Trabajo 1

4,5

Estudiantes

Luis Miguel Doria Rodríguez
Diana Sirley García Quintero
Juan Stiven Giraldo Rua
Luis Miguel Martínez Zapata

Equipo 19

Docente

Julieth Veronica Guarín Escudero

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 . . . 9

4.2.1. Datos atípicos . . . 10

4.2.2. Puntos de balanceo . . . 11

4.2.3. Puntos influenciales . . . 12

4.3. Conclusiones . . . 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

20pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

Acá especificamos el nombre de las variables:

- 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	-2.4118
β_1	0.0386
β_2	0.0674
β_3	0.0601
β_4	0.0165
β_5	0.0021

3pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -2.4118 + 0.0386 X_{1i} + 0.0674 X_{2i} + 0.0601 X_{3i} + 0.0165 X_{4i} + 0.0021 X_{5i},\ 1 \leqslant i \leqslant 59$$

1.2. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{bajoH_0}{\sim} f_{5,53} \tag{1}$$

Ahora, se presenta la tabla Anova:

5 pt

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	48.2535	5	9.650700	11.1709	2.2166e-07
Error	45.7875	53	0.863915		

De la tabla Anova, se observa un valor P muy pequeño y menor a un alfa de 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 $\beta_j \neq 0$ con $1 \leqslant j \leqslant 5$, 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	-2.4118	1.6259	-1.4834	0.1439
β_1	0.0386	0.0860	0.4492	0.6551
β_2	0.0674	0.0293	2.2975	0.0256
β_3	0.0601	0.0175	3.4427	0.0011
β_4	0.0165	0.0074	2.2468	0.0288
β_5	0.0021	0.0007	2.9787	0.0044

6 pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_2, β_3, β_4 y β_5 son significativos, pues sus P-valores son menores a α. De este modo, concluimos que el parámetro β_1 no es significativo, pues su P-valor es mayor a α.

3pt

1.4. Interpretación de los parámetros

$\hat{\beta}_2$**:** Por cada unidad de rutina de cultivos a pacientes sin síntomas que aumenta durante el periodo de estudio, el riesgo de infección aumenta en promedio 0.0674, cuando las demás están fijas.

$\hat{\beta}_3$**:** Por cada cama que aumenta en el hospital durante el periodo de estudio, el riesgo de infección aumenta en promedio 0.0601, cuando las demás están fijas.

$\hat{\beta}_4$**:** Por cada paciente adicional por día durante el periodo de estudio, el riesgo de infección aumenta en promedio 0.0165, cuando las demás están fijas

$\hat{\beta}_5$**:** Por cada enfermera que aumente durante el periodo de estudio, el riesgo de infección aumenta en promedio 0.0021, cuando las demás están fijas

1.5. Coeficiente de determinación múltiple R^2

3pt

$$R^2 = \frac{SSR}{SST} = \frac{48.2535}{48.2535+45.7875} = 0.5131$$

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5131$, lo que significa que aproximadamente el 51.31 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

2. Pregunta 2

5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más pequeño en el modelo fueron X_2, 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:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	45.788	X1 X2 X3 X4 X5
Modelo reducido	65.550	X1 X4

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

2.2. Estadístico de prueba y conclusión

3pt

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_4) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,53} \\ &= \frac{(65.550 - 45.788)/3}{0.863915} \\ &= 7.624978 \end{aligned} \tag{2}$$

2pt

Ahora, comparando el F_0 con $f_{0.95,3,53} = 2.7791$, se puede ver que $F_0 > f_{0.95,1,45}$ y por tanto se rechaza la hipótesis nula, y podemos concluir que al menos uno de los betas es diferente de cero.

¿Es posible o no descartar las variables del subconjunto?
Debido a que rechazamos la hipótesis nula y afirmamos que al menos uno de los betas es significativo, no resulta conveniente descartar alguna de las variables del subconjunto, puesto que nos aportan información relevante para dar explicación al modelo propuesto.

3. Pregunta 3

4pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hacen las siguientes tres preguntas:

- ¿El efecto de la duración de la estadía sobre el riesgo de infección es igual al doble del efecto de número de enfermeras sobre el riesgo de infección?
- ¿El efecto del número de enfermeras sobre el riesgo de infección es igual a tres veces el efecto del censo promedio diario sobre el riesgo de infección?
- ¿El efecto de la rutina de cultivos sobre el riesgo de infección es igual al efecto del censo promedio diario sobre el riesgo de infección?

Por consiguiente, se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 2\beta_5;\ \ \beta_5 = 3\beta_4;\ \ \beta_2 = \beta_4 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

, 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 **L** dada por

$$L = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & -3 & 1 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{bmatrix}$$

2 pt

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_4 X_{4i}^{*} + \varepsilon_i, \;\; \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \;\; 1 \leqslant i \leqslant 59$$

0 pt

Donde

$$X_{4i}^{*} = X_{2i} + X_{4i} + 3X_{5i}$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/3}{MSE(MF)} \overset{bajo H_0}{\sim} f_{3,53} \tag{3}$$

Con los valores conocidos, reemplazamos y resulta:

$$F_0 = \frac{(SSE(MR) - 45.788)/3}{9.6507}$$

2 pt

4. Pregunta 4

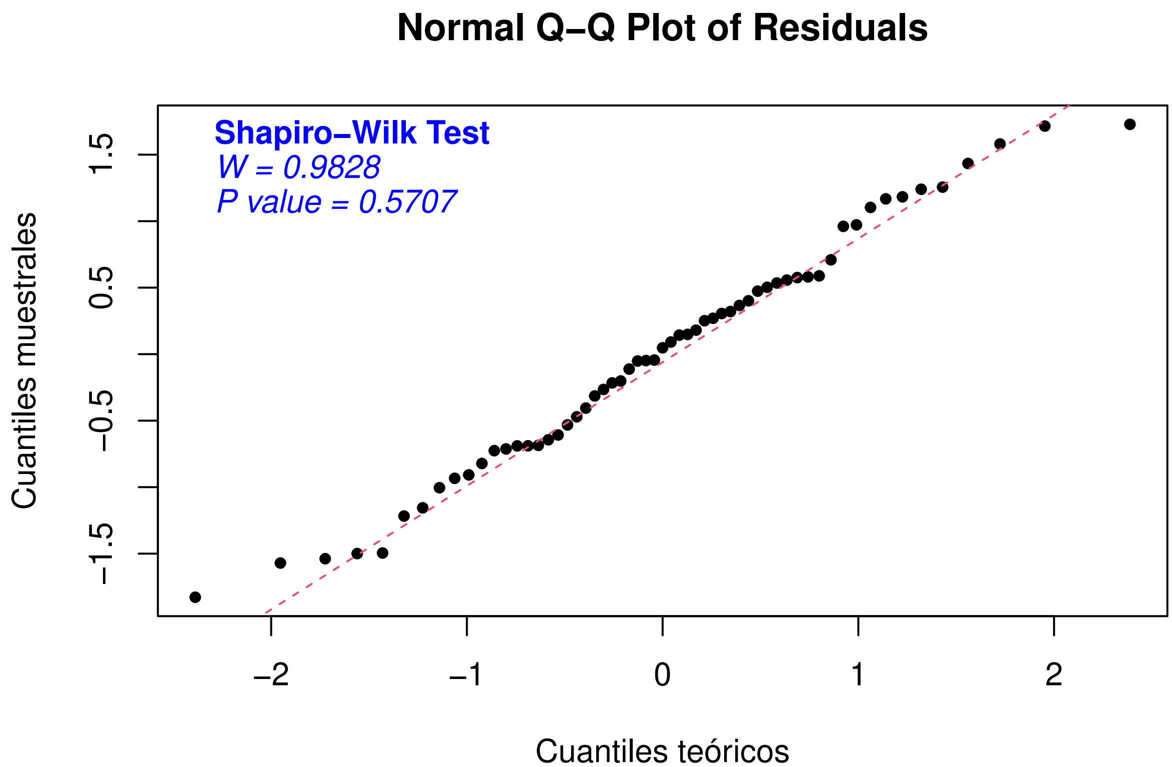
16 pt

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 analizará por medio de la gráfica Q-Q Norm y de la prueba Shapiro-Wilk:

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



4pt

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.5707 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 normal con media 0 y varianza σ^2, adicionalmente, la gráfica de comparación de cuantiles permite ver colas no tan pesadas y que la mayoría de los datos tienden a estar en el centro y ajustados a la recta. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

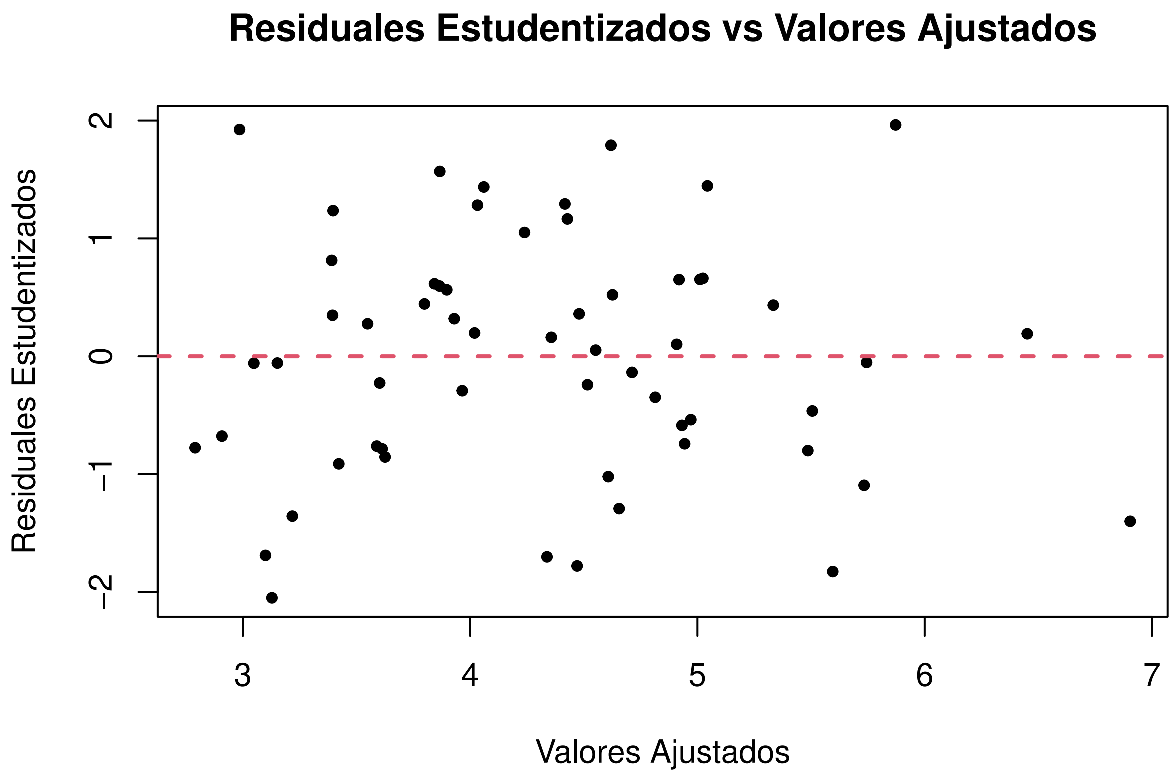


Figura 2: Gráfico residuales estudentizados vs valores ajustados

3pt

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente, decrezca ni un comportamiento que permita descartar una varianza constante, al notar que sigue un patrón constante podemos afirmar con certeza que estamos ante un caso de varianza constante y, de igual forma, no presenta falta de ajuste.

4.2. Verificación de las observaciones

-1pt

Tengan cuidado acá, modifiquen los límites de las gráficas para que tenga sentido con lo que observan en la tabla diagnóstica. También, consideren que en aquellos puntos extremos que identifiquen deben explicar el qué causan los mismos en el modelo.

4.2.1. Datos atípicos

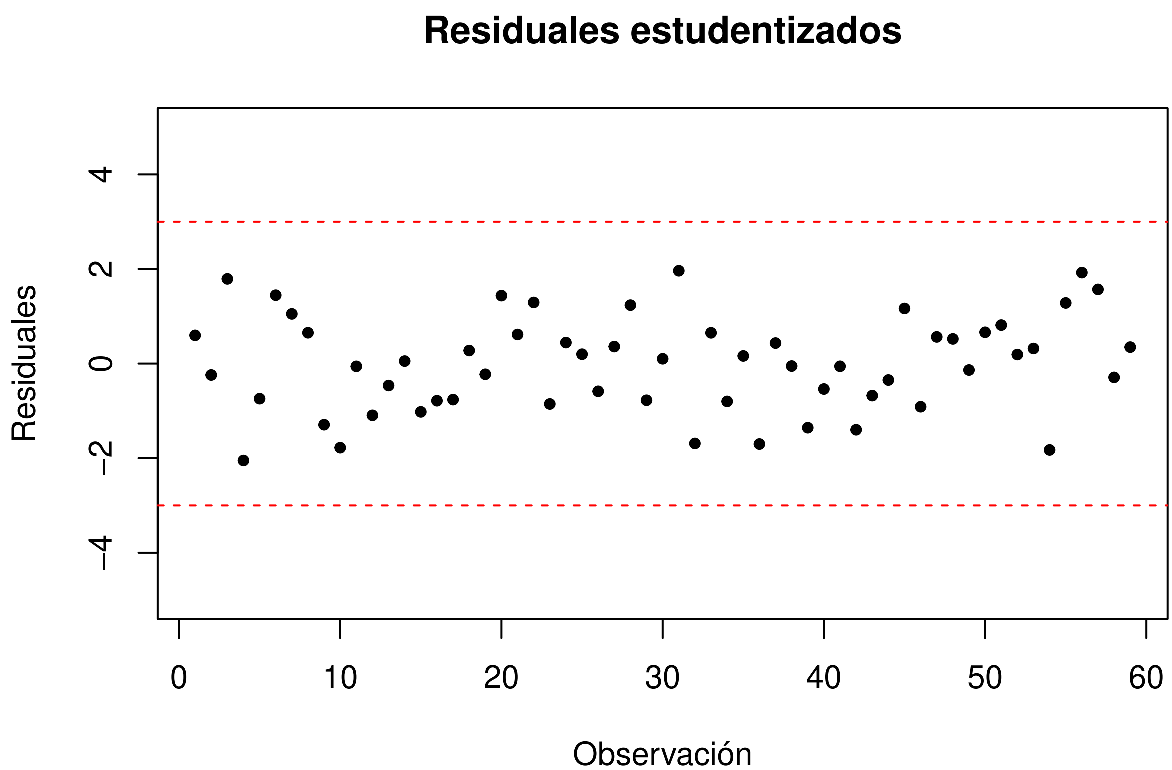


Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

3p7

4.2.2. Puntos de balanceo

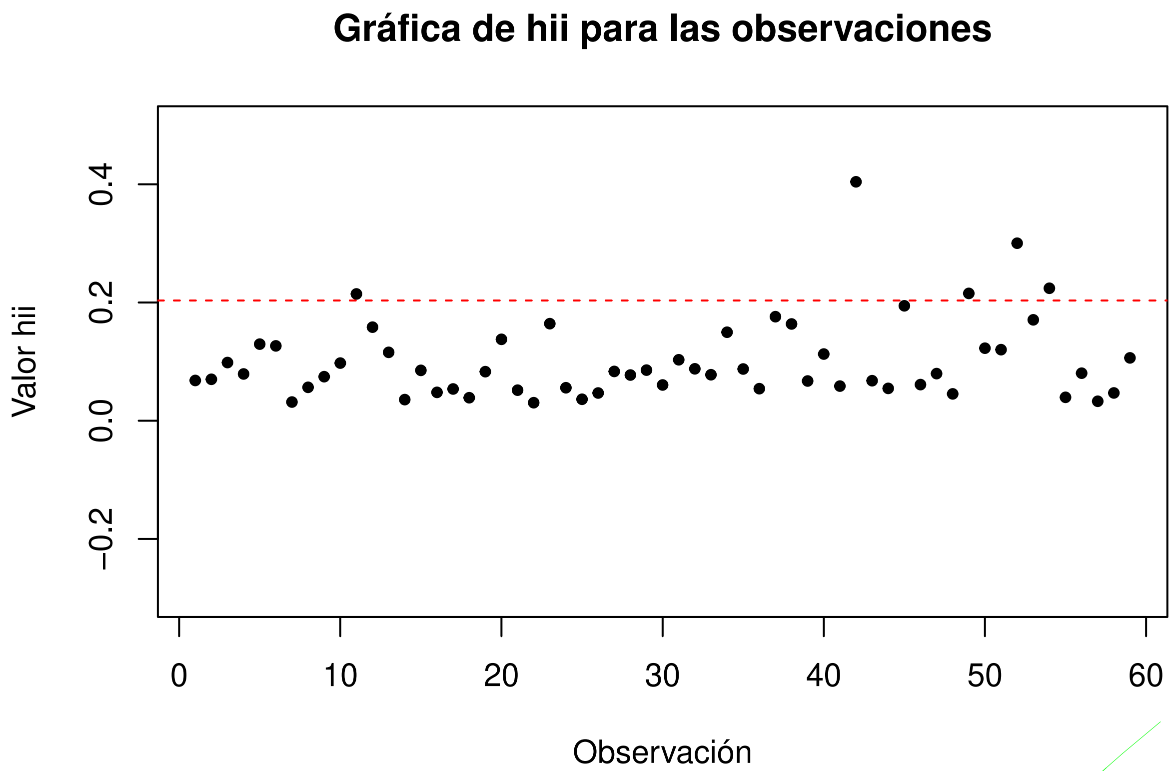


Figura 4: Identificación de puntos de balanceo

```
##     res.stud Cooks.D hii.value  Dffits
## 11  -0.0586  0.0002    0.2145 -0.0303
## 42  -1.3999  0.2216    0.4042 -1.1639
## 49  -0.1365  0.0009    0.2154 -0.0708
## 52   0.1911  0.0026    0.3003  0.1241
## 54  -1.8264  0.1605    0.2241 -1.0043
```

3pt

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 $h_{ii} > 2\frac{p}{n}$, los cuales son los presentados en la tabla. Estos puntos de balanceo nos indican datos que se alejan de los demás en el mundo de las x.

4.2.3. Puntos influenciales

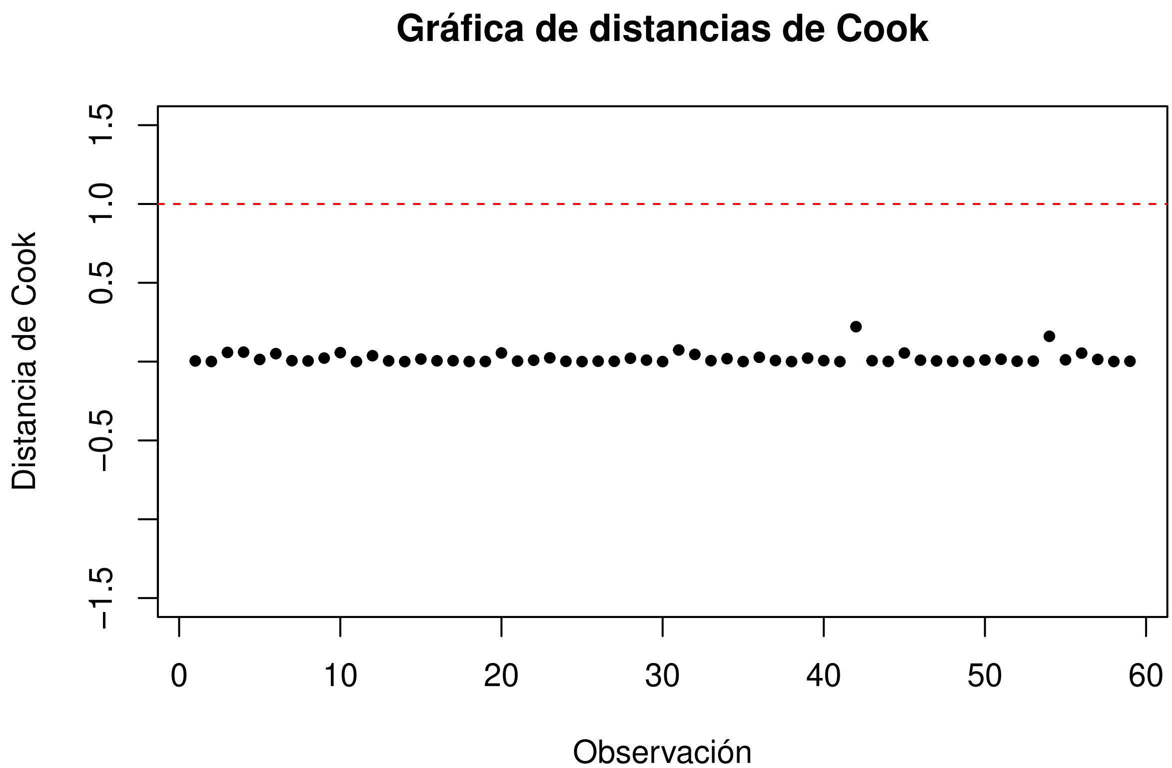


Figura 5: Criterio distancias de Cook para puntos influenciales

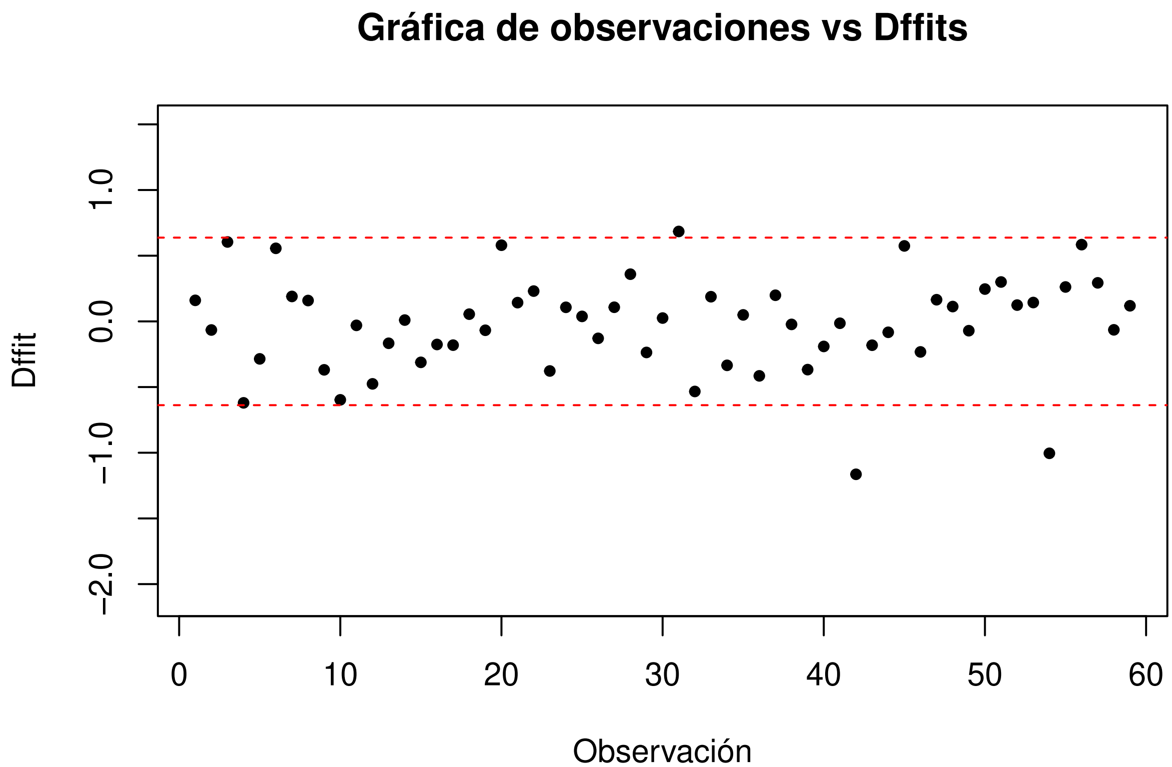


Figura 6: Criterio Dffits para puntos influenciales

```
##      res.stud Cooks.D hii.value  Dffits
## 31    1.9630  0.0738    0.1031  0.6846
## 42   -1.3999  0.2216    0.4042 -1.1639
## 54   -1.8264  0.1605    0.2241 -1.0043
```

4 pt

Como se puede ver, las observaciones 31, 42 y 54 son 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. 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. Con base en estos diagnósticos, identificamos que tres observaciones de la muestra son puntos influenciales, y que, de estos, las 42 y 54 son también puntos de balanceo. Estos puntos nos indican que tienen alto grado de influencia en los coeficientes de la regresión, se considera que su exclusión del modelo causaría cambios importantes en la ecuación de regresión ajustada

4.3. Conclusiones

0 pt   Validez sólo dada por los supuestos

- Podríamos afirmar que el modelo en cuestión es parcialmente válido, debido a que al observar el R^2 vemos que el modelo sólo explica el 51.31 % de los datos; mientras que el

49.69 % restante es explicado por el error, por lo cual consideramos que faltan variables que nos ayuden a explicar mejor el modelo.

- Las observaciones 42 y 54 son los que mayor número de enfermeras datan, por lo tanto, jalonan el modelo hacia estos puntos de balanceo. Y alteran el valor de la estimación de los betas.