Trabajo 1

Estudiantes

Isabela García Salazar
Miguel Fernando Olave Riascos
Albeiro Jose Burbano Tobar

Docente

Veronica Guarin

Asignatura

Estadística II

Sede Medellín
30 de Marzo de 2023

Índice

1. Pregunta 1 **2**

1.1. Modelo de regresión . . . 2

1.2. Significancia de la regresión . . . 2

1.3. Significancia de los parámetros . . . 3

1.4. Interpretación de los parámetros . . . 3

1.5. Coeficiente de determinación múltiple R^2 . . . 4

1.6. Comentarios . . . 4

2. Pregunta 2 **4**

2.1. Planteamiento prueba de hipotesis y modelo reducido . . . 4

2.2. Estadístico de prueba y conclusiones . . . 5

3. Pregunta 3 **5**

3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . 5

3.2. Estadístico de prueba . . . 5

4. Pregunta 4 **6**

4.1. Supuestos del modelo . . . 6

4.1.1. Normalidad de los residuales . . . 6

4.1.2. Media 0 y Varianza constante . . . 7

4.2. Observaciones extremas . . . 8

4.2.1. Datos atípicos . . . 8

4.2.2. Puntos de balanceo . . . 9

4.2.3. Puntos influenciales . . . 9

4.3. Conclusiones . . . 11

Índice de figuras

1. Gráfico cuantil-cuantil y normalidad de los residuales . . . 6

2. Gráfico residuales estudentizados vs valores ajustados . . . 7

3. Identificación de datos atípicos . . . 8

4. Identificación de puntos de balanceo . . . 9

5. Criterio distancias de Cook para puntos influenciales . . . 10

6. Criterio Dffits para puntos influenciales . . . 11

Índice de tablas

1. Tabla de valores de los coeficientes estimados 2
2. Tabla anova significancia de la regresión 3
3. Resumen de los coeficientes 3
4. Resumen de todas las regresiones 4
5. Tabla de puntos de Balanceo 9
6. Tabla del criterio DFFITS para encontrar puntos influenciales 11

1. Pregunta 1

18 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 **Equipo05.txt**, las variablespara el modelo son

Y RI Riesgo de infección en porcentaje: Probabilidad promedio estimada de adquirir infección en el hospita. l X1 DEH 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 NO.ENF Número de enfermeras: Promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio.

El modelo que se propone es:

$$RI_i = \beta_0 + \beta_1 DEH_i + \beta_2 RC_i + \beta_3 NC_i + \beta_4 CPD_i + \beta_5 NO.ENF_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

$i = 1, 2, \ldots, 55$

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 1: Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	-1.40236
$\hat{\beta}_1$	0.24687
$\hat{\beta}_2$	0.01490
$\hat{\beta}_3$	0.05429
$\hat{\beta}_4$	0.01626
$\hat{\beta}_5$	0.00164

No va el ec. ajustada

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

$$\widehat{RI}_i = -1.40236 + 0.24687 DES_i + 0.0149 RC + 0.05429 NC_i + 0.01626 CPD_i + 0.00164 NENF_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Donde las variables se mueven deacuerdo $1 \leq i \leq 55$

1.2. Significancia de la regresión

4 pt

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 2: Tabla anova significancia de la regresión

	Sumas de cuadrados	g.l	Cuadrado medio	F_0	Valor-P
Modelo de regresión	58.8204	5	11.76407	10.8888	4.43254e-07
Error	52.9389	49	1.08039		

Tras examinar los resultados de la Tabla ANOVA, se puede concluir que la hipótesis nula ha sido rechazada en base a la evidencia muestral. Por lo tanto, se puede inferir que la regresión es significativa según la evidencia muestral.

1.3. Significancia de los parámetros

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

A continuación, se presentará información acerca de los criterios para evaluar la significancia de los parámetros de forma individual.

Tabla 3: Resumen de los coeficientes

	Estimación β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-1.4024	1.9536	-0.7178	0.4763
β_1	0.2469	0.1112	2.2192	0.0311
β_2	0.0149	0.0318	0.4690	0.6411
β_3	0.0543	0.0169	3.2159	0.0023
β_4	0.0163	0.0086	1.8863	0.0652
β_5	0.0016	0.0009	1.8993	0.0634

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_0,β_2,β_4 y β_5 no son significativos cada uno en presencia de los demás parametros Por el contrario los parametros β_1,β_3 individualmente son significativos en presencia de los demas parametros

1.4. Interpretación de los parámetros

- $\hat{\beta}_0 = -1.40236$: El parámetro $\hat{\beta}_0$ carece de interpretación, ya que no es significativo y además no está contenido en ninguna variable predictora.
- $\hat{\beta}_1 = 0.24687$: Si se aumenta la duración de la estancia en el hospital en un día, manteniendo constantes las demás variables predictoras, se espera que el promedio del porcentaje de riesgo de infección aumente en 0.24687 unidades de medida.
- $\hat{\beta}_2 = 0.0149$:El parámetro $\hat{\beta}_2$ no tiene una interpretación significativa, ya que no es significativo en los análisis realizados.

- $\hat{\beta}_3$= 0.05429:Si se incrementa en una unidad el número promedio de camas en el hospital durante el periodo de estudio, manteniendo constantes las demás variables predictoras, se espera que el promedio del riesgo de infección aumente en un 0.05429.
- $\hat{\beta}_4$= 0.01626: El parámetro $\hat{\beta}_4$ no tiene una interpretación significativa, ya que no es significativo en los análisis realizados.
- $\hat{\beta}_5$= 0.00164: El parámetro $\hat{\beta}_5$ no tiene una interpretación significativa, ya que no es significativo en los análisis realizados.

1.5. Coeficiente de determinación múltiple R^2

El valor de R^2 del modelo es de 0.5263, lo que indica que alrededor del 52.63 % de la variabilidad total en el porcentaje de riesgo de infección es explicada 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) y Número de camas (NC) y por parte su de R^2 nos indica un valor bajo para del porcentaje que es explicado por la regresion por lo que podemos decir que esta regresion se puede mejorar.

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=0.6411, β_4 con VP= 0.0652, β_5 con VP= 0.0634. Por tanto, se plantea la siguiente prueba de hipótesis:

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

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

$$\text{MR: } Rinf_i = \beta_0 + \beta_1 DEH_i + \beta_3 NC_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

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

Tabla 4: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	52.939	X1 X2 X3 X4 X5
Modelo reducido	60.421	X1 X3

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSR(\beta_0,\beta_1,\beta_3|\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,49}$$

$$F_0 = \frac{(SSE(\beta_0,\beta_3,\beta_5) - SSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5))/2}{MSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)} \overset{H_0}{\sim} f_{2,55}$$

$$= \frac{(60.421 - 52.939)/2}{52.939/49} = 3.462646$$

Al comparar el valor de la estadística de prueba $F_0 = 3.462646$ con el valor crítico $F_{\alpha=0.05,2,49} = 3.186582$ de la distribución F para un nivel de significancia del 5 %, y considerando que el valor p obtenido es pequeño, se sugiere que se debe rechazar la hipótesis nula H_0. Por lo tanto, en base a la evidencia muestral, se concluye que al menos un parámetro 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:

$$\begin{cases} \mathrm{H}_0 : \beta_3 = \beta_4, \beta_3 = \beta_5 \\ \mathrm{H}_a : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

Reescribendo matricialmente:

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

Donde **L** está dada por:

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

Donde el modelo reducido está dado por:

$$Rinf = \beta_0 + \beta_1 DES_i + \beta_2 RC_i + \beta_3(NC_i + CPD_i) + \beta_5(NO.ENF_i + NC_i)\varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

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

Obteniendo esto podemos definir la region de rechazo de la hipotesis nula como $F_0 > F_{0.05,2,49} = 3.186582$ y con valor p:P($F_{2,49} > |F_0|$)

4. Pregunta 4 13 pt

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 1 pt

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

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

→ No saben si la var es constante σ^2

Grafico cuantil-cuantil:

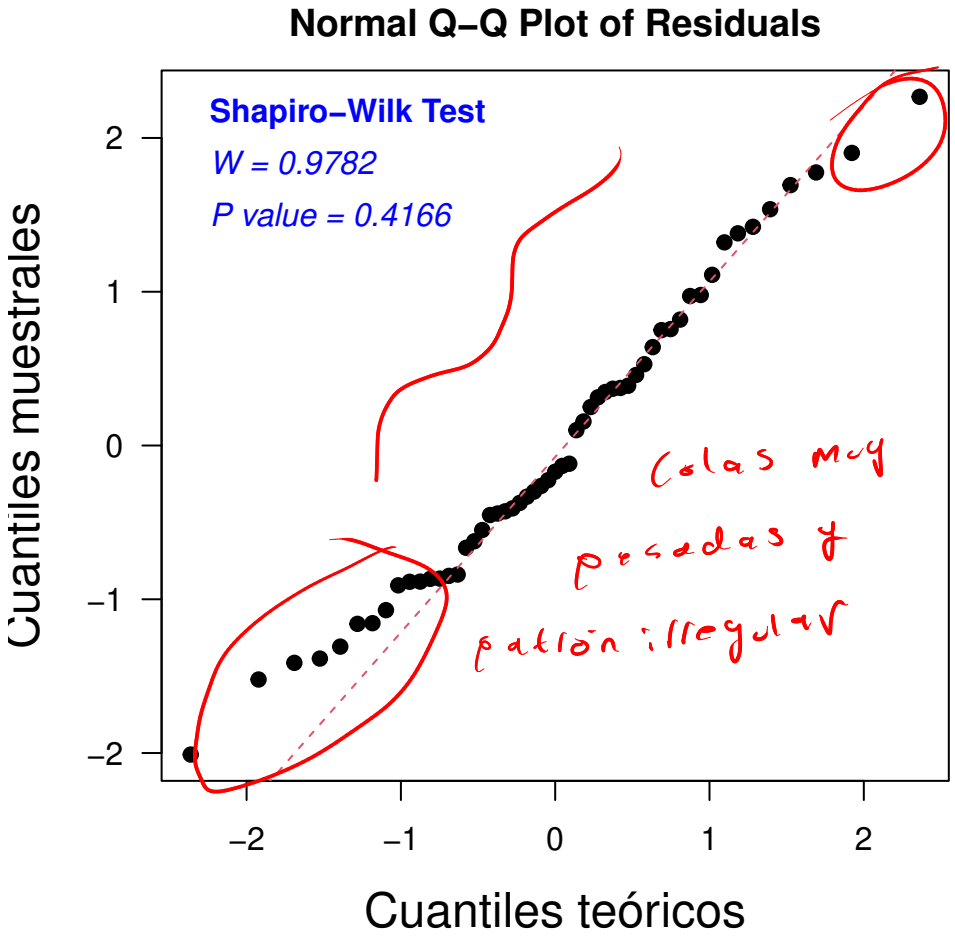

Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

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.

Les faltó analizar lo más importante: el gráfico, de lo que se nota que no distribuye normal y por tanto modelo no es válido

4.1.2. Media 0 y Varianza constante

En esta prueba se quiere probar

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

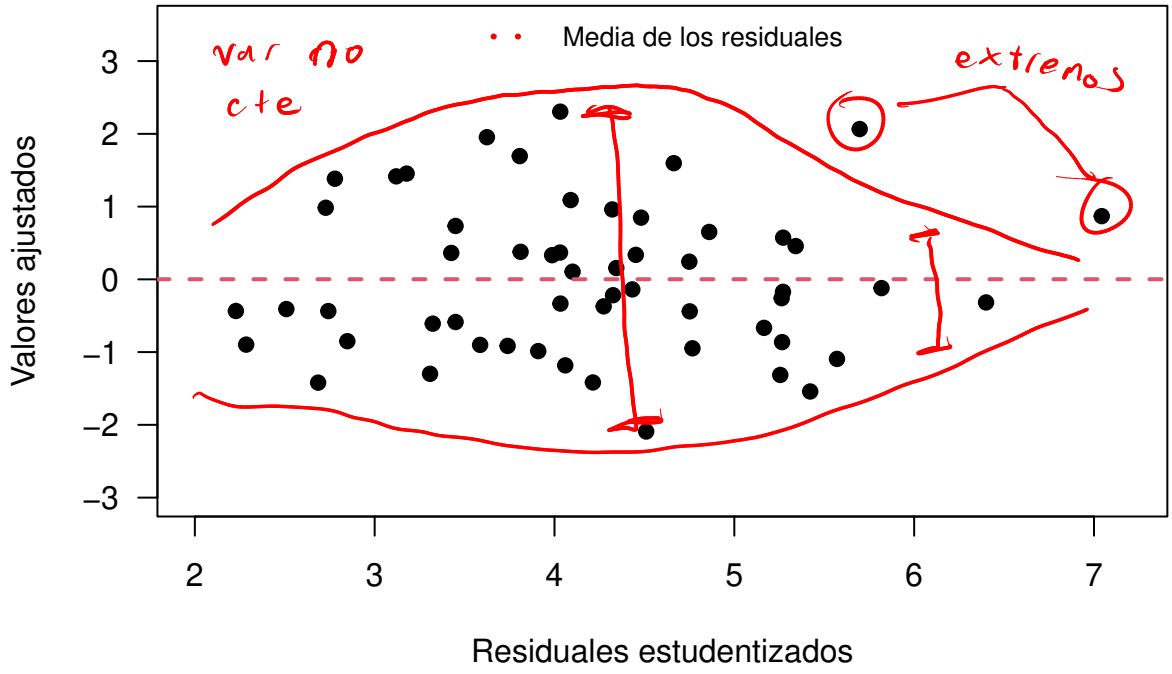

Figura 2: Gráfico residuales estudentizados vs valores ajustados

SSe puede apreciar en el gráfico que la línea punteada roja, la cual representa la media de los residuos, se sitúa cerca o en cero. Esto sugiere que los residuos tienen una media aproximadamente igual a cero. Además, al examinar la distribución de los residuos, no se puede distinguir ningún patrón claro, lo que indica que la varianza de los errores es constante a través de todo el rango de los valores observados. En consecuencia, se puede concluir que el modelo cumple con el supuesto de que los errores tienen una media cero y una varianza constante (homocedasticidad).

4.2. Observaciones extremas

4.2.1. Datos atípicos

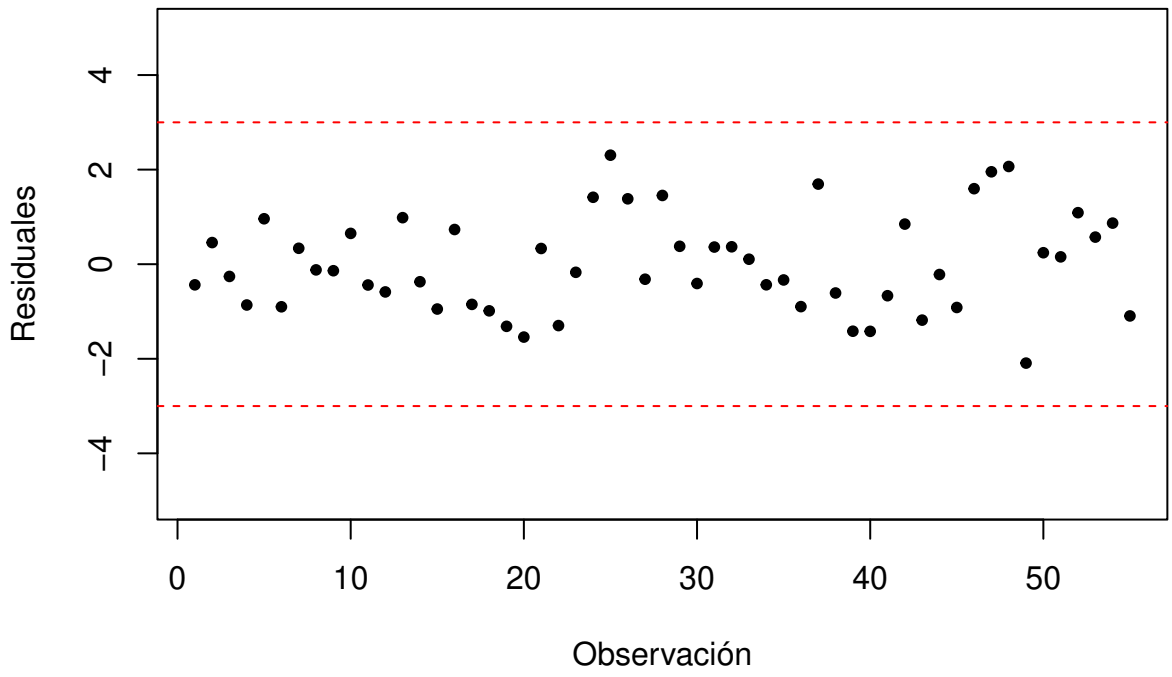

Figura 3: Identificación de datos atípicos

La figura muestra que no hay valores atípicos en el modelo según el criterio de los residuos estandarizados ($|r_i| > 3$). Sin embargo, es necesario considerar que existen otros métodos para detectar valores atípicos y se recomienda realizar un análisis más completo antes de descartar su presencia en el modelo.

4.2.2. Puntos de balanceo

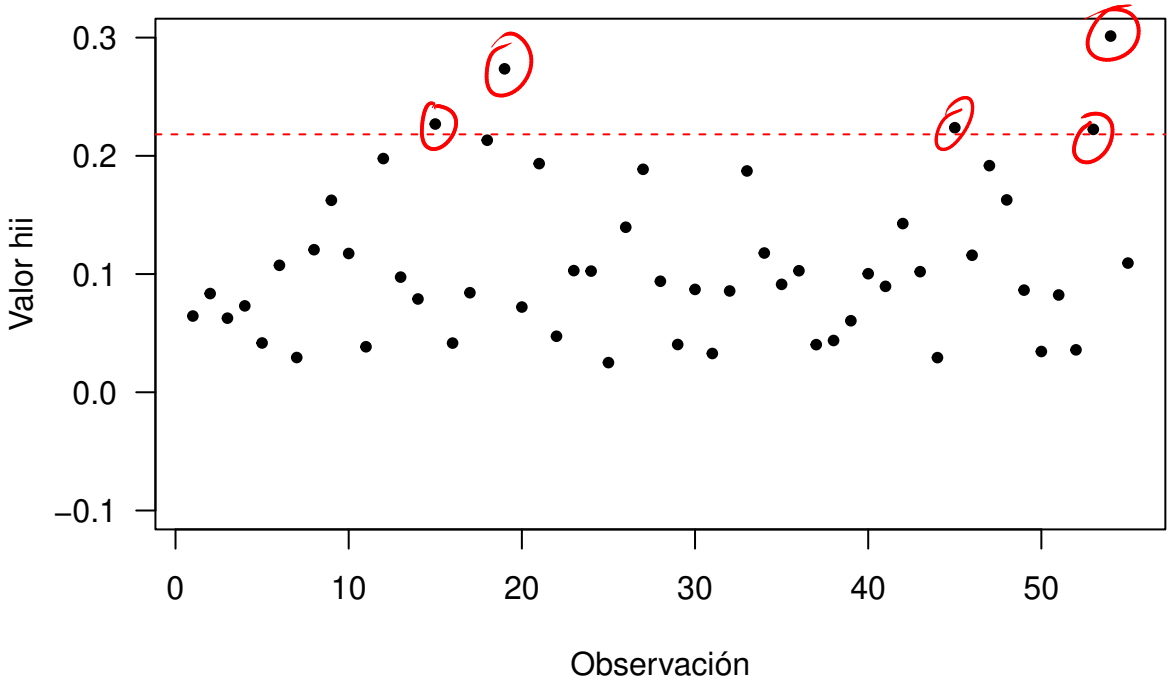

Figura 4: Identificación de puntos de balanceo

Tabla 5: Tabla de puntos de Balanceo

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
15	-0.9478	0.0440	0.2269	-0.5135
19	-1.3140	0.1068	0.2736	-0.8065
45	-0.9150	0.0404	0.2238	-0.4913
53	0.5729	0.0159	0.2224	0.3064
54	0.8690	0.0545	0.3013	0.5706

El modelo tiene 5 puntos de balanceo según el criterio hii $>$ 2p/n y el gráfico de la diagonal principal de la matriz Hat. Estos puntos pueden influir en el ajuste y las propiedades del modelo, por lo que es importante analizar su impacto antes de llegar a conclusiones. Los datos de balanceo son 15, 19, 45, 53 y 54, ya que superan el valor de 2p/n.

4.2.3. Puntos influenciales

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

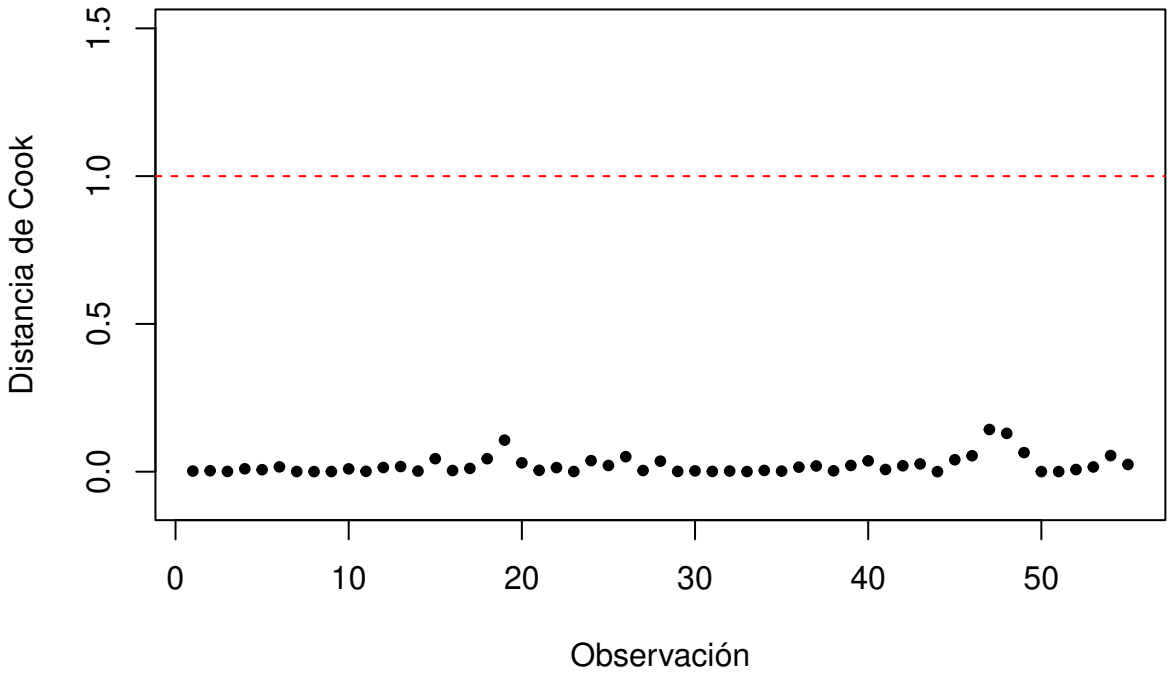

Figura 5: Criterio distancias de Cook para puntos influenciales

La gráfica del criterio de Cook sugiere que no existen puntos influyentes en el modelo de acuerdo con este criterio.

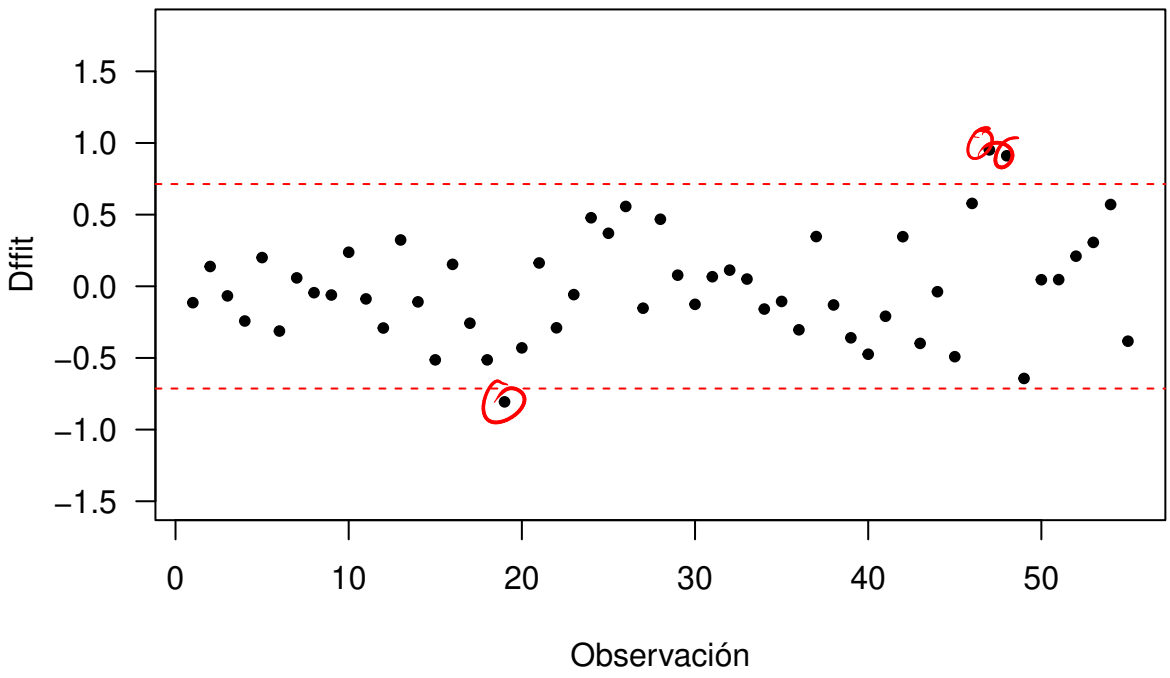

Figura 6: Criterio Dffits para puntos influenciales

Tabla 6: Tabla del criterio DFFITS para encontrar puntos influenciales

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
19	-1.3140	0.1068	0.2736	-0.8065
47	1.9542	0.1428	0.1917	0.9518
48	2.0663	0.1296	0.1627	0.9109

Se ha obtenido la gráfica utilizando el criterio de Dffits, la cual muestra la presencia de varios valores influyentes en el modelo, específicamente en las observaciones 19, 47 y 48. Es importante realizar un análisis detallado de la influencia de estos valores en el modelo de regresión y determinar si es necesario corregirlos o excluirlos. Por lo tanto, se requiere realizar un análisis adicional para evaluar la influencia de estos puntos en el modelo.

¿Qué análisis? ¿Qué causan los influenciales según este criterio? 1,5 pt

En estadística no se habla de buenas señales

4.3. Conclusiones

El modelo de regresión parece estar en línea con los supuestos de normalidad y homocedasticidad, lo que es una buena señal. Sin embargo, se han identificado algunos puntos influyentes utilizando el criterio de Dffits y también hay algunos puntos de balanceo que parecen tener un impacto significativo en los resultados del modelo. Por lo tanto, es importante tener en cuenta estos puntos y evaluar la posibilidad de eliminarlos o ajustarlos en futuros análisis para mejorar la calidad de las predicciones del modelo. Se debe llevar a cabo

un análisis adicional para determinar la mejor estrategia para manejar estos puntos influyentes y mejorar la validez y precisión del modelo.

Ughh, qué análisis?

No dicen si el modelo es válido o no.