Trabajo 1

4,6

Estudiantes

Miller Johan Chica Acero
Estefanía Ríos Cordero
Jerónimo Ledesma Patiño
Catalina Restrepo Salgado

Equipo # 4

Docente

Julieth Verónica Guarín Escudero

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

Durante la elaboración del presente informe se tendrá en cuenta la información proporcionada en la base de datos Equipo04.txt, donde puede encontrarse una muestra de 64 datos recopilados en un estudio a gran escala realizado en diferentes hospitales de EE.UU, a fin de analizar la eficacia en el control de infecciones hospitalarias.

En dicho volumen de datos puede identificarse una variable respuesta y 5 variables regresoras, con las cuales se ha planteado el siguiente modelo de regresión lineal múltiple.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Donde:

- Y: Riesgo de infección (en porcentaje)
- X_1: Duración de la estadía (en días)
- 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

En primer lugar, se realiza el ajuste del modelo de regresión lineal, donde se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-0.7559
β_1	0.1246
β_2	0.0290
β_3	0.0559
β_4	0.0139
β_5	0.0017

Dado lo anterior, el modelo de regresión ajustado se muestra a continuación:

$$\hat{Y}_i = -0.7559 + 0.1246 X_{1i} + 0.029 X_{2i} + 0.0559 X_{3i} + 0.0139 X_{4i} + 0.0017 X_{5i}$$

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Donde se puede definir el siguiente estadístico de prueba:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,58} \tag{1}$$

Es necesario, entonces, observar los resultados presentados en la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	49.9446	5	9.98891	9.06443	2.06947e-06
Error	63.9154	58	1.10199		

Teniendo en cuenta lo anterior, se llega a la conclusión de que, con un nivel de significancia $\alpha = 0.05$, puede rechazarse la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$. De esta manera, se acepta la hipótesis alternativa en la que algún $\beta_j \neq 0$, y puede decirse que el modelo de regresión es significativo, aunque todavía no es posible afirmar que este sea válido.

1.3. Significancia de los parámetros

Posteriormente al análisis de la significancia de la regresión, es conveniente determinar la significancia individual de los parámetros. Para esto es muy importante apoyarse de la información proporcionada en el siguiente cuadro.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-0.7559	1.5495	-0.4879	0.6275
β_1	0.1246	0.0765	1.6301	0.1085
β_2	0.0290	0.0289	1.0062	0.3185
β_3	0.0559	0.0169	3.3164	0.0016
β_4	0.0139	0.0081	1.7071	0.0932
β_5	0.0017	0.0008	2.2593	0.0276

De acuerdo con los valores-P que se observan en la tabla, es posible concluir con un nivel de significancia $\alpha = 0.05$, que los parámetros β_3 y β_5 son significativos, puesto que poseen un valor-P menor al α determinado anteriormente.

1.4. Interpretación de los parámetros

En consecuencia con el análisis de significancia realizado, puede realizarse la siguiente interpretación de los parámetros relacionados a dos de las variables predictoras presentes en el modelo.

- $\hat{\beta}_3$**:** Este parámetro indica que, por cada unidad que aumenta el número de camas del hospital, el promedio del porcentaje de riesgo de infección también aumenta en 0.0559 unidades; esto ocurre cuando las demás variables predictoras consideradas se mantienen fijas.
- $\hat{\beta}_5$**:** Por su parte, este parámetro da a entender que, por cada unidad que aumenta el número de enfermeras durante el período del estudio, el promedio del porcentaje de riesgo de infección aumenta en 0.0017 unidades, teniendo en cuenta también que las demás variables predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

Para finalizar esta sección del informe, es necesario realizar un análisis de bondad de ajuste al modelo, que puede observarse a continuación.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} = \frac{49.9446}{49.9446 + 63.9154} = 0.4386492$$

Teniendo en cuenta el resultado anterior, se puede decir que el modelo tiene un coeficiente de determinación múltiple $R^2 = 0.4386492$, lo cual significa que, aproximadamente el 43.86 % de la variabilidad total observada en el porcentaje de riesgo de infección es explicada por el modelo propuesto. Sin embargo, es importante tener en cuenta que el R^2 ajustado es una medida más adecuada para analizar la bondad de ajuste del modelo.

$$R^2_{adj} = 1 - \frac{(n-1)\cdot MSE}{SST} = 1 - \frac{(n-1)\cdot MSE}{SSR + SSE} = 1 - \frac{63 \cdot 1.10199}{49.9446 + 63.9154} = 0.3902567$$

Dicho esto, y observando el resultado obtenido anteriormente, se puede evidenciar que el R^2 ajustado presenta un valor menor al R^2 calculado, por lo que es posible concluir que este estadístico ha penalizado al modelo por la existencia de variables predictoras que no logran reducir la suma cuadrática del error SSE y, por el contrario, le restan grados de libertad.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

De acuerdo con el análisis de significancia realizado a los parámetros del modelo, X_3, X_4 y X_5 se pueden identificar como las covariables con los valores-P más pequeños, de manera que se realizará la prueba de hipótesis que se muestra a continuación.

$$\begin{cases} \mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 3, 4, 5 \end{cases}$$

De la tabla de todas las regresiones posibles, se deben extraer los datos correspondientes al modelo completo (es decir, que contiene todas las variables predictoras) y al modelo reducido (que corresponde al modelo completo bajo la hipótesis nula). Esta información puede observarse en el siguiente cuadro:

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	63.915	X1 X2 X3 X4 X5
Modelo reducido	87.243	X1 X2

De esta manera, se llega al siguiente modelo reducido para la prueba de significancia del subconjunto dado anteriormente.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

2.2. Estadístico de prueba y conclusión

Para evaluar la hipótesis planteada anteriormente, se calcula el estadístico de prueba F_0 de la siguiente manera:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{(87.243 - 63.915)/3}{63.915/58} \\ &= 7.0563717 \end{aligned} \tag{2}$$

Observando este valor y llevándolo a comparación con el estadístico crítico, que corresponde al siguiente valor: $f_{0.05,3,58} = 2.7636$, es posible notar que $F_0 > f_{0.05,3,58}$; por lo tanto, se rechaza la hipótesis nula planteada anteriormente a un nivel de significancia $\alpha = 0.05$.

Teniendo en cuenta este resultado, se puede concluir que el conjunto de variables predictoras X_3, X_4 y X_5 no pueden descartarse del modelo en presencia de las demás variables.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

En esta sección del informe se evaluarán las siguientes preguntas con respecto al comportamiento del modelo ajustado, en términos de la relación entre los efectos de algunas variables predictoras sobre la variable respuesta. Los interrogantes a analizar son los siguientes:

- ¿El efecto en el porcentaje del riesgo de infección en los pacientes ocasionado por la duración de la estadía en el hospital en días, corresponde a cuatro veces el efecto causado por la rutina de cultivos realizada?.
- ¿El efecto causado por el censo promedio diario de pacientes en el hospital sobre el porcentaje del riesgo de infección corresponde a ocho veces el efecto ocasionado por el número de enfermeras durante el período del estudio?

Dadas las preguntas establecidas previamente, se plantea la prueba de hipótesis que se muestra a continuación.

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 4\beta_2;\ \beta_4 = 8\beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

Para analizar esta prueba de hipótesis lineal general es necesario reescribir el razonamiento anterior, llegando a este resultado:

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

Con la siguiente matriz $\mathbf{L}$.

$$L = \begin{bmatrix} 0 & 1 & -4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -8 \end{bmatrix}$$

Además, el vector $\underline{\beta}$ puede definirse de la siguiente manera:

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

Por lo tanto, es posible llegar al modelo reducido aquí presentado.

$$Y_i = \beta_0 + \beta_2 X^*_{2i} + \beta_3 X_{3i} + \beta_5 X^*_{5i} + \varepsilon_i;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Donde:

- $X^*_{2i} = 4X_{1i} + X_{2i}$
- $X^*_{5i} = 8X_{4i} + X_{5i}$

3.2. Estadístico de prueba

En concordancia con lo anterior, se plantea el estadístico de prueba F_0 como se observa seguidamente.

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} = \frac{(SSE(MR) - 63.9154)/2}{63.9154/58} \overset{H_0}{\sim} f_{2,58} \tag{3}$$

Finalmente, es posible decir que para rechazar H_0 con un nivel de significancia $\alpha = 0.05$, el estadístico de prueba deberá encontrarse en la región de rechazo que se muestra a continuación, aunque esto no pueda evaluarse con la información que se tiene actualmente.

$$RR = \left\{ F_0 > f_{0.05,2,58} \right.$$

4. Pregunta 4

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

El primer paso para realizar la validación de los supuestos del modelo es evaluar la normalidad de los residuales, en este sentido, se hace el planteamiento de la siguiente prueba de hipótesis a través de una prueba Shapiro-Wilk acompañada de un gráfico cuantil - cuantil.

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

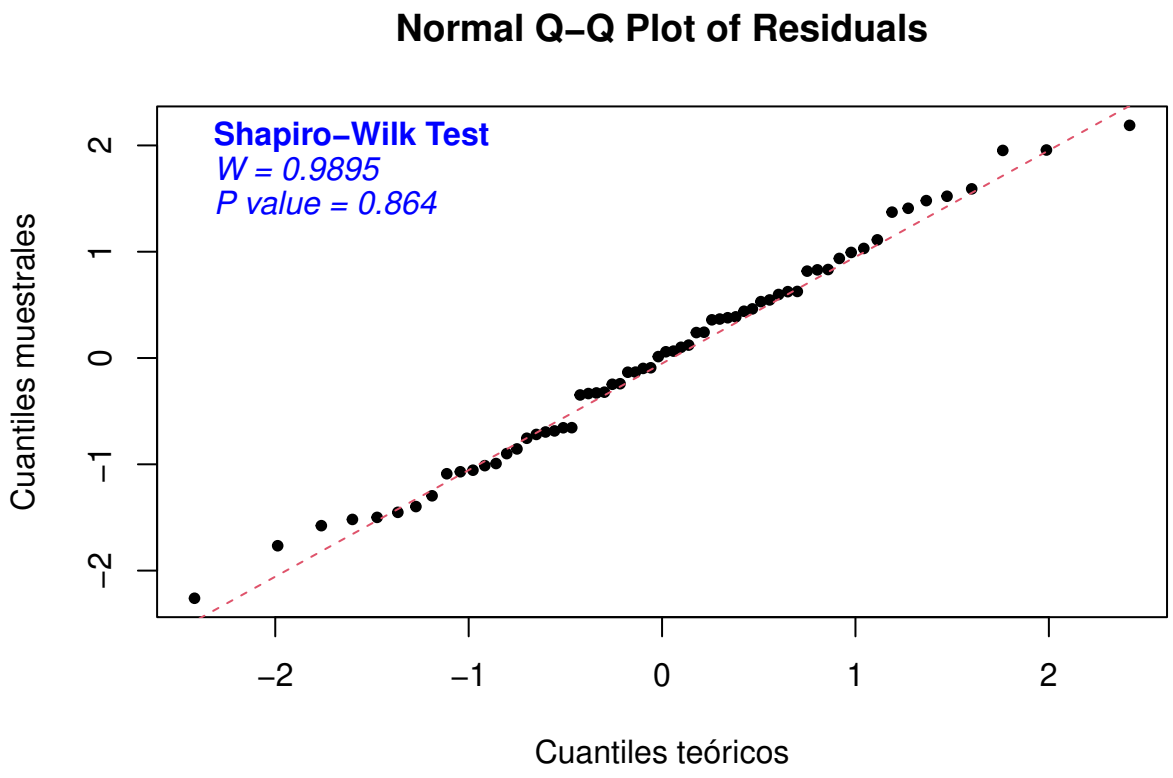

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Dado el resultado de la prueba Shapiro-Wilk, puede observarse con un nivel de significancia $\alpha = 0.05$, que el valor-P $= 0.864$ es considerablemente mayor a este valor de α (de hecho, es posible considerarlo bastante cercano a 1); por lo tanto, no se rechazaría la hipótesis nula establecida anteriormente.

Ahora bien, teniendo en cuenta los resultados proporcionados por el criterio gráfico, es posible decir que, aunque los residuales no se ajustan de manera perfecta a la forma normal esperada (puesto que presentan algunos patrones irregulares y faltas de ajuste que se concentran al inicio y al final), estos no son lo suficientemente notables como para rechazar el cumplimiento de este supuesto. Por consiguiente, no se rechaza la hipótesis nula y se valida que los datos distribuyen normal con media μ y varianza σ^2.

4.1.2. Varianza constante

Como segundo paso, es necesario validar si la varianza cumple con el supuesto de ser constante, para lo cual se utilizará el criterio gráfico (sabiendo que tiene más poder que otro tipo de pruebas).

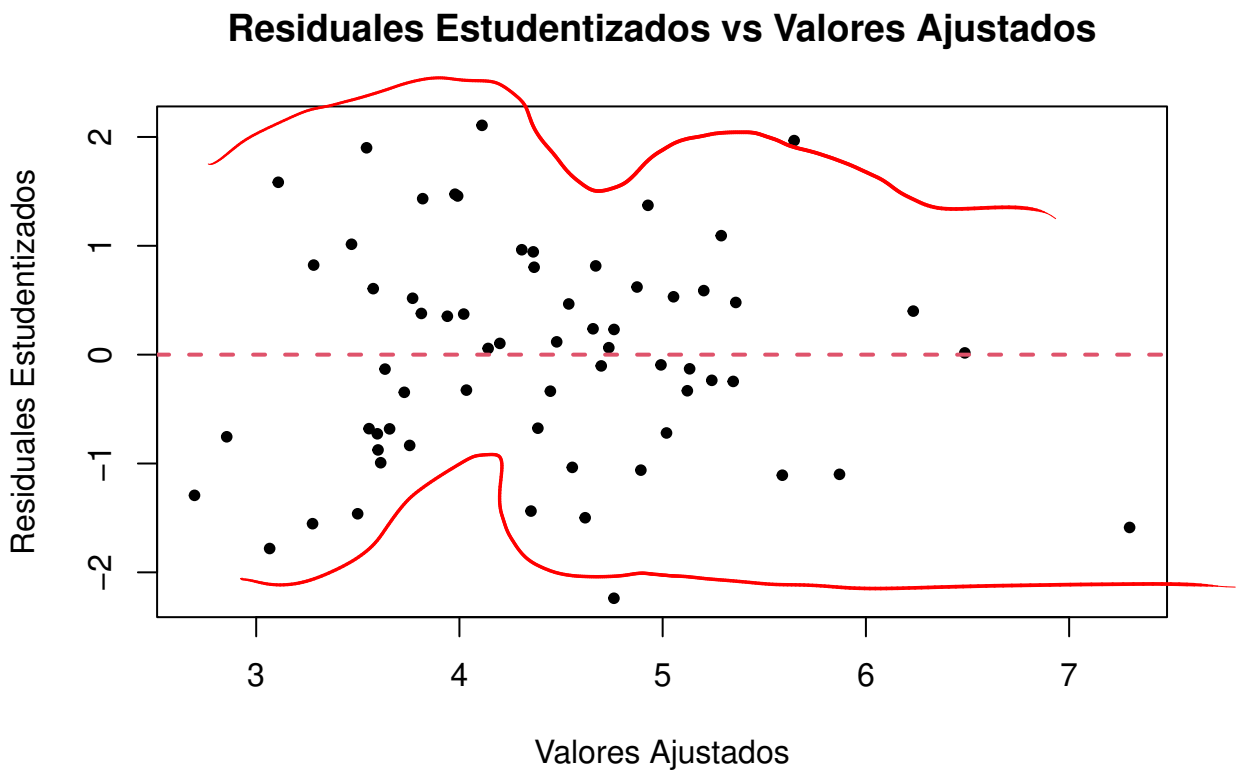

Figura 2: Gráfico residuales estudentizados vs valores ajustados

Como puede observarse en la gráfica que relaciona los residuales estudentizados con los valores ajustados, la dispersión de los datos no presenta patrones que pudieran sugerir un aumento o disminución de la varianza, a la misma vez que no se detectan efectos no lineales de variables que no se estuvieran considerando. De esta manera, es posible afirmar que no hay información suficiente para descartar el supuesto de varianza constante, por lo que se acepta y se confirma, adicionalmente, media cero en los datos analizados.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

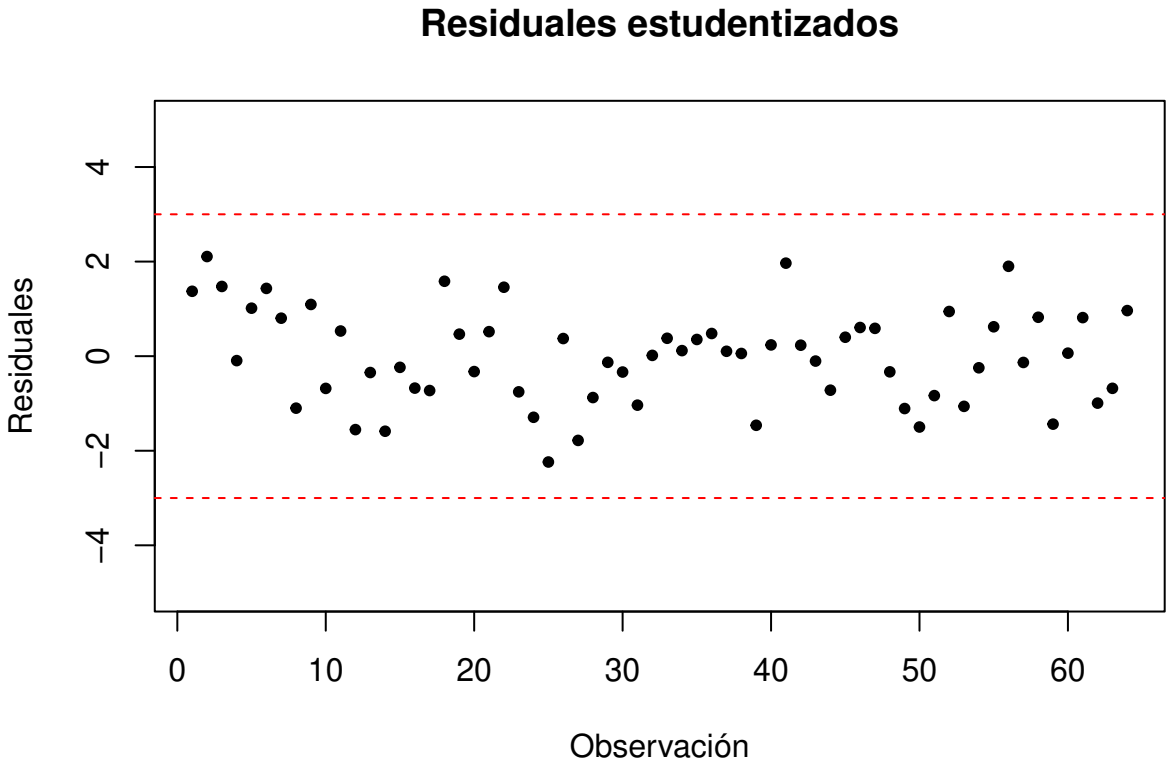

Figura 3: Identificación de datos atípicos

Teniendo en cuenta la gráfica anterior, y considerando que una observación es **atípica** cuando su residual estudentizado r_i sobrepasa el criterio de $|r_i| > 3$, es posible determinar que el conjunto de datos evaluado no contiene datos atípicos. Esto significa que ninguna de las observaciones tomadas durante el estudio está separada en su valor del porcentaje de riesgo de infección de las demás, por lo que no afectan los resultados del ajuste del modelo de regresión lineal múltiple.

4.2.2. Puntos de balanceo

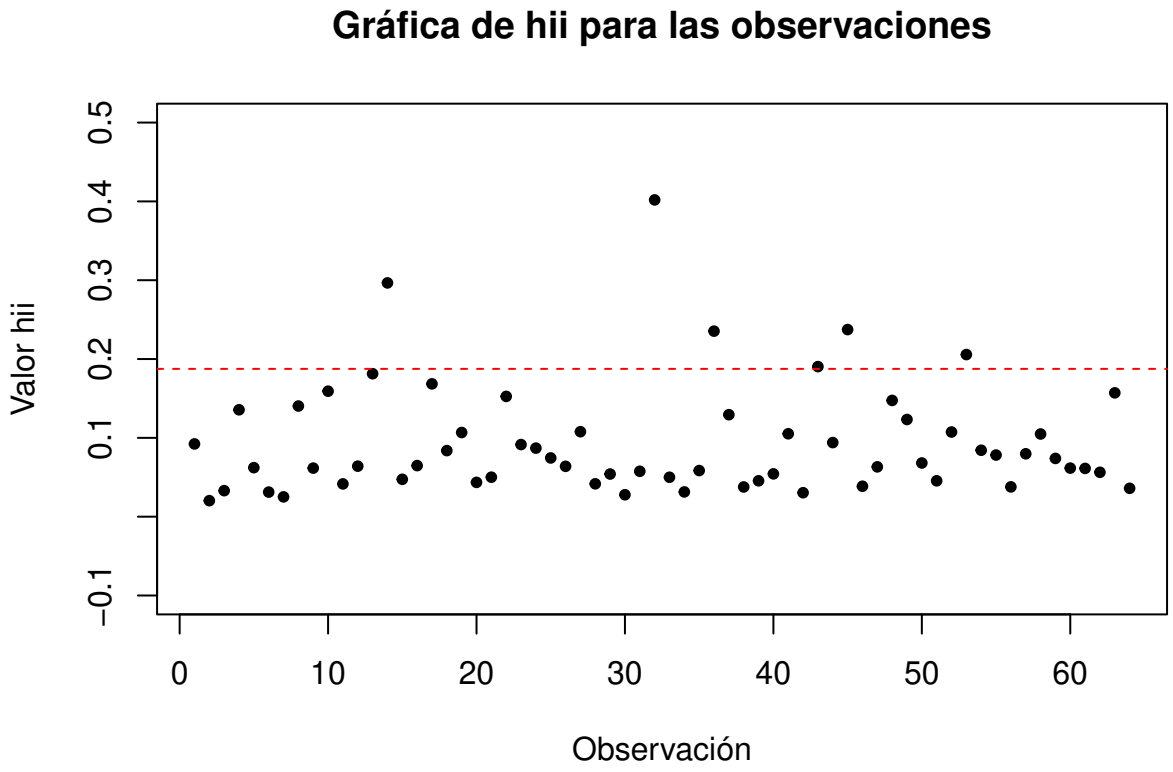

Figura 4: Identificación de puntos de balanceo

Cuadro 5: Identificación de observaciones que son puntos de balanceo

	res.stud	Cooks.D.	hii.value	Dffits
14	-1.5875	0.1771	0.2966	-1.0450
32	0.0163	0.0000	0.4019	0.0133
36	0.4795	0.0118	0.2354	0.2643
43	-0.1035	0.0004	0.1904	-0.0497
45	0.4002	0.0083	0.2374	0.2217
53	-1.0611	0.0486	0.2057	-0.5406

Sabiendo que una observación i se puede considerar como **punto de balanceo** si cumple que su valor h_{ii} es mayor al $h_{ii} = 2\frac{p}{n} = 2\frac{6}{64} = 0.1875$, representado en la línea punteada roja del gráfico que relaciona las observaciones con los valores hii.

Dicho lo anterior, es posible apreciar que 6 observaciones del conjunto de datos analizado son puntos de balanceo en los que se cumple que $h_{ii} > 0.1875$. Estos datos son:

- Observación $= 14$, $h_{ii} = 0.2966$

- Observación = 32, $h_{ii} = 0.4019$
- Observación = 36, $h_{ii} = 0.2354$
- Observación = 43, $h_{ii} = 0.1904$
- Observación = 45, $h_{ii} = 0.2374$
- Observación = 53, $h_{ii} = 0.2057$

¿Qué causan?

2pt

4.2.3. Puntos influenciales

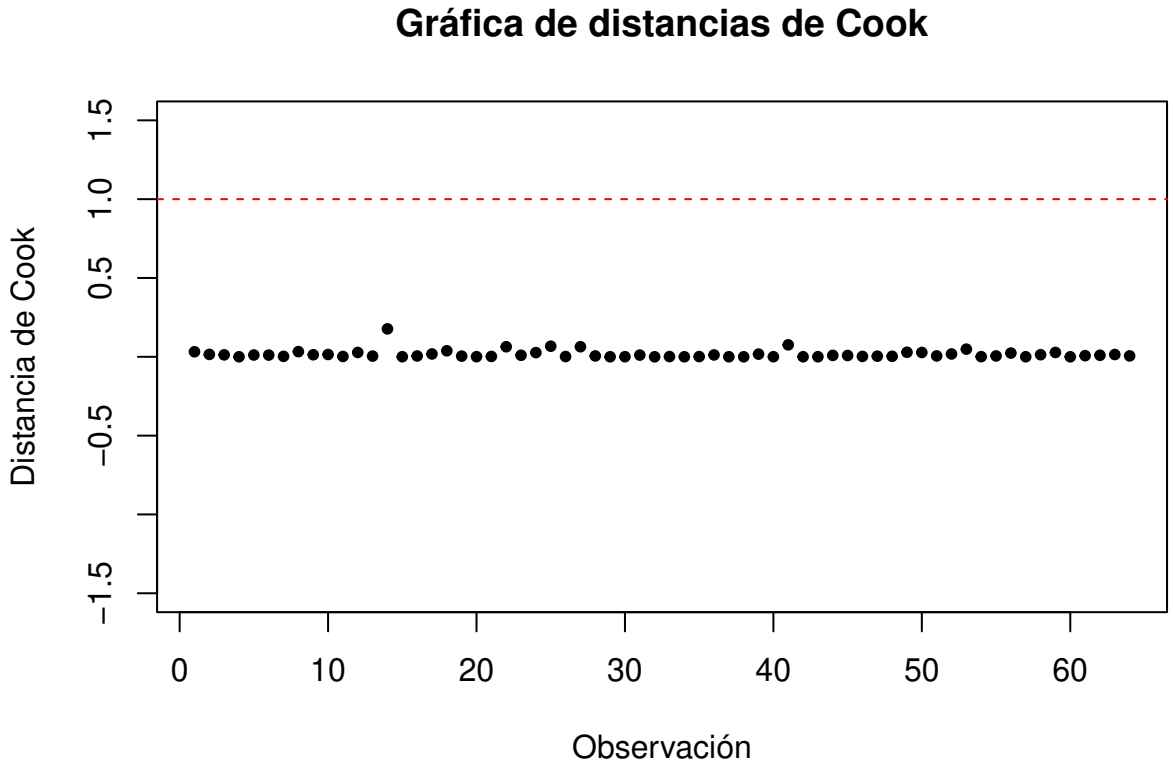

Figura 5: Criterio distancias de Cook para puntos influenciales

Por un lado, puede observarse que según el criterio de las **distancias de Cook**, en el cual se evalúa si alguna observación cumple $D_i > 1$, no hay observaciones influenciales en el modelo de regresión ajustado.

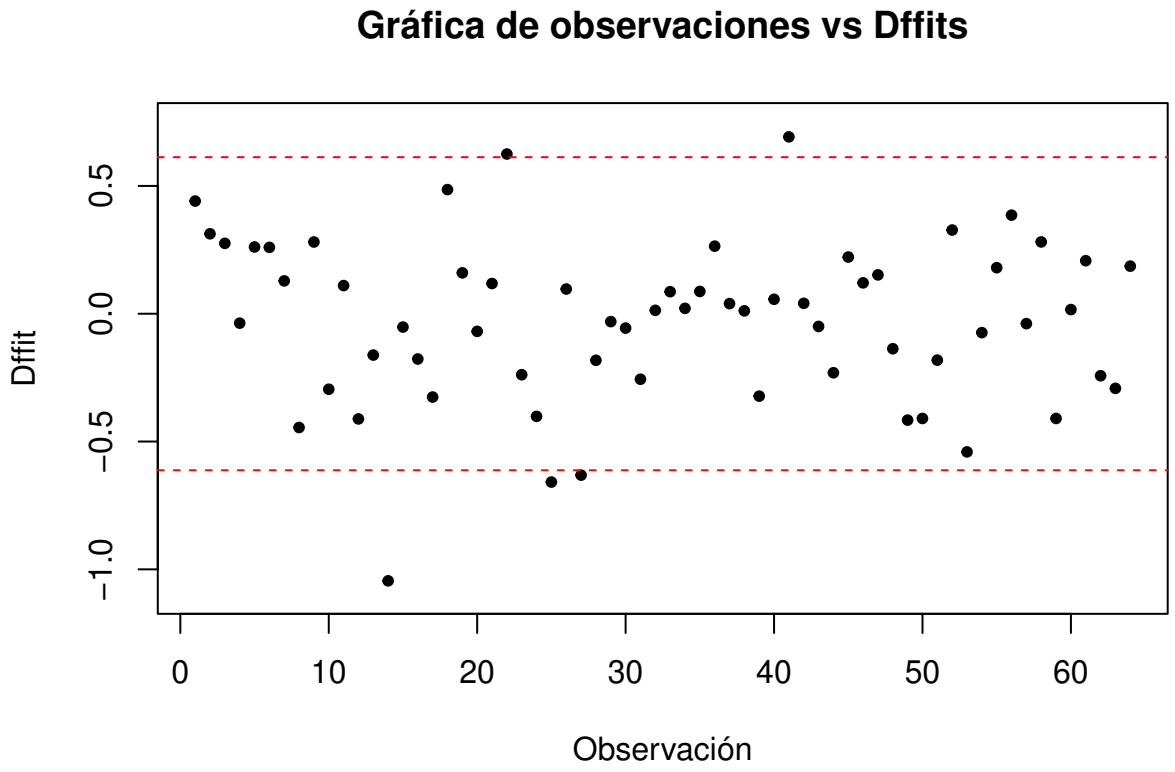

Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Identificación de observaciones influenciales

	res.stud	Cooks.D.	hii.value	Dffits
14	-1.5875	0.1771	0.2966	-1.0450
22	1.4577	0.0638	0.1526	0.6247
25	-2.2380	0.0672	0.0745	-0.6585
27	-1.7812	0.0639	0.1078	-0.6314
41	1.9667	0.0758	0.1052	0.6919

Teniendo en cuenta que una observación puede considerarse influencial según el **criterio de Dffits** si cumple que su valor $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{6}{64}} = 0.6123724$, se identifican 5 observaciones que son influenciales para el modelo ajustado. Dichos datos son los siguientes:

- Observación $= 14$, $D_{ffits} = -1.0450$
- Observación $= 22$, $D_{ffits} = 0.6247$
- Observación $= 25$, $D_{ffits} = -0.6585$
- Observación $= 27$, $D_{ffits} = -0.6314$

- Observación = 41, $D_{ffits} = 0.6919$

4.3. Conclusión

Para finalizar la elaboración del presente informe, en el que se ajustó un modelo de regresión lineal múltiple para analizar la eficacia en el control de infecciones hospitalarias, de acuerdo con una muestra de 64 datos obtenidos en un estudio a gran escala realizado en EE.UU, es posible determinar las siguientes conclusiones en términos de la validez del modelo.

- **Normalidad de los errores:** Teniendo en cuenta los resultados proporcionados por la prueba de Shapiro-Wilk y el criterio gráfico, se llega a la conclusión de que el modelo cumple este supuesto.
- **Varianza constante y media cero:** De acuerdo con el análisis realizado a través de la prueba gráfica, se llega a la conclusión de que el modelo cumple con el supuesto de varianza constante. Por otro lado, se asume el cumplimiento del supuesto de media cero.
- **Independencia:** El supuesto de independencia de los errores siempre se asume como cierto, en términos del análisis realizado en este curso.

Sumado a lo anterior, es posible decir que no se encontraron datos atípicos que afecten significativamente al modelo, así como 6 observaciones del conjunto de datos (14, 32, 36, 43, 45 y 53) pudieron identificarse como puntos de balanceo; por su parte, 5 observaciones (14, 22, 25, 27 y 41) se identificaron como puntos influenciales. No obstante, su número es relativamente pequeño comparado con el tamaño total del conjunto de datos y no se observó que incidieran de forma negativa en los resultados obtenidos en el análisis realizado en la sección anterior (aunque posiblemente pudieron haber afectado la medida de bondad de ajuste del modelo, que tiene un valor $R^2 = 0.4386492$).

Por lo tanto, como el modelo de regresión lineal múltiple cumple con todos los supuestos de los errores, es posible concluir que se considera **válido** para hacer predicciones de valores futuros en puntos de interpolación.

7 en este curso

7 No necesariamente es negativo

2,5 pt