93,5

Trabajo 1

Equipo #14

Valeria Hincapié Uribe
Christopher Andrés Obando Rivera
Valeria Vásquez Hernández

Docente

Julieth Verónica Guarín Escudero

Asignatura

Estadística II

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1 18pt

Teniendo en cuenta la base de datos “Equipo 14”, en la cual hay 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

Entonces, se plantea el modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 45$$

1.1. Modelo de regresión 3pt

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Valores de los coeficientes

	Valor del parámetro
β_0	-1.5988
β_1	0.1297
β_2	0.0333
β_3	0.0855
β_4	0.0170
β_5	0.0015

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.5988 + 0.1297X_{1i} + 0.0333X_{2i} + 0.0855X_{3i} + 0.017X_{4i} + 0.0015X_{5i}$$

1.2. Significancia de la regresión 4pt

Para analizar la significancia de la regresión, se plantea el siguiente juego de hipótesis:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a: \text{Algún } \beta_j \neq 0 \text{ ; } 0 \leq \text{j} \leq 5 \end{cases}$$

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

El estadístico de prueba que se usará es:

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

Ahora, se presenta la tabla ANOVA:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas cuadráticas	Grados de libertad	Cuadrados medios	F_0	Valor P
Regresión	58.5472	5	11.709437	15.6379	1.9336e-08
Error	29.2026	39	0.748784		

De la tabla ANOVA anterior, se obtienen los valores del estadístico de prueba $F_0 = 15.6379$ y también su correspondiente $Valor - P = 1.9336e-08$, lo que es aproximadamente 0, también se define un $\alpha = 0.05$.

Como $Valor - P < \alpha$ se rechaza H_0 concluyendo que el modelo de regresión lineal múltiple propuesto es significativo. Por lo tanto, se puede decir, que el riesgo de infección depende significativamente de al menos una de las variables predictoras del modelo.

1.3. Significancia de los parámetros del modelo

6pt

Para esta prueba, se establece el siguiente juego de hipótesis:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases} \quad \text{para} \quad j = 0, 1, \ldots, 6.$$

Ahora 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	-1.5988	1.9173	-0.8339	0.4094
β_1	0.1297	0.0779	1.6648	0.1040
β_2	0.0333	0.0352	0.9467	0.3496
β_3	0.0855	0.0168	5.0966	0.0000
β_4	0.0170	0.0084	2.0267	0.0496
β_5	0.0015	0.0007	2.1384	0.0388

A un nivel de significancia $\alpha = 0.05$, los parámetros individuales β_3, β_4 y β_5 son significativos, porque sus Valores-P son menores a α.

la probabilidad

1.4. Interpretación de los parámetros 2pt

?

$\hat{\beta}_3$: Indica que por cada unidad que se aumente en el número de camas, ~~el~~ promedio del riesgo de infección aumenta en 0.0855 unidades, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_4$: Indica que por cada unidad que se aumente en el censo promedio diario, el promedio del riesgo de infección aumenta en 0.0170 unidades, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_5$: Indica que por cada unidad que se aumente en el número de enfermeras, el promedio del riesgo de infección aumenta en 0.0015 unidades, cuando las demás variables predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2 3pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6672$, lo que significa que aproximadamente el 66.72 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto y aproximadamente el 33.28 % de la variablidad total observada en la respuesta es explicada por el error.

¿Cómo se calcula?

2. Pregunta 2

3,5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

→ 1 es predicción 3

Las covariable con el P-valor más alto en el modelo fueron X_1, X_2, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

1,5pt

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	29.203	X1 X2 X3 X4 X5
Modelo reducido	33.200	X3 X4 X5

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

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

Muy bien

$$
\begin{aligned}
F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_4, \beta_5) - SSE(\beta_0,\ \cdots\ ,\ \beta_5))/2}{MSE(\beta_0,\ \cdots\ ,\ \beta_5)} \overset{H_0}{\sim} f_{2,39} \\
&= \frac{33.200 - 29.203}{33.200/39} \\
&= 4.695271
\end{aligned} \tag{2}
$$

Ahora, comparando el F_0 con $f_{0.95,2,39} = 3.2381$, se puede ver que $F_0 > f_{0.95,1,39}$ por lo que se rechaza la hipótesis H_0, entonces el subconjunto es significativo. Lo anterior indica que no es posible descartar las variables del subconjunto del modelo, ya que al menos una de las variables del subconjunto es distinta de 0.

2pt

3. Preguntа 3

4pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si $\beta_2 = 2\beta_3;\ \beta_4 = 3\beta_5$ por lo tanto se plantea la siguiente prueba de hipótesis:

No es una pregunta

$$
\begin{cases}
\mathrm{H}_0 : \beta_2 = 2\beta_3;\ \ \beta_4 = 3\beta_5 \\
\mathrm{H}_1 : \text{Alguna de las igualdades no se cumple}
\end{cases}
$$

En forma matricial se puede expresar como:

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

Con $\mathbf{L}$ dada por

$$
L = \begin{bmatrix} 0 & 0 & 1 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -3 \end{bmatrix}
$$

2pt

El modelo reducido está dado por:

$$
Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X^*_{2i} + \beta_4 X^*_{4i} + \varepsilon_i;\quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45
$$

0pt

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

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2\left(X_{2i} + \frac{1}{2}X_{3i}\right) + \beta_4\left(X_{4i} + \frac{1}{3}X_{5i}\right) + \varepsilon_i$$

$$X^*_{2i} = X_{2i} + \frac{1}{2}X_{3i};\quad X^*_{4i} = X_{4i} + \frac{1}{3}X_{5i}$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$\begin{aligned} F_0 &= \frac{[SSE(MR) - SSE(MF)]/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,39} \\ &= \frac{[SSE(MR) - 29.2926]/2}{0.748784} \overset{H_0}{\sim} f_{2,39} \end{aligned} \quad (3)$$

2pt

4. Pregunta 4

18pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

4pt

Para probar esta hipótesis, se presenta la siguiente prueba de hipótesis de ~~Shapiro-Wilk~~, con gráficos cuantil-cuantil de residuales:

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

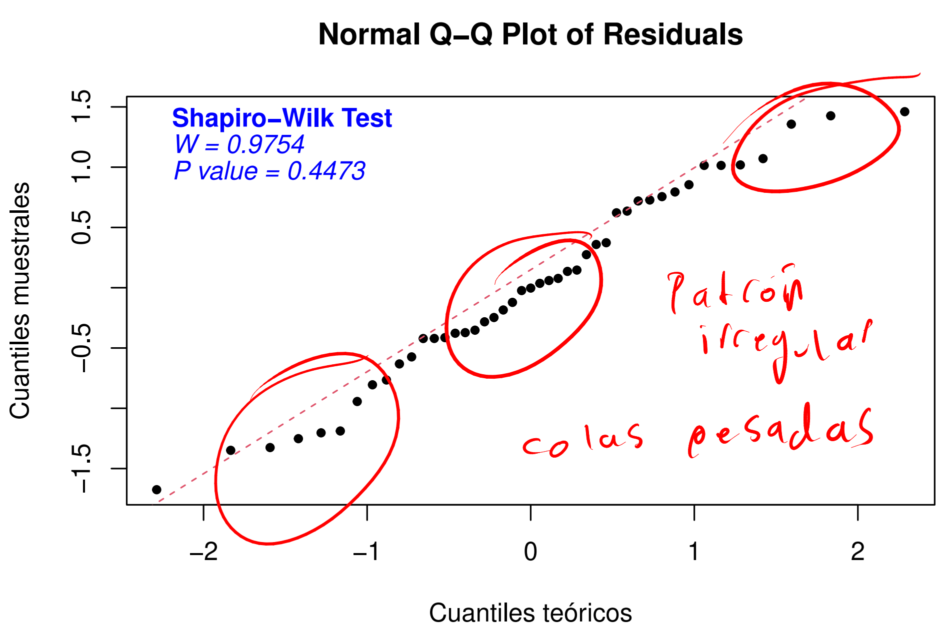


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al considerar un nivel de significancia $\alpha = 0.05$ y el P-valor obtenido $P - valor = 0.4473$, no se rechazaría la hipótesis H_0, puesto que el P-valor es mucho mayor que el nivel de significancia, lo que significa que los datos estarían normalmente distribuidos con respecto a la media μ y la varianza σ^2; sin embargo, la gráfica de comparación normal de cuantil-cuantil sugiere una desviación con respecto a la normal con una ~~cola~~ ~~inferior~~ más pesada, lo que implica que los valores en la ~~cola~~ ~~inferior~~ de la distribución de la muestra son menores que los valores esperados en una distribución teórica; esto podría evidenciar una distribución asimétrica o sesgada. Teniendo en cuenta lo anterior, se puede rechazar el cumplimiento del supuesto de normalidad de los residuales. ✓

4.1.2. Varianza constante 3pt

A continuación, se validará si se cumple con el supuesto de varianza constante.

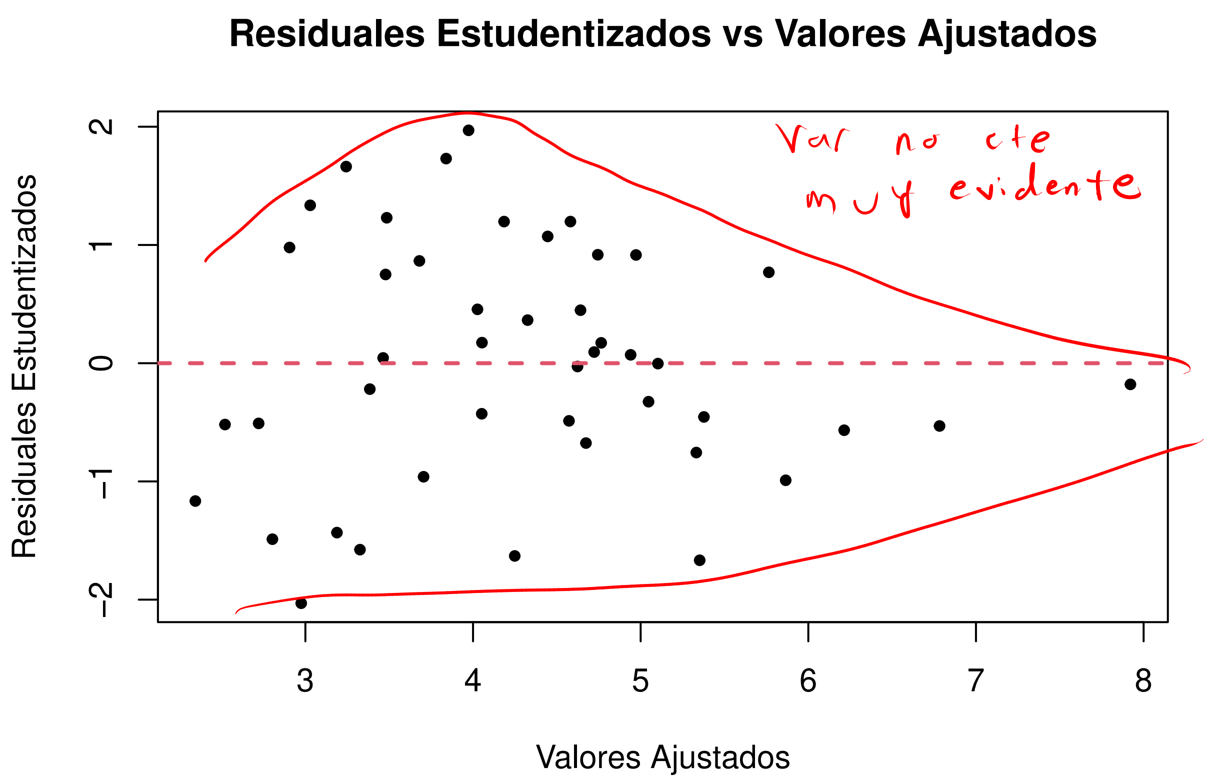


Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados, se observa que los residuales estudentizados se distribuyen en un patrón de dispersión aleatorio alrededor de cero a lo largo de la línea horizontal, lo que sugiere que el modelo ajustado es adecuado; sin embargo, la densidad de la dispersión es mayor en el lado izquierdo de la gráfica, lo que indica que existen problemas con la homogeneidad de la varianza y esta podría no ser constante. El incumplimiento del supuesto de homocedasticidad puede conducir a una sobreestimación o subestimación de la significancia estadística de los coeficientes del modelo y, por ende, puede afectar las inferencias y conclusiones que se puedan hacer. ✓

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3 pt

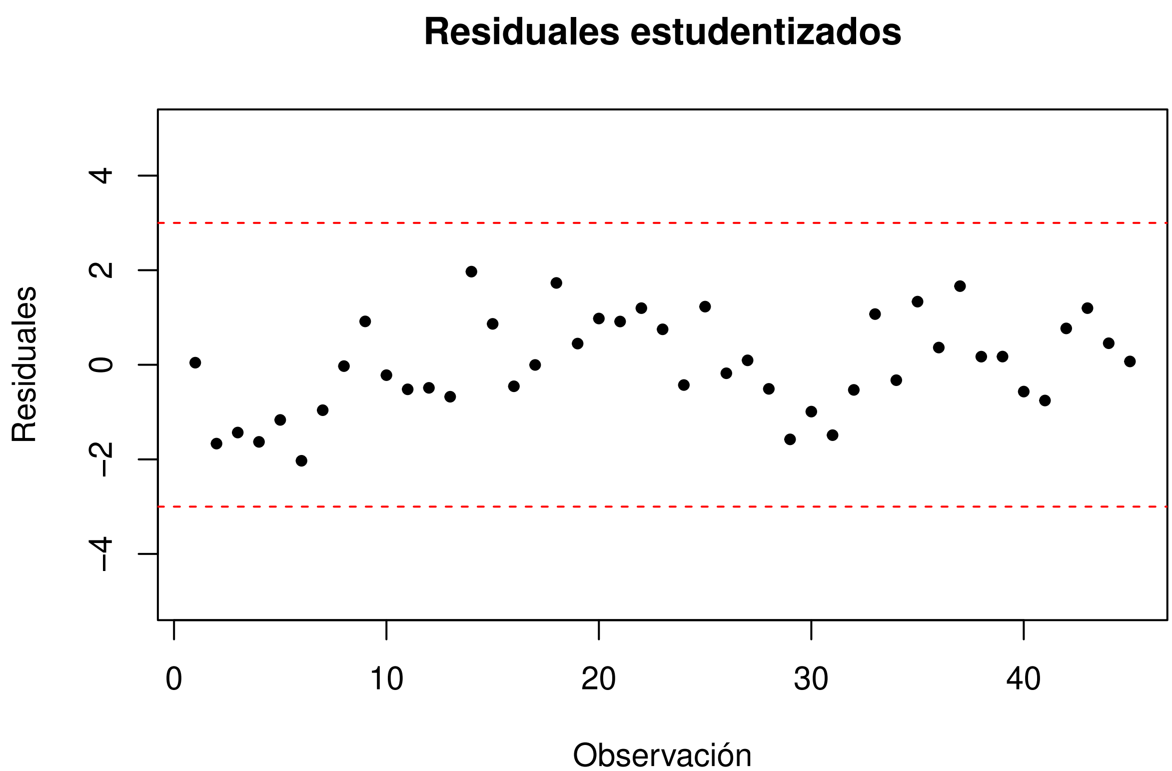


Figura 3: Identificación de datos atípicos

No se observan en la gráfica de residuales estudentizados datos atípicos en el conjunto de datos, pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$. ✓

4.2.2. Puntos de balanceo

3pt

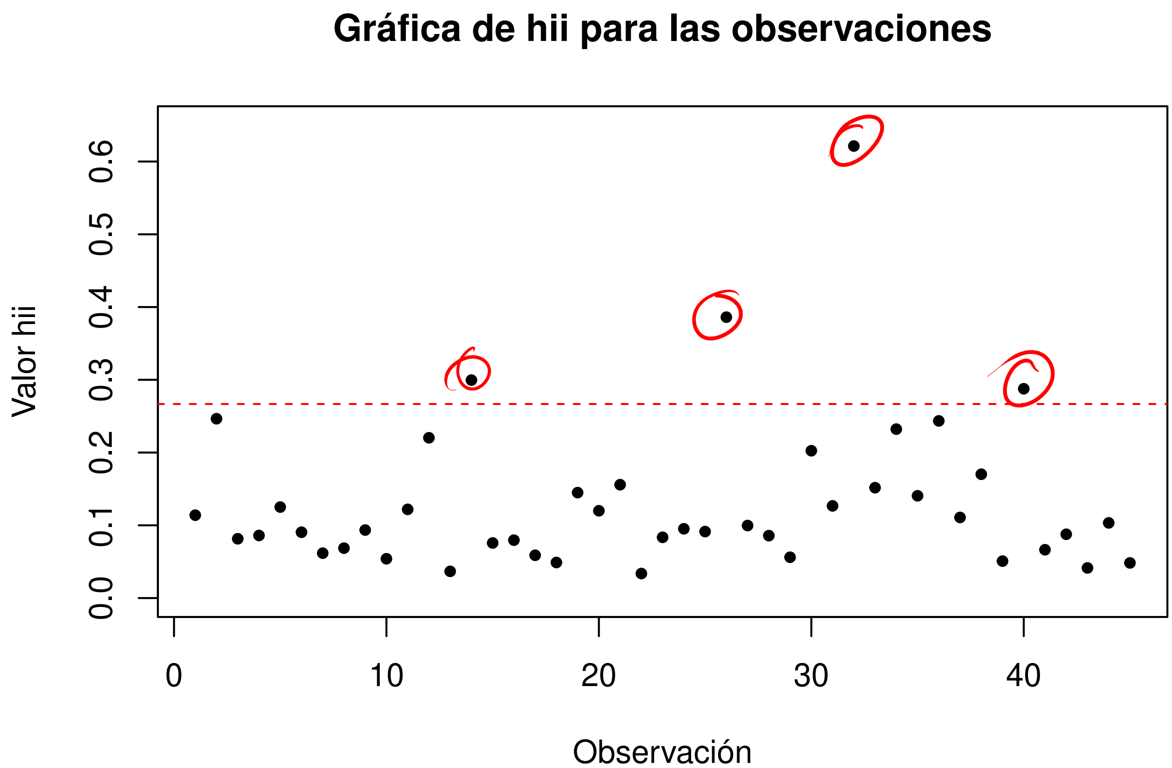


Figura 4: Identificación de puntos de balanceo

Al observar la gráfica de observaciones vs valores h_{ii}, se puede evidenciar que existen 4 datos del conjunto ubicados por encima de la línea punteada roja que representa el valor $h_{ii} = 2\frac{6}{45} = 0.2666667$; estos son puntos de balanceo según el criterio $h_{ii} > 2\frac{p}{n}$. Las observaciones y su respectivo valor h_{ii} se representan en la siguiente tabla:

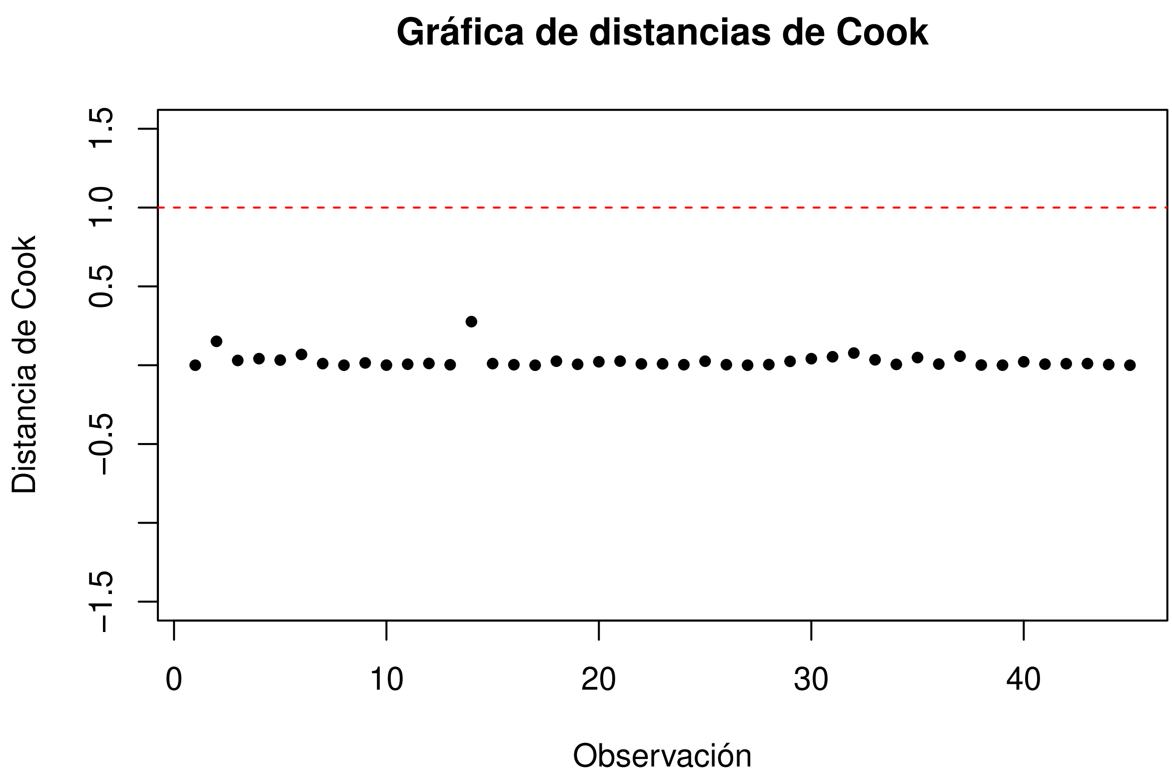
Cuadro 5: Puntos de balanceo

	Valor h_{ii}
14	0.2996
26	0.3861
32	0.6213
40	0.2877

Dos de los puntos de balanceo están muy lejos del valor $h_{ii} = 2\frac{6}{45} = 0.2666667$, lo que indica que estos valores extremos tienen una gran influencia en la estructura de la relación entre variables. La presencia de estos puntos de balanceo podría indicar que no se está cumpliendo el supuesto de varianza constante.

↳ No sólo eso, normalidad, R^2, etc también se ven afectados

4.2.3. Puntos influenciales



1,5 et

Figura 5: Criterio distancias de Cook para puntos influenciales

No se observan puntos infleunciales en la gráfica de distancias de Cook que sean mayores a $D_i > 1$, lo que significa que no hay observaciones en el conjunto de datos con gran influencia en los resultados del modelo. Esto podría indicar que el modelo ajustado es ~~adecuado~~ y que no hay ~~valores atípicos~~ que afecten significativamente al modelo; la ausencia ~~de estos~~ puntos en el gráfico no garantiza que el modelo sea ~~correcto~~, por lo que se hace necesario usar otros análisis para determinar la validez de los supuestos.

? puntos de balanceo

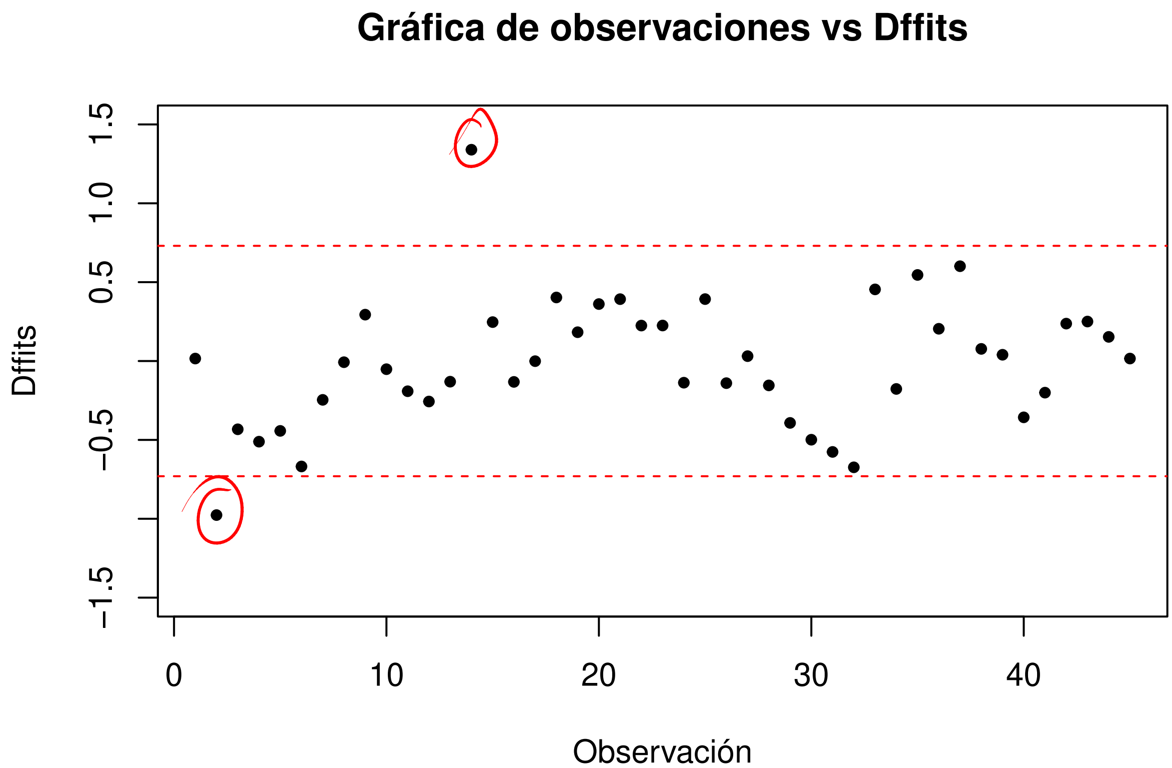


Figura 6: Criterio D_{ffits} para puntos influenciales

Como se puede ver en la gráfica de Observación vs D_{ffits}, las observaciones 2 y 14 son puntos influenciales según el criterio de D_{ffits}, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{6}{45}} = 0.7302967$, es un punto influencial.

Cuadro 6: Puntos influenciales según criterio de D_{ffits}

1 pt

	Valor D_{ffits}
2	-0.9766
14	1.3397

¿Qué causan estos puntos?

4.3. Conclusión

2,5 pt

del modelo

Dado que no se cumplen todos los supestos de ~~validez~~, es posible concluir que el modelo no es ~~adecuado~~, puesto que existen algunas limitaciones importantes que se deben considerar, aunque se puede aceptar que el modelo ajustado es ~~adecuado en algunas de las pruebas~~, por ejemplo, dado que el gráfico de residuales estudentizados en relación con los

válido

valores ajustados sugiere un patrón de dispersión aleatoria alrededor del número cero y no hay puntos significativos en el gráfico de distancias de Cook.

Por un lado, la gráfica de comparación cuantil-cuantil normal sugiere que los datos pueden no estar distribuidos normalmente debido a la cola inferior más pesada, lo que implica una posible distribución sesgada. Además, la presencia de cuatro puntos de balanceo en el gráfico de observaciones vs h_{ii} indica que la varianza podría no ser constante y, por lo tanto, no se cumplirán por completo los supuestos del modelo.

Por otro lado, la presencia de dos puntos de influencia en el gráfico de Observaciones vs D_{ffits} muestra que estos valores extremos tienen una gran influencia en la estructura de la relación entre variables, lo que puede afectar en gran medida las inferencias y los resultados del argumento del modelo. Además, la gráfica de residuales estudentizados vs valores ajustados sugiere problemas de homogeneidad de la varianza, lo que puede llevar a sobreestimar o subestimar la significancia estadística de los coeficientes del modelo y, por lo tanto, afectar las inferencias y conclusiones que se puedan hacer.