9,25

Trabajo 1

Estudiantes

Sergio Alejandro Alvarez Quijano
Freddy Quintero Colorado
Angélica María Arce Parra
Yesica Andrea Henao Ceballos

Equipo #40

Docente

Francisco Javier Rodriguez Cortes

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

19,5 pt

Correspondiente a la base de datos asignada al equipo 40, la cual es Equipo40.txt, se consideran las 5 variables regresoras denominadas por:

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

Por lo que se plantea el siguiente modelo de regresión lineal múltiple (RLM):

$$Y_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,$$
$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$
$$donde\ i = 1, 2, 3....., 65$$

1.1. Modelo de regresión

Se obtienen los siguientes valores para los parámetros estimados del modelo de RLM:

Cuadro 1: Tabla de parámetros estimados del modelo

	Valor del parámetro
$\hat{\beta}_0$	-1.3025
$\hat{\beta}_1$	0.1908
$\hat{\beta}_2$	0.0300
$\hat{\beta}_3$	0.0379
$\hat{\beta}_4$	0.0152
$\hat{\beta}_5$	0.0014

3 pt

Por lo que, se obtiene que la ecuación ajustada de regresión es:

$$\hat{Y}_i = -1.3025 + 0.1908 X_{i1} + 0.03 X_{i2} + 0.0379 X_{i3} + 0.0152 X_{i4} + 0.0014 X_{i5}$$

$$donde\ i = 1, 2, 3...., 65$$

1.2. Significancia de la regresión

Se plantea el siguiente juego de hipótesis, con el fin de probar la significancia de la regresión:

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

Donde su estadístico de prueba esta denotado por:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,59} \tag{1}$$

A continuación, se ilustra la correspondiente tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	S.cuadrados	Grados de libertad	S.Cuadrados Medios	F_0	Valor P
Regresión	40.1651	5	8.033028	9.36639	1.31078e-06
Error	50.6010	59	0.857644		

Según la tabla ANOVA, se concluye que el modelo de regresión es significativo, ya que el valor P es menor que el nivel de significancia $\alpha = 0.05$, lo que implica el rechazo de la hipotesis nula H_0.Por tanto, es posible afirmar que el riesgo de infección depende significativamente de al menos una de las variables predictoras del modelo propuesto.

1.3. Significancia de los parámetros

En el cuadro 3 se muestra la información acerca de los valores de los parámetros estimados, junto con sus respectivos valores P. Por consiguiente, se plantea el siguiente juego de hipotesis:

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

A continuación se presenta el cuadro de resumen de los coeficientes

Cuadro 3: Resumen de los parámetros estimados

	$\hat{\beta}_j$	$se(\hat{\beta}_j)$	T_{0j}	Valor P
β_0	-1.3025	1.6626	-0.7834	0.4365
β_1	0.1908	0.1018	1.8734	0.0660
β_2	0.0300	0.0292	1.0262	0.3090
β_3	0.0379	0.0137	2.7558	0.0078
β_4	0.0152	0.0066	2.3211	0.0238
β_5	0.0014	0.0007	2.1235	0.0379

En base a los resultados obtenidos en el cuadro de coeficientes,los parámetros β_3, β_4 y β_5 son significativos individualmente, debido a que sus respectivos valores P son menores que el nivel de significancia $\alpha = 0.05$. Por otro lado, β_0,β_1 y β_2, no son significativos individualmente en presencia de los demás dado que se acepta su hipótesis nula.

1.4. Interpretación de los parámetros

Para la interpretación de los parámetros, se toman en cuenta únicamente los parámetros significativos del modelo, los cual corresponden a β_3,β_4 y β_5.

$\hat{\beta}_3$**: 0.0379**. Indica que por cada unidad que aumente el número de camas en el hospital, el riesgo promedio de infección en el hospital aumenta significativamente en 0.0379 unidades, cuando las demás variables predictoras están fijas.

$\hat{\beta}_4$**: 0.0152**. Señala que por cada unidad que aumente el censo promedio diario,el riesgo promedio de infección en el hospital aumenta significativamente en 0.0152 unidades, cuando las demás variables predictoras están fijas.

$\hat{\beta}_5$**: 0.0014**. Indica que por cada unidad de aumento en el número de enfermeras del hospital, el promedio del riesgo de adquirir una infección en el hospital aumenta significativamente en 0.0014 unidades, cuando las demás variables predictoras están fijas.

1.5. Coeficiente de determinación múltiple R^2

$$R^2 = (40.1651/90.7661) = 0.4425$$

El modelo de regresión lineal múltiple propuesto explica el 44.25 % de la variabilidad total del riesgo de adquirir una infección en el hospital.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las variables con el P-valor más alto en el modelo fueron X_1, X_2, X_5, por lo tanto a través de la tabla de todas las regresiones posibles, se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Variables en el modelo
Modelo completo	50.601	X1 X2 X3 X4 X5
Modelo reducido	63.836	X3 X4

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

$$Y_i = \beta_0 + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i; \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,59} \\ &= \frac{(63.836 - 50.601)/3}{0.857644} \\ &= 5.143936956 \end{aligned} \tag{2}$$

Dado que el estadístico de prueba F_0 es mayor a $f_{0.05,3,59} = 2.7608$, se debe rechazar la hipótesis nula, lo que indica que no es posible descartar las variables del subconjunto, es decir, el riesgo de infección depende de al menos una de las variables del subconjunto escogido (Duración de la estadía, Rutina de cultivos, Número de enfermeras).

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se quiere probar si las variables X_1 (Duración de la estadía), X_2(Rutina de cultivos) y X_3(Número de camas), X_5 (Número de enfermeras), son linealmente dependientes o no, es decir, si son redundantes o no.

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

De forma matricial, tenemos que:

$$\begin{cases} \mathrm{H}_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ \mathrm{H}_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Con **L** y β dados por:

$$L = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -2 \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1(X_{i1} + X_{i2}) + \beta_4 X_{i4} + \beta_5(2X_{i3} + X_{i5}) + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

$$Y_i = \beta_o + \beta_1 X^*_{i1} + \beta_4 X_{i4} + \beta_5 X^*_{i5} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

Donde $X^*_{i1} = X_{i1} + X_{i2}$ y $X^*_{i5} = 2X_{i3} + X_{i5}$.

3.2. Estadístico de prueba

De acuerdo con los datos conocidos,el estadístico de prueba F_0 es el siguiente:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} = \frac{(SSE(MR) - 50.6010)/2}{0.857644} \overset{H_0}{\sim} f_{2,59} \qquad (3)$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Se planteará la siguiente prueba de hipótesis Shapiro-Wilk, acompañada de un Q-Q Plot de residuales, para verificar la normalidad de los residuales:

errores

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

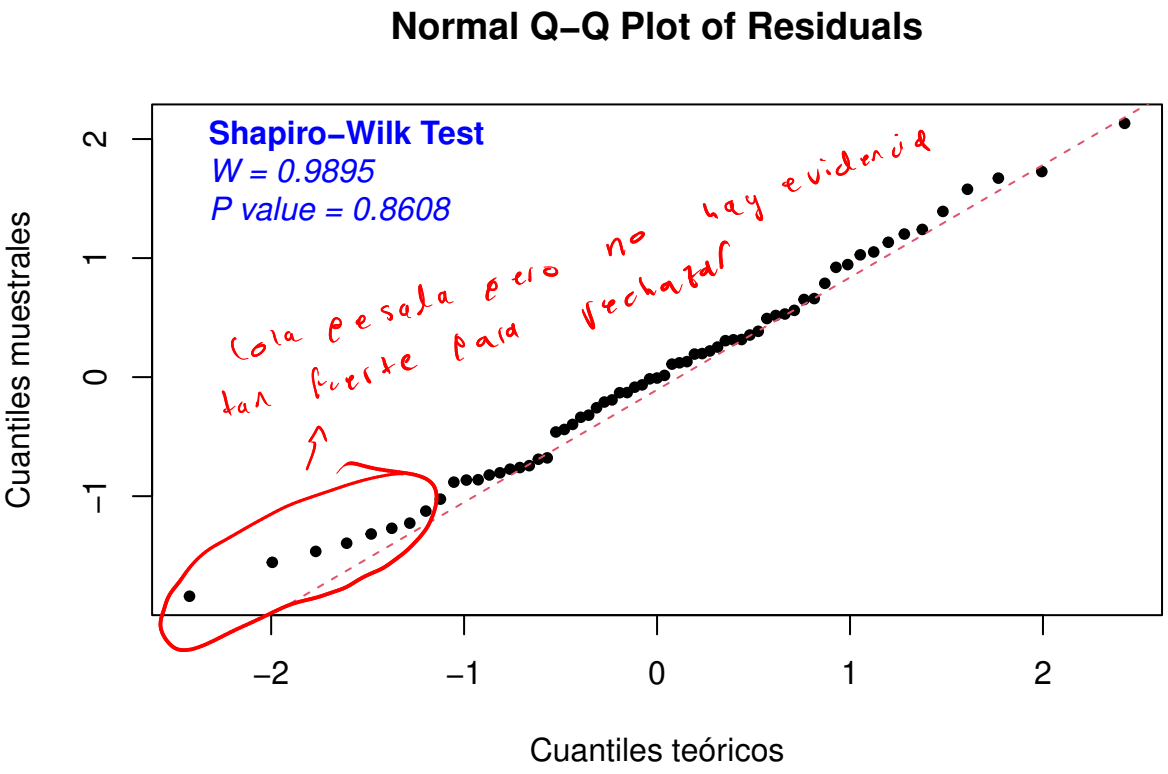

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

De la prueba de normalidad, se obtuvo un estadístico de prueba con un valor de 0.9895 y un valor P de 0.8608. Teniendo en cuenta esto, se puede concluir a un nivel de significancia $\alpha = 0.05$, que la hipótesis nula no se rechaza, por ende, el supuesto de normalidad para los errores del modelo se cumple.

No tienen en cuenta el gráfico, que es más importante

opt

4.1.2. Varianza constante

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} H_0: \text{Varianza} = \sigma^2 \\ H_1: \text{Varianza} \neq \sigma^2 \end{cases}$$

de los ε_i

$$\begin{cases} H_0: Var[\varepsilon_i] = \sigma^2 \\ H_1: Var[\varepsilon_i] \neq \sigma^2 \end{cases}$$

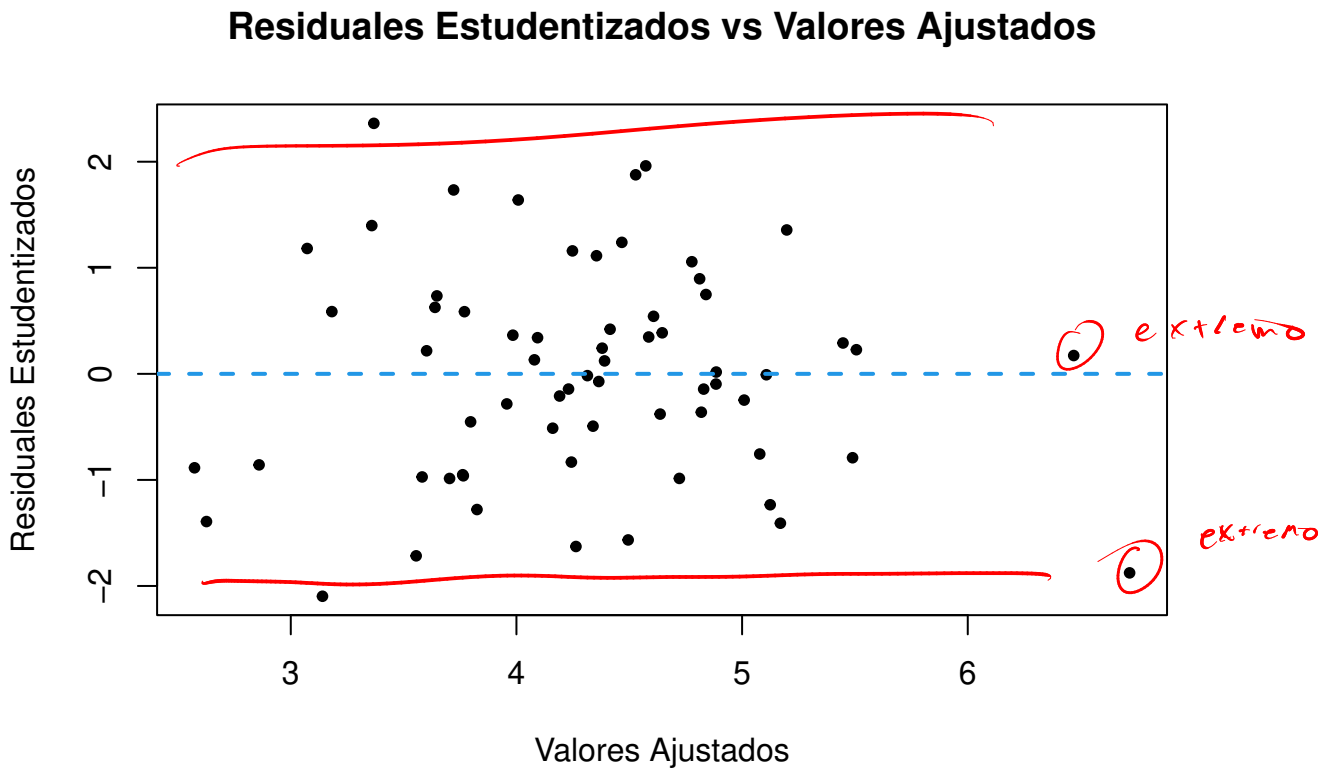

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En estadística afirmen no anden suponiendo cuando tienen evidencia.

A partir del gráfico anterior, se puede suponer que la varianza de los errores no es constante, dado que se evidencia una heterocedasticidad de los residuales estudentizados versus los valores ajustados.

Sus residuales no tienen evidencia fuerte en contra de este supuesto. Análisis muy poco profundo

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

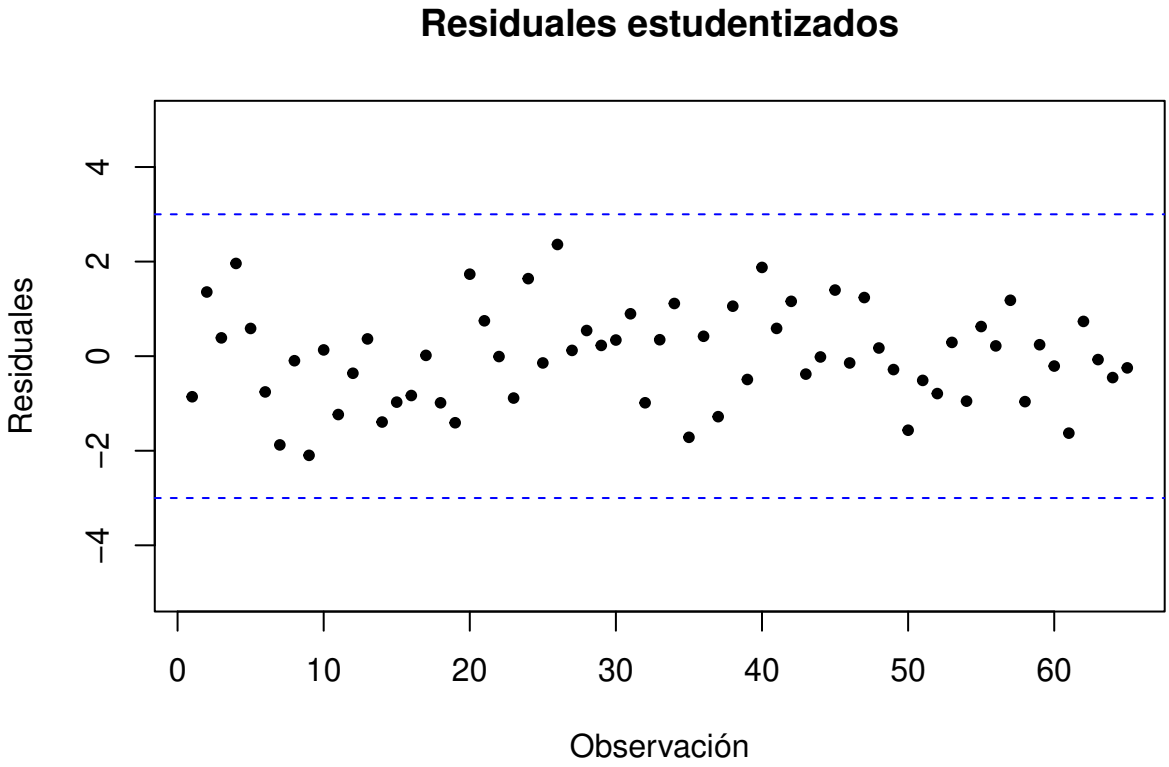

Figura 3: Identificación de datos atípicos

De acuerdo a lo observado en la gráfica anterior, no hay datos atípicos en el conjunto de datos dado que ningún residual estudentizado está fuera del invervalo I=(-3,3).

4.2.2. Puntos de balanceo

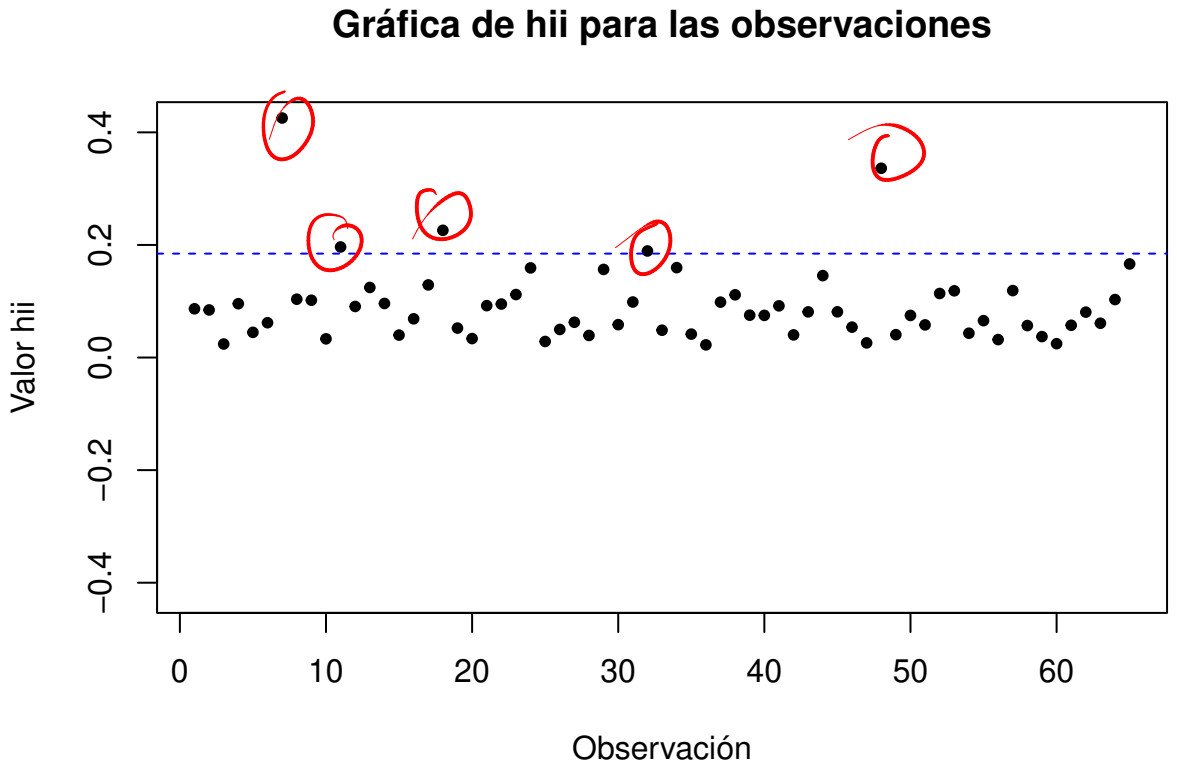

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Tabla de puntos de balanceo.

	Residuales estudentizados	Cook.D	hii	Dffits
7	-1.8767	0.4346	0.4254	-1.6511
11	-1.2343	0.0620	0.1964	-0.6129
18	-0.9865	0.0474	0.2260	-0.5330
32	-0.9859	0.0378	0.1893	-0.4763
48	0.1727	0.0025	0.3362	0.1219

Utilizando la gráfica de valores h_{ii} vs valores observados y los resultados obtenidos en el cuadro 5 , es posible notar que existen 5 puntos de balanceo, ya que estos se encuentran por encima de la recta $2\frac{p}{n} = 0.1846$ en color azul. Dichos puntos corresponden a las observaciones 7, 11,18,32 y 48. Estas observaciones no afectan los coeficientes de regresión ajustados, pero sí algunas propiedades del modelo como lo son el coeficiente de determinación múltiple y los errores estándar de dichos coeficientes.

4.2.3. Puntos influenciales

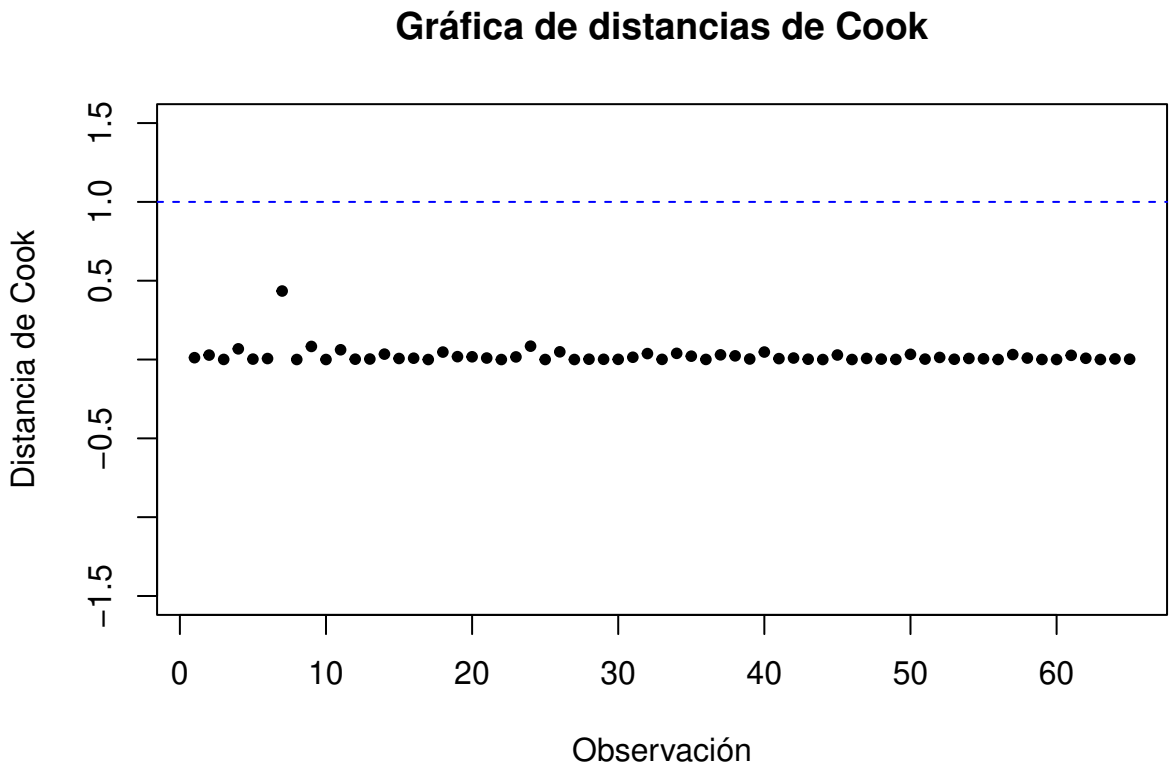

Figura 5: Criterio distancias de Cook para puntos influenciales

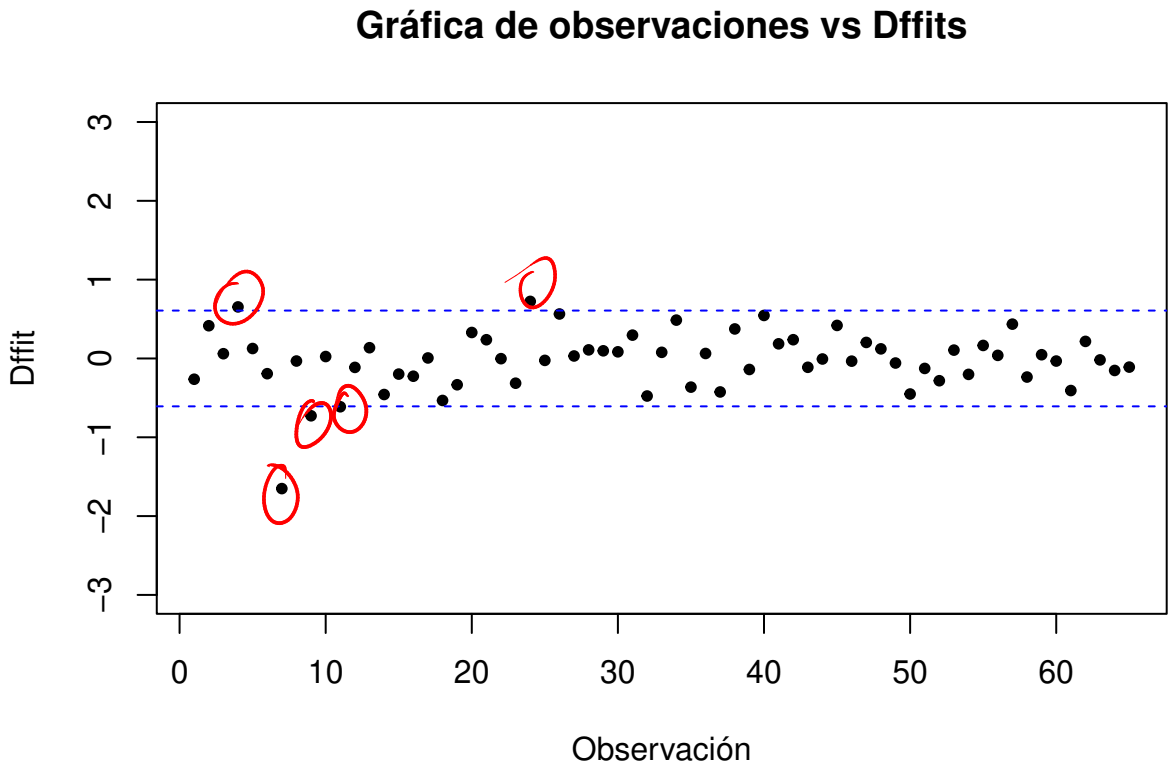

Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Tabla de puntos influenciales según los criterios Dffits y distancias de Cook.

	Residuales estudentizados	Cook.D	hii	Dffits
4	1.9609	0.0678	0.0956	0.6539
7	-1.8767	0.4346	0.4254	-1.6511
9	-2.0970	0.0830	0.1018	-0.7275
11	-1.2343	0.0620	0.1964	-0.6129
24	1.6392	0.0848	0.1592	0.7239

3,5 pt

A partir del criterio de distancias de Cook, $D_i > 1$, se puede concluir que no hay ninguna observación influencial; esto de acuerdo con la figura 5 y la columna Cooks. D de la tabla anterior. Por otro lado, con el criterio Dffits y los resultados obtenidos en la tabla anterior, es posible afirmar que las observaciones 4,7,9,11 y 24 son influenciales, dado que su $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.6076$.Estas observaciones tienen un impacto significativo sobre los coeficientes de regresión ajustados, por ende se sugiere que sean investigados.

4.3. Conclusión

2,5 pt

Finalmente, se puede afirmar que el supuesto de normalidad para los errores del modelo se cumple,sin embargo, no ocurre lo mismo con el supuesto de varianza constante.

Entonces es válido el modelo o no?

Además,en cuanto a las observaciones extremas se tiene que no hay observaciones atípicas, hay 5 de balanceo y 5 influenciales. Por tanto, el modelo de RLM propuesto no tiene validez para realizar estimaciones o predicciones acerca del riesgo de adquirir una infección en un hospital en Estados Unidos. Se recomienda aplicar las técnicas vistas en clase, como las transformaciones, que permitan estabilizar la varianza y tener un modelo adecuado para explicar el comportamiento del riesgo de infección.

La validez no la dan los puntos extremos. Debieron concluir que no era válido por no cumplir supuestos, respecto a los extremos, pueden afectar el análisis de supuestos pero como no la mencionaron...