3,5

Trabajo 1

Estudiantes

Angelica Maria Chaparro Rojas
Santiago Macías Ruíz
Samuel Palacio Morales
Josué Duque Gutierrez

Equipo #1

Docente

Julieth Veronica Guarín Escudero

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

Pregunta 1 11,5 pt

Teniendo en cuenta la base de datos 1, en la cual hay 5 variables regresores, determinados por:

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

Entonces, se plantea el modelo de regresión lineal múltiple:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ \ 1 \leqslant i \leqslant 55$$

Modelo de regresión 1,5 pt

¿?

Ajustando el modelo, el calculo nos arroja los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	2.9148
β_1	0.0875
β_2	-0.0193
β_3	0.0628
β_4	0.0026
β_5	0.0022

Supuestos no van en ec. ajustada

Al ajustar el modelo, tenemos que:

$$\hat{Y}_i = 2.9148 + 0.0875X_{1i} - 0.0193X_{2i} + 0.0628X_{3i} + 0.0026X_{4i} + 0.0022X_{5i} + \varepsilon_i\ \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 55$$

Modelo de regresión

¿cuál?

Una vez realizado el análisis, se logran obtener los coeficientes correspondientes al modelo ajustado, ~~los cuales son~~ los siguientes:

¿otra vez?

Cuadro 2: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	2.9148
β_1	0.0875
β_2	-0.0193
β_3	0.0628
β_4	0.0026
β_5	0.0022

~~Por lo tanto, el modelo de regresión ajustado es:~~

$\hat{Y}_i = 2.9148 + 0.0875X_{1i} - 0.0193X_{2i} + 0.0628X_{3i} + 0.0026X_{4i} + 0.0022X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 55$

Significancia de la regresión 3 pt

Con el fin de examinar la importancia estadística de la regresión, se formula el siguiente conjunto de hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,3,4, 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} = \frac{\frac{SSR}{5}}{\frac{SSE}{55-6}} \overset{H_0}{\sim} F_{5,49}$$

$$F_0 = \cancel{\frac{MST}{MSE}} \overset{H_0}{\sim} f_{5,49} \tag{1}$$

A continuación se muestra la tabla de Anova:

Cuadro 3: Tabla ANOVA para el modelo

	Sumas Cuadraticas	Grados de Libertad	Cuadrado medio	F_0	P-valor
Regresión	32.9947	5	6.59895	7.93706	1.53888e-05
Error	40.7391	49	0.83141		

Al analizar los resultados de la tabla Anova presentada, se puede observar que el valor P es aproximadamente igual a cero. Este resultado sugiere que la hipótesis nula, en la que todos los coeficientes de regresión son iguales a cero, puede ser rechazada. Por lo tanto, se puede aceptar la hipótesis alternativa de que al menos uno de los coeficientes de regresión no es cero.

Esta conclusión implica que la regresión es significativa en términos estadísticos y que al menos una de las variables predictoras tiene una relación lineal significativa con la variable

↳ marginal

 ¿Por qué?

de respuesta. Este resultado puede ser útil en la interpretación del modelo de regresión y ~~en la toma de decisiones basadas en la relación entre las variables incluidas en el modelo.~~ En resumen, la tabla Anova proporciona información valiosa para evaluar la significancia estadística de los coeficientes de regresión y para determinar la importancia de la regresión ~~en su conjunto.~~

Significancia de los parámetros

A continuación se muestra una tabla con información detallada de los parámetros, la cual será útil para identificar aquellos que resultan significativos.

Cuadro 4: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	2.9148	1.8579	1.5688	0.1231
β_1	0.0875	0.1123	0.7790	0.4397
β_2	-0.0193	0.0306	-0.6292	0.5321
β_3	0.0628	0.0169	3.7108	0.0005
β_4	0.0026	0.0083	0.3135	0.7553
β_5	0.0022	0.0007	2.8909	0.0057

5,5 pt

La evidencia estadística sugiere que la variable predictora asociada a β_3 y la variable predictora asociada a β_5 tienen una relación significativa con la variable ~~de~~ respuesta. Por lo tanto, es importante considerar estas variables ~~al interpretar los resultados del modelo~~ y al tomar decisiones basadas en las relaciones identificadas en el análisis de regresión.

Opt

Interpretación de los parámetros

Aumento en x provoca cambio en Y, no al revés.

$\hat{\beta}_3$**:** Por cada unidad de incremento en el porcentaje de riesgo de infección, el promedio de porcentaje del número de camas aumenta significativamente en 0.0643 unidades cuando los valores en las demás predictoras se mantienen fijos.

$\hat{\beta}_5$**:** Observamos que por cada unidad de incremento en el porcentaje de riesgo de infección, el promedio de porcentaje del número de enfermeras aumenta en 0.0019 unidades cuando los valores en las demás predictoras se mantienen fijos.

Coeficiente de determinación múltiple R^2

1,5 pt

No tiene nada que ver con la significancia

El valor del coeficiente de determinación múltiple R^2 en el modelo de regresión es de 0.4475, lo que indica que aproximadamente el 44.75 % de la variabilidad observada en la variable de respuesta puede ser explicada por el modelo propuesto. Aunque este valor no es muy alto, ~~sugiere que el modelo es significativo y~~ puede explicar una parte importante de la variabilidad observada en la variable de respuesta. Sin embargo, todavía puede haber factores adicionales que contribuyan a la variabilidad no explicada en la variable de respuesta.

s $\boxed{\text{es}}$ explicada

5 pt

Pregunta 2

Planteamiento pruebas de hipótesis y modelo reducido

Las variables X_1, X_2, X_4 presentaron los valores P más altos en el modelo de regresión, lo que indica que tienen una menor significancia estadística en relación a la variable de respuesta. Debido a esto, se planteó una prueba de hipótesis utilizando una tabla de todas las posibles combinaciones de regresión para determinar si se pueden eliminar estas variables del modelo sin afectar significativamente la calidad de ajuste del modelo. En resumen, se busca confirmar si estas variables pueden ser eliminadas del modelo sin afectar su capacidad para explicar la variabilidad observada en la variable de respuesta:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_4 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 1, 2, 4 \end{cases}$$

Cuadro 5: Resumen tabla de todas las regresiones

	SSE	variables explicativas en el modelo
Modelo completo	40.739	X1 X2 X3 X4 X5
Modelo reducido	41.663	X3 X5

Podemos observar que un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 55$$

Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_3, \beta_5) - SSE(\beta_0,\ \cdots,\ \beta_5))/3}{MSE(\beta_0,\ \cdots,\ \beta_5)} \overset{H_0}{\sim} f_{3,49} \\ &= \frac{0.308}{0.8314} \\ &= 0.37045 \end{aligned} \tag{2}$$

¿menor a qué?

Ahora, comparando el F_0 con $f_{0.95,3,49} = 2.7939$, se puede ver que $F_0 < f_{0.95,3,54}$. De esta manera observamos que, el subconjunto no es significativo, pues el valor de $F_0 = 0.370$ es menor al valor del cuantil $f_0 = 2.7939$, es decir que el subconjunto tiene una menor significancia estadística, por lo tanto, se puede determinar que estas variables pueden descartarse porque se rechaza H_0, por lo tanto las variables eliminadas no afectan la capacidad del modelo para explicar la variabilidad de la respuesta.

Pregunta 3 5 pt

Prueba de hipótesis y prueba de hipótesis matricial

Se plantea la siguiente prueba de hipótesis:

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

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 **L** dada por

$$L = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \checkmark \quad 2\,pt$$

El modelo completo se denota por:

$$Y_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i, \;\; \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \;\; 1 \leqslant i \leqslant 55$$

El modelo reducido se denota por:

$$Y_i = \beta_o + \beta_1 (X_{i1} + X_{i2}) + \beta_3 X_{i3} + \beta_4 (X_{i4} + X_{i5}) + \varepsilon_i, \;\; \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \;\; 1 \leqslant i \leqslant 55 \checkmark$$

~~#~~Donde $X_{2i}^* = X_{2i} + X_{4i}$ y $X_{3i}^* = 3X_{1i} + X_{3i}$ 1 pt

Estadístico de prueba

2 pt

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} = \frac{(SSE(MR) - 40.739)/2}{0.8314082} \overset{H_0}{\sim} f_{2,49} \checkmark \qquad (3)$$

Pregunta 4 13,5 pt

Supuestos del modelo

Normalidad de los residuales 2 pt

Uno de los supuestos importantes en un modelo de regresión es que los ~~residuos~~ errores siguen una distribución normal. Para validar este supuesto, se puede utilizar la prueba de hipótesis de Shapiro-Wilk, que se basa en la hipótesis nula de que los residuos siguen una distribución normal. En caso de que la prueba de hipótesis indique que no se puede rechazar la hipótesis nula, se puede concluir que los residuos tienen una distribución normal, y se podra observar a continuación:

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

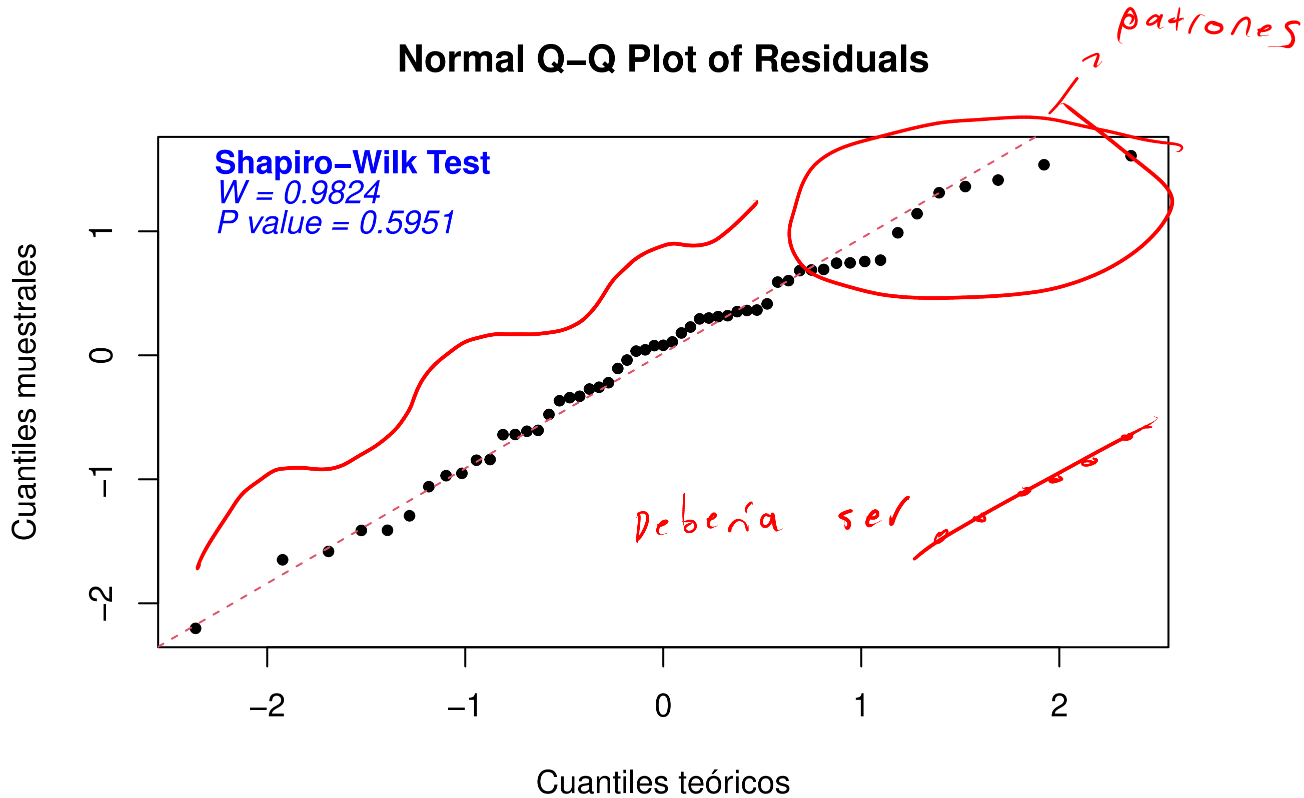


Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Después de realizar la prueba de hipótesis de Shapiro-Wilk, se obtuvo un P-valor de aproximadamente 0.5951. Al comparar este valor con el nivel de significancia establecido de $\alpha = 0.05$, se encontró que el P-valor es significativamente mayor, lo que indica que no se puede rechazar la hipótesis nula de que los datos se distribuyen normalmente con media μ y varianza σ^2, esto es respaldado por el resultado del valor estadístico de prueba del test Shapiro-Wilk, ya que cuanto más cercano sea este valor a 1, más se asemejarán los datos a una distribución normal,también mediante la gráfica podemos observar como el patrón de los residuales sigue la linea de ajuste.Por lo tanto, se procederá a validar el supuesto de homogeneidad de varianzas.

↳ hace falta un análisis más exhaustivo, recuerden que es más importante el criterio gráfico

Varianza constante 3 pt

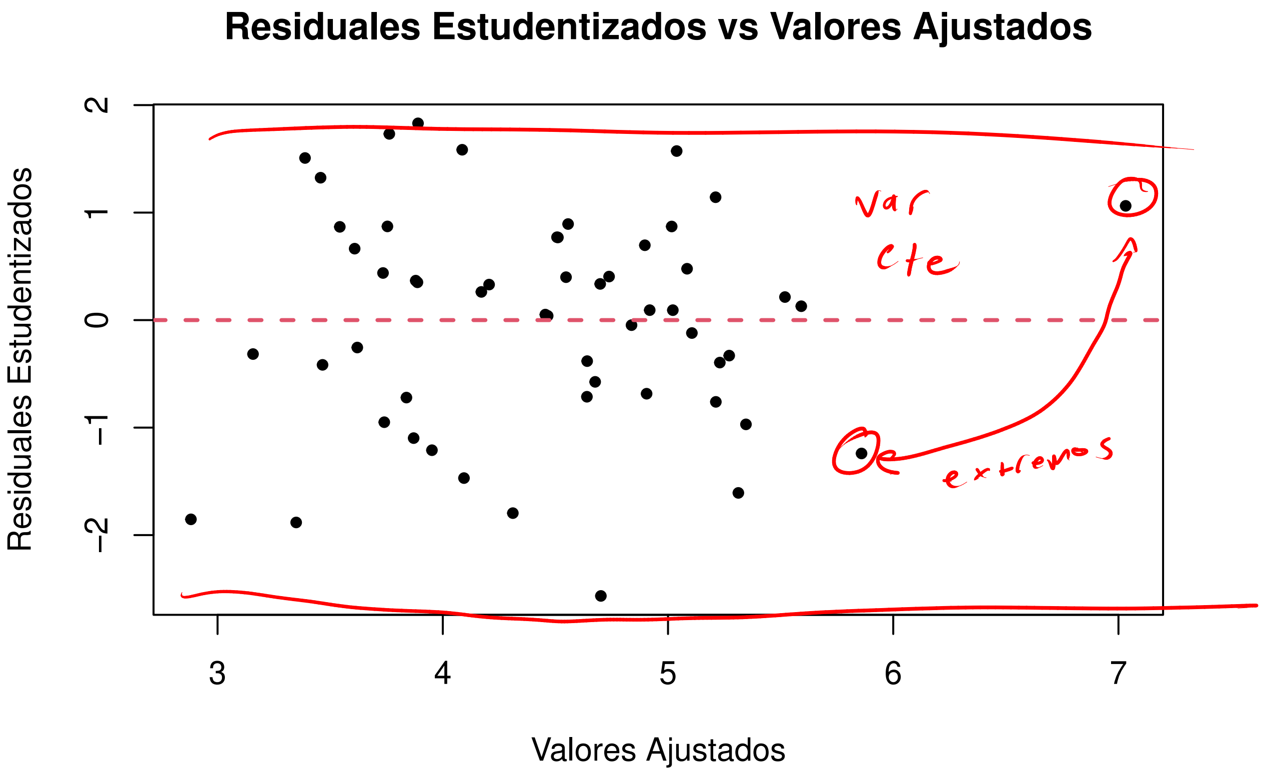


Figura 2: Gráfico residuales estudentizados vs valores ajustados

En el presente caso, al analizar el gráfico de residuales estudentizados vs valores ajustados, se puede observar que no hay patrones que sugieran un cambio en la varianza. Los puntos se distribuyen aleatoriamente alrededor de la línea horizontal en cero, lo que indica que la varianza ~~de los residuos~~ es constante y homocedástica en toda la gama de valores ajustados.

constante y homocedástica es lo mismo

Además, se observa que la media de los residuos estudentizados es cero, lo cual es un requisito importante para poder confiar en la validez de los intervalos de confianza y las pruebas de hipótesis basadas en los residuos. Una media de cero en los residuos indica que el modelo está capturando adecuadamente ~~la media~~ de la variable respuesta.

¿cuál es ese grado?

En conclusión, se puede afirmar con un alto grado de confianza que el supuesto de constancia de la varianza es válido en el modelo ajustado.

los residuales estudentizados siempre tienen media 0

Verificación de las observaciones

3 pt

Datos atípicos

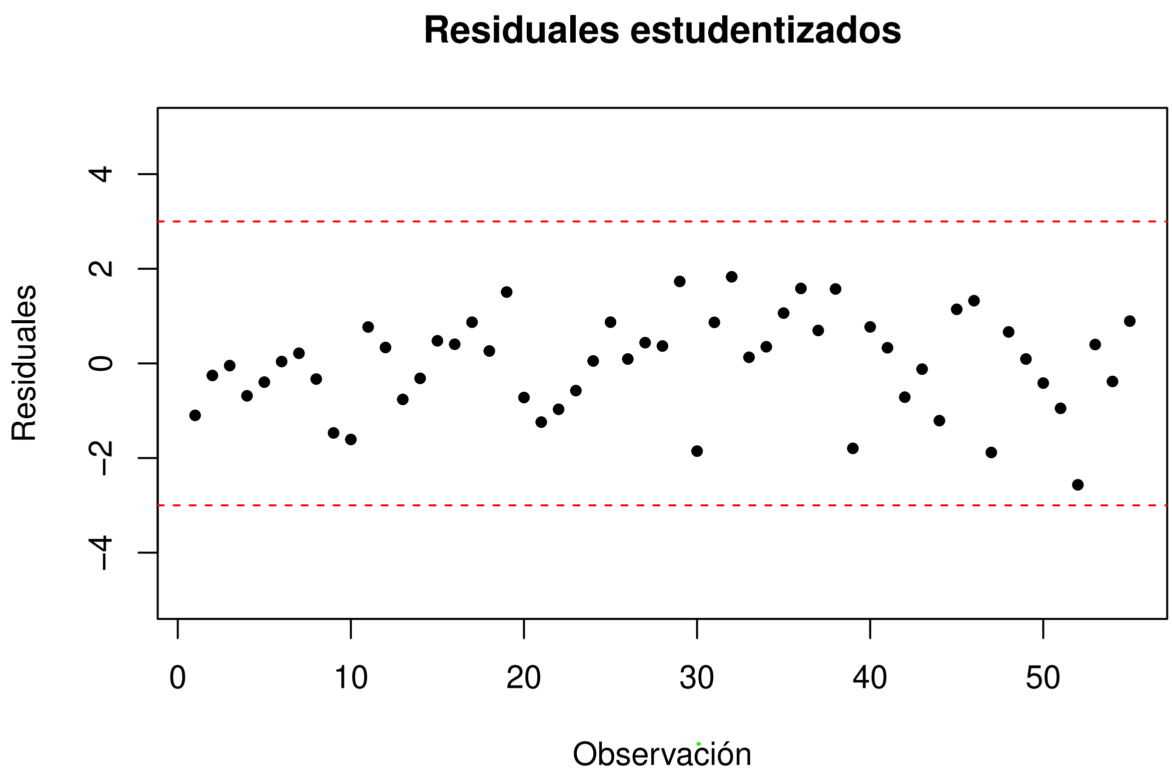


Figura 3: Identificación de datos atípicos

A partir de la gráfica presentada, se puede concluir que no hay valores extremos o inusuales en el conjunto de datos, ya que ninguno de los residuos estandarizados supera el criterio establecido de $|r_{estud}| > 3$. ✓

Puntos de balanceo 1,5 pt

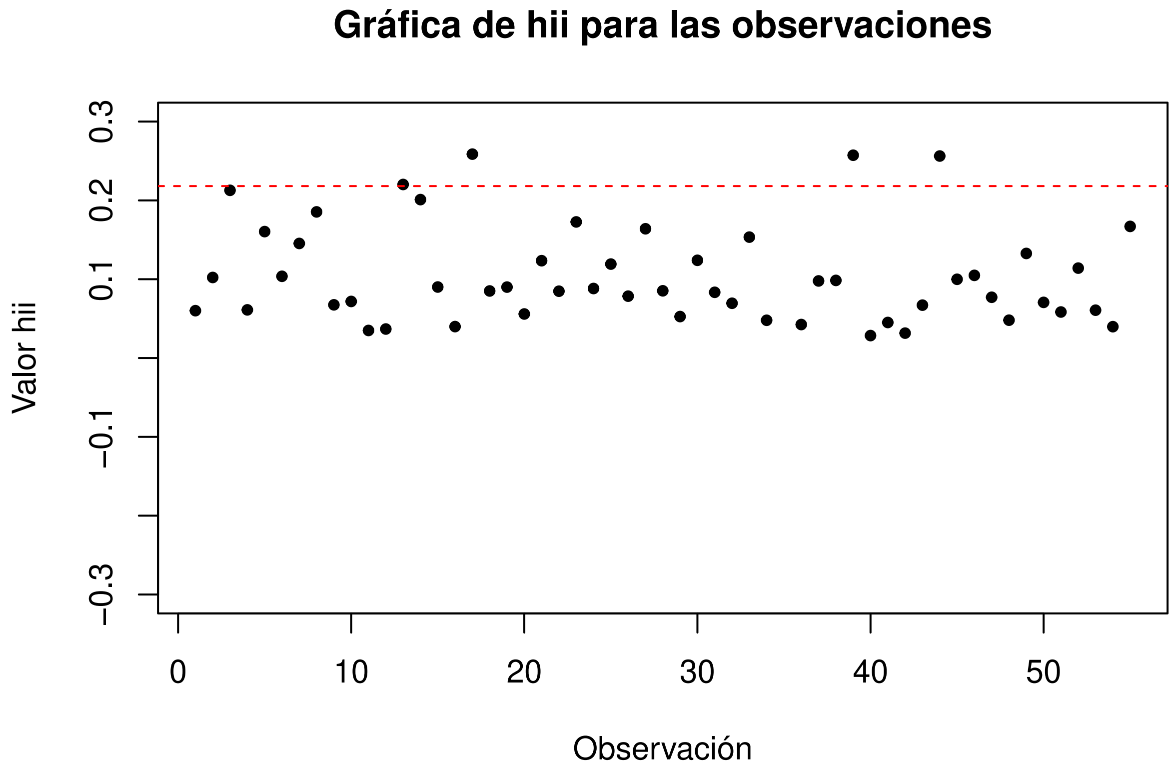
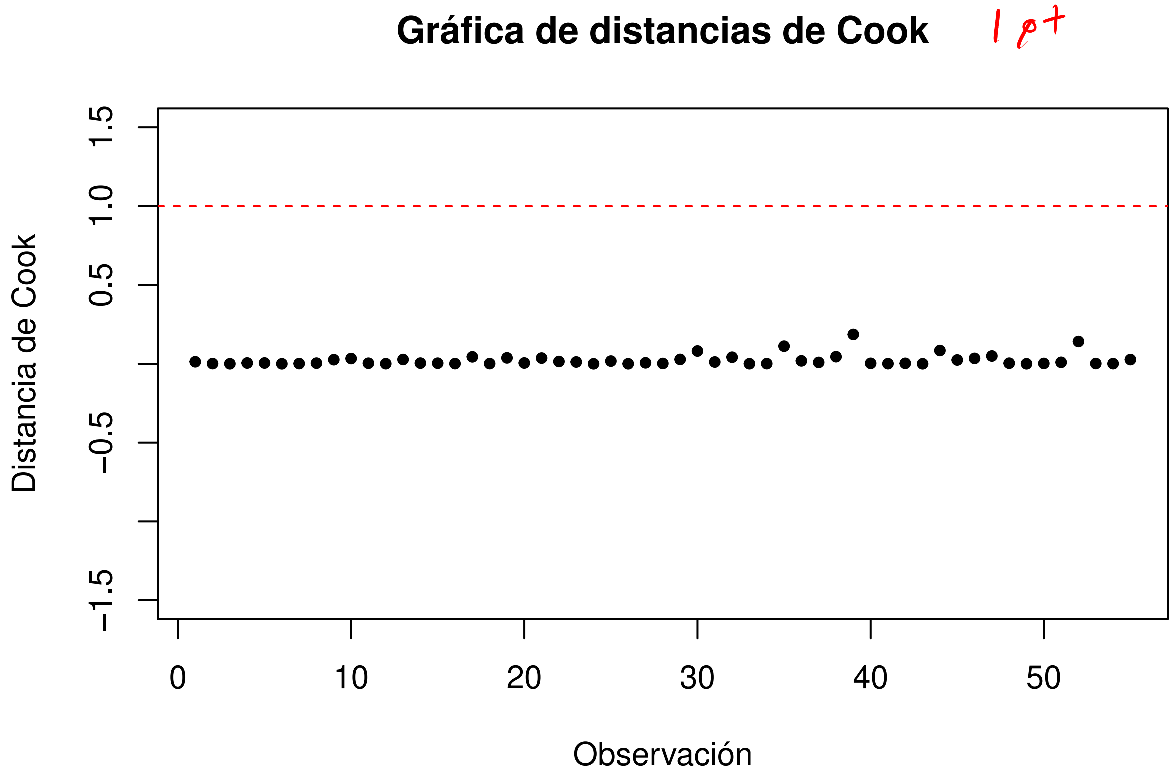


Figura 4: Identificación de puntos de balanceo

Al analizar la gráfica de observaciones vs valores h_{ii}, se observa que la línea punteada roja, que representa el valor $h_{ii} = 0.2181818$, indica el umbral para identificar puntos de balanceo en los datos. Se identificaron 5 puntos que superan este umbral, lo que significa que son puntos críticos para la regresión. Los puntos críticos se presentan en la tabla correspondiente.

¿Qué índice de dato tienen estos puntos de balanceo? ¿cuáles son? ¿Qué causan en el modelo?

Puntos influenciales



1 pt

Figura 5: Criterio distancias de Cook para puntos influenciales

Observamos que no hay ningún punto que se pueda considerar influencial en el modelo de regresión, puesto que ninguno es mayor a 1.

según este criterio

ninguna distancia de Cook

1, 5 pt

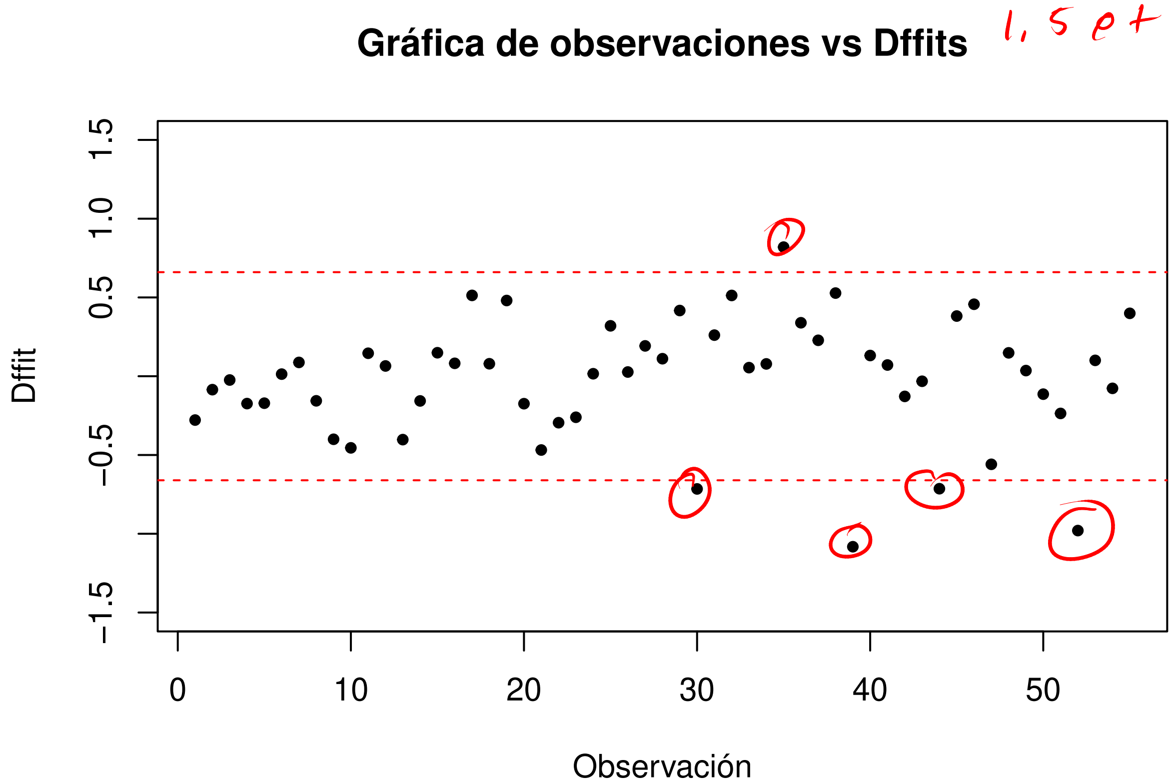


Figura 6: Criterio Dffits para puntos influenciales

Cuadro 6: Hii value

	Valor del parámetro
‘30‘	0.1240
‘35‘	0.3722
‘39‘	0.2573
‘44‘	0.2563
‘52‘	0.1141

Muy bien por hacer la tabla :3

estimación de los $\hat{y}_i$, para los parámetros es D.cook.

La tabla de diagnóstico muestra que hay algunas observaciones que son puntos influenciales según el criterio de Dffits. Este criterio se utiliza para detectar puntos que tienen un efecto ~~importante en la estimación de los parámetros del modelo~~, y se define como cualquier punto cuyo $|D_{ffit}| > 2\sqrt{0.2181818}$, siendo p el número de parámetros del modelo y n el número de observaciones. En este caso, hay algunas observaciones que superan este umbral, lo que indica que tienen un impacto significativo en el modelo. Sin embargo, cabe destacar que según el criterio de distancias de Cook, ninguna de las observaciones es un punto influencial. Este criterio mide la distancia que se movería la estimación de los parámetros si se eliminara una observación en particular, y se considera que cualquier punto cuya $D_i > 1$ es un punto influencial. En resumen, aunque hay algunas observaciones que son puntos influenciales según

¿No estaban hablando de Dffits para los parámetros?

el criterio de Dffits, no hay ningún punto que cumpla con el criterio de distancias de Cook para ser considerado influencial. → No tienen por qué cumplir ambas para ser influencial.

Conclusión 1,5 pt

Se realizó un análisis de la eficacia en el control de infecciones hospitalarias utilizando un modelo estadístico. Para validar la aplicabilidad de este modelo, se realizaron pruebas de supuestos y se evaluó si se cumplían las condiciones necesarias para la validez de los resultados.

dijeron que sí

En primer lugar, se llevó a cabo el test de Shapiro-Wilk para evaluar la normalidad de los datos. Se encontró que los datos no presentan una distribución normal significativa, ya que no se rechazó la hipótesis nula H_0 de normalidad en la prueba. Por lo tanto, se asume que los errores del modelo ~~son distribuidos normalmente~~ con media 0 y varianza constante. → esto no lo prueban con normalidad

→ No que no?

En segundo lugar, se examinó la ~~linealidad~~ del modelo, se observó ~~que los residuos se ajustan adecuadamente a la linealidad del modelo~~. Además, se evaluó la homogeneidad de la varianza de los residuos mediante el gráfico de residuales estudentizados vs valores ajustados, la cual no rechazó la hipótesis nula de homogeneidad de varianzas, indicando que la varianza de los residuos es constante en todo el rango de valores de la variable respuesta. → ¿?

errores

En tercer lugar, se verificó que los residuos estandarizados siguieran una distribución normal y se evaluó su media, se encontró que la media de los residuos estandarizados es aproximadamente 0. Este resultado indica que el modelo captura adecuadamente la media de la variable respuesta.

siempre lo es, se debe revisar los residuales con cludos

Finalmente, se confirmó la validez del modelo al cumplir con los supuestos necesarios para la inferencia estadística. Los residuos del modelo son identicamente independientes, tienen una distribución normal con media 0 y varianza constante, y ~~se ajustan a la linealidad del modelo.~~ Por lo tanto, se pueden realizar pruebas de hipótesis y construir intervalos de confianza con validez estadística.