4,3

Trabajo 1

Estudiantes

Elizabeth Laverde Sierra
Maria Clara González Arismendi
Paula Andrea Cifuentes David
Daniel Felipe Zapata Patron
Equipo 11

Docente

Julieth Verónica Guarín Escudero

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

Estime un modelo de regresión lineal múltiple que explique el riesgo de infección en términos de las variables restantes (actuando como predictoras). Analice la significancia de la regresión y de los parámetros individuales. Interprete los parámetros estimados. Calcule e interprete el coeficiente de determinación múltiple R2.

1.1. Modelo de regresión

Teniendo en cuenta la base de datos asignada, la cual es **Equipo11.txt**,en la que las covariables son:
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
el modelo de regresión lineal múltiple que se propone es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2);\ 1 \leq i \leq 55$$

al ajustar el modelo anterior, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores de los coeficientes estimados

	valor del parámetro
$\hat{\beta}_0$	-2.2385
$\hat{\beta}_1$	0.0579
$\hat{\beta}_2$	0.0484
$\hat{\beta}_3$	0.0317
$\hat{\beta}_4$	0.0284
$\hat{\beta}_5$	0.0021

por lo tanto, el modelo de regresión ajustado es:

$$\widehat{Y}_i = -2.2385 + 0.0579 X_{1i} + 0.0484 X_{2i} + 0.0317 X_{3i} + 0.0284 X_{4i} + 0.0021 X_{5i}$$

1.2. Significancia de la regresión

Para la significancia de la regresión se hará uso de la siguiente tabla anova:

Cuadro 2: Tabla anova de significancia de la regresión

	Sumas de cuadrados	g.l	Cuadrado medio	F_0	Valor-P
Modelo de regresión	52.4755	5	10.495096	11.668	1.87366e-07
Error	44.0743	49	0.899476		

Para analizar la significancia de la regresión se establece la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \text{ para } j = 1, 2, ..., 5 \end{cases}$$

cuyo estadístico de prueba es:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,49} \tag{1}$$

Con un valor de significancia de $\alpha = 0,05$ y a partir de la tabla Anova se concluye que el vP < 0.05, por lo que se rechaza la hipótesis nula por lo cual es posible afirmar que el modelo de regresión lineal múltiple sí es significativo. Esto quiere decir, que el riesgo de infección depende significativamente de al menos una de las predictoras del modelo.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta la información de los parámetros, la cuál permitirá determinar los parámetros individuales que son significativos o no para el modelo.

Cuadro 3: Resumen de los coeficientes

	Estimación β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	-2.2385	1.4607	-1.5325	0.1318
β_1	0.0579	0.1028	0.5638	0.5755
β_2	0.0484	0.0267	1.8081	0.0767
β_3	0.0317	0.0152	2.0904	0.0418
β_4	0.0284	0.0096	2.9460	0.0049
β_5	0.0021	0.0008	2.4583	0.0175

Estas pruebas establecen la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_j = 0 \\ \mathrm{H}_a : \beta_j \neq 0 \ j = 1, 2, ..., 5 \end{cases}$$

De acuerdo al valor p, en el cuadro 3 y teniendo una significancia de $\alpha = 0.05$ se puede concluir que β_3, β_4, β_5 son significativos cada uno en presencia de los demás parámetros.

Asimismo, es posible deducir que β_0, β_1, β_2 son parámetros individualmente no significativos cada uno en presencia de los demás parámetros.

1.4. Interpretación de los parámetros

- $\hat{\beta}_3$ indica que por cada unidad que aumenta de número de camas, el promedio del riesgo de infección aumenta en 0.0317 cuando las demás se mantienen fijas.
- $\hat{\beta}_4$ indica que por cada unidad que aumenta de censo promedio diario, el promedio del riesgo de infección aumenta en 0.0284 cuando las demás se mantienen fijas.
- $\hat{\beta}_5$ indica que por cada unidad que aumenta de número de enfermeras, el promedio del riesgo de infección aumenta en 0.0021 cuando las demás se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

A partir de la tabla ANOVA calculamos el

$$R^2 = \frac{SSR}{SST} = \frac{52.4755}{52.4755 + 44.0743} = 0.5435$$

lo que quiere decir que aproximadamente el 54.35 % de la variabilidad total en el riesgo de infección, es explicado por las variables del modelo de regresión lineal múltiple propuesto.

2. Pregunta 2

Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más grandes del punto anterior. Según el resultado de la prueba, ¿es posible descartar del modelo las variables del subconjunto? Explique su respuesta.

2.1. Planteamiento prueba de hipótesis y modelo reducido

Los parámetros cuyos valores P fueron los más altos corresponden a $\beta_1, \beta_2, \beta_3$. Por tanto, se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \neq 0 \text{ para } j = 1, 2, 3 \end{cases}$$

El modelo completo es el definido en la sección 1.1, y el modelo reducido es:

$$\text{MR: } Y_i = \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2);\ 1 \leq i \leq 55$$

Se presenta la siguiente tabla con el resumen de todas las regresiones para plantear el estadístico de prueba:

Cuadro 4: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	44.074	X1 X2 X3 X4 X5
Modelo reducido	50.501	X4 X5

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSE(\beta_3 \beta_4, \beta_5) - SSE(\beta_0, \beta_1, \cdots, \beta_5))/3}{MSE(\beta_0, \beta_1, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,49}$$
$$= \frac{(50.501 - 44.074)/3}{44.074/49} = 2.38$$

entonces $f_{0.95,3,49} = 2.7939$, Ahora bien, teniendo el valor crítico podemos compararlo con el estadístico de prueba: $F_0 = 2.38 < f_{0.95,3,49}$. Finalmente, se concluye que no se rechaza la hipótesis nula H_0, por lo que las variables no son significativas en presencia de los demás parámetros, es decir, se pueden retirar del modelo.

3. Pregunta 3

Plantee una pregunta donde su solución implique el uso exclusivo de una prueba de hipótesis lineal general de la forma $H_0 : L\beta = 0$ (solo se puede usar este procedimiento y no sumas de cuadrados extra).Especifique claramente la matriz L, el modelo reducido y la expresión para el estadístico de prueba (no hay que calcularlo).

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_5 \ \text{ y } \ \beta_1 = \beta_3 \\ \mathrm{H}_a : \beta_2 \neq \beta_5 \ \text{ y } \ \beta_1 \neq \beta_3 \end{cases}$$

Reescribiendo matricialmente:

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

Donde $\mathbf{L}$ está dada por:

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

Por lo que H_0 se puede plantear de la siguiente manera:

$$\mathbf{H_0} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$ ✓

En donde el modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_1 X^*_{1i} + \beta_2 X^*_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i; \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2);\ 1 \leq i \leq 55$$ 0,5 pt

Donde

$$X^*_{1i} = X_{1i} + X_{3i} \quad \text{y} \quad X^*_{2i} = X_{2i} + X_{5i}$$ ✓

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SEE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,49}$$ ✓

2 pt

$$F_0 = \frac{(SSE(MR) - 44.074)/2}{0.8995} \overset{H_0}{\sim} f_{2,49}$$ ✓

4. Pregunta 4 16,5

Realice una validación de los supuestos en los errores y examine si hay valores atípicos, de balanceo e influenciales. ¿Qué puede decir acerca de la validez de éste modelo?. Argumente.

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales 4 pt

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \text{H}_0 : \varepsilon_i \sim Normal. \\ \text{H}_1 : \varepsilon_i \not\sim Normal \end{cases}$$ ✓

Muy bien que se dieron cuenta que shapiro-wilk no es la prueba de hipótesis sino un test.

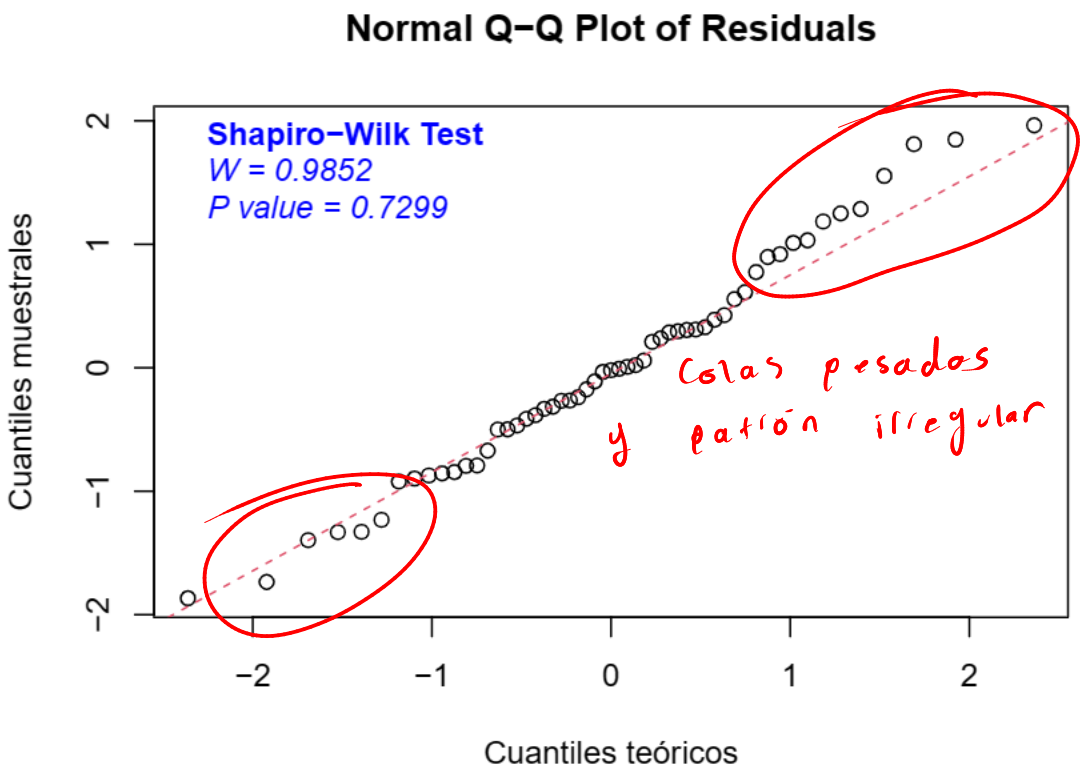

Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Mediante el test de Shapiro-Wilk podemos concluir que el valor-p=0.7299 $> \alpha(0.05)$ por lo que la hipótesis nula no se rechaza, lo que quiere decir que los datos distribuyen de manera normal con media μ y varianza σ^2 sin embargo,es posible notar del análisis de la gráfica anterior, que existen datos dispersos los cuales no siguen la línea roja que representa el ajuste a la distribución de los residuales a una distribución normal. Así, no distribuye de manera normal. → analicen más a fondo q-e hay un patrón y colas pesadas, pero concluyen bien

4.1.2. Varianza constante

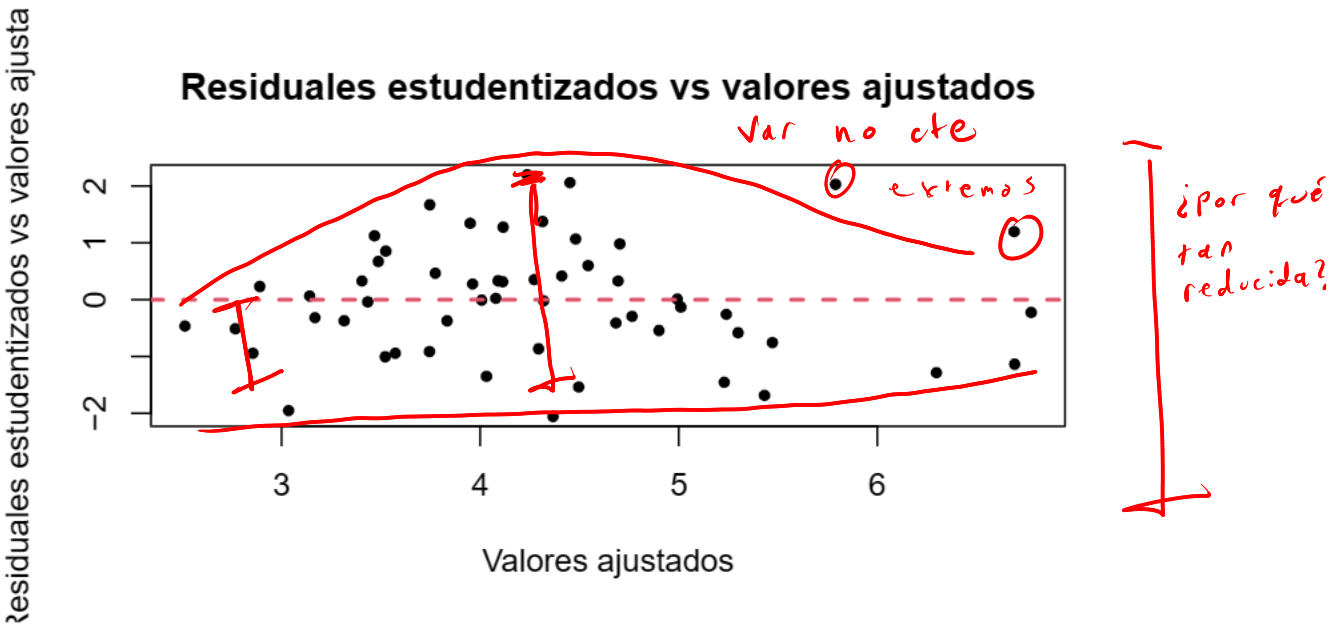

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Después de analizar el gráfico, no encontramos ningún patrón que sugiera un aumento o disminución en la variabilidad de los datos dado que los puntos en la gráfica se encuentran muy dispersos, por lo que no encontramos ninguna señal que indique que la variable es constante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

3 pt

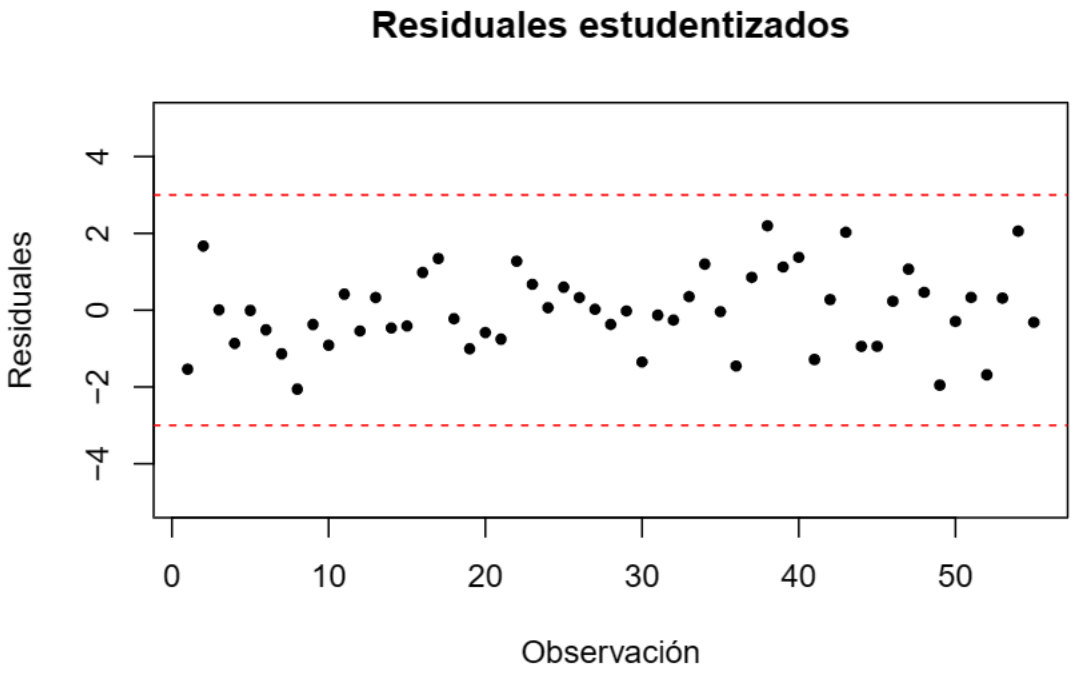

Figura 3: Identificación de datos atípicos

Según el criterio para evaluar datos atípicos, se considera que una observación es atípica cuando $|r_{estud}| > 3$ (el valor absoluto del residual estudiantizado es mayor que 3). En este caso, al observar la gráfica se puede evidenciar que los valores del residual están entre -3 y 3, por lo que se considera que no hay datos atípicos. ✓

4.2.2. Puntos de balanceo

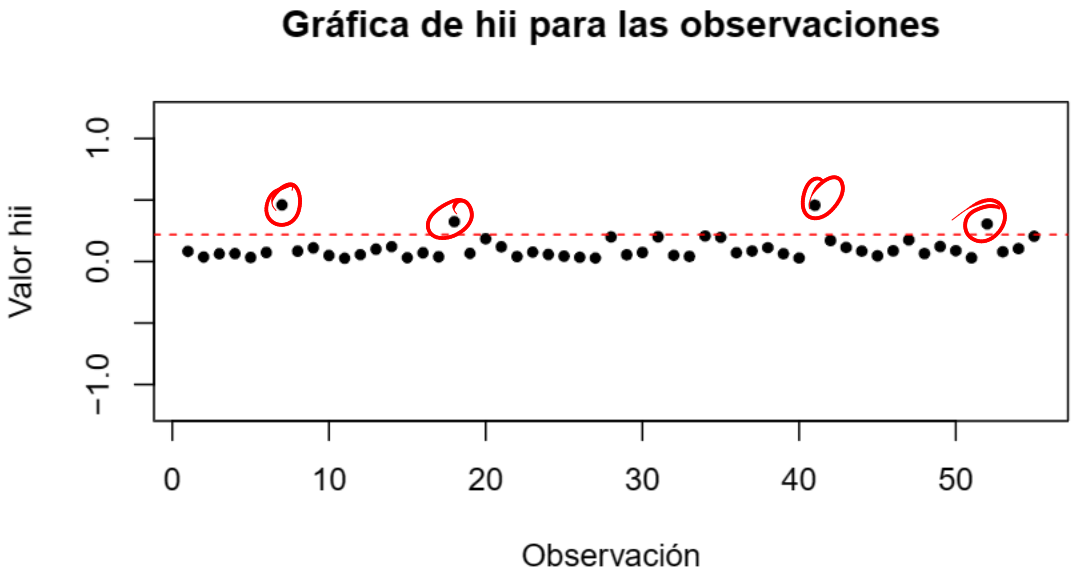

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Tabla de puntos de balanceo

	Residuales estudentizados	D.cook	Valor hii	Dffits
7	-1.1362	0.1835	0.4603	-1.0524
18	-0.2243	0.0040	0.3242	-0.1538
41	-1.2850	0.2321	0.4575	-1.1882
52	-1.6845	0.2075	0.3049	-1.1377

Al observar la gráfica anterior h_{ii} donde la línea punteada roja representa el valor

$$hii > 2p/n = 2p/n = 2(6/55) = 0.21818$$

nos permite notar que existen 4 puntos (7,18,41,52) de balanceo los cuales se pueden evidenciar en el cuadro 5 "tabla de puntos de balanceo".

4.2.3. Puntos influenciales

Bajo el criterio de Cook, se hace la siguiente gráfica:

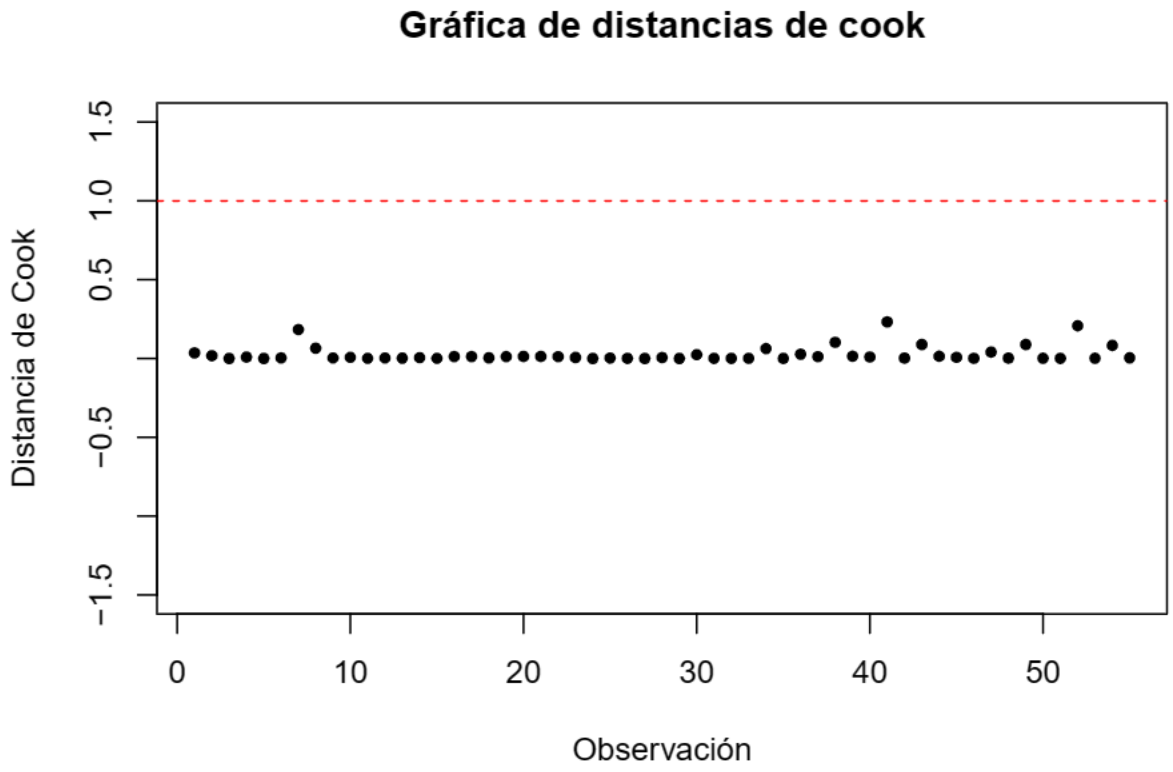

Figura 5: Criterio distancias de Cook para puntos influenciales

Según el criterio de Cook que dice que la observación i será influencial si $Di > 1$, podemos concluir que no hay puntos influenciales.

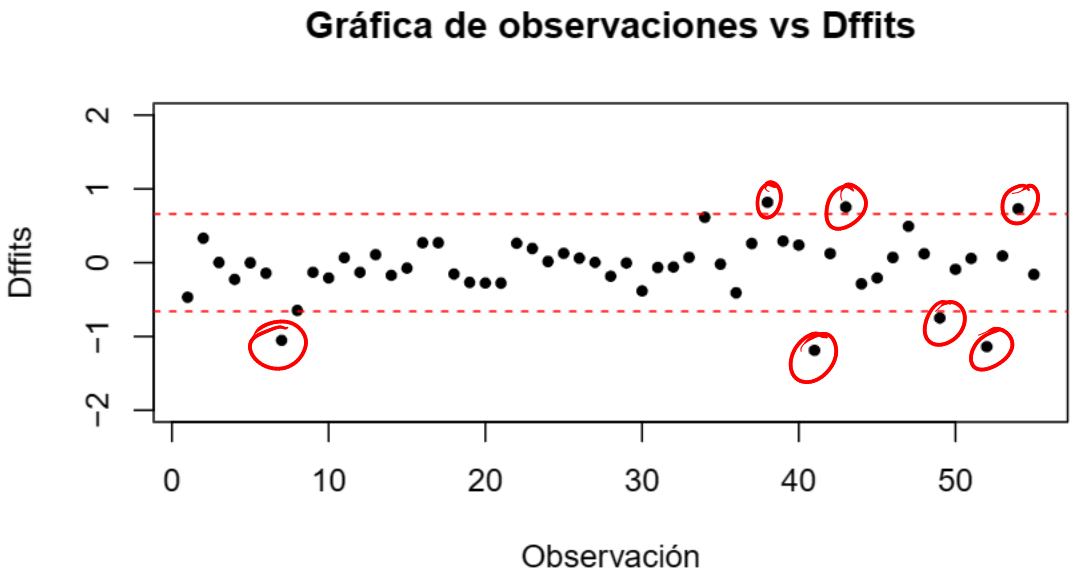

Figura 6: Criterio Dffits para puntos influenciales

De acuerdo con la columna Dffits del cuadro 6 “Tabla de puntos influenciales bajo el criterio

de Dffits" mostrada a continuación y según el criterio de este mismo el cual dice que para cualquier punto cuyo $|Dffits| > 2\sqrt{\frac{p}{n}}$ la observación i será un punto influencial, es decir, si $|Dffits| > 2\sqrt{\frac{p}{n}} = 0.6606$, por lo que las observaciones 7,38,41,43,49,52 y 54 son puntos influenciales.

Cuadro 6: Tabla de puntos influenciales bajo el criterio de Dffits

	Residuales estudentizados	D.cook	Valor hii	Dffits
7	-1.1362	0.1835	0.4603	-1.0524
38	2.1980	0.1026	0.1130	0.8180
41	-1.2850	0.2321	0.4575	-1.1882
43	2.0288	0.0889	0.1147	0.7554
49	-1.9526	0.0884	0.1221	-0.7504
52	-1.6845	0.2075	0.3049	-1.1377
54	2.0593	0.0829	0.1050	0.7303

5. Conclusiones

1. Del modelo de regresión

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2);\ 1 \leq i \leq 55$$

es posible concluir que las variables β_3 β_4 y β_5 correspondientes al número de camas, censo promedio diario y número de enfermeras respectivamente, son indispensables para estimar el porcentaje del riesgo de infección en los hospitales, puesto que estos parámetros son significativos y a su vez interpretables.

2. Al realizar la prueba de significancia para las tres variables predictoras con los valores P más grandes, se encontró que la duración de la estadía $\beta_1 =$, la rutina de cultivos $\beta_2 =$ y el número de camas $\beta_3 =$, no son estadísticamente significativos para predecir el promedio del riesgo de infección hospitalaria, lo que lleva a la conclusión de que estas variables se descartan del modelo por su falta de aporte significativo.

3. Tras la realización de la prueba de hipótesis en forma matricial, fue posible resolver el sistema de ecuaciones que nos llevó a una nueva expresión del modelo reducido utilizando solamente las variables que son significativas

4. Después de realizar un análisis de la validez de los supuestos del modelo, incluyendo la normalidad de los residuales y la varianza constante, se encontró que no se cumplen dichos supuestos. Además, se detectaron puntos de balanceo y puntos influenciales, lo que sugiere que el modelo no es el más adecuado para describir el comportamiento de la variable respuesta (riesgo de infección).