4,15

Trabajo 1

Estudiantes

Alejandra Diego Ramirez Agudelo
Andres Felipe Ramirez Suarez
Juan Diego Ramirez Agudelo
Sara Paulina Aguirre Restrepo
Equipo # 37

Docente

Fransisco Javier Rodriguez Cortes

Asignatura

Estadistica II

Sede Medellin
30 de marzo de 2023

Índice

1. Pregunta 1 **3**

1.1. Modelo de regresion . . . 3

1.2. Significancia de la regresion . . . 3

1.3. Significancia de los parametros . . . 3

1.4. interpretación de los parámetros . . . 4

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

2. Pregunta 2 **5**

2.1. Planteamiento prueba de hipótesis y modelo reducido . . . 5

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 . . . 6

4. Pregunta 4 **6**

4.1. Supuestos del modelo . . . 6

4.1.1. Normalidad de los residuales . . . 6

4.1.2. Varianza constante . . . 8

4.2. Verificación de las observaciones . . . 9

4.2.1. Datos atípicos . . . 9

4.2.2. Puntos de balanceo . . . 10

4.2.3. Puntos influenciales . . . 11

4.3. Conclusión . . . 12

Índice de figuras

1. Gráfico cuantil-cuantil y normalidad de residuales . . . 7
2. Gráfico residuales estudentizados vs valores ajustados . . . 8
3. Identificación de datos atípicos . . . 9
4. Identificación de puntos de balanceo . . . 10
5. Criterio distancias de Cook para puntos influenciales . . . 11
6. Criterio Dffits para puntos influenciales . . . 12

Índice de tablas

1. Tabla de valores de los coeficientes estimados . . . 3
2. Tabla ANOVA significancia de la regresion . . . 3
3. Resumen de los coeficientes . . . 4
4. Resumen de todas las regresiones . . . 5
5. Valor covariables (*) . . . 6
6. TablaPuntos influenciales tomando en cuenta hii.value . . . 10
7. Tabla Puntos influenciales tomando en cuenta Diffts . . . 12

1. Pregunta 1

17 pt

Teniendo en cuenta la base de datos asignada. La cual es **equipo37.txt**. Las covariables son: duración de la estadía (X_1), rutina de cultivos (X_2), número de camas (X_3), censo promedio diario (X_4) y número de enfermeras (X_5).

El modelo propuesto seria, entonces:

$$Y_j = \beta_0 + \beta_i X_{ij} + \varepsilon_j, \quad \varepsilon_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2); \quad 1 \leq i \leq 5; \quad 1 \leq j \leq 55$$

$Y_j = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{2i} + \beta_3 \ldots$

1.1. Modelo de regresion

Al ajustar el modelo, se obtienen los siguiente coeficientes:

Tabla 1: Tabla de valores de los coeficientes estimados

	Valor del parametro
$\hat{\beta}_0$	-2.8475
$\hat{\beta}_1$	0.2774
$\hat{\beta}_2$	0.0370
$\hat{\beta}_3$	0.0565
$\hat{\beta}_4$	0.0122
$\hat{\beta}_5$	0.0020

3 pt

El modelo ajustado es:

$$\widehat{Y}_j = -2.8475 + 0.2774X_{1j} + 0.037X_{2j} + 0.0565X_{3j} + 0.0122X_{4j} + 0.002X_{5j}$$

1.2. Significancia de la regresion

2,5 pt

Para la interpretación se utilizará la tabla de análisis de varianza.
La cual da como resultado lo siguiente:

¿Prueba de hipótesis? ¿Estadístico de prueba y cómo distribuye?

Tabla 2: Tabla ANOVA significancia de la regresion

	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor-P
Modelo de regresion	84.8460	5	16.96920	16.518	1.55928e-09
Error	50.3384	49	1.02731		

¿Dónde está?

Tomando a $\alpha = 0.05$.
Se tiene que $1.55928 * 10^{-9} < 0.05$, por lo $\boxed{\text{cual se rechaza la } H_0}$ (el modelo no es significativo), concluyendo que el modelo de RLM propuesto es significativo. Eso significa que el riesgo de infección hospitalarias dependen de forma significativa de por lo menos una de las variables predictoras.

1.3. Significancia de los parametros

6 pt

En el siguiente cuadro se presenta información de los parámetros. Lo cual permitirá calcular la significancia de ~~las variables predictoras.~~

los parámetros

Tabla 3: Resumen de los coeficientes

	Estimacion β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-2.8475	1.7659	-1.6125	0.1133
β_1	0.2774	0.1225	2.2637	0.0281
β_2	0.0370	0.0333	1.1089	0.2729
β_3	0.0565	0.0166	3.4056	0.0013
β_4	0.0122	0.0086	1.4280	0.1596
β_5	0.0020	0.0007	2.9134	0.0054

parámetros

Los valores P permiten concluir con una significancia $\alpha = 0.05$ que $\hat{\beta}_1$, $\hat{\beta}_3$ y $\hat{\beta}_5$ son los únicos ~~valores~~ significativos. Por otro lado, en el caso $\hat{\beta}_0$ el valor 0 no se encuentra incluido en el ajuste. Por ende, no es interpretable y como su valor es mayor al alfa, tampoco es significativo.

1.4. interpretación de los parámetros

2,5 pt

- $\hat{\beta}_1$ su valor $P = 0.0281 < 0.05$, por lo que es un parámetro significativo en presencia de los demás en el modelo. Se tiene que su valor estimado es $\hat{\beta}_1 = 0.2774$, además X_1 representa duración de la estadía de cada paciente en el hospital. Se podría interpretar que el aumento en 1 unidad de este número representa un aumento de 0.2774 en el riesgo de infección hospitalarias, mientras las otras variables permanecen constantes.

  la probabilidad promedio.

- $\hat{\beta}_3$ su valor $P = 0.0013 < 0.05$, por lo que es un parámetro significativo en presencia de los demás en el modelo. Se tiene que su valor estimado es $\hat{\beta}_3 = 0.0565$, además X_3 representa el numero promedio de camas en el hospital durante el periodo de estudio. Se podría interpretar que el aumento en 1 unidad de este número representa un aumento de 0.0565 en el riesgo de infección hospitalarias, mientras las otras variables permanecen constantes.
- $\hat{\beta}_5$ su valor $P = 0.0054 < 0.05$, por lo que es un parámetro significativo en presencia de los demás en el modelo. Se tiene que su valor estimado es $\hat{\beta}_5 = 0.0020$, además X_5 representa el numero promedio de enfermeras equivalentes a tiempo completo. Se podría interpretar que el aumento en 1 unidad de este número representa un aumento de 0.0020 en el riesgo de infección hospitalarias, mientras las otras variables permanecen constantes.

1.5. Coeficiente de determinación R^2

3 pt

Basado en la tabla **ANOVA** se tiene que:

$$R^2 = \frac{84.8460}{84.8460 + 50.3384} = 0.6276316$$

El R^2 mide la proporción de la variabilidad total observada en la respuesta que es explicada por el modelo propuesto. El modelo tiene un $R^2 = 0.6276316$ lo cual significa que las variables independientes explican aproximadamente el 62.76 % de la variabilidad de Y (Riesgos de infeccion hospitalarias).

2. Pregunta 2

3pt

2.1. Planteamiento prueba de hipótesis y modelo reducido

Pedian las 3 covariables.

Los parámetros cuyos valores P fueron los más altos corresponden a $\boxed{\beta_2, \beta_4, \beta_0}$. Por lo tanto, se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_4 = \beta_0 = 0 \\ \mathrm{H}_a : \text{Algun } \beta_i \text{ distinto a 0 con i = 2, 4, 0} \end{cases}$$

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

$$\text{MR: } Y_j = \beta_1 X_{1j} + \beta_3 X_{3j} + \beta_5 X_{5j} + \varepsilon_j, \quad \varepsilon_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

i = 1, 2 ...

Tabla 4: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	50.338	X1 X2 X3 X4 X5
Modelo reducido	54.036	X1 X3 X5

→ esto fue ajustado con β_0

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba

$$F_0 = \frac{(SSE(\beta_1, \beta_3, \beta_5) - SSE(\beta_0, \beta_1, \cdots, \beta_5))/3}{MSE(\beta_0, \beta_1, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,49}$$

1pt

$$F_0 = \frac{(35.623 - 34.832)/3}{34.832/49} = 1.199902$$

Al menos son congruentes con el error

Ahora, comparando con un nivel de significancia $\alpha = 0.05$, F_0 con $f_{0.05,3,49} = 2.7939489$ y valor $P = 0.3196538$.

Como F_0 es MENOR que 2.7939489. Entonces, no es posible rechazar H_0, por lo que el subconjunto de las variables no es significativo. Teniendo en cuenta lo anterior, es posible descartar las variables. Debido a que son insignificantes para el modelo. Es decir, su valores son iguales a 0.

2pt

↳ estadisticamente iguales a 0

3. Pregunta 3

5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 - \beta_5, \beta_4 = \beta_2 + \beta_5 \\ \mathrm{H}_a : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

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:

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

2 pt

Con $r = 2$ filas linealmente independientes

1 pt

Donde el modelo reducido está dado por:

$$\begin{aligned} \text{MR: } Y_j &= \beta_0 + (\beta_3 - \beta_5)X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + (\beta_2 + \beta_5)X_{4j} + \beta_5 X_{5j} + \varepsilon_j \\ &= \beta_0 + \beta_2(X_{2j} + X_{4j}) + \beta_3(X_{1j} + X_{3j}) + \beta_5(X_{4j} + X_{5j} - X_{1j}) + \varepsilon_j \\ &= \beta_0 + \beta_2 X^*_{2j} + \beta_3 X^*_{3j} + \beta_5 X^*_{5j} + \varepsilon_j \end{aligned}$$

Suponiendo que: $\varepsilon_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$

En donde los valores de las covariables (*) son:

Tabla 5: Valor covariables (*)

	X_{2j}*	X_{3j}*	X_{5j}*
Valor	$X_{2j} + X_{4j}$	$X_{1j} + X_{3j}$	$X_{4j} + X_{5j} - X_{1j}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,49}$$

2 pt

$$F_0 = \frac{(SSE(MR) - 50.338/2}{50.338/49} \overset{H_0}{\sim} f_{2,49}$$

Perfecto

4. Pregunta 4

16,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

3,5 pt

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis ~~shapiro-wilk~~, acompañada de un gráfico cuantil-cuantil:

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

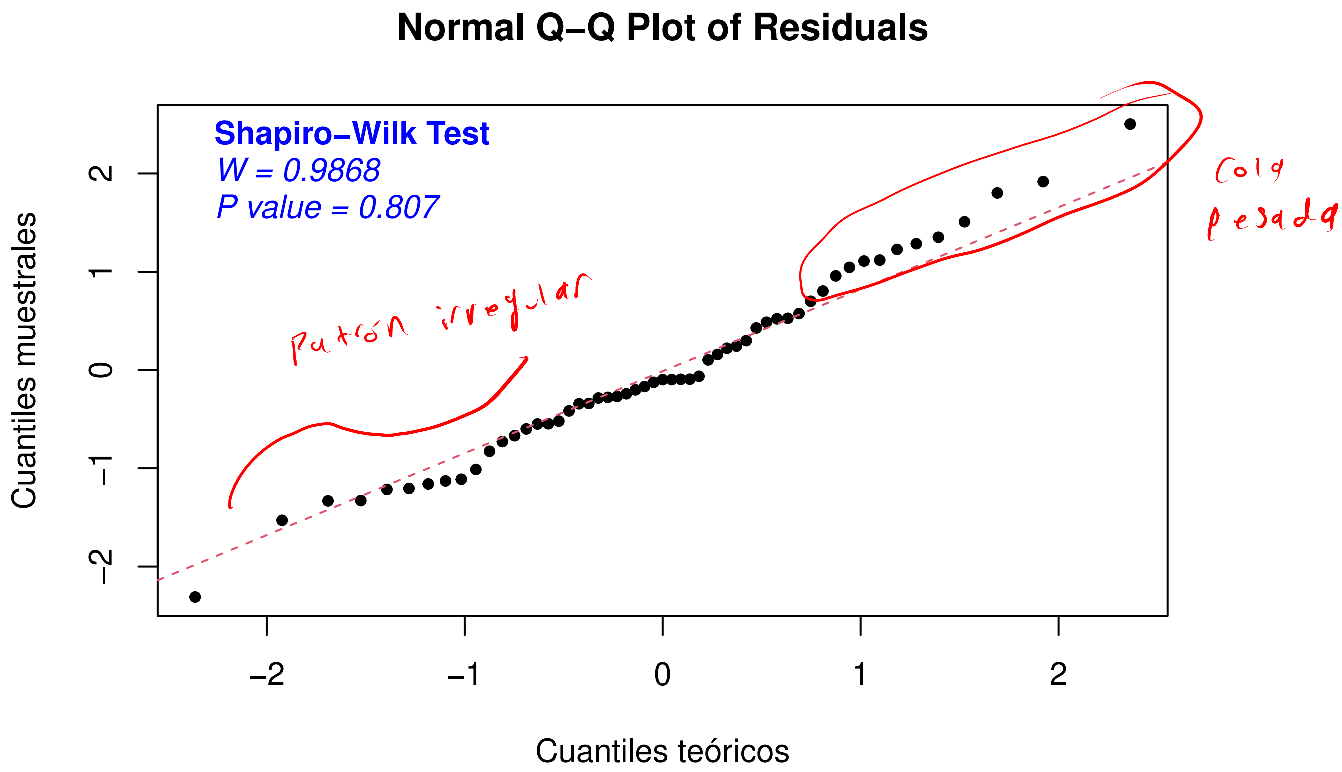


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

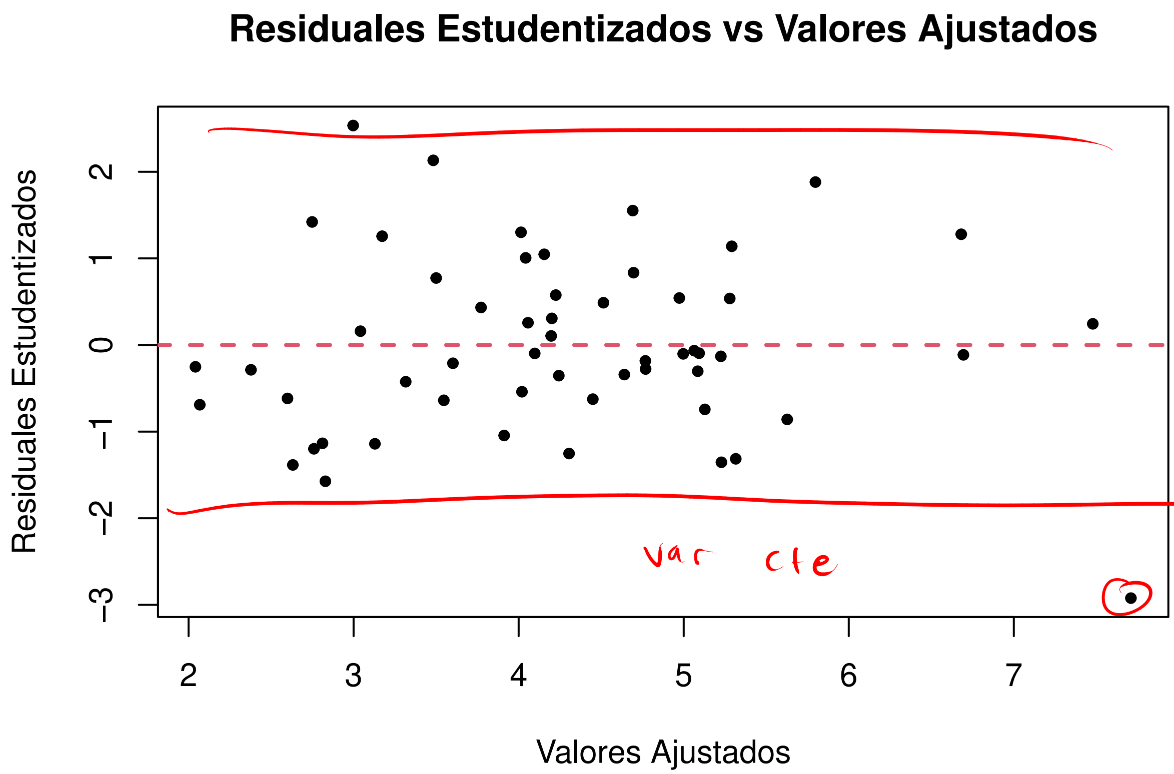
Al ser el P-valor aproximadamente igual a 0.807 y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, no se rechazaría la hipótesis nula, es decir que los datos distribuyen normal con media μ y varianza σ^2, sin embargo la gráfica de comparación de cuantiles permite ver patrones irregulares, al tener más poder el análisis gráfico, se termina por rechazar el cumplimiento de este supuesto.

↳ y cola pesada

¿No están probando media constante μ ni var cte σ^2

3pt

4.1.2. Varianza constante



extremo

Figura 2: Gráfico residuales estudentizados vs valores ajustados

El modelo cumple con media 0 y ademas podemos ver en el gráfico de residuales estudentizados vs valores ajustados se puede observar que $\boxed{\text{no hay patrones}}$ en los que la varianza muestre un comportamiento que permita descartar una varianza constante, al no haber evidencia suficiente en contra de este supuesto se acepta como cierto

no cualquier patrón es adjudicado a eso

El análisis pudo haber sido hecho mucho mejor.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3 pt

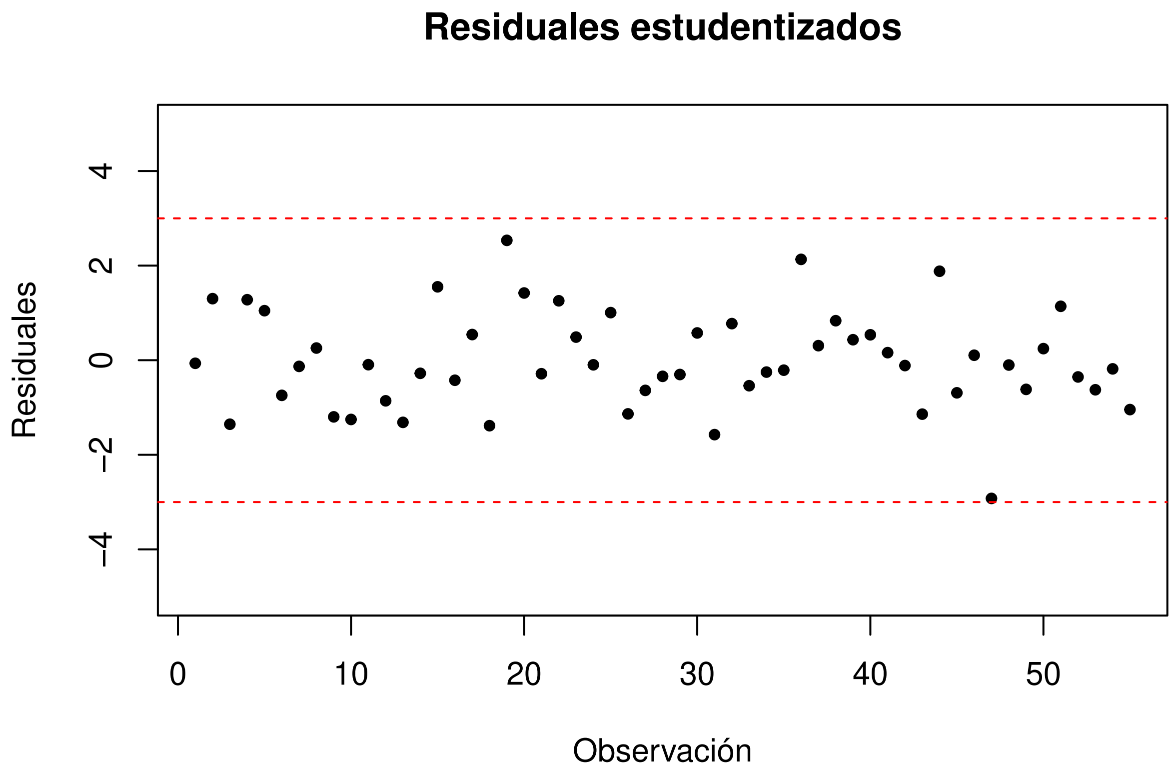


Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica anterior, no hay datos atípicos en el conjunto de datos pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$,solo hay dato que se acerca mucho a este criterio, el cual seria el dato Numero 47 teniendo un residual estudentizado de -2.9225.

4.2.2. Puntos de balanceo

2 pt

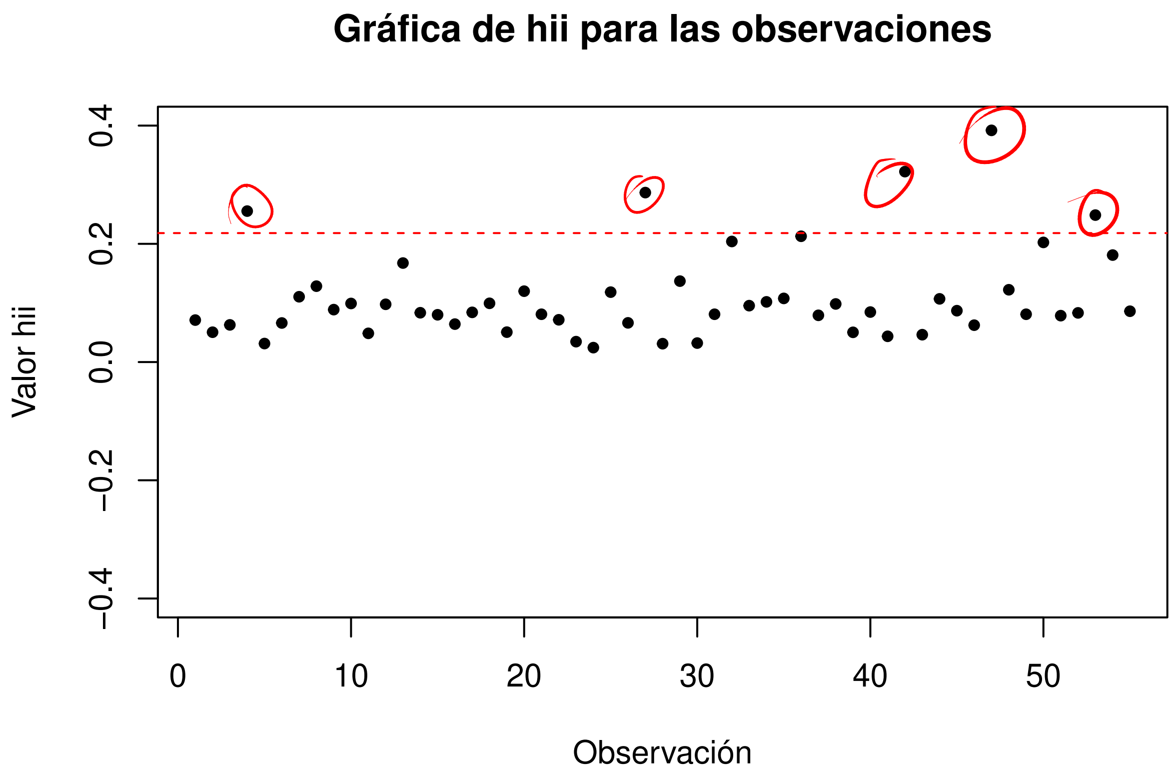


Figura 4: Identificación de puntos de balanceo

balanceo

Tabla 6: TablaPuntos ~~influenciales~~ tomando en cuenta hii.value

	hii.value
punto 4	0.2554
punto 27	0.2867
punto 42	0.3222
punto 47	0.3920
punto 53	0.2486

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.2181$, se puede apreciar que existen 5 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n} = 0.2181$, los cuales son los presentados en la tabla.

¿Qué causan?

4.2.3. Puntos influenciales

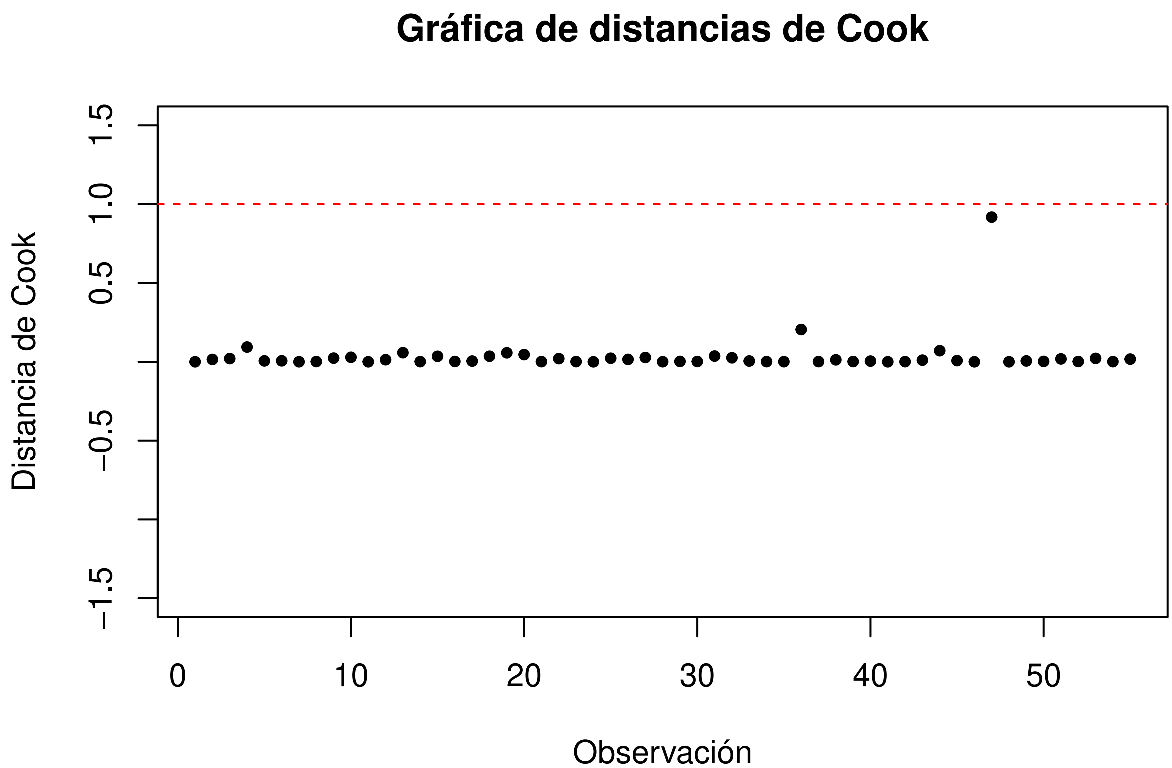
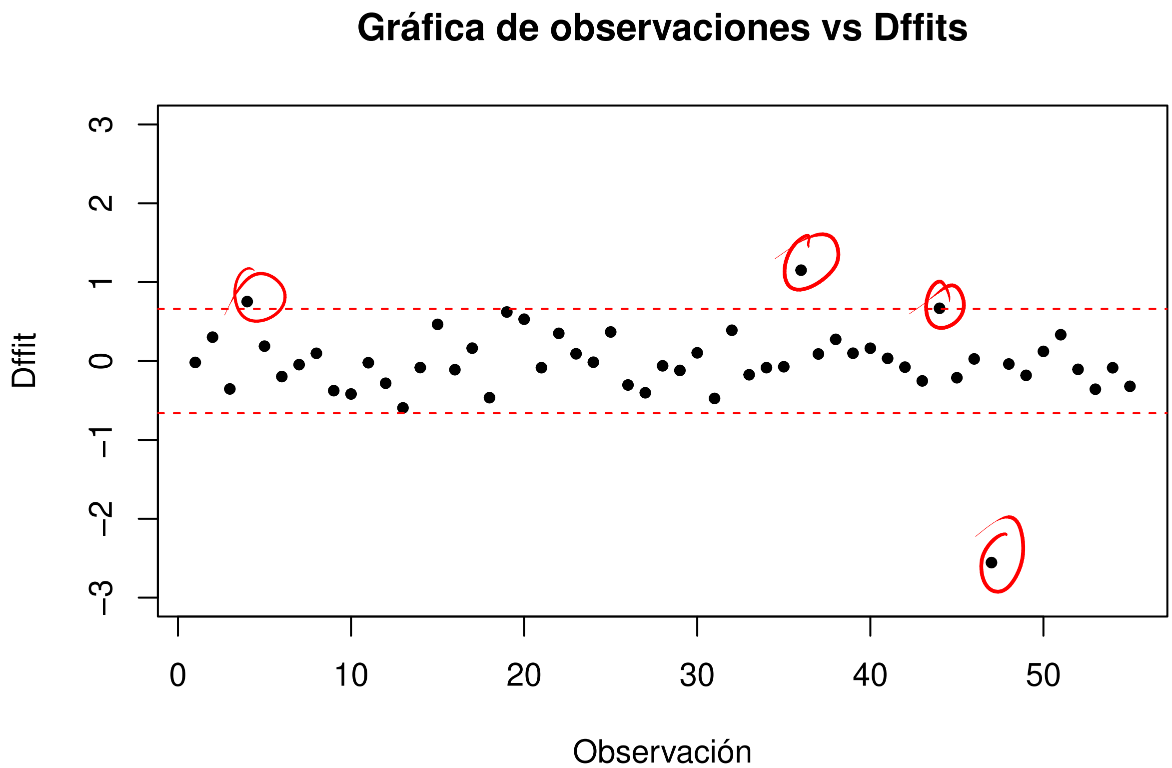


Figura 5: Criterio distancias de Cook para puntos influenciales

Como se observa en la grafica no tenemos ningun punto influencial tomando en cuenta el criterio de Cook, con lo cual se podria decir que los datos no están siendo afectados de manera significativa por ningún punto ~~particular~~ en la gráfica. influencial

2 pt



1,5 pt

Figura 6: Criterio Dffits para puntos influenciales

Tabla 7: Tabla Puntos influenciales tomando en cuenta Diffts

	Dffits
punto 4	0.7540
punto 36	1.1518
punto 44	0.6686
punto 47	-2.5559

¿cuánto da?

Como se puede ver,hay 4 puntos influenciales según el criterio de Dffits, los cuales son 4, 36, 44 y 47, respectivamente,este criterio dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial. Cabe destacar también que con el criterio de distancias de Cook, en el cual para cualquier punto cuya $D_i > 1$, es un punto influencial, y como se dijo anteriormente, ninguno de los datos cumple con serlo.

¿Qué causan?

4.3. Conclusión

1,5 pt

1. El modelo no sigue una distribución normal, ya que se observan patrones irregulares en la gráfica de comparación de cuantiles.
2. El modelo cumple con la suposición de que la media es 0 y que la varianza es constante, ya que en la gráfica de residuales estudentizados vs valores ajustados no se observan patrones que sugieran lo contrario.

3. No hay valores atípicos en el conjunto de datos, ya que ningún residual estudentizado sobrepasa el criterio de $|\text{restud}| > 3$.

4. Basándonos en la tabla de análisis de varianza se rechaza la H0 (el modelo no es significativo), concluyendo que el modelo de RLM propuesto es significativo.

En general, estos resultados sugieren que el modelo es adecuado para analizar los datos, aunque es importante tener en cuenta que la falta de normalidad en la distribución de los datos puede afectar la precisión y la validez de los resultados del análisis estadístico.

Falso, su modelo no es válido por no cumplir todos los supuestos