Technical Assistance
HYDROLOGY PROJECT
World Bank & Government of The Netherlands funded

Training module # SWDP - 37

How to do hydrological data validation using regression

New Delhi, February 2002

CSMRS Building, 4th Floor, Olof Palme Marg, Hauz Khas,
New Delhi – 11 00 16 India
Tel: 68 61 681 / 84 Fax: (+ 91 11) 68 61 685
E-Mail: hydrologyproject@vsnl.com

DHV Consultants BV & DELFT HYDRAULICS

with
HALCROW, TAHAL, CES, ORG & JPS

Table of contents

1. Module context

While designing a training course, the relationship between this module and the others, would be maintained by keeping them close together in the syllabus and place them in a logical sequence. The actual selection of the topics and the depth of training would, of course, depend on the training needs of the participants, i.e. their knowledge level and skills performance upon the start of the course.

2. Module profile

Title	:	How to do hydrological data validation using regression
Target group	:	Hydrologists, Data Processing Centre Managers
Duration	:	Two sessions of 60 min each.
Objectives	:	After training, the participants will be able to 1. carry out hydrological data validation using regression 2. filling in missing data using regression
Key concepts	:	• regression analysis & double mass • multiple & step-wise regression • analysis of variance • establishment of relationship • percent variation explained • standard error
Training methods	:	Lecture, exercises
Training tools required	:	Board, OHP, Computer
Handouts	:	As provided in this module
Further reading and references	:	

3. Session plan

No	Activities	Time	Tools
1	***Introduction***	10 min	
	• Use of regression analysis		OHS 1
	• Linear and non-linear regression equations		OHS 2
	• Suitable regression model		OHS 3
	• General form of rainfall-runoff relation		OHS 4
	• Use of regression model for discharge validation		OHS 5
	• Regression data vector		OHS 6
2	***Simple linear regression***	25 min	
	• General features		OHS 7
	• Estimation of regression coefficients		OHS 8
	• Measure for goodness of fit		OHS 9
	• Confidence limits		OHS 10
	• Example for annual rainfall and runoff data		OHS 11
	• Records of rainfall and runoff		OHS 12
	• Regression fit rainfall-runoff		OHS 13
	• Plot of residual versus rainfall		OHS 14
	• Plot of residual versus time		OHS 15
	• Plot of accumulated residual		OHS 16
	• Double mass analysis observed and computed runoff		OHS 17
	• Plot of rainfall versus corrected runoff		OHS 18
	• Plot of residual (corrected) versus runoff		OHS 19
	• Plot of residual (corrected) versus time		OHS 20
	• Plot of regression line with confidence limits		OHS 21
	• Extrapolation		OHS 22
			OHS 23
3	***Multiple and stepwise regression***	15 min	
	• Multiple linear regression models		OHS 24
	• Estimation of regression coefficients		OHS 25
	• Analysis of variance table (ANOVA)		OHS 26
	• Coefficient of determination		OHS 27
	• Comments on use and stepwise regression		OHS 28
4	***Non-linear models***	5 min	
	• Effects of transformation		OHS 29
5	***Filling in missing data***	5 min	
	• Various uses of regression for infilling missing data		OHS 30
	• Type of regression model for filling-in missing flows		OHS 31
6	***Exercise***		
	• Create monthly and annual rainfall and runoff series for Bilodra	20 min	
	• Make runoff regression models for a monsoon month and annual series	20 min	
	• Generate runoff data and compare with observed series	20 min	

4. Overhead/flipchart master

5. Handout

Add copy of the main text in chapter 7, for all participants

6. Additional handout

These handouts are distributed during delivery and contain test questions, answers to questions, special worksheets, optional information, and other matters you would not like to be seen in the regular handouts.

It is a good practice to pre-punch these additional handouts, so the participants can easily insert them in the main handout folder.

7. Main text

Contents

How to do hydrological data validation using regression

1 Regression Analysis

1.1 Introduction

In regression analysis a relation is made between a dependent variable Y (i.e. the one one wants to estimate) and one or a number of independent variables X_i. The objective(s) of establishing a regression model may be manifold, like:

1. Making forecasts/predictions/estimates on Y based on data of the independent variable(s)
2. Investigation of a functional relationship between two or more variables
3. Filling in missing data in the Y-series
4. Validation of Y-series

In data processing at a number of occasions regression analysis is applied:

- for validation and in-filling of missing water level data a relation curve is established based on a polynomial relation between the observations at two water level gauging stations either or not with a time-shift
- for transformation of water levels into discharge series a discharge rating curve is created. The commonly used discharge rating curves are of a power type regression equation, where for each range of the independent variable (gauge reading) a set of parameters is established.
- for estimation of rainfall (or some other variable) on the grid points of a grid over the catchment as a weighted average of observations made at surrounding stations with the aid of kriging also falls into the category of regression.

For validation of rainfall data use is made of a linear relation between observations at a base station and surrounding stations. The weights given to the surrounding stations is inverse distance based. Because the weights are not determined by some estimation error minimization criterion as is the case in regression analysis but rather on the geographical location of the observation stations those relations are not regression equations.

In the above examples of applications of regression analysis linear as well as non-linear relations have been mentioned:

- a **linear** regression equation is an equation which is linear in its coefficients:

  $$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$$

  How the variables X_i behave does not matter and they may for example form an i^{th} order polynomial; hence the relation between Y and X may be non-linear.

- in a **non-linear** regression equation the coefficients also appear as a power, like e.g.:

  $$Y = \alpha X_1^{\beta 1} X_2^{\beta 2} .....X_i^{\beta i}$$

  By considering a logarithmic transformation on the equation an non-linear equation as shown above can be brought back to a linear one. Then, the error minimisation is carried out on the logarithm rather than on the original values. Note that far more complex non-linear regression models can be considered but this is outside the scope of hydrological data processing.

In this module at first attention will be given to linear regression equations. Dependent on the number of independent variables in the regression equation a further distinction is made between:

- **simple linear regression**, where the dependent variable is regressed on one independent variable, and
- **multiple and stepwise linear regression**: the dependent variable is regressed on more than one independent variable. The difference between multiple and stepwise regression is that in multiple linear regression all independent variables brought in the analysis will be included in the regression model, whereas in stepwise regression the regression equation is built up step by step 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 entered 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.

The type of regression equation that is most suitable to describe the relation depends naturally on the variables considered and with respect to hydrology on the physics of the processes driving the variables. Furthermore, it also depends on the range of the data one is interested in. A non-linear relation may well be described by a simple linear regression equation, within a particular range of the variables in regression, as applies for example to annual runoff regressed on annual rainfall. In Figure.1 the general nature of of such a relationship is shown.

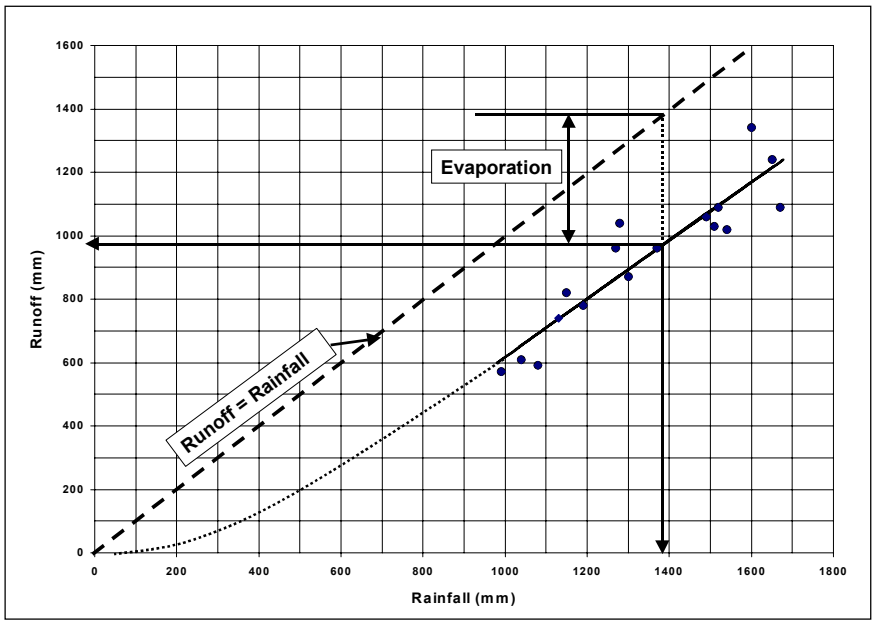

Figure 1: General form of relation between annual rainfall and runoff

For low rainfall amounts the relation is highly non-linear in view of the strong varying rainfall abstractions due to evaporation. For very high rainfalls the abstraction is constant as it has reached its potential level; then the rainfall-runoff relation runs parallel to a line under 45^{o} with an offset equal to the potential evaporation and becomes a true linear relation. In between reaches may approximately be described by a linear equation. As long as the application of the relation remains within the observed range then there is no harm in using a linear relation, provided that the residuals distribute randomly about the regression equation over the range considered.

Another application of regression, which has not been discussed previously, is for validation of discharge data. A regression model is developed where runoff is regressed on rainfall (when monthly data are considered on rainfall in the same and in the previous month). By investigating the time-wise behaviour of the deviations from the regression line (i.e. the residuals) an impression is obtained about the stationarity of the rainfall-runoff relation (note: not of the stationarity of either rainfall or runoff!). Provided that the rainfall data are free of observation errors any non-stationary behaviour of the residuals may then be explained by:

- change in the drainage characteristics of the basin, or
- incorrect runoff data, which in turn can be caused by:
  - errors in the water level data, and/or
  - errors the discharge rating curve

Experience has shown that by applying double mass analysis on the observed and computed runoff (derived from rainfall) a simple but effective tool is obtained to validate the discharge data. (Alternatively, instead of using a regression model, also a conceptual rainfall-runoff model can be used but at the expense of a far larger effort.) Hence, a very important aspect of judging your regression model is to look carefully at the behaviour of the residuals, not only about the regression line as a function of X but also as a function of time. An example has been worked out on this application.

2 Simple Linear Regression

The most common model used in hydrology is based on the assumption of a linear relationship between two variables. Such models are called simple linear regression models, which have the following general form:

$$\hat{Y} = \alpha + \beta X \qquad (1)$$

Where: $\hat{Y}$ = dependent variable, also called response variable (produced by the regression model)

X = independent variable or explanatory variable, also called input, regressor, or predictor variable

α, β = regression coefficients

The actual observations on Y do not perfectly match with the regression equation and a residual ε is observed, see also Figure 2:

$$Y = \alpha + \beta X + \varepsilon \qquad (2)$$

Hence:

$$Y_1 - \hat{Y}_i = \varepsilon_i \qquad (3)$$

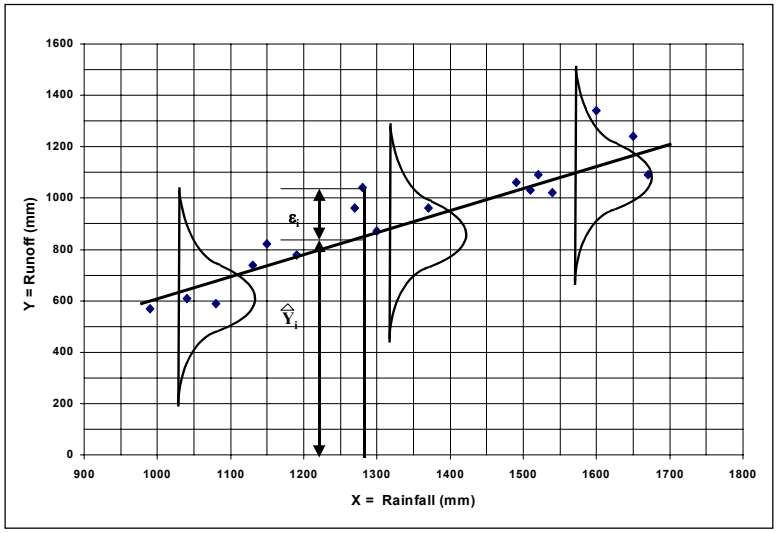

Figure 2:
Explained $_i$ and unexplained part ε_i of Y under the assumption of a constant error distribution

The regression line will be established such that $E\left[Y - \hat{Y}\right] = E[\varepsilon] = 0$, i.e. that it produces unbiased results and further that the variance of the residual σ_ε^2 is minimum. With respect to the residual it is assumed that its distribution about the regression line is normal and independent of X, hence for all values of X the distribution $F(\varepsilon)$ about the regression-line is the same, see Figure 2.

Now consider the following partitioning:

$$(Y - \overline{Y}) = (Y - \hat{Y}) + (\hat{Y} - \overline{Y}) = \varepsilon + (\hat{Y} - \overline{Y}) \text{ so}: (Y - \overline{Y})^2 = (\varepsilon + (\hat{Y} - \overline{Y}))^2 \text{ and since}: E[\varepsilon(\hat{Y} - \overline{Y})] = 0$$

$$E[(Y - \overline{Y})^2] = E[(\hat{Y} - \overline{Y})^2] + E[\varepsilon^2] \quad \text{or}:$$

$$\sigma_Y^2 = \sigma_{\hat{Y}}^2 + \sigma_\varepsilon^2 \tag{4}$$

Above equation expresses: **Total variance = explained variance + unexplained variance**

Hence, the smaller the unexplained variance (variance about regression) is, the larger the explained variance (or variance due to regression) will be. It also shows that the explained variance is always smaller than the total variance of the process being modelled. Hence the series generated by equation (1) will only provide a smoothed representation of the true process, having a variance which is smaller than the original, unless a random error with the characteristics of the distribution of the residual is added. Nevertheless, for individual generated values the estimate according to (1) is on average the best because $E[\varepsilon] = 0$. The root of the error variance is generally denoted as standard error.

In the following we will discuss:

- estimation of the regression coefficients
- measure for the goodness of fit
- confidence limits for the regression coefficients
- confidence limits for the regression equation
- confidence limits for the predicted values
- application of regression to rainfall-runoff analysis

Estimation of the regression coefficients

The estimators for the regression coefficients α and β , denoted by a and b respectively are determined by minimising $\Sigma\varepsilon^2$ Denoting the observations on X and Y by x_i and y_i this implies, that for:

$$M = \sum \varepsilon_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2 \tag{5}$$

to be minimum, the first derivatives of M with respect to a and b be set equal to zero:

$$\frac{\partial M}{\partial a} = -2\sum (y_i - a - bx_i) = 0 \tag{6a}$$

$$\frac{\partial M}{\partial b} = -2\sum x_i (y_i - a - bx_i) = 0 \tag{6b}$$

Above equations form the so called **normal equations**. From this it follows for a and b:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})} = \frac{S_{XY}}{S_{XX}} \quad \text{and:} \quad a = \bar{y} - b\bar{x} \tag{7}$$

Since the procedure is based on minimising $\Sigma\varepsilon^2$, the estimators a and b for α and β are commonly called **least squares estimators**. This solution also satisfies $\Sigma\varepsilon = 0$ as is observed from (6a)

With 7 the simple regression equation can also be written in the form:

$$\left(\hat{Y} - \overline{Y}\right) = b\left(X - \overline{X}\right) \tag{8}$$

or with the definition of the correlation coefficient $r = S_{XY}/\sigma_X.\sigma_Y$:

$$\hat{Y} - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X}) \tag{9}$$

Measure for goodness of fit

By squaring (9) and taking the expected value of the squares it is easily observed by combining the result with (4) that the error variance can be written as:

$$\sigma_\varepsilon^2 = \sigma_Y^2 (1 - r^2) \tag{10}$$

Hence, the closer r^2 is to 1 the smaller the error variance will be and the better the regression equation is in making predictions of Y given X. Therefore r^2 is an appropriate measure for the quality of the regression fit to the observations and is generally called the **coefficient of determination.**

It is stressed, though, that a high coefficient of determination **is not sufficient**. It is of great importance to investigate also the behaviour of the residual about the regression line and its development with time. If there is doubt about the randomness of the residual about regression then a possible explanation could be the existance of a non-linear relation. Possible reasons about absence of randomness with time have to do with changes in the relation with time as was indicated in the previous sub-chapter.

Confidence limits of the regression coefficients and model estimates

It can be shown, that, based on the sampling distributions of the regression parameters, the following estimates and confidence limits hold (see e.g. Kottegoda and Rosso, 1998).

Error variance

An unbiased estimate of the error variance is given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-(a+bx_i))^2 = \frac{1}{n-2}\left(S_{YY}-\frac{S_{XY}^2}{S_{XX}}\right) \tag{11}$$

Note that n-2 appears in the denominator to reflect the fact that two degrees of freedom have been lost in estimating (α, β)

Regression coefficients

A (100-α) percent confidence interval for b is found from the following confidence limits:

$$CL_\pm = b \pm t_{n-2,1-\alpha/2}\frac{\hat{\sigma}_\varepsilon}{\sqrt{S_{XX}}} \tag{12}$$

A (100-α) percent confidence interval for a results from the following confidence limits:

$$CL_\pm = a \pm t_{n-2,1-\alpha/2}\hat{\sigma}_\varepsilon\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}} \tag{13}$$

Regression line

A (100-α) percent confidence interval for the mean response to some input value x_0 of X is given by:

$$CL_\pm = a+bx_0 \pm t_{n-2,1-\alpha/2}\hat{\sigma}_\varepsilon\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}} \tag{14}$$

Note that the farther away x_0 is from its mean the wider the confidence interval will be because the last term under the root sign expands in that way.

Prediction

A (100-α) percent confidence interval for a predicted value Y when X is x_0 follows from:

$$CL_\pm = a+bx_0 \pm t_{n-2,1-\alpha/2}\hat{\sigma}_\varepsilon\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}} \tag{15}$$

It is observed by comparing (15) with (14) that in (15) full account of the error variance is added to last term. Hence, these confidence limits will be substantially wider that those for the mean regression line. Note however, since the multiplier of the standard error is under the root sign, the confidence limits in (15) are not simply obtained by adding t-times the standard error to the confidence limits of the regression line.

Example 1

In Table 1 some 17 years of annual rainfall and runoff data of a basin are presented. Regression analysis will be applied to validate the runoff series as there is some doubt about the rating curves applied before 1970. No changes took place in the drainage characteristics of the basin.

Year	Rainfall (mm)	Runoff (mm)	Year	Rainfall (mm)	Runoff (mm)	Year	Rainfall (mm)	Runoff (mm)
1961	1130	592	1967	1670	872	1973	1650	1240
1962	1280	832	1968	1540	816	1974	1510	1030
1963	1270	768	1969	990	456	1975	1600	1340
1964	1040	488	1970	1190	780	1976	1300	870
1965	1080	472	1971	1520	1090	1977	1490	1060
1966	1150	656	1972	1370	960			

Table 1: Rainfall and runoff data (in mm) for the period 1961 to 1977

A time series plot of the rainfall and runoff series is presented in Figure 3a. A simple linear regression equation is established for R = f(P), see Figure 3b. The regression equation reads:

R = -530 + 1.025xP, with σ_ε = 130 mm and the coefficient of determination r^2 = 0.75.

From Figure 3c it is observed that the trend line for the residuals runs exactly parallel to the axis of the independent variable (=rainfall) and is zero throughout meaning that the regression was properly performed mathematically. It appears, though, that the assumption of a constant error distribution is not fulfilled: the variation about regression clearly increases with increase in the independent variable. The time series plot of the residuals when subjected to a trend analysis shows a clear upward trend. This looks like a gradual change in the rainfall-runoff relation in the period of observation. However, as stated above, no changes took place in the drainage characteristics of the basin. The plot of accumulated residuals shown in Figure 3e features a distinct change in the residuals as from 1970 onward. A double mass analysis on the observed runoff against the runoff computed by regression on the rainfall also shows a distinct break around 1970, see Figure 3f. From this analysis it is revealed that the runoff data prior to 1970 have been underestimated by 20%. Accordingly, a correction was applied to the runoff.

The corrected time series is shown in Figure 4a. The results of the regression analysis on the corrected data are presented in the Figures 4b to 4e. The regression equation now reads:

R = -303 + 0.920xP, with σ_ε = 88.3 mm and the coefficient of determination r^2 = 0.84. It is observed that the coefficient of determination has increased substantially and consequently the standard error has decreased; its value is now over 30% less. The behaviour of the residual as a function of the dependent variable and as a function of time are shown in Figures 4c and d. Figure 4c shows that the variance of the residual is now fairly constant with X. From Figure 4d it is observed that no time effect is present anymore. In Figure 4e the 95% confidence limits about the regression line and of the predictions are shown. The computations are outlined in Table 2.

Year	X=Rainfall	Y=Runoff	$(X-Xm)^2$	Yest	CL1	CL2	UC1	LC1	UC2	LC2
1	2	3	4	5	6	7	8	9	10	11
1961	1130	740	44100	737	64	199	801	673	936	538
1962	1280	1040	3600	875	47	194	922	827	1069	681
1963	1270	960	4900	866	48	194	914	818	1060	671
1964	1040	610	90000	654	78	204	732	576	858	450
1965	1080	590	67600	691	72	201	762	619	892	489
1966	1150	820	36100	755	61	198	816	694	953	557
1967	1670	1090	108900	1234	84	206	1317	1150	1440	1028
1968	1540	1020	40000	1114	62	198	1176	1052	1312	916
1969	990	570	122500	608	87	207	695	521	815	400
1970	1190	780	22500	792	56	196	848	736	988	596
1971	1520	1090	32400	1096	60	198	1155	1036	1293	898
1972	1370	960	900	958	46	194	1004	911	1152	764
1973	1650	1240	96100	1215	80	205	1295	1135	1420	1011
1974	1510	1030	28900	1086	58	197	1145	1028	1284	889
1975	1600	1340	67600	1169	72	201	1241	1098	1371	968
1976	1300	870	1600	893	46	194	940	847	1087	699
1977	1490	1060	22500	1068	56	196	1124	1012	1264	872
Xm	1340	SXX	790200							

Table 2: Example computation of confidence limits for regression analysis

In the computations use is made of equations (14) and (15). In Column 2 $\bar{t}$he mean of X is computed and the sum of Column 4 is S_{XX}. In the Columns 6 and 7 the last term of equations (14) and (15) are presented. Note that $t_{n-2,1-\alpha/2}$ = 2 131 and σ_ε = 88.3 mm. Column 6 and 7 follow from:

$$CL1 = t_{n-1,1-\alpha/2}\sigma_\varepsilon\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{XX}}} = 2.131x88.3\sqrt{\frac{1}{17}+\frac{(1130-1340)^2}{790200}} = 2.131x88x0.34 = 64\,mm$$

$$CL2 = t_{n-1,1-\alpha/2}\sigma_\varepsilon\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})}{S_{XX}}} = 2.131x88.3\sqrt{1+\frac{1}{17}+\frac{(1130-1340)^2}{790200}} = 2.131x88.3x1.06 = 199\,mm$$

The upper and lower confidence limits of the mean regression line then simply follow from Column 5 + 6 and Column 5 – 6, whereas the confidence limits for the predicted value (Columns 10 and 11) are derived from Column 5 + 7 and Column 5 – 7. It may be observed that the width of the confidence interval is minimum at the mean value of the independent variable. The variation of the width with the independent variable is relatively strongest for the confidence limits of the mean relation. The confidence limits for the prediction are seen to vary little with the variation in the independent variable, since the varying part under the root (i.e. the last term) is seen to be small compared to 1.

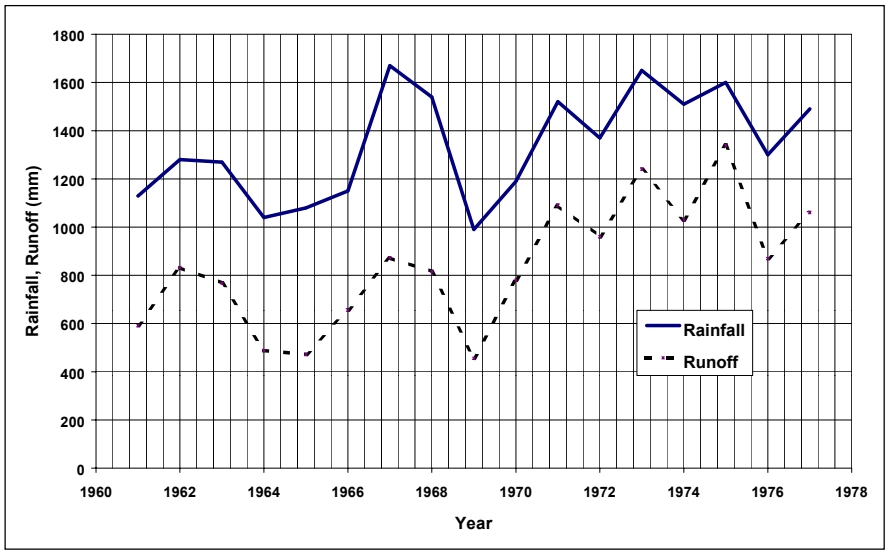

Figure 3a: Rainfall-runoff record 1961-1977

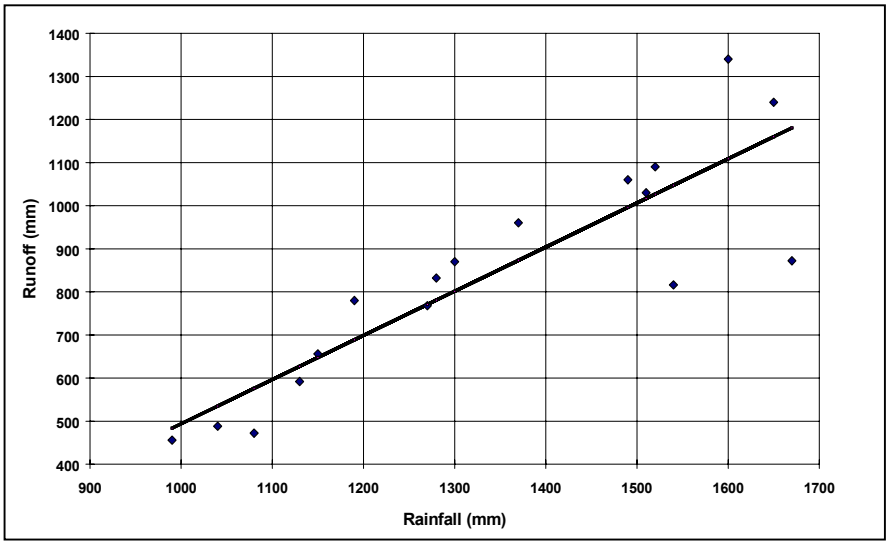

Figure 3b: Regression fit Rainfall-runoff

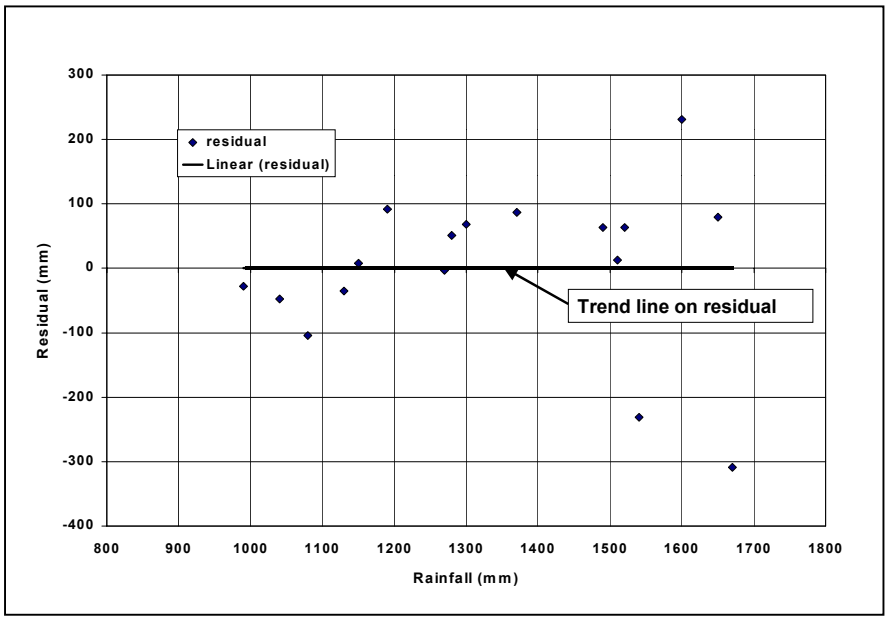

Figure 3c: Plot of residual versus rainfall

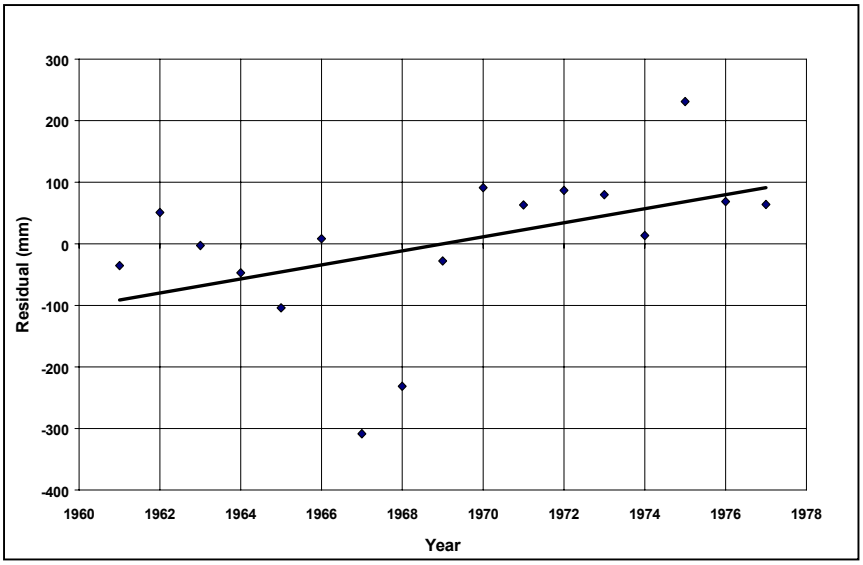

Figure 3d:
Plot of residual versus time

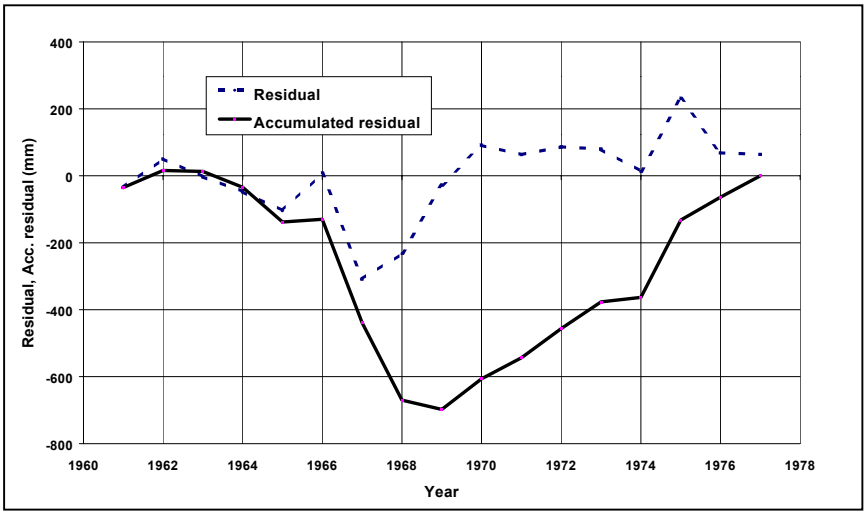

Figure 3e:
Plot of accumulated residual

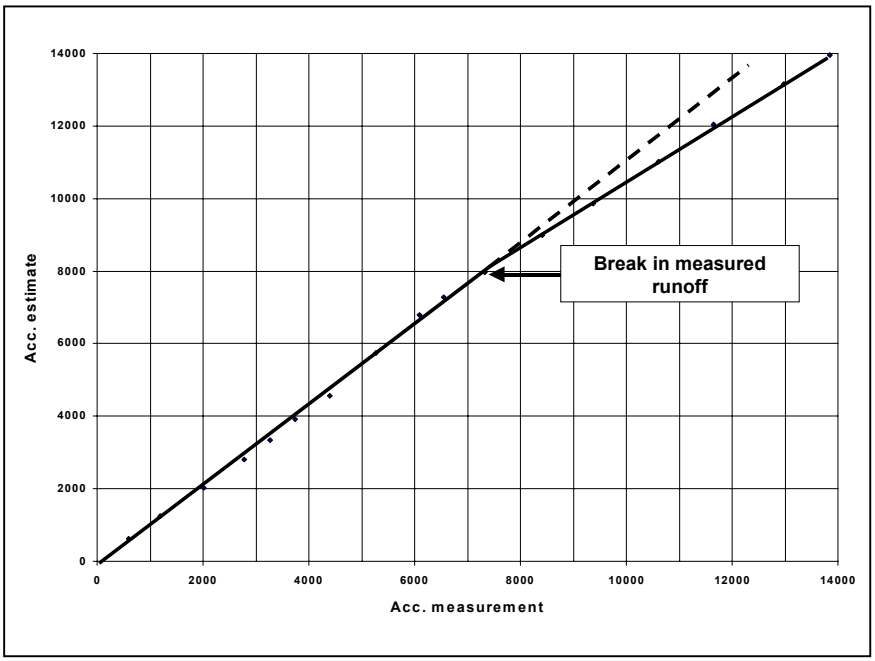

Figure 3f:
Double mass analysis observed versus computed runoff

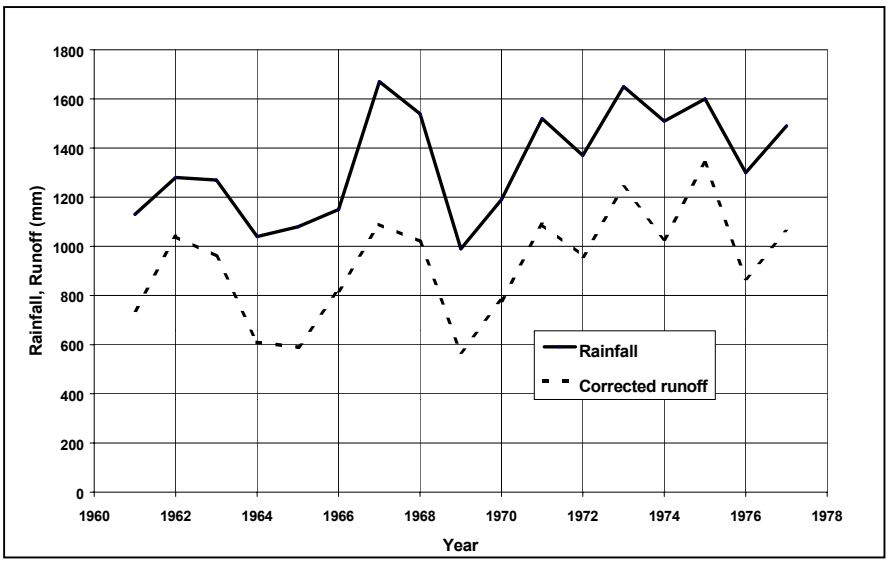

Figure 4a:
Plot of rainfall and corrected runoff

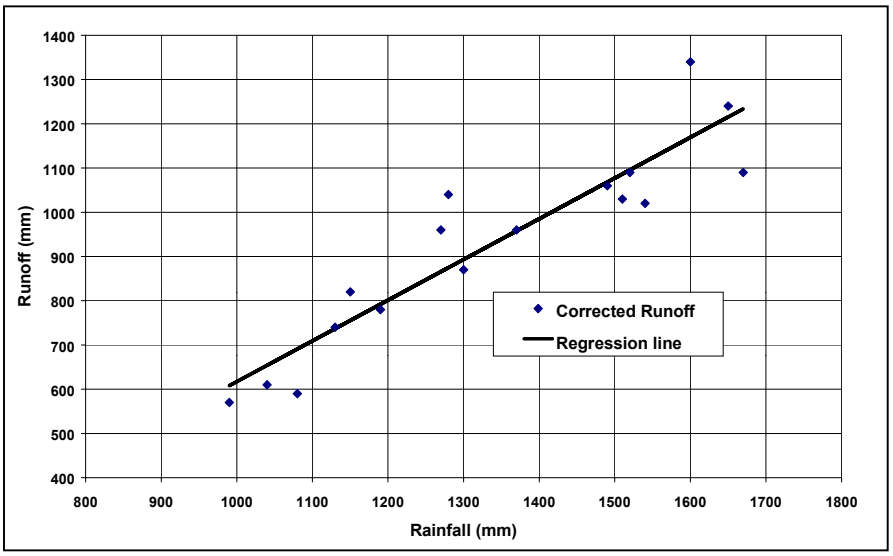

Figure 4b:
Plot of rainfall runoff regression, corrected runoff data

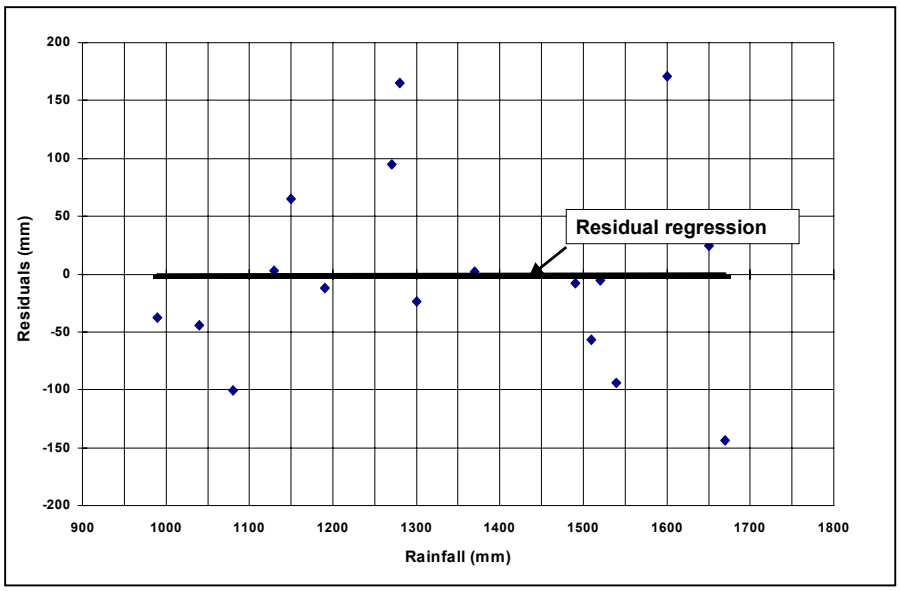

Figure 4c:
Plot of residual versus rainfall

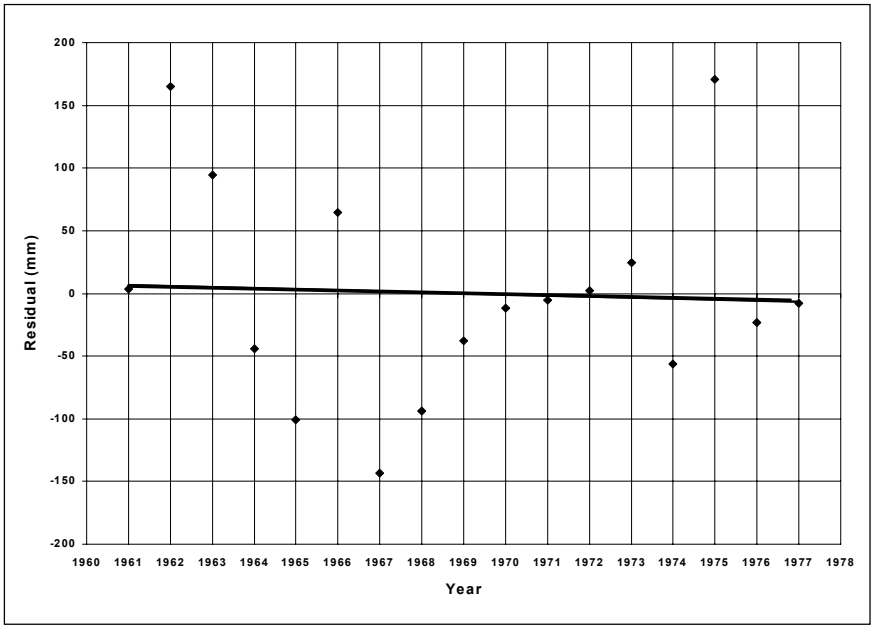

Figure 4d:
Plot of residual versus time

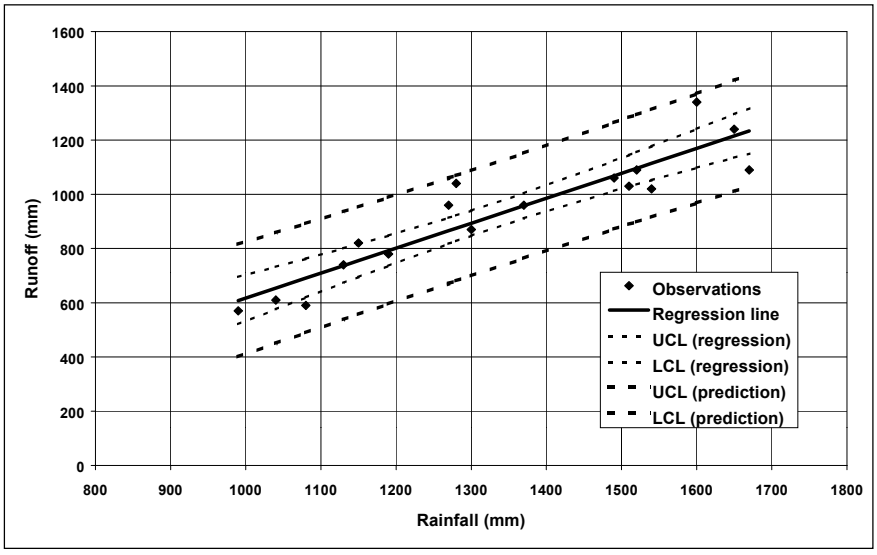

Figure 4e:
Regression line with confidence limits for the mean regression and predicted values

Extrapolation

The extrapolation of a regression equation beyond the range of X used in estimating α and β is discouraged for two reasons. First as can be seen from Figure 4e the confidence intervals on the regression line become wide as the distance from $\overline{X}$ is increased. Second the relation between Y and X may be non-linear over the entire range of X and only approximately linear for the range of X investigated. A typical example of this is shown in Figure 1.

Multiple Linear Regression

Often we wish to model the dependent variable as a function of several other quantities in the same equation. In extension to the example presented in the previous sub-chapter monthly runoff is likely to be dependent on the rainfall in the same month and in the previous month(s) Then the regression equation would read:

$$R(t) = \alpha + \beta_1 P(t) + \beta_2 P(t-1) + \ldots. \quad (16)$$

In this section the linear model is extended to include several independent variables.

A general linear model of the form:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots. \beta_p X_p + \varepsilon \quad (17)$$

is discussed, where Y is a dependent variable, X_1, X_2, …. X_p are independent variables and β_1, β_2, ….. β_p are unknown parameters. This model is linear in the parameters β_j .Note that the form (16) can always be brought to the form (17) with the constant α by considering the variables Y and X_i centered around their mean values, similar to (8).

In practice n observations would be available on Y with the corresponding n observations on each of the p independent variables. Thus n equations can be written, one for each observation. Essentially we will be solving n equations for the p unknown parameters. Thus n must be equal to or greater than p. In practice n should be at least 3 or 4 times as large as p. The n equations then read

$$\mathbf{Y} = \mathbf{X}\,\beta + \varepsilon \quad (18)$$

where $\mathbf{Y}$ = (n x 1)-data column vector of the centred dependent variable $(y_i - y)$
$\mathbf{X}$ = (n x p)-data matrix of the centred independent variables $(x_{i1}-x_1),\ldots,(x_{ip}-x_p)$
β = (p x 1)- column vector, containing the regression coefficients
ε = (n x 1)-column vector of residuals

The residuals are conditioned by:

$$E[\mathbf{e}] = 0 \quad (19)$$

$$Cov(\mathbf{e}) = \sigma_\varepsilon^2\,\mathbf{I} \quad (20)$$

Where: $\mathbf{I}$ = (n x n) diagonal matrix with diagonal elements = 1 and off-diagonal elements = 0
σ_ε^2 = variance of (Y|X)

According to the least squares principle the estimates $\mathbf{b}$ of β are those which minimise the residual sum of squares $\varepsilon^T\varepsilon$. Hence:

$$\varepsilon^T\varepsilon = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \quad (21)$$

is differentiated with respect to $\mathbf{b}$, and the resulting expression is set equal to zero. This gives:

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y} \quad (22)$$

called the **normal equations**, where β is replaced by its estimator $\mathbf{b}$. Multiplying both sides with $(\mathbf{X}^T\mathbf{X})^{-1}$ leads to an explicit expression for $\mathbf{b}$:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{Y} \quad (23)$$

The properties of the estimator $\mathbf{b}$ of β are:

$$E[\mathbf{b}] = \beta \quad (24)$$

$$Cov(\mathbf{b}) = \sigma_\varepsilon^2(\mathbf{X}^T\mathbf{X})^{-1} \quad (25)$$

By (21) and (22) the total adjusted sum of squares $\mathbf{Y}^T\mathbf{Y}$ can be partitioned into an explained part due to regression and an unexplained part about regression, as follows:

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{b}^T\mathbf{X}^T\mathbf{Y} + \mathbf{e}^T\mathbf{e}. \qquad (26)$$

Where: $(\mathbf{Xb})^T\mathbf{Y}$ = sum of squares due to regression

$\mathbf{e}^T\mathbf{e}$. = sum of squares about regression, with ε replaced by **e** due to the replacement of β by **b.**

In words this reads:

Total sum of squares about the mean = regression sum of squares + residual sum of squares

The mean squares values of the right hand side terms in (26) are obtained by dividing the sum of squares by their corresponding degrees of freedom. If **b** is a (p x 1)-column vector, i.e. there are p-independent variables in regression, then the regression sum of squares has p-degrees of freedom. Since the total sum of squares has (n-1)-degrees of freedom (note: 1 degree of freedom is lost due to the estimation of $\bar{y}$), it follows by subtraction that the residual sum of squares has (n-1-p)-degrees of freedom. It can be shown that the residual mean square s_e^2:

$$s_e^2 = \frac{e^T e}{n-1-p}$$

Is an unbiased estimate of σ_ε^2 .The estimate s_e of σ_ε is the **standard error of estimate**.

The analysis of variance table (ANOVA) summarises the sum of squares quantities

Source	**Sum of squares**	**Degrees of freedom**	**Mean squares**
Regression (b_1, …., b_p)	$S_R = \mathbf{b^TX^TY}$	p	$MS_R = \mathbf{b^TX^TY/p}$
Residual (e_1, …., e_n)	$S_e = \mathbf{e^Te = Y^TY - b^TX^TY}$	n-1-p	$MS_e = s_e^2 = \mathbf{e^Te}/(n-1-p)$
Total (adjusted for $\bar{y}$)	$S_Y = \mathbf{Y^TY}$	n-1	$M_{SY} = s_Y^2 = \mathbf{Y^TY}/(n-1)$

Table 3: Analysis of variance table (ANOVA)

As for the simple linear regression a measure for the quality of the regression equation is the **coefficient of determination**, defined as the ratio of the explained or regression sum of squares and the total adjusted sum of squares.

$$R_m^2 = \frac{b^T X^T Y}{Y^T Y} = 1 - \frac{e^T e}{Y^T Y}$$

The coefficient should be adjusted for the number of independent variables in regression. Then, instead of the sum of squares ratio in the most right-hand side term the mean square ratio is used. So with the adjustment:

$$R_{ma}^2 = 1 - \frac{MS_e}{MS_Y} = 1 - \frac{e^T e}{Y^T Y}.\frac{(n-1)}{(n-p-1)} = 1 - (1-R_m^2)\left(\frac{n-1}{n-p-1}\right) \qquad (29)$$

From this it is observed that $R_{ma}^2 < R_m^2$ except for $R_m = 1$ (i.e. a perfect model) where R_m is the multiple correlation coefficient and R_{ma} the adjusted multiple correlation coefficient.

Reference is made to the annex to the HYMOS manual for statistical inference on the regression coefficients.

Confidence Intervals on the Regression Line

To place confidence limits on Y_0 where $Y_0 = \mathbf{X}_0\mathbf{b}$ it is necessary to have an estimate for the variance of $\hat{Y}_0$. Considering Cov(**b**) as given in (25) the variance $Var(\hat{Y}_0)$ is given by (Draper and Smith 1966):

$$Var(\hat{Y}_0) = s_e^2 \mathbf{X}_0(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}_0^T \qquad (30)$$

The confidence limits for the mean regression equation are given by

$$CL_{\pm} = X_0 b \pm t_{1-a/2,n-p}\sqrt{Var(\hat{Y}_0)} \qquad (31)$$

Comments

A common situation in which multiple regression is used is when one dependent variable and several independent variables are available and it is desired to find a linear model that is developed does not necessarily have to contain all of the independent variables. Thus the points of concern are: (1) can a linear model be used and (2) what independent variable should be included?

A factor complicating the selection of the model is that in most cases the independent variables are not statistically independent at all but are correlated. One of the first steps that should be done in a regression analysis is to compute the correlation matrix.

Retaining variables in a regression equation that are highly correlated makes the interpretation of the regression coefficients difficult. Many times the sign of the regression coefficient may be the opposite of what is expected if the corresponding variable is highly correlated with another independent variable in the equation.

A common practice in selecting a multiple regression model is to perform several regressions on a given set of data using different combinations of the independent variables. The regression that “best” fits the data is then selected. A commonly used criterion for the ”best” fit is to select the equation yielding the largest value of R_{ma}^2.

All of the variables retained in a regression should make a significant contribution to the regression unless there is an overriding reason (theoretical or intuitive) for retaining a non-significant variable. The variables retained should have physical significance. If two variables are equally significant when used alone, but are not both needed, the one that is easiest to obtain should be used.

The number of coefficients estimated should not exceed 25 to 35 percent of the number of observations. This is a rule of thumb used to avoid “over-fitting” whereby oscillations in the equation may occur between observations on the independent variables.

Stepwise Regression

One of the most commonly used procedures for selecting the “best” regression equations is **stepwise regression**. This procedure consists of building the regression equation one variable at a time by adding at each step the variable that explains the largest amount of the

remaining unexplained variation. After each step all the variables in the equation are examined for significance and discarded if they are no longer explaining a significant variation. Thus the first variable added is the one with the highest simple correlation with the dependent variable. The second variable added is the one explaining the largest variation in the dependent variable that remains unexplained by the first variable added. At this point 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 variables already in the equation. The variables in the equation are then tested for significance. This procedure is continued until all of the variables not in the equation are found to be insignificant and all of the variables in the equation are significant. This is a very good procedure to use but care must be exercised to see that the resulting equation is rational.

The real test of how good is the resulting regression model, depends on the ability of the model to predict the dependent variable for observations on the independent variables that were not used in estimating the regression coefficients. To make a comparison of this nature, it is necessary to randomly divide the data into two parts. One part of the data is then used to develop the model and the other part to test the model. Unfortunately, many times in hydrologic applications, there are not enough observations to carry out this procedures.

Transforming Non Linear Models

Many models are not naturally linear models but can be transformed to linear models. For example

$$Y = \alpha X^{\beta} \tag{32}$$

is not a linear model. It can be linearized by using a logarithmatic transformation:

$$\ln Y = \ln \alpha + \beta \ln X \tag{33}$$

or

$$Y_T = \alpha_T + \beta_T X_T \tag{34}$$

where

$Y_T = \ln Y$
$\alpha_T = \ln \alpha$
$\beta_T = \beta$
$X_T = \ln X$

Standard regression techniques can now be used to estimate α_T and β_T for the transformed equation and α and β estimated from the logarithmatic transformation. Two important points should be noted.:

Firstly, the estimates of α and β obtained in this way will be such $\sum(Y_{T_i} - \hat{Y}_{T_i})^2$ that is a min. and not such that $\sum(Y_i - \hat{Y}_i)^2$ is a minimum.

Secondly, the error term on the transformed equation is additive ($Y_T = \alpha_T + \beta_T X_T + \varepsilon_T$) implying that it is multiplicative on the original equation i.e. $Y = \alpha X^{\beta}\varepsilon$. These errors are related $\varepsilon_T = \ln \varepsilon$. The assumptions used in hypothesis testing and confidence intervals must now be valid for ε` and the tests and confidence intervals made relative to the transformed model.

In some situations the logarithmic transformation makes the data conform more closely to the regression assumptions. The normal equations for a logarithmic transformation are based on a constant percentage error along the regression line while the standard regression is based on a constant absolute error along the regression line

Filling in missing data

An important application of regression analysis is the use of a regression equation to fill in missing data. In Part II of Volume 8 attention has been given to fill in missing rainfall and water level data. In this section attention will be given to filling in missing runoff data using rainfall as input. Typically, such techniques are applied to time series with time intervals of a decade, a month or larger.

Generally a regression of the type presented in equation (16) is applicable. Assume that the objective is to fill in monthly data. The regression coefficients are likely to be different for each month, hence the discharge in month k of year m is computed from:

$$Q_{k,m} = a_k + b_{1k}P_{k,m} + b_{2k}P_{k-1,m} + s_{e,k}e \qquad (35)$$

It is observed that the regression coefficients are to be determined for each month in the year. The last term is added to ensure that the variance of the discharge series is being preserved. It represents the unexplained part of the functional relationship. Omitting the random component will result in a series with a smaller variance, which creates an inhomogeneity in the series and certainly does affect the overall monthly statistics. Dependent on the application it has to be decided whether or not to include the random components. If, however, a single value is to be estimated the random component should be omitted as the best guess is to rely on the explained part of equation (35); E[e] = 0. Note that the calibration of such a model will require at least some 15 to 20 years of data, which might be cumbersome occasionally.

Experience has shown that for a number of climatic zones the regression coefficients do not vary much from month to month, but rather vary with the wetness of the month. Two sets of parameters are then applied, one set for wet conditions and one for dry conditions with a rainfall threshold to discriminate between the two parameter sets The advantage of such a model is that less years with concurrent data have to be available to calibrate it, with results only slightly less than with (2.35) can be achieved. The use of a threshold is also justifiable from a physical point of view as the abstractions from rainfall basically create a non-linearity in the rainfall-runoff relationship.

Concurrent rainfall and runoff data should be plotted to investigate the type of relationship applies. One should never blindly apply a particular model.