Trabajo No.1

3,9

Equipo Nro 31

Karla Orozco Velez
Mateo Murcia Valles
Juan Diego Espinosa Hernandez
Angie Dayana Palacio Rodriguez

Profesor

Mateo Ochoa Medina

Asignatura

Estadística 2

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
05 de octubre de 2023

Índice

1. Punto 1

Estime un modelo de regresión lineal múltiple que explique el riesgo de infección en términos de las variables restantes (actuando como predictoras) Analice la significancia de la regresión y de los parámetros individuales. Interprete los parámetros estimados. Calcule e interprete el coeficiente de determinación múltiple R2.

Solución

Teniendo en cuenta la base de datos brindada, en la cual hay 5 variables regresoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 69$$

Donde;

- 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

1.1. Modelo de regresión

```
model <- lm(Y ~ ., data = DB)
summary(model)
```

Variable	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-0.7746802	1.5739233	-0.492	0.62429
X1	0.0986650	0.0838726	1.176	0.24387
X2	0.0385808	0.0307500	1.255	0.21424
X3	0.0528568	0.0150538	3.511	0.00083***
X4	0.0099527	0.0060499	1.645	0.10493
X5	0.0019262	0.0006988	2.757	0.00763**

Cuadro 1: Resumen de coeficientes de regresión

Por lo tanto, el modelo de regresión ajustado es:

$\hat{\beta}_0 = -0.7746802,\ \hat{\beta}_1 = 0.0986650,\ \hat{\beta}_2 = 0.0385808,\ \hat{\beta}_3 = 0.0528568,\ \hat{\beta}_4 = 0.0099527,\ \hat{\beta}_5 = 0.0019262$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \hat{\varepsilon}_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 69$$

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,63} \tag{1}$$

Estadística	Valor
Error estándar residual	0.9062
Grados de libertad (residual)	63
R^2 múltiple	0.4499
R^2 ajustado	0.4062
Estadístico F	10.3
Grados de libertad (modelo)	5
Grados de libertad (residual)	63
Valor p	2.999×10^{-7}

Cuadro 2: Resumen de estadísticas del modelo

Dados los valores-p individualmente, podemos notar que las variables X_3(Número promedio de camas) y X_5(Número promedio de enfermeras) son las unicas variables significativas individualmente del modelo, con un $\alpha = 0.05$

Con un valor $p = 0.000000299$ de la regresion del modelo propuesto, y con un $\alpha = 0.05$ podemos decir que el modelo propuesto es significativo. Esto quiere decir que la probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje) es afectada significativamente por al menos una de las predictoras consideradas.

1.3. Interpretación de los parámetros

Lo primero es identificar aquellos parámetros susceptibles de interpretación, esto es, solo se podrán interpretar parámetros que resultaron significativos individualmente, en este caso son: $\hat{\beta}_3$ y $\hat{\beta}_5$

$\hat{\beta}_3 = 0.0528568$ indica que por cada unidad de aumento en el número de camas el promedio estimado de adquirir infección en el hospital (en porcentaje) aumenta en 0.0528568, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_5 = 0.0019262$ indica que por cada enfermera adicional de tiempo completo el promedio estimado de adquirir infección en el hospital (en porcentaje) aumenta en 0.0019262, cuando las demás variables predictoras se mantienen fijas.

1.4. Coeficiente de determinación multiple R^2

Con R^2 múltiple igual a 0.4499 podimos decir que el 44.9 % de la variabilidad en el riesgo de infección es explicada por el modelo propuesto.

2. Punto 2

Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más pequeños del punto anterior. Según el resultado de la prueba este subconjunto de parámetros son todos significativos? Explique su respuesta.

2.1. Planteamiento prueba de hipotesis y modelo reducido

Las covariables con el P- Valor mas bajo en el modelo fueron X_3,X_4,X_5, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

	SSE	Covariables del modelo
Modelo completo	51.739	$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$
Modelo reducido	69.772	$\beta_1 = \beta_2$

Cuadro 3: Resumen de SSE y Covariables del Modelo

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 69$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \ldots, \beta_5))/3}{MSE(\beta_0, \ldots, \beta_5)} \overset{H_0}{\sim} f_{3,63} \\ &= \frac{(69.772 - 51.739)/3}{51.739/63} \\ &= \frac{6.011}{0.821} \\ &= 7.321 \end{aligned} \tag{2}$$

Ahora, al comparar F_0 con $f_{3,63} = 2.360$, podemos observar que $F_0 > f_{3,63}$,y por lo tanto, rechazamos la hipótesis nula, lo que indica significancia estadística.

3. Pregunta 3

3.1. Prueba de hipotesis y prueba de hipotesis matricial

Se quiere probar $H_0 : \beta_1 = \beta_4, \beta_2 = \beta_5$, versus una hipótesis alternativa, esta hipotesis se puede plantear en termino de $H0 : L\beta = 0$ asi que se tiene una prueba de hipotesis lineal general, asi que tendremos lo siguiente:

3.2. Prueba de hipotesis

$$H_0 : \begin{cases} \beta_1 - \beta_4 = 0 \\ \beta_2 - \beta_5 = 0 \end{cases}$$

$$H_1 : \begin{cases} \beta_1 - \beta_4 \neq 0 \\ \beta_2 - \beta_5 \neq 0 \end{cases}$$

Que en forma matricial se puede expresar de la siguiente forma:

$$L\beta = 0$$

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

Note que la matriz L, tiene $r = 2$ filas linealmente independientes, entonces:

Modelo resultante(reducido):

$$\begin{aligned} Y &= \beta_0 + \beta_1(X_1 + X_4) + \beta_2(X_2 + X_5) + \beta_3 X_3 \\ &= \beta_0 + \beta_1 X_{1,4} + \beta_2 X_{2,5} + \beta_3 X_3 \end{aligned}$$

donde $X_{1,4} = X_1 + X_4$, y $X_{2,5} = X2 + X5$

3.3. Estadistico de prueba

Para el estadistico de prueba tendremos la siguiente expresion:

$$F_0 = \frac{MSH}{MSE} = \frac{\frac{SSE(RM)-SSE(FM)}{2}}{MSE}$$

2 pt

$$F_0 = \frac{\frac{SSE(RM)-51.739}{2}}{0.8212}$$

Con respecto a el $SSE(RM)$, ocurre lo siguiente, en primera instancia con lo visto en el curso aun no se puede calcular, y segundo en la tabla de todas las regresiones posibles no se puede obtener este valor ya que no admite sumas de variables entre sus opciones, por lo tanto solo lo dejaremos expresado.

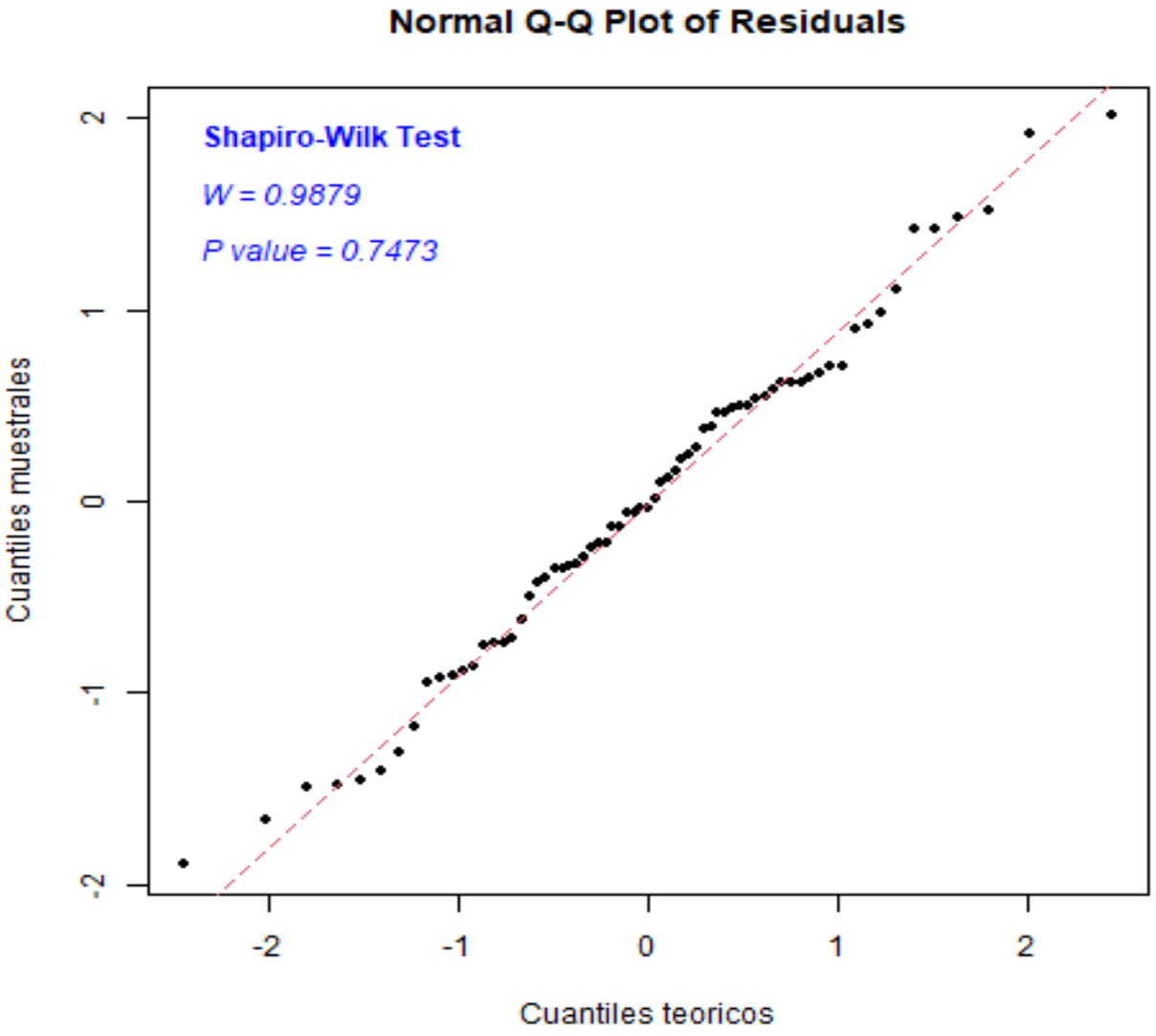
4. Pregunta 4

14,5 pt

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

$$H_0 : \epsilon_i \sim \text{Normal} \quad \text{vs} \quad H_1 : \epsilon_i \text{Normal}$$

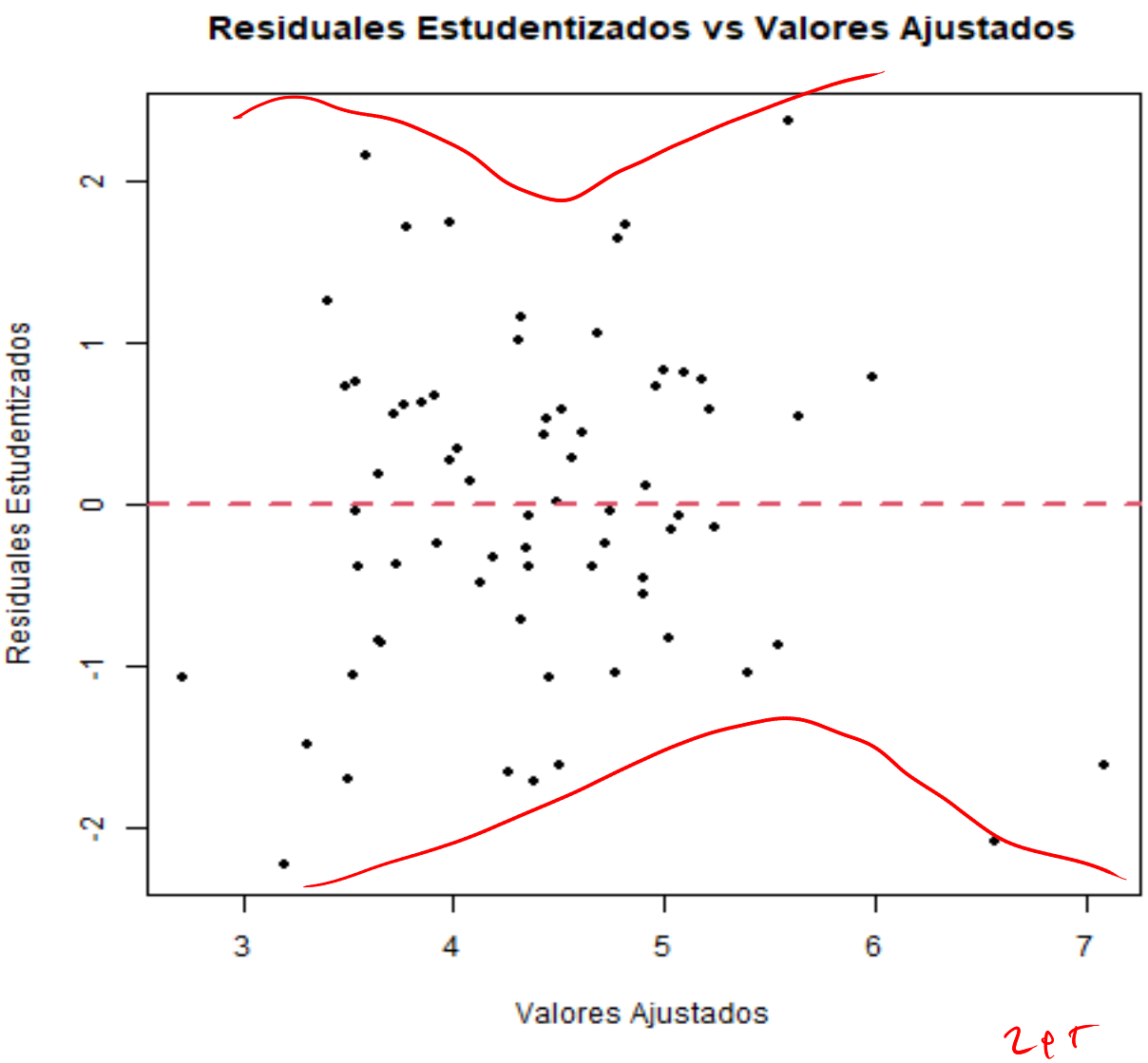


3,5 pt

Al analizar la gráfica detalladamente, podemos ver como la mayoria de los puntos se acercan bastante a linea roja (linea teorica) y esto nos puede indicar normalidad en nuestro datos, por el lado de las colas, nos indica que la distribucion de nuestros datos tiene una cola mas pesada por lo tanto es sesgada a la izquierda, teniendo en cuenta el test de S-W, tomando un $\alpha = 0.05$ se acepta la hipotesis nula ya que nuestro valor-P es mayor a 0.05.

4.1.2. Varianza constante

$$H_0 : V[\epsilon_i] = \sigma^2 \; VS \; H_1 : V[\epsilon_i] \neq \sigma^2$$

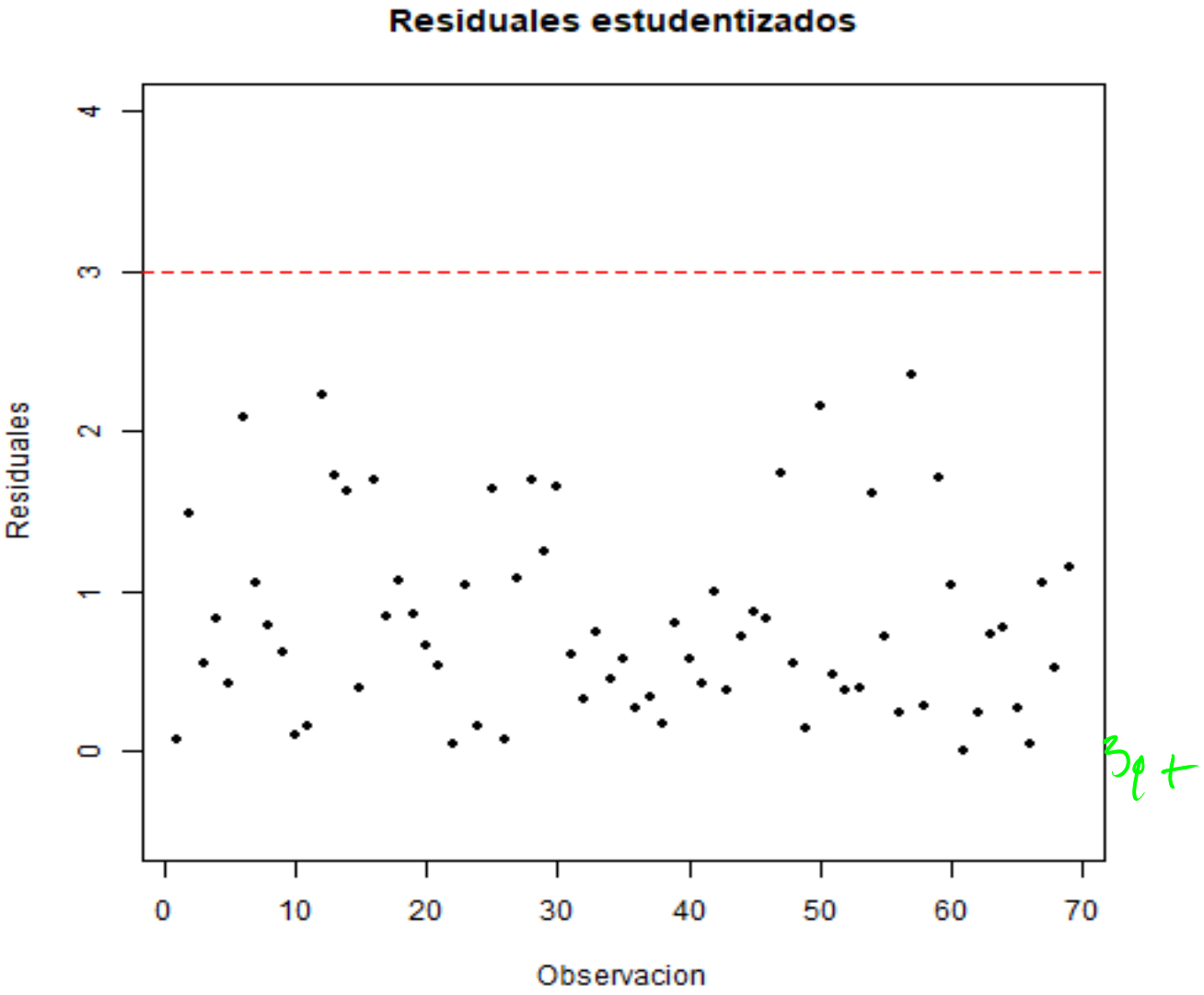


En este caso, podemos dibujar una especia de rectangulo con nuestros puntos y esto nos indicaria que tenemos varianza constante y por lo tanto se cumple este supuesto.

4.2. Verificacion de las observaciones

4.2.1. Datos atípicos

$$|r_i| = \left| \frac{e_i}{\sqrt{MSE(1-h_{ii})}} \right| > 3$$



Tomamos el valor absoluto de cada uno de los residuales estudentizados, los graficos y trazamos una recta en Y = 3, con el objetivo de identificar graficamente los valores atipicos ($|r_i| > 3$).

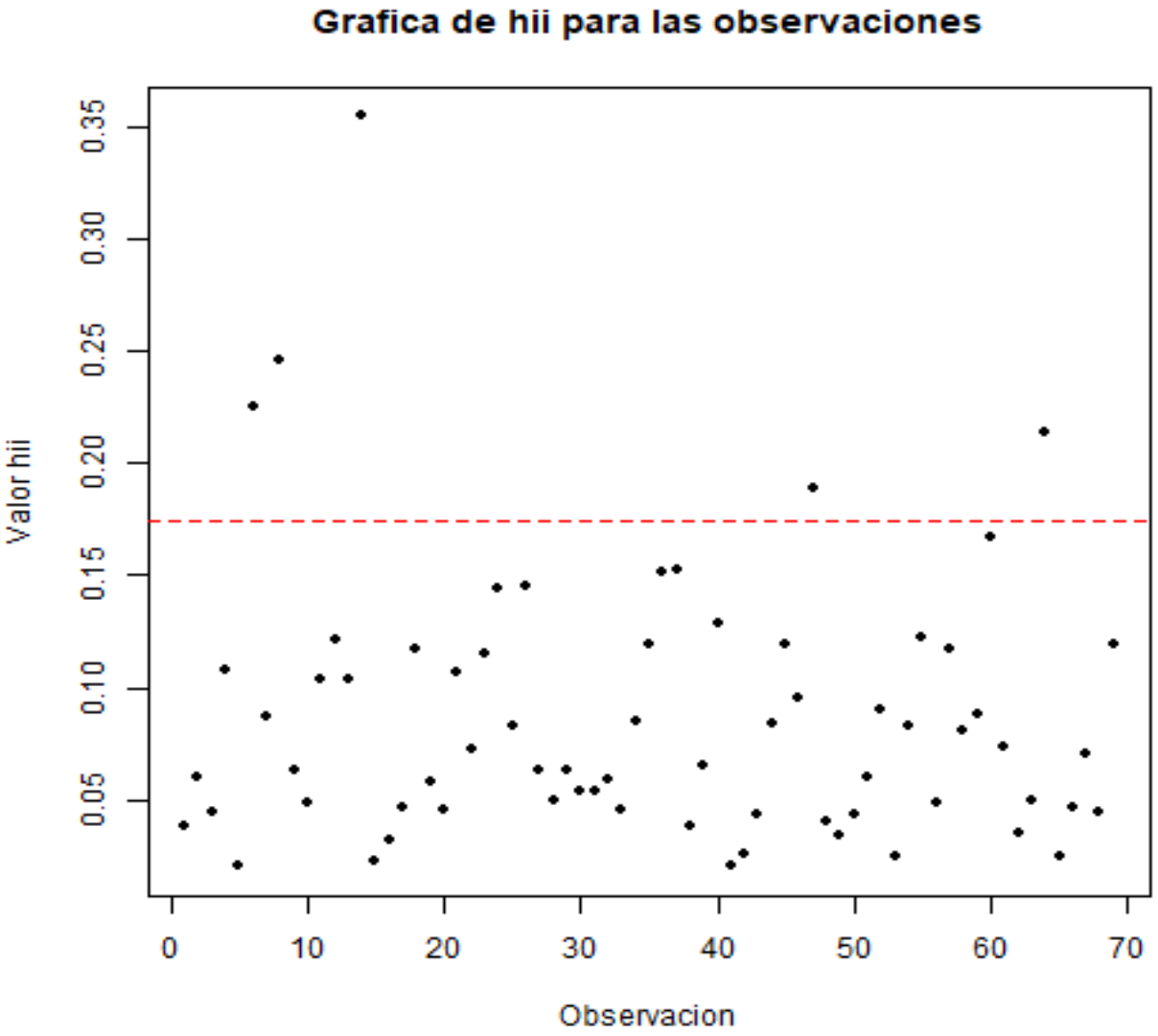
En esta grafica podemos observar que no tenemos datos atípicos ya que el valor absoluto de los residuales estudentizados no superan el valor de 3, ademas de esto hicimos la siguiente tabla, en la cual se confirma esto.

Var1	Freq
FALSE	69

4.2.2. Puntos de balanceo

$$h_{ii} > \frac{2p}{n}$$

$$Donde: \ \frac{2p}{n} = 0.173913$$



Respecto a la gráfica encontramos 5 puntos los cuales se salen de nuestra banda la cual equivale al calculo $\frac{2p}{n}$ y ya la superan, podemos decir que corresponden a puntos de balanceo los datos de 6, 8, 14, 47 y 64.

En la siguiente tabla, correspondiente al diagnostico realizado para el balanceo. Y encontramos tambien 5 puntos de balanceo, sin embargo la tabla nos dice exactamente a cual observacion corresponde.

	hii.value
6	0.2249
8	0.2464
14	0.3545
47	0.1892
64	0.2141

¿cuáles...?

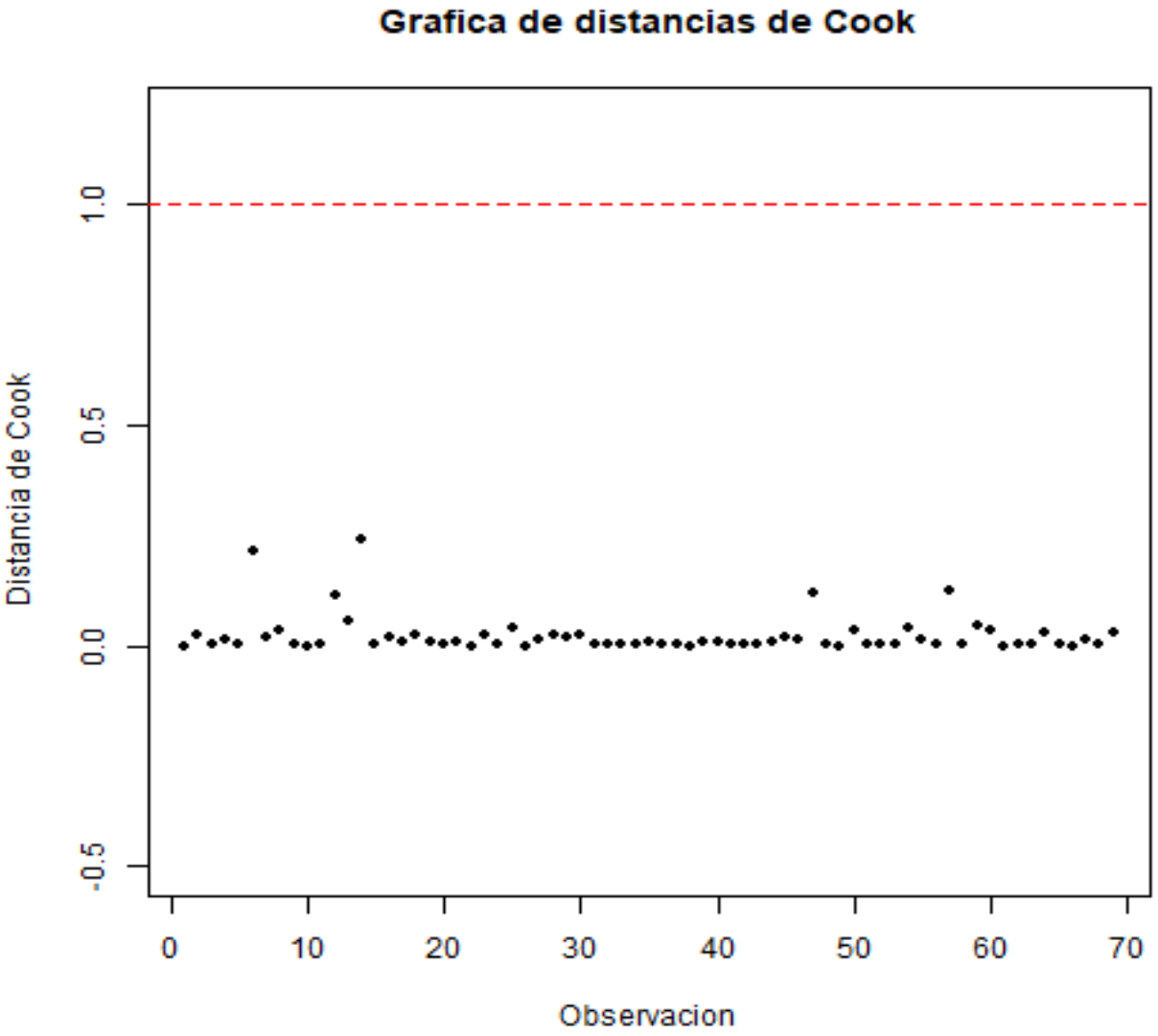
20 +

4.2.3. Puntos influenciales

Criterio de cook:

$$D_i > 1$$

Donde D_i se refiere al valor de la distancia de cook calculado en cada uno de nuestros valores observados



Var1	Freq
FALSE	69

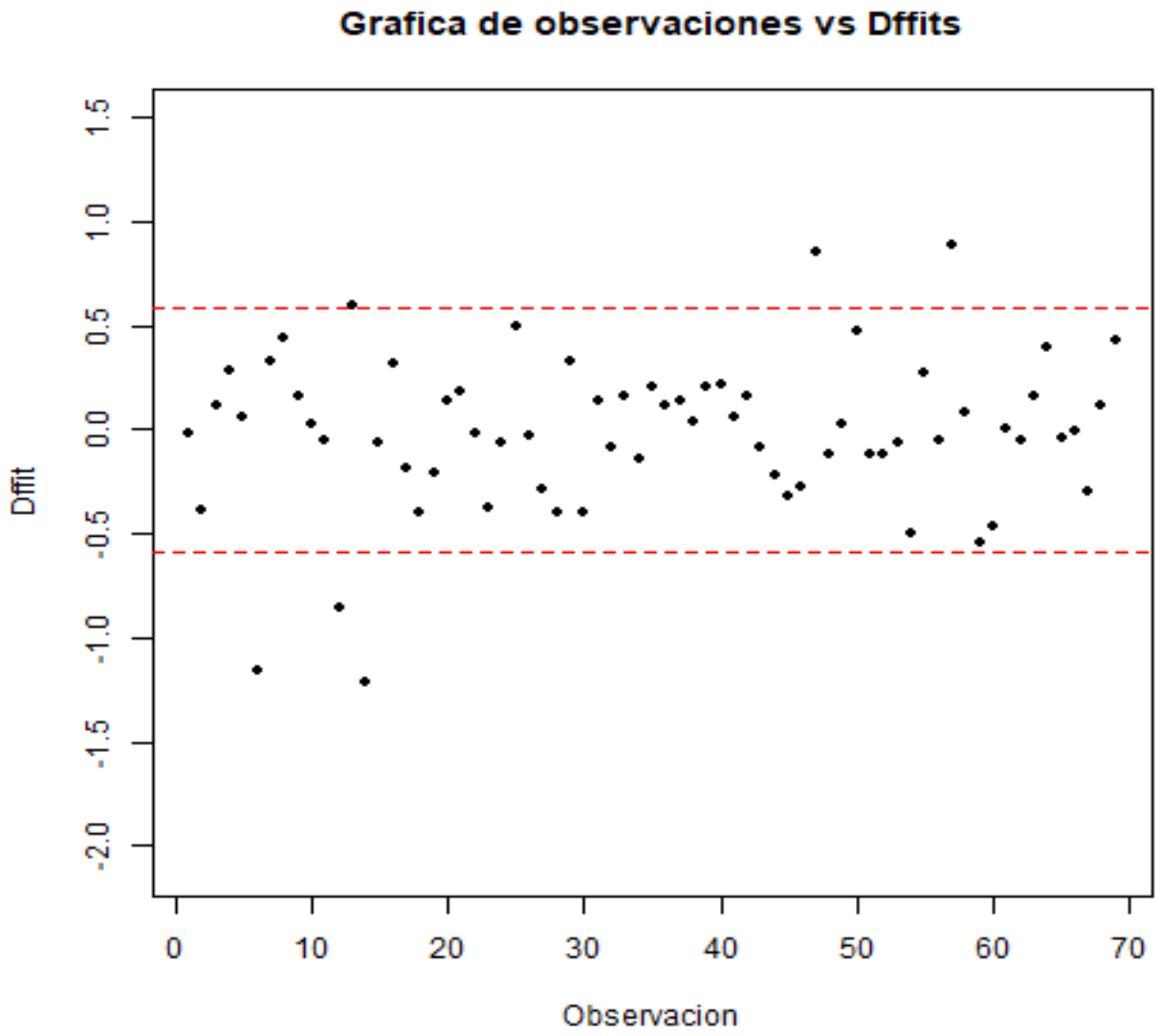
2 pt

Como podemos ver en la grafica y ademas haciendo una tabla en la cual se buscaron distancias de cook las cuales superaran el valor de 1, y encontramos que bajo este criterio ningun valor es influenciable, asi que proseguiremos con el criterio DFFITS.

Criterio DFFITS:

$$|DFFITS| > 2\sqrt{\frac{p}{n}}$$

$$Donde: \ 2\sqrt{\frac{p}{n}} = 0.5897678$$



Como vemos en la grafica hay un valor en justamente sobre la frontera que dibujamos para identificar valores influenciables, por lo tanto la tabla nos sera de gran ayuda para saber realmente si es mayor, menor o incluzo igual.

	Dffits
6	-1.1597
12	-0.8590
13	0.5965
14	-1.2183
47	0.8520
57	0.8934

Bajo el análisis de DFFITS, observamos una variación en 6 puntos, corresponden a los datos 6, 12, 13, 14, 47 y 57, los cuales se pueden afirmar, son influenciables.

4.3. Conclusión

Para el caso de este modelo, encontramos que se cumplio la validez de todos los supuestos, sin embargo debemos de tener en cuenta que encontramos puntos de balanceo este punto influyente no afecta las estimaciones de los coeficientes de regresión, pero ciertamente sí tiene un efecto marcado sobre las estadísticas de resumen del modelo, como R^2 y sobre los errores estándar de los coeficientes de regresión.

Por otro lado, la presencia de dos puntos de influencia en el gráfico de Observaciones vs Dffits muestra que estos valores extremos tienen una gran influencia en la estructura de la relación entre variables, lo que puede afectar en gran medida las inferencias y los resultados del argumento del modelo.

Aunque cumplimos con los supuestos, la presencia de puntos influyentes afectara las inferencias que querramos hacer con este modelo, asi que habria que tratar cada caso (punto influyente) por aparte y tomar decisiones al respecto, por el lado de las predicciones el modelo nos podria ayudar, sin embargo lo ideal seria compararlo con otros y asi saber cual es el mejor entre los modelos planteados.

Nunca dijeron si era valido o no, la red por cumplir supestos