Kendall-Theil Robust Line (KTRLine—version 1.0)—A Visual Basic Program for Calculating and Graphing Robust Nonparametric Estimates of Linear-Regression Coefficients Between Two Continuous Variables

By Gregory E. Granato

Chapter 7
Section A, Statistical Analysis,
Book 4, Hydrologic Analysis and Interpretation

In cooperation with the
U.S. Department of Transportation
Federal Highway Administration
Office of Natural and Human Environment

Techniques and Methods 4-A7

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
Dirk Kempthorne, Secretary

U.S. Geological Survey
P. Patrick Leahy, Acting Director

U.S. Geological Survey, Reston, Virginia: 2006

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

For more information on the USGS--the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment:
World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS

Suggested citation:
Granato, G.E., 2006, Kendall-Theil Robust Line (KTRLine—version 1.0)—A visual basic program for calculating and graphing robust nonparametric estimates of linear-regression coefficients between two continuous variables: Techniques and Methods of the U.S. Geological Survey, book 4, chap. A7, 31 p.

Contents

Figures

Conversion Factors and Water-Quality Units

Multiply	By	To obtain
	Flow rate	
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)
cubic foot per second per square mile [$(ft^3/s)/mi^2$]	0.01093	cubic meter per second per square kilometer [$(m^3/s)/km^2$]

WATER-QUALITY UNITS

Chemical concentration is given in units of milligrams per liter (mg/L). Milligrams per liter are units expressing the mass of solute per unit volume (liter) of water. Milligrams per liter are equivalent to parts per million.

ACRONYMS

BCF	bias correction factor
CPU	computer processing unit
IQR	interquartile range
KTRLine	Kendall-Theil Robust Line
RAM	random access memory
MAD	median absolute deviation
NWIS	National Water Information System
NWiz	National Water Information System Wizard
OLS	ordinary least squares
PRESS	prediction error sum of squares
RMSE	root mean square error
USGS	U.S. Geological Survey

Kendall-Theil Robust Line (KTRLine—version 1.0)—A Visual Basic Program for Calculating and Graphing Robust Nonparametric Estimates of Linear-Regression Coefficients Between Two Continuous Variables

By Gregory E. Granato

Abstract

The Kendall-Theil Robust Line software (KTRLine—version 1.0) is a Visual Basic program that may be used with the Microsoft Windows operating system to calculate parameters for robust, nonparametric estimates of linear-regression coefficients between two continuous variables. The KTRLine software was developed by the U.S. Geological Survey, in cooperation with the Federal Highway Administration, for use in stochastic data modeling with local, regional, and national hydrologic data sets to develop planning-level estimates of potential effects of highway runoff on the quality of receiving waters. The Kendall-Theil robust line was selected because this robust nonparametric method is resistant to the effects of outliers and nonnormality in residuals that commonly characterize hydrologic data sets. The slope of the line is calculated as the median of all possible pairwise slopes between points. The intercept is calculated so that the line will run through the median of input data. A single-line model or a multisegment model may be specified.

The program was developed to provide regression equations with an error component for stochastic data generation because nonparametric multisegment regression tools are not available with the software that is commonly used to develop regression models. The Kendall-Theil robust line is a median line and, therefore, may underestimate total mass, volume, or loads unless the error component or a bias correction factor is incorporated into the estimate. Regression statistics such as the median error, the median absolute deviation, the prediction error sum of squares, the root mean square error, the confidence interval for the slope, and the bias correction factor for median estimates are calculated by use of nonparametric methods. These statistics, however, may be used to formulate estimates of mass, volume, or total loads.

The program is used to read a two- or three-column tab-delimited input file with variable names in the first row and data in subsequent rows. The user may choose the columns that contain the independent (X) and dependent (Y) variable. A third column, if present, may contain metadata such as the sample-collection location and date. The program screens the input files and plots the data. The KTRLine software is a graphical tool that facilitates development of regression models by use of graphs of the regression line with data, the regression residuals (with X or Y), and percentile plots of the cumulative frequency of the X variable, Y variable, and the regression residuals. The user may individually transform the independent and dependent variables to reduce heteroscedasticity and to linearize data. The program plots the data and the regression line. The program also prints model specifications and regression statistics to the screen. The user may save and print the regression results. The program can accept data sets that contain up to about 15,000 XY data points, but because the program must sort the array of all pairwise slopes, the program may be perceptibly slow with data sets that contain more than about 1,000 points.

Introduction

Definition of relations between variables in hydrologic data sets can be crucial for understanding environmental processes and identifying management measures that will minimize adverse effects of anthropogenic activities. For example, the ability to estimate concentrations and loads of water-quality constituents as a function of discharge (defined herein as the volume of streamflow passing a specific point during a given time interval) is considered important in studies to assess the effectiveness of programs for abating nonpoint-source constituents and to understand the transport and fate of sediment-borne constituents (Crawford, 1991). Estimation of concentrations and loads of sediment and other water-quality constituents from discharge measurements is of particular interest because continuous records of daily discharge are available at more than 20,000 streamflow-gaging stations across the United States for periods of years to decades (U.S. Geological Survey, 2004). Availability of water-quality measurements, however, is limited by the effort and expense of collection and analysis of concentration data. Many

studies have used regression methods to analyze water-quality constituents such as sediment (Miller, 1951; Glysson, 1987; Clarke, 1990a, b; Gilroy and others, 1990; Crawford, 1991; Nash, 1994; Syvitski and others, 2000; Vogel and others, 2003), nutrients (Smith and others, 1982; Clarke, 1990a; Cohn and others, 1992; Cohn, 1995; House and Warwick, 1998; Vogel and others, 2005), major ions (O'Connor, 1976; House and Warwick, 1998), and trace elements (Driver and Tasker, 1990). Relations between discharge and water quality commonly follow a power law rather than a linear trend. Therefore, investigators commonly use logarithmic transformation to linearize the relation.

Linear regression analyses also have many potential uses in urban and highway runoff studies (Tasker and Granato, 2000). Regression analysis is commonly used to develop estimates of runoff coefficients (Schueler, 1987; Driscoll and others, 1990). Regression has been used to examine relations between different constituents in runoff (Driscoll and others, 1990; Thomson and others, 1997). Regression has also been used to predict concentrations and loads of water-quality constituents in runoff (Driscoll and others, 1990; Thomson and others, 1996; Charbeneau and Barrett, 1998). Estimates of the performance of structural best management practices for reducing concentrations and loads in runoff have also been made using regression analysis (Martin and Smoot, 1986; Barrett, 2005). These studies have all used parametric-regression techniques, such as Ordinary Least-Squares (OLS) regression, and the authors of these studies have reported varying levels of success for different applications.

Parametric-regression techniques, such as OLS regression, commonly are available in spreadsheets, graphing, and statistical software. Parametric techniques, however, may not be the best methods for some hydrologic data sets. Helsel and Hirsch (2002) featured the Kendall-Theil robust line as a nonparametric alternative to OLS methods for statistical analysis of water-resources data. In their discussion on regression, Helsel and Hirsch (2002) noted that investigators commonly apply parametric-regression techniques a priori with a "blind reliance on the computer software." Therefore, availability of OLS routines within statistical and spreadsheet software and the interface design of such software may encourage only a cursory examination of regression statistics, which may lead to poorly specified regression models.

Regression analysis of hydrologic data is further complicated by various factors that may make a multisegment regression model the most suitable choice for analysis of data. Multisegment models are appropriate when different environmental processes are dominant over different ranges of the explanatory variable(s). For example, O'Connor (1976) demonstrates use of a theoretical two-line mixing model for river basins that have a relatively constant base-flow concentration (C_g) below some threshold discharge (Q_o) and either a positive or negative slope as rainfall-runoff processes either contribute or dilute water-quality constituents in the receiving water (fig. 1). Similarly, Glysson (1987) and Simon (1989) suggest use of two- or three-segment models to predict suspended-sediment concentrations and loads to account for changing relations in relative sediment availability and transport capacity with increasing stream discharge.

A search of the literature and of the documentation for statistical and spreadsheet software commonly cited in the hydrologic literature did not indicate that a computer implementation of the Kendall-Theil slope estimator and the median-intercept estimator were readily available. This type of nonparametric regression is not commonly implemented for various reasons including the scientific community's general familiarity with and prevalence of parametric statistics and the computational intensity required to calculate the median-slope estimator. The OLS regression components of commonly cited software did not facilitate the graphical approach to the analysis of data and residuals that is recommended for regression analysis of water-resources data (Helsel and Hirsch, 2002). Furthermore, available commercial software is commonly designed to produce a single regression line for each data set rather than the multisegment models that are recommended for analysis of suspended sediments, sediment-borne constituents, and dissolved constituents as a function of stream discharge (O'Connor, 1976; Glysson, 1987).

This report describes the implementation, use, and interpretation of results from the Kendall-Theil Robust Line (KTRLine—version 1.0) software. The KTRLine software was developed by the U.S. Geological Survey, in cooperation with the Federal Highway Administration, for use in the analysis of local, regional, and national hydrologic data sets. The software was developed for data generation in support of a stochastic empirical loading- and dilution-model for planning-level estimates of the effects of highway runoff on the quality of receiving waters. The graphical process of developing a non-parametric-regression model is described with the governing equations and numerical methods. Methods for transformation of data and selection of multisegment models are documented. The formats of input data and output regression results are described. Step-by-step use of the program's graphical user interface is illustrated. The program source-code written in Microsoft Visual Basic 6.0 is documented in individual files in a Visual Basic project directory on the computer disk accompanying this report.

Statistical Theory and Governing Equations

Linear regression is considered a fundamental tool in the analysis of water-resources data (Helsel and Hirsch, 2002). Linear regression is the process of fitting a straight line to a data set that consists of an explanatory (independent, predictor, or X) variable and a response (dependent, predicted, or Y) variable. The process of regression yields an estimate of the slope of the line, an estimate of the Y-intercept (the value of the line when the predictor variable X equals zero), and regression statistics that indicate how well the line describes

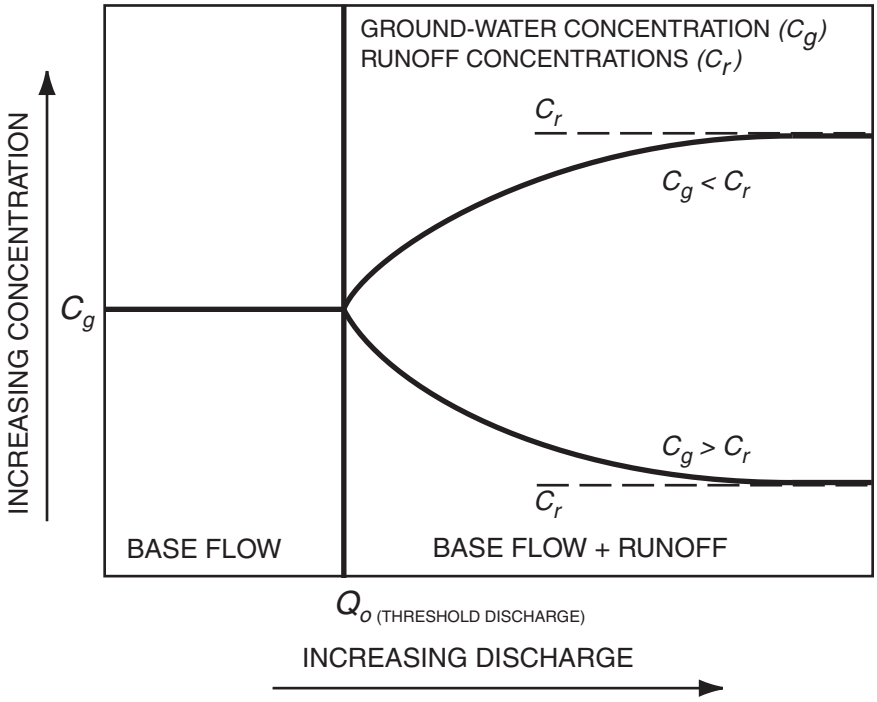

Figure 1. Simplified relations between increasing discharge and the concentration of water-quality constituents in a stream or river during base flow and runoff events. (Modified from O'Connor, 1976.)

variability in the data. Regression methods commonly are used to predict values of *Y* from measurements of *X*, detect trends, explain variability caused by one variable (for example, discharge) so that other trends (for example, seasonality) may be detected, or to detect changes in an environmental system (for example, changes in the equation of the line caused by implementing a management measure within a river basin) (Helsel and Hirsch, 2002). Regression models of concentration and discharge are commonly used to estimate continuous water-quality records from continuous discharge measurements (water-quality transport curves) during periods between periodic water-quality measurements. Regression models also are used for data generation in the development of stochastic models (Koch and Smillie, 1986). Success of a modeling effort depends on the premise that the data are representative of the system of interest and the linear model that is used is appropriate for describing those data.

Parametric Regression

A number of regression techniques may be suitable for analysis of hydrologic data. OLS regression is the parametric-regression technique that is commonly used for analysis of hydrologic data (Tasker and Granato, 2000). OLS regression is prevalent because OLS regression parameters, including estimates of the slope, the intercept, and regression diagnostic statistics, are based on parametric-statistical methods that are familiar to most investigators and, in comparison to other statistical methods, are relatively simple to calculate (Helsel and Hirsch, 2002). As such, OLS is commonly the standard regression technique available in statistical-software packages and spreadsheets.

Proper application of OLS regression techniques depends on several restrictive assumptions about the structure of input-data sets, which may not be attainable for hydrologic data sets. Hydrologic data sets commonly have statistical properties, such as outliers and skewed distributions, that are non-ideal for application of parametric statistical techniques and that may not meet the theoretical underpinnings of an OLS model (Hirsch and others, 1982; Koch and Smillie, 1986; Hirsch and others 1991; Helsel and Hirsch, 2002). For example, hydrologic data are commonly characterized by measurement error or uncertainty that is a function of the magnitude of the measurement. OLS regression, however, is based on the assumption that the predictor variable *X* is known without error and therefore OLS regression equations are designed to minimize residual errors in the predicted *Y* variable without regard to potential errors in *X* (Helsel and Hirsch, 2002). Serial correlation and seasonality also violate OLS assumptions about the independence of subsequent measurements (Glysson, 1987; Cohn and others, 1992; Helsel

and Hirsch, 2002). For example, rainfall-runoff processes cause hysteresis in discharge-concentration relations so that there may be one relation between concentration and discharge on the rising limb of the hydrograph and a different relation on the falling limb of the hydrograph (O'Connor, 1976; Glysson, 1987; House and Warwick, 1998). Thus, the OLS technique may not provide the best linear equation for hydrologic data. Furthermore, OLS regression is not well suited to regional studies that include the combination and analysis of multiple data sets because the properties of the mixed distribution are not suitable for use with parametric techniques (Hirsch and others, 1991).

Transformation may be used to linearize data and to reduce heteroscedasticity in residuals, but this method does not address other problems that make some hydrologic data sets non-ideal for OLS analysis. For example, Vogel and others (2005) determined that a lognormal transformation provides a good first-order approximation to the relation between concentration and flow, but that residuals from such regression models commonly fail standard tests for normality. Also, environmental measurements (such as discharge, water-quality, and sediment-quality measurements) commonly are reported as being below one or more detection limits or greater than one or more maximum reporting limits. These hydrologic data points are commonly considered high-leverage points because the results of OLS analysis commonly are sensitive to assumptions made about these data in the tails of the distributions (Helsel and Hirsch, 2002). Robust methods, because they are not sensitive to values at the tails of the population distributions, minimize the effect of assumptions about data below detection limits and the effect of outliers on the determination of relations between variables (Helsel and Hirsch, 2002).

Nonparametric Regression

The need for robust regression techniques have led some researchers to apply the nonparametric techniques described by Kendall (1938) to the problem of estimating the slope of a best-fit line through data that are not well suited for OLS regression (Theil, 1950; Sen, 1968; Conover, 1980; Brauner, 1997). The nonparametric estimate of slope, identified as the Kendall-Theil or Sen slope in the literature (Conover, 1980; Hirsch and others, 1982; Dietz, 1987; Hirsch and others, 1991; Brauner, 1997; Nevitt and Tam, 1998; Helsel and Hirsch, 2002), is the median of all the slopes that can be calculated between each data point and every other data point in the data set (Theil, 1950; Sen, 1968). The nonparametric estimate of intercept has been proposed as either the median intercept calculated by use of each data pair and the median slope (Theil, 1950) or the intercept calculated by use of the median slope and the median of the X and Y variables (Conover, 1980). Dietz (1987) concluded that the Conover (1980) estimate was more robust than other estimates of the intercept because it was consistently one of the most efficient estimators under all the conditions tested. Nevitt and Tam (1998) concluded that the Theil (1950) estimate was slightly more robust than the Conover (1980) estimate of the intercept under the range of conditions they tested with their hypothetical data sets. Helsel and Hirsch (2002), however, recommended using the estimate of intercept produced by placing the line through the point defined by the medians of the predictor (X) and response (Y) variables (Conover, 1980) because it is robust, efficient, relatively simple to compute, and is analogous to OLS regression in which the intercept is calculated by placing the line through the point defined by the averages of the X and Y variables. Therefore, the Kendall-Theil robust line regression is defined as having the median slope and intercept (Theil, 1950; Sen, 1968; Conover, 1980; Helsel and Hirsch, 2002).

A primary consideration for selection of a nonparametric regression method is the potential effect of outliers on regression statistics. The slope and intercept of an OLS regression line is based on the means and sum of squares of the X and Y data sets, which are substantially influenced by outliers in the data set (Helsel and Hirsch, 2002). The nonparametric KTRLine statistics, however, commonly are not influenced by these outliers. For example, even one high-leverage outlier may bias an OLS regression estimate as shown in figure 2. The data set in figure 2, with the exception of one outlier, follows a line with a slope of 1.1, an intercept of 2, and an alternating error component of plus or minus 0.5 units. The slope estimated by OLS regression is more than twice that of the majority of the data set, whereas the KTRLine slope matches that of the "good" data in this example. The KTRLine estimate of the intercept in this example is slightly biased by the number of points and the systematic error component, but this estimate is much better than the biased OLS estimate of the intercept for the majority of data points. In this example, the user could easily identify and eliminate the outlier to produce a better OLS estimate for the "good data." Identification of outliers and the resultant effect on OLS estimates of slope and intercept can be difficult or impossible in the analysis of hydrologic data sets, which commonly range over two or more orders of magnitude and include a substantial random error component. Furthermore, a priori elimination of outliers may be misleading, will misrepresent the predictive capability of the OLS regression model, and is not recommended for analysis of hydrologic data sets (Helsel and Hirsch, 2002). Investigation of such outliers may identify a different environmental process (for example, fig. 1) or population of data, which may be characterized by more data collection and(or) a separate equation.

Use of OLS regression provides the best estimate of the linear slope, intercept, and regression statistics under the ideal conditions described by the assumptions underlying the OLS technique, but OLS can provide estimates that are substantially biased when data violate one or more of these assumptions (Theil, 1950; Sen, 1968; Dietz, 1987; Hirsch and others 1991; Nevitt and Tam, 1998; Helsel and Hirsch, 2002). Numerical experiments indicate that the Kendall-Theil slope estimator is almost as efficient as OLS regression under ideal conditions for OLS and is much more efficient than OLS even when conditions do not depart substantially from the ideal (Hussain

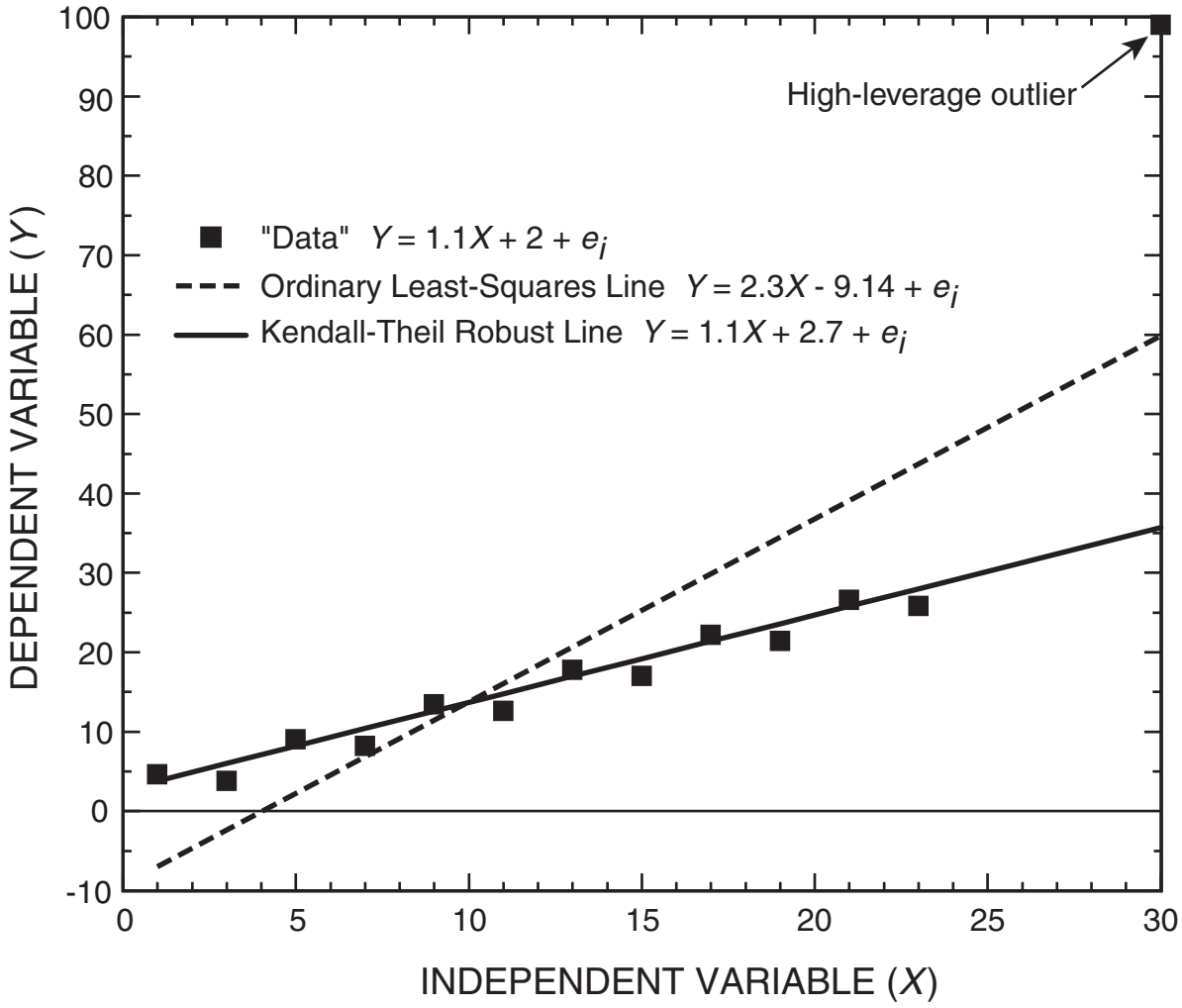

Figure 2. Effect of one high-leverage outlier on estimates of ordinary least-squares regression statistics.

and Sprent, 1983; Dietz, 1987; Hirsch and others, 1991; Brauner, 1997; Nevitt and Tam, 1998; Helsel and Hirsch, 2002). Similarly, the literature indicates that the OLS estimate of intercept is most efficient under ideal conditions, but if data are not ideal, the nonparametric estimates of the intercept of a line are more efficient than the OLS estimate (Dietz, 1987; Nevitt and Tam, 1998).

The arithmetic average has special meaning in hydrologic studies designed to determine constituent loads because the arithmetic average constituent load is, in theory, the total mass of the constituent divided by the total amount of water that flows past the measurement point. Although a parametric regression equation may overestimate the value of a high proportion of values in a skewed data set, the parametric line, which provides an estimate of the average response, will provide a better estimate of the total load. The Kendall-Theil robust line is a median line and, therefore, may underestimate the total mass, volume, or load unless the error component or a bias correction factor is incorporated into the estimate. Because it is not biased by the skew in the data set, the nonparametric line will provide a more robust estimate of individual values over the full range of data.

Governing Equations for Kendall-Theil Robust Line Regression

Linear regression is based on the algebraic expression of a straight line. In a regression model, however, none of the terms are known, each term must be estimated from available data, and an error term is added to account for individual departures from the estimated linear equation. Therefore, the equation or a linear regression model may be written as

$$Y_i = m \times X_i + b + e_i \quad \text{for } i = 1 \text{ to } n \ , \tag{1}$$

where

X_i is the explanatory (independent, predictor, or X) variable for each data point (i);

Y_i is the response (dependent, predicted, or Y) variable for each data point (i);

e_i is the residual error or uncertainty in the predicted Y value for each data point (i);

m is the estimated slope;

b is the estimated intercept;

and

n is the number of XY data points in the sample.

Slope

The slope of the line (m) is estimated as the median of all pairwise slopes between each pair of points in the data set (Theil, 1950; Sen, 1968; Helsel and Hirsch, 2002). Each individual slope estimate (m_{ij}) for the line connecting the ith and jth data point is calculated by use of the equation

$$m_{ij} = \frac{(Y_j - Y_i)}{(X_j - X_i)} \text{ for } i = 1 \text{ to } n\text{-}1 \text{ and } j = 2 \text{ to } n \ . \tag{2}$$

The number of possible slopes between data pairs is calculated by use of the equation

$$N_p = \frac{n \times (n-1)}{2} \ . \tag{3}$$

After each slope is calculated, all the slope estimates (m_{ij}) are sorted and ranked from lowest to highest. Sorting is a computationally intensive process because each slope estimate in the array of slopes must be compared to other values and put in the proper order. If N_p is an odd number, the median slope is selected as the middle value of the array; otherwise, the median is calculated as the arithmetic average of the two center points. Helsel and Hirsch (2002) provide a graphical example of the process (fig. 3).

There commonly are multiple measurements of the response variable Y_i that share a single value of the predictor variable X_i in hydrologic data sets. In the limit, if a data set consisted of one value of X and multiple values of Y, there would be no relation between X and Y and the slope would be zero. In practice, there is some measurement uncertainty in each measured value of X. This uncertainty, however, poses a problem because as one value X_j approaches another X_i, the slope calculated by equation 2 approaches infinity. In a computer implementation of this equation, the situation in which X_j equals X_i produces a division by zero error. In the KTRLine software, the situation in which X_j equals X_i is handled by alternately setting m_{ij} equal to plus or minus 10 million. The number of such values is tracked and the median of only the finite slopes is selected. This strategy produces the correct median estimate as long as at least two X variables have different values. Two distinct X values are necessary to produce at least one finite slope. The problem of ties in the independent variable is illustrated with three hypothetical data sets in figure 4. Each data set has eight points with six tie values between adjacent points in the sorted array of X values. In example A, 7 of the points share one X value, 7 of the 28 potential pairwise slopes are finite, and 21 of the potential

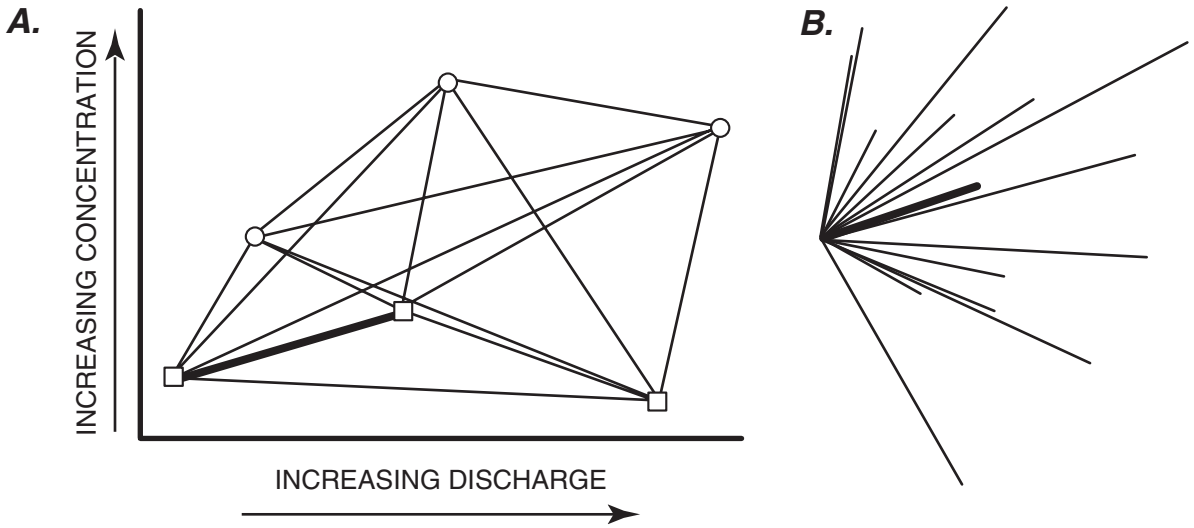

Figure 3. The manual method of determining the median slope: *A,* all possible pairwise slopes between six data points; and *B,* all possible slopes rearranged to meet at a common origin. The thick line is the median of the N_p slopes. (Modified from Helsel and Hirsch, 2002.)

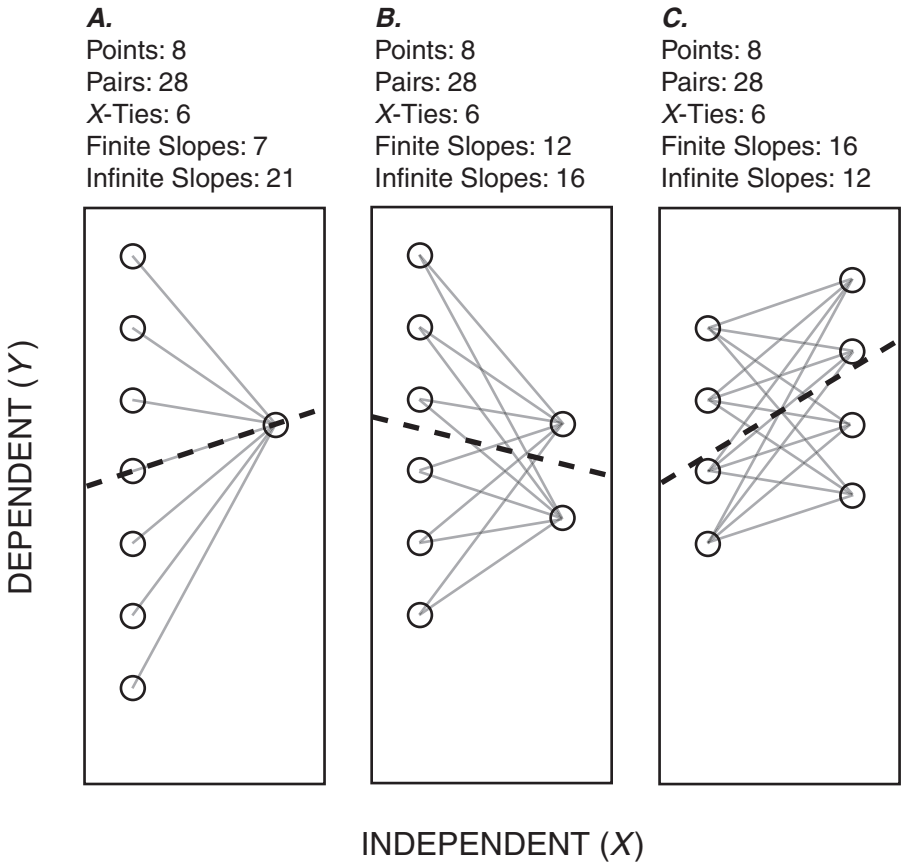

Figure 4. The effect of ties in the independent variable on the number of finite slopes that may be calculated for the Kendall-Theil Robust Line (the bold dashed-line) and the representativeness of the median of finite slopes for indicating relations between the *X* and *Y* variables.

pairwise slopes would approach infinity. Examples *B* and *C* demonstrate the effect of multiple ties in two *X*-axis locations, which reduces the proportion of infinite slopes. In each case, the median of the finite slopes provides the best representation of the relation between data at each *X* value. The KTRLine software indicates the number of tie values between adjacent points in the sorted array of *X* values. These ties do not affect the estimate of slope but may affect the estimate of the confidence limit of the slope. If the KTRLine software outputs an estimate of the confidence limit of the slope that includes a value of plus or minus 10 million, then the user must evaluate the data set because the number of ties may be excessive.

The confidence limits for the KTRLine slope are calculated by use of the large-sample approximation equations described by Helsel and Hirsch (2002). These equations are

$$R_u = \frac{N_p + Z \times \sqrt{\frac{n \times (n-1) \times (2n+5)}{18}}}{2} + 1 \qquad (4)$$

and

$$R_l = \frac{N_p - Z \times \sqrt{\frac{n \times (n-1) \times (2n+5)}{18}}}{2} \quad , \qquad (5)$$

where

R_u is the rank order of the slope that is the upper confidence interval,

R_l is the rank order of the slope that is the lower confidence interval,

N_p is the number of pairwise slopes calculated by use of equation 3,

n is the number of *XY* data pairs,

and

Z is the critical *Z* value taken from a table of standard normal quantiles.

The KTRLine software rounds the estimates of rank to the nearest integer, and the value of that rank is selected from the array of pairwise slopes. The KTRLine software uses a *Z* value of 1.96 to calculate the 95-percent confidence interval of the slope. The example provided by Helsel and Hirsch (2002) indicates that the large-sample approximation is appropriate for samples that include at least 20 pairs of data. A sample size of five *XY* points is, algebraically, the minimum sample size that will produce meaningful ranks. Use of equations 4 and 5 with five *XY* points will produce a 95-percent confidence interval including all 10 pairwise slopes. A user may specify a multisegment model that includes one or more individual segments with less than five *XY* data points. Therefore, the software is designed to reset the confidence interval to the maximum and minimum ranks of the *XY* data points that are used to calculate the slope of each segment if the ranks calculated by use of equations 4 or 5 exceed the number of *XY* pairs in the data set for any given segment. In this case, the program calculates and reports the confidence interval for the maximum and minimum ranks of the *XY* data points.

Intercept

The estimate of the intercept is calculated by use of the Conover (1980) equation

$$b = Y_{median} - m \times X_{median} \quad , \tag{6}$$

where

b is the estimated intercept,

Y_{median} is the median of the response variables,

m is the estimated slope,

and

X_{median} is the median of the explanatory variables.

Residual Error

The residual error or uncertainty in the predicted *Y* value for each data point (e_i) can be calculated by use of the equation

$$e_i = Y_i - (m \times X_i + b) \quad . \tag{7}$$

The residual error around a linear model is a random variable, and, in theory (with a perfectly specified regression model), should be normally distributed with a mean and median of zero, and a constant variance that is independent of the value of X_i (Helsel and Hirsch, 2002). A Kendall-Theil regression model, however, is not bound to these assumptions. For hydrologic data, these errors are related to measurement error (including sample collection, preservation, and analysis) and natural variability caused by processes not evaluated in the regression model. Natural processes may include seasonality (Glysson, 1987; Cohn and others, 1992), hysteresis in concentrations in the rising and falling limb of the hydrograph (O'Connor, 1976; Glysson, 1987; House and Warwick, 1998), and changes in basin characteristics with time (Glysson, 1987; Cohn and others, 1992). Differences among basin characteristics that cause differences in the concentration-discharge relations among basins used to formulate a regional regression model also are a source of error (or uncertainty) in the regression model. Version 1.0 of the KTRLine software does not support multivariate linear-regression models, but the software has a data-identification utility that may be used to identify different populations within a data set. The user may use this information to segregate the data, by season, for example, for separate analysis.

Regression Statistics

A number of regression statistics are based on population characteristics of the residual error including the median deviation, the median absolute deviation (MAD), the prediction error sum of squares (PRESS), and the root mean square error (RMSE). The median deviation, which is the median of e_i values, is a location estimator for the population of residual errors (Helsel and Hirsch, 2002). If the line is a good fit, the median deviation would be zero. Departures in the median deviation indicate that the population of values are skewed above (for positive median deviations) or below the best-fit line (for negative median deviations). Substantial departures from zero may indicate that transformation of the data and(or) a multisegment model may be necessary to properly characterize the relation(s) between *X* and *Y*. The MAD, which is the median of the absolute value of all e_i values, is an estimator of spread in the population of residual errors. The MAD is analogous to the standard deviation value used in parametric statistics, but the MAD is not affected by outliers that are common in hydrologic data sets (Helsel and Hirsch, 2002). If the population of residual errors is normally distributed, the MAD will equal about two-thirds of the estimated standard deviation and about one-half of the interquartile range (IQR) of the residual error in the data set (Helsel and Hirsch, 2002).

A nonparametric version of the PRESS statistic is implemented for comparison of the predictive capability of competing regression models developed with the same data-transformation method (D.R. Helsel, U.S. Geological Survey, written commun., 2005). The PRESS statistic is based on the residual divided by the leverage for each point (Helsel and Hirsch, 2002). The nonparametric version of the PRESS statistic is based on the nonparametric KTRLine residual and a leverage statistic h_i, which describes how far each data point is from the center-of-mass of all *X* values. The KTRLine program calculates h_i by use of the median value of *X* rather than the mean. Thus, the nonparametric version of the PRESS statistic is calculated as

$$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_i} \right)^2 \quad , \tag{8}$$

where e_i is defined as the residual error of the KTRLine (eq. 8), and the leverage statistic h_i is defined as

$$h_i = \left(\frac{1}{n} \right) + \frac{\left(X_i - X_{median} \right)^2}{\sum_{i=1}^{n} \left(X_i - X_{median} \right)^2} \quad . \tag{9}$$

The RMSE, also known as standard error of the regression or standard deviation of residuals, indicates lack of precision (spread) in the population of residual errors (Helsel and Hirsch, 2002). The RMSE is calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - 2}} \quad . \tag{10}$$

The RMSE gives more weight to outliers than the MAD because the error term is squared and then it is summed to calculate the RMSE.

The Bias Correction Factor

Methods that produce a biased estimate of the average response of a system will produce a biased estimate of the total load. The arithmetic average has special meaning in hydrologic studies designed to determine constituent loads because the arithmetic average constituent load is, in theory, the total mass of the constituent divided by the total amount of water that flows past the measurement point. Even a small bias in individual predictions may produce a substantial error in calculated loads. Bias occurs in OLS regression of transformed variables because this OLS line is a prediction of the average response of the *Y* variable for a given *X* variable, and the retransformed average of the transformed data does not equal the arithmetic average of the untransformed data (Koch and Smillie, 1986; Glysson, 1987; Gilroy and others, 1990; Crawford, 1991; Cohn and others, 1992; Hirsch and others, 1992; Helsel and Hirsch, 2002). Ignoring bias correction can result in underprediction of total loads, but use of bias correction can result in overprediction of total loads (Koch and Smillie, 1986). Several bias correction methods have been proposed in the literature (Duan, 1983; Koch and Smillie, 1986; Cohn and others, 1992; Helsel and Hirsch, 2002).

The nonparametric smearing estimator (average of the retransformed log-regression residuals) proposed by Duan (1983) was selected for implementation with the KTRLine software because it performs reasonably well and is not sensitive to statistical assumptions of residual population characteristics (Gilroy and others, 1990; Crawford, 1991; Hirsch and others, 1992; Helsel and Hirsch, 2002). The Duan (1983) bias correction factor is called the smearing estimator because the method applies or "smears" the average retransformed error over all measurements. This approach was developed for log-normal transformations with the natural log. With the appropriate retransformation method, however, the smearing estimator should be applicable for other transformations. A generalized expression of the smearing estimator, applicable for any log-based transformation is

$$BCF = \sum_{i=1}^{n} \frac{G(e_i)}{n} \quad , \tag{11}$$

where

BCF is the bias correction factor,
G is the retransformation function,
e_i is the residual error or uncertainty in the predicted *Y* value for each data point (*i*),

and

n is the number of *XY* data pairs.

A *BCF* should be used when calculating total loads by use of predictions from a Kendall-Theil robust line even if a transformation is not used, because the Kendall-Theil robust line produces an estimate of the median response rather than the mean response (D.R. Helsel, U.S. Geological Survey, oral commun., 2004). The smearing estimate should be added to KTRLine estimate when a *BCF* is used with a regression model of untransformed data if mass or load estimates are required. Three equivalent approaches can be used to estimate masses or loads in log-based transformations: (1) the regression result may be calculated in log space, transformed, and multiplied by the *BCF*; (2) the equation may be expressed as an exponential function in real (untransformed) space and multiplied by the *BCF*; and (3) the *BCF* may be retransformed to log space and added to the regression equation expressed in log space. Any of these approaches may work because values that are multiplied in real space are added in log space. Application of bias correction for other transformations, which are not commonly used in hydrology, is more complex because other transformations will produce truncated normal distributions. These transformations in the ladder of powers will require polynomial approximations for retransformation (T.A. Cohn, U.S. Geological Survey, written commun., 2005).

Bias correction for estimates of mass or loads can be generalized to any transformation taking the average of all regression predictions for a given *X* value with each individual error value (e_i) in the transformed space, retransforming each

value, and computing the arithmetic average of these values (Hirsch and others, 1992). This approach can be expressed mathematically as

$$Y_i = \frac{\sum_{i=1}^{n} G(m \times X_i + b + e_i)}{n} \quad , \tag{12}$$

where

Y_i is the untransformed value of *i*th data point;
X_i is the value of *i*th data point (which may or may not be transformed depending on the model specified);
G is the retransformation function;
e_i is the residual error or uncertainty in the predicted *Y*-value for each data point (*i*);

and

n is the number of *XY* data pairs.

The Y_i in equation 12 may be calculated by the user outside the KTRLine program in two ways: (1) by calculating the result of the regression equation with errors determined from the original input-data set and averaging the result, or (2) by using a random number generator that will produce a normal distribution of errors that have the same median error as the input-data set and a standard deviation of errors calculated from the MAD. These methods are illustrated with the example data set in the file "BCFExamples.xls" on the CD-ROM accompanying this report.

The KTRLine software was developed for data generation in support of a stochastic empirical loading and dilution model for planning-level estimates of the effects of highway runoff on the quality of receiving waters. No bias correction is necessary for a regression model that is used for stochastic data generation (for example, in a Monte Carlo model) as long as the random error component of the relation between *X* and *Y* is included in estimates from the model (Koch and Smillie, 1986). Bias correction is not necessary for use in stochastic data generation because application of the regression model with the random error component produces a population of individual estimates that may be retransformed and then averaged.

Cunnane Plotting Position Formula

The KTRLine software provides for graphical analysis of the data and the residuals. A probability plot of the data and residuals is produced by use of the Cunnane (1978) plotting position formula. This formula was selected for implementation in the KTRLine software because the Cunnane (1978) plotting position formula is acceptable for normal probability plots, it has been found useful for flow-duration and flood-frequency curves, and it is recommended for hydrologic applications by Helsel and Hirsch (2002). The general equation for the plotting position formula is

$$p = \frac{i - a}{n + 1 - 2a} \quad , \tag{13}$$

where

p is the calculated plotting position;
i is the rank of the *i*th data point;
a is the plotting position factor, which is different for each plotting position formula;

and

n is the number of data points in the population of interest (Helsel and Hirsch, 2002).

The plotting position factor (a) is 0.4 for the Cunnane formula. In comparison, the plotting position factor (a) is 0 for the Weibull formula, 0.375 for the Blom formula, 0.44 for the Gringorten formula, and 0.5 for the Hazen formula (Helsel and Hirsch, 2002). Equation 13 produces a plotting position greater than zero and less than one. If the percentiles are desired, the result from equation 13 may be multiplied by 100.

The Point of Convergence for Multisegment Models

If a user chooses to implement a multisegment regression model, the KTRLine program calculates the slope and intercept of each segment of the model by use of the data in the interval that is specified by the user. It is, however, necessary to provide a continuous function that provides a unique value of the dependent variable for each unique value of the independent variable (Glysson, 1987). Therefore, it is necessary to calculate the point where two adjacent regression line segments converge and to set the convergence point as the break point of each interval. Regression statistics for each line segment are calculated on the basis of the points in the interval between calculated convergence points. The convergence point is where both the independent (X) and dependent (Y) variables are equal, and, therefore, the equations of both line segments

$$Y_{seg1} = Y_{seg2} \tag{14}$$

and

$$X_{seg1} = X_{seg2} \tag{15}$$

are equal.

The combination of equations 1 and 14 yields

$$m_{seg1} \times X_{seg1} + b_{seg1} = m_{seg2} \times X_{seg2} + b_{seg2} \quad . \tag{16}$$

Two straight lines have one point of intersection. The combination of equation 16 with the equivalence of equation 15 yields the equation for the point of convergence as

$$X_{convergence} = \frac{b_{seg2} - b_{seg1}}{m_{seg1} - m_{seg2}} \quad . \tag{17}$$

Potential problems occur when the user specifies adjacent intervals that have lines with either the same (or similar) *Y*-intercepts or the same (or similar) slopes. The program assigns *XY* points to each segment in descending segment order from right to left (higher to lower *X* values) on the graph and calculates regression statistics for each segment with the data points within the intervals defined by the convergence calculations. If the adjacent lines have the same *Y*-intercept but different slopes, the program applies equation 17 and sets the point of convergence at the intercept irrespective of the user-specified interval for each line. Therefore, if a two-line model is specified and intercepts are equal, the point of convergence is at the *Y* intercept because the point of convergence equals zero (eq. 17). If the adjacent lines have the same slope and the same intercept, the lines are equal. This condition of equivalent lines would create a division by zero error when equation 17 is applied in the KTRLine software. If the adjacent lines have the same slope but different intercepts, the lines are parallel. Parallel lines also would create a division by zero error when equation 17 is applied in the KTRLine software. Therefore, in each of these cases, the KTRLine software attributes all values in both intervals to the highest order segment for the purposes of calculating regression statistics and plotting the multisegment model on the graph. If any of these conditions occur, however, the program warns the user that there is a discontinuity in the intersection of the regression lines and annotates the regression results to indicate that the segments do not converge. If a discontinuity is indicated, the user should reduce (or increase) the number of segments being specified and(or) adjust the specified break point of each interval until a regression model is specified that does not include any discontinuities.

Development of a Regression Model

Helsel and Hirsch (2002) describe the process for development of a regression model. The KTRLine software was designed to follow this process, which will be summarized herein. The first step is to graph the data to visually inspect the relation or relations between the predictor variable and the response variable. The second step in the development of a regression model is calculation of regression statistics and examination of the calculated values to determine if the estimated model is a reasonable representation of the system of interest. The third step in the regression process is to examine the residuals. The fourth and final step, which is much more important for application of OLS regression than for application of the Kendall-Theil robust line, is to examine the effect of outliers on estimates of the slope and the intercept.

Graphing the data allows the user to assess the linearity of the relation(s) between *X* and *Y*, to examine if the variability in scatter of the *Y* variable above and below the line is constant with increasing *X*, and to identify outliers in the data set (Helsel and Hirsch, 2002). Graphing and visual inspection also provide the best method for assessing if more than one regression relation may be necessary to characterize different relations between discharge and water quality that may occur in different flow regimes (O'Connor, 1976; Glysson, 1987). The KTRLine software provides an initial estimate of a one-line linear model. The user may graph the data with the line, the residuals of the data, or a probability plot of the *X* data, the *Y* data, or the residuals. If a multisegment model is selected, the residuals are calculated by use of each segment of the model in the interval of *X* values defined by application of equation 17. Graphs of the data, the residuals, and probability plots of data and residuals facilitate examination of the data and the regression-model results.

Examination of the graph may indicate if data transformation is necessary. If nonlinearity is a problem, a transformation of the *X* data may be warranted. If, however, nonlinearity and non-constant variance in the *Y* data is present, then transformation of both the *X* and *Y* variables may be necessary. Helsel and Hirsch (2002) describe the ladder of powers (fig. 5) as a method to classify potential transformations and the bulging rule as a method to examine curvature in the data (fig. 6) to choose the appropriate transformation. The KTRLine software provides a method to independently specify transformation of the *X* and *Y* data set within the ladder of powers (Velleman and Hoaglin, 1981) from a reciprocal transformation to a cubed transformation. Logarithmic transformations (natural logarithm or base-10 logarithm) are commonly used for hydrologic data analysis (Helsel and Hirsch, 2002). As each transformation is tested, the user may regraph the data and residuals to find the best regression model.

The KTRLine software automatically calculates the slope, the intercept, and the regression statistics for the model as specified by the user. Examination of the slope and intercept (and the graph of the line with the data points) is necessary to determine if the regression model is a reasonable representation of the relation between *X* and *Y*. For discharge and water-quality data, a positive slope would indicate that concentrations increase with increasing discharge and a negative slope would indicate that concentrations decrease (are diluted) with increasing discharge. The user must examine the data and to determine if the slope is reasonable for the

LADDER OF POWERS

(modified from Velleman and Hoaglin, 1981; Helsel and Hirsch, 2002)

Use	Power	Transformation	Name	Comment
		• • •		higher powers can be used
for (-) skewness	3	x^3	cube	
	2	x^2	square	
	1	x	original units	no transformation
	1/2	$\sqrt{x}$	square root	commonly used
for (+) skewness	1/3	$\sqrt[3]{x}$	cube root	commonly used
	0	log(x)	logarithim	commonly used; holds the place of x^0
	-1/2	$-1/\sqrt{x}$	reciprocal root	minus sign preserves order of observations
	-1	$-1/x$	reciprocal	
	-2	$-1/x^2$	reciprocal square	
		• • •		lower powers can be used

Figure 5. The ladder of powers for use in transforming the independent (*X*) and(or) dependent (*Y*) variables to improve a regression model. (Modified from Helsel and Hirsch, 2002.) All powers except for the reciprocal root and reciprocal square are available in the Kendall-Theil Robust Line software. The line separates transformations for negative (-) and positive (+) skewness.

hydrologic system being examined. For example, a regression model that produces negative values of concentration for reasonable values of discharge may not be considered reasonable for application to hydrologic data (Helsel and Hirsch, 2002). Multisegment models may be used to resolve potential problems with a one-line model. For example, examination of national water-quality data sets indicates that a high percentage of samples commonly are collected during base-flow conditions; this condition presumably reflects the proportion of time that streams flow under these conditions (Smith and others, 1982; Alexander and others, 1997). Therefore, the slope and intercept of a one-line model may be determined by a large number of base-flow measurements and may not indicate the steeper slope(s) commonly detected for rainfall-runoff events (O'Connor, 1976; Glysson, 1987).

Examination of residual plots allows the user to visually assess the predictive capability of a regression model to ensure that the data and model meet assumptions used for the regression analysis (Helsel and Hirsch, 2002). The KTRLine software allows the user to plot the residuals in a probability plot and in relation to the input *X* value or the *Y* value calculated with the regression equation. Examination of the residuals in relation to the input *X* or calculated *Y* value will indicate whether or not there is substantial curvature or

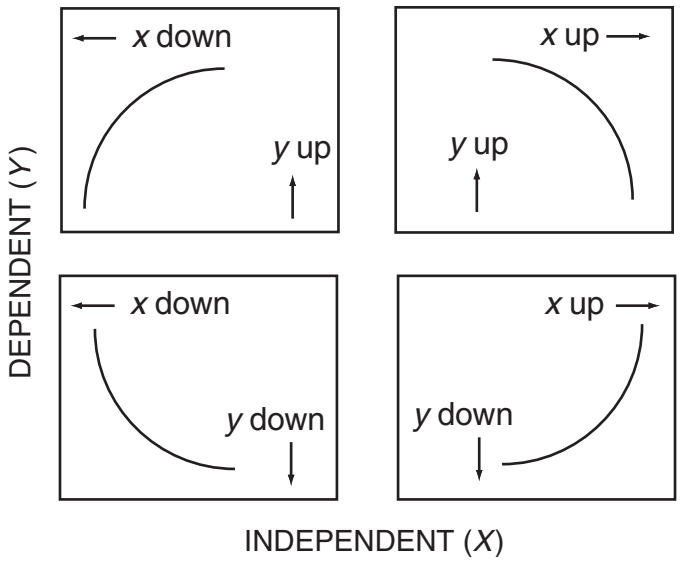

Figure 6. The bulging rule for transforming curvature to linearity. (Modified from Helsel and Hirsch, 2002.)

heteroscedasticity. The user may try different transformations, and if there are multiple processes (such as base flow and stormwater runoff) in the data set, a multi-line model to address these problems in the residual plot. Examination of the probability plot of the residuals will indicate if the residuals are approximately normally distributed above and below the Kendall-Theil robust line.

Outliers, which appear on probability plots as departures from the pattern of the rest of the data, are not uncommon in water-quality data sets (Helsel and Hirsch, 2002). Outliers are commonly omitted by OLS regression users so that the OLS regression line will provide the best fit for the bulk of the data, but these outliers may represent the most important information in a data set (Helsel and Hirsch, 2002). The Kendall-Theil robust line is resistant to outliers in the data set and, therefore, it is not necessary to remove or adjust outliers. The KTRLine software, however, provides a means for identification and documentation of outliers. To identify an outlier, the user can left-click the computer mouse on the point of interest. The KTRLine software provides a dialog box that lists the *X* and *Y* values of the data point, the associated Kendall-Theil robust line prediction, the residual error value, and any metadata stored for that point in the input data file. (It is, therefore, recommended that the user concatenate the date, the time, and station number in the metadata column of the KTRLine input file.) The point-selection dialog box provides the option of saving the point and associated information to the output file, which records the results of analysis. This information may be used to reexamine input data to determine if outliers represent a meaningful process that should (or should not) be included in the data analysis. For example, suspended-sediment concentrations in the rivers draining the Mount St. Helens area in Washington commonly exceeded 100,000 milligrams per liter (mg/L) and periodically exceeded 1,000,000 mg/L after the volcanic eruption in 1980 (Dinehart, 1997). These data are good for characterizing effects of an eruption, but would exceed, by three to six orders of magnitude, the normal sediment concentrations expected for streams in this area of the country.

Use of the KTRLine Software

The KTRLine software is a Visual Basic 6.0 program designed for use in developing regression models of hydrologic data with nonparametric methods. The user interface consists of an Input-Data Specification Form, an Interpretive-Graphing and Model-Specification Form, and a Multisegment Regression-Model Output Form. The KTRLine software, if installed properly, should be compatible with Microsoft operating systems from Windows 98 through Windows XP. At least 1,024 by 768 pixels are needed to display the entire Interpretive Graphing and Model Specification Form at a scale that would be useful for graphical analysis of hydrologic data. The program reads a tab-delimited input file and allows the user to specify, save, and(or) print one or more regression models for each data set. Suspended-sediment concentrations and associated discharge measurements collected from USGS monitoring station 01197500 on the Housatonic River near Great Barrington, MA (Bent, 2000), are available on the computer disk accompanying this report. These data are used for the examples in the following discussion on installation and use of the software.

Installation and Removal

The KTRLine software depends on a number of software drivers and dynamic-link libraries that may not be installed and available on the user's computer. Therefore, this software must be installed by someone with administrative rights on the user's computer. The folder KTInstall on the computer disk accompanying this report contains a readme.txt file with installation instructions and includes the two installation files. The file setup.exe file is the installation program. The setup program is a standard Microsoft installation wizard that should be familiar to the user or their system administrator. All the standard choices for software installation should be followed. The user may uninstall the KTRLine software and its support files by use of the standard Microsoft Windows Add or Remove Programs wizard found on the control panel.

The installation program creates the directory C:\Program Files\KTRLine_V01.0\ and includes the KTRLine software in the computer's registry. The file KTRLv1.CAB is the file that contains the KTRLine software, support files, example data files, and a shortcut to the program. This shortcut may be copied to the user's desktop. The installation package does not, however, put a shortcut on the desktop or the program

bar. If desired, these shortcuts can be added manually upon installation. If the directory C:\Program Files\KTRLine_V01.0\ is not used, then the shortcut provided will need to be modified. Although sample files are saved in the program-files directory, security settings may cause a fatal error if the user attempts to establish an output file in this (or other directories) in which they do not have read, write, and execute rights. In this case, the user may select a different output directory or the system administrator can reset write rights in this directory for the user.

Creating an Input-Data File

The KTRLine software requires a tab-delimited plain-text file as the input-data set. The software reads only the first three columns of any input-data file and will ignore any information beyond a third tab character. The expected format for the input-data file has the independent (X) variable in the first column, the dependent (Y) variable in the second column, and metadata (explanatory information) into the third tab-delimited column. The third (metadata) column is optional, but if a third column is specified there must be at least one character (even if it is just a space character) after the second tab character in every line of the file.

The KTRLine software reads each column and allows the user to specify any of the three columns as X or Y data as long as the entire column contains numeric data. The KTRLine software will not read date formats as an X or Y data set, but regression with dates can be done by converting dates into an equivalent number of Julian days from a common origin. The metadata column may contain any combination of numbers, text, and spaces (with the exception of a tab character) and may be any length. For example, the user may concatenate the station number, date, and time of sample collection from a National Water Information System (NWIS) tab-delimited file to provide metadata for water-quality samples. Example input-data sets are provided in the primary directory on the computer disk documenting the KTRLine software.

The KTRLine software automatically reads the first line of the input file as a header (or explanation line) and the remaining lines as data. Therefore, only one header line is allowed in the input-data set. The input header in each column may contain any number of characters, but only the first 65 characters will appear in the text box on the user interface, and a text-string beyond this length will not be displayed properly on the graph. Therefore, it is recommended that the header text be shorter than 65 characters. If the first line of an input file contains data, the user may specify that the first line has data and enter explanatory text on the KTRLine Input-Data Specification Form. Any data descriptions provided by the user on the Input-Data Specification Form will appear on the graph and will be recorded in the output file. This information, however, will not be added to the input file by the KTRLine software; therefore, it is recommended that a header line be added when the input file is created.

The KTRLine software requires the simple input format described above because the focus of the development effort was to (rapidly) develop a simple and robust analysis tool. Powerful data-handling utilities, such as text editors, spreadsheets, and database software, already are available and can easily be used to create a tab-delimited KTRLine input file by entering and manipulating data. For example, Microsoft Excel or a Microsoft Access query may be used to import a NWIS-Web water-quality file; select two data columns of interest; concatenate station number, date, and time (by use of the ampersand character); and export a KTRLine format text file. The user should ensure that there are no blank lines at the end of the input file. Users may create their own programs or spreadsheet macros to automate the process of creating a KTRLine input file. For example, the NWIS Wizard (NWiz) program, a data-handling utility designed to download NWIS-Web water-quality files, explore the data, and create KTRLine input files, is included on the computer disk documenting the KTRLine software.

Input-Data Specification Form

The Input-Data Specification Form (fig. 7) is the opening form of the KTRLine software and is activated by left-clicking the file "C:\Program Files\KTRLine_V01.0\KTRLv1.exe" or a Windows shortcut to this file on the user's hard drive. The Input-Data Specification Form is the initial program interface; this form provides information about the program and guides the user through the data-specification process. The data-specification process is designed to be a linear progression through the steps necessary to define the input-file characteristics, to specify and initialize the output filename, and to begin the regression analysis. As such, most of the controls on the Input-Data Specification Form are hidden when the form is initialized and are activated in a step-by-step process as they are required in the input-data process. Once a step is complete, each control is disabled so that the user may follow the proper input process. If a mistake is made in the process, the user may close the program at any point and restart the process with only a small investment in time.

Open and Test Input File

The "Open and Test Input File" button activates the file-handling features of the KTRLine software. The user must specify both the input and output filenames. During the file-specification process, the KTRLine software checks for the existence of the input and output files, and does a preliminary examination of the input file. Once the "Open and Test Input File" button is clicked, this button activates a standard file-specification dialog box (commonly used in Microsoft software). This file-specification dialog box is used to navigate to the directory containing input data and to specify the input-data filename. Once a filename is specified, the program checks for the existence of the file. If the file does not exist,

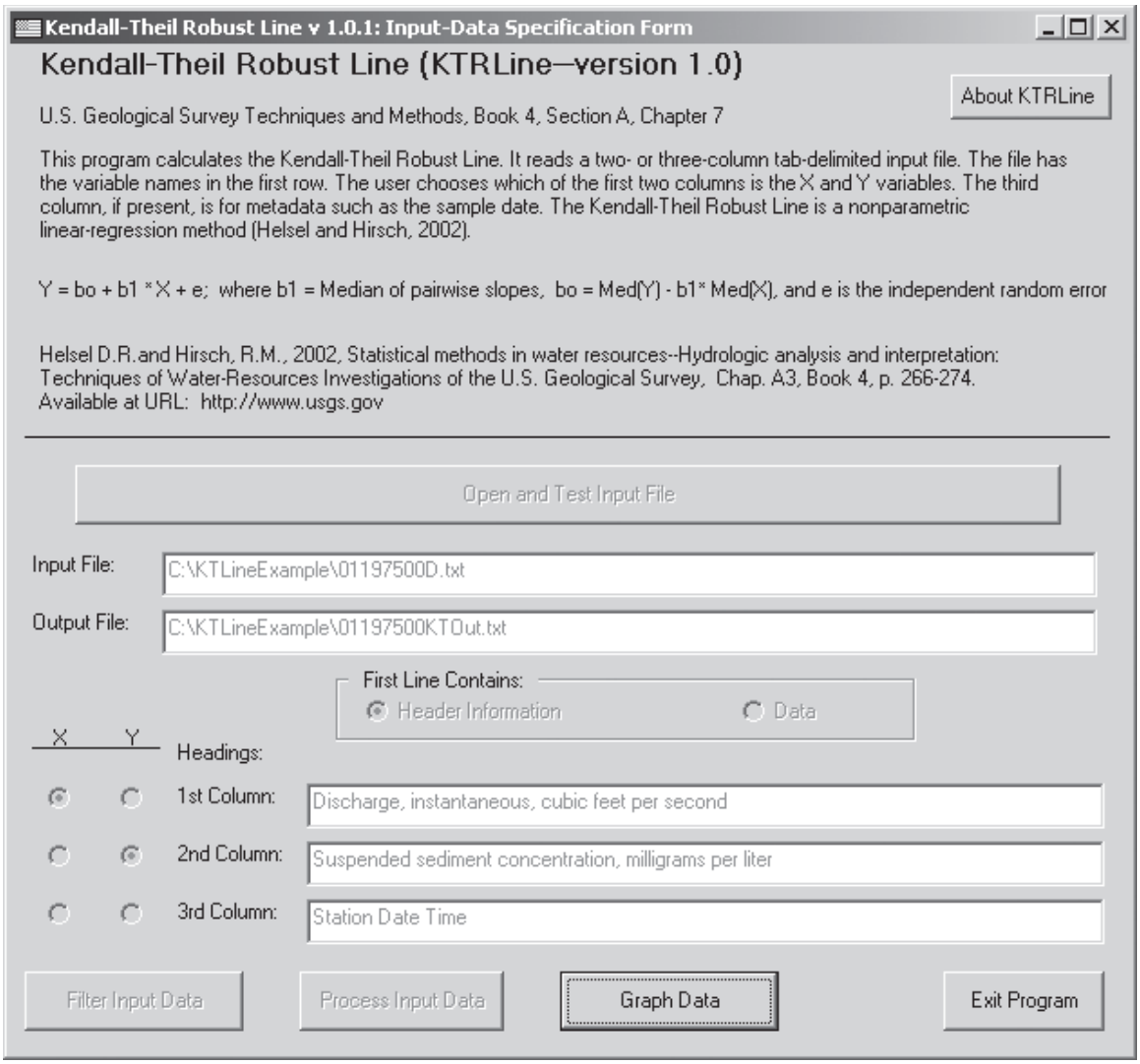

Figure 7. Example of the Kendall-Theil Robust Line Input-Data Specification Form as it appears when the user is preparing to graph the data.

a warning is generated and the input file-specification dialog box is reactivated. The user may choose to cancel the process on the file-specification dialog box. If the user selects cancel, the choice is given to exit the program or restart the file-specification process. The next step is to specify an output filename in the output file-specification dialog box that appears on the screen. If the user selects the input filename as the output file, an error will be generated and the user will be required to respecify the output filename. The program will test for the existence of the output file. If the output file exists, the program will add a sequence-number suffix to the filename. If the user specifies a read-only directory, however, an error will be generated and the user must specify a different directory. As with the input file-specification dialog box, the choice to cancel will result in a message box giving the user the choice to exit the program or restart the entire file-specification process.

Once the user specifies the input and output filenames, the program makes a preliminary analysis of the input-data set. The KTRLine software counts the number of columns and rows, and reads the first line of the data set. If there are less than two tab-delimited columns in the data file, an error is generated and the user must restart the process. If there are more than three tab-delimited columns, a warning is given and the user may proceed with the first three columns in the input data file. If the input file contains two or more columns, the program prints the input and output file paths and names to the form, activates the input-data specification options, and provides information about the input-data set in a series of message boxes.

Input-Data Specification Options

The input-data specification options include a series of *X*-*Y* selection buttons, selection buttons for header-line specifications, and text boxes for specification of explanatory information about each variable in the input-data set (fig. 7). The program provides one option line for each column of data that is recognized in the input file. By default, the first column is assigned as the independent (X) variable, and the second column is assigned as the dependent (Y) variable. The user may reselect any of the available columns as the *X* or *Y* variable, but the program will not allow the user to specify one column for both variables. The user must first select the *X* variable and then select the *Y* variable. By default, the first line of data is read as an explanatory header line. If the first line contains numeric data, the user may select the option indicating that the first line contains data. The user may edit the header information in the text box provided for each active column. This information will be used in the graph and will be saved with the output file. If the input-data specification options are correct, the user may proceed by choosing the "Filter Input Data" button.

Filter Input Data

The "Filter Input Data" button (fig. 7) activates a process in which the input-data specifications are applied to the input file as the data are read into the program. If format errors are detected, a message box will appear informing the user that errors have occurred and that the errors are summarized in an error file. The error file will be saved in the output file directory and will be named KTRLErrFile#.txt (where "#" is the sequence number of the error file in the output-file directory). The error file provides a line-by-line error report, including the line number and type of each error in the input-data set, so that the user may identify the problems and reformat the input data for use with the program. The last line of the error file summarizes the number of problems and the path and filename of the input-data set. If the software detects errors, it reactivates the "Open and Test Input File" button so that the user may restart the process after correcting errors in the input file. If the program reads the input data without error, it disables all file input and header options and activates the "Process Input Data" button.

Process Input Data

The "Process Input Data" button (fig. 7) reads the filtered data, assigns the data to arrays, and calculates initial estimates of the Kendall-Theil robust line statistics. The program assigns the *X* and *Y* data to a two-dimensional array that is then sorted in order of increasing *X* values. A secondary array of *Y* values is assigned and sorted in order of increasing *Y* values. An array (eq. 3) of pairwise slopes is calculated (eq. 2) and sorted. The KTRLine software uses a standard quicksort procedure (Microsoft, 1998). Sorting is a lengthy process that may require as many as N^2 comparisons (where N is the number of points in the data being sorted). Potential problems for processing input data are described in the section on program performance and numerical limitations. A text box becomes visible on the form to indicate the current program status. Large data sets (more than a few hundred points) may take more than a minute to process depending on the size of the data set and the characteristics of the user's computer. The "Process Input Data" button also writes the initial KTRLine statistics to the output file. Once processing is complete, a "Graph Data" button replaces the status box, and the user may proceed with the graphical analysis of the data.

Graph Data

The "Graph Data" button (fig. 7) activates the Interpretive Graphing and Model Specification Form and closes the Input-Data Specification Form. The data and initial estimates of regression statistics are passed to the Graphical Analysis Interface Form by use of public (global) variables within the KTRLine software as it is opened. The Input-Data Specification Form will close automatically as part of this process.

Exit Program

The "Exit Program" button (fig. 7) closes the Input-Data Specification Form and exits the program. If the "Exit Program" button is selected, the user will be prompted to confirm the selection because the "Exit Program" button terminates the program without saving input options. The user may exit the program at each step in the process by use of the "Exit Program" button. If, however, the "Exit Program" button is selected while the program is processing input data, it will not terminate the program immediately. To terminate the program while it is processing data, the user must use the Windows task manager (for example, clicking the Ctrl, Alt, and Delete keys at the same time; selecting the KTRLine_V01.0; and clicking the "End Task" button) .

Interpretive Graphing and Model Specification Form

The Interpretive Graphing and Model Specification Form (fig. 8) includes a graphical display, tools to manipulate the appearance of data on the graph, three tab-strip menus, a summary of one-line regression model statistics, and controls to save information, print the form, and exit the program. The three tab-strip menus labeled "Plot," "Transform," and "Specify," provide the user with menus to plot the data and residuals, transform the data, and specify multisegment regression models, respectively. The Interpretive Graphing and Model Specification Form is designed as an interactive interface so that the user may explore the data and try different options in

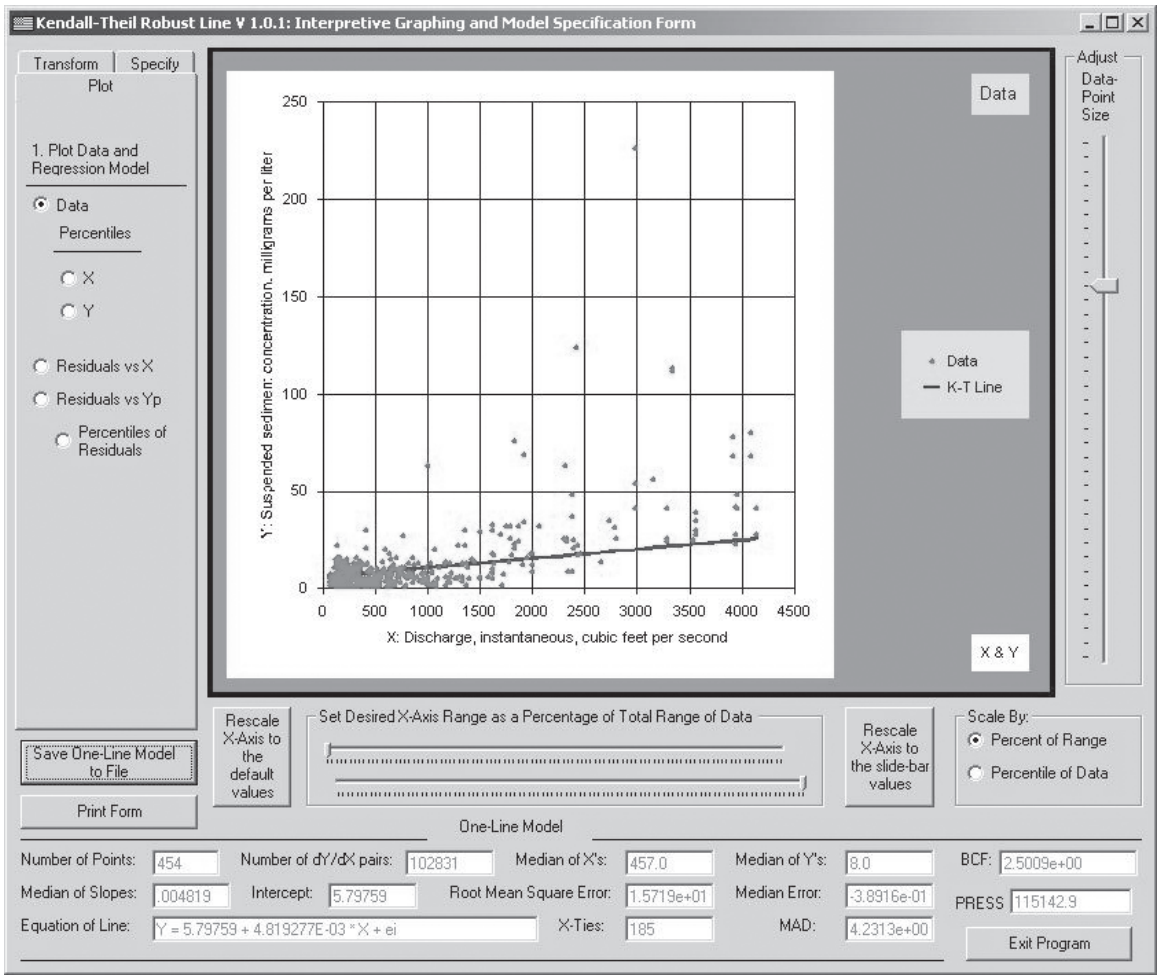

Figure 8. Example of the Kendall-Theil Robust Line Interpretive Graphing and Model Specification Form with the plot menu selected.

an effort to specify the best possible regression model for a given data set.

When the Interpretive Graphing and Model Specification Form is loaded, the input data and the initial KTRLine model is plotted on the graphical display. The one-line regression model statistics also are printed in text boxes at the bottom of the form as it is loaded. The regression model statistics are defined as discussed in the Governing Equations for Kendall-Theil Robust Line Regression section of this report. When the form is loaded, the initial focus is on the "Save One-Line Model to File" button. It is, however, unnecessary to choose this option at this time because KTRLine statistics for the initial model have been saved in the output file.

Graphical-Display Interface

Regression model specification is facilitated by the graphical display and several tools that are designed to manipulate the appearance of data on the graphical-display interface (fig. 8). The graphical-display interface consists of a square graph with the independent (X) values on the horizontal axis and the dependent (Y) values on the vertical axis. The graphical-display interface also includes a graph title on the upper-right corner, an explanation block on the middle-right side, and a transformation-status title on the lower-right side. The heading information read from the input file or the input screen is displayed as the respective X and Y axis titles. The title in the upper-right corner of the graphical display indicates the type of plot (for example, data, residuals, or plotting positions). An explanation block on the middle right of the graphical-display interface indicates what is being plotted on the graph. The transformation-status title in the lower-right corner of the graphical-display interface indicates the transformation status of the X and Y data on the plot. If the data are not transformed, the transformation-status title will read "X & Y."

There are four related tools immediately below the graphical-display interface that control the range of X-axis values displayed. These tools are, from right to left, the "Scale

By:" option box, the "Rescale *X*-axis to the slide-bar values" button, the "Set Desired *X*-axis Range as a Percentage of the Total Range of Data" slide-bar box, and the "Rescale *X*-axis to the default values" button. The first three tools are used in conjunction with each other to graph a subset of the range of the *X*-data points. The slide bars are used to specify the maximum and minimum percentiles of the *X*-axis to be graphed. The top slide bar controls the lower limit of the data on the plot, and the bottom slide bar controls the upper limit of data on the plot; this information is conveyed by the program if the mouse is positioned over the slide-bar control. The "Scale By" options allow the user to specify whether the percentiles selected on the slide bars will be scaled by the percentage of the total range in *X* data, or the percentile-rank of sorted *X* values. The "Rescale *X*-axis to the slide-bar values" button is used to regraph the data once the selections are made. The "Rescale *X*-axis to the default values" button is used to regraph the data to include all *X*-axis data points.

Manipulation of the *X*-axis range can be helpful for data exploration and model specification. For example, visual examination of the model graphed in figure 8 seems to suggest a linear model that underpredicts water-quality values for most of the range (stream discharges from 2,000 to 4,000 cubic feet per second) of the *X*-axis. Use of the "Scale By Percentile of Data" option, however, indicates that 50 percent of the *X* data are associated with a discharge of less than about 460 cubic feet per second; more than 90 percent of the data are associated with a discharge of less than about 2,300 cubic feet per second. The graph of the line with data indicates a better fit for the data below the 90th percentile of discharges (fig. 9). The plot in figure 9 was generated by moving the bottom (upper limit) slide bar to the 90-percent value, selecting the "Percentile of Data" option and clicking the "Rescale *X*-axis to the slide-bar values" button.

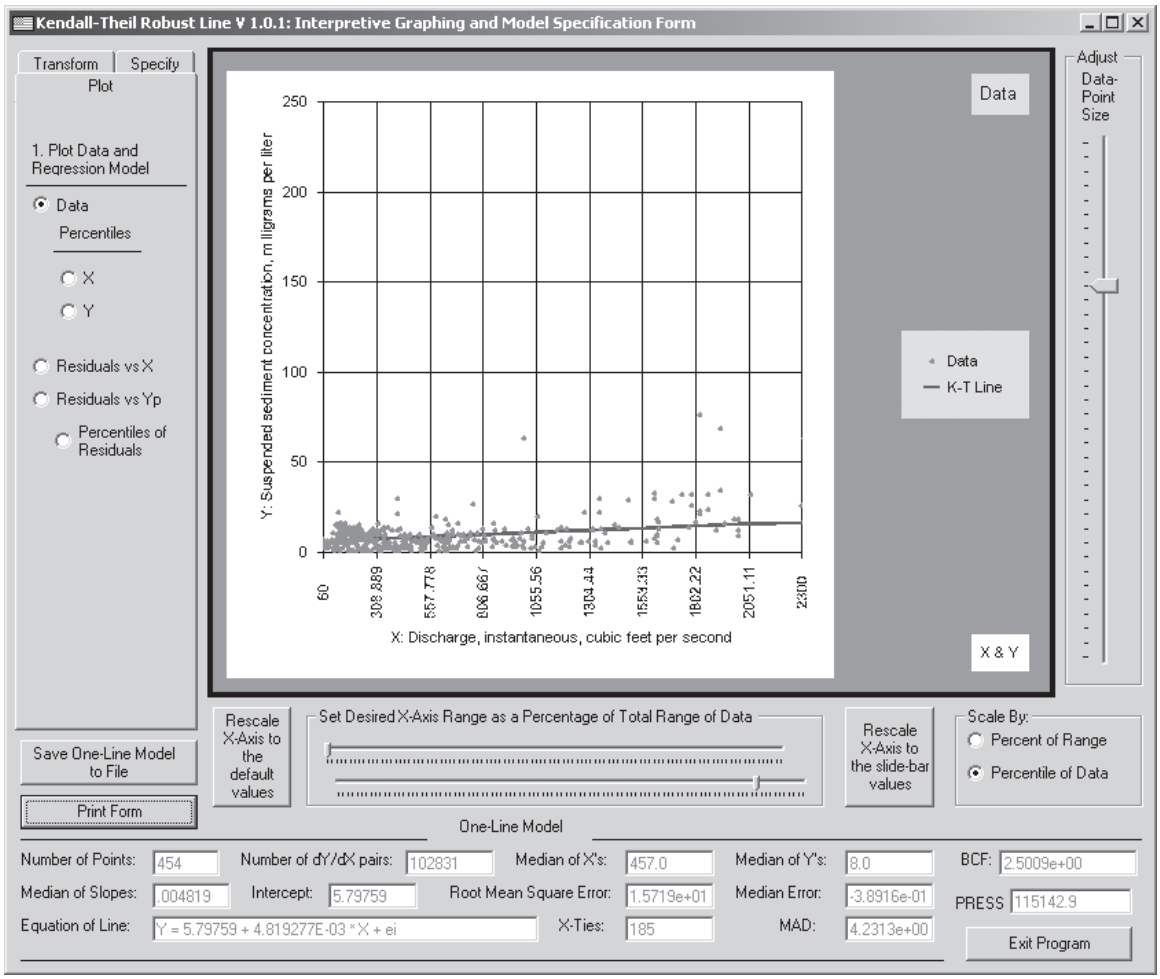

Figure 9. Example of the Kendall-Theil Robust Line Interpretive Graphing and Model Specification Form demonstrating the use of the *X*-axis range tool and the data-point size selector.

The "Adjust Data-Point Size" slide bar on the right side of the Interpretive Graphing and Model Specification Form allows the user to change the size of the data points on the graph. The user may increase the size of the points by moving the slide bar downward on the form. The user may want to increase data-point size to better distinguish between adjacent points with similar but not equal *XY* values. The user may wish to decrease data-point size to better distinguish among a group of points clustered in one part of the graph. The user also may wish to increase or decrease data-point size to facilitate point selection.

Data-Identification Tool

The data-identification tool is not visible on the form, but it is useful for identification and documentation of individual data points on the graph in the graphical-display interface. The data-identification tool is activated by left-clicking on a data point on the graph. The data-identification tool activates a message box that provides information about the data point on the plot. If the graph is a data plot, residuals versus *X* plot, or residuals versus *Y* plot, the data-identification tool provides basic information (fig. 10). This information includes the *X* value, *Y* value, KT line value, *Y*-residual value, and, if it is available, metadata from the input file. If the graph is displaying a probability plot, the data-identification tool provides the percentile of the point selected and the *X*, *Y*, or residual (e_i) value. In any case, the data-identification tool gives the user the choice of whether or not to save the point data to the output file for further analysis. The data-identification tool may be used to identify a group of data (for example, by season, or, in a regional model, by station number) that does not fit the general regression model. This information may be used to segregate data and generate multiple models for different situations. The user, however, must be careful when selecting points to click on the point of interest, because misspecification of the point of interest can easily occur, especially with large data sets with many closely spaced points. Rescaling the *X*-axis may help the user to separate individual points for identification.

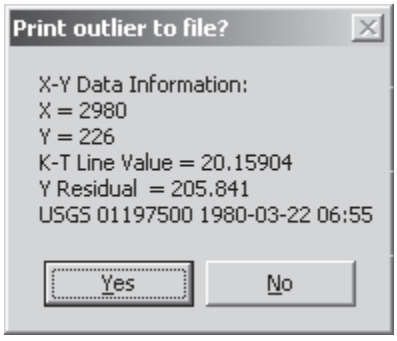

Figure 10. Example of the Kendall-Theil Robust Line data-identification tool message box. The box lists available information about the point of interest and provides the ability to record outliers or other points of interest in the record of analysis.

Plot Tab-Strip Menu

The "Plot" tab-strip menu (fig. 8) is the default tab-strip menu on the left side of the Interpretive Graphing and Model Specification Form. "Plot" is the title of this tab strip. If either the "Transform" or "Specify" tab-strips are active, the user may choose the "Plot" tab-strip menu by left-clicking the "Plot" tab. The "Plot" tab-strip menu has six options. The user may choose to plot data with the regression line, plot the *X* or *Y* data with their respective Cunnane plotting positions, plot the residuals with respect to the *X* data, plot the residuals with respect to the predicted *Y* value, or plot the residuals with respect to their Cunnane plotting positions. An example of a plot of the residuals with respect to the *X* data is shown in figure 11. The residual plots can help the user in the selection of an appropriate transformation and the specification of a multisegment model. An example of a probability plot of the residual values and their ranked percentile values is shown in figure 12. The Cunnane plotting position formula (Helsel and Hirsch, 2002) is used to calculate the percentiles of the data. The ability to instantly view the distribution of *X*, *Y*, and error data also can help the user in the selection of an appropriate transformation and the specification of a multisegment model.

Transform Tab-Strip Menu

The "Transform" tab-strip menu (fig. 13) is a secondary choice on the left side of the Interpretive Graphing and Model Specification Form. "Transform" is the title of this tab strip. The user can independently choose any of the transformations available within the "Ladder of Powers" for the *X* and *Y* data set. Once the transformation is selected, the user must activate the "Transform Data" button to do the actual transformation. At this point, the arrays are transformed and the slopes are recalculated and resorted to find the median value. This process may take more than a minute for large data sets. A progress bar is provided, but the progress for each step is not evenly distributed in time. Once the transformation is complete, the KTRLine program updates all the statistics in the "One-Line Model" section of the Interpretive Graphing and Model Specification Form. The data are replotted with the transformed variables, and the annotation on the graphical display interface is refreshed. The user is prompted to save the one-line model results to the output file. The "Plot" tab-strip menu is automatically selected so that the user may examine the results of the transformation on the data, the line, and the residuals. At any point, the user may retransform the data to find a better regression model. If a multisegment model had been specified, the transformation option would reset the regression to a one-line model. The user, however, may then respecify a multisegment model by use of the "Specify" (multisegment model) tab-strip.

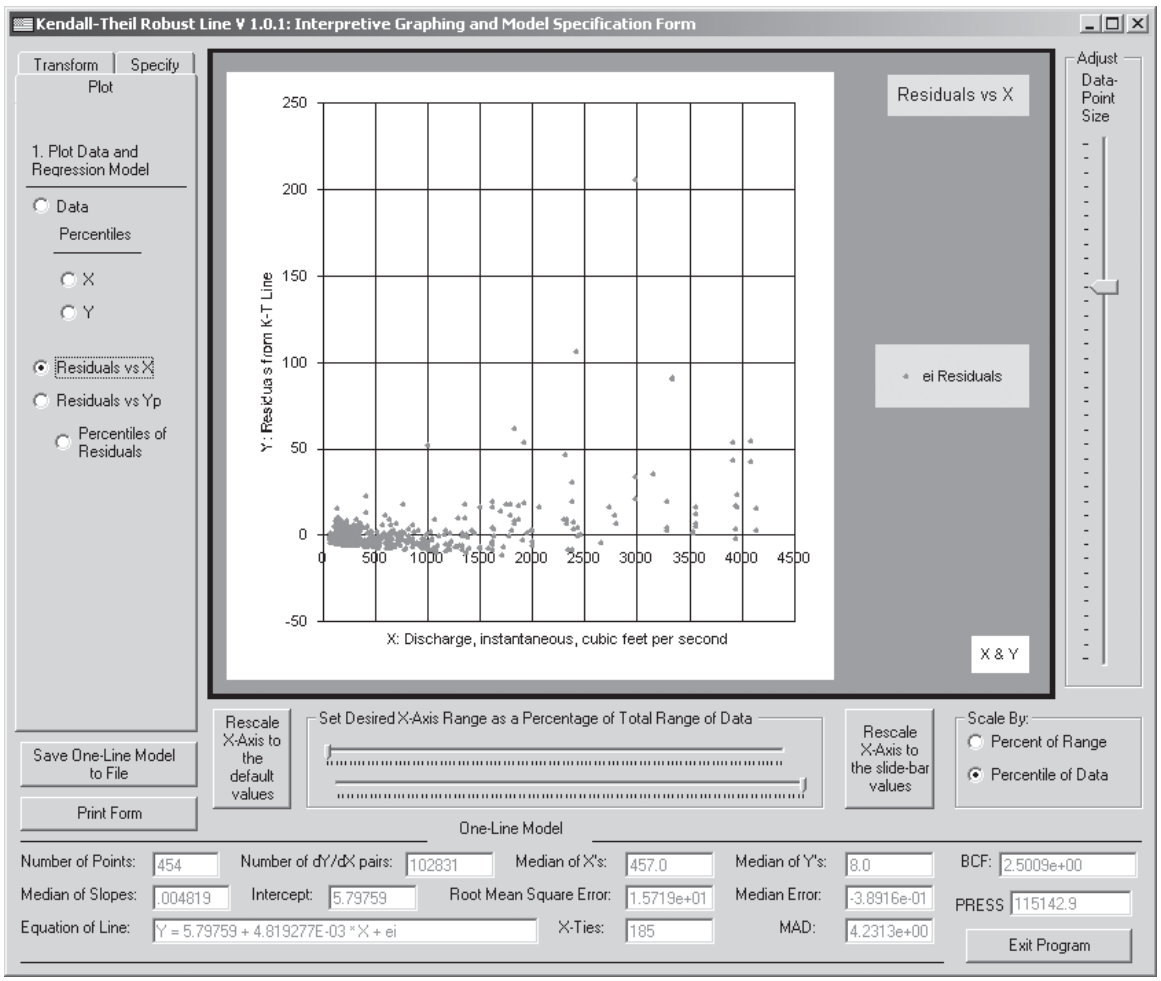

Figure 11. Example of the Kendall-Theil Robust Line Interpretive Graphing and Model Specification Form with a residual plot selected.

Specify (Multisegment Model) Tab-Strip Menu

The "Specify" (Multisegment Model) tab-strip menu (fig. 14) is a secondary choice on the left side of the Interpretive Graphing and Model Specification Form. "Specify" is the title of this tab strip. In the first step, the user can select a model with up to five line segments if there are enough data points in the input-data set. The number of segments available on the Specify Multisegment Model Tab-Strip Menu is one if there are less than 20 points, is two if there are 20–29 points, and increases to five with each additional decade of available data. The number of segments is limited for small data sets to provide stability in slopes and resultant convergence points in multisegment models. The user, however, may specify one or more segments that have fewer than 10 data points as long as the total number of points in the data set exceeds 20.

The next step is to specify the break points between segments. For most options, the number of break points between segments equals the number of segments minus one. The user may specify a multisegment model by using an even interval of *X*-values, by left-clicking data points on the graph, or by manually entering *X*-axis break points between segments in an input box. It is, however, advisable to examine the plotting positions of *X* values and break points in the residuals plots before specifying a multisegment model. The "Enter Break Points" option allows the user to specify overlapping data sets.

If the "Use Even Interval" button is selected, the KTRLine software automatically divides the number of points in the data set by the number of segments selected to calculate the ranks of the break points within the sorted array. Within the even-interval calculations, the program adds an additional point from each adjoining segment to help provide stability in slopes and resultant convergence points.

If the "Select Break Points From Graph" button is selected, the KTRLine software prompts the user to select the break points between segments by left clicking on points

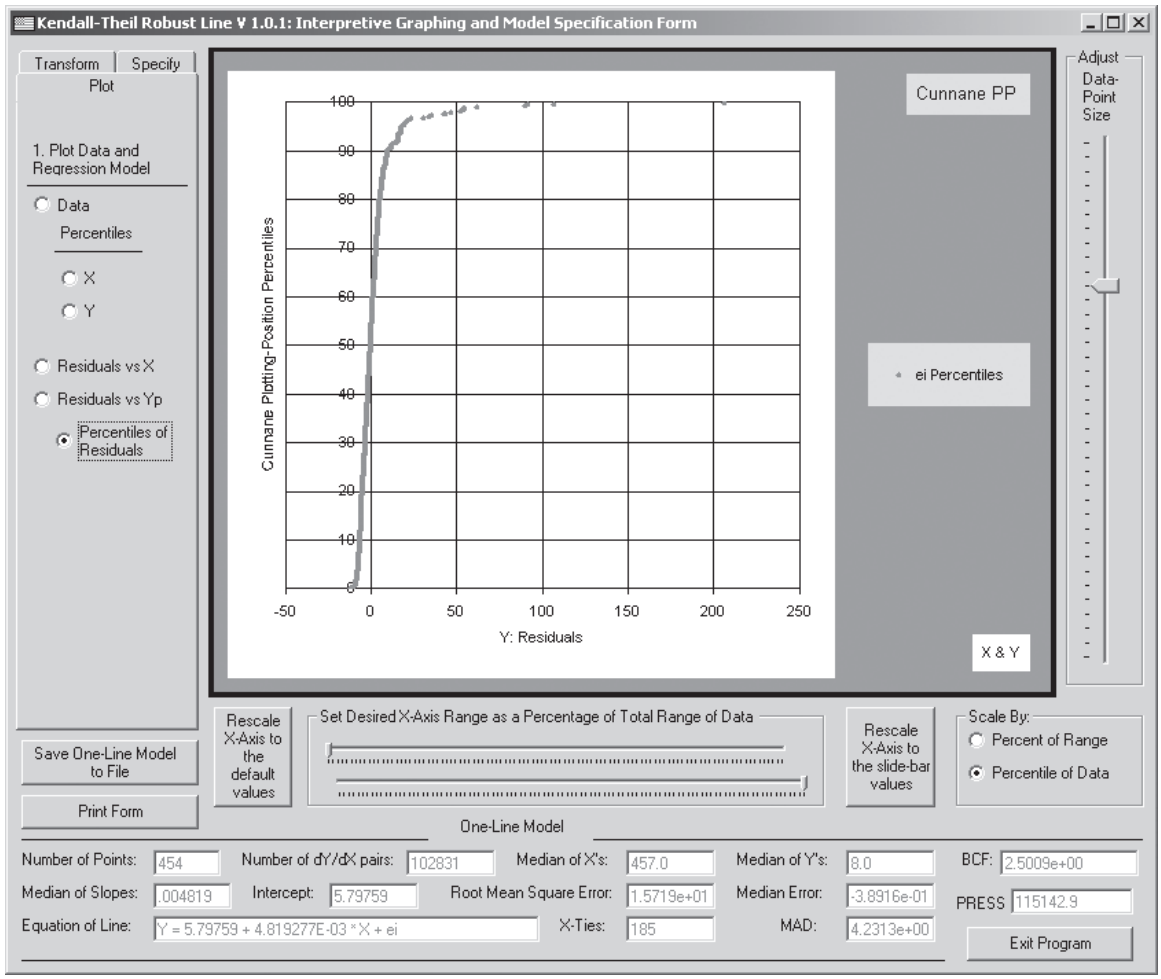

Figure 12. Example of the Kendall-Theil Robust Line Interpretive Graphing and Model Specification Form with a probability plot of the residuals and their ranked percentiles.

on the graph from the lowest break point to the highest break point. The program confirms the selection of each break point by indicating the percentile rank of the *X*-value selected in a message box. The program generates an error message if the *X*-axis value of one break point is less than the previous break point and prompts the user to reselect that point.

The "Enter Break Points" button is different from the other options in that this option allows for the specification of overlapping data intervals to formulate a regression model for each segment. Use of overlapping data intervals is commonly used in transport-curve development (G.D. Glysson, U.S. Geological Survey, written commun., 2005). If the "Enter Break Points" button is selected, the KTRLine software prompts the user to determine if overlapping break points will be used. If the user selects "Yes," the user is prompted to enter the numerical *X*-axis value of each break point in an input box that appears on screen. The program uses an actual data-point value from the input file for each break point. Therefore, the value entered by the user is automatically modified to the next lower *X*-value in the data set. The input box indicates the sequence number of the segments being entered, the maximum value in the data set, the minimum value in the current interval, and, for two or more break points, the actual *X*-value in the data set of the previous break point. If the point entered is less than or equal to the minimum value or it is greater than or equal to the maximum value, the program generates an error and requires that the user reselect the break point. The program generates a confirmation message, which indicates if there are less than 10 data points in a segment of the multisegment model. The Cunnane plotting-position percentiles of the maximum and minimum data points used to calculate each segment also are calculated and presented with the confirmation message.

Once the user selects the number of segments and specifies the break point(s), the user must click the "Generate Multisegment Model" button to proceed.

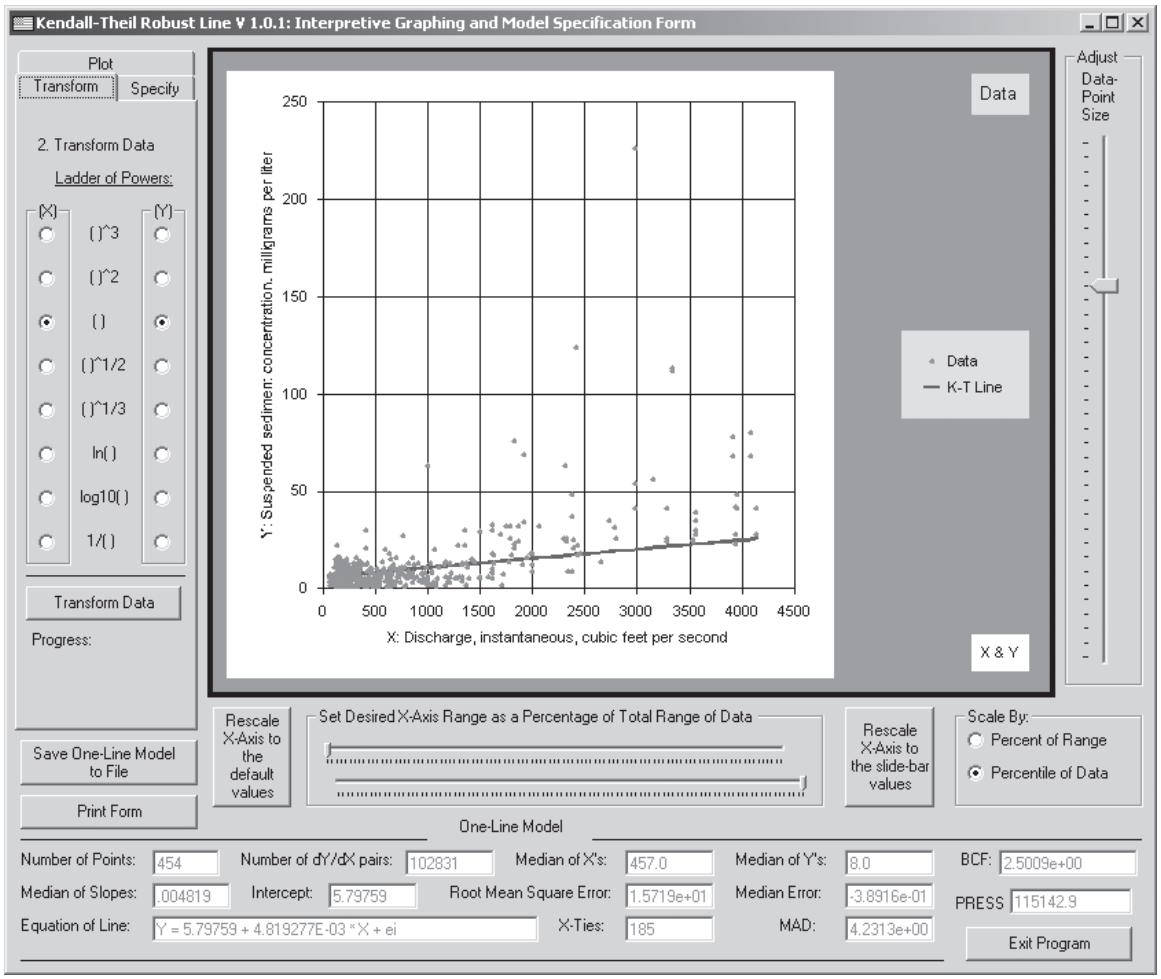

Figure 13. Example of the Kendall-Theil Robust Line Interpretive Graphing and Model Specification Form graphics and analysis screen with the transformation tab-strip menu selected.

When this button is selected, the program does the following tasks:

- Sets up an array for each segment;
- Determines the slope and intercept for each segment based on the points in the specified intervals;
- Calculates the point of convergence between adjacent segments by use of equation 17;
- Calculates regression statistics based on the points that fall within the points of convergence between adjacent segments;
- Resets the graphical interface to show the data and the multisegment model;
- Displays the Plot-Data Tab-Strip menu; and
- Launches the "Multisegment Regression-Model Output" Form.

Multisegment models are plotted on the graphics and analysis screen. An example of a two-line model plotted on the software graphics and analysis screen is shown in figure 15. The user should be aware that the graphical-display interface plots the regression line value for each *X* in the data set and automatically connects each calculated *Y* value with a straight-line segment. This limitation in the graphical-display interface can produce artifacts when graphing multisegment models. If a multisegment model is selected that has a point of convergence that falls between two data points with a perceptible space, the user may see what looks like additional segments in the multisegment model. The graphical-display interface plots the equation of one line segment up to the maximum point, plots the equation of the next line segment starting at the next consecutive data point, and plots a line between these values. The user can draw the actual point of convergence by printing the form and extending each line to their intersection. Conversely, the KTRLine software will not plot segments that do not include data between convergence points that are calculated as being outside the range of the input-data set.

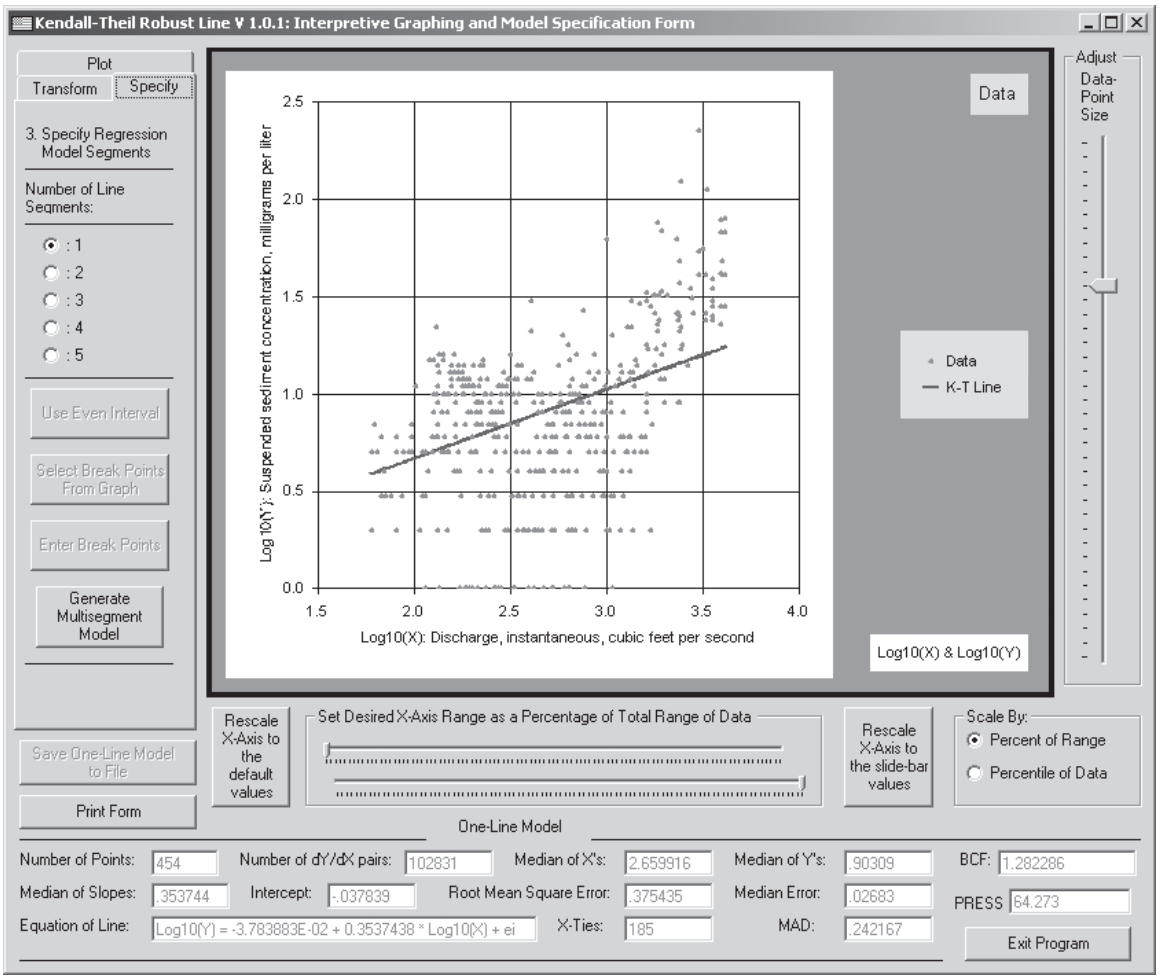

Figure 14. Example of the Kendall-Theil Robust Line Interpretive Graphing and Model Specification Form graphics and analysis screen with the specify (multisegment model) tab-strip menu selected.

Multisegment Regression-Model Output Form

The Multisegment Regression-Model Output Form (fig. 16) displays detailed results of a multisegment model. The Multisegment Regression-Model Output Form (fig. 16) appears in front of the Interpretive Graphing and Model Specification Form (fig. 15) when the regression calculations are complete. The user can move or minimize the Multisegment Regression-Model Output Form to view data and residual plots before saving or printing model results.

The Multisegment Regression-Model Output Form (fig. 16) consists of a text display area and three control buttons. The "Save Model Output" button prints the multisegment model details as they appear in the text display area to the current KTRLine output file. The user may try several transformations or try to optimize a multisegment model by trying different break points and examining the residual error population characteristics. Once the model output has been saved, the button is disabled; this feature ensures that only one copy of each model is saved to the output file. The "Print Model Output" button prints the multisegment model details as they appear in the text display area to the printer through a standard Microsoft Windows print-dialog box. The user also may copy the contents of the text window into a text-processing package. The "Close" button closes the window after confirming whether the user has or wants to save the results of the multisegment regression.

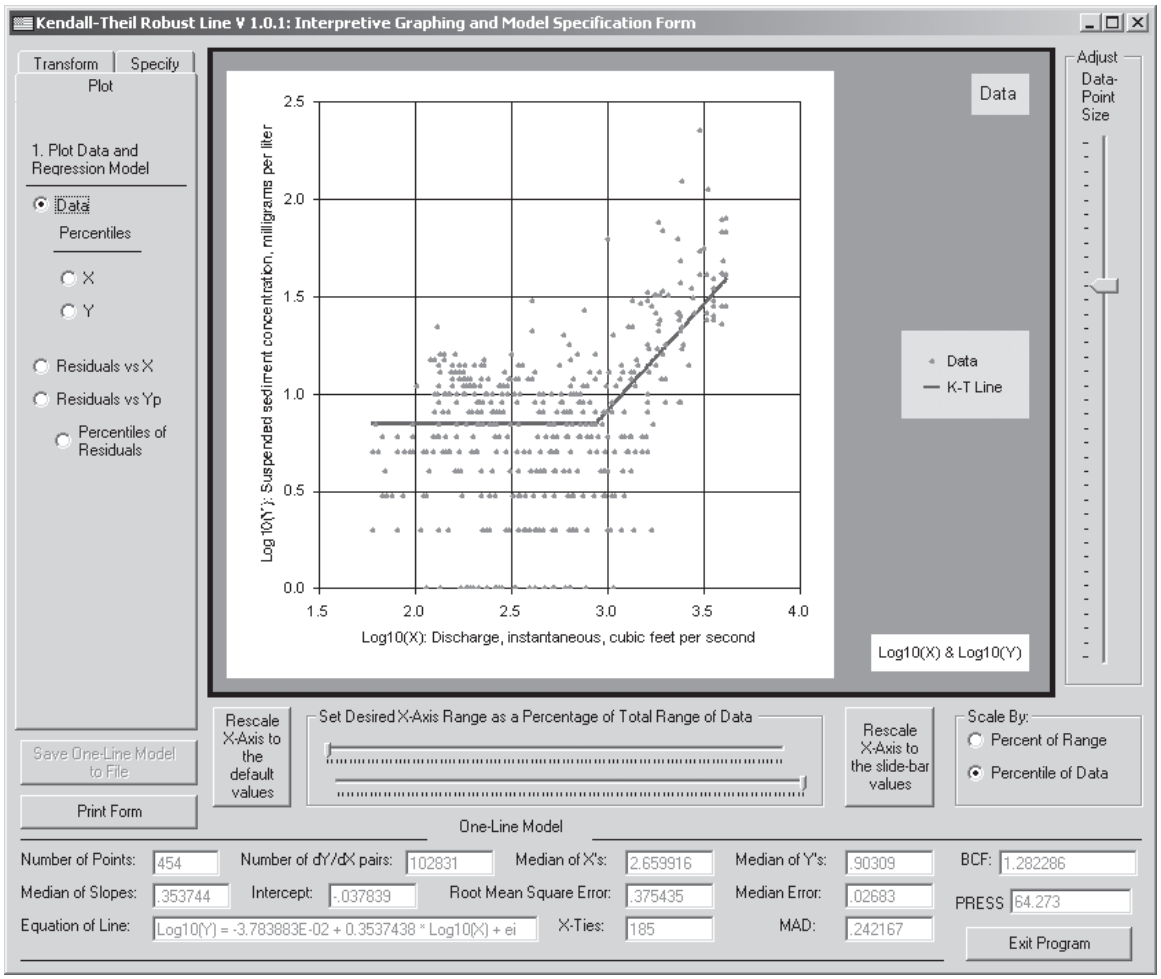

Figure 15. Example of a two-segment model plotted on the Kendall-Theil Robust Line Interpretive Graphing and Model Specification Form.

KTRLine Output-File Format

The KTRLine output-file format (fig. 17) is designed to document the results of analysis and to facilitate evaluation and use of the regression model(s) generated. The KTRLine program automatically prints information about the initial analysis and then provides the user with the opportunity to record results of subsequent analyses if they are deemed useful. The bulk of the information is designed as explanatory information for the user or a reviewer to read. The program, however, prints the data on outliers and the results of the multi-line regression models in a tab-delimited format to facilitate a copy-and-paste operation for use with quantitative spreadsheet or graphics software packages. The KTRLine output-file is in a plain-text ASCII file format that can be easily read or printed with a number of commonly accessible software packages including word processors.

The KTRLine program automatically prints the time and date of the analysis, the input file name, and the title of the independent and dependent variable. Raw input-data statistics and the initial one-line model based on raw input data also are printed to the output file as the values are calculated in the Input-Data Specification Form and graphed on the Interpretive Graphing and Model Specification Form. The program prints this information to help identify the contents of an output file and to record the first regression-model statistics (fig. 17*A*).

Results of a data transformation may be recorded in the output file by clicking the "Save One-Line Model to File" button on the Interpretive Graphing and Model Specification Form (fig. 8). The type of transformation and all relevant regression-line statistics are then printed to the output file (fig. 17*A*). All statistics are printed in the transformed space to better document the results of analysis. The user may then retransform the results of analysis into the original data units.

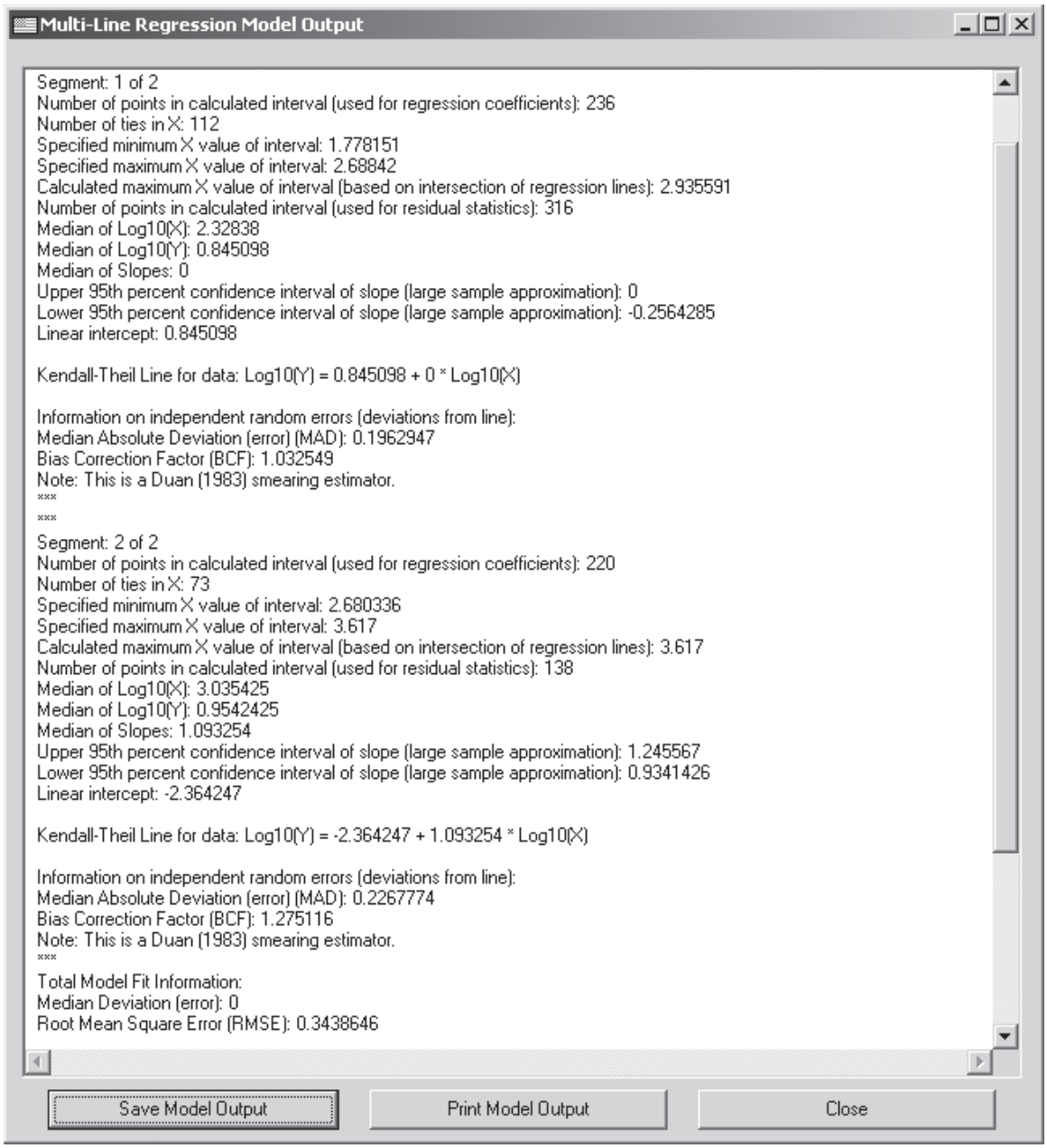

Figure 16. Example of the Kendall-Theil Robust Line multisegment model results screen. This screen provides regression statistics and allows the user to save and print multisegment model results.

A.

```
Kendall-Theil Robust Line (KTRLine--version 1.0) Output
Analysis done on: 2005-10-15 07:56:19
Input File Name: C:\KTLineExample\01197500D.txt
Independent Variable (X): Discharge, instantaneous, cubic feet per second
Dependent Variable (Y): Suspended sediment concentration, milligrams per liter
**********
Linear model all data:
 Number of points: 454 Number of pairs: 102831
 Number of ties in X:       185
 Minimum X:    60
 Maximum X:    4140
 Minimum Y:    1
 Maximum Y:    226
 Median of X's: 457
 Median of Y's: 8
 Median of Slopes: 4.819277E-03
  Upper 95th percent confidence interval of slope (large sample approximation): 5.924672E-03
  Lower 95th percent confidence interval of slope (large sample approximation): 3.743316E-03
 Linear intercept: 5.79759
 Kendall-Theil Line for all linear data: Y = 5.79759 + 4.819277E-03 * X
Information on independent random errors (deviations from line):
 Median Deviation (error):  -0.3891568
 Median Absolute Deviation (error) (MAD):             4.231325
 Root Mean Square Error (RMSE):          15.71856
 NonParametric PRediction Error Sum of Squares (NPPRESS):             115142.9
 Bias Correction Factor (BCF):           2.500924
Note: This is a Duan (1983) smearing estimator.
**********
Model all data with: (Log10(X)) & (Log10(Y)) with one line
 Number of points: 454 Number of pairs: 102831
 Number of ties in X:       185
 Median of (Log10(X)): 2.659916
 Median of (Log10(Y)): 0.90309
 Median of Slopes: 0.3537438
  Upper 95th percent confidence interval of slope (large sample approximation): 0.4299055
  Lower 95th percent confidence interval of slope (large sample approximation): 0.2752349
 Linear intercept: -3.783883E-02
 Kendall-Theil Line for all data: (Log10(Y)) = -3.783883E-02 + 0.3537438 * (Log10(X))
Information on independent random errors (deviations from line):
 Median Deviation (error):   2.682987E-02
 Median Absolute Deviation (error) (MAD):             0.2421672
 Root Mean Square Error (RMSE):          0.3754346
NonParametric PRediction Error Sum of Squares (NPPRESS):             64.2733
 Bias Correction Factor (BCF):           1.282286
Note: This is a Duan (1983) smearing estimator.
**********
Tab Delimited information for Export:
Yvar          XVar  Segments Line    Intercept           Slope       MAD           MaxX   Number of Points
(Log10(Y))  (Log10(X)) 1   1       -3.783883E-02    0.3537438  0.2421672     3.617    454
```

Figure 17. Example of a Kendall-Theil Robust Line output file including *A,* information about the analysis and results of the preliminary regression line and information about the results of a regression of the log-transformed data; and *B,* information about the results of a multisegment regression of the log-transformed data.

B.

```
**********
Model all data with: Log10(X) & Log10(Y) with 2 segment(s)
***
Segment: 1 of 2
Number of points in calculated interval (used for regression coefficients): 236
Number of ties in X: 112
Specified minimum X value of interval: 1.778151
Specified maximum X value of interval: 2.68842
Calculated maximum X value of interval (based on intersection of regression lines): 2.935591
Number of points in calculated interval (used for residual statistics): 316
Median of Log10(X): 2.32838
Median of Log10(Y): 0.845098
Median of Slopes: 0
Upper 95th percent confidence interval of slope (large sample approximation): 0
Lower 95th percent confidence interval of slope (large sample approximation): -0.2564285
Linear intercept: 0.845098

Kendall-Theil Line for data: Log10(Y) = 0.845098 + 0 * Log10(X)

Information on independent random errors (deviations from line):
Median Absolute Deviation (error) (MAD): 0.1962947
Bias Correction Factor (BCF): 1.032549
Note: This is a Duan (1983) smearing estimator.
***
***
Segment: 2 of 2
Number of points in calculated interval (used for regression coefficients): 220
Number of ties in X: 73
Specified minimum X value of interval: 2.680336
Specified maximum X value of interval: 3.617
Calculated maximum X value of interval (based on intersection of regression lines): 3.617
Number of points in calculated interval (used for residual statistics): 138
Median of Log10(X): 3.035425
Median of Log10(Y): 0.9542425
Median of Slopes: 1.093254
Upper 95th percent confidence interval of slope (large sample approximation): 1.245567
Lower 95th percent confidence interval of slope (large sample approximation): 0.9341426
Linear intercept: -2.364247

Kendall-Theil Line for data: Log10(Y) = -2.364247 + 1.093254 * Log10(X)

Information on independent random errors (deviations from line):
Median Absolute Deviation (error) (MAD): 0.2267774
Bias Correction Factor (BCF): 1.275116
Note: This is a Duan (1983) smearing estimator.
***
Total Model Fit Information:
Median Deviation (error): 0
Root Mean Square Error (RMSE): 0.3438646
NonParametric PRediction Error Sum of Squares (NPPRESS): 53.87493
**********
Tab Delimited information for Export:
Yvar    XVar      Segments Line   Intercept       Slope     MAD          MaxX     Number of Points
Log10(Y)  Log10(X) 2    1      0.845098        0         0.1962947    2.935591 316
Log10(Y)  Log10(X) 2    2      -2.364247       1.093254  0.2267774    3.617    138
```

Figure 17—Continued. Example of a Kendall-Theil Robust Line output file including *A,* information about the analysis and results of the preliminary regression line and information about the results of a regression of the log-transformed data; and *B,* information about the results of a multisegment regression of the log-transformed data.

An example of the KTRLine program output for a two-line log-transformed model is printed in figure 17*B*. The two-line model results include the specified data intervals and the data intervals calculated by use of equation 17. The information about each segment of the model is followed by the total-model fit information, which includes the median deviation (or error), a nonparametric estimate of the PRESS statistic, and the RMSE (fig. 17*B*). These statistics indicate the combined fit for all data of the multi-line model. The tab-delimited information for export is the basic information necessary to implement the regression model generated by the KTRLine program in a spreadsheet, in graphing software, or in a Monte Carlo model.

Program Performance and Numerical Limitations

The nonparametric techniques that are the basis for the KTRLine program impose a computational burden that affects the performance of the program and limits the maximum size of the data sets to be processed. The KTRLine program calculates the equation of a line by calculating the median of all pairwise slopes, the *X* data, the *Y* data, and the residuals. Calculating the median requires that each array be sorted, which is a computationally intensive process. Manipulation of the pairwise-slope array is the limiting factor that controls the processing speed and the maximum data-set size because of the relation between the number of points in the data set and the number of pairwise slopes (eq. 3). For example, a data set with 1,000 data points has 499,500 pairwise slopes. Sorting the slope array for this data set may take as many as about 0.25 trillion computer operations because the sorting operation may require as many as N^2 comparisons (where N is the number of points in the data being sorted).

The maximum number of data points that may be processed is about 15,000 points (about 112 million pairwise slope values). If the maximum array size is exceeded, the program will produce an error message and the program must be terminated. Visual Basic 6.0 will handle a process with an array size of 500 million double-precision decimal numbers (Microsoft Corporation, 1998). The KTRLine program, however, is designed to maintain the original and transformed arrays of *X* points, *Y* points, pairwise slopes, and residuals. The graph object also requires an *XY* array for the points being graphed and an array for the line objects. Examination of water-quality data sets retrieved from the NWIS Web in 2004 indicates that most surface-water-quality monitoring stations do not contain extremely large data sets for the water-quality constituents of interest. The percentage of sites in NWIS with more than 100 measurements of instantaneous discharge paired with water-quality-constituent concentrations was relatively low. For example, the percentage of sites that had more than 100 paired measurements was about 8 percent with total hardness, 9 percent with total nitrogen, 20 percent with total phosphorus, and 34 percent with suspended sediment. Maximum sample sizes at individual monitoring stations for paired measurements of instantaneous discharge with water-quality-constituent concentrations were 1,422 with total hardness, 2,390 with total nitrogen, 10,308 with total phosphorus, and 26,934 with suspended sediment. Only 3 of 7,477 suspended-sediment monitoring stations in the database contained more than 15,000 samples, and about 90 percent of these stations contained fewer than 500 paired measurements of instantaneous discharge and suspended-sediment concentration.

Substantial redesign efforts would be required to optimize array management within the program so that input-data-set sizes could be increased to a maximum of about 31,000 double-precision decimal numbers. This increase in allowable array size, however, would be accompanied by a large performance penalty because individual arrays would have to be written to disk or recalculated and resorted at each step of the analysis. Use of the computer disk to swap arrays in and out of memory also would perceptibly increase processing time.

The time required to complete processing, transformation, and multisegment modeling efforts slows perceptibly once the data-set size exceeds about 1,000 data points. Results of performance-time experiments indicate that processor clock speed rather than random access memory (RAM) controls KTRLine processing time. For example, a computer with a 1.99-gigahertz processor speed will process about 150,000 pairwise slopes per second, whereas a computer with a processing speed of about 0.5 gigahertz will process about 30,000 pairwise slopes per second. The results of a number of test runs that were done to illustrate relations between the number of samples and the time required to process input data for different computers are shown in figure 18. The time required to transform the input-data set is equivalent to the initial processing time, and the time necessary to calculate statistics for a multi-line model is about 50 to 75 percent of the initial processing time. Therefore, the minimum time required to process, transform, and generate a multi-line model is about 3 times the number of seconds shown on figure 18. For example, it is estimated that a user with a computer with a 1.99-gigahertz processor would use about 13 minutes of computer processing time to generate a transformed multi-line model for a data set with 10,000 points (or about an hour of computer processing time for a user with a 0.5-gigahertz processor).

It is important to note that the KTRLine program will use all available computer resources. It is advisable to close other programs when analyzing very large data sets (over 5,000 points) to facilitate analysis. Computer processing unit (CPU) usage will be at or near 100 percent while the KTRLine program is processing data. If the user invokes the Windows task manager during processing, the task manager may indicate that the KTRLine program is not responding. The KTRLine program, however, will complete the process if the data set does not exceed the maximum array size (as indicated by the aforementioned error message) as long as the program is allowed to continue until completion. Therefore, the user may wish to use the example data sets provided on the accompanying CD-ROM to become familiar with the operation of the program before using the program for large data sets.

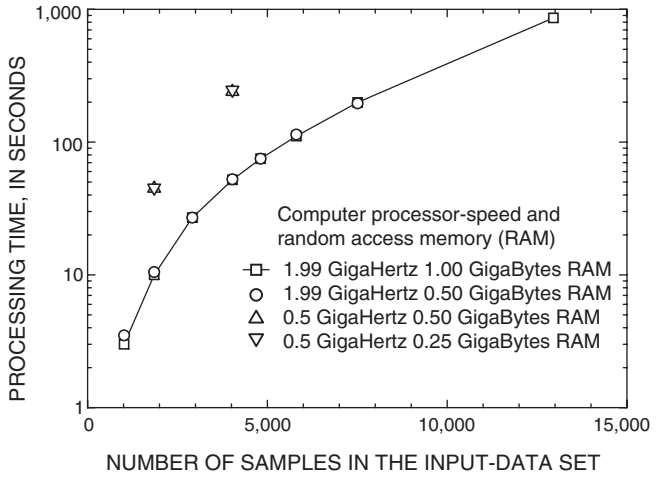

Figure 18. Graph showing relations between the number of samples in the input-data set and processing time in seconds from experiments with four different computers.

Summary and Conclusions

The Kendall-Theil Robust Line software (KTRLine—version 1.0) is a Visual Basic program that may be used with the Microsoft Windows operating system to calculate parameters for robust, nonparametric estimates of linear-regression coefficients between two continuous variables. The KTRLine software was developed by the U.S. Geological Survey, in cooperation with the Federal Highway Administration, to analyze local, regional, and national hydrologic data sets. The software was developed for data generation in support of a stochastic empirical loading and dilution model for planning-level estimates of the effects of highway runoff on the quality of receiving waters. The KTRLine software was developed to facilitate development of water-quality transport curves, for analysis of surrogate-parameter relations among water-quality constituents, for analysis of runoff coefficients, and for assessment of the effectiveness of structural best management practices. The Kendall-Theil Robust Line software is designed for stochastic analysis of hydrologic data because it provides a robust regression equation with information about the random variation in data above and below the regression line.

The Kendall-Theil robust line method was selected for this application because this nonparametric method is resistant to the effects of outliers and nonnormality in residuals that characterize water-quality data. The program was developed because nonparametric multisegment regression tools are not available with software commonly used for development of regression models. The slope of the line is calculated as the median of all possible pairwise slopes between points. The intercept is calculated so that the line will run through the median of input data. A single-line model or a multisegment model may be specified. Regression statistics, such as the median error, the median absolute deviation, the prediction error sum of squares, the root mean square error, the confidence interval for the slope, and the bias correction factor for median estimates, are calculated by use of nonparametric methods. Because the Kendall-Theil robust line is a nonparametric median line, the regression equation may underestimate total mass, volume, or loads. If the random error component or the bias correction factor is incorporated into the estimate, however, the Kendall-Theil robust line method may be used for estimates of total mass, volume, or loads.

The KTRLine software was designed to help the user follow the standard process for development of a regression model. The first step is to graph the data to visually inspect the relation or relations between the independent variable (X) and the dependent variable (Y). The second step in the development of a regression model is calculation of regression statistics and examination of the calculated values to determine if the estimated model is a reasonable representation of the system of interest. The third step in the regression process is to examine the residuals. The fourth and final step is to examine the effect of outliers on estimates of the slope and the intercept.

The program provides a relatively simple graphical user interface for development of regression models. The program is used to read a two- or three-column tab-delimited input file with variable names in the first row and data in subsequent rows. The user may choose the columns that contain the independent (X) and dependent (Y) variable. A third column, if present, may contain metadata such as the sample-collection location and date. The program screens the input files and plots the data. The KTRLine software is a graphical tool that facilitates development of regression models by use of graphs of the regression line with data, the regression residuals (with X or Y), and percentile plots of the cumulative frequency of the X variable, Y variable, and the regression residuals. The user may individually transform the independent and dependent variables to reduce heteroscedasticity and to linearize data. The program plots the data and the regression line. The program also prints model specifications and regression diagnostics to the screen. The user may save and print the regression results. The program can accept data sets that contain up to about 15,000 XY data points, but because the program must sort the array of all pairwise slopes, the program may be perceptibly slow with data sets that contain more than about 1,000 points.

Acknowledgments

The author thanks Gardener C. Bent, G. Douglas Glysson, and John R. Gray of the U.S. Geological Survey for the benefit of their experiences on the process for developing water-quality-transport curves. Dennis R. Helsel, Timothy A. Cohn, and Gary D. Tasker of the U.S. Geological Survey

provided input and information on application of regression statistics. Carl S. Carlson, G. Douglas Glysson, Dennis R. Helsel, and Becca S. Sniderman of the U.S. Geological Survey and Patricia A. Cazenas of the Federal Highway Administration, provided thoughtful and thorough technical reviews of this report, the associated CD-ROM, and the KTRLine software.

References Cited

Alexander, R.B., Slack, J.R., Ludtke, A.S., Fitzgerald, K.K., Schertz, T.L., Briel, L.I., Buttleman, K.P., 1997, Data from selected U.S. Geological Survey National Stream Water-Quality Networks (WQN): U.S. Geological Survey Digital Data Series 37, CD-ROM.

Barrett, M.E., 2005, Performance comparison of structural stormwater best management practices: Water Environment Research, v. 77, no. 1, p. 78–86.

Bent, G.C., 2000, Suspended-sediment characteristics in the Housatonic River Basin, western Massachusetts and parts of eastern New York and northwestern Connecticut, 1994–96: U.S. Geological Survey Water-Resources Investigations Report 2000-4059, 121 p.

Brauner, J.S., 1997, Nonparametric estimation of slope—Sen's method, *in* Gallagher, Daniel (ed.), Environmental sampling and monitoring primer, accessed on June 30, 2004, at URL http://www.cee.vt.edu/

Charbeneau, R.J., and Barrett, M.E., 1998, Evaluation of methods for estimating stormwater pollutant loads: Water Environment Research, v. 70, no. 7, p. 1295–1302.

Clarke, R.T., 1990a, Statistical characteristics of some estimators of sediment and nutrient loadings: Water Resources Research, v. 26, no. 9, p. 2229–2233.

Clarke, R.T., 1990b, Bias and variance of some estimators of suspended sediment load: Hydrological Sciences Journal, v. 35, no. 3, p. 253–261.

Cohn, T.A., 1995, Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers: U.S. National Report to International Union of Geodesy and Geophysics 1991–1994, Reviews of Geophysics, Supplement, v. 33, p. 1117–1123.

Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., and Summers, R.M., 1992, The validity of a simple log-linear model for estimating fluvial constituent loads—An empirical study involving nutrient loads entering Chesapeake Bay: Water Resources Research, v. 25, no. 9, p. 2353–2363.

Conover, W.L., 1980, Practical nonparametric statistics, 2d ed.: New York, John Wiley and Sons, 493 p.

Crawford, C.G., 1991, Estimation of suspended-sediment rating curves and mean suspended-sediment loads: Journal of Hydrology, v. 129, p. 331–348.

Cunnane, C., 1978, Unbiased plotting positions—A review: Journal of Hydrology, v. 37, p. 205–222.

Dietz, E.J., 1987, A comparison of robust estimators in simple linear regression: Communication in Statistics-Simulation, v. 16, p. 1209–1227.

Dinehart, R.L., 1997, Sediment transport at gaging stations near Mount St. Helens, Washington, 1980–90, Data Collection and Analysis: U.S. Geological Survey Professional Paper 1573, 111 p.

Driscoll, E.D., Shelley, P.E., and Strecker, E.W., 1990, Pollutant loadings and impacts from highway stormwater runoff volume III—Analytical investigation and research report: Federal Highway Administration Final Report FHWA-RD-88-008, 160 p.

Driver, N.E., and Tasker, G.D., 1990, Techniques for estimation of storm-runoff loads, volumes and selected constituent concentrations in urban watersheds in the United States: U.S. Geological Survey Water-Supply Paper 2363, 44 p.

Duan, Naihua, 1983, Smearing estimate—A nonparametric retransformation method: Journal of the American Statistical Association, v. 78, no. 383, p. 605–610.

Gilroy, E.J., Hirsch, R.M., and Cohn, T.A., 1990, Mean-square error of regression-based constituent transport estimates: Water Resources Research, v. 26, no. 9, p. 2069–2077.

Glysson, G.D., 1987, Sediment-transport curves: U.S. Geological Survey Open-File Report 87-218, 47 p.

Helsel D.R., and Hirsch, R.M., 2002, Statistical methods in water resources—Hydrologic analysis and interpretation: Techniques of Water-Resources Investigations of the U.S. Geological Survey, chap. A3, book 4, 510 p.

Hirsch, R.M., Alexander, R.B., and Smith, R.A., 1991, Selection of methods for the detection and estimation of trends in water quality: Water Resources Research, v. 27, no. 5, p. 803–813.

Hirsch, R.M., Helsel, D.R., Gilroy, E.J., and Cohn, T.A., 1992, Chapter 17—Statistical analysis of hydrologic data, *in* Maidment, D.R., (ed.), Handbook of Hydrology: New York, McGraw-Hill Book Company, p. 17.1–17.55.

Hirsch, R.M., Slack, J.R., and Smith, R.A., 1982, Techniques of trend analysis for monthly water quality data: Water Resources Research, v. 18, no. 1, p. 107–121.

House, W.A., and Warwick, M.S., 1998, Hysteresis of the solute concentration discharge relationship in rivers during storms: Water Research, v. 32, no. 8, p. 2279–2290.

Hussain, S.S., and Sprent, P., 1983, Nonparametric regression: Journal of the Royal Statistical Society, Series A, v. 146, p. 182–191.

Kendall, M.G., 1938, A new measure of rank correlation: Biometrika, v. 30, p. 81–93.

Koch, R.W., and Smillie, G.M., 1986, Bias in hydrologic prediction using log-transformed regression models: Water Resources Bulletin, v. 22, no. 5, p. 717–723.

Martin, E.H., and Smoot, J.L., 1986, Constituent-load changes in urban stormwater runoff routed through a detention pond-wetlands system in central Florida: U.S. Geological Survey Water-Resources Investigations Report 85-4310, 74 p.

Microsoft Corporation, 1998, Microsoft Office 2000/Visual Basic Programmer's Guide: Redmond, WA, Microsoft Corporation, accessed on January 1, 2005, at URL http://msdn.microsoft.com/

Miller, C.R., 1951, Analysis of flow-duration, sediment-rating curve method of computing sediment yield: U.S. Bureau of Reclamation unnumbered report, 65 p.

Nash, D.B., 1994, Effective sediment-transporting discharge from magnitude-frequency analysis: Journal of Geology, v. 102, p. 79–95.

Nevitt, Jonathan, and Tam, H.P., 1998, A comparison of robust and nonparametric estimators under the simple linear regression model: Multiple linear regression viewpoints, v. 25, p. 54–69.

O'Connor, D.J., 1976, The concentration of dissolved solids and river flow: Water Resources Research, v. 12, no. 2, p. 279–294.

Schueler, T.R., 1987, Controlling urban runoff: A practical manual for planning and designing urban BMP's: Metropolitan Washington Council of Governments, Department of Environmental Programs, Washington, D.C., 275 p.

Sen, P.K., 1968, Estimates of the regression coefficient based on Kendall's tau: Journal of the American Statistical Association, v. 63, p. 1379–1389.

Simon, Andrew, 1989, The discharge of sediment in channelized alluvial streams: Water Resources Bulletin, v. 25, no. 6, p. 1177–1188.

Smith, R.A., Hirsch, R.M., and Slack, J.R., 1982, A study of trends in total phosphorus measurements at stations in the NASQAN network: U.S. Geological Survey Water-Supply Paper 2190, 34 p.

Syvitski, J.P., Morehead, M.D., Bahr, D.B., and Mulder, T., 2000, Estimating fluvial sediment transport—The rating parameters: Water Resources Research, v. 36, p. 2747–2760.

Tasker, G.D., and Granato, G.E., 2000, Statistical interpretation of local, regional, and national highway runoff and urban stormwater data: U.S. Geological Survey Open-File Report 00-491, 59 p.

Theil, H., 1950, A rank-invariant method of linear and polynomial regression analysis, Part 3: Proceedings of Koninalijke Nederlandse Akademie van Weinenschatpen A.53, p. 1397–1412.

Thomson, N.R., McBean, E.A., and Mostrenko, I.B., 1996, Prediction and characterization of highway stormwater runoff quality: Research and Development Branch, Ministry of Transportation, Ontario, Canada, 98 p.

Thomson, N.R., McBean, E.A., Snodgrass, W., and Monstrenko, I.B., 1997, Highway stormwater runoff quality—Development of surrogate parameter relationships: Water, Air, and Soil Pollution, v. 94, p. 307–347.

U.S. Geological Survey, 2004, National Water Information System (NWIS) Web data for the Nation, accessed on May 31, 2004, at URL http://waterdata.usgs.gov/nwis/

Velleman, P.F., and Hoaglin, D.C., 1981, Applications, basics, and computing of exploratory data analysis: Boston, MA, Duxbury Press, 354 p.

Vogel, R.M., Rudolph, B.E., and Hooper, R.P., 2005, Probabilistic behavior of water-quality loads: Journal of Environmental Engineering, v. 131, no. 7, p. 1081–1089.

Vogel, R.M., Stedinger, J.R., and Hooper, R.P., 2003, Discharge indices for water-quality loads: Water Resources Research, v. 39, no. 10, p. 1273–1281.