3,95

Trabajo Corto 01 - Equipo 30

Estudiantes

Santiago Poveda Grajales
Juan David Calle Gallego
Jesús Andrés Álvarez Alvarado
Daniela Toro Arteaga

Docente

Francisco Javier Rodríguez Cortés

Asignatura

Estadística II

Sede Medellín
30 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

Teniendo en cuenta la base de datos brindada y la descripción del problema, en donde se busca estimar el riesgo de infección en un hospital ~~en~~ con base a 5 variables regresoras, dadas por:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

En donde: Y = Riesgo Infección, X_1 = Duración de la estadía , X_2 = Rutina de cultivos, X_3 = Número de camas, X_4 = Censo promedio diario y X_5 = Número de enfermeras. Por su parte ε_i es el término del error, que distribuye normal con media cero y varianza constante.

1.1. Estimación Modelo de regresión

Al ajustar el modelo anterior, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	0.2449
β_1	0.1556
β_2	0.0073
β_3	0.0615
β_4	0.0106
β_5	0.0011

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i}$$

Reemplazando los coeficientes encontrados se tiene:

$$\hat{Y}_i = 0.2449 + 0.1556 X_{1i} + 0.0073 X_{2i} + 0.0615 X_{3i} + 0.0106 X_{4i} + 0.0106 X_{5i}$$

1.2. Significancia de los parámetros individuales

Interpretación de β_0:

Para interpretar el intercepto es necesario estudiar el rango de las variables:

Cuadro 2: Tabla rangos de las Variables

	Y	X_1	X_2	X_3	X_4	X_5
Mínimo	1.3	6.70	42.0	1.6	39.6	29
Máximo	7.8	19.56	65.9	52.4	133.5	833

Como ninguna de las coordenadas X_1, X_2, ..., $X_5 = (0, 0, 0, 0, 0)$, es decir que en el rango de estas no está contenido el cero, entonces el intercepto no es interpretable.

En la siguiente tabla se puede observar la información de los parámetros estimados, mediante la cual se analiza la significancia de los mismos:

Cuadro 3: Tabla resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	0.2449	1.4600	0.1677	0.8674
β_1	0.1556	0.0702	2.2155	0.0306
β_2	0.0073	0.0269	0.2692	0.7887
β_3	0.0615	0.0139	4.4406	0.0000
β_4	0.0106	0.0068	1.5488	0.1268
β_5	0.0011	0.0007	1.6590	0.1024

Los P-valores permiten concluir, con un nivel de significancia $\alpha = 0.05$, que los parámetros β_1 y β_3 son significativos, pues sus P-valores son menores a 0.05. Por otra parte, con β_2, β_4 y β_5 no se pueden hacer interpretaciones, pues no son estadisticamente diferentes de 0.

1.3. Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,59} \qquad (1)$$

Y considerando la tabla Anova:

Cuadro 4: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	58.258	5	11.651605	13.722	6.89619e-09
Error	50.098	59	0.849118		

Se puede observar que la prueba F tiene un valor p aproximadamente igual a 0, por lo que se rechaza la hipótesis nula, y se concluye que al menos una de las variables predictoras es diferente de 0, de esta manera se puede afirmar que el modelo de regresión en su conjunto es significativo.

1.4. Interpretación de los parámetros significativos

Dado que solo β_1 y β_3 son significativos, se procede con la interpretación de las variables relacionadas a estos:

$X1$: Mide la duración promedio, en días, de la estadía de todos los pacientes en el hospital, la estimación que se obtuvo de este regresor fue $\beta = 0.1556$, lo que quiere decir que por cada día más de estadía en el hospital, la probabilidad promedio de infectarse aumenta en aproximadamente 15.6 %, manteniendo los demás regresores fijos.

$X3$: Mide el numero promedio de camas en el hospital durante el periodo de estudio, la estimación que se obtuvo de este regresor fue $\beta = 0.0615$, lo que quiere decir que por cada unidad de cama adicional en el hospital aumenta la probabilidad de infectarse en aproximadamente 6.2 %, manteniendo los demás regresores fijos.

1.5. Coeficiente de determinación múltiple R^2

R^2: 0.5376539

R^2 *Ajustado*: 0.498472

El modelo tiene un coeficiente de determinación múltiple de $R^2 = 0.5377$, lo que significa que aproximadamente el 53.77 % de la variabilidad total observada en la variable respuesta es explicada por el modelo de regresión propuesto, pero como se está trabajando en una regresión lineal múltiple lo que interesa es el R^2 *ajustado* que fue de 0.4985, en este caso, baja con respecto al R^2 normal, ya que está castigando la inserción de variables no explicativas: X_2,X_4 y X_5. En este caso la regresión explica el 49.9 % de la variabilidad del porcentaje promedio de infecciones en el hospital.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más alto en el modelo fueron X_2, 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:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de 0 para } j = 2, 4, 5 \end{cases}$$

Cuadro 5: Tabla resumen de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	50.098	X1 X2 X3 X4 X5
Modelo reducido	54.517	X1 X3

Para este caso el R^2 disminuye 0.309 con respecto al R^2 del modelo original (0.5377) y el R^2 *Ajustado* cae a 0.275 con respecto al del modelo original (0.4985). Según la teoría, como se excluyeron las variables significativas que eran X_1 y X_3, el R^2 *Ajustado* debería caer y así lo es.

Entonces como R^2 *Ajustado* no aumentó si no que disminuyo, se comprueba aún mas que las regresoras X_2 , X_4 *y* X_5 no son significativas.

Luego un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,59} \\ F_0 &= \frac{54.517 - 50.098}{\frac{54.517}{59}} \\ F_0 &= 4.78238 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,59} = 2.7608$, se puede ver que $F_0 > f_{0.95,1,59}$. Entonces, se rechaza la hipotesis nula y por lo tanto el modelo es significativo y no se puede descartar.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hace la pregunta si los regresores X_2, X_4 y X_5 son o no significativos, por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_2 = \beta_4 = \beta_5 = 0 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

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

El modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \ 1 \leqslant i \leqslant 65$$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)} \overset{H_0}{\sim} f_{2,59} \tag{3}$$

Se espera que H_0 no se rechace, ya que los regresores escogidos para esta pregunta son los no significativos del modelo, identificados anteriormente.

4. Pregunta 4

12 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

1 pt

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis Shapiro-wilk, acompañada de un gráfico Cuantil-Cuantil:

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

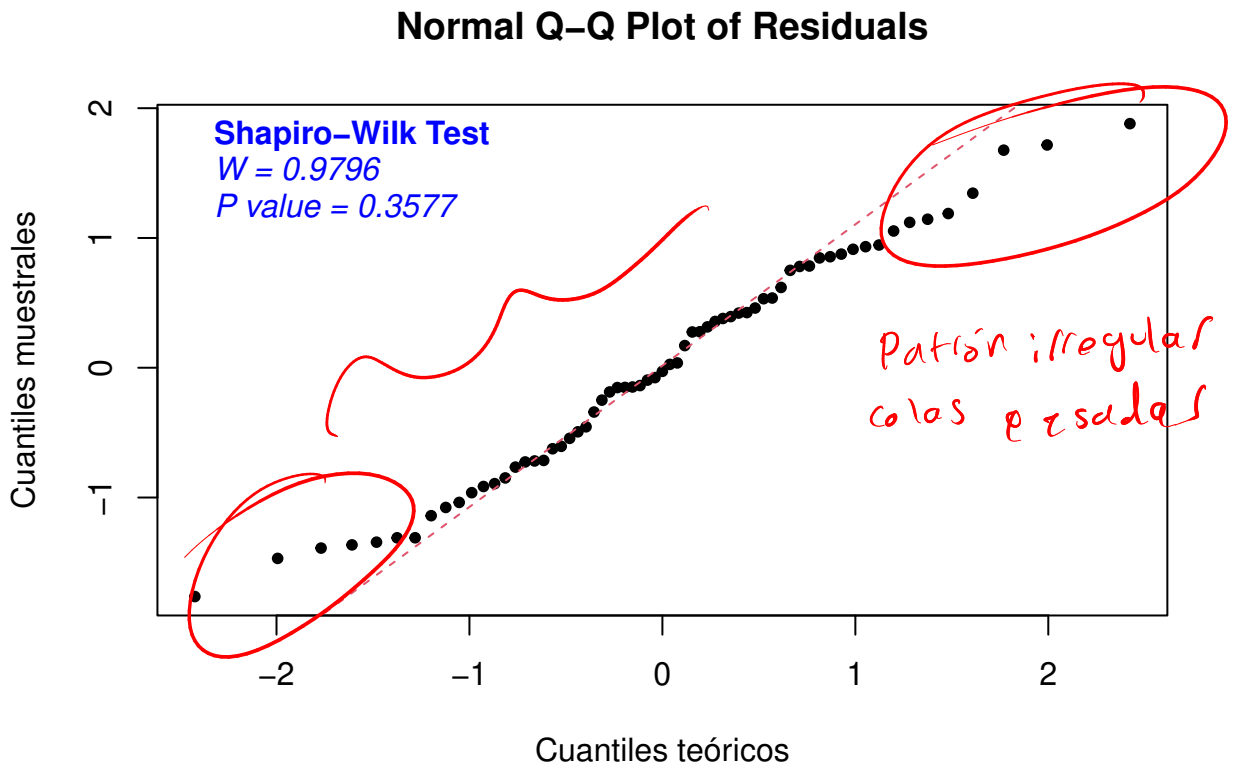

Figura 1: Gráfico Cuantil-Cuantil y Normalidad de Residuales

Teniendo el P-valor aproximadamente igual a 0.3577 y sabiendo que el nivel de significancia $\alpha = 0.05$, no se rechaza la hipótesis nula, es decir que los datos distribuyen normal con media μ.

→ No se está probando que tengan media cte. No evalúan gráfica, la cual es más importante, No distribuyen normal

4.1.2. Varianza constante

Para comprobar la varianza constante se utiliza la grafica de residuales estudentizados vs valores ajustados.

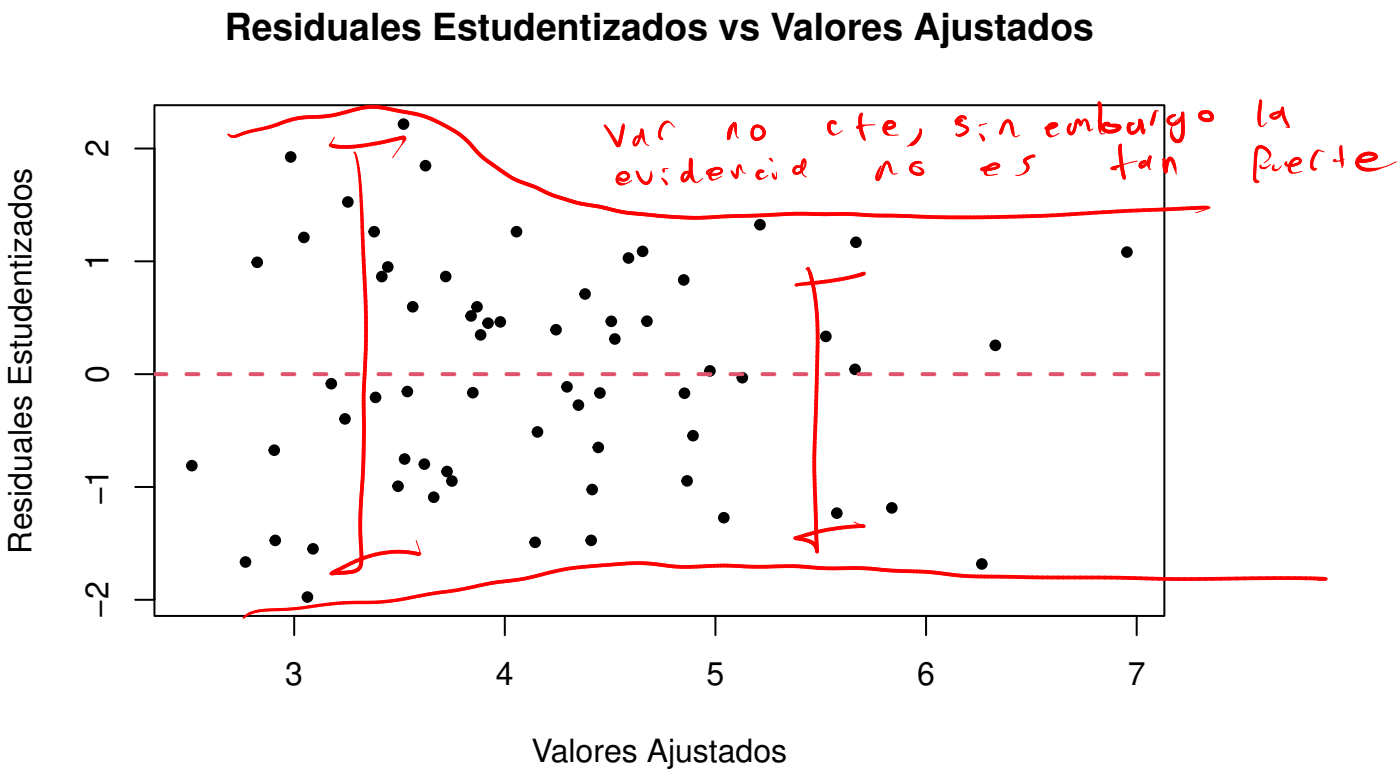

Figura 2: Gráfico Residuales Estudentizados vs Valores Ajustados

Como en el gráfico no se ven formaciones de conos o figuras que nos indiquen varianza no constante, se podría inferir que el modelo cumple con el supuesto de varianza constante, al no haber evidencia suficiente en contra de este supuesto se acepta como cierto. Además es posible observar media 0.

Media Residuales Estudentizados: -0.0314

Se puede ver cómo la media de los errores es aproximadamente cero.

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

Analizando los valores de los residuales estudentizados los cuales se estandarizan, se dice que valores mayores a 3 o menores que -3 son considerados valores atípicos.

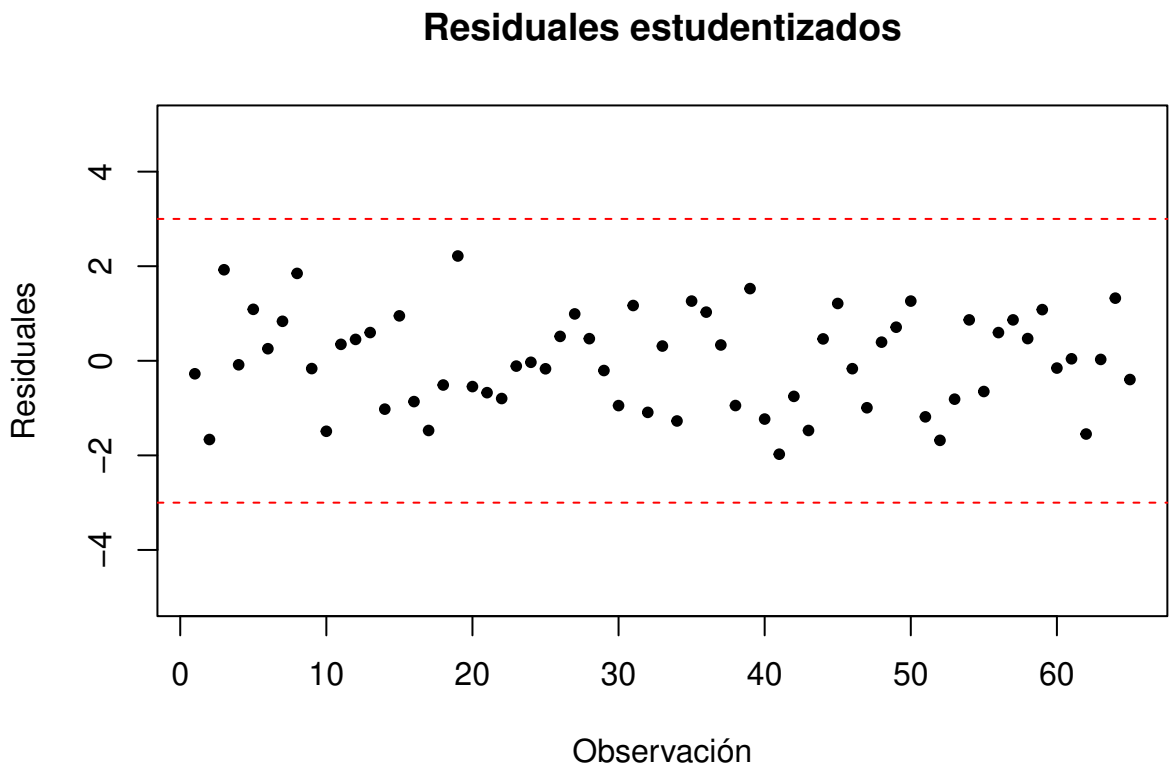

Figura 3: Gráfico Identificación de datos atípicos

Para el caso de este modelo no se encontró ningún residuo que sobrepasara estos valores, pues ningún residual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$, por lo tanto se concluye que no existen valores atípicos en el modelo.

4.2.2. Puntos de balanceo

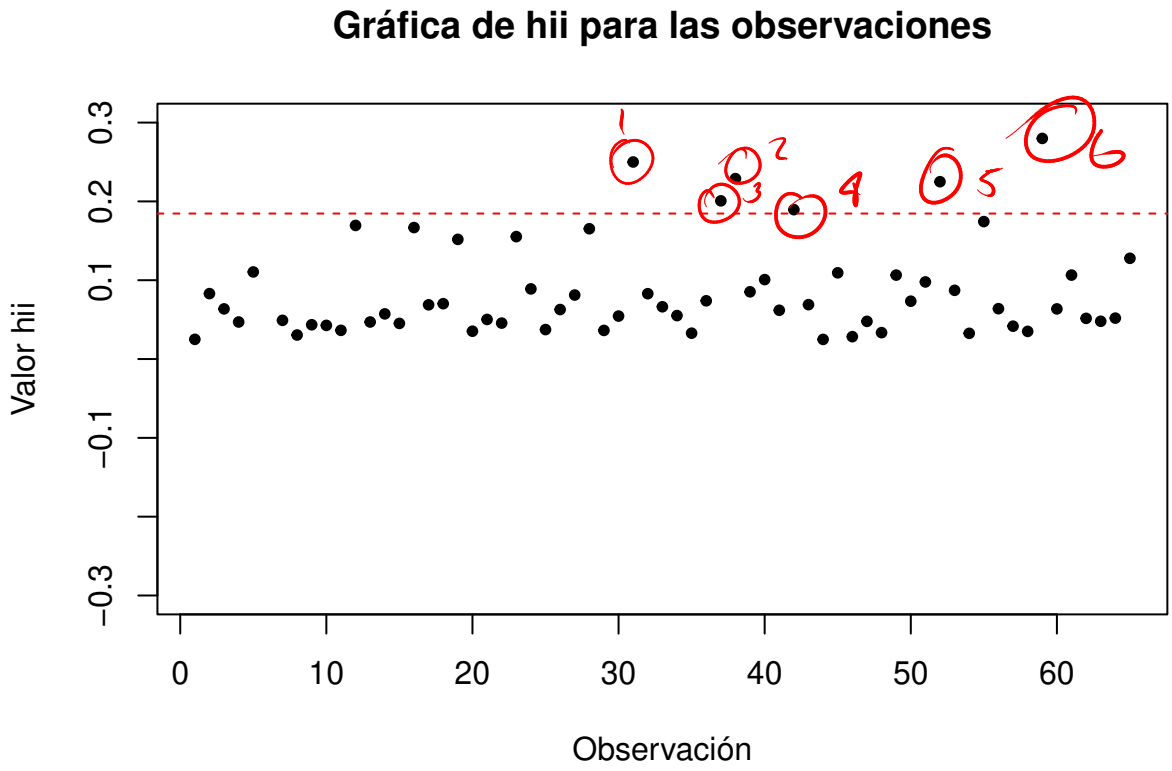

Figura 4: Gráfico Identificación de puntos de balanceo

Cuadro 6: Tabla diagnostico hii puntos de balanceo

	Res.stud	Cooks.D	hii.value	Dffits
6	0.2552	0.0097	0.4718	0.2392
31	1.1688	0.0759	0.2500	0.6771
37	0.3345	0.0047	0.2007	0.1664
38	-0.9460	0.0443	0.2289	-0.5149
42	-0.7524	0.0221	0.1895	-0.3625
52	-1.6824	0.1369	0.2250	-0.9211
59	1.0821	0.0759	0.2799	0.6757

En la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n}$, se puede apreciar que existen 7 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 2\frac{6}{65} \; -> \; h_{ii} > 0.1846154$, los cuales son los presentados en la tabla.

4.2.3. Puntos influenciales

Para detectar los valores influenciables se utiliza el DFFITS que es una manera de medir cuantas desviaciones estándar se mueve el valor ajustado si se elimina la observación i, para este caso todo valor $DFFITS > 0.60764$ será considerado una observación capaz de influir en el ajuste del modelo.

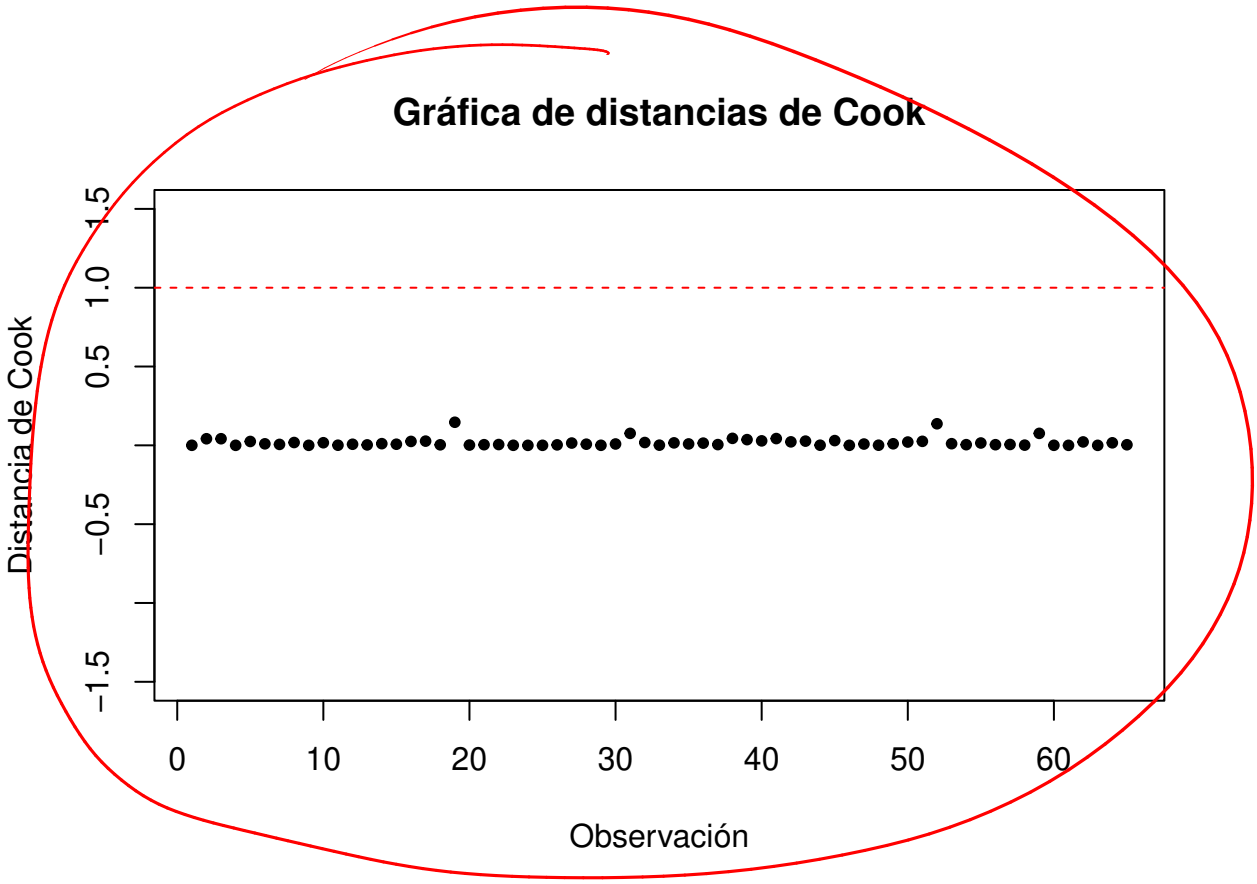

Figura 5: Gráfico distancias de Cook para puntos influenciales

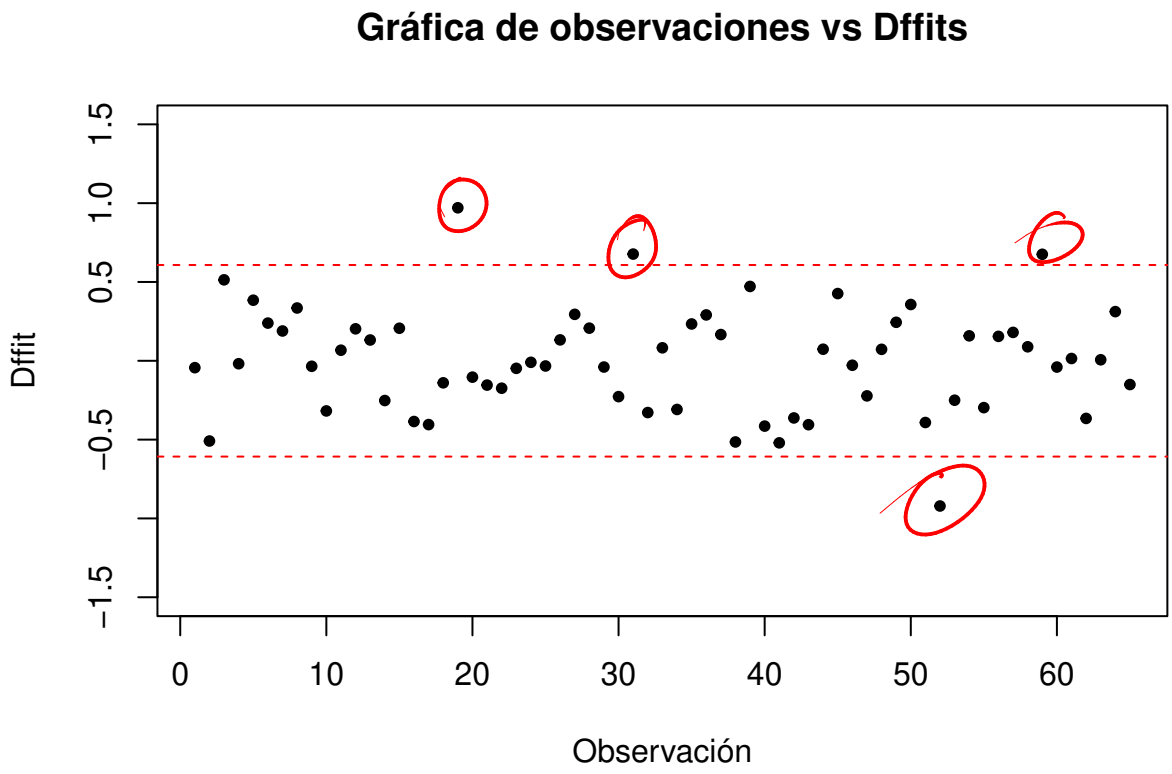

Figura 6: Gráfica Dffits para puntos influenciales

Cuadro 7: Tabla Criterio Dffits para puntos influenciales

	Res.stud	Cooks.D	hii.value	Dffits
19	2.2161	0.1465	0.1518	0.9707
31	1.1688	0.0759	0.2500	0.6771
52	-1.6824	0.1369	0.2250	-0.9211
59	1.0821	0.0759	0.2799	0.6757

Como se puede ver, las observaciones 19,31,52 y 59 son puntos influenciales según el criterio de Dffits, $|D_{ffit}| > 2\sqrt{\frac{p}{n}}$, los cuales pueden cambiar el valor ajustado considerablemente.

4.3. ¿Qué puede decir acerca de la validez del modelo?

Como se observó en el inicio, el modelo en su conjunto con la prueba F es significativo, se encontró que los regresores X_1 y X_3 son significativos, se podría considerar eliminar los

regresores X_2 , X_4 *y* X_5 pero se debe tener cuidado ya que a veces existen regresores que no tienen significancia estadística pero tienen importancia teórica para el modelo lo cual podría dañar su ajuste.

Se puede observar que el modelo cumple los supuestos de normalidad en los errores, media cero de los errores y varianza constante, lo cual garantiza que las estimaciones del modelo sean precisas y confiables. → Acá deberían decir que es válido.

Analizando los valores atípicos, puntos de balanceo y valores influenciables , se obtiene que el modelo no tiene valores atípicos y la única observación que podría cambiar un poco el ajuste del modelo es la observación #59 que fue detectada como punto de balanceo y valor influenciable, por lo tanto se podría considerar eliminar esta observación para mejorar el ajuste del modelo.

Haciendo los análisis anteriores se podría considerar que el modelo en términos generales es confiable para predecir la probabilidad promedio estimada de infección de una persona en un hospital de EE. UU, teniendo en cuenta el número de días de su estadía y el número de camas promedio que hallan en el hospital. → ¿No sólo teniendo en cuenta esas

Nota: es importante aclarar que solo se podrá hacer predicciones según el rango de las variables predictoras ya que no es posible extrapolar con el método de regresión lineal.

→ Tienen 7 de balanceo y 9 influenciales y dún así solo importa el dato 59 ?