11.4.4 Comments on Multiple Regression

As demonstrated through the example, multiple regression is suitable in situations where one dependent variable and several independent variables are available and it is desired to fit a linear model containing all of the significant independent variables. Two questions that may be asked are: (1) Is a linear model suitable for the problem? and (2) which independent variables should be included in the model?

The first question can be answered by plotting the data and computing the statistical performance indices. Regarding the second question, a factor to be considered in selection of the variables is that in most cases, the independent variables are not statistically independent and are correlated. Hence while using regression analysis, the correlation matrix should be computed in the beginning.

If the regression equation contains independent variables that are highly correlated among themselves then (besides the mathematical difficulties in determining the coefficients) the interpretation of the regression coefficients becomes difficult. Many times the sign and magnitudes of coefficient of a variable may be different than what is expected if the corresponding variable is highly correlated with another independent variable in the equation.

A common practice in multiple regression analysis is to perform several regressions with the given set of data using different combinations of the independent variables. Finally the regression equation that best fits the data is selected. A commonly used criterion for the best fit is to select the equation that gives the smallest value of the sum of squares of errors. A scatter plot of observed and computed values is always helpful.

All of the variables included in a regression equation should make a significant contribution to the model unless there is an overriding reason (theoretical or intuitive) for retaining a particular variable. The variables retained should have physical significance. If two variables are equally significant when used alone and are highly correlated than the one that is easiest to obtain should be used.

The number of variables retained in regression should be small compared to the number of observation and should not exceed 25 to 30 percent of the number of observations. This is a rule of thumb to avoid “over-fitting” whereby oscillations in the prediction may occur.

11.5 Stepwise Regression

Stepwise regression is a procedure that is commonly used to select the "best" regression equation (by including only relevant variables) from amongst a number of independent variables. This approach consists of building the regression equation by adding one variable at a time. At each step, all the variables in the regression equation are examined for significance and at any stage, a particular variable is removed if it is no longer explaining a significant variation of the dependent variable.

To begin with, the first variable to be added in the regression equation is the one which has the highest correlation with the dependent variable. The second variable to be added is the one that explains the largest remaining unexplained variation in the dependent variable. At this stage the first variable is tested for significance and retained or discarded depending on the results of this test. The third variable added is the one that explains the largest portion of the variation that is not explained by the two regression variables already in the equation. The variables in the equation are then tested for significance. This procedure is repeated till a situation is reached when all the variables that are not in the equation are insignificant and all the variables that are in the equation are significant. This is a very good approach to use but care must be exercised to ensure that the resulting equation is rational.

The steps of stepwise regression are:

i. The variable which has the highest correlation with the dependent variable is picked up as the first independent variable.
ii. The variable which explains the largest of the residual variation in the dependent variable after the first step is added as the next variable.
iii. Test the significance of the new variable and retain or discard it depending on the results of this test.
iv. Repeat steps (ii) & (iii) until each of the variables that are not in the equation are found to be insignificant and all the variables in the equation are significant.

The real test of how good is the regression model (or any other model), is the ability of the model to predict the dependent variable by using the observations of the independent variables that were not used in estimating the regression coefficients. For this purpose, the data are divided into two parts. One part of the data is used to develop the model and the other part to test the model. Transformation of independent variables may significantly improve the regression relationship.

The difference between multiple and stepwise regression is that in multiple linear regression all independent variables are included in the regression model, whereas in stepwise regression, the equation is built up step by step by taking those independent variables into consideration first, which reduce the error variance most. The entry of new independent variables is continued until the reduction in the error variance falls below a certain limit. In some stepwise regression tools a distinction is made between free and forced independent variables. A forced variable will always be included into the equation no matter what error variance reduction it produces, whereas a free variable enters only if the error variance reduction criterion is met.

11.6 Transforming Non Linear Relations

Relationship between some variables may be non-linear but can be transformed to linear form so that the technique of linear regression can be applied. For example, consider that two variables X and Y are non-linearly related as follows:

$$Y = \alpha X^{\beta} \tag{11.57}$$

This non-linear relation can be linearized by logarithmic transformation of the equation

$$\text{Ln } Y = \text{Ln } \alpha + \beta \text{Ln } X \tag{11.58}$$

or

$$A = a + b*B \tag{11.59}$$

where A = Ln Y, a = Ln α, b = β, and B = Ln X. Now, one can use the regression technique to estimate parameters *a* and *b* and thereby α and β. In this procedure, two important points are worth noting.

a) The values of a and b are estimated by minimizing $\Sigma(A - A_{reg})^2$ and not by minimizing $\Sigma(Y - Y_{reg})^2$. Here A_{reg} and Y_{reg} are the value of A and Y estimated by the regression equation.
b) In the log-transformed equation, the error term is additive (A = a + b B+ c) which means that it is multiplicative in the original equation

$$Y = \alpha X^{\beta} \varepsilon \tag{11.60}$$

The errors are related as c = ln ε. Hence, the assumptions in hypothesis testing and confidence intervals should be valid for c.

In some cases, it has been observed that the log transformed data follows the regression assumptions more closely than the original data. The standard regression is based on a constant absolute error along the regression line whereas the normal equations for a logarithmic transformation are based on a constant percentage error along the regression line.

11.7 Closure

This module has discussed two statistical tools with are very commonly employed in hydrology for variety of tasks. Correlation is commonly used to study the degree of statistical relationship between a set of variables. Regression has found diverse applications when the variables have a cause-and-effect relationship.

References

Davis, J.C. (1986). Statistics and Data Analysis in Geology, John Wiley & Sons, New York.

Haan, C.T. (2002). Statistical Methods in Hydrology. Iowa State Press, Ames, U.S.A.

Hirsch, R.M., Helsel, D.R., Cohn, T.A., and Gilroy, E.J. (1993). Statistical Analysis of Hydrologic Data, in Handbook of Hydrology, edited by D.R. Maidment. McGraw-Hill Inc., New York.

Hosking, J.R.M. (1986). The theory of probability-weighted moments. Technical Report RC 12210. Mathematics, IBM Thomas J. Watson Research Center, Yorktown Heights, New York.

Hosking, J.R.M. (1990). L-Moments: Analysis and estimation of distribution using linear combination of order statistics. J. of Royal Statistical Soc., Series B, 52(1), 105-124.

Kite, G.W. (1977). Frequency and Risk Analysis in Hydrology. Water Resources Publications, Colorado.

McCuen, R.H. (1993). Statistical Hydrology. Prentice-Hall, New Jersey.

Rao, A.R. and Hamed, K.H., (1994. Frequency analysis of upper Cauvery flood data by L-moments. Water Resources Management. Vol.8, 183-201.

Salas, J.D., Delleur, J.R., Yevjevich, Y., and Lane, W.I. (1980). Applied Modeling of Hydrologic Time Series. Water Resources Publications, Colorado.

Salas, J.D. (1993). Analysis and Modeling of Hydrologic Time Series, in Handbook of Hydrology, edited by D.R. Maidment. McGraw-Hill Inc., New York.

Yevjevich, V. (1972). Probability and Statistics in Hydrology. Water Resources Publications, Fort Collins, Colorado.