Trabajo 1

Estudiantes

Alejandro Montoya Vargas
Juan Pablo Castaño Uribe
Johan Estiven Martinez López
Luisa Fernanda Mazo Perez

Equipo 29

Docente

Mateo Ochoa Medina

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

Donde:

- Y: Riesgo de infección (en porcentaje).
- X_1: Duración de la estadía (en dias).
- X_2: Rutina de cultivos (por cada 100)
- X_3: Número de camas (camas)
- X_4: Censo promedio diario (pacientes)
- X_5: Número de enfermeras (enfermeras)

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.5866
β_1	0.0712
β_2	0.0298
β_3	0.0731
β_4	0.0130
β_5	0.0020

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.5866 + 0.0712 X_{1i} + 0.0298 X_{2i} + 0.0731 X_{3i} + 0.013 X_{4i} + 0.002 X_{5i}$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,63} \tag{1}$$

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	77.0901	5	15.418012	17.0595	1.23617e-10
Error	56.9381	63	0.903779		

De la tabla Anova, se observa un valor P aproximadamente igual a 0, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto el modelo de RLM propuesto es significativo. Esto quiere decir que el riesgo de infeccion depende significativamente de al menos una de las predictoras del modelo.

1.3. Significancia de los parámetros

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}	P-valor
β_0	-0.5866	1.5892	-0.3691	0.7133
β_1	0.0712	0.0719	0.9902	0.3259
β_2	0.0298	0.0295	1.0112	0.3158
β_3	0.0731	0.0162	4.5124	0.0000
β_4	0.0130	0.0072	1.7981	0.0769
β_5	0.0020	0.0007	2.9218	0.0048

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_3 y β_5 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

Solo interpretaremos los significativos.

$\hat{\beta}_3$**:** Indica que por cada unidad que se aumente el número de camas en el hospital (X_3), la probabilidad promedio de adquirir una infección en el hospital aumenta en 0.0731 unidades, cuando las demas predictoras se mantienen constantes

$\hat{\beta}_5$**:** Indica que por cada unidad que se aumente en el número de enfermeras (X_5), la probabilidad promedio de adquirir una infección en el hospital aumenta en 0.0020 unidades, cuando las demas predictoras se mantienen constantes

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5751783$, lo que significa que aproximadamente el 57.5 % de la variabilidad total observada en el riesgo de infección es explicada por el modelo de regresión multiple propuesto en el presente informe.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más bajo en el modelo fueron X_3, X_4, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 3, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	56.938	X1 X2 X3 X4 X5
Modelo reducido	101.693	X1 X2

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \ \cdots, \ \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{14.9183333}{0.9037778} \\ &= 16.5066388 \end{aligned} \tag{2}$$

Para el criterio de decision se requiere obtener el valor critico de una distribucion $f_{3,63}$ a un nivel de significancia $\alpha = 0.05$, esto es, $f_{0.05,3,63} = 0.1167$, se puede ver que $F_0 > f_{0.05,3,63}$ (0.1167) por tanto se rechaza la hipotesís nula H_0, y se concluye que el riesgo de infección depende de al menos una de las variables del subconjunto $\{B_3, \ B_4, \ B_5\}$, por lo tanto, no es posible descartar estas variables.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Preguntas: (compruebe si estas suceden a la vez).

- ¿El efecto del número de enfermeras (X_5) sobre el riesgo de infección, es igual a 5 veces el efecto de la rutina de cultivos (X_2) sobre el riesgo de infección.?
- ¿El efecto de la duración de la estadia (X_1) sobre el riesgo de infección es igual al efecto del censo promedio diario (X_4) sobre el riesgo de infección.?
- ¿El efecto del número de camas (X_3) sobre el riesgo de infección es 2 veces el efecto de la duración de la estadia (X_1) sobre el riesgo de infección.?

$$\begin{cases} \mathrm{H}_0 : \beta_5 = 5\beta_2; \ \beta_4 = \beta_1; \ \beta_3 = 2\beta_1 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

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

Con $\mathbf{L}$ dada por

$$L = \begin{bmatrix} 0 & 0 & -5 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & -2 & 0 & 1 & 0 & 0 \end{bmatrix}$$

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_2 X^*_{2i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

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

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(RM) - SSE(FM))/3}{MSE(FM)} \overset{H_0}{\sim} f_{3,63} \tag{3}$$

$$F_0 = \frac{(SSE(RM) - 56.938)/3}{0.90377} \overset{H_0}{\sim} f_{3,63} \tag{4}$$

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

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

Se validara por medio de un gráfico cuantil-cuantil, acompañada de shapiro-wilk:

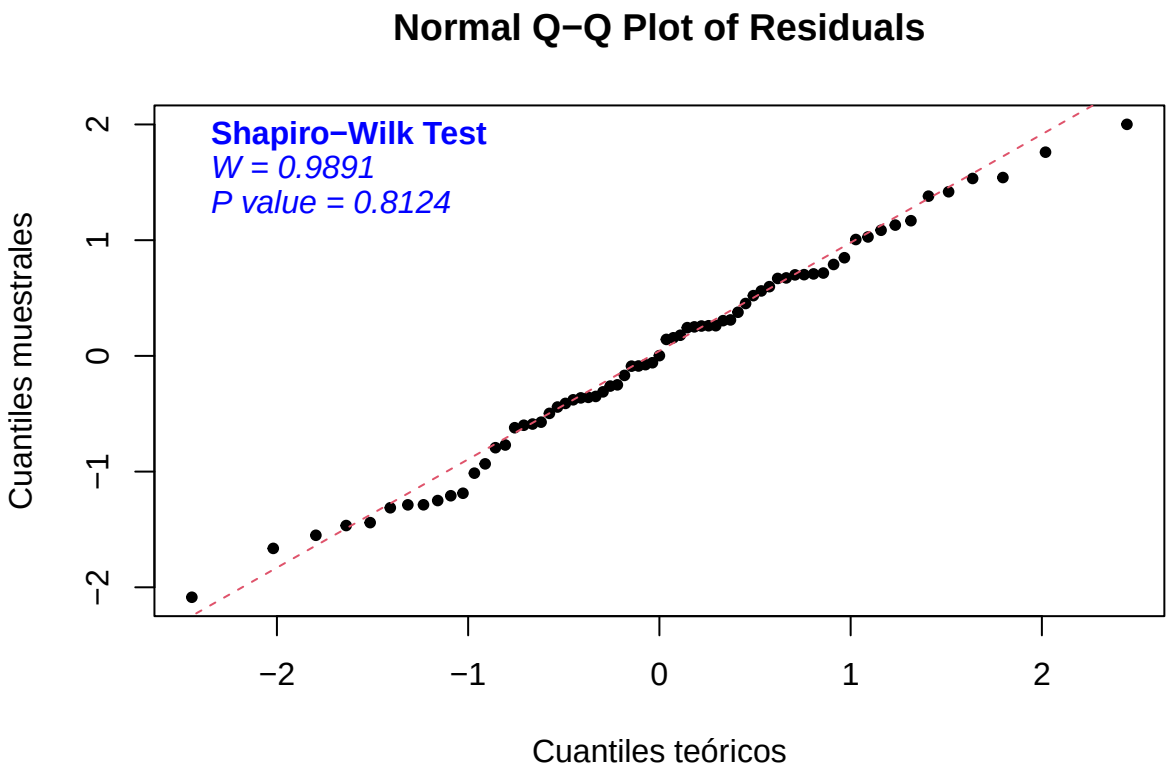

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Como se puede observar en la gráfica Q-Q Plot, se puede ver que la mayoría de las observaciones se encuentran sobre o muy cerca de la línea roja, que representa el ajuste de la distribución de los residuos a una distribución normal. Sin embargo, hay algunas observaciones en las colas (al comienzo y al final), que a pesar de que están muy cerca de la línea roja, si se les nota un poco diferentes a las demás. Estas observaciones podrían deberse a puntos extremos que presenta el modelo.

Sin embargo, viendo que la gran mayoría de los puntos si siguen la línea roja, y debido al alto p-valor de del test de Shapiro Wilk (0.8124) como respaldo, y teniendo en cuenta que el nivel de significancia $\alpha = 0.05$. Concluimos que el supuesto de normalidad en los errores se cumple.

4.1.2. Varianza constante

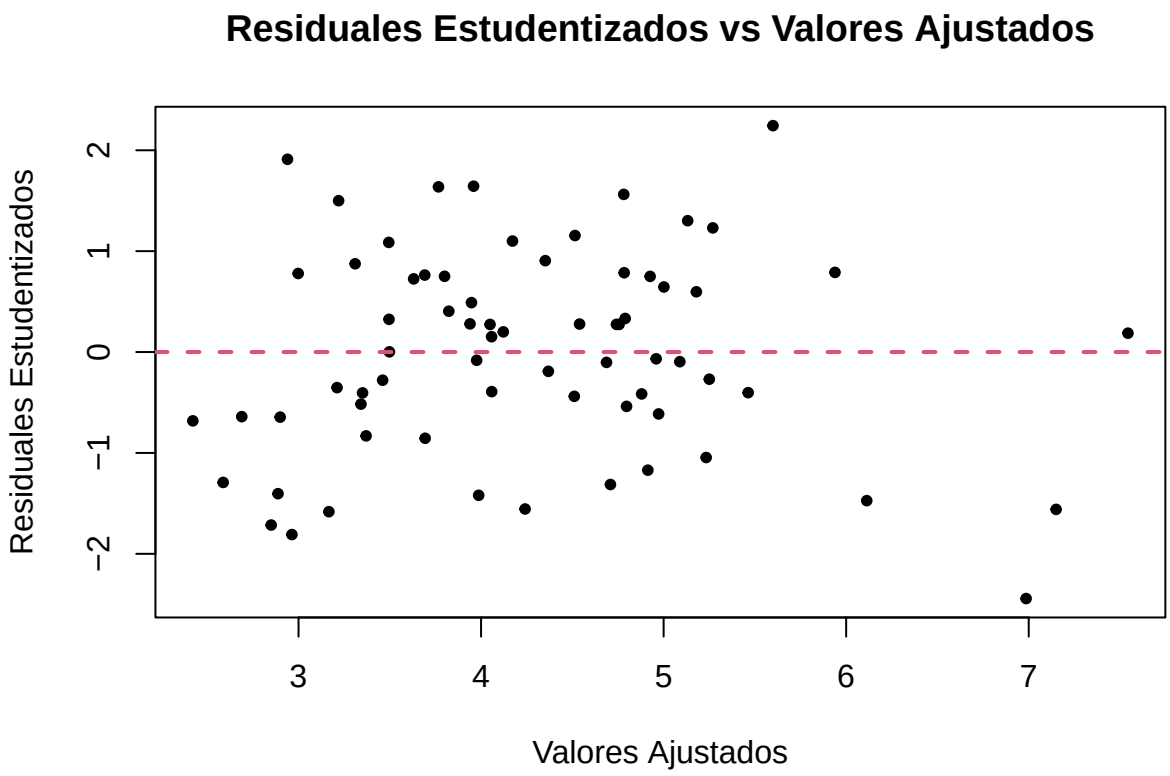

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Del gráfico de residuales estudentizados vs valores ajustados, se puede observar que en general, el supuesto de varianza constante se cumple,ya que no hay comportamientos de que decrecimiento o aumento, y por tanto, no hay forma de decir que no se cumple. Sin embargo, hay que notar que algunos puntos están un poco alejados de la nube de puntos, lo cual podría ser un indicio de la existencia de valores extremos o falta de ajuste.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

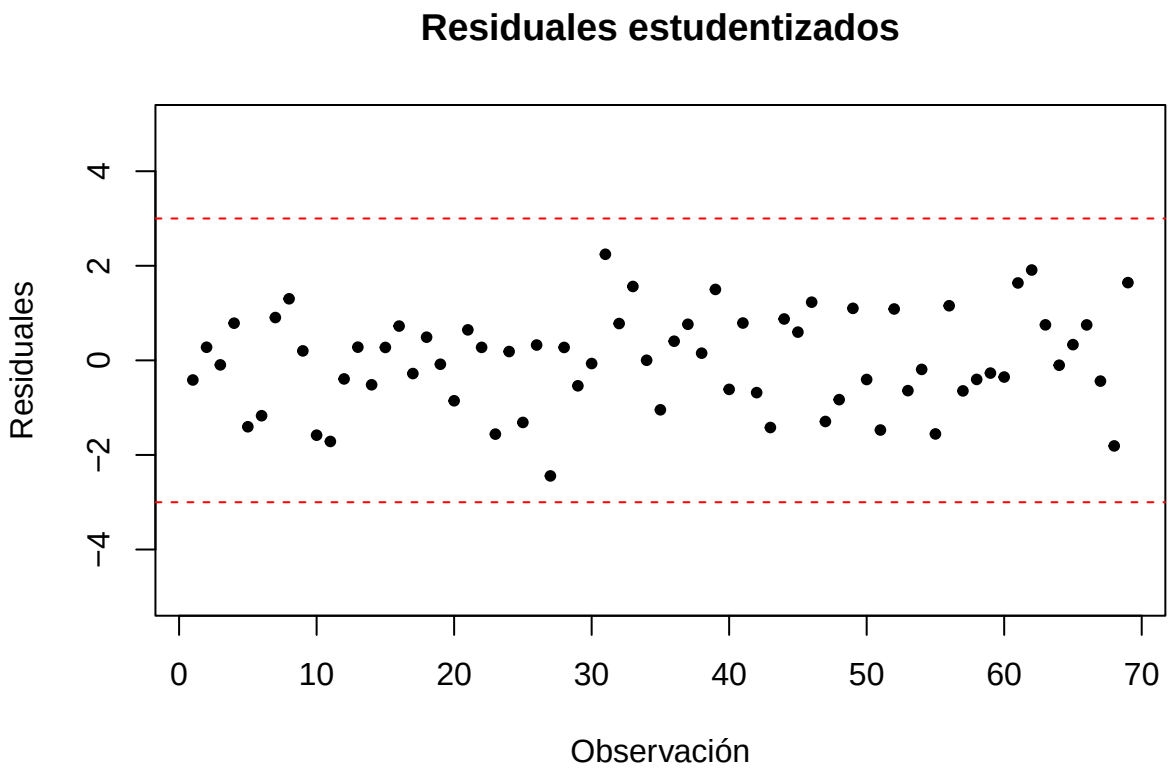

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

Esto podria indicar que el modelo es bastante estable, ya que al no haber valores atipicos, nuestros supuestos no se veran muy afectados. Sin embargo esto no indica que no se afecten en lo absoluto, pues aun podrian haber faltas de ajuste en el modelos y puntos influenciales.

4.2.2. Puntos de balanceo

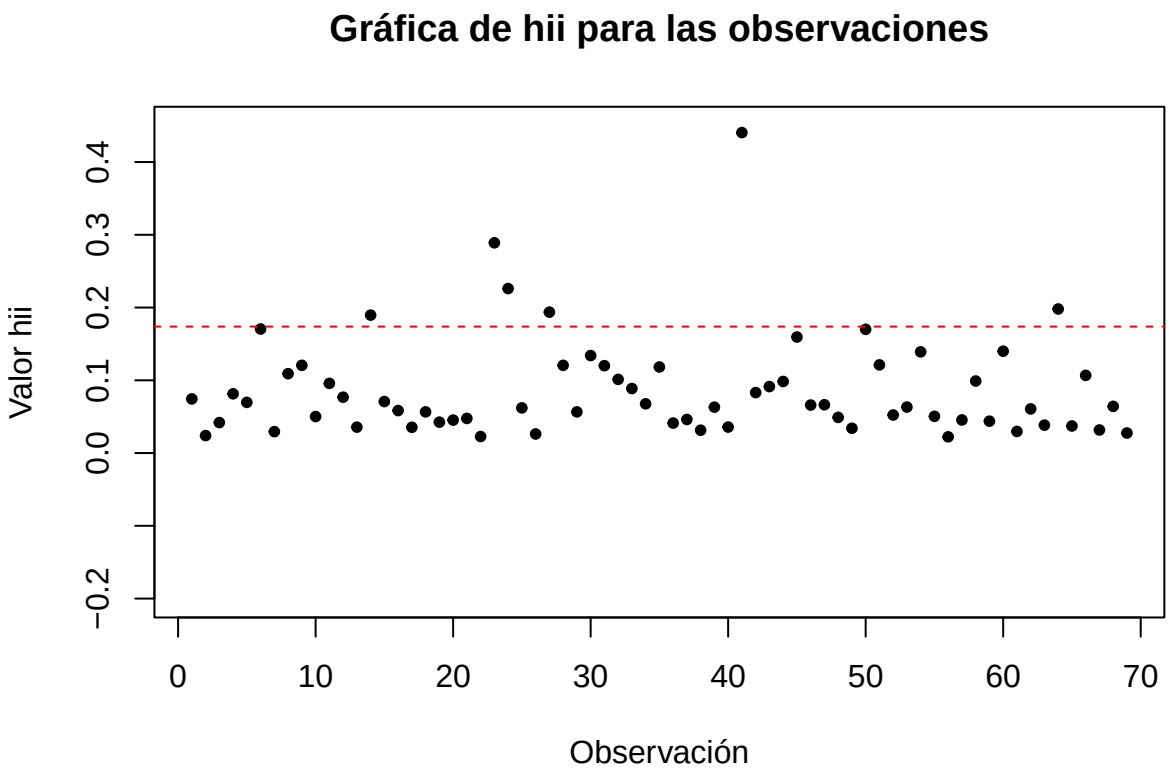

Figura 4: Identificación de puntos de balanceo

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$ (0.173913), se puede apreciar que existen 6 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$ (0.173913), los cuales son los presentados en la tabla:

Cuadro 5: Puntos de balanceo

	hii.value
14	0.1896
23	0.2890
24	0.2261
27	0.1937
41	0.4405
64	0.1980

Estos puntos podrían tener un impacto grande en los datos, interpretaciones y predicciones de nuestro modelo, como en la estimacion de nuestros parametros y los interalos de confianza o de prediccion.

4.2.3. Puntos influenciales

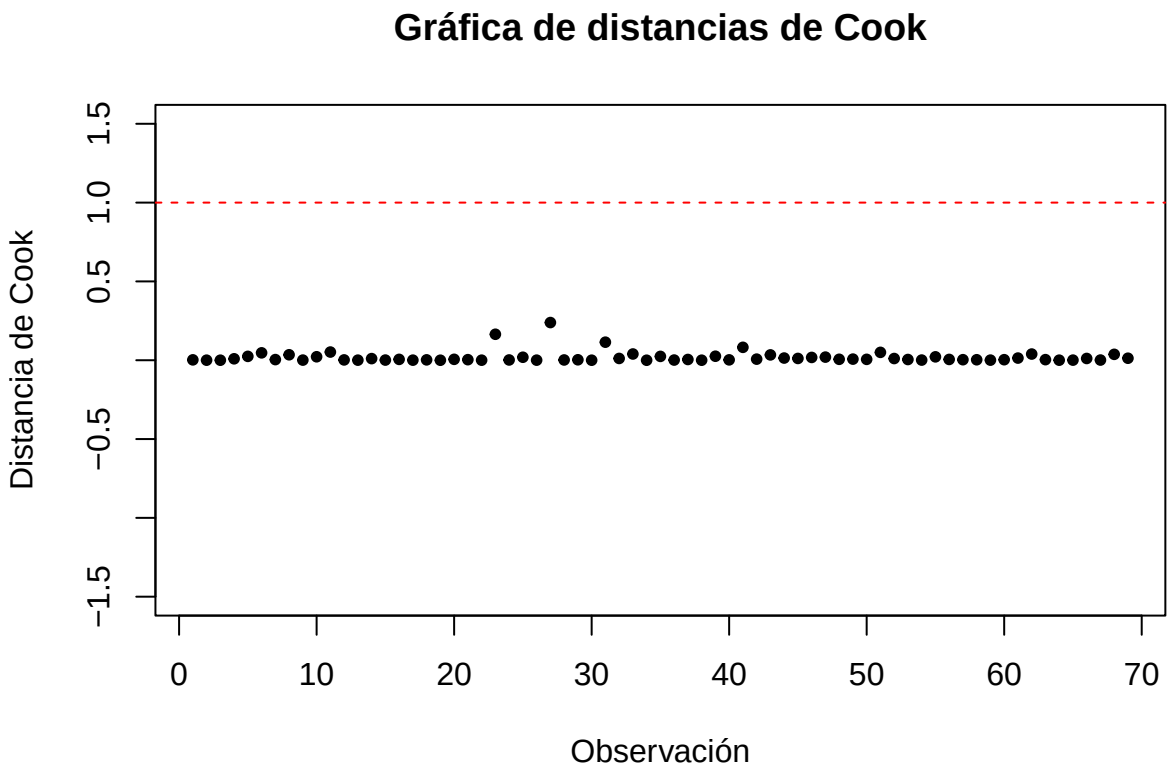

Figura 5: Criterio distancias de Cook para puntos influenciales

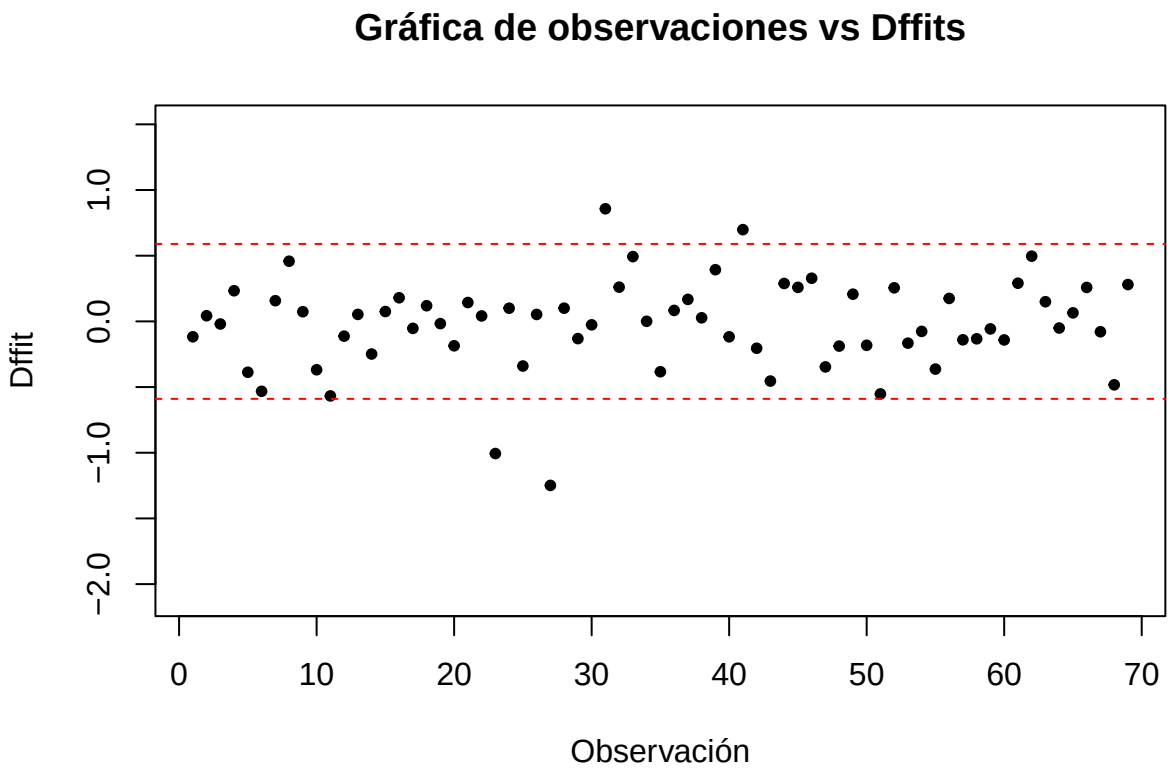

Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Puntos influenciales con Dffits

	Dffits
23	-1.0060
27	-1.2487
31	0.8574
41	0.6983

Como se puede ver,las observaciones 23, 27, 31 y 41 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffits}| > 2\sqrt{\frac{p}{n}}$ (0.5897678), 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, ninguno de los datos cumple con serlo.

Estos puntos nos pueden indicar por ejemplo que: - Cambios de significancia estadistica en nuestros B_j, lo que puede afectar las conclusiones sobre la importancia de una variable predictora en la explicacion de la varaible de respuesta.

- Predicciones y estimaciones erroneas.
- Ya que los puntos influencias tienen la capacidad de por asi decirlo “jalar” el modelo hacia ellos, la estabilidad del modelo se ve comprometida, afectando asi la confiabilidad de las conclusiones.

4.3. Conclusión

Asi como esta el modelo en este momento, nosotros no lo tomaríamos como válido, ya que a pesar de cumplir los 2 supuestos, hay muchas observaciones extremas que no han sido analizados, los cuales probablemente están cambiando el comportamiento del modelo de muchas formas:

Con respecto a la significancia del modelo, tendríamos un modelo de 5 variables predictoras, el cual solo depende de 2, por lo que valdría la pena considerar cambiar el modelo. Sin embargo, al no saber cómo están afectando exactamente esas observaciones extremas, podría ser que realmente mi respuesta dependa de más de 2 variables predictoras.

Con respecto a los supuestos, si bien es cierto que en este momento se cumplen, esto podría ser debido a, de nuevo, las observaciones extremas extremos, por lo cual no se tiene una confianza para decir que realmente si se están cumpliendo, sin antes haber analizado a fondo las observaciones extremas.

Y en general, no es posible decir con confianza que un modelo sea viable, debido principalmente a las observaciones extremas que pueden o no estar modificándolo.