Trabajo 1

4,2

Estudiantes

Denilson Alvarez Guzman
Mariana Zapata Duque
Juan Jose Pacheco Arias
Nelson Fernando Imbacuan Chapuel

Equipo 30

Docente

Mateo Ochoa Medina

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín

5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

Nota: Para el presente trabajo, cada una de las pruebas de hipótesis que se desarrollan se realizaron con un nivel de significancia estadística de $\alpha = 0.05$.

1. Pregunta 1

17 pt

Teniendo en cuenta la base de datos brindada (en este caso del equipo 30), en la cual hay 5 variables regresoras dadas por:

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

Dónde tenemos qué:

- 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

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	0.2596
β_1	0.1873
β_2	-0.0052
β_3	0.0493
β_4	0.0151
β_5	0.0017

3 pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 0.2596 + 0.1873 X_{1i} - 0.0052 X_{2i} + 0.0493 X_{3i} + 0.0151 X_{4i} + 0.0017 X_{5i}, \quad i = 1, 2, \ldots, 74$$

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:

MSR

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

3pt

Ahora, presentamos la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	76.9169	5	15.383386	18.8415	1.04554e-11
Error	55.5194	68	0.816462		

De la tabla Anova, se observa que el valor-p $< \alpha = 0.05$, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $0 \leqslant j \leqslant 5$, 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	0.2596	1.4686	0.1768	0.8602
β_1	0.1873	0.0668	2.8047	0.0066
β_2	-0.0052	0.0272	-0.1925	0.8479
β_3	0.0493	0.0115	4.2968	0.0001
β_4	0.0151	0.0067	2.2599	0.0270
β_5	0.0017	0.0006	2.7314	0.0080

6pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3, β_4 y β_5 son significativos individualmente en presencia de los demás, pues sus P-valores respectivos son menores a α. Sin embargo para una confianza del 98 %, es decir un $\alpha = 0.02$, el parámetro β_4 no sería significativo en presencia de los demás. Por otro lado los parámetros β_0 y β_2 no son significativos.

1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** Ante un cambio de una unidad en la duración promedio de la estadía de todos los pacientes en el hospital (en días), la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje), aumenta en promedio 0.1873 %, cuando las demás se mantienen constantes.

$\hat{\beta}_3$**:** Ante un cambio de una unidad en el numero promedio de camas en el hospital durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje), aumenta en promedio 0.0493 %, cuando las demás se mantienen constantes.

3pt

$\hat{\beta}_4$**:** Ante un cambio de una unidad en el numero promedio de pacientes en el hospital por dia durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje), aumenta en promedio 0.0151 %, cuando las demas se mantienen constantes.

$\hat{\beta}_5$**:** Ante un cambio de una unidad en el numero promedio de enfermeras, equivalente a tiempo completo, durante el periodo de estudio, la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje), aumenta en promedio 0.0017 %, cuando las demas se mantienen constantes.

$\hat{\beta}_0$**:** Para este parámetro tenemos que no es significativo, pero además que el 0 no está incluido en el intervalo.

1.5. Coeficiente de determinación múltiple R^2

2pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.5808$, lo que significa que aproximadamente el 58.08 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital, es explicada por la RLM propuesta. En sentido inverso el 41.92 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital, es explicado por el error.

¿Cómo se calcula!

2. Pregunta 2

5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las 3 covariables con el P-valor más pequeño en el modelo fueron X_1, X_3, 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	55.519	X1 X2 X3 X4 X5
Modelo reducido	104.960	X2 X4

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

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

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

3pt

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,68} \\ &= \frac{(104.960 - 55.519)/3}{0.81645588} \\ &= 20.1852099 \end{aligned} \tag{2}$$

Ahora, comparando el F_0=20.1852099 con $f_{0.95,3,68} = 2.7395$, se puede ver que $F_0 > f_{0.95,3,68}$ y por tanto se rechaza la hipotesis nula (H_0) y se concluye que la probabilidad promedio de adquirir infección en el hospital está influenciada significativamente por la duración promedio de la estadía de los pacientes (X_1), el numero de camas promedio en el hospital (X_3), así como por el numero de enfermeras (X_5). Por tanto no es posible descartar las variables de este subconjunto (X_1, X_3, X_5) del modelo. 2pt

3. Pregunta 3 5pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Teniendo en cuenta el contexto de la regresión, se plantea la siguiente pregunta: ¿Si se presenta un aumento de una unidad en el número promedio de pacientes en el hospital por día (X_4), la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) incrementará el doble que si aumentara en una unidad el número promedio de enfermeras presentes a tiempo completo en el hospital durante el periodo del estudio (X_5)?

Adicionalmente podríamos preguntarnos si ¿el efecto en la probabilidad promedio estimada de adquirir infección en el hospital que causa el aumentar en un día la duración promedio de la estadía de todos los pacientes en el hospital (X_1), es igual al efecto ocasionado por reducir en una unidad el número promedio de camas en el hospital(X_3) durante el mismo periodo?

A partir de ello planteamos la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_5 = 2\beta_4;\ \ \beta_1 = -\beta_3 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

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

Para nuestra matriz $\mathbf{L}$ es útil saber qué

2pt

$$H_0 : \begin{cases} \beta_5 - 2\beta_4 = 0 \\ \beta_1 + \beta_3 = 0 \end{cases}$$

Con $\mathbf{L}$ dada por

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

Obteniendo entonces:

$$\begin{aligned} Y &= \beta_0 + \beta_1 \left(X_{1i} - X_{3i}\right) + \beta_2 X_2 + \beta_4 \left(X_{4i} + 2X_{5i}\right) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \\ &= \beta_0 + \beta_1 X_{1,3,i} + \beta_2 X_2 + \beta_4 X_{4,5,i} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \end{aligned}$$

Por tanto el modelo reducido está dado por:

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

1pt

Dónde $X_{1i}^* = X_{1i} - X_{3i}$ y $X_{4i}^* = X_{4i} + 2X_{5i}$

3.2. Estadístico de prueba

Con el estadístico de prueba F_0 dado por:

2pt

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,68} \tag{3}$$

El cuál reemplazando los valores conocidos queda de la forma:

$$F_0 = \frac{(SSE(MR) - 55.5194/2}{0.816462} \overset{H_0}{\sim} f_{2,68} \tag{4}$$

4. Pregunta 4

15pt

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:

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

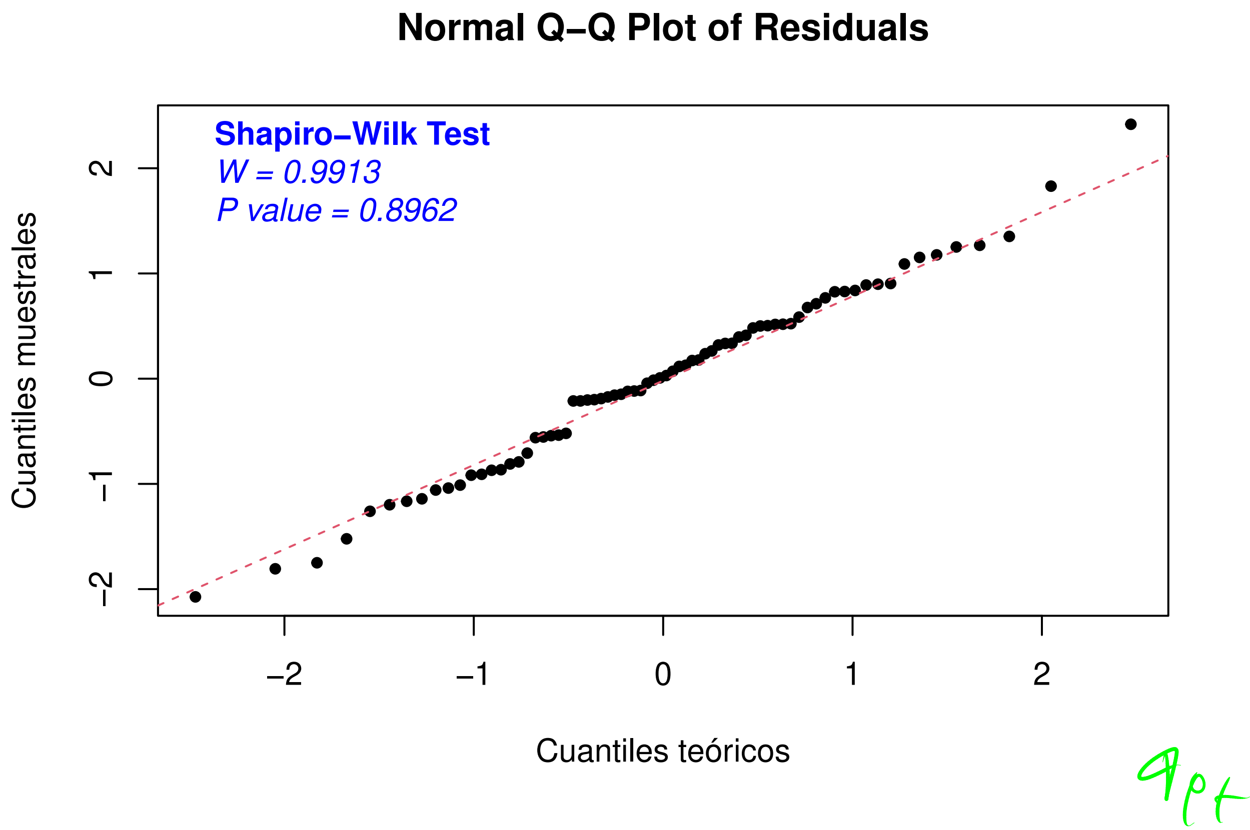
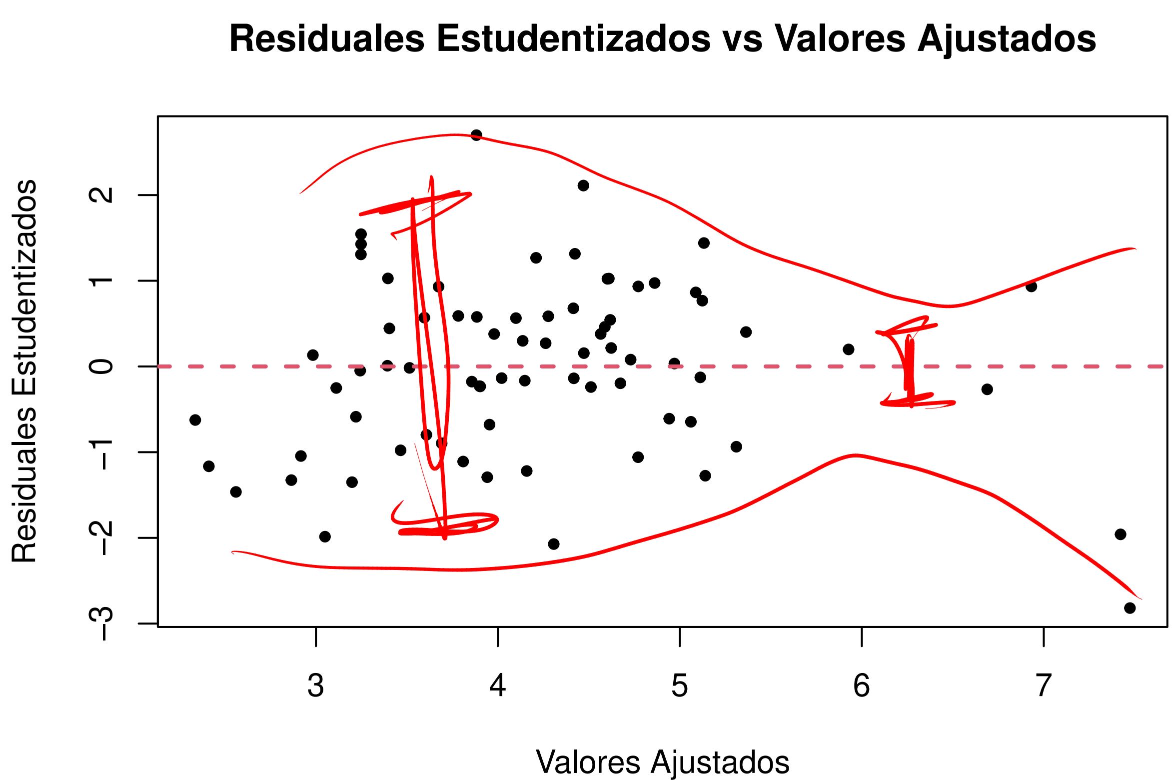


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.8962 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 se acepta Ho llevando a que los datos distribuyen normal con media μ y varianza σ^2, además la gráfica de comparación de cuantiles permite ver patrones regulares siguendo la linea roja, se termina por aceptar el cumplimiento de este supuesto. Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante



3pt

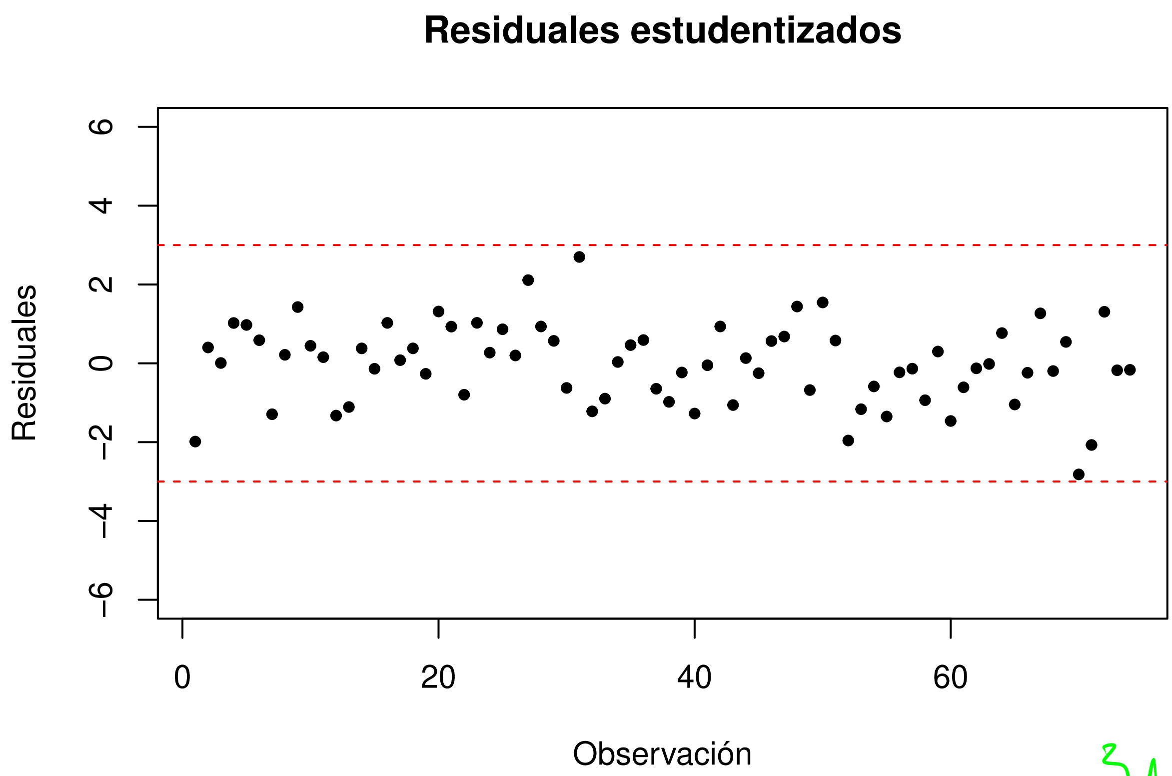
Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados, se puede observar que no hay patrones regulares cercanos a cero y tiene un patrón de los puntos que indica un decrecimiento de la dispersión en el comienzo de la grafica, luego un aumento y finalmente un decrecimiento de la dispersión, al haber evidencia suficiente en contra, el supuesto de varianza constante no se cumple, es posible que algunas observaciones extremas estén afectando este análisis.

4.2. Verificación de las observaciones

Para identificar si en el modelo hay observaciones extremas, se deben calcular los estadísticos que nos permiten aplicar criterios en ese sentido, los cuales incluyen: residuales estudentizados, los valores de la diagonal de la matriz H (los h_{ii}), la distancia de Cook (D_i) y los DFFITS.

4.2.1. Datos atípicos



3pt

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 $|r_{estud}| > 3$. Graficamente observamos que ningun valor observado corresponde a un residual por fuera de los limites ubicados en -3 y 3, es decir las lineas horizontales de color rojo.

4.2.2. Puntos de balanceo

Se asume que la observación i es un punto de balanceo si $h_{ii} > 2\frac{p}{n}$. En este caso tenemos que: $h_{ii} > 2\frac{6}{74}$ con $2\frac{6}{74} = 0,162$. Por lo tanto, en la tabla de valores para el diagnóstico de valores extremos se tiene que:

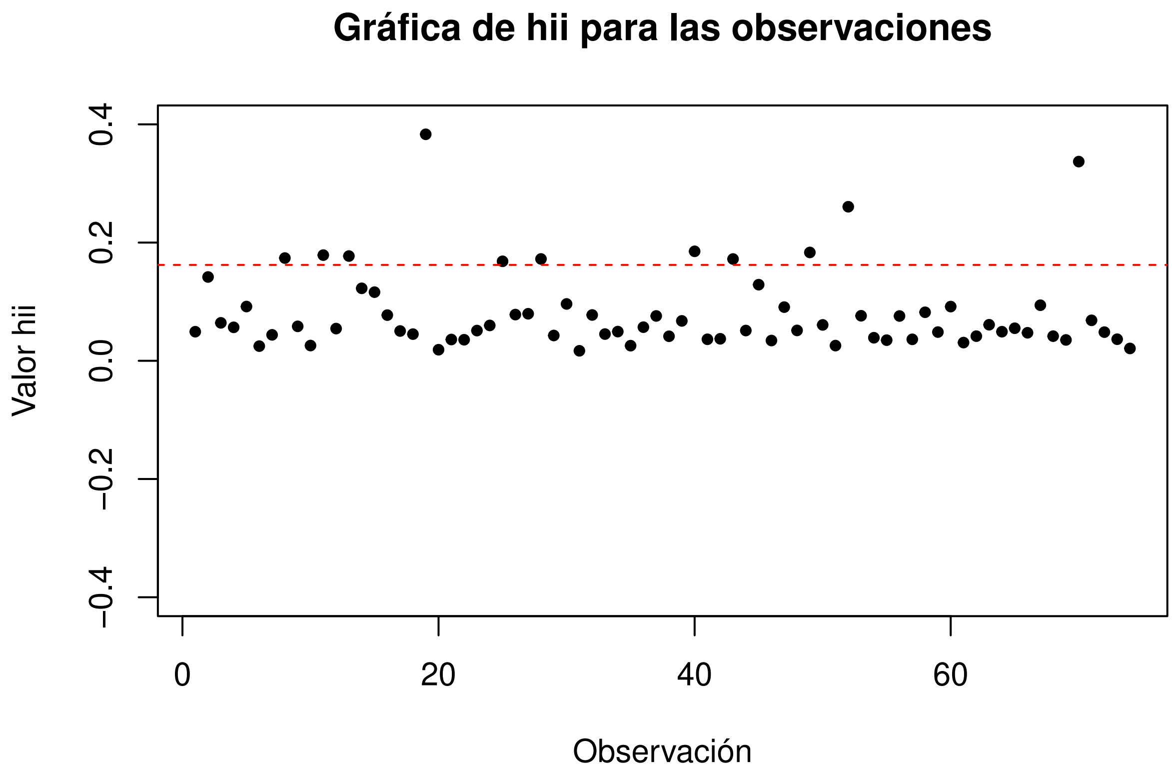


Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 8    0.2158  0.0016    0.1738  0.0983
## 11   0.1557  0.0009    0.1788  0.0721
## 13  -1.1091  0.0442    0.1772 -0.5156
## 19  -0.2668  0.0074    0.3832 -0.2088
## 25   0.8644  0.0252    0.1683  0.3881
## 28   0.9342  0.0303    0.1723  0.4258
## 40  -1.2748  0.0616    0.1852 -0.6107
## 43  -1.0588  0.0389    0.1722 -0.4834
## 49  -0.6785  0.0172    0.1833 -0.3202
## 52  -1.9589  0.2254    0.2606 -1.1884
## 70  -2.8188  0.6732    0.3370 -2.1228
```

2p+

Al examinar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$ en este caso $0,162$, se puede apreciar que existen 11 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{p}{n}$, los cuales estan tambien representados en la tabla con su respetivo # de dato.

¿Causan...?

4.2.3. Puntos influenciales

Los puntos influenciales se encuentran mediante dos pruebas:

1.Se dice que la observación i sera influencial si $D_i > 1$.

2.Una observacion sera influencial si $\|DFFITS\| > 2\frac{p}{n}^{1/2}$, en este caso $2 * \frac{6}{74}^{1/2} = 0,569$ en valor absoluto.

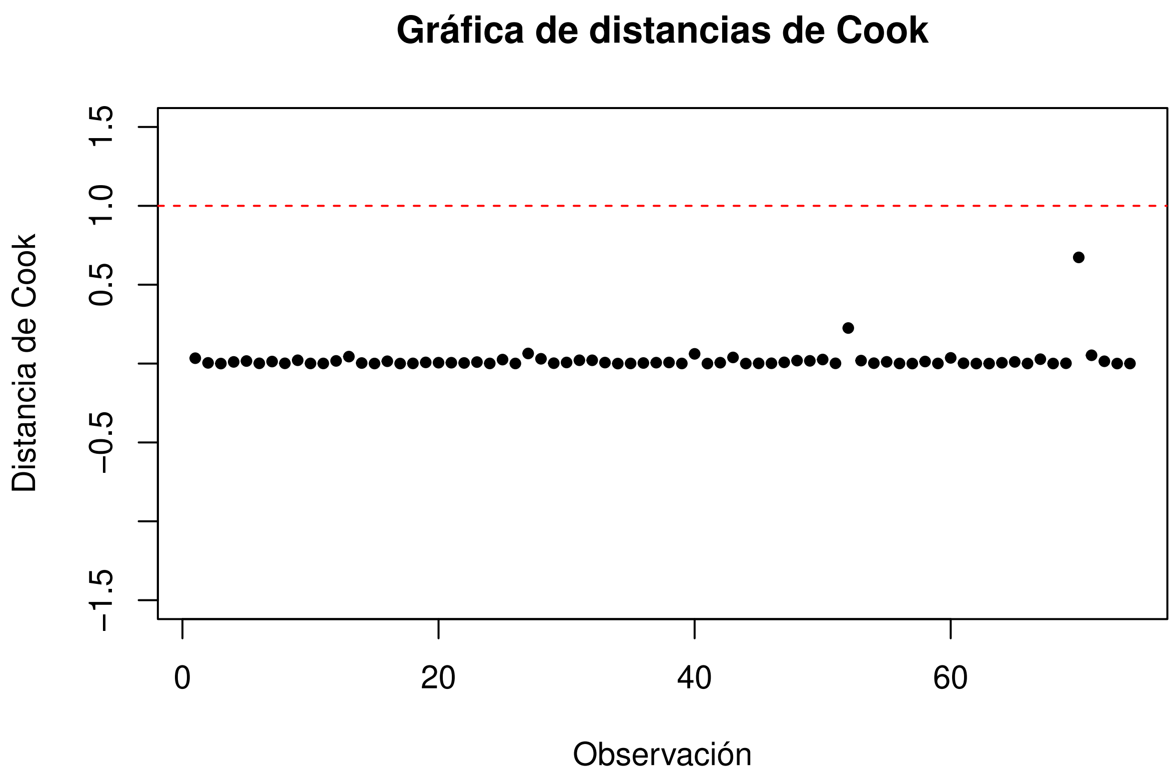


Figura 5: Criterio distancias de Cook para puntos influenciales

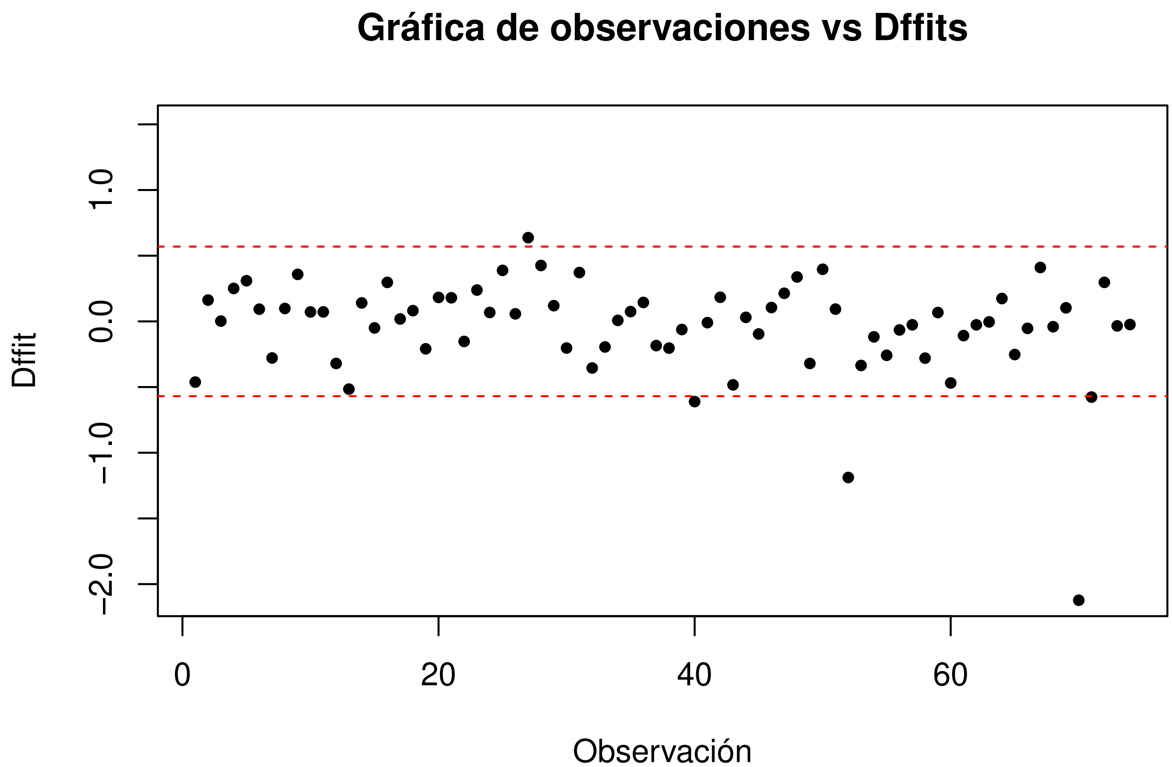


Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 27   2.1110  0.0642    0.0796  0.6374
## 40  -1.2748  0.0616    0.1852 -0.6107
## 52  -1.9589  0.2254    0.2606 -1.1884
## 70  -2.8188  0.6732    0.3370 -2.1228
## 71  -2.0720  0.0526    0.0685 -0.5760
```

3pt

Como se puede ver,las observaciones y la tabla se pueden encontrar 5 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 en el ajuste de la recta. Sin embargo De la primera prueba gráfica podemos decir que no hay valores D_i superiores a 1. Por tanto el criterio de distancias de cook nos dice que no hay valores influenciales en las estimaciones de los párametros.

¿Causan... ?

4.3. Conclusión

0pt

Con todos los estudios del modelo realizado anteriormente se logra concluir que el modelo es válido de la siguiente forma:

1. la regresión múltiple es válida con una confianza de 0,05 → why?

De hecho no...

2. Las variables X1,X3,X4,X5 son significativas individualmente en presencia de los demas
3. Apróximadamente el 58.08 % de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital,es explicada por la RLM propuesta.
4. se concluye que la probabilidad promedio de adquirir infección en el hospital está influenciada significativamente por la duración promedio de la estadía de los pacientes (X_1), el numero de camas promedio en el hospital (X_3), así como por el numero de enfermeras (X_5)
5. Los Ei se distribuyen de forma normal sin varianza constante.
6. Se encuentran cero datos atipicos, 11 puntos de balanceo y 5 puntos influenciales.

por tanto no es válido.