XIV SEMINARIO NACIONAL
DE HIDRAULICA E HIDROLOGIA

ANÁLISIS DEL AJUSTE DE MODELOS DE REGRESIÓN EN HIDROLOGIA

Ricardo A. Smith Q. y Claudia Campuzano
Posgrado en Aprovechamiento de los Recursos Hidráulicos
Facultad de Minas, Universidad Nacional de Colombia, Medellín
cpcampuz@andromeda.unalmed.edu.co

RESUMEN

Los modelos de regresión son unos de los modelos más utilizados en hidrología. Se usan de manera intensiva en procedimientos de regionalización, en representaciones de funciones, en relleno de información faltante, en modelamiento de variables hidrológicas y en muchos otros casos. Estos modelos se usan en muchos casos sin realizar un análisis completo sobre su ajuste, y en muchos casos violando las suposiciones básicas de su desarrollo teórico. Esta situación puede llevar al uso de modelos inadecuados con las consecuentes consecuencias sobre los resultados que se obtengan con ellos. Se presenta en este trabajo los aspectos que se deben revisar en el análisis del ajuste de modelos de regresión en hidrología. Aspectos como la recionalidad física del modelo, la multicoliniaridad, los análisis residuales, su capacidad predictiva y la presencia de observaciones influyentes son revisadas en este trabajo. Se presenta luego un caso de aplicación con un análisis detallado sobre la bondad del ajuste del modelo de regresión. Finalmente se presentan algunas conclusiones y recomendaciones.

ABSTRACT

Regression models are widely used in hydrology. They are used for regionalization procedures, filling in missing data, in modeling of hydrologic variables and in many other applications. These models are used in many cases without a detailed goodness of fit analysis, even violating the main assumptions made for the developing of the regression models. This situation can lead to unacceptable results. In this work the main aspects that must be reviewed in a goodness of fit analysis of regression models are discussed. Aspects such as the model rationality, the multicolinearity problem, residual analysis, prediction capability, and the presence of influential observations, are reviewed in this work. A example with a detailed discussion of the goodness of fit analysis of a regression model is presented. Finally some conclusions and recommendations are presented.

1. INTRODUCCION

Los modelos de regresión lineal son unos de los modelos más utilizados en hidrología para establecer relaciones lineales entre dos variables o entre una variable y un grupo de variables. Estos modelos son los modelos más utilizados en hidrología para diferentes propósitos y se usan incluso con diferentes formas de transformación de las variables pero manteniendo la relación lineal en los parámetros del modelo. Ejemplos de usos de estos modelos en hidrología se dan en aspectos como: predicción hidrológica, relleno y extensión de datos faltantes, modelamiento hidrológico, análisis de relaciones entre variables, y muchos más.

Estos modelos se usan de manera extensiva y en mucha de sus aplicaciones no se hace una adecuada evaluación de la bondad de ajuste del modelo. Casi siempre la única prueba que se hace para aceptar o rechazar el modelo es la prueba de significancia del coeficiente de correlación. Aunque esta es una de las herramientas que debe ser utilizada en la evaluación del modelo, no debe usarse sola ya que puede existir correlación significativa entre dos variables que no estén relacionadas. Es entonces muy importante que se haga una adecuada evaluación de los modelos de regresión propuestos de tal manera que no se utilicen finalmente modelos inadecuados. La herramienta de análisis de regresión y correlación puede ser muy útil en hidrología pero mal utilizada puede llevar a resultados absurdos o muy costosos.

Cuando un modelo de regresión lineal se ajusta a un grupo de variables se debe entonces evaluar la bondad del ajuste del modelo de una manera completa. Para este propósito hay varios aspectos del ajuste del modelo que deben ser analizados tales como:

- Racionalidad del modelo ajustado
- Capacidad de Predicción del Modelo
- Análisis Residual
- Presencia de observaciones influyentes

La racionalidad del modelo trata de evaluar el significado físico de las relaciones establecidas en el mismo. La capacidad de predicción del modelo intenta analizar que tan bueno es el modelo prediciendo las variables dependientes. El análisis residual comprueba que los residuos calculados usando el modelo propuesto cumplan con las suposiciones hechas sobre ellos. La presencia de observaciones influyentes trata de detectar la presencia de observaciones atípicas que puedan tener una influencia significativa en la estimación del modelo y en las pruebas de hipótesis sobre los parámetros. Cada uno de estos aspectos se discuten a continuación en algún detalle.

2. RACIONALIDAD DEL MODELO AJUSTADO

Cuando un modelo de regresión se ajusta a un grupo de variables los parámetros estimados asociados a cada variable indican de que manera una variable independiente particular afecta la variable dependiente. El parámetro estimado asociado a una variable independiente específica representa la cantidad de cambio en la variable dependiente debido a una unidad de cambio en esa variable independiente cuando todas las otras variables independientes se mantienen fijas. Este cambio puede ser positivo (un aumento o disminución en la variable independiente significa un aumento o disminución en la variable

dependiente) o negativo (un aumento en la variable independiente significa una disminución en la variable dependiente, o una disminución en la variable independiente significa un aumento en la variable dependiente). El parámetro intercepto representa el valor de la variable dependiente cuando todas las variables independientes son cero. Cuando se estima un modelo de regresión, el analista debe inferir si los parámetros estimados reflejan adecuadamente el efecto de las variables independientes sobre la variable dependiente. Éste es uno de los análisis más importantes a realizarse cuando se ajusta un modelo de regresión. Los parámetros deben tener significado físico además de un significado estadístico.

La racionalidad estadística del modelo se garantiza cuando se usa alguno de los procedimientos de selección de modelos de regresión. Estos procedimientos definen el mejor modelo de regresión en un sentido estadístico. En general, todos los procedimientos de selección de modelos de regresión disponibles en la literatura definen el mismo modelo, seleccionan el mismo conjunto de variables independientes, sin embargo, pueden presentarse casos en donde algunos de ellos definen modelos diferentes. En este caso el analista tendría que definir cual modelo debe usarse y esto podría hacerse analizando la justificación física de las variables independientes en el modelo.

Si los parámetros estimados reflejan adecuadamente el efecto de las variables independientes sobre la variable dependiente, y si el signo asociado con el parámetro refleja adecuadamente la dirección de ese efecto, se puede concluir que el modelo tiene un significado físico adecuado.

El analista debe tener cuidado cuando funciones de variables derivadas se usan en el modelamiento ya que se podrían incluir relaciones espurias en el modelo de regresión. Un modelo de regresión múltiple tiene relaciones espurias o multi-colinealidad (Helsel y Hirsch, 1992) cuando por lo menos una de las variables independientes esta relacionada con una o más de las otras variables independientes. La presencia de relaciones espurias o multi-colineales en el modelo puede llevar a que un modelo de regresión no sea adecuado.

Una prueba que puede usarse para probar multi-colinealidad es el factor de inflación de la varianza (VIF) dado como (Helsel y Hirsch, 1992):

$$VIF_j = \frac{1}{(1-R_j^2)} \qquad (1)$$

en donde R_j representa el coeficiente de correlación múltiple entre la j-ava variable independiente y todas las otras variables independientes. Si $R_j = 0$ entonces $VIF_j = 1$ y éste es el caso ideal. Cuando $VIF_j < 10$, la multi-colinearidad en el modelo es aceptable, pero cuando $VIF_j > 10$ ésta no es aceptable y el modelo de regresión tiene que ser redefinido, eliminando variables multi-colineales o espurias.

3. CAPACIDAD DE PREDICCIÓN DEL MODELO

Una vez que se define un modelo de regresión (la racionalidad física y estadística del modelo ha sido analizada), se está interesado en analizar que tan bueno es el modelo prediciendo las variables dependientes. Para evaluar la capacidad de predicción del modelo la serie de residuos o residual pueden definirse como:

$$\hat{\epsilon}_t = Y_t - \hat{Y}_t \quad , \quad t = 1,.....,N \qquad (2)$$

en donde Y_t representa las observaciones de la variable dependiente, $\hat{Y}_t$ son los valores estimados de la variable dependiente usando el modelo de regresión y las observaciones de las variables independientes, $\hat{\epsilon}_t$ es el error estimado o la observación t de la serie residual, y N es el número de observaciones. En el caso de un modelo de regresión lineal múltiple, el error o la serie residual puede determinarse como:

$$\begin{aligned} \hat{Y}_t &= \hat{a} + \hat{b}_1 x_t^{(1)} + \hat{b}_2 x_t^{(2)} + ... + \hat{b}_m x_t^{(m)} \\ \hat{\epsilon}_t &= Y_t - (\hat{a} + \hat{b}_1 x_t^{(1)} + \hat{b}_2 x_t^{(2)} + ... + \hat{b}_m x_t^{(m)}) \\ \hat{\epsilon}_t &= Y_t - \hat{a} - \hat{b}_1 x_t^{(1)} - \hat{b}_2 x_t^{(2)} - ..... - \hat{b}_m x_t^{(m)}) \quad , t = 1,.....,N \end{aligned} \qquad (3)$$

donde $\hat{a}$ y $\hat{b}_j$, j = 1,..., m son los parámetros estimados de la regresión, y m es el número de variables independientes.

La serie residual puede usarse para evaluar que tan bueno es el modelo con el que se está prediciendo la variable dependiente. Por ejemplo, se espera que la serie residual tenga media cero con una variabilidad aleatoria alrededor de ella. La media residual puede calcularse como:

$$\hat{\mu}_\epsilon = \frac{1}{N}\sum_{t=1}^{N} \hat{\epsilon}_t = \frac{1}{N}\sum_{t=1}^{N} (Y_t - \hat{Y}_t) \qquad (4)$$

Si la media residual calculada da un valor positivo o negativo esta puede ser una indicación de un sesgo en el modelo. En este caso puede pasar que el modelo este sistemáticamente sobrestimando o subestimando la variable dependiente. Si el analista quiere puede probar la hipótesis nula de media cero o calcular los límites de confianza en la media. Si la hipótesis nula de media cero se acepta o los límites de confianza calculados incluyen el valor cero, puede concluirse que la media residual puede considerarse igual a cero.

La varianza residual $\hat{\sigma}_\epsilon^2$ puede calcularse como

$$\hat{\sigma}_\epsilon^2 = \frac{1}{N-p}\sum_{t=1}^{N} (Y_t - \hat{Y}_t) = \frac{1}{N-p}\sum_{t=1}^{N} \hat{\epsilon}_t^2 \qquad (5)$$

donde p representa el número de parámetros en el modelo de regresión. La desviación estándar residual $\hat{\sigma}_\epsilon$ se denomina como el error estándar estimado. El error estándar

estimado es entonces la medida de la dispersión o variabilidad del valor predecido $\hat{Y}_t$ alrededor del valor observado Y_t. Lo que se quiere es que $\hat{\sigma}_{\epsilon}$ sea tan pequeño como sea posible. En el caso de predicción perfecta, es decir, $\hat{Y}_t = Y_t$, t=1,..., N entonces $\hat{\sigma}_{\epsilon} = 0$. Si $\hat{Y}_t = \hat{\mu}_Y$, t=1,... ,N y p=1, entonces $\hat{\sigma}_{\epsilon} = \hat{\sigma}_Y$, y el mejor modelo será usar el promedio o la media para hacer todas las predicciones. La relación $\hat{\sigma}_{\epsilon} / \hat{\sigma}_Y$ da una indicación de lo adecuado de las predicciones con el modelo de regresión cuando se compara con el modelo de valor promedio. Valores de $\hat{\sigma}_{\epsilon} / \hat{\sigma}_Y$ menores que uno indican una mejoría con el uso del modelo de regresión. Valores pequeños de $\hat{\sigma}_{\epsilon}$ indican una buena capacidad de predicción.

El coeficiente de determinación R^2 ha sido propuesto para analizar la cantidad de variabilidad en los datos explicado por el modelo de regresión. Un estimador de R^2 es el cuadrado del coeficiente de correlación. R^2 puede tomar valores entre cero y uno ($0 < R^2 < 1$) y mientras más cercano a uno es mayor la cantidad de variabilidad en los datos explicados por el modelo de regresión. Cuando se usan transformaciones y varios modelos de variables transformadas se comparan, el uso de R^2 para seleccionar el modelo podría ser engañoso.

4. ANÁLISIS RESIDUAL

Se asume que la serie residual estimada $\hat{\epsilon}_t$, t=1,..., N es una serie aleatoria normalmente distribuida con varianza constante. Se puede entonces verificar el ajuste del modelo probando si la serie residual es independiente y normalmente distribuida con varianza constante. Para este propósito, pueden construirse varias gráficas y también pueden realizarse varias pruebas. Se puede inicialmente realizar un análisis gráfico sobre el comportamiento de la serie residual. Algunos de las gráficas que se pueden hacer son:

- Residuos contra la variable dependiente predecida
- Residuos contra tiempo y espacio
- Residuos contra los residuos rezagados
- Residuos contra las variables independientes
- Características periódicas residuales
- Gráfica de probabilidad Normal de los residuos

El análisis gráfico de los residuos puede dar una indicación de qué esperar en el análisis confirmatorio basado en las pruebas estadísticas de independencia, normalidad, y varianza constante de los residuos. Por ejemplo, la gráfica de los residuos contra las variables independientes podría indicar la necesidad de transformar la variable independiente si la gráfica no sigue una línea recta (Helsel y Hirsch, 1992).

Probar que los residuos vinieron de una población con varianza constante es más complejo. El lector interesado puede consultar el artículo de Hipel et al. (1977). Si la hipótesis nula de que los residuos son independientes se rechaza, significa que el modelo de regresión falla

para representar adecuadamente la variabilidad en la variable dependiente. Uno puede corregir este problema introduciendo variables independientes adicionales en el modelo o transformando algunas de las variables independientes. Si los residuos son autocorrelacionados, los datos disponibles podrían ser reestructurados agrupándolos (Helsel y Hirsch, 192). Si la hipótesis nula de que los residuos son normalmente distribuídos se rechaza, generalmente una transformación simple de la serie puede corregir el problema. En este caso, uno debe ser consciente que el modelo de regresión está en el dominio transformado así las inferencias hechas con tal modelo solo se aplican a las variables transformadas. Si la hipótesis nula de que los residuos vinieron de una población con varianza constante se rechaza uno puede corregir el problema usando un procedimiento de regresión tipo Mínimos Cuadrados Generalizados (GLS) o transformando las observaciones de la muestra. En el último caso uno debe ser consciente que el modelo se deriva en el dominio transformado. Otra solución puede ser incorporar variables adicionales en el modelo de regresión.

Si los residuos muestran una variación periódica en la media y/o en la varianza, pueden necesitarse variables independientes adicionales o las observaciones originales deben ser "desestacionalizadas" para construir el modelo de regresión.

5. PRESENCIA DE OBSERVACIONES INFLUYENTES

Los valores anormalmente extremos o las observaciones atípicas pueden tener una influencia significativa en la estimación del modelo de regresión y las pruebas de hipótesis sobre los parámetros del mismo. Estas observaciones necesitan ser identificadas y removidas de los datos de la muestra. La gráfica de Serie de tiempo de la variable dependiente o las gráficas de la variable dependiente contra la variable independiente pueden ayudar a identificar esos valores anormales en muestra. Cuando hay dudas sobre la influencia de un dato específico en la estimación del modelo de regresión, uno puede estimar al modelo de regresión con y sin ese dato sospechoso y comparar los resultados. Si los modelos son similares uno puede concluir que el dato específico no tiene influencia en la estimación del modelo de regresión. Para realizar una comparación más confiable, los límites de confianza sobre los parámetros sin usar el dato anormal pueden obtenerse. Si los parámetros estimados incluyendo el dato anormal caen dentro de los límites, uno puede concluir que los dos grupos de parámetros son similares y que el dato anormal no tiene influencia en la estimación del modelo de regresión.

Varios procedimientos para la detección de observaciones influyentes se presentan brevemente a continuación. Las pruebas presentadas aquí son todas relaciondas con los modelos de regresión lineal. Se presenta una diferencia entre las observaciones influyentes y los outliers. Los outliers podrían ser o no observaciones influyentes, mientras que una observación influyente siempre es un outlier.

Outliers en la dirección de X: Influencia. La influencia es una medida de un outlier en la dirección X y es dada por el i^{th} término diagonal de la matriz $X(X'X)^{-1} X'$. Para un modelo de regresión lineal simple está dada por (Helsel y Hirsch, 1992)

$$h_i = \frac{1}{N} + \frac{(x_i - \hat{\mu}_x)^2}{\sum_{i=1}^{N}(x_i - \hat{\mu}_x)^2} \quad (6)$$

Un punto de influencia alta es uno donde (Helsel y Hirsch, 1992; Kottegoda y Rosso, 1997)

$$h_i \geq 2.5\frac{p}{N} \quad (7)$$

donde p es el número de parámetros en el modelo de regresión y N es el tamaño de la muestra. Un punto de influencia alta podría ejercer una influencia importante en las estimaciones de los parámetros del modelo de regresión o en las pruebas de hipótesis sobre el modelo de regresión.

Distancia de Cook. Ésta es una medida ampliamente usada para detectar observaciones influyentes. Usando los residuos $\hat{\epsilon}_t$ y la serie de influencia h_i, la serie de distancia de Cook puede calcularse como (Helsel y Hirsch, 1992; Kottegoda y Rosso, 1997)

$$D_i = \frac{\hat{\epsilon}_i^2 \, h_i}{p(1-h_i^2)\hat{\sigma}_{\epsilon}^2} \quad (8)$$

donde p es el número de parámetros del modelo de regresión. Se considera que la i^{th} observación tuvo una influencia alta en el modelo de regresión si (Helsel y Hirsch, 1992)

$$D_i > f_{0.10}(p+1,\ N-p) \quad (9)$$

donde $f_{0.10}$ (p+1, N-p) es el 0.10 quantil de la distribución F con p+1 y N-p grados de libertad.

Residuos Studentized. La serie residual modificada $\hat{\epsilon}_i'$ puede determinarse usando la serie de influencia h_i como (Helsel y Hirsch, 1992):

$$\epsilon_i' = \frac{\epsilon_i}{1-h_i} \quad (10)$$

y su varianza puede calcularse como (Helsel y Hirsch, 1992):

$$\hat{\sigma}_{\epsilon}^{2}(i) = \frac{(N-p)\hat{\sigma}_{\epsilon}^{2} - \frac{(\epsilon_i')^2}{(1-h_i)}}{N-p-1} \qquad (11)$$

y la serie residual studentized se obtiene como (Helsel y Hirsch, 1992):

$$\xi_i = \frac{\epsilon_i}{\hat{\sigma}_{\epsilon'}(i)\sqrt{1-h_i}} = \epsilon_i' \frac{\sqrt{1-h_1}}{\hat{\sigma}_{\epsilon'}(i)} \qquad (12)$$

Además, una serie de diagnóstico puede definirse entonces como

$$D_i = \frac{\epsilon_i \sqrt{h_i}}{\hat{\sigma}_{\epsilon'}(i)(1-h_i)} = \frac{\epsilon_i' \sqrt{h_i}}{\hat{\sigma}_{\epsilon'}(i)} \qquad (13)$$

Entonces, la i^{th} observación tuvo una influencia alta en el modelo de regresión lineal si

$$\left|D_i\right| \geq 2\sqrt{\frac{p}{N}} \qquad (14)$$

Los residuos Studentized también pueden usarse para probar un outlier en el modelo de regresión. En este caso, la prueba estadística se define como (Kottegoda y Rosso, 1997):

$$T = \max\ (\xi_i, \quad i = 1,.....,N) \qquad (15)$$

donde ξ_j son las series de residuos studentized. El valor crítico para la prueba al nivel de significancia α es

$$C_{\alpha} = \sqrt{\frac{(N-p)f_{\alpha/N}(1, N-p-1)}{N-p-1+f_{\alpha/N}(1, N-p-1)}} \qquad (16)$$

donde $f_{\alpha/N}(1, N\text{-}p\text{-}1)$ es el α/N quantil de la distribución F con 1 y (N-p-1) grados de libertad. La hipótesis nula de un outlier en el modelo de regresión se rechaza al nivel de significancia α si (Kottegoda y Rosso, 1997):

$$T \leq C_{\alpha}$$

6. CASO DE APLICACIÓN

Se uso un procedimiento de selección de modelos de regresión múltiple para definir un modelo de predicción de caudales medios mensuales en el río San Lorenzo en Jaguas, relacionando los caudales del río con los mismos caudales y variables macroclimáticas relacionadas con el fenómeno de El Niño. El modelo finalmente seleccionado relaciona la descarga media mensual estandarizada del río San Lorenzo en Jaguas, Colombia, con las mismas descargas rezagadas un mes y con la serie SST estandarizada en la región Niño 4 rezagada dos meses. El modelo estimado puede escribirse como:

$$Y_t = 0.009 + 0.561x_t^{(1)} - 0.205x_t^{(4)} + e_t$$

El modelo anterior con dos variables independientes fue evaluado usando las pruebas y procedimientos descritos anteriormente.

La racionalidad física del modelo es muy aceptable ya que incluye los caudales del mes anterior que es de esperarse dada la persistencia que se esperaría en una serie de caudales medios mensuales estandarizados. Igualmente incluye una serie directamente relacionada con la ocurrencia del fenómeno del Niño que es determinante en la hidrología de esta región del país (Antioquia).

Con respecto a la prueba de multi-colinealidad basada en el factor de inflación de la varianza (VIF), se obtuvieron valores de 1.132 para la relación entre las variables independientes, valor que esta muy por debajo del límite de 10 mostrando que en este caso las variables no están correlacionadas y que pueden usarse en el modelo de regresión definido.

A continuación se analizó la capacidad predictiva del modelo de regresión y se obtuvieron los siguientes valores:

Desviación típica de la variable dependiente	= 0.978
Relación desviación típica residuales y variable dependiente	= 0.781
Coeficiente de correlación	= 0.621
Coeficiente de Determinación	= 0.390

Indicando que hay una correlación significativa. El valor obtenido de la relación de la desviación típica de los residuos a la variable dependiente es menor que uno indicando que se obtiene una mejoría con el uso del modelo de regresión con respecto a solo usar el valor medio. El coeficiente de determinación da un valor relativamente pequeño indicando que es baja la cantidad de variabilidad en los datos explicados por el modelo de regresión.

A las series de residuos obtenida con el modelo de regresión se le hicieron pruebas de independencia y normalidad. Todas ellas rechazaron la hipótesis de independencia y normalidad indicando que tal vez sea necesario transformar la muestra antes de modelarla o incluir parámetros adicionales en el modelo.

Con respecto a la presencia de observaciones influyentes se realizaron las pruebas descritas anteriormente una de estas pruebas identifico la presencia de una observación influyente en los datos y las otras dos no identificaron ninguna observación influyente en los datos. Al realizar una prueba de puntos anormalmente extremos en los datos la hipótesis de un punto anormalmente alto en los datos fue aceptada pero por un muy pequeña diferencia entre el estadístico calculado para la prueba y el de la distribución teórica.

Los resultados muestran que el modelo es en términos generales adecuado. El problema de independencia y normalidad de los residuos podría solucionarse con una transformación de las observaciones y repitiendo todo el procedimiento para la serie obtenida. Para complementar este análisis la Figura 1 muestra la serie de tiempo de los residuos, la Figura 2 los residuos contra los valores de la variable dependiente, y la Figura 3 el correlograma de los residuos. Observando estas figuras puede observarse que la serie residual no es independiente mostrando una clara dependencia del tiempo.

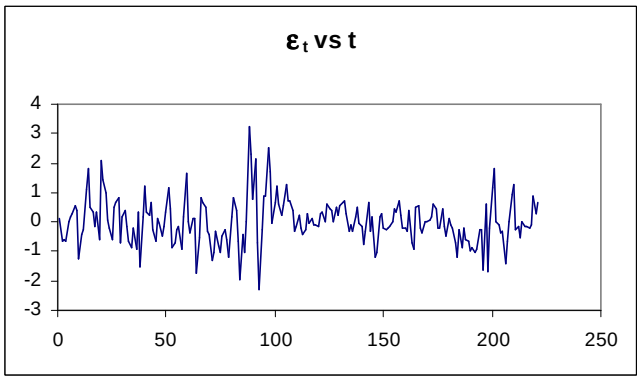

Figura 1. Serie de tiempo de los residuos

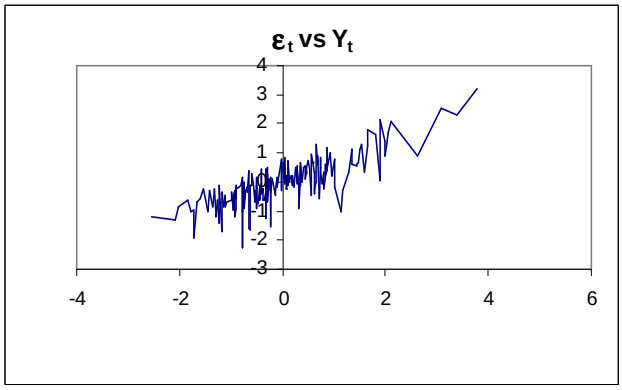

Figura 2. Figura 2 Residuos vs Variable Dependiente

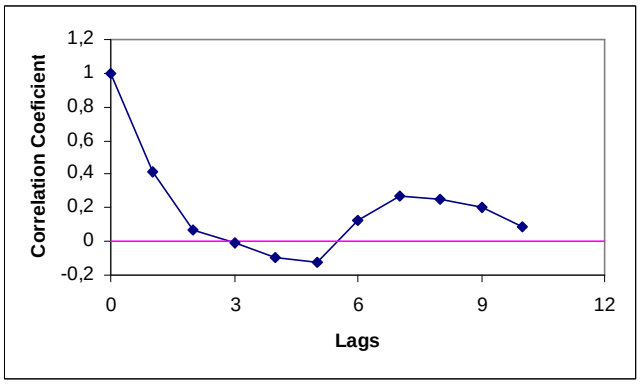

Figura 3. Correlograma de la serie Residual

7. CONCLUSIONES Y RECOMENDACIONES

Los modelos de regresión son una herramienta poderosa extensivamente usados en hidrología, recursos hidráulicos y en problemas de ingeniería ambiental. Sin embargo, se debería tener cuidado con el uso de modelos de regresión. El modelo refleja las condiciones de los datos usados para derivar y estimar el modelo y como tal no debería usarse cuando las condiciones con las que el modelo se derivó cambian. Por ejemplo, cuando un modelo se desarrolla para una cierta región geográfica éste debería aplicarse

solamente para esa región. Si va a ser usado para otra región, exige un análisis extenso para determinar su aplicabilidad.

Los modelos de regresión son tan buenos como los datos usados para derivarlos. Cuando una cantidad pequeña de datos es usada para definir y estimar un modelo de regresión, puede pasar que el modelo muestre un ajuste excelente, pero cuando hay más datos disponibles puede mostrar que el modelo no era tan bueno como se pensó. En el caso de grupos de datos pequeños, hay siempre dudas sobre la bondad del modelo estimado y el modelo tiene siempre que ser re-estimado cuando más datos están disponibles.

Las pruebas de bondad de ajuste de un modelo de regresión presentadas en este trabajo se deben usar de manera sistemática cuando se analiza el posible uso de un modelo de regresión. Estas pruebas son complementarias y en algunos casos algunas de ellas podrían indicar un modelo aceptable y otras podrían resaltar algunos problemas en el modelo que se esta tratando de desarrollar.

Cuando estas pruebas de ajuste indican problemas en el modelo es posible que en algunos casos esos problemas tengan solución. Por ejemplo los problemas de normalidad de los residuos normalmente se solucionan con una transformación simple de las observaciones. Los problemas de independencia pueden solucionarse incluyendo variables adicionales en el modelo. Los problemas de observaciones influyentes se solucionan identificando y removiendo del análisis esas observaciones. En algunos casos esos problemas no son solucionables y en ocasiones no se podrá desarrollar un modelo de regresión aceptable para el problema en cuestión.

Los análisis de bondad de ajuste de modelos de regresión deben hacerse con consideraciones amplias y flexibles. La racionalidad física del modelo que se esta ajustando es tal vez una de las consideraciones más importantes en ese análisis. La capacidad predictiva del modelo es otra consideración importante ya que evalúa el objetivo para el cual se desarrolló el modelo, si el modelo hace adecuadamente el trabajo que se supone debe hacer. Es posible que el modelo no cumpla con algunas pruebas estadísticas y sin embargo el analista decida usarlo basado en la recionalidad física del modelo y en su capacidad predictiva.

8. REFERENCIAS

Helsel, D.R. y Hirsch, R.M., 1993. Statistical Methods in Water Resources. Elsevier, Amsterdam.

Hipel, K.W., McLeod, A.I. y Lennox, W.C., 1977. Advances in Box - Jenkins Modeling - 1. Modeling Construction. Water Resources Research, 13(3): 567-575.

Kottegoda, N.T. y Rosso, R., 1997. Probability, Statistics and Reliability for Civil and Environmental Engineers. The McGraw Hill Book Co., New York.