

CHAPTER 9

REGRESSION ANALYSIS AND APPLICATION TO REGIONAL STUDIES

9-1. Nature and Application.

a. General. Regression analysis is the term applied to the analytical procedure for deriving prediction equations for a variable (dependent) based on given values of one or more other variables (independent). The dependent variable is the value sought and is to be related to various explanatory variables which will be known in advance, and which will be physically related to the dependent variable. For example, the volume of spring-season runoff from a river basin (dependent variable) might be correlated with the depth of snow cover in the watershed (explanatory variable). Recorded values of such variables over a period of years might be graphed and the apparent relation sketched in by eye. However, regression analysis will generally permit a more reliable determination of the relation and has the additional advantage of providing a means for evaluating the reliability of the relation or of estimates based on the relation.

b. Definitions. The function relating the variables is termed the "regression equation," and the proportion of the variance of the dependent variable that is explained by the regression equation is termed the "coefficient of determination," which is the square of the "correlation coefficient." Correlation is a measure of the association between two or more variables. Regression equations can be linear or curvilinear, but linear regression suffices for most applications, and curvilinear regression is therefore not discussed herein. Often a curvilinear relation can be linearized by using a logarithmic or other transform of one or more of the variables.

9-2. Calculation of Regression Equations.

a. Simple Regression. In a simple regression (one in which there is only one independent, or explanatory, variable), the linear regression equation is written:

$$Y = a + bX \quad (9-1)$$

in which Y is the dependent variable, X is the independent variable, "a" is the regression constant, and "b" is the regression coefficient. The coefficient "b" is evaluated from the tabulated data by use of the following equations:

$$b = \frac{\sum(yx)}{\sum(x)^2} \quad (9-2a)$$

or

$$b = RS_y/S_x \quad (9-2b)$$

in which y is the deviation of a single value y_i from the mean (\bar{Y}) of its series, x is similarly defined, S_y and S_x are the respective standard deviations and R is computed by

Equation 9-11. The regression constant is obtained from the tabulated data by use of the following equation:

$$a = \bar{Y} - b\bar{X} \quad (9-3)$$

All summations required for a simple linear regression can be obtained using Equations 9-8 and 9-9a.

b. Multiple Regression. In a multiple regression (one in which there is more than one explanatory variable) the linear regression equation is written:

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_NX_N \quad (9-4)$$

In the case of two explanatory variables, the regression coefficients are evaluated from the tabulated data by solution of the following simultaneous equations:

$$\begin{aligned} \sum(x_1)^2b_1 + \sum(x_1x_2)b_2 &= \sum(yx_1) \\ \sum(x_1x_2)b_1 + \sum(x_2)^2b_2 &= \sum(yx_2) \end{aligned} \quad (9-5)$$

In the case of three explanatory variables, the b coefficients can be evaluated from the tabulated data by solution of the following simultaneous equations:

$$\begin{aligned} \sum(x_1)^2b_1 + \sum(x_1x_2)b_2 + \sum(x_1x_3)b_3 &= \sum(yx_1) \\ \sum(x_1x_2)b_1 + \sum(x_2)^2b_2 + \sum(x_2x_3)b_3 &= \sum(yx_2) \\ \sum(x_1x_3)b_1 + \sum(x_2x_3)b_2 + \sum(x_3)^2b_3 &= \sum(yx_3) \end{aligned} \quad (9-6)$$

For cases of more than three explanatory variables, the appropriate set of simultaneous equations can be easily constructed after studying the patterns of the above two sets of equations. In such cases, solution of the equations becomes tedious, and considerable time can be saved by use of the Crout method outlined in reference (51) or (52). Also, programs are available for solution of simple or multiple linear regression problems on practically any type of electronic computer. For multiple regression equations, the regression constant is determined as follows:

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - \dots - b_N\bar{X}_N \quad (9-7)$$

In Equations 9-2, 9-5 and 9-6, the quantities $\sum(x)^2$, $\sum(yx)$ and $\sum(x_1x_2)$ can be determined by use of the following equations:

$$\sum(x)^2 = \sum(X)^2 - (\sum X)^2/N \quad (9-8)$$

$$\sum(yx) = \sum(XY) - \sum X \sum Y/N \quad (9-9a)$$

$$\sum(x_1x_2) = \sum(X_1X_2) - \sum X_1 \sum X_2/N \quad (9-9b)$$

9-3. The Correlation Coefficient and Standard Error.

a. General. The correlation coefficient is the square root of the coefficient of determination, which is the proportion of the variance of the dependent variable that is explained by the regression equation. A correlation coefficient of 1.0 would correspond to a coefficient of determination of 1.0, which is the highest theoretically possible and indicates that whenever the values of the explanatory variables are known exactly, the corresponding value of the dependent variable can be calculated exactly. A correlation coefficient of 0.5 would correspond to a coefficient of determination of 0.25, which would indicate that 25 percent of the variance is accounted for and 75 percent unaccounted for by the regression equation. The remaining variance (error variance) would be 75 percent of the original variance and the remaining standard error would be the square root of 0.75 (or 87 percent) multiplied by the original standard deviation of the dependent variable. Thus, with a correlation coefficient of 0.5, the average error of estimate would be 87 percent of the average errors of estimate based simply on the mean observed value of the dependent variable without a regression analysis.

b. Determination Coefficient. The sample coefficient of multiple determination (R^2) can be computed by use of the following equation:

$$R^2 = \frac{b_1 \sum(yx_1) + b_2 \sum(yx_2) \dots + b_N \sum(yx_n)}{\sum(y)^2} \quad (9-10)$$

In the case of simple correlation, Equation 9-10 resolves to:

$$R^2 = \sum(yx)^2 / \sum(y)^2 \sum(x)^2 \quad (9-11)$$

An unbiased estimate of the coefficient of determination is recommended for most applications, and is computed by the following equation:

$$\bar{R}^2 = 1 - (1 - R^2)(N - 1) / df \quad (9-12)$$

The number of degrees of freedom (df), is obtained by subtracting the number of variables (dependent and explanatory) from the number of events tabulated for each variable.

c. Standard Error. The adjusted standard error (S_e) of a set of estimates is the root-mean-square error of those estimates corrected for the degrees of freedom. On the

average, about one out of three estimates will have errors greater than the standard error and about one out of 20 will have errors greater than twice the standard error. The adjusted error variance is the square of the adjusted standard error. The adjusted standard error or error variance of estimates based on a regression equation is calculated from the data used to derive the equation by use of one of the following equations:

$$S_e^2 = \frac{\sum(y)^2 - b_1 \sum(yx_1) - b_2 \sum(yx_2) \dots - b_n \sum(yx_n)}{df} \quad (9-13a)$$

$$= (1 - \bar{R}^2) \sum(y)^2 / (N-1) \quad (9-13b)$$

$$= (1 - \bar{R}^2) S_y^2 \quad (9-13c)$$

Inasmuch as there is some degree of error involved in estimating the regression coefficients, the actual standard error of an estimate based on one or more extreme values of the explanatory variables is somewhat larger than is indicated by the above equations, but this fact is usually neglected.

d. Reliability. In addition to considering the amount of variance that is explained by the regression equation, as indicated by the determination coefficient or the standard error, it is important to consider the reliability of these indications. There is some chance that any correlation is accidental, but the higher the correlation and the larger the sample upon which it is based, the less is the chance that it would occur by accident. Also, the reliability of a regression equation decreases as the number of independent variables increases. Ezekiel (8) gives a set of charts illustrating the reliability of correlation coefficients. It shows, for example, that an unadjusted correlation coefficient (R) of 0.8 based on a simple linear correlation with 12 degrees of freedom could come from a relationship that has a true value as low as 0.53 in one case out of 20. On the other hand, the same unadjusted correlation coefficient based on a multiple linear correlation with the same number of degrees of freedom but with seven independent variables, could come from a relationship that has a true value as low as zero in one case out of 20. With only 4 degrees of freedom, an unadjusted correlation coefficient of 0.97 would one time in 20 correspond to a true value of 0.8 or lower, in the case of simple correlation, and as low as zero in a seven-variable multiple correlation. Accordingly, extreme care must be exercised in the use of multiple correlation in cases based on small samples.

9-4. Simple Linear Regression Example.

a. General. An example of a simple linear regression analysis is illustrated on Figures 9-1 and 9-2. The data for this example are the concurrent flows at two stations in Georgia for which a two-station comparison is desired (see Section 3-7). The long record station is the Chattooga, so the flows for this station are selected as X; therefore the flows for the short record station (Tallulah) are assigned to Y.

b. Physical Relationship. The values in the table are the annual peak flows for the water years 1965-1985 (21 values). These two stations are less than 20 miles apart and are likely to be subject to the same storm events; therefore, the first requirement of a

Year	Chattooga River		Tallulah River	
	Flow X'	Log X	Flow Y'	Log Y
1965	27200	4.434568	7440	3.871572
1966	13400	4.127104	5140	3.710963
1967	15400	4.187520	2800	3.447158
1968	5620	3.749736	3100	3.491361
1969	14700	4.167317	2470	3.392696
1970	3480	3.541579	2010	3.303196
1971	3290	3.517195	976	2.989449
1972	7440	3.871572	2160	3.334453
1973	19600	4.292256	8500	3.929418
1974	6400	3.806179	4660	3.668385
1975	6340	3.802089	2410	3.382017
1976	18500	4.267171	6530	3.814913
1977	13000	4.113943	3580	3.553883
1978	7850	3.894869	4090	3.611723
1979	14800	4.170261	6240	3.795184
1980	10900	4.037426	2880	3.459392
1981	4120	3.614897	1600	3.204119
1982	5000	3.698970	1960	3.292256
1983	7910	3.898176	3260	3.513217
1984	4810	3.682145	2000	3.301029
1985	4740	3.675778	1010	3.004321

$$\begin{aligned} \Sigma X &= 82.55075 & \Sigma Y &= 73.07071 \\ \Sigma X^2 &= 325.93995 & \Sigma Y^2 &= 255.61486 \\ \bar{X} &= 3.93099 & \bar{Y} &= 3.47956 \\ (\Sigma XY)^2 &= 288.37484 \\ \frac{(\Sigma X)^2}{N} &= 324.50602 \\ \frac{(\Sigma Y)^2}{N} &= 254.25371 \\ \frac{\Sigma X \Sigma Y}{N} &= 287.24008 \\ x^2 &= 1.43393 & & \text{(by equation 9-8)} \\ xy &= 1.13351 & & \text{(" " 9-9a)} \\ y^2 &= 1.26115 & & \text{(" " 9-8)} \end{aligned}$$

Computations for a, b, R², and R̄²:

$$b = 1.13351/1.433922 \quad \text{(by equation 9-2a)} \\ = 0.79049$$

$$a = 3.47956 - (0.79049)(3.93099) \quad \text{(by equation 9-3)} \\ = 0.37213$$

$$R^2 = (1.13351)^2 / (1.43393)(1.36115) \quad \text{(by equation 9-11)} \\ = 0.658290$$

$$\bar{R}^2 = 1 - (1 - 0.65892)(21 - 1) / (21 - 2) \quad \text{(by equation 9-12)} \\ = 0.64031$$

Computations for standard error:

$$s_e^2 = (1 - 0.64031^2)(1.36115) / (21 - 1) \quad \text{(by equation 9-13b)} \\ = 0.02448$$

$$s_e = 0.15646$$

Regression equation: $Y = 0.37213 + 0.79049X$ (by equation 9-1)
 $Y' = 2.356X^{0.79}$ (without logarithms)

Figure 9-1. Computation of Simple Linear Regression Coefficients.

regression analysis (logical physical relationship) is satisfied. Because runoff is a multiplicative factor of precipitation and drainage area, the logarithmic transformation is likely to be appropriate when comparing two stations with different drainage areas. A linear correlation analysis was made, as illustrated on Figure 9-1, using equations given in Section 9-2. The annual peaks for the each station are plotted against each other on Figure 9-2.

c. Regression Equation. The regression equation is plotted as Curve A on Figure 9-2. This curve represents the best estimate of what the annual peak Tallulah River would be given the observed annual peak on the Chattooga River. Although not computed in Figure 9-1, Curve B represents the regression line for estimating the annual peak flow for the Chattooga River given an observed annual peak on the Tallulah River.

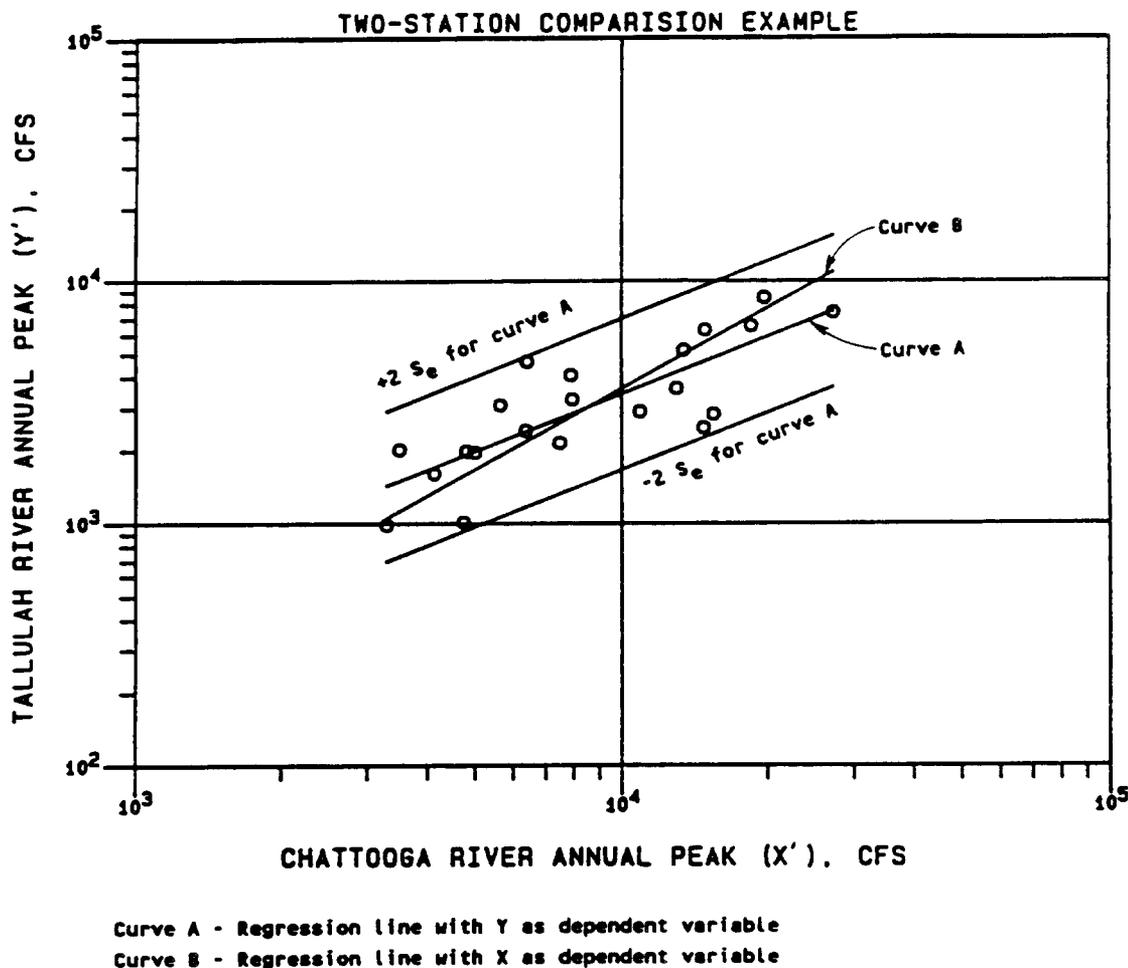


Figure 9-2. Illustration of Simple Regression.

d. Reliability. In addition to the curve of best fit, an approximate confidence interval can be established at a distance of plus and minus 2 standard errors from Curve A. Because logarithms are used in the regression analysis, the effect of adding (or subtracting) twice the standard error to the estimate is equivalent to multiplying (or dividing) the annual peak values by the antilogarithm of twice the standard error. In this case, the standard error is 0.156, and the antilogarithm of twice this quantity is 2.05. Hence, values of annual peak flow represented by the confidence interval curves are those of Curve A multiplied and divided respectively by 2.05. There is a 95 percent chance that the true value of the dependent variable (Y) for a single observed independent value (X) will lie between these limits. The confidence interval is not correct for repeated predictions using the same sample (6).

9-5. Factors Responsible for Nondetermination.

a. General. Factors responsible for correlations being less than 1.0 (perfect correlation) consist of pertinent factors not considered in the analysis and of errors in the measurement of those factors considered. If the effect of measurement errors is appreciable, it is possible in some cases to evaluate the standard error of measurement of each variable (see Paragraph 9-3c) and to adjust the correlation results from such effects.

b. Measurement Errors. If an appreciable portion of the variance of Y (dependent variable) is attributable to measurement errors and these errors are random, then the regression equation would be more reliable than is indicated by the standard error of estimate computed from Equation 9-13. This is because the departure of some of the points from the regression line on Figure 9-2 is artificially increased by measurement errors and therefore exaggerates the unreliability of the regression function. In such a case, the curve is generally closer to the true values than to the erroneous observed values. Where there is large measurement error of the dependent variable, the standard error of estimate should be obtained by taking the square root of the difference of the error variance obtained from Equation 9-13 and the measurement error variance. If well over half of the variance of the points from the best-fit line is attributable to measurement error in the dependent variable, then the regression line would actually yield a better estimate of a value than the original measurement. If appreciable errors exist in the values of an explanatory variable, the regression coefficient and constant will be affected, and erroneous estimates will result. Hence, it is important that values of the explanatory variables be accurately determined, if possible.

c. Other Factors. In the example used in Section 9-4 there may well be factors responsible for brief periods of high intensities that do not contribute appreciably to annual precipitation. Consequently, some locations with extremely high mean annual precipitation may have maximum short-time intensities that are not correspondingly high, and vice versa. Therefore, the station having the highest mean annual precipitation would not automatically have the highest short-time intensity, but would in general have something less than this. On the other hand, if mean annual precipitation were made the dependent variable, the station having the highest short-time intensity would be expected to have something less than the highest value of mean annual precipitation. Thus, by interchanging the variables, a change in the regression line is effected. Curve B of Figure 9-2 is the regression curve obtained by interchanging the variables Y and X. As there is a considerable difference in the two regression curves, it is important to use the variable whose value is to be calculated from the regression equation as the dependent variable in those cases where important factors have not been considered in the analysis.

d. Average Slope. If it is obvious that all of the pertinent variables are included in the analysis, then the variance of the points about the regression line is due entirely to measurement errors, and the resulting difference in slope of the regression lines is entirely artificial. In cases where all pertinent variables are considered and most of the measurement error is in one variable, that variable should be used as the dependent variable. Its errors will then not affect the slope of the regression line. In other cases where all pertinent variables are considered, an average slope should be used. An average slope can be obtained by use of the following equation:

$$b = S_y/S_x \quad (9-14)$$

9-6. Multiple Linear Regression Example.

a. General. An example of a multiple linear regression analysis is illustrated on Figure 9-3. In this case, the volume of spring runoff is correlated with the water equivalent of the snow cover measured on April 1, the winter low-water flow (index of ground water) and the precipitation falling on the area during April. Here again, it was determined that logarithms of the values would be used in the regression equation. Although loss of 4 degrees of freedom of 12 available, as in this case, is not ordinarily desirable, the adjusted correlation coefficient attained (0.94) is particularly high, and the equation is consequently fairly reliable. The computations in Figure 9-3 were made with the HEC computer program MLRP (reference 50).

b. Logarithmic Transformation. In determining whether logarithms should be used for the dependent variable as above, questions such as the following should be considered: "Would an increase in snow cover contribute a greater increment to runoff under conditions of high ground water (wet ground conditions) than under conditions of low ground water?" If the answer is yes, then a logarithmic dependent variable (by which the effects are multiplied together) would be superior to an arithmetic dependent variable (by which the effects are added together). Logarithms should be used for the explanatory variables when they would increase the linearity of the relationship. Usually logarithms should be taken of values that have a natural lower limit of zero and a natural upper limit that is large compared to the values used in the study.

c. Function of Multiple Regression. It should be recognized that multiple regression performs a function that is difficult to perform graphically. Reliability of the results, however, is highly dependent on the availability of a large sampling of all important factors that influence the dependent variable. In this case, the standard error of an estimate as shown on Figure 9-3 is approximately 0.038, which, when added to a logarithm of a value, is equivalent to multiplying that value by 1.09. Thus, the standard error is about 9 percent, and the 1-in-20 error is roughly 18 percent. As discussed in Paragraph 9-3d, however, the calculated correlation coefficient may be accidentally high.

9-7. Partial Correlation. The value gained by using any single variable (such as April precipitation) in a regression equation can be measured by making a second correlation study using all of the variables of the regression equation except that one. The loss in correlation by omitting that variable is expressed in terms of the partial correlation coefficient. The square of the partial correlation coefficient is obtained as follows:

INPUT DATA

OBS NO	OBS ID	LOG Q	LOG SNO	LOG GW	LOG PRCP
1	1936	.939	.399	.325	.710
2	1937	.945	.343	.385	.634
3	1938	1.052	.369	.408	.886
4	1939	.744	.246	.428	.581
5	1940	.666	.181	.316	1.027
6	1941	1.081	.297	.460	1.315
7	1942	1.060	.299	.511	1.097
8	1943	.892	.354	.379	.707
9	1944	1.021	.295	.395	1.240
10	1945	.920	.321	.376	1.091
11	1946	.755	.168	.413	1.038
12	1947	.960	.280	.410	.979

STATISTICS OF DATA

VARIABLE	AVERAGE	VARIANCE	STANDARD DEVIATION	
LOG SNO	.2960	.0050	.0704	
LOG GW	.4005	.0028	.0531	
LOG PRCP	.9421	.0572	.2392	
LOG Q	.9196	.0181	.1346	DEPENDENT VARIABLE

UNBIASED CORRELATION COEFFICIENTS (R)

VARIABLE	LOG SNO	LOG GW	LOG PRCP	LOG Q
LOG SNO	1.0000	.0000	-.0459	.6308
LOG GW	.0000	1.0000	.1275	.4170
LOG PRCP	-.0459	.1275	1.0000	.2011
LOG Q	.6308	.4170	.2011	1.0000

REGRESSION RESULTS

INDEPENDENT VARIABLE	REGRESSION COEFFICIENT	PARTIAL DETERMINATION COEFFICIENT	
LOG SNO	1.621806	.9106	
LOG GW	1.012912	.6814	
LOG PRCP	.273390	.7451	

REGRESSION CONSTANT	R SQUARE	UNBIASED R SQUARE	STANDARD ERROR OF ESTIMATE
-.223698	.9437	.9226	.0375

Figure 9-3. Example Multiple Linear Regression Analysis.

$$r_{Y3.12}^2 = 1 - (1 - R_{Y.123}^2) / (1 - R_{Y.12}^2) \quad (9-15)$$

in which the subscript to the left of the decimal indicates the variable whose partial correlation coefficient is being computed, and the subscripts on the right of the decimal indicate the independent variables. An approximation of the partial correlation can sometimes be made by use of beta coefficients. After the regression equation has been calculated, beta coefficients are very easy to obtain by use of the following equation:

$$\beta_n = b_n S_n / S_Y \quad (9-16)$$

The beta coefficients of the variables are proportional to the influence of each variable on the result. While the partial correlation coefficient measures the increase in correlation that is obtained by addition of one more explanatory variable to the correlation study, the beta coefficient is a measure of the proportional influence of a given explanatory variable on the dependent variable. These two coefficients are related closely only when there is no interdependence among the various explanatory variables. However, some explanatory variables naturally correlate with each other, and when one is removed from the equation, the other will take over some of its weight in the equation. For this reason, it must be kept in mind that beta coefficients indicate partial correlation only approximately.

9-8. Verification of Regression Results. Acquisition of basic data after a regression analysis has been completed will provide an opportunity for making a check of the results. This is done simply by comparing the values of the dependent variable observed, with corresponding values calculated from the regression equation. The differences are the errors of estimate, and their root-mean-square is an estimate of the standard error of the regression-equation estimates (Paragraph 9-3). This standard error can be compared to that already established in Equation 9-13. If the difference is not significant, there is no reason to suspect the regression equation of being invalid, but if the difference is large, the regression equation and standard error should be recalculated using the additional data acquired.

9-9. Regression by Graphical Techniques. Where the relationships among variables used in a regression analysis are expected to be curvilinear and a simple transformation cannot be employed to make these relationships linear, graphical regression methods may prove useful. A satisfactory graphical analysis, however, requires a relatively large number of observations and tedious computations. The general theory employed is similar to that discussed above for linear regression. Methods used will not be discussed herein, but can be found in references 8 and 27.

9-10. Practical Guidelines. The most important thing to remember in making correlation studies is that accidental correlations occur frequently, particularly when the number of observations is small. For this reason, variables should be correlated only when there is reason to believe that there is a physical relationship. It is helpful to make preliminary examination of relationships between two or more variables by graphical plotting. This is particularly helpful for determining whether a relationship is linear and in selecting a transformation for converting curvilinear relationships to linear relationships. It should also be remembered that the chance of accidentally high

correlation increases with the number of correlations tried. If a variable being studied is tested against a dozen other variables at random, there is a chance that one of these will produce a good correlation, even though there may be no physical relation between the two. In general, the results of correlation analyses should be examined to assure that the derived relationship is reasonable. For example, if streamflow is correlated with precipitation and drainage area size, and the regression equation relates streamflow to some power of the drainage area greater than one, a maximum exponent value of one should be used, because the flow per square mile usually does not increase with drainage area when other factors remain constant.

9-11. Regional Frequency Analysis.

a. General. In order to improve flood frequency estimates and to obtain estimates for locations where runoff records are not available, regional frequency studies may be utilized. Procedures described herein consist of correlating the mean and standard deviation of annual maximum flow values with pertinent drainage basin characteristics by use of multiple linear regression procedures. The same principles can be followed using graphical frequency and correlation techniques where these are more appropriate.

b. Frequency Statistics. A regional frequency correlation study is based on the two principal frequency statistics: the mean and standard deviation of annual maximum flow logarithms. Prior to relating these frequency statistics to drainage basin characteristics, it is essential that the best possible estimate of each frequency statistic be made. This is done by adjusting short-record values by the use of longer records at nearby locations. When many stations are involved, it is best to select long-record base stations for each portion of the region. It might be desirable to adjust the base station statistics by use of the one or two longest-record stations in the region, and then adjust the short-record station values by use of the nearest or most appropriate base station. Methods of adjusting statistics are discussed in Section 3-7.

c. Drainage-Basin Characteristics. A regional analysis involves the determination of the main factors responsible for differences in precipitation or runoff regimes between different locations. This would be done by correlating important factors with the long-record mean and with the long-record standard deviation of the frequency curve for each station (the long-record values are those based on extension of the records as discussed in Section 3-7). Statistics based on precipitation measurements in mountainous terrain might be correlated with the following factors:

- Elevation of station
- General slope of surrounding terrain
- Orientation of that slope
- Elevation of windward barrier
- Exposure of gage
- Distance of leeward controlling ridge

Statistics based on runoff measurements might be correlated with the following factors:

- Drainage area (contributing)
- Stream length
- Slope of drainage area or of main channel
- Surface storage (lakes and swamps)
- Mean annual rainfall
- Number of rainy days per year
- Infiltration characteristics
- Urbanized Area

d. Linear Relationships. In order to obtain satisfactory results using multiple linear regression techniques, all variables must be expressed so that the relation between the independent and any dependent variable can be expected to be linear, and so that the interaction between two independent variables is reasonable. An illustration of the first condition is the relation between rainfall and runoff. If the runoff coefficient is sensibly constant, as in the case of urban or airport drainage, then runoff can be expected to bear a linear relation to rainfall. However, in many cases initial losses and infiltration losses cause a marked curvature in the relationship. Ordinarily, it will be found that the logarithm of runoff is very nearly a linear function of rainfall, regardless of loss rates, and in such cases, linear correlation of logarithms would be most suitable. An illustration of the second condition is the relation between rainfall, D, drainage area, A, and runoff, Q. If the relation used for correlation is as follows:

$$Q = aD + bA + c \quad (9-17)$$

then it can be seen that one inch change in precipitation would add the same amount of flow, regardless of the size of drainage area. This is not reasonable, but again a transformation to logarithms would yield a reasonable relation:

$$\log Q = d \log D + e \log A + \log f \quad (9-18)$$

or transformed:

$$Q = fD^d A^e \quad (9-19)$$

Thus, if logarithms of certain variables are used, doubling one independent quantity will multiply the dependent variable by a fixed ratio, regardless of what fixed values the other independent variables have. This particular relationship is reasonable and can be easily visualized after a little study. There is no simple rule for deciding when to use

logarithmic transformation. It is usually appropriate, however, when the variable has a fixed lower limit of zero. The transformation should provide for near-uniform variance throughout the range of data.

e. Example of Regional Correlation. An illustrative example of a regional correlation analysis for the mean log of annual flood peaks (Y) with several basin characteristics is shown on Figure 9-4. In this example, the dependent variable is primarily related to the drainage area size, but precipitation and slope added a small amount to the adjusted determination coefficient. The regression equation selected for the regional analysis included only drainage area as an independent variable.

f. Selection of Useful Variables. In the regression equations shown on Figure 9-4, the adjusted determination coefficient increases as variables are deleted according to their lack of ability to contribute to the determination. This increase is because there is a significant increase in the degrees of freedom as each variable is deleted for this small sample of 20 observations. Both the adjusted determination coefficient and standard error of estimate should be reviewed to determine how many variables are included in the adopted regression equation. Even in the case of a slight increase in correlation obtained by adding a variable, consideration of the increased unreliability of R as discussed in Section 9-3 might indicate that the factor should be eliminated in cases of small samples. The simplest equation that provides an adequate predictive capability should be selected. In this example, there is some loss in determination in only using drainage area, but this simple equation is adopted to illustrate regional analysis. The adopted equation is:

$$\log Y = 1.586 + 0.962 \log (\text{AREA}) \quad (9-20)$$

or

$$Y = 38.5 \text{ AREA}^{.962} \quad (9-21)$$

The \bar{R}^2 for this equation is 0.839.

g. Use of Map. Many hydrologic variables cannot be expressed numerically. Examples are soil characteristics, vegetal cover, and geology. For this reason, numerical regional analysis will explain only a portion of the regional variation of runoff frequencies. The remaining unexplained variance is contained in the regression errors, which varies from station to station. These regression errors are computed by subtracting the predicted values from the observed values for each station. These errors can then be plotted on a regional map at the centroid of each station's area, and lines of equal values drawn (perhaps using soils, vegetation, or topographic maps as a guide). Combining this regional error with the regression equation should be much better than using the single constant for the entire region. In smoothing lines on such a map, consideration should be given to the reliability of computed statistics. Equations 8-1 and 8-2 can be used to compute the standard errors of estimating means and standard deviations. In Figure 9-5 for example, Station 5340 (observation 11) had 66 years of record and the standard error for the mean was 0.028. There is about one chance in three that the mean is in error by more than 0.029 or about one chance in twenty that the mean is in error by more than 0.056 (twice the standard error). Figure 9-6 shows a map of the errors and Figure 9-5 shows the regional map values for each station and evaluates the worth of the map. The map has a mean square error of 0.0112 compared to that of 0.0356 for the regression equation alone.

INPUT DATA

OBS NO	OBS ID	AREA	SLOPE	LENGTH	LAKES	ELEV	FOREST	PRECIP	SOILS	MEAN
1	5090	292.0	6.3	31.5	1.5	1.230	35.0	40.0	3.5	3.783
2	5140	185.0	14.3	30.3	1.0	1.024	52.0	38.2	3.2	3.783
3	5180	282.0	44.0	29.8	1.0	1.740	64.0	35.0	3.3	4.030
4	5200	298.0	20.1	37.3	1.0	1.600	30.0	36.5	3.3	4.044
5	5205	771.0	24.4	44.8	1.0	1.447	33.0	36.0	3.2	4.333
6	5260	114.0	35.8	17.5	1.0	1.383	30.0	34.0	3.0	3.751
7	5270	52.2	29.9	17.4	1.0	1.489	26.0	33.0	3.0	2.637
8	5280	66.8	12.6	20.1	1.0	1.305	41.0	33.6	3.0	3.186
9	5305	77.5	45.2	18.9	1.0	1.123	27.0	34.6	3.0	3.348
10	5320	215.0	17.7	27.5	1.1	.966	57.0	35.5	3.0	3.995
11	5340	383.0	21.3	36.7	3.5	1.370	54.0	42.0	3.3	4.122
12	5375	15.7	291.0	5.6	1.0	1.350	81.0	40.0	3.5	2.722
13	5380	43.8	52.2	22.1	1.0	1.300	85.0	43.0	3.5	3.078
14	5390	274.0	39.6	32.3	1.6	1.200	70.0	43.0	2.8	3.930
15	5445	136.0	37.4	22.7	1.0	1.800	94.0	43.0	4.9	3.590
16	5485	604.0	22.8	44.5	1.0	1.900	83.0	37.0	3.2	4.092
17	5495	37.7	54.8	15.2	1.0	1.350	67.0	37.8	3.2	3.284
18	5500	173.0	45.7	24.2	1.0	1.700	65.0	39.0	3.9	3.816
19	5520	443.0	24.4	56.0	1.3	1.600	89.0	44.0	3.2	4.275
20	5525	23.8	115.7	10.0	1.0	1.800	80.0	47.0	4.3	3.249

STATISTICS OF DATA

VARIABLE	AVERAGE	VARIANCE	STANDARD DEVIATION	
AREA	2.1488	.2249	.4743	
SLOPE	1.5105	.1255	.3542	
LENGTH	1.3832	.0550	.2345	
STORAGE	.0540	.0171	.1308	
ELEV	1.4339	.0707	.2659	
FOREST	1.7293	.0353	.1879	
PRECIP	1.5845	.0020	.0444	
SOILS	.5230	.0034	.0581	
MEAN	3.6524	.2455	.4955	DEPENDENT VARIABLE

UNBIASED CORRELATION COEFFICIENTS

VARIABLE	AREA	SLOPE	LENGTH	LAKES	ELEV	FOREST	PRECIP	SOILS	MEAN
AREA	1.0000	-.6327	.9304	-.2749	.0000	.0000	.0000	.0000	.9159
SLOPE	-.6327	1.0000	-.7144	-.1318	.1187	.3635	.1867	.2053	-.4521
LENGTH	.9304	-.7144	1.0000	.2345	.0000	.0000	.0000	-.1096	.8263
STORAGE	.2749	-.1318	.2345	1.0000	.0000	.0000	.2812	.0000	.2596
ELEV	.0000	.1187	.0000	.0000	1.0000	.2791	.0000	.5297	.0000
FOREST	.0000	.3635	.0000	.0000	.2791	1.0000	.6877	.4304	.0000
PRECIP	.0000	.1867	.0000	.2812	.0000	.6877	1.0000	.5412	.0000
SOILS	.0000	.2053	-.1096	.0000	.5297	.4304	.5412	1.0000	.0000
MEAN	.9159	-.4521	.8263	.2596	.0000	.0000	.0000	.0000	1.0000

SUMMARY OF REGRESSION ANALYSES FOR MEAN LOG OF ANNUAL PEAKS

REGRESSION CONSTANT	AREA (LOG)	SLOPE (LOG)	LENGTH (LOG)	LAKES (LOG)	ELEV (NONE)	FOREST (LOG)	PRECIP (LOG)	SOILS (LOG)	ADJUSTED DETERMINATION COEFFICIENT	STANDARD ERROR OF ESTIMATE	MEAN SQUARE ERROR
-1.522	1.261	0.182	-0.328	-0.272	-0.184	0.055	1.739	0.140	0.8097	0.2162	0.0257
-1.633	1.269	0.169	-0.364	-0.289	-0.169	0.054	1.874	-----	0.8254	0.2070	0.0257
-1.808	1.267	0.179	-0.350	-0.301	-0.165	-----	2.023	-----	0.8386	0.1990	0.0258
-1.668	1.130	0.243	-----	-0.251	-0.167	-----	1.753	-----	0.8469	0.1939	0.0263
-1.130	1.104	0.250	-----	-----	-0.129	-----	1.399	-----	0.8537	0.1896	0.0269
-1.034	1.069	0.198	-----	-----	-----	-----	1.319	-----	0.8584	0.1865	0.0278
-1.134	0.975	-----	-----	-----	-----	-----	1.699	-----	0.8553	0.1885	0.0302
1.586	0.962	-----	-----	-----	-----	-----	-----	-----	0.8390	0.1988	0.0356

Figure 9-4. Regression Analysis for Regional Frequency Computations.

REGIONAL ANALYSIS WITH REGRESSION ON DRAINAGE AREA ONLY

OBS NO	STATION	OBSERVED	COMPUTED	ERROR	MAP VALUE	DIFF	DIFF ²	YEARS OF RECORD	STANDARD DEVIATION	S _y
1	5090	3.783	3.957	-0.174	-0.18	0.006	0.000036	43	0.190	0.029
2	5140	3.783	3.766	0.017	0.01	0.007	0.000049	52	0.195	0.027
3	5180	4.030	3.942	0.088	0.09	-0.002	0.000004	39	0.289	0.046
4	5200	4.044	3.965	0.079	0.07	0.009	0.000081	27	0.256	0.049
5	5205	4.333	4.362	-0.029	0.08	-0.109	0.011881	50	0.251	0.035
6	5260	3.751	3.564	0.187	0.11	0.077	0.005929	35	0.293	0.050
7	5270	2.637	3.238	-0.601	-0.22	-0.381	0.145161	31	0.206	0.037
8	5280	3.186	3.341	-0.155	-0.17	0.015	0.000225	45	0.186	0.028
9	5305	3.348	3.403	-0.055	-0.04	-0.015	0.000225	44	0.128	0.019
10	5320	3.995	3.829	0.166	0.16	0.006	0.000036	66	0.288	0.035
11	5340	4.122	4.070	0.052	0.02	0.032	0.001024	66	0.227	0.028
12	5375	2.722	2.736	-0.014	-0.05	0.036	0.001296	40	0.323	0.051
13	5380	3.078	3.164	-0.086	-0.08	-0.006	0.000036	60	0.226	0.029
14	5390	3.930	3.930	0.000	0.00	0.000	0.000000	41	0.261	0.041
15	5445	3.590	3.638	-0.048	-0.15	0.102	0.010404	39	0.278	0.045
16	5485	4.092	4.260	-0.168	-0.08	-0.088	0.007744	61	0.242	0.031
17	5495	3.284	3.102	0.182	0.04	0.142	0.020164	39	0.242	0.039
18	5500	3.816	3.738	0.078	0.08	-0.002	0.000004	66	0.237	0.029
19	5520	4.275	4.131	0.144	0.17	-0.026	0.000676	54	0.277	0.038
20	5525	3.249	2.910	0.339	0.20	0.139	0.019321	39	0.291	0.047
Sum					0.06	-0.058	0.224296			
Average					0.003	-0.003	0.0112			

Figure 9-5. Regional Analysis Computations for Mapping Errors.

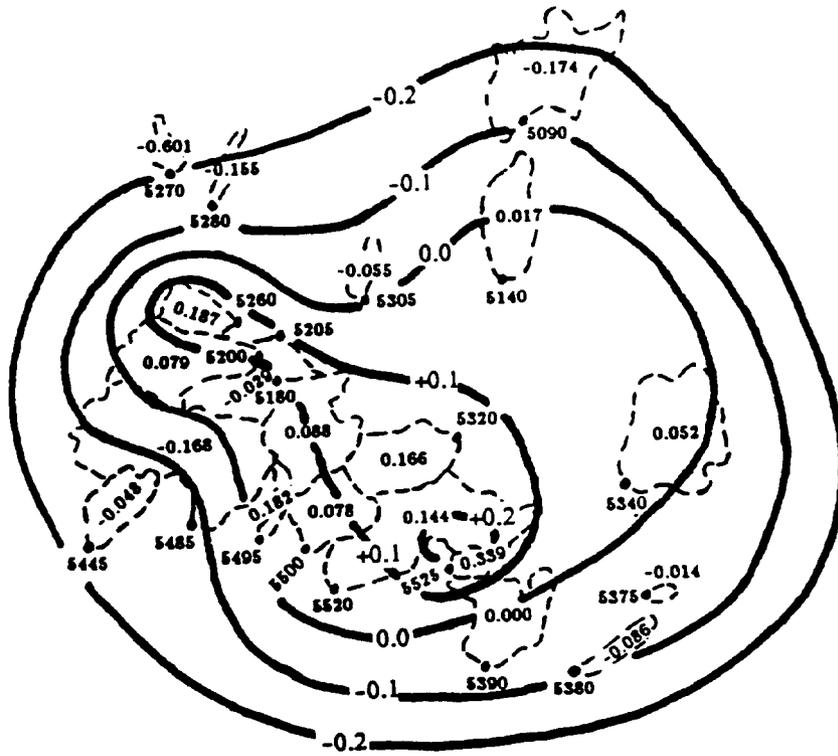


Figure 9-6. Regional Map of Regression Errors.

h. Summary of Procedure. A regional analysis of precipitation or flood-flow frequencies is generally accomplished by performing the following steps:

- (1) Select long-record base stations within the region as required for extension of records at each of the short-record stations.
- (2) Tabulate the maximum events of each station.
- (3) Transform the data to logarithms and calculate \bar{X} , S and, if appropriate, G (Equations 3-1, 3-2 and 3-3) for each base station.
- (4) Calculate \bar{X} and S for each other station and for the corresponding values of the base station, and calculate the correlation coefficient (Equation 3-16).
- (5) Adjust all values of \bar{X} and S by use of the base station, (Equations 3-17 and 3-19). (If any base station is first adjusted by use of a longer-record base station, the longer-record statistics should be used for all subsequent adjustments.)
- (6) Select meteorological and drainage basin parameters that are expected to correlate with \bar{X} and S, and tabulate the values for each drainage basin or representative area.
- (7) Calculate the regression equations relating \bar{X} and S to the basin characteristics, using procedures explained in Section 9-2, and compute the corresponding determination coefficients.
- (8) Eliminate variables in turn that contribute the least to the determination coefficient, recomputing the determination coefficient each time, and select the regression equation having the highest adjusted determination coefficient, or one with fewer variables if the adjusted determination coefficient is nearly the same.
- (9) Compute the regression errors for each station, plot on a suitable map, and draw isopleths of the regression errors for the regression equations of \bar{X} (see Figures 9-5 and 9-6 for an example) and S considering the standard error for each computed, or adjusted, \bar{X} and S. Note that an alternate procedure is to add the regression constant to each error value and develop a map of this combined value. This procedure eliminates the need to keep the regression constant in the regression equation as the mapped value now includes the regression constant.
- (10) A frequency curve can be computed for any ungaged basin in the area covered within the mapped region by using the adopted regression equations and appropriate map values to obtain \bar{X} and S, and then using the procedures discussed in Section 3-2 to compute several points to define the frequency curve. (It may also be necessary to develop regional (generalized) values of the skew coefficient if the Pearson type III distribution is considered appropriate. The next section describes the necessary steps to compute a generalized skew coefficient.)

i. Generalized Skew Determinations. Skew coefficients for use in hydrologic studies should be based on regional studies. Values based on individual records are highly unreliable. Figure 9-7 is a plot of skew coefficients sequentially recomputed after adding the annual peak for the given year. Note that, after 1950, the skew coefficient was at a minimum of about 0.5 in 1954 and maximum of about 1.9 in 1955, only one year

apart. The procedures for developing generalized skew values are generally set forth in Bulletin 17B (pages 10-15).

In summary, it is recommended that:

- (1) the stations used in the study have 25 or more years of data,
- (2) at least 40 stations be used in the analysis, or at least all stations surrounding the area within 100 miles should be included,
- (3) the skew values should be plotted at the centroid of the basins to determine if any geographic or topographic trends are present,
- (4) a prediction equation should be developed to relate the computed skew coefficients to watershed and climate variables,
- (5) the arithmetic mean of at least 20 stations, if possible, in an area of reasonably homogeneous hydrology should be computed, and
- (6) then select the method that provides the most accurate estimation of the skew coefficient (smallest mean-square error).

In addition to the above guidelines, care should be taken to select stations without significant man-made changes such as reservoirs, urbanization, etc.

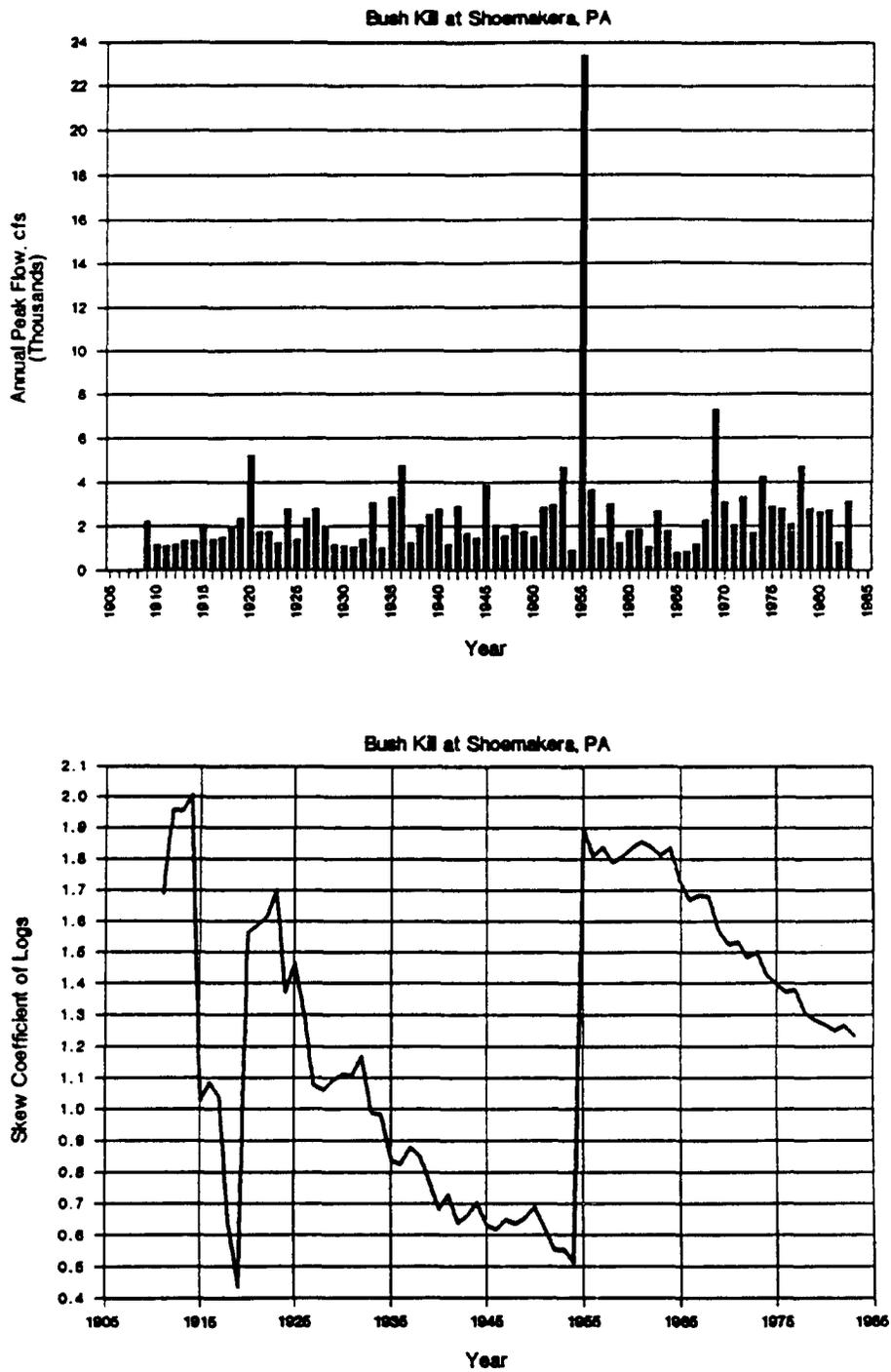


Figure 9-7. Annual Peaks and Sequential Computed Skew by Year.