Trabajo 1

4,5

Estudiantes

Luis Fernando Lopez Echeverri
Libardo Jose Navarro Pedrozo
Efrain Gomez Ramirez
Santiago Acevedo Cacua

Equipo

35

Docente

Mateo Ochoa Medina

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 **5**

2.1. Planteamiento pruebas de hipótesis y modelo reducido 5

2.2. Estadístico de prueba y conclusión 6

3. Pregunta 3 **6**

3.1. Prueba de hipótesis y prueba de hipótesis matricial 6

3.2. Estadístico de prueba 7

4. Pregunta 4 **7**

4.1. Supuestos del modelo 7

4.1.1. Normalidad de los residuales 7

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 4
4. Resumen tabla de todas las regresiones 5

1. Pregunta 1

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

Donde:

- 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

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	-1.5602
β_1	0.0227
β_2	0.0399
β_3	0.0649
β_4	0.0246
β_5	0.0028

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.5602 + 0.0227 X_{1i} + 0.0399 X_{2i} + 0.0649 X_{3i} + 0.0246 X_{4i} + 0.0028 X_{5i}\ 1 \leqslant i \leqslant 54$$

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:

$$F_0 = \frac{MSR}{MSE} \overset{bajo H_0}{\sim} f_{5,48} \tag{1}$$

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	56.3009	5	11.260184	13.503	3.12399e-08
Error	40.0274	48	0.833904		

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 $0 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que almenos 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

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-1.5602	1.6500	-0.9456	0.3491
β_1	0.0227	0.1003	0.2263	0.8219
β_2	0.0399	0.0290	1.3754	0.1754
β_3	0.0649	0.0157	4.1214	0.0001
β_4	0.0246	0.0083	2.9644	0.0047
β_5	0.0028	0.0009	3.1542	0.0028

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_3, β_4 y β_5 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

$\hat{\beta}_3$: El riesgo de infección promedio aumenta en 0.0649 por cada unidad de número de camas que aumentan si el resto de variables regresoras permanecen constantes.

$\hat{\beta}_4$: El riesgo de infección promedio aumenta en 0.0246 por cada unidad del censo promedio diario que aumentan si el resto de variables regresoras permanecen constantes.

$\hat{\beta}_5$: El riesgo de infección promedio aumenta en 0.0028 por cada unidad de número de enfermeras que aumentan si el resto de variables regresoras permanecen constantes.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5844689$, lo que significa que aproximadamente el 58.45 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión 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	40.027	X1 X2 X3 X4 X5
Modelo reducido	69.312	X1 X2

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \ \cdots, \ \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,48} \\ &= \frac{9.761}{0.833904} \\ &= 11.70518 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,48} = 2.7981$, se puede ver que $F_0 > f_{0.95,3,48}$ y por tanto se rechaza la hipotesis nula y se puede concluir que al menos una de las siguientes variables regresoras: X_3 (numero de camas), X_4 (censo promedio diario) y X_5 (numero de enfermeras) es una variable singnificativa para explicar el riesgo de infeccion. Sin embargo, gracias al punto 1.3 sabemos que las tres son signficativas, con esto concluimos que no se recomienda descartar estas variables del modelo.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hacen las preguntas si ¿La duración de la estadia y la rutina de cultivos no son significativos para el modelo de regresión? y ¿La tasa de aumento promedio del riesgo de infección respecto al censo promedio diario es igual a la del numero de enfermeras? por consiguiente se plantea la siguiente prueba de hipótesis:

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

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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_3 X_{3i} + \beta_4 X^*_{4i} + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 54$$

Donde $X^*_{4i} = X_{4i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/3}{MSE(MF)} \overset{H_0}{\sim} f_{3,48} \tag{3}$$

$$F_0 = \frac{(SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0, \beta_1, \beta_2\beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0, \beta_1, \beta_2\beta_3, \beta_4, \beta_5)} \overset{H_0}{\sim} f_{3,48} \tag{4}$$

4. Pregunta 4

4.1. Supuestos del modelo

Los supuestos del modelo sobre independencia y media de los errores igual a cero, los asumiremos como ciertos, esto debido a que no contamos con la información de como fueron tomadas las muestras en el tiempo.

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:

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

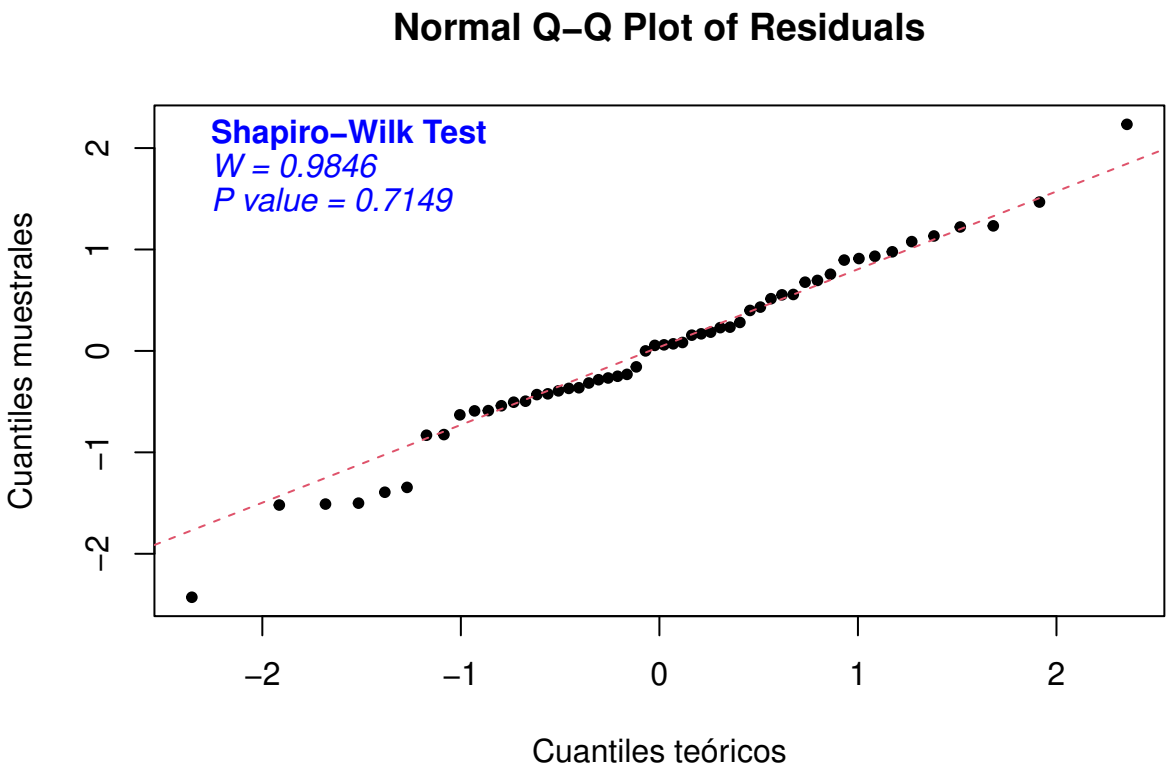

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.7149 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, los datos distribuyen normal, ademas con el analisis grafico podemos llegar a la misma conclusion, incluso considerando de que la cola izquierda contiene algunos puntos alejados, estos podrian ser candidatos a influenciales o atipicos, no obstante, a falta de pruebas considerablemente significativas en contra de la normalidad, se termina por aceptar el cumplimiento de este supuesto. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

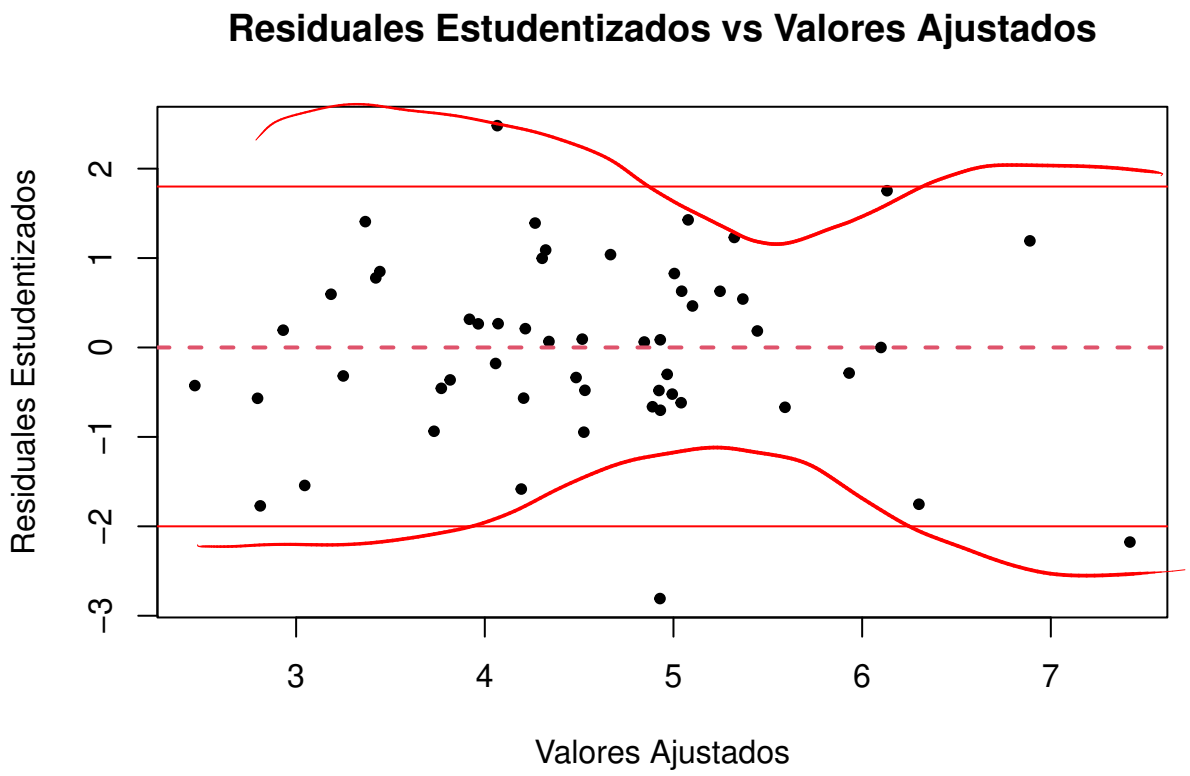

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados versus valores ajustados, no se aprecian patrones que indiquen un aumento o disminución en la variabilidad. Tampoco se observa ningún comportamiento que sugiera rechazar la hipótesis de varianza constante. Dado que no hay suficiente evidencia en contra de esta suposición, se acepta como válida. Además, se nota que la media de los residuales es igual a cero.

El grafico tambien permite apreciar que existen aproximadamente 6~7 puntos que se encuentran muy alejados a la derecha respecto a los demas, estos pueden ser candidatos a puntos de balanceo.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

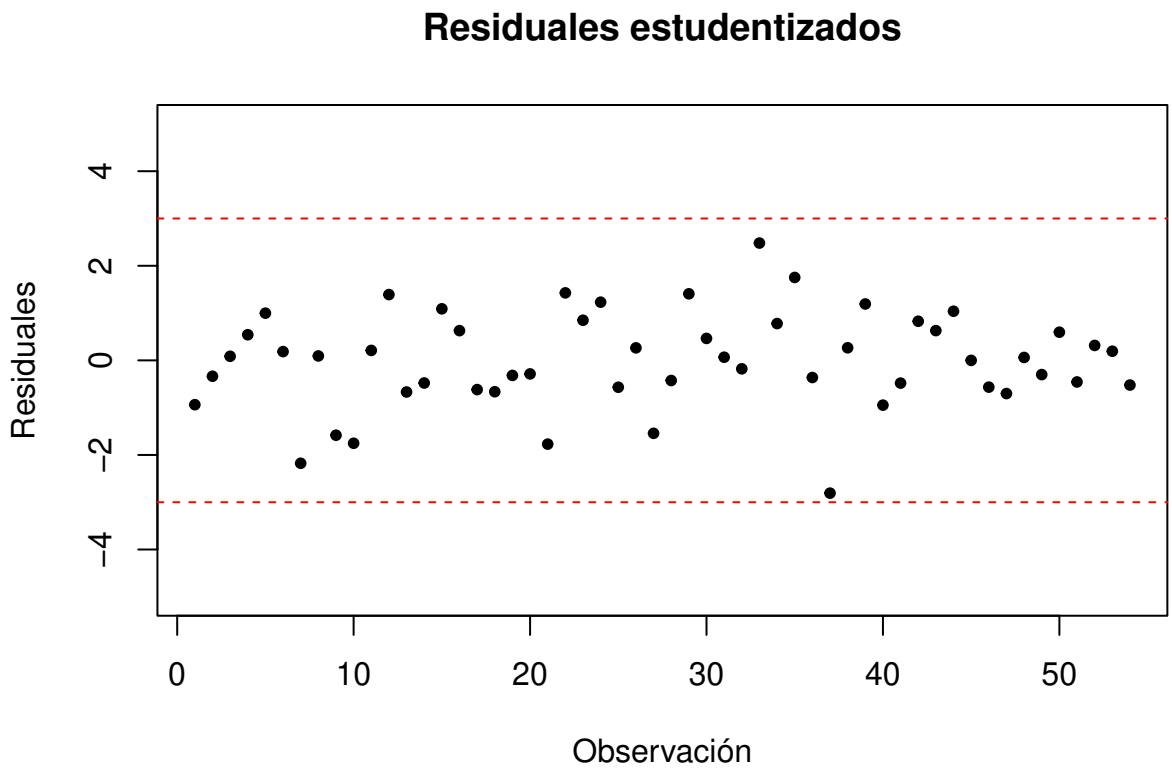

Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

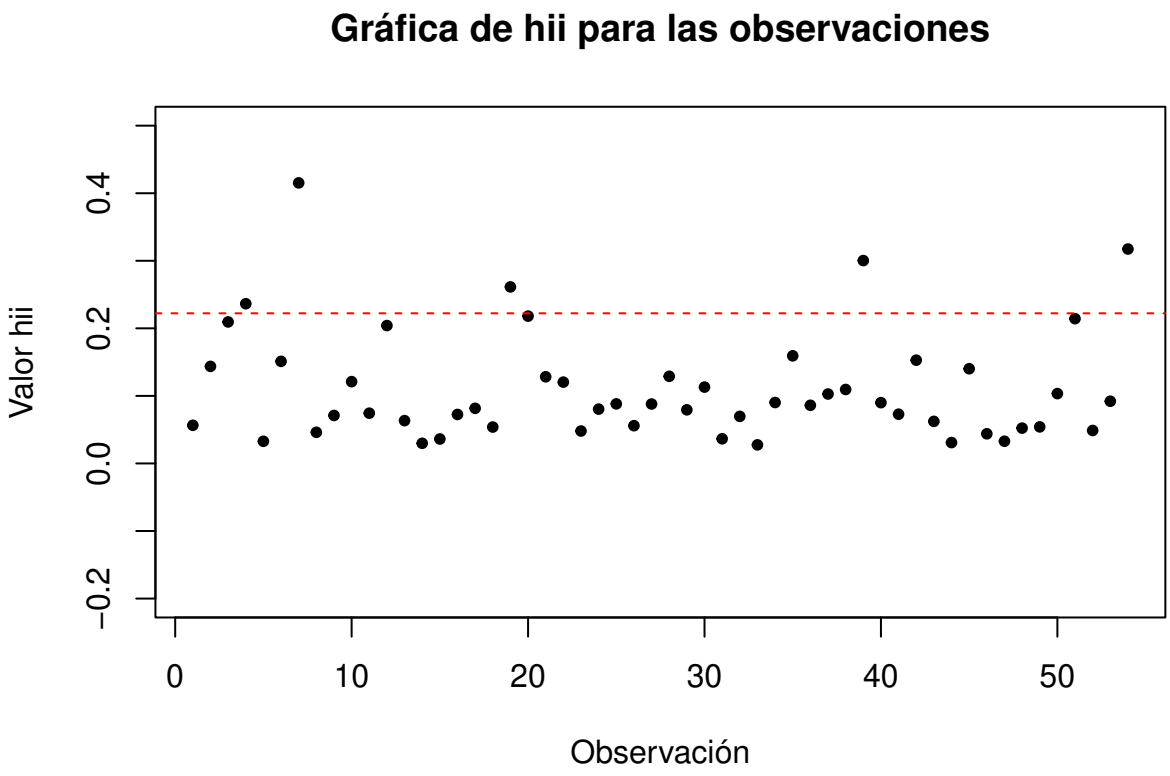

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 4    0.5417  0.0151    0.2364  0.2992
## 7   -2.1759  0.5604    0.4153 -1.9113
## 19  -0.3184  0.0060    0.2613 -0.1876
## 39   1.1924  0.1017    0.3003  0.7847
## 54  -0.5216  0.0211    0.3174 -0.3530
```

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.2222$, se puede apreciar que existen 5 datos (4,7,19,39 y 54) del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$, los cuales son los presentados en la tabla.

4.2.3. Puntos influenciales

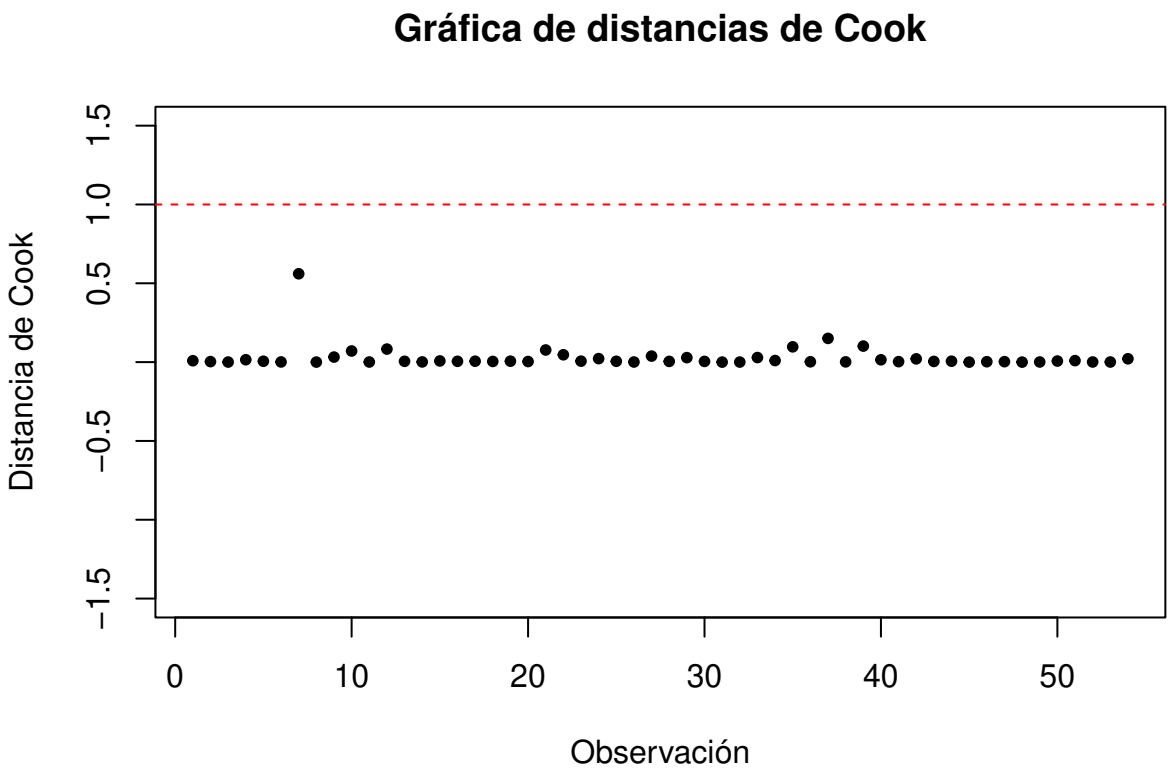

Figura 5: Criterio distancias de Cook para puntos influenciales

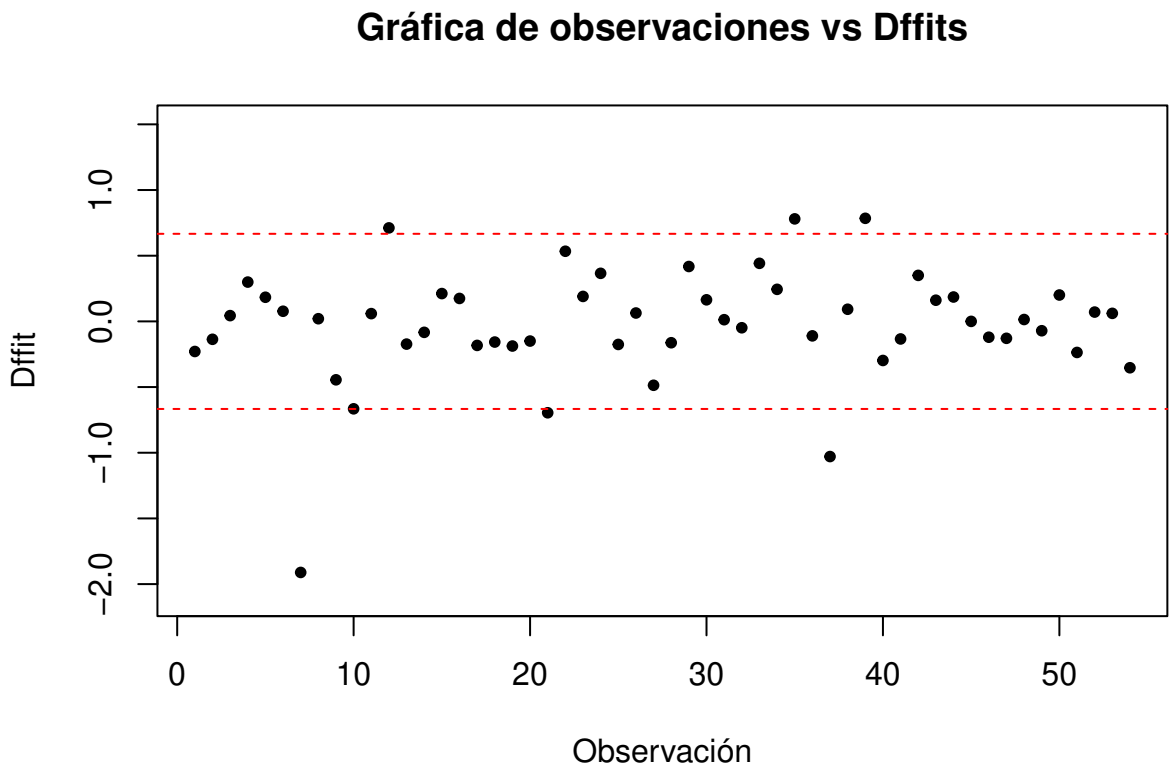

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 7   -2.1759  0.5604    0.4153 -1.9113
## 12   1.3912  0.0827    0.2041  0.7117
## 21  -1.7712  0.0770    0.1283 -0.6956
## 35   1.7531  0.0970    0.1593  0.7804
## 37  -2.8078  0.1506    0.1028 -1.0290
## 39   1.1924  0.1017    0.3003  0.7847
```

Como se puede ver,las observaciones 7,12,21,35,37 y 39 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.666$, 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.

4.3. Conclusión

El modelo cumple con los supuestos de varianza constante y normalidad, por lo tanto el modelo es valido, sin embargo debemos considerar la presencia de los puntos de balanceo

e influenciales los cuales pueden estar afectando la validez, por esta razon es importante examinar cómo estos estan influyendo en nuestro modelo y considerar técnicas de corrección o transformación si es necesario.