3,5

Trabajo 1

Estudiantes

Juan José Montoya Lopera
Leonardo Jesus Mendoza Luna
Samuel Zapata Cuervo

Docente

Veronica Guarin

Asignatura

Estadística II

Sede Medellín
27 de Marzo de 2023

Índice

Índice de figuras

Índice de tablas

1. Pregunta 1

15 pt

Estime un modelo de regresión lineal múltiple que explique el riesgo de infección en términos de las variables restantes (actuando como predictoras) Analice la significancia de la regresión y de los parámetros individuales. Interprete los parámetros estimados. Calcule e interprete el coeficiente de determinación múltiple R^2

Teniendo en cuenta la base de datos asignada a nuestro equipo, la cual es **Equipo20.txt**, las variablespara el modelo son

Variable	Abreviatura	Descripción
Y	RI	Riesgo de infección en porcentaje:Probabilidad promedio estimada de adquirir infección en el hospital
X1	DEST	Duración de la estadía en días: Duración promedio de la estadía de todos los pacientes en el hospital
X2	RC	Rutina de cultivos: Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100 pacientes
X3	NC	Número de camas :Promedio de camas en el hospital durante el periodo del estudio
X4	CPD	Censo promedio diario :Número promedio de pacientes en el hospital por día durante el periodo del estudio
X5	NENF	Número de enfermeras:Promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio

El modelo que se propone es:

$$RI_i = \beta_0 + \beta_1 DEST_i + \beta_2 RC_i + \beta_3 NC_i + \beta_4 CPD_i + \beta_5 NENF_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

$i = 1, 2, \ldots, 50$

1.1. Modelo de regresión

2 pt

Al ajustar el modelo de regresion para el riesgo de infeccion en un hospital, se obtienen los siguientes coeficientes:

Tabla 2: Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	-2.06216
$\hat{\beta}_1$	0.18267
$\hat{\beta}_2$	0.02907
$\hat{\beta}_3$	0.06504
$\hat{\beta}_4$	0.01791
$\hat{\beta}_5$	0.00195

Por lo que el modelo con los respectivos valores de los parametos es:

$$\widehat{RI}_i = -2.06216 + 0.18267 DES_i + 0.02907 RC + 0.06504 NC_i + 0.01791 CPD_i + 0.00195 NENF_i + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Donde las variables se mueven deacuerdo $1 \leq i \leq 50$

1.2. Significancia de la regresión

Se Plantea el siguiente Juego de Hipotesis

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \ para \ j = 1, 2, 3, 4, 5 \end{cases}$$

Se utilizará la siguiente tabla ANOVA para evaluar la significancia de la regresión:

Tabla 3: Tabla anova significancia de la regresión

	Sumas de cuadrados	g.l	Cuadrado medio	F_0	Valor-P
Modelo de regresión	82.8877	5	16.57754	16.4497	3.89801e-09
Error	44.3421	44	1.00777		

Al observar los resultados de la Tabla Anova,La evidencia muestral nos dice que se rechaza la hipotesis nula, por tanto la evidencia muestral nos indica que la regresion es significativa

1.3. Significancia de los parámetros

$$\begin{cases} \mathrm{H}_0 : \beta_j = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \ para \ j = 1, 2, 3, 4, 5 \end{cases}$$

En la siguiente tabla se presentará información sobre los criterios para evaluar la significancia los parámetros de manera individual:

Tabla 4: Resumen de los coeficientes

	Estimación β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-2.0622	1.6415	-1.2563	0.2156
β_1	0.1827	0.0834	2.1898	0.0339
β_2	0.0291	0.0312	0.9331	0.3559
β_3	0.0650	0.0181	3.6008	0.0008
β_4	0.0179	0.0084	2.1283	0.0389
β_5	0.0020	0.0008	2.4771	0.0172

Los resultados de las pruebas: valor del estadístico de prueba y el valor p para la prueba se obtiene en las dos últimas columnas de la tabla de los parámetros estimados.

Con un nivel de significancia$\alpha = 0.05$ se concluye que los parámetros individuales β_1,β_3,β_4,β_5 son significativos cada uno en presencia de los demás parametros Por el contrario los parametros β_0,β_2 individualmente no son significativos en presencia de los demas parametros

1.4. Interpretación de los parámetros

- $\hat{\beta}_0 = -2.06216$: El parámetro $\hat{\beta}_0$ no es interpretable porque no es significativo, ademas ninguna variable predictora lo contiene.
- $\hat{\beta}_1 = 0.18267$: Si la Duración de la estancia en el hospital (en días) aumenta en un día, manteniendo constantes las demás variables predictoras, se espera que el promedio del porcentaje del Riesgo de infección aumente en 0.18267 % unidades de medida.
- $\hat{\beta}_2 = 0.02907$: El parámetro β_2 no es interpretable porque no es significativo.
- $\hat{\beta}_3 = 0.06504$: si el número promedio de camas en el hospital durante el periodo de estudio aumenta en una unidad, manteniendo constantes las demás variables predictoras, se espera que el promedio del Riesgo de infección aumente en un 0.02907 %
- $\hat{\beta}_4 = 0.01791$: si el número censo del promedio Diario del paciente en el hospital durante el periodo de estudio aumenta en una unidad, manteniendo constantes las demás variables predictoras, se espera que el promedio del Riesgo de infección aumente en un 0.06504 %
- $\hat{\beta}_5 = 0.00195$: Por cada unidad que aumenta el numero promedio de enfermeras en el hospital, el riesgo de infeccion aumenta promedio en 0.00195 % cuando las demás variables predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un R^2 de 0.5899 lo cual significa que aproximadamente el $\$ 65.15 \%\$$ de la variabilidad total en el porcentaje de Riesgo de infeccion es explicado por el modelo RLM

1.6. Comentarios

En el modelo, se observa que las variables que tienen una contribución significativa en la regresión son Duración de la estadía en el hospital (DE),Censo promedio diario de pacientes en el hospital(CDP), Número de camas (NC), y Número de enfermeras (NEnf) y la significancia de los parametros

2. Pregunta 2

Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más grandes del punto anterior. Según el resultado de la prueba es posible descartar del modelo las variables del subconjunto? Explique su respuesta.

2.1. Planteamiento prueba de hipotesis y modelo reducido

Los parametros cuyos valores P fueron los más altos corresponden a β_2 con VP=0.3559, β_1 con VP= 0.0339, β_4 con VP= 0.0389. Por tanto, se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_4 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \; para \; j = 1, 2, 4 \end{cases}$$

El modelo completo es el definido en la sección 1.1, y el modelo reducido es:

$$\text{MR: } Rinf_i = \beta_0 + \beta_3 NC_i + \beta_5 NENF_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Se presenta la siguiente tabla con el resumen de todas las regresiones para plantear el estadístico de prueba:

Tabla 5: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	44.342	X1 X2 X3 X4 X5
Modelo reducido	67.973	X3 X5

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSR(\beta_0,\beta_3,\beta_5|\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,44}$$

$$F_0 = \frac{(SSE(\beta_0,\beta_3,\beta_5) - SSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5))/2}{MSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)} \overset{H_0}{\sim} f_{2,44}$$

$$= \frac{(67.973 - 44.342)/2}{44.342/44} = 11.724$$

Ahora, comparando a un nivel de significancia $\alpha = 0.05$, F_0 con $f_{0.05,2,44} = 3.209278$. Con valor P=8.2916164×10^{-5}

Se observa que el valor de la estadística de prueba $F_0 = 11.724$ es mayor que el valor crítico $F_{\alpha=0.05,2,44} = 3.209278$ de la distribución F con un nivel de significancia del 5 %, y el valor p obtenido es pequeño. Por lo tanto, la evidencia sugiere que se debe rechazar la hipótesis nula H_0. Esto implica que, con base a la evidencia muestral, hay al menos un parámetro que es significativo en el subconjunto de datos considerado.

3. Pregunta 3

Plantee una pregunta donde su solución implique el uso exclusivo de una prueba de hipótesis lineal general de la forma H0 : L$\beta = 0$ (solo se puede usar este procedimiento y no SSextra). Especifique claramente la matriz L, el modelo reducido y la expresión para el estadístico de prueba (no hay que calcularlo).

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \text{H}_0 : \beta_1 = \beta_3, \beta_2 = \beta_5 \\ \text{H}_a : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

Reescribendo matricialmente:

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

Donde $\mathbf{L}$ está dada por:

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

Donde el modelo reducido está dado por:

$$Rinf = \beta_0 + \beta_1(DES_i + NC_i) + \beta_2(RC_i + NENF_i) + \beta_3(NC_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$$

3.2. Estadístico de prueba

14,5 pt

El estadístico de prueba F_0 está dado por:

1,5 pt

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,54}$$

Obteniendo esto podemos definir la region de rechazo de la hipotesis nula como $F_0 > F_{0.05,2,44} = 3.209278$ y con valor p:P($F_{2,44} > |F_0|$)

Reemplazar lo que conocen

4. Pregunta 4

Realice una validación de los supuestos en los errores y examine si hay valores atípicos, de balanceo e influenciales. Qué puede decir acerca de la validez de éste modelo?. Argumente su respuesta.

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

2 pt

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis (~~shapiro wilk~~)

$$\begin{cases} \text{H}_0: & \varepsilon_i \sim N(\mu, \sigma^2) \\ \text{H}_a: & \varepsilon_i \nsim N(\mu, \sigma^2) \end{cases}$$

→ con normalidad no van a mirar media constante ni
→ tampoco Var constante σ^2

acompañado de un grafico cuantil-cuantil:

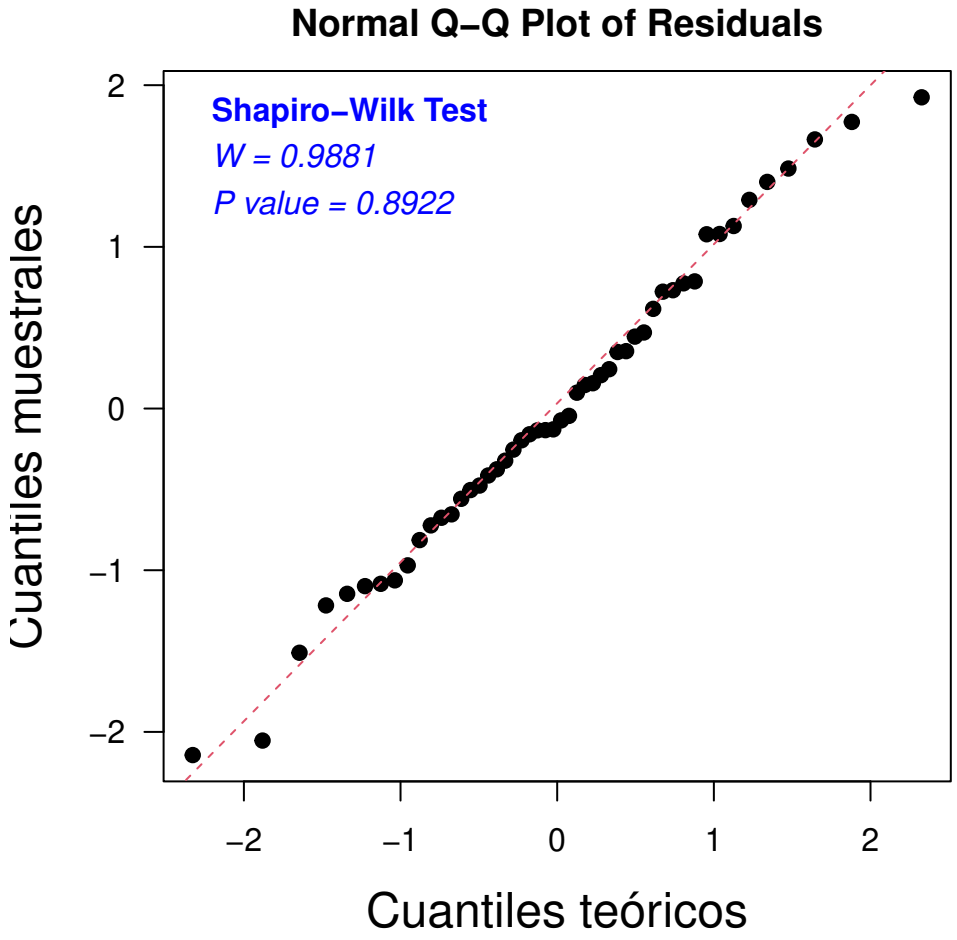

Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Si el valor P es grande, esto sugiere que no hay suficiente evidencia para rechazar la hipótesis nula H_0. En consecuencia, se puede concluir que el modelo es consistente con el supuesto de normalidad de los residuos.

No hacen análisis gráfico, que es más importante

2,5 pt

4.1.2. Media 0 y Varianza constante

En esta prueba se quiere probar

$$H_0 : \mathbf{V}[\varepsilon_i] = \sigma^2 \qquad \textbf{vs} \qquad \mathbf{V}[\varepsilon_i] \neq \sigma^2$$

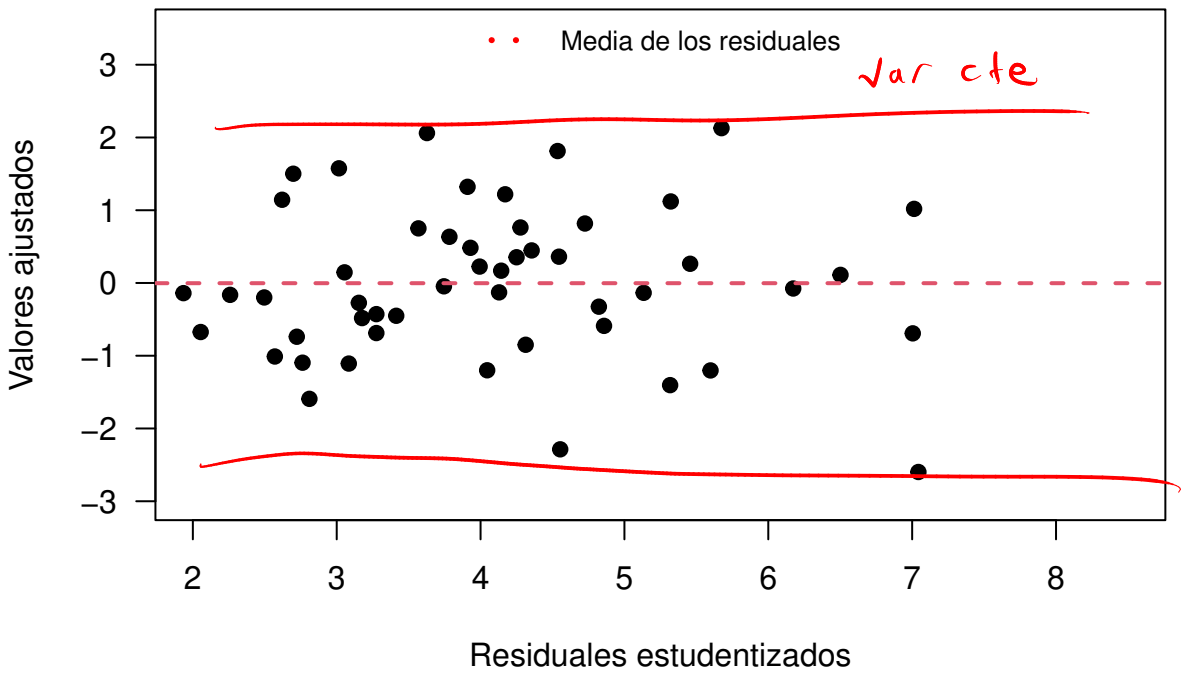

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Se puede observar que la línea punteada roja, que representa la media de los errores, se encuentra en cero o muy cerca de cero. A partir de esta evidencia, se puede concluir que los errores tienen una media cercana a cero. X residuales estudentizados siempre van a tener media 0.

Además, al examinar los residuos, no se observa ningún patrón discernible. Por lo tanto, se puede concluir que la varianza de los errores es constante en todo el rango de los valores observados. → Análisis poco profundo

En resumen, la media cercana a cero de los errores y la constancia de la varianza de los residuos indican que el modelo se ajusta adecuadamente a los datos y cumple con los supuestos básicos de la regresión lineal.

4.2. Observaciones extremas

4.2.1. Datos atípicos

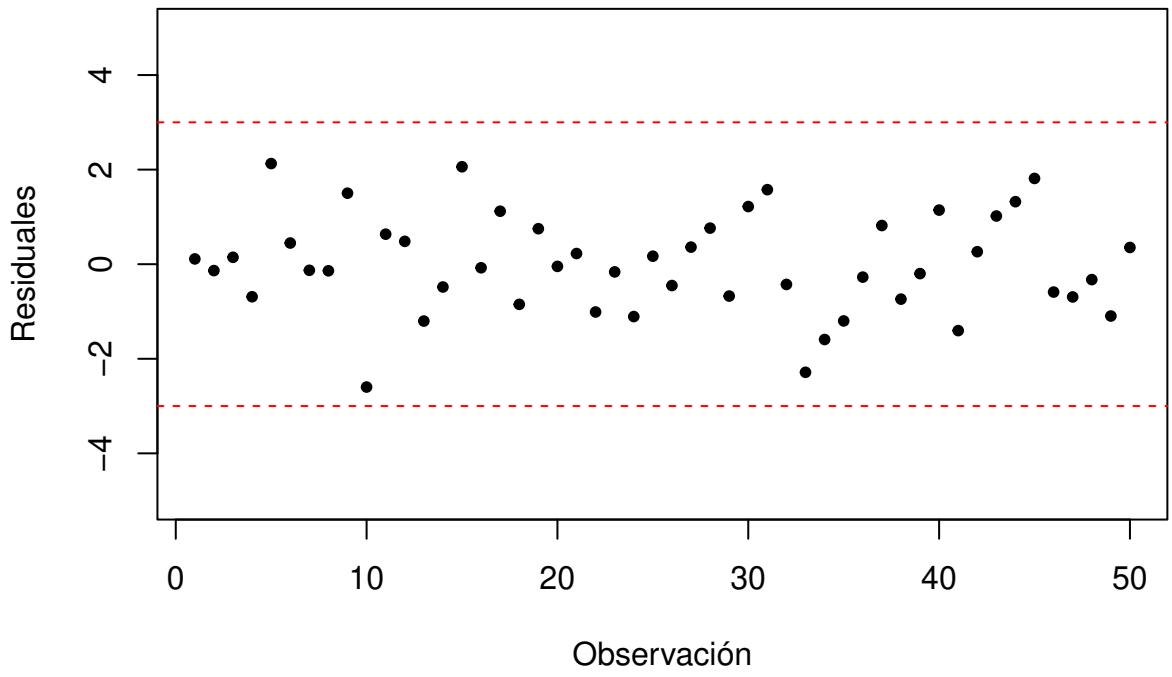

Figura 3: Identificación de datos atípicos

De acuerdo a la figura presentada, se puede observar que no hay valores de residuos estandarizados $|r_i| > 3$, lo cual sugiere que no hay datos atípicos en el modelo bajo el criterio de los residuos estandarizados.

Es importante tener en cuenta que el criterio de $|r_i| > 3$ es solo una regla general, y existen otros criterios y técnicas que pueden ser utilizados para detectar valores atípicos en los datos. Por lo tanto, es recomendable realizar un análisis más detallado y exhaustivo antes de descartar completamente la presencia de valores atípicos en el modelo.

4.2.2. Puntos de balanceo

2 pt

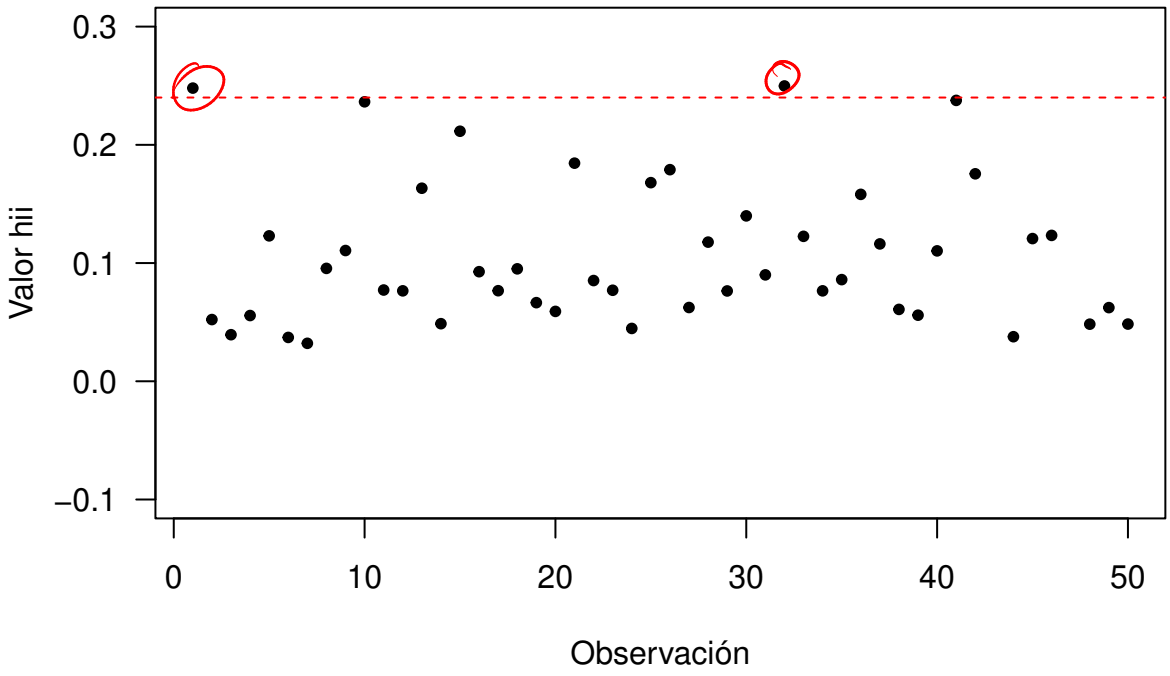

Figura 4: Identificación de puntos de balanceo

¿cuánto da?

Basándonos en el criterio hii > 2p/n y en el gráfico de la diagonal principal de la matriz Hat, podemos concluir que el modelo tiene 5 puntos de balanceo. Estos puntos de balanceo pueden ejercer una influencia importante en el ajuste del modelo y en sus propiedades. Por lo tanto, es importante tener en cuenta los puntos de balanceo y realizar un análisis detallado de su impacto en el modelo antes de sacar conclusiones definitivas.

Muy bien por decir qué ocasionan los puntos de balanceo

Tabla 6: Tabla de puntos de Balanceo

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
1	0.1114	0.0007	0.2480	0.0640
32	-0.4281	0.0104	0.2498	-0.2470
43	1.0187	0.1192	0.4081	0.8459
47	-0.6924	0.0747	0.4801	-0.6654

Notese que los datos de balanceo que deben ser investigados son los datos 1,32,43,47, ya que estos son mayores a $\frac{2p}{n}$

Dicen que hay 5, muestran 2 en la gráfica y en la tabla presentan 4. Tienen 4! Debieron ampliar límites de la gráfica.

4.2.3. Puntos influenciales

Bajo el criterio de Cook, se hace la siguiente gráfica:

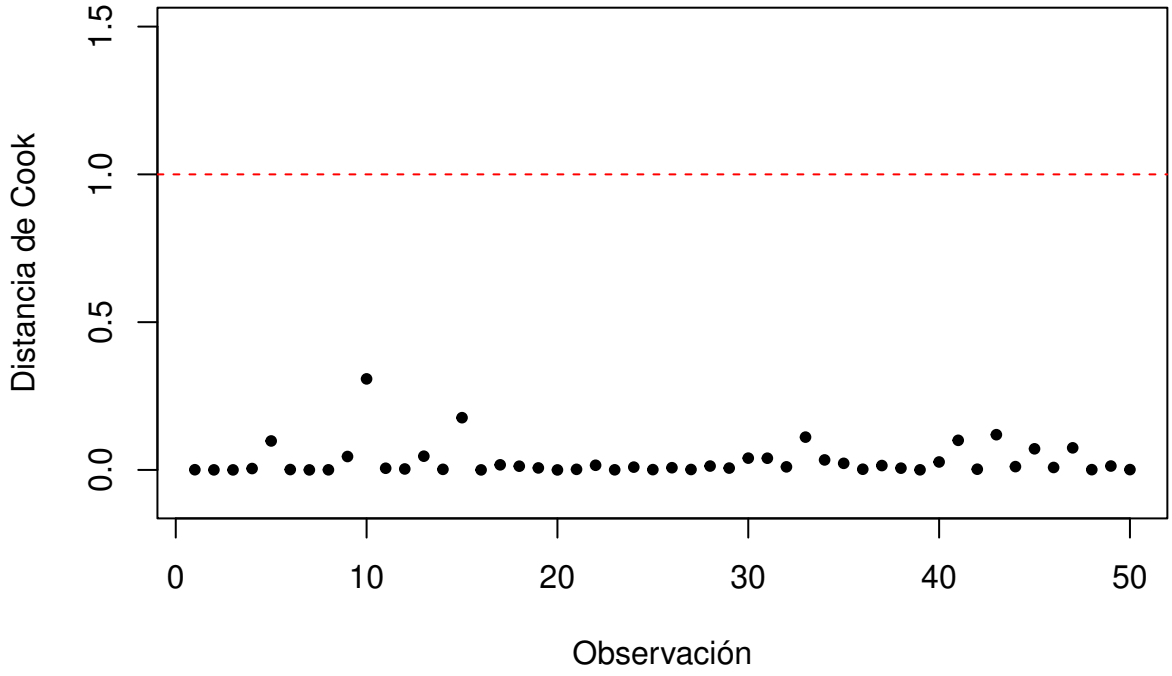

Figura 5: Criterio distancias de Cook para puntos influenciales

Bajo el criterio de cook, se obtuvo la anterior gráfica. A partir de la gráfica podemos concluir que no existen puntos influenciales bajo este criterio

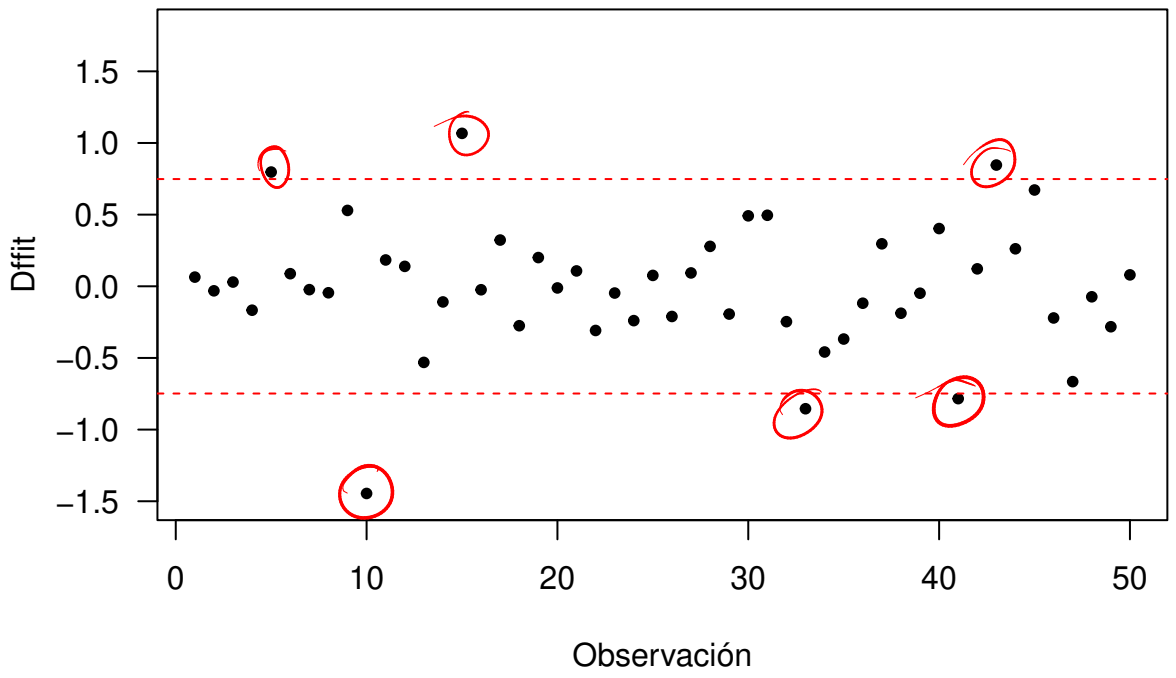

Figura 6: Criterio Dffits para puntos influenciales

Tabla 7: Tabla del criterio DFFITS para encontrar puntos influenciales

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
5	2.1283	0.0980	0.1230	0.7972
10	-2.5981	0.3081	0.2364	-1.4458
15	2.0607	0.1768	0.2115	1.0672
33	-2.2863	0.1110	0.1226	-0.8545
41	-1.4044	0.1002	0.2376	-0.7839
43	1.0187	0.1192	0.4081	0.8459

Utilizando el criterio de Dffits, se ha obtenido la gráfica anterior, a partir de la cual se puede concluir que existen varios valores influyentes en el modelo, específicamente en las observaciones 5, 10, 15, 33, 41 y 43. Estos valores influyentes deben ser investigados más detalladamente para determinar su impacto en el modelo de regresión y si deben ser excluidos o corregidos. Por lo tanto, se debe realizar un análisis adicional para evaluar la influencia de estos puntos en el modelo.

4.3. Conclusiones

El modelo de regresión parece cumplir con los supuestos de normalidad y homocedasticidad, lo cual es una buena señal. Sin embargo, se han identificado algunos puntos influyentes utilizando los criterios de Cook

y Dffits, los cuales parecen tener un impacto significativo en los resultados del modelo. Por lo tanto, es importante tomar en cuenta estos puntos y evaluar su posible eliminación o ajuste en futuros análisis para mejorar la calidad de las predicciones del modelo. Es necesario realizar un análisis adicional para determinar la mejor estrategia a seguir para manejar estos puntos influyentes y mejorar la validez y precisión del modelo.

No hablan sobre validez