Trabajo #1
Regresión Lineal Múltiple

3,3

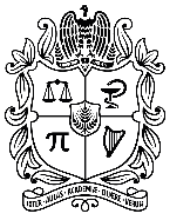
Estudiantes:

Neila Sirley Perilla Perdomo
Carlos Mario Cardona Moscote
Diego Andrés Gracia Granados
Roger Albeiro Meneses Carmona

Equipo 34

Docente:
Francisco Javier Rodríguez Cortez

Asignatura:
Estadística II



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
Marzo de 2023

INDICE

Contenido

PREGUNTA 1

Teniendo en cuenta la base de datos asignada, la cual es Equipo34.txt, las variables del riesgo de infección (Y).
El modelo propone lo siguiente:

$$Y_i = \beta 0 + \beta 1X1_i + \beta 2X2_i + \beta 3X3_i + \beta 4X4_i + \beta 5X5_i + \varepsilon i \quad , \quad \varepsilon i \text{ iid} \sim N(0, \sigma^2)$$

1.1 MODELO DE REGRESIÓN

Al ajustar el modelo se obtienen los siguientes parámetros ajustados:

Valor del parámetro	
$\hat{\beta}0$	-1.865137244
$\hat{\beta}1$	0.012422596
$\hat{\beta}2$	0.051767745
$\hat{\beta}3$	0.054182451
$\hat{\beta}4$	0.023579025
$\hat{\beta}5$	0.002540102

Tabla 1: Tabla de valores de los parámetros ajustados

Por lo tanto, el modelo de regresión ajustado es:
$\hat{Y}i$= -1.865137244 + 0.012422596Xi1 + 0.051767745Xi2+ 0.054182451Xi3 + 0.023579025Xi4 +0.002540102Xi5.

Donde *1 ≤ i ≤ 50*

1.2. SIGNIFICANCIA DE LA REGRESIÓN

Para analizar la significancia de la regresión, tomamos como hipótesis:

$$\begin{cases} H_0 : \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=0 \\ H_a : \text{Algún } \beta_i \neq 0, \quad \text{para i = 1,2,3,4,5} \end{cases}$$

Para la significancia de la regresión se hará uso de la siguiente tabla anova:

	Suma de cuadrados	***g.l***	***Cuadrado medio***	***F0***	***Valor-P***
Modelo de regresión	66.8814	5	13.376283	14.2052	2.80362e-08
Error	41.4324	44	0.941645		

Tabla 2: Tabla anova significancia de la regresión

Al observar los resultados de la Tabla Anova, la evidencia muestral nos dice que se rechaza la hipótesis nula, que sugeriría que los parámetros no son significativos. Por tanto, la prueba de anova muestra que los resultados de la regresión son significativos. es decir, al menos un parámetro es significativo.

Con una significancia del 0.05, podemos rechazar la hipótesis nula ya que el valor P es muy pequeño, y por lo tanto, nuestra regresión es significativa.

1.3. SIGNIFICANCIA DE PARÁMETROS

Ahora analicemos nuestros parámetros teniendo en cuenta el siguiente juego de hipótesis:

$$\begin{cases} H_0: \beta_i = 0 \\ H_a: \text{Algún } \beta_i \neq 0 \text{ para } i = 1, 2, 3, 4, 5 \end{cases}$$

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar la significancia individual de estos.

	Estimación βi	***se(β̂i)***	***T0i***	***Valor-P***
β0	-1.865137244	1.5800655227	-1.1804177	0.2441763236
β1	0.012422596	0.0976393102	0.1272295	0.8993384758
β2	0.051767745	0.0289253171	1.7897036	0.0803853399
β3	0.054182451	0.0143172039	3.7844297	0.0004623008
β4	0.023579025	0.0081762694	2.8838366	0.0060593376
β5	0.002540102	0.0008608907	2.9505516	0.0050676181

Tabla 3: Resumen de los coeficientes

con un nivel de significancia del 0.05, podemos decir que los parámetros β_0, β_1 y β_2 no son significativos. Por ende, los parámetros β_3, β_4 y β_5 si son significativos en presencia de los demás parámetros.

1.4. INTERPRETACIÓN DE PARÁMETROS

Ahora podemos hacer el siguiente análisis de cada variable:

- $\hat{\beta}_0$**=-1.865137244** *(Intercepto)*: como $X_i = 0 / \in [X_{i,min}, X_{i,max}] \forall i$ entonces este valor no es interpretable.
- $\hat{\beta}_1$= **0.012422596** **(***Duración de la estadía***):** El parámetro $\hat{\beta}_1$, no podemos interpretar nada, ya que no es significativo.
- $\hat{\beta}_2$**=0.051767745** *(Rutina de cultivos)***:** El parámetro$\hat{\beta}_2$**,** no podemos interpretar nada, ya que no es significativo.
- $\hat{\beta}_3$= **0.054182451***(Número de camas)***:** El parámetro $\hat{\beta}_3$**,** por cada unidad que aumente el número de camas, la probabilidad promedio de adquirir una infección aumenta 0.0542 unidades mientras las demás variables regresoras permanecen constantes.
- $\hat{\beta}_4$=**0.023579025** *(Censo promedio diario)***:** Por cada unidad que aumente el número de censos promedio diario, la probabilidad promedio de adquirir una infección aumenta 0.0236 unidades mientras las demás variables regresoras permanecen constantes.
- $\hat{\beta}_5$= **0.002540102** *(Número de enfermeras)***:** Por cada unidad que aumente el número de enfermeras , la probabilidad promedio de adquirir una infección aumenta 0.0025 unidades mientras las demás variables regresoras permanecen constantes.

1.5 COEFICIENTE DE DETERMINACIÓN MÚLTIPLE R^2

Para calcular el R^2 usamos la siguiente fórmula:

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

$$R^2 = \frac{66.8814}{66.8814 + 41.4324} = 0.6174781053$$

Ahora, al analizar el R^2 del modelo, nos da un valor de 0.6174 lo cual nos dice que el modelo explica un 61.74% de la variabilidad total, lo cual nos parece que no es un mal modelo.

PREGUNTA 2

2.1. PLANTEAMIENTO PRUEBA DE HIPÓTESIS Y MODELO REDUCIDO

Los valores P más grandes del modelo son los de β_0 con *0.2441763236* , β_1 con *0.8993384758* y β_2 con *0.0803853399*. Ahora queremos analizar si estas variables en conjunto se pueden descartar del modelo, por lo que haremos un estadístico de prueba F, al cual sometemos al siguiente juego de hipótesis:

$$\begin{cases} H_0: \beta_0 = \beta_1 = \beta_2 = 0 \\ H_a: \beta_i \neq 0, \quad para\ i = 0,1,2 \end{cases}$$

El modelo completo es el definido en la sección 1.1, y el modelo reducido es:

$$RM:\ Y = \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon i\ ,\ \varepsilon i\ iid \sim N(0, \sigma^2)$$

A partir de la tabla de todas las regresiones, se construye la siguiente tabla donde se evidencia la suma de cuadrados del error del modelo completo y reducido lo cual nos permitirá hacer cálculos posteriores.

	SSE	***Covariables en el modelo***
Modelo completo	41.432	*X1 X2 X3 X4 X5*
Modelo reducido	44.888	*X3 X4 X5*

Tabla 4: Resumen de la suma de cuadrados del error.

2.2. ESTADÍSTICO DE PRUEBA Y CONCLUSIONES

Para la construcción del estadístico de prueba es necesario calcular la suma de cuadrados extra de la siguiente forma:

$$SSextra = SSE(\beta 0, \beta 3, \beta 4, \beta 5) - SSE(\beta 0, \beta 1, \beta 2, \beta 3, \beta 4, \beta 5)$$

$$SSextra = 44{,}888 - 41{,}432 = 3{,}456$$

Se construye el estadístico de prueba como:

$$Fo = \frac{SSextra/gl\ SSextra}{MSE(FM)}\ H0\ \sim f2{,}44$$

$$Fo = \frac{3{,}456(2)}{0.941645} = 1.835$$

Ahora, teniendo en cuenta el criterio de rechazo para esta prueba, el c $F_0 > f_{0,05,2,44}$ y al calcular el percentil obtenemos un resultado de 3.20928 lo que indica que la hipótesis nula (Ho) no se rechaza, con esto podemos decir que al menos una de las variables X1: Duración de la estadía, X2: Rutina de cultivos, no afecta de forma significativa a la variable de respuesta Y.

PREGUNTA 3

3.1. PLANTEAMIENTO DE LA PREGUNTA

Verificar si el efecto que tiene X1: Duración de la estadía, sobre la variable respuesta Y: Riesgo de infección es el mismo que tiene X2: Rutina de cultivos, y a su vez, determinar si el efecto que tiene X3: Número de camas, sobre la variable respuesta Y es el mismo que tiene X4: Censo promedio diario.

3.2. PLANTEAMIENTO PRUEBA DE HIPÓTESIS Y MODELO REDUCIDO

Para solucionar dicha incógnita se plantean las siguientes hipótesis.

$$\begin{cases} H_0: \beta_1 - \beta_2 = 0\,,\ \beta_3 - \beta_4 = 0 \\ H_a: \beta_1 - \beta_2 \neq 0\,,\ \beta_3 - \beta_4 \neq 0 \end{cases}$$

Al reescribir dichas hipótesis de forma matricial obtenemos lo siguiente

$$H_0: \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad H_a: \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

donde **L** es la matriz $\begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix}$ de constantes de *MxP*, donde *M=2* y *P=6*

Partiendo de las igualdades en las hipótesis, el modelo reducido para esta prueba de hipótesis es el siguiente

RM: $Y = \beta_0 + \beta_1 (X_1 + X_2) + \beta_3 (X_3 + X_4) + \beta_5 X_5 + \varepsilon i$, εi *iid* $\sim N(0, \sigma^2)$

3.3. ESTADÍSTICO DE PRUEBA Y REGIÓN DE RECHAZO

Para plantear el estadístico de prueba es necesario conocer la suma de cuadrados debido a la hipótesis (SSH), el cual calculamos de la siguiente manera.

$$SSH = SSE(RM) - SSE(FM)$$

Con esto planteamos el siguiente estadístico de prueba *F0*

$$Fo = \frac{(SSE(MR) - SSE(MF))/gl\ SSH}{MSE(FM)} = \frac{MSH}{MSE(FM)} \quad H0 \sim f_{m,n-p}$$

$$Fo = \frac{(SSE(MR) - 41{,}432)/2}{0.941645} \quad H0 \sim f_{2,44}$$

Donde se rechaza la hipótesis nula si se cumple la siguiente condición $F_0 > f_{\propto,2,44}$

PREGUNTA 4

Para identificar si en el modelo hay observaciones extremas, se deben calcular los estadísticos que nos permiten aplicar criterios en ese sentido, los cuales incluyen: *la validación de los supuestos* (la validación de los supuestos de normalidad "Normal Q - Q plot", la validación de supuestos con media 0 y varianza constante "residuales estudentizados vs valores ajustado") y *las observaciones extremas* (puntos atípicos "residuales vs observación", puntos de balanceo "los valores de la diagonal de la matriz H (los hii)" y puntos influenciales "la distancia de Cook (Di) y los DFFITS").

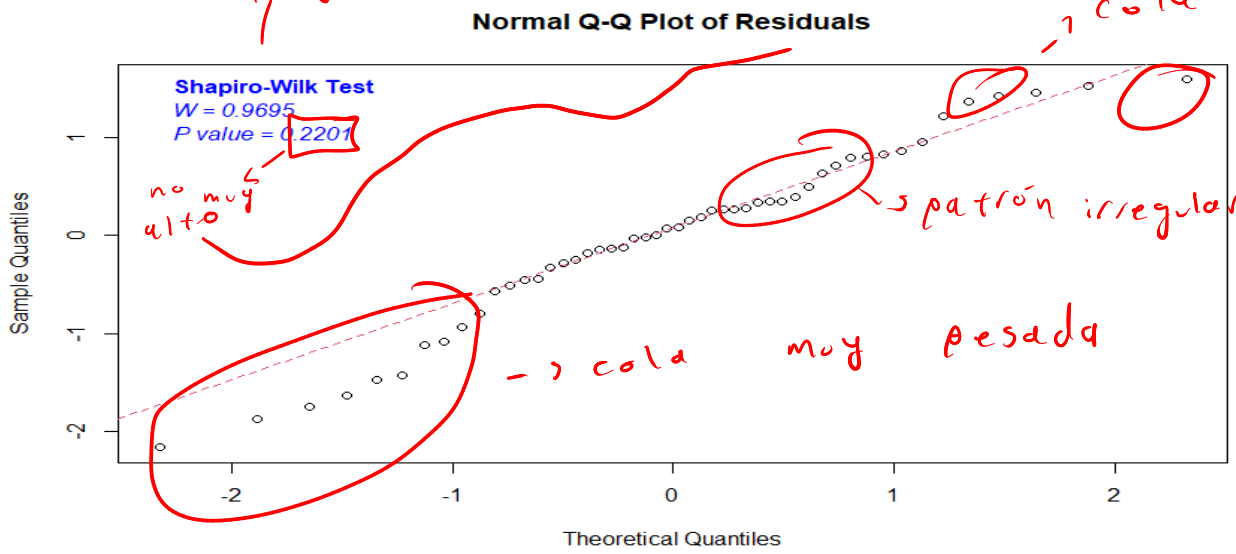
4.1 VALIDACIÓN DE LOS SUPUESTOS

4.1.1 VALIDACIÓN SUPUESTO DE NORMALIDAD

Para validar el supuesto de normalidad se proponen las siguientes hipótesis.

$$H0: \ \varepsilon i \sim Normal$$
$$Ha: \ \varepsilon i \ No \sim Normal$$

Realizamos la prueba de Shapiro Wilk y obtenemos la siguiente gráfica.

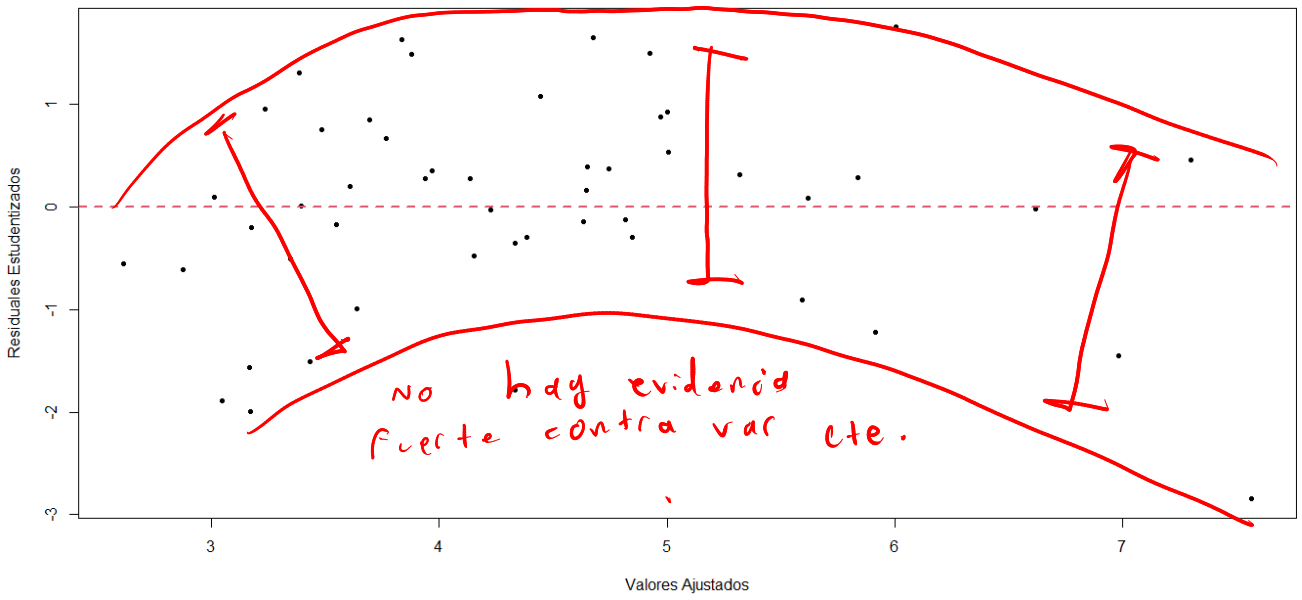


Gráfica 1: Normal Q - Q plot

A partir de esta gráfica podemos concluir que, como el valor P de dicha prueba es mayor a nivel de significancia α no rechazamos la hipótesis nula, es decir, el supuesto de normalidad no se cumple bajo esta prueba. Podemos ver en la ***Gráfica 1:*** *Normal Q-Q plot* que el patrón de los residuales no sigue la recta del ajuste de la distribución de los residuales a una distribución normal, esto puede ser una consecuencia de la presencia de observaciones influenciales en los datos; debido a esto el supuesto de normalidad no se cumple ya que no hay un buen ajuste.

4.1.2 VALIDACIÓN SUPUESTO MEDIA 0 Y VARIANZA CONSTANTE

De acuerdo con la siguiente gráfica.

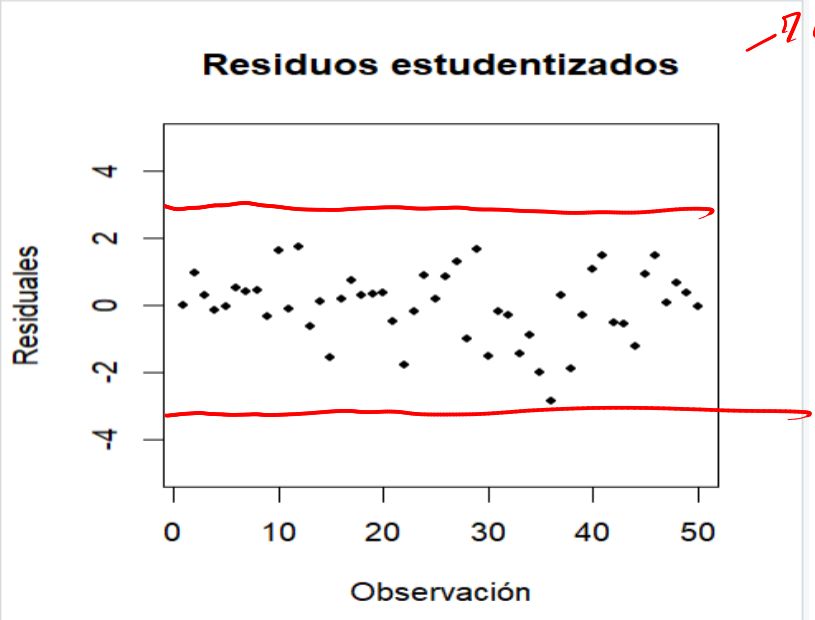


Gráfica 2: Residuales estudentizados vs valores ajustados

Analizando la gráfica de *residuales estudentizados vs valores ajustados*, se puede notar que no es clara una dispersión de datos homogénea, se puede concluir que la varianza de los errores no es constante y la media no tiende a 0. Es posible que algunas observaciones extremas estén afectando este análisis.

4.2. OBSERVACIONES EXTREMAS

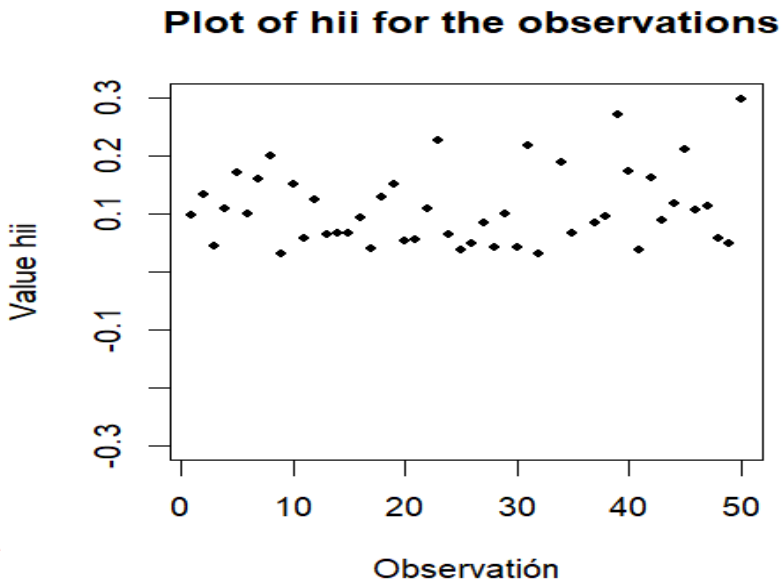
4.2.1. PUNTOS ATÍPICOS



Gráfica 3: Identificación de datos atípicos

Según la ***Gráfica 3:*** *Identificación de datos atípicos*, no hay datos atípicos bajo el criterio de *|restud| > 3*

4.2.2. PUNTOS DE BALANCEO



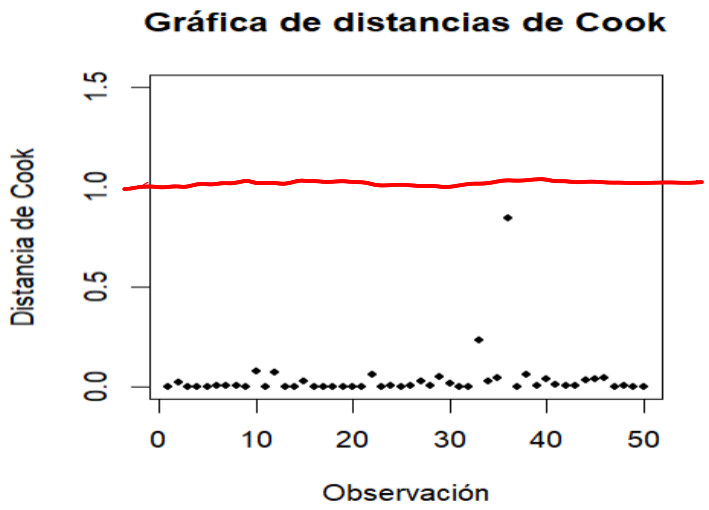
Gráfica 4: Identificación de puntos de balanceo

Analizando los elementos de la diagonal principal de la matriz Hat vs las observaciones obtenemos el gráfico anterior. A partir de este gráfico y a su vez del criterio de punto de balanceo que propone $Hii > 2\frac{p}{n}$, donde *n=50 y p=7*, podemos concluir que el modelo tiene 4 puntos de balanceo. Estos puntos de balanceo controlan ciertas propiedades del modelo, como el R^2 y los errores estándar de los coeficientes estimados. Es decir, estos puntos causan una sobreestimación en el R^2 y pueden afectar el supuesto de varianza, media y normalidad.

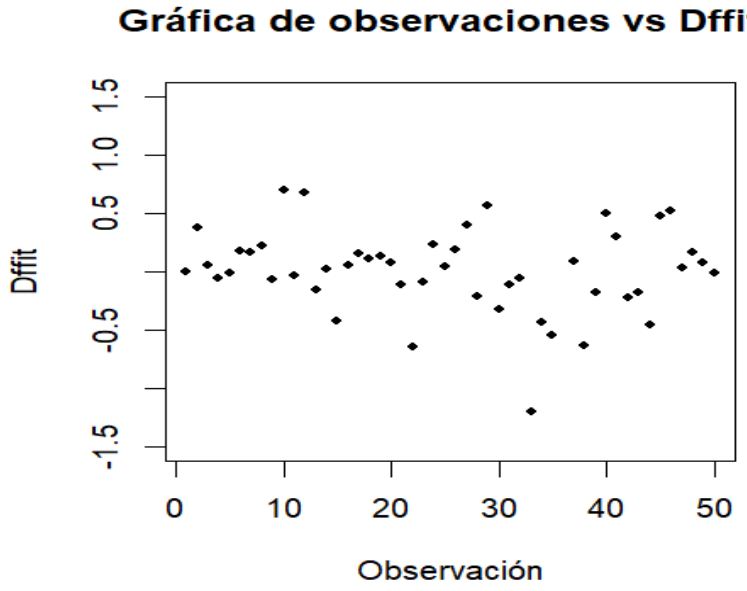
	Errores Estudentizados	***D.Cook***	***Valor hii***	***DFFITS***
33	-1.4446	0.2349	0.4031	-1.2024
36	-2.8462	0.8445	0.3848	-2.4636
39	-0.2983	0.0055	0.2710	-0.1800
50	-0.0209	0.0000	0.2976	-0.0134

Tabla 5: Tabla de puntos de Balanceo

4.2.3. PUNTOS INFLUENCIALES



Gráfica 5: Criterio distancias de Cook para puntos influenciales



límites!!!
¿cómo se supone que veo ahí los influenciales?

lpt

Gráfica 6: Observaciones vs DFFITS

	Errores Estudentizados	***D.Cook***	***Valor hii***	***DFFITS***
10	1.6356	0.0791	0.1507	0.7028
33	-1.4446	0.2349	0.4031	-1.2024
36	-2.8462	0.8445	0.3848	-2.4636

Tabla 6: Criterio Dffits para puntos influenciales

este vale menos que el límite inferior, ni se ve y es muy importante por ese valor tan alto.

este párrafo se parece mucho al del equipo 33, hicieron plagio?

No se ve nada

Bajo el criterio de DFFITS, se obtuvo la anterior gráfica. A partir de la gráfica podemos concluir que existen varios valores influenciales en el modelo *(Yi, i = 10 ,33 y 36)*. Como los coeficientes de DFFITS son mayores a $|DFFIT| > 2\sqrt{\frac{p}{n}}$ podemos afirmar que los *Yi, i=10, 33 y 36* tienen influencia sobre los β_i , *i = 0,. . . ,50.* Por ende, tales datos deben ser investigados.

No hablan sobre Cook

El equipo 33 dijo exactamente el mismo error. es sobre $\hat{Y}$.

5. CONCLUSIONES 3pt

El modelo de regresión lineal múltiple no es válido debido a que no se cumplen los supuestos del error (Distribución normal, varianza constante y media 0), dichos supuestos se comprobaron mediante la prueba Shapiro-Wilk y gráfica de Residuales estudentizados vs valores ajustados. Al no cumplirse dichas hipótesis indica que en el modelo puede existir la presencia de puntos influenciales.

Al verificar la presencia de observaciones extremas en el modelo, se obtuvo el resultado de que no existían puntos atípicos en el modelo, esto indica que las observaciones no están separadas en su valor de respuesta respecto a Y; Ahora bien, se identificaron cuatro puntos de balanceo en el modelo, esto implica afectaciones en las estadísticas de resumen como lo es el R^2 y los errores estándar de los coeficientes estimados generando una falsa explicación por parte de las variables predictoras a la variable respuesta "*Y: Riesgo de infección*" .

No necesariamente significa eso

Para determinar el impacto de dichas observaciones (atípica y de balanceo) sobre los coeficientes ajustados, se calcularon las observaciones influenciales. Al usar los diferentes diagnósticos para identificar las observaciones se obtuvo que existe la presencia de tres observaciones influenciales la cuales "jalan" el modelo en su dirección y tienen un mayor efecto sobre la recta de regresión generando errores de predicción.

ajuste

Debido a lo anterior se plantea y considera que los resultados de este modelo no deben tomarse como válidos, ya que es necesario reconstruir el modelo sin las observaciones atípicas, de balanceo e influenciales que modifican los resultados y a su vez los supuestos del modelo.

Aunque es delicado de afirmar eso