IMPORTANCIA DE LOS REGISTROS HIDROLOGICOS EN EL DISEÑO Y PROYECCION DE ESTRUCTURAS HIDRAULICAS

LOS METODOS DE REGRESION

(Segunda Parte)

*José Iván Cárdenas Montoya**

Continuación del artículo iniciado sobre este tema en el No. 1 de la Revista. En esta parte se describen los métodos de regresión lineal (simple y múltiple), multivariada y el método de la "Tormenta Similar".

Dichos métodos no consideran las características físicas y geomorfológicas de las cuencas hidrográficas y sus implicaciones sobre los resultados obtenidos, considerando más el aspecto matemático-estadístico. Más bien, se analizan detenidamente los efectos de la reconstrucción o de la extensión de los registros sobre los parámetros estadísticos de las series de tiempo hidrológicas.

Fundamentalmente los métodos que se describen son: Métodos de regresión lineal (Simple y múltiple) a través de enfoques analítico y gráfico, métodos de regresión multivariadas, y el método de la "Tormenta Similar".

Los métodos de regresión son los que han recibido la mayor atención, por ser los más utilizados para estos propósitos y además, por tener muchas implicaciones de tipo Matemático-Estadístico.

---

* Ingeniero Civil de la Universidad Militar "Nueva Granada" Ingeniero de Mantenimiento en comunicación de datos, en la División Informática Banco del Estado.

Obviamente este modelo debe estar de acuerdo con las leyes físicas que gobiernan los fenómenos, pero sus resultados dependen de los datos utilizados.

Limitándose al caso de los métodos de regresión con tres variables, es decir, dos independientes y una dependiente se pueden obtener entre otras, las configuraciones que se observan en la figura 1, con sus ecuaciones analíticas correspondientes.

También se pueden utilizar combinaciones de las ecuaciones mencionadas para describir relaciones complejas entre las variables y, además, las ecuaciones se pueden extender para incluir un mayor número de variables independientes.

Tal vez el de mayor utilización en el campo de los fenómenos hidrológicos es el modelo **y= a + bx**, pero su aplicación debe ser debidamente controlada ya que no siempre es el apropiado para describir un determinado fenómeno.

El método gráfico de correlación entre dos variables, el cual se analizará adelante, puede dar importantes pautas sobre el tipo de regresión que se debe emplear.

Una vez seleccionado el modelo apropiado, se pueden de-terminar los coeficientes de la regresión **(a,b,c,d,etc.)** por cualquiera de los métodos existentes, siendo el de los mínimos cuadrados el de mayor utilización en los problemas de regresión lineal.

2.1.2.1 Transformación de variables.

Muchas veces los datos originales de las variables involucradas en una ecuación de regresión se transforman mediante artificios matemáticos. Existen tres razones principales para hacer esta transformación:

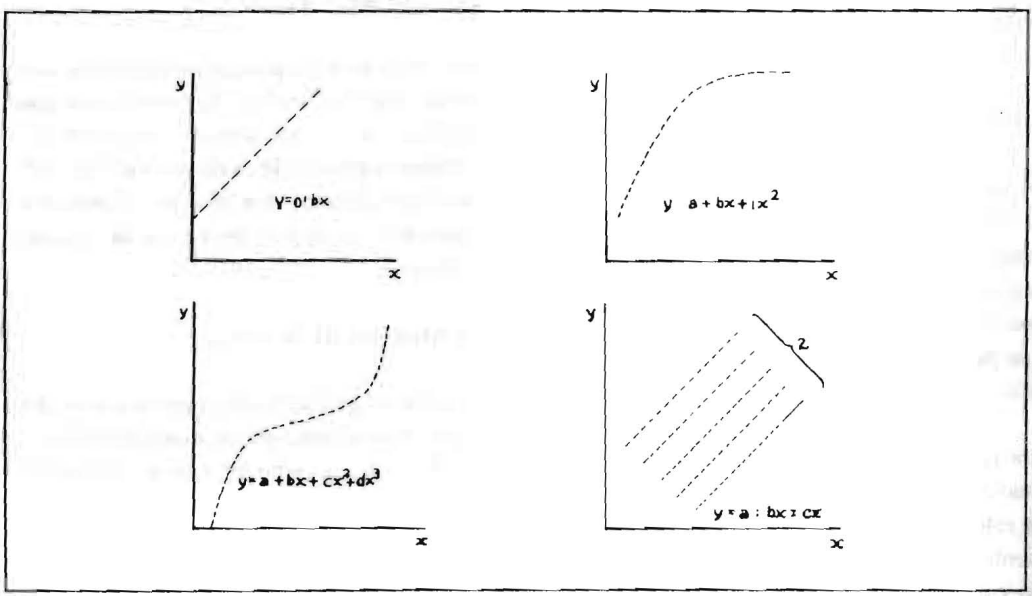


FIGURA 1. Representación de funciones de regresión

- Aproximar la distribución marginal de las variables transformadas a la distribución normal.

- La variación de los puntos a lo largo de la linea de re gresión es más homogénea.

- Obtener una varianza igual alrededor de la línea de regresión a través de todo el rango de aplicación.

Algunas transformaciones, a manera de ejemplo, para linealizar funciones son las siguientes:

Tipo de función	Ecuación en forma lineal	
$y = a + bx$	$y = a + bx$	(1)
$y = be^{ax}$	$\log y = \log b + a\,(\log e)x$	(2)
$y = ax^b$	$\log y = \log a + b \log x$	(3)

Sobre este último tipo de transformación, Hirsch (1979, p. 1783) hizo una comparación con el método clásico de regresión lineal. Con respecto a esto se hablará en la parte dedicada a la regresión en el espacio logarítmico.

En el texto de Chow (1964,p.8-49 Tabla 8-II-2.) se puede encontrar una lista de tales transformaciones.

2.1.2.2 Regresión lineal simple.(Ajuste analítico.)

Posiblemente el modelo comúnmente utilizado en hidrología se basa en la suposición de una relación lineal entre dos variables, y su objetivo principal radica en estimar una variable (dependiente) a partir del conocimiento de otra variable (independiente).

El resultado final de la regresión lineal de una variable **Y** contra otra variable **X** es una línea recta que da el mejor estimado de **Y** para un valor dado de **X**. También se puede determinar la línea recta que da el mejor estimado de **X**, para un valor dado de **Y**. esas dos rectas no necesariamente son iguales.

En hidrología, generalmente solo se utiliza la primera de estas regresiones, dada por la ecuación (4):

$$y^3(i) = a + bx(i) \qquad (4)$$

en donde:

$y^3(i)$: Valor estimado de la variable dependiente **y**.
a y b : Intercepto y pendiente de la línea de regresión.
x(i) : Valor de la variable dependiente **x**.

Las premisas en que basa el método de regresión lineal, desde el punto de vista estadístico, son las siguientes:

- *La variable independiente está exenta de errores, mientras que éstas solamente ocurren en la variable dependiente.*

- *La varianza de la variable dependiente no depende de los valores de la variable independiente.*

- *Los valores registrados de la variable dependiente son variables aleatorias sin correlación. Sharp et al (1960, p. 1284).*

Para la aplicación de las pruebas de significancia estadística de la regresión, se supone que la población de la variable dependiente se distribuye normalmente alrededor de la línea de regresión para cualquier valor constante de la variable independiente.

Más adelante, dentro del análisis de regresión línea múltiple se mirarán estas suposiciones a la luz de los fenómenos hidrológicos, y se verá en que medida se ajustan a la realidad física.

La aplicación de la ecuación (1) al tema del presente trabajo, resulta directa. Para el caso de caudales medios mensuales y niveles de agregación mayores, se extiende el registro de la estación que presenta un registro corto (estación satélite), a la que se asocia la variable **Y** mediante el registro extenso de otra estación que se escoja dentro de un grupo de estaciones vecinas y a la que se asocia la variable **X** (estación base o pivote) .

El criterio de selección de la estación base, que parece razonable y que es el utilizado en la literatura sobre el tema, es el de la estación que tenga el alto coeficiente de correlación muestral con la estación satélite. También se puede llenar el vacío de un dato faltante, simplemente entrando a la ecuación (4) de regresión con el valor de **X**. "Cuando algunas observaciones están perdidas, (es decir que son vacíos en el registro), el ajuste de un modelo lineal basado en el principio de los mínimos cuadrados, es el procedimiento desarrollado para estimar los valores de las observaciones perdidas". (Wilkinson,1958, p.257).

Hay que anotar que el mismo procedimiento puede aplicarse para registros incompletos de precipitación, con las mismas suposiciones y con el mismo análisis de resultados. Este procedimiento se conoce con el nombre de transferencia de información hidrológica de una estación a otra.

Muchos autores abordan este tema, sobre todo en el sentido de mejorar los estimadores de los parámetros estadísticos de las series de registros reconstruidos

En el caso específico de las series de caudales medios la base de las hipótesis sobre los que se apoya todo el enfoque de la transferencia de la información según Fiering (1963, p. 2), es la siguiente:

-Se supone un régimen hidrológico estable, o sea que se puede esperar una correlación significativa entre las series para diferentes sitios. Básicamente se espera que no ocurran cambios en los regímenes hidrológicos con los cuales están asociadas las series.

-Los caudales anuales, o alguna transformación de éstos, se suponen normalmente distribuidos. Los eventos concurrentes para el caso de dos series siguen distribución normal conjunta.

-Los eventos se distribuyen independientemente en el tiempo, de tal forma que el coeficiente de autocorrelación se considera nulo.

De no hacerse esta última suposición el análisis sería muy complicado ya que debido al fenómeno llamado "persistencia hidrológica (definido como la tendencia que existe en la naturaleza que a caudales altos anuales tienden a seguir caudales altos) disminuye el contenido de información que tenga sobre un evento hidrológico". (Matalas y Langbein, 1962, p. 3442)

Estas suposiciones limitan la aplicabilidad de los resultados obtenidos mediante la transferencia de información y tal vez la última mencionada sea la que presenta los mayores inconvenientes en las series de caudales debido a la autoregresión existente dentro de ellas.

Muchos autores han tratado este tema, ya que "intuitivamente los procesos hidrológicos poseen un carácter autoregresivo (lo que significa que el valor de una variable hidrometeorológica en el momento presente depende de los valores precedentes)" , Múnera (1983, p.26). Además de los resultados de las investigaciones sobre los caudales de diversos ríos han demostrado esta apreciación, Yevjevich (1964, p.12) haciendo un análisis de los correlogramas (el correlograma es un gráfico del coeficiente de correlación Vs. diferentes rezagos de la serie) de los caudales

anuales en una muestra de 140 estaciones de medición encontró que 124 tenían un coeficiente de correlación de primer orden, positivo.

Para valores de caudales medios diarios, se analizaron re-gistros de 17 ríos (Quimpo, 1967, p.18), y se encontraron correlaciones seriales que variaban entre 0.56 y 0.97.

Otros aspectos de la autocorrelación de los caudales, sobre los análisis hidrológicos han sido analizados en distintas investigaciones entre las que se pueden mencionar los trabajos de Leopold (1959), Matalas y Langbein (1962) y Lloyd (1963), que han demostrado que la autocorrelación reduce la confiabilidad de los otros parámetros estadísticos de las series de caudales e incrementa los requerimientos de almacenamiento para objetos de regulación de los caudales de un río.

Además de los trabajos anteriores, Fiering (1967, p. 29) desarrolló un estimador de la función de autorrelación de los caudales anuales, basado en la suposición de estacionariedad de las series de caudales. También se desarrolló un modelo para estimar la estructura de correlación de los caudales mensuales, suponiendo la no estacionariedad en el proceso.

El resumen anterior, sobre las investigaciones realizadas a cerca de la autocorrelación, se ha hecho con el propósito de indicar hasta qué punto son válidas las suposiciones que se aceptan para efectuar la transferencia de información sin la cual sería muy dispendioso el cálculo matemático de los modelos.

Estimación de los parámetros de la regresión. Retomando la ecuación (4):

$$y^3(i) = a + bx(i) \qquad (4)$$

y aplicando el método de estimación de mínimos cuadrados, se tiene que los coeficientes **a** y **b** son aquellos valores que minimizan la suma de los cuadrados de las desviaciones (o residuales), lo que se expresa matemáticamente como:

$$\min Z = \sum_{i-1}^{N} [\, g(i) - y(i) \,]^a \qquad (5)$$

$$= \sum_{i-1}^{n} [\, g(i) - a - bx(i) \,]^a \qquad (6)$$

en donde:

y^3 **(i)** : Valor estimado por la regresión.
y (i) : Valor muestral **i** de la serie **y.**
x (i) : Valor muestral **i** de la serie **x.**
N : Número de términos de la regresión.

Es decir que **Z** es una función de **a** y **b**. Para que esta función tenga un mínimo es necesario que se cumplan las igualdades:

$$\frac{n_E}{n_a} = 0 \quad : \quad \frac{n_E}{n_b} = 0 \qquad (7)$$

Resolviendo el sistema anterior se obtiene:

$$a = \bar{y} - b\,\bar{x} \qquad (8)$$

$$b = \frac{\sum_{i-1}^{u} x_{(i)}\, y_{(i)} - N\overline{xy}}{\sum_{i-1}^{a} x^2_{(i)} - N\bar{x}^2} \qquad (9)$$

en donde **x** e **y** son los valores medios de las series x e y para la muestra de tamaño N.

De acuerdo con ésto, si se desea extender el registro en una estación y cuenta con n_1 datos de caudal o precipitación, medios mensuales, trimestrales, anuales etc., a partir de un registro más extenso en una estación x que cuenta con n_1+n_2 datos de la misma variable, se aplica la regresión lineal al siguiente arreglo de datos:

x(1), x(2),x(3),.....,x(n1),..... x(n1+n2)
y(1), y(2),y(3),.....,y(n1)

No es necesario que las dos series comiencen o terminen en forma simultanea, ni tampoco que todos los registros sean consecutivos.

La forma óptima de la ecuación (1), esta dada como:

$$y^3_{1\text{-}0} = \bar{y}(n_1) + i \frac{S[y(n_1)]}{S[x(n_1)]} [x(n_1) - \bar{x}(n_1)] \qquad (10)$$

En donde:

$y^3(1)$: Valor estimado i de la serie y a partir de la regresión.

$\bar{y}(a_1)$, $\bar{x}(a_1)$: Medias muestrales de los n_1 valores de y y x respectivamente.

$S[y(n_1)]$, $S[x(n_1)]$: Desviaciones estándar muestrales de los n_1 valores de y y x respectivamente

r : Valor estimado del coeficiente de correlación entre x e y para los n_1 valores.

De esta manera, para un valor faltante $y_{(i)}$, se entra en la ecuación con el valor de $x_{(i)}$ de intervalo de tiempo de interés y se halla el valor estimado $y_{(i)}$.

La técnica de regresión lineal ha sido ampliamente estudiada por muchos autores como instrumento eficaz en la extensión de series con pocos registros.

Los factores de los cuales depende el mejoramiento de los parámetros son la longitud del registro original y el coeficiente de correlación entre las series.

Existe un parámetro conocido como **Información relativa I**, definido como "la relación entre la varianza de un parámetro estadístico estimado con el registro original y la varianza estimada a partir del registro extendido (combinado)" (Fiering,1962 p. 21). Este parámetro se da en términos de las varianzas debido a que es una medida de confiabilidad de un estimador que esta dada por su varianza.

En otros términos **Información Relativa I**:

I= [Varianza(serie original)]/[Varianza(serie extendida)]

Cuando **I > 1**, se concluye que es más preciso el estimador de la serie extendida, ya que la varianza de parámetro de ésta es menor que la varianza de la serie original.

Por otro lado cuando la varianza del parámetro de la serie extendida es mayor que la varianza del parámetro, el valor de **I** (Información relativa) se hace **menor que uno**, esto da a entender que es menos preciso, y por lo anterior el pro-cedimiento de regresión, entendido como la aplicación de la ecuación (4) no debe ser utilizado

Otro criterio para la evaluación de la aplicación de los métodos anteriores se dispone del **Error Medio Cuadrático (EMC)** de cualquier parámetro estadístico θ_i, definido como el valor esperado de la desviación del estimado sesgado y el estimador insesgado de parámetro.

$$E\,M\,C\,(8_i) = |(\phi_i - 8_i)^2| \qquad (11)$$

Es deseable desde el punto de vista físico que el **EMC** se aproxime a **CERO**, cuando T (El intervalo de tiempo muestral) aumente.

Entonces, "para un **T** grande, cualquier estimador ϕ_i, necesariamente tendería a un valor muy cercano al de un estimador verdadero θ_i. " Los estimadores que tienen esta propiedad se denominan estimadores consistentes. Puede verse que el **EMC** se reduce a:

$$E[(\phi_i - 8_i)3] - E\,[(\phi_i - E[\phi_i] + E[\phi_i] - 8_i|^2|$$

$$- E[(\phi_i - E[\phi_i])^3] + 2E[(\phi_i - E[\phi_i])\,(E[\phi_i] -$$

$$\phi_i)| + E[(E[\phi_i] - \phi_i)^2]$$

El segundo término se anula ya que:

$$E[\phi_i - E[\phi_i]| - E[\phi_i] - E[\phi_i] - 0$$

$$E[(\phi_i - 8_i)^2 = E[(\phi_i - E[\phi_i])^2] + E[(E[\phi_i] -$$

$$8_i)^2] \qquad (12)$$

La ecuación (12) indica que el **EMC** es la suma de dos partes: la primera es la varianza del estimado, que se expresa como:

$$Var[\phi_i] = 0^2_{\phi i} = E[(\phi_i - E[\phi_i])^2] = E[\phi^2_i] - E^2[\phi_i]$$

$$(13)$$

y la segunda es el cuadrante del sesgo del estimador:

$$b^2[\phi_i] - E[b^2[\phi_i]| - E[(E[\phi_i] - \phi_i)^2] \qquad (14)$$

Es decir que, en términos de varianza y el cuadrado del sesgo, el **EMC** es:

$$E[(\phi_i - 8_i)^2] - Var[\phi_i] + b^2[\phi_i] \qquad (15)$$

Esta presentación breve, de la teoría sobre **EMC** se justifica debido a que es uno de los criterios utilizados con más frecuencia en la mayoría de las técnicas de reconstrucción de registros, para comparar la calidad de los parámetros observados con los parámetros estimados.

Otros aspectos del análisis de regresión simple. La confiabilidad de la regresión está medida por el error estándar o error típico, que se define como la desviación estándar de la distribución (normal) de los residuales al rededor de la línea de regresión, tal como lo muestra la **Figura 2.**

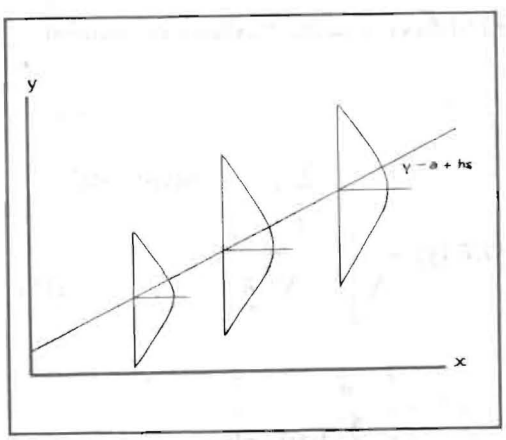


FIGURA 2. Distribución **NORMAL** de los puntos alrededor de la línea de regresión.

Por definición, el error estándar es el mismo a travez de todo el rango de **x**. También se conoce como error estándar del estimado, error estándar de la regresión o desviación estándar de la regresión.

El error estándar del dato calculado mediante la técnica de regresión se compone de tres partes: el error de la media, el error de la pendiente de la línea y el error estándar del estimado. De esta forma el error estándar de la predicción es:

$$SEP = SEE\,(y)\sqrt{1 + \frac{1}{N} + \left(\frac{(x - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right)} \qquad (16)$$

en donde:

SEP : Error estándar de la precipitación.
SEE(y) : Error estándar del estimado.
N : Número de datos de la regresión.

El **SEE(y)** se define mediante la ecuación:

$$SEE\,(y) = \sqrt{\frac{\sum_{i-1}^{N} |y - a - b[x(i) - \bar{x})]^3}{N - K}} \qquad (17)$$

$$\sqrt{\frac{\sum_{i-1}^{n} |y(i) - \bar{y}|^3}{N - K}} \qquad (18)$$

en donde:

SEE(y) : Error estándar del estimado.
k : Grados de libertad, en este caso el número de coeficientes de la línea de regresión **(k=2)**.

El **SEE(y)** mediante la ecuaciones **17 y 18**, se mide el porcentaje de la varianza de la variable dependiente no explicada por la regresión. El error estándar del estimado de la regresión puede utilizarse como un estimador razonable del error estándar de la predicción, ya que la inexactitud de la ecuación de regresión generalmente es pequeña en comparación con la dispersión de los datos alrededor de la línea de regresión.

En otros términos el error total de estimación es:

$$\Sigma\,(y - \bar{y})^2 = \Sigma\,(y - y^3)^2 + \Sigma\,(y^3 - \bar{y})^2 \qquad (19)$$

Error total = Error no explicado + Error explicado

Gráficamente, esta relación se puede ver en la **figura 3**

Otra de las características importantes que se pueden obtener de la regresión es el coeficiente de determinación, R^2, y se define por la relación entre el error explicado y el error total.

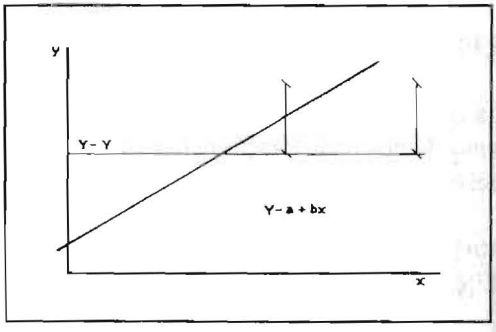


FIGURA 3. Composición del error total

R^2 = **{Error Explicado} / {Error total}**
en donde **R** es el coeficiente de correlación.

Con el fin de poder determinar la confiabilidad de la aplicación de este método, se pueden analizar algunas características importantes. Con el análisis de éstas, se determinan pautas importantes a seguir para determinar la confiabilidad de los resultados al ser aplicados en las series de datos, tales como:

Coeficiente de determinación. Este coeficiente se puede estudiar de dos maneras:

- Medida de mejoramiento en la estimación de **y**, utilizando la línea de regresión.

Este mejoramiento se mide respecto al error total, el valor de $R^2 = 1$ indica que la reducción del error total es completa al ser estimado **y** mediante este método.

- Si $R^2 = 0.7$, la reducción del error total debido a la determinación de la línea de regresión es del 70%.

- Calidad del ajuste y medida de linealidad. El valor de R^2 **tiende a 1** cuando los puntos se acercan a la línea de regresión y se acercan a una línea recta.

Se puede utilizar otro criterio dado por la relación entre R^2 **y** SEE^2**(y)** dada por:

$$SEE^2(y) = (1 - R^2)\, S^2(y) \qquad (20)$$

de donde S^2**(y)** es el estimado de la varianza de **y**.

De la anterior relación se deduce que estos criterios son complementarios debido a que por medio de ellos se puede obtener el porcentaje de varianza de **y** no explicado por la regresión, indicando una medida de la calidad del ajuste de ellas.

Analizando los resultados de la aplicación de este criterio se deduce lo siguiente:

En la medida en que aumente el valor de **R**, es mas confiable la regresión, y en complemento el valor de **SEE(y)** al ser mas pequeño.

No necesariamente los valores de estos dos parámetros **R y SEE**, son buenos indicadores de la calidad de la ecuación de estimación para valores muy alejados de los valores medios.

Otra manera de obtener una idea de la bondad del ajuste lo-grado en la regresión es mediante el análisis gráfico de los residuos.

Se efectúa una gráfica de los valores de **y(i)** contra los valores estimados. Para un ajuste perfecto se debe obtener una línea recta de **45°** que pasa por el origen sobre la que se dispersan los puntos a lado y lado. Cualquier variación de esta línea puede atribuirse a que las suposiciones hechas sobre los errores (media cero, varianza constante y distribución normal) no se ha cumplido, lo cual debe ser corregido efectuando algún tipo de transformación sobre los datos originales o mediante otros artificios.

Otro gráfico de interés en el análisis es el de residuos contra tiempo, colocando los errores en un gráfico en su orden cronológico para detectar posibles ciclos o estacionalidades en los registros.

En seguida se presentan temas concernientes a la teoría de decisiones y a las pruebas de hipótesis.

Una hipótesis estadística es un juicio o suposición, que puede ser o no cierta, acerca de dos o mas poblaciones.

En este tipo de estudios se proponen diferentes hipótesis H_0 y H_1 esto a fin de probar la nulidad de la primera, conocida como **Hipótesis de Nulidad.** El rechazo de H_0 conduce a aceptar su hipótesis alternativa H_1.

A fin de probar la validez de una regresión se debe demostrar que el coeficiente **b** de la ecuación (1) difiere significativamente de cero.

Es decir:

$$A_0 \mid B - 1)$$

$$A_1 \mid B - 1)$$

En donde la hipótesis nula esencialmente dice que: la variación en **Y** no está explicada por la línea recta, sino que ocurre en forma aleatoria.

En estadística existe una prueba denominada como **Prueba F** que señala en una forma relativamente indirecta la existencia o no de una relación entre la variable independiente(o independientes en el caso de una regresión múltiple) y la variable dependiente.

Se define como:

$$r = \frac{SS_{reg/gl1}}{SS_{rmm/gl2}} = \frac{\text{Varianza debida a la regresión}}{\text{Varianza residual}}$$

en donde:

SS_{reg}: Suma de cuadrados de la Regresión
SS_{res}: Suma de cuadrados residuales
gl_1: Grados de libertad: **k=1**
gl_2: Grados de libertad **N-k-1**

Este factor **F** tiene una distribución de probabilidades, conocida como **distribución F,** con $gl_1 = 1$ y $gl_2 = N-2$ para el caso de la regresión lineal simple.

Por lo tanto:

$$F = \frac{SS_{reg/1}}{SS_{res/(N-2)}} \qquad (22)$$

Se rechaza H_0, significativo al nivel de confianza α, cuando:

$$F > F_0\,(1, n-2)$$

El valor de F_α **(1, n-2)** se lee de la tabla de la distribución **F** para un nivel de significancia de α con **1** y **n-2** grados de libertad.

En este caso la regresión es estadísticamente significativa y en caso contrario NO.

La prueba **t (de Student)** para el valor **b** con base en su desviación estándar, consiste en calcular los límites de confianza de **b** y determinar si el valor de cero queda dentro de estos límites.

En el caso de la regresión lineal simple,las pruebas **t** y **F** dan el mismo resultado, Para la regresión múltiple la prueba **F** se aplica a la significancia de la regresión total, en tanto que la prueba **t** es una verificación de los coeficientes.

Este tema de la calidad del ajuste se puede profundizar en los libros de estadística especialmente de Walpole y Myers (1978), Kreyszig(1973),Haan (1977) y Draper y Smith (1966).

2.1.2.3 Ajuste Gráfico de la línea de regresión. La línea de regresión que establece el tipo de correlación entre los valores del caudal o de la precipitación en una estación con los valores en otra estación, también puede ajustarse por medio de un procedimiento gráfico muy diferente al que utiliza la regresión analítica (mínimo-cuadrático). Sobre este último ya se ha hecho amplia referencia y ahora se tratará el del análisis gráfico de la regresión. Este método también puede utilizarse para efectuar la regresión lineal entre dos variables (simple), como también entre tres o más variables (múltiple). Esto último se realiza a través de una serie de ábacos que relacionan las variables involucradas.

Fundamentalmente la referencia bibliográfica consultada en esta parte del trabajo son: Langbein y Handison (1955), Searcy (1960),Riggs (1968) y Benson(1965)

Estos métodos, analíticos y gráficos se diferencian fundamentalmente en cuanto a como se ajusta la gráfica a los puntos del diagrama.

Cuando es imposible establecer la selección de un modelo analítico sobre una base física que describa la correlación entre dos estaciones es conveniente establecer un tipo de regresión gráfica a priori que pueda indicar el modelo apropiado.

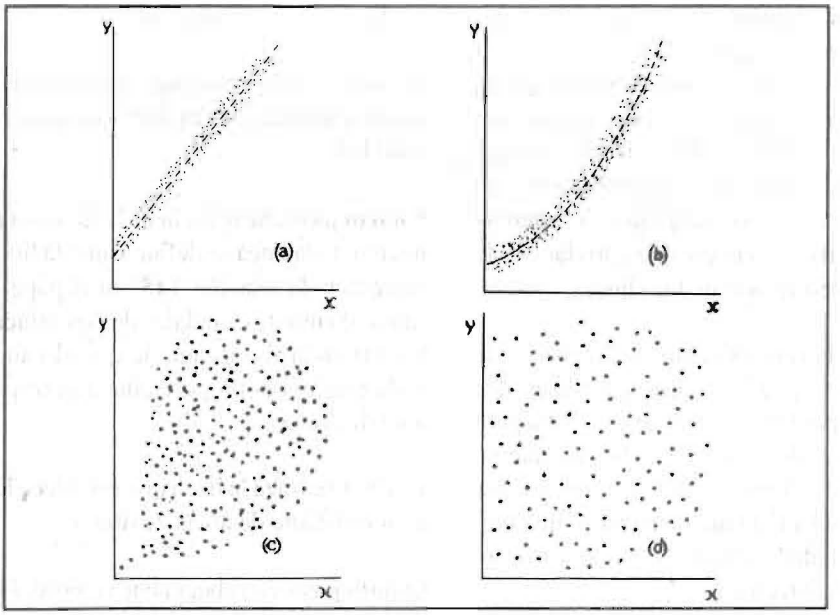


FIGURA 4 Diagramas de dispersión de los datos

En la **figura 4** se pueden establecer criterios gráficos para ajustar un determinado tipo de ecuación particular a la nube de puntos de cada gráfico. La figura **4a** indica que se debe utilizar un modelo lineal de la forma:

$$y = a + bx \qquad (23)$$

La figura **4b** de un modelo:

$$y = a + bx + cx^2 \qquad (24)$$

en el que el sentido de la curvatura viene dado por el signo de **c**

Las figuras **4(c)** y **4(d)** muestran un diagrama de puntos muy disperso por lo tanto requieren de una transformación de variables ya que parece que no existiera ningún tipo de correlación entre las variables originales. Introduciendo una tercera variable, y en algunos casos más variables entre la variable que se considera independiente y la que se considera variable dependiente, se pueden obtener altas correlaciones entre las variables transformadas y ajustar el modelo analítico apropiado a esos puntos.

En el procedimiento gráfico de la línea de regresión, la forma de la gráfica (lineal o curvilínea) ajustada depende, en una gran parte, del tipo de papel empleado. Tal vez el papel logarítmico resulta más útil debido a que la transformación logarítmica, como ya se vio, tiende a convertir los valores de los caudales en variables aleatorias con distribución normal (Normalizar los caudales) y además tiende a convertir una posible correlación de tipo curvilíneo en una de tipo lineal.

En general la relación entre los registros de dos estaciones, puede expresarse mediante dos líneas: Una para los caudales bajos y otra para los caudales altos. Si el ángulo con que se interceptan es grande se puede construir una curva de transición para suavizar el quiebre, a menos que dicho ángulo sea suave, no hay necesidad de hacerla.

- Procedimiento. El proceso gráfico de ajuste de la línea de regresión para el caso de dos estaciones (dos variables) es el siguiente:

Primer paso. Se grafican los valores de los caudales, generalmente caudales mensuales en **m^3/seg**, de las estaciones en un papel **Logarítmico** (Log-Log), colocando la variable independiente en el eje de las **X** y la variable dependiente en **Y**

Segundo paso. Se divide el rango de distribución a lo largo del eje **x**, de los caudales, utilizando líneas verticales en un grupo de cinco a diez intervalos. Lo mismo que a lo largo del eje **y**

Tercer paso. Se determina gráficamente el punto medio de cada intervalo en ambas direcciones. Si el número de puntos en una banda es par se promedian gráficamente los dos puntos medios, si es impar se promedian los tres puntos del medio, ponderado dos veces el punto del medio. Si una banda contiene menos de tres puntos, no se determina el punto medio para esa cuadrícula.

Lo mismo se hace a lo largo del eje y, recomendándose utilizar colores diferentes para los dos sentidos.

Cuarto paso. Se traza la línea de igual rendimiento. Esta línea se define como la línea que representa la relación (**45°** en el papel logarítmico) entre los caudales de dos estaciones, basada en la suposición de que el caudal en cada estación es proporcional a su respectiva área de drenaje.

Es decir se traza la línea que establece la relación constante de áreas de drenaje.

Quinto paso. Se dibujan líneas rectas a través del promedio de los puntos medios trazados en los intervalos (línea de relación) dando poco peso a los valores extremos. El extremo superior de ésta(s) línea(s) casi siempre resulta paralelo a la línea de igual rendimiento. El extremo inferior generalmente es otra línea recta que puede conectarse con la línea superior mediante una curva de transición suave.

Sexto paso. Se dibujan dos líneas, cada una de ellas equidistante (paralelas) a la curva de relación, de tal forma que una sexta parte de los puntos queden por encima de la curva superior y otra sexta parte quede por debajo de la inferior. Rara vez quedan a la misma distancia de la línea de relación.

Séptimo paso. Obviamente, el área comprendida entre estas dos líneas dibujadas contiene las dos terceras partes restantes de los puntos. Por definición el ancho vertical de esta banda es de dos veces el error estándar estimado.

Este se calcula obteniendo el logaritmo del cociente entre un caudal que caiga sobre la línea superior y un caudal que caiga sobre la línea inferior. El error estándar puede escribirse en unidades logarítmicas pero con mayor frecuencia se expresa en términos porcentuales.

La confiabilidad del valor estándar estimado gráficamente está influenciada por dos factores que tienen efectos contrarios. Si la línea de regresión gráfica tiene una mayor pendiente que la línea obtenida por el método de los mínimos cuadrados, el error estándar que se obtiene por el primer método será mayor que el calculado mediante el segundo. Si se supone que la línea de regresión gráfica es la misma que la línea de regresión mínimo cuadrática, el error estándar obtenido gráficamente subestimará al error estándar analítico cuando algunos puntos están lejos de la línea pero la mayoría están cerca de ella.

"En cualquier caso el error estándar determinado gráficamente es solamente un valor aproximado pero es adecuado para muchos problemas ". (Riggs, 1968,p.20).

Octavo paso Se dibujan las líneas horizontales que excluyan una sexta parte de los puntos (en la dirección y) por encima y otra sexta parte por debajo. La mitad de la distancia entre estas líneas se multiplica por el factor:

$$\frac{N}{(N-1)}$$

para obtener la desviación estándar de y. El término N es el número de observaciones concurrentes.

Noveno paso. Se puede estimar el coeficiente de correlación de la línea de regresión gráfica mediante la expresión:

$$r = 1 - \left(\frac{Se}{Sy}\right)^2 \qquad (25)$$

en donde

r: Coeficiente de correlación
Se: Error estándar determinado gráficamente
Sy: Desviación estándar de la variable Y

- Regresión múltiple gráfica. Cuando el error estándar estimado mediante la regresión simple por el método gráfico es grande, se deben utilizar más estaciones en la correlación (regresión múltiple) con el fin de reducir el error estándar estimado.

A veces también se pueden agregar otro tipo de datos en el análisis de correlación, como pueden ser los registros de precipitación de una estación cercana.

El método gráfico también puede efectuar este procedimiento mediante la utilización de ábacos; pero no se incluye en este trabajo ya que el método analítico de regresión múltiple puede resultar óptimo en el sentido de obtener mejores coeficientes de correlación múltiple utilizado, por ejemplo el método de

regresión por pasos ó método de regresión escalonada.

El método gráfico de correlación múltiple se encuentra analizado ampliamente en los trabajos anteriormente mencionados de Riggs (1955) y Searcy (1960), como también en el texto de Linsley, Kohler y Paulus (1977, p.359).

2.1.2.4. Comparación entre los métodos gráficos y analítico de regresión lineal.

METODO GRAFICO

VENTAJAS

- Es de rápida ejecución
- Es de mucha utilidad en el proceso de definición del modelo de ajuste apropiado.
- Es un método que señala la necesidad de hacer transformaciones sobre los datos originales si es necesario.
- Es un método que considera la existencia de puntos falsos dentro de la regresión y llama la atención sobre ellos. Más adelante se hará referencia con detalle sobre la presencia de esa clase de puntos dentro de un registro hidrológico.

DESVENTAJAS

- No se pueden identificar los efectos de los cambios pequeños en la variables independientes sobre la variable dependiente.
- El número de variables independientes debe restringirse a tres ya que los errores acumulados durante el tratamiento manual del gráfico le disminuye su exactitud.
- La relación resultante que incluya tres o más variables es confusa a menos de que se exprese matemáticamente o se establezca otra relación de tipo gráfico.

METODO ANALITICO

VENTAJAS

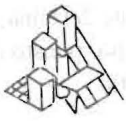
- Proporciona los mejores estimados de los coeficientes de regresión y del error estándar de estimación, del modelo utilizado.
- Permite efectuar pruebas sobre la significancia estadística de los coeficientes de la regresión.
- Los resultados se presentan en forma clara y concisa de manera que son faciles de interpretar.
- Para la muestra de datos y el modelo utilizados, los resultados son los mismos y no dependen de quién haga el procedimiento.

DESVENTAJAS

- Para modelos con muchas variables independientes se requiere bastante tiempo de computador, si no en el cálculo de los resultados entonces en la preparación de los datos.
- No se detiene a reflexionar sobre la presencia de puntos espurios en los datos de regresión y sobre el posible efecto que éstos tienen sobre la línea (plano ó Hiperplano) de ajuste, final.
- Se corre la incertidumbre que el modelo de regresión escogido, o la transformación sobre los datos no sean apropiados.

En general debería primero efectuarse un trabajo gráfico para explorar cual tipo de modelo emplear, y luego desarrollar el método analítico para obtener la conclusiones finales.

STRUTEC
LIMITADA



PEDRO JOSE HERRERA ROCA

Carrera 6a. No. 86-20 Tels. 218 7221 - 218 7227 - Fax. 610 0267
Santafé de Bogotá, D.C. Colombia

industrial elecivil ingenieros ltda

CESAR ASHLEY MORA BARNEY

Carrera 9a. No. 53-58 Of. 314 Tels. 249 8027 - 212 0709
224 9525 Santafé de Bogotá, D.C.