GUIDELINES FOR THE USE OF STRUCTURAL VERSUS REGRESSION ANALYSIS IN GEOMORPHIC STUDIES

U.S. GEOLOGICAL SURVEY

Water-Resources Investigations 78-135

S_d

$y = a + bx + e_y$

$SE_s = \left[\Sigma(\log_{10} y - \log_{10} y_s)^2/(n-2)\right]^{1/2}$

SE_r

$y = ax^b e_y$

$\bar{x}$

b_r

R^2

$\bar{y}$

U.S. DEPARTMENT OF THE INTERIOR
March 3, 1849

BIBLIOGRAPHIC DATA SHEET	1. Report No.	2.	3. Recipient's Accession No.

4. Title and Subtitle

GUIDELINES FOR THE USE OF STRUCTURAL VERSUS REGRESSION ANALYSIS IN GEOMORPHIC STUDIES

5. Report Date
November 1978

6.

7. Author(s)
W. R. Osterkamp, J. M. McNellis, and P. R. Jordan

8. Performing Organization Rept. No. WRI 78-135

9. Performing Organization Name and Address

U.S. Geological Survey, Water Resources Division
1950 Avenue "A" - Campus West
Lawrence, Kansas 66045

10. Project/Task/Work Unit No.

11. Contract/Grant No.

12. Sponsoring Organization Name and Address

U.S. Geological Survey, Water Resources Division
1950 Avenue "A" - Campus West
Lawrence, Kansas 66045

13. Type of Report & Period Covered
Final

14.

15. Supplementary Notes

16. Abstracts

Simple-regression and structural analysis are similar methods of developing a linear relation from a bivariant group of data. Regression analysis is a useful curve-fitting technique, but often is misapplied to geomorphic data sets. When error components can be identified for both variables, the statistical technioue of structural analysis is preferred. If regression results are available, conversion to a structural analysis can be made either manually or by computer.

Use of computer-generated data sets permits the construction of curves relating variation between regression and structural analyses to the range of data of the independent variable. The data have randomly imposed error components of specified standard deviation of a slope of the linear relation that simulates gradient-discharge relations of natural alluvial streams. The empirically developed curves can be used to determine the need for structural analysis of real geomorphic data sets.

17. Key Words and Document Analysis. 17a. Descriptors

Structural analysis, regression analysis, curve-fitting technioues, statistics

17b. Identifiers/Open-Ended Terms

Power-function equations, artificial data, geomorphology, alluvial streams

17c. COSATI Field/Group

18. Availability Statement No restriction on distribution	19. Security Class (This Report) UNCLASSIFIED	21. No. of Pages
	20. Security Class (This Page UNCLASSIFIED	22. Price

GUIDELINES FOR THE USE OF STRUCTURAL VERSUS REGRESSION ANALYSIS IN GEOMORPHIC STUDIES

By W. R. Osterkamp, J. M. McNellis, and P. R. Jordan

U.S. GEOLOGICAL SURVEY

WATER-RESOURCES INVESTIGATIONS 78-135

November 1978

UNITED STATES DEPARTMENT OF THE INTERIOR

CECIL D. ANDRUS, SECRETARY

GEOLOGICAL SURVEY

H. WILLIAM MENARD, DIRECTOR

For additional information write to:

U.S. Geological Survey
1950 Ave. A - Campus West
University of Kansas
Lawrence, Kansas 66045

CONTENTS

ILLUSTRATIONS

TABLES

ABSTRACT

Simple-regression and structural analysis are similar methods of developing a linear relation from a bivariant group of data. Regression analysis is a useful curve-fitting technique, but often is misapplied to geomorphic data sets. When error components can be identified for both variables, the statistical technique of structural analysis is preferred. If regression results are available, conversion to a structural analysis can be made either manually or by computer.

Use of computer-generated data sets permits the construction of curves relating variation between regression and structural analyses to the range of data of the independent variable. The data have randomly imposed error components of specified standard deviation and a slope of the linear relation that simulates gradient-discharge relations of natural alluvial streams. The empirically developed curves can be used to determine the need for structural analysis of real geomorphic data sets.

INTRODUCTION

Attempts to recognize relations between two or more variables are a fundamental operation of scientific empiricism. The earliest technique probably was a graphical analysis that served well despite the absence of a well-defined criterion for "best fit," tests of significance, and confidence limits. After mathematical techniques of simple regression and multiple regression had been developed, graphical examination remained useful for judging the applicability of various statistical techniques and as a means of presenting data and results. For relations that are not clearly defined by the data, graphical examination sometimes has led to the suspicion that the regression relation is biased and that the bias is associated with errors in the data.

For an equation relating two variables, errors associated with individual observations can be of two types: (1) error of measurement and (2) error of estimation of observed value by the equation. The second type of error may be considered as the net result of the influences of many variables not included in the equation. Each of these influences is believed to be too small to be predictable from commonly available data. Thus the error may be called "stochastic" error. Error of measurement is possible for either variable, but stochastic error occurs only in the variable being predicted by the equation (the dependent variable). In the following discussion and analysis, distinction between measurement and stochastic errors of the dependent variable generally is unnecessary, and the term "error" will refer to the net result of the two types.

When considering two variables, three general conditions of error can be identified: (1) both variables have little or no error, (2) one variable has little or no error, and (3) both variables have significant error. It has been pointed out that the third condition often applies to the earth sciences and to geomorphic studies in particular. Commonly, it is impossible to distinquish between the dependent and independent variables (Miller and Kahn, 1962, p. 186; Brooks, Hart, and Wendt, 1972; Till, 1973; Mark and Church, 1977, p. 64; and Cox, 1977, p. 1200).

The first condition, when the independent and dependent variables are x and y, respectively, is represented by the equation:

$$y = f(x) \ . \tag{1}$$

This invariable correspondence between y and x results in a correlation coefficient of 1.00. When the independent variable is measured without significant error, but the dependent variable has associated error (condition 2), the relation can be written as:

$$y = f(x) + e_y \ . \tag{2}$$

Regression analysis was developed specifically for this relation, and application of regression analysis to other conditions produces biased results (Mark and Church, 1977, p. 63-64; Snedecor and Cochran, 1967, p. 164-166). In the third condition, common in geomorphic studies, the true value of x is unknown, and the result of the measurement is:

$$x' = x + e_x \ , \tag{3}$$

where e_x is error associated with the independent variable. Thus, for condition (3):

$$y = f(x' - e_x) + e_y \ . \tag{4}$$

Mark and Church (1977, p. 66-67) discuss the misuse of regression analysis and describe a modification, which they term structural analysis, that considers the errors associated with both variables. Unless e_x is small relative to e_y, or both e_x and e_y are small relative to x and y, respectively, structural analysis is the preferred statistical technique for developing a relation for the third condition.

The purpose of the paper presented here is to discuss the practical application of simple-structural analysis, as described and advocated by Mark and Church (1977). Depending on the relative magnitudes of the two variables and their concomitant errors, a decision should be made on the worth of extending an analysis to obtain a structural relation. In many situations the equations developed through the two techniques, regression and structural analysis, are so closely similar that application of structural analysis is unnecessary. This paper extends the discussion of Mark and Church (1977), provides a specific description of the application of structural analysis, and offers guidelines, as well as precautions, on the practical use of the technique.

DEVELOPMENT OF GUIDELINES

The recent paper, "On the misuse of regression in earth science," by Mark and Church (1977), provides insight in the proper use of regression analysis. It briefly describes the conditions when least-squares regression is used properly and conditions when the related technique of structural analysis should be applied. This paper is valuable in guiding the proper selection of a curve-fitting technique for relating two or more variables and in providing an easily applied method to convert a regression relation to a structural relation.

For sets of two linearly related variables, x and y, the purpose of least-squares regression analysis is the determination of the intercept (constant a) and the slope (coefficient b) for the linear relation:

$$y = a + bx + e_y \ . \tag{5}$$

When subscripts r and s, respectively, refer to regression and structural relations, the pertinent equation given by Mark and Church (1977, p. 67) to convert a regression to a structural relation is:

$$b_s = \frac{(b_r{}^2/R^2 - \lambda) + [(b_r{}^2/R^2 - \lambda)^2 + 4\lambda b_r{}^2]^{\frac{1}{2}}}{2b_r} \ . \tag{6}$$

R is the coefficient of correlation, and λ describes the relative errors of the two variables:

$$\lambda = \frac{e_y{}^2}{e_x{}^2} \ . \tag{7}$$

Inspection of equation (6) shows that, if regression results including the correlation coefficient are available, only a value for λ is required to convert to a structural relation. As previously shown, λ is a measure of the relative error between the dependent and independent variables; whereas, the correlation coefficient (R) and the coefficient of determination (R^2) can be regarded as indicators of data scatter relative to the range of the data. Mark and Church (1977, p. 66) point out that when values of λ are large, b_s and b_r are nearly equal; as values of λ decrease, the difference between the two slopes becomes increasingly greater. As R^2 approaches one, the relative scatter becomes minimal, and the two slopes are nearly equal. For very low values of R^2, the deviation between b_s and b_r may be very large; that is, b_r is near zero, but the slope of the structural line may approach the vertical. Equation (6) shows that, assuming R^2 does not equal zero, the absolute value of the structural coefficient must exceed that of the regression coefficient.

The remainder of the paper by Mark and Church (1977) discusses the need for knowledge of error variances when using structural analysis and gives examples of procedures and problems when applying curve-fitting techniques to real geomorphic data. A basic difficulty when converting from a simple-regression to a structural relation is the estimation of λ (Mark and Church, 1977, p. 68-70). In the proper use of least-squares regression analysis, all error is in the dependent variable (y), which permits the calculation of residuals from predicted values of the dependent variable. Structural analysis, however, by accounting for the effect of errors in both variables, shows correspondence between two variables rather than being merely a predictive technique. Because errors in geomorphic measurements are often difficult to define, Mark and Church (1977, p. 68-70) provide only limited suggestions for the estimation of λ.

STRUCTURAL RELATIONS FROM ARTIFICIAL DATA

Procedure

The practical guidelines for the use of structural analysis that are advocated here result from comparisons of regression and structural analyses made from artificial data. The data were generated by digital computer to simulate previously determined gradient-discharge relations of alluvial stream channels (Lane, 1957; Osterkamp, in press) that are expressed as power-function relations. For convenience and uniformity with the other studies, logarithmic transformations were made on the data of this study. Modifying equation (5) into logarithmic form,

$$\log y = \log a + b \log x + \log e_y \ . \qquad (8)$$

Taking antilogs yields:

$$y = ax^b e_y \ , \qquad (9)$$

and the process of regression analysis considers the error component, providing predicted values for y_r:

$$y_r = ax^b \ . \qquad (10)$$

On logarithmic-coordinate paper, equation (10) plots as a straight line. The data generated, therefore, represent the logarithms of gradient and discharges, and the coefficient in the linear relation, equation (5), becomes the exponent or slope in the power-function relation, equation (10).

The synthetic data were generated in three sets of 600 data pairs. A standard exponent of -0.25 was used, but different error components were imposed on each set. The errors were generated randomly from normal distributions of mean value 0.0 and of specified standard deviation. The error standard deviations for the dependent ($\log_{10}$ of gradient) and independent variables ($\log_{10}$ of average discharge) for the three data sets were 0.154 and 0.09, 0.154 and 0.20, and 0.154 and 0.30. These deviations gave λ values of approximately 2.9, 0.59, and 0.26, respectively, for the three data sets.

Variation for the coefficient of determination, R^2, was obtained by selecting differing groups of 50 data pairs from each set. From a set of 600 data pairs ordered from low to high discharge (but subject to variation owing to the error component), groups of 50 data pairs were selected to provide variation in the range of data. Because the error components are applied in a consistent manner through any set, variation of the range of data alters the relation of range to error; hence, R^2 increases as the range of data increases. By imposing different values of λ on data sets and by selecting differing ranges of data from a set, both of the variables that influence the structural exponent (for a power-function equation), other than the regression exponent itself, can be selectively adjusted.

From each data set, the smallest ranges were obtained by selecting 50 consecutive data pairs. Larger ranges were acquired by using 50 consecutive even-numbered or odd-numbered data pairs. The greatest ranges resulted from the use of each twelfth data pair in a set. Because the calculation of a structural exponent from a regression relation (equation 6) is independent of the number of data pairs, analyses also were made for the entire 600 pairs of each data set.

Simple-regression analyses were made by digital computer on the various groups of 50 data pairs, as well as on each of the three sets of 600 data pairs. The program used for the regression analyses was BMD02R of the Biomedical Computer Programs, a series of programs that was developed by the University of California School of Medicine (Dixon, 1965). Output of the program includes the linear (or power-function) relation that results in the lowest possible error sum of squares, the standard error of estimate (SE_r), the correlation coefficient (R), the coefficient of determination (R^2), and residuals for the individual data pairs.

Assuming that a reasonable estimate can be made for λ, conversion from a regression to a structural relation can be made either manually or by computer. Manual computations are simple, although the calculation of the structural standard error of estimate is laborious. The structural exponent, b_s, is calculated by use of equation (6). The coefficient, a_s, for the structural equation is computed easily by applying b_s and mean values of the logarithms of dependent and independent variables to the equation:

$$\log_{10} a_s = (\log_{10} y)_m - b(\log_{10} x)_m \ , \qquad (11)$$

where $(\log_{10} y)_m$ and $(\log_{10} x)_m$ are the mean logarithms of the dependent and independent variables. Mean values of the two variables are provided as output when using BMD02R.

The standard error of estimate for a structural or regression analysis is a measure of deviations from a relation line that requires the recalculation of a dependent-variable value for each value of the independent variable used in the regression.

$$SE_s = [\Sigma(\log_{10} y - \log_{10} y_s)^2/(n - 2)]^{\frac{1}{2}} , \qquad (12)$$

where all values of the difference between $\log_{10} y$, an input or observed value of the dependent variable, and $\log_{10} y_s$, the corresponding value of the dependent variable calculated from the structural power-function relation and an individual input value of the independent variable, are squared and summed; n is the number of data pairs used in the analysis, generally 50 for the computer-generated data groups of this study. In all cases SE_s will exceed SE_r, although the differences may be very small.

Structural relations and the standard errors of estimate given in this paper were calculated by use of a computer program developed for that purpose. The program, which is depicted as a flow chart in figure 1, uses equations (6), (11), and (12) and requires the appropriate input values and regression-analysis results for those equations. In this study the conversion program was run independently, although the program could be incorporated directly into a computer program such as BMD02R.

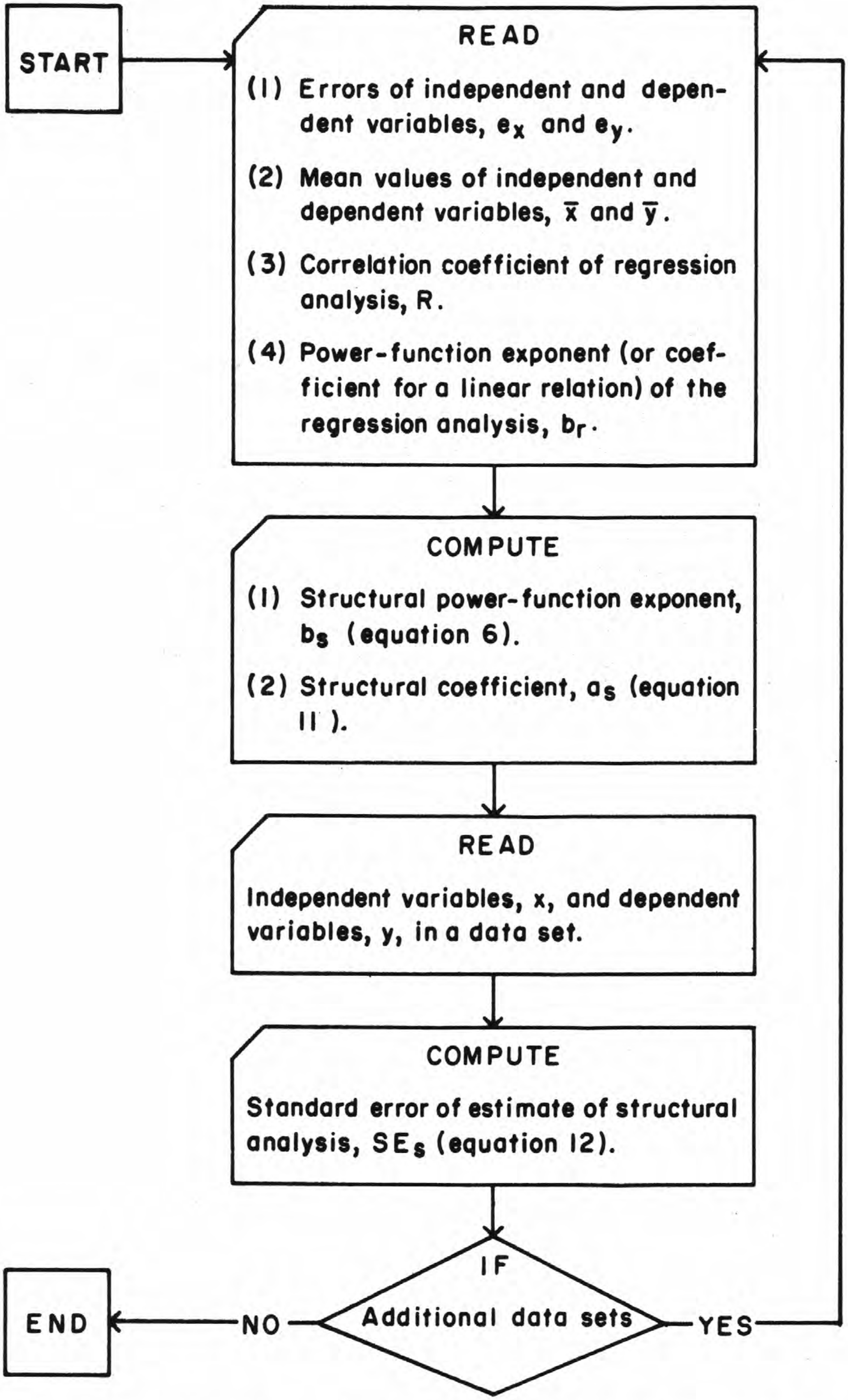
START
READ
(1) Errors of independent and dependent variables, e_x and e_y.
(2) Mean values of independent and dependent variables, $\overline{x}$ and $\overline{y}$.
(3) Correlation coefficient of regression analysis, R.
(4) Power-function exponent (or coefficient for a linear relation) of the regression analysis, b_r.
COMPUTE
(1) Structural power-function exponent, b_s (equation 6).
(2) Structural coefficient, a_s (equation 11).
READ
Independent variables, x, and dependent variables, y, in a data set.
COMPUTE
Standard error of estimate of structural analysis, SE_s (equation 12).
IF
Additional data sets
NO
YES
END

Figure 1.--Generalized flow chart for calculation of power-function exponent and standard error of a structural analysis.

Results

Differences between the regression and structural analyses for various groups and sets of artificial data are listed in table 1. Because an exponent largely defines a power-function expression, perhaps the most significant variation is the one that occurs between regression and structural exponents for the same data groups. If the data groups are of limited range and hence give a low coefficient of determination, the variation between the regression and structural exponents, expressed here as percent difference (table 1), can be very large. As the range of the data increases and the coefficient of determination increases, the percentage difference between the two exponents decreases. For example, the data for groups 2-4, 2-20, and 3-11 (table 1) are plotted, respectively, in figures 2, 3, and 4. The discharge-gradient values for group 2-4 (fig. 2) show almost no correlation ($R = 0.042$), which results in a very large difference between the two slopes or exponents. The data range and coefficient of determination, R^2, for group 2-20 (fig. 3) are large enough that the difference between the regression and structural exponents is only 9.04 percent (table 1). For group 3-11 (fig. 4), the difference in the two slopes appears relatively small, although it is 227 percent (table 1). If instead of being nearly horizontal the regression slope (exponent) were about 1.0, a 227-percent difference would appear highly significant. The percentage difference between the regression and structural exponents is, as previously indicated, dependent largely on the values of λ and the coefficient of determination, and partly on the value of the regression exponent itself. The angular difference between the two slopes, however, is greatest when the absolute value of the slope is 1.0 and approaches zero as the slope approaches either the horizontal or vertical.

The results listed in table 1 are generalized in figure 5. The solid lines show the manner in which percentage difference between the exponents decreases with increasing range of data and coefficient of determination for each of the three data sets. The zonations shown by dotted lines for the coefficients of determination (fig. 5) are not well defined but, nevertheless, can provide an estimated difference between the two exponents if λ and the coefficient of determination for the regression analysis are known. The positions of the three curves, relating percent difference between the two exponents to the logarithm of the range of the independent variable (average discharge), are well defined. They show, for example, that, if λ is less than about 0.25, a large range of data (in this case at least three and a half orders of magnitude) will be required to have no less than a 10-percent difference between the exponents. If λ exceeds 3, less than 1-percent difference between the regression and structural exponents can be expected for a similar range of data. Depending on the accuracy requirements of a study, therefore, figure 5 can be used to show when conversion to a structural relation might be warranted.

Table 1.--Regression- and structural-analysis results of the artificial data.

Group	Range $\bar{Q}$ (m^3/s)		Range $\bar{Q}$ Log units	b_r	b_s	a_r $(\times 10^5)$	a_s $(\times 10^5)$	SE_r	SE_s	S_d	R	R^2	$\frac{b_s - b_r}{b_r}$ $(\times 100)$
Set 1 ($\lambda = 0.154^2/0.09^2 \simeq 2.9$)													
1-1	0.0166	0.0494	0.475	0.05805	0.15863	271	385	0.1527	0.1531	0.1513	0.042	0.002	173
1-2	.0377	.130	.538	.13470	.25878	266	371	.1518	.1526	.1512	.112	.013	92.1
1-3	.0912	.307	.527	.08166	.11431	181	192	.1290	.1291	.1282	.089	.008	40.0
1-4	.191	.933	.689	-.11698	-.17039	112	107	.1631	.1634	.1626	.121	.015	45.7
1-5	.486	1.94	.601	-.10372	-.23284	96	95	.1864	.1873	.1850	.081	.007	124
1-6	1.07	3.92	.564	-.33263	-.43903	102	110	.1263	.1273	.1340	.360	.129	32.0
1-7	2.70	7.86	.547	-.17824	-.28291	93	110	.1509	.1516	.1515	.167	.028	58.7
1-8	5.78	25.2	.639	-.27661	-.35980	108	132	.1358	.1364	.1415	.314	.099	30.1
1-9	11.2	46.6	.618	-.19000	-.29726	80	113	.1382	.1389	.1390	.180	.032	56.5
1-10	33.1	133	.604	-.17812	-.26366	78	111	.1504	.1510	.1513	.178	.032	48.0
1-11	63.2	314	.696	-.26405	-.35413	100	156	.1298	.1305	.1342	.288	.083	34.1
1-12	53	780	.707	-.06015	-.09010	34	41	.1441	.1442	.1429	.061	.004	49.8
1-13	0.0166	0.120	.862	-.09625	-.11916	155	144	.1723	.1724	.1719	.127	.016	23.8
1-14	.0912	.933	1.01	-.22567	-.25468	102	98	.1558	.1560	.1653	.361	.130	12.9
1-15	.486	3.92	0.907	-.19355	-.21910	94	94	.1366	.1368	.1424	.312	.098	13.2
1-16	2.53	20.9	.918	-.24721	-.28443	100	108	.1526	.1529	.1624	.367	.135	15.1
1-17	10.2	111	.996	-.25690	-.30235	103	122	.1601	.1604	.1696	.357	.127	17.7
1-18	63.2	764	1.08	-.15033	-.16768	175	63	.1305	.1306	.1339	.262	.069	11.5
1-19	0.017	0.780	1.66	-.20423	-.21401	109	106	.1729	.1730	.1957	.485	.235	4.79
1-20	.492	20.9	1.63	-.25302	-.26180	105	106	.1436	.1437	.1819	.624	.389	3.47
1-21	11.2	428	1.58	-.26880	-.28069	107	113	.1577	.1578	.1954	.602	.362	4.42
1-22	0.017	3.01	2.24	-.26945	-.27629	90	90	.1774	.1774	.2469	.703	.494	2.54
1-23	2.54	540	2.33	-.23043	-.23226	101	102	.1003	.1003	.1793	.833	.693	0.794
1-24	0.021	13.5	2.79	-.24765	-.25012	89	89	.1469	.1469	.2548	.821	.675	.997
1-25	.027	96.8	3.55	-.25198	-.25402	94	94	.1631	.1631	.3082	.852	.726	.810
1-26	.0245	685	4.45	-.24657	-.24787	90	100	.1630	.1630	.3562	.892	.796	.527
Set 1	.0166	780	4.66	-.24210	-.24328	98	98	.1477	.1477	.3422	.902	.814	.487

Table 1.--Regression- and structural-analysis results of the artificial data--continued.

Group	Range $\bar{Q}$ (m^3/s)		Log units	b_r	b_s	a_r ($\times 10^5$)	a_s ($\times 10^5$)	SE_r	SE_s	S_d	R	R^2	$\frac{b_s - b_r}{b_r}$ ($\times 100$)
					Set 2 ($\lambda = 0.154^2/0.02^2 \approx 0.59$)								
2-1	0.00874	0.104	1.08	-.10987	-.40448	167	61	0.1661	0.1809	0.1665	0.159	0.025	179
2-2	.0535	.552	0.983	-.05966	-.22532	151	112	.1667	.1718	.1656	.089	.008	278
2-3	.114	1.15	1.01	-.06421	-.15552	119	109	.1458	.1475	.1452	.107	.012	142
2-4	.366	3.71	1.01	.03371	5.2674	108	146	.1873	1.250	.1855	.042	.002	15,500
2-5	.564	4.77	0.927	-.18131	-0.41668	91	108	.1286	0.1377	.1327	.283	.080	130
2-6	1.44	17.5	1.08	-.04326	-.18550	72	90	.1643	.1678	.1629	.064	.004	329
2-7	3.86	35.2	0.960	-.13235	-1.1400	79	957	.1719	.2743	.1424	.161	.026	761
2-8	8.28	95.9	1.06	-.02774	-0.10874	53	70	.1482	.1493	.1468	.042	.002	292
2-9	58.5	834	1.15	-.23776	-.56825	82	471	.1587	.1780	.1672	.343	.118	139
2-10	67.4	1180	1.24	-.13419	-.44360	54	325	.1609	.1771	.1624	.195	.038	231
2-11	0.0120	0.249	1.32	-.22113	-.46608	109	51	.1689	.1822	.1780	.343	.118	111
2-12	.0535	.770	1.16	-.18729	-.34257	117	94	.1412	.1468	.1477	.325	.105	82.9
2-13	.366	4.29	1.07	-.23278	-.39668	103	109	.1553	.1625	.1678	.401	.161	70.4
2-14	1.44	24.4	1.23	-.12702	-.23180	78	97	.1517	.1547	.1544	.233	.054	82.5
2-15	8.28	117	1.33	-.04447	-.06441	49	53	.1326	.1327	.1320	.103	.011	44.8
2-16	58.5	1180	1.31	-.20185	-.35128	73	163	.1607	.1669	.1701	.355	.126	74.0
2-17	0.0161	0.774	1.68	-.29619	-.35667	94	82	.1530	.1553	.2000	.652	.425	20.4
2-18	.555	24.4	1.64	-.21646	-.26299	96	101	.1533	.1547	.1800	.538	.290	21.5
2-19	10.5	572	1.74	-.28388	-.33211	118	147	.1466	.1483	.1950	.668	.447	17.0
2-20	0.0221	4.45	2.30	-.24774	-.27014	97	94	.1564	.1571	.2254	.727	.529	9.04
2-21	2.23	617	2.44	-.25345	-.28807	104	119	.1843	.1857	.2449	.667	.445	13.7
2-22	0.0138	13.2	2.98	-.24137	-.25569	106	105	.1668	.1673	.2656	.784	.614	5.93
2-23	.0186	884	3.92	-.24279	-.24927	95	95	.1428	.1430	.2989	.881	.776	2.67
2-24	.0087	422	4.68	-.21282	-.21789	101	101	.1635	.1637	.3255	.868	.753	2.38
Set 2	.00874	1180	5.13	-.24477	-.25020	102	103	.1618	.1619	.3563	.891	.794	2.22

Table 1.--Regression- and structural-analysis results of the artificial data--concluded.

Group	Range $\overline{Q}$												
	(m^3/s)		Log units	b_r	b_s	a_r ($\times10^5$)	a_s ($\times10^5$)	SE_r	SE_s	S_d	R	R^2	$\frac{b_s - b_r}{b_r}$ ($\times100$)
					Set 1 ($\lambda = 0.154^2/0.09^2 \simeq 2.9$)								
3-1	.00681	0.211	1.49	-.01703	-.04045	223	206	.1356	.1358	.1343	.044	.002	138
3-2	.0105	.455	1.64	-.07688	-.31355	171	91	.1595	.1789	.1600	.163	.026	308
3-3	.0227	.753	1.52	.05376	.19911	170	224	.1565	.1645	.1560	.119	.014	270
3-4	0.418	1.64	1.60	-.15186	-.65759	109	68	.1727	.2390	.1778	.276	.076	333
3-5	.133	4.85	1.56	.06191	.12877	108	94	.1365	.1387	.1369	.163	.026	108
3-6	.655	12.2	1.27	.05515	.56718	84	52	.1676	.2362	.1669	.106	.011	928
3-7	1.32	17.5	1.12	.11898	.91749	55	17	.1766	.3042	.1786	.204	.042	671
3-8	2.26	37.9	1.27	.09730	.72809	43	9	.1658	.2554	.1667	.178	.032	648
3-9	6.13	384	1.80	-.10246	-.40818	56	156	.1692	.2008	.1713	.210	.044	298
3-10	11.2	245	1.34	-.09993	-.44223	60	232	.1576	.1925	.1593	.201	.040	343
3-11	14.9	713	1.68	-.04533	-.14818	34	56	.1402	.1441	.1396	.104	.011	227
3-12	0.00878	0.212	1.38	-.05610	-.09591	190	168	.1266	.1275	.1271	.166	.027	71.0
3-13	.0226	1.02	1.65	-.07079	-.18185	127	108	.1616	.1674	.1623	.170	.029	157
3-14	.261	6.68	1.41	.05799	.15204	89	86	.1470	.1508	.1470	.140	.020	162
3-15	1.32	41.8	1.50	-.01250	-.13242	60	77	.1908	.1964	.1889	.026	.001	959
3-16	6.13	384	1.80	-.06443	-.32354	51	127	.1757	.2000	.1755	.134	.018	402
3-17	14.9	2020	2.13	-.07117	-.13548	39	54	.1521	.1545	.1535	.194	.038	90.4
3-18	0.0102	1.02	2.00	-.15627	-.21167	117	102	.1467	.1500	.1665	.489	.239	35.5
3-19	.473	37.9	1.90	-.14863	-.23510	92	102	.1636	.1691	.1775	.409	.168	58.2
3-20	11.1	1070	1.98	-.18733	-.26319	76	106	.1631	.1682	.1902	.529	.280	40.5
3-21	0.0101	7.25	2.86	-.23023	-.31891	100	89	.1986	.2068	.2459	.601	.361	38.5
3-22	1.57	515	2.52	-.15389	-.18508	68	76	.1698	.1715	.2045	.570	.325	20.3
3-23	0.00878	41.8	3.68	-.22889	-.27015	106	104	.2117	.2153	.3003	.716	.531	18.0
3-24	.0147	525	4.55	-.25394	-.26942	101	103	.1853	.1865	.3839	.878	.772	6.10
Set 3	**.00681**	**2020**	**5.47**	**-.23472**	**-.24583**	**98**	**99**	**.1779**	**.1785**	**.3546**	**.865**	**.749**	**4.73**

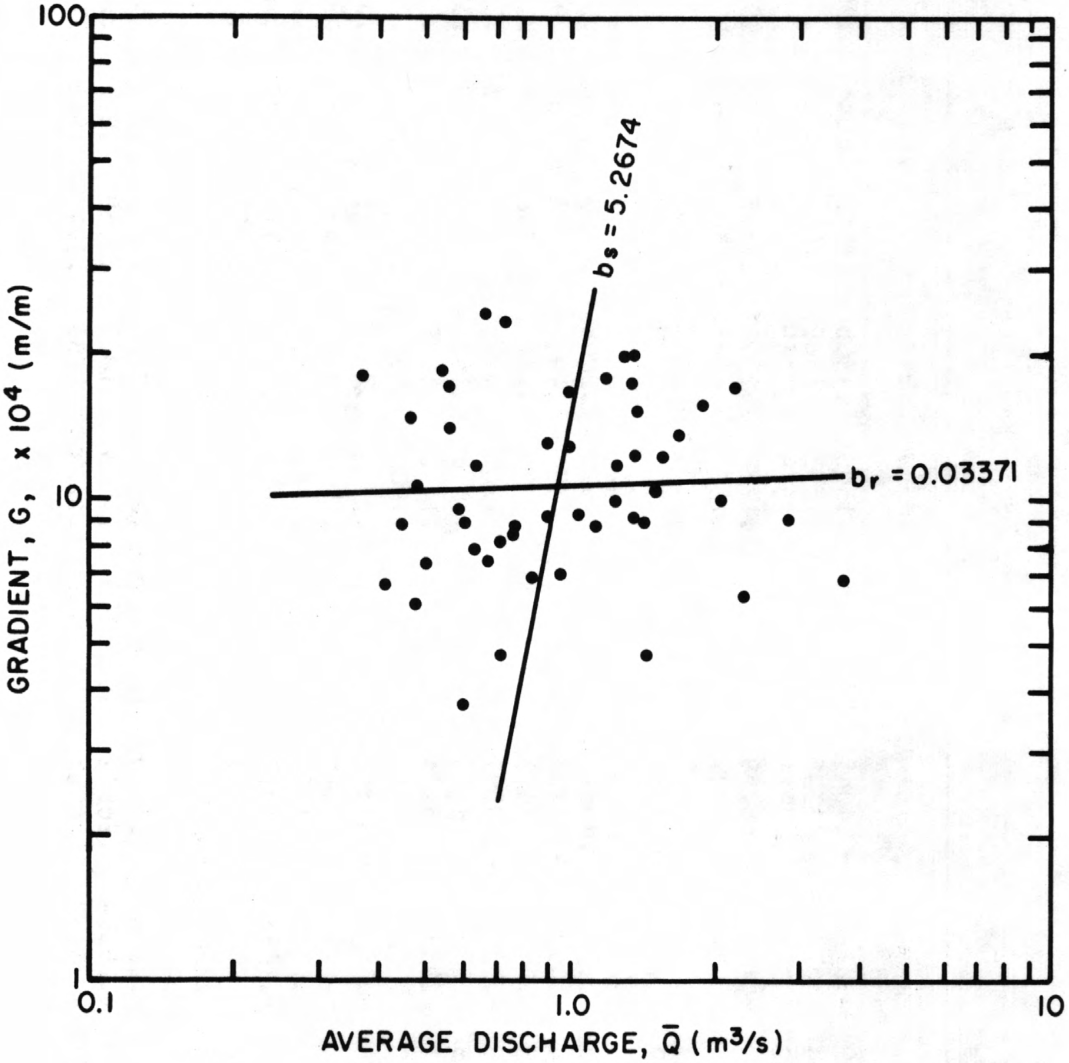

Figure 2.--Gradient-discharge data and linear slopes for group 2-4 (table 1).

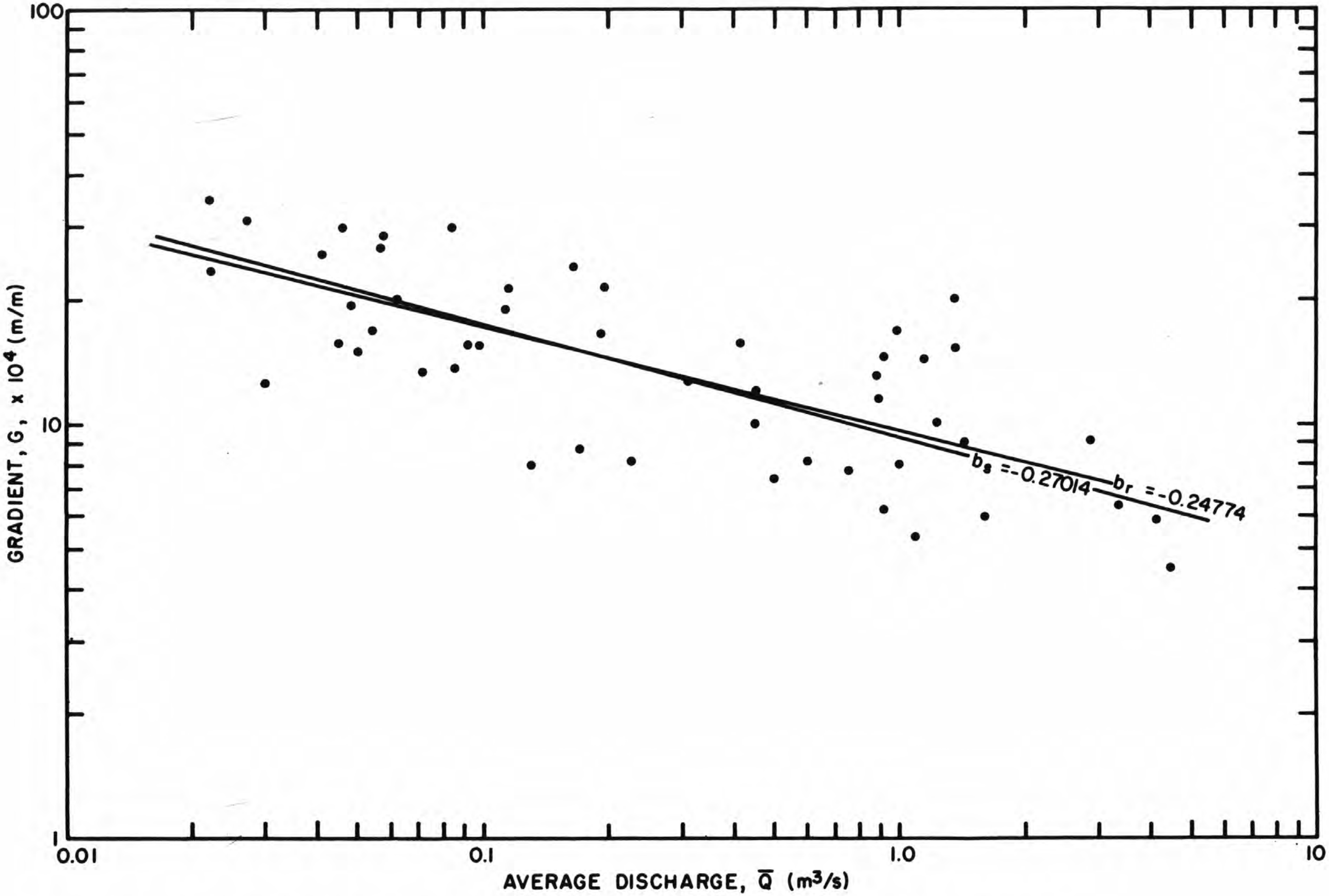

Figure 3.--Gradient-discharge data and linear slopes for group 2-20 (table 1).

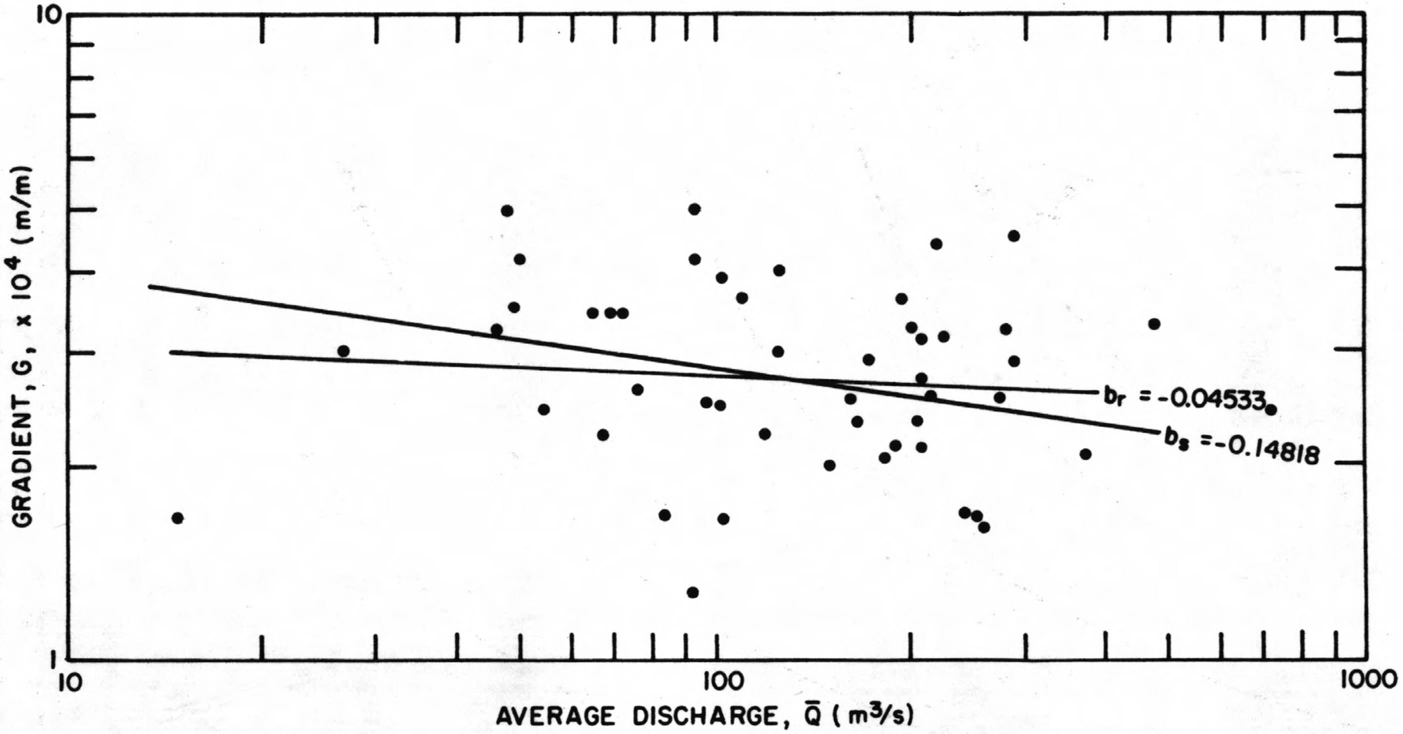

Figure 4.--Gradient-discharge data and linear slopes for group 3-11 (table 1).

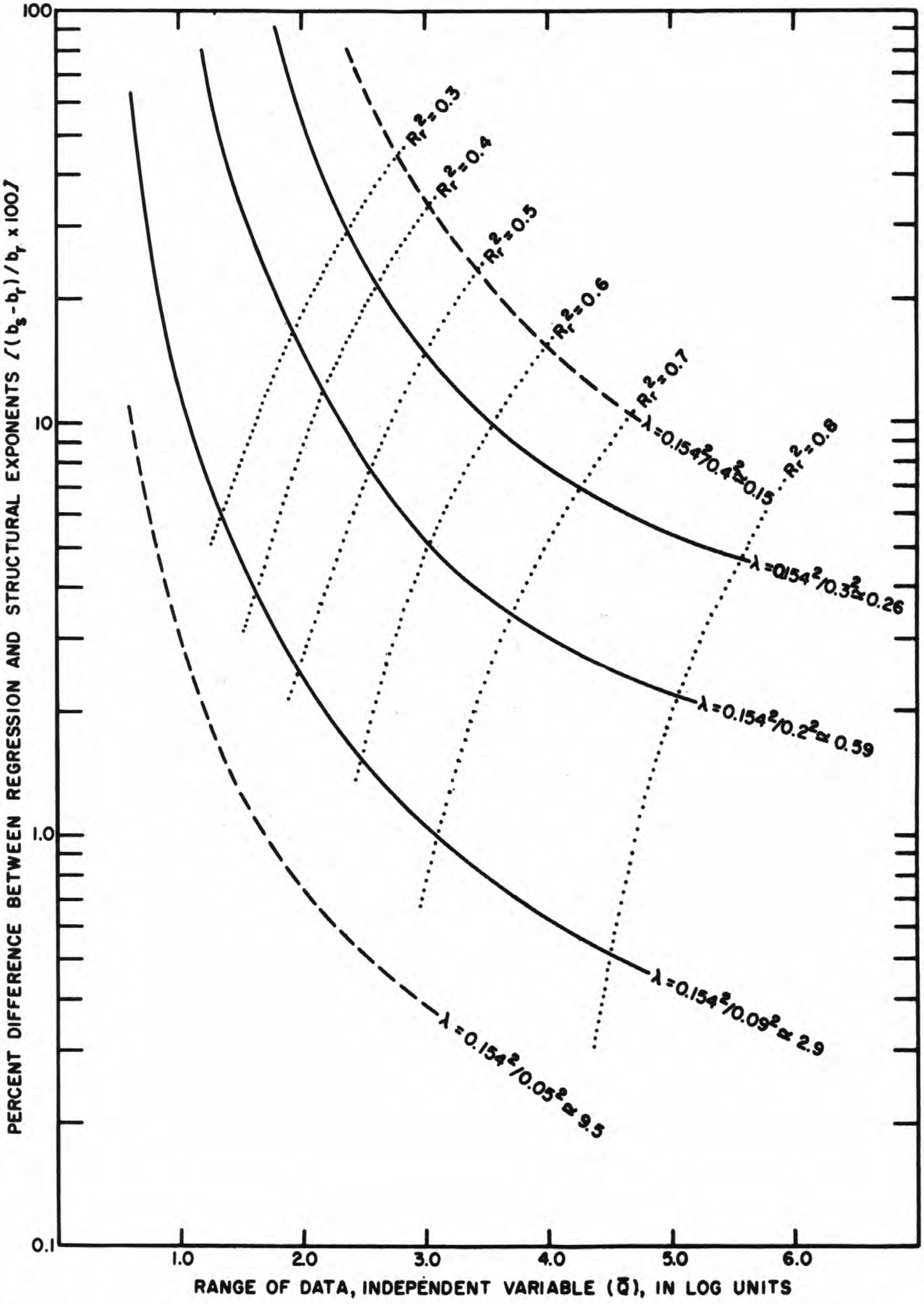

Figure 5.--Variation of percent difference between regression and structural exponents with range of data of the independent variable for different values of λ.

The three values of λ used for structural calculations (table 1) were varied by changing the imposed error components of the independent variables only. In this manner, for two different λ values, ratios between the error component and the data range of the independent variable (both expressed as logarithms) can be compared with the associated percentage differences of the two exponents. Therefore, if the error component of the dependent variable does not change for various values of λ, as is the case for the data represented by table 1, then:

$$\left[\left(\frac{b_s - b_r}{b_r}\right)100\right]_{p_1} \approx \left[\left(\frac{b_s - b_r}{b_r}\right)100\right]_{p_2} , \qquad (13)$$

where p is the ratio between the error component and the range of data (both expressed as logarithms) for the independent variable, and p_1 and p_2, being the ratios for two different values of λ, are selected to be equal. As an example (fig. 5), for λ values of 2.9 and 0.59, the error components for $\log_{10} \bar{Q}$ are 0.09 and 0.2, respectively. For data ranges chosen at 0.9 and 2.0 log cycles, respectively, p_1 and p_2 both equal 0.1, and the corresponding difference between the two exponents in both cases is 15 percent. By using equation (13), a curve relating the range of data to the difference between the regression and structural exponents can be generated for any value of λ. Two such curves, shown as dashed lines, are given in figure 5 for λ values of 9.5 (error component of 0.05 $\log_{10}$ units) and 0.15 (error component of 0.4 $\log_{10}$ units).

When developing a power-function relation between two variables, if it is concluded that a structural analysis is appropriate (that is, significant error is associated with both variables and can be reasonably estimated for both), it would seem preferable to convert to a structural relation in all cases when the necessary computer capability is available. Figure 5 is of utility, however, if computations for the conversion must be made manually, particularly if the standard error of estimate is included. Assuming that reasonable values for λ, R^2, and the range of data are available, the difference between the regression and structural exponents can be estimated quickly by use of figure 5. If that difference is less than the accuracy requirements of the study, it may be impractical to make the conversion to a structural relation.

The standard error of estimate is a measure of the deviation or scatter of the dependent variable about a linear relation and is minimized in regression analysis. For a structural analysis, therefore, the standard error of estimate must be greater than that of the corresponding regression analysis (table 1). The results of this empirical study, however, indicate that the change is generally small, except when the coefficient of determination is very small. Specifically, the standard error of estimate measures the deviations of values of the dependent variable from the regression relation (e.g. in the y direction). Thus, the standard error of estimate, which is provided as output when using BMD02R (Dixon, 1965), can serve as a good estimate of the error component for the dependent variable. The value of 0.154, which is the imposed error component of $\log_{10}$ gradient for the three computer-generated data sets, was the standard error of estimate of a representative regression analysis of actual gradient-discharge data.

The coefficient of determination, R^2, may be calculated from the standard error of estimate (SE) and the standard deviation of the dependent variable (S_d):

$$R^2 = [1-(\frac{SE_r}{S_d})^2] [\frac{n-2}{n-1}] . \tag{14}$$

Both SE and S_d are subject to variation (table 1) depending on the data comprising a group, and the resulting R^2 can show considerable variation for data groups with similar range and error components. The contours given in figure 5, therefore, are approximations or indications of the manner in which the coefficient of determination changes with change of data range and λ. More importantly, however, it should be recognized that, by imposing a constant gradient error component of 0.154 on the artificial data, values not varying greatly from 0.154 also are imposed for the standard errors of estimate of the regressions. Because R^2 is largely a function of SE and S_d, holding the standard error to a limited range makes R^2 mostly dependent on S_d, and the contours for R^2 (fig. 5) must be considered specific for a range of standard errors not deviating greatly from 0.154.

GRADIENT-DISCHARGE RELATIONS OF KANSAS STREAMS

The analyses for the computer-generated data groups have imposed values of λ that are known and consistent. In practice, however, the error components must be estimated; they rarely can be assumed to be accurate. It is the consistency of the artificial data and the imposed error components and slope of the linear relation that result in the relatively well-defined curves of figure 5. The computer data were generated to be analogous to actual gradient-discharge relations because field data and the resulting regression and structural analyses are available for comparison. Table 2 lists regression- and structural-analysis results for gradient and discharge data collected from 76 gaging-station sites on perennial streams in Kansas (Osterkamp, in press). The 76 data pairs (values for channel gradient and average discharge) were separated into five groups according to channel sinuosity and into four groups based on variations of the sediment characteristics of the bed material. Thus, analyses were made for nine data groups, plus a tenth which included all 76 of the data pairs. In all cases λ was calculated using SE_r to approximate the error component of gradient and 0.09 for the error component of mean discharge. Most of the 76 gaging stations used in the study have been in operation in excess of 20 years, which corresponds to a standard error of about 0.09 ($\log_{10}$ units) for mean flow of Kansas streams (Jordan and Hedman, 1970, p. 16). The regression standard error of group 2D (table 2) was considered representative, and therefore 0.154 was used for the imposed error component of gradient in all of the structural analyses of table 1.

Table 2.--Summary of gradient-discharge relations for streams of Kansas.

Data group	Data pairs	Range $\overline{Q}$ (log units)	b_r	b_s	SE_r	SE_s	R^2	$(b_s-b_r)/b_r$ (x100)
1A	17	2.68	-0.35161	-0.35829	0.1526	0.1527	0.689	1.90
1B	14	2.49	-.30328	-.30797	.1080	.1081	.797	1.55
1C	17	2.60	-.17723	-.18154	.2017	.2017	.208	2.43
1D	18	2.36	-.23398	-.23736	.1404	.1404	.607	1.44
1E	10	2.20	-.30334	-.30789	.0865	.0865	.859	1.50
2A	17	1.93	-.23650	-.24492	.1475	.1475	.372	3.56
2B	13	2.03	-.31149	-.31662	.1338	.1338	.721	1.65
2C	23	3.77	-.25611	-.25863	.1702	.1702	.648	0.984
2D	23	2.72	-.22747	-.23156	.1536	.1537	.496	1.80
3	76	3.77	-.23634	-.24023	.2063	.2063	.394	1.65

Differences between the regression and structural exponents, with values of R^2, for the Kansas streams are plotted in figure 6, which otherwise is identical to part of figure 5. Although the Kansas data were the basis for generating the artificial data, figure 6 shows inconsistencies occurring because results for the artificial and real data are not completely comparable. The imposed and estimated values of λ probably are closely similar, but the Kansas results represent dynamic hydrologic and geomorphic systems that are influenced by other variables. Measurement errors for gradient and average discharge might be estimated with reasonable accuracy, but a gradient-discharge relation commonly is affected by unknown (natural) variables, resulting in the introduction of stochastic errors. Thus, the structural relation probably will be based on incorrect values of λ . Because most of the points indicate a λ value less than 2.9 (fig. 6), it appears reasonable that the gradient-discharge relations are being influenced by unexplained causes. Owing to similar reasons and to the fact that R^2 zones of figures 5 and 6 are based specifically on gradient error components of 0.154, the plotted values of R^2 (fig. 6) are not directly comparable.

Figure 6 does show that, if even crude estimates of λ can be made, the amount of variation between regression and structural relations can be anticipated. This empirical observation is consistent with the statement of Mark and Church (1977, p. 70) that the relation between b_r and b_s "is not a rapidly changing function of λ, hence approximations of λ may be reasonable."

DISCUSSION AND CONCLUSIONS

Simple-regression and structural analysis are two similar statistical methods of developing a linear power-function relation from a bivariant group of data. Choice of the proper technique should depend on whether error is associated with one or both of the variables. Mark and Church (1977), in noting that regression analysis often is misapplied to geomorphic studies, discussed the theory of structural analysis and when it should be used. The discussion presented here identifies the extent of error that can be expected when regression is applied improperly and gives specific suggestions for the implementation of structural analysis.

Mark and Church (1977) point out that structural analysis is the proper method of determining a linear relation for most geomorphic studies because few geomorphic (or hydrologic) variables can be measured without error. To convert a regression relation to a structural relation, the relative error components (λ) for the two variables must be considered. The independent variable is subject to measurement error; whereas, the dependent variable is subject to both measurement and stochastic error. For the computer-generated data, both types of error were imposed, and λ was known accurately. In dynamic systems, however, error components only can be estimated, particularly for small groups of data. Because of the manner in which the error components for both the independent and dependent variables were estimated, the values of λ for the structural analyses of the Kansas data are necessarily inaccurate. The inaccuracies are inferred to be the cause of the scatter for the Kansas data in figure 6.

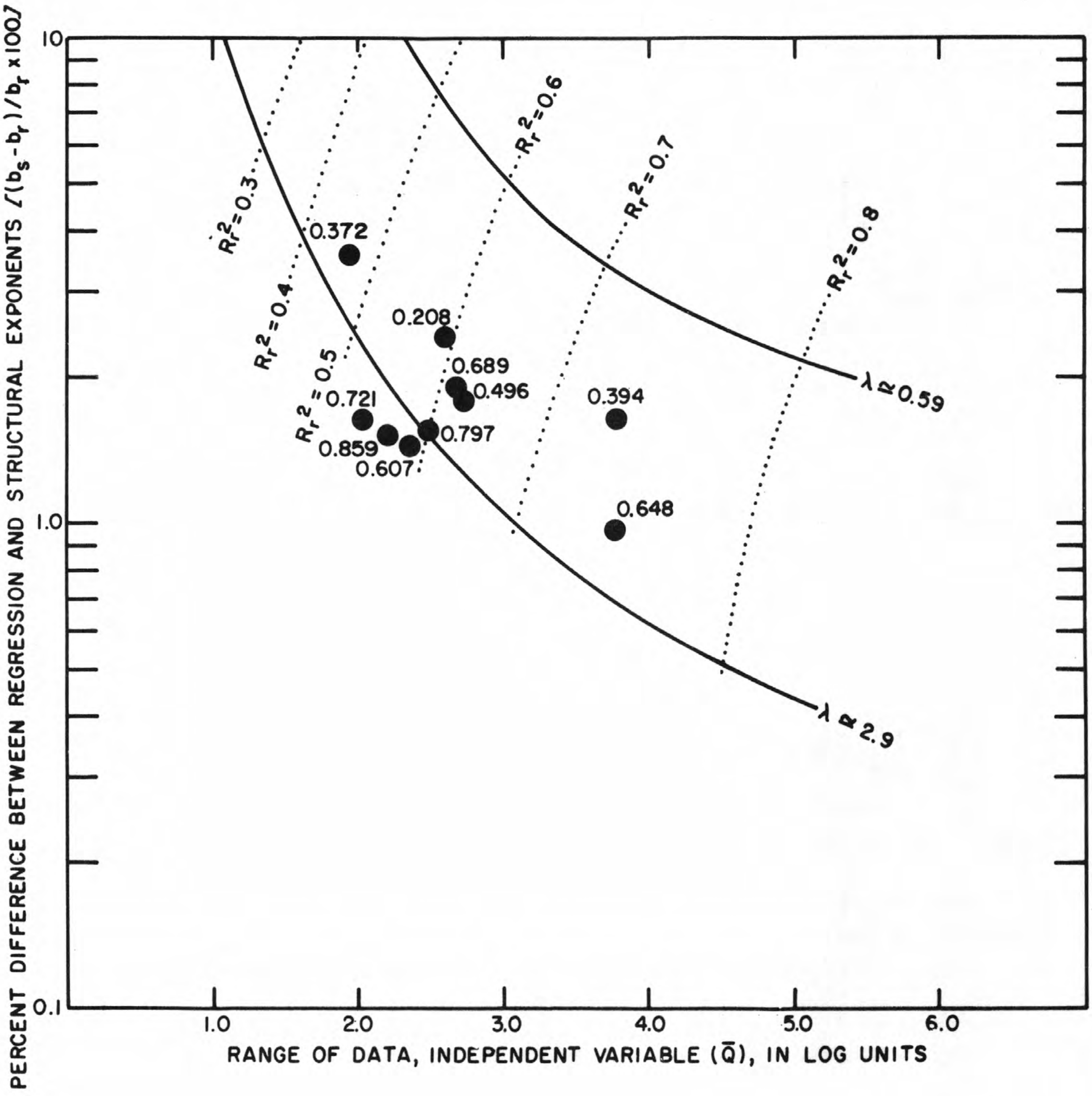

Figure 6.--Comparison of results for Kansas data with the computer-generated regression and structural analyses.

An advantage of the information provided by figure 5 is that the abscissa is the range of data for the independent variable. The graph can be used before regression results are available. If λ can be estimated reasonably, figure 5 permits an assessment of the probable difference that would occur if both regression and structural analyses were made. Use of an analogous graph by Mark and Church (1977, p. 70, fig. 2) requires values for the regression exponent and coefficient of determination, as well as an estimate of λ. This graph was developed by assuming values for a modified form of equation (6) and, therefore, gives accurate relations. Figure 5 presented here was derived empirically and does not give precise relations, but does provide a general utility, regardless of the R^2 values.

Results of this study suggest that conversion to a structural relation is justified except when R^2 is very small or large. When there is very little correlation between two variables, a curve-fitting technique has little meaning. As shown by figure 2 and table I, the difference in slope for the two techniques can be very large. In addition, the regression exponent may be of opposite sign from that which is correct when the line of relation approaches horizontal and conversion to a structural relation compounds the error (table 1, fig. 2). If R^2 approaches 1, the difference between the regression and structural exponents is very small. Depending on the accuracy required for a study, conversion may serve the sole purpose of making a linear relation theoretically unbiased.

Confirming the work of Mark and Church (1977), the computer-generated data of this study show that a structural exponent is not highly sensitive to changes in λ. That is not to minimize, however, the importance of selecting reasonable values for the error components when making a structural analysis. After regression analysis, a standard error of estimate generally is available. Estimates for error of the independent variable occasionally are available in the literature, as is the case for mean discharge in this study. More often, however, it would seem necessary to determine how the values of the independent variable were determined, which should lead to a reasonable estimate of the measurement error.

REFERENCES

Brooks, C., Hart, S. R., and Wendt, I., 1972, Realistic use of two-error regression treatments as applied to rubidium-strontium data: Review of Geophysics and Space Physics, v. 10, no. 2, p. 551-577.

Cox, N. J., 1977, Allometric change of landforms--Discussion: Geological Society America Bulletin, v. 88, p. 1199-1202.

Dixon, W. J., ed., 1965, BMD, Biomedical computer programs: University of California School of Medicine, Los Angeles, 620 p.

Jordan, P. R., and Hedman, E. R., 1970, Evaluation of the surface-water data program in Kansas: Kansas Water Resources Board Bulletin 12, 49 p.

Lane, E. W., 1957, A study of the shape of channels formed by natural streams flowing in erodible material: U.S. Army Engineer Division, Missouri River, M.R.D. Sediment Series No. 9, 106 p.

Mark, D. M., and Church, Michael, 1977, On the misuse of regression in earth science: Mathematical Geology, v. 9, p. 63-75.

Miller, R. L., and Kahn, J. S., 1962, Statistical analysis in the geological sciences: John Wiley, New York, 483 p.

Osterkamp, W. R., Gradient, discharge, and particle-size relations of alluvial channels in Kansas, with observations on braiding: American Journal of Science (in press).

Snedecor, G. W., and Cochran, W. G., 1967, Statistical methods (6th ed.): The Iowa State University Press, 593 p.

Till, R., 1973, The use of linear regression in geomorphology: Area, v. 5, no. 4, p. 303-308.

★U.S. GOVERNMENT PRINTING OFFICE: 1978--666453/240