Trabajo 1

4,7

Estudiantes

Luiza Fernanda Alvarez Londoño
Valentina Cardoso Bran
Willy Manco Galeano
Luis Angel Paredes Contreras

Equipo 57

Docente

Javier Armando Lozano

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

1. Pregunta 1 **3**

1.1. Modelo de regresión 3

1.2. Significancia de la regresión 4

1.3. Significancia de los parámetros 4

1.4. Interpretación de los parámetros 5

1.5. Coeficiente de determinación múltiple R^2 5

2. Pregunta 2 **5**

2.1. Planteamiento pruebas de hipótesis y modelo reducido 5

2.2. Estadístico de prueba y conclusión 6

3. Pregunta 3 **6**

3.1. Prueba de hipótesis y prueba de hipótesis matricial 6

3.2. Estadístico de prueba 7

4. Pregunta 4 **7**

4.1. Supuestos del modelo 7

4.1.1. Normalidad de los residuales 7

4.1.2. Varianza constante 9

4.2. Verificación de las observaciones 10

4.2.1. Datos atípicos 10

4.2.2. Puntos de balanceo 11

4.2.3. Puntos influenciales 12

4.3. Conclusión 14

Índice de figuras

1. Gráfico cuantil-cuantil y normalidad de residuales 8
2. Gráfico residuales estudentizados vs valores ajustados 9
3. Identificación de datos atípicos 10
4. Identificación de puntos de balanceo 11
5. Criterio distancias de Cook para puntos influenciales 12
6. Criterio Dffits para puntos influenciales 13

Índice de cuadros

1. Tabla de valores coeficientes del modelo 3
2. Tabla ANOVA para el modelo 4
3. Resumen de los coeficientes 4
4. Resumen tabla de todas las regresiones 5

1. Pregunta 1

19pt

Teniendo en cuenta la base de datos brindada acerca de la eficacia en el control de infecciones hospitalarias, en la cual hay 5 variables regresoras, que explican a través de un modelo lineal el riesgo de infección. el modelo esta construido de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

Las variables son definidas de la siguiente forma

- Y: Riesgo de Infección
- X_1: Duración de la Estadía
- X_2: Rutina de Cultivo
- X_3: Número de camas
- X_4: Censo Promedio Diario
- X_5: Número de Enfermeras

1.1. Modelo de regresión

Tras el ajuste del modelo obtenemos los parametros respectivos a cada variable y el intercepto, los cuales se ven a continuación:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	1.5143
β_1	0.1628
β_2	-0.0165
β_3	0.0230
β_4	0.0159
β_5	0.0016

3pt

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = 1.5143 + 0.1628 X_{1i} - 0.0165 X_{2i} + 0.023 X_{3i} + 0.0159 X_{4i} + 0.0016 X_{5i}\ 1 \leqslant i \leqslant 59$$

1.2. Significancia de la regresión

Para s:

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,53} \tag{1}$$

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	39.7829	5	7.956577	8.5934	5.17217e-06
Error	49.0724	53	0.925894		

Al realizar la tabla Anova, la cual nos permite analizar la significancia del modelo y corrovorar el contraste de las hipótesis através del estadístico F_0 establecemos un nivel de significancia del 0.05. Tras obtener un valor-P menor a nuestro nivel de significancia, decidimos rechazar la hipótesis nula ($\beta_j = 0$ 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

Luego de comprobar que existe algún $\beta_j \neq 0$, comprobaremos la significancia individual de cada parámetro, establecemos un valor de sifnificancia de 0.05.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	1.5143	1.8985	0.7977	0.4286
β_1	0.1628	0.0927	1.7572	0.0847
β_2	-0.0165	0.0338	-0.4869	0.6283
β_3	0.0230	0.0149	1.5440	0.1285
β_4	0.0159	0.0078	2.0422	0.0461
β_5	0.0016	0.0007	2.3116	0.0247

Luego de establecer el $\alpha = 0.05$, podemos concluir, que los únicos P-valores que son menores a α, son los estimadores de β_4 y β_5, por lo tanto, se asume que estos son los estimadores significativos.

1.4. Interpretación de los parámetros

3 pt

Al ser β_4 y β_5 los unicos parámetros significantes para el modelos, es indispensable su interpretación.

$\hat{\beta}_4$: Por cada aumento en el número promedio de pacientes diarios en el estudio, el riesgo de infección aumenta un 0.0159, cuando las demás variables se mantienen constantes

$\hat{\beta}_5$: Por cada aumento en el número promedio de enfermeras durante el periodo de estudio, el riesgo de infección aumenta un 0.0016, cuando las demás variables se mantienen constantes

1.5. Coeficiente de determinación múltiple R^2

2 pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.44772$, lo que significa que aproximadamente el 44.77 % de la variabilidad total observada en la variable respuesta, es explicada por el modelo de regresión lineal múltiple que es propuesto en el presente informe.

cómo se calcula?

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más bajo en el modelo fueron X_1, X_4, X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 4, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	49.072	X1 X2 X3 X4 X5
Modelo reducido	71.464	X2 X3

Luego para construir un modelo reducido para la prueba de significancia del subconjunto, donde asumimos que la hipótesis inicial es cierta, planteamos el RM.

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_3) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,53} \\ &= \frac{7.464}{0.925894} \\ &= 8.061398 \end{aligned} \tag{2}$$

Para concluir, recordamos que establecimos un $\alpha = 0.05$ y teniendo en cuenta que el F_0=8.061398 es necesario compararlo con $f_{0.05,3,53} = 2.7791$, por lo que se puede ver que $F_0 > f_{0.05,3,53}$, es decir, rechazamos H_0, por lo tanto hay algún β_j distinto de 0 para j=1, 4, 5, por lo cual no se pueden descartar del modelo.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Con la intención de saber si existen redundancias, o relaciones entre la variables, planteamos las siguientes preguntas:

- ¿ La duración promedio de la estadía de todos los pacientes sobre el riesgo de infección, es igual a diez veces menos la razón del numero de cultivos realizados en pacientes sin síntomas de infección sobre el riesgo de infección?
- ¿La relación que hay entre el riesgo de infección y el numero promedio de pacientes en el hospital por día es igual a la relación entre el riesgo de infección y nueve veces el número promedio de enfermeras durante el periodo de estudio?

por ende se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = -10\beta_2;\ \beta_4 = 9\beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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 $\mathbf{L}$ dada por

$$L = \begin{bmatrix} 0 & 1 & 10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -9 \end{bmatrix}$$

2 pt

El modelo reducido, en donde asumimos que nuestra hipótesis inicial es cierta, está dado por:

$$Y_i = \beta_0 + \beta_1 X_1 - \frac{\beta_1}{10} X_2 + \beta_3 X_3 + \beta_4 X_4 + \frac{\beta_4}{9} X_5 \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 59$$

1 pt

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,53} \tag{3}$$

2 pt

$$F_0 = \frac{(SSE(MR) - 49.0724)/2}{0.925894} \overset{H_0}{\sim} f_{2,53} \tag{4}$$

4. Pregunta 4

18 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:

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

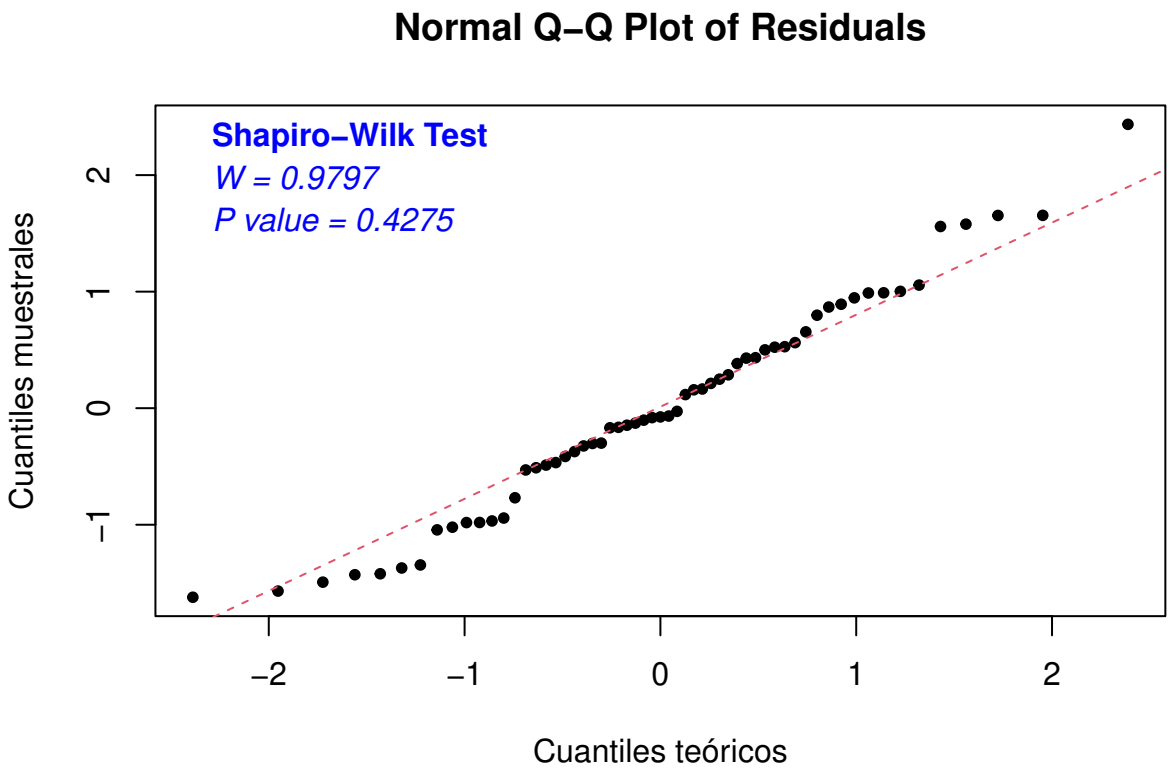

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al analizar los resultados obtenidos por la pueba shapiro-wilk,obtenemos resultados numericos y graficos, sin embargo, por medio del análisis gráfico obtenemos resultados que se priorizan ante el análisis numérico. En este caso en particular el p-valor es aproximadamente 0.4275 el cual es mayor a $\alpha = 0.05$, por lo cual, aceptamos la hipótesis inicial, es decir, según esto se distribuye de manera normal, pero, al analizar la parte gráfica, se evidencia que tiene colas pesadas, hay comportamientos ciclicos, y la mayoría de sus puntos no logran ajustarse a la recta, por lo que concluimos finalmente que el modelo no se distribuye de manera normal.

4.1.2. Varianza constante

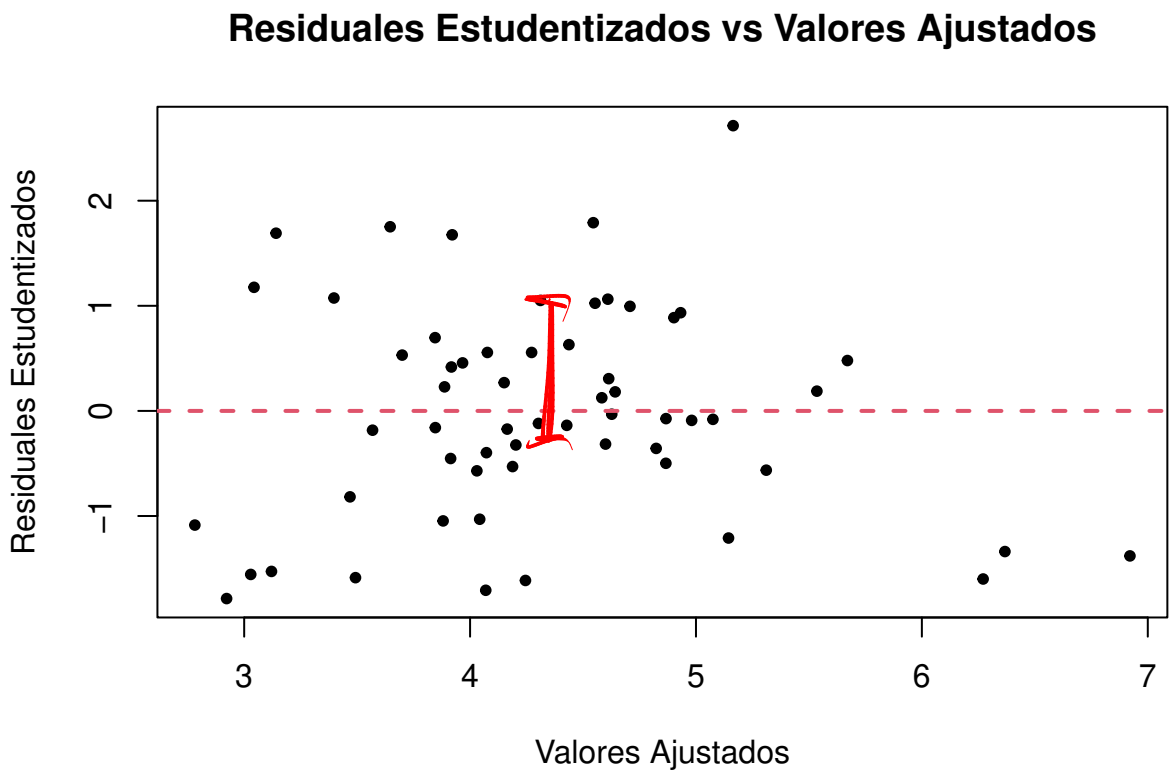

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Para analizar este gráfico *ResidualesEstudentizadosVsValoresAjustados*, podemos observar que no existe ningún patrón, o algún comportamiento en donde se refleje que la varianza crezca o decrezca, por lo tanto, no se puede descartar que hay una varianza constante, por lo que asumimos que es constante.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

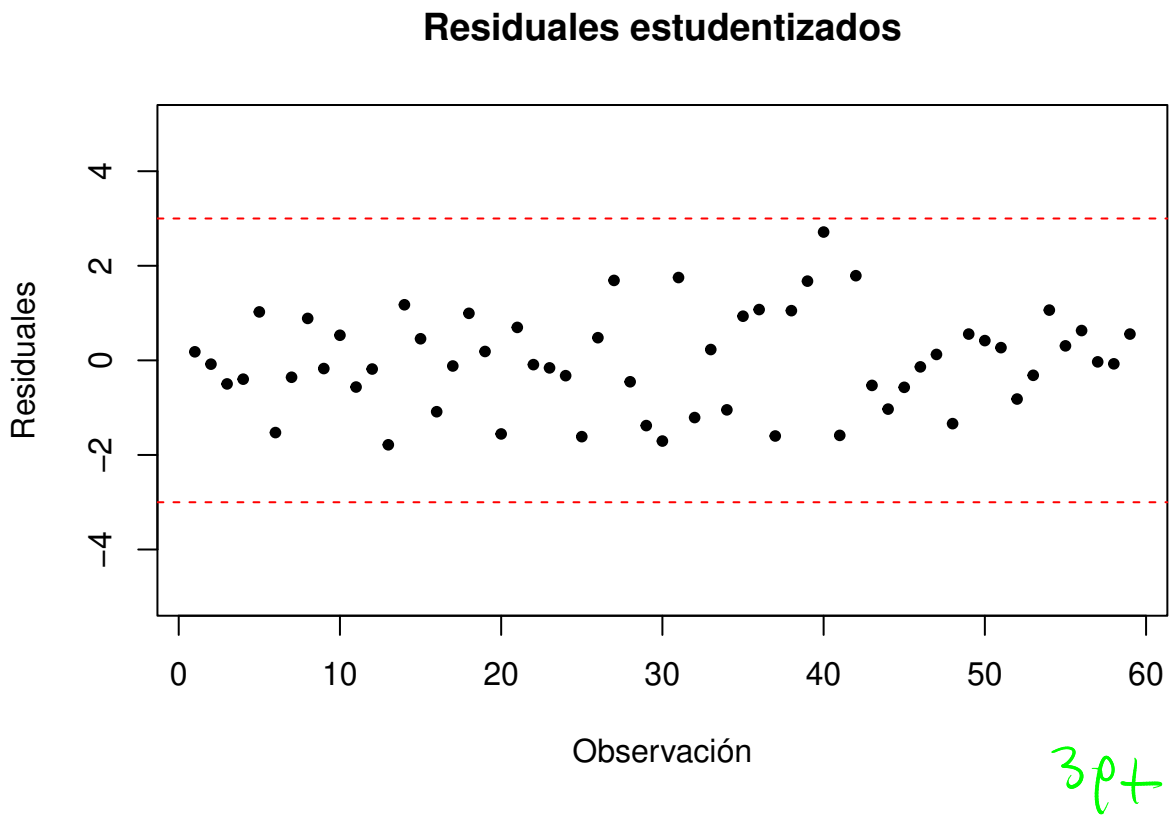

Figura 3: Identificación de datos atípicos

Una observación atípica es aquella que está separada en su valor respuesta del resto de observaciones, ésta puede afectar los resultados obtenidos del ajuste del modelo de regresión. 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 $|r_{estud}| > 3$, por lo que no hay observaciones alejadas en su valor de respuesta Y.

4.2.2. Puntos de balanceo

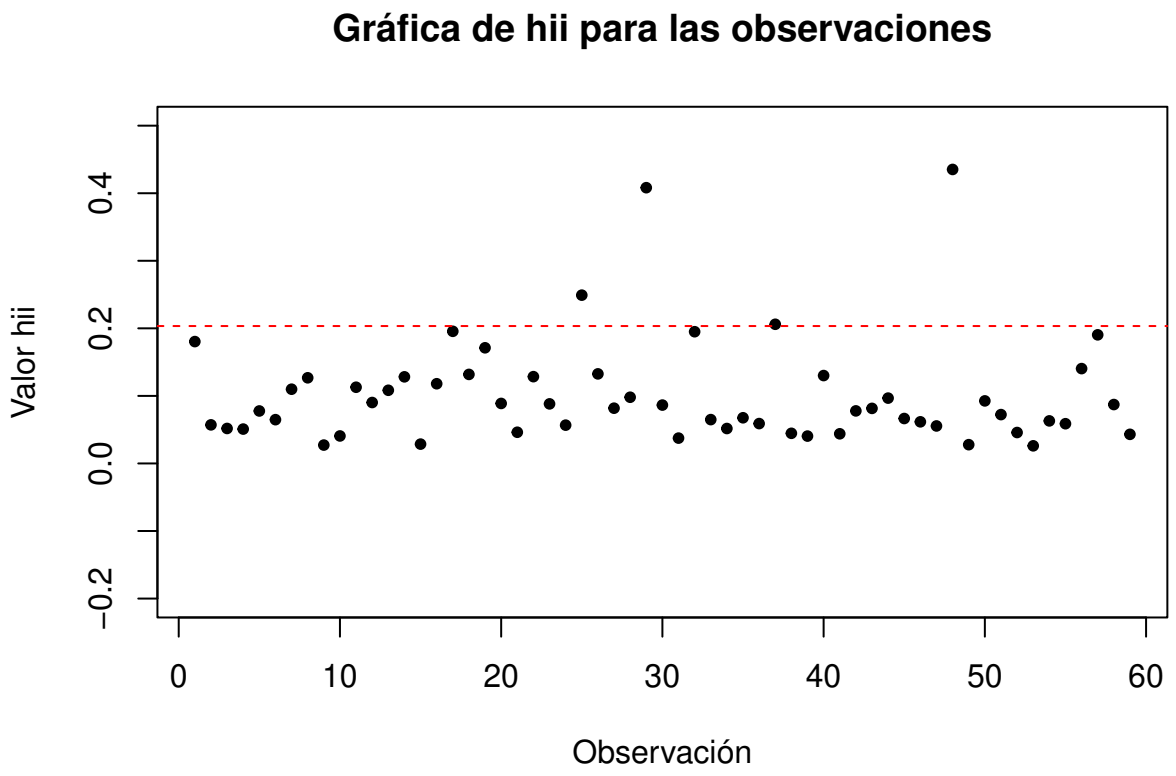

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 25  -1.6132  0.1439    0.2491 -0.9437
## 29  -1.3795  0.2188    0.4082 -1.1558
## 37  -1.5992  0.1106    0.2060 -0.8272
## 48  -1.3380  0.2299    0.4352 -1.1835
```

Los Puntos de Balanceo, nos indica una observación en el espacio de las predictoras que esta alejada del resto de la muestra. Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$ el cual es igual a 0.203, y teniendo en cuenta que una observacion es un punto de balanceo si$h_{ii} > 2\frac{p}{n}$, se puede apreciar que existen 4 puntos de balanceo, según esta condición, los cuales pueden verse con claridad no solo graficamente sino también en la tabla, en donde se tiene información detallada de estos.

4.2.3. Puntos influenciales

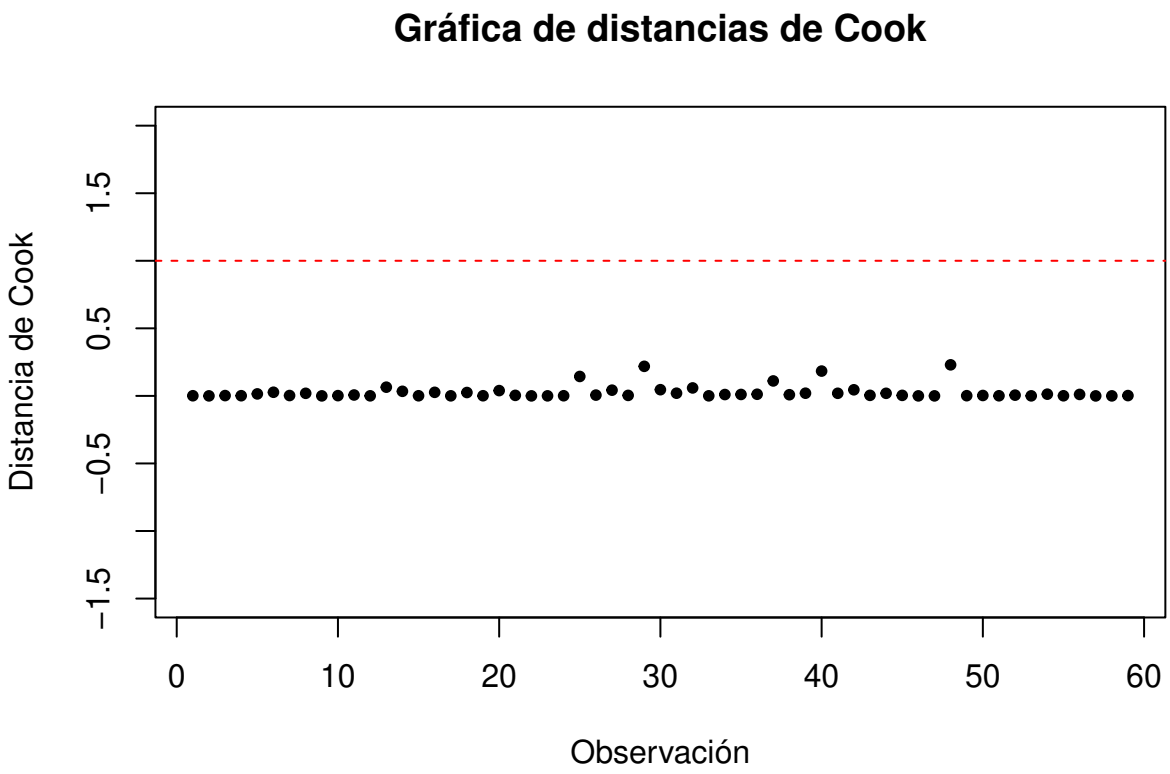

Figura 5: Criterio distancias de Cook para puntos influenciales

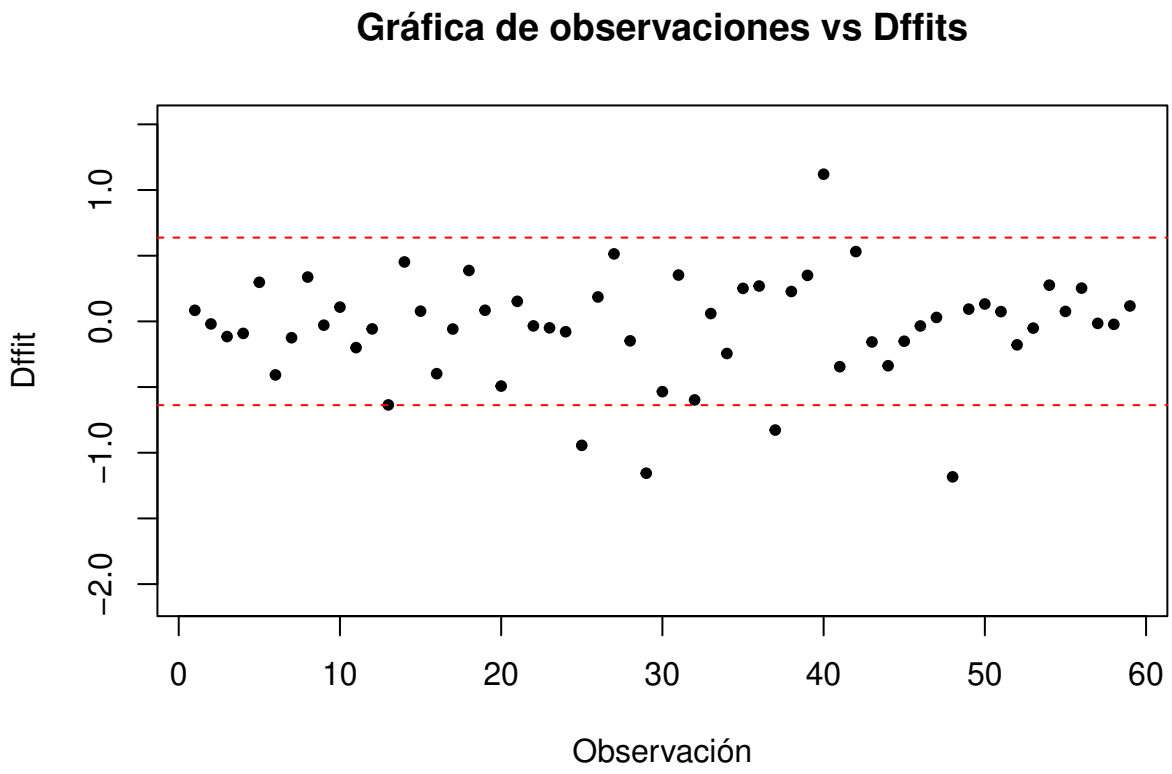

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 25  -1.6132  0.1439    0.2491 -0.9437
## 29  -1.3795  0.2188    0.4082 -1.1558
## 37  -1.5992  0.1106    0.2060 -0.8272
## 40   2.7138  0.1836    0.1301  1.1205
## 48  -1.3380  0.2299    0.4352 -1.1835
```

4 e 7

Las observaciones influenciales son aquellas que tienen un impacto notable sobre los coeficientes de regresión ajustados; y para identificar aquellas observaciones implementaremos dos métodos, *Distanciadecook* y *diagnósticoDFFITS*,Como se puede ver,las observaciones 25, 29, 37, 40, 48 son puntos influenciales según el criterio de Dffits, el cual dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$,por lo que podemos decir que son observaciones que halan el modelo en su dirección, es decir, su exclusión del modelo causa cambios importantes en la ecuación de regresión ajustada, por lo tanto, 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 pts

Para concluir, asumimos que nuestro modelo de regresión lineal multiple, que buscaba modelar la probabilidad promedio estimada de adquirir infección en el hospital, (por medio de las variables :$ Duración promedio de la estadía de todos los pacientes en el hospital(X_{1}),la razón del número de cultivos realizdos en pacientes sin síntomas de infección hospitalaria(X_2), número promedio de camas en el hospital(X_3), número promedio de pacientes en el hospital por día(X_4), y numero promedio de enfermeras(X_5)$) no es un modelo válido, no cumple con todos los supuestos del error, porque si bien su varianza es constante y no muestra ningún indicio de no linealidad, no se distribuye normal, y para que un modelo pueda ser considerado como válido es necesario que cumpla con todos los supuestos, además que no cumpla con el supuesto de normalidad puede estar directamente relacionado con los puntos influenciales obtenidos.