Trabajo 1

9,7

Estudiantes
Julián Pareja Toro
Nicole Juliana Mejía Montes
Maria Camila Agudelo Espinosa
Juan Miguel Cadavid Jimenez

Equipo
33

Docente
Mateo Ochoa Medina

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
5 de octubre de 2023

Índice

1. Pregunta 1 **3**

1.1. Modelo de regresión . . . 3

1.2. Significancia de la regresión . . . 3

1.3. Significancia de los parámetros . . . 4

1.4. Interpretación de los parámetros . . . 4

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 . . . 5

3. Pregunta 3 **6**

3.1. Prueba de hipótesis y prueba de hipótesis matricial . . . 6

3.2. Estadístico de prueba . . . 6

4. Pregunta 4 **7**

4.1. Supuestos del modelo . . . 7

4.1.1. Normalidad de los residuales . . . 7

4.1.2. Varianza constante . . . 8

4.2. Verificación de las observaciones . . . 8

4.2.1. Datos atípicos . . . 9

4.2.2. Puntos de balanceo. . . . 10

4.2.3. Puntos influenciales. . . . 11

4.3. Conclusión. . . . 12

Índice de figuras

1. Gráfico cuantil-cuantil y normalidad de residuales 7
2. Gráfico residuales estudentizados vs valores ajustados 8
3. Identificación de datos atípicos 9
4. Identificación de puntos de balanceo 10
5. Criterio distancias de Cook para puntos influenciales 11
6. Criterio Dffits para puntos influenciales 12

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
5. Diagnóstico puntos de balanceo 11
6. Diagnóstico Criterio Dffits 13

1. Pregunta 1

20pt

Teniendo en cuenta la base de datos brindada, en la cual hay 5 variables regresoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i \; ; \;\; \varepsilon_i \sim iid\ N(0, \sigma^2), i = 1,2, \ldots, 69$$

Donde

- 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 periodo del estudio
- X_4 = Censo promedio diario: Número promedio de pacientes en el hospital por día durante el periodo del estudio
- X_5 = Número de enfermeras: Número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio

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	-0.4315
β_1	0.1307
β_2	0.0168
β_3	0.0575
β_4	0.0142
β_5	0.0023

3pt

Por lo tanto, el modelo de regresión ajustado es:

$$\widehat{Y}_i = -0.4315 + 0.1307 X_{i1} + 0.0168 X_{i2} + 0.0575 X_{i3} + 0.0142 X_{i4} + 0.0023 X_{i5}; \;\; i = 1,2, \ldots, 69$$

1.2. Significancia de la regresión

Para analizar la significancia de la regresión, se plantea el siguiente juego de hipótesis:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1: \text{Algún } \beta_j \text{ distinto de 0 para } j = 1,..,5 \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \sim H_0 \, f_{5,63}$$

5pt

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	58.5877	5	11.71754	14.8726	1.26639e-09
Error	49.6352	63	0.78786		

De la tabla Anova, se observa un valor P aproximadamente igual a 0, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leq j \leq 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$ con $1 \leq j \leq 5$, 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

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	$T_{0,j}$	P-valor
β_0	-0.4315	1.4759	-0.2923	0.7709
β_1	0.1307	0.0766	1.7048	0.0931
β_2	0.0168	0.0266	0.6311	0.5302
β_3	0.0575	0.0136	4.2176	0.0000
β_4	0.0142	0.0077	1.8359	0.0710
β_5	0.0023.	0.0007	3.3399	0.0014

6pt

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$ los parámetros β_3 y β_5 son significativos, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

Dado que el 0 no se encuentra en el rango de los valores observados de las covariables del modelo ajustado, el parámetro β_0 no cuenta con una interpretación para el modelo. Por su parte, los parámetros significativos seleccionados arriba, indican lo siguiente:

$\hat{\beta}_3$: Dado un aumento en una unidad en el número promedio de camas en el hospital durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital aumenta en 0.0575 (5.75%), cuando las demás covariables permanecen constantes.

3pt

$\hat{\beta}_5$: Dado un aumento en una unidad en el promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio, la probabilidad promedio estimada de adquirir infección en el hospital aumenta en 0.0023 (0.23%), cuando las demás covariables permanecen constantes.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple R^2 dado por:

3pt

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{58.5877}{58.5877 + 49.6352} = \frac{58.5877}{108.2229} = \mathbf{0.5413614}$$

Lo que significa que aproximadamente el 54.13% de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

Por su parte, el modelo tiene un coeficiente de determinación múltiple R_{adj}^2 dado por

$$R_{adj}^2 = 1 - \frac{(n-1)MSE}{SST} = 1 - \frac{(69-1)(0.78786)}{58.5877 + 49.6352} = \frac{53.57448}{108.2229} = \mathbf{0.49504}$$

Que es un valor menor R^2 lo que indica que en el modelo puede haber variables que no aporten significativamente a la respuesta $\hat{Y}_i$.

2. Pregunta 2

5pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más pequeño en el modelo fueron X_3, X_4 *y* X_5, por lo tanto, a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} H_0: \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1: \text{Algún } \beta_j \text{ distinto de 0 para } j = 3,4,5 \end{cases}$$

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	49.6352	X_1 X_2 X_3 X_4 X_5
Modelo reducido	78.443	X_1 X_2

Así, el modelo completo estará dador por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i \,; \;\; \varepsilon_i \sim iid\ N(0, \sigma^2), i = 1,2, \dots ,69$$

Y el modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \,; \;\; \varepsilon_i \sim iid\ N(0, \sigma^2), i = 1,2, \dots ,69$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

3pt

$$F_0 = \frac{\frac{SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)}{3}}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \sim H_0\, f_{3,63}$$

$$= \frac{\frac{78.443 - 49.635}{3}}{0.78786}$$

$$= 12.18833$$

Ahora, comparando el estadístico F_0 con el cuantil $f_{0.05,3,63} = 2.750541$, podemos corroborar que $F_0 > f_{0.05,3,63}$, y por ende rechazamos la hipótesis nula. Por tanto, alguno de los parámetros β_3 , β_4 o β_5 es significativo y no puede descartarse del modelo.

2pt

3. Pregunta 3

4pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Queremos probar si el efecto de la duración de la estadía sobre el riesgo de infección promedio de un paciente es el doble que el efecto que tiene el número de camas en el hospital sobre el mismo riesgo de infección promedio. Además, por medio de la misma prueba queremos saber si la cantidad de pacientes diaria en el hospital afecta 5 veces más al riesgo de infección promedio que el número de enfermeras. Por consiguiente, se plantea la siguiente prueba de hipótesis:

$$\begin{cases} H_0: \beta_1 = 2\beta_3;\ \beta_4 = 5\beta_5 \\ H_1: \beta_1 \neq 2\beta_3\ \vee\ \beta_4 \neq 5\beta_5 \end{cases}$$

Reescribiendo matricialmente:

$$\begin{cases} H_0: \boldsymbol{L}\beta = 0 \\ H_1: \boldsymbol{L}\beta \neq 0 \end{cases}$$

Con **L** dada por

2pt

$$\begin{bmatrix} 0 & 1 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -5 \end{bmatrix}$$

El modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3}{}^* + \ \beta_5 X_{i5}{}^* + \varepsilon_i\ ;\ \ \varepsilon_i \sim iid\ N(0, \sigma^2), i = 1,2, \ldots ,69$$

1pt

Donde:

$$X_{i3}{}^* = 2X_{i1} + X_{i3}$$

$$X_{i5}{}^* = 5X_{i4} + X_{i5}$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

1pt

$$F_0 = \frac{\frac{SSE(\cancel{\beta_0, \beta_2, \beta_3}, \beta_5) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)}{2}}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \sim_{H_0} f_{2,63}$$

$$= \frac{\frac{SSE(\cancel{\beta_0, \beta_2, \beta_3, \beta_5}) - 49.6352}{2}}{0.78786}$$

4. Pregunta 4

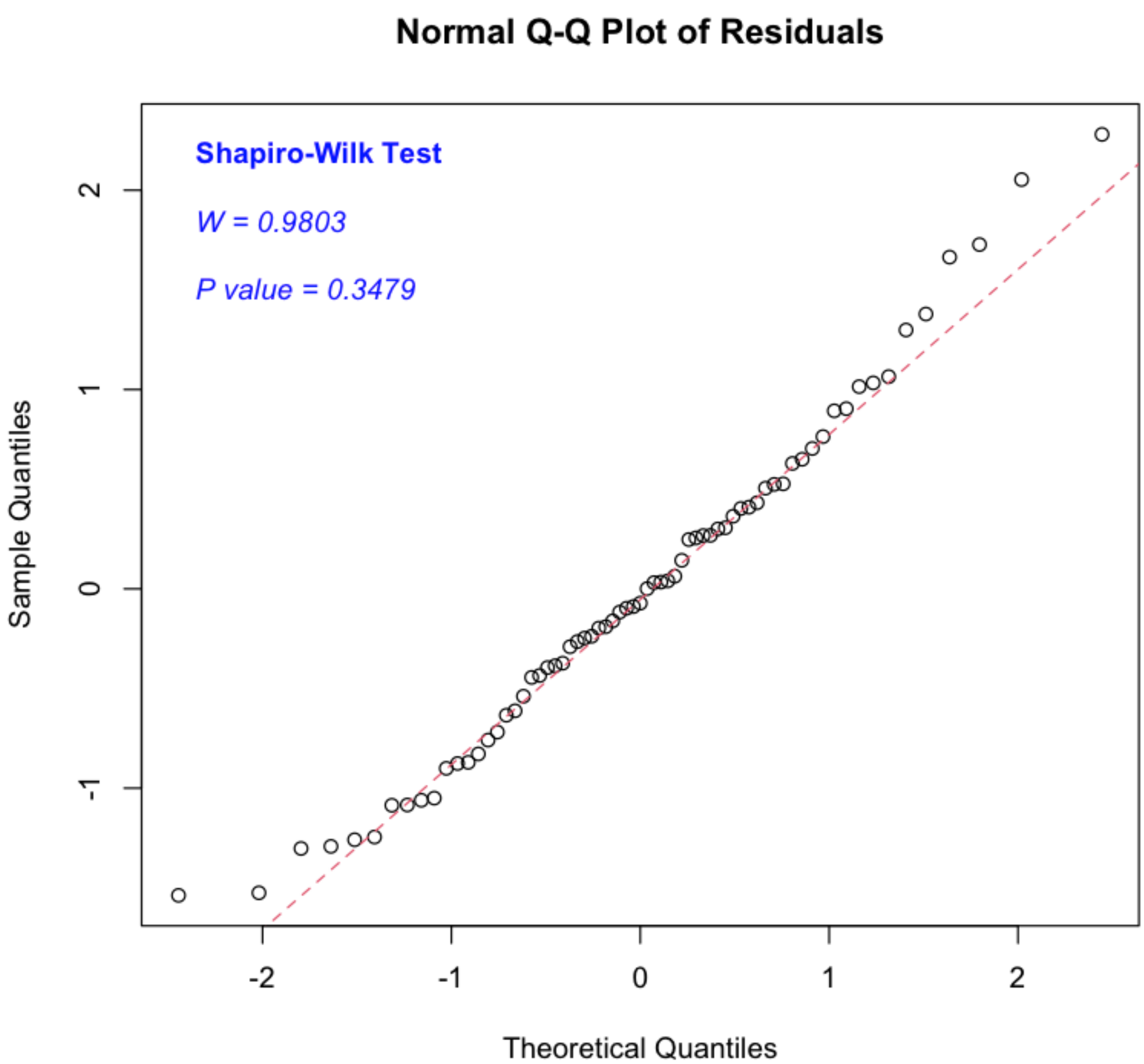
18pt

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:

$$\begin{cases} H_0: \varepsilon_i \sim \text{Normal} \\ H_1: \varepsilon_i \nsim \text{Normal} \end{cases}$$

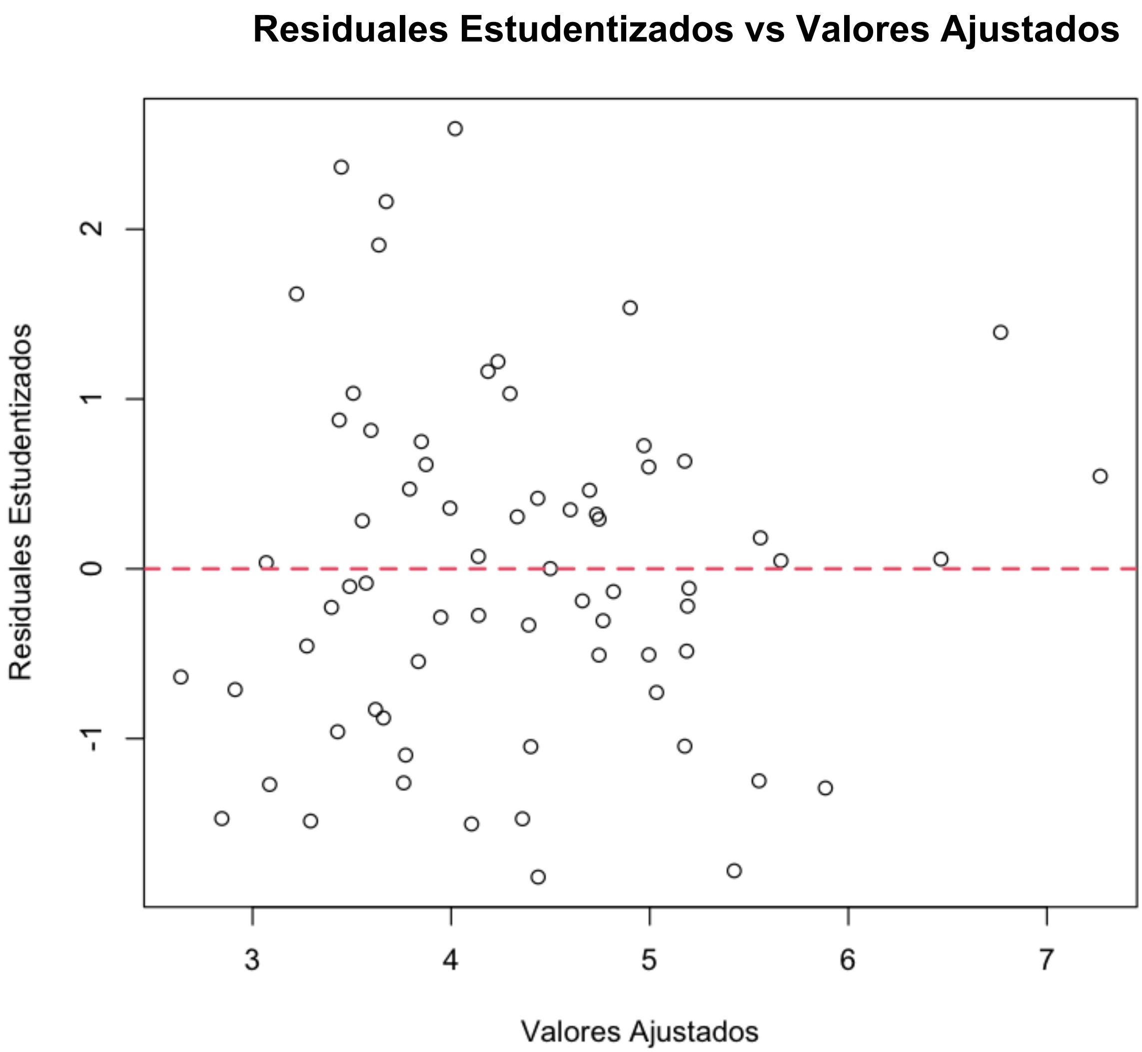


4pt

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.3479 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 que asumimos que los errores distribuyen normal con media μ y varianza σ^2. Sin embargo, la gráfica de comparación de cuantiles permite ver colas más pesadas y patrones irregulares que no se ajustan completamente a la línea recta de cuantiles teóricos, por tanto, y al tener más poder el análisis gráfico, rechazamos la hipótesis nula y afirmamos que el supuesto de normalidad no se cumple para los errores del modelo. 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 existe un patrón que muestra un "achicamiento" en la distancia de los valores en forma de embudo que no permite dar cuenta de una varianza que se mantenga constante para todos los valores ajustados. Por tanto, rechazamos el supuesto de varianza constante de los errores.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

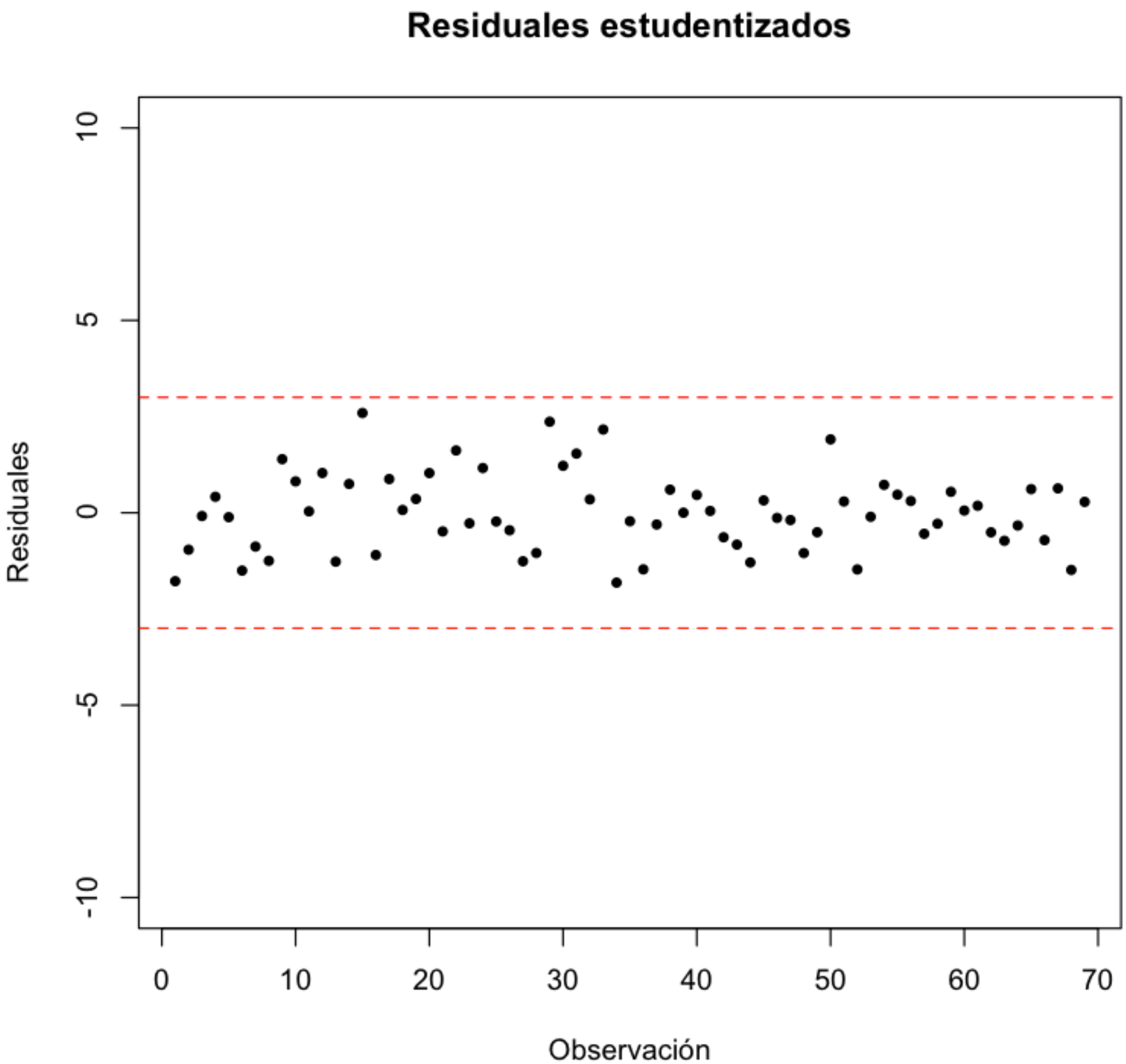


Figura 3: Identificación de datos atípicos

3 pt

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 $|r_{estud}| > 3$ delimitado por las líneas horizontales de color rojo.

4.2.2. Puntos de balanceo

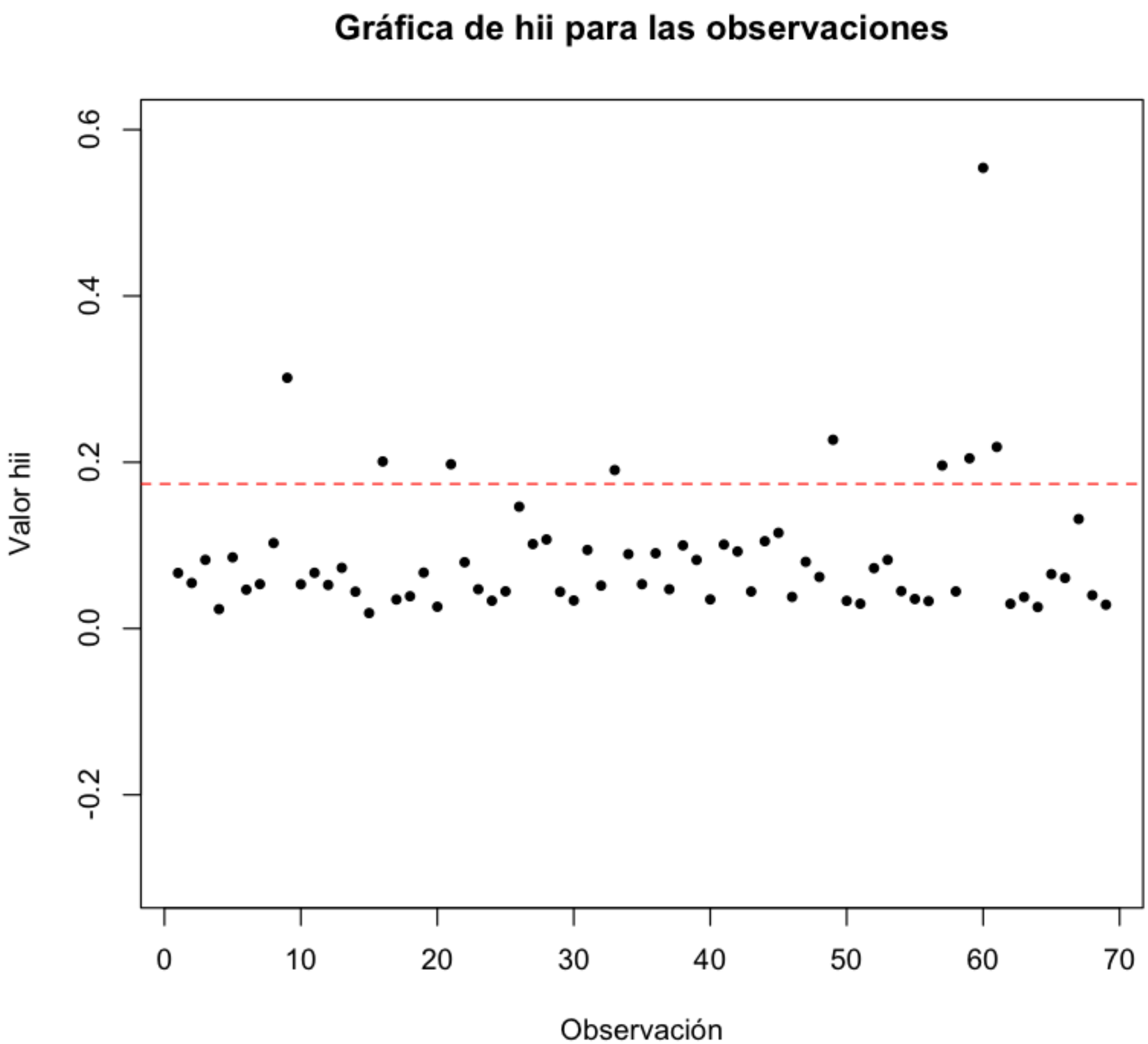


Figura 4: Identificación de puntos de balanceo

Cuadro 5: Diagnóstico puntos de balanceo

	res.stud	*Cooks.D*	*hii.value*	*Dffits*
9	13.929	0.1396	0.3015	0.9223
16	-10.975	0.0505	0.2009	-0.5512
21	-0.4851	0.0097	0.1976	-0.2393
33	21.629	0.1836	0.1906	10.821
49	-0.5064	0.0126	0.2270	-0.2728
57	-0.5463	0.0121	0.1961	-0.2683
59	0.5455	0.0128	0.2047	0.2752
60	0.0570	0.0007	0.5541	0.0630
61	0.1822	0.0015	0.2184	0.0956

3 ℓ +

Al observar la gráfica de *observaciones* vs *valores* h_{ii}, donde la línea punteada rojarepresenta el valor $h_{ii} = \frac{2p}{n} = \frac{2(6)}{69} \approx 0.174$, se puede apreciar que existen 9 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > \frac{2p}{n}$, los cuales son los presentados en la tabla.

4.2.3. Puntos influenciales

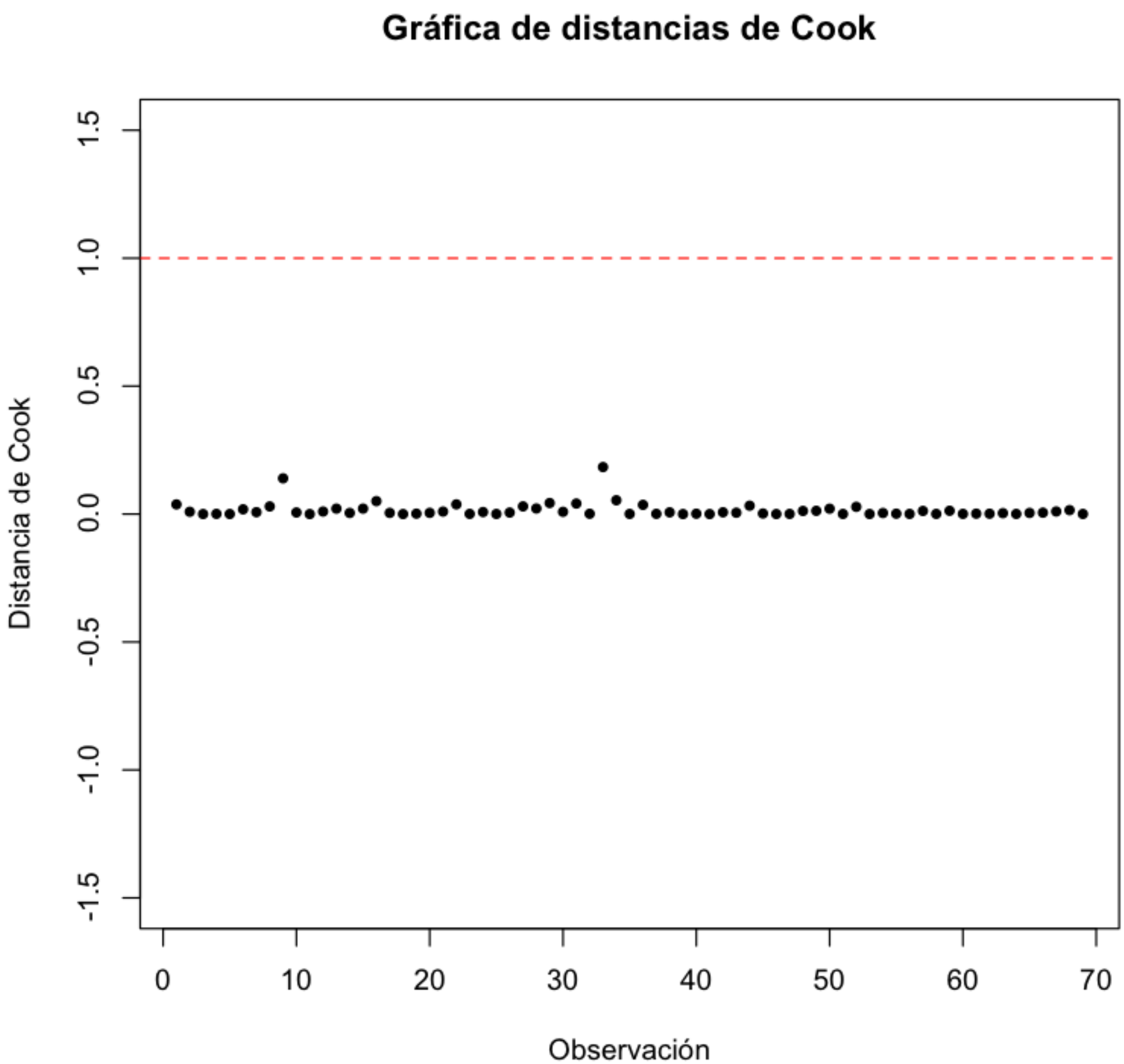


Figura 5: Criterio distancias de Cook para puntos influenciales

Respecto del criterio de distancias de Cook, gráficamente podemos corroborar que ninguna observación está por encima del criterio $D_i > 1$; por tanto, no se identifican observaciones influénciales.

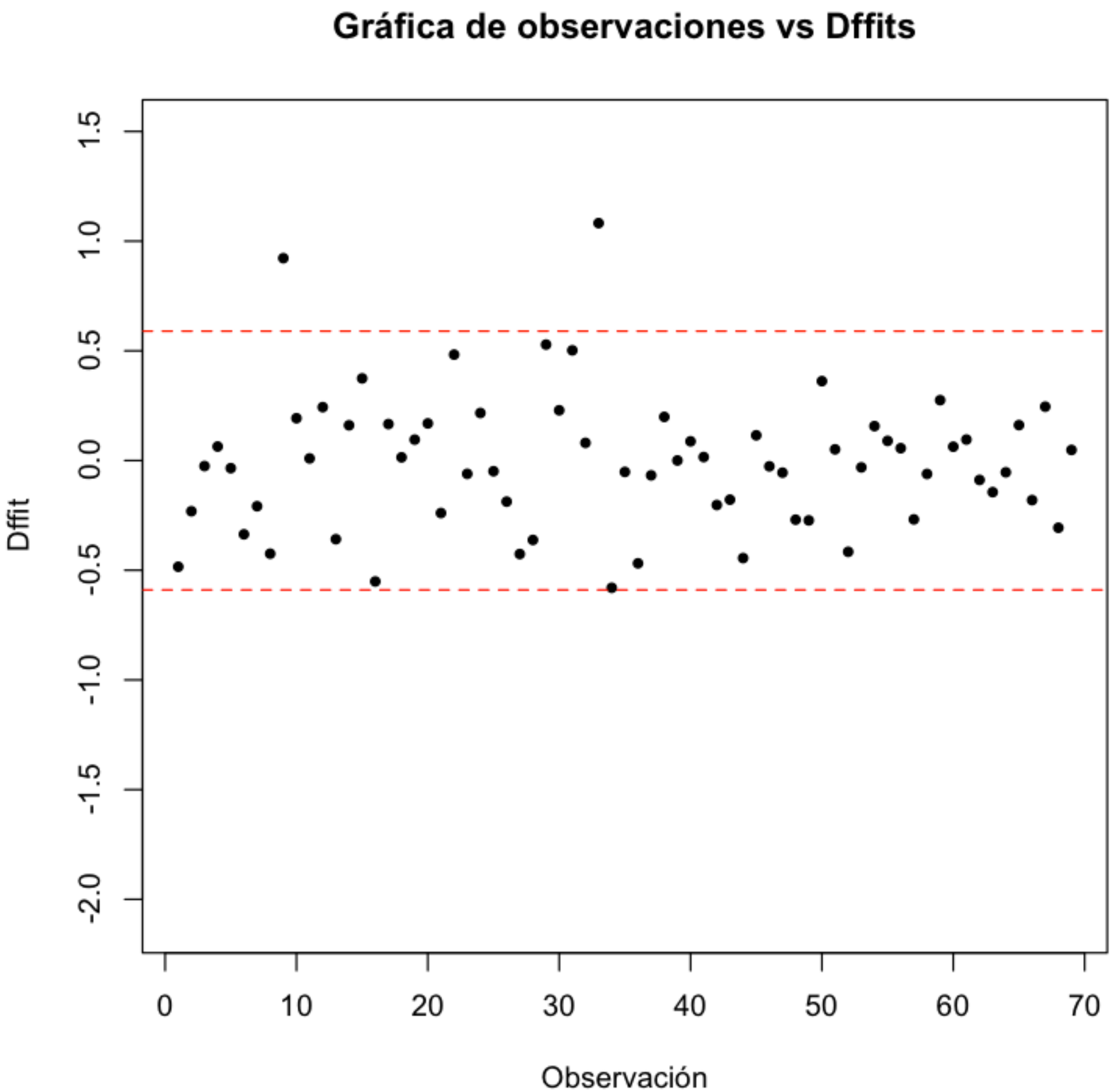


Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Diagnóstico Criterio Dffits

	res.stud	*Cooks.D*	*hii.value*	*Dffits*
9	13.929	0.1396	0.3015	0.9223
33	21.629	0.1836	0.1906	10.821

Por su parte, con el criterio Dffits podemos ver que existen dos observaciones influénciales según el criterio $|DFFITS_i| > 2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{6}{69}} \approx 0.5897$.

4.3. Conclusión

lpr

En resumen, tenemos que:

- No hay datos atípicos en el modelo, por tanto, no se cuenta con observaciones que estén superadas en la respuesta (Y) por el resto de las observaciones.

- Existen 9 datos del conjunto que son puntos de balanceo, por tanto, se cuenta con observaciones en el espacio de las predictoras alejadas del resto de la muestra.

- De acuerdo con las distancias de Cook, no se identificaron observaciones influénciales, pero el criterio Dffits identificó dos observaciones que sí lo son. Por tanto, el modelo cuenta con observaciones que tienen impacto sobre los coeficientes de regresión ajustados que los "hala" en su dirección.

Estos resultados nos permiten afirmar que el modelo no es válido para explicar la probabilidad promedio estimada de adquirir infección en un hospital, en tanto que el supuesto de normalidad y de varianza constante no parecen cumplir en este caso (de acuerdo al criterio fuerte de las gráficas arriba analizadas). Igualmente, se cuenta solo con dos parámetros significativos y un R^2 de 54.13% .

) Falso, es únicamente por los supuestos