3,85

Trabajo 1

Estudiantes

Jalan Howard Hudgson
Juan José Lopez Taborda
Diego Quintero Gaítan

Docente

Francisco Javier Rodríguez Cortes

Asignatura

Estadística II

Sede Medellín
27 de Marzo de 2023

Índice

Índice de figuras

Índice de tablas

1. Pregunta 1

17 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 **Equipo42.txt**, las variables para el modelo son

Y RI Riesgo de infección en porcentaje: Probabilidad promedio estimada de adquirir infección en el hospital.

X1 DEHOS 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 NCP 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 ENFP Número de enfermeras: Promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio.

El modelo que se propone es:

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

$i = 1, 2, \dots, 65$

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
$\hat{\beta}_0$	-1.84471
$\hat{\beta}_1$	0.18354
$\hat{\beta}_2$	0.02896
$\hat{\beta}_3$	0.04178
$\hat{\beta}_4$	0.02269
$\hat{\beta}_5$	0.00121

No va en ec ajustada

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

$$\widehat{RI}_i = -1.84471 + 0.18354 DES_i + 0.02896 RC + 0.04178 NC_i + 0.02269 CPD_i + 0.00121 NENF_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Donde las variables se mueven deacuerdo $1 \leq i \leq 65$

1.2. Significancia de la regresión

4 pt

Se Plantea el siguiente Juego de Hipotesis

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

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	70.1988	5	14.039762	14.3115	3.61266e-09
Error	57.8797	59	0.981011		

Los resultados obtenidos de la Tabla Anova indican que la hipótesis nula debe ser rechazada. Esto nos lleva a concluir que la regresión es significativa según la evidencia muestral analizada.

1.3. Significancia de los parámetros

6 pt

¿Y β_0?

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

La tabla a continuación muestra los criterios utilizados 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	-1.8447	1.6298	-1.1319	0.2623
β_1	0.1835	0.0750	2.4472	0.0174
β_2	0.0290	0.0294	0.9844	0.3289
β_3	0.0418	0.0129	3.2407	0.0020
β_4	0.0227	0.0080	2.8505	0.0060
β_5	0.0012	0.0007	1.7369	0.0876

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 son significativos cada uno en presencia de los demás parametros Por el contrario los parametros β_0,β_2,β_5 individualmente no son significativos en presencia de los demas parametros

1.4. Interpretación de los parámetros

2 pt

A continuacion se hara la interpretacion de los parametros que son significativos, ya que los otros parametros no tiene interpretacion ~~y no aportan al modelo.~~

- $\hat{\beta}_1$= 0.18354:Si se mantiene constante el efecto de las demás variables predictoras, un incremento de un día en la Duración de la estancia en el hospital (medida en días) resultaría en un aumento esperado del promedio del Riesgo de infección en un ~~0.18354~~ %, según los resultados del análisis de regresión
  18,354%

probabilidad

- $\hat{\beta}_3$= 0.04178: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 ~~0.04178~~ %

4,178%

- $\hat{\beta}_4$= 0.02269: si el número censo del promedio Diario del paciente en el hospital durante el periodo de estudio se incrementa en una unidad, cuando las demas variables se mantienen constantes, se espera que el promedio del Riesgo de infección se incrementa en un ~~0.02269~~ %

2,269 %

3 pt

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un R^2 de 0.5481lo cual significa que aproximadamente el 54.81 % de la variabilidad total en el porcentaje de Riesgo de infeccion es explicado por el modelo RLM

¿Cómo se calcula?

1.6. Comentarios

En el modelo de regresión, se puede observar que las variables que tienen un aporte significativo son la Duración de la estadía en el hospital (DE), el Censo promedio diario de pacientes en el hospital (CDP) y el número de camas, lo cual se ve reflejado en la significancia de los parámetros.

2. Pregunta 2

3 pt

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=0.3284, β_5 con VP= 0.087, β_1 con VP= 0.0174. Por tanto, se plantea la siguiente prueba de hipótesis:

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

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

$$\text{MR: } Rinf_i = \beta_0 + \beta_3 NCP_i + \beta CPD4_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

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	57.880	X1 X2 X3 X4 X5
Modelo reducido	75.342	X3 X4

Así no se llaman sus var's

2.2. Estadístico de prueba y conclusiones

$$\frac{SSR(\beta_1, \beta_2, \beta_5 \mid \beta_0, \beta_3, \beta_4)/3}{MSE(MF)}$$

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSR\cancel{(\beta_0, \beta_3, \beta_4 | \beta_0, \beta_1, \beta_2, \beta_3}, \beta_4, \beta_5/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,59}$$

3

1,5 pt

3

$$F_0 = \frac{(SSE(\beta_0,\beta_3,\beta_4) - SSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5))/2}{MSE(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)} \overset{H_0}{\sim} f_{2,59}$$

3

$$= \frac{(75.342 - 57.880)/2}{57.880/59} = 8.899948$$

?

1,5 pt

Al comparar a un nivel de significancia $\alpha = 0.05$, F_0 con $f_{0.05,2,59} = 3.153123$. Con valor P=4.1878942×10^{-4}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.Por tanto concluimos que no que no se puede descartar este subconjunto de datos dle modelo

La conclusión de la PH es que son significativas y luego ahí sí

3. Pregunta 3

9 pt

Plantee una pregunta donde su solución implique el uso exclusivo de una prueba de hipótesis lineal general de la forma H0 : Lβ = 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:

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

1,5 pt

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:

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

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

Donde el modelo reducido está dado por:

$$RI = \beta_0 + \beta_1(DES_i + ENFP_i) + \beta_2(RC_i + NCP_i) + \beta_4 CPD_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$$

1 pt

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

1,5 pt

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,59}$$

Obteniendo esto podemos definir la region de rechazo de la hipotesis nula como $F_0 > F_{0.05,2,59} = 3.153123$ y con valor p:P($F_{2,59} > |F_0|$)

Reenplazar lo que conocen

F

4. Pregunta 4

19,5 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

2 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 se está probando media cte μ y var cte σ²

acompañado de un grafico cuantil-cuantil:

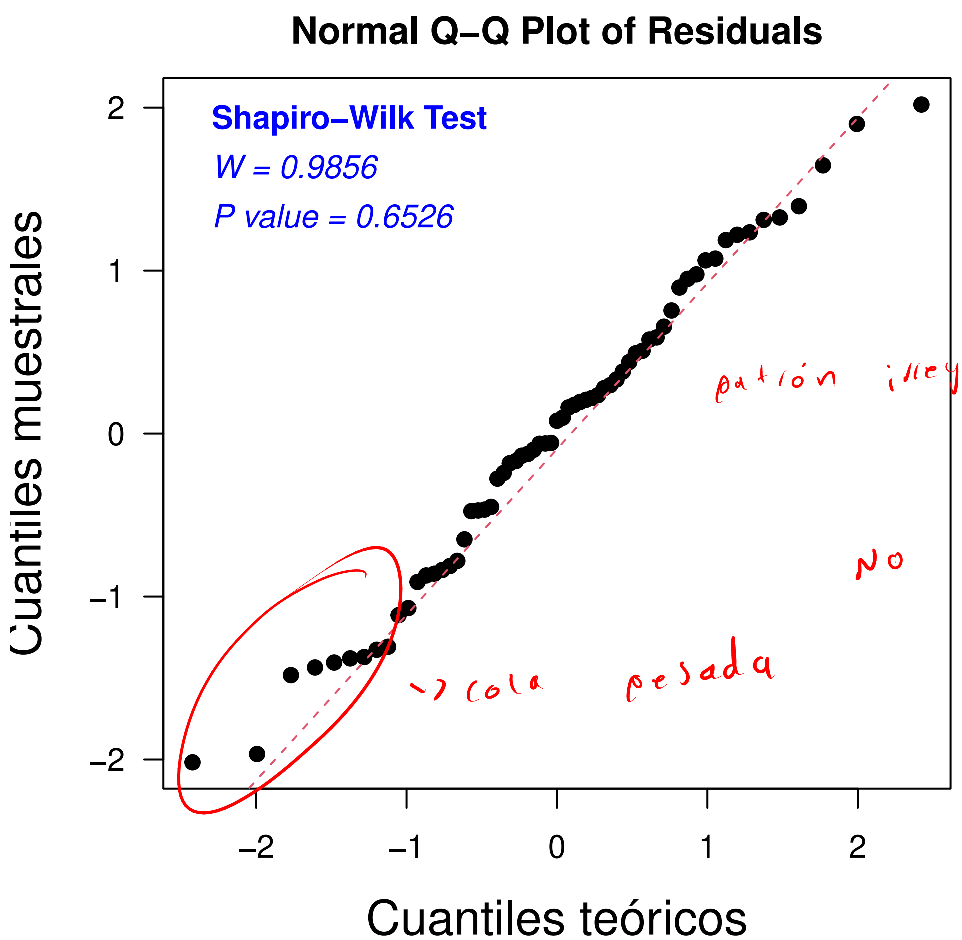


Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Ya queel valor de P es alto, se puede inferir que no hay suficiente evidencia para rechazar la hipótesis nula H_0. Por lo tanto, se puede concluir que el modelo es congruente con la asunción de que los residuos se distribuyen normalmente.

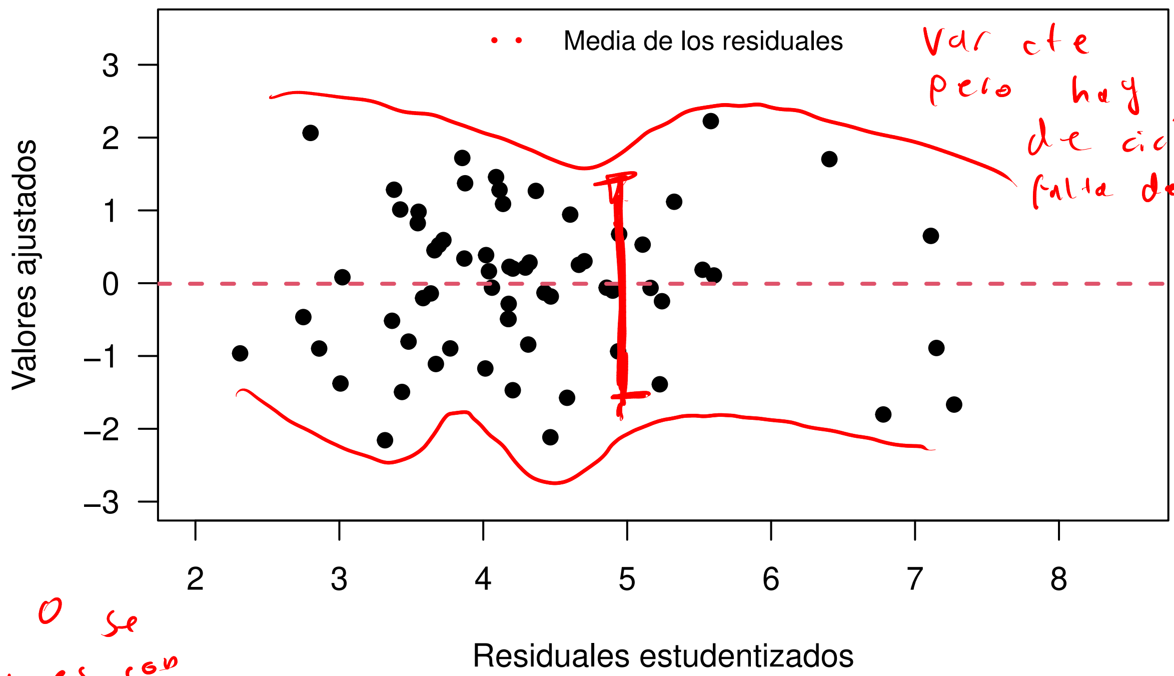
Es más importante el análisis gráfico y no lo hicieron

4.1.2. Media 0 y Varianza constante

2pt

En esta prueba se quiere probar

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



Var cte pero hay patrón de ciclos, posible falta de ajuste.

media 0 se prueba es con residuales crudos

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Podemos notar que la línea punteada roja, que muestra la media de los errores, está en cero o cerca de cero. Esto sugiere que los errores tienen una media cercana a cero. Al analizar los residuos, no se observa ningún patrón discernible, lo que indica que la varianza de los errores es constante en todo el rango de los valores observados. En resumen, la media cercana a cero de los errores y la constancia de la varianza de los residuos sugieren que el modelo se ajusta bien a los datos y cumple con los supuestos básicos de la regresión lineal.

no cualquier patrón es por var no cte

4.2. Observaciones extremas

4.2.1. Datos atípicos

2,5 e+

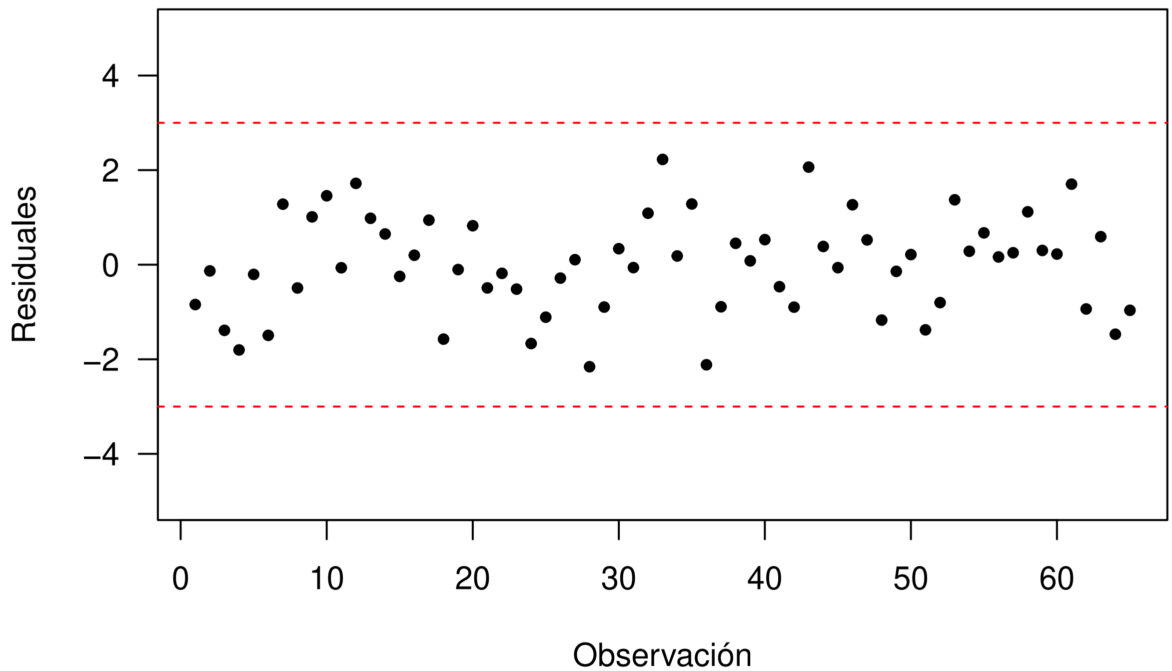


Figura 3: Identificación de datos atípicos

Notese que segun este criterio no existen puntos atipicos que deban ser investigados

¿Qué dice el criterio?

4.2.2. Puntos de balanceo

2,5 pt

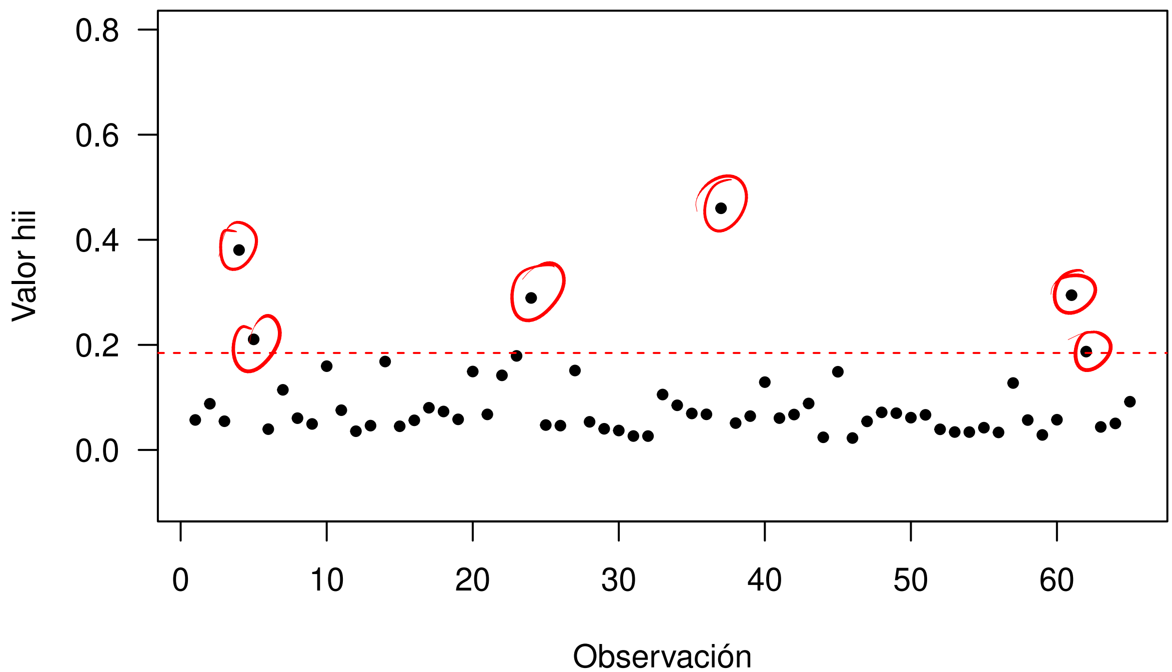


Figura 4: Identificación de puntos de balanceo

Tabla 5: Tabla de puntos de Balanceo

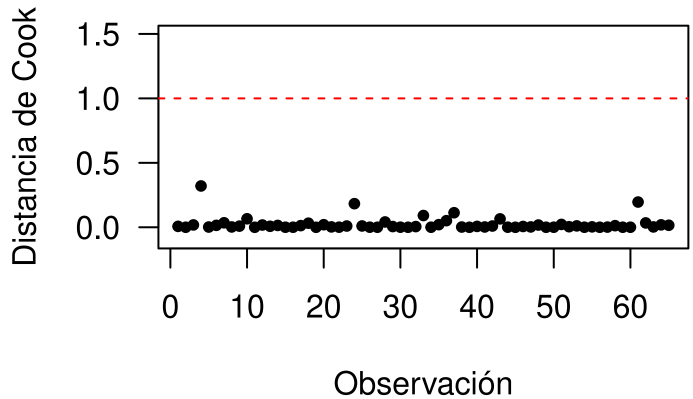
	Errores Estudentizados	D.Cook	Valor hii	DFFITS
4	-1.8031	0.3206	0.3805	-1.4131
5	-0.2050	0.0019	0.2105	-0.1058
24	-1.6672	0.1831	0.2893	-1.0638
37	-0.8901	0.1129	0.4600	-0.8215
61	1.7038	0.1958	0.2946	1.1012
62	-0.9362	0.0337	0.1872	-0.4493

Es importante destacar que hay seis datos que deben ser cuidadosamente analizados en términos de su impacto en el ajuste del modelo y sus propiedades. Estos datos corresponden a los puntos 4, 5, 24, 37, 61 y 62, ya que son mayores que el valor crítico $\frac{2p}{n}$. Estos puntos de balanceo pueden tener una gran influencia en el modelo y, por lo tanto, es crucial llevar a cabo un análisis detallado de su impacto antes de sacar conclusiones definitivas.

¿En qué influyen específicamente?

4.2.3. Puntos influenciales

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



2 pt

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

redundantes

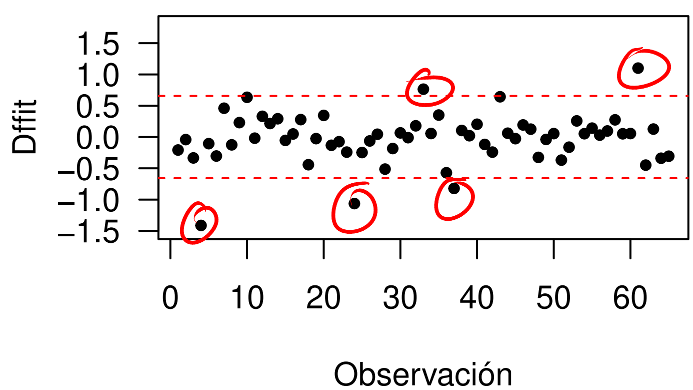


Figura 6: Criterio Dffits para puntos influenciales

1,5 pt

Tabla 6: Tabla del criterio DFFITS para encontrar puntos influenciales

	Errores Estudentizados	D.Cook	Valor hii	DFFITS
4	-1.8031	0.3206	0.3805	-1.4131
24	-1.6672	0.1831	0.2893	-1.0638
33	2.2262	0.0913	0.1055	0.7644
37	-0.8901	0.1129	0.4600	-0.8215
61	1.7038	0.1958	0.2946	1.1012

Dijeron lo mismo de los de balanceo, ¿qué afectan?

Puntos

Usando el criterio de Dffits, se ha generado el gráfico anterior, el cual sugiere que ~~hay varios valores~~ influyentes en el modelo. Específicamente, las observaciones 4, 24, 33, 37 y 61 parecen tener un impacto significativo en el modelo y deben ser investigadas con más detalle. Es necesario realizar un análisis adicional para determinar si estos valores influyentes deben ser eliminados o ajustados de alguna manera y evaluar su impacto en el modelo de regresión.

4.3. Conclusiones

2 pt — Y normalidad?

El modelo cumple con los supuestos básicos de la regresión lineal, es decir, que la media de los residuos es cercana a cero y la varianza es constante. Sin embargo, se observa un gran número de datos de balanceo e influenciadores en el modelo, lo que indica la necesidad de investigar si estos datos están afectando significativamente el modelo y sus supuestos, incluyendo la normalidad de los residuos. En conclusión, aunque el modelo cumple con los supuestos básicos, no es adecuado para hacer predicciones y se debe llevar a cabo un análisis adicional para evaluar la influencia de los datos de balanceo e influyentes y determinar si el modelo es una buena estimación.

¿Por qué no si cumple los supuestos? Ni siquiera dijeron si era válido según lo que encontraron.