Trabajo 1

4,5

Estudiantes

Amilder Stewin Ospina Tobón
Leydi Torres Chamorro
Esteban Hernandez Taborda

Equipo 40

Docente

Carlos Mario Lopera

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

20pt

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),\ 1 \leqslant i \leqslant 64$$

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	-0.7438
β_1	0.1984
β_2	0.0157
β_3	0.0625
β_4	0.0107
β_5	0.0018

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -0.7438 + 0.1984 X_{1i} + 0.0157 X_{2i} + 0.0625 X_{3i} + 0.0107 X_{4i} + 0.0018 X_{5i};\ 1 \leqslant i \leqslant 64$$

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:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,58} \qquad (1)$$

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	78.2602	5	15.652041	16.1105	6.19811e-10
Error	56.3496	58	0.971545		

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 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	-0.7438	1.6933	-0.4393	0.6621
β_1	0.1984	0.0818	2.4257	0.0184
β_2	0.0157	0.0317	0.4952	0.6224
β_3	0.0625	0.0141	4.4366	0.0000
β_4	0.0107	0.0069	1.5426	0.1284
β_5	0.0018	0.0008	2.1631	0.0347

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 para el modelo, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

Interpreten sólo los parámetros significativos, respecto a β_0 ya saben que se debe cumplir que el 0 esté en el intervalo

3pt

- $\hat{\beta}_1$: Indica el cambio en la respuesta media Riesgo de infección"por unidad de incremento en la variable "Duración de la estadía", cuando las demás variables predictoras permanecen constantes.
- $\hat{\beta}_3$:Indica el cambio en la respuesta media Riesgo de infección"por unidad de incremento en la variable "Numero de camas", cuando las demás variables predictoras permanecen constantes.
- $\hat{\beta}_5$:Indica el cambio en la respuesta media Riesgo de infección"por unidad de incremento en la variable "Numero de enfermeras", cuando las demás variables predictoras permanecen constantes.

1.5. Coeficiente de determinación múltiple R^2

3pt

El modelo tiene un coeficiente de determinación múltiple

$$\mathbf{R^2} = \frac{SSR}{SST} = \frac{78.2602}{78.2602 + 56.3496} = 0.5814$$

lo que significa que aproximadamente el 58.14 % de la variabilidad total observada del porcentaje de Riesgo de infección 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_4 y X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	56.350	X1 X2 X3 X4 X5
Modelo reducido	98.712	X2 X4

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

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{14.12066667}{0.97155172} \\ &= 14.53413783 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,58} = 2.7636$, se puede ver que $F_0 > f_{0.95,3,58}$ y por tanto qué se rechaza la hipítesis nula y se concluye que al menos una de las variables del subconjunto es significativa. En este sentido no es posible descartar las variables del subconjunto.

A pesar de que se sabe a priori (desde la Tabla 4) que los parámetros de este subconjunto son todos significativos, desde esta prueba solamente podemos concluir que al menos uno de los parámetros es significativo.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la prueba de hipótesis donde:

¿El efecto de la "duración de la estadía" sobre el "riesgo de infección" es equivalente al efecto que la "rutina de cultivos" tiene sobre el "riesgo de infección"?

¿El efecto que el "número de camas" tiene sobre el "riesgo de infección" es equivalente a 2 veces el efecto que el "censo promedio diario" tiene sobre el "riesgo de infección"?

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

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

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

2 pt

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_3 X^*_{3i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

6 pt

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

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

3.2. Estadístico de prueba

Bajo un alpha = 0.05, el estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{0.950,2,58} \tag{3}$$

2 pt

$$F_0 = \frac{(SSE(MR) - 78.2602)/2}{15.652041} \overset{H_0}{\sim} f_{0.950,2,58} \tag{4}$$

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

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

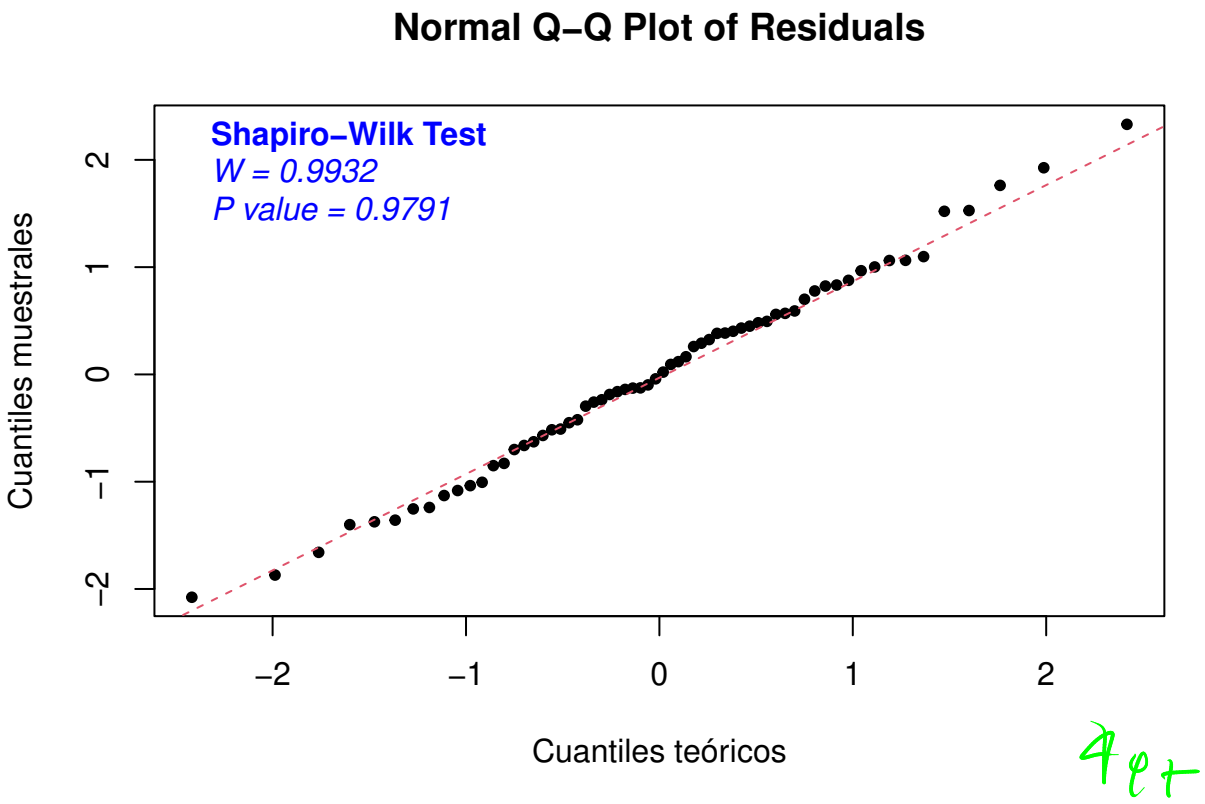
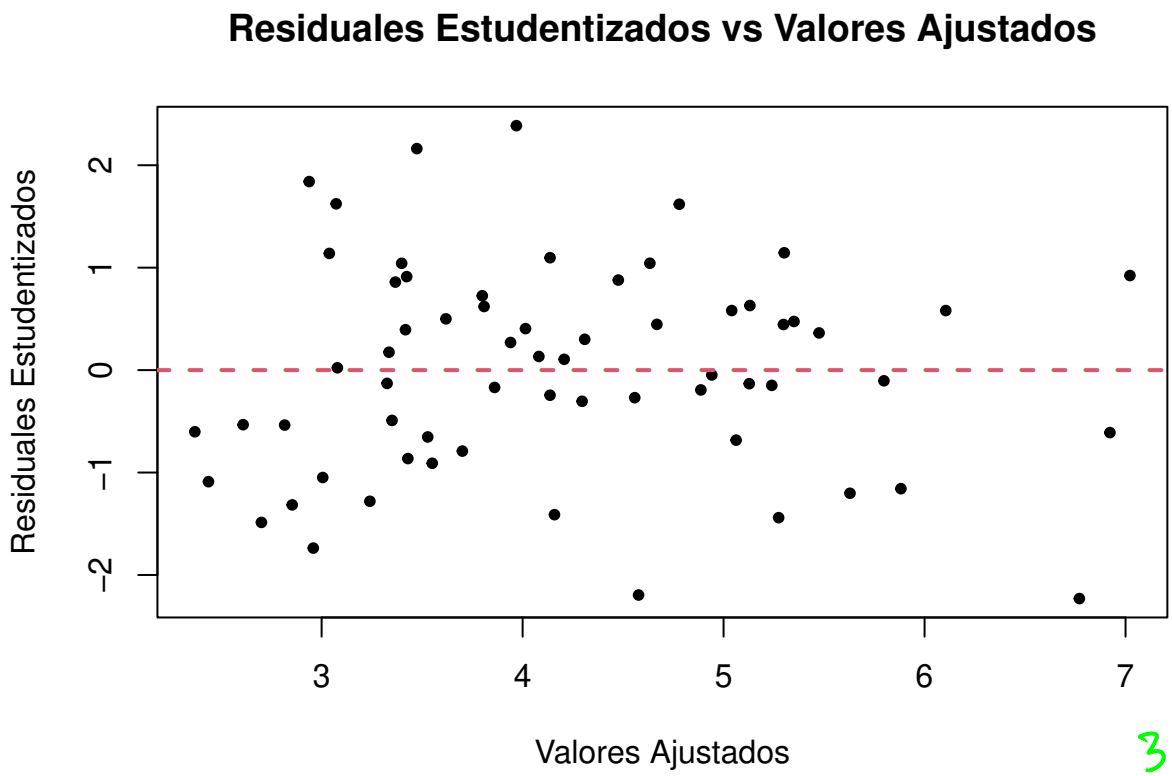


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.9791 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. A pesar de que la grafica denota ciertas desviaciones en las colas, se determina que esto no es suficiente como para rechazar este supuesto. Por lo tanto no rechazamos H0.

Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante



3 pt

Figura 2: Gráfico residuales estudentizados vs valores ajustados

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 no haber evidencia suficiente en contra de este supuesto se acepta como cierto. Además es posible observar media 0.

-1 pt

4.2. Verificación de las observaciones

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

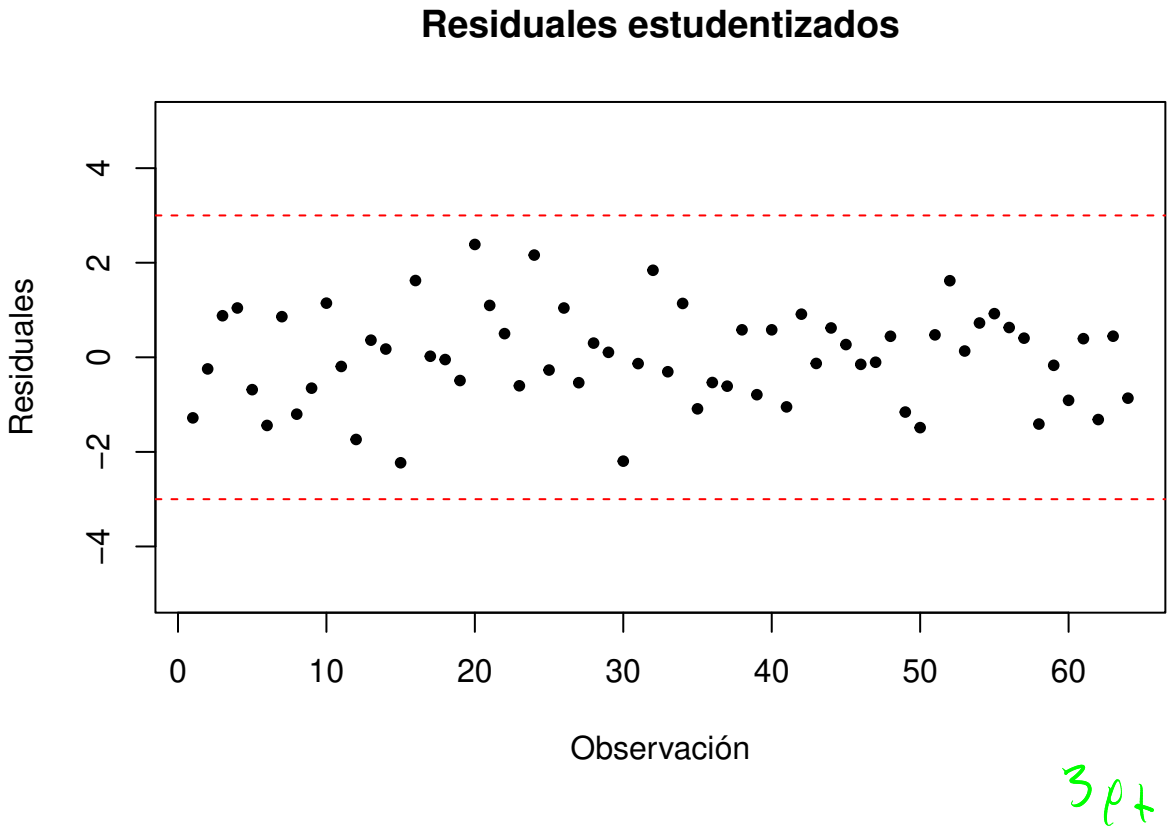


Figura 3: Identificación de datos atípicos

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 $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

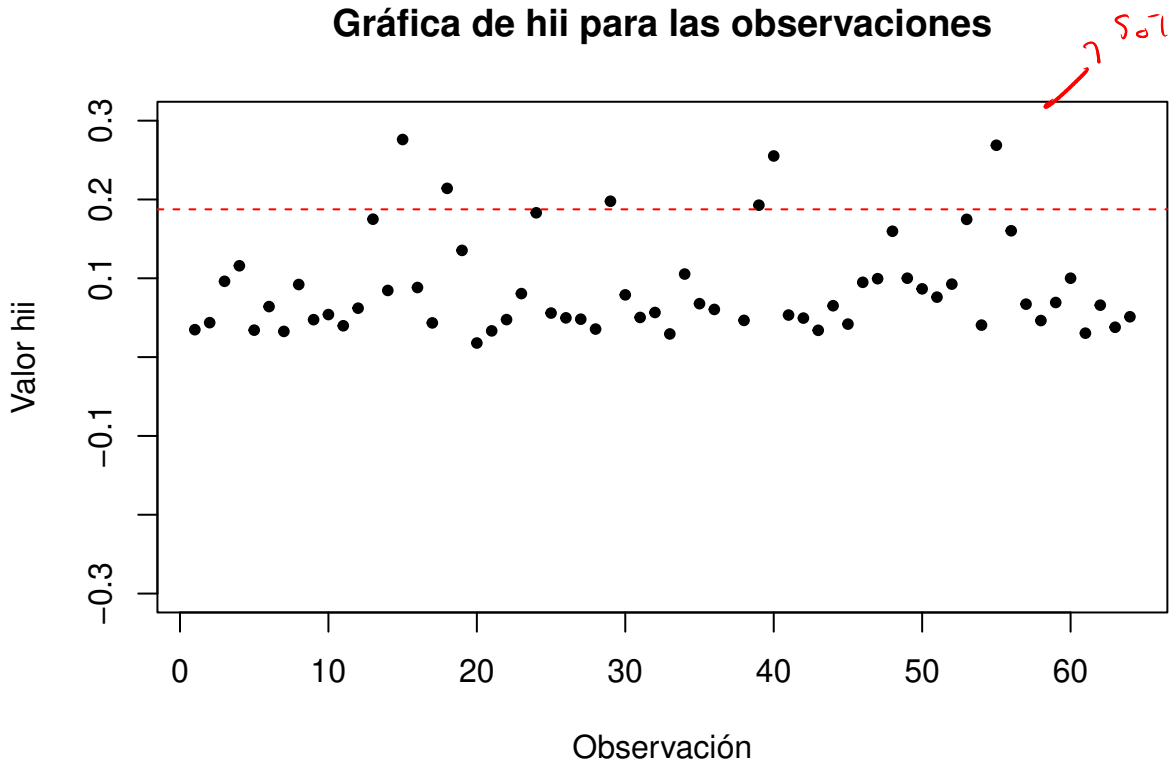


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 15  -2.2303  0.3163    0.2761 -1.4283
## 18  -0.0472  0.0001    0.2141 -0.0244
## 29   0.1054  0.0005    0.1978  0.0519
## 37   0.6101  0.0636    0.5063 -0.6144
## 39  -0.7907  0.0249    0.1928 -0.3852
## 40   0.5812  0.0193    0.2552  0.3383
## 55   0.9230  0.0522    0.2688  0.5588
```

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 2\frac{6}{64} = 0.1875$, se puede apreciar que existen 7 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.

4.2.3. Puntos influenciales

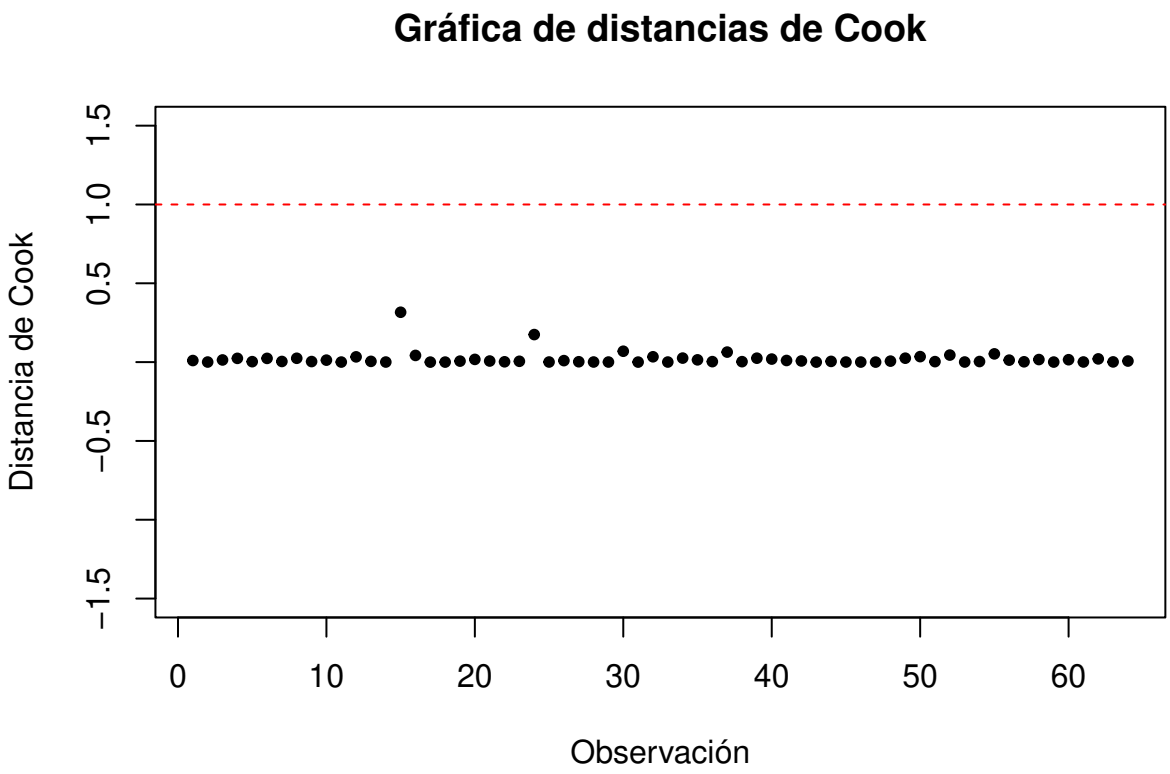


Figura 5: Criterio distancias de Cook para puntos influenciales

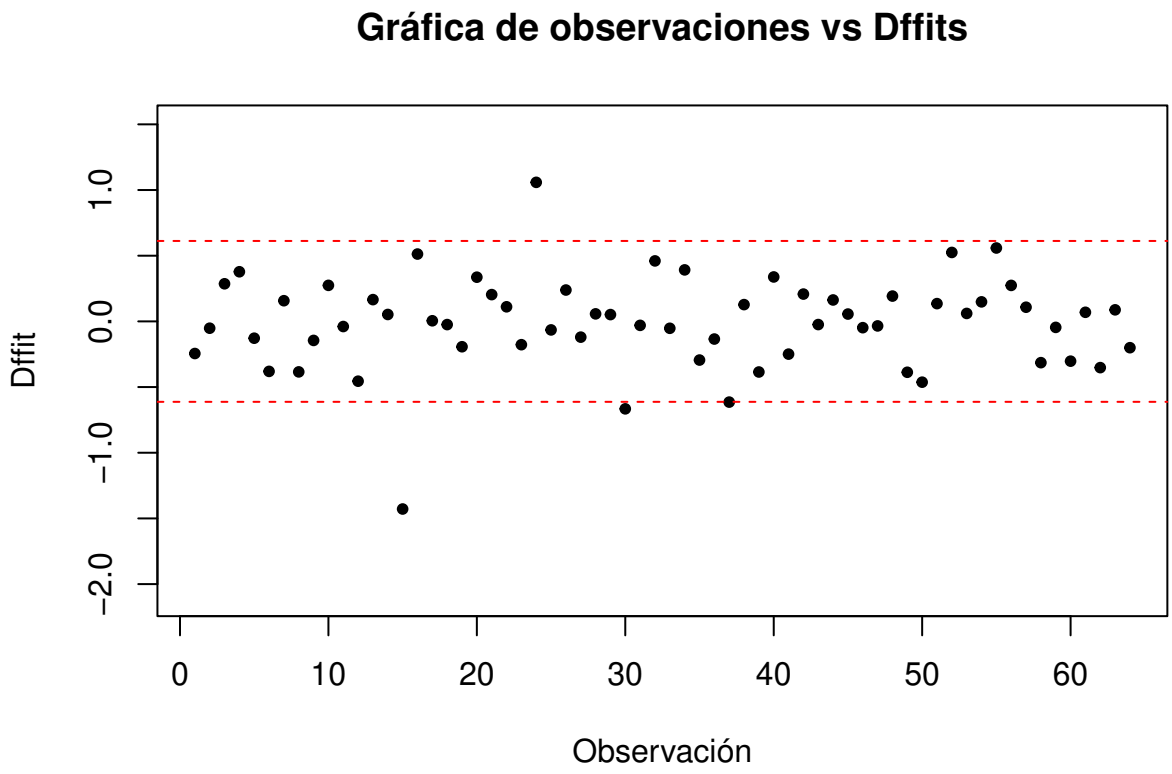


Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 15  -2.2303  0.3163    0.2761 -1.4283
## 24   2.1623  0.1747    0.1831  1.0586
## 30  -2.1958  0.0690    0.0790 -0.6659
## 37  -0.6101  0.0636    0.5063 -0.6144
```

Causan...?

3 pt

Como se puede ver,las observaciones 15, 24, 30 y 37 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, ninguno de los datos cumple con serlo.

4.3. Conclusión

3 pt

Según los resultados obtenidos anteriormente, podemos concluir que el modelo es válido pues se cumplen los supuestos de los errores en los criterios de normalidad y de varianza constante. Adicionalmente, gracias al no rechazo de la hiótesis núla de significancia corroboramos que el modelo es significativo con 3 variables que son individualmente significativas frente a las demás. Estas son: X1: “Duración de la estadía”, X3: “Número de camas” y X5:

"Número de enfermeras". Observamos que no hay valores atípicos en el modelo más si 7 puntos de balanceo y 4 influenciales. Consideramos que estos no son suficientes como para invalidar este modelo.