Trabajo 1

4,5

Estudiantes

Sara Gabriela Muñoz Cabrera
Daniel Giraldo Vanegas
Sebastián Orjuela Alfonso
Simón Pedro Serna Cardona

Equipo 58

Docente

Javier Armando Lozano Rodriguez

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

5 de octubre de 2023

Índice

1. Pregunta 1 **3**

1.1. Modelo de regresión . . . 3

1.2. Significancia de la regresión . . . 4

1.3. Significancia de los parámetros . . . 4

1.4. Interpretación de los parámetros . . . 5

1.5. Coeficiente de determinación múltiple R^2 . . . 5

2. Pregunta 2 **6**

2.1. Planteamiento pruebas de hipótesis y modelo reducido . . . 6

2.2. Estadístico de prueba y conclusión . . . 6

3. Pregunta 3 **7**

3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . 7

3.2. Estadístico de prueba . . . 7

4. Pregunta 4 **8**

4.1. Supuestos del modelo . . . 8

4.1.1. Normalidad de los residuales . . . 8

4.1.2. Varianza constante . . . 9

4.2. Verificación de las observaciones . . . 10

4.2.1. Datos atípicos . . . 10

4.2.2. Puntos de balanceo . . . 11

4.2.3. Puntos influenciales . . . 12

4.3. Conclusión . . . 13

Índice de figuras

1. Gráfico cuantil-cuantil y normalidad de residuales 8
2. Gráfico residuales estudentizados vs valores ajustados 9
3. Identificación de datos atípicos 10
4. Identificación de puntos de balanceo 11
5. Criterio distancias de Cook para puntos influenciales 12
6. Criterio Dffits para puntos influenciales 13

Índice de cuadros

1. Tabla de valores coeficientes del modelo 3
2. Tabla ANOVA para el modelo 4
3. Resumen de los coeficientes 5
4. Resumen tabla de todas las regresiones 6

1. Pregunta 1

17pt

Teniendo en cuenta la base de datos brindada, en la cual hay 5 variables regresoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

Donde cada variable regresora representa lo siguiente:

- 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_1:Número de enfermeras

1.1. Modelo de regresión

Al ajustar el modelo, para obtener la relación de la variable respuesta con cada una de las variables regresoras se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	0.3199
β_1	0.2511
β_2	-0.0009
β_3	0.0443
β_4	0.0064
β_5	0.0017

3pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.3199 + 0.2511 X_{1i} - 9 \times 10^{-4} X_{2i} + 0.0443 X_{3i} + 0.0064 X_{4i} + 0.0017 X_{5i};\ 1 \leqslant i \leqslant 74$$

1.2. Significancia de la regresión

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

No va acá

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,68} \tag{1}$$

3pt

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	70.4484	5	14.08968	12.9464	7.46148e-09
Error	74.0051	68	1.08831		

De la tabla Anova, se observa un valor P muy cercano a 0, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $\cancel{0} \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto tiene significancia el modelo de regresión.

1.3. Significancia de los parámetros

Después de realizar la prueba general del modelo y de concluir su significancia, se realizará una prueba de hipótesis sobre los coeficientes individuales del modelo con el fin de saber cuáles son significativos o no, se establece primero el juego de hipotesis:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \;\; j = 0, 1, 2, ..., 5 \end{cases}$$

El estadístico es el siguiente:

$$T_{j,0} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \overset{H_0}{\sim} t_{68} \tag{2}$$

En el siguiente cuadro se presenta información de los parámetros, el 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	0.3199	1.6056	0.1993	0.8427
β_1	0.2511	0.0827	3.0344	0.0034
β_2	-0.0009	0.0300	-0.0287	0.9772
β_3	0.0443	0.0138	3.1972	0.0021
β_4	0.0064	0.0077	0.8319	0.4084
β_5	0.0017	0.0007	2.4375	0.0174

6pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3 y β_5 son significativos, pues sus valores son menores a α. En cuanto a β_0, no se hubiera podido interpretar puesto que ninguna de las $X_{j,i}$ contienen al 0.

1.4. Interpretación de los parámetros

3pt

$\hat{\beta}_1$**:** Por cada unidad de incremento en X_1 (duración de la estadía) el porcentaje promedio en el riesgo de infección aumenta en 0.2511 unidades cuando las demás variables predictoras permanecen constantes, es decir, a medida que los pacientes duren más días en el hospital hay mayor riesgo de infección.

$\hat{\beta}_3$**:** Por cada unidad de incremento en X_3 (número de camas) el porcentaje promedio en el riesgo de infección aumenta en 0.0443 unidades cuando las demás variables predictoras se mantienen constantes, entonces mientras hayan más camas ocupadas más aumenta el riesgo de infección.

$\hat{\beta}_5$**:** Por cada unidad que aumente X_5 (número de enfermeras) el porcentaje promedio en el riesgo de infección aumenta en 0.0017 unidades cuando las demás variables predictoras se mantienen constantes, en otras palabras es que según aumente el número de enfermeras también lo hace el riesgo de infección.

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.48768$, lo que significa que aproximadamente el 48.77 % de la variabilidad total observada en la variable respuesta (Y) es explicada por el modelo de regresión ajustado.

Cómo se calcula?

2. Pregunta 2

5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las 3 covariables con el valor P más pequeño en el modelo fueron X_1, X_3, X_5.Por lo tanto, a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	74.005	X1 X2 X3 X4 X5
Modelo reducido	116.611	X2 X4

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,68} \\ &= \frac{[116.611 - 74.005]/3}{74.005/68} \\ &= 13.049 \end{aligned} \tag{3}$$

Ahora, comparando a un nivel de significancia $\alpha = 0.05$, se tiene el F_0 con $f_{0.95,3,68} = 2.7395$ y se puede ver que $F_0 > f_{0.95,1,68}$, por tanto se rechaza la hipótesis nula lo que dice que el subconjunto mencionado es significativo.

Entonces se llega a la conclusión de que no puede considerarse apropiado descartar las variables incluidas en el subconjunto del modelo puesto que el riesgo de infección depende de al menos una de las covariables de este. 2pt

3. Pregunta 3

5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Suponga que se quiere probar las siguientes igualdades:

$$\beta_1 = 3\beta_2;\ 6\beta_3 = \beta_4;\ \beta_1 = \beta_5$$

Para ello se usa el siguiente juego de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 3\beta_2;\ 6\beta_3 = \beta_4;\ \beta_1 = \beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

Escribiendo de otra forma la hipótesis se obtiene lo siguiente:

$$\begin{cases} \mathrm{H}_0 : \beta_1 - 3\beta_2 = 0;\ 6\beta_3 - \beta_4 = 0;\ \beta_1 - \beta_5 = 0 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

Que matricialmente sería:

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

Con $\mathbf{L}$ dada por

$$L = \begin{bmatrix} 0 & 1 & -3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}$$

2pt

Para obtener el módelo reducido operamos:

$$Y_i = \beta_0 + 3\beta_2 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + 6\beta_3 X_{4i} + 3\beta_2 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

Que si se quiere ver más simplificada sería:

$$Y_i = \beta_o + \beta_2 X^*_{1i,2i,5i} + \beta_3 X^*_{3i,4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

1pt

Donde $X^*_{1i,2i,5i} = 3X_{1i} + X_{2i} + 3X_{5i}$ y $X^*_{3i,4i} = X_{3i} + 6X_{4i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

2pt

$$F_0 = \frac{(SSE(MR) - SSE(MF))/3}{MSE(MF)} \overset{H_0}{\sim} f_{3,68} \tag{4}$$

Ahora reemplazando con los valores de SSE(MF) y el MSE(MF) que conocemos:

$$F_0 = \frac{(SSE(MR) - 74.0051)/3}{1.08831} \overset{H_0}{\sim} f_{3,68} \tag{5}$$

4. Pregunta 4

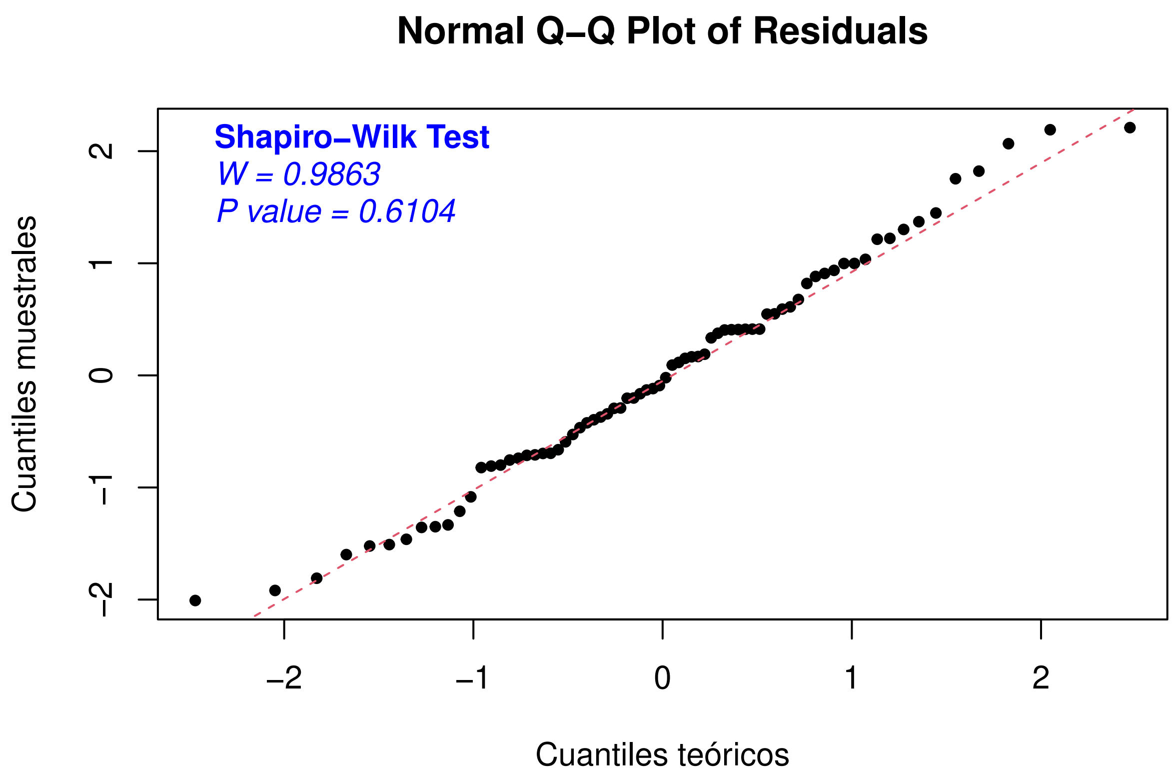
18 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

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

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



4 pt

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Se evidencia que se tiene un valor p aproximadamente a 0.6104 y asumiendo un nivel de significancia del $\alpha = 0.05$, no se rechaza la hipótesis nula, por lo que, se concluye que el supuesto de distribución de los datos es normal con media µ y varianza σ^2, debido a que el valor de p es mucho mayor. Más importante que ésta prueba analítica, en el gráfico Cuantil-Cuantil se puede observar que hay una falta de ajuste de los residuales, no existen colas pesadas pero si patrones irregulares; aunque los datos tratan de seguir una tendencia lineal se opta por rechazar el supuesto de normalidad

4.1.2. Varianza constante

Para validar el supuesto de varianza constante analizaremos el gráfico de los residuales estudentizados vs los valores ajustados:

$$\begin{cases} \mathrm{H}_0 : V[\varepsilon_i] = \sigma^2 \\ \mathrm{H}_1 : V[\varepsilon_i] \neq \sigma^2 \end{cases}$$

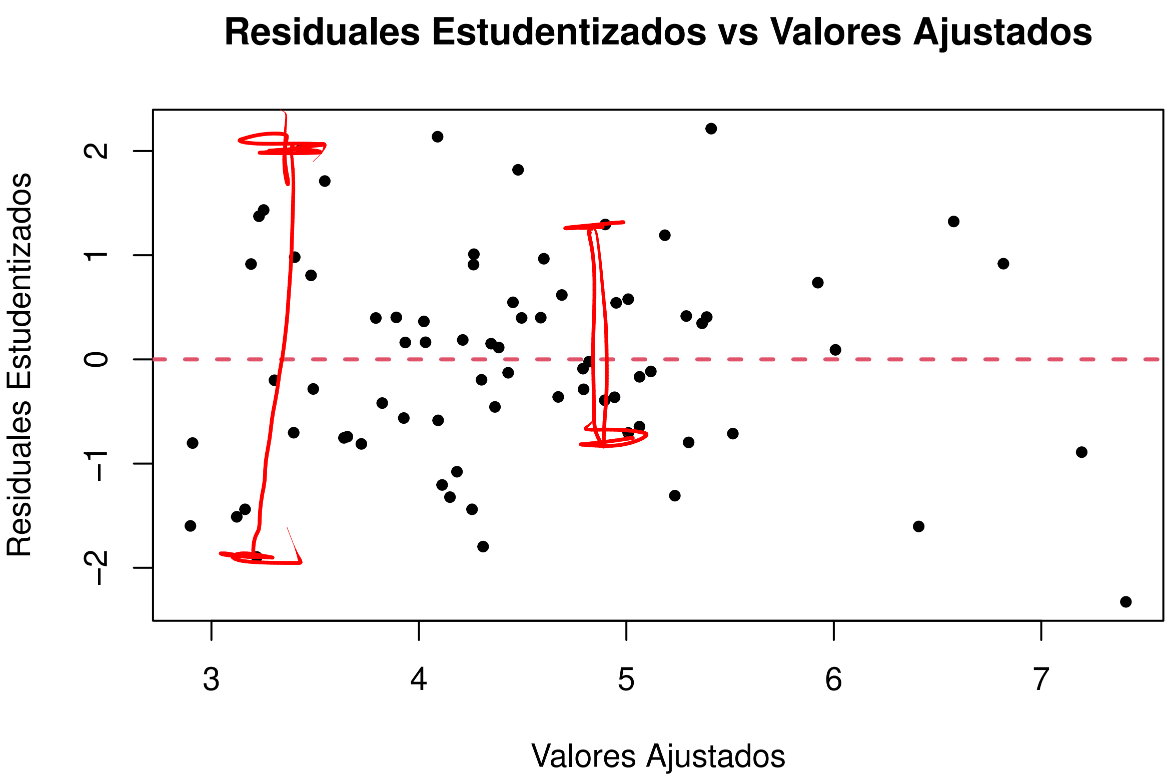


Figura 2: Gráfico residuales estudentizados vs valores ajustados

Sólo se ve con e_i

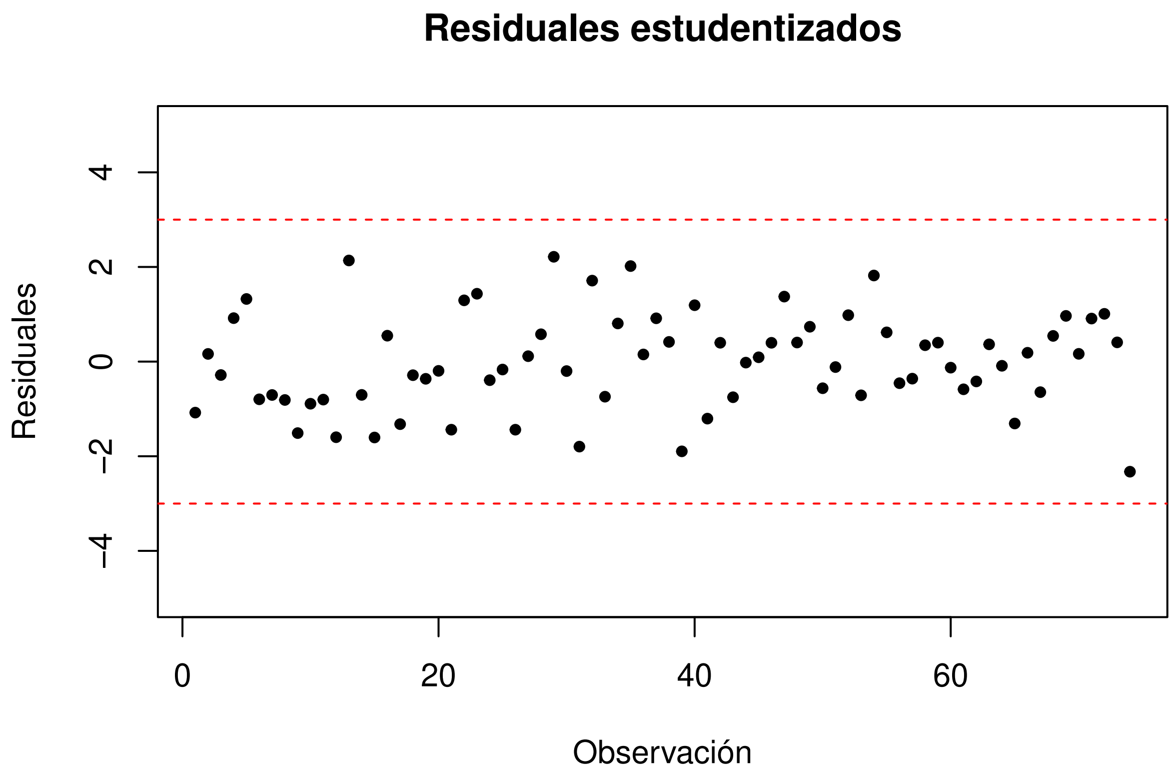
Se observa de la gráfica de residuales estudentizados vs valores ajustados que no se tiene suficiente información que nos permita concluir una varianza constante o no constante, ya que, no se evidencia una dispersión creciente o decreciente de los puntos a medida que aumente la variable independiente, pero si se observa media 0. Sin embargo, si se traza una línea recta por encima y por debajo de los datos, finalmente se puede ver que si hay varianza constante, es decir, que se acepta este supuesto.

Sí hay patrón

2pt

4.2. Verificación de las observaciones

4.2.1. Datos atípicos



3pt

Figura 3: Identificación de datos atípicos

Se observa que el conjunto de datos del modelo de regresión lineal múltiple, no evidencia algún dato atípico, esto porque ningún residual estudentizado sobrepasa el criterio correspondiente $|r_{estudentizados}| > 3$, es decir, ninguna de las observaciones se encuentra por encima de 3 o por debajo de -3.

4.2.2. Puntos de balanceo

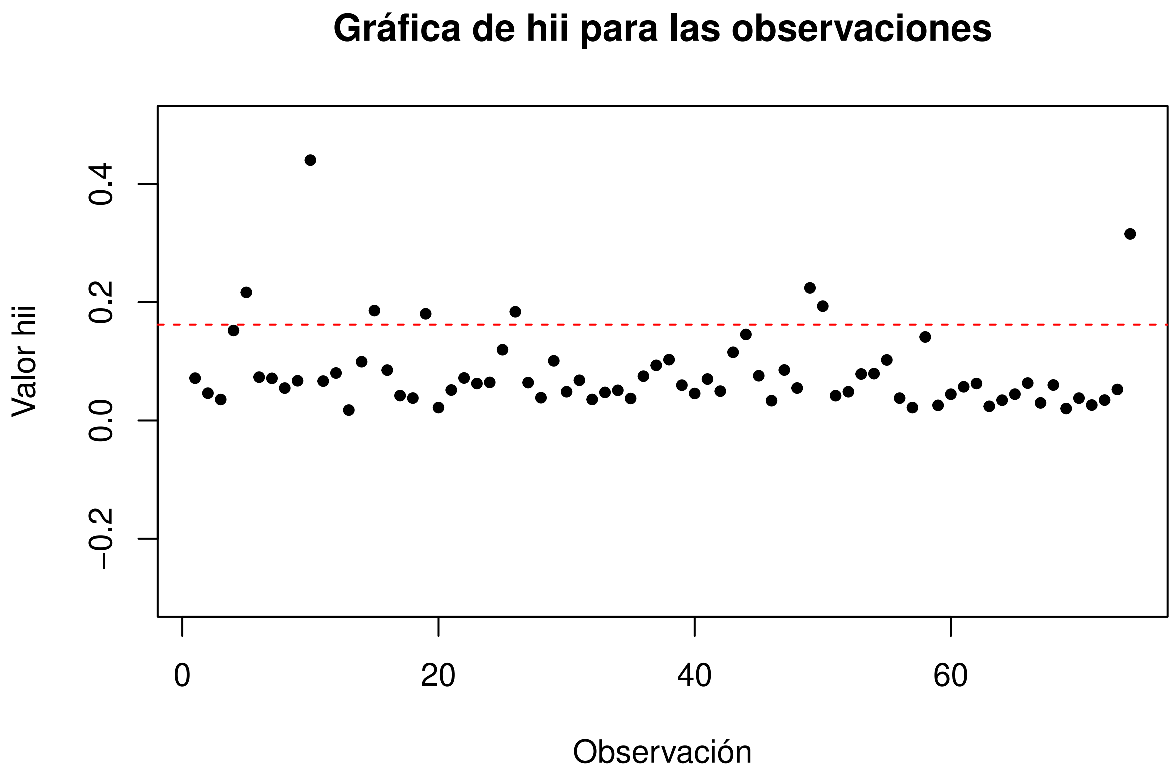


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 5    1.3235  0.0807    0.2167  0.7000
## 10  -0.8905  0.1040    0.4404 -0.7887
## 15  -1.6035  0.0978    0.1859 -0.7753
## 19  -0.3635  0.0049    0.1805 -0.1695
## 26  -1.4390  0.0778    0.1839 -0.6886
## 49   0.7367  0.0262    0.2243  0.3948
## 50  -0.5628  0.0127    0.1934 -0.2741
## 74  -2.3269  0.4162    0.3156 -1.6351
```

347

Con base en el gráfico, se evidencian 8 puntos de balanceo, teniendo en cuenta que la línea roja que está discontinua es representada como el valor $h_{ii} = 2\frac{p}{n} = 2\frac{6}{74} = 0.16$, siendo p el número de parámetros del modelo y n el número de observaciones. Entonces estos puntos de balanceo se deben a que su valor h_{ii} es mayor que 0.16, lo que indica un alejamiento importante del centro del espacio definido por las variables predictoras.

4.2.3. Puntos influenciales

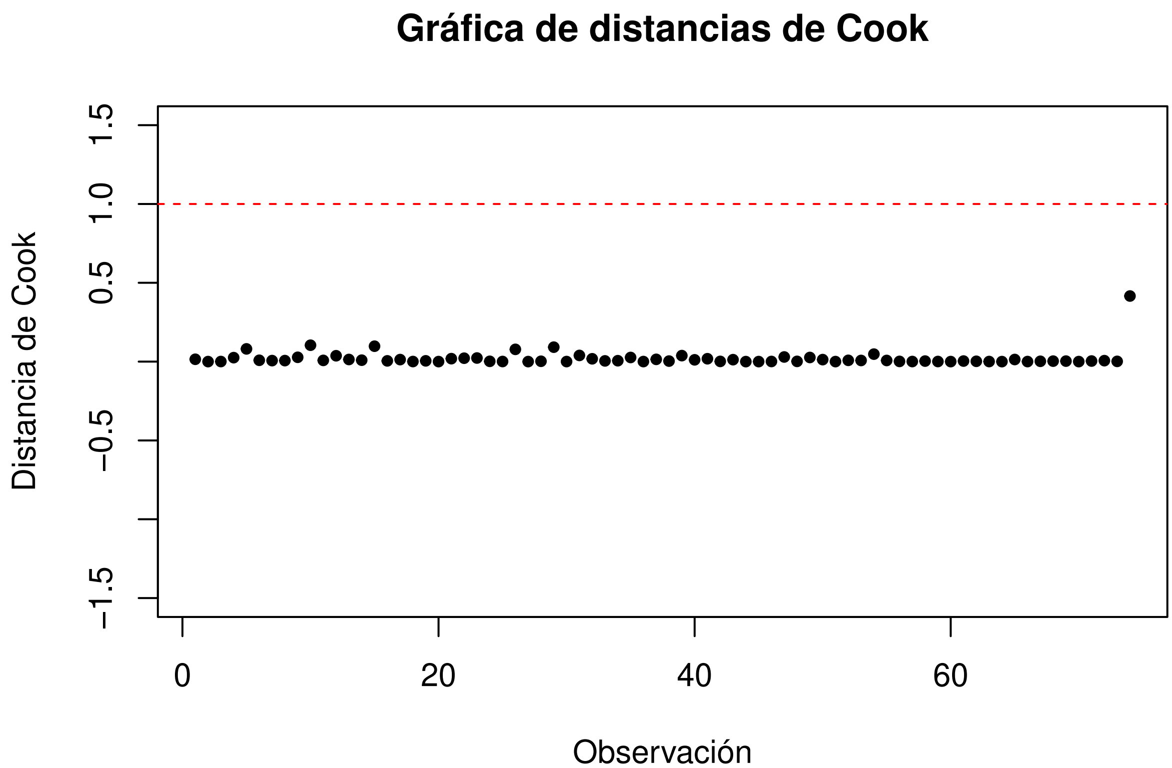


Figura 5: Criterio distancias de Cook para puntos influenciales

Se observa que no hay presencia de ningún punto de influencia porque ninguno de los valores de distancia de Cook asociados a los datos supera al valor de 1.

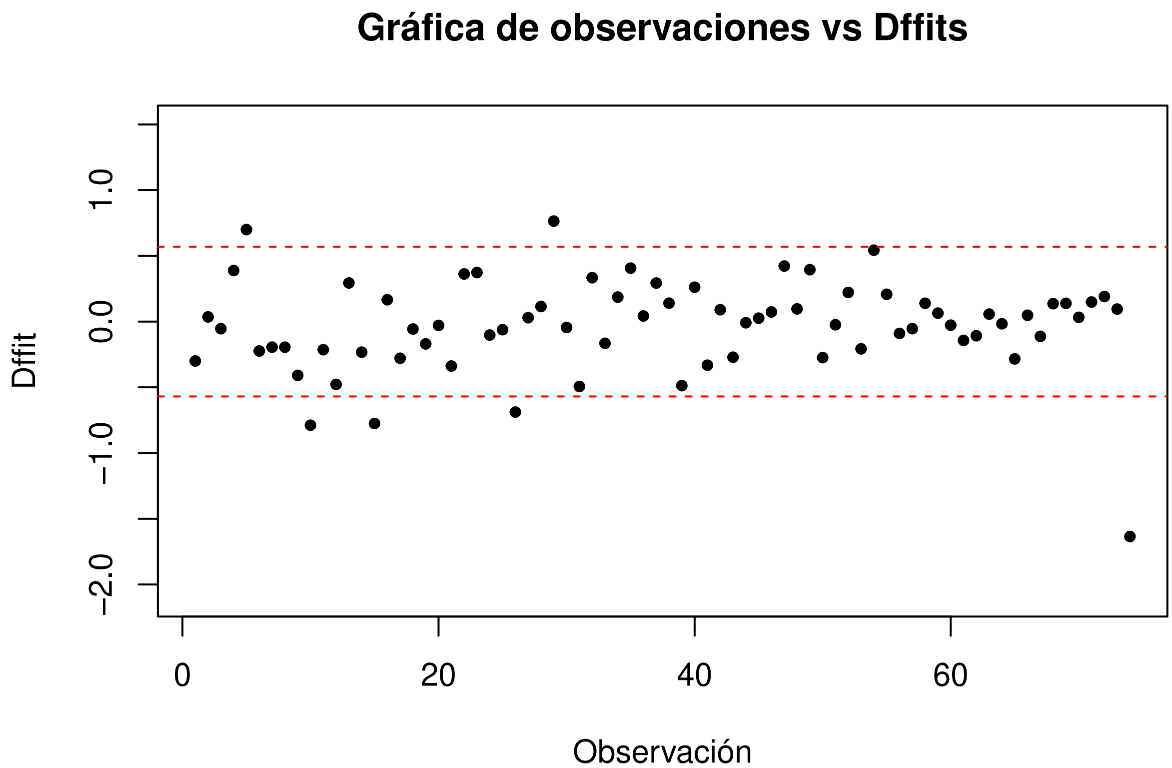


Figura 6: Criterio Dffits para puntos influenciales

Causan...?

##	res.stud	Cooks.D	hii.value	Dffits
## 5	1.3235	0.0807	0.2167	0.7000
## 10	-0.8905	0.1040	0.4404	-0.7887
## 15	-1.6035	0.0978	0.1859	-0.7753
## 26	-1.4390	0.0778	0.1839	-0.6886
## 29	2.2148	0.0917	0.1009	0.7645
## 74	-2.3269	0.4162	0.3156	-1.6351

3pt

Tal como nos lo muestra la gráfica y la tabla, tenemos 6 puntos influenciales, provenientes de las observaciones 5, 10, 15, 26, 29 y 74, que cumplen el criterio del diágnostico DFFITS el cual establece que es un punto influencial si $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$. 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, ninguno de los datos cumple con serlo.

4.3. Conclusión

3pt

Se llega a la conclusión que el modelo propuesto no es válido, pues no se cumple el supuesto de normalidad, aunque si el de varianza constante y media cero.

Además, tampoco se encontraron datos atipicos, y es importante recalcar que al tener 8 puntos de balanceo y 6 datos de influencia puede existir una afectación en la validación de los supuestos para este modelo de regresión lineal múltiple. Por consiguiente, es necesario analizar estas observaciones individualmente, y realizar el análisis respectivo que permita explicar el por qué ocurre esto.