Trabajo 1

4,5

Estudiantes

Marcela Katerine Rodriguez Gaviria
Luis Fenando Montoya Rodriguez
Angie Gabriela Medina Ramirez
Jose David Corredor Mesa

Equipo 49

Docente

Carlos Mario Lopera

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

11 pt

Teniendo en cuenta la base de datos de un estudio sobre el control de infecciones hospitalirias en EEUU, en la cual hay 5 variables predictoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

Donde las variables representan:

- Y: Riesgo de infección (Pocentaje).
- 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	-0.8690
β_1	0.1340
β_2	0.0182
β_3	0.0543
β_4	0.0184
β_5	0.0022

3 pt

Por lo tanto, el modelo de regresión ajustado está dado por:

$$\hat{Y}_i = -0.869 + 0.134X_{1i} + 0.0182X_{2i} + 0.0543X_{3i} + 0.0184X_{4i} + 0.0022X_{5i}$$

1.2. Significancia de la regresión

Para analizar la significancia de la regresión, se plantea el siguiente juego de hipótesis con k=5 parámetros iguales o diferentes de 0:

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

Cuyo estadístico de prueba es:

5pt

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,68} \tag{1}$$

A partir de lo anterior, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	88.0454	5	17.609076	20.9438	1.31488e-12
Error	57.1730	68	0.840779		

A partir de la prueba de hipótesis, rechazo H_0 si Val-P < α. Según la tabla ANOVA, Val-P es un número apróximadamente 0, por lo que se rechaza la hipótesis nula en la que $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$; probando la significancia del modelo, con un nivel de significancia de $\alpha = 0.05$.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información sobre los parámetros, la cual permitirá determinar cuáles de ellos son significativos, teniedo en cuenta el resultado de los Val-P correspondientes a cada parámetro.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.8690	1.4509	-0.5989	0.5512
β_1	0.1340	0.0749	1.7890	0.0781
β_2	0.0182	0.0258	0.7064	0.4824
β_3	0.0543	0.0114	4.7797	0.0000
β_4	0.0184	0.0069	2.6481	0.0101
β_5	0.0022	0.0007	3.1353	0.0025

6pt

Los Val-P arrojados por la tabla, permiten concluir que con un nivel de significancia de $\alpha = 0.05$, los parámetros β_3 , β_4 y β_5 son significativos. Pues sus P-valores cumplen con el criterio de Val-P $< \alpha$, para la prueba de significancia de los parámetros estimados en el modelo de regresión.

1.4. Interpretación de los parámetros

3 pt

Los parámetros significativos son interpretados como:

$\hat{\beta}_3$**:** Por cada cama que se aumenta en el hospital, el promedio de infección aumenta en 0.0543 %, cuando las demás variables permanecen constantes.

$\hat{\beta}_4$**:** Por cada nuevo paciente, el promedio de infección aumenta en 0.0184 %, cuando las demás variables permanecen constantes.

$\hat{\beta}_5$**:** Por cada enfermera de tiempo completo contratada en el hospital, el promedio de infección aumenta en un 0.0022 %, cuando las demás variables permanecen constantes.

1.5. Coeficiente de determinación múltiple R^2

2 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.6063$, el cual expresa que aproximadamente el 60.63 % de la variabilidad total observada en la respuesta, es explicada por el modelo de regresión propuesto.

¿cómo se calcula?

2. Pregunta 2

5 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más pequeño 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:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	57.173	X1 X2 X3 X4 X5
Modelo reducido	99.975	X1 X2

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,68} \\ &= \frac{(99.975 - 57.173)/3}{0.840779} \\ &= 16.9691837 \end{aligned} \tag{2}$$

2pt

Ahora, comparando el F_0 con $f_{0.95,3,68}$, se puede ver que $16.9691837 > 2.739502$, por tanto se rechaza Ho. Y se concluye que el riesgo promedio de infección es explicado por al menos una de las varaibles del subconjunto; ya sea X3(Numero de camas), X4(Censo promedio diario) y/o X5(Numero de enfermeras), donde estas son significativas simultaneamente.

Es posible o no descartar las variables del subconjunto?

R: No es posible descartarlas, pues explican significativamente el modelo.

3. Preguntа 3

2pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

¿El efecto del número de camas (X_3)) sobre el riesgo de infección (Y) es igual a 2 veces el número de enfermeras(X_5)?. Simultaneamente, ¿El efecto sobre el riesgo de infección (Y) causado por los días de duración de la estadía(X_1) es igual a 6 veces el efecto de rutina de cultivos (X_2) igual a 0? Por lo anterior, se plantea siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_3 = 2\beta_5;\ \boxed{\beta_1 = 6\beta_2 = 0} \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$?

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:

$$L = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & -2 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \end{bmatrix}$$

0pt

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_3 X_{3i}^* + \beta_4 X_{4i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 74$$

0pt

Donde $X_{3i}^* = \cancel{X_{3i} + 2X_{5i}}$.

$= 2X_{3i} + X_{5i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 57.1730/3}{0.840779} \overset{H_0}{\sim} f_{3,68} \quad (3)$$

2pt

4. Pregunta 4

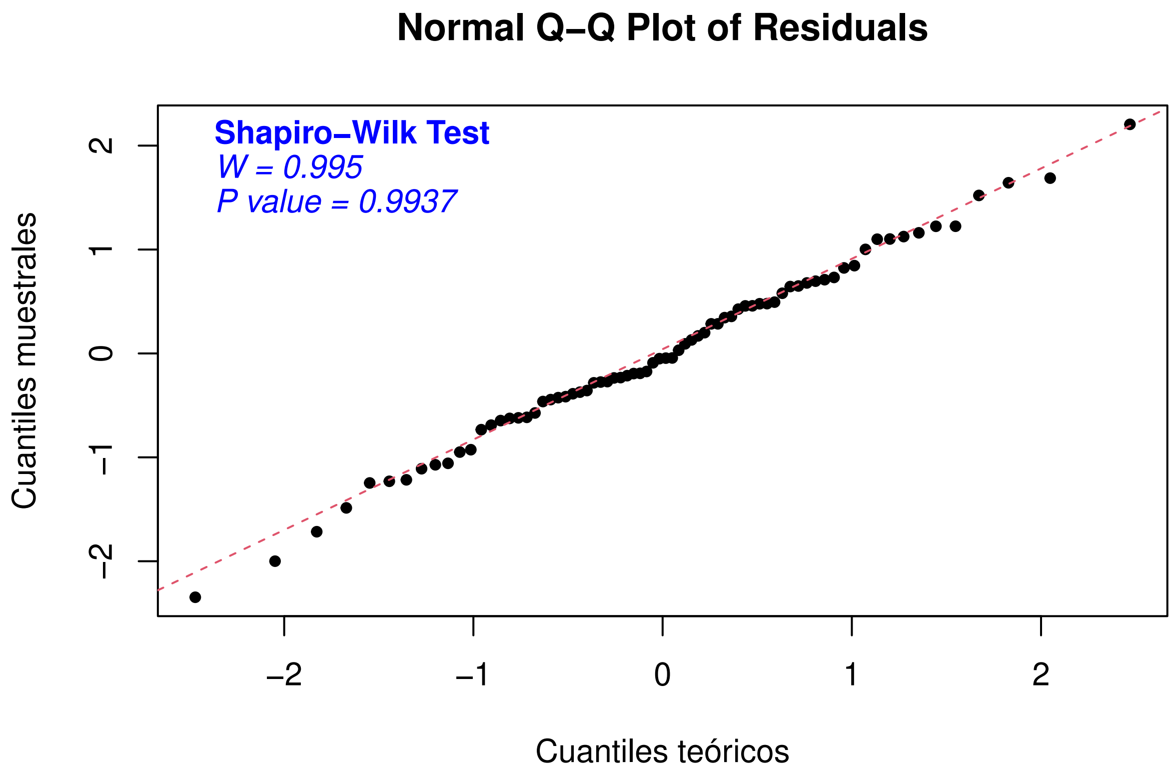
19 pt

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 realizará por medio de la prueba de normalidad de shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

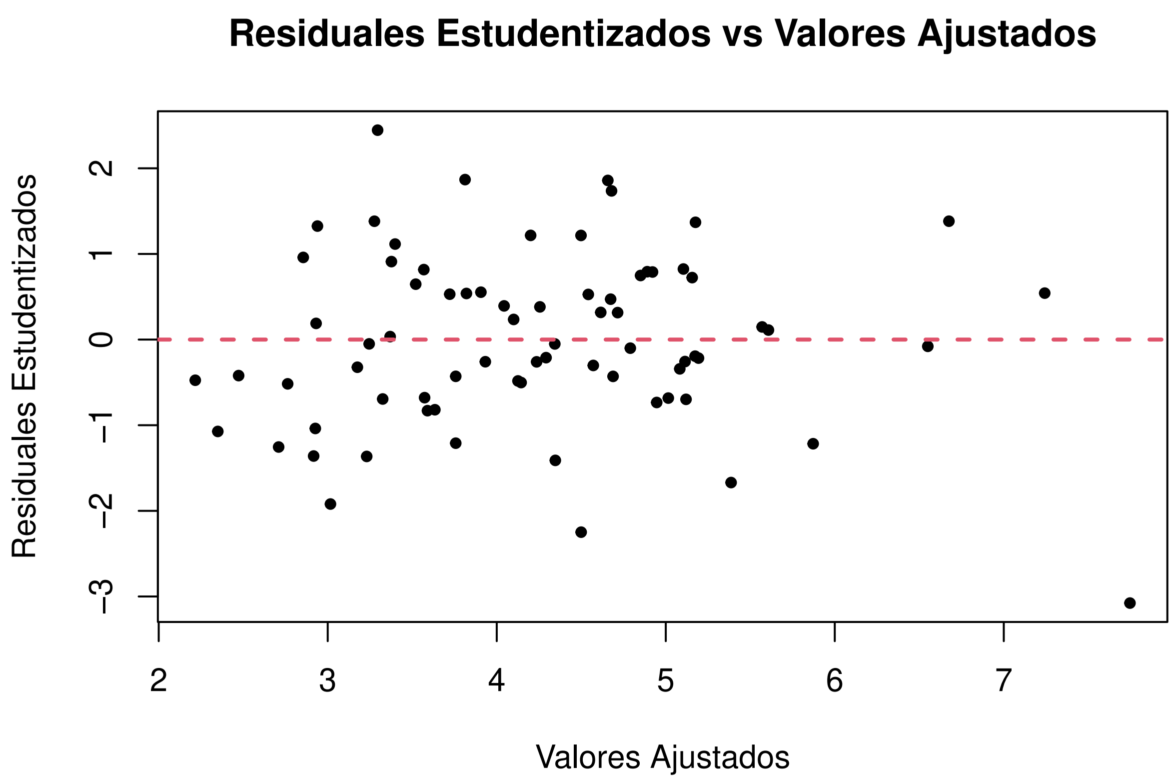


4et

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.9937 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 se distribuyen normal con media μ y varianza σ^2. En respaldo a esta conclusion, la gráfica de comparación de cuantiles permite ver en su medio qué tan cerca se ajustan las observaciones a la recta roja, y que la cola superior también lo hace. Solo hay una pequeña y despreciable irregularidad que está en la cola inferior,y son las 3 primeras observaciones que se desprende del patron, aproximadamente un 4 % de los datos. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante



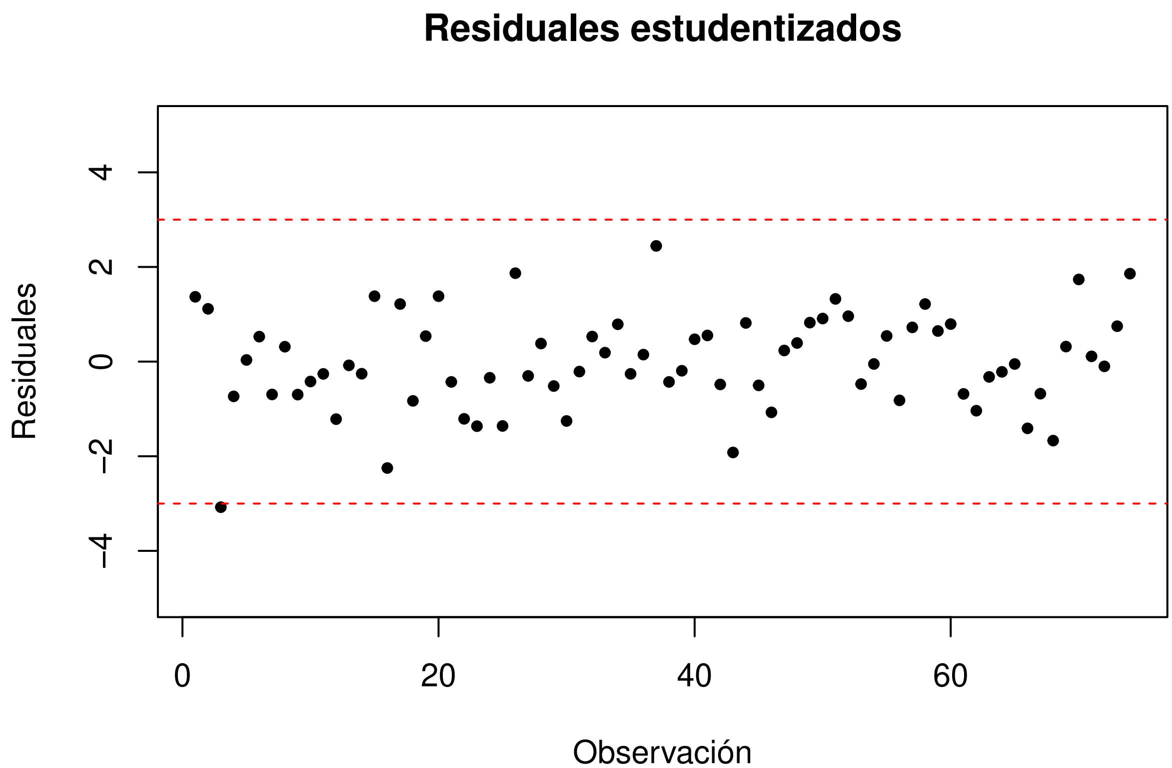
3ext

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar unos valores atípicos, los cuales se salen de la tendencia. Al mirar las demás observaciones, podemos ver como la gráfica se asemeja a una forma rectangular en más de la primera mitad indicando que el supuesto se cumple, en el resto de la gráfica se observan unos pocos valores alejados de la nube. Se concluye que se cumple el supuesto de varianza constante, sin embargo hay algunos valores alejados de la nube de tendencia.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos



3pt

Figura 3: Identificación de datos atípicos

Como se puede observar en la gráfica anterior, en el conjunto de datos hay un dato atípico, pues este residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$. Dicho valor corresponde a la observacion 3 de nuestros datos.

Esta observación posiblemente sea la responsable de dispersar la gráfica de residuales vs valores ajustados, y afectar el patrón de la varianza.

4.2.2. Máximos y mínimos por Variable

	Y	X1	X2	X3	X4	X5
min	1.3	6.70	38.8	1.6	39.6	29
max	7.8	19.56	63.9	60.5	122.8	752

4.2.3. Puntos de balanceo

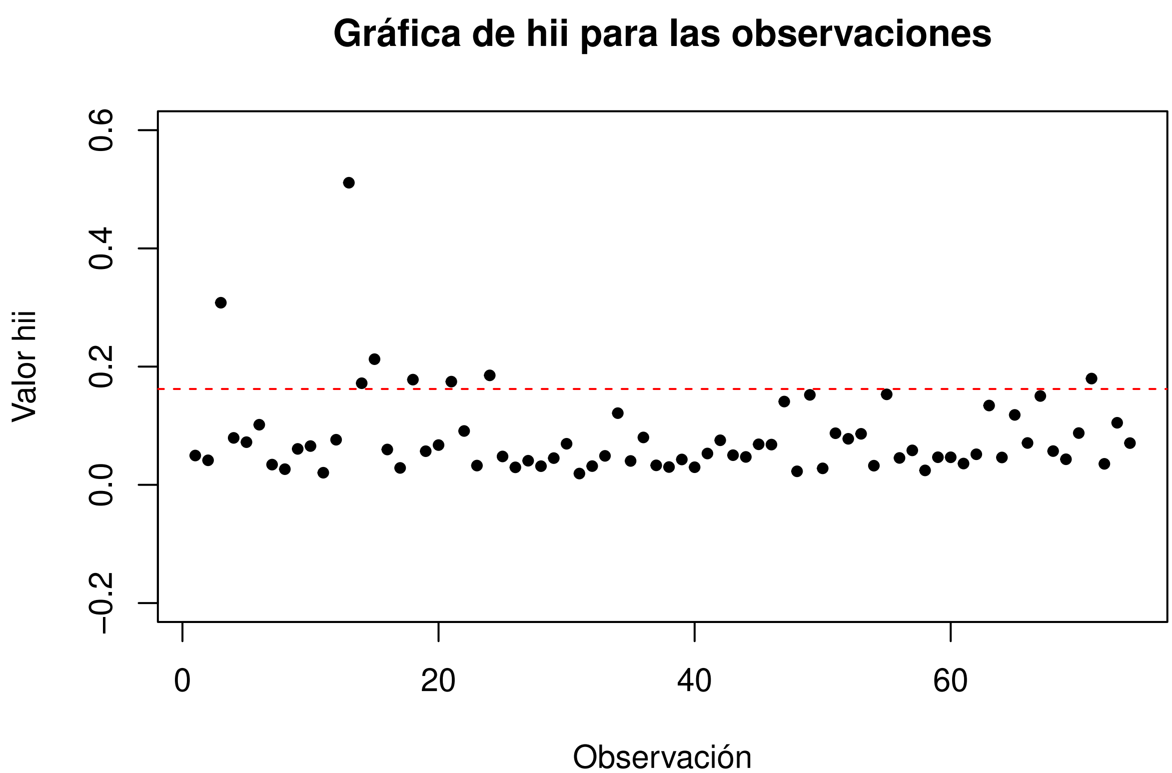


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 3   -3.0767  0.7024    0.3081 -2.1963
## 13  -0.0787  0.0011    0.5111 -0.0799
## 14  -0.2566  0.0023    0.1719 -0.1161
## 15   1.3822  0.0860    0.2126  0.7232
## 18  -0.8303  0.0249    0.1780 -0.3855
## 21  -0.4291  0.0065    0.1746 -0.1962
## 24  -0.3419  0.0044    0.1852 -0.1620
## 71   0.1113  0.0005    0.1798  0.0517
```

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 8 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{6}{74}$, quienes están presentados en la tabla. Estos valores pueden afectar algunas propiedades del modelo, como los errores estándar de los coeficientes estimados y el valor de R^2.

4.2.4. Puntos influenciales

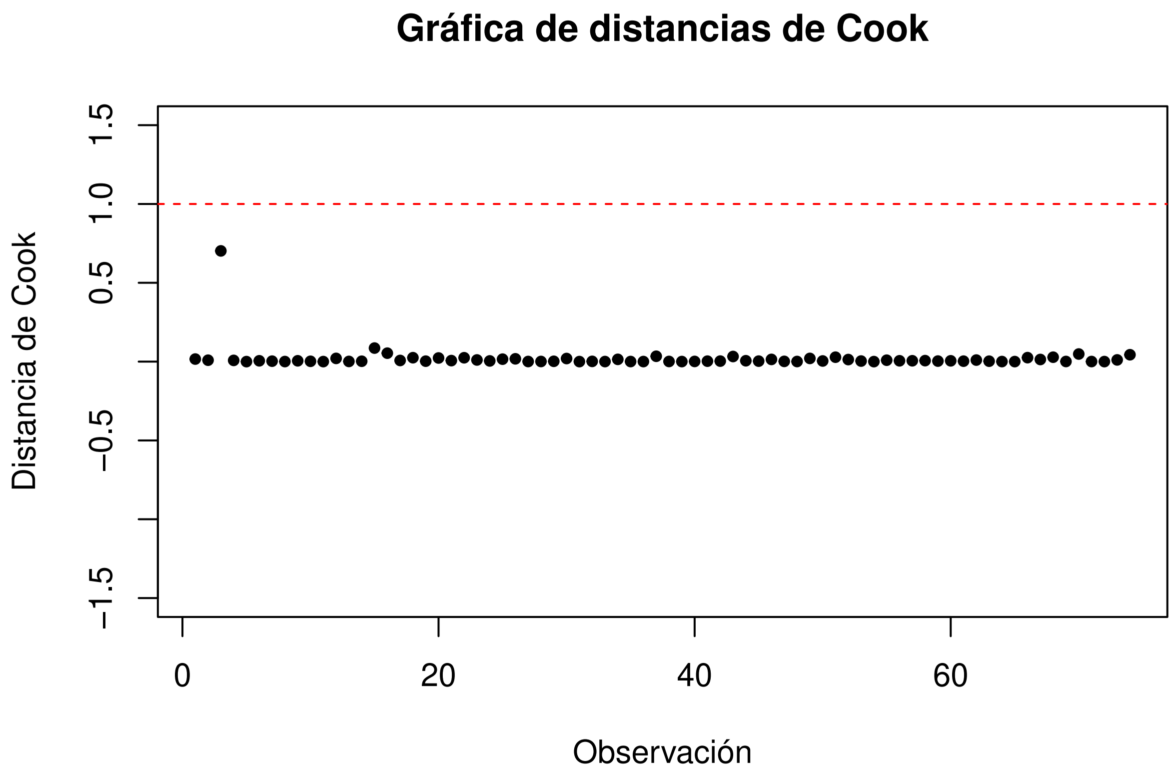


Figura 5: Criterio distancias de Cook para puntos influenciales

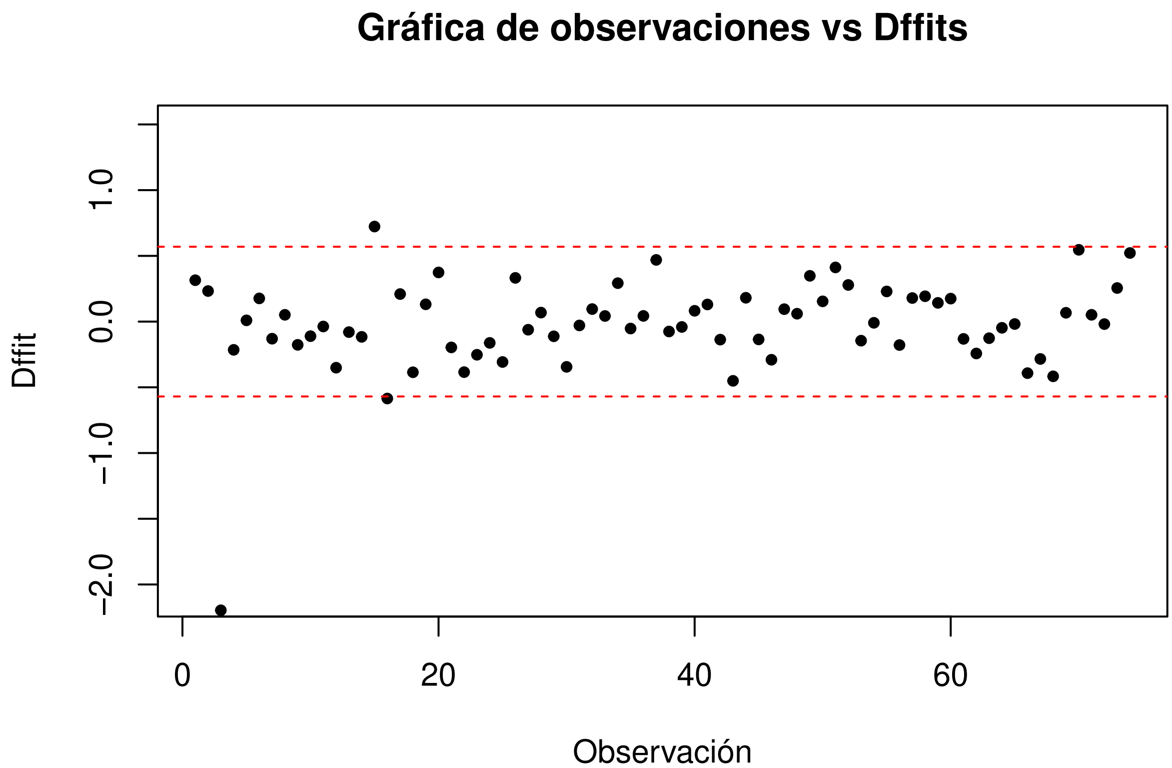


Figura 6: Criterio Dffits para puntos influenciales

4pt

```
##    res.stud Cooks.D hii.value  Dffits
## 3   -3.0767  0.7024    0.3081 -2.1963
## 15   1.3822  0.0860    0.2126  0.7232
## 16  -2.2488  0.0536    0.0598 -0.5853
```

Como se puede observar en la gráfica, las observaciones 3, 15 y 16 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, este se define como observación 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 el criterio, por lo que según este, no se odsevan puntos influenciales. Las observaciones influenciales tienen impacto sobre los coeficientes de la regresión ajustada, causando cambios importantes principalmente en la dirección del modelo.

4.3. Conclusión

2pt

De acuerdo a los supuestos de normalidad y de varianza constante que se deben cumplir para que el modelo sea valido, podemos decir que la normalidad se cumple bajo el analisis de la gráfica Normal QQplot de residuales, y que la varianza se cumple pero bajo el estudio de los valores extremos que esta presenta en la gráfica de Residuales Estudentizados vs Valores Ajustados. Por lo que se concluye que el modelo ajustado explica el riesgo promedio de infeccion en los hospitales en unidades de porcentaje, con parámetros significativos β_3 , β_4 y β_5.

Pero es válido o no finalmente?