# HERRAMIENTAS ESTADISTICAS EN HIDROLOGIA

Por H. C. Riggs

Técnicas de Investigación de Recursos Hidráulicos
UNITED STATES GEOLOGICAL SURVEY

Capítulo A1

# HERRAMIENTAS ESTADÍSTICAS EN HIDROLOGÍA

Por H. C. Riggs

Libro 4

ANALISIS E INTERPRETACION HIDROLOGICOS

Traducido y prep. por: A. Eduardo R.
Dirección General de Recursos Hidráulicos
M. O. P.

## CONTENIDO

## TABLAS

## FIGURAS

--°--

# HERRAMIENTAS ESTADISTICAS EN HIDROLOGIA

Por H. C. Riggs

Este capítulo sobre "Tecnicas de Investigación de los Recursos Hidrológicos" provee el material básico necesario para poder comprender los procedimientos estadísticos más usados en hidrología, proporciona procedimientos detallados de análisis de regresión, con ejemplos, describe el análisis de la varianza y de la covarianza y, discute las características de los datos hidrológicos.

## INTRODUCCION

Como el análisis hidrológico se hace cada vez más sofisticado, el diseño adecuado y la interpretación de estos análisis exige un conocimiento mayor de los métodos estadísticos. En realidad, existen dos herramientas estadísticas usadas durante largo tiempo: la curva de frecuencia de crecida y la curva de duración, que requieren un conocimiento de la teoría estadística para poder hacer una evaluación adecuada.

Los métodos estadísticos más complejos son los matemáticos, pe

ro los métodos gráficos son extremadamente útiles y adecuadamente exactos para muchos propósitos, si se hacen en base a un conocimiento de las hipótesis que los sostienen y si se interpretan correctamente. Hasta hace pocos años los textos de estadística recalcaban los procedimientos aplicables a los datos distribuídos normalmente debido a que la suposición de normalidad es apropiada para muchos tipos de datos biológicos y agrícolas. Pero gran parte de los datos usados en hidrología, o no estan distribuídos normalmente o no tienen en lo absoluto una distribución de probabilidad. La mayoría de los hidrólogos aprendieron estadística en textos o cursos dirigidos hacia el análisis de datos normalmente distribuídos. En consecuencia, algunos primeros análisis hidrológicos, o estaban incorrectamente hechos o se interpretaban incorrectamente.

Este capítulo de "Técnicas de Investigación de los Recursos Hidrológicos" suministra el material básico necesario para comprender los procedimientos estadísticos más útiles en hidrología. Aunque comienza con el concepto básico de distribución, se omiten muchos detalles elementales. Se supone que el lector tiene cierta familiaridad con la terminología estadística, con los procedimientos de computación y con la teoría elemental de probabilidades al nivel obtenido en los cursos de estadística para ingenieros o a partir del curso por correspondencia.

"La Estadística Elemental en la Hidrología" del Geological Survey de los E. E. U. U."

Aunque en este capítulo se recalca la teoría, la forma de exposición es más intuitiva que rigurosa. Se presenta muchos enfoques prácticos de la regresión gráfica y se localizan las trabas asociadas al cómputo de las líneas de mínimos cuadrados. El capítulo finaliza con una discusión de las características estadísticas de los datos hidrológicos.

## DISTRIBUCIONES

El concepto de una población de objetos que tiene una distribución según el tamaño (ó según cualquier otra características) es básico para el método estadístico. No es posible recoger una cantidad suficiente de datos para definir exactamente una distribución de frecuencia, pero se puede probar la existencia de una distribución determinada con el grado de confianza deseado repitiendo un experimento muchas veces.

Kendall (1. 952, P. 23) informó acerca de los resultados de una experiencia que hizo tirando 12 dados simultáneamente y anotando los seis que salían en cada tirada, En la tabla 1 se muestra el resultado obtenido despúes de haber tirado los dados 4. 096 veces, además se señala la frecuencia relativa computada a partir de los resultados experimentados y la frecuencia relativa teórica computada a partir de la distribución binomial. El estrecho acuerdo entre la frecuencia téorica y la experimental indica que

la distribución binomial es aplicable a este problema.

TABLA 1.- RESULTADOS DEL EXPERIMENTO CON LOS DADOS (Según Kendall, 1952).

| N° de Seis | Frecuencia | Frecuencia Relativa. | Frecuencia relativa Teórica. |
|---|---|---|---|
| 0 | 447 | 0,109 | 0,112 |
| 1 | 1.145 | 0,280 | 0,269 |
| 2 | 1.181 | 0,288 | 0,296 |
| 3 | 796 | 0,194 | 0,197 |
| 4 | 380 | 0,093 | 0,089 |
| 5 | 115 | 0,028 | 0,029 |
| 6 | 24 | 0,006 | 0,007 |
| 7 y más | 8 | 0,002 | 0,001 |
| TOTAL | 4.096 | 1, 00 | 1, 00 |

La distribución binomial es discreta, es decir, puede tomar valores solamente en puntos específicos de la escala. Es posible, en el experimento, conseguir solamente un número entero de seis; no hay cifras tales como 5,5 o 3,2 seis.

Comúnmente una variable puede tomar cualquier valor de la escala; a esta variable y a su distribución se les denomina contínuas. Se puede clasificar una variable contínua como tal si puede tomar cualquier valor de la escala aún cuando la limitación impuesta por las condiciones restrinja las observaciones a valores discretos. Esta condición está presente en la mayoría de los fenomenos naturales.

Para ayudarlos a comprender una distribución, considérense 1.000 anillos divisores de los troncos de varios árboles con un tamaño variable entre 2 y 240 unidades de espesor.

Si se agrupan por cada incremento de tamaño de seis unidades se obtiene un histograma o distribución de frecuencia (fig. 1). La irregularidad del perfil de esta distribución se debe (estadísticamente) al número de divisiones usadas en su preparación, mientras el número usado sea mayor más uniforme será el perfil de la distribución de frecuencia. Si el número de observaciones se aproxima al infinito y la magnitud del incremento se aproxima a cero, la evolvente de la distribución de frecuencia tendrá como límite una curva suave.

Luego, si se divide el valor de cada ordenada por un número tal que el área subtendida por la curva sobre el eje de abcisas sea uno, la curva resultante es una curva de densidad de probabilidad o de distribución de probabilidades, según se muestra en la figura 2.

El proceso que se acaba de describir requiere una suposición adicional de que la variable pueda tomar cualquier valor dentro del intervalo en que la variable es contínua, no discreta.

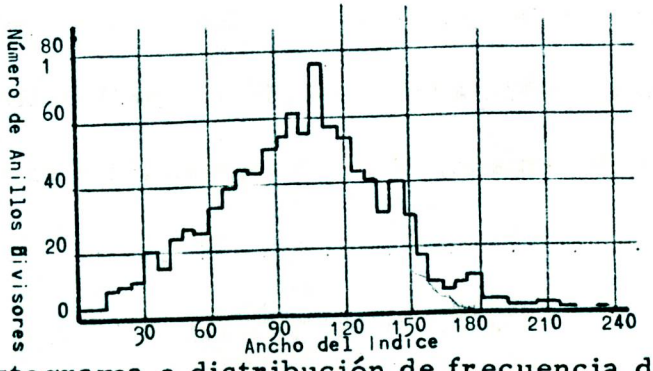


Figura N° 1. - Histograma o distribución de frecuencia de 1.000 divisores anulares de troncos de árboles.

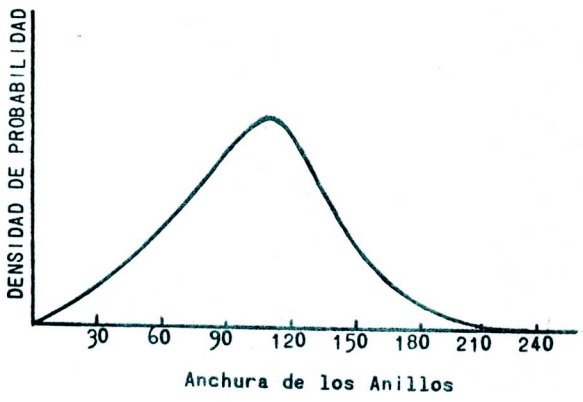


Figura N° 2.- Curva de densidad de probabilidad de 1.000 divisores anulares de árboles.

La distribución de probabilidad teórica describe la relación entre el tamaño (o alguna otra característica) y la probabilidad. Para que esta relación sea válida hay que tomar los elementos al azar, o estos deben surgir de esta manera. El tamaño de cualquier elemento tomado no debe depender del tamaño del que se ha tomado anteriormente. De acuerdo con las distribuciones de frecuencia se define la probabilidad como frecuencia relativa. La distribución de los seis obtenidos al arrojar repetidamente 12 dados se puede ilustrar llevando a un gráfico los valores de las frecuencias teóricas relativas de la Tabla 1. Las frecuencias relativas de cada uno de los incrementos de 6 unidades de la figural, se podrían computar de igual manera. En el primero de estos ejemplos, con cada posible resultado se asocia una probabilidad; en el segundo, con cada aumento de tamaño. Aqui la probabilidad es la de obtener no una

pieza específica sino una pieza cualquiera comprendida dentro del incremento del tamaño. Para la distribución contínua se requiere esta interpretación debido a que existe un número infinito de valores posibles y, por lo tanto, no hay la probabilidad de ocurrencia de un elemento en particular.

Refiriéndonos de nuevo a la figura 2, podemos ver que la probabilidad está relacionada con la distribución contínua de la manera siguiente: el área que subtiende la curva representa la suma de todas las probabilidades y debe de ser por lo tanto igual a la unidad. Debido al hecho de que para definir la distribución para la cual el área total es uno, la probabilidad de que cualquier elemento caiga dentro de cualquier segmento de la distribución es igual a la relación que existe entre el área de ese segmento y el área total.

Las distribuciones que se acaban de describir, la contínua y la discreta, se denominan distribuciones relativas de frecuencia, distribuciones de probabilidades o sólo distribuciones. Sin embargo la interpretación de probabilidad es válida solamente si se toman los datos al azar, por ejemplo, el flujo diario promedio de un río está estrechamente relacionado con el flujo de los días anteriores, luego a la distribución del caudal diario no se aplica estrictamente la interpretación de probabilidad.

También es posible aproximarse a una distribución que simplemente describe a la muestra. Por ejemplo, la distribución del tamaño de los granos de una muestra tomada de un lecho fluvial se mide para caracterizar

al material, no hay interés en averiguar la probabilidad de obtener un grano de un tamaño X por muestreo adicional. Aquí la muestra no es un grano sino el conjunto de granos de diferentes tamaños.

Sólo se usan ampliamente unas cuantas distribuciones teóricas standard. La inferencia y la teoría de muestreo se basan en su mayor parte, en la distribución normal con la cual se supone que el lector está familiarizado.

En este capítulo y en los siguientes se introducirán otros tipos de distribuciones teóricas, cuando se considere apropiado.

## DISTRIBUCIONES ACUMULATIVAS

Supongamos que conocemos la curva de densidad de probabilidad (distribución de probabilidad) para una variable y que estamos interesados en la probabilidad de un evento al azar mayor que un valor particular E. Esta probabilidad se puede obtener midiendo o computando la relación del área total por encima del valor básico. Por ejemplo, la curva de la izquierda de la figura 3 muestra una curva que subtiende un área de 0, 1 a la derecha del punto E, es decir P=0, 1. Luego la probabilidad de un evento al azar mayor que E es 0, 1.

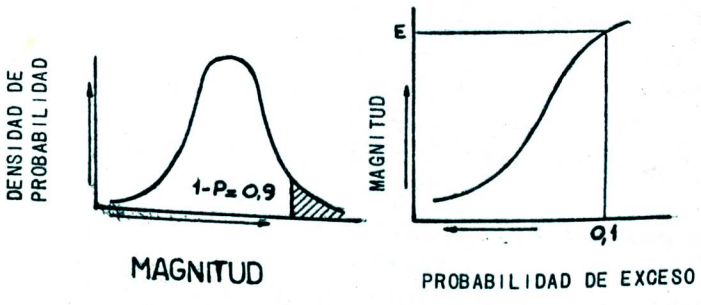


Figura N° 3. -Curva de densidad de probabilidad (a la izquierda) y su forma acumulativa (a la derecha)

Se puede preparar otra forma de la curva de probabilidades acumulando las probabilidades desde un extremo de la curva y llevando cada una de estas probabilidades acumuladas a un gráfico en función de la magnitud del suceso apropiado. Generalmente, la acumulación se realiza matemáticamente. El resultado de esto es la curva de la derecha de la figura 3. Las distribuciones acumulativas se llevan comúnmente a una escala gráfica de probabilidades tal que la curva teórica resulta ser una línea recta. Esta escala se puede improvisar para cualquier distribución biparamétrica. Se conoce y se usa mucho el papel normal de gráficos de probabilidad. En muchos análisis de frecuencia hidrológica se usa el papel Gumbel. (Aunque la distribución Gumbel del valor extremo es una distribución triparamétrica; un parámetro, la oblicuidad, es constante para la forma usada y permite la construcción de una escala que da un gráfico de línea recta). Los papeles de probabilidad normal y Gumbel se encuentran en el mercado con ambas escalas. Por lo tanto, existen cuatro tipos de distribuciones, a saber: normal, normal-log, Gumbel y Bumbel-log.

Cuando la curva de densidad de probabilidad se acumula desde el extremo derecho, se obtiene la probabilidad de exceder a las diferentes magnitudes. Si se acumula desde la izquierda, se obtienen las probabilidades de no exceder esas magnitudes. La curva acumulativa apropiada, más comunmente llamada curva de frecuencia, depende del uso deseado. Las diferentes distribuciones teóricas acumulativas usadas en hidrología y los métodos de estimación de sus parámetros a partir de una muestra de datos se

discute en otro capítulo de esta serie.

## INFERENCIA ESTADISTICA

Hemos tomado registros, a través de 54 años, de datos del río Rappahannock de Virginia y se nos antojan dos preguntas acerca del flujo promedio. Primeramente, Cúal es el flujo promedio para este período?. Este es un valor único que se puede computar fácilmente. La segunda pregunta es: Cúal es el flujo promedio del río?. Sólo podemos suponer que la media del ejemplo de 54 años es una estimación de la verdadera media (de población). En otras palabras, a partir de las características de una muestra de esa población podemos deducir las características de ésta.

La inferencia estadística se basa en la teoría del muestreo. De una población de características conocidas se sacan muchos ejemplos, (ya sean reales o conceptuales) y se define así la relación que existe entre las características de la muestra y las de la población. La teoría del muestreo exige el uso del concepto de una distribución de probabilidad. Supongamos que la distribución de alguna variable cualquiera es normal con una media $\mu$ y una desviación standard $\sigma$, según se muestra en la figura 4. (El término "cualquiera", en la forma que se usa aquí, quiere decir que la probabilidad de sacar un elemento de la población es la misma que para otro elemento de ésta).

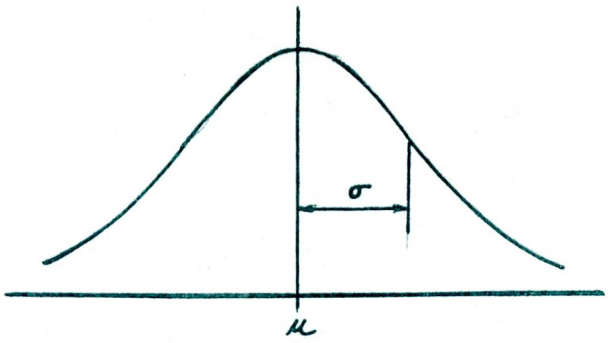


Figura N° 4. - Distribución normal.

Supongamos ahora que tomamos muchas muestras de tamaño N de esta distribución, que computamos la media de cada una de estas muestras, y la media, y la varianza de las medias de estas muestras. En la figura N° 5, la distribución de las medias de muestras de tamaño N se sobrepone a la distribución original. Se puede demostrar que la distribución de la media está centrada en $\mu$ y que la desviación standard de la distribución de la media es: $\sigma/\sqrt{N}$ . Por lo tanto, la media de las medias de las muestras de tamaño N es una estimación insesgada de $\mu$ . En consecuencia inferimos que la media de muestra $\overline{X}$, es una estimación de la media poblacional. Obviamente si usásemos otras muestras obtendríamos estimaciones diferentes de la media de población.

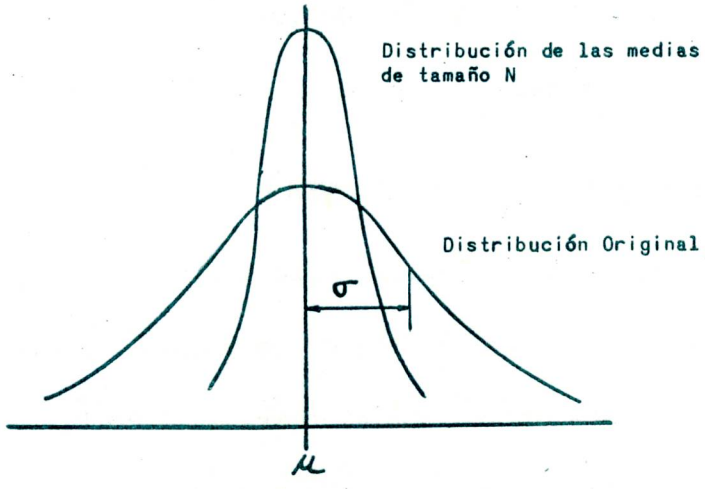


Figura N° 5. - Distribución de las medias de muestras de distribución normal. -

Podemos apreciar la solidez de la estimación, $\bar{X}$, de la media de población a partir de una muestra sencilla. La distribución de las medias de los valores sacados de una distribución normal es normal. En consecuencia, los dos tercios del valor deben estar comprendidos dentro de una desviación standard $(\sigma/\sqrt{N})$ a cada lado de la media.

Sin embargo, no conocemos al valor de $\sigma$, por lo tanto, tenemos que reemplazarla por S (siendo S la desviación standard computada a partir de las muestras). La distribución de $\bar{X}$ con una desviación standard de $S/\sqrt{N}$ se conoce como la distribución t de STUDENT ~~del estudiante~~, cuyos valores se tabulan en los textos de Estadística para diferentes tamaños de N.

Supongamos ahora que tenemos K muestras de tamaño N y que hemos definido K diferentes distribuciones de muestreo de la media. Para cada distribución de muestreo, podemos definir una media y una escala de solidez, e interesarnos por saber si tal escala incluye una media verdadera. Considerando la escala como un intervalo tomado al azar podemos decir que la probabilidad (P) de que este intervalo incluya a $\mu$ es 1 - e, en donde "e", es el nivel de significación. Matemáticamente, para e = 0,32;

$$P\left((\bar{X} - \sigma/\sqrt{N}) < \mu < (\bar{X} + \sigma/\sqrt{N})\right) = 1 - e = 0,68$$

El intervalo que está entre parentesis se denomina intervalo de confidencia, y los extremos, límites de confianza. Nótese que la relación anterior es verdadera solamente si usamos $\sigma$ en lugar de S, entonces 1 - e, es una función del tamaño de la muestra y el enunciado apropiado de probabilidad es:

$$P\left((\bar{X} - tS/\sqrt{N}) < \mu < (\bar{X} + tS/\sqrt{N})\right) = 1 - e = 0,68$$

en donde la t de estudiante es 1.09 para 10 grados de libertad, por ejemplo. La amplitud del intervalo confidencial aumenta a medida que disminuye el nivel de significación:por ejemplo, los límites de confidencia de 95% (e=0.05) son:

$$(\bar{X} - 2,23S/\sqrt{N}) \qquad (\bar{X} + 2,23S/\sqrt{N})$$

en donde el valor 2,23 corresponde al t de la tabla para 10 grados de libertad.

El intervalo confidencial descrito en el enunciado de probabilidad es un intervalo cualquiera, no uno específico. El enunciado de probabilidad (para e=0.05) significa que el 95 por ciento de un número grande de intervalos similarmente obtenidos incluiría a la media verdadera. Este enunciado de probabilidad no se puede extender a un intervalo específico debido a que éste contiene o no a la verdadera media y la probabilidad es uno a cero. La verdadera media no es una variable, es una constante.

Pero estamos interesados en hacer un enunciado de probabilidad, con respecto a un intervalo específico. Podemos decir que la probabilidad de que obtengamos un intervalo cualquiera que incluya a la verdadera media, es del 95 por ciento. Ordinariamente, en los informes hidrológicos, solo se necesita enunciar el intervalo obtenido y su nivel, no interpretar el significado. Ver el artículo de Mood (1.950 p. 221-222) para tener un enunciado preciso de la interpretación de un intervalo de confidencia. De acuerdo con esta teoría, se puede hacer una estimación de la media de población y tam-

bién una medición de su fiabilidad, a partir de una muestra cualquiera. Este es un ejemplo de inferencia estadística.

Volviendo a la teoría de muestreo, considérese la distribución de varianzas de una muestra de tamaño N de una distribución normal. Esta distribución no está centrada en $\sigma^2$ sino que está a la izquierda de ella, según se muestra en el gráfico superior de la figura 6. Por lo tanto, se conoce a $S^2$ como un estimador parcializado sesgado de $\sigma^2$ . Se le puede transformar en insesgado dividiendo entre N/(N-1) según el esquema inferior de la figura N° 6. La desviación standard de la distribución de muestreo de $S^2 \cdot N/(N-1)$ también se puede computar.

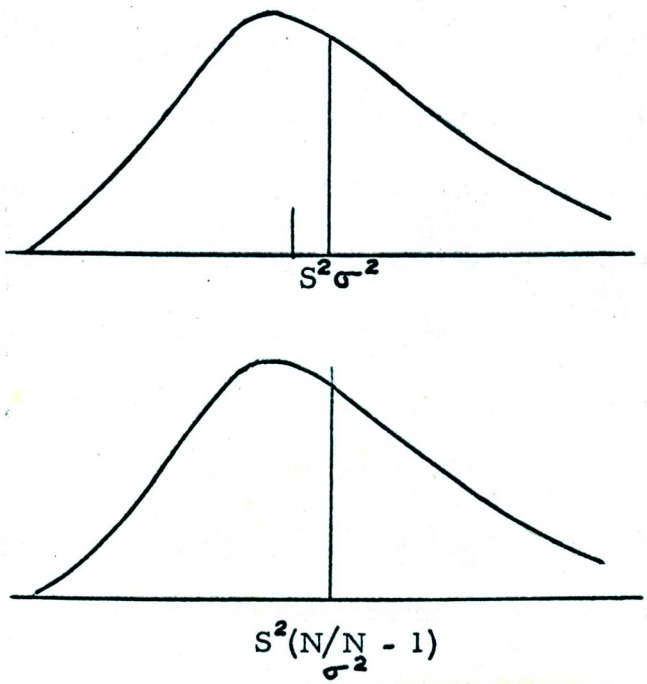


Figura N° 6. - Distribución de varianzas de muestras.

La prueba de hipótesis es otro uso de la inferencia, a continuación daremos un ejemplo. Supongamos que establecemos la hipótesis nula, Ho, de que la media de una población es cero, es decir, hacemos la hipótesis de que estadísticamente no hay diferencia entre la media y cero. Esta hipótesis nula se escribe:

$$Ho : \mu = 0$$

De esta población sacamos una muestra, computamos la estadística de la distribución de muestreo de la media y la definimos como normal con desviación media cero y standard igual a $S/\sqrt{N}$ según la muestra (Fig. 7). Ahora bién, si $\bar{X}$ ( según se calculó a partir de la muestra), se encuentra dentro de una desviación standard del cero, concluiríamos que no hay base para dudar de la hipótesis. Si, por otra parte $\bar{X}$ estuviese a dos o tres desviaciones standard de cero, concluiríamos que no es probable que la media de la población sea cero. Por esta última condición la probabilidad de obtener una $\bar{X}$ de tal tamaño, a partir de una población de media 0, es pequeña, por lo tanto rechazaríamos la hipótesis y diríamos que el resultado tiene significado a un cierto nivel de probabilidad, queriendo decir

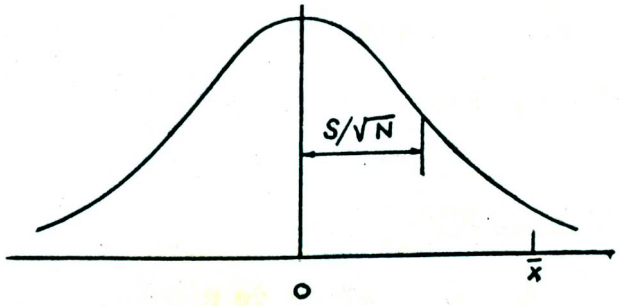


Figura N° 7. - Distribución hipotética de muestreo de la media.

que los resultados obtenidos difieren notablemente de la hipótesis. Un problema común es la prueba de significación de un coeficiente de regresión. La hipótesis nula es nuevamente que el valor real del coeficiente de regresión sea cero, y se puede realizar la prueba igual que antes. Sin embargo, el procedimiento que se usa comunmente es algo diferente. Se computan los límites de confianza con respecto al valor teórico. Si el coeficiente de regresión es b, su error standard S, y su valor de población $\beta$, se encuentra que los límites son:

$$b - tSb < \beta < b + tSb$$

en donde t es el valor apropiado para el nivel de significación y para el tamaño de la muestra. Si los límites incluyen el cero, se acepta la hipótesis, es decir que el coeficiente de regresión no es muy diferente de cero. Si los límites se encuentran a un solo lado del cero, se rechaza la hipótesis y se considera que el coeficiente de regresión es bastante diferente de cero, es decir, tiene cierto significado.

Existen muchas otras pruebas de significación, pero todas las pruebas paramétricas están basadas en la teoría de muestreo y siguen el procedimiento general descrito anteriormente. Cuando no se conoce la distribución de probabilidad de la variable en estudio se puede usar un grupo menos fuertes de pruebas no paramétricas (Ver Siegel, l. 956).

CORRELACION Y REGRESION. -

Las distinciones entre correlación y regresión deben ser conocidas para aplicar e interpretar cualquiera de los métodos. Estas distinciones son muy marcadas aunque puede parecer que tienen poca importancia debido a la similaridad existente entre los procedimientos de computación.

Dixon y Massey )1. 957, p. 189) hicieron la siguiente distinción entre las dos : " Un problema de regresión considera la distribución de frecuencia de una variable cuando otra se mantiene fija a diferentes niveles. Un problema de correlación considera la variación conjunta de dos mediciones,ninguna de las cuales se restringe durante el experimento".

La correlación es un proceso por medio del cual se define el grado de asociación entre las muestras de dos variables. El coeficiente de correlación es una definición matemática de esa asociación. Naturalmente, es posible computar un coeficiente de correlación a partir de cualquier par de juego de datos. La definición matemática de la asociación no implica ninguna relación de causa y efecto, ni siquiera , que la relación entre las dos variables es el resultado de una causa común.

La teoría de la correlación exige que se tomen los datos al azar de una distribución normal bivariada. Sin embargo McDonald (en 1. 957)

informó que los estudios experimentales de muestreo señalan que los efectos de no normalidad que los estadístas consideran generalmente como perturbadores, son de magnitudes geofísicamente inconsecuentes.

Otro requisito de la correlación es que las variables X e Y no conlleven errores de medición. Nada se puede medir sin error, luego este requisito es solo de grado. La cuestión del error permisible está sujeta a decisiones arbitrarias ya que no se sabe el verdadero error de los datos. El producto final del proceso de correlación es el coeficiente respectivo; no una ecuación. Las ecuaciones que describen a Y como función de X, y a X como función de Y son ecuaciones de regresión, no ecuaciones de correlación. Otra manera de establecer la distinción entre correlación y regresión es considerando que la correlación mide el grado de asociación entre dos variables, mientras que la regresión proporciona ecuaciones para estimar los valores individuales de una variable a partir de los valores dados de otra.

La fiabilidad de los resultados de una correlación depende del número de elementos usados para computar el coeficiente de correlación y la magnitud del mismo. Para muestras de 30 ó menos elementos, a menos que el coeficiente de correlación sea muy grande. Por ejemplo Bennett y Franklin (de 1. 954 p. 275) presentaron una carta en donde se indica que un coeficiente de correlación de + 0, 8 calculado a partir de una muestra de 20 elementos tendría una región de aceptación que se extendería desde 0, 6

hasta 0,9 para una probabilidad del 95%. Debido a esta inseguridad no se puede interpretar a plenitud la comparación de dos coeficientes de correlación que difieren solamente en unas cuantas centésimas. Tampoco parece haber justificación para proporcionar coeficientes de correlación con más de dos cifras significativas.

Si razonablemente se puede suponer que los datos se sacan de una distribución normal bivariada, entonces son apropiados ambos análisis. Es bajo esta suposición que se analizan los ejemplos de la mayoría de los textos de estadística. Sin embargo, la regresión es también apropiada en otras condiciones en que la correlación no lo es. Las únicas suposiciones que exige la regresión son:

1. - Las desviaciones de la variable dependiente con respecto a la línea de regresión (para cualquier X prefijado) están normalmente distribuídas y presentan la misma varianza a través de todo el campo de definición.

2. - Se sabe que los valores de la variable independiente no tienen error. Se considera a la variable dependiente como observación de una variable cualquiera, y a la variable independiente como una constante conocida asociada a esta variable cualquiera.

3. - Los valores observados de la variable dependiente son acontecimientos cualesquiera no relacionados.

4. - Cada una de las variables es homogénea, es decir, todos los valores individuales de la variable miden la misma cosa. Se consideran los datos

homogéneos si cualquier subgrupo, al cual se le puede asignar lógica - mente algunos de estos datos, tiene la misma varianza y la misma media esperada que cualquier otro subgrupo de la población. En regresión, nin - guna de las variables necesita tener una distribución de probabilidad (pero naturalmente, se supone que los valores de Y que corresponden a un X fi - jo están normalmente distribuídos).

Los productos finales de un análisis de regresión son dos ecuacio - nes: $Y = f(Y)$ y $X = f(Y)$ (generalmente sólo se calcula uno), debido a que- ésta es direccional. En oposición a esto, la correlación proporciona un ín- dice de la relación entre las variables.

La ecuación de regresión dá la cantidad promedio de cambio de la variable dependiente que corresponde a un cambio unitario de la variable - independiente. Luego da una información más específica que la que dá la - correlación. Se puede probar el coeficiente de regresión para determinar - si es apreciablemente diferente de cero; y esta prueba es idéntica a la prue- ba de significación del coeficiente de correlación (siempre que los datos se saquen de una distribución bivariada normal).

La fiabilidad de un error se mide por el error standard, que es la - desviación standard de la distribución (supuesta normal) de los residuos con respecto a la línea de regresión (La figura 8 muestra la distribución de los residuos). Por definición, el error standard es el mismo en todo el campo

de X. Ezekiel (1. 950, p. 131) llamó este error, error standard de estimación. También se le denomina error standard de regresión y desviación - standard de la regresión.

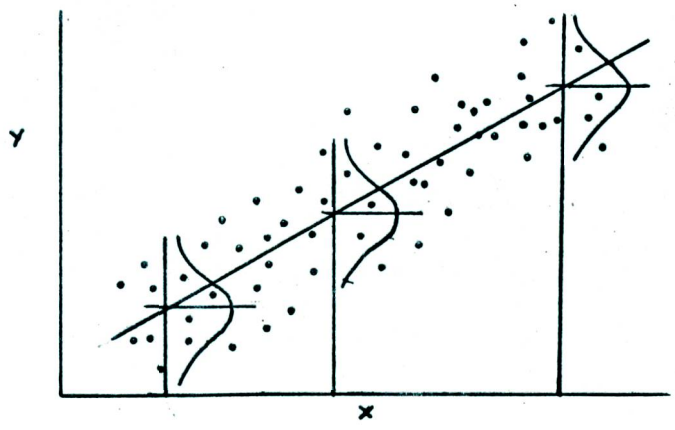


Figura N° 8. - Distribución normal de los puntos trazados con respecto a la línea de regresión.

El error standard de una predicción a partir de la regresión está - formado por tres partes: el error de la media, el error de la pendiente de la línea y el error standard de estimación, de tal manera que el error standard de una predicción (Sp) es:

$$Sp = Se\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(X-\bar{X})^2}}$$

en donde Se es el error standard de estimación; n, el número de elementos de la muestra y X la variable independiente. Luego el error de una predic - ción aumenta con la distancia a la media (Snedecor, 1. 948. p. 120). La mayoría de los análisis requieren el uso de la correlación múltiple o de la regre

sión. Una correlación múltiple se evalúa por medio de coeficientes parciales de correlación y de un índice de correlación total. Un coeficiente parcial de correlación es un índice del grado de asociación entre una variable dependiente y una variable independiente después que se han eliminado los efectos de otras variables independientes.

En una ecuación de regresión múltiple los coeficientes de regresión se denominan coeficientes parciales de regresión. Cada uno de ellos muestra el efecto que surte un cambio unitario de una determinada variable independiente sobre Y, cuando se mantienen constantes los efectos de otras variables independientes.

Si las variables independientes de una regresión están relacionadas entre sí, los coeficientes parciales de regresión serán de una magnitud diferente a los coeficientes sencillos de regresión. (Las variables independientes de una regresión están generalmente relacionadas entre sí así como también con la variable dependiente). Ver la sección sobre "Aplicación del Método de Regresión" para la elaboración de este asunto (p. 19).

Las suposiciones necesarias para la correlación se encuentran raras veces en los problemas de ingeniería y no se presentan, por lo general, en los problemas de hidrología. Muchos de estos problemas a los que no se puede aplicar el método de correlación se pueden tratar por el método de regresión debido a suposiciones menos restrictivas. Luego se puede utilizar el método de regresión para relaciones tales como la de la resistencia del-

concreto con el tiempo de fraguado, en donde ninguno de los dos valores se selecciona al azar y ninguna variable tiene una probabilidad de distribución. Obviamente el campo de tal relación está limitado al campo de los datos seleccionados.

En estas condiciones el coeficiente de correlación no se aplica, - pero, naturalmente, se puede computar a partir de la relación:

$$r = \sqrt{1 - (S_e/S_y)^2} \quad ,$$

en donde r, es el coeficiente de correlación; Se, el error standard de estimación y Sy la desviación standard de los valores de la variable inde - pendiente. De esta fórmula se deduce que r depende de Sy el que a su vez depende del campo de datos seleccionados para problemas tales como la relación que tiene la resistencia del concreto con el tiempo del fraguado. Por lo tanto, si las variables usadas en una regresión no son muestras tomadas al azar, el valor computado de r cambia con el campo de la mues - tra seleccionada arbritariamente y no tiene entonces significado. La verificación empírica de este enunciado se dá por los datos llevados al gráfico de la figura N° 9. (Estos datos se seleccionaron para demostrar este principio; la relación no tiene significado hidrológico). Usando todos los puntos, se computa la relación.

$$\log FMA = 2.27 + 0.59 \log AD;$$

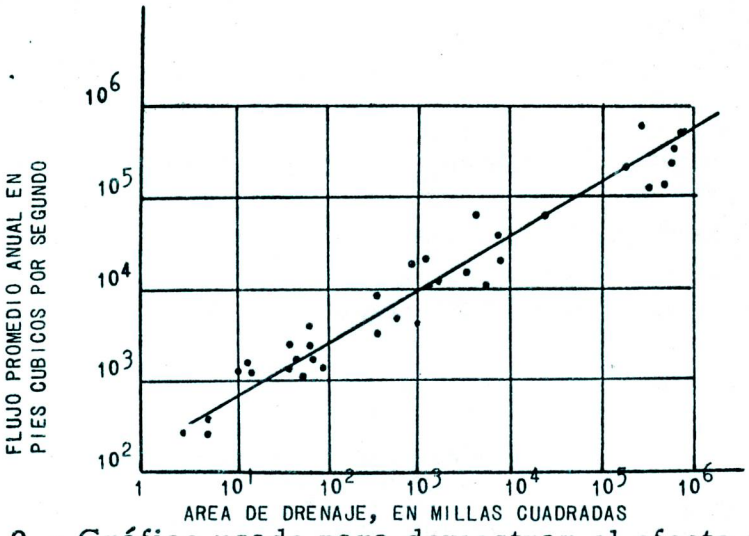


Figura N° 9. - Gráfico usado para demostrar el efecto del campo de muestreo sobre los coeficientes computados de correlación. La línea de trazos es la relación para 14 áreas de drenaje. que van desde 103 hasta 5180 $Km^2$.

Si se usan solamente los 14 puntos para las áreas de drenaje que - van desde 103 hasta 5180 Kms. la relación es:

$$\text{Log FMA} = 2.31 + 0.57 \log AD$$

Esta relación tiene un error standard de 0,23 unidades logarítmicas (casi lo mismo que el error previo standard), pero el coeficiente computa - do de correlación es 0,83 mucho más bajo que el obtenido usando muestras de un campo mayor. Obviamente, tal variabilidad del coeficiente de corre - lación lo hará inapropiado como medida del grado de relación para este ti - po de aplicación.

Este manual dá mayor énfasis a la regresión que a la correlación , no sólo porque esta no es comunmente aplicable a datos hidrológicos deter minados sino porque la regresión proporciona respuestas cuantitativas pa-

ra problemas específicos. En general, se prefiere la regresión a la co - rrelación, para los problemas hidrológicos aún cuando los datos sean - buenos para hacer un análisis de correlación. Los usos de los análisis - de regresión son:

1. - Para estimar los valores individuales de la variable dependiente que corresponde a los valores seleccionados de las variables independientes.
2. - Para determinar la cantidad de cambio de la variable dependiente asociada con un cambio unitario de la variable independiente.
3. - Para saber si ciertas variables (que no tienen distribuciones de probabilidad) están relacionadas con una variable dependiente.
4. - Para mejorar las estimaciones de los parámetros que definen la distribución de probabilidad de la variable dependiente.

La correlación es muy útil en los estudios teóricos y en análisis de serie cronológica.

CORRELACION SERIAL. -

Se ha señalado que para que una distribución de probabilidad sea válida los individuos deben aparecer o ser sacados al azar. Los datos hidrológicos tales como las descargas diarias de los ríos forman una serie cronológica, es decir una secuencia de valores arreglados por orden de aparición.

Las características y el análisis de las series hidrológico-cronológicas han sido descritos por Dawdy y Matalas (1.964). Una característica común de una serie cronológica es la existencia de un elemento determinado que produce una dependencia entre las observaciones que difieren en K unidades , a esta dependencia se le denomina correlación-serial, y su grado se mide por el coeficiente de correlación serial.

Una correlación serial de primer orden sería la dependencia entre observaciones adyacentes en el tiempo; el orden K-esimo es la dependencia entre observaciones separadas K unidades. El gráfico en donde se lleva el coeficiente de correlación serial contra el orden es un correlograma (Dawdy y Matalas, 1.964).

Para determinar la correlación serial, la serie cronológica se relaciona consigo misma K unidades aparte. Por ejemplo, la serie cronológica de la primera columna está relacionada consigo desplazada una observación para obtener un coeficiente de correlación serial de primer orden.

| | |
|---|---|
| $X_1$ | - |
| $X_2$ | $X_1$ |
| $X_3$ | $X_2$ |
| . | . |
| . | . |
| . | . |
| $X_n$ | $X_{n-1}$ |

Los detalles de cálculo son los mismos que para la relación entre dos variables, y se dan en la sección "Regresión Lineal Simple".

Dawdy y Matalas muestran una prueba de significado de un coeficiente de correlación serial de primer orden.

METODOS DE REGRESION. -

La sección previa describía la regresión en términos generales y concluía con algunos usos de éstas. Esta sección describe el computo y la interpretación de las ecuaciones de regresión, analítica y gráficamente, y algunas características del método de regresión.

MODELOS DE REGRESION. -

Un problema de regresión se comienza con una variable dependiente que se quiere predecir a partir de una o varias variables independientes. Las variables independientes son valores o características que parecen estar fisícamente relacionadas con la variable dependiente.

A continuación, necesitamos un modelo que describa la manera en que las variables independientes están relacionadas con las variables dependientes. El modelo debe estar de acuerdo con los principios físicos conocidos, pero su forma exacta debe de estar dictada por los datos usados.

En la figura 10, se muestran las ecuaciones y los gráficos de algunos modelos comunes de regresión, usando como variable dependiente a Y, y como variable independiente a X y Z. Las relaciones conjun -

tas, las que incluyen una variable que es el producto de otras dos, fueron discutidas detalladamente por Ezekiel y Fox (1.959).

El producto de dos variables se llama límite de interacción. Las combinaciones de los modelos que se muestran en la figura 10, se pueden usar para describir relaciones más complicadas, y las ecuaciones pueden ampliarse para incluir variables independientes adicionales.

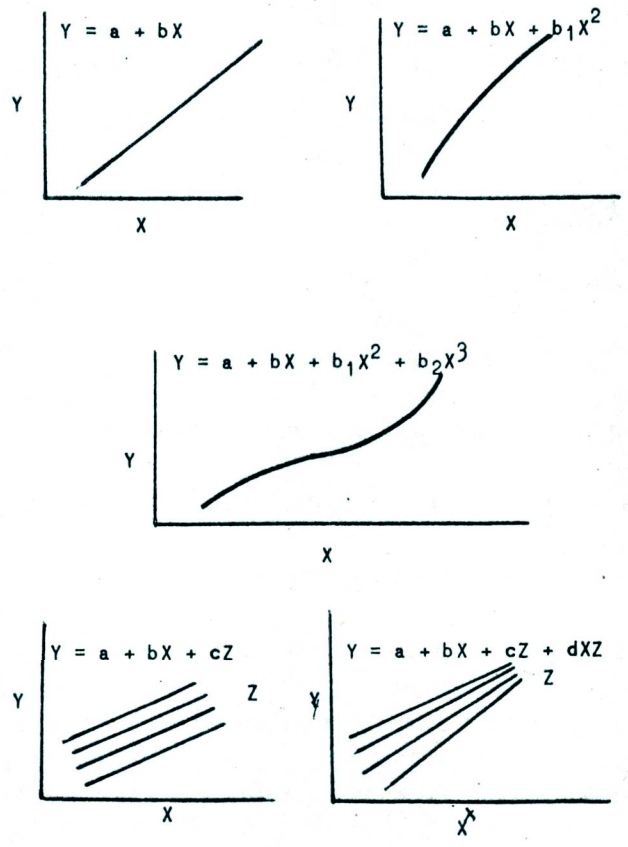


Figura N° 10. - Ecuaciones y gráficos de algunos modelos comunes de-regresión.

Al seleccionar un modelo conveniente, los coeficientes de esa ecuación del modelo se computan a partir de los datos de muestreo - por el método del mínimo cuadrático. Esta linealidad de la ecuación- del modelo es un requisito para la solución directa del mínimo cuadrado. La linealidad se puede conseguir transformando las variables.

TRANSFORMACIONES. -

Hay dos razones principales para transformar los datos antes del análisis: (1) Para obtener un modelo lineal de regresión y

(2) Para lograr igual varianza con respecto a la línea de regresión en todo el campo.

En la figura 10 vimos que ciertas regresiones de dos variables se pueden linealizar sin transformar las variables. El método se conoce como regresión polinómica en la que las variables adicionales se añaden al modelo en potencias sucesivamente crecientes de la variable - independiente. Pero supongamos que sabemos o hacemos la hipótesis- de que una relación debe ser de la forma: $Y = a X^b$

Tomando logarítmos a ambos lados se obtiene la ecuación lineal siguiente:

$$\text{Log } Y = \log a + b \log X$$

en donde a y b son constantes que se pueden computar por medio del análisis del cuadrado mínimo usando como variables log Y, y log X. De igual manera la relación $Y = ab^x$, se puede transformar en:

$$\log Y = \log a + X \log b$$

en donde a y b son constantes y log Y y log X, las variables. Algunas veces se usan otras transformaciones, pero la transformación logarítmica es la más común, la segunda razón, y la más importante, para realizar la transformación de los datos es la de conseguir igual varianza con respecto a la línea de regresión. Una de las suposiciones básicas para el método de regresión es la de que la distribución de errores con respecto a la línea de regresión es normal y constante en todo el campo (fig. 8). Se usa de nuevo una transformación logarítmica.

Por ejemplo, el gráfico de la izquierda de la figura 11, (sacado del Geological Survey de E. E. U. U. de 1. 949 p. 488) muestra un aumento en la dispersión de los puntos al aumentar las lluvias. Pero cuando se hace un gráfico logarítmico de las variables ( el de la derecha de la figura 11), la dispersión de los puntos es casi uniforme en todo el campo. Luego si se hubiese deseado llevar el análisis más allá de la presentación gráfica se hubiera tenido que hacer una transformación de las variables. (ordinariamente se invertirían las variables en la carta si hubiese que hacer una regresión debido a que el desague es la variable dependiente).

Otras razones para transformar los datos son las de introducir actividad al modelo y alcanzar la normalidad. Acton (1. 959 p. 219-223) , discute el uso de las transformaciones. En los ejemplos anteriores solamente se ha usado la transformación logarítmica porque es la más común y la más útil. Otras transformaciones, tales como la de la raíz cuadrada, pueden ser adecuadas para ciertos datos.

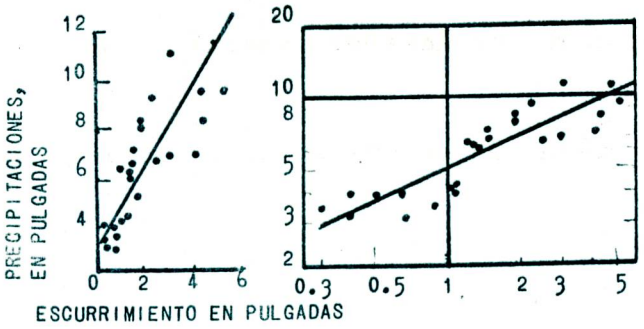


**Figura N° 11. -** Datos del Geological Survey de E. E. U. U. (1. 949, p. 488) - en escala logarítmica y natural mostrando los resultados de varianzas iguales con respecto a la línea de regresión usando la transformación logarítmica.

| AÑO | Desague (1) Y | Precipitación (2) X | X Y | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1. 928 | 125 | 110 | | | |
| 1. 929 | 67 | 73 | | | |
| 1. 930 | 68 | 74 | | | |
| 1. 931 | 71 | 91 | | | |
| 1. 932 | 118 | 108 | | | |
| 1. 933 | 144 | 130 | | | |
| 1. 934 | 169 | 152 | | | |
| 1. 935 | 138 | 134 | | | |
| 1. 936 | 102 | 98 | | | |
| 1. 937 | 91 | 90 | | | |
| 1. 938 | 125 | 119 | | | |
| 1. 939 | 87 | 77 | | | |
| 1. 940 | 84 | 100 | | | |
| 1. 941 | 58 | 84 | | | |
| 1. 942 | 79 | 85 | | | |
| 1. 943 | 124 | 115 | | | |
| 1. 944 | 62 | 70 | | | |
| 1. 945 | 87 | 91 | | | |
| | 1.799 | 1.801 | 192, 042 | 189, 291 | 197.373 |
| MEDIA | 99.44 | 100.06 | | | |

**Tabla N° 2. -** Datos y cómputos para el ejemplo de regresión de dos variables. -

1) Porcentaje de la media de desague anual (Río Buniping, cerca del Nile en Washington.

2) Porcentaje de la media de precipitación anual (Lago Bumping).

REGRESION LINEAL SIMPLE. -

Usando los datos dados en la tabla N° 2, se demuestra como se-computa la ecuación de una regresión por medio del modelo Y= a + bX. Esa tabla también muestra el cómputo de las medias, de los productos, y de los cuadrados. No hay que registrar el producto ni la raíz; la suma-de los productos o de los cuadrados se pueden acumular en una calcula-dora de mesa. Estos cálculos se verifican a menudo repitiendo la operación. Los coeficientes a y b de la ecuación de regresión y el error promedio de la estimación se computan según señalamos a continuación:

$$b = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}} = \frac{\Sigma XY - N\bar{X}\bar{Y}}{\Sigma X^2 - N\bar{X}^2}$$

$$b = \frac{192.042 - \frac{(1.801)(1.799)}{18}}{189.291 - \frac{(1.801)^2}{18}} = 1,325$$

Coeficiente de regresión. -

$$a = \bar{Y} - b\bar{X} = 99,94 - 1,325\ (100,06) = -32,6$$

Intercepta. entonces a:

$$Y = a + bX = -32,6 + 1,32\ X$$

ó:

$$Y = \bar{Y} - b ( X - \bar{X} ) = 99,94 - (1,325) ( X - 100,06)$$

$$Y = -32,6 + 1,32 X$$

La ecuación de la línea del mínimo cuadrado

$$S_x^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N - 1} = \frac{189,291 - \frac{(1.801)^2}{18}}{17}$$

$$= 534,76 \text{ La variancia de X}$$

$$S_y^2 = \frac{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}{N - 1} = \frac{197.373 - \frac{(1.799)^2}{18}}{17}$$

$$= 1.033,71. \text{ La variancia de Y}$$

$$S_{yx}^2 = \frac{N - 1}{N - 2} \left( S_y^2 - b^2 S_x^2 \right) = \frac{17}{16} \cdot \left[ 1.033,71 - (1,325)^2 (534,76) \right]$$

$$= 100,8$$

$S_{yx} = 10.0$ El error standard de la estimación de Y.

Se puede probar la significación del coeficiente de regresión de la manera siguiente (Bennett y Franklin, 1.954, p. 228):

$$S_b^2 = \frac{S_{y.x}^2}{(\Sigma x)^2} = \frac{100,8}{189.291-(1,801)^2/18} = 0,056$$

Probando la hipótesis de que $\beta = 0$

$$t_{n-2} = \frac{b-\beta}{S_b} \quad \frac{1,325-0}{0,236} = 5,6$$

De una tabla de t se obtiene para $t_{16:0,01=2,92}$: por lo tanto, b es bastante diferente de cero. Los límites de confianza de $\beta$, del 99 por ciento, son: $1,325-2,92(0,236)$ $\quad$ $1,325+2,92(0,236)$ ó:

$0,63$ $\quad$ $2,02$

El lugar geométrico de la ecuación de regresión y los datos usados se muestran en la figura N° 12.

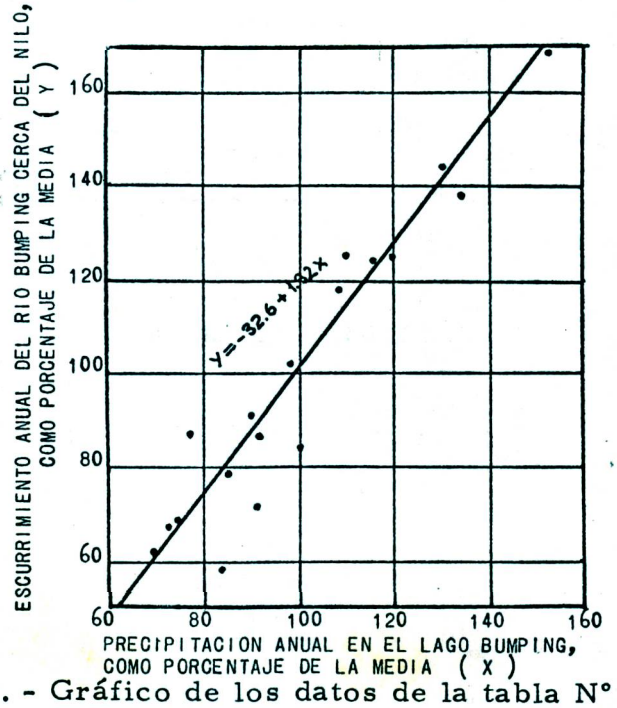


Figura N° 12. - Gráfico de los datos de la tabla N° 2, mostrando las lineas de regresión.

Otro ejemplo que muestra el cómputo detallado de la ecuación de la regresión lo dan Ezekiel y Fox (1.959, p. 57-63).

REGRESION LINEAL MULTIPLE. -

Las constantes de un modelo de regresión lineal múltiple se computan a partir de las ecuaciones normales. Para dos variables independientes las ecuaciones normales son:

$$\Sigma(x_2^2)\, b_2 + \Sigma(x_2 x_3)\, b_3 = \Sigma(x_1 x_2)$$

$$\Sigma(x_2 x_3)\, b_2 + \Sigma(x_3^2)\, b_3 = \Sigma(x_1 x_3)$$

y:

$$a = \bar{X}_1 - b_2 \bar{X}_2 - b_3 \bar{X}_3$$

en donde el símbolo $\bar{X}i$, representa la media de la i-ésima variable, $\bar{X}i$ representa un valor particular de la i-ésima variable y $x_i$, ( Xi - $\bar{X}i$), o sea la desviación de la media de la variable. Es más sencillo calcular los cuadrados y los productos de las variables en función de X y convertir los resultados en función x, que empezar con las desviaciones de la media.

Las ecuaciones de conversión son:

$$\Sigma(x_1 x_2) = \Sigma X_1 X_2 - N \bar{X}_1 \bar{X}_2 ,$$

$$\Sigma x_2^2 = \Sigma X_2^2 - N \bar{X}_2^2 ,$$

$$\Sigma x_1 x_3 = \Sigma X_1 X_3 - N \bar{X}_1 \bar{X}_3$$

$$\Sigma(x_2 x_3) = \Sigma X_2 X_3 - N \bar{X}_2 \bar{X}_3$$

y:

$$\Sigma(x_3^2) = \Sigma X_3^2 - N \bar{X}_3^2$$

en donde el último término de cada ecuación se denomina elemento de corrección y N es el número de elementos de la muestra. En esta notación, $X_1$ es la variable dependiente.

Para tres variables independientes las ecuaciones normales son:

$$\Sigma(x_2^2)\, b_2 + \Sigma(x_2 x_3)\, b_3 + \Sigma(x_2 x_4)\, b_4 = \Sigma(x_1, x_2)$$

$$\Sigma(x_2 x_3)\, b_2 + \Sigma(x_3^2)\, b_3 + \Sigma(x_3 x_4)\, b_4 = \Sigma(x_1, x_3)$$

$$\Sigma(x_2 x_4)\, b_2 + \Sigma(x_3 x_4)\, b_3 + \Sigma(x_4^2)\, b_4 = \Sigma(x_1 x_4)$$

y:

$$a = \bar{X}_1 - b_2\bar{X}_2 - b_3\bar{X}_3 - b_4\bar{X}_4$$

en donde los símbolos son los mismos de antes.

El método de computación se describe mejor por medio de un ejemplo. El modelo, los datos y los cómputos preliminares se muestran en la tabla N° 3. Nótese que los logarítmos de los valores originales son las variables que se relacionan. La necesidad de hacer esta transformación fué indicada en un análisis gráfico preliminar.

Sólo las sumas acumulativas de los productos y de los cuadrados se sacan en la calculadora y se registran en la tabla N° 3, no se necesitan los valores individuales. Los cálculos se realizan tomando cinco cifras decimales cuando las variables son logarítmos debido a que las sumas convertidas pueden ser pequeñas en relación con los números que se restan de otros. Los elementos de corrección que se muestran en la tabla N° 3, se obtienen a partir del último término de la ecuación de conversión apropiada. Para $X_2X_3$ la ecuación apropiada es:

$$\Sigma(x_2 x_3) = \Sigma(X_2 X_3) - N\bar{X}_2\bar{X}_3$$

y el elemento de corrección de la tabla N° 3 es: $N\bar{X}_2\bar{X}_3$

Si a $X_2X_3$ le restamos el elemento de corrección obtenemos que es la suma corregida de la tabla N° 3.

**Table 3.— Ejemplo de Regresión Múltiple: Características de flujo bajo en Tennessee**

[Model is Log $Q_{20} = \log a + b_2 \log Q_2 + b_3 \log A + b_4 \log S$]

| Estación | Descarga con un intervalo de $Q_2$ años | Area de drenaje en mi. cuadradas $A$ | Pendiente de la curva de recesión | Descarga Intervalo de recurrencia 20 años $Q_{20}$ | Log $Q_2$ $X_2$ | Log $A$ $X_3$ | Log $S$ $X_4$ | Log $Q_{20}$ $X_1$ | $X_1X_2$ | $X_2X_3$ | $X_1X_3$ | $X_2^2$ | $X_1X_4$ | $X_1X_3$ | $X_3^2$ | $X_3X_4$ | $X_4^2$ | $X_2^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hatchie-Stanton | 330 | 1940 | 83 | 240 | 2.51851 | 3.28780 | 1.91908 | 2.38021 | | | | | | | | | | |
| Hatchie-Bolivar | 184 | 1470 | 78 | 86 | 2.26482 | 3.15534 | 1.89051 | 1.93450 | | | | | | | | | | |
| Wolf-Rossville | 150 | 803 | 89 | 100 | 2.17609 | 2.79157 | 1.94939 | 2.00000 | | | | | | | | | | |
| S. F. Forked Deer-Jackson | 96 | 574 | 88 | 64 | 1.98227 | 2.75891 | 1.94448 | 1.80618 | | | | | | | | | | |
| S. F. Forked Deer-Chestnut Bluff | 153 | 1100 | 86 | 107 | 2.18469 | 3.04139 | 1.93450 | 2.02938 | | | | | | | | | | |
| M. F. Forked Deer-Alamo | 93 | 410 | 93 | 65 | 1.96848 | 2.61278 | 1.96848 | 1.81291 | | | | | | | | | | |
| Obion-Obion | 315 | 1830 | 86 | 210 | 2.49831 | 3.27416 | 1.93450 | 2.32222 | | | | | | | | | | |
| Rutherford Fork Obion-Bradford | 21.5 | 203 | 86 | 13 | 1.33244 | 2.30750 | 1.93450 | 1.11394 | | | | | | | | | | |
| S. F. Obion-Greenfield | 95 | 431 | 94 | 66 | 1.97772 | 2.63448 | 1.97313 | 1.81954 | | | | | | | | | | |
| N. F. Obion-Union City | 99 | 490 | 90 | 82 | 1.99564 | 2.69020 | 1.95424 | 1.91381 | | | | | | | | | | |
| Barren-Trousdale | 45 | 132 | 87 | 31 | 1.65321 | 2.12057 | 1.93952 | 1.49136 | | | | | | | | | | |
| Buffalo-Flatwoods | 138 | 447 | 81 | 85 | 2.13989 | 2.65031 | 1.90849 | 1.92942 | | | | | | | | | | |
| Buffalo-Lobelville | 223 | 707 | 80 | 140 | 2.34830 | 2.84942 | 1.90309 | 2.14612 | | | | | | | | | | |
| Big Sandy-Bruceton | 48 | 206 | 80 | 25.5 | 1.68124 | 2.31175 | 1.90309 | 1.40654 | | | | | | | | | | |
| Calfkiller-Sparta | 25 | 178 | 84 | 13.5 | 1.39794 | 2.25042 | 1.92428 | 1.13033 | | | | | | | | | | |
| Clear For.-Robbins | 4.5 | 278 | 25 | .4 | .65321 | 2.44404 | 1.39794 | −.39794 | | | | | | | | | | |
| Clinch-Tazewell | 212 | 1474 | 75 | 108 | 2.32634 | 3.16850 | 1.87506 | 2.03242 | | | | | | | | | | |
| Collins-McMinnville | 83 | 624 | 70 | 46 | 1.91925 | 2.79518 | 1.84510 | 1.66276 | | | | | | | | | | |
| Cypress-Florence | 61 | 209 | 90 | 47 | 1.78533 | 2.32015 | 1.95424 | 1.67210 | | | | | | | | | | |
| Doe-Elizabethton | 57 | 137 | 78 | 28 | 1.75587 | 2.13672 | 1.89209 | 1.44716 | | | | | | | | | | |
| Duck-Manchester | 16.5 | 107 | 75 | 9 | 1.21748 | 2.02938 | 1.87506 | .95424 | | | | | | | | | | |
| Elk-Estill Springs | 55 | 282 | 87 | 36 | 1.74036 | 2.45025 | 1.93952 | 1.55630 | | | | | | | | | | |
| Falling Water-Cookeville | 4.1 | 73.2 | 80 | 1.9 | .61278 | 1.86510 | 1.90309 | .27875 | | | | | | | | | | |
| Flint-Chase, Ala | 82 | 342 | 85 | 59 | 1.91381 | 2.53403 | 1.92942 | 1.77085 | | | | | | | | | | |
| French Broad-Newport | 850 | 1858 | 78 | 370 | 2.92942 | 3.26905 | 1.89209 | 2.56820 | | | | | | | | | | |
| L. Pigeon-Sevierville | 68 | 233 | 72 | 26 | 1.83251 | 2.54777 | 1.85733 | 1.41497 | | | | | | | | | | |
| Nolichucky-Embreeville | 355 | 806 | 60 | 140 | 2.55023 | 2.90580 | 1.77815 | 2.14613 | | | | | | | | | | |
| Piney-Vernon | 62 | 193 | 85 | 33 | 1.79239 | 2.28556 | 1.92942 | 1.51778 | | | | | | | | | | |
| Richland-Pulaski | 23.7 | 366 | 67 | 10 | 1.37475 | 2.56349 | 1.82607 | 1.00000 | | | | | | | | | | |
| Powell-Arthur | 109 | 685 | 76 | 64 | 2.03743 | 2.83569 | 1.92081 | 1.80618 | | | | | | | | | | |
| Red-Adams | 61 | 678 | 85 | 33 | 1.78533 | 2.83123 | 1.92942 | 1.51851 | | | | | | | | | | |
| Roaring-Hilham | 6 | 70.8 | 80 | 2.3 | .77815 | 1.85003 | 1.90309 | .36173 | | | | | | | | | | |
| Shoal-Iron City | 100 | 348 | 68 | 52 | 2.00000 | 2.54158 | 1.83251 | 1.71600 | | | | | | | | | | |
| Sequatchie-Whitwell | 48 | 384 | 74 | 24 | 1.68124 | 2.58433 | 1.86923 | 1.38021 | | | | | | | | | | |
| S. Chickamauga-Chickamauga | 105 | 428 | 78 | 73 | 2.02119 | 2.63141 | 1.89209 | 1.86332 | | | | | | | | | | |
| Sewee-Decatur | 19 | 117 | 79 | 10 | 1.27875 | 2.06819 | 1.89763 | 1.00000 | | | | | | | | | | |
| Stones-Smyrna | 15 | 552 | 59 | 4 | 1.17609 | 2.74194 | 1.77085 | .60206 | | | | | | | | | | |
| Toccoa-Dial, Ga | 150 | 177 | 85 | 110 | 2.17609 | 2.24797 | 1.92942 | 2.04139 | | | | | | | | | | |
| Turtletown-Turtletown | 13 | 26.9 | 86 | 10 | 1.23527 | 1.42975 | 1.94453 | 1.00000 | | | | | | | | | | |
| Tellico-Tellico Plains | 41 | 113 | 74 | 22 | 1.61278 | 2.07188 | 1.86923 | 1.34242 | | | | | | | | | | |
| Sums | | | | | 72.82342 | 101.78064 | 70.63122 | 61.85001 | 130.44215 | 137.37714 | 123.04731 | 203.96512 | 102.40038 | 163.22378 | 143.23712 | 117.54904 | 109.03084 | 140.97063 |
| Means | | | | | 1.80814 | 2.54469 | 1.89078 | 1.53858 | | | | | | | | | | |
| Correction items | | | | | | | | | 134.09980 | 136.75160 | 111.30040 | 259.05630 | 102.47020 | 156.65120 | 143.00200 | 118.28720 | 94.73600 | 130.77600 |
| Corrected sums | | | | | | | | | 6.33123 | .62554 | 11.74691 | 6.90632 | −.07852 | 6.57453 | .23512 | 1.15244 | 14.31084 | 10.21063 |

N = 40.

Esta y las otras sumas corregidas son sustituídas luego en la ecuación normal. A continuación se muestra como se computan los coeficientes de regresión, con la subsiguiente explicación:

N = 40

Ecuaciones Normales (ver Ezekiel, 1.950, p. 198):

$$\text{I} \quad \Sigma(x_2^2)\, b_2 + \Sigma(x_2 x_3)\, b_3 + \Sigma(x_2 x_4)\, b_4 = \Sigma(x_1 x_2)$$

$$\text{II} \quad \Sigma(x_2 x_3)\, b_2 + \Sigma(x_3)^2\, b_3 + \Sigma(x_3 x_4) b_4 = \Sigma(x_1 x_3)$$

$$\text{III} \quad \Sigma(x_2 x_4)\, b_2 + \Sigma(x_3 x_4) b_3 + \Sigma(x_4^2)\, b_4 = \Sigma x_1 x_4$$

$$\text{I} \quad 10,20183 b_2 + 6,38133 b_3 + 0,62554 b_4 = 11,74691$$

$$-b_2 - 0,625508 b_3 - 0,061316 b_4 = -1,5145$$

$$\text{II} \quad 6,38133 b_2 + 6,90632 b_3 - 0,06952 b_4 = 6,57458$$

$$(-0,625508) \quad \text{I} \quad -6,38133 b_2 - 3,99157 b_3 - 0,39128 b_4 = -7,34779$$

$$\Sigma_2 \quad 2,91475 b_3 - 0,46080 b_4 = -0,77321$$

$$\text{II}' \quad -b_3 + 0,158092 b_4 = 0,2657$$

$$\text{III} \quad 0,62554 b_2 - 0,06952 b_3 + 0,33512 b_4 = 1,16244$$

$$(-0,061316) \quad \text{I} \quad 0,62554 b_2 - 0,39128 b_3 - 0,03836 b_4 = -0,72027$$

$$(0,158092) \quad \Sigma_2 \quad 0,46080 b_3 - 0,07285 b_4 = -0,12224$$

$$\Sigma_3 \quad 0,22391 b_4 = 0,31993$$

$$b_4 = 1,42883$$

$$\text{II}' \quad -b_3 + (0,158092)(1.42883) = 0,26527$$

$$b_3 = -0,03938$$

$$\text{I}' \quad -b_2 - (0,625508)(-0,03938) - (0,061316)(1,42883) = 1,15145$$

$$-b_2 + 0,02463 - 0,08761 = -1,15145$$

III $b_2 = 1,08847$

$(0,62454)\ (1,08847) - (0,06952)\ (-0,03938) + (0,33512)(1,42883)$

$= 1,16245 \qquad 1,6244$ verificación.

Este cómputo hace uso del método de Doolittle, método simplificado para resolver ecuaciones simultáneas que tienen cierta simetría.

Las ecuaciones normales son las de las primeras tres líneas, luego viene la primera ecuación normal con sumas convertidas de la tabla N° 3, y sustitúidas en la ecuación. La línea 5 se obtiene dividiendo la ecuación precedente entre el coeficiente de $b_2$ con signo contrario, en la sexta línea está la segunda ecuación normal con sumas convertidas en la tabla N° 3 y sustituídas en la ecuación. La línea 7 se obtiene multiplicando la ecuación de la línea 4 por el coeficiente de $b_3$ en la línea 5. La línea 8 se obtiene restando la línea 7 de la 6. La línea 9, es la línea 8 dividida entre el coeficiente de $b_3$ con signo contrario. La línea 11, es la línea 4 multiplicada por el coeficiente de $b_4$ de la línea 5 con signo contrario. La línea 12, es la línea 8 multiplicada por el coeficiente de $b_4$ en la línea 9 con signo contrario. La línea 13 es la suma de las líneas 10, 11, y 12. En las líneas 14 a 18 se completan los cómputos de los coeficientes de regresión.

Las líneas 20 y 21 se usan para verificar los resultados, sólo la tercera ecuación normal proporciona una verificación completa..

La constante de regresión se obtiene de:

$$a = \overline{X}_1 - b_2 \overline{X}_2 - b_3 \overline{X}_3 - b_4 \overline{X}_4$$

$a = 1,53888 - (1,0884)(1,8014) - (-0,03938)(2,54489)$

$-(1,42883)(1,89078)$

$a = -3,03061$

Substituyendo las constantes computadas en el modelo de regresión se obtiene:

$$\log Q_{20} = 3,03 + 1,09 \log Q_2 - 0,04 \log A + 1,43 \log S$$

Tomando antilogarítmos esto se transforma en:

$$Q_{20} = 0,00093 Q_2^{1,09} A^{-0,04} S^{1,43}$$

El error standard de la estimación S, se computa de la manera siguiente:

$$S^2 = \frac{\Sigma(x_1)^2 - b_2 \Sigma(x_1 x_2) - b_3 \Sigma(x_1 x_3) - b_4 \Sigma(x_1 x_4)}{N - M}$$

En donde N es el número de elementos de la muestra y M es el número de grados de libertad que se han perdido. (En una ecuación de regresión se pierde un grado de libertad para cada constante. Substituyendo: $14,31085 - (1,088847)(11,74691)$

$$S^2 = \frac{(-0,03938)(6,57458) - (1,42883)(1,16244)}{40 - 4}$$

$S^2 = 0,00341$

$S = 0,0584$

= error standard en unidades logarítmicas.

El error standard de una regresión que tenga una variable dependiente logarítmica es un porcentaje constante del valor de la curva en todo el campo de Y antes que una magnitud constante en función de la variable no transformada como en el ejemplo de la tabla N° 2.

Para computar el error standard en porcentaje véanse los antilogarítmos de 1 + S y de 1 - S. Estos antilogarítmos son submultiplos de diez, de donde es obvia la desviación de porcentaje. Consideremos el error standard de la unidad logarítmica 0,0584 que computamos anteriormente:

1 + S = 1,0584 antilog: 11,4

1 - S = 0,9416 antilog: 8,75

Los errores de porcentaje son:

100(11,4 - 10)/ 10 = + 14 por ciento

y: 100(10 - 8,75)/10 = - 12,5 por ciento

El cómputo se puede hacer mucho más rápidamente con una regla de cálculo. Para este problema no se computa un coeficiente de correlación porque (1) el propósito del problema es obtener una ecuación de estimación y (2) no se pueden considerar los datos usados como sacados de una distribución normal multivariada; por lo tanto, la correlación no es apropiada y no tendría significado el coeficiente de correlación computado.

El error standard de estimación de esta regresión es una medida de su fiabilidad y se puede usar para estimar la fiabilidad de las predicciones de ducidas de las ecuaciones de regresión según se describió enla

sección sobre "Correlación y Regresión". Pero puede surgir la duda de si se podría obtener un resultado tan bueno usando pocas variables, o si cada una de las variables independientes está relacionada con la variable dependiente. Podríamos responder a la primera pregunta recomputando las ecuaciones de regresión y los errores standard usando menos variables; pero para responder a la segunda, necesitamos hacer una prueba de significación de cada coeficiente de regresión, para hacer esto hay que calcular la regresión de manera algo diferente, según se describe en la siguiente sección.

CALCULO DE REGRESION USANDO MULTIPLICADORES "C". -

En este método las ecuaciones normales se expresan en función de multiplicadores "c", en vez de usar los coeficientes de regresión. El método tiene dos ventajas: 1) Las pruebas de significación de los coeficientes de regresión se realizan de manera sencilla y, 2) se pueden obtener los coeficientes de regresión para diferentes variables dependientes a partir de los mismos multiplicadores "c".

Ezekiel y Fox (1.959, p.499-503), Fisher (1.950, p.156-166), y Bennett y Franklin (1.954, p.248-255) han descrito este método. Las ecuaciones normales son:

$$C_{22} \Sigma (x_2)^2 + C_{23} \Sigma (x_2 x_3) + C_{24} \Sigma (x_2 x_4) = 1$$

$$C_{22} \Sigma (x_2 x_3) + C_{23} \Sigma (x_3)^2 + C_{24} \Sigma (x_3 x_4) = 0$$

y: $$C_{22} \Sigma (x_2x_4) + C_{23} \Sigma (x_3x_4) + C_{24} \Sigma (x_4^2) = 0$$

Se puede obtener $C_{32}, C_{33}, C_{34}$ y $C_{42}, C_{43}, C_{44}$ se obtienen resolviendo ecuaciones similares reemplazando los miembros de la derecha por: 0, 1, 0 y 0, 0, 1 respectivamente.

Entonces se pueden evaluar los coeficientes de regresión por medio de las ecuaciones:

$$b_2 = C_{22} \Sigma (x_1x_2) + C_{23} \Sigma (x_1x_3) + C_{24} \Sigma (x_1x_4)$$

$$b_3 = C_{32} \Sigma (x_1x_2) + C_{33} \Sigma (x_1x_3) + C_{34} \Sigma (x_1x_4)$$

y: $$b_4 = C_{42} \Sigma (x_1x_2) + C_{43} \Sigma (x_1x_3) + C_{44} \Sigma (x_1x_4)$$

en donde $X_1$ es la variable dependiente.

Para probar la significación de los coeficientes de regresión compútese primero la varianza $S^2$ de las observaciones de $X_1$ con respecto a la superficie de regresión. Esta varianza resulta ser el cuadrado del error standard de estimación y se obtiene con la misma fórmula usada en los cómputos previos, es decir:

$$S^2 = \frac{\left[\Sigma x_1^2 - b_2 \Sigma (x_1x_2) - b_3 \Sigma (x_1x_3) - b_4 \Sigma (x_1x_4)\right]}{N - M}$$

luego,

varianza de $b_2 = S^2 (C_{22})$

varianza de $b_3 = S^2 (C_{33})$

y: varianza de $b_4 = S^3 (C_{44})$

y los errores standard son las raices cuadradas de las varianzas.

El intervalo de confidencia para $\beta_2$, coeficiente de regresión de la población, a un nivel de probabilidad se puede expresar como:

$$b_2 - t_{N-4} S \sqrt{C_{22}} < \beta_2 < b_2 + t_{N-4} S \sqrt{C_{22}}$$

En donde $t_{N-4}$ procede de la distribución t con N - 4 grados de libertad al nivel seleccionado.

El coeficiente de regresión, $b_2$, es bastante diferente de cero si los límites de confidencia no incluyen al cero.

A continuación damos el cómputo de regresión usando los multiplicadores "c" y los datos de la tabla N° 3. Las soluciones de las ecuaciones normales siguen el mismo modelo según se describió previamente. Hallando $C_{22}$, $C_{23}$, $C_{24}$.

$$\text{I} \quad \Sigma (x_2^2) C_{22} + \Sigma (x_2 x_3) C_{23} + \Sigma (x_2 x_4) C_{24} = 1$$

$$\text{II} \quad \Sigma (x_2 x_3) C_{22} + \Sigma (x_3^2) C_{23} + \Sigma (x_3 x_4) C_{24} = 0$$

$$\text{III} \quad \Sigma (x_2 x_4) C_{22} + \Sigma (x_3 x_4) C_{23} + \Sigma (x_4^2) C_{24} = 0$$

$$\text{I} \quad 10,20183 C_{22} + 6,38133 C_{23} + 0,62554 C_{24} = 1$$

$$\begin{array}{lll}
I' & -C_{22} - 0,625508C_{23} - 0,061316C_{24} = -0,0980216 & \\
II & 6,38133C_{22} + 6,90632C_{23} - 0,06952C_{24} = 0 & \\
(-0,625508)\ I & -6,38133C_{22} - 3,99157C_{23} - 0,39128C_{24} = -0,625508 & \\
\Sigma_2 & 2,91475C_{23} - 0,46080C_{24} = -0,625508 & \\
II' & -C_{23} + 0,158092C_{24} = 0,214601 & \\
III & 0,62554C_{22} - 0,06952C_{23} + 0,33512C_{24} = 0 & \\
(-0,061316)\ I & -0,03836C_{24} = -0,061316 & \\
(0,158092)\ \Sigma_2 & -0,07285C_{24} = -0,098888 & \\
\Sigma_3 & 0,22391C_{24} = -0,160204 &
\end{array}$$

$$C_{24} = -0,71548$$

$$II' \quad -C_{23} + 0,158092(-0,71548) = 0,214601$$

$$C_{23} = -0,32771$$

$$I' \quad -C_{22} - (0,625508)(0,32771) - (0,061316)(-0,71548) = -0,0980216$$

$$-C_{22} + 0,204985 + 0,043870 = -0,0980216$$

$$C_{22} = 0,34688$$

$$III \quad (0,62554)(0,34688) - (0,06952)(-0,32771) + (0,33512)(-0,71548) = 0$$

$0 \quad 0$ Se verifica (hasta cinco cifras).

Para hallar a $C_{32}$, $C_{33}$ y a $C_{34}$:

$$I \quad \Sigma(x_2^2)C_{32} + \Sigma(x_2 x_3)C_{33} + \Sigma(x_2 x_4)C_{34} = 0$$

$$II \quad \Sigma(x_2 x_3)C_{32} + \Sigma(x_3^2)C_{33} + \Sigma x_3 x_4 C_{34} = 1$$

| | | | | | |
|---|---|---|---|---|---|
| | III | $\Sigma (x_2x_4)C_{32}$ + | $\Sigma x_3x_4C_{33}$ + | $\Sigma (x_4)^2 C_{34}$ | = 0 |
| | I | 10,20183 $C_{32}$ + | 6,38133 $C_{33}$ + | 0,62554 $C_{34}$ | = 0 |
| | I′ | $-C_{32}$ - | 0,625508 $C_{33}$ - | 0,061316 $C_{34}$ | = 0 |
| | II | 6,38133 $C_{32}$ + | 6,90632 $C_{33}$ - | 0,06952 $C_{34}$ | = 1 |
| (-0,625508) | I | -6,38133 $C_{32}$ - | 3,99157 $C_{33}$ - | 0,39128 $C_{34}$ | = 0 |
| | $\Sigma_2$ | | 2,91475 $C_{33}$ - | 0,46080 $C_{34}$ | = 1 |
| | II′ | | $-C_{33}$ + | 0,158092 $C_{34}$ | = -0,343082 |
| | III | 0,62554 $C_{32}$ - | 0,06952 $C_{33}$ + | 0,33512 $C_{34}$ | = 0 |
| (-0,061316) | I | | | - 0,03836 $C_{34}$ | = 0 |
| (0,158092) | $\Sigma_2$ | | | - 0,07285 $C_{34}$ | = 0,158092 |
| | $\Sigma_3$ | | | 0,22391 $C_{34}$ | = 0,158092 |
| | | | | $C_{34}$ | = 0,70605 |

II′ $-C_{33}+(0,158092)(0,70605) = 0,343082$

$C_{33} = 0,45470$

I′ $-C_{32}-(0,625508)(0,45470)-(0,061316)(0,70605)=0$

$-C_{32}-0,28442-0,04329=0$

$C_{32} = -0,32771$

III $(0,62554)(-0,32771)-(0,06952)(0,45470)+(0,33512)(0,70605)=0$

$0 = 0$ verificación (hasta cinco cifras)

Para hallar a $C_{42}$, $C_{43}$ y $C_{44}$ :

$$\text{I} \quad \Sigma(x_2^2)C_{42} + \Sigma(x_2x_3)C_{43} + \Sigma(x_2x_4)C_{44} = 0$$

$$\text{II} \quad \Sigma(x_2x_3)C_{42} + \Sigma(x_3^2)C_{43} + \Sigma x_3x_4C_{44} = 0$$

$$\text{III} \quad \Sigma x_2x_4C_{42} + \Sigma(x_3x_4)C_{43} + \Sigma(x_4)^2C_{44} = 1$$

$$\text{I} \quad 10,20183C_{42} + 6,38133C_{43} + 0,62554C_{44} = 0$$

$$\text{I}' \quad -C_{42} + 0,625508C_{43} - 0,061316C_{44} = 0$$

$$\text{II} \quad 6,38133C_{42} + 6,90632C_{43} - 0,06952C_{44} = 0$$

$$(-0,625508)\ \text{I} \quad -6,38132C_{42} - 3,99157C_{43} - 0,39128C_{44} = 0$$

$$\Sigma_2 \quad 2,91475C_{43} - 0,46080C_{44} = 0$$

$$\text{II}' \quad -C_{43} + 0,158092C_{44} = 0$$

$$\text{III} \quad 0,62554C_{42} - 0,06952C_{43} + 0,33512C_{44} = 1$$

$$(-0,061316)\ \text{I} \quad -0,03836C_{44} = 0$$

$$(0,158092)\ \Sigma_2 \quad -0,07285C_{44} = 0$$

$$\Sigma_3 \quad 0,22391C_{44} = 1$$

$$C_{44} = 4,46608$$

$$\text{II}' \quad -C_{43} + 0,158092\,(4,46608) = 0$$

$$C_{43} = 0,70605$$

$$\text{I}' \quad -C_{42} - (0,625508)(0,70605) - (0,061316)(4,46608) = 0$$

$$-C_{42} - 0,44164 - 0,27384 = 0$$

$$C_{42} = -0,71548$$

$$\text{III} \quad (0,62554)(-0,71548) - (0,06952)(0,70605) + (0,33512)(4,46608) = 1$$

$$1,00003 \cong 1 \text{ verificación}$$

Computemos ahora los coeficientes b y comparemos estos resultados con los que habíamos computado previamente:

$b_2 = C_{22} \Sigma (x_1 x_2) + C_{23} \Sigma (x_1 x_3) + C_{24} \Sigma (x_1 x_4)$

$= (0,34688)(11,74691) + (-0,32771)(6,57458) + (-0,71548)(1,16244),$

$b_2 = 1,08851$ ( que verifica el valor previo de 1,08847)

$b_3 = C_{32} \Sigma (x_1 x_2) + C_{33} \Sigma (x_1 x_3) + C_{34} \Sigma (x_1 x_4),$

$= (-0,32771)(11,74691) + (0,45470)(6,57458) + (0,70605)(1,16244),$

$b_3 = -0,03938$ (que verifica el valor -0,03938 calculado previamente)

$b_4 = C_{42} \Sigma (x_1 x_2) + C_{43} \Sigma (x_1 x_3) + C_{44} \Sigma (x_1 x_4),$

$= (-0,71548)(11,74691) + (0,70605)(6,57458) + (4,46608)(1,16244),$

$b_4 = 1,42885$ (que verifica el valor 1,42883 de la computación previa).

El coeficiente a se computaría como se describió previamente.

Computemos el error standard de las b:

$S_{1,234} = 0,0584$ del computo previo

$S_{b_2} = S_{1,234}\sqrt{C_{22}} = o,0584\sqrt{0,34688} = (0,0584)(0,589),$

$= 0,0344$ ; error standard de $b_2$.

$S_{b_3} = S_{1,234}\sqrt{C_{33}} = 0,0584\sqrt{0,4547} = (0,0584)(0,6743)$

$= 0,0394$; error standard de $b_3$

$S_{b_4} = S_{1,234}\sqrt{C_{44}} = 0,0584\sqrt{4,466} = (0,0584)(2,113),$

$= 0,1234$; error standard de $b_4$.

Cómputo de los intervalos de confidencia de los coeficientes (Bennet y Franklin; 1.954, p. 250)

$t_{36;\,0,95} = 2,93$ (Dixon y Massey; 1957, tabla A -5 , p. 384).

La tabla presentada por Dixon y Massey (1,957; tabla A-5, p.384)

da los valores de la mitad de la distribución. Para límites del 95 por ciento sería 0,025 en cada extremo, y se toma el valor de la columna marcada $t_{0,975}$. Nótese que para infinitos grados de libertad la t y las distribuciones normales son las mismas. En la distribución normal, el valor 1,96 a ambos lados de la media incluye el 96 por ciento de los elementos. En la tabla A-5, el valor $1,96\sigma$ aparece en la columna $t_{0,975}$.

Los límites de confidencia son:

$$b_2 - (t_{36;\,0,95})(S_{b_2}) < \beta_2 < b_2 + (t_{36;\,0,95})S_{b_2}$$

$$1,08851 - (2,03)(0,0344) < \beta_2 < 1,08851 + (2,03)(0,0344)$$

$$1,0187 < \beta_2 < 1,1583$$

$$-0,03938 - (2,03)(0,0394) < \beta_3 < -0,03938 + (2,03)(0,0394)$$

$$-0,1194 < \beta_3 < 0,0406$$

$$1,42885 - (2,03)(0,1234) < \beta_4 < 1,42885 + (2,03)(0,1234)$$

$$-1,1783 < \beta_4 < 1,6793$$

Se considera a $\beta$ como la verdadera inclinación. Por lo tanto, los límites de confianza dan el campo dentro del cual está $\beta$ con un 96 por ciento de probabilidad. En este ejemplo los límites de $\beta_3$ incluyen al cero.

Esto indica que $\beta_3$ no es muy diferente de cero a un nivel de 95 por ciento y que el parámetro A debe ser sacado de la regresión.

## REGRESIONES QUE TIENEN VARIAS VARIABLES INDEPENDIENTES

Se han dado ejemplos de cómputos de regresión que tienen una y

tres variables independientes, y de las ecuaciones normales para una regresión con dos variables independientes. El método de solución para el caso de las dos variables fué dado por Ezekiel y Fox (1.959, p. - 181 - 183). Debido al hecho de que el cómputo de tales regresiones, realizado con una calculadora corriente, toma mucho tiempo, se usan los computadores digitales.

USO DE LOS COMPUTADORES DIGITALES. -

Existen programas para el cómputo de regresiones para la mayoría de los computadores, y, generalmente, el cómputo de las regresiones de más de dos variables independientes se debe realizar con un computador digital en vez de usar una calculadora corriente. Las regresiones sencillas y las de dos variables independientes se podrían hacer con relativa rapidez en una calculadora corriente; en este caso podría ser ventajoso su uso.

Los programas de regresión para los computadores digitales varían, pero generalmente requieren el alistamiento de los datos en notación de punto flotante. Estos valores se perforan entonces en tarjetas que se introducen a la computadora. La computadora imprime los resultados. Se dispone de una amplia variedad de opciones en lo que respecta a la salida.

Se deben obtener instrucciones detalladas para la preparación de-

datos, e instrucciones para el computador en cuestion y para el programa usado.

Aunque no es necesario ningún conocimiento acerca del análisis de regresiones en la preparación de datos de un programa para un computador, se necesita cierta experiencia para apreciar los resultados. La oportunidad de que en el proceso se introduzcan errores existe en el alistamiento de datos y en su transferencia a las tarjetas.

También se pueden obtener resultados cuestionables si en los computos se usan pocas cifras significantes. Sólo aquellos que han realizado cómputos de regresión por el método más difícil pueden juzgar adecuadamente si los resultados de un análisis de regresión realizado con computador (o cualquier otro método) son correctos. La disponibilidad de los computadores digitales ha permitido el cómputo inmediato de las regresiones usando muchas variables, lo que muchas veces ha dado como resultado la sustitución del computador por el cerebro del analista. El problema debe ser resuelto por el analista; el computador realiza la aritmética.

APLICACION DEL METODO DE REGRESION. -

Un problema analítico que deba ser resuelto por regresión incluye:

1. -La selección de los factores que se espera que influyan en la variable dependiente.
2. - La descripción cuantitativa de estos factores.

3. - La selección del modelo de regresión.

4. - El cómputo de la ecuación de regresión, el error standard de estimación y del significado de los coeficientes de regresión.

5. - De los resultados de la evaluación.

La selección de los factores apropiados no debiera ser un problema estadístico, sino que en el proceso deben introducirse los conceptos estadísticos. Si el analista solo quiere saber la relación entre la precipitación anual y el desague puede proceder directamente a la selección de un modelo. Pero si su problema es el de hacer la mejor estimación posible del desague, incluirá otros factores, algunos de los cuales pueden estar relacionados entre sí o con el desague. El problema de la determinación de si ciertos factores están relacionados con la variable independiente necesita de la cuidadosa selección de los índices que describen a estos factores cuantitativamente.

Estos índices deben reflejar los efectos con exactitud, y no deben existir dos que señalen lo mismo. Una característica de la regresión es que si un factor está relacionado con la variable dependiente y se introduce dos veces en el modelo de regresión (como dos variables diferentes), el efecto causado sobre la variable dependiente estará dividido por igual entre las dos.

Por lo tanto, si el efecto total es pequeño, el resultado de dividirlo en dos partes puede ser el de ocasionar falta de significado en cada una de las partes.

De igual manera, varias variables que estén muy estrechamente relacionadas se pueden computar como insignificantes, mientras que un índice adecuadamente seleccionado mostraría un efecto real. Luego las variables independientes se deben seleccionar con extremo cuidado, no se debe usar el enfoque de escopeta.

Otra precaución que se debe tomar al seleccionar la variable es la de evitar tener una variable, o parte de ella, a ambos lados de la ecuación. Tal condición puede ser aceptable para ciertos problemas, pero se deben evaluar cuidadosamente los resultados. Puede resultar una relación espúrea, o la relación puede ser correcta pero es díficil apreciar su fiabilidad. Bensón describió (en 1.965) la manera de construir relaciones espúreas en una regresión.

El que usa el método de regresión debe de entender el efecto de variables independientes relacionadas entre sí sobre los coeficientes de regresión computados. Si las variables independientes no están relacionadas bajo ningún aspecto, los coeficientes simples de regresión y los correspondientes coeficientes parciales serían los mismos. Sin embargo raras veces se presentan estas condiciones en la naturaleza. El método de regresión múltiple proporciona la manera de separar el efecto total de las variables independientes y de un efecto no explicado. Considérese la regresión simple:

$$Y = a + b_1 X_1 \pm \text{error}, \qquad (1)$$

en donde Y también está afectada por otra variable, $X_2$, que está relacionada con $X_1$. La regresión que emplea $X_1$ y $X_2$ será:

$$Y = a' + b_1' X_1 + b_2 X_2 \pm \text{error}, \qquad (2)$$

en donde $b_1' \neq b_1$. Si $X_1$ y $X_2$ son las únicas variables que afectan la Y (efectos lineales), entonces la ecuación (2) describe a Y por completo y, $b_1$ y $b_2$ son los valores verdaderos de los coeficientes de regresión (excepto por los errores de muestreo). Si $X_1$ y $X_2$ están correlacionadas positivamente entre sí y con Y, considérese el efecto de la magnitud de $b_1$. Para cada valor de $X_1$ de la ecuación 1, la Y parecerá más estrechamente relacionada de lo que está en realidad puesto que $X_2$ aumenta con $X_1$ y su influencia sobre Y es real aunque no medida. Por lo tanto, el coeficiente de regresión $b_1$ es mayor que su verdadero valor $b_1'$...Si se incluyese en la regresión otro factor, relacionado con $X_1X_2$ e Y, ocurrían cambios similares en $b_1$ y $b_2$. Estos cambios en la magnitud de los coeficientes de regresión, debido a la adición o supresión de una variable, son característicos de la regresión, algunas veces se interpretan como indicación de la insignificancia física de los coeficientes parciales de regresión. Estas interpretaciones no son necesariamente correctas. Si las variables usadas en la regresión se seleccionan de acuerdo con principios físicos y es apreciable el efecto de cada variable, entonces los coeficientes parciales de regresión deben de concordar con los principios físicos. En realidad, es buena práctica comparar el signo y la magnitud general de cada coeficien-

te parcial de regresión con lo que se espera. Benson hizo (en1. 962-p. 52-55) una comparación completa de esta clase.

Los coeficientes de regresión de ciertas variables pueden cambiar de signo cuando otra variable relacionada se añade o se suprime a la regresión. Este efecto puede resultar porque (1) la variable no es un buen índice de la característica física representada, (2) el efecto de la variable es pequeño en relación con el error de muestreo, (3) una variable está tan correlacionada con una o más de las otras variables de la regresión que el verdadero efecto se divide entre ellas y ninguna muestra un efecto notable y (4) el campo de la variable del muestreo puede ser demasiado pequeño para definir a un efecto importante.

Una ecuación de regresión no implica una relación de causa y efecto entre las variables independientes y la variable dependiente. Ambas pueden estar influenciadas por algún otro factor que no se mida en el acto. Sin embargo, debe existir algún lazo físico entre las variables si se puede considerar que los resultados tienen algún significado.

La selección de un modelo de regresión comienza generalmente con un análisis gráfico. Comunmente se usa un modelo que se representa como una línea recta a menos que haya evidencia de lo contrario.

Si próximo a la asíntota o a un punto de máximo o mínimo de la curva existen datos de la muestra, puede no ser adecuado un modelo sen

cillo para describir la relación y puede que no se justifique uno más sofisticado a menos que existan muchos datos disponibles. En la figura 13 se da un ejemplo que muestra las características de tres modelos comunes cuando se aplican a datos definidos cerca del cero. Las consideraciones físicas sugieren que ni b ni $Q_7$ deben ser menores que cero y que la línea debería curvarse.

La limitación del cero se puede conseguir usando las variables - log b y log $Q_7$ haciendo así a la curva asintótica con respecto a ambos ejes. La adición del término $(\log Q_7)^2$ proporcionará la curvatura necesaria. La ecuación de regresión usando tres variables es la superior de la figura 13. No se ajuste bien a los datos.

Supongamos ahora que no es necesario que la curva sea asíntota de $Q_7 = 0$, entonces sería apropiado un modelo semilogarítmico usando- las variables log b, $Q_7$, y $Q_7^2$. Pero la ecuación basada en este modelo- alcanza su máximo muy pronto y no se ajusta muy bien a los datos. Como último recurso supongáse un modelo sencillo con las variables b, $Q_7$, y $Q_7^2$. Esta ecuación si se ajusta a los datos, debido en gran parte a la- posición de éstos. Un punto adicional b=0 cuando $Q_7$=10, ó más, habría hecho descender la curva por debajo del punto b=0 $Q_7$=7. La curva que se- muestra en la figura 13 alcanza un mínimo en $Q_7$ =7 y de allí en adelante aumenta. Se ha descrito la mecánica del cómputo de la ecuación de regresión, del error standard y de las pruebas de significancia. Queda una labor importante, la de evaluar los resultados.

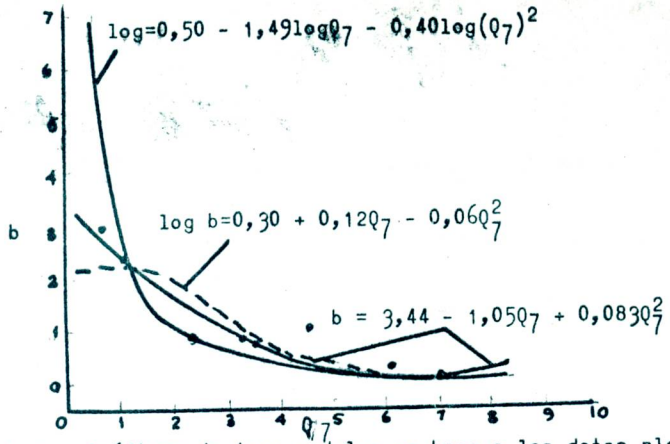


FIGURA 13. Ecuaciones y Gráficos de tres modelos en base a los datos ploteados.

Los analistas deben de reconocer que la ecuación de regresión desarrollada, aunque se ajusta a los datos, no es necesariamente correcta para hacer una extrapolación. Por ejemplo, la curva que corresponde a la ecuación inferior de la figura 13 se ajusta bien a los siete puntos pero aumenta en proporción directa a $Q_7$ para los valores de $Q_7$ mayores que 7.

Por otra parte, la curva de trazos se ajusta a los 5 puntos inferiores pero se hace asíntota de cero a medida que $Q_7$ aumenta. La información disponible no indica cual es la extrapolación más correcta.

Los signos de todos los coeficientes significativos de regresión deben de estar de acuerdo con los principios físicos. La regresión no es necesariamente incorrecta si los signos no lo son, la inconformidad puede deberse a las interrelaciones entre las variables independientes. Tal regresión es útil en la estimación de los valores de las variables dependientes a partir de valores conocidos de las variables independientes, y la fiabilidad de los resultados se puede computar si caen dentro del campo que se ha definido para la regresión.

El problema más difícil, el de establecer si en particular una variable.

ble determinada está relacionada con la variable dependiente, puede no tener una respuesta definida. Aún cuando un coeficiente de regresión - sea estadísticamente importante hay poca probabilidad de que este resultado se obtenga al azar.

Otros ejemplos podrían producir resultados conflictivos. Por otra parte, si muchas regresiones dan coeficientes no significativos de- una variable determinada, teniendo todos los coeficientes el mismo - signo, podríamos concluir que el efecto de esa variable es real pero, - naturalmente, escaso.

Se debe hacer distinción entre el significado estadístico y el práctico. El coeficiente de regresión de una variable puede probar ser de mucho significado, y sin embargo el efecto de esa variable sobre la variable independiente puede ser despreciable.

Riggs (2n. 1. 960) y, Amorocho y Hart (2n. 1. 964) han discutido el uso y las interpretaciones del análisis de regresión en la hidrología.

REGRESION GRAFICA. -

Las suposiciones que exige la regresión gráfica son las mismas - que las de la regresión analítica. Los resultados de una regresión gráfica se pueden expresar matemáticamente si no se añaden restricciones - al análisis gráfico, y se puede estimar el error standard.

La regresión gráfica es menos restrictiva que la analítica en el-

hecho de que el modelo no tiene que especificarse completamente por adelantado. En efecto, si un modelo analítico no se puede seleccionar en base a una característica física, es convencional preparar una regresión preliminar gráfica que indicará un modelo apropiado. Por ejemplo considérese los cuatro grupos de datos de la figura 14. El primero (el superior de la izquierda indica el uso del modelo).

$$Y = a + bX$$

El segundo (el de la derecha ) exige:

$$Y = a + bX + b_1X^2$$

en donde la dirección de la curvatura determina el signo de $b_1$. El tercero (el inferior de la izquierda) indica la necesidad de una transformación a menos que se pueda corregir la divergencia con una variable adicional, el cuarto gráfico no señala ninguna relación entre Y y X, y, si se considerase solamente una relación con dos variables no se haría ningún análisis posterior. Sin embargo, el efecto de otra variable Z que no se ha incluído podría oscurecer la relación entre Y y X en el cuarto gráfico .

Este aspecto se discute más adelante.

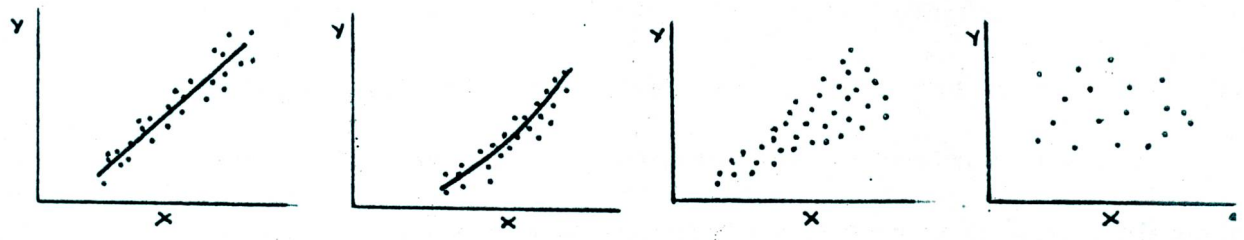


Figura N° 14. -Cuatro posibles resultados del gráfico Y contra X. -

La preparación de relaciones lineales simples entre dos variables se conoce muy bien. La línea de regresión no es necesariamente la misma que se trazaría siguiendo los puntos marcados. Hay dos líneas de regresión para $Y = f(X)$ y otra para $X = f(Y)$ (Fig. 15).

La línea estructural que balancea a los puntos marcados en ambas direcciones tiene una pendiente aproximada de valor intermedio al de las dos líneas de regresión. La Diferencia de pendiente entre las tres líneas depende del grado de correlación de las variables. Para que exista una correlación perfecta, las tres líneas deben de tener la misma pendiente. A pesar de la correlación, ambas líneas de regresión pasan por el punto que representa a la media; la línea estructural puede o no pasar por ese punto.

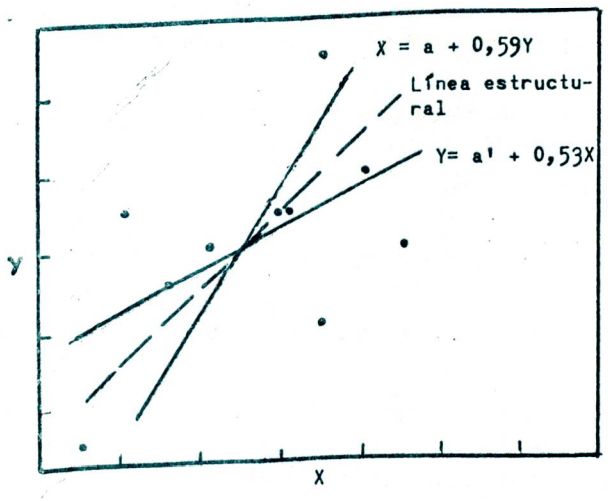


Figura 15.- Gráfico de las dos líneas de regresión y de la línea Estructural.

Para aproximar la regresión Y = f (X), (1) agrúpense los puntos en pequeños incrementos de X, (2) estímese la media de cada grupo en la dirección Y, y (3) trácese una línea que dé el promedio de estas medias. Este procedimiento se puede comprender refiriéndose a la figura 15 y recordando que la distribución de los puntos con respecto a la línea de regresión en la dirección de Y se supone que sea la misma en todo el campo. Obviamente esa suposición no puede ser cierta para un número escaso de puntos pero es la condición a la cual tratamos de aproximarnos.

La línea de regresión de Y=f (X) tendrá una pendiente menor que la de la línea trazada para balancear los puntos en ambas direcciones X e Y.

El error standard de estimación de una regresión gráfica se puede calcular inmediatamente. Recordando (1) que el error standard de la estimación es una desviación standard de puntos trazados con respecto a la línea de regresión, (2) que dos tercios de los puntos deben caer dentro de una desviación standard a ambos lados de la media de una distribución normal, y (3) que una línea de regresión pasa teoricamente por el valor medio de Y que corresponde a cualquier valor de X, luego dos líneas, paralelas a la línea de regresión y a una desviación standard por encima y por debajo (en la dirección de Y), debe abarcar los dos tercios

de los puntos trazados. En la práctica es más simple trazar líneas excluyendo un sexto de los puntos que se encuentran por encima y por debajo, y luego usar el promedio de estas desviaciones de la media como el error standard estimado. En la figura 16 se ilustra el procedimiento para una relación logarítmica. El error standard se puede describir en unidades logarítmicas pero más comunmente se expresa como un porcentaje. Este valor se obtiene inmediatamente usando divisores para establecer un error standard por debajo y por encima de una separación de ciclo si se traza la relación en papel logarítmico.

Los porcentajes se miden desde el 1 según se muestra en la figura N° 16.

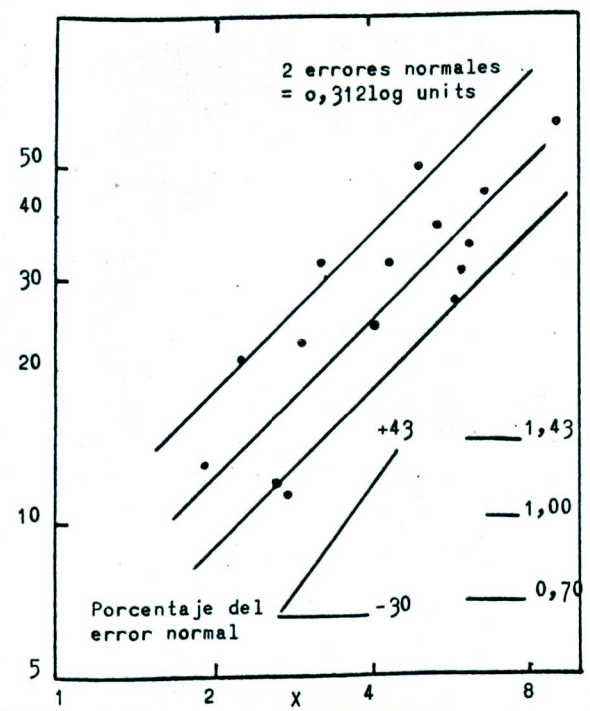


Figura 16.- Método de estimación del error standard de una regresión gráfica.

Para las regresiones con gráficos logarítmicos, el error standard estará en las mismas unidades que Y y, se puede ver en el gráfico.

La fiabilidad del error standard determinado gráficamente está influenciada por dos factores que tienen efectos opuestos. Si la línea de regresión gráfica tiene mayor pendiente que la línea de regresión del mínimo cuadrado, el error gráfico standard será mayor que el error - standard computado. Si ahora suponemos que la línea de relación gráfica es la misma que la línea de mínimos cuadrados, el error standard computado gráficamente subestimará al error standard computado cuando algunos de los puntos trazados están lejos de la línea, pero la mayoría está cerca.

En cualquier caso el error standard determinado gráficamente es solo una aproximación, pero es adecuado para muchos problemas.

El coeficiente de correlación se puede estimar tambien a partir de una regresión gráfica por la relación.

$$r = \sqrt{1 - S_e^2/S_d^2}$$

en donde $S_e$ es el error standard determinado gráficamente y $S_d$, la desviación standard de las variables Y con respecto a la media determinada de la misma manera que el error standard. Obviamente, el coeficiente de correlación debería estimarse solamente para las relaciones entre variables que razonablemente se puede suponer que se han - sacado de una distribución normal bivariada.

## REGRESION GRAFICA MULTIPLE

Existen dos métodos generales de regresión gráfica múltiple. El método de las desviaciones se basa en el modelo:

$$Y = a + b_1X_1 + b_2X_2 + \ldots\ldots\ldots\ldots\ldots + b_nX_n ,$$

o en un modelo similar que permita la curvilinealidad. Este método es probablemente el más simple y el más útil que se tenga a disposición. El método coaxial de regresión gráfica múltiple usado en las relaciones desague precipitación es más flexible que el de desviaciones porque permite la interacción y la curvilinealidad. Sin embargo, estas ventajas se obtienen a expensas de gran cantidad de trabajo adicional y con la pérdida de un método simple para evaluar la fiabilidad del resultado. Linsley y otros describieron (1.949, p. 650-655) el procedimiento detalladamente. A menos que se diga otra cosa, las descripciones de regresión gráfica múltiple de esta sección se refieren al método de desviaciones.

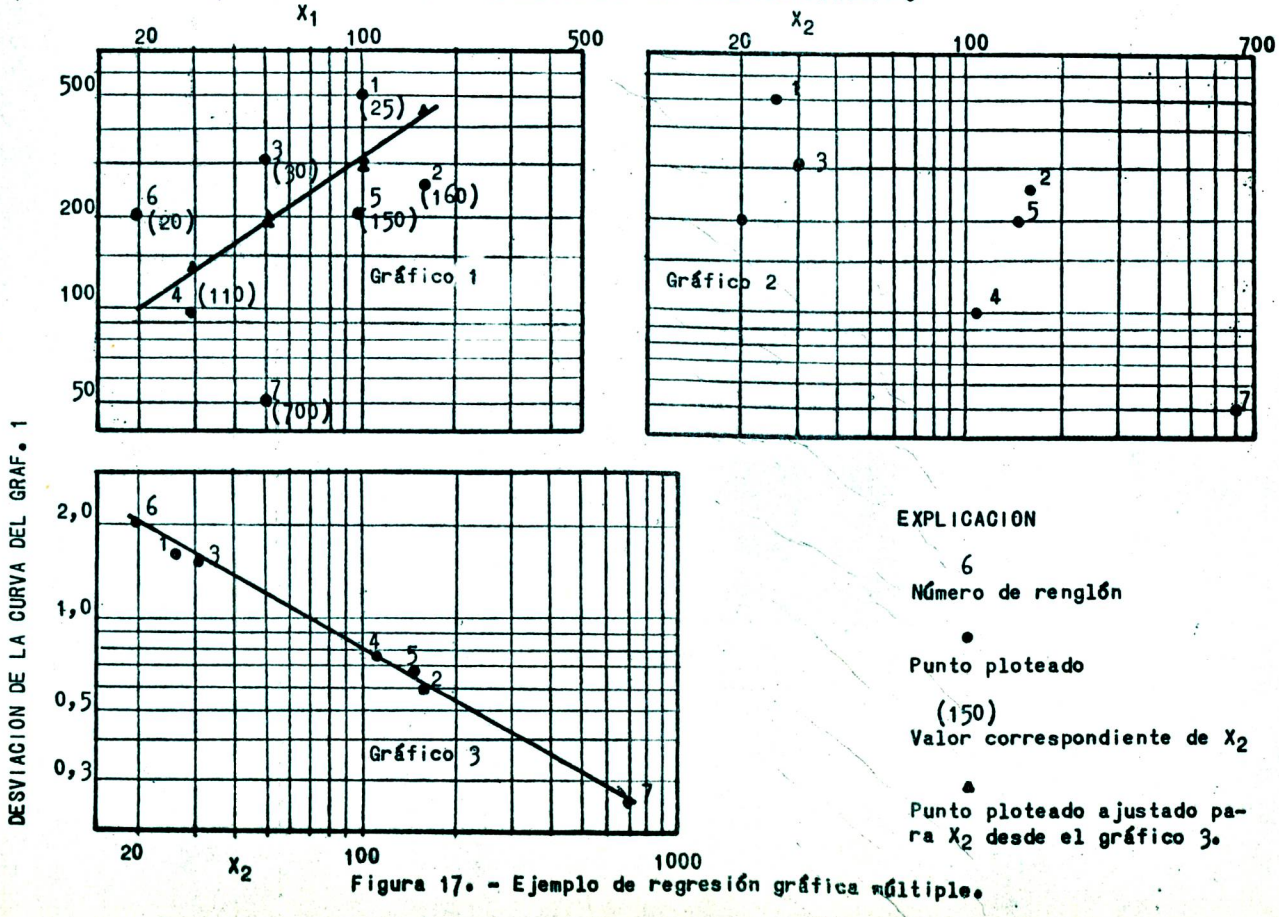


Figura 17. - Ejemplo de regresión gráfica múltiple.

El propósito de la regresión múltiple es el de determinar como cambia una variable dependiente cuando cambian dos ó más variables. Este problema no puede ser resuelto considerando una variable independiente aisladamente pués las variables independientes están generalmente correlacionadas entre sí de cierta manera. Este enunciado se puede verificar analizando los siguientes datos sintéticos:

| N° | Y | $X_1$ | $X_2$ |
|---|---|---|---|
| 1 | 500 | 100 | 25 |
| 2 | 250 | 150 | 160 |
| 3 | 300 | 50 | 30 |
| 4 | 100 | 30 | 110 |
| 5 | 200 | 100 | 150 |
| 6 | 200 | 20 | 20 |
| 7 | 50 | 50 | 700 |

Supongamos que los logarítmos de las variables, están linealmente relacionados. Esta relación exije un gráfico logarítmico. Hágase primero una comparación gráfica entre Y y $X_1$ llevando al papel los datos apropiados (véase el gráfico 1 de la fig. 17). (En estadística la variable dependiente se lleva generalmente sobre el eje de ordenadas). Hágase también un gráfico de Y contra $X_2$ (gráfico 2 de la figura 17). Estos gráficos indican que no se puede estimar confiadamente a Y a partir de cualquiera de los dos parámetros.

Determinemos ahora la relación entre Y y las otras dos variables. El procedimiento es el siguiente:

1. - En el gráfico de la figura 17, escriba al lado de cada punto el correspondiente valor de $X_2$. Se verá que los valores altos de $X_2$ tienden a estar a un lado del grupo y los valores bajos del otro lado. Esta condición es in-

dicación de que los valores de $X_2$ están relacionados con Y. Trace una línea recta que pasa por los puntos de tal manera que represente groseramente algún valor constante de $X_2$. La línea probablemente no balanceará a los puntos.

2. - Lleve al gráfico las desviaciones de Y (también llamadas residuos) desde la línea recta del gráfico 1 contra $X_2$ como abcisa en el gráfico 3 (Fig. N° 17). Las desviaciones se pueden trazar a escala a partir del gráfico 1 o se pueden transferir con divisores. Se deben medir por encima o por debajo del 1,00 en el gráfico 3, ya que son relaciones.

3. - Trácese una línea recta que promedie los puntos del gráfico 3.

4. - Midánse las desviaciones de los puntos de la curva del gráfico 3 y llévense al gráfico 1. Estas desviaciones se miden con respecto a la línea recta del gráfico 1 y definen la relación entre Y y $X_1$ habiéndose suprimido el efecto de $X_2$. Algunas veces estos puntos trasladados no se distribuyen al azar con respecto a la línea, en cuyo caso se debe trazar esta de nuevo y repetir todo el proceso. Cuando se consigue un balance satisfactorio se completa la regresión. La dispersión de los puntos ajustados con respecto a la línea del gráfico 1 es una medición del error. El error standard de una regresión gráfica múltiple se puede aproximar usando los puntos ajustados, según se describió en la sección sobre " Regresión Gráfica". La línea del gráfico 1 es la relación entre Y y $X_1$ para aquel valor de $X_2$ para el cual la línea del gráfico 3 cruza a la línea 1,0 ($X_2$ = 66). La relación de Y y $X_1$ para cualquier otro valor de $X_2$ será una línea paralela a la línea del gráfico 1, en una posición definida por la curva del gráfico 3, para el valor deseado de $X_2$.

El ejemplo usado dió mucho mejor resultado de lo que ordinariamente se esperaría en un análisis hidrológico. Se elaboraron los datos (1) para ilustrar el procedimiento, (2) para señalar que una buena relación no se puede reconocer si se estudian dos variables solamente.

Se pueden hacer regresiones gráficas que incluyen a más de dos variables independientes. Los residuos de cada línea son llevados a un gráfico en función de la siguiente variable hasta que se usen todas estas. Luego los residuos de la últi-

ma línea se vuelven a llevar a un gráfico desde el principio según se describió en el paso 4. En el trabajo práctico es generalmente difícil definir los efectos que originan más de tres variables independientes, particularmente cuando la influencia de una o dos es pequeña.

Se debe usar la regresión lineal siempre que los puntos llevados al gráfico no definan definitivamente a una curva y cuando no se conozca ninguna razón física para esperar una curvatura en la relación. Si una o varias curvas están indicadas por ambos criterios, se deben emplear dichas curvas. Las curvas complicadas exigen cuatro o más puntos para su definición. Se deben evitar cuando sólo se dispone de relativamente pocos puntos para definir la relación.

Las regresiones gráficas múltiples no tienen que ser elaboradas necesariamente en papel logarítmico. Los gráficos aritméticos se pueden manejar con la misma eficiencia. En la figura 18 se relaciona el desague de verano con el contenido acuoso de primavera en las nieves y con las precipitaciones en el verano. El procedimiento gráfico es el mismo que el del primer ejemplo, excepto que las desviaciones se miden en las mismas unidades que la variable dependiente y la escala de desviación debe tener su centro en cero con los valores positivos por encima y los negativos por debajo. Obviamente el modelo matemático que describe esta relación sería diferente de uno cuando la relación gráfica se desarrolla en papel logarítmico.

El papel seleccionado para un problema determinado debe ser aquel-

en el cual la distribución de la variable dependiente sea aproximadamente la misma para cualquier valor de la variable independiente.

Este criterio es más importante que el de conseguir la linealidad.

REGRESION GRAFICA MULTIPLE CUANDO LAS VARIABLES INDEPENDIENTES ESTAN MUY CORRELACIONADAS ENTRE SI. -

La figura N° 19 demuestra una técnica que es algunas veces útil en la regresión gráfica. En la tabla cuatro se dan los datos. La curva 1 (Fig. - 19) es la relación entre una creciente de 100 años ($Q_{100}$) y el flujo medio anual ($Q_{2.33}$) para 17 estaciones. El numerador de la fracción marcado en cada punto es la descarga media anual ($Q_{av}$) del río. Es imposible predecir por inspección si el uso de $Q_{av}$ mejorará la relación ya que esta descarga aumenta con $Q_{2,33}$. Para definir el efecto de $Q_{av}$, si es que existe alguno, sobre la dispersión de puntos con respecto a la curva 1, se puede usar el procedimiento siguiente:

1. - Llévense los valores de $Q_{av}$ contra los de $Q_{2,33}$ (como abcisa y trácese la línea media (curva 2).

2. - Divídase cada $Q_{av}$ por su valor en la curva 2 para el mismo valor de $Q_{2,33}$ estas divisiones se muestran para cada punto de la curva 1 del gráfico. Se podrían haber obtenido directamente midiendo las desviaciones con respecto a la curva 2 en porcentaje con relación a los divisores (sólo en papel logarítmico); en la práctica se obtendrían así:

3. - Use los dividendos obtenidos en el paso 2 como la tercera variable.

4. - Continúe con la regresión gráfica múltiple según se describió previamente. Los símbolos triangulares cerca de la curva 1 son los puntos a

justados por el efecto $Q_{av}$. El hecho de que muestren menos dispersión que los puntos originales, indica que las estimaciones de $Q_{100}$ se mejoran usando $Q_{av}$ como variable adicional. Computando la ecuación de la relación gráfica se puede demostrar que es de la forma:

$$\log Q_{100} = \log a + b_1 \log Q_{2,33} - b_2 \log Q_{av}$$

La introducción del dividendo $Q_{av}$ es sólo un artificio; no aparece en la relación final.

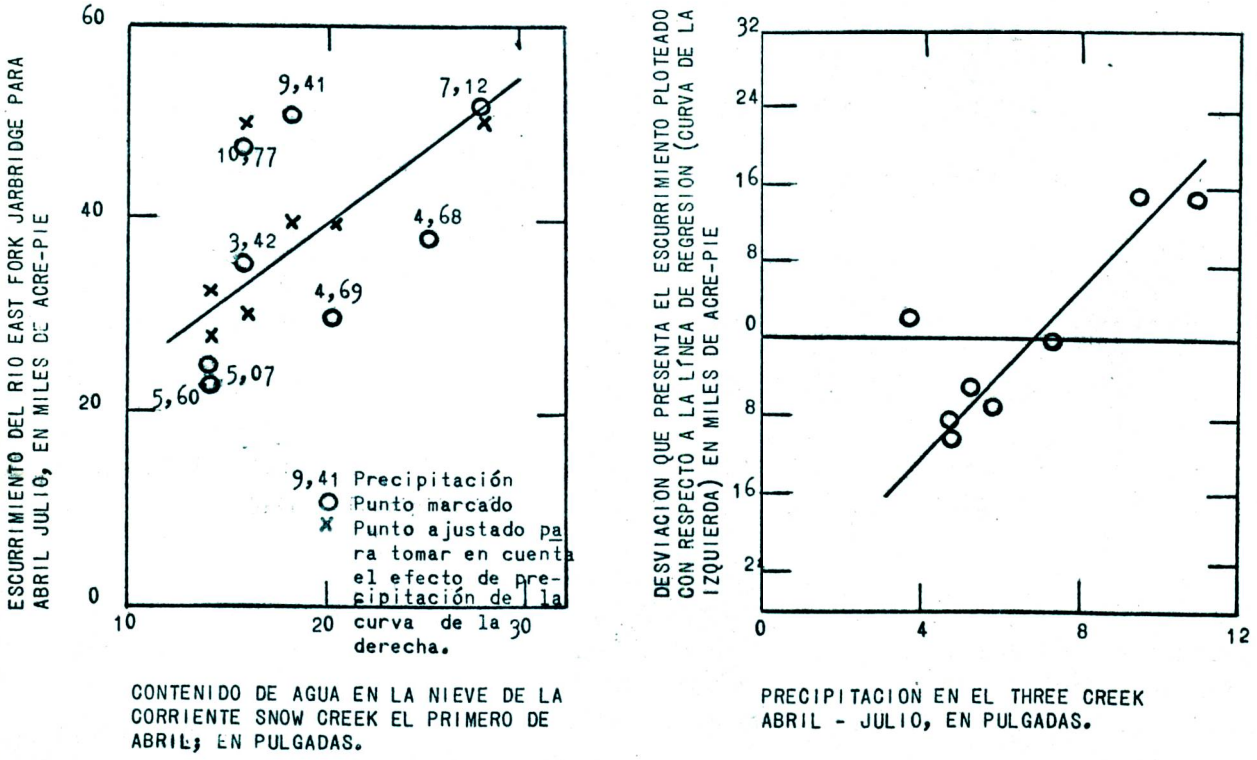


Figura N° 18. - Ejemplo de regresión gráfica múltiple usando escalas aritméticas.

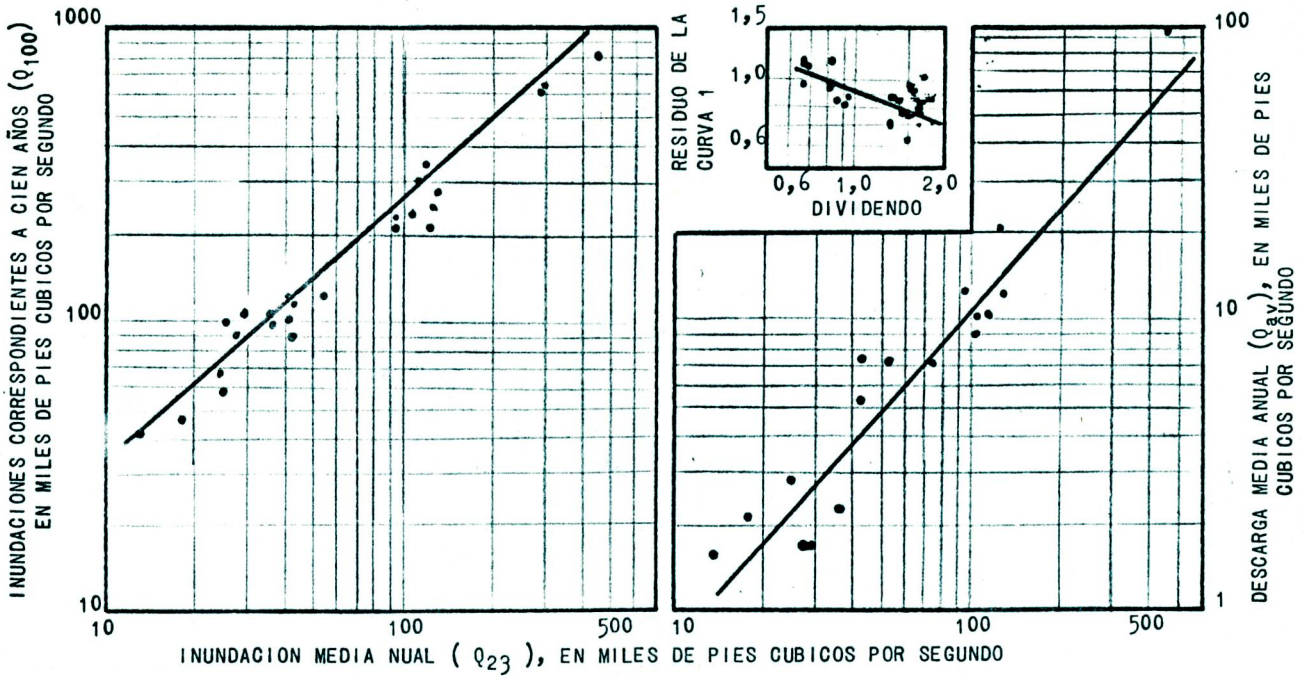


Figura Nº 19.- Regresión gráfica usando variables independientes muy correlacionadas. Basados en los datos de la tabla Nº 4.

Tabla Nº 4.- Datos de regresión gráfica usando variables independientes muy correlacionadas.

| | Río y situación. | Crecida en 100 años. (mcs) | Flujo medio anual (mcs) | Descarga promedio. |
|---|---|---|---|---|
| 1. | Neosho-Iola Kans | 2.974 | 801 | 48 |
| 2. | Big-Blue-Randolph Kans. | 2.738 | 782 | 48 |
| 3. | Miami-Dayton Ohio | 3.058 | 1.017 | 64 |
| 4. | Savannah-Augusta, Ga. | 9.911 | 1.197 | 298 |
| 5. | West Branch Susquehanna-Williamsport, Pa. | 7.362 | 1.948 | 278 |
| 6. | Susquehanna Towanda, Pa. | 6.683 | 3.033 | 294 |
| 7. | Susquehanna-Harrisburg, Pa. | 16.820 | 8.003 | 983 |
| 8. | Kanawha-Kanawha-Falls- W. Va. | 7.816 | 3.557 | 359 |
| 9. | Allegheny-Red House, N. Y. | 1.597 | 694 | 79 |

| | | | | |
|---|---|---|---|---|
| 10. | Iowa. Iowa City, Iowa. | 1.597 | 694 | 79 |
| 11. | Tennesse-Knoxville, Tenn. | 6.456 | 2.682 | 44 |
| 12. | French Broad-Asheville, N. C. | 1.288 | 459 | 60 |
| 13. | Des Moines, Keosangna-Iowa. | 2.917 | 1.164 | 152 |
| 14. | Connecticut-White River, Junction, Vt. | 3.455 | 1.504 | 204 |
| 15. | Cumberland-Nashville, Tenn. | 5.890 | 3.460 | 578 |
| 16. | Hudson. Mechanicville, N. Y. | 2.529 | 1.203 | 210 |
| 17. | Ohio-Cincinnatti, Ohio. | 22.653 | 12.564 | 2.766 |

---

ELECCION DEL METODO GRAFICO O ANALITICO PARA EL CALCULO DE LA REGRESION MULTIPLE. -

Para realizar el trabajo exploratorio y para hacer las estimaciones preliminares existe un método gráfico standard. El método gráfico tiene las siguientes ventajas:

1. - Es rápido
2. - Ayuda a definir el modelo apropiado
3. - Señala la necesidad de transformaciones, si las hay
4. - Llama la atención sobre puntos muy aislados si existen en los datos (ver los puntos aislados de la figura 18).

Las DESVENTAJAS del método gráfico son:

1. - No se puede identificar los efectos pequeños que ocasionan las variables independientes.

2. - El número de variables independientes está limitado a tres debido al efecto acumulativo de las inexactitudes del trazado y de la situación de las líneas.

3. - Las pruebas de significación de los efectos de las variables individuales no están a disposición.

4. - La relación resultante de la inclusión de tres o más variables es confusa para el usuario a menos que se exprese matemáticamente o se lleve a un gráfico siguiendo una forma distinta.

Un método analítico tiene las siguientes VENTAJAS:

1. - Para el modelo usado, da la mejor estimación de las constantes de la ecuación, y del error standard.

2. - Permite la prueba de los coeficientes que expresan la diferencia con el cero.

3. - Los resultados se pueden presentar de una forma clara y concisa que todos los hidrólogos pueden comprender.

4. - Los resultados son exclusivos para el modelo y la muestra usados;cualesquiera que sean los investigadores los resultados serán los mismos.

Las DESVENTAJAS del método analítico son:

1. - El cómputo lleva tiempo, especialmente cuando hay varias variables o modelos complicados; el uso de los computadores reduce el tiempo de computación, exige considerable tiempo para la preparación de los datos.

2. - La existencia de puntos aislados está oculta como lo estaría la existencia de un grupo de puntos diferentes a la mayoría (a menos que se compute la diferencia existente entre todos los puntos y sus estimaciones).

3. - El modelo seleccionado puede no ser el apropiado.

## DETERMINACION DE LAS ECUACIONES DE LAS RELACIONES GRAFICAS. -

Los análisis gráficos son a menudo adecuados para cierto tipo de problemas. Los resultados se pueden dar suministrando copias de los gráficos, pero las interpretaciones de los gráficos con más de dos variables es difícil para cualquiera que no esté familiarizado con el procedimiento. Por ejemplo, consideremos la relación de tres variables de la figura 18. Cuál es el desague

posible que corresponde a un contenido de agua en forma de nieve de 51 cm. y de una precipitación sobre el Three Creek de 25 cm. ? -- Es de 4. 938 Ha-metro de curva de la izquierda más 1. 728 Ha-metro de la curva de la derecha para un total de 6. 666 Ha-metro. Un método mejor de representación sería el uso de una familia de curvas. Otra manera sería la de escribir la ecuación de la relación gráfica, que para la relación de la figura 18 es:

$$R = -22 + 1,6S + 4,4P \qquad \begin{matrix} 12 < S < 28 \\ 4 < P < 11 \end{matrix}$$

Estos límites de definición para S y P le dice al lector que la aplicación de la ecuación fuera de ellos es riesgosa.

Otra ventaja de definir las ecuaciones de las relaciones gráficas aparece cuando se desea comparar relaciones del mismo tipo pero deducidas de datos diferentes. Por ejemplo Riggs ( en 1. 965 ) relacionó las descargas básicas de flujo de nueve pequeños ríos con el agua de drenaje y el porcentaje de la cuenca vaciada. Hizo ocho relaciones diferentes, cada una basada en mediciones realizadas en los mismos ríos pero en momentos diferentes. El interés estaba en la variabilidad del efecto del porcentaje vaciado de la cuenca; esta variabilidad fué aparente cuando se definieron las ecuaciones de las relaciones y cuando se compararon los coeficientes de regresión del porcentaje vaciado de la cuenca.

Un uso más de la ecuación de una relación gráfica es el de reducir una relación a su forma más simple. Esta reducción es un procedimiento deseable si el análisis gráfico hace uso de variables compuestas.

La regresión gráfica de la figura 20 es el resultado de un estudio exploratorio de datos procedentes de una cuenca del occidente de los Estados Unidos, e indica que se puede estimar con bastante seguridad el FMA a partir del área de drenaje y del flujo promedio en metros cúbicos por segundo por Kilómetro - cuadrado. Debido al hecho de que el área de drenaje se usa dos veces, el efecto real de ella debe ser evaluado. Empezamos por escribir la ecuación de la relación (por un método subsecuentemente), que es:

$$\log FMA = 1,00 + \log A + 1,02 \log \bar{Q}$$

en donde A es el área de drenaje en Kilómetros cuadrados y $\bar{Q}$ es el flujo promedio en metros cúbicos por segundo por Kilómetro cuadrado. Sea $\bar{q}$ el flujo - medio en metros cúbicos por segundo, entonces:

$$\bar{Q} = \bar{q} / A$$

Substituyendo en la primera ecuación resulta:

$$\log MAF = 1,00 + \log A + 1,02 \log ( \bar{q} / A )$$

$$= 1,00 + \log A + 1,02 \log \bar{q} - 1.02 \log A$$

Luego el coeficiente neto de regresión de log A es -0, 02; el cual es despreciable y, si se elimina queda:

$$\log MAF = 1,00 + 1,02 \log \bar{q} \quad ; \text{ ó sea:}$$

$$MAF = 10 \; \bar{q}^{1,02}$$

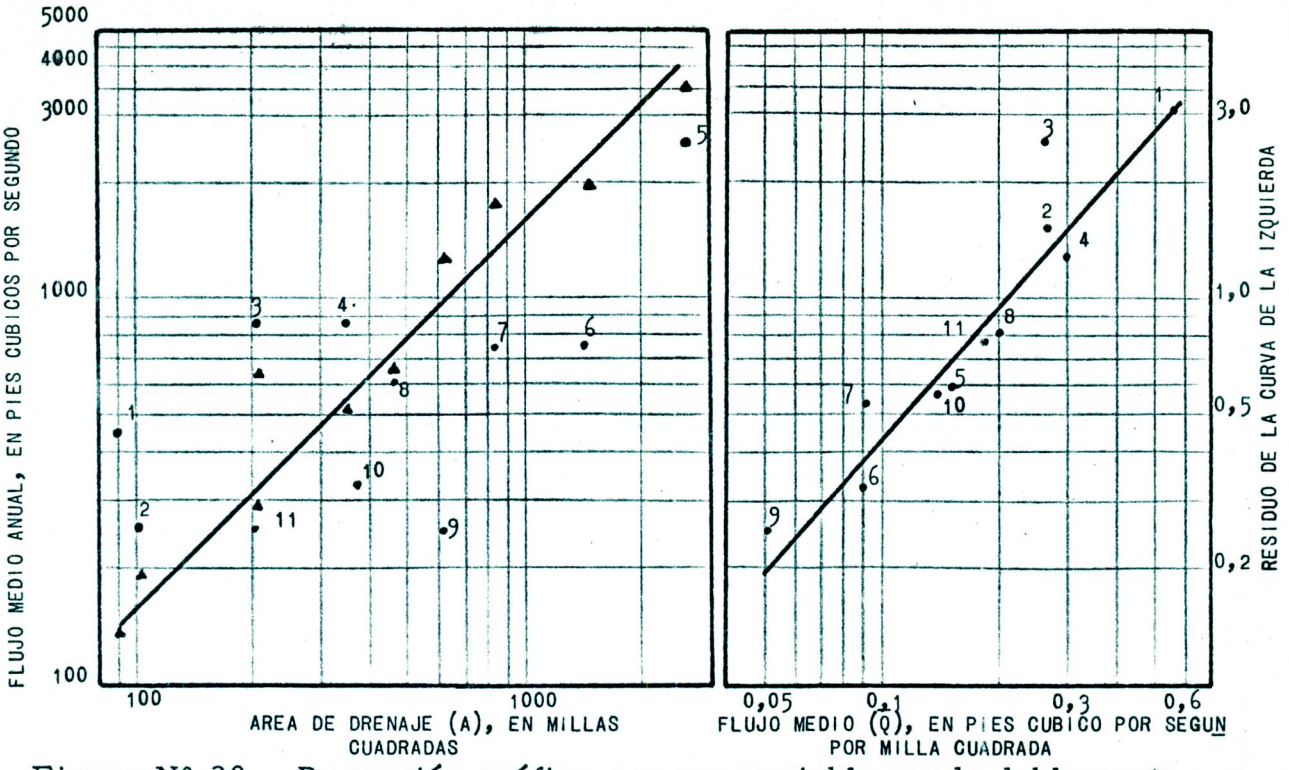


Figura N° 20. - Regresión gráfica con una variable usada doblemente.

METODOS GENERALES. -

Toda ecuación lineal en dos variables es de la forma: $Y = a + bX$ y esta forma general es la ecuación de una línea recta en coordenadas rectangulares. La forma lineal en escala logarítmica:

$$\log Y = \log a + b \log X$$

que cuando se expresa en las variables originales es la ecuación:

$$Y = aX^{b}$$

Una línea recta en papel semilogarítmico tiene la forma lineal:

$$\log Y = \log a + b X$$

que se reduce a la ecuación exponencial:

$$Y = a ( 10 )^{bX}$$

Si b = c log K en esta última ecuación, se obtiene:

$$Y + aK^{cX}$$

A veces, los puntos que se llevan a un papel logarítmico definen a una curva corriente en vez de una recta. El lugar geométrico de los puntos puede hacerse algunas veces lineal añadiéndole o restándole una constante a una de las variables. La relación sería de la forma:

$$\log Y = \log a + b \log ( X + c ) \text{ ; ó:}$$

$$Y = a (X + c)^{b}$$

Para determinar la ecuación de cualquier relación lineal de dos variables calcúlese la pendiente, b, de la línea como la distancia vertical dividida entre la distancia horizontal. Estas distancias siempre se miden en unidades aritméticas aunque el gráfico sea en papel logarítmico (las b no se transforman). El intervalo de escala debe de ser el mismo en ambos ejes o se debe hacer un ajuste aritmético adecuado. La intercepción, a, se mide generalmente de la hoja del gráfico sobre el eje de ordenadas para un valor adecuado de X. -

Para la relación:

$$Y = a + bX$$

$$Y = a \text{ cuando } X = 0$$

y para la relación:

$$\log Y = \log a + b \log X$$

$$Y = a \text{ cuando } X + 1$$

Si no se puede extender la línea convenientemente hasta X = 1 ó X=0, se puede sustituir en la ecuación las coordenadas de un punto de la curva y deter -

minar la intercepción.

Las ecuaciones standard dadas en los textos de geometría analítica son de poco uso en el análisis empírico. Se necesitan expresiones matemáticas - más flexibles, y son deseables aquellas que se puedan poner en forma lineal - por su facilidad de computación. Si no se puede hallar una transformación que- haga a esta relación lineal, entonces un modelo del tipo:

$$Y = a + b_1X + b_2X^2 + b_3X^3 + \text{----------} + b_nX^n$$

o alguna porción del mismo, se ajustará a la mayoría de las curvas suaves. Si- la curvatura tiene un solo sentido, el término $X^2$ introducirá la curvatura nece saria. Cuando una curva tiene un punto de inflexión se necesitan los términos - $X^2$ y $X^3$. En el trabajo empírico raras veces se usan términos con exponen- tes mayores.

El modelo anterior es igualmente aplicable cuando se reemplaza a X por log X. Una línea de curvatura en un sólo sentido se expresa en escala logarítmi ca como:

$$\log Y = \log z + b_1 \log X + b_2 (\log X)^2$$

que reducida a la forma exponencial es:

$$Y = a X^{b_1} X^{b_2 \log X} = aX^{b_1 + b_2 \log X}$$

La forma general de las relaciones lineales en varias variables es:

$$Y = a + b_1X_1 + b_2X_2 + \text{----------} + b_nX_n$$

Algunas veces el coeficiente de regresión para una variable cambia -

con otra. A esto se le conoce con el nombre de interacción. En el modelo:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2$$

se denomina al término $b_3X_1X_2$; producto de interacción. Su uso proporciona un cambio sistemático de pendiente. Las relaciones curvilíneas en varias variables se pueden describir añadiendo términos con potencias de las variables independientes. La ecuación de una curva o de una relación múltiple que implica una interacción no se puede computar con facilidad. La ventaja de reconocer la forma general de la ecuación que representaría una relación gráfica particular yace en la necesidad de un modelo si hay que computar una regresión de mínimo cuadrado. La definición de la ecuación de una regresión gráfica se limita a las regresiones lineales.

DEFINICION DE ECUACIONES. -

Por medio de dos ejemplos demostraremos los métodos para definir la ecúación de una regresión gráfica. Los procedimientos usados en estos ejemplos se pueden adaptar, sin más, a otros problemas. El primer ejemplo, que se muestra en la figura 21, es una regresión lineal múltiple por el método de los residuos. La ecuación de esta relación se obtiene de la manera siguiente. - Considere primero la relación entre $Y_c$ y $X_1$ en donde $Y_c$ es el valor de la curva de la parte derecha de la figura 21. Esta relación es de la forma $Y = a + bX_1$ - en donde a, es la intercepción en $X_1 = 0$ y b, es la pendiente de la línea.

Para este ejemplo: $Y_c = 5,4 + 0,86\ X_1$

La ecuación de la segunda línea se obtiene de manera similar, y es:

$$\text{Residuo} = -10 + 2,78\ X_2$$

El Residuo (denominado R) es el valor puntual individual, Y, menos el valor - obtenido de la primera ecuación, es decir:

$$R = Y - Y_c = -10 + 2,78\ X_2$$

Substituyendo a $Y_c$ por su valor expresado en la ecuación anterior correspondiente:

$$Y = (5,4 + 0,86\ X_1)$$

a partir de la cual se obtiene la relación deseada:

$$Y = -4,6 + 0,86\ X_1 + 2,78\ X_2$$

El segundo ejemplo (Fig. N° 22) es una regresión gráfica múltiple coaxial simple adaptada de una hecha por el Departamento Hidráulico de Investigación de la Oficina de Vialidad (Hydraulic Research Branch of the Bureau of Public Roads). Esta regresión es lineal, y las líneas correspondientes a S y a P están espaciadas sistemáticamente y son paralelas. Bajo estas condiciones, se puede determinar la ecuación de la relación gráfica.

Estudiando la figura 22 se evidencia, lo siguiente:

1. - $Q_{10}$ es la variable dependiente
2. - A, es la variable independiente principal
3. - Las líneas con un P, igual, están espaciadas linealmente en el papel logarítmico y son paralelas.
4. - Las líneas con un S igual están espaciadas logarítmicamente al doble de la escala logarítmica del papel y son paralelas.

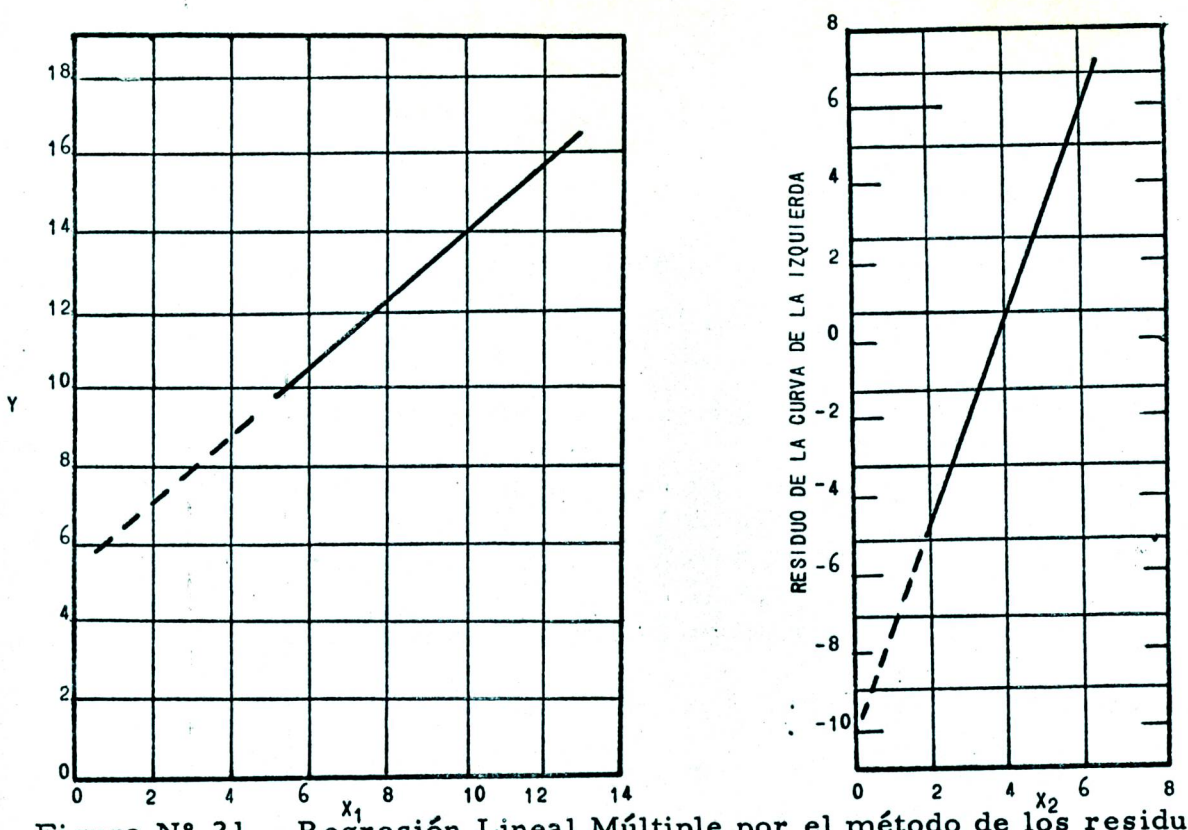


Figura N° 21. - Regresión Lineal Múltiple por el método de los residuos.

Para encontrar la solución, separe la regresión en dos partes introduciendo una variable intermedia, $Q_{adj}$; ( a una escala arbitraria ) de tal manera que:

$$Q_{adj} = f\ (A, P) \qquad y:$$

$$Q_{adj} = f\ (S, Q_{10})$$

Consideremos la primera relación. Para un P fijo, el modelo sería:

$$Q_{adj} = KA^{n}$$

en donde K, es la intersección sobre la escala $Q_{adj}$ ( para A = 1) y n es la pendiente de la línea. En este ejemplo, n=1, 28 que es la relación lineal entre las longitudes vertical y horizontal. Para P = 1, 20 la intersección es K= 78.

Para obtener esta intersección graficamente hay que prolongar bastante la curva. Es más sencillo computar la intersección de cualquier otro valor

de A diferente de uno, por ejemplo, para $Q_{adj} = 1.000$ , $A = 7,3$; luego

$$1,000 = K(7,3)^{1,28} \qquad \text{de donde } K = 78.$$

La introducción de P como variable hace necesaria la definición de la intersección K en función de P (pues la intersección es diferente para cada valor de P). Cuando la diferencia de intervalos de diez unidades se proyecta sobre el eje $Q_{adj}$ resulta equivalente a 1,36; es decir: por cada aumento de P en 10 unidades la intersección aumenta 1,36 veces (está en escala logarítmica). Este aumento se puede medir cuando los intervalos son individuales o, computar a partir del aumento total. Por ejemplo para A = 10 $Q_{adj} = 1.480$ cuando P = 1,2 y
$= 17.000$ cuando P = 2

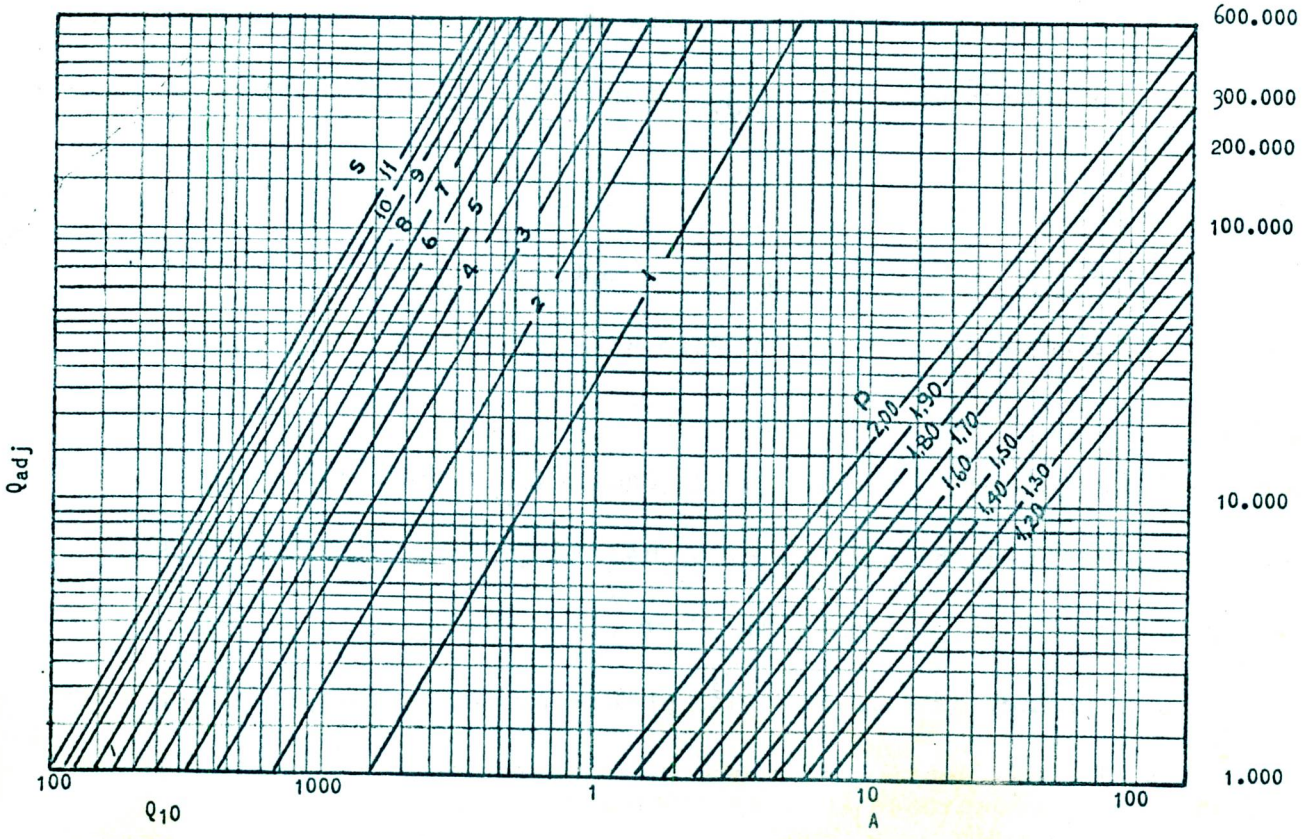


Figura N° 22.- Regresión gráfica múltiple coaxial.

El aumento correspondiente a ocho intervalos es igual a 17.000/1.480 = 11,5.

Luego: $$K = 78\ (1,36)^{10\ (P-1,2)}$$

en donde 78 es la intersección para P=1,2, el aumento de K por décimas es- de 1,36 y el factor 10 (P-1,2) es el número de décimas por encima de 1,2 .

Substituyendo los valores de K y n resulta para la primera relación :

$$Q_{adj} = 78\ (1,36)^{10\ (P-1,2)}\ A^{1,28} \qquad (1)$$

La segunda relación:

$$Q_{adj} = \quad (S,\ Q_{10})$$

se maneja considerando de nuevo que $Q_{adj}$ es la variable dependiente. Des - preciando a S, el modelo sería:

$$Q_{adj} = K\ Q_{10}^{m}$$

La pendiente se mide a escala y es igual a 1,78. La intersección K se compu - ta para S = 1, $Q_{adj}$= 100.000 y $Q_{10}$=19.000, de acuerdo con:

$$100.000 = K\ (19.000)^{1,78} \qquad \text{ó:}$$

$$\log 100.000 = \log K + 1,78 \log 19.000$$

$$\log K = 5 - 7,60 = -2,60 = 7,40 - 10$$

$$K = 0,00025$$

el espaciamiento de las líneas en una dirección vertical (paralela al eje $Q_{adj}$ )- es el doble de las escalas del papel. Por lo tanto, la intersección K varía direc - tamente con $S^2$, y la ecuación es:

$$Q_{adj} = 0,0025\ S^2\ Q_{10}^{1,78} \qquad (2)$$

Combinando las ecuaciones 1 y 2 se obtiene:

$$0,0025\, S^2\, Q_{10}^{1,78} = 78\,(1,36)^{10\,(P-1,2)}\, A^{1,28}$$

que en forma logarítmica es:

$$\log 0,0025 + 2 \log S + 1,78 \log Q_{10} = \log 78 + 10\,(P - 1,2)\log 1,36 + 1,28 \log A$$

Resolviendo para $Q_{10}$ dá:

$$\log Q_{10} = 1,61 + 0,72 \log A - 1,12 \log S + 0,75\, P$$

que es el resultado deseado.

OTRAS HERRAMIENTAS. -

Análisis de las varianzas. -

El Análisis de las varianzas es un procedimiento por medio del cual se pueden transformar las variaciones enmarcadas en los datos en variaciones componentes debidas a los factores independientes. Está estrechamente ligado a la correlación pero es aplicable a problemas en donde algunos de los factores sólo se pueden describir por clases, y no como variedades numéricas.

El análisis depende del carácter aditivo de las varianzas. Su propósito es el de probar si medios diferentes son parecidos o nó. Los fundamentos básicos del problema son (1) la medición de la varianza de datos experimentales multiplicados por la suma de los cuadrados de las observaciones de las desviaciones de la media (2) la partición de la suma total de las desviaciones al cuadrado en parte independientes, estando cada parte asociada a alguna propiedad física del experimento (3) la estimación de los parámetros en las distribuciones que se han postulado para dar fundamento seguro a los da-

tos y (4) pruebas de significado de los parámetros. Los resultados de las pruebas dan la probabilidad de la existencia de una diferencia apreciable - entre los efectos de un factor o de varios factores a diferentes niveles.

Un ejemplo muy simple de análisis de varianza se ocupa de averiguar si el desague promedio para dos períodos de registro en una estación de aforo con estimaciones de la misma población media. A continuación se dan los desagues anuales:

| Período 1 | Período 2 |
|---|---|
| 17,3 | 6,4 |
| 21,9 | 15,2 |
| 13,6 | 9,7 |
| 10,8 | 4,4 |
| 9,7 | 9,9 |
| 20,7 | 11,9 |
| 16,3 | 11,9 |
| 16,2 | 15,4 |
| 12,5 | 9,4 |
| 11,3 | 7,0 |
| 14,0 | 16,0 |
| 16,5 | 17,0 |
| 15,3 | 11,2 |
| 19,2 | 13,2 |
| 13,0 | 11,5 |
| 238,3 | 170,1 |

Los cómputos se hacen de la manera siguiente:

Gran Total = T = 238,3 + 170,1 = 408,4

Número total de elementos = N = 30

Número de elementos en cada período = n = 15

$$T^2/N = (408,4)^2/30 = 5.559,7$$

Suma de los cuadrados = $\Sigma Y_{ia}^2 = 6.067,3$

dividida entre n = $\Sigma T_i^2/n = [(238,3)^2 + (170,1)^2]/15 = 5.714,7$

Suma de los cuadrados interperíodos = $\Sigma T_i^2/n$

$-T^2/N = 5.714,7 - 5.559,7 = 155,0$

Suma de cuadrados dentro de los períodos = $\Sigma Y_i^2$

$- \Sigma T_i^2/n = 6.067,3 - 5.714,7 = 352,6.$

Suma total de los cuadrados = $\Sigma Y_i^2 - T^2/N = 6.067,3$

$- 5.559,7 = 507,6$

La tabla del análisis de varianza es:

| Origen | Suma de cuadrados | Grados de libertad | Cuadrado de la media | Cuadrado promedio de la media. |
|---|---|---|---|---|
| Entre períodos | 155,0 | 1 | ** 155 | $\sigma^2 + n\sigma_s^2$ |
| Dentro de los períodos. | 352,6 | 28 | 12,6 | $\sigma^2$ |
| Total. | 507,6 | 29 | ------- | -------- |

Los grados de libertad, G.L. , son en número, uno menos que el número de períodos, p, para las sumas entre períodos; y N-1 para el total; luego los grados de libertad asociados a la fuente dentro de períodos es N-p. El cuadrado medio se obtiene dividiendo la suma de los cuadrados por G.L.

La última columna de la tabla de análisis de varianza señala los posibles valores del cuadrado de la media. Si las medias de cada período fuesen iguales, el término $n\sigma_s^2$ sería nulo. Las estimaciones de la relación:

$[\sigma^2 + n\sigma^2_\xi{}^2] / \sigma^2$ pueden ser mayores que uno, debido a la casualidad o debido a que hay una diferencia real. La relación tiene la distribución F y se puede probar estadísticamente. La relación de la tabla es de: - 155/126 = 12,3. El valor F para 1 y 28 grados de libertad y una probabilidad de 0,01 es de 7,6 por la tabla de distribución F. Como la proporción de la muestra excede a la proporción tabular, concluímos que hay una diferencia real entre períodos; es decir, la probabilidad de que tal diferencia de las medias haya ocurrido por casualidad si no había diferencia verdadera entre los períodos es menor de 0,01. El doble asterisco, del cuadrado de la - media entre períodos (del análisis de la tabla de varianza), denota el significado estadístico por encima del nivel 0,01.

Consideremos ahora un problema similar, para determinar si la precipitación media anual de tres estaciones es diferente. En la tabla siguiente - se dan los datos:

| Año | Precipitación en pulgadas en el Puesto 1 | Puesto 2 | Puesto 3 |
|---|---|---|---|
| 1.945 | 40,6 | 48,2 | 47,5 |
| 1.946 | 36,1 | 40,2 | 34,8 |
| 1.947 | 37,5 | 37,8 | 42,2 |
| 1.948 | 52,3 | 58,2 | 59,9 |
| 1.949 | 42,2 | 43,3 | 51,7 |
| 1.950 | 40,6 | 41,4 | 42,5 |
| 1.951 | 38,3 | 42,3 | 40,5 |
| 1.952 | 45,8 | 48,2 | 47,8 |
| SUMAS | 333,4 | 359,6 | 366,9 |
| $\bar{Y}$ | 41,7 | 45,0 | 45,9 |

Con los datos de la tabla podemos hacer los siguientes cálcu-

los:

$$T = 1.059,9$$

$$T^2/N = 46.807,8$$

$$Y_{ia}^2 = 47.784,0$$

$$T_{ia}^2/n = 46.885,4$$

La tabla del análisis de varianza es:

| Origen | Sumas de cuadrados | Grados de libertad | Cuadrados de la media |
|---|---|---|---|
| Entre puestos | 77,6 | 2 | 38,8 |
| En los puestos | 898,6 | 21 | 42,9 |
| Total | 976,2 | 23 | ---- |

$F_{2,21} = 38,8/42,9$ 1 , por lo tanto, estadísticamente, no hay diferen-

cia entre las medias.

Una cuidadosa lectura de la literatura hidrológica revelará - muy pocas aplicaciones del análisis de la varianza. La mayoría de es - tos análisis se basan en datos procedentes de experimentos preparados; y es para esta aplicación para la cual se obtienen los mejores resulta- dos. Los datos hidrológicos son generalmente parte de una serie de tiempo que no puede ser estacionaria. Por lo tanto, los valores individua- les no pueden ser completamente independientes como lo exigiría un ana- lisis valido de varianza.

En el ejemplo donde se comparan los desagues medios de dos períodos de registro, se concluyó que entre períodos había una diferencia real. Pero no hay razón física para esperar un cambio en esta cuenca. El período inicial fué de grandes precipitaciones; el último incluyó las sequias de los años treinta. También es posible que algunos de los desagues anuales estuviesen correlacionados en serie.

Luego el carácter de los datos tiende a desacreditar los resultados de esta aplicación particular del análisis de varianza.

En el último ejemplo, la precipitación en el puesto 2 es mayor que en el 1 en cada uno de los años señalados, sin embargo el análisis de la varianza no indica diferencia de las medias. (Un análisis de la varianza entre el puesto 1 y el puesto 2, solamente muestra una diferencia a un nivel de probabilidad de aproximadamente 0, 25).

Las precipitaciones anuales en un determinado puesto pueden ser independientes, pero las precipitaciones en los diferentes puestos para el mismo año no lo son. Por tanto, no se cumplen los requisitos del método, y los resultados se deben de aceptar con reserva.

Los dos ejemplos dados utilizan un modelo estadístico muy simple. Para problemas más complicados, se pueden considerar varios modelos. La selección del modelo apropiado es difícil para el estadísta "ocasional". Muchos textos de estadística tratan detalladamente el análisis de la varianza. Ver a Bennett y Franklin (1. 954), Brownlee (1. 960), y, Dixon y Massey (1. 957). En gene-

ral, un análisis de varianza hecho por alguién que no esté completamente familiarizado con el proceso debe ser revisado por un estadísta para que verifique la conveniencia del modelo y la exactitud de la interpretación.

ANALISIS DE COVARIANZA. -

El análisis de la varianza de los datos de desague para dos períodos- (Ver pág. ) indica que las medias de la población eran probablemente diferentes, sin embargo otra información, particularmente los registros de precipitaciones, lleva a la conclusión contraria. Los datos de precipitación, se pueden incorporar al análisis usando un análisis de covarianza. Este método incluye conceptos del análisis de la varianza y de la regresión y es aplicable en aquellas partes en donde una variable representa una medición para cada individuo en contraposición con una variable que solamente puede ser separada en unas cuantas categorías.

En la figura 23 se muestran dos condiciones generales para las cuales un análisis de covarianza originará conclusiones diferentes de las de un análisis de varianza. En cada condición, Y, es la variable, que se analiza, y X, es la variable dependiente. El gráfico A de la figura 23 muestra las medias de Y para que los dos períodos sean practicamente iguales.

En estas condiciones un análisis de varianza no mostraría una diferencia apreciable entre las medias. Pero entre los períodos 1 y 2 ocurrió un cambio importante en la relación de Y a X, y es a este cambio al que el análisis de covarianza puede identificar.

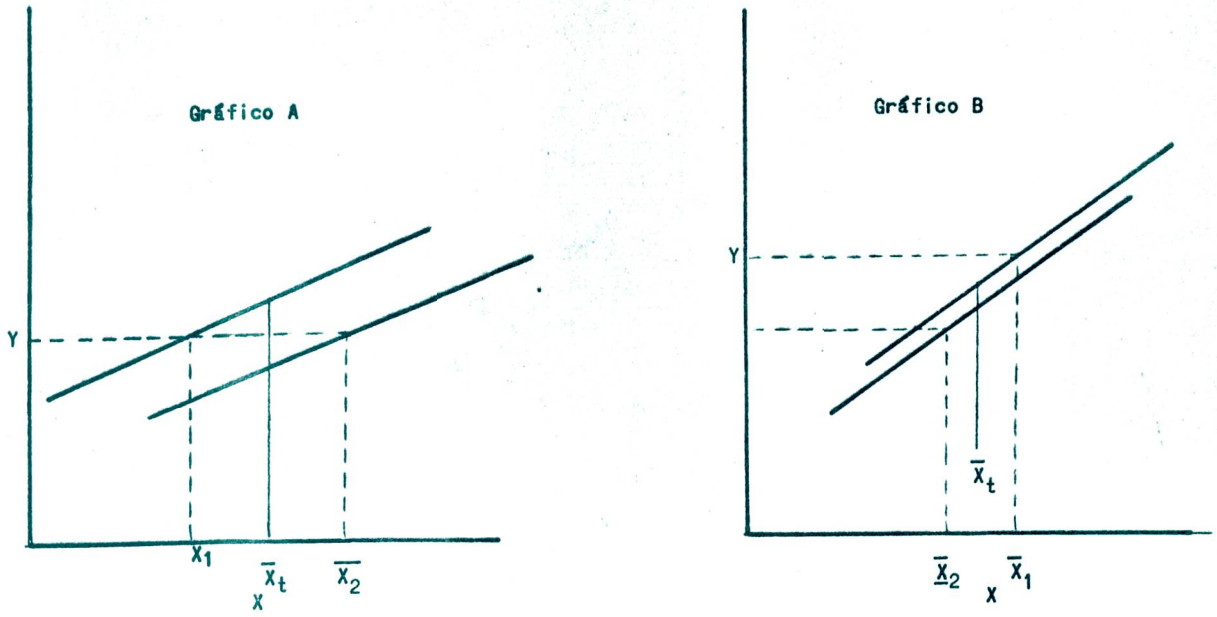


Figura N° 23.- Dos condiciones para las cuales el análisis de la covarianza dará lugar a conclusiones diferentes de las del análisis de la varianza.

El análisis de las pruebas de covarianza se hace en base a desviaciones de la regresión en vez de en base a las medias. La prueba implica la suma de los cuadrados de las desviaciones de la regresión definida por todos los puntos marcados con respecto a las propias medias de su período y a la suma de los cuadrados de las desviaciones de una línea de regresión total (Dixon Y Massey, 1.957, p. 210). En efecto la prueba indica si los dos períodos son diferentes cuando se ajustan al mismo valor de X. Como se dijo previamente, un análisis de la varianza de los datos de la condición del gráfico A, Fig. 23, no indicaría diferencia entre períodos porque las media $\overline{Y}_1$ e $\overline{Y}_2$ son casi iguales. Pero el análisis de covairanza indicaría una diferencia apreciable de los valores de Y de la media total $\overline{X}_t$.

El gráfico B de la figura 23 muestra dos períodos que tienen valores medios, Y, muy diferentes pero no tienen ninguna diferencia real en las regresiones de Y sobre X para los dos períodos. Los dos resultados no se contradicen. Hay una diferencia para las medias de los dos períodos, pero esta se debe a una diferencia de los valores de X para dichos períodos.

El análisis de la covarianza exige que las pendientes de las líneas de regresión de los períodos individuales sean virtualmente paralelas. Dixon y Massey describieron (1.957, p.218) una prueba de paralelismo.

Tabla N° 5. - Indice anual de precipitación y desague anual para el ejemplo de análisis de covarianza.

| Período 1 | | Período 2 | |
|---|---|---|---|
| Indice de Precipitación. (X) | Desague (Y) | Indice de Precipitación. (X) | Desague (Y) |
| 27 | 17,3 | 14 | 6,4 |
| 36 | 21,9 | 26 | 15,2 |
| 26 | 13,6 | 15 | 9,7 |
| 18 | 10,8 | 11 | 4,4 |
| 27 | 19,7 | 19 | 9,9 |
| 30 | 20,7 | 21 | 11,9 |
| 25 | 16,3 | 18 | 11,9 |
| 28 | 16,2 | 22 | 15,4 |
| 19 | 12,5 | 20 | 9,4 |
| 22 | 11,3 | 17 | 7,0 |
| 22 | 14,0 | 29 | 16,0 |
| 29 | 16,5 | 30 | 17,0 |
| 26 | 15,3 | 16 | 11,2 |
| 29 | 19,2 | 23 | 13,2 |
| 24 | 13,0 | 23 | 11,5 |
| 388 | 238,3 | 304 | 170,1 |

Tx = 692; Ty = 408,4

A continuación damos los detalles de cómputo de un análisis de co - varianza usando (1) los mismos datos de desague de la sección previa para - el ejemplo del análisis de varianza y (2) algunos valores supuestos de un ín - dice de precipitación, y que están todos en la tabla N° 5 y en el gráfico de la figura N° 24.

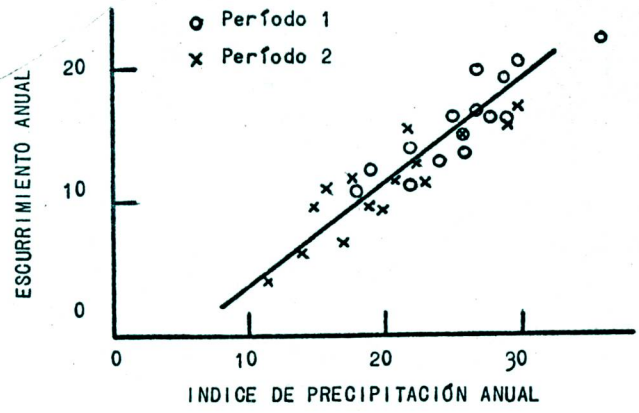


Figura N° 24. -Gráfico con datos de la tabla N° 5.

El gráfico indica que no hay cambios de relación entre períodos y que la diferencia de las medias de desague entre períodos se debe a diferencias - de precipitaciones. Con estos datos no sería necesario hacer un análisis de - covarianza. Sin embargo si los puntos marcados indicasen regresiones sepa - radas, podría ser ideal hacer un análisis de covarianza para probar si las - dos regresiones eran estadística y apreciablemente diferentes.

El siguiente cómputo ilustra el procedimiento: la suma total de pro - ductos $= \Sigma Xij\ Y = TxTy / nK$, en donde Tx y Ty son los totales absolutos de X e Y; n, el número de elementos en cada período y K el número de periódos.

La suma de productos entre las medias $= \Sigma Txi\ Tyi / n - TxTy / nK$, en donde Txi y Tyi, son totales (de períodos).

Por ejemplo: la suma total de productos = 27 (17, 3)

+ 36 (21, 9)...+ 23 (11, 5)

- 692 (408, 4) / 15 (2)

- 10.054, 7 - 9.420, 5 = 634, 3

Suma de productos entre medias:

= 388 (328, 3) / 15 + 304 ( 170, 1) / 15

- (692) (408, 4) / 30 = 9.611, 4

- 9.420, 4 = 191, 0

La suma total de cuadrados $X = \Sigma X^2_{ia}$

$$- T_x^2 / N = 16.898 - 15.962 = 235$$

Suma de cuadrados de X entre períodos:

$$= \Sigma T_{xi}^2 / n - T_x^2 / N = 16.197 - 15.962 = 235$$

Las sumas de cuadrados de Y se toman del ejemplo del análisis de varianza.

Las desviaciones de regresión se computan con la fórmula:

$$\Sigma y^2 - ( \Sigma xy)^2 / \Sigma x^2$$

que para los totales es: $507, 6 - (634, 3)^2 / 936 = 507, 6$

$$-429, 8 = 77, 8$$

porque dentro de los períodos las desviaciones de la regresión son:

$$352, 6 - (443, 3)^2 / 701 = 352, 6 - 280, 3 = 72, 3$$

y la desviación entre medias de la regresión se obtiene por sustracción.

| Origen | Datos | | | | Desviaciones | | |
|---|---|---|---|---|---|---|---|
| | Grados de Libertad. | $x^2$ | xy | $y^2$ | Grados de Libertad. - | $(Y-\bar{Y})^2$ | Cuadrado de la media. |
| Entre Medias | 1 | 235 | 191 | 155, 0 | 1 | 5, 5 | 5, 5 |
| Dentro de Períodos. | 28 | 701 | 443, 3 | 352, 6 | 27 | 72, 3 | 2, 7 |
| Total. | 29 | 936 | 634, 3 | 507, 6 | 28 | 77, 8 | -- |

La suma de cuadrados dentro de períodos (de la primera parte de la tabla) se obtiene por sustracción. Los grados de libertad para las desviaciones de regresión, $(Y-\bar{Y})^2$, dentro de períodos y totales son uno menos que para las medias.

La prueba de significado compara a F (relación de los cuadrados de las medias de la tabla de covarianza), con los valores de distribución de ésta a niveles del 5 y 10 por ciento. Para éste ejemplo son:

$$F = 5, 5/2, 7 = 2, 0$$

$$F_{1, 27;0, 05} = 4, 2 \qquad y$$

$$F_{1, 27;0, 10} = 2, 9$$

Como 2, 0 es menor que 2, 9 la diferencia de períodos no es apreciable al nivel del 10 por ciento cuando los desagues se ajustan a la precipitación.

Ver el artículo de Wilm (1, 943, que incluye una discusión de Davenport, para una aplicación del análisis de covarianza a un problema hidrológico).

## ANALISIS MULTIVARIADO.

Las relaciones múltiples de variables independientes que están relacionadas entre sí producen algunas veces resultados incongruentes a partir de grupos diferentes de datos. Por ejemplo, el coeficiente de regresión de una variable independiente, puede ir de positivo a negativo en regresiones diferentes, y probar ser, sin embargo, estadísticamente importante.

En estas condiciones, las conclusiones referentes al efecto de esa variable sobre la variable dependiente podrían ser erradas si sólo se analizase un grupo de datos. El uso del análisis multivariado ha sido propuesto como medio de solución para este dilema.

El análisis multivariado se ocupa de la relación de grupos de variables aleatorias dependientes e incluye varios procedimientos diferentes, - cada uno con el próposito de cumplir un objetivo diferente. Snyder (1.962) investigó el uso del análisis multivariado en hidrología en donde la estructura de la solución era de primordial importancia. Kendall (1.957) describió esta teoría. En su estado actual de desarrollo, el análisis multivariado no es una herramienta útil para definir las relaciones de causa y efecto en hidrología, el análisis de regresión es aún el mejor método a disposición.

## CARACTERISTICA DE LOS DATOS HIDROLÓGICOS

El flujo de los ríos es un proceso contínuo que varía con el tiempo, y se dice que los datos correspondientes forman una serie cronológica. Un gráfico del caudal de un río en función del tiempo mostraría un patrón de variación recurrente cada año; es decir, los flujos abundantes tienden a ocurrir

en ciertos momentos particulares del año, y los escasos, en otros, en respuesta a las características climatológicas que también varian estacionariamente.

Debido al hecho de que el caudal no tiene valores discretos, necesitamos cortar la carta hidrográfica en piezas que consideraremos como corrientes individuales. Estas piezas particulares tienen ciertas características que se deben de considerar en el análisis. La pieza más común es la correspondiente a la descarga media diaria, la cual está relacionada con la descarga del día anterior y cae en un campo que depende de la época del año. En estadística, la descarga media diaria es una variable correlacionada en serie;es decir no es casual. Además,este tipo de descarga no es homogénea durante el año; es muy probable que en un momento del año sean mayores que en otros.

Los datos se consideran homogéneos si cualquier subgrupo, al cual se puedan asignar lógicamente algunos de estos datos, tiene la misma media posible y la misma varianza que cualquier otro subgrupo de la población.

Las descargas medias mensuales para diferentes meses también están correlacionadas en serie y no son homogéneas. Las descargas medias anuales pueden ser valores homogéneos;pueden o no estar correlacionadas en serie, dependiendo de la cantidad almacenada en la cuenca en el momento de iniciación del año hidrológico.

En lugar de considerar una variable de caudal constituída por segmentos adyacentes de una carta hidrográfica, podemos considerar variables tales-

como la media de julio, descarga pico anual o el flujo mínimo anual. Estas variables están formadas por un individuo de cada año y son por lo tanto independientes del ciclo anual de la corriente. También son independientes entre sí ( con la posible excepción de flujos mínimos anuales que incluyen el flujo del año previo desde el sitio de recarga de agua subterránea.

Las precipitaciones, la temperatura, la descarga de sedimentos, la calidad del agua, la transpiración, la evaporación y la radiación solar varían en el curso del año, los índices que la describen pueden no ser casuales ni homogéneos.

Es aparente que la distinción entre los datos casuales y los no casuales y entre los homogéneos y los no homogéneos no es siempre muy clara. El analista tendrá que determinar si los efectos de la posible no existencia moderada de la casualidad y de la no homogeneidad invalidará las conclusiones de su análisis particular. Es importante que el cáracter de los datos se considere al diseñar el análisis y al interpretar los resultados.

Hasta aquí hemos descrito las variables que se pueden considerar como muestras de la población si los individuos son homogéneos; si son además casuales, podemos estimar la distribución de frecuencia de la variable de la muestra. Otro tipo de variable usado extensivamente en la hidrología no se puede considerar como posesora de una distribución de probabilidad, o ni que ha sido extraída de una población en la manera que se considera usual.

Las características de una cuenca tales como área de drenaje, pendiente, elevación, e índice vegetal se encuentran en esta categoría. (Es posible con

cebir ciertos parámetros fisiográficos, como variables aleatorias fortuitas, pero raras veces se puede considerar a la muestra en cuestión como seleccionada al azar o como representativa).

En la regresión se usa a veces el tiempo como variable. Este no tiene distribución y se usa solamente como sustituto del factor o de los factores reales (que se desconocen o no pueden expresarse mediante índices), asociados a cambios de la variable dependiente.

EFECTOS DE LAS CARACTERISTICAS DE LOS DATOS EN EL ANALISIS. -

La preparación de la distribución de frecuencia de las descargas medias diarias se hace a partir de datos correspondientes a varios años, y la curva resultante se denomina curva de duración. Los valores individuales ni son fortuitos ni son homogéneos. Por lo tanto, no se puede considerar a la curva de duración como curva de frecuencia. La probabilidad de exceder un cierto valor de un determinado día futuro depende del valor precedente y de la época del año. Por lo tanto, la curva de duración es simplemente la distribución de las medias diarias que han ocurrido. Se puede considerar como estimación de la distribución durante un período futuro de varios años de duración.

Por otra parte, las curvas de frecuencia correspondientes a los picos anuales de crecida se pueden interpretar como curvas de probabilidad debido al hecho de que los individuos no están relacionados y son homogéneos.

La mayoría de las curvas de frecuencia, correspondientes a flujos escasos, se pueden interpretar similarmente, pero ocasionalmente se encontrará una muestra correlacionada en serie.

El efecto de usar datos no homogéneos en un problema de regresión se muestra en la figura N° 25 en la cual se hace un gráfico correspondiente a 4 años de descarga media mensual, para cada mes sin excepción, uno de Turkey y otro de Idaho. La relación parece bastante buena, pero en realidad no existe ninguna entre los dos ríos para un determinado mes. La relación aparente usando todos los meses del año se origina por que el caudal en Idaho es semejante al de Turkey. Las descargas de inviernos son menores y en los meses de deshielo de primavera son abundantes.

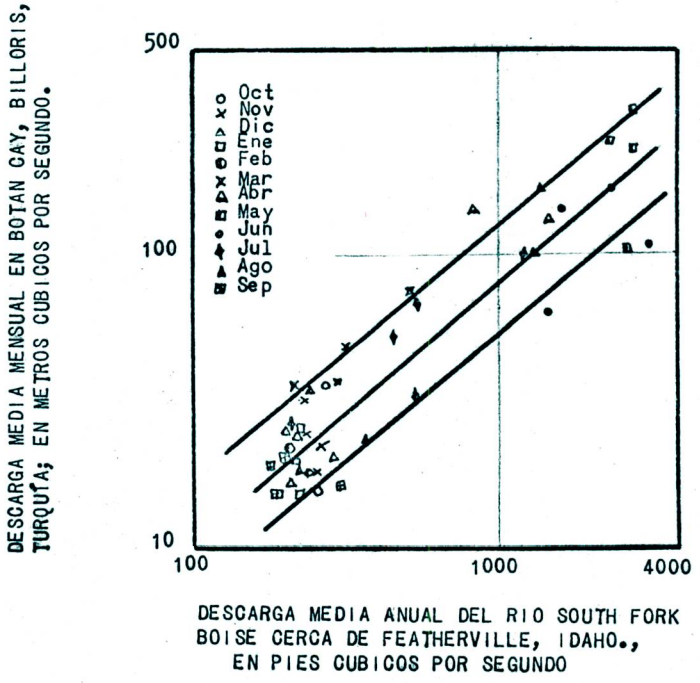


Figura N° 25. - Relación espúrea usando datos no-homogéneos.

Las relaciones entre descargas medias mensuales de dos cuencas contíguas muestran condiciones menos extremas. Por ejemplo, no hay relación entre las descargas medias mensuales del Lago Fork por encima del Lago Moon, en Utah y del río Duchesne en el sendero del río Provo, también en Utah, para el mes de enero, en cambio, hay una buena relación para el mes de junio (figura N° 26). Con pocas excepciones, la relación de dos cuencas de drenaje adyacentes para un determinado mes no es la misma que para cualquier otro mes. Cuando las descargas mensuales correspondientes a todo el año se usan juntas, el coeficiente computado de correlación es muy elevado y el error standard computado será un promedio de los errores standard para las relaciones mensuales individuales.

TESTIGOS. -

Hay muchos factores que influencian el flujo de un río;algunos ejercen mucha influencia en una época y ninguna en otra; la mayoría ejerce efectos interrelacionados con efectos de otros factores. En una regresión que se use para estimar el caudal sólo se puede incluir unos pocos factores, y los efectos de estos son sólo aproximados. En consecuencia hay una dispersión de puntos con respecto a la línea de regresión y en ocasiones se encuentra un punto aislado (ver fig. N° 18, como ejemplo). A tales puntos aislados se les denomina en estadística testigos, y existen pruebas estadísticas que se pueden usar para determinar si un cierto punto se debe o no rechazar por no

pertenecer al grupo. Parece cuestionable el rechazo o no de testigos en los análisis hidrológicos en base a pruebas estadísticas. Considerése el punto-aislado de la figura N° 18. Si la precipitación hubiese sido de 7 pulgadas en lugar de 3, 4; el punto no estaría aislado. Es posible que las precipitaciones en las partes más altas de la cuenca del río Jarbidge fuese mayor que en - Three Creek, en cuyo caso, podría pasar nuevamente lo mismo y se habría - dado más peso en el análisis a ese punto. Sin embargo, si se encuentra que- algunos de los datos para ese año no son dignos de confianza se podría rechazar el punto.

Acton dedicó (1. 959) un breve capítulo al rechazo de los datos indeseables. En parte, él dice: "Pero la pura verdad es que los científicos y los ingenieros no necesitan estímulo para ignorar datos aislados que persisten-más bién necesitan mantenerse en observación".

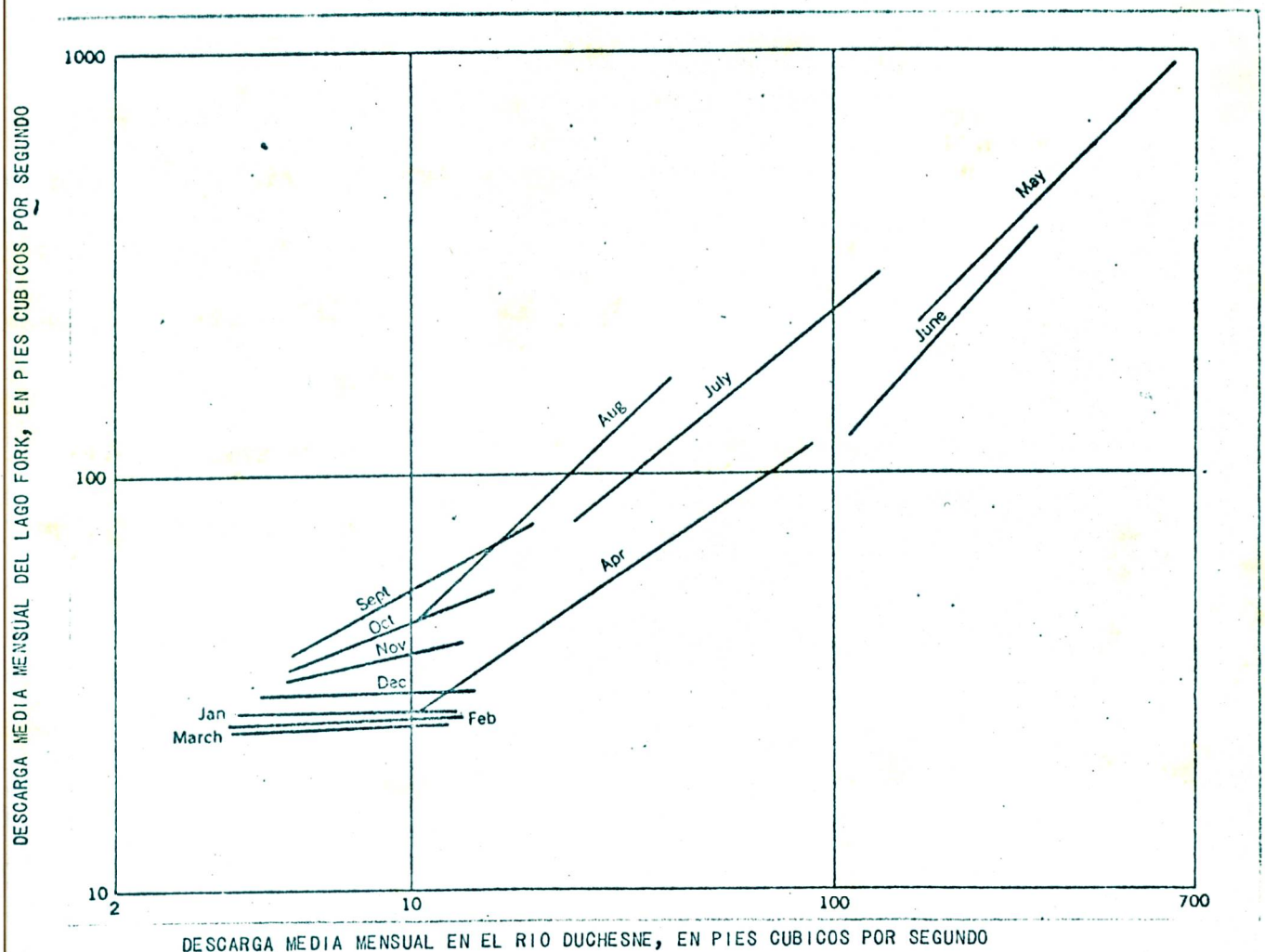


Figura 26.- Relaciones de descarga para meses particulares en dos estaciones de Utah.

## REFERENCIAS SELECCIONADAS


Acton, F. S., 1959, Analysis of straight-line data: New York, John Wiley & Sons, Inc., 265 p.

Amorocho, J. and Hart, W. E., 1964, A critique of current methods in hydrologic systems investigation: Trans. Am. Geophys. Union, v. 45, no. 2, p. 307-321.

Bennett, C. A., and Franklin, N. L., 1954, Statistical methods in chemistry and the chemical industry: New York, John Wiley & Sons, Inc., 724 p.

Benson, M. A., 1960, Characteristics of frequency curves based on a theoretical 1,000-year record, *in* Dalrymple, Tate, Flood-frequency analysis: U.S. Geol. Survey Water-Supply Paper 1543-A, p. 51-74.

——— 1962, Factors influencing the occurrence of floods in a humid region of diverse terrain: U.S. Geol. Survey Water-Supply Paper 1580-B, 64 p.

——— 1965, Spurious correlation in hydraulics and hydrology: Am. Soc. Civil Engineers Proc., v. 91, no. HY4, p. 35-42.

Brownlee, K. A., 1960, Statistical theory and methodology in science and engineering: New York John Wiley & Sons, Inc., 570 p.

Dawdy, D. R., and Matalas, N. C., 1964, Analysis of variance, covariance, and time series, *in* Chow, V. T., Handbook of applied hydrology: New York, McGraw-Hill Book Co., p. 8-68 to 8-90.

Dixon, W. J., and Massey, F. J., 1957, Introduction to statistical analysis: New York, McGraw-Hill Book Co., Inc., 488 p.

Ezekiel, M., 1950, Methods of correlation analysis: 2d ed., New York, John Wiley & Sons, Inc., 531 p.

Ezekiel, M., and Fox, K. A., 1959, Methods of correlation and regression analysis: New York, John Wiley & Sons, Inc., 548 p.

Fisher, R. A., 1950, Statistical methods for research workers: New York, Hafner Publishing Co., 355 p.

Gumbel, E. J., 1958, Statistics of extremes: New York, Columbia Univ. Press, 371 p.

Hazen, Allen, 1930, Flood flows: New York, John Wiley & Sons, Inc., 199 p.

Kendall, M. G., 1952, The advanced theory of statistics: London, Charles Griffin and Co., v. 1, 457 p.

——— 1957, A course in multivariate analysis: London, Charles Griffin and Co., Ltd., 185 p.

Linsley, R. K., Kohler, M. A., and Paulhus, J. L. H., 1949, Applied hydrology: New York, McGraw-Hill Book Co., Inc., 689 p.

McDonald, J. E., 1957, A critical evaluation of correlation methods in climatology and hydrology: Arizona Univ. Inst. Atmospheric Physics Sci. Rept. 4, 35 p.

Mood, A. M., 1950, Introduction to the theory of statistics: New York, McGraw-Hill Book Co., Inc., 433 p.

Riggs, H. C., 1958, Discussion of paper by E. Kuiper, "100 frequency curves of North American rivers": Am. Soc. Civil Engineers Proc., v. 84, no. HY1, paper 1558, p. 61–63.

——— 1960, Discussion of paper by A. L. Sharp, A. E. Gibbs, W. J. Owen, and B. Harris, "Application of the multiple regression approach in evaluating parameters affecting water yields of river basins": Jour. Geophys. Research, v. 65, no. 10, p. 3509–3511.

——— 1965, Effect of land use on low flows in Rappahannock County, Virginia *in* Geological Survey Research 1965, U.S. Geol. Survey Prof. Paper 525–C, p. C196–C198.

Siegel, Sidney, 1956, Nonparametric statistics: New York, McGraw-Hill Book Co., Inc., 312 p.

Snedecor, G. W., 1948, Statistical methods: 4th ed., Iowa State Coll. Press, 485 p.

Snyder, W. M., 1962, Some possibilities for multivariate analysis in hydrologic studies: Jour. Geophys. Research, v. 67, no. 2, p. 721–729.

U.S. Geological Survey, 1949, Floods of August 1940 in the Southeastern States: U.S. Geol. Survey Water-Supply Paper 1066, 554 p.

Wilm, H. G., 1943, Statistical control of hydrologic data from experimental watersheds: Am. Geophys. Union Trans., 1943, pt. 2, p. 618–624 [with discussion by Davenport].

U.S. GOVERNMENT PRINTING OFFICE : 1967—O-275-662