United States Environmental Protection Agency | Office Of Water (4503F) | EPA-841-B-97-007 June 1997

Linear Regression for Nonpoint Source Pollution Analyses

INTRODUCTION

The purpose of this fact sheet is to demonstrate an approach for describing the relationship between variables using regression. The fact sheet is targeted toward persons in state water quality monitoring agencies who are responsible for nonpoint source assessments and implementation of watershed management.

Regression can be used to model or predict the behavior of one or more variables. The general regression model, where ϵ is an error term, is given as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon \quad (1)$$

In this equation, the behavior of a single dependent variable (y) is modeled with one or more independent variables ($x_1, \ldots, x_n$). The x's may be linear or nonlinear (e.g., x_i can represent x^2, x^3, x^{-1}, etc.). $\beta_0, \ldots, \beta_n$ are numerical constants that are computed using equations described later. Nonlinear models are commonly applied to physical systems, but they are somewhat more difficult to analyze because iterative techniques are involved when the model cannot be transformed to a linear model. The use of two or more independent variables (x) in a linear function to describe the behavior of y is referred to as multiple linear regression. In either case, regression techniques attempt to explain as much of the variation in the dependent variable as possible.

In nonpoint source analyses, linear regression is often used to determine the extent to which the value of a water quality variable (y) is influenced by land use or hydrologic factors (x) such as crop type, soil type, percentage of land treatment, rainfall, or stream flow, or by another water quality variable. Practical applications of these regression results include the ability to predict the water quality impacts due to changes in the independent variables.

SIMPLE LINEAR REGRESSION

The simplest form of regression is to consider one dependent and one independent variable using

$$y = \beta_0 + \beta_1 x + \epsilon \quad (2)$$

where y is the dependent variable, x is the independent variable, and β_0 and β_1 are numerical constants representing the y-intercept and slope, respectively. Helsel and Hirsch (1995) summarize the key assumptions regarding application of linear regression (Table 1). The uses of a regression analysis should not be extended beyond those supported by the assumptions met. Note that the normality assumption (assumption 5) can be relaxed when testing hypotheses and estimating confidence intervals if the sample size is relatively large.

The first step in applying linear regression (assumption 1 in Table 1) is to examine the data to see if linear regression makes sense—that is, to use a bivariate scatter plot to see if the points approximate a straight line. If they fall in a straight line, linear regression makes sense; if they do not, a data transformation might be needed, or perhaps a nonlinear relationship should be used.

To illustrate the use of linear regression, the fraction of water (split) collected by a water and sediment sampler for a plot-sized runoff sampler is used (Dressing et al., 1987). In this data set the sampling percentage (split) was measured for a range of flow rates. The scatter plot (Figure 1) shows that linear regression can be applied.

Presuming that the data are representative (assumption 2 in Table 1), the next step is to develop the regression line using the method of least squares (Freund, 1973). To determine the values of β_0 and β_1 in Equation 2, the following equations can be used (Helsel and Hirsch, 1995):

$$\beta_1 = \frac{S_{xy}}{SS_x} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2} \quad (3)$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (4)$$

where n, $\bar{x}$, and $\bar{y}$ are the number of observations, the mean of the independent variable (e.g., flow rate), and the mean of the dependent variable (e.g., split), respectively. S_{xy} is the sum of the xy cross products and SS_x is the sum of the squares x.

Table 1. Assumptions necessary for the purposes of linear regression.

Assumption	Purpose			
	Predict y given x	Predict y and a variance for the prediction	Obtain best linear unbiased estimator of y	Test hypotheses, estimate confidence or prediction intervals
(1) The model form is correct: y is linearly related to x	✓	✓	✓	✓
(2) The data used to fit the model are representative of data of interest	✓	✓	✓	✓
(3) The variance of the residuals is constant and does not depend on x or anything else		✓	✓	✓
(4) The residuals are independent			✓	✓
(5) The residuals are normally distributed				✓
✓ Indicates that assumption is required.				

Reprinted from Helsel and Hirsch, *Statistical Methods in Water Resources*, 1995, page 225, with kind permission from Elsevier Science - NL, Sara Burgerhartstraat 25, 1055 KV Amsterdam, The Netherlands.

For the data in the first two columns of Table 2 (same as those displayed in Figure 1), Equations 3 and 4 were used to compute a slope (β_1) of -0.0119 and an intercept (β_0) of 3.1317. ($\bar{x}$, $\bar{y}$, S_{xy}, and SS_x were computed as 28.89, 2.79, -40.8175, and 3423.7373, respectively.) Thus, the linear model for predicting split versus flow rate is

$$Split = 3.1317 - 0.0119 \cdot Flow\ rate \qquad (5)$$

ASSUMPTION EVALUATION

The analyst must make sure that β_0 and β_1 make sense. In this case, perhaps the best approach is to plot the regression line with the raw data, as shown in Figure 1. The third column in Table 2 contains the predicted split, $\hat{y}_i$, computed using Equation 5 for each flow rate. The

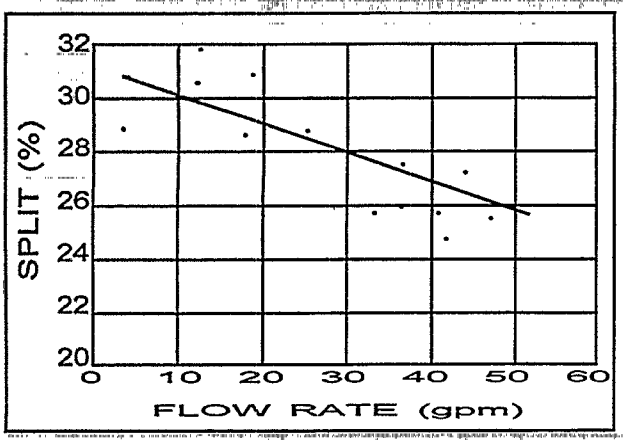

Figure 1. Split versus flow rate.

Table 2. Runoff sampler calibration data.

Flow Rate, x_i (gpm)	Split, y_i (%)	Predicted Split, $\hat{y}_i$	Residual $e_i = y_i - \hat{y}_i$
52.1	2.65	2.5106	0.1394
19.2	3.12	2.9028	0.2172
4.8	3.05	3.0745	-0.0245
4.9	2.86	3.0733	-0.2133
35.2	2.72	2.7121	0.0079
44.4	2.70	2.6024	0.0976
13.2	3.04	2.9743	0.0657
25.8	2.83	2.8241	0.0059
17.6	2.84	2.9219	-0.0819
37.6	2.60	2.6835	-0.0835
41.4	2.54	2.6382	-0.0982
40.1	2.58	2.6536	-0.0736
47.4	2.49	2.5666	-0.0766
35.7	2.60	2.7061	-0.1061
13.9	3.19	2.9660	0.2240

predicted split, $\hat{y}_i$, is plotted as the regression line in Figure 1. By visual inspection, β_0 and β_1 seem reasonable.

Residuals plotted as a function of predicted values of y, residuals plotted as a function of time, and normal probability plots of residuals are the most effective approaches to evaluate the last three assumptions listed in Table 1, respectively. The fourth column of Table 2 presents the residuals, e_i, which are computed as the observed split minus the predicted split ($y_i - \hat{y}_i$).

The plot of residuals should appear to be a uniform band of points around 0, as shown in Case A of Figure 2 (Ponce, 1980). In Figure 2, residuals are plotted as a function of predicted values of y. The analyst should look for two types of patterns when evaluating assumption 3 from Table 1 (e.g., constant variance). The first is a pattern of increasing or decreasing variance with predicted values of y, as depicted in Case B of Figure 2. The second is a pattern (e.g., a trend, a curved line) of the residual with predicted values of y. Both characteristics are usually assessed based on a review of the residual plots and professional judgment alone. The analyst may also need to examine other variables besides predicted values of y to fully evaluate assumption 3.

Independence of residuals (assumption 4 from Table 1) can be evaluated by examining residuals plotted as a function of time. The analyst should look for the same patterns as before. As an alternative for evaluating independence, the analyst can also plot the ith residual, e_i, as a function of the (i-1)th residual, e_{i-1}. One word of caution is in order when reviewing any residual plot: If there are more points in a certain section of the residual plot, the residuals might not appear to be a uniform band of points around 0 (as suggested in Case A of Figure 2); instead, that section might have a somewhat wider band (Helsel and Hirsch, 1995). This is an expected result.

The normality of residuals can be assessed by examining a probability plot. Two problems with nonnormal residuals are the loss of power in subsequent hypothesis tests and increased prediction intervals together with the impression of symmetry (Helsel and Hirsch, 1995).

Figure 3 displays all three of these plots for the split data analyzed from Table 2. From Figure 3, A and B, the split residuals appear to be independent of predicted values of y and time, as well as having constant variance. Thus, the regression meets assumptions 3 and 4 listed in Table 1. In this analysis, testing for residual independence is important since the testing apparatus was calibrated initially, The pumps or other equipment could

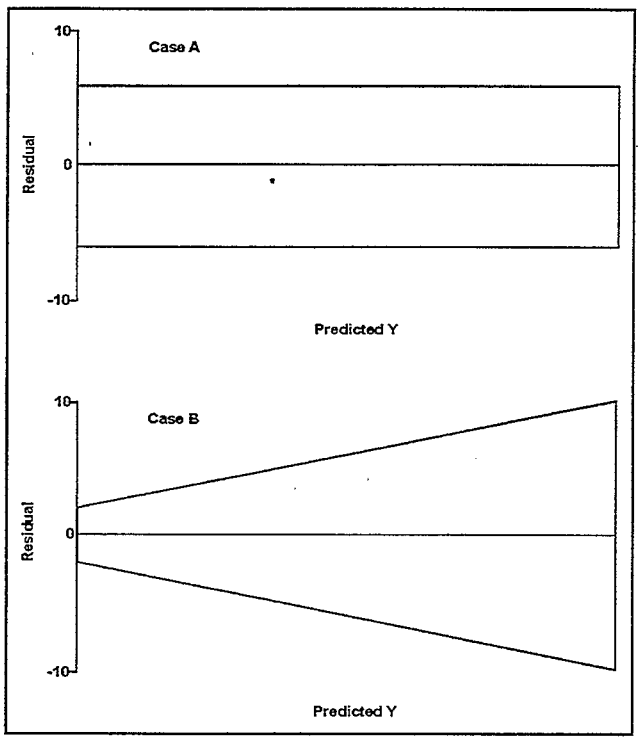

Figure 2. Plot of residuals versus predicted values. (Source: Ponce, 1980)

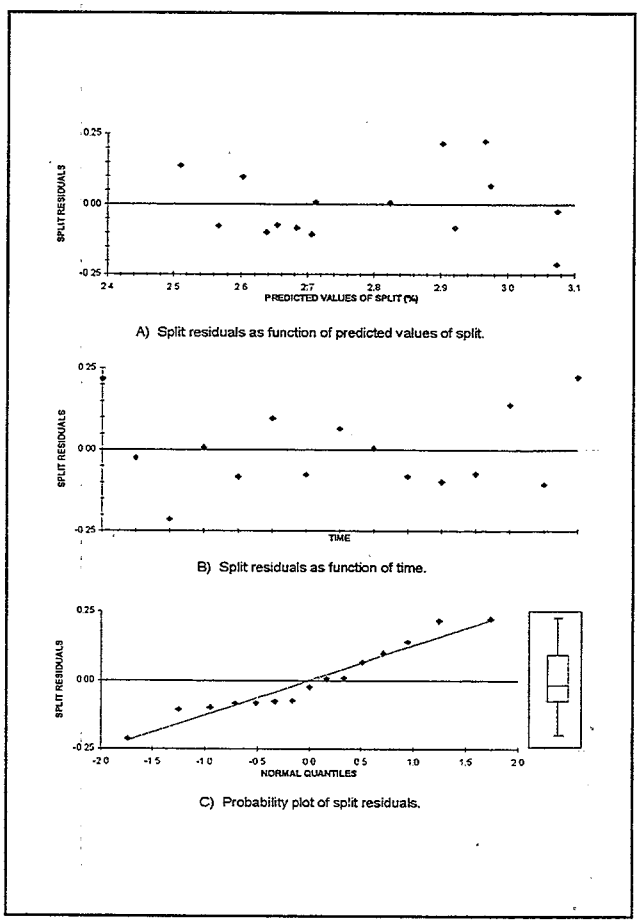

Figure 3. Plot of split residuals.

have differed in performance over time, which in turn would affect the results. Figure 3C, the probability plot, suggests that the data might not rigorously follow the normality assumption, although upon inspection any normality violation appears to be relatively minor. The points in Figure 3C would fall along the plotted line if the residuals were normally distributed. The Shapiro-Wilk W statistic can also be computed to evaluate normality.

Had this analysis violated any of these assumptions, using a different regression technique, transforming the data, or adding variables to the regression would have to be considered. Alternatively, the uses of the regression results could be limited to those identified in Table 1 as restricted by the assumptions met.

MODEL EVALUATION

To determine how well the regression line fits the data, several things can be evaluated:

- Evaluate the proportion of variation in y explained by the model.
- Test whether β_0 is zero.
- Test whether β_1 is zero.
- Compute the confidence interval for β_0.
- Compute the confidence interval for β_1.

As one might imagine, many of these evaluations have already been integrated into standard spreadsheet software. Tables 3 and 4 present a common format that spreadsheets use to present the results from a regression analysis. The top portion of Table 3 also presents the equations used in computing the analysis of variance (ANOVA) summary. Note that S_{xy} and SS_x, the sum of the xy cross products and the sum of the squares x, are defined in Equation 3.

The coefficient of determination, R^2, can be used to evaluate what proportion of the variation can be explained by the model (Gaugush, 1986). R^2 can be computed as (Helsel and Hirsch, 1995)

$$R^2 = \frac{[SS_y - s^2(n-2)]}{SS_y} = 1 - \frac{SSE}{SS_y} \qquad (6)$$

where

$$SS_y = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 \qquad (7)$$

and

$$SSE = \sum_{i=1}^{n} e_i^2 \qquad (8)$$

The residual, e_i, is defined as $y_i - \hat{y}_i$. Values for R^2 range between 0 and 1, with 1 representing the case where all observed y values are on the regression line. The correlation coefficient, r, measures the strength of linear relationships (Freund, 1973) and is computed as the square root of R^2. The sign of r should be the same as the sign of the slope. It ranges from -1 to 1, with the extreme values representing the strongest association and 0 rrepresenting no correlation.

Using the split data from above, the sum of residuals-squared (SSE) is equal to 0.2227 and the sum of the

Table 3. ANOVA summary for runoff sampler calibration data.

Source of Variation	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	SSR = $(Sxy)^2/SS_x$	MSR = SSR/1	MSR/MSE	p
Residual	n-2	SSE	MSE = SSE/(n-2)		
Total	n-1	SSR + SSE			
Application to Runoff Sampler Calibration Data					
Regression	1	0.486623	0.486623	28.410248	0.0001366
Residual	13	0.222670	0.017128		
Total	14	0.709293			

Table 4. Regression analysis of runoff sampler calibration data.

	Coefficients	Standard Error	*t* Statistic	*p* value	Lower 95%	Upper 95%
Intercept (β_0)	3.1317	0.072914	42.950756	2.14E-15	2.97420	3.28924
Flow Rate (β_1)	-0.0119	0.002237	-5.330126	0.00014	-0.01675	-0.00709

squares y (SS_y) is 0.7093; thus, R^2 is equal to 1-(0.2227/0.7093) = 0.686, or 68.6 percent of the variance is explained by the model. The correlation coefficient, r, is then equal to -0.828. The overall model can also be evaluated with the F statistic (28.41), which is computed in Table 3. The F statistic is a measure of the variability in the data set that is explained by the regression equation in comparison to the variability that is not explained by the regression equation. Since the p value of 0.0001366 is less than 0.05, the overall model is significant at the 95 percent confidence level.

Are β_0 and β_1 significantly different from zero? The standard error for β_0 and β_1 in Table 4 can be calculated as (Helsel and Hirsch, 1995)

$$SE(\beta_0) = s\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{SS_x}} \quad (9)$$

$$SE(\beta_1) = \frac{s}{\sqrt{SS_x}} \quad (10)$$

where

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} \quad (11)$$

The value s is equal to the standard error of the regression (which is the same as the standard deviation of the residuals). The corresponding t statistics (with n - 2 degrees of freedom) for β_0 and β_1 are then equal to β_0 and β_1 divided by their respective standard error. The t statistics may then be compared to values from the t distribution to determine whether β_0 or β_1 are significantly different from zero. In this case, β_0 and β_1 are both significantly different from zero based on inspection of their associated p values in Table 4.

The confidence intervals for β_0 and β_1 can be computed using the following formulas (Helsel and Hirsch, 1995):

$$\beta_0 \pm t_{\alpha/2,n-2}\, SE(\beta_0) \quad (12)$$

$$\beta_1 \pm t_{\alpha/2,n-2}\, SE(\beta_1) \quad (13)$$

where $t_{\alpha/2,n-2}$ is the t statistic with n - 2 degrees of freedom. In Table 4, the 95 percent confidence limits are computed. Since the 95 percent confidence limit was selected, α is equal to 0.05 (=1-0.95) and $\alpha/2$ is equal to 0.025. There are 13 degrees of freedom since n is equal to 15. Based on this information, the t statistic can be selected from a look-up table; in this case the analyst would look up $t_{0.025,13}$. Table 5 presents percentiles of the t distribution that can be used for this purpose (more complete tables are available in most introductory statistics books). From Table 5, the appropriate t statistic is estimated as 2.1604. The lower and upper 95 percent confidence limits for β_0 and β_1 are provided in Table 4, using Equations 12 and 13. Had the analyst elected to compute the 90 percent confidence interval, $\alpha/2$ would be equal to 0.05 and $t_{0.05,13}$ would be estimated as 1.7709.

Table 5. Percentiles of the $t_{\alpha,df}$ distribution (values of t such that 100(1-α)% of the distribution is less than t).

df	α=0.10	α=0.05	α=0.025	α=0.010	α=0.005
10	1.3722	1.8125	2.2281	2.7638	3.1593
15	1.3406	1.7531	2.1315	2.6025	2.9467
20	1.3253	1.7247	2.0860	2.5280	2.8453
25	1.3163	1.7081	2.0595	2.4851	2.7874
30	1.3104	1.6973	2.0423	2.4573	2.7500
40	1.3031	1.6839	2.0211	2.4233	2.7045
50	1.2987	1.6759	2.0086	2.4033	2.6778
75	1.2929	1.6654	1.9921	2.3771	2.6430
100	1.2901	1.6602	1.9840	2.3642	2.6259

USING THE REGRESSION LINE

The most obvious use of the regression line is to predict y values for selected values of x. For example, using the regression established above (Equation 5), the split for any flow rate can be estimated. (It is not good practice, however, to predict values beyond the range of test conditions.) For a flow rate of 10 gpm, the predicted split is 3.01 percent; for a flow rate of 50 gpm, the predicted split is 2.53 percent.

Since in most cases the regression line will not fit the data perfectly, the uncertainty associated with the predicted values should be quantified. The regression line can be used either to establish the confidence interval for the population mean of y or to determine the prediction interval for a single value of y. The limits for the single value of y are wider than the corresponding limits on the mean of y (Remington and Schork, 1970) because single observations vary more than means.

The equation for the confidence interval for the population mean y at $x = x_0$ is (Helsel and Hirsch, 1995)

$$\hat{y} \pm t_{\alpha/2,n-2}\, s\sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x}} \quad (14)$$

In this example, $\hat{y}_i$ is equal to the predicted split using Equation 5 and the flow rate equal to x_0. SS_x and s can be estimated by using Equations 3 and 11, respectively. This interval is most narrow at $\bar{x}$ and widens as x_0 moves farther from $\bar{x}$. By calculating the interval at each point along the regression line, a curve like the dashed line in Figure 4 for the example data can be plotted. The equation for the prediction interval for individual values of y at $x = x_0$ is (Helsel and Hirsch, 1995)

$$\hat{y} \pm t_{\alpha/2,n-2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x}} \qquad (15)$$

Figure 4 also shows this interval for the example data.

One of the simplest (in theory) nonpoint source control applications of linear regression is the regression of a water quality indicator against an implementation indicator. For example, flow-adjusted total suspended solids (TSS) concentration could be regressed against a sediment control variable such as the total combined erosion rate of all cropland for which delivery to the stream is likely to be 50 percent or greater. A significant negative slope would suggest (but not prove) that water quality has improved because of implementation of sediment control practices.

Another possible use of simple linear regression is to model a water quality parameter versus time. In this application a significant slope would indicate change over time. The sign of the slope would indicate either improvement or degradation depending on the parameter used. For nonpoint source studies, a simple regression versus time will most likely be confounded by the variability in precipitation and flows. Thus, considerable data manipulation (transformation, stratification, etc.) might be required before regression analysis can be successfully applied. In these cases, it might be more appropriate to apply one of the alternatives to regression described by Helsel and Hirsch (1995).

In many cases water quality parameters are regressed against flow. This approach is particularly relevant in nonpoint source studies. In analysis of covariance, regressions against flow are often performed prior to an ANOVA. One of the implicit goals of nonpoint source control is to change the relationship between flow and pollutant concentration or load. In paired watershed studies, measured parameters from paired samples are often regressed against each other to compare the watersheds (USEPA, 1993). These regression lines can be compared over time to test for the impact of nonpoint source control efforts (Spooner et al., 1985). The reader is referred to *Paired Watershed Study Design* (USEPA, 1993) for an example that demonstrates this technique.

NONLINEAR REGRESSION AND TRANSFORMATIONS

Nonlinear regression (as discussed here) involves transformation to linear equations, followed by simple linear regression. Helsel and Hirsch (1995) provide a detailed discussion on transformations using the "bulging rule" described by Mosteller and Tukey (1977), which can be used to select appropriate transformations. Crawford et al. (1983) list the numerous regression models most often applied by the U.S. Geological Survey for flow-adjusting concentrations. The selection of which transformation to use is ultimately based on an inspection of the residuals and whether the assumptions described earlier are met. Typical transformations include x^2, x^3, $\ln x$, $1/x$, $x^{0.5}$, etc.

When the residuals do not exhibit constant variance (heteroscedasticity), one of several common transformations should be used. Logarithmic transformations are used when the standard deviation in the original scale is proportional to the mean of y. Square root transformations are used when the variance is proportional to the mean of y. In many instances, the right

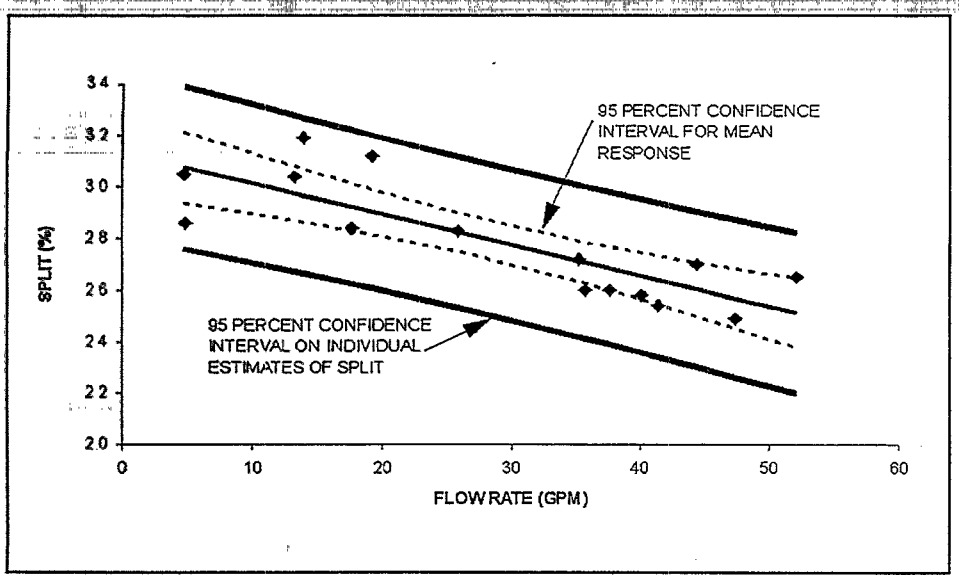

Figure 4. Plot of split versus flow rate with confidence limits for mean response and individual estimates.

transformation will "fix" the nonlinear and heteroscedastic problem. With data that are percentages or proportions (between the values of 0 and 1), the variances at 0 and 1 are small. The arcsin of the square root of the individual values is a common transformation that helps spread out the values near 0 and 1 to increase their variance (Snedecor and Cochran, 1980).

There are several disadvantages when applying transformations to regression applications. The most important issue is that the regression line and confidence intervals are symmetric in the transformed form of the variables. When these lines are transformed back to their normal units, the lines will no longer be symmetrical. The most notable time in hydrology when this creates a problem is when estimating mass loading. To estimate the mass, the means for short time periods are regressed and summed to estimate the total mass over a longer period. This approach is acceptable if no transformations are used—the analyst is summing the means. However, if a log transformation was used, summing the mass over the back-transformed values results in summing the median, which will result in an estimate that is biased low for the total mass (Helsel and Hirsch, 1995).

As an example of nonlinear regression, consider a common relationship that is used to describe load (L) as a function of discharge (Q):

$$L = aQ^{b} \tag{16}$$

Taking the logarithms of both sides yields

$$\ln(L) = \ln(a) + b \ln(Q) \tag{17}$$

which has the same form as Equation 2, introduced at the beginning of this document, where $\ln(L)$ corresponds to y, $\ln(a)$ corresponds to β_0, b corresponds to β_1, and $\ln(Q)$ corresponds to x. By taking the logarithms of both sides, the nonlinear problem has been reduced to a simple linear model. The only additional step that the analyst must perform is to convert L and Q to $\ln(L)$ and $\ln(Q)$ before using standard software. The analyst should be aware that all of the confidence limits are in transformed units; when they are plotted in normal units, the confidence intervals will not be symmetric.

Figure 5 demonstrates how transforming the data may improve the regression analysis. In Figure 5A, sulfate concentrations (in milligrams per liter) are plotted as a function of stream flow (in cubic feet per second). The apparent downward trend is typical of a stream dilution effect; however, the trend is clearly nonlinear. The trend line plotted in this figure, as well as the residuals plotted in Figure 5C, demonstrate that a linear model would tend to over- and underestimate sulfate concentrations depending on the flow. Figure 5B displays the same data after computing the logarithms (base 10) of the sulfate and flow data. A trend line fitted to these data and the residual plot (Figure 5D) clearly demonstrate that applying linear regression after log-transformation would be appropriate for these data.

CONCLUSION

When properly used, regression analysis can be an important tool for evaluating nonpoint source data. However, the analyst should pay close attention that the application of regression does not exceed the uses that are met in Table 1. In some instances it might be necessary to select distribution-free approaches that tend to be more robust. The reader is referred to *Statistical Methods in Water Resources* (Helsel and Hirsch, 1995) for a more complete discussion regarding distribution-free approaches.

REFERENCES

Crawford, C.G., J.R. Slack, and R.M. Hirsch. 1983. *Nonparametric tests for trends in water-quality data using the statistical analysis system.* USGS Open File Report 83-550. U.S. Geological Survey, Reston, Virginia.

Dressing, S., J. Spooner, J.M. Kreglow, E.O. Beasley, and P.W. Westerman. 1987. Water and sediment sampler for plot and field studies. *J. Environ. Qual.* 16(1):59-64.

Freund, J.E. 1973. *Modern elementary statistics.* Prentice-Hall, Englewood Cliffs, New Jersey.

Gaugush, R.F., ed. 1986. *Statistical methods for reservoir water quality investigations.* Instruction Report E-86-2. U.S. Army Engineer Waterways Experiment Station, Vicksburg, Mississippi.

Helsel, D.R., and R.M. Hirsch. 1995. *Statistical methods in water resources.* Elsevier, Amsterdam.

Mosteller, F., and J.W. Tukey. 1977. *Data analysis and regression.* Addison-Wesley Publishers, Menlo Park, California.

Ponce, S.L. 1980. *Statistical methods commonly used in water quality data.* WSDG Technical Paper WSDG-TP-00001. U.S. Department of Agriculture, Forest Service.

Remington, R.D., and M.A. Schork. 1970. *Statistics with applications to the biological and health sciences.* Prentice-Hall, Englewood Cliffs, New Jersey.

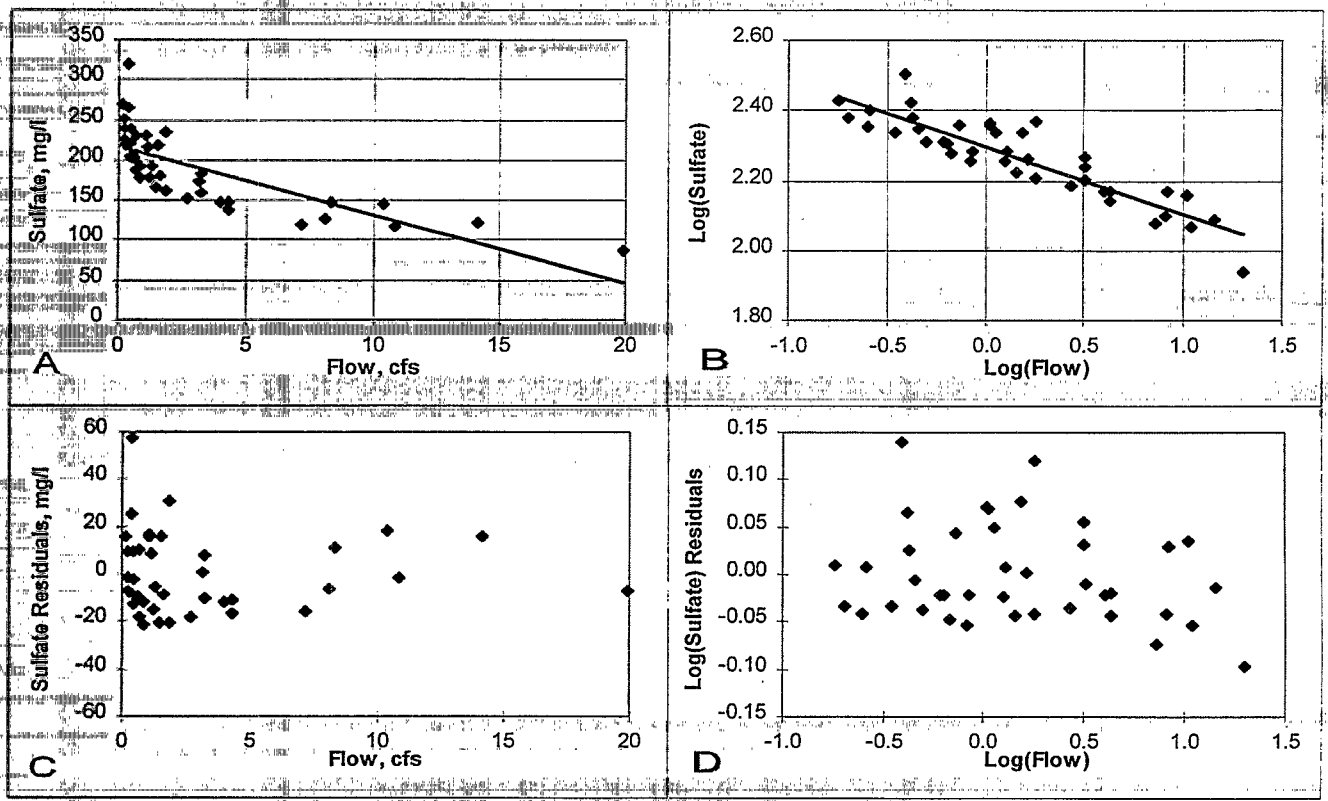

Figure 5. Comparison of regression analyses using raw and log-transformed data.

Snedecor, G.W., and W.G. Cochran. 1980. *Statistical methods.* 7th ed. The Iowa State University Press, Ames, Iowa.

Spooner, J., R.P. Maas, S.A. Dressing, M.D. Smolen, and F.J. Humenik. 1985. Appropriate designs for documenting water quality improvements from agricultural NPS control programs. In *Perspectives on nonpoint source pollution,* proceedings of a national conference, May 19-22, Kansas City, Missouri. EPA 440/5-85-001. U.S. Environmental Protection Agency, Washington, DC.

USEPA. 1993. *Paired watershed study design.* 841-F-93-009. U.S. Environmental Protection Agency, Office of Water, Washington, DC.