24

INLAND WATERS BRANCH

Application of Regression Analysis in Hydrology

N. TYWONIUK
K. WIEBE

TECHNICAL BULLETIN NO. 24

DEPARTMENT OF ENERGY,
MINES AND RESOURCES

CANADA

TECHNICAL BULLETIN NO. 24

Application of Regression Analysis in Hydrology

N. TYWONIUK
K. WIEBE

INLAND WATERS BRANCH
DEPARTMENT OF ENERGY, MINES AND RESOURCES
OTTAWA, CANADA, 1970

Contents

Illustrations

SECTION 1

Introduction

OBJECTIVE

Regression analysis is one of the oldest statistical tools used in the field of hydrology. Its main applications have been in the study of the relationship between two or more hydrologic variables and the investigation of dependence between the successive values of a series of hydrologic data. For example, in hydrology we are frequently concerned with the dependence of a given variable, Y, on one or more independent variables, X_1, X_2 ...X_n, so that we may predict the value of Y when the values X_1, X_2 ...X_n are known.

In spite of its frequent use in hydrology, a great deal of confusion appears to exist in its application. Regression analysis techniques and correlation analysis techniques are generally discussed together. The distinctions between regression and correlation must therefore be recognized in order to successfully apply and interpret either of the methods. These distinctions are very marked, but are frequently not emphasized because of the similarity of the computation procedures.

In a discussion of the hydrologic application of regression analysis, therefore, the distinction between regression and correlation analysis must first be made clear and an object of this paper is to point out these distinctions. The primary object, however, is to describe the application of regression analysis in hydrology and to point out its limitations. Reference is made periodically to correlation analysis to compare it with regression analysis and to describe how it can be used and interpreted when used together with regression analysis.

BASIC ASSUMPTIONS

There are several requirements or assumptions on which the regression method is based. In most analyses these are as follows (Hahn and Shapins, 1966):

1) the principal requirement of the regression method is that the deviations of the dependent variable about the regression line be normally distributed with the same variance for each value of the independent variable. The assumption of normality is required in establishing confidence limits and in conducting tests of significance, but not in obtaining the least squares fit itself. In contrast, for linear correlation the assumptions are made that the data are drawn from a bivariate or multivariate normal distribution,

2) observations from which the regression is developed are statistically independent,

3) the independent variables are measured (or are known) without error,

4) a correct form of the model has been chosen, and

5) the data are typical, that is, they represent a random sample of the situations concerning which one wishes to generalize.

It should be noted that these assumptions are not completely restrictive, that is, the analysis can be made on data which do not completely comply with the assumptions. With experience in the use of regression analysis the procedures for recognizing and correcting for departures from these assumptions will become evident.

SECTION 2

Linear Regression

SIMPLE LINEAR REGRESSION

In a simple linear regression, a straight-line relationship between two variables, one independent and the other dependent, is assumed. The dependent variable is the one whose values are distributed at random about the regression line, so that its expected value is some function of the observed value of the independent variable.

The general equation is of the form:

$$E_X(Y) = A_o + A_1 X$$

that is, the expected value of each value of Y is a linear function of the corresponding value of X.

A_o and A_1 are the regression co-efficients with the intercept $Y = A_o$ and the slope of the line A_1. There will be one equation for the regression of X upon Y and another for Y upon X.

MULTIPLE LINEAR REGRESSION

In a statistical treatment of a complex model of three or more variables, a multiple regression analysis is used. In this analysis, one variable is considered to be dependent and the remainder are independent.

In the general case of a linear relationship between the mean value of the dependent variable Y and independent variables X_1, X_e, ...X_k; the equation takes the form $Y = A_o + A_1X_1 + A_2X_2 + A_3X_3 \ldots\ldots A_kX_k$.

The parameter A_o is the intercept; the parameter A_1 is the multiple regression co-efficient of Y on X_1 when the effect of all other variables is not considered; the parameter A_2 is the multiple regression co-efficient of Y on X_2 when the effect of all other variables is not considered.

One of the most commonly used mathematical analysis for the multiple linear regression is the method of least squares. This method and other methods of analysis are discussed in detail in Section 4.

SECTION 3

Non-linear Regression

SIMPLE NON-LINEAR REGRESSION

In Section 2 a straight line equation is used to describe the relationship between two variables even when the relationship is more complex than the straight line can portray. When it is important to know the exact form of the relationship it often becomes necessary to express the relationship by means of a non-straight or curvilinear line. There is practically no limit to the different types of curves which are possible and which can be described by mathematical equations. Examples of some equations which are useful in statistical analysis and their corresponding curve shapes (on an arithmetic scale) are shown in Figure 1. The curves shown in Figures 1(a), (d) and (e) illustrate that the curves may have a negative or positive slope depending on the sign of the constants in the mathematical equations.

At this point the use of transformation of variables should be mentioned and perhaps strongly emphasized because of their frequent application in hydrology. There are a number of advantages which may be obtained from transformations (Riggs, 1960):

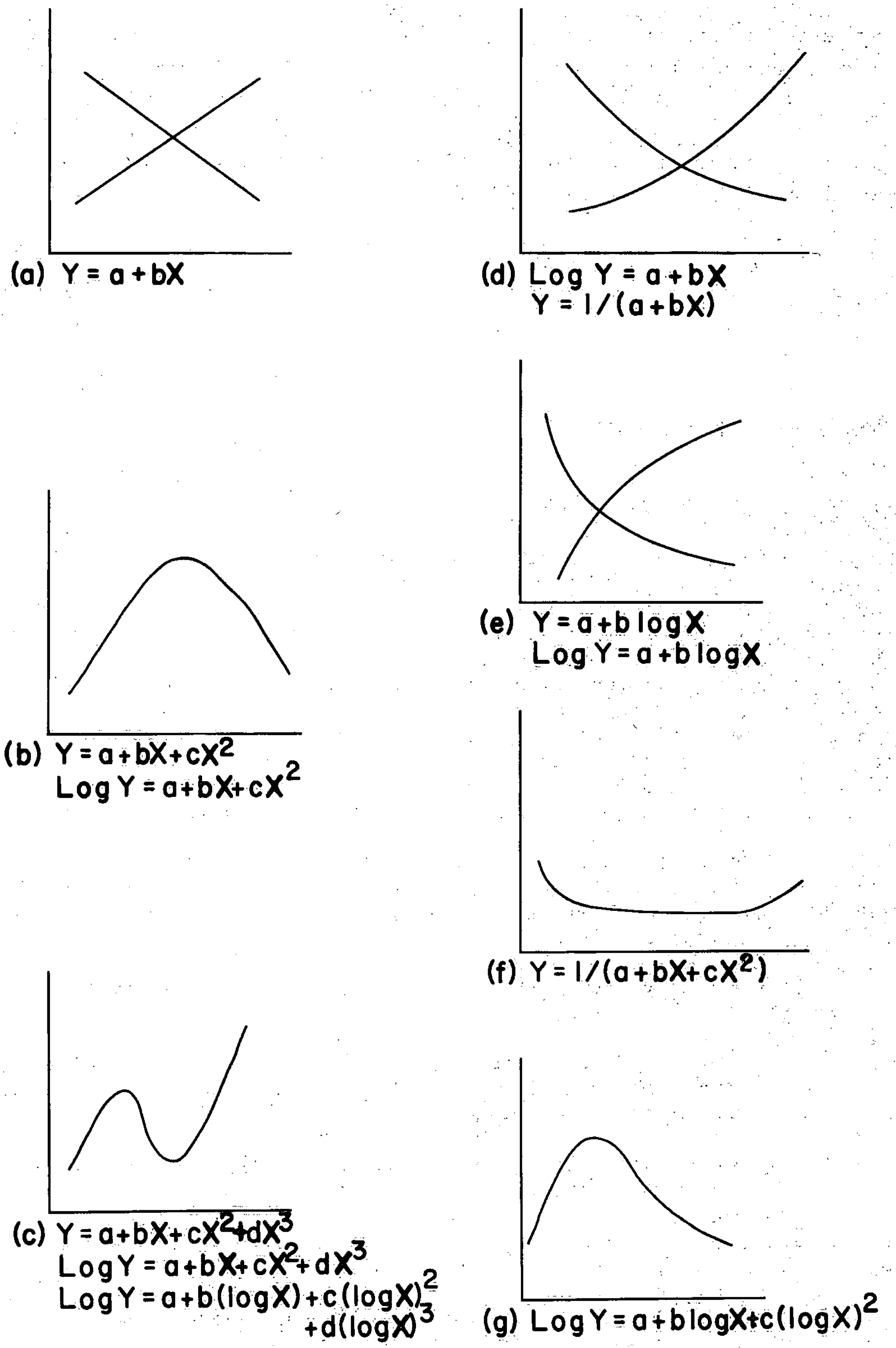
1) a simple linear relationship is obtained among the transformed variables,

2) the marginal distributions of the transformed variables approach more closely to the normal distribution than do those of the untransformed variables, and

3) the variation of the points along the regression line is more homogeneous (that is, the variance is stabilized).

These advantages may or may not all occur in a given transformation.

For example, the function $Y = aX^b$ is linearized by the equation $\log Y = \log a + b \log X$; $Y = a + bX + cX^2$ is linearized by the equation



(a) $Y = a + bX$

(b) $Y = a + bX + cX^2$
$\mathrm{Log}\, Y = a + bX + cX^2$

(c) $Y = a + bX + cX^2 + dX^3$
$\mathrm{Log}\, Y = a + bX + cX^2 + dX^3$
$\mathrm{Log}\, Y = a + b(\log X) + c(\log X)^2 + d(\log X)^3$

(d) $\mathrm{Log}\, Y = a + bX$
$Y = 1/(a + bX)$

(e) $Y = a + b \log X$
$\mathrm{Log}\, Y = a + b \log X$

(f) $Y = 1/(a + bX + cX^2)$

(g) $\mathrm{Log}\, Y = a + b \log X + c(\log X)^2$

Figure 1. Curves illustrating different types of mathematical functions.

$\left[\frac{Y - Y_o}{X - X_o}\right] = a + 2bX_o + c\left[(X - X_o)\right]$; etc. Yevdjevich (1965) has shown that the following additional equations may be transformed into linear form:

1) $Y = be^{aX}$

2) $Y = a + b/X$

3) $Y = X/(a + bX)$

4) $Y = a/(b + cX)$

5) $Y = c + be^{aX}$

6) $Y = c + aX^{b}$

7) $Y = c + \frac{b}{X-a}$

8) $Y = c + \frac{X}{a+bX}$

9) $Y = d + cX + be^{aX}$

10) $Y = dc^{X}\ b^{aX}$

11) $Y = de^{cX} + be^{aX}$

12) $Y = e^{aX}\ (d \cos bX + c \sin bX)$

in which a,b,c,d and e are constants and X and Y are variables between which a relationship is being sought. In hydrology, the logarithmic transformation is used more frequently than the others.

Functions which cannot be readily transformed to a linear relationship or fitted linearly to the data may be developed in the form of a power series. When the higher order terms of the power series are neglected the power series are reduced to a polynomial regression as follows:

$$Y = A_o + A_1X + A_2X^2 + \ldots\ldots A_mX^m$$

in which A_o, A_1 .... A_m are co-efficients which must be evaluated in the analysis and m the order of the variable X of the last term. (m is selected so as to minimize the sum of squares of departure from the line. It should be small to keep the analysis simple and to leave enough degrees of freedom, N-m, for making a reliable estimate of the standard deviation.

The polynomial thus selected from curvilinear regression may be analysed in two ways (Riggs, 1960):

1) the parameters are computed by the least squares method directly; (this involves the solution of m + 1 equations),

2) the values of X, X^2, X^3 ... are treated as the variables u, v ... z and a multiple linear regression procedure is used to determine the parameters.

MULTIPLE NON-LINEAR REGRESSION

In this analysis the curvilinear relationship between a dependent variable and two or more independent variables is determined. The analytical method, the least squares fitting of the regression function, is similar to that used for multiple linear regression. The equations used for determining the regression co-efficients, however, are more complicated, and determination of parameters becomes more difficult as more complex functions are used and as the number of variables increases. The analytical method is usually not practical without the use of a digital computer to carry out the voluminous amount of computational work.

In contrast, the coaxial graphical method of regression analysis, which is described in Section 4, has been widely used in hydrology. Similarily, other graphical techniques can be used (Sharp et al., 1960); however, use of such techniques is generally laborious and reliable statistical inferences become difficult to make.

SECTION 4

Methods of Analysis

CURVE FITTING AND ESTIMATION OF PARAMETERS

Graphical: The most common application of graphical regression analysis is in the case of the simple linear regression. The data are plotted on graph paper and a line of best fit is estimated by eye and drawn through the points. From this regression line, the co-efficients A_o and A_1 are obtained.

If a regression line is non-linear on arithmetic graph paper, it may be linearized by plotting the data on semi-log or log-log graph paper. In this manner the variables are transformed automatically. However, a graphical analysis of the transformed variety should be approached with a certain amount of caution. The data plotted on semi-log or log-log graph paper may appear to have very little scatter, but this is not necessarily true because graph paper has a tendency to make relatively large deviations appear to be minimal.

In the case of the multi-linear regression, the analysis becomes more complex. The graphical analysis is commonly done by the coaxial graphical method. Yevdjevich (1965) illustrates an analysis of this nature in relating basin recharge as the dependent variable to the antecedent precipitation, rainfall amount, etc., as the independent variables.

Graphical analysis is frequently used in hydrology for reasons expressed by Lyons and Cavadias in their discussion of the paper by Solomon (1966). Some of these reasons are:

1) graphical relationships are visible and thus can be more easily understood,

2) little or no knowledge of statistical methods is necessary in applying graphical techniques,

3) graphical techniques do not require the large number of manual computations necessary for analytical methods,

4) graphical regression curves can take any reasonable shape that can be drawn, whereas the analytical methods must assume a shape for the curves at the outset.

Analytical: Analytical methods, however, do have advantages over the graphical method and, by the use of electronic computers, the otherwise tedious computations are easily performed.

The most commonly used analytical method is the least squares linear regression analysis. This analysis, as outlined by Neville and Kennedy (1964) for the simple linear regression, is based on the hypothesis; if Y is a linear function of an independent variable X, the most probable position of a line $Y = A_o + A_1X$ is such that the sum of squares of deviations of all points (X_i, Y_i) from the line is a minimum. The deviations are measured in the direction of the Y-axis.

The general form of analysis is as follows:

Assume n pairs of observations. Applying the general equation $Y = A_o + A_1X$, the problem is to find the values of A_o and A_1 for the line of "best fit". For a point, X_iY_i, on the line we have; $Y_i - (A_o + A_1X_i) = 0$. However, if there is error in the measurement, there will be a residual E_i and the equation becomes; $Y_i - (A_o + A_1X_i) = E_i$. Thus, for n observations we have n equations;

$$Y_1 - (A_o + A_1X_1) = E_1$$

$$Y_2 - (A_o + A_1X_2) = E_2$$

$$\begin{matrix} ' & ' & ' \\ ' & ' & ' \end{matrix}$$

$$Y_n - (A_o + A_1X_n) = E_n$$

The sum of the residuals is defined as $P = \sum Ei^2$ or

$$P = \sum_{i=1}^{n} \left[Y_i - (A_o + A_1X_i)\right]^2$$

The hypothesis is satisfied when the sum of the squares of residuals is a minimum, that is, when P is a minimum. This occurs when

$$\frac{\partial P}{\partial A_o} = 0 \quad \text{and} \quad \frac{\partial P}{\partial A_1} = 0$$

or $\sum_{i=1}^{m} [Y_i - (A_o + A_1X_i)] = 0$, and $\sum_{i=1}^{m} X_i [Y_i - (A_o + A_1X_i)] = 0$

By removing the subscripts, we get the normal equations:

$$\sum Y = NA_o + A_1\sum X$$

$$\sum XY = A_o\sum X + A_1\sum X^2$$

Thus, by solving the normal equations for A_o and A_1, we obtain:

$$A_o = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$$

$$A_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

And finally, by substituting these values into the general equation we get;

$$Y = \frac{\sum X^2 \sum Y - \sum Y \sum XY}{n\sum X^2 - (\sum X)^2} + \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} X$$

The more complex models such as multiple linear, non-linear and multiple non-linear regressions are analysed in a similar fashion. If the model contains n variables, we obtain n normal equations similar to the equations which were derived for the simple linear regression model. Equations for multiple linear regressions of up to six variables are shown by Solomon (1966).

Another analytical method occasionally used for the analysis of parameters in a regression model is the Maximum Likelihood Method. By this method, the value of a parameter is determined to make the probability of obtaining the observed outcome as high as possible. In mathematical terms it is stated as, $\partial \log p(X)/\partial u = 0$, where $p(X)$ is probability density and u is a statistical parameter. The use of this method is rather infrequent due to its complexity.

MEASURES OF ASSOCIATION

The most commonly used statistical parameter for measuring the degree of association of two linearly dependent variables is the correlation co-efficient which is mathematically defined as follows:

$$r = \frac{\sum(\Delta X_i \ \Delta Y_i)}{\sqrt{\sum(\Delta X_i)^2 \sum(\Delta Y_i)^2}}$$

or $$r = \frac{\sum X_i Y_i - N\bar{X}\bar{Y}}{s_X \ s_Y}$$

where s_X and s_Y are the standard deviations of X_i and Y_i respectively, $\Delta X_i = X_i - \bar{X}$ and $\Delta Y_i = Y_i - \bar{Y}$. If there is no linear relationship between variables, then $r = 0$. If there is a functional linear relationship, then $r = \pm 1$.

In Section 2, the regression co-efficients are defined. The co-efficient A_1 may be expressed in terms of the correlation co-efficient r and the standard deviations s_X and s_Y in the following manner.

$A_{1Y} = r(s_Y/s_X)$ for regression Y versus X. For regression X versus Y the equation becomes;

$$A_{1X} = r(s_X/s_Y)$$

From these equations we obtain the co-efficient of determination;

$$D = r^2 = A_{1X}A_{1Y}$$

The co-efficient of determination is a measure of the difference between the variance of the observed values Y_i and the variance of the values determined for given values of X_i by the use of the linear regression line. As D increases, the difference between the variance of observed values and the variance of the determined values decreases.

Another measure of association is the standard deviation of residuals which is expressed mathematically as follows:

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{n}} = s_Y \sqrt{1 - r^2}$$

$$\Delta Y_i = Y_i - Y$$

where Y_i is the observed value and Y is the value determined by the regression line for a given $X = X_i$.

The unbiased standard deviation of residuals is given as:

$$\hat{S}_Y = \sqrt{\frac{n - 1}{n - 2}} \quad s_Y \sqrt{1 - r^2}$$

Since $\hat{S}_Y$ is the spread of points around the regression line, a large value of $\hat{S}_Y$ indicates that the spread is large.

In a multiple linear regression, the degree of correlation of a dependent variable to two or more externally independent variables can be measured by the use of variations of the previously described parameters. The multiple correlation co-efficient takes the form

$$R_1 = \frac{s_{ei}}{s_1} = \sqrt{1 - \frac{S_1{}^2}{s_1}}$$

where s_{ei} is the standard deviation of estimated values, s_1 is the standard deviation of actual values of X_1 and S_1 is the standard deviation of residuals.

The co-efficient of multiple determination D_1 then becomes equal to $R_1{}^2$.

The unbiased standard deviation of residuals takes the form

$$\hat{S}_1 = \sqrt{\frac{n\ S_1{}^2}{n-m}}$$

where $\hat{S}_1$ is the unbiased and S_1 is the biased standard deviation of residuals and m denotes the number of parameters in the regression model.

STATISTICAL INFERENCE

Because the regression lines and the degrees of association are measured by parameters, the reliability of these estimated parameters is important. This reliability is measured by the procedures commonly referred to as confidence limits and tests of significance.

The confidence interval is that interval around the estimated parameter within which a given percentage of parameters of a large number of samples is expected to be found. This given percentage is referred to as the level of confidence. For example, the confidence limit at the 95 percent level means that, out of 100 samples of equal size, it is expected that 95 values of a parameter would be inside that interval.

If the estimated parameter falls inside the confidence limits, it may be significant or non-significant depending on the problem. If it falls outside, it may be considered as non-significant or significant depending on the problem. The limits in hydrology are generally chosen at between 80 and 95 percent either by convention or by judgment.

The computation of confidence limits will first be illustrated for the case of simple linear regression defined by $Y = A_o + A_1X$. The variance of an estimate enables us to form confidence limits of the estimate. In considering the variance about a regression line, the deviations are taken from the line instead of from the mean. Thus the variance of Y, estimated by the regression line, is the sum of squares of deviations divided by the number of degrees of freedom, ν, available for calculating the regression line:

$$s_Y{}^2 = \frac{\sum E_i{}^2}{\nu}$$

in which $E_i = Y_i - (A_o + A_1X_i)$. For n observations there are n - 2 degrees of freedom (since two constraints determine the regression line: the centroidal point $(\bar{X},\bar{Y})$ and either slope A_1 or intercept A_o). Hence,

$$s_Y{}^2 = \frac{\sum E_i{}^2}{n-2}$$

The variance of the mean value of Y, that is, $\bar{Y}$, is given by

$$s_{\bar{Y}}{}^2 = \frac{sY^2}{n}$$

and confidence limits are written for $\bar{Y}$ for some desired level of significance of t and for the appropriate number of degrees of freedom. Thus the true value of $\bar{Y}$ lies within the interval $\bar{Y} \pm ts_{\bar{Y}}$.

To obtain a closer estimate of the confidence limits, the limits for $\hat{Y}_i$ corresponding to a specific value of X_i can be calculated. To do this we need the variance of $\hat{Y}_i$ (denoted by Y). The limits are for the mean estimated value for Y_i, that is, $Y_{\bar{i}}$. The variance of this mean value is:

$$s_{\bar{Y}_i}{}^2 = sY^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum(X - \bar{X})^2} \right]$$

The confidence interval for the mean estimated value of $Y_{\bar{i}}$ is then

$$Y_{\bar{i}} \pm ts_{Y_{\bar{i}}}$$

In predicting the confidence interval of a single estimated value $\hat{Y}_i$, the variance of a single value is used as follows:

$$s_{Y_i}{}^2 = sY^2 \left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum(X - \bar{X})^2} \right]$$

The confidence interval then becomes

$$Y \pm ts_{Y_i}$$

The variance of the slope, A_1, can be shown to be

$$s_{A_1}{}^2 = \frac{sY^2}{\sum(X-\bar{X})^2}$$

The confidence band for the slope is therefore represented by a double fan-shaped area with slopes of $A_1 \pm ts_{A_1}$ and apex $(\bar{X}, \bar{Y})$. The effect of the slope confidence bands is that the confidence area of the regression now becomes bounded by smooth curves asymptotic to the confidence interval of the slope near the ends of the range of the observations. This is graphically illustrated in Figure 2.

To define the regression line by use of the intercept A_o, we must find the variance of A_o as a particular case of the variance of any mean estimated value $\hat{Y}_{\bar{i}}$. In this case, $X_i = 0$ and the variance of A_o is given by

$$s_{A_o}{}^2 = sY^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X - \bar{X})^2} \right]$$

When the theoretical value, A_1^o, of the slope A_1 is known, a t-test

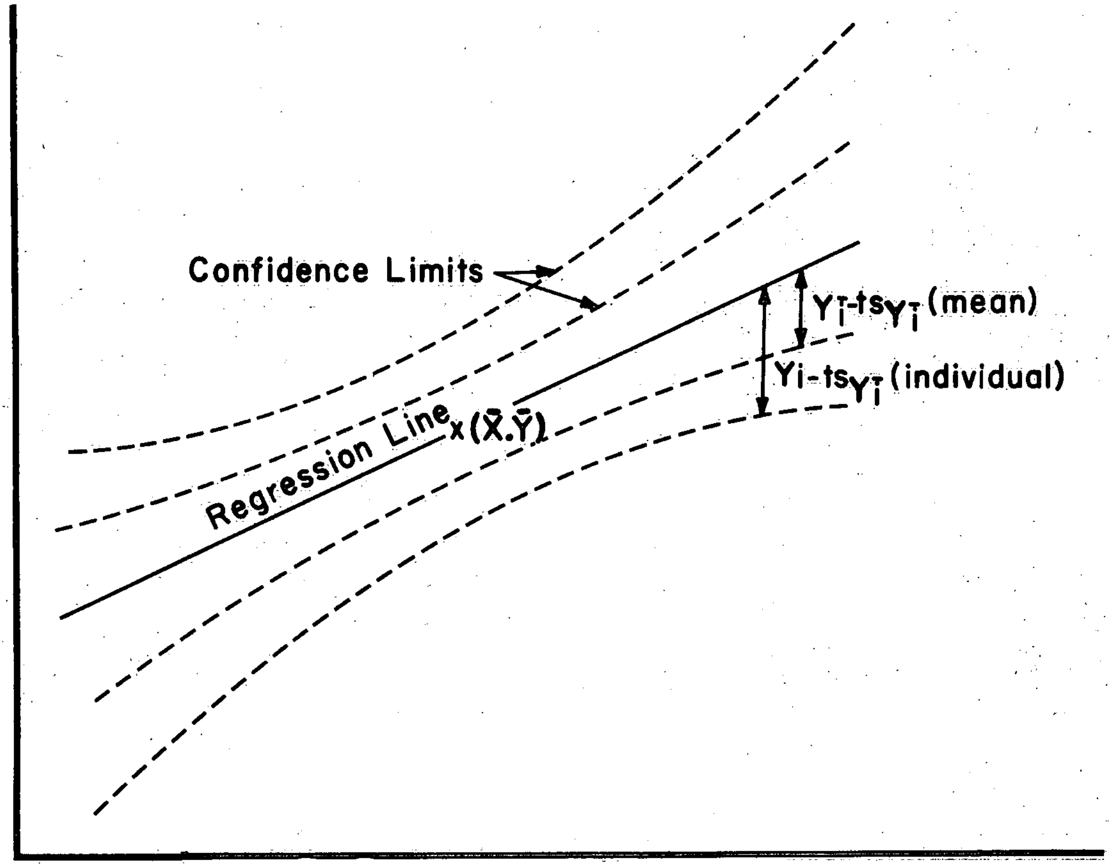


Figure 2. Confidence limits illustrating the confidence area of the mean and the individual points for simple linear regression.

may be used to determine whether there is a significant difference between A_{o_1} and the value of A_1 given by the regression line. The test is applied to $|A_1 - A_{o_1}|$. The standard deviation of $|A_1 - A_{o_1}|$ is equal to the standard deviation of A_1, s_{A_1}, because A_{o_1} is free from error. Thus

$$t = \frac{|A_1 - A_{o_1}|}{s_{A_1}}$$

and if the calculated t is greater than that tabulated for the required level of significance then we conclude there is a significant difference between A_{o_1} and A_1. A similar test can be applied to the intercept A_o in which

$$t = \frac{|A_o - A_{o_o}|}{s_{A_o}}$$

In the case of multiple linear regression, expressed as

$$y = A_o + A_1X_1 + A_2X_2 + \ldots A_kX_k$$

in which A_o is a constant and A_1, A_2 ... A_k are partial regression co-efficients, the confidence limits can be established in a manner similar to that described for simple linear regression. In the analysis, the partial regression co-efficients are used together with the standard deviations of the partial regression co-efficients, s_{A_1}, s_{A_2}, etc. The tests of significance of the regression co-efficients are

$$t = \frac{A_1}{s_{A_I}} \quad \text{for the first regression co-efficient}$$

$$t = \frac{A_2}{s_{A_2}} \quad \text{for the second, etc.}$$

The number of degrees of freedom for this case is $\nu = n-k-1$. This test effectively tests to see if the selected independent variables significantly influence the variable Y. If the regression equation is found not to be statistically significant, then the regression equation must be revised.

To test if the regression co-efficient, A_i, significantly differs from a known theoretical value, A^o_j, the t test is applied:

$$t = \frac{\left|A_j - A^o_j\right|}{s_A}$$

The null hypothesis is rejected if t for some required level of significance is larger than the critical tabulated value for n-k-1 degrees of freedom.

The test to determine if the multiple regression as a whole is statistically significant is done by comparison of the variance contributed by the regression and the error variance $s_Y{}^2$. This is done by means of the F-test. For example, for k=2, that is, three co-efficients are used (A_o, A_1 and A_2), the sum of the squares of deviations, $\sum c^2$, in Y accounted by the regression is

$$\sum c^2 = A_1\sum(Y-\bar{Y})(X_1-\bar{X}_1) + A_2\sum(Y-\bar{Y})(X_2-\bar{X}_2)$$

Therefore,

$$F = \frac{\sum c^2/k}{s_Y{}^2}$$

with the number of degrees of freedom $\nu = k = 2$ for the numerator, and $\nu_2 = n-k-1 = n-3$ for the denominator. If the computed F is greater than the value tabulated, then the hypothesis that all the true partial regression co-efficients are equal to zero is rejected.

The analytical method for simple curvilinear and multiple curvilinear regression is similar to that used for linear regression. The equations used become relatively more complicated and use of digital computers is usually made in their analysis.

SECTION 5

Conclusions

In conclusion a number of general items can be summarized. In the regression analysis the following basic procedure of analysis may be used: the data must first be selected and examined. On the basis of physical consideration and/or convention a regression model is established. Either graphical or analytical methods can then be used in fitting the curve and estimating the parameters of the association. Calculation of the regression co-efficients, correlation co-efficients, etc. may be made to determine the degree of association of the variables and finally, significance tests may be made to determine the reliability of the analysis.

Hydrologic data rarely, if ever, completely meet the requirements or assumptions on which the regression method is based. Its main difference from correlation analysis should again be emphasized: the regression method does not require that the variables be normally distributed as does the correlation method. Although the correlation co-efficient has been shown to be a measure of association between variables, it may not be a reliable measure if at least one variable does not conform to a normal distribution. Limited confidence should be placed on t-tests and co-efficients of determination and correlation until the absence of skewness in the data is established (Sharp and others, 1960). Hydrologists sometimes tend to evaluate results of regression analysis as if they were correlation analysis. The regression problem must be recognized if an appropriate interpretation of the analysis is to be made.

A number of other precautions must be observed in the application of the regression method:

1) indexes are often used to describe quantitatively one or more variables of a regression model. Selection of indexes is difficult; they must accurately reflect the effects, no two should describe the same thing, and a variable or any portion thereof should not appear on both sides of the equation. Several different indexes are often tried in the analysis.

2) a variable that seems unimportant over a limited range of variation may be very important over a broader range.

3) if many variables are used in the analysis spurious statistical significance (due to chance) may result. The variables should therefore be carefully scrutinized on the basis of physical considerations.

4) highly correlated variables should not both be used in the regression equation. Only the variable which has a true effect on the dependent variable should be used.

5) omissions of important variables from the model (that is, variables for which no data are available) may result in biased estimates of the terms included in the model. Such omissions can sometimes be detected by examination of the calculated residual values and then closer examination of those actual values which differ the most from the predicted.

6) a correct form of the regression model must be selected. Linear models are relatively simple and should be used whenever possible. When a linear model is found inadequate a more complex model should be tried. An F test can be used to evaluate if use of a more complex model results in significant improvement over the simpler model. It is often also useful to plot each independent variable against the dependent variable and from these determine if the relationship should be linear or non-linear. It should be noted, however, that inclusion of too many variables in a multiple regression analysis often leads to difficulty in the physical interpretation of the results and is often not good practice.

7) although it was assumed that the deviations of the dependent variable about the regression line are normally distributed with constant variance, the F-test for significance and the t-tests for setting confidence limits are frequently insensitive to slight deviations from these assumptions. The normality assumption may be examined by plotting the calculated residuals on normal probability paper.

8) if the observations from which the regression is developed are not statistically independent, use may be made of "dummy variables" in the model to account for the different items or other identifiable variables which affect the dependent variable but which cannot be expressed on a quantitative scale.

9) errors in the measurement of independent variables are usually ignored if they are slight. Appreciable errors, however, may disqualify the use of the regression method.

10) although there can be no linear relationship among the independent variables (such as one variable being a multiple of another), non-linear relationship (such as X_1, and $X_1{}^2$) may be used in complex models.

11) the number of observations must exceed the number of constants in the regression model. At least 10 more data points than

unknown co-efficients should be used to assure a satisfactory number of degrees of freedom for the estimation of the standard error of the regression.

Graphical and analytical regression analyses are among the most useful statistical tools in hydrology. The use of the digital computer considerably reduces the work of the analytical analysis making it possible to include variables which would otherwise go unnoticed. The analytical analysis must be done with care and with the help of graphical plots, otherwise it is possible to get nonsense correlations or to draw unwarranted inferences from the data.

APPENDIX A

Examples of Regression Analysis in Hydrology

Regression analysis is a valuable tool in many kinds of studies. The analyses have been used in studying streamflow, forecasting water supplies from snowmelt, estimating flood peaks, detecting changes in streamflow resulting from changes in the use and treatment of land, and in solving many other types of operational problems. Some examples of regression analysis with which the authors are familiar are described below in greater detail.

Examples of simple regression analysis (two variables) are as follows:

1) runoff related to precipitation for a drainage basin,

2) discharge related to stage for a particular cross-section of a stream,

3) suspended sediment discharge related to water discharge for a particular cross-section of a stream,

4) snowmelt related to radiation.

In the case of multiple regression analysis, a more complex model is used. The dependent variable is related to two or more independent variables. Examples of multiple regression analysis are as follows:

1) surface runoff related to precipitation, antecedent precipitation, groundwater in storage, temperature, percentage of watershed in row crops, percentage in pasture and terraces (Sharp et al., 1960),

2) evaporation related to temperature, wind velocity and humidity,

3) suspended sediment discharge related to water discharge, basin condition (availability of sediment) and time in days since the preceding peak flow,

4) basin recharge related to antecedent precipitation, week of year, storm duration and storm precipitation.

These examples are only a few of the many that are found in literature.

References

Abraham, Charles E. 1969. Suspended Sediment Discharges in Streams. Paper presented at AGU Golden Anniversary Meeting in Washington, D.C. April 21-25, 1969.

Ezekiel, M. and K.A. Fox. 1966. Methods of Correlation and Regression Analysis, John Wiley and Sons, Inc. New York.

Hahn, G.J. and S.S. Shapins. 1966. The Use and Misuse of Multiple Regression, Ind. Quality Control, Oct. 1966, pp. 184-186.

Neville, A.M. and J.B. Kennedy. 1964. Basic Statistical Methods for Engineers & Scientists. Int. Text Book Co., Scranton, Penn.

Renard, K.G. 1969. Sediment Rating Curves in Ephemeral Streams. Trans. ASAE, 12, 80-85

Riggs, H.C. 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*, J.G.R., 65, 3509-3511.

Sharp, A.L., A.E. Gibbs, W.J. Owen, B. Harris. 1960. Application of the Multiple Regression Approach in Evaluating Parameters Affecting Water Yields of River Basins. J.G.R., 63, 4.

Solomon, S. 1966. Statistical Association between Hydrologic Variables. Proceedings of Hydrology Symposium No. 5, Feb. 1966, pp. 55-114.

Yevdjevich, V.M. 1965. Regression and Correlation Analysis, *in* Handbook of Applied Hydrology, V.T. Chow, Ed., McGraw Hill Book Co., New York.

Current Technical Bulletins

No. 12 Sediment surveys in Canada. W. Stichling and T.F. Smith, 1969.

An outline of the Sediment Survey Program of the Water Survey of Canada, including methods, instrumentation and data available.

No. 13 Climatology studies of Baffin Island, Northwest Territories. R.G. Barry and S. Fogarasi, 1969.

A report of the results of a program of climatological investigations in Baffin Island.

No. 14 Hydrology of the Good Spirit Lake Drainage Basin, Saskatchewan: A preliminary analysis. R.A. Freeze, 1969.

A report on the first water balance for the Good Spirit Lake Drainage Basin in a discussion of the methods used in the investigations.

No. 15 Digitizing hydrographs and barographs. T.W. Maxim and J.A. Gilliland, 1969.

A discussion of the conversion of analogue water level recorder graphs and barographs to digital form using a pencil follower and key punch.

No. 16 The computation and interpretation of the power spectra of water quality data. A. Demayo, 1969.

A discussion of the concept of spectral analysis and the method of calculating power spectra of water quality data, including a practical example and the computer program used to perform this type of calculation.

No. 17 Groundwater Investigation - Mount Kobau, British Columbia. E.C. Halstead, 1969.

A report of the results obtained from a study of the groundwater storage system at the summit of Mount Kobau, British Columbia.

No. 18 The effects of the W.A.C. Bennett Dam on downstream levels and flows. A. Coulson and R.J. Adamcyk, 1969.

A report summarizing the expected effects of the W.A.C. Bennett Dam on levels and flows in the Mackenzie River basin.

No. 19 Airborne techniques in climatology; oasis effects above prairie surface features. R.M. Holmes, 1970.

A report describing a pilot study of oasis effects in southern Alberta using a specially-instrumented aircraft and a mobile ground station.

No. 20 Hydrogeological Reconnaissance of the North Nashwaaksis River Basin, New Brunswick. J.E. Charron, 1969.

A description of a hydrogeological reconnaissance carried out as part of an International Hydrological Decade study of the hydrology of the North Nashwaaksis Basin.

No. 21 An instrumented experimental site for the investigation of soil moisture, frost and groundwater discharge. R.A. Freeze and J.A. Banner, 1970.

A report describing an instrumented experimental site at Calgary, Alberta, to provide integrated measurements of the subsurface moisture regime in saturated and unsaturated zones. A summary of the first year's operation is included.

No. 22 Detergents, phosphates and water pollution. W.J. Traversy, P.D. Goulden and G. Kerr, 1970.

A report on the results of chemical analyses of phosphate content in detergents and washing products. The report traces the development of washing products from organic soaps to modern phosphate-based detergents and describes the relationship between phosphates and the eutrophication process.

No. 23 Regional Groundwater Flow Between Lake Ontario and Lake Simcoe. C.J. Haefeli, 1970.

A report on the hydrogeological conditions to the north of Toronto with a view to determining if the terrestrial water balance of the Lake Ontario Basin is affected by a major seepage from Lake Simcoe.

Copies may be obtained from the Director, Inland Waters Branch, Department of Energy, Mines and Resources, 588 Booth Street, Ottawa, Ontario.