Trabajo Final Integrador

Carrera de Especialización en Estadística Matemática

Profesor: Dr. Ricardo Maronna

Alumna: María Soledad Pelliza
(mspelliza@gmail.com)

2018

1. Introducción

Proteger ríos y arroyos mediante el monitoreo de concentraciones químicas y comunidades de algas.

Técnicas inteligentes para monitorear la calidad del agua usando químicos indicadores y población de algas.

Los últimos años se han caracterizado por una preocupación creciente en el impacto que el hombre está teniendo en el medio ambiente. El impacto en el medio ambiente de los residuos tóxicos, de una amplia variedad de procesos de fabricación, es bien conocido. Más recientemente, sin embargo, ha quedado claro que los efectos más sutiles del nivel de nutrientes y los cambios en el equilibrio químico derivados de la escorrentía de tierras agrícolas y el tratamiento de aguas residuales también tienen un efecto grave, pero indirecto, en los estados de ríos, lagos e incluso el mar.

En climas templados de todo el mundo los veranos se caracterizan por numerosos informes de crecimiento excesivo de algas de verano que resulta en una pobre claridad del agua, muertes masivas de peces de río por niveles reducidos de oxígeno y el cierre de instalaciones recreativas de agua a causa de los efectos tóxicos de esta floración anual de algas. La reducción del impacto de estos cambios provocados por el hombre en los niveles de nutrientes del río ha estimulado gran parte de la investigación biológica con el objetivo de identificar las variables cruciales de control químico para los procesos biológicos.

Los datos utilizados en este problema provienen de uno de esos estudios. Durante el estudio de investigación, se tomaron muestras de calidad del agua en diferentes ríos europeos en un período de aproximadamente un año. Estas muestras se analizaron para diversas sustancias químicas, incluyendo: nitrógeno en forma de nitratos, nitritos y amoníaco, fosfato, pH, oxígeno, cloruro. Paralelamente, se tomaron muestras de algas para determinar la distribución de la población de las mismas. Es sabido que la dinámica de la comunidad de algas está determinada por un entorno químico externo con uno o más factores que predominan.

Si bien el análisis químico es barato y se automatiza fácilmente, la parte biológica implica un examen microscópico, requiere mano de obra capacitada y, por lo tanto, es costosa y lenta.

El objetivo será predecir lo mejor posible la abundancia de la primera especie de algas sobre la base de las concentraciones medidas de las sustancias químicas y la información global sobre la temporada en que se tomó la muestra, el tamaño del río y su velocidad de flujo.

Los primeros 11 valores de cada conjunto de datos son la estación, el tamaño del río, la velocidad del fluido y 8 concentraciones químicas que deberían ser relevantes para la distribución de la población de algas. Los últimos valores de cada conjunto de datos son la distribución de diferentes tipos de algas. El valor 0 significa que la frecuencia es muy baja.

2. Análisis exploratorio de los datos

Para comenzar con el análisis exploratorio de los datos defino primero las variables del problema;

Covariables

Categóricas;

E: Estación del año *T*: Tamaño del río *V*: Velocidad del río

Cuantitativas;
cqA, cqB, cqC, cqD, cqE, cqF, cqG, cqH: Cada una de las diferentes concentraciones químicas analizadas en los ríos.

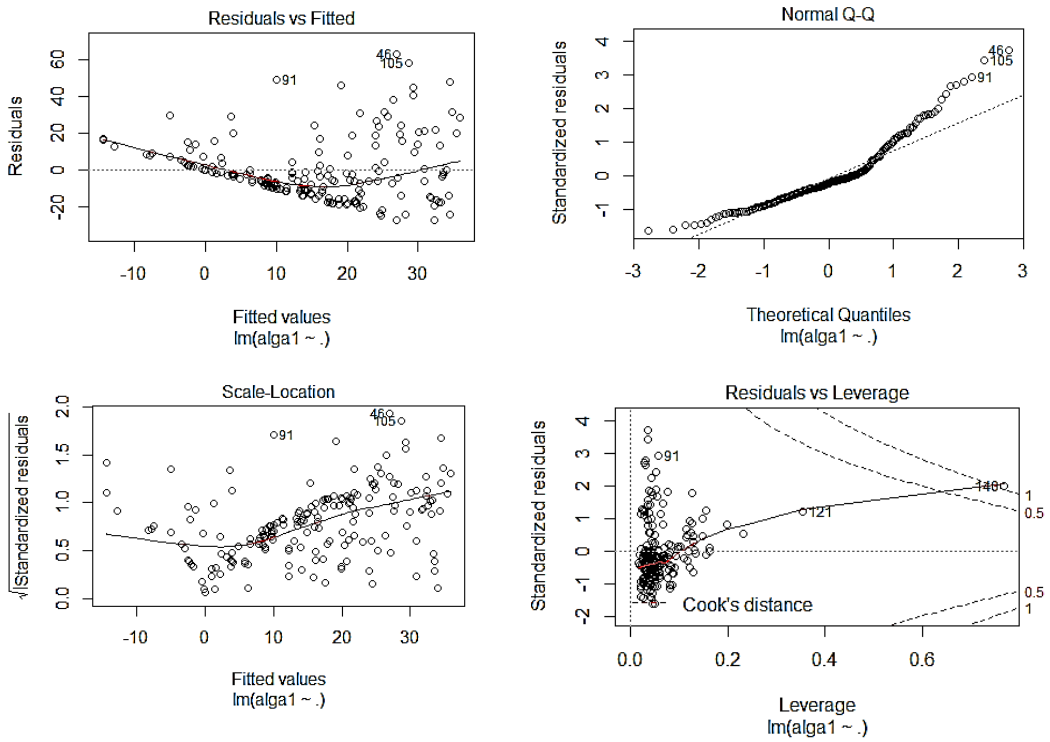
Variable respuesta

alga1: Distribución del alga de tipo 1
(Los restantes tipos de algas no serán analizados en el problema, ya que se pide predecir la cantidad da algas del primer tipo en relación a las covariables mencionadas).

En una primera aproximación a los datos, propongo un modelo lineal con todas las covariables, definiendo a las categóricas como factores.

Obtengo que las únicas covariables significativas son la temperatura del rio y el compuesto químico "D". Los supuestos de homoscedasticidad y normalidad de los residuos no se verifican y se obtiene un valor del error cuadrático medio muy elevado.

Gráfico 1: Verificación de supuestos para el modelo lineal con todas las covariables



En el *Gráfico 1* se observa que los supuestos no se verifican. Los residuos forman una estructura de abanico, los cuantiles no se ajustan a la normal y existe una observación con una distancia de Cook elevada, con lo cual podemos asumir que es un outlier muy influyente (observación número 140).

3. Selección de variables predictoras

- CP de Mallows y R^2 ajustado

Para seleccionar las variables predictoras utilizo el criterio del CP de Mallows y el del R cuadrado ajustado.

Según el criterio del CP de Mallows, debería elegir el modelo cuyo valor de Cp sea pequeño y a la vez esté cerca de la cantidad de covariables seleccionadas más la intercept. Seleccioné la fila del análisis que tiene validez para lo que quiero mostrar: el mejor modelo según este criterio incluye a las covariables Temperatura, CqC, CqD y CqG, además de la inercept. Este modelo, además, es el que posee el valor del R^2 ajustado más cercano a 1:

Tabla 1: Selección de variables según el criterio del CP de Mallows

n	Est	Tem	Vel	CqA	CqB	CqC	CqD	CqE	CqF	CqG	CqH	n+int.	Cp	R2 ajustado
3	0	0	0	1	0	0	0	1	0	1	0	4	16,32	0,24
4	0	1	0	0	0	0	1	1	0	1	0	5	3,76	0,29
4	0	1	0	0	0	1	1	0	0	1	0	5	4,20	0,29
4	0	1	0	0	0	0	1	0	0	1	1	5	5,67	0,28
4	0	1	0	0	1	0	1	0	0	1	0	5	6,43	0,28
4	0	1	0	1	0	0	1	0	0	1	0	5	6,46	0,28
4	0	1	0	0	0	1	0	1	0	1	0	5	6,46	0,28
4	0	1	1	0	0	0	1	0	0	1	0	5	7,08	0,28

n : Cantidad de covariables consideradas para el modelo

n+int. : Cantidad de covariables consideradas para el modelo más la inercept.

- Stepwise

A continuación se muestran los mejores modelos para cada cantidad de covariables seleccionadas.

E2, E3 y E4 son variables *dummys* que indican la consideración de las estaciones 2, 3 y 4. Análogamente, T2 y T3 son variables *dummys* que indican la consideración de las temperaturas 2 y 3, y V2 y V3 son variables *dummys* que indican la consideración de las velocidades 2 y 3.

Tabla 2: Selección de variables aplicando stepwise

n	fac_E2	fac_E3	fac_E4	fac_T2	fac_T3	fac_V2	fac_V3	cqA	cqB	cqC	cqD	cqE	cqF	cqG	cqH	R2 adj.
1														X		0,2056
2											X			X		0,2330
3				X	X									X		0,2565
4				X	X						X			X		0,2807
5				X	X						X	X		X		0,2885
6				X	X					X	X	X		X		0,2944
7				X	X					X	X	X		X	X	0,2973
8				X	X	X				X	X	X		X	X	0,2977
9	X			X	X	X				X	X	X		X	X	0,2977
10	X	X		X	X	X				X	X	X		X	X	0,2968
11	X	X		X	X	X			X	X	X	X		X	X	0,2937
12	X	X		X	X	X		X	X	X	X	X		X	X	0,2905
13	X	X		X	X	X	X	X	X	X	X	X		X	X	0,2873
14	X	X	X	X	X	X	X	X	X	X	X	X		X	X	0,2834
15	X	X	X	X	X	X	X	X	X	X	X	X		X	X	0,2795

Considerando el valor del R cuadrado ajustado, el mejor modelo incluiría 9 covariables. El valor del R cuadrado ajustado en este caso sería 0,297747:

Gráfico 2: Cantidad de predictores vs. R^2 ajustado

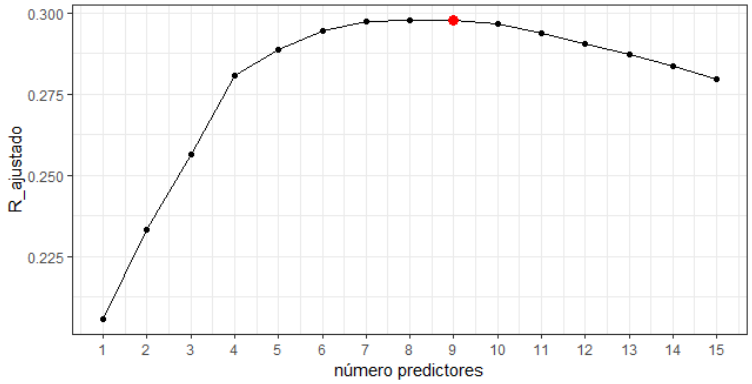


Tabla 3: Coeficientes de las 9 variables predictoras seleccionadas con el método stepwise

(Intercept)	fac_E2	fac_T2	fac_T3	fac_V2	cqC	cqD	cqE	cqG	cqH
37.071	-2.994	-6.085	-9.741	-2.878	-0.040	-0.768	-0.013	-0.049	-0.099

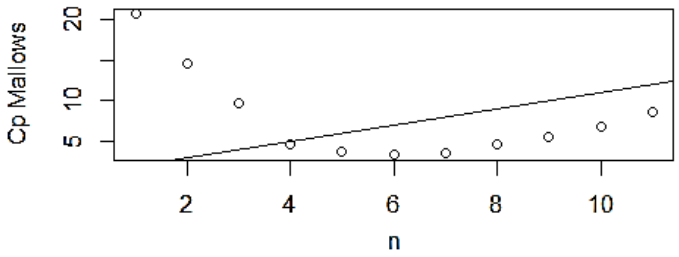
Sin embargo, como puede observarse en el *Gráfico 2*, el valor del R^2 ajustado en principio aumenta cuando se aumenta la cantidad de predictores pero entra en una meseta alrededor del valor máximo a partir de considerar cuatro predictores.
Si bien el modelo con mayor R^2 ajustado es el formado por 9 predictores, la mejora conseguida a partir de 4 predictores es mínima, pasando de 0,2807 a 0,297747.
Si consideramos el principio de parsimonia entonces, podríamos considerar sólo los predictores fac_T2, fac_T3, cqD y cqG.

Tabla 4: Coeficientes de las 4 variables predictoras seleccionadas con el método stepwise, si aplicamos el principio de parsimonia

(Intercept)	fac_T2	fac_T3	cqD	cqG
34.49809452	-8.7047	-12.0790	-0.8999	-0.0645

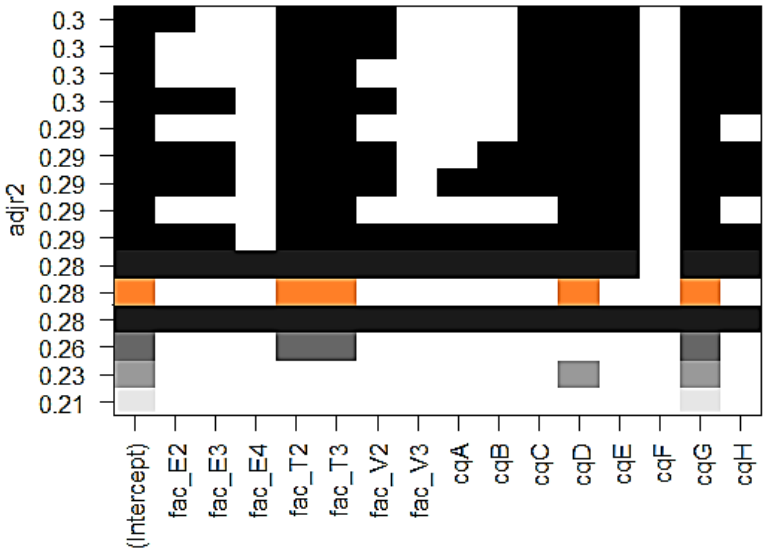
Si consideramos el Cp de Mallows para cada uno de los modelos considerados, observamos que el mejor modelo es el que considera cuatro variables predictoras, ya que en este caso el valor de ese coeficiente es el más similar a la cantidad de variables predictoras seleccionadas;

Gráfico 3: Cantidad de predictores vs. Cp de Mallows



Utilizando una muestra de los datos como prueba y otra como validación, ponemos a prueba los modelos propuestos y obtenemos nuevamente que el modelo más adecuado si tenemos en cuenta que su valor del R^2 sea el más próximo a 1 y a su vez no sea parsimonioso, entonces el modelo elegido será el que considera las cuatro variables predictoras mencionadas anteriormente (marcado en naranja);

Gráfico 4: Predictores seleccionados según el R2 poniendo a prueba los modelos con una muestra de entrenamiento



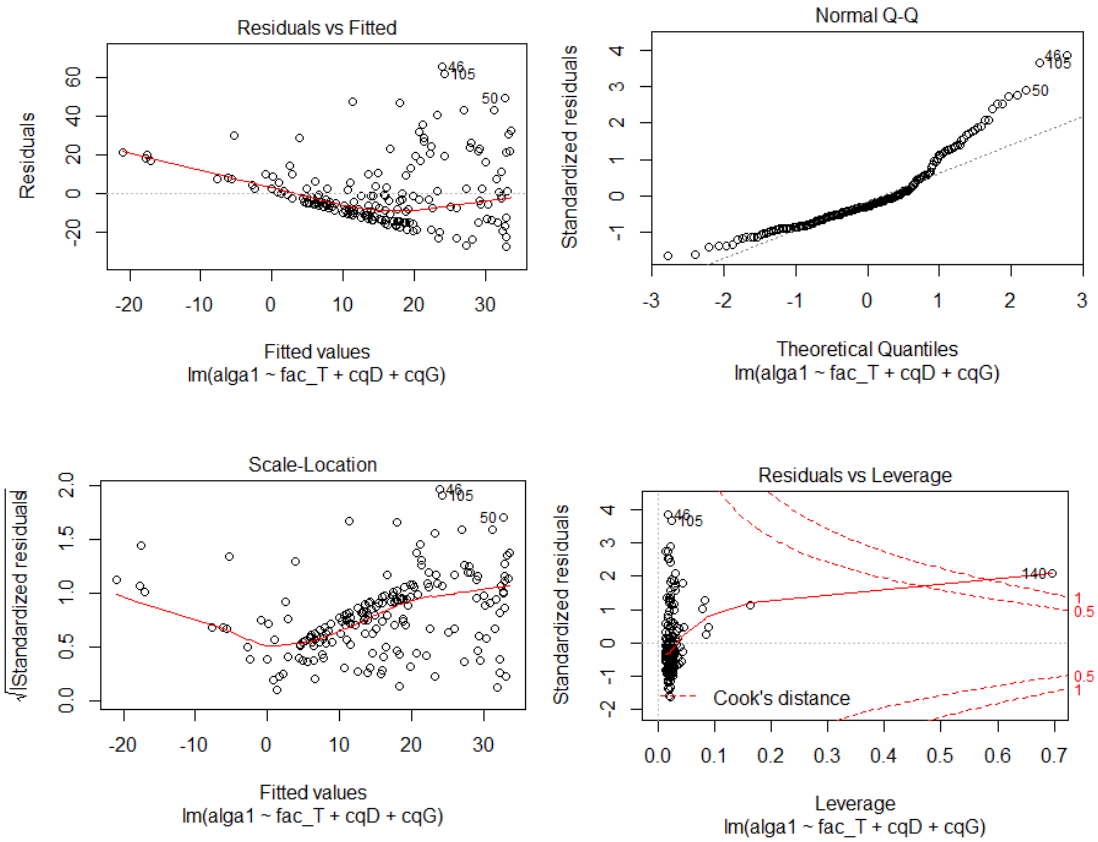
En el Gráfico 4 se observa en naranja el modelo con valor de R2 más próximo al máximo (0.30), que es al mismo tiempo el menos parimonioso puesto que considera sólo cuatro covariables. Los demás modelos incorporan covariables sin mejorar notablemente el valor del R2, o bien tienen un valor de R2 menor.

Proponemos un modelo lineal solamente con esas variables y todas resultan significativas al 0,01%. El modelo resulta;

$$\hat{y} = 34{,}5 - 8{,}7*\text{fac\_T2} - 12{,}08*\text{fac\_T3} - 0{,}9*\text{cqD} - 0{,}06*\text{cqG}$$

Verificamos los supuestos para el modelo propuesto:

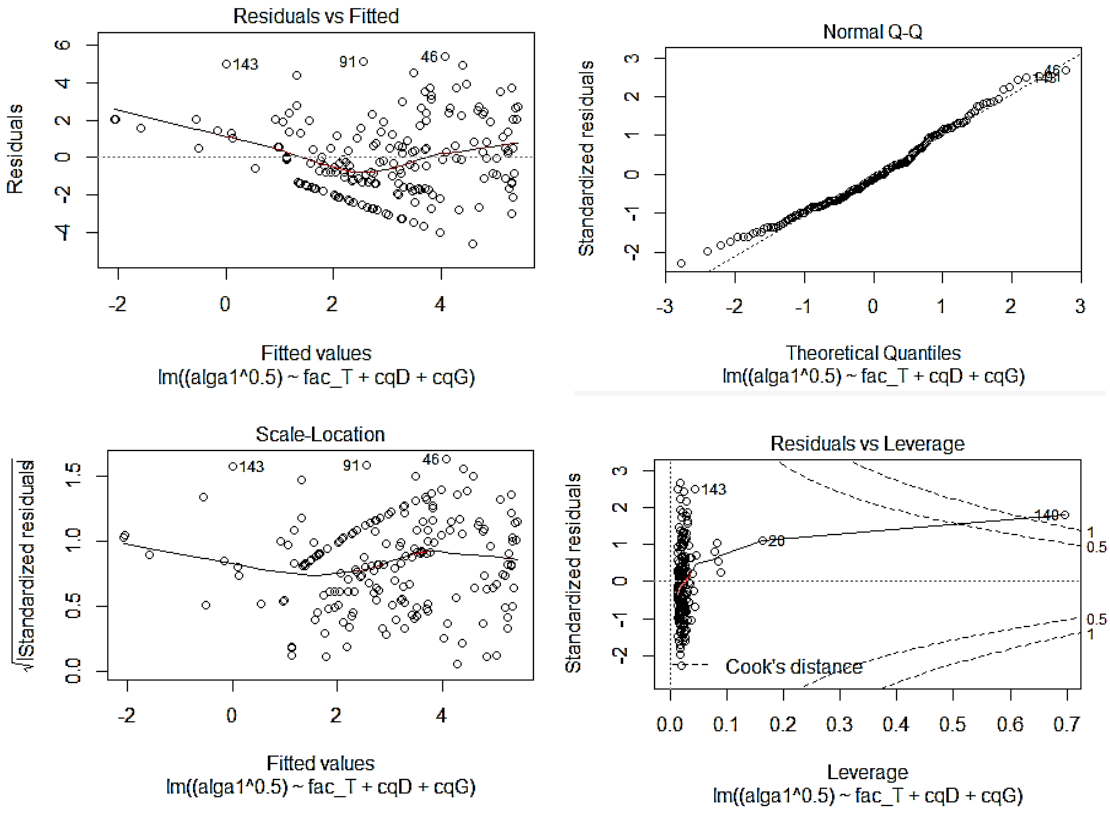
Gráfico 5: Verificación de supuestos para el modelo lineal con las covariables fac_T2, fac_T3, cqD y cqG



En el *Gráfico 5* podemos observar lo siguiente:
- Los residuos forman un abanico. Esto indica que hay heteroscedasticidad sistémica y por lo tanto no podremos evitarla por medio de transformaciones.
- El qqplot indica que no hay normalidad en los residuos. Esto lo verificamos también realizando un test de Shapiro-Wilks, en el cual obtuvimos que se rechaza la normalidad.
- La observación 140 sigue siendo muy influyente según lo indica el gráfico de la distancia de Cook.

Buscamos una transformación de BoxCox adecuada y, a pesar de que no corrige la heteroscedasticidad, sí mejora la normalidad de los residuos. Para lambda=0.5 obtenemos;

Gráfico 6: Verificación de supuestos para el modelo lineal con las covariables fac_T2, fac_T3, cqD y cqG, con transformación de BoxCox con lambda=0.5



En el *Gráfico 6* podemos observar lo siguiente:
- Se mantiene la heteroscedasticidad.
- Se redujeron los residuos.
- En el qqplot se observa que se mejora la normalidad.
- La observación 140 sigue siendo muy influyente según lo indica el gráfico de la distancia de Cook.

Con esta transformación, además, logramos llevar el R2 a 0.34, que si bien sigue siendo un valor bajo, es el máximo que pudimos hallar para todos los modelos probados.

El modelo propuesto entonces resulta:

$$\widehat{y^{1/2}} = 5.53 - 0.94*\text{fac\_T2} - 1{,}60*\text{fac\_T3} - 0{,}10*\text{cqD} - 0{,}01*\text{cqG}$$

Al poner a prueba la selección de variables y la transformación elegida mediante distintas muestras de la base de datos, obtengo que los coeficientes se mantienen estables. Por otra parte los valores predichos me resultan satisfactorios.

Los coeficientes se modifican algunos décimos cuando ajusto el modelo en aquellas muestras que no contienen al outlier. Elijo ajustar el modelo con la base de datos sin incluir a ese outlier, puesto que de esa manera considero que resulta más representativo.

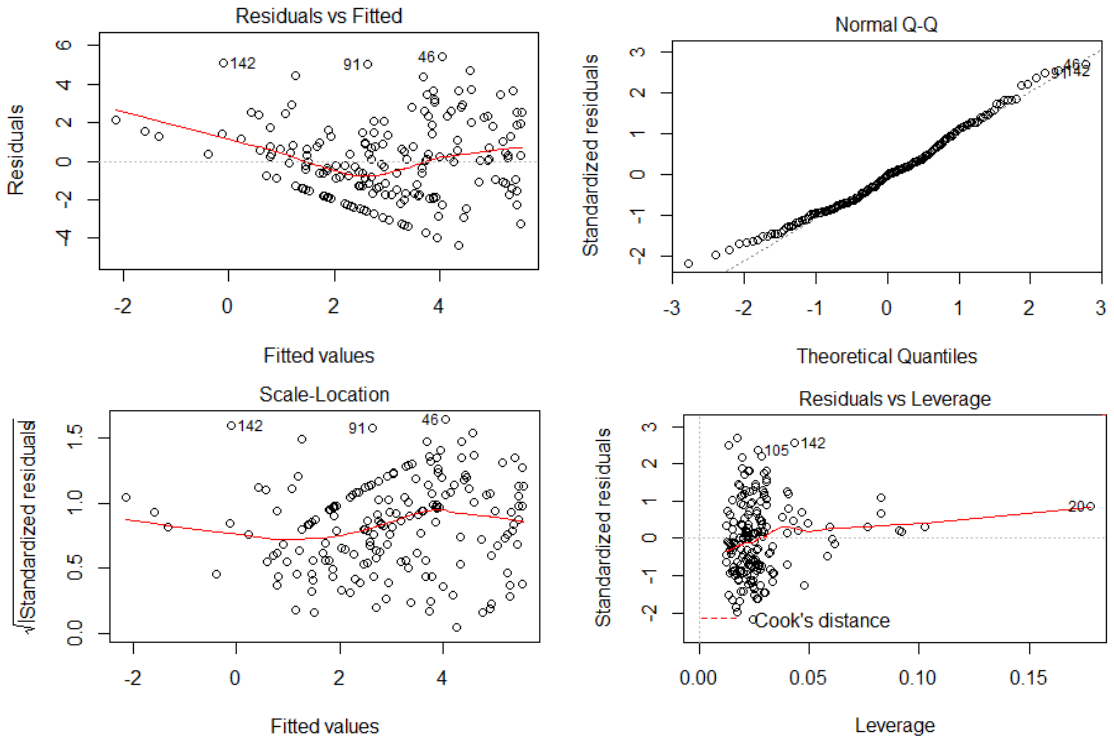
Cabe aclarar que si bien el outlier se presenta en todos los modelos presentados anteriormente, si analizamos los datos que corresponden a esa observación, veremos que pre-

senta una drástica diferencia en la concentración del compuesto químico "G", con respecto a las demás observaciones; todos los datos de la muestra presentan una concentración de este químico inferior a 10,5 mientras que para esa observación la concentración es de 45,65. Considerando que esta covariable, nombrada cqG, resulta significativa en todos los modelos, aún si no tenemos en cuenta esa observación, decido ajustar el modelo sin tenerla en cuenta.

Por lo tanto, finalmente el modelo elegido resulta:

$$\widehat{y^{1/2}} = 5.748 - 0.9*\text{fac\_T2} - 1{,}64*\text{fac\_T3} - 0{,}21*\text{cqD} - 0{,}008*\text{cqG}$$

Gráfico 7: Verificación de supuestos para el modelo lineal con las covariables fac_T2, fac_T3, cqD y cqG, con transformación de BoxCox con lambda=0.5, sin considerar el outlier



Si comparamos el *Gráfico 7* con el *Gráfico 6* sólo observamos que se reducen las distancias de Cook. Por otra parte este modelo logra llevar el valor de R^2 a 0.35.