Trabajo 1

Equipo # 27

9,2

Docente

Mateo Ochoa Medina

Asignatura

Estadística II

Sede Medellín

Fecha

1. Pregunta 1

19 pt

Teniendo en cuenta la base de datos asignada, la cual es **Equipo27.txt**, las covariables son: Duración de la estadía (X_1), Rutina de cultivos (X_2), Número de camas (X_3), Censo promedio diario (X_4) y Número de enfermeras (X_5). El modelo que se propone es:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,\ \ \varepsilon_i \sim N\left(0, \sigma^2\right),\ i = 1, 2, \ldots, 54$$

Donde Y_i es la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).

1.1 Modelo de regresión

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

Tabla 1. Tabla de valores de los coeficientes estimados

	Valor del parámetro
$\hat{\beta}_0$	-1.00520093
$\hat{\beta}_1$	0.14804727
$\hat{\beta}_2$	0.02710202
$\hat{\beta}_3$	0.06298611
$\hat{\beta}_4$	0.01489569
$\hat{\beta}_5$	0.00105300

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.00520093 + 0.14804727X_{i1} + 0.02710202X_{i2} + 0.06298611X_{i3} + 0.01489569X_{i4} + 0.00105300X_{i5}$$

Donde $i = 1, 2..., 54$

1.2 Significancia de la regresión

Para la significancia de la regresión se hará uso de la siguiente tabla Anova y se establece la siguiente prueba de hipótesis:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_5 = 0, \quad vs. \quad H_1: algún\ \beta_j \neq 0,\ j = 1, 2, \ldots, 5.$$

Tabla 2. Tabla ANOVA para el modelo

	Suma de cuadrados	g.l	Cuadrados medios	F_0	Valor-P
Modelo de regresión	33.3216	5	6.664319	9.87581	1.54523e-06
Error	32.3910	48	0.674812		

De la Tabla 2 se obtienen el valor del estadístico de prueba $F_0 = 9.87581$ y su correspondiente valor-P

$vp = 1.54523 * 10^{-6}$.

El valor p, al ser tan pequeño, permite rechazar la hipótesis nula con mucha seguridad, por lo que el modelo de regresión lineal múltiple propuesto es significativo. Esto quiere decir que el riesgo de infecciones hospitalarias depende significativamente de al menos una de las predictoras del modelo.

1.3 Significancia de los parámetros del modelo

Para la significancia de los parámetros se presenta información sobre los parámetros y la siguiente prueba de hipótesis:

$$H_0: \beta_j = 0, \quad vs. \quad H_1: \beta_j \neq 0 \ para\, j = 0, 1, \ldots, 5$$

Tabla 3. Resumen de los coeficientes

	Estimación β_j	$se\left(\hat{\beta}_j\right)$	T_0	Valor-P
β_0	-1.00520093	1.5842203057	-0.6345083	0.5287619045
β_1	0.14804727	0.1121482175	1.3201036	0.1930613448
β_2	0.02710202	0.0286236436	0.9468405	0.3484625569
β_3	0.06298611	0.0174368340	3.6122445	0.0007245836
β_4	0.01489569	0.0067798385	2.1970574	0.0328763504
β_5	0.00105300	0.0006012114	1.7514645	0.0862537487

6 pt

Donde el estadístico de prueba T_0 para cada coeficiente β_j es de la forma

$$T_{j,0} = \frac{\hat{B}_j}{se(\hat{B}_j)} \sim t_{48}, \quad j = 0, 1, \ldots, 5$$

Los valores-P permiten concluir que los parámetros individuales β_3, β_4 son significativos, cada uno de estos en presencia de otros parámetros.

Además, notamos que los parámetros $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_5$ son individualmente no significativos en presencia de los demás parámetros.

1.4 Interpretación de los parámetros

3 pt

- $\hat{\beta}_3$: Por cada unidad que aumente el número de camas (X_3), la probabilidad promedio de riesgo de infección aumenta en 0.06298611 unidades, cuando las demás predictoras se mantienen fijas, en otras palabras el riesgo aumenta un 6.3% aproximadamente.

- $\hat{\beta}_4$: Por cada unidad que aumente el censo promedio diario (X_4), la probabilidad promedio de riesgo de infección aumenta en 0.01489569 unidades, cuando las demás predictoras se mantienen fijas, en otras palabras el riesgo aumenta un 1.5% aproximadamente..

1.5 Coeficiente de determinación múltiple R^2

¿cómo se calcula? 2 pt

El modelo tiene un $R^2 = 0.5070$ lo que significa que aproximadamente el 50,70% de la variabilidad total del riesgo de infección está explicado por el modelo de regresión múltiple propuesto.

Adicionalmente, calculamos el R^2_{adj}= 0.4557.

2. Significancia de un subconjunto

9 pt

Para esto, el subconjunto elegido es un subconjunto formado por los tres valores B_j j=0,1…5 los cuales tienen el valor p más pequeño entre el resto de parámetros. Dichos parámetros son β_3, β_4 *y* β_5. Luego, generamos el siguiente juego de hipótesis.

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0 \qquad \text{vs} \qquad H_1: Al\ Menos\ un\ \beta_J \neq 0,\ j = 3, 4, 5$$

Lo que nos daría 2 subconjuntos de la siguiente forma:

$$A = \{\beta_3, \beta_4, \beta_5\}$$

$$B = \{\beta_0, \beta_1, \beta_2\}$$

Ahora calcularemos el SSR:

$$SSR(\beta_3, \beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2)$$

$$SSR(\beta_3, \beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2) = SSE\left(\beta_0, \beta_1, \beta_2\right) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$$

De la tabla de todas las regresiones posibles obtenemos que $SSE\left(\beta_0, \beta_1, \beta_2\right) = 51.209$ y $SSE\left(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5\right) = 32.3910$.

$$SSE(\beta_3, \beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2) = 51.209 - 32.3910 = 18.818$$

Y sus grados de libertad son:

$$DF = (n - 3) - (n - 6)$$

$$DF = 3$$

Ahora bien, el MSR seria:

3 pt

$$MSR(\beta_3, \beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2) = \frac{SSR\left(\beta_0, \beta_1, \beta_2\right)}{3} = 6.272667$$

Y nuestro estadístico de prueba se calcula de la siguiente forma:

$$F_0 = \frac{MSR(\beta_3, \beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2)}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} = \frac{6.27667}{0.674812} = 9.2954 \sim f_{3,48}$$

6

1 pt

Como $F_0 = 9.2954 > f_{0.05,3,48} = 2.798061$, entonces se rechaza H_0 y se concluye que la probabilidad promedio de adquirir una infección en el hospital, está influenciada por el número de camas, el censo promedio diario y el número de enfermeras . Con base a estos resultados, podemos afirmar que es posible aceptar del modelo anterior las variables de este subconjunto que involucran a β_3, β_4 y β_5.

3. Prueba de hipótesis general lineal

4 pt

¿Existe alguna relación significativa entre la probabilidad de adquirir una infección en el hospital y las características hospitalarias, como la duración de la estadía de los pacientes, la rutina de cultivos, el censo promedio diario, agregado a esto, la significancia entre el número de camas y el número de enfermeras es la misma?

H_0:No existe una relación significativa entre el riesgo de infección (Y) y ninguna de las características hospitalarias (Duración de la estadía (X_1), Rutina de cultivos (X_2) y el censo promedio (X_4)), además la significancia del número de camas (X_3) y la significancia del número de enfermeras (X_5) es la misma.

H_1:Existe una relación significativa entre el riesgo de infección (Y) y las características hospitalarias (Duración de la estadía (X_1), Rutina de cultivos (X_2) y el censo promedio (X_4)), además la significancia del número de camas (X_3) y la significancia del número de enfermeras (X_5) no es la misma.

Es decir:

$$H_0: \beta_1 = \beta_2 = \beta_4 = 0,\ \beta_3 = \beta_5 \ \ vs \ \ H_1: Algún\ \beta_j \neq 0\ j = 1,\ 2,\ 4,\ \ \beta_3 \neq \beta_5$$

Lo que nos quedaría un sistema de ecuaciones con m=4 y una hipótesis nula tal que así:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Entonces, la matriz **L** queda de la siguiente forma:

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

Que tiene un r=4, el cual es la cantidad de filas linealmente independientes.

A partir de lo anterior, el modelo reducido quedaría tal que así:

$$RM = Y = \beta_0 + \beta_3(X_3 + X_5) + \varepsilon_i$$

En este modelo se tiene una suma de cuadrados del error $SSE(RM) = SSE(\beta_0, \beta_3)$ con n-2 grados de libertad

Ahora para calcular el SSH primero calcularemos el SSE (FM) que sería:

$$SSE(FM) = SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) \; con\, n - 6 \; grados\; de\; libertad$$

Entonces nuestro SSH nos quedaría de la siguiente forma

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

Y sus grados libertad serían:

2 pt

$$DF = (n - 2) - (n - 6)$$

$$DF = 4$$

Entonces nuestro estadístico de prueba quedaría de la siguiente forma:

$$F_0 = \frac{MSH}{MSE} = \frac{SSH/4}{MSE} = \frac{[SSE(RM) - SSE(FM)]/4}{MSE(FM)} = \frac{[SSE(RM) - 32.3910]/4}{0.674812}$$

4. Validación de los supuestos

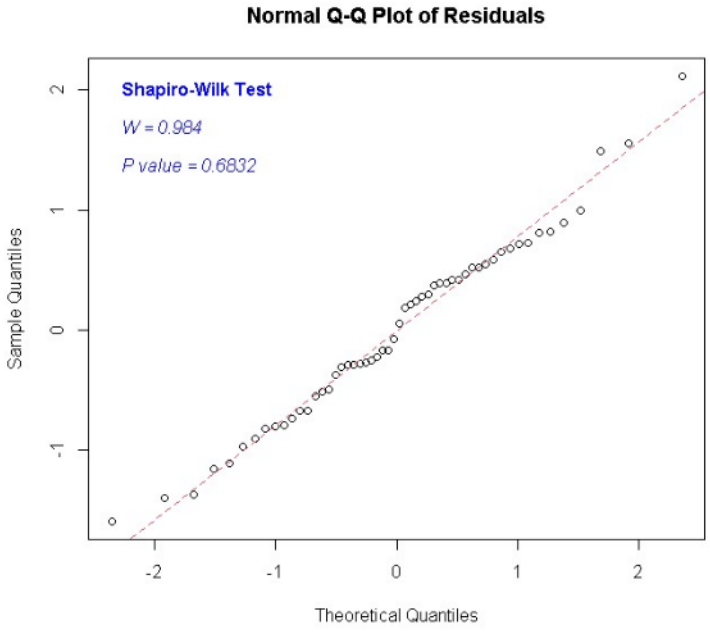
4.1. Supuesto de normalidad 14,5 pt

Para confirmar que los ε_i se distribuyen de forma normal, efectuamos una prueba de Shapiro Wilk, en el cual tenemos el siguiente juego de hipótesis:

$$H_0: \varepsilon_i \sim N(0, \sigma^2) \;\; vs. \;\; H_1: \varepsilon_i \; no\; se\; distribuyen\; N(0, \sigma^2)$$

De la prueba Shapiro Wilk obtenemos el siguiente resultado:

y esto son literalmente todos los supuestos.



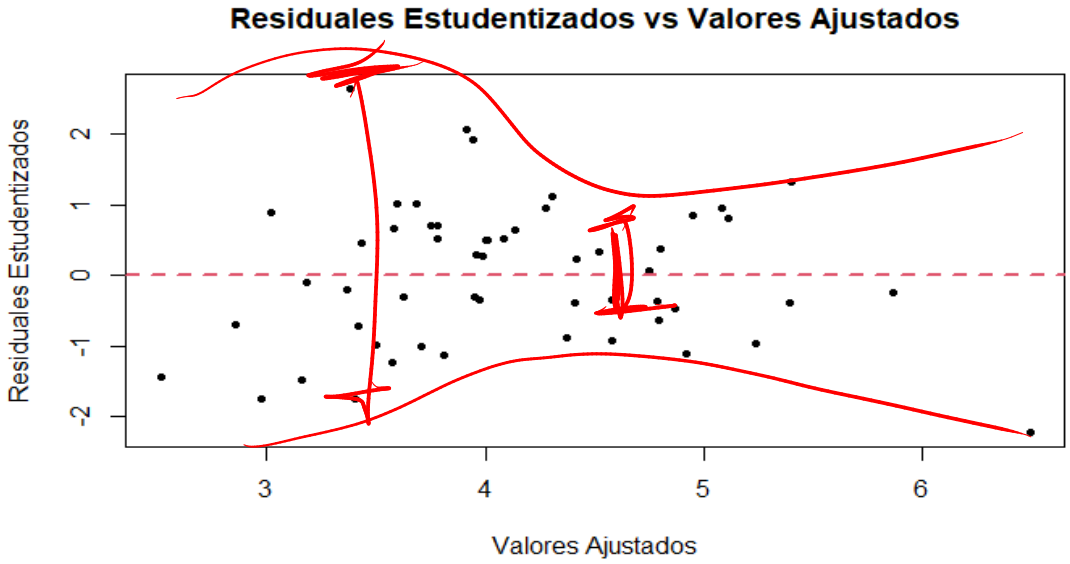
La prueba de normalidad Shapiro-Wilk indica que los errores son normales (valor-P = 0.6832 > 0.05). A pesar de tener cierta tendencia curvilínea, no posee colas notoriamente pesadas entonces diremos que si se distribuye con normalidad.

4.2 Supuesto de varianza constante

Para verificar si la varianza es constante o no, se hará la prueba de forma gráfica, a través del gráfico de residuales vs. valores ajustados, donde tenemos que probar:

$$H_0: Var[\varepsilon_i] = \sigma^2 \quad vs. \quad H_1: Var[\varepsilon_i] \neq \sigma^2$$

Para ello se usa la siguiente gráfica:



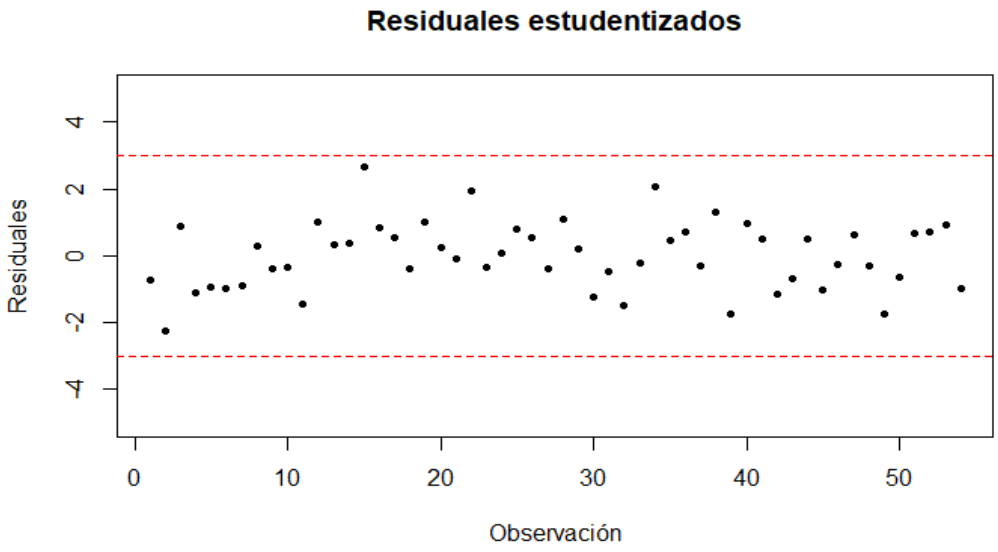
Analizando el grafico concluimos que la varianza es constante por que no se observa un cambio notorio de esta.

4.3 Verificación de las observaciones

4.3.1. Valores atípicos

Se considera que una observación es atípica cuando su residual estandarizado r_i es :

$|r_i| > 3$



Graficamentes, vemos que ningún valor es menor que -3 o mayor a 3

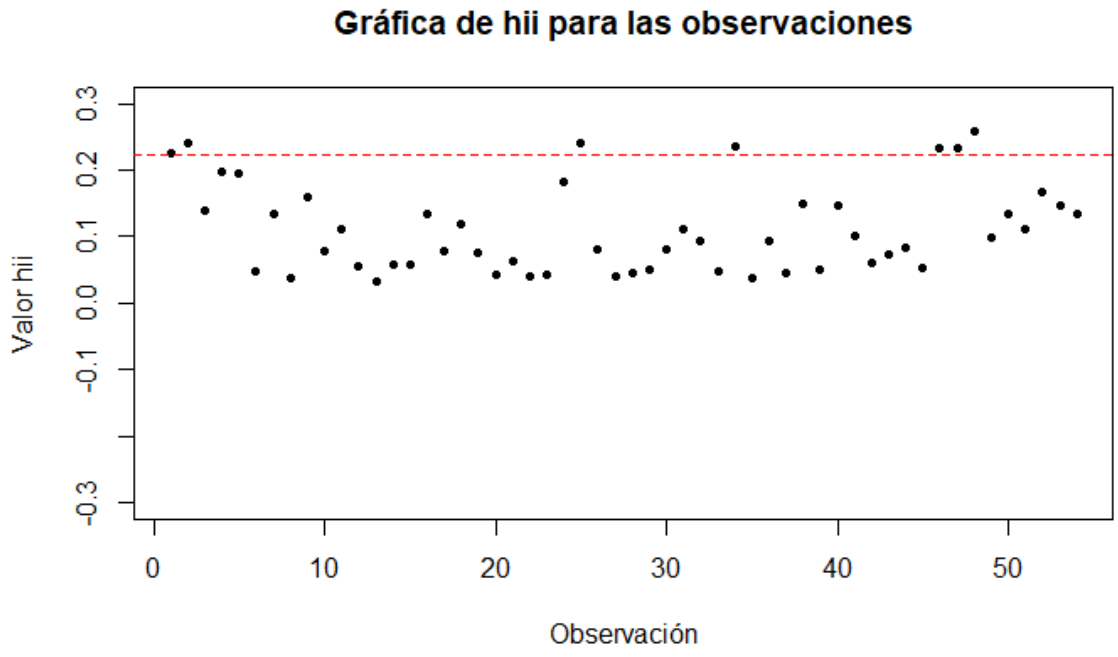
4.3.2 Puntos de balanceo

Una observación es un punto de balanceo sí $hii > 2P/n = 12/54 = 0.22$. Donde los puntos que sobrepasan esta, se consideran puntos de balanceo, donde "P" es el número de parámetros y "n" la cantidad de muestras:

Tabla 4. Resumen puntos de balanceo

Observación	Valor hii
1	0.2249
2	0.2416
25	0.2408
34	0.2364
46	0.2320
47	0.2336
48	0.2579

De acuerdo a la Tabla 4 se puede observar que los puntos 1, 2, 25, 34, 46, 47 y 48 son puntos de balanceo, además en la gráfica vemos que se resaltan las 7 observaciones ya mencionadas:



4.3.3 Puntos influenciales

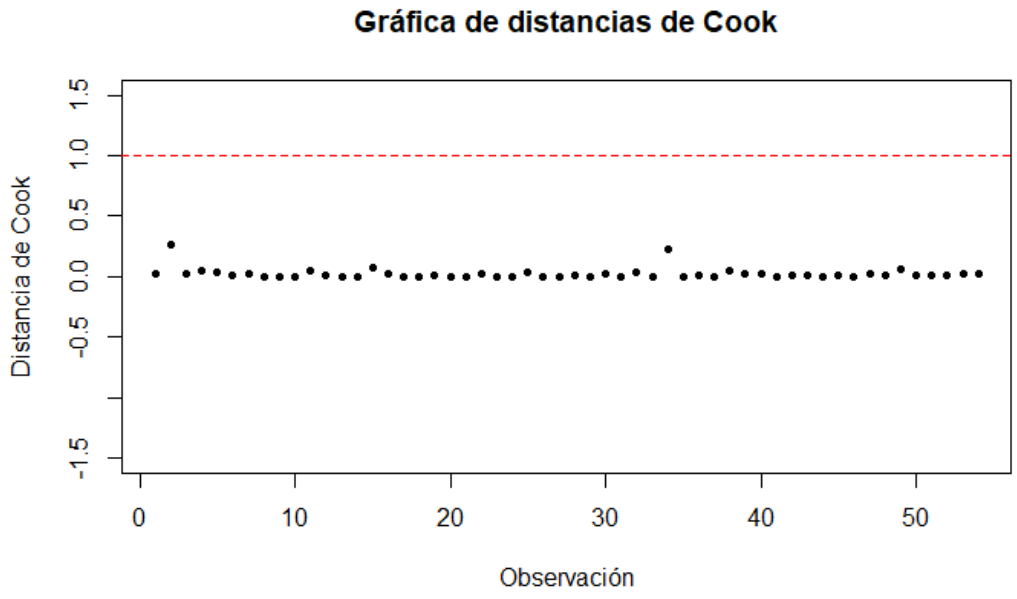
Una observación será influencial si $D_i > 1$ y si $|Dffits| > 2\sqrt{\frac{P}{n}}$. En este caso en Dffits debe superar en valor absoluto a $2\sqrt{\frac{6}{54}} = 0.6667$

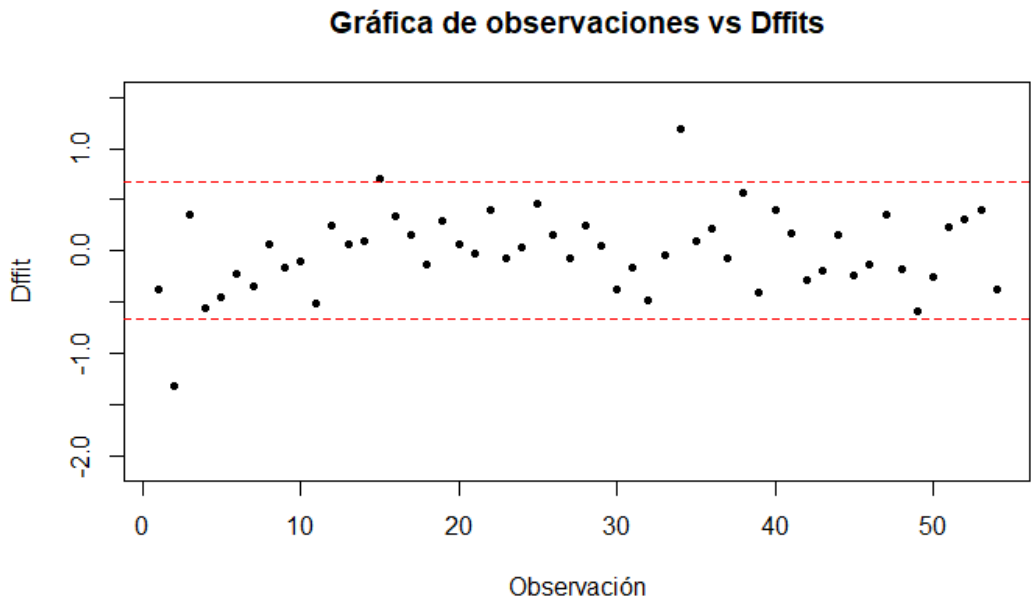
Según el criterio de Dffits $|Dffit| > 2\sqrt{(P/n)} = 0.667$, las observaciones 2, 15 y 34 son puntos influenciales. Ningún dato cumple el criterio de distancias de Cook tal que $D(i) > 1$, por lo que a partir de esta medida no añadimos ningún otro valor influencial adicional.

Tabla 5. Resumen observaciones influenciales

Observación	Cooks (D_i)	Dffits
2	0.2644	-1.3166
15	0.0731	0.7094
34	0.2219	1.1966

De acuerdo a la columna Cooks (D_i) no tenemos una observación influencial y de acuerdo a la columna Dffits, tenemos que las observaciones 2, 15 y 34 son influenciales. Esto lo podemos confirmar mediante los siguientes gráficos:





4.4 Conclusiones

Los supuestos de normalidad y varianza constante se cumplen, por lo tanto, el modelo es válido para hacer estimaciones y predicciones.

Al analizar las observaciones extremas que pueden alterar el modelo se obtuvo que ninguna de las observaciones es atípica, pero si evidenciamos que 7 observaciones son puntos de balanceo, 3 observaciones son influenciales por el criterio de Dffits y que ninguna observación es influencial por el criterio de distancia de Cook

Teniendo en cuenta la presencia de dichas observaciones extremas, lo mejor es estudiarlas y ver como influyen antes de usar el modelo.

Estudiantes:

- Nicolas Orozco Medina / 1011510119
- David Iral Roldan / 1025881354
- Juan Jose Jimenez / 1025641141
- Santiago Molina Velásquez /1000292934