Trabajo 1

4,5

Estudiantes

Eber David Baracaldo Galeano
Salomon Correa Espinosa
Angie Valentina Cubillos Barragan
Jhon Steven Marin Tapasco

Equipo 28

Docente

Mateo Ochoa Medina

Asignatura

Estadistica II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellin
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: probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).
- X_1: Duración de la estadía: duración promedio de la estadía de todos los pacientes en el hospital (en días).
- X_2: Rutina de cultivos: razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria por cada 100.
- X_3: Número de camas: número promedio de camas en el hospital durante el periodo del estudio.
- X_4: Censo promedio diario: número promedio de pacientes en el hospital por día durante el periodo del estudio.
- X_5: Número de enfermeras: número promedio de enfermeras, equivalentes a tiempo completo durante el periodo del estudio.

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	1.3314
β_1	0.1543
β_2	-0.0100
β_3	0.0567
β_4	0.0094
β_5	0.0014

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 1.3314 + 0.1543 X_{1i} - 0.01 X_{2i} + 0.0567 X_{3i} + 0.0094 X_{4i} + 0.0014 X_{5i};\ 1 \leqslant i \leqslant 69$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

5 pt

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

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	50.7325	5	10.146491	11.9269	3.87619e-08
Error	53.5957	63	0.850725		

De la tabla Anova, se observa un valor P menor a 0.05, 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 la regresión es significativa.

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

6 pt

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	1.3314	1.4875	0.8951	0.3741
β_1	0.1543	0.0635	2.4307	0.0179
β_2	-0.0100	0.0264	-0.3782	0.7066
β_3	0.0567	0.0144	3.9456	0.0002
β_4	0.0094	0.0068	1.3782	0.1730
β_5	0.0014	0.0006	2.2156	0.0303

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 P-valores son menores a α.

1.4. Interpretación de los parámetros

3pt

$\hat{\beta}_1$**:** La probabilidad promedio de adquirir una infección en el hospital aumenta un 15.43 % aproximadamente por un aumento de un día en el promedio de la estadía de todos los pacientes en el hospital cuando las demás variables se mantienen constantes.

$\hat{\beta}_3$**:** La probabilidad promedio de adquirir una infección en el hospital aumenta un 5.67 % por un aumento de una cama en el promedio de camas en el hospital cuando las demás variables se mantienen constantes.

$\hat{\beta}_5$**:** La probabilidad promedio de adquirir una infección en el hospital aumenta un 0.14 % por el aumento de una enfermera en el promedio de enfermeras equivalentes a tiempo completo cuando las demás variables se mantienen constantes.

Además de su insignificancia estadística, dado que en el conjunto de datos no se encuentra al 0 en el intervalo de estos, β_0 carece de interpretación.

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.94657$, lo que significa que aproximadamente el 94.66 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

¿Cómo se calcula?

2. Pregunta 2

1pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más bajo 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:

what?

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	53.596	X1 X2 X3 X4 X5
Modelo reducido	55.393	X1 X3 X5

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

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

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3\beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{0.599}{0.850725} \\ &= 0.70410532 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,63} = 2.7505$, se puede ver que $F_0 < f_{0.95,3,63}$ y por tanto, no se puede rechazar la hipótesis nula, por lo que hay suficiente evidencia para adoptar el modelo reducido ya que se pudo demostrar la inexistencia de un efecto en la variable dependiente por el subconjunto de dos variables excluidas.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se pregunta si el efecto sobre la probabilidad promedio de adquirir una infección será: ¿En el aumento de una cama en el promedio de estas un tercio del aumento de un día en el promedio de un días de estadía? ¿En el aumento de un cultivo en la razón de estos por cada 100 pacientes el negativo del aumento de un paciente en el promedio de estos por día? ¿En el aumento de una enfermera en el número promedio de estas 0.1 veces el aumento de un día en el promedio de un dias de estadía? Por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_3 = (1/3)\beta_1;\ \beta_2 = -\beta_4;\ \beta_5 = (0.1)\beta_1 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

$$L = \begin{bmatrix} 0 & -1/3 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & -0.1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_4 X^*_{4i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 69$$

Donde $X^*_{1i} = X_{1i} + (1/3)X_{3i} + (0.1)X_{5i}$ y $X^*_{4i} = -X_{2i} + X_{4i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF)/3}{MSE(MF)} \overset{H_0}{\sim} f_{3,63} = \frac{(SSE(MR) - 53.596/3}{0.850725} \overset{H_0}{\sim} f_{3,63} \qquad (3)$$

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 que se realizará por medio de shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

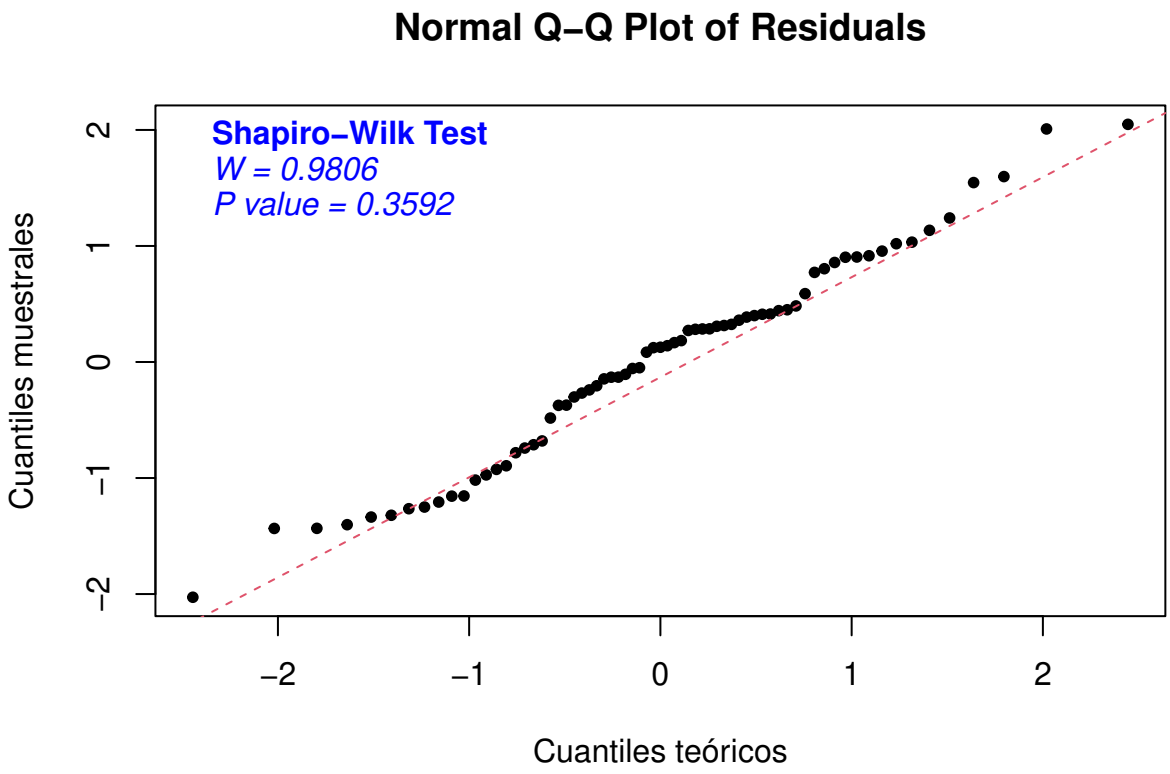

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.3592 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 como una normal, sin embargo, la gráfica de comparación de cuantiles permite ver colas más pesadas y patrones irregulares similares a ondas, al tener más poder el análisis gráfico se termina por rechazar el cumplimiento de este supuesto. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

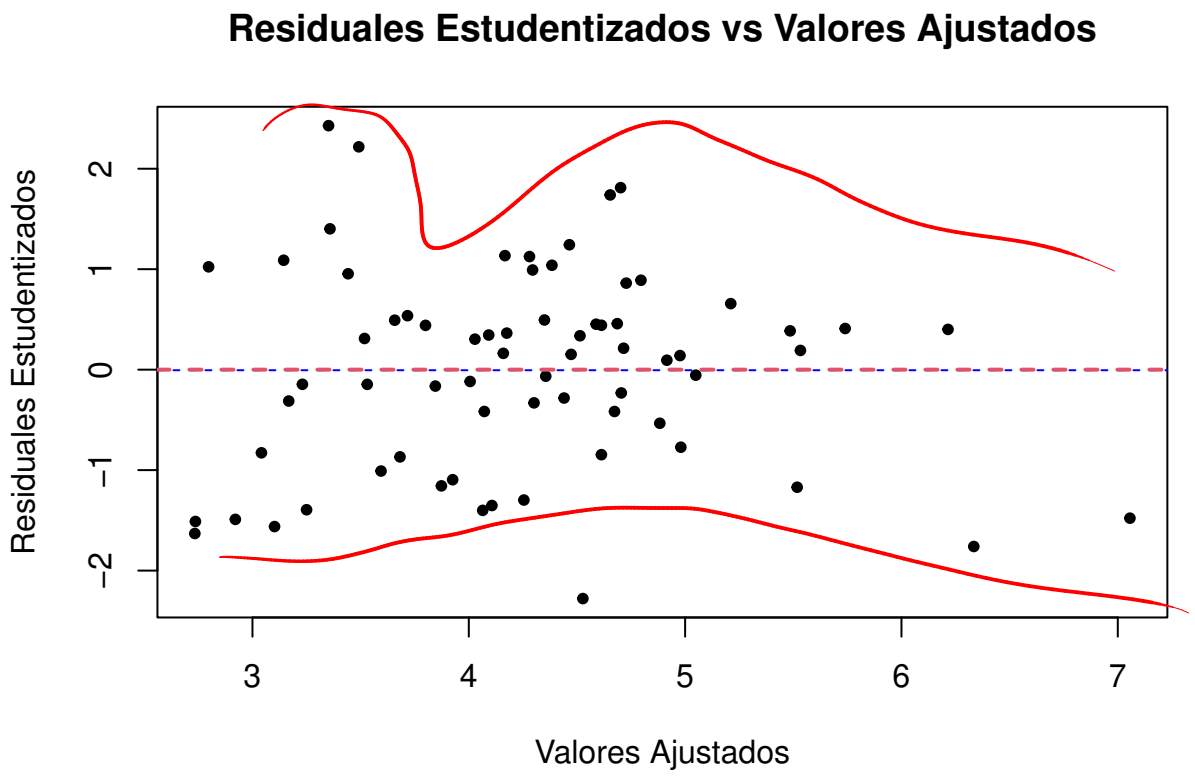

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar patrones de aumento y decenso de la varianza por lo cual se puede descartar una varianza constante. Por otro lado, es posible observar que la media de los residuales (linea azul) corresponde a la media 0 (linea roja) cumpliendo este supuesto.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

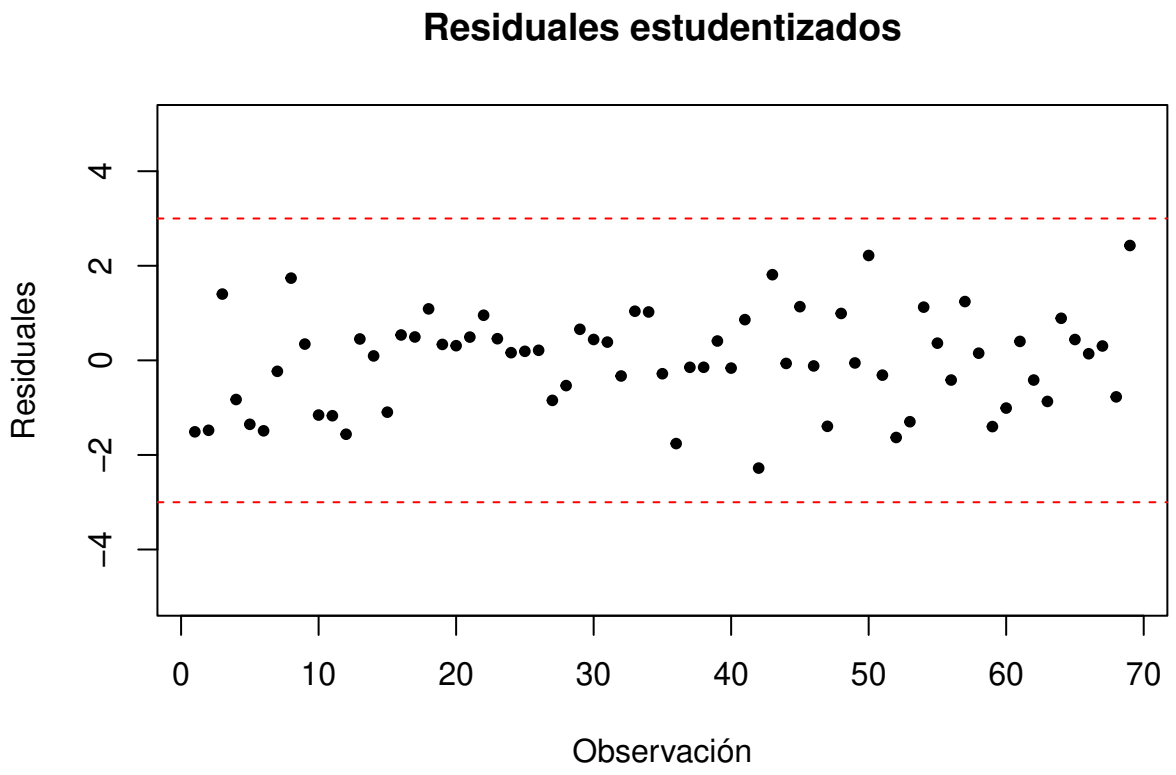

Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica anterior, no hay datos atípicos ú outliers que puedan afectar el ajuste del modelo dado que en el conjunto de datos ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

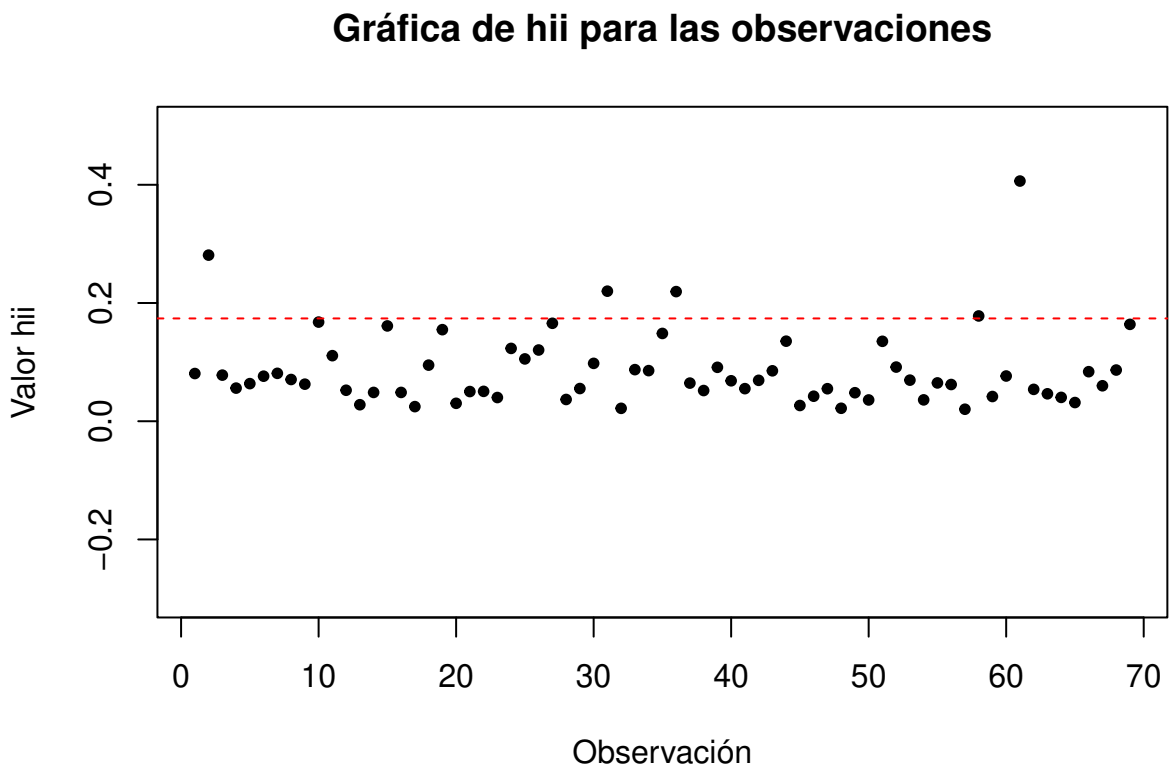

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 2   -1.4783  0.1424    0.2810 -0.9332
## 31   0.3861  0.0070    0.2200  0.2036
## 36  -1.7602  0.1450    0.2193 -0.9490
## 58   0.1519  0.0008    0.1781  0.0702
## 61   0.4007  0.0183    0.4063  0.3292
```

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, se puede apreciar que existen 5 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n} = 0.173913$, los cuales son los presentados en la tabla anterior. Así las observaciones 2, 31, 36, 58 y 61 pueden estar afectando los estádisticos de la tabla ANOVA, el cumplimiento de los supuestos o el R^2.

4.2.3. Puntos influenciales

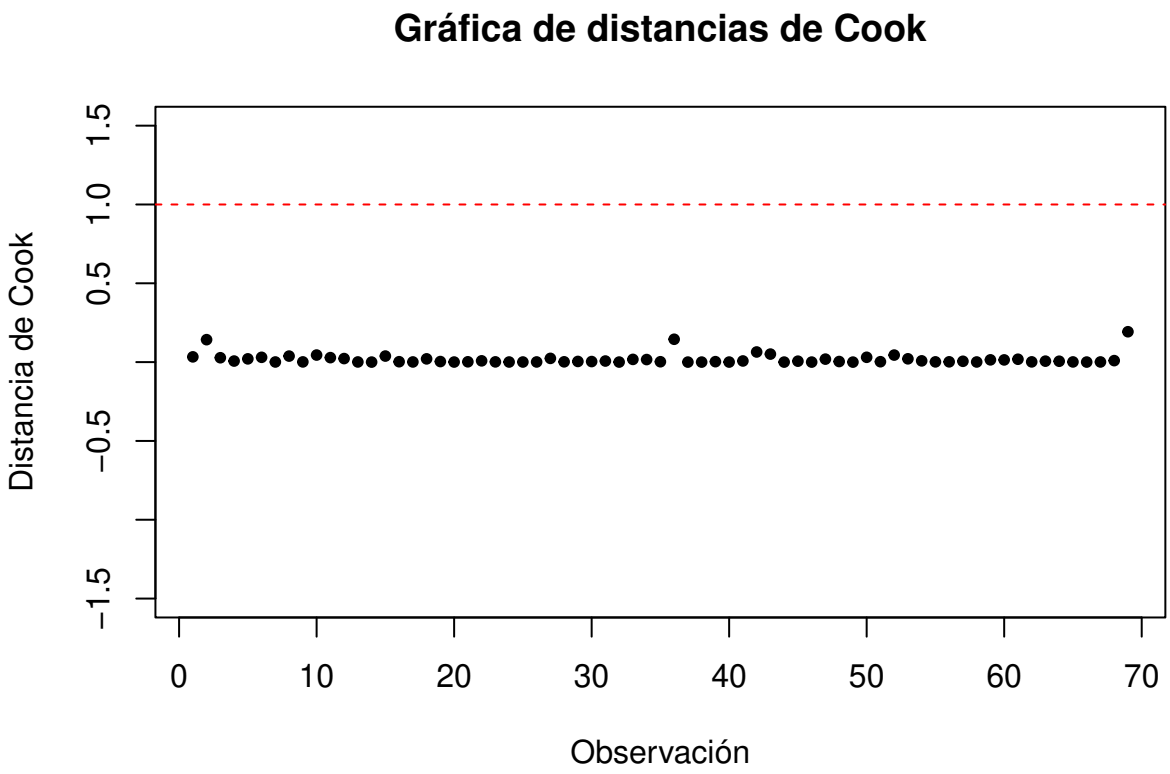

Figura 5: Criterio distancias de Cook para puntos influenciales

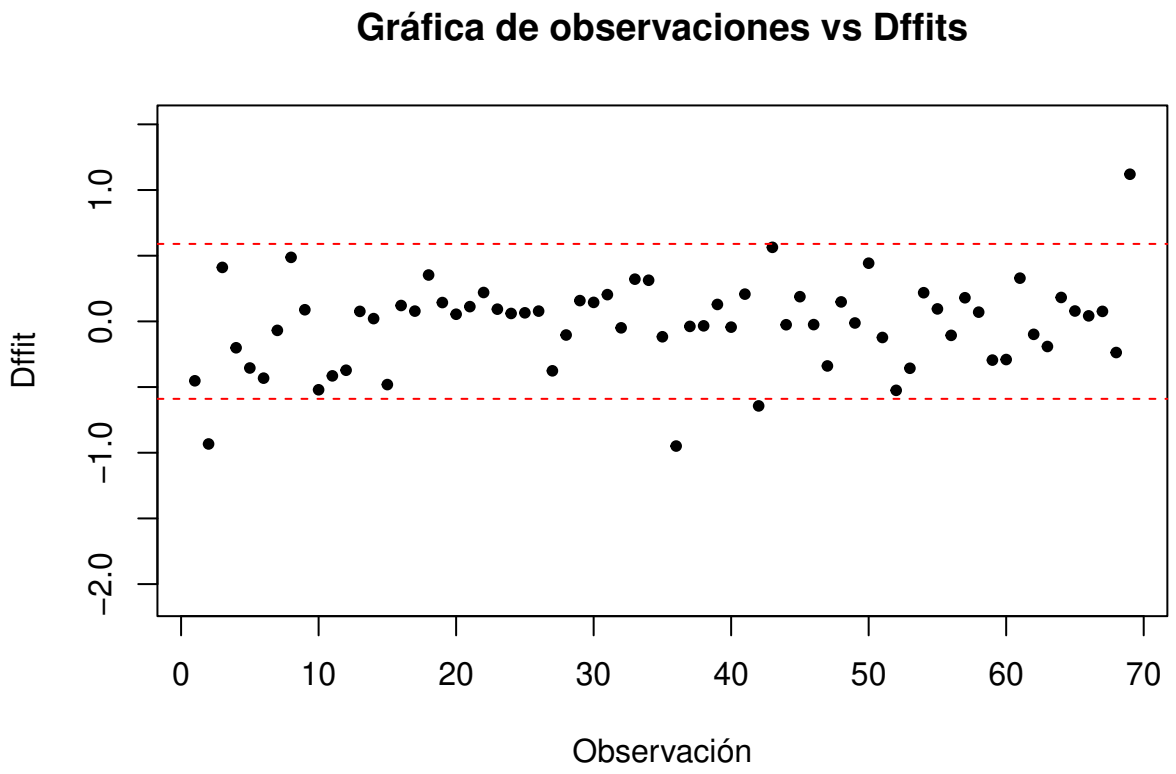

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 2   -1.4783  0.1424    0.2810 -0.9332
## 36  -1.7602  0.1450    0.2193 -0.9490
## 42  -2.2784  0.0643    0.0692 -0.6434
## 69   2.4285  0.1927    0.1639  1.1203
```

4pt

Como se puede ver, ninguna observación corresponde a un punto influencial según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.5897678$, es un punto influencial, por tanto, no se está afectando el espacio de la respuesta. Sin embargo, con el criterio de distancias de Cook, en el cual para cualquier punto cuya $D_i > 1$, las observaciones 2, 36, 42 y 69 corresponden a puntos influenciales. Así estas observaciones se caracterizan por tener un valor moderadamente inusual en el espacio de las predictoras, por tanto, su exclusión del modelo causaría cambios importantes en la estimación de estas.

4.3. Conclusión

3pt

En conclusión, el modelo presentado no es válido dado que a pesar de que se cumplió que la media de los residuales es igual a cero, no se cumplieron los otros supuestos: En cuanto

a la normalidad de los residuales a pesar de que por medio de Shapiro-Wilk se aceptó en un principio, el criterio más fuerte que es el análisis gráfico rechazó esté supuesto como se presentó en la sección 4.1.1; por otra parte, según lo expuesto en la sección 4.1.2, se identificó un patrón irregular en la varianza de los datos. Teniendo en cuenta lo anterior, el modelo no es idóneo para el conjunto de datos presentados.

En cuanto a los puntos extremos, se detectaron 5 datos como puntos de balanceo que pudieron alterar la no validación de los supuestos de normalidad y varianza constante además del R^2 y los estadísticos presentados en el ANOVA. Luego, mediante el criterio de las distancias de Cook se identificaron 4 observaciones que corresponden a puntos influenciales, los cuales bajo este criterio al ser excluidos de la base de datos, deberían tener un gran efecto en los coeficientes estimados de las variables independientes.