Trabajo 1

4,6

Estudiantes

Manuel Rivera Estrada
Manuela Usme Martinez
Luis Alejandro Varela Ojeda
Leonardo Echavarria Cardona

Equipo 52

Docente

Carlos Mario Lopera

Asignatura

Estadística II

Sede Medellín
05 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Donde cada variable representa lo siguiente:

- **Y**: Riesgo de infección
- **X1**: Duración de la estadía
- **X2**: Rutina de cultivos
- **X3**: Número de camas
- **X4**: Censo promedio diario
- **X5**: Número de enfermeras

1.1. Modelo de regresión

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	-1.8132
β_1	0.2326
β_2	0.0332
β_3	0.0682
β_4	0.0060
β_5	0.0023

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.8132 + 0.2326 X_{1i} + 0.0332 X_{2i} + 0.0682 X_{3i} + 0.006 X_{4i} + 0.0023 X_{5i};\ 1 \leqslant i \leqslant 64$$

1.2. Significancia de la regresión

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

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

El estadístico de prueba se expresa como:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5,58} \tag{1}$$

Se aprecia que el estadístico de prueba sigue una distribución F con 5 grados de libertad.

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	69.2508	5	13.850165	14.6625	2.77533e-09
Error	54.7867	58	0.944598		

En la tabla ANOVA, se observa un valor de P extremadamente pequeño de 2.77533e-09 que es menor a un nivel de significancia $\alpha = 0.05$, por otro lado también tenemos un valor de $F_0 = 14.6625$ que es mayor que $f_{0.95,5,58} = 2.7636$, lo que respalda la afirmación de que al menos una de las variables predictoras tiene un efecto significativo en la variable de respuesta. Esto conduce al rechazo de la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto la regresión es significativa.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar cuáles de ellos son significativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-1.8132	1.5329	-1.1829	0.2417
β_1	0.2326	0.0802	2.9000	0.0053
β_2	0.0332	0.0289	1.1511	0.2544
β_3	0.0682	0.0165	4.1234	0.0001
β_4	0.0060	0.0081	0.7388	0.4630
β_5	0.0023	0.0007	3.1813	0.0024

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia $\alpha = 0.05$, los parámetros β_1, β_3 y β_5 son significativos, ya que tienen P-valores de 0.0053, 0.0001 y 0.0024, respectivamente, que son menores que un nivel de significancia de α.

1.4. Interpretación de los parámetros

$\hat{\beta}_1$**:** El coeficiente β_1 relacionado con la variable X_1, que representa la duración promedio de la estadía de los pacientes en los hospitales (en días). Su valor estimado es 0.2326, con un error estándar de 0.0802. Esta estimación nos indica que, cuando la duración de la estadía (X_1) aumenta en una unidad (días) y los demás parámetros se mantienen constantes, el riesgo de infección (Y) tiene un aumento promedio de 0.2326 unidades. Además, el P-valor asociado a β_1 es 0.0053, por lo que se concluye que la duración e la estadía de los pacientes en el hospital tiene un efecto significativo en el riesgo de infección.

$\hat{\beta}_3$**:** El coeficiente β_3, relacionado con la variable X_3 que representa el número promedio de camas en el hospital, tiene un valor estimado de 0.0682 y un error estándar de 0.0165. Lo cual significa que cuando aumenta en una unidad el número de camas (X_3) en el hospital y los demás parámetros se mantienen constantes, el riesgo de infección (Y) tiene un incremento promedio de 0.0682 unidades. Además, se tiene en cuenta el P-valor asociado a β_3 ya que es muy pequeño (0.0001), lo que indica una alta significancia estadística.

$\hat{\beta}_5$**:** El coeficiente β_5, relacionado con la variable X_5 que representa el número promedio de enfermeras en el hospital durante el tiempo de estudio, tiene un valor estimado de 0.0023 y un error estándar de 0.0007. Lo cual significa que cuando aumenta en una unidad el número de enfermeras (X_5) en el hospital y los demás parámetros se mantienen constantes, el riesgo de infección (Y) tiene un incremento promedio de 0.0023 unidades. Además, se tiene en cuenta el P-valor asociado a β_5 ya que es bastante pequeño (0.0024), lo que indica una alta significancia estadística.

1.5. Coeficiente de determinación múltiple R^2

El modelo presenta un coeficiente de determinación múltiple $R^2 = 0.5583$. Esto indica que alrededor del 55.83 % de la variabilidad total en la variable de respuesta puede ser explicada por las variables predictoras incluidas en el modelo de regresión propuesto en este informe.

2. Pregunta 2

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariables con el P-valor más alto en la tabla de coeficientes son:

β_2 asociado a la variable X_2 (Rutina de cultivos) con un P-valor de 0.2544. β_4 asociado a la variable X_4 (Censo promedio diario) con un P-valor de 0.4630. Estos son los coeficientes que tienen los P-valores más altos en el modelo por lo que su significancia es baja, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

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

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	54.787	X1 X2 X3 X4 X5
Modelo reducido	56.765	X1 X3 X5

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

2.2. Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/r}{(SSE(\beta_0, \cdots, \beta_5))/(n-p)} \overset{H_0}{\sim} f_{3,58} \\ &= \frac{(56.765 - 54.787)/2}{(54.787)/(64-6)} \\ &= \frac{0.989}{0.9446} \\ &= 0.9506 \end{aligned} \tag{2}$$

Ahora, comparando el F_0 con $f_{0.95,3,58} = 2.7636$, se puede ver que $F_0 < f_{0.95,1,45}$ y por tanto se rechaza H_1 aceptando la hipotesis incial, que dice que $\beta_2 = \beta_4 = 0$

Es posible descartar las variables x_2 y x_4 ya que al estar multiplicadas por β_2 y β_4 respectivamente,tendrán un valor de 0 ya que los betas son 0, por lo tanto no aportan o no influyen en el modelo.

3. Pregunta 3

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Si 4 veces $\beta_2 = 3$ veces β_3, $\beta_1 = \beta_5$ y 3 veces $\beta_4 = 7$ veces β_5 por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : 4\beta_2 = 3\beta_3;\ \beta_1 = \beta_5;\ 3\beta_4 = 7\beta_5 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

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

$$L = \begin{bmatrix} 0 & 0 & 4 & -3 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -3 & 7 \end{bmatrix}$$

El modelo reducido está dado por:

$$Y_i = \beta_o + \beta_1 X^*_{1i} + \beta_2 X^*_{2i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 64$$

Donde $X^*_{1i} = X_{5i} + X_{1i} + \frac{7}{3}X_{4i}$ y $X^*_{2i} = \frac{4}{3}X_{3i} + X_{2i}$

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$\begin{aligned} F_0 &= \frac{(SSE(MR) - SSE(MF))/3}{MSE(MF)} \overset{H_0}{\sim} f_{3,58} \\ F_0 &= \frac{(SSE(MR) - (54.7867)/3}{0.944598} \overset{H_0}{\sim} f_{3,58} \end{aligned} \tag{3}$$

4. Pregunta 4

4.1. Supuestos del modelo

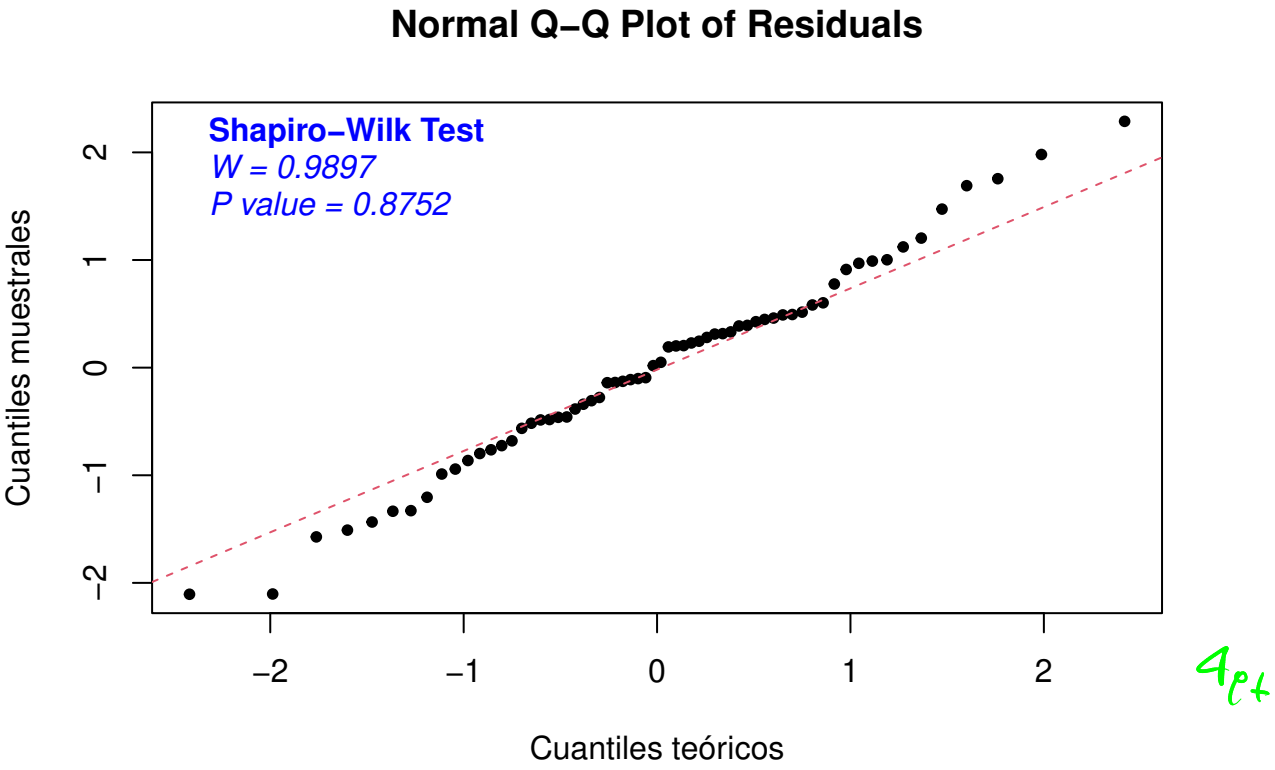
Teniendo en cuenta los supuestos para los errores, los cuales son: media cero, varianza constate, independendientes y siguen una distribución normal. A continuación se procede a probar cada uno de estos supustos por medio de los residuales, los cuales sí son observables.

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se plantea el siguiente juego de hipótesis:

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

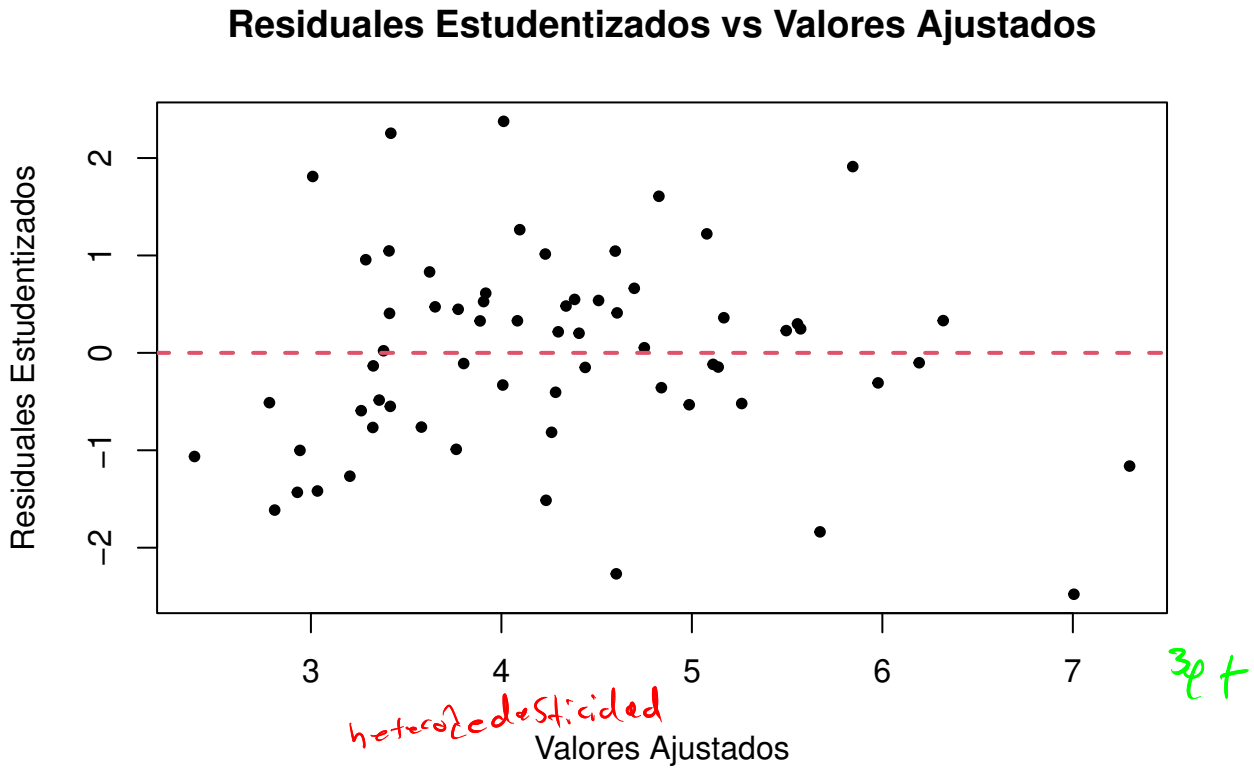
También se usará el criterio del gráfico QQplot y claramente el valor p de la prueba anterior, presentado a continuación:



Al ser el P-valor aproximadamente igual a 0.8752 y teniendo en cuenta que el nivel de significancia es $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, no se rechazaría la hipótesis nula, es decir, los residuales se distribuyen normal según este criterio. Sin embargo, la gráfica de comparación de cuantiles permite ver colas más pesadas y patrones irregulares, al tener más poder el análisis gráfico, se concluye que no hay suficiente información muestral para determinar que los residuales se distribuyen normal. Además de esto, se pueden observar posibles candidatos como datos atípicos debido a los puntos que están alejados considerablemente en los extremos de la recta.

4.1.2. Varianza constante

Ahora, los segundo que se validará es el supuesto de varianza constante



En el gráfico de residuales estudentizados vs. valores ajustados se puede observar que hay un problema serio de homocedasticidad, se evidencia que para los primeros valores del eje horizontal la varianza es pequeña y a medida que aumentan los valores de dicho eje la varianza aumenta consigo, la cual nos da indicios de carencia de ajuste por la forma de U invertida, como se marca en la figura. Además de ello, se evidencia que la nube de puntos está contenida alrededor de cero, por lo tanto se podría decir que el supuesto de media cero se cumple.

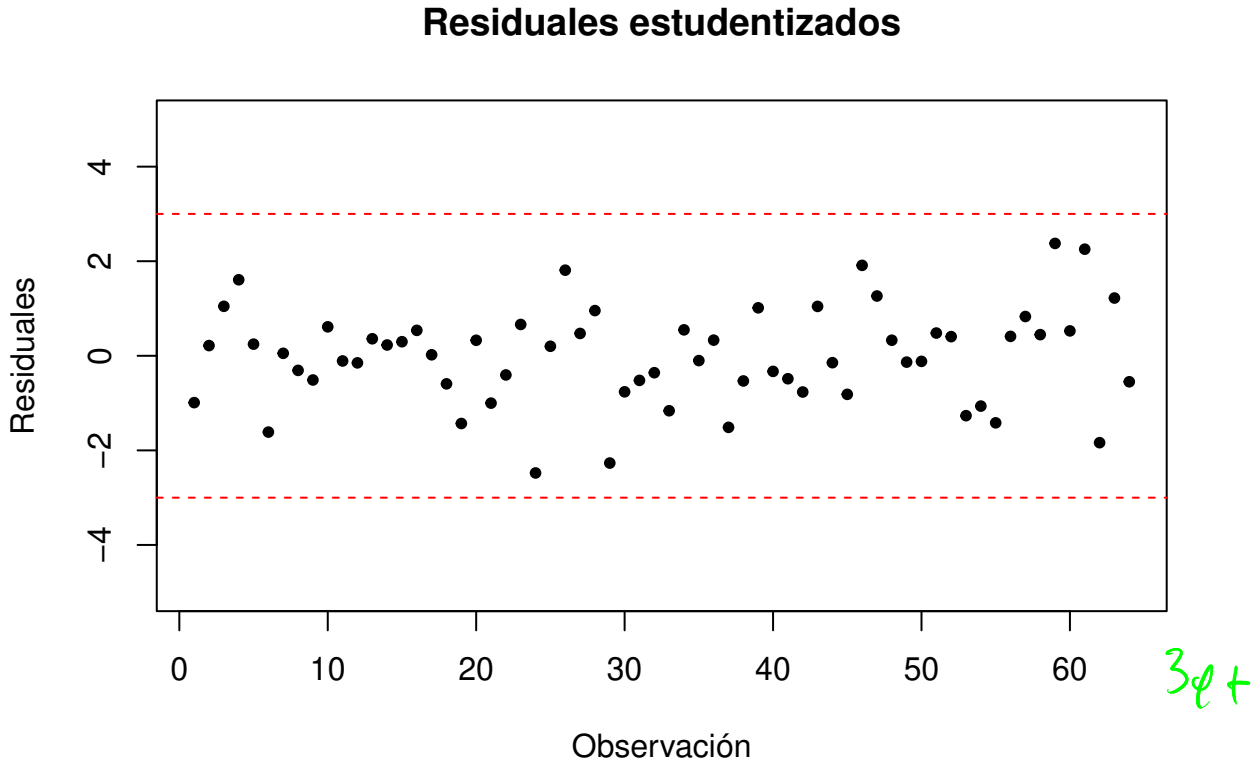
A modo de conclusión, se observan claros problemas sobre la distribución normal de los residuales, problemas de relacionados con heterocedasticidad y aparte de esto, indicios de carencia de ajuste.

4.2. Verificación de las observaciones

Además de la validación de supuestos se debe chequear la presencial de observaciones atípicas, puntos de balanceo y observaciones influyentes.

4.2.1. Observaciones atípicas

Es importante recordar que una observación atípica en la respuesta Y, es aquella que en algún aspecto está separada de los datos y por lo tanto puede afectar los resultados del ajuste del modelo de regresión.



Un e_i grande ($|e_i| > 3$) es indicio de una observacción atípica potencial. Ahora, según la gráfica anterior de residuales estudentizados, no se observan observaciones atípicas potenciales, las cuales deberían salir de la región encerrada por las líneas rojas $(-3, 3)$ y se evidencia que ninguna observación sobre pasa dichos límites.

```
## [1] res.stud  Cooks.D   hii.value Dffits
## <0 rows> (or 0-length row.names)
```

No tener observaciones atípicas es realmente positivo, ya que, tener dichas observaciones en los datos causan niveles de confianza menores a lo esperado.

4.2.2. Puntos de balanceo

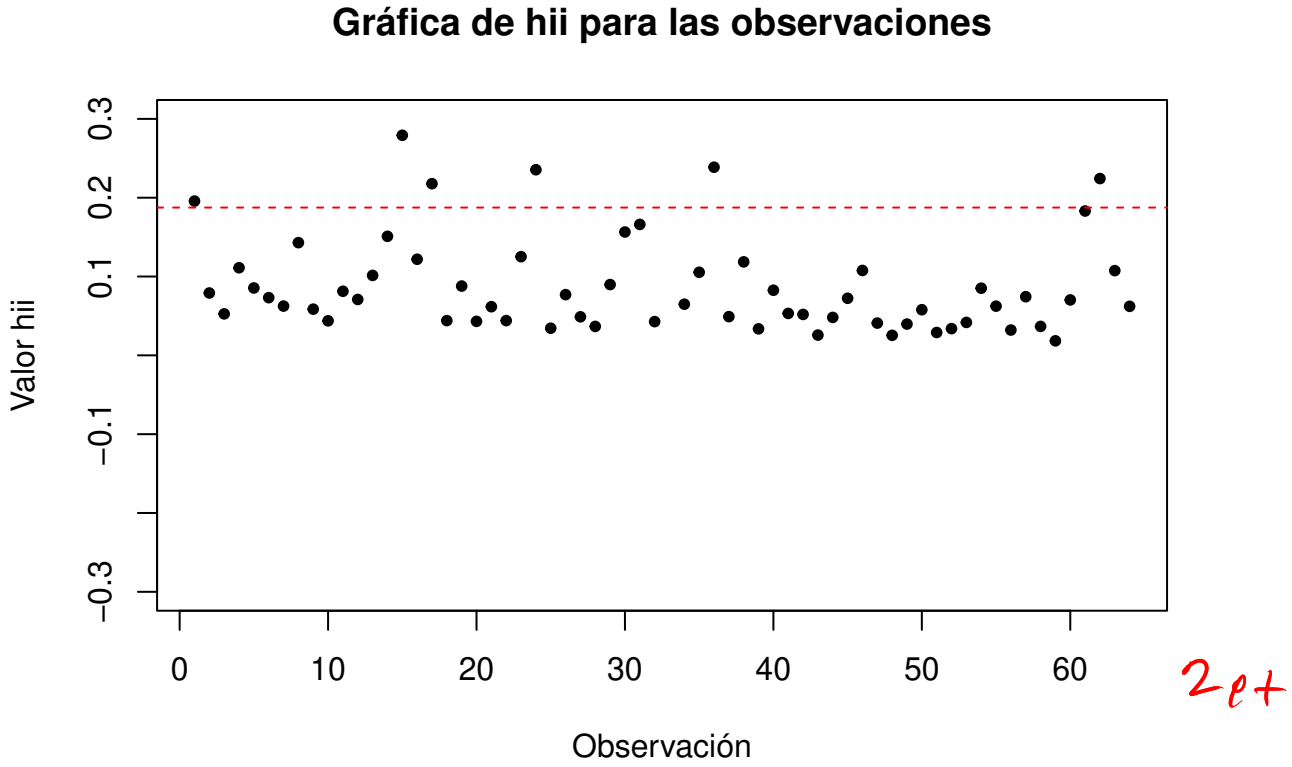
Los puntos de balanceo son observaciones en el espacio de las predictoras, alejada del resto de la muestra y que puede controlar ciertas propiedades del modelo ajustado. Encontrar este tipo de observaciones en los datos, no afecta los coeficientes de regresión estimados, pero sí directamente las estadísticas de resumen como el R^2 y los errores estándar de los coeficientes estimados. Se sabe que un punto es de balanceo si se cumple que $h_{ii} > 2\frac{p}{n}$, para este caso serían los $h_{ii} > 0.1875$ y $0.1875 < 1$.

Sabiendo esto y con el vector de h_{ii} se presenta la siguiente tabla en la que se filtran las observaciones que cumplen que $h_{ii} > 0.1875$:

```
##    res.stud Cooks.D hii.value  Dffits
## 1   -0.9900  0.0397    0.1957 -0.4882
## 15   0.2976  0.0057    0.2792  0.1838
## 17   0.0217  0.0000    0.2178  0.0113
```

```
## 24  -2.4778  0.3152    0.2355 -1.4417
## 33  -1.1621  0.2257    0.5006 -1.1672
## 36   0.3309  0.0057    0.2387  0.1839
## 62  -1.8374  0.1627    0.2243 -1.0092
```

Por lo tanto, se confirma que hay puntos de balanceo en las observaciones 1, 15, 17, 24, 33, 36 y 62. De manera gráfica se ve de la siguiente manera:



Notese que hay un punto en la parte derecha que parece estar tocando la línea roja superior que representa la cota para identificar si la observación es de balanceo o no, dicho punto es incluido en la tabla anteior.

4.2.3. Puntos influenciales

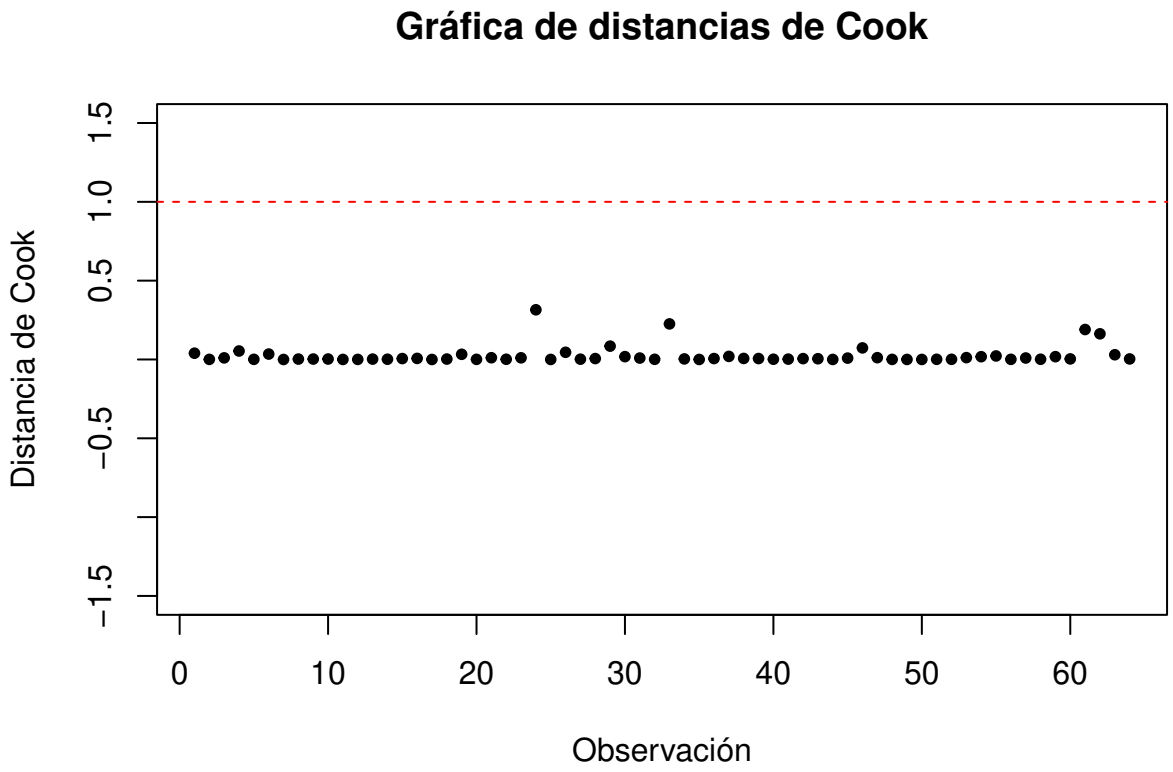
Los puntos incluenciales sí tienen un efecto directo sobre la estimación de los coeficientes de regresión ajustados, es decir, tener una observación influencial arrastra al modelo en su dirección. Esto quiere decir que si un punto es influencial, su exclusión del modelo causa cambios importantes en la ecuación de regresión ajustada.

Así, después de identificar las observaciónes que están alejadas de Y (observaciones atípicas) y con respecto a sus valores en X (puntos de balanceo), evaluamos si éstas son influenciales.

Para probar esto, dichos puntos influenciales se identifican usando las siguientes medidas: (1) distancia de Cook y (2) $DFFITS_i$.

(1)*Cook*: para la distancia de Cook, se dice que una observación es influencial si $D_i > 1$.

Para ello se gráfican los D_i:

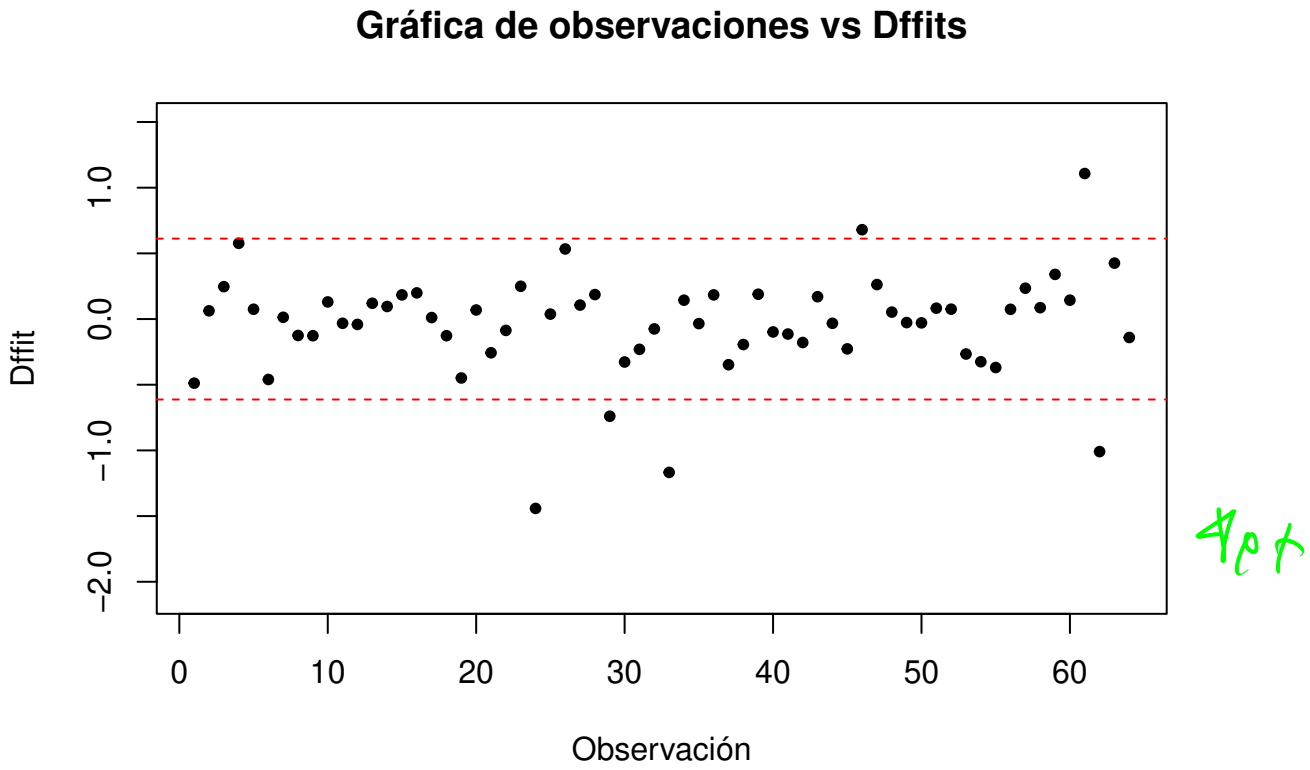


Se observa que según este criterio, no hay puntos influenciales en la muestra.

(2)$DFFITS_i$: para este criterio se dice que una observación es influencial si $|DFFITS_i| > 2\sqrt{(\frac{p}{n})}$, es decir, en este caso sería $|DFFITS_i| > 0.61237$. Con esto, se presenta la siguiente tabla con las observaciónes en las que se cumple el criterio para este método:

Observando en la tabla, según $DFFITS_i$, las observacciones 24, 29, 33, 46, 61 y 62 son influenciales.

Gráficamente:



Se confirma lo mostrado por la tabla anterior, se evidencian seis observaciones influenciales en los datos, mencionadas anteiormente.

4.3. Conclusión

Con lo mostrado en la sección de validación de supuestos de los errores y en la verificación de las observaciones, es importante anotar que:

- El supuesto de normalidad para los errores no se cumple apropiadamente.
- El supuesto de varianza constante para los errores presenta problema, asimismo como una posible carencia de ajuste.
- El supesto de media cero para los errores se cumple mientras que para verificar el supuesto de independencia entre los mismos, se necesita tener el orden temporal (lo cual no se tiene), por lo que no podemos hablar sobre el supuesto de independencia.

Lo anterior se deduce sobre los errores con una muestra grande. Se concluye sobre los errores pero el análisis es hecho sobre los residuales, los cuales sí son observables.

Ahora, sobre la verificación en las observaciones se concluye que:

- No se evidencian puntos atípicos aunque en el gráfico QQplot sobre los residuales se notaron posibles candidatos.
- Se evidencian siete puntos de balanceo según el criterio aplicado, los cuales afectan directamente el R^2 del modelo y los errores estándar de los coeficientes estimados para el modelo.

- No hay presencia de puntos influenciales según el método de la distancia de Cook, pero sí se observan seis puntos influenciales según el método $DFFITS_i$, los cuales afectan directamente las estimaciones de los coeficientes estimados para el modelo de regresión.

Sobre el modelo se concluye que hay problemas generales en los cuales es posible proponer transformaciones para corregir dichos problemas. Asimismo, la presencia de puntos de balanceo (aunque no son muchos), genera desconfianza en el R^2 hallado y se puede pensar que el modelo no está ajustando correctamente; de la misma manera para la estimación de los parámetros ajustados del modelo, es necesario calcular dichos parámetros de nuevo sin los puntos influenciales presentados, y comparar si hay gran diferencia con los hallados.

Válido o no?