Trabajo 1

4,8

Estudiantes

Valeria Moncada Sanchez
Jose Manuel Miranda Pitalua
Josselyn Fernandez Cifuentes
Santiago Loaiza Gomez

Equipo 09

Docente

Veronica Guarin Escudero

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 5
4. Resumen tabla de todas las regresiones 6

1. Pregunta 1

19 pt

Teniendo en cuenta la base de datos dada, la cual representa una muestra de un estudio realizado en EE.UU sobre la eficacia en el control de infecciones hospitalarias, cuya muestra corresponde a 54 hospitales, se plantea a continuación un modelo de regresión lineal multiple (RLM) donde hay 5 variables regresoras, y además se tiene como supuestos:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

Donde las variables mencionadas corresponden a:

- Y: Riesgo de infección (probabilidad promedio estimada de adquirir infección en el hospital, en porcentaje)
- X_1: Duración de la estadía (Duración promedio de la estadía de todos los pacientes en el hospital, en días)
- X_2: Rutina de cultivos (Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100)
- X_3: Número de camas (Número promedio de camas en el hospital durante el tiempo de estudio)
- X_4: Censo promedio diario (Número promedio de pacientes en el hospital por día durante el tiempo de estudio)
- X_5: Número de enfermeras (Número promedio de enfermeras, equivalentes a tiempo completo, durante el tiempo del estudio)

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes valores estimados de cada parámetro:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-1.1062
β_1	0.2038
β_2	0.0333
β_3	0.0498
β_4	0.0069
β_5	0.0018

3 pt

Con base en la tabla de valores coeficientes del modelo, se obtiene que la ecuación del modelo de regresión ajustado con la estimacion de los respectivos parámetros es:

$$\hat{Y}_i = -1.1062 + 0.2038 X_{1i} + 0.0333 X_{2i} + 0.0498 X_{3i} + 0.0069 X_{4i} + 0.0018 X_{5i}\ 1 \leqslant i \leqslant 54$$

1.2. Significancia de la regresión

La prueba de significancia de la regresión establece lo siguiente:

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,48} \tag{1}$$

Para ello necesitamos la información contenida en la tabla Anova o también llamada tabla de análisis de varianza del modelo de RLM:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	36.6619	5	7.33238	6.46898	0.000113708
Error	54.4064	48	1.13347		

De la tabla Anova anterior, se observa el valor del estadístico de prueba $F_0 = 6.46898$ y su correspondiente valor P, el cual es aproximadamente igual a 0, como Vp $< 0.05 = \alpha$ se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto se concluye que el modelo de RLM propuesto es significativo. Esto quiere decir, que el riesgo de infección depende significativamente de al menos una de las predictorias.

1.3. Significancia de los parámetros

Luego de probar que el modelo propuesto es significativo, se procede a identificar aquellos parámetros susceptibles de interpretación, esto significa, solo se interpreta aquellos parámetros que son significativos individualmente. El siguiente cuadro de resumen de los coeficientes, presenta la información para determinar cuales de ellos son significativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-1.1062	2.1147	-0.5231	0.6033
β_1	0.2038	0.0909	2.2422	0.0296
β_2	0.0333	0.0383	0.8691	0.3891
β_3	0.0498	0.0221	2.2509	0.0290
β_4	0.0069	0.0091	0.7614	0.4501
β_5	0.0018	0.0010	1.8406	0.0719

El valor del estadístico de prueba y el valor-P para la prueba, se pueden observar en las dos últimas columnas de la tabla de resumen de los coeficientes. A un nivel de significancia $\alpha = 0.05$ se concluye que los parámetros individuales β_1 y β_3 son significativos cada uno en presencia de los demás parámetros, pues sus P-valores son menores a α. Por otro lado, para el resto de valores, se encuentra que son individualmente no significativos en presencia de los demás parámetros.

1.4. Interpretación de los parámetros

Solo se podrán interpretar parámetros que resultaron significativos individualmente, en este caso son:

$\hat{\beta}_1$**:** 0.2038 indica que por cada unidad (días) que aumente la duración promedio de la estadia de todos los pacientes en el hospital (X1) el promedio del riesgo de infección aumenta en 0.2038 unidades, cuando las demás predictorias se mantienen fijas.

$\hat{\beta}_3$**:** 0.0498 indica que por cada unidad que aumenta el número promedio de camas en el hospital (X3) el promedio del riesgo de infección aumenta en 0.0498 unidades, cuando las demás predictorias se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.40258$, lo que significa que aproximadamente el 40.26 % de la variabilidad total observada en el riesgo de infección es explicado por el modelo de regresión lineal multiple propuesto en el presente trabajo.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Con el resultado del punto anterior, tenemos que las tres variables con el P-valor más pequeño son X_1, X_3, X_5, teniendo esto en cuenta y usando la tabla de todas las regresiones

posibles, se busca realizar la siguiente prueba de hipotesis sobre la significancia simultanea del subconjunto:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 3, 5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	54.406	X1 X2 X3 X4 X5
Modelo reducido	79.959	X2 X4

Segun lo anterior, un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

2.2. Estadístico de prueba y conclusión

Con el siguiente estadístico de prueba:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_4) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0,\ \beta_1\beta_2, \beta_3, \beta_4\beta_5)} \overset{H_0}{\sim} f_{3,48} \\ &= \frac{8.5177}{1.1335} \\ &= 7.5145 \end{aligned} \tag{2}$$

Haciendo la comparación del F_0 con $f_{0.05,3,48} = 0.1165$, se evidencia que $F_0 > f_{0.05,3,48}$ por lo que se rechaza la hipotesis nula que afirma que $\beta_j = 0$ con $j = 1,3,5$. Con esto podemos afirmar que el subconjunto es significativo y no es posible descartar las variables del subconjunto.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Según los datos del estudio ¿la duracion de la estadía y el censo promedio diario tienen el mismo efecto en el riesgo de infección? ¿ el efecto que produce el numero de enfermeras

es el doble del efecto del numero de camas? para dar respuesta a las preguntas, se plantea la siguiente prueba de hipótesis:

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

Al escribirlo 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

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

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_2 X_{2i} + \beta_3 X^*_{3i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 54$$

Donde $X^*_{1i} = X_{1i} + X_{4i}$ y $X^*_{3i} = 2X_{5i} + X_{3i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF)/2}{MSE(MF)} \sim f_{2,48} \tag{3}$$

$$F_0 = \frac{(SSE(MR) - 54.4064/2}{1.13347} \sim f_{2,48} \tag{4}$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

El supuesto de normalidad se validará por medio de la prueba analítica de normalidad de shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

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

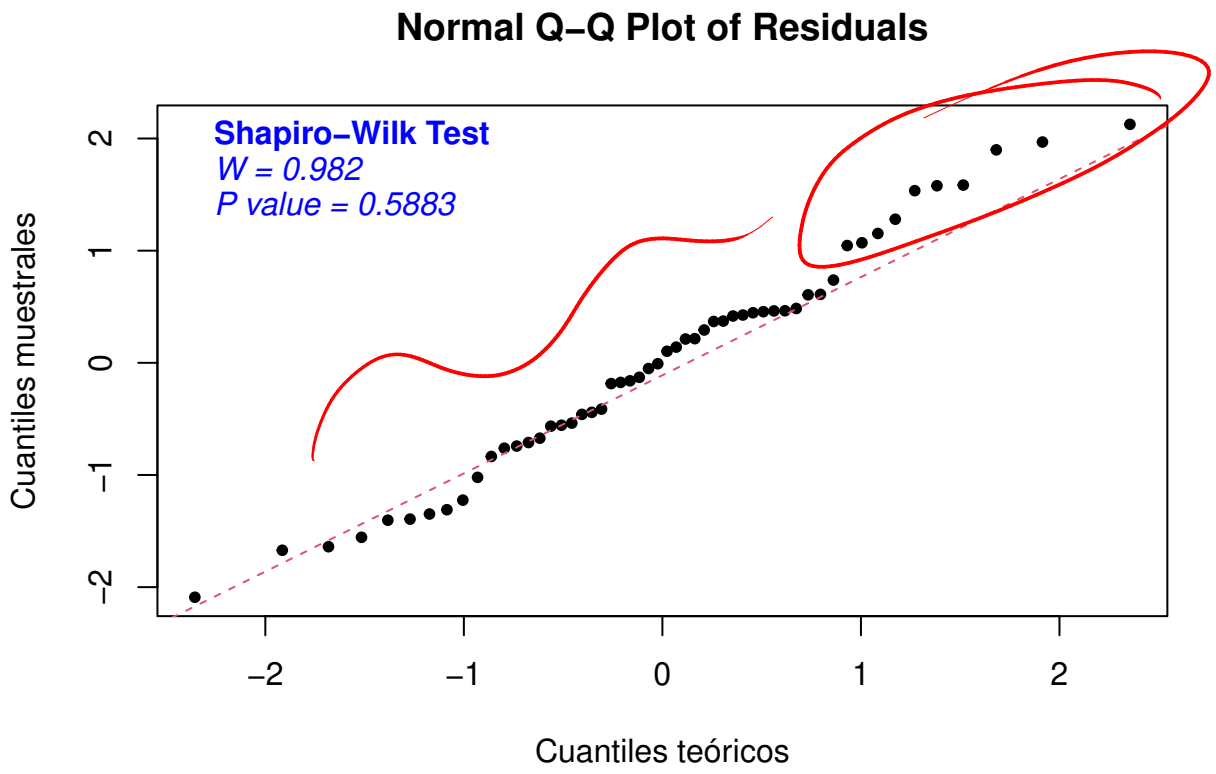

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Aunque la prueba de normalidad S-W indica que los errores son normales. (valor-P mayor a 0.05), podemos notar en el análisis gráfico de comparación de cuantiles cómo el patron de los residuales no sigue la línea roja que representa el ajuste de la distribución de los residuales a una distribución normal, se observa una cola superior pesada y patrones irregulares que podrian deberse a la presencia de observaciones influenciales. Debido a esto, se infiere que el supuesto de normalidad no se cumple.

4.1.2. Varianza constante

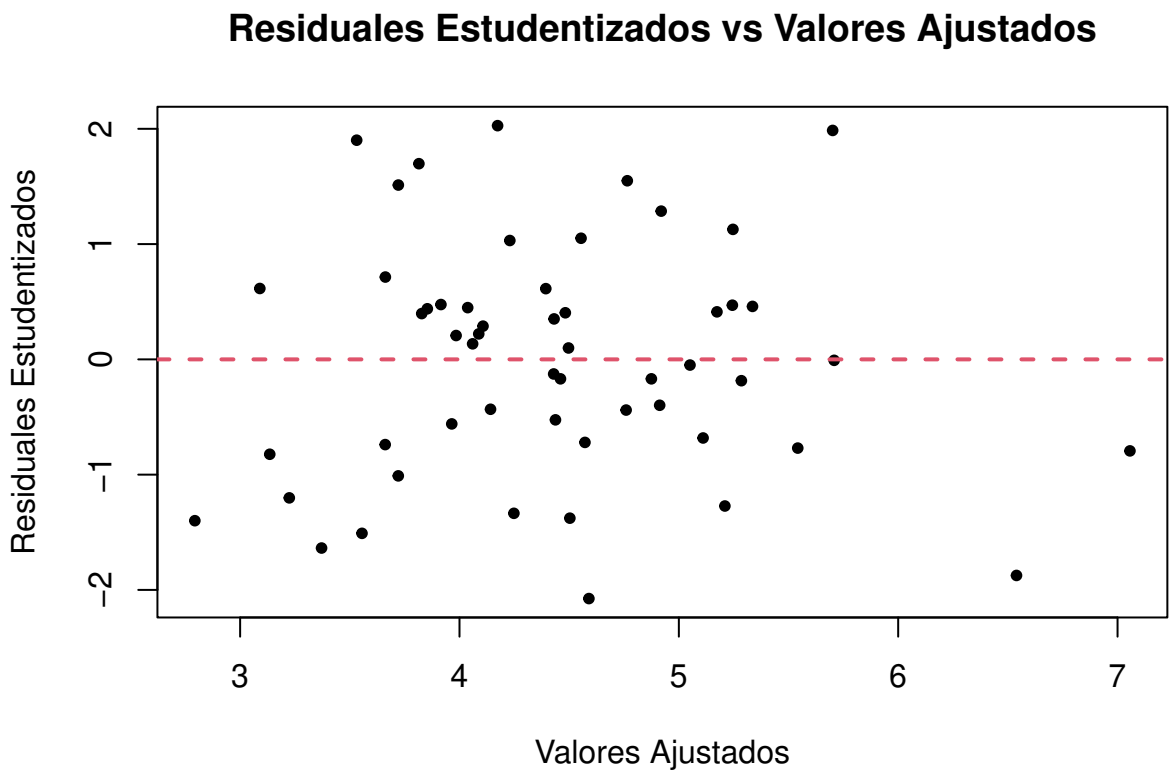

Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se observa que no hay patrones en los que la varianza aumente o decrezca, ni un comportamiento que permita descartar una varianza constante, pues los puntos de la grafica obedecen el patron rectangular esperado. Como no se tiene evidencia suficiente para rechazar este supuesto, se acepta como cierto, por lo tanto asumimos una varianza constante. Además es posible observar en la gráfica media 0.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

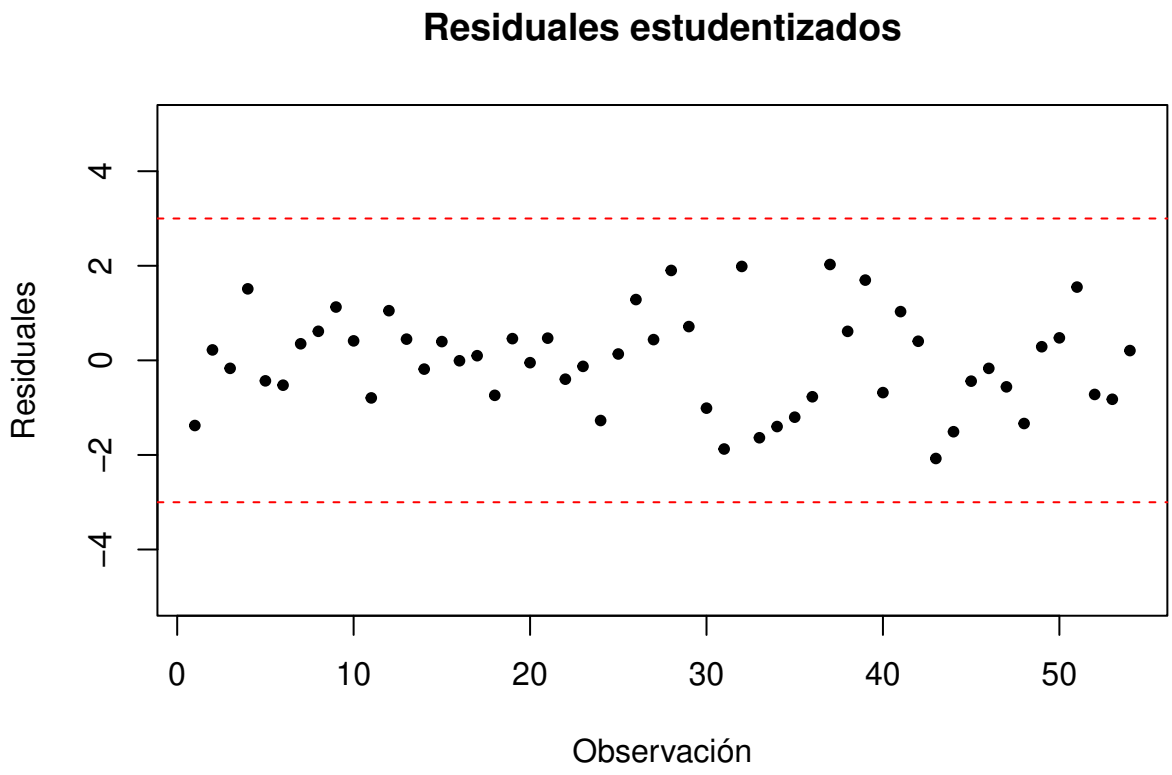

Figura 3: Identificación de datos atípicos

Se considera que una observación es atípica cuando su residual estudentizado, es tal que: $|r_{estud}| > 3$. Como se puede observar en la gráfica anterior, no hay datos atípicos en el conjunto de datos, pues ningún residual estudentizado sobrepasa el criterio establecido, esto es, no hay observaciones separadas (en su valor de la respuesta Y) del resto de las observaciones.

4.2.2. Puntos de balanceo

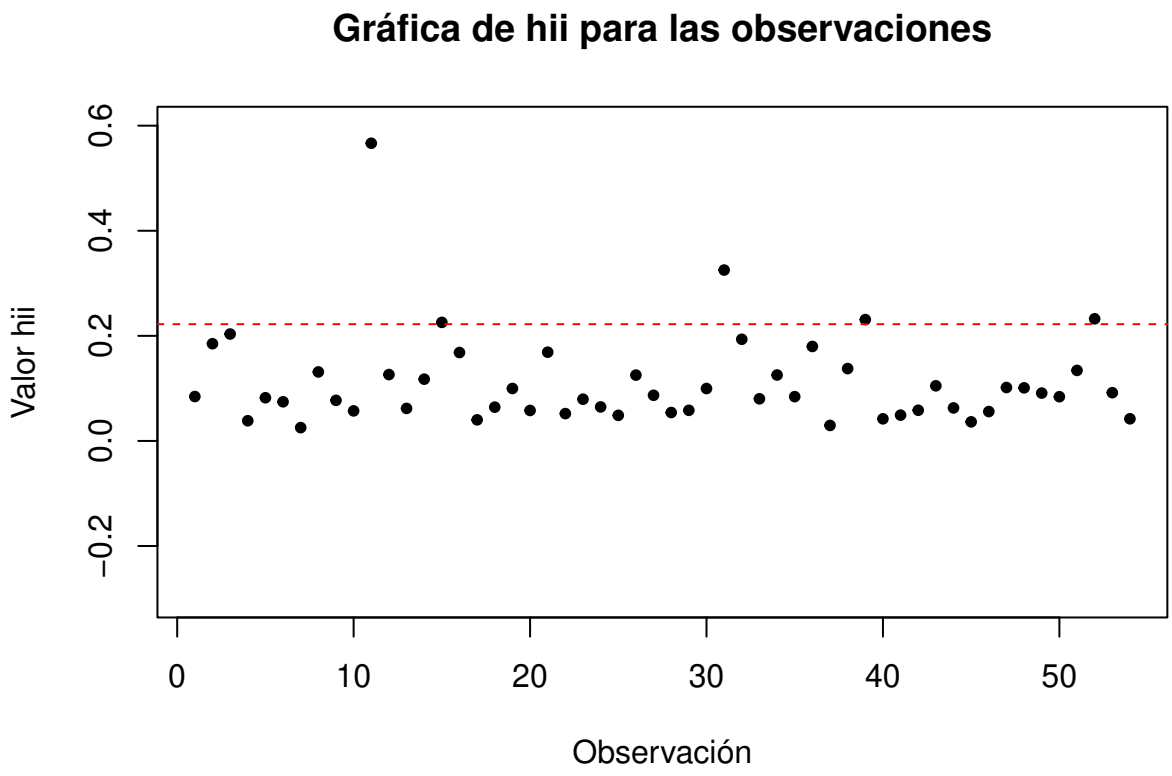

Figura 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 11  -0.7940  0.1373    0.5665 -0.9040
## 15   0.3972  0.0077    0.2256  0.2125
## 31  -1.8745  0.2822    0.3252 -1.3376
## 39   1.6975  0.1442    0.2309  0.9492
## 52  -0.7204  0.0262    0.2324 -0.3943
```

Un punto de balanceo es una observación en el espacio de las predictoras, alejada del resto de la muestra y que puede controlar ciertas propiedades del modelo ajustado. Analizando la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, se pueden apreciar 5 datos del conjunto que son puntos de balanceo bajo el criterio $h_{ii} > 2\frac{p}{n}$, estos puntos son los datos 11, 15, 31, 39, 52, como se observa en la tabla.

4.2.3. Puntos influenciales

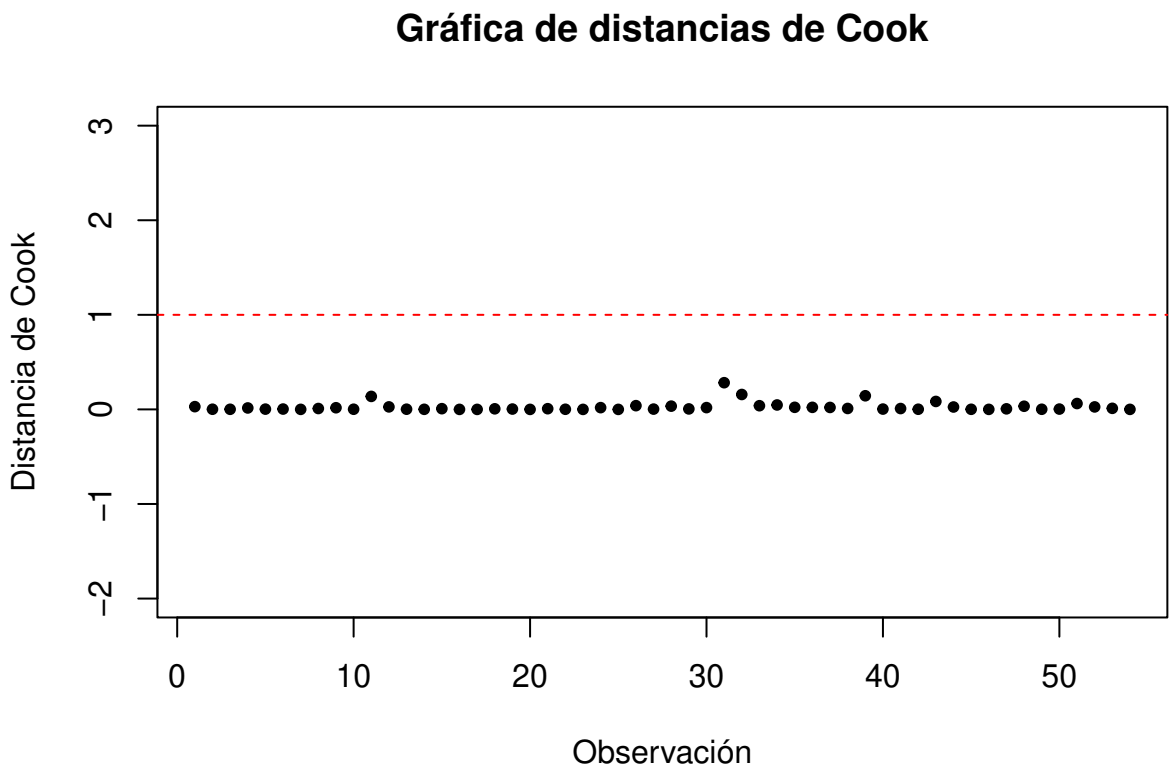

Figura 5: Criterio distancias de Cook para puntos influenciales

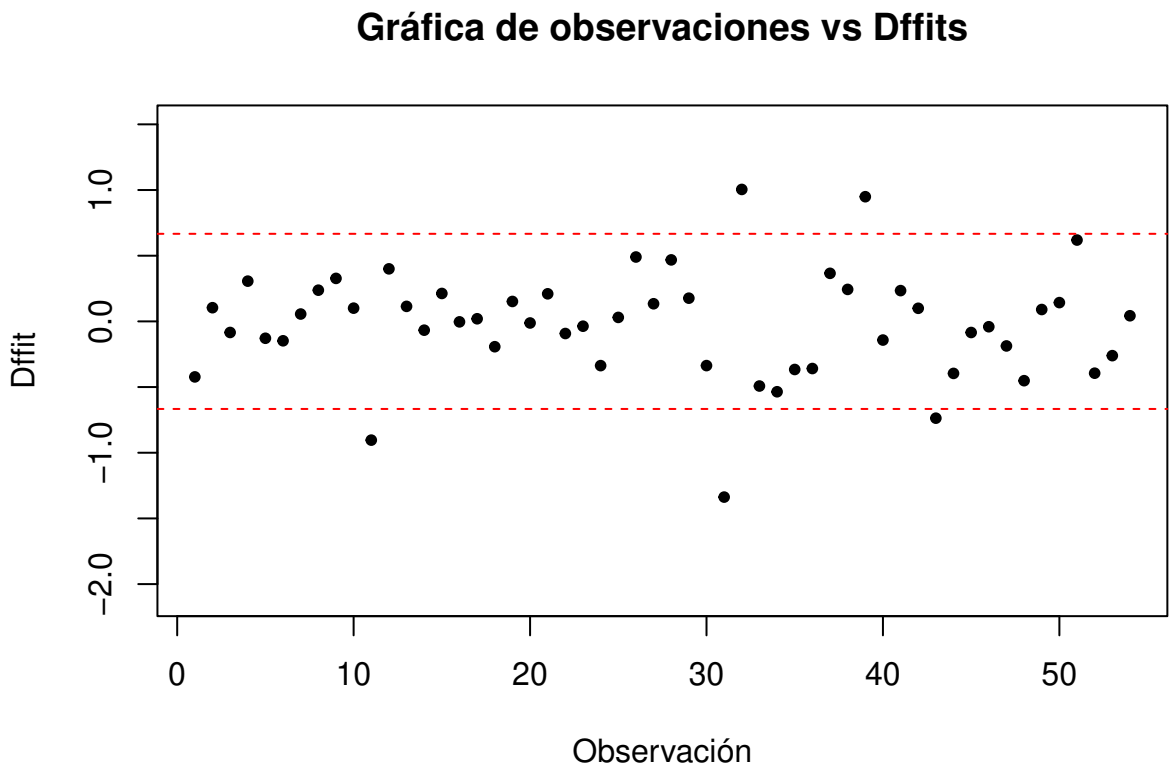

Figura 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 11  -0.7940  0.1373    0.5665 -0.9040
## 31  -1.8745  0.2822    0.3252 -1.3376
## 32   1.9859  0.1577    0.1935  1.0046
## 39   1.6975  0.1442    0.2309  0.9492
## 43  -2.0750  0.0841    0.1049 -0.7369
```

Una observación es influencial si tiene un impacto notable sobre los coeficientes de regresión ajustados, es decir, es influencial si su exclusión del modelo causa cambios importantes en la ecuación de regresión ajustada. Para saber la existencia de datos influenciales, se analizan dos criterios: el criterio de Dffits y el criterio de distancias de Cook.

El primer criterio dice que para cualquier punto cuyo $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, es un punto influencial; en este caso los puntos 11, 31, 32, 39, 43 cumplen con este criterio.

El criterio de distancias de Cook afirma que cualquier punto cuya $D_i > 1$, es un punto influencial, como se observa ninguno de los datos cumple con este criterio.

En resumen, para el análisis de observaciones extremas se tiene que:

- No hay valores atípicos.

- Las observaciones 11, 15, 31, 39, y 52 son puntos de balanceo.
- Las observaciones 11, 31, 32, 39, y 43 son influenciales.

4.3. Conclusión

2,5 pt

Según los datos asignados, el modelo ajustado que representa el riesgo de infección como respuesta a 5 variables predictoras que son: duración de la estadía, rutina de cultivos, número de camas, censo promedio diario y número de enfermeras, no es un modelo válido, a pesar de que cumple con los supuestos de varianza constante y media cero, luego de un análisis a profundidad, se observa que no cumple con el supuesto de normalidad de los errores, lo cual puede derivar de la presencia de 5 observaciones influenciales encontradas mediante el diagnóstico DFFITS, las cuales causan cambios importantes en la ecuación de regresión ajustada, además de tener en cuenta la presencia de 5 puntos de balanceo (3 de estos presentes en los puntos influenciales) que pueden afectar los errores estándar de los coeficientes estimados.

La presencia de los DFFITS se puede deber tanto por un error de ajuste grande como por un gran balanceo, por eso, los puntos detectados por este criterio deben ser investigados.

no son puntos, es un método para identificarlos.