Trabajo 1

Estudiantes

Andres Mauricio Paez Martinez
Andres Felipe Devia Orrego
Manuela Ferrer Cuervo
Juan Esteban Vasquez Mesa

Equipo #3

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

Teniendo en cuenta la base de datos 03, en la cual hay 5 variables regresoras, denominadas por:

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_5: Número de enfermeras.

Entonces, se plantea el modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 45$$

1.1. Modelo de regresión

Al hacer el ajuste del modelo, se obtuvieron los siguientes coeficientes:

Cuadro 1: Tabla de valores de coeficientes del modelo.

	Valor del parámetro.
β_0	-0.0905
β_1	0.1919
β_2	0.0090
β_3	0.0520
β_4	0.0118
β_5	0.0012

Entonces, el modelo de regresión ajustado es el siguiente:

$$\hat{Y}_i = -0.0905 + 0.1919 X_{1i} + 0.009 X_{2i} + 0.052 X_{3i} + 0.0118 X_{4i} + 0.0012 X_{5i} + \varepsilon_i; \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 45$$

1.2. Significancia de la regresión

Ahora, para analizar la significancia de la regresión, se plantean las siguientes hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2, 3, 4, 5.} \end{cases}$$

Esto nos da un estadístico de prueba:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,39} \tag{1}$$

done la tabla Anova de la regresion es la siguiente:

Cuadro 2: Tabla ANOVA para el modelo

	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Vp
Regresión	40.1227	5	8.024534	9.19656	7.6235e-06
Error	34.0298	39	0.872558		

Ahora, a partir de los resultados de la tabla Anova, podemos identificar un Valor P muy cercano a 0, por lo tanto rechazamos la hipotesis nula H_0 en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, y se acepta H_1 en la que algún $\beta_j \neq 0$, a partir de esto podemos concluir que la regresión es significativa.

1.3. Significancia de los parámetros

Para el cálculo de la significancia individual de los parámetros, se tiene el siguiente juego de hipótesis:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2, 3, 4, 5.} \end{cases}$$

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar cuáles de ellos son significativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$Se(\hat{\beta}_j)$	T_{0j}	Vp
β_0	-0.0905	1.8742	-0.0483	0.9617
β_1	0.1919	0.0797	2.4076	0.0209
β_2	0.0090	0.0324	0.2765	0.7836
β_3	0.0520	0.0161	3.2257	0.0025
β_4	0.0118	0.0086	1.3732	0.1775
β_5	0.0012	0.0010	1.2620	0.2145

Los Vp mostrados en la tabla nos permiten concluir si un parametro es significativo cuando su Vp$<\alpha$ (Rechazando la hipotesis nula), Por lo tanto, podemos concluir con un nivel de significancia $\alpha = 0.05$, que los parámetros β_1 y β_3 Son significativos, ya que estamos

rechazando H_0, porque sus Vp son menores a $\alpha = 0.05$. Por otra parte se acepta H_0 para los coeficientes β_0, β_2, β_4, β_5, ya que Vp>α, por tanto no se puede rechazar la hipotesis nula y se interpretan como no significativos.

1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** Indica, por cada aumento de unidad en la variable X_1 (duración de la estadía), el promedio de riesgo de infección aumenta en 0.1919 unidades cuando las demás predictoras permanecen constantes.

$\hat{\beta}_3$**:** Indica, por cada aumento de unidad en la variable X_3(número de camas), el promedio de riesgo de infección aumenta en 0.0520 unidades cuando las demás predictoras permanecen constantes.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5411$, lo que significa que aproximadamente el 54.11 % de la variabilidad total del riesgo de infección observada en la respuesta es explicada por el modelo de regresión múltiple propuesto en el presente informe. Notese tambien que $R^2_{Aj} = 0.4822$, si comparamos se puede observar que $R^2_{Aj} < R^2$, esto nos indica que en el modelo existen variables que no aportan significativamente al modelo, como se pudo comprobar anteriormente.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el Valor P más alto en el modelo fueron X_2, X_4, X_5, por lo tanto a través de la tabla de todas las regresiones posibles planteamos el siguiente juego de hipotesispara probar la significancia simultánea del subconjunto:

$$\begin{cases} \text{H}_0 : \beta_2 = \beta_4 = \beta_5 = 0 \\ \text{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	Suma Cuadratica Del Error	Covariables
Modelo completo	34.030	X1 X2 X3 X4 X5
Modelo reducido	37.126	X1 X3

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,39} \\ &= \frac{37.126 - 34.030}{0.872558} \\ &= 3.54816 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,39} = 2.8451$, se puede ver que $F_0 > f_{0.95,1,45}$.Entonces el subconjunto es significativo.

A partir de esto podemos Rechazar H_0 y dar como cierta la hipotesis alternativa.Concluyendo asi que Y depende de almenos una de las variables asociadas a los parametros del subconjunto estudiado y no es posible descartarlas.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta mediante el planteamiento de que no debe haber significancias por consiguiente se plantea la siguiente prueba de hipótesis igualando los parametros:

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

reescribiendo matricialmente:

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

donde la matriz $\mathbf{L}$ segun lo anterior esta dada por

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

El modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_1 X^{2}_{2i} + \beta_3 X^{3}_{3i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 45$$

Donde $X^{1i} = X1i + X_{3i}$ y $X^{2i} = X2i + X_{4i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,39} \tag{3}$$

$$F_0 = \frac{(SSE(MR) - (34.030))/2}{0.872558)} \overset{H_0}{\sim} f_{2,39} \tag{4}$$

4. Pregunta 4

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 shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

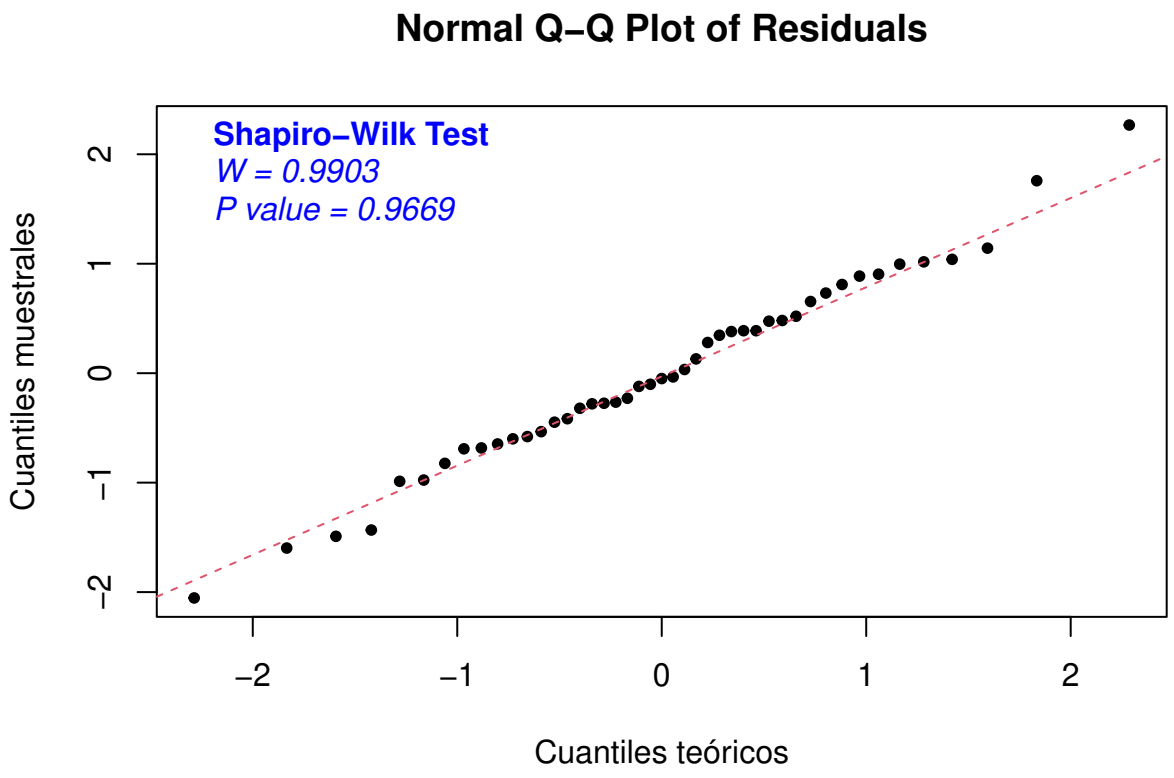

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Dado que el valor de Shapiro-Wilk es 0.9903 y el P-valor es del 99 %, podemos concluir que no hay suficiente evidencia para rechazar la hipótesis nula de que los datos siguen una distribución normal, es decir es decir que los datos se distribuyen normal con media μ y varianza σ^2.La línea gráfica también se puede ver que los puntos se alinean bien,Sin embargo, es importante tener que nos muestra patrones irregulares, lo que indica que puede ser necesario considerar otras distribuciones o realizar pruebas adicionales para evaluar si los datos siguen una distribución normal; Además, es importante validar si la varianza cumple con el supuesto de ser constante para asegurar la precisión de los resultados del análisis.

Excelente análisis!

4.1.2. Varianza constante

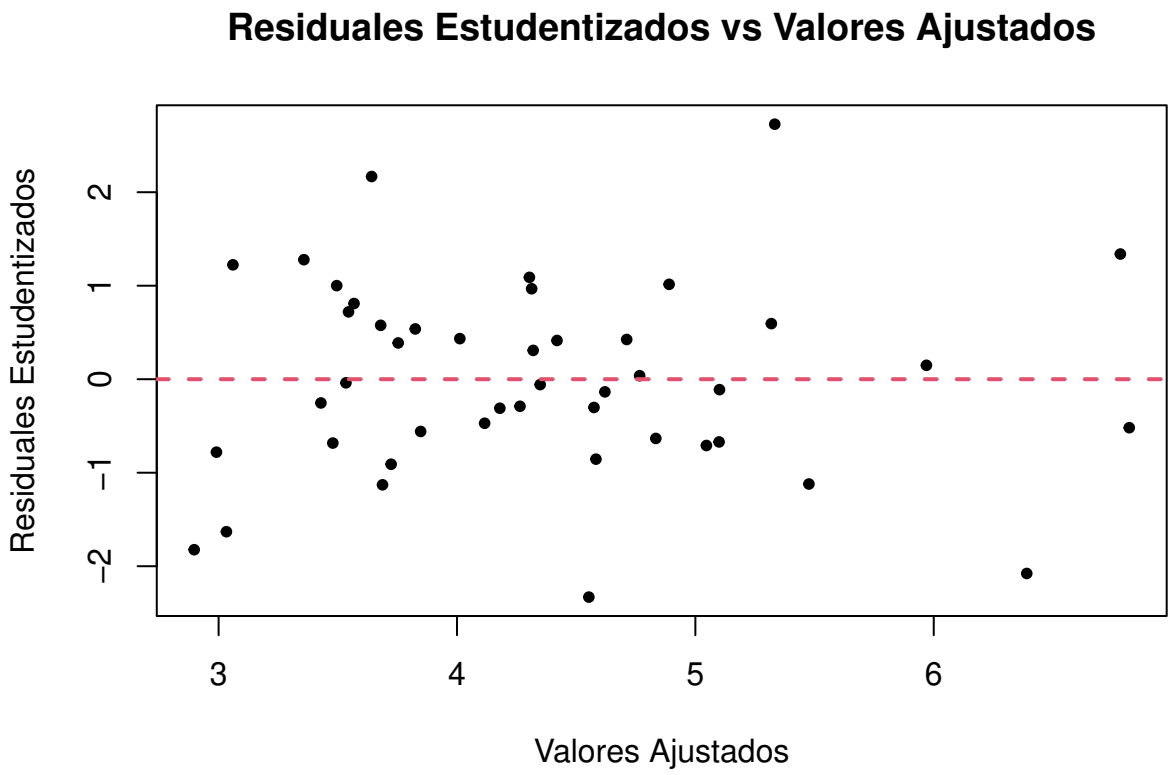

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el análisis de los residuales estudentizados vs valores ajustados se puede concluir que no hay patrones claros en los que la varianza aumente o disminuya, lo que sugiere que es plausible asumir que la varianza es constante. Además, el hecho de que la media de los residuales estudentizados esté cerca de cero es una señal positiva de que el modelo se ajusta adecuadamente a los datos. En general, al no haber suficiente evidencia para rechazar el supuesto de varianza constante, se puede aceptar como válido, lo que aumenta la confianza en los resultados del análisis.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

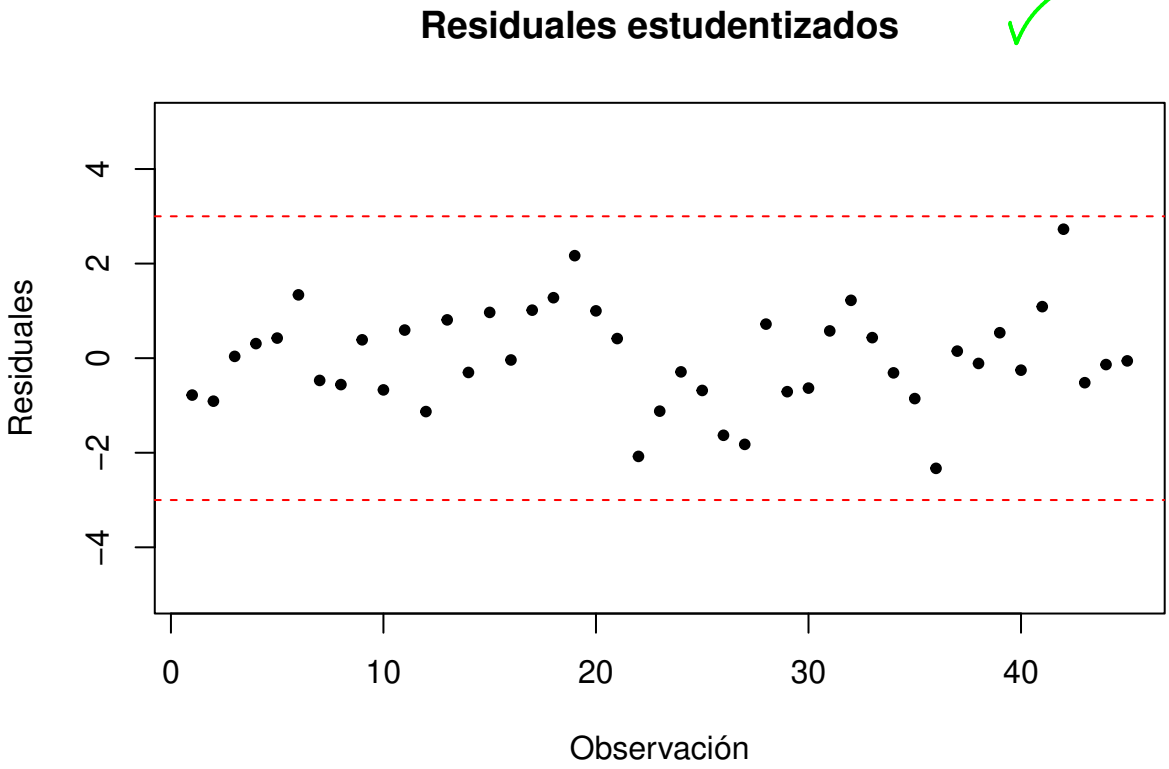

Figura 3: Identificación de datos atípicos

La gráfica anterior indica que no hay valores extremos o inusuales (atipicos) en los datos porque ninguno de los residuos estudiantizados supera $|r_{estud}| > 3$ o queda por debajo $|r_{estud}| < 3$ que son el umbral establecido. Esto sugiere que los datos son coherentes y no hay valores que se desvíen significativamente de la tendencia general.

4.2.2. Puntos de balanceo

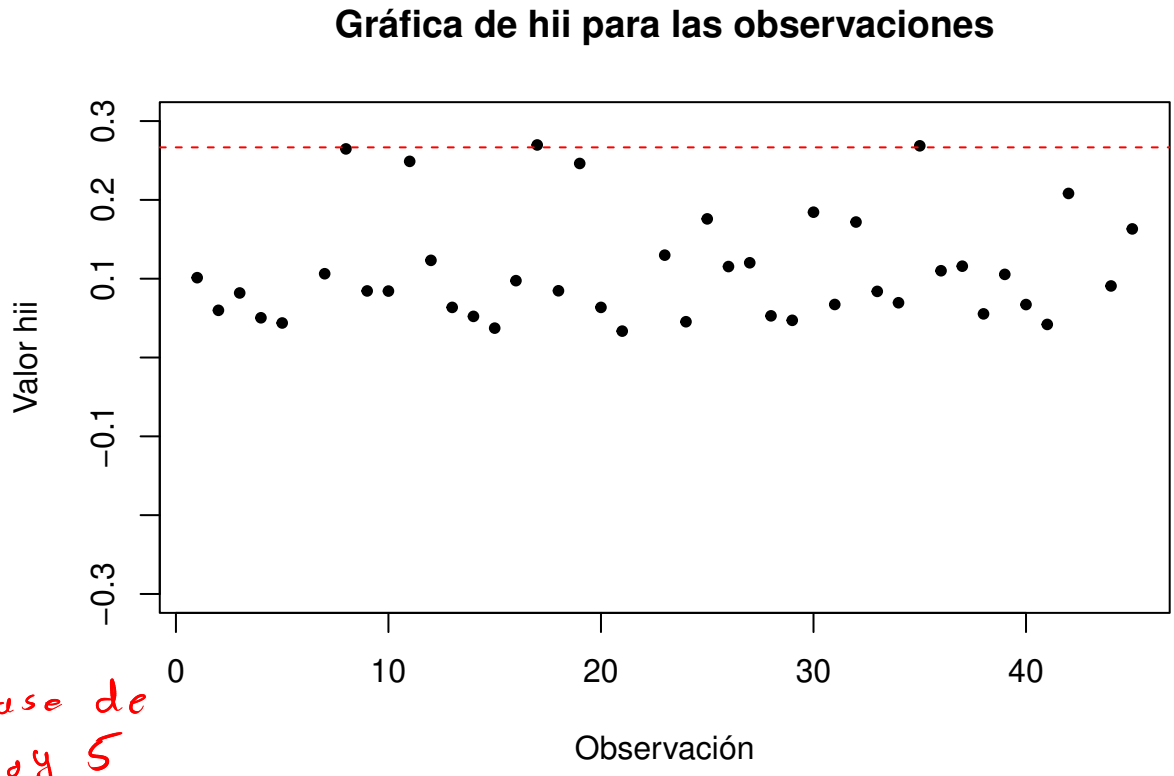

Figura 4: Identificación de puntos de balanceo

Al observar la gráfica h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, Se puede evidenciar la presencia de 3 puntos de balanceo en la gráfica, lo que puede ser indicativo de que hay tres observaciones que están afectando significativamente el modelo. Estos 3 puntos son considerados como puntos de balanceo, ya que si se eliminan, la influencia en el modelo disminuiría significativamente, permitiendo obtener una mejor ajuste y predicciones más precisas.

1,5 pt

4.2.3. Puntos influenciales

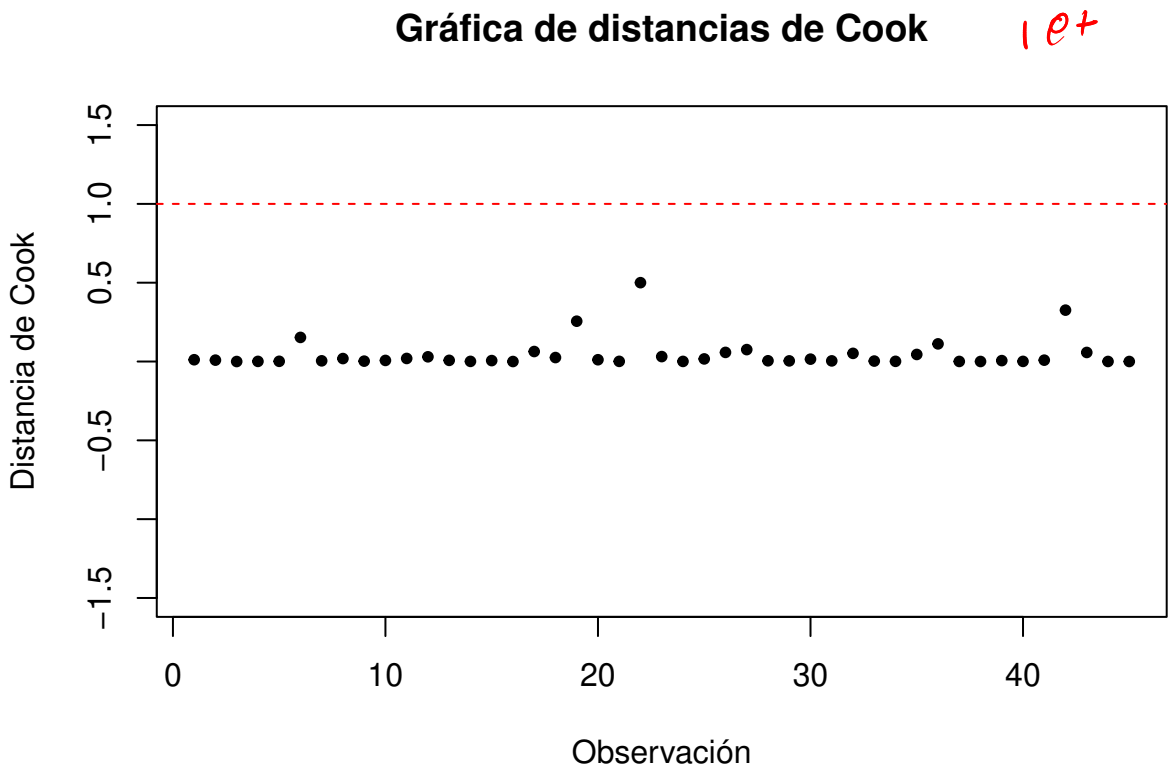

Figura 5: Criterio distancias de Cook para puntos influenciales

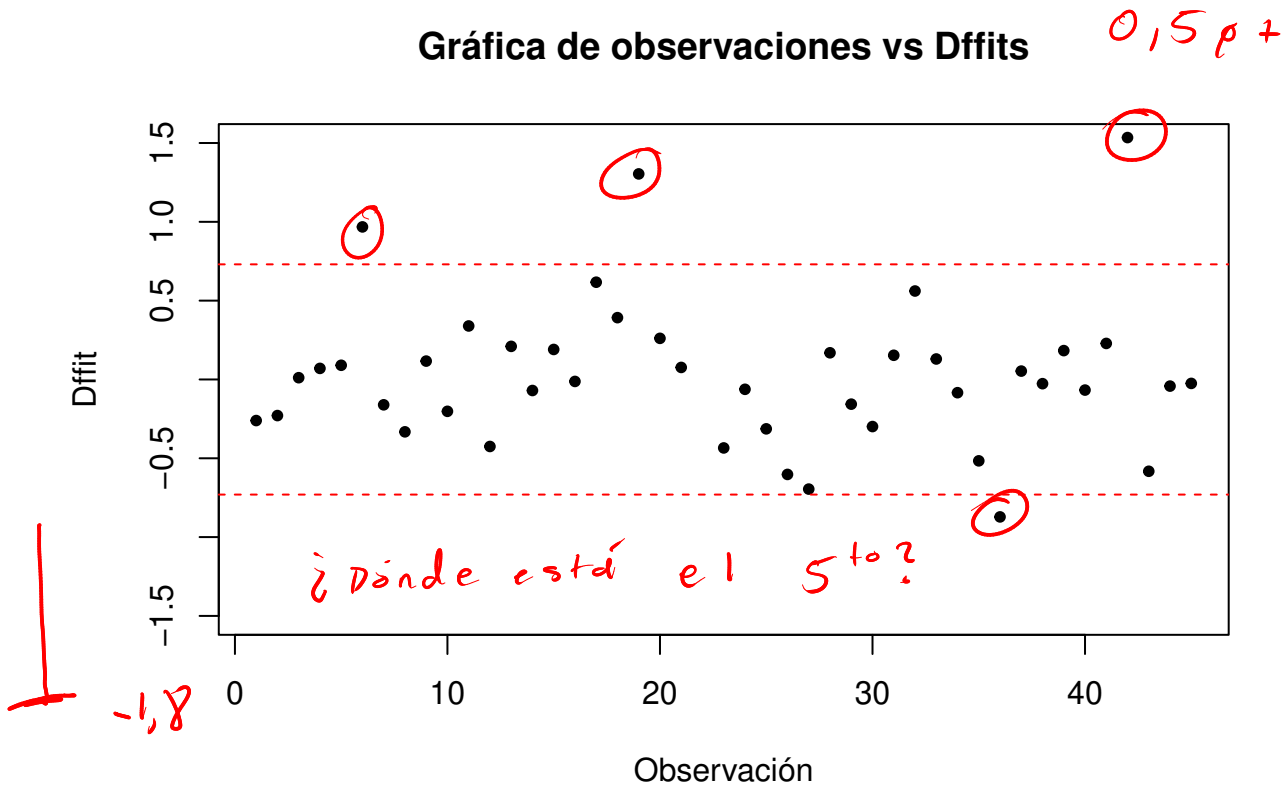

Figura 6: Criterio Dffits para puntos influenciales

```
##     res.stud Cooks.D hii.value  Dffits
## 6     1.3387  0.1529    0.3386  0.9679
## 19    2.1679  0.2558    0.2462  1.3040
## 22   -2.0774  0.5003    0.4102 -1.8134
## 36   -2.3301  0.1120    0.1101 -0.8720
## 42    2.7273  0.3260    0.2082  1.5346
```

Dada la muestra obtenida en la gráfica de Observación vs DFFITS, se puede concluir que hay 5 puntos de influencia, uno muy cerca a la frontera y otros cuatro mas alejados; el que esta cerca a la frontera ~~puede ser el que mas influencia tenga en el modelo,~~ es decir que al excluirlo altere los resultados ajustados ya planteados por el modelo de regresión.

Como se dijo anteriormente, son 5 los puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial, pues tendrian un gran efecto en la forma en que se ajusta el modelo y podria afectar significativamente los resultados.

4.3. Conclusión

Basándonos en el análisis exhaustivo presentado en los distintos puntos del documento, se puede concluir que el modelo de regresión lineal múltiple cumple con los supuestos necesarios y, por lo tanto, es válido. Sin embargo, es importante tener en cuenta algunos aspectos que pueden afectar el modelo, como los puntos influenciables. Además, se destaca la relevancia

de las estimaciones presentadas en la otras secciones, las cuales indican la significancia del modelo y la posibilidad de reducirlo eliminando parámetros insignificantes.