VOL. 6, NO. 6 WATER RESOURCES RESEARCH DECEMBER 1970

Definition and Uses of the Linear Regression Model

M. H. DISKIN*

USDA Southwest Watershed Research Center, Agricultural Research Service Tucson, Arizona 85705

Abstract. A simple three-element model is proposed as an interpretation of the regression equation for the relationship between annual rainfall and annual runoff from watersheds. The model employs two parameters that are related to the constants used in the regression equation. The parameters of the model can be evaluated by the usual least squares equations if the runoff data do not include zero or near zero values. For arid or semiarid watersheds where runoff may be zero for some years, a special procedure is proposed for evaluating the parameters. The procedure seeks the minimum of an objective function defined as the sum of squared deviations between observed data and prediction lines defined by the regression model.

INTRODUCTION

One of the first tools adopted by early hydrologists in their investigations was the linear regression equation. The equation in its simplest form expresses the relationship between two variables, X and Y, in a simple linear form

$$Y = AX + B \tag{1}$$

where A and B are constants expressing the slope of the straight line and the Y axis intercept, respectively. Estimates of the values of the two parameters are obtained by a least squares procedure leading to the equations

$$A = \frac{\sum (X_i Y_i) - N\langle X\rangle\langle Y\rangle}{\sum (X_i^2) - N\langle X\rangle^2} \tag{2}$$

$$B = \langle Y\rangle - A\langle X\rangle \tag{3}$$

where X_i and Y_i are individual values of the variables, N is the number of observations, $\langle X\rangle$ and $\langle Y\rangle$ are the means of the N observations, and the summations are carried out over the N observations.

The linear regression equation has been used fairly extensively in hydrology and, with proper care, has provided a useful tool for prediction purposes. One of the uses that gave satisfactory results was the relationship between annual precipitation P and annual runoff R of watersheds. The use of the linear regression equation for annual or seasonal rainfall–runoff relationships is described in a number of technical papers, and it is included in some textbooks on hydrology [*Johnstone and Cross*, 1949; *Linsley et al.*, 1958; *Wilson*, 1969]. The value of the constant B is invariably negative, and if the precipitation P and runoff R are expressed in the same units, the constant A is usually less than unity, leading to an equation of the form

$$R = AP - B \tag{4}$$

The purpose of this paper is to present an interpretation of equation 4 in terms of a simple conceptual model for the rainfall–runoff relationship and to discuss some problems associated with evaluating the parameters of the model for arid or semiarid watersheds. The model proposed will be called the linear regression model to emphasize its relationship to equation 4.

LINEAR REGRESSION MODEL

The conceptual model proposed is composed of three elements as shown in Figure 1, and its operation may be described in terms of the three operators that characterize these three elements. The definition of the first two elements involves the use of one parameter for each element. The third element does not need a parameter for its definition. The two parameters thus defined are related to the two constants that appear in the linear regression equation.

* On leave from the Technion-Israel Institute of Technology, Haifa, Israel.

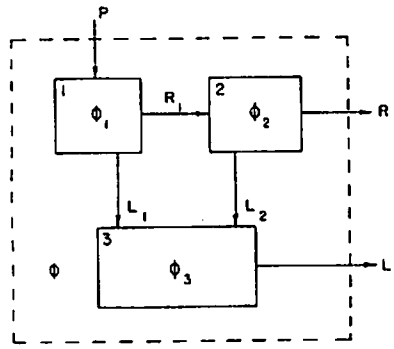

Fig. 1. The linear regression model.

The first element receives as input the annual precipitation P and divides it into two outputs R_1 and L_1, which may be called initial runoff and initial loss, respectively. The operator ϕ_1 describing this operation is specified as follows:

if $P \leq C$

$$L_1 = P \quad \text{and} \quad R_1 = 0 \tag{5}$$

if $P > C$

$$L_1 = C \quad \text{and} \quad R_1 = P - C \tag{6}$$

where C is the parameter needed to specify the operator ϕ_1 numerically.

The second element receives as input the initial runoff R_1 produced as output of the first element and divides it into two outputs R and L_2. The quantity L_2 may be considered to represent the losses incurred during the flow of the initial runoff R_1 to the watershed outlet. The output R represents the annual runoff from the watershed remaining after the abstractions of the two types of losses. The operator ϕ_2 describing the operation of the second element is specified by

$$R = AR_1 \quad \text{and} \quad L_2 = (1 - A)R_1 \tag{7}$$

where A is a constant parameter needed to specify the operator ϕ_2 numerically.

The third element receives two inputs, L_1 and L_2, and produces one output L equal to the sum of the two inputs. The operator ϕ_3 is specified by the summation operation

$$L = L_1 + L_2 \tag{8}$$

The output of the element L represents the total losses of the watershed.

Instead of breaking down the model into three elements, it may also be described in terms of one operator or one system. The complete system represented by the dashed line box in Figure 1 and by the operator ϕ receives an input P equal to the annual depth of precipitation and produces two outputs, the annual volume of runoff R and the annual volume of losses L. Combining the operations of the first two elements (1 and 2) leads to the following direct relationships between annual volume of runoff and annual depth of precipitation:

if $P \leq C$

$$R = 0 \tag{9}$$

if $P > C$

$$R = A(P - C) = AP - B \tag{10}$$

where $B = AC$ is a constant parameter. Similarly, the relationship between annual losses and annual precipitation is given by the following expressions:

if $P \leq C$

$$L = P \tag{11}$$

if $P > C$

$$L = (1 - A)P + B \tag{12}$$

A few examples of the use of the linear regression equation are summarized in Table 1. The examples selected refer to watersheds ranging in size from 0.6 to 152,000 square miles. The values of the parameters given in the table are either those suggested in the original references listed or in some cases values recomputed from data given in them. The range of values for the parameter A is from 0.37 to 0.99; for B from 11.9 to 34.0 inches; and for C from 16.3 to 36.8 inches. The small range of these values is remarkable, but it is not intended to imply that values outside these ranges cannot be obtained.

The physical significance of the parameter C is related to the presence of surface and subsurface storage on the watershed that does not contribute to runoff at its outlet. It represents approximately the sum of the annual evapotranspiration from the surface of the watershed and the annual contribution to regional groundwater. The different values of the parameter for differ-

ent watersheds reflect variations in climatic conditions, soil properties, vegetation characteristics, as well as different temporal distributions of rainfall throughout the year. The physical significance of the parameter $(1 - A)$ is probably related to the characteristics of the soil and vegetation along the channel system of the watershed as well as to the variability of flow in the channels. The parameter A may also include a correction factor needed to convert the observed annual precipitation to the true input to the watershed [*Amorocho and Orlob*, 1961]. In such cases it is possible to obtain values of A greater than 1.0. An example for such a case is given by *Amorocho and Orlob* [1961] when they used a rain gage outside the French Dry Creek watershed to estimate the mean rainfall over that watershed. The value of $A = 0.70$ given for this watershed in Table 1 was obtained when data for a rain gage inside the watershed were used.

Another theoretical possibility for values of A larger than 1 is for a watershed receiving large contributions of groundwater from other watersheds. If these contributions appear at the outlet of the watershed, annual runoff may exceed annual rainfall over the area of the watershed, leading to a value of A larger than 1.0. The model proposed herein is not suitable to represent watersheds where existing conditions give rise to values of A larger than 1.0.

EVALUATION OF MODEL PARAMETERS

The evaluation of the parameters of the linear regression model is a simple application of equations 2 and 3, provided that all values of runoff are larger than zero, as was the case for all watersheds listed in Table 1. If some of the runoff values are zero or nearly zero, as may happen on arid and semiarid watersheds, the use of the above equations with all the data included leads to erroneous results. The simple remedy of removing from the data all zero values still does not produce the optimal values of the parameters. This is due to the presence of the deviations of the observed data from the lines represented by equations 9 and 10.

The values of the parameters may be estimated fairly closely by a graphical procedure of passing a straight line (by eye) through the plotted data. If an objective estimate of the optimal parameters is desired, a special procedure described below may be followed. The procedure will be described with reference to a set of precipitation and runoff data given in Table 2. The data have been prepared for illustrating the procedure used and do not represent observed values from a real watershed.

The parameter evaluation procedure, which can be carried out by a digital computer, starts by arranging the data in increasing values of the observed precipitation P, as in Table 2. The observed data are then divided into two

TABLE 1. Examples of the Use of Linear Regression Equation for Annual Rainfall-Runoff Relationship

No.	Watershed Location	Area, square miles	Parameters A	B, inches	C, inches	Reference
1	Green Acre Branch, Missouri	0.6	0.66	15.0	22.6	*Whipkey* [1960]
2	Little Beaver Creek, Missouri	6.4	0.71	14.2	19.9	*Whipkey* [1960]
3	Beaver Creek, Missouri	14.0	0.75	14.7	19.7	*Whipkey* [1960]
4	Taylor Creek (W-3), Florida	15.7	0.83	29.1	35.0	*Stephens* [1970]
5	Ahoskie Creek, North Carolina	57.0	0.61	13.7	22.4	*Stephens* [1970]
6	French Dry Creek, California	72.0	0.70	25.7	36.8	*Amorocho and Orlob* [1961]
7	Taylor Creek (W-2), Florida	99.0	0.97	34.0	35.0	*Stephens* [1970]
8	Hurricane Creek near Alma, Georgia	150.0	0.50	12.5	25.0	*Stephens* [1970]
9	Merrimack River above Lawrence, Massachusetts	4460.0	0.80	13.0	16.3	*Linsley et al.* [1958]
10	Cedar River above Cedar Rapids, Iowa	6510.0	0.99	25.1	25.4	*Foster* [1949]
11	Volta River, Africa	152090.0	0.37	11.9	32.6	*Amorocho and Orlob* [1961]

TABLE 2. Data Used in Example and Predicted Values

Observation No.	Precipitation P	Runoff R	Losses L	Predicted Runoff
1	2.79	0.00	2.79	0.00
2	3.61	0.01	3.60	0.00
3	4.14	0.08	4.06	0.00
4	4.63	0.00	4.63	0.00
5	5.11	0.20	4.91	0.00
6	5.20	0.00	5.20	0.00
7	5.58	0.00	5.58	0.00
8	6.29	0.02	6.27	0.00
9	6.45	0.42	6.03	0.00
10	6.87	0.25	6.62	0.17
11	7.09	0.16	6.93	0.32
12	7.12	0.59	6.53	0.35
13	7.54	0.40	7.14	0.64
14	7.93	1.03	6.90	0.90
15	8.41	0.22	8.19	1.24
16	8.60	1.82	6.78	1.37
17	9.20	1.80	7.40	1.78
18	9.55	2.91	6.64	2.02
19	10.19	2.63	7.56	2.46
20	10.98	2.83	8.15	3.01
21	11.60	3.65	7.95	3.44
22	12.60	2.99	9.61	4.13
23	12.63	4.36	8.27	4.15
24	13.21	4.20	9.01	4.55
25	13.78	5.61	8.17	4.94
26	14.51	5.30	9.21	5.45
27	15.99	6.64	9.35	6.47
Mean	8.578	1.782	6.796	

Measurements are given in inches.

groups by choosing one of the values of P as a separation point P_0, and the least squares equations (equations 2 and 3) are applied only to observations in the group for which the precipitation is higher than the separation point ($P > P_0$). The straight line thus obtained will be the best fitting line for the observations within the group so defined. The sum S of the squared deviation of all the observed values of runoff and the values predicted by the line

$$R = AP + B \quad \text{for} \quad P > P_0 \tag{13}$$

or by the line

$$R = 0 \quad \text{for} \quad P \leq P_0 \tag{14}$$

is then computed. This sum will obviously be a function of the separation point P_0

$$S = f(P_0) \tag{15}$$

If the point P_0 is varied systematically, it is possible to plot the functional relationship between the sum S defined as the objective function and the separation point P_0 and to determine the value of P_0 that will minimize the function S. The values of the parameters computed for this division point will be the optimal in the sense that the objective function S has obtained its minimal value consistent with the adopted structure of the model. The curve obtained for the data given in Table 2 is shown in Figure 2. The second minimum in Figure 2 reflects the effects of the large deviations of the points in the vicinity of the separation point. Instead of plotting the curve represented by equation 15, the value of P_0 giving the smallest sum of squares and the corresponding values of the parameters can, of course, be chosen by the computer program or by inspection from a print-out of the results.

The regression lines obtained by the above procedure for the data in Table 2 are shown in Figure 3, lines A and B. The equations of the

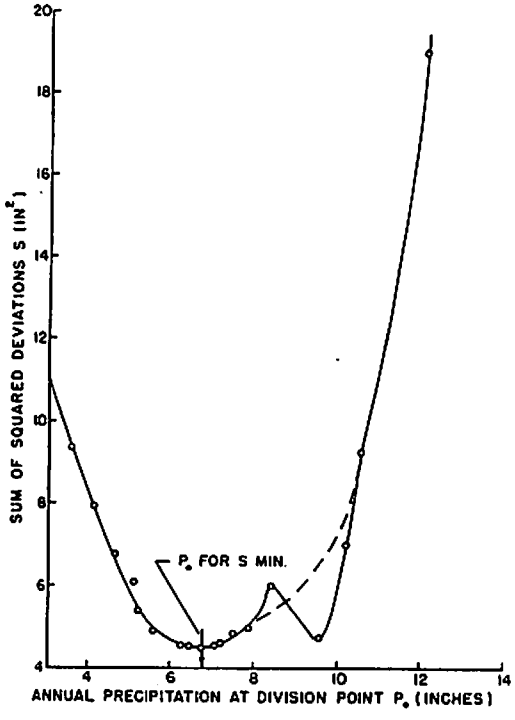

Fig. 2. Objective function for the linear regression model.

lines are

$$R = 0.69P - 4.57 \quad \text{for} \quad P > 6.62 \quad (16)$$

$$R = 0 \quad \text{for} \quad P \leq 6.62 \quad (17)$$

The line obtained by applying equations 2 and 3 to all data in Table 2 is also shown in Figure 3, line C, and it is obvious that it does not give a good representation of the data. Values of runoff predicted by the linear regression model (equations 16 and 17) are given in Table 2 for comparison with the original data.

An alternative procedure for evaluation of the parameters is a mapping technique in which the values of the parameters are varied systematically, and the sum of the squares of the deviations is computed for each set of assumed values of the parameters. The resulting map for the data used in the above example is shown in Figure 4. The optimal point obtained by the procedure discussed above is also shown on the map. Hill climbing procedures, such as those given by *Green* [1970], *DeCoursey and Snyder* [1969], or *Dawdy and O'Donnel* [1965], may also be employed. The procedure proposed in this paper and illustrated by Figure 2 appears to be the simplest and most economical in terms of computer time for the model proposed herein.

CONCLUSIONS

The linear regression equation may be interpreted in terms of a simple conceptual model for rainfall–runoff relationships. The model is composed of three elements that receive the annual precipitation as input and produce the annual runoff and annual losses as output.

The parameters of the model can be evaluated by the usual least squares equations as long as all runoff values are well above zero. A special procedure is presented for evaluating the parameters for watersheds in arid or semiarid locations where some of the runoff values are zero or nearly zero.

The linear regression model or equations that can be represented by the model have been used

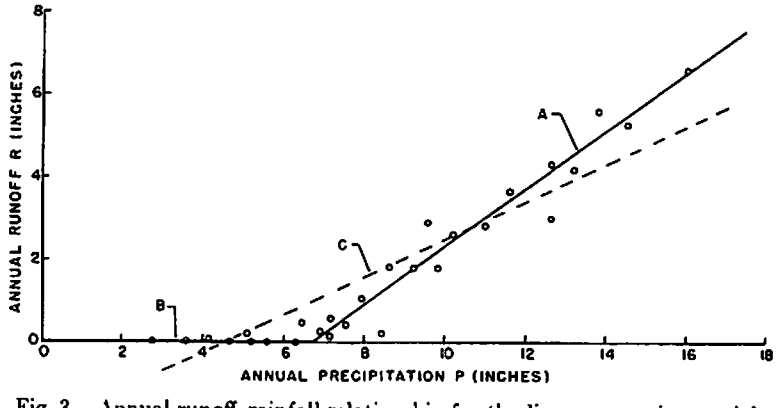

Fig. 3. Annual runoff–rainfall relationship for the linear regression model.

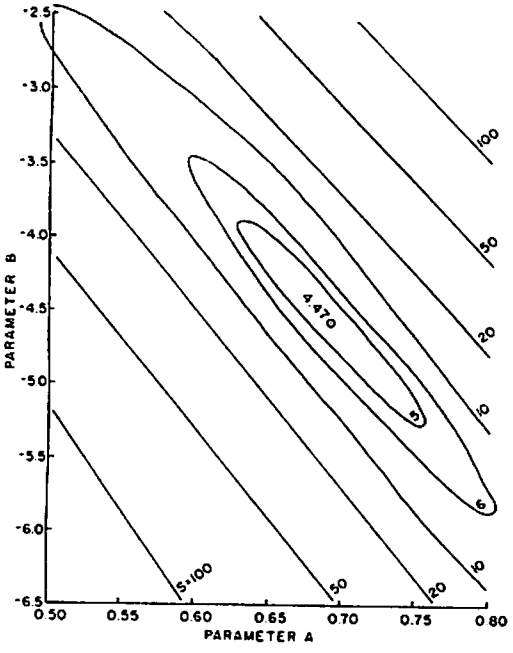

Fig. 4. Sum of squared deviations for the linear regression model $S = f(A, B)$.

in the past for representing annual or seasonal precipitation–runoff relationships. The proposed model provides a convenient tool for prediction of annual runoff and may be useful in cases where the use of more complicated models is not justified.

The linear regression model is not applicable in cases where there is appreciable carryover or lag between rainfall and runoff. If the time interval chosen to be represented by the model is too short, rainfall during any one interval will cause runoff during two or more such intervals. The model in its present form cannot accommodate such a time distribution procedure, but it can be changed to do so by adding a fourth element. The linear regression model is also not applicable in cases where a large part of the annual runoff is derived from groundwater inflow from adjacent watersheds.

The best application of the linear regression model appears to be for annual rainfall–runoff relationships in watersheds where two or more distinct seasons are present. These watersheds exhibit similar storage conditions just before the beginning of the rainy season every year. Adopting this point as the start of the hydrologic year results in a time interval for which there is minimum of carryover from one year to the next so that the linear regression model produces satisfactory results.

Acknowledgments. The paper is based on material partly presented at the ARS-SCS Workshop on Watershed Modeling held on March 16–18, 1970, in Tucson, Arizona [*Diskin*, 1970]. The paper is a contribution of the Southwest Watershed Research Center, U.S. Department of Agriculture, Soil and Water Conservation Research Division.

REFERENCES

Amorocho, J., and G. T. Orlob, An evaluation of the inflow-runoff relationships in hydrologic studies, *Water Resour. Center Contrib. 41*, University of California, Berkeley, California, 1961.

Dawdy, D. R., and T. O'Donnel, Mathematical models of catchment behavior, *J. Hydraul. Div., Amer. Soc. Civil Eng., 91*(HY4), 123–137, 1965.

DeCoursey, D. G., and W. M. Snyder, Computer oriented method of optimizing hydrologic model parameters, *J. Hydrol., 9*, 34–56, 1969.

Diskin, M. H., Objectives and techniques of watershed modeling, paper presented at the Watershed Modeling Workshop, U.S. Department of Agriculture, Agricultural Research Service, Soil Conservation Service, Tucson, Arizona, March 16–18, 1970.

Foster, E. E., *Rainfall and Runoff*, 487 pp., MacMillan, New York, 1949.

Green, R. F., Optimization by the pattern search method, 73 pp., *T.V.A. Hydraul. Data Br. Res. Pap. 7*, Knoxville, Tennessee, January 1970.

Johnstone, D., and W. P. Cross, *Elements of Applied Hydrology*, 276 pp., Ronald, New York, 1949.

Linsley, R. K., M. A. Kohler, and J. L. H. Paulhus, *Hydrology for Engineers*, 340 pp., McGraw-Hill, New York, 1958.

Stephens, J. C., Watersheds and hydrologic models in the southeast, pp. 19-1 to 19-39, paper presented at the Watershed Modeling Workshop, U.S. Department of Agriculture, Agricultural Research Service, Soil Conservation Service, Tucson, Arizona, March 16–18, 1970.

Whipkey, R. Z., Annual water yield from small watersheds in the Ozarks, paper presented at the Hydrology Workshop of the U.S. Department of Agriculture, New Orleans, Louisiana, October 24–27, 1960.

Wilson, E. M., *Engineering Hydrology*, 182 pp., Macmillan, London, 1969.

(Manuscript received June 8, 1970.)