3,25

Trabajo 1

Estudiantes

Santiago Alejandro Picon Vargas
Santiago Barrada González
Maria Alejandra Jaramillo Arroyave
María José Vargas Moreno
Equipo # 18

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

Sede Medellín
27 de Marzo de 2023

Índice

Índice de figuras

Índice de tablas

1. Pregunta 1

Teniendo en cuenta la base de datos asignada, la cual es **Equipo18.txt**, las covariables son: Duración de la estadía (X_1), Rutina de cultivos (X_2), Número de camas (X_3), Censo promedio diario (X_4) y Número de enfermeras (X_5). El modelo que se propone es:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

1.1. Modelo de regresión

Al ajustar el modelo tal, se obtienen los siguientes coeficientes:

Tabla 1: Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	0.2478
$\hat{\beta}_1$	0.2539
$\hat{\beta}_2$	-0.0066
$\hat{\beta}_3$	0.0407
$\hat{\beta}_4$	0.0102
$\hat{\beta}_5$	0.0013

Por lo tanto, el modelo de regresión ajustado es:

$$\widehat{Y}_i = 0.2478 + 0.2539 X_{i1} - 0.0066 X_{i2} + 0.0407 X_{i3} + 0.0102 X_{i4} + 0.0013 X_{i5}$$

Donde $1 \leq i \leq 65$

1.2. Significancia de la regresión

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

Tabla 2: Tabla anova significancia de la regresión

	Sumas de cuadrados	g.l	Cuadrado medio	F_0	Valor-P
Modelo de regresión	61.8836	5	12.376729	13.4831	8.99672e-09
Error	54.1585	59	0.917941		

De la tabla anova,vemos que el valor-P es de 8.99672e-09 $< \alpha = 0.05$; por tanto, podemos concluir que se rechaza la hipótesis nula para la significancia de los parámetros, por lo tanto la regresión es significativa y algún paramétro es distinto de cero, por tanto es significativo.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar que parametros son significativos:

Tabla 3: Resumen de los coeficientes

	Estimación β_j	$se(\hat{\beta}_j)$	T_{0j}	Valor-P
β_0	0.2478	1.8795	0.1319	0.8955
β_1	0.2539	0.0887	2.8641	0.0058
β_2	-0.0066	0.0315	-0.2109	0.8337
β_3	0.0407	0.0124	3.2932	0.0017
β_4	0.0102	0.0071	1.4407	0.1550
β_5	0.0013	0.0006	2.0365	0.0462

los valores-P permiten concluir con una significancia $\alpha = 0.05$ que los parámetros individuales $\hat{\beta}_1$, $\hat{\beta}_3$, $\hat{\beta}_5$ son significativos cada uno de estos en presencia de los demás parámetros.

Adicionalmente, podemos notar que los parámetros $\hat{\beta}_0$,$\hat{\beta}_2$,$\hat{\beta}_4$ son individualmente no significativos en presencia de los demás parámetros.

1.4. Interpretación de los parámetros

- $\hat{\beta}_1$: Por cada unidad que aumente la duración de la estadía, el promedio de riesgo de infección aumenta en 0.2539 unidades, cuando las demás predictorias se mantienen constantes.
- $\hat{\beta}_3$: Por cada unidad que aumente el número de camas, el promedio de riesgo de infección aumenta en 0.0407 unidades, cuando las demás predictoras se mantienen constantes.
- $\hat{\beta}_5$: Por cada unidad que aumente el número de enfermeras, el promedio de riesgo de infección aumenta en 0.0013 unidades, cuando las demás predictoras se mantienen constantes.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un $R^2 = 0.5332$ lo cual significa que aproximadamente el 53.32 % de la variabilidad total del riesgo de infección está explicado por el modelo de regresión lineal múltiple (RLM) propuesto.

1.6. Comentarios

Con los resultados obtenidos para la base de datos **Equipo18.txt**, podemos observar que el modelo para determinar el riesgo de infección, teniendo en cuenta las variables como variables independientes es muy bueno, ya que obtenemos un Valor-P = 8.9967e-09 significativamente pequeño, además de tener 3 parámetros que son $\hat{\beta}_1$,$\hat{\beta}_3$,$\hat{\beta}_5$ muy significativos y un $R^2 = 0.5332$, así pues podemos afirmar que el modelo es aceptable.

2. Pregunta 2

2.1. Planteamiento prueba de hipotesis y modelo reducido

Para probar la significancia simultánea de las tres variables con los valores-P más grandes, equivale a plantear la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_a : \text{Algún } \beta_j \text{ distinto de 0 para } j = 0, 2, 4 \end{cases}$$

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

$$\text{MR: } Y_i = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

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

Tabla 4: Resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	54.159	X1 X2 X3 X4 X5
Modelo reducido	59.509	X1 X3

1,5 pt

2.2. Estadístico de prueba y conclusiones

Se construye el estadístico de prueba como:

$$F_0 = \frac{(SSE(\beta_0,\beta_1,\beta_3) - SSE(\beta_0,\beta_1,\cdots,\beta_5))/2}{MSE(\beta_0,\beta_1,\cdots,\beta_5)} \overset{H_0}{\sim} f_{2,59}$$
$$= \frac{(59.509 - 54.159)/2}{0.9179} = 0.049391$$

0,5 pt

Ahora, comparando a un nivel de significancia $\alpha = 0.05$, F_0 con $f_{0.05,2,59} = 3.153123$. Con valor-P $=0.9518482$

Se concluye que con un valor-P $= 0.9518482$ mayor que un nivel de significancia $\alpha = 0.05$, no rechazamos H_0 por lo tanto, concluimos que las variables X_2,X_4,X_5 se pueden descartar del modelo.

3,5 pt

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 + \beta_4,\ \beta_3 = \beta_5 \\ \mathrm{H}_a : \text{Alguna de las desigualdades no se cumple} \end{cases}$$

Reescribendo 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:

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

2 pt

Donde el modelo reducido está dado por:

$$\text{MR: } Y_i = \beta_0 + (\beta_2+\beta_4)X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i,\quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$$

0,5 pt

$$\text{MR: } Y_i = \beta_0 + \beta_2(X_1+X_2) + \beta_3(X_3+X_5) + \beta_4(X_1+X_4) + \varepsilon_i,\quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,59} \qquad (1)$$

1 pt

Reemplazar lo que conocen

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se plantea la siguiente prueba de hipótesis ~~(shapiro wilk)~~... acompañado de un grafico cuantil-cuantil:

$$\begin{cases} \mathrm{H}_0 : X \sim Normal(\mu, \sigma^2) \\ \mathrm{H}_a : X \nsim Normal(\mu, \sigma^2) \end{cases}$$

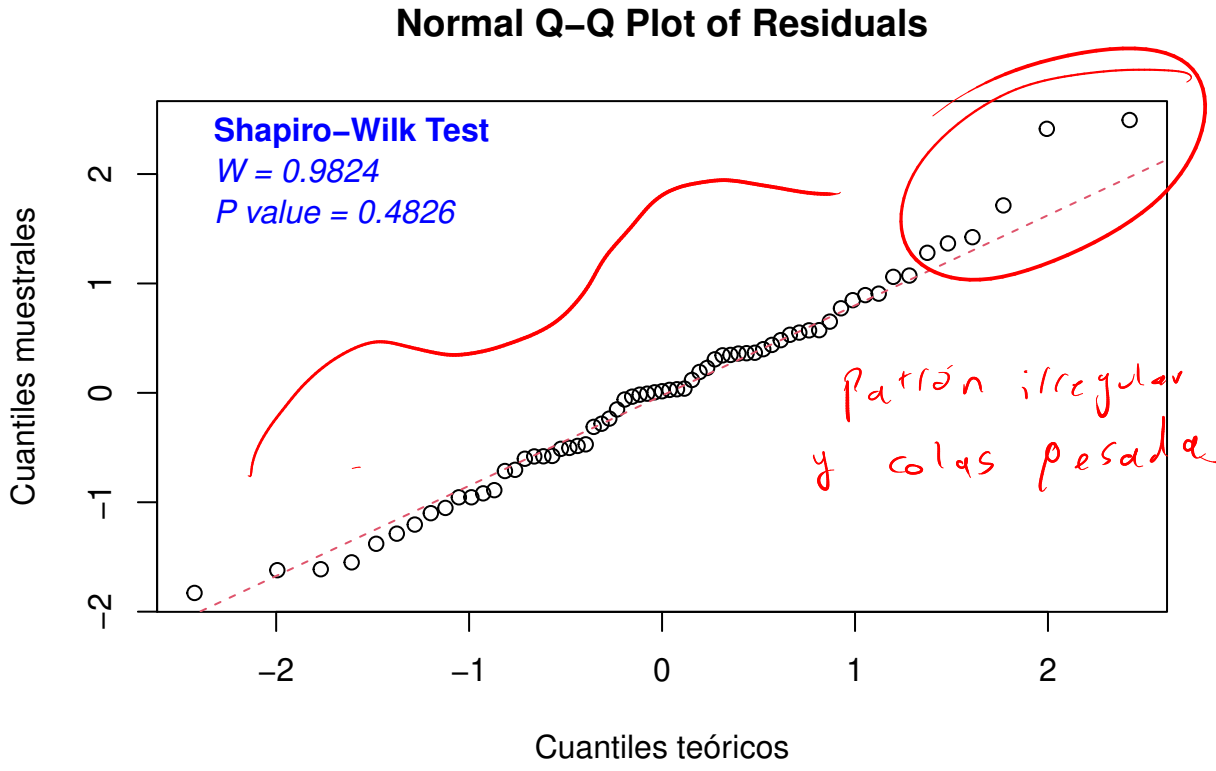


Figura 1: Gráfico cuantil-cuantil y normalidad de los residuales

Como el valor P es demasiado grande, entonces no se rechaza H_0, entonces si cumplen con el supuesto de normalidad de residuales. Además, podemos apreciar que la nube de puntos se ajusta bien a la recta, lo cual confirma nuestra conclusión.

lpt

Var cte pero hay falta de ajuste

4.1.2. Media 0 y Varianza constante

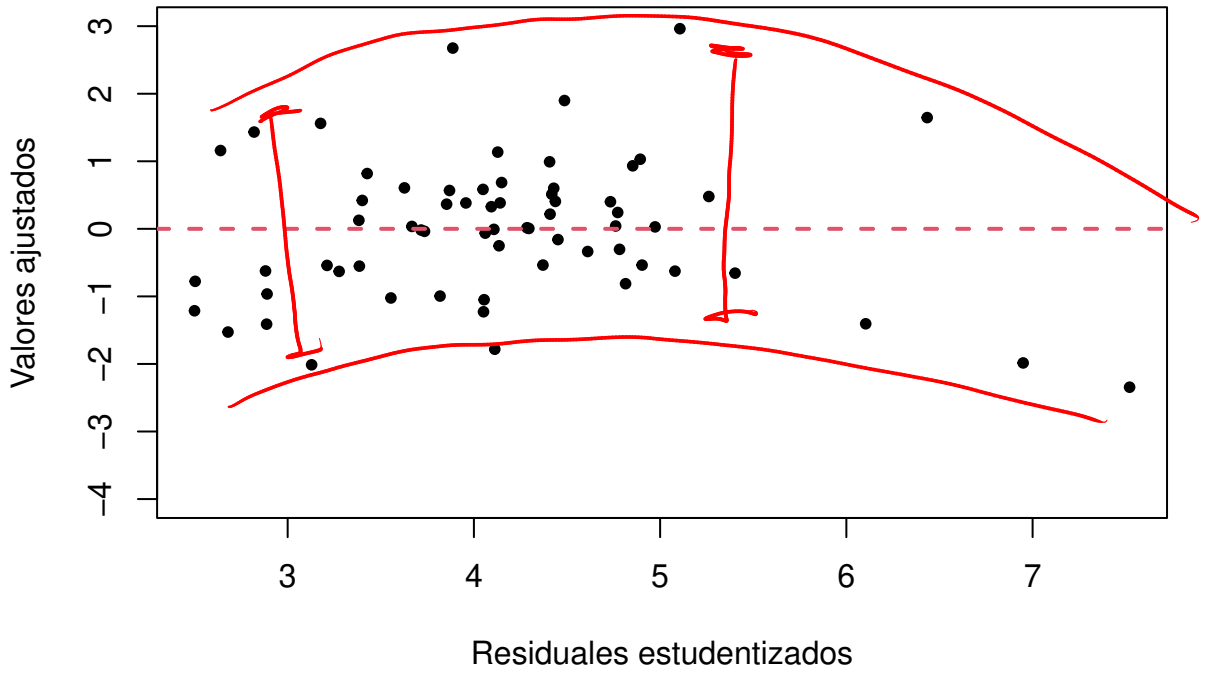


Figura 2: Gráfico residuales estudentizados vs valores ajustados

El modelo cumple con media 0 y varianza constante.

Argumenten por qué.

Res. estud. siempre tienen media 0, este supuesto se ve es en residuales crudos

Debieron también explicar aparte de var cte el patrón de falta de ajuste

4.2. Observaciones extremas

4.2.1. Datos atípicos

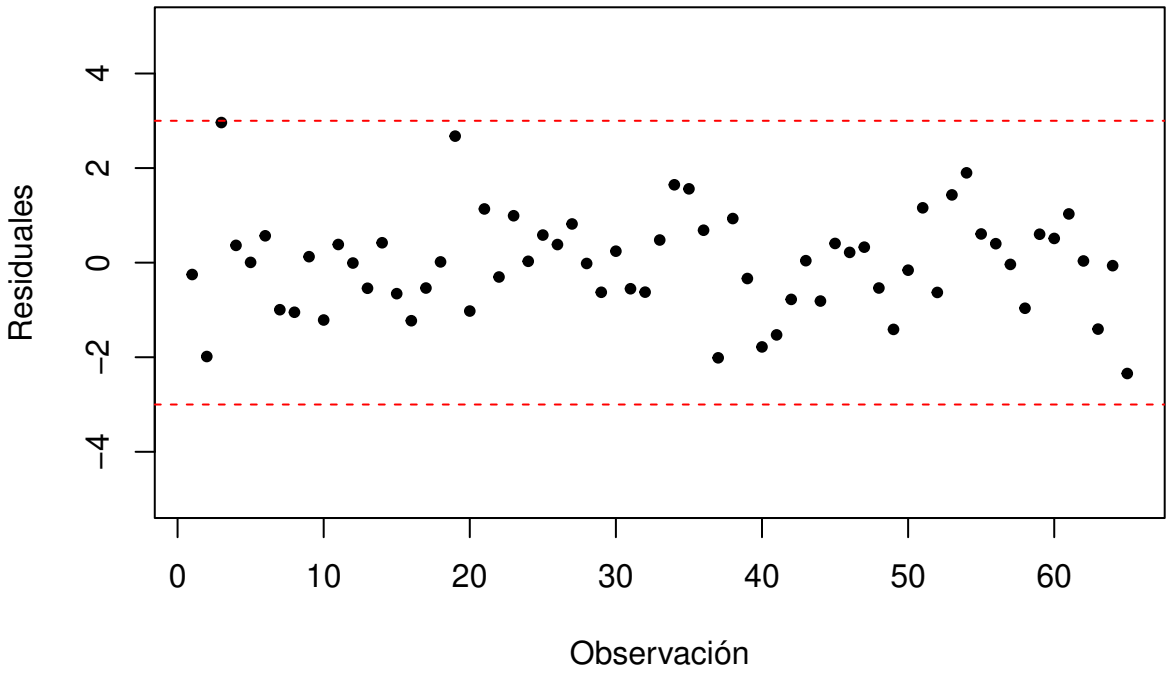


Figura 3: Identificación de datos atípicos

Según la figura de identificación de datos atipicos no hay datos atípicos; bajo el criterio de $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

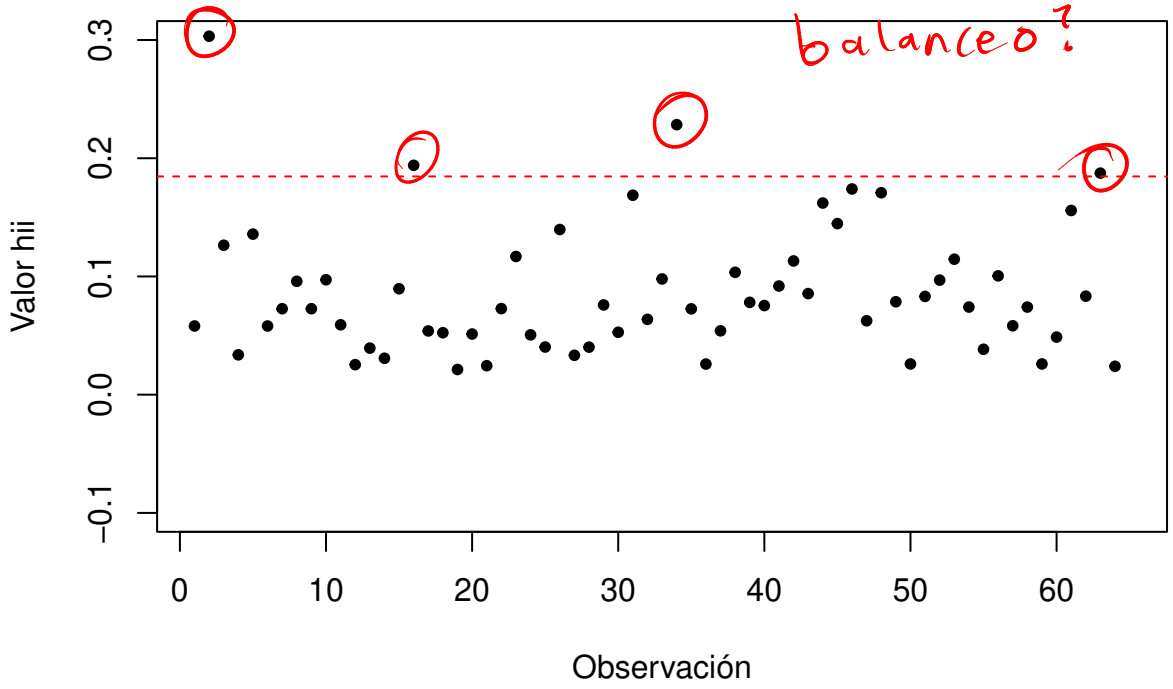


Figura 4: Identificación de puntos de balanceo

El criterio para hallar puntos de balanceo es que el hii.value $> 2(\frac{p}{n})$, donde p = 6 es el número de parámetros y n = 65 es el número de datos, $2(\frac{6}{65}) = 0.184615$

```
##    res.stud Cook.D hii.value  Dffits
## 2   -1.9849 0.2722    0.3032 -1.3094
## 16  -1.2280 0.0600    0.1940 -0.6025
## 34   1.6464 0.1299    0.2284  0.8957
## 63  -1.4055 0.0747    0.1874 -0.6749
## 65  -2.3442 0.6671    0.4394 -2.0755
```

Los datos de punto de balanceo que cumplen con el criterio son: 2, 16, 34, 63, 65.

4.2.3. Puntos influenciales

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

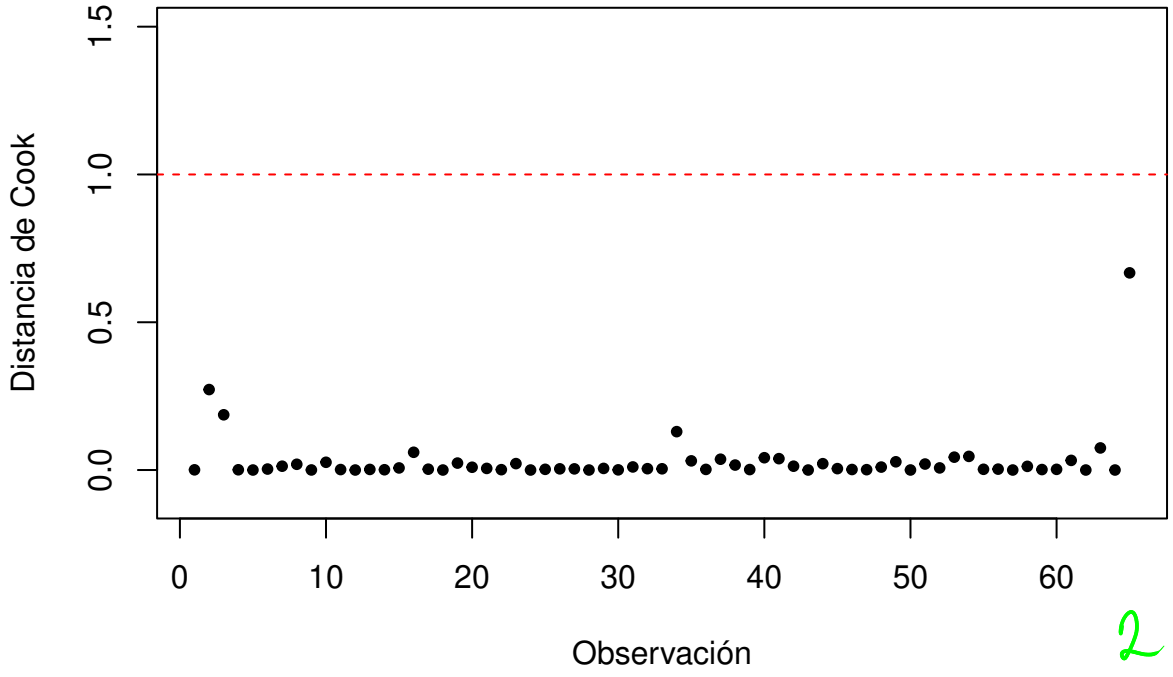
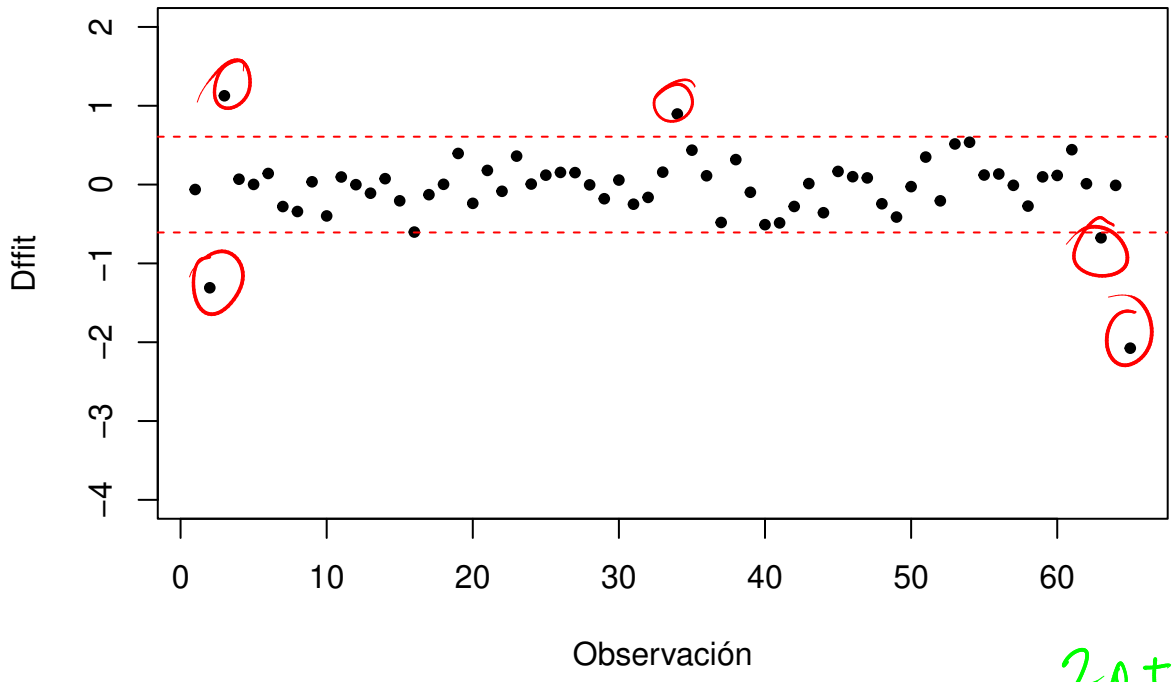


Figura 5: Criterio distancias de Cook para puntos influenciales

Podemos apreciar en la gráfica que, según el criterio de distancias de Cook donde los $D_i > 1$,no hay datos influénciales que afecten las estimaciones de los parámetros.

Bajo el criterio dffits, se hace la siguiente gráfica



2pt

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cook.D hii.value  Dffits
## 2   -1.9849 0.2722    0.3032 -1.3094
## 3    2.9626 0.1869    0.1264  1.1267
## 34   1.6464 0.1299    0.2284  0.8957
## 63  -1.4055 0.0747    0.1874 -0.6749
## 65  -2.3442 0.6671    0.4394 -2.0755
```

¿Cuánto da?

Según este criterio donde Dffit debe de estar por fuera del intervalo $\pm 2 * \sqrt{(\frac{p}{n})}$ donde p = 6 y n = 65, encontramos que los datos 2, 3, 34, 63, 65 son los puntos influénciales que afectan las estimaciones de $\hat{y}$

1pt

4.3. Conclusiones

Al analizar los supuestos del modelo, probamos mediante una prueba de normalidad Shapiro Wilk que los residuales si se distribuyen de forma normal, además de cumplir con media cero y varianza constante. No se encontraron datos atípicos, pero si 5 puntos de balanceo. Por otro lado, se evidenció que no existen puntos influénciales por el criterio de distancia de Cook, pero si encontramos 5 puntos influénciales por el criterio de Dffits.

En conclusión, a pesar de tener un $R^2 = 0.5332$ que no es muy grande, teniendo en cuenta la validación de los supuestos del modelo podemos determinar que este es un modelo óptimo para explicar el porcentaje de riesgo de infección en los hospitales de EE.UU, según la muestra dada.

¿Qué es óptimo en un MRLM?

Ninguna de las 2, además no dicen si el modelo es válido o no.