vs valores ajustados:

$$\begin{cases} \mathrm{H}_0 : V[\varepsilon_i] = \sigma^2 \\ \mathrm{H}_1 : V[\varepsilon_i] \neq \sigma^2 \end{cases}$$



3 pt

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Del gráfico se puede observar que la nube de puntos muestra un comportamiento lineal donde su dispersión con respecto a la linea roja no aumenta ni disminuye. Lo anterior nos lleva a concluir que el supuesto de varianza constante se cumple.

4.2. Verificación de las observaciones

Se procede a identificar observaciones extremas.

4.2.1. Datos atípicos

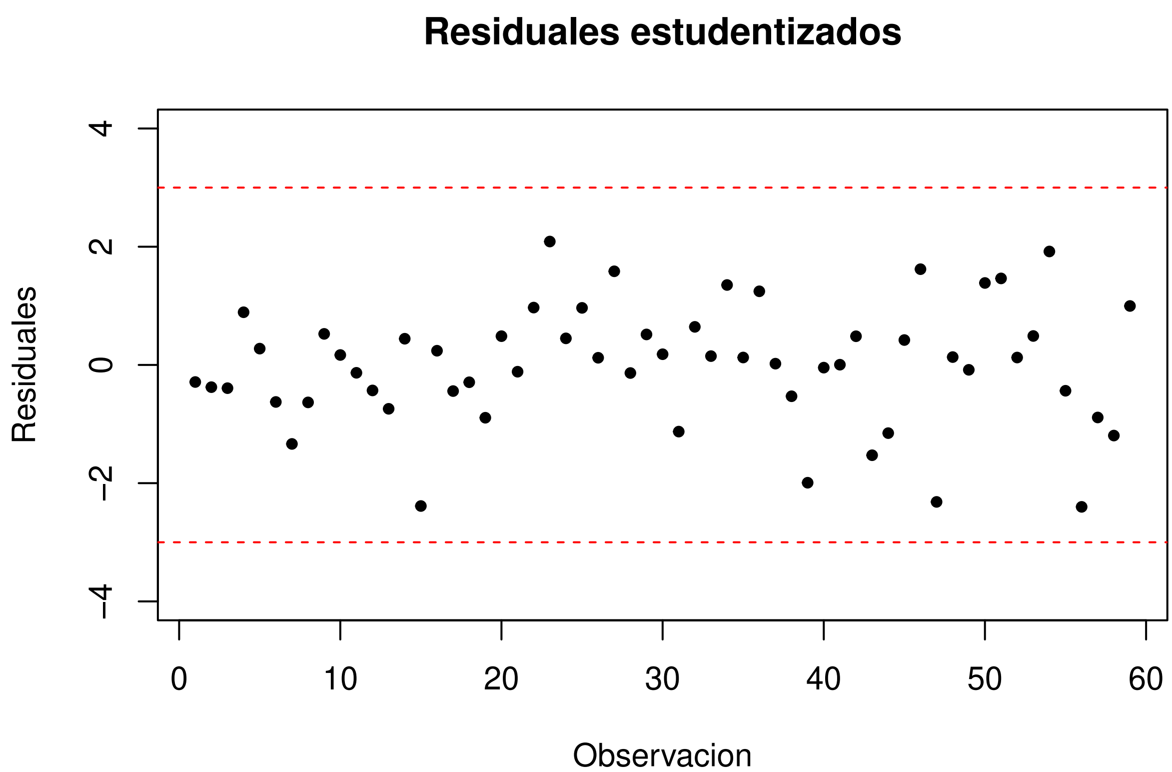


Figura 3: Identificación de datos atípicos

De acuerdo al gráfico anterior, se tiene que no hay datos atípicos en el conjunto de datos, pues ningún residual estudentizado sobrepasa el criterio de $|r_i| > 3$.

3pt

4.2.2. Puntos de balanceo

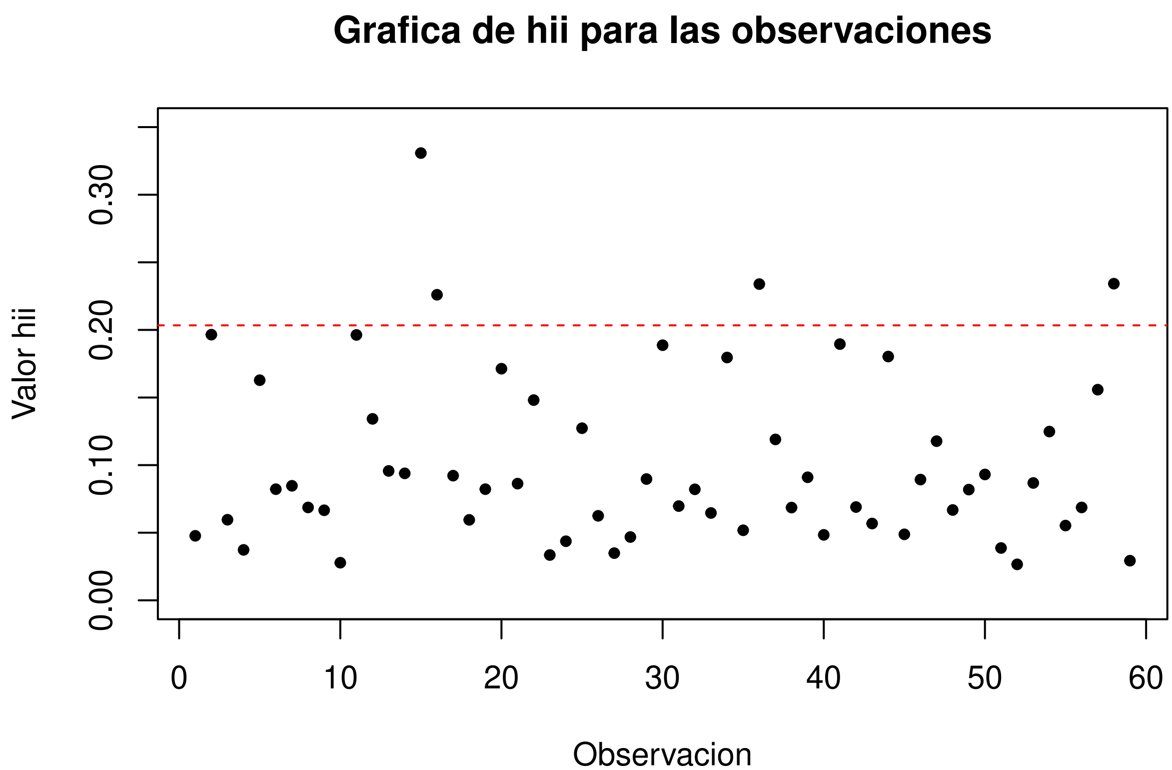


Figura 4: Identificación de puntos de balanceo

Al observar el gráfico anterior, donde la línea punteada roja representa el valor de $h_{ii} = 2\frac{p}{n} = 2\frac{6}{59} = 0.2034$, 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 el siguiente cuadro.

Cuadro 5: Resumen puntos de balanceo

	r_i	Dist. de Cook	h_{ii}	$DFFITS_i$
Observación n°15	-2.3863	0.4692	0.3308	-1.7592
Observación n°16	0.2405	0.0028	0.2260	0.1288
Observación n°36	1.2461	0.0790	0.2339	0.6922
Observación n°58	-1.1946	0.0727	0.2342	-0.6634

3pt

Los puntos de balanceo posiblemente no afecten los coeficientes de regresión estimados $\hat{\beta}_i$, pero sí las estadísticas de resumen como el R^2 y los errores estándar de los coeficientes estimados $se(\hat{\beta}_i)$.

4.2.3. Puntos influenciales

Para identificar si una observación es influencial utilizaremos los siguientes criterios:

- La observación i será influencial si $D_i > 1$.
- La observación i será influencial si $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$.

Donde D_i es la distancia de Cook de la observación i y $DFFITS_i$ es el diagnóstico DFFITS de la observación i.

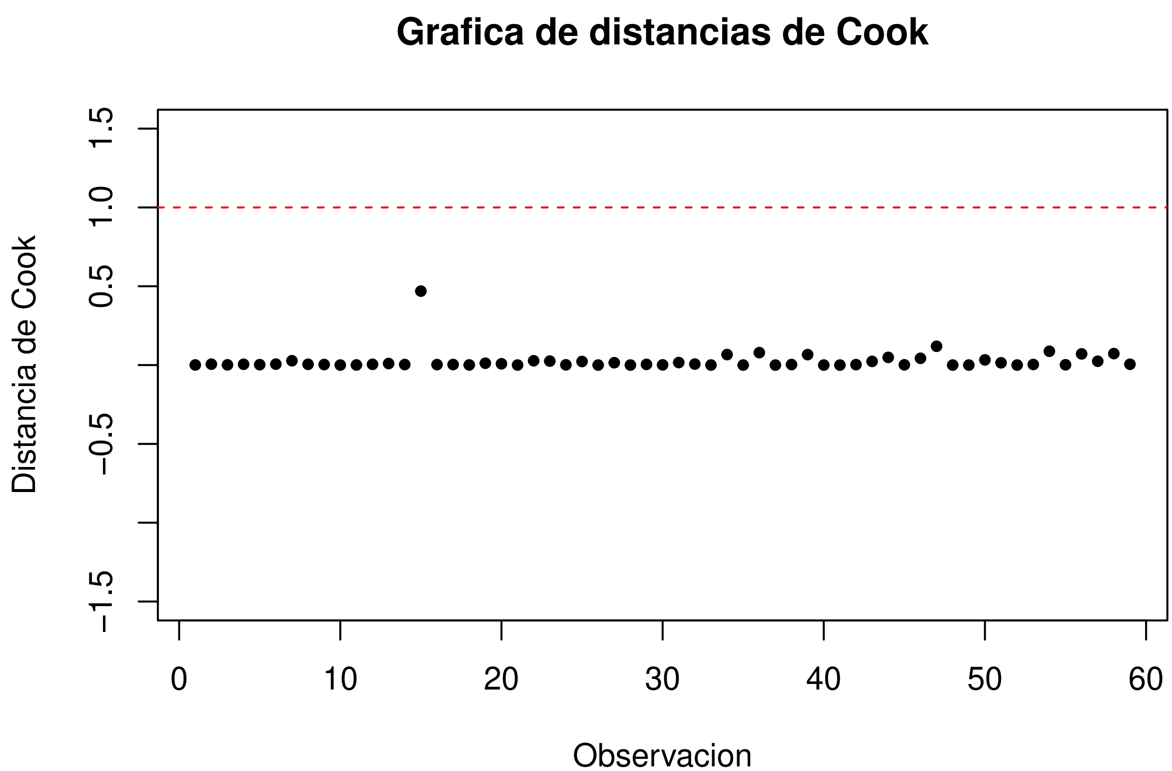


Figura 5: Criterio distancias de Cook para puntos influenciales

De la anterior figura, se puede deducir que, a partir del criterio de la distancia de Cook, no se encuentran observación es influencial, pues $D_i < 1$ para $1 \leqslant i \leqslant 59$.

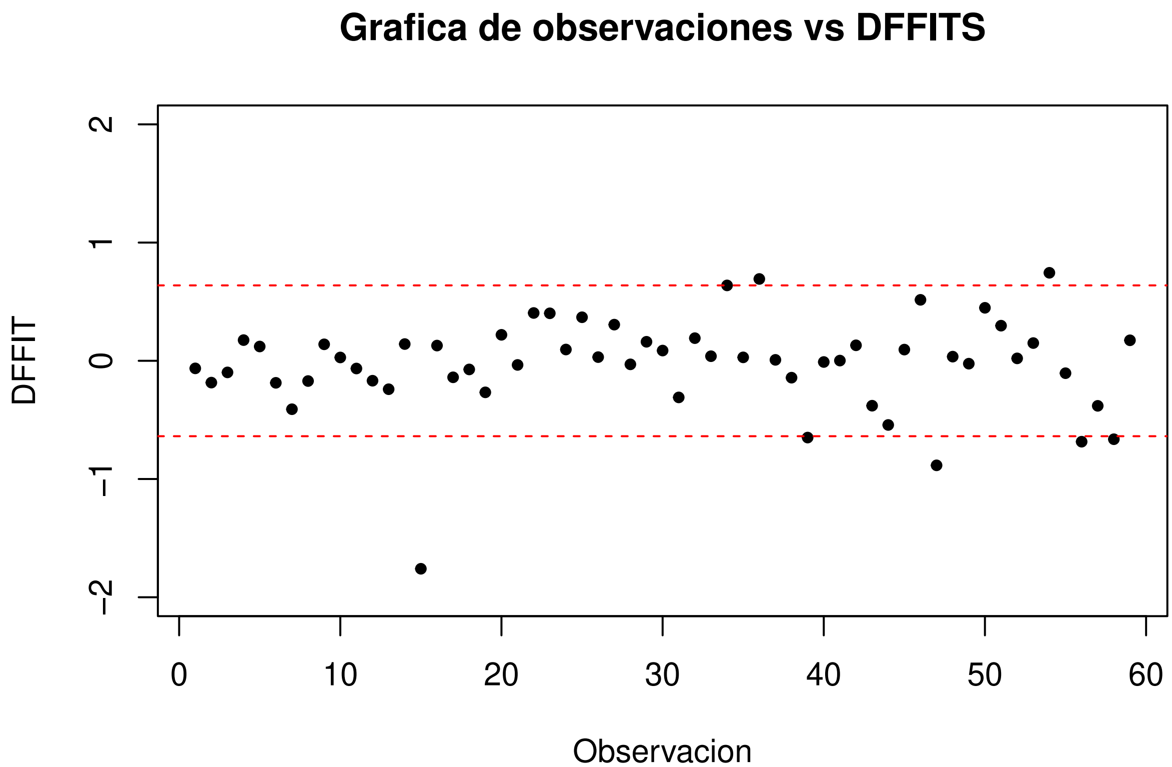


Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Resumen puntos influenciales

	r_i	Dist. de Cook	h_{ii}	$DFFITS_i$
Observación n°15	-2.3863	0.4692	0.3308	-1.7592
Observación n°36	1.2461	0.0790	0.2339	0.6922
Obsservación n°39	-1.9926	0.0663	0.0910	-0.6493
Observación n°47	-2.3170	0.1193	0.1177	-0.8841
Observación n°54	1.9197	0.0876	0.1248	0.7444
Observación n°56	-2.4011	0.0709	0.0687	-0.6843
Observación n°58	-1.1946	0.0727	0.2342	-0.6634

3 pt

De la figura y el cuadro anteriores, se puede observar que, según el criterio del diagnóstico DFFITs, las observaciónes 15, 36, 39, 47, 54, 56 y 58 son influenciales, pues $2\sqrt{\frac{p}{n}} = 0.6378$ y $|DFFITS_i| > 0.6378$; i=15, 36, 39, 47, 54, 56 y 58.

. . . ¿? ¿Qué causan?

En resumen, para el análisis de observaciones extremas se tiene que:

- No hay valores atípicos.
- Las observaciones 15, 16, 36 y 58 son puntos de balanceo.
- Las observaciones 15, 36, 39, 47, 54, 56 y 58 son influenciales.

4.3. Conclusión

3pt

Aunque el supuesto de varianza constante se cumple, al ver que el supuesto de normalidad no se cumple, se concluye que el modelo no es válido. Además, la presencia de las observaciones que son puntos de balanceo y, a su vez, puntos influenciales pueden tener un impacto notable sobre los coeficientes de regresión ajustados, su exclusión del modelo podría causar cambios importantes en la ecuación de regresión ajustada.