U.S. Geological Survey,
Texas Tech University, Center for
Multidisciplinary Research in Transportation

ALTERNATIVE REGRESSION EQUATIONS FOR ESTIMATION OF ANNUAL PEAK-STREAMFLOW FREQUENCY FOR UNDEVELOPED WATERSHEDS IN TEXAS USING PRESS MINIMIZATION

Research Report 0–4405–2

Texas Department of Transportation
Research Project 0–4405

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/TX–05/0–4405–2		

4. Title and Subtitle	5. Report Date
Alternative Regression Equations for Estimation of Annual Peak-Streamflow Frequency for Undeveloped Watersheds in Texas Using PRESS Minimization	August 2005
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
William H. Asquith and David B. Thompson	0–4405–2
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
U.S. Geological Survey Texas Water Science Center 8027 Exchange Drive Austin, Texas 78754	11. Contract or Grant No. Project 0–4405
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Texas Department of Transportation Research and Technology Implementation Office P.O. Box 5080 Austin, TX 78763–5080	Technical Report September 2003–August 2005
	14. Sponsoring Agency Code

15. Supplementary Notes

Project conducted in cooperation with the Texas Department of Transportation and the Federal Highway Administration.

16. Abstract

Peak-streamflow frequency estimates are needed for flood-plain management; for objective assessment of flood risk; for cost-effective design of dams, levees, other flood-control structures; and for design of roads, bridges, and culverts. Peak-streamflow frequency represents the collective peak streamflow for recurrence intervals of 2, 5, 10, 25, 50, and 100 years. A common model for estimation of peak-streamflow frequency is based on the regional regression method. The current (2005) regional regression equations for 11 regions of Texas are based on log10 transformations on all regression variables (drainage area, main-channel slope, watershed shape). The log10-exclusive transformation does not fully linearize the relations between the variables, and the effect is demonstrated. As a result, some systematic bias remains in the current equations. The bias results in overestimation of peak streamflow for both the smallest and largest watersheds. The bias increases with increasing recurrence interval. The primary source of the bias is the discernible curvilinear relation between peak streamflow and drainage area in log10 space. The bias is demonstrated by selected residual plots with superimposed LOWESS (LOcally WEighted Scatterplot Smoothing) trend lines. To address the bias, a statistical framework based on minimization of the PRESS (PRediction Error Sum of Squares) statistic through power transformation on drainage area is described, implemented, and the resulting regression equations reported. Compared to log10-exclusive equations, the equations derived from PRESS minimization have PRESS statistics and residual standard errors less than those of the log10-exclusive equations. Selected residual plots for the PRESS-minimized equations demonstrate that the systematic bias in regional regression equations for peak-streamflow frequency estimation in Texas can be removed. Because the overall error is similar to the overall error associated with the equations currently in use and bias is removed, the PRESS-minimized equations provide an alternative technique for peak-streamflow frequency estimation. A promising line of research into peak-streamflow frequency estimation through the regional regression method is demonstrated.

17. Key Words	18. Distribution Statement
Surface Water, Peak Streamflow, Recurrence Interval, Flood Frequency, Texas	No restrictions.

19. Security Classif. (of report)	20. Security Classif. (of this page)	21. No. of pages	22. Price
Unclassified	Unclassified	36	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

ALTERNATIVE REGRESSION EQUATIONS FOR ESTIMATION OF ANNUAL PEAK-STREAMFLOW FREQUENCY FOR UNDEVELOPED WATERSHEDS IN TEXAS USING PRESS MINIMIZATION

by

William H. Asquith, Research Hydrologist
U.S. Geological Survey, Austin, Texas

David B. Thompson, Associate Professor
Department of Civil Engineering, Texas Tech University

Research Report 0–4405–2

Texas Department of Transportation Research Project Number 0–4405
Research Project Title: "Identify Appropriate Size Limitations for Hydrologic Models"

Performed in cooperation with the
Texas Department of Transportation
and the
Federal Highway Administration

August 2005

U.S. Geological Survey
Austin, Texas 78754-4733

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official view or policies of the Federal Highway Administration (FHWA) or the Texas Department of Transportation (TxDOT). This report does not constitute a standard, specification, or regulation. The United States government and the State of Texas do not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the object of this report. The researcher in charge of this project was Dr. David B. Thompson, Texas Tech University.

No invention or discovery was conceived or first actually reduced to practice in the course of or under this contract, including any art, method, process, machine, manufacture, design, or composition of matter, or any new useful improvement thereof, or any variety of plant, which is or may be patentable under the patent laws of the United States of America or any foreign country.

ACKNOWLEDGMENTS

The authors recognize the contributions of George Herrmann, San Angelo District, Project Director 0–4405; David Stolpa, Design Division, Program Coordinator for Project 0–4405; and Amy Ronnfeldt, Design Division, Project Advisor 0–4405.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

This page intentionally left blank.

ABSTRACT

Peak-streamflow frequency estimates are needed for flood-plain management; for objective assessment of flood risk; for cost-effective design of dams, levees, and other flood-control structures; and for design of roads, bridges, and culverts. Peak-streamflow frequency represents the collective peak streamflow for recurrence intervals of 2, 5, 10, 25, 50, and 100 years. A common model for estimation of peak-streamflow frequency is based on the regional regression method. The current (2005) regional regression equations for 11 regions of Texas are based on $\log_{10}$ transformations on all regression variables (the peak-streamflow values and the watershed characteristics of drainage area, main-channel slope, watershed shape). The $\log_{10}$-exclusive transformation does not fully linearize the relations between the variables. As a result, some systematic bias remains in the current equations. The bias results in overestimation of peak streamflow for both the smallest and largest watersheds. The bias increases with increasing recurrence interval. The primary source of the bias is the discernible curvilinear relation between peak streamflow and drainage area in $\log_{10}$ space. The bias is demonstrated by selected residual plots with superimposed LOWESS (LOcally WEighted Scatterplot Smoothing) trend lines. To address the bias, a statistical framework based on minimization of the PRESS (PRediction Error Sum of Squares) statistic through power transformation on drainage area is described and implemented, and the resulting regression equations are reported. Compared to $\log_{10}$-exclusive equations, the equations derived from PRESS minimization have PRESS statistics and residual standard errors less than those of the $\log_{10}$-exclusive equations. Selected residual plots for the PRESS-minimized equations demonstrate that the systematic bias in regional regression equations for peak-streamflow frequency estimation in Texas can be removed. Because the overall error is similar to the overall error associated with the equations currently in use and bias is removed, the PRESS-minimized equations provide an alternative technique for peak-streamflow frequency estimation. A promising line of research into peak-streamflow frequency estimation through the regional regression method is demonstrated.

INTRODUCTION

Peak-streamflow frequency estimates are needed for flood-plain management; for objective assessment of flood risk; for cost-effective design of dams, levees, and other flood-control structures; and for design of roads, bridges, and culverts. Peak-streamflow frequency represents the peak streamflow for recurrence intervals of 2, 5, 10, 25, 50, and 100 years.

In 2003 as part of Texas Department of Transportation (TxDOT) Research Project 0–4405, the U.S. Geological Survey (USGS) and Texas Tech University began a 3-year investigation of the influence of hydrologic scale (represented by watershed drainage area for this report) on hydrologic model performance. Hydrologic models for estimation of design floods are in widespread use by TxDOT engineers and the broader hydrologic engineering community. A common model for estimation of peak-streamflow frequency is based on the regional regression method. This method is the subject of this report.

Bias exists in current (2005) regional regression equations for estimation of peak-streamflow frequency in Texas (Asquith and Slade, 1997, hereinafter referred to as AS1997). The source of the bias is the discernible curvilinear relation between peak streamflow and drainage area—the bias is graphically illustrated herein. In general, the current regional regression equations are expected to overestimate peak streamflow for both the smallest and largest watersheds. The bias is scale-dependent (depends on the size of the drainage area) and can be removed.

Purpose and Scope

The primary purpose of this report is to evaluate an alternative statistical framework for developing regression equations with enhanced prediction capabilities for small watersheds. The alternative framework uses a technique involving the minimization of the PRESS (PRediction Error Sum of Squares) statistic (Helsel and Hirsch, 1992, p. 248). A secondary purpose of this report is to present regression equations based on PRESS minimization for the estimation of peak-streamflow frequency at ungaged sites in undeveloped watersheds in Texas. A tertiary purpose of this report is to present regression equations for Texas independent of geographic region. The scope of the report is limited to the at-site peak-streamflow frequency values for 664 USGS streamflow-gaging stations used in AS1997 and digitally tabulated in Asquith and Slade (1999, file `tx664.dat`). The alternative regression equations presented here are based on the entire study area (Texas and slight overlap with surrounding states) of AS1997 and are not based on geographic regions of Texas, unlike the approach of AS1997.

Current (2005) Regional Regression Equations for Peak-Streamflow Frequency Estimation in Texas

AS1997 provides regional regression equations to estimate the 2-, 5-, 10-, 25-, 50-, and 100-year annual peak discharge (the peak-streamflow frequency curve) for undeveloped watersheds in Texas. The equations use the watershed characteristics of drainage area, main-channel slope, and a basin shape factor as predictor variables. AS1997 divides Texas into 11 regions. The mean number of stations used for each equation is 36. For each region, six or 12 weighted-least-squares regression equations were developed using a forward stepwise procedure. A distinction between the six- and 12-equation regions is elaborated upon later in this section.

Although the AS1997 statistical analysis is sound with innovative methods of equation development and presentation, and the report is popular and in widespread use, three observations regarding the AS1997 procedural framework are important for this report. The observations are important because the observations relate to application or implementation of the AS1997 equations by end users involved in public and private infrastructure design. The observations have gradually developed over the years since publication of AS1997 and were refined with progress on TxDOT-sponsored research into the relation between hydrologic scale and technology appropriate for watershed analysis (TxDOT research project 0–4405). The three observations are described in separate sections that follow.

Inconsistent Peak-Streamflow Frequency Curves by Regional Regression

For a given region, watershed characteristics used for developing regression equations from AS1997 for the 2- through 100-year equations are inconsistent, the reason for a statistically inconsistent peak-streamflow frequency curve for some watersheds. By definition, a peak-streamflow frequency curve must monotonically increase with increasing recurrence interval. The term “inconsistent” in this context means that the computed discharge for a recurrence interval exceeds the computed discharge for a larger recurrence interval. For example, the 50-year peak streamflow is computed to be greater than the 100-year peak streamflow. The source of the peak-streamflow inconsistency is the inconsistent use of watershed characteristics within an equation ensemble (a set of equations for a given region).

The inconsistency in watershed characteristics exists because AS1997 used a forward stepwise regression procedure and did not specifically force predictor variables into the equations. For example, the equations for region 11 of AS1997 (southeastern Texas) are listed in table 1. Main-channel slope is not used for the 2-year recurrence interval, but it is for larger recurrence intervals. Watershed shape is used for the 2- through 10-year recurrence intervals, but it is not used for larger recurrence intervals. Although difficult to visualize, combinations of watershed characteristics can be substituted into the equations listed in table 1 to produce an inconsistent frequency curve.

Table 1. Asquith and Slade (1997) regression equations for region 11 (southeastern Texas).

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; A, drainage area in square miles; S, main-channel slope in feet per mile; H, dimensionless basin shape factor. Sixty-six stations used in the regression development.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$
$Q_2 = 159A^{0.669}H^{-0.262}$	0.91	0.18
$Q_5 = 191A^{0.696}S^{0.130}H^{-0.186}$	.91	.18
$Q_{10} = 199A^{0.718}S^{0.221}H^{-0.151}$	.90	.20
$Q_{25} = 201A^{0.713}S^{0.313}$	.88	.22
$Q_{50} = 207A^{0.735}S^{0.380}$	.86	.24
$Q_{100} = 213A^{0.755}S^{0.442}$	.85	.26

AS1997 (AS1997, p. 11) explicitly remarks on the potential for inconsistent peak-streamflow frequency curves from the equations. When the equations and the guidance on equation application originally were developed, the authors (Asquith and Slade) anticipated that end users would apply “hydrologic engineering judgement” to manually mitigate the peak-streamflow inconsistencies. However, numerous end users have communicated to the senior author a degree of confusion regarding application of the AS1997 equations, which indicates a need for alternative equations that will not produce, or have a reduced potential to produce, inconsistent peak-streamflow frequency curves.

Regional Regression Equation Applicability and Implementation

AS1997 provides numerous figures (AS1997, figs. 4–14) in which the relations between drainage area, main-channel slope, and basin shape factor are graphically depicted for each of the 11 regions. Superimposed on these plots are generalized convex hulls representing the “approximate [parameter space] defined by [watershed] characteristics” for each region. For watersheds having coordinates of drainage area, main-channel slope, and basin shape factor outside the convex hull, the applicability of the equations for the region is uncertain, and the potential for an inconsistent peak-streamflow frequency curve increases.

Since publication of AS1997, the senior author has learned from interaction with end users that the convex hulls presented in AS1997 commonly are underutilized. Furthermore, some end users abstracted (reproduced) for application only the equations from AS1997. As a result, important context that contributes to optimum use of the equations is lost.

The apparent lack of full adherence to the entire procedural framework and caveats of the AS1997 regional regression equations is understandable given that AS1997 provides 96 separate equations and considerable detail. Therefore, a simpler regional regression method for estimation of peak-streamflow frequency in Texas might be considered an enhancement over AS1997.

Biased Peak-Streamflow Frequency Values

The multiple linear regional regression equations of AS1997 are exclusively based on $\log_{10}$ transformations of observed peak-streamflow frequency values, drainage area, main-channel slope, and basin shape factor. Multiple linear regression is based on a linear relation between the regressor variable (peak-streamflow frequency) and the predictor variables (drainage area, main-channel slope, and others). AS1997 (AS1997, p. 8) notes that, for some regions, the peak-streamflow values (for example, the 100-year peak streamflow) have a discernible curvilinear relation with drainage area in $\log_{10}$ space. AS1997 addresses the nonlinearity (and thus mitigates the bias) by classifying watersheds into two ranges of drainage area. Separate regional regression analyses were done for watersheds with drainage areas less than 32 square miles and for watersheds with areas greater than 32 square miles. The 32-square-mile break point was determined by data interpretation. The drainage-area distinction and bias mitigation is explicitly discussed in AS1997 (AS1997, p. 13).

The drainage-area classification was not made for six of the 11 regions because of either the small number of watersheds (degrees of freedom for regression) within a region or the perceived absence of a discernible curvilinear relation between $\log_{10}$-transformed peak streamflow and drainage area. For a region in which the drainage-area classification was made, 12 equations were developed—six equations for watersheds with drainage areas less than 32 square miles and six equations for watersheds with drainage areas greater than 32 square miles. Conversely, six equations were developed for regions in which nonlinearity was not apparent and no drainage-area classification was made.

The drainage-area classification greatly complicates application of the equations near the 32-square mile break point. AS1997 (AS1997, p. 12) provides an ad hoc procedure to prorate estimates for watersheds of 10 to 100 square miles between the equation ensemble for drainage areas less than or equal to 32 square miles and the ensemble for drainage areas greater than or equal to 32 square miles. If the proration procedure is not followed, “jumps” in peak streamflow at 32 square miles will result.

The nonlinearity is apparent in the graphical depiction of the 32-square-mile classification technique to mitigate for nonlinearity (AS1997, figs. 3 and 15). Despite the measures to address the nonlinearity and thus mitigate bias, the AS1997 equations still have the potential to overestimate peak-streamflow frequency values for both the smallest and largest watersheds. As noted, eliminating or reducing the potential for inconsistent peak-streamflow frequency curves and making the regional regression equation method easier for end users to apply are reasons for developing alternative equations; but the primary reason for developing alternative equations is to remove the bias inherent in the AS1997 equations.

Typical regression practice to reduce overestimation and underestimation of peak-streamflow frequency values is to seek an alternative transformation on the regressor variable (for example, Maindonald and Braun, 2003, p. 126–127). Statistical practitioners might question why an alternative transformation on drainage area (a predictor variable) is sought rather than an alternative transformation on the 2- through 100-year peak streamflow values (regressor variables). The authors chose to assess an alternative transformation on drainage area so that the residual standard errors ($\log_{10}$ units of streamflow) reported are directly comparable to those from AS1997.

REGRESSION EQUATIONS BASED ON LOG$_{10}$ TRANSFORMATION OF DRAINAGE AREA

The traditional practice for development of regression equations to estimate peak-streamflow frequency is to $\log_{10}$ transform the regressor variables (the at-site peak-streamflow frequency values, such as the 2- through 100-year peak streamflows) and all the predictor variables (Stedinger and others, 1992, p. 18.35). Drainage area, a measure of watershed slope, and other characteristics are common predictor variables. AS1997 considered six characteristics: 2-year 24-hour precipitation, mean annual precipitation, drainage area, stream length, basin shape factor, and main-channel slope. The precipitation statistics reported in AS1997 are for the approximate watershed centroid. However, for the equations reported in AS1997, only drainage area, main-channel slope, and basin shape factor are used.

Because of the ubiquitous nature of $\log_{10}$ transformation in hydrologic analysis, important comparative analysis for this report is facilitated by developing regression equations using $\log_{10}$ transformation on the same data used in AS1997. However, no designation of geographic region is used for this report. The data are digitally available through Asquith and Slade (1999). AS1997 considered the data for 664 USGS streamflow-gaging stations. From preliminary data analysis, eight stations were identified as outliers (results not presented here) and eliminated from further analysis (table 2).

Table 2. U.S. Geological Survey streamflow-gaging stations identified as outliers and removed from analysis.

Station no.	Station name	Drainage area (square miles)
08039900	Little Sandy Creek tributary near Jasper, Texas	0.46
08080700	Running Water Draw at Plainview, Texas	382
08089100	Elm Creek tributary near Graford, Texas	1.10
08210400	Lagarto Creek near George West, Texas	155
08383200	Pintada Arroyo tributary near Clines Corners, New Mexico	29.20
08393600	North Spring River at Roswell, New Mexico	19.50
08405050	Last Chance Canyon tributary near Carlsbad Caverns, New Mexico	.20
08434000	Toyah Creek below Toyah Lake near Pecos, Texas	3,709

Weighted-least squares regression on the 2-, 5-, 10-, 25-, 50-, and 100-year peak-streamflow values for the remaining 656 streamflow-gaging stations is accomplished using drainage area, mean annual precipitation, and main-channel slope as predictor variables. For comparison, the mean number of stations per equation in AS1997 is 36. Therefore, the degrees of freedom for the regression equations reported here are about 18 times larger than those of AS1997.

Analysis of collinearity through variance inflation factors and statistical significance (results not reported here) strongly indicated that inclusion of watershed shape in the regression equations in addition to drainage area, mean annual precipitation, and main-channel slope is not appropriate.

Table 3. Regression equations based on $\log_{10}$ transformation of drainage area using three predictor variables.

[Q_T, peak streamflow for *T*-year recurrence interval in cubic feet per second; PRESS, PRediction Error Sum of Squares; *A*, drainage area in square miles; *P*, mean annual precipitation in inches; *S*, main-channel slope in feet per mile.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic
$Q_2 = 10^{-0.5240} A^{0.6565} P^{1.474} S^{0.3525}$	0.8282	0.2866	54.36
$Q_5 = 10^{-0.2204} A^{0.6790} P^{1.376} S^{0.4828}$	.8414	.2686	47.76
$Q_{10} = 10^{-0.04207} A^{0.6896} P^{1.317} S^{0.5421}$	.8310	.2778	51.12
$Q_{25} = 10^{0.1501} A^{0.7005} P^{1.256} S^{0.6005}$	.8086	.2993	59.34
$Q_{50} = 10^{0.2748} A^{0.7073} P^{1.218} S^{0.6359}$	.7887	.3186	67.25
$Q_{100} = 10^{0.3879} A^{0.7133} P^{1.183} S^{0.6660}$	.7675	.3393	76.25

The six regression equations are listed in table 3. For all six equations, the p-values for the coefficients on the watershed characteristics are less than .0001.

A simple comparison between 100-year residual standard error in table 3 and the weighted-mean 100-year residual standard error from AS1997 provides perspective. The weighted-mean 100-year residual standard error from AS1997 is computed by weighting the errors in AS1997 (AS1997, table 2) by the number of stations for each region. The weighted-mean 100-year residual standard error from AS1997 is about 0.27; the 100-year residual standard error in table 3 is 0.34. These two residual standard errors are of similar magnitude. Additional comparisons of residual standard errors in table 3 to those in AS1997 indicate that all have about the same magnitude, although overall the errors are greater for the equations reported here. The conclusion from this comparison is that the six equations in table 3 have approximately the same residual standard error as the 96 equations reported in AS1997.

For the equations in table 3, inclusion of mean annual precipitation for the watershed is useful. Mean annual precipitation becomes a surrogate spatial location variable that takes the place of geographic region designation associated with the equations in AS1997. Mean annual precipitation was not used in AS1997 for the final equations shown in that report.

Bias in multiple linear regression is well depicted in a residual (observed minus predicted) graph in which the residual for a particular data point is plotted on the vertical axis and the corresponding fitted value is plotted on the horizontal axis. If there is a discernible trend or shape in the graph—that is, a tendency for residuals to plot above or below the zero-residual line then bias in the equation exists.

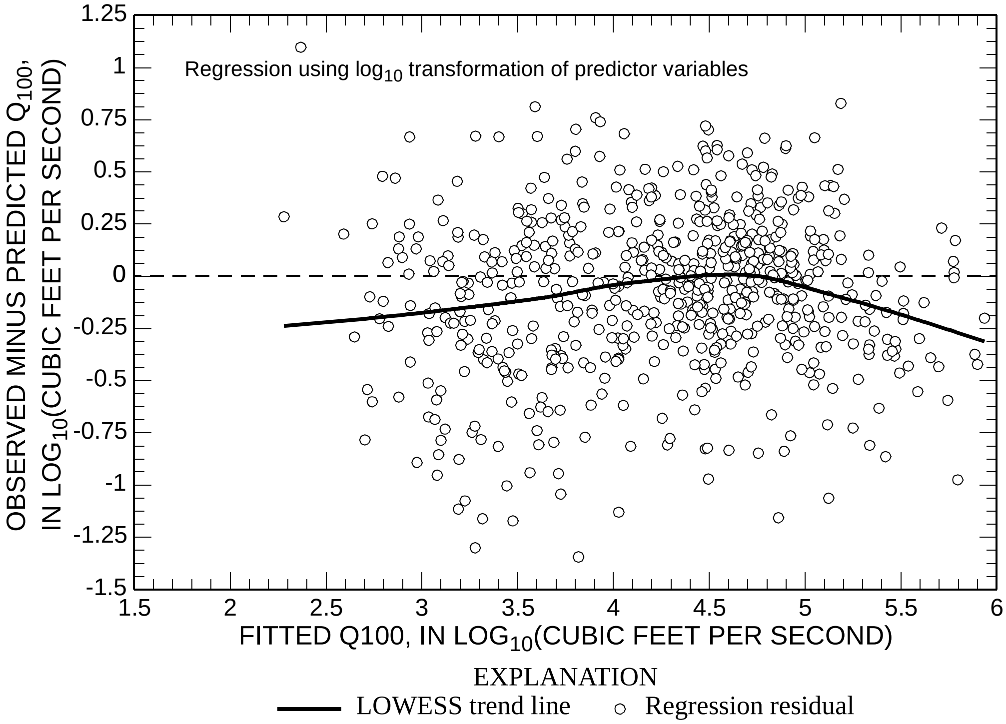

Figure 1. Residual plot of regression of 100-year peak streamflow using $\log_{10}$ transformation of drainage area using three predictor variables.

Residuals for the 100-year peak-streamflow equation listed in table 3 are graphed in figure 1. A LOWESS (LOcally WEighted Scatterplot Smoothing) trend line (Cleveland, 1979) through the data is superimposed. The "`lowess()`" function of the R System software package (R Development Core Team, 2004) with default settings was used. The concave-down shape of the LOWESS trend line indicates systematic bias in the regression. The negative magnitudes of the left and right segments of the LOWESS trend line indicate that overestimation of the 100-year peak streamflow occurs for watersheds with small and large fitted values, respectively (the smallest and largest watersheds). The LOWESS trend line is only an indicator of bias and does not represent a true bias correction; however, interpretation of the line as a bias measure is useful. For example, referring to

Table 4. Regression equations based on $\log_{10}$ transformation of drainage area using drainage area as the only predictor variable.

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; PRESS, PRediction Error Sum of Squares; A, drainage area in square miles.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic
$Q_2 = 10^{2.339} A^{0.5158}$	0.7642	0.3357	74.22
$Q_5 = 10^{2.706} A^{0.5111}$	.7889	.3099	63.28
$Q_{10} = 10^{2.892} A^{0.5100}$	.7820	.3156	65.63
$Q_{25} = 10^{3.086} A^{0.5093}$	.7612	.3343	73.67
$Q_{50} = 10^{3.209} A^{0.5092}$	.7414	.3525	81.88
$Q_{100} = 10^{3.318} A^{0.5094}$	.7199	.3724	91.39

figure 1, for a fitted value of about 2.5 (316 cubic feet per second) and a LOWESS-indicated bias of about -0.25, a more appropriate value might be 2.5–0.25=2.25 (178 cubic feet per second). Therefore, the bias-corrected value is about 44 percent less than the fitted value. In general, the $\log_{10}$-exclusive regressions in table 3 have a concave-down trend lines through the residuals. The concavity of the LOWESS trend line (interpreted as bias in the equations) increases with increasing recurrence interval (results not presented here).

Hydrologic scale typically is measured by drainage area. Therefore, it is informative to develop a second set of $\log_{10}$ transformed regression equations on the same 656 stations using only drainage area as a predictor variable (table 4). For all six equations, the p-values for the coefficients on the intercept and drainage area are less than .0001. The residual standard errors associated with the equations in table 4 are all greater than those listed in table 3 because two fewer predictor variables are in the equations in table 4.

Residuals for the 100-year peak-streamflow equations in table 4 are shown in figure 2. A LOWESS trend line is superimposed on the figure. The LOWESS trend line has considerable downward concavity like that in figure 1. The interpretations of the regressions in table 4 using the LOWESS trend line on the residual plot are the same as those for the regression equations in table 3. Specifically, peak streamflow is overestimated for watersheds with small fitted values (the smallest watersheds) and for watersheds with large fitted values (the largest watersheds). The bias is considerable. The concavity of the LOWESS trend line increases with increasing recurrence interval (results not presented here).

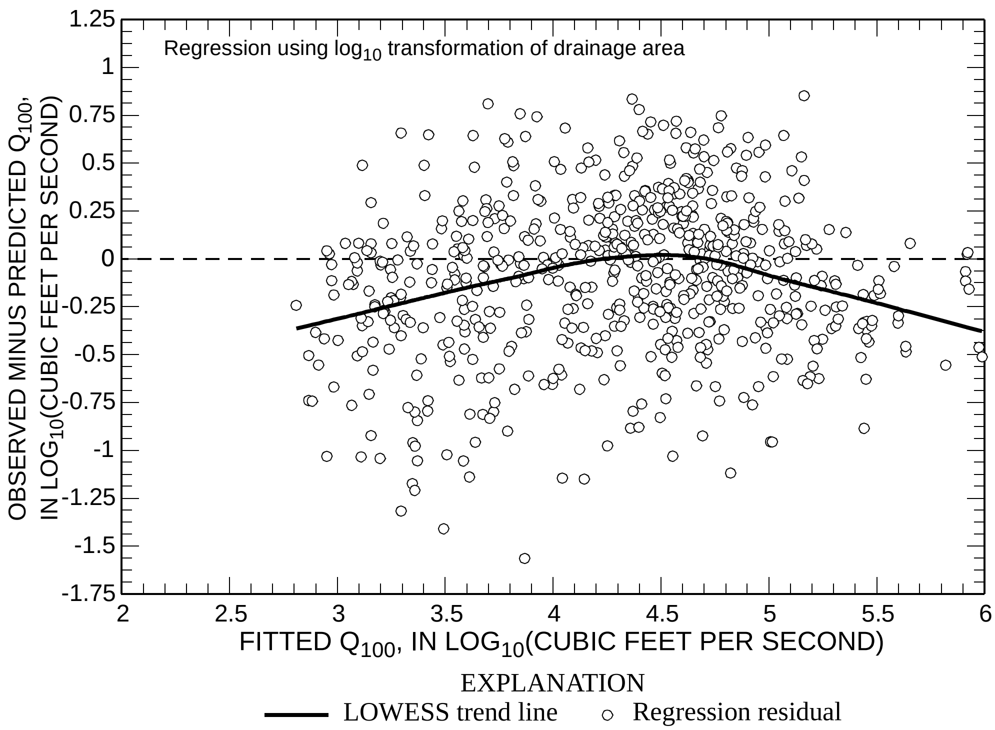

Figure 2. Residual plot of regression of 100-year peak streamflow using $\log_{10}$ transformation of drainage area using drainage area as the only predictor variable.

In conclusion, systematic bias is in the regression equations reported in tables 3 and 4, and by general method association, bias is in the AS1997 equations. The bias exists because of the curvilinear relation between $\log_{10}$-transformed peak streamflow and drainage area. The bias is mitigated in the AS1997 analysis by separating regressions into two groups on the basis of watershed drainage area, less than or greater than 32 square miles. The relation between $\log_{10}$-transformed peak streamflow and drainage area becomes increasingly curvilinear with increasing recurrence interval.

REGRESSION EQUATIONS BASED ON PRESS MINIMIZATION AND POWER TRANSFORMATION OF DRAINAGE AREA

The PRESS statistic generally is regarded as a measure of regression performance when the model is used to predict new data (Montgomery and others, 2001, p. 153). Prediction of new data is what analysts and engineers are doing when they estimate peak streamflow from a regression equation. Regression equations with small PRESS values are desirable. Thus, PRESS minimization is an appropriate goal. Helsel and Hirsch (1992, p. 248) state that, “Minimizing PRESS means that the equation produces the least error when making new predictions.” Conceptually, PRESS minimization identifies the appropriate transformation to “press” the bias out of the equations (fig. 3).

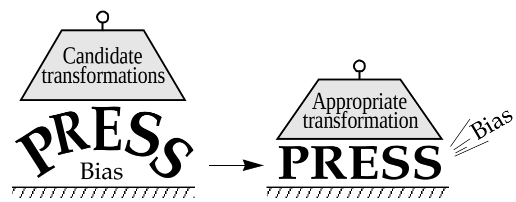

Figure 3. Conceptual display of the PRESS statistic minimization.

The PRESS statistic is computed from the PRESS residuals, which are defined as

$$e_{(i)} = y_i - y_i' , \tag{1}$$

where $e_{(i)}$ is the PRESS residual, y_i is the observed ith peak-streamflow value, and y_i' is the predicted value based on the remaining $n-1$ sample points. In other words, the ith station (data point) is not used to generate the ith regression equation. Thus, PRESS residuals are regarded as validation statistics. The PRESS statistic, with inclusion of the regression weight factor (w_i), is

$$PRESS = \sum_{i=1}^{n} w_i e_{(i)}{}^2 . \tag{2}$$

Eq. 2 is computationally intensive (n regressions are required). A more efficient computation of PRESS is made using regression residuals (e_i) and leverage (h_{ii}). These values are readily available from modern regression software packages. The efficient PRESS computation is made by

$$PRESS = \sum_{i=1}^{n} w_i \left(\frac{e_i}{1-h_{ii}}\right)^2 . \tag{3}$$

Because the PRESS statistic is an overall measure of regression fit (like residual standard error) and is a validation statistic (unlike residual standard error), minimization of PRESS is desirable. The most “valid” regression is produced when the PRESS statistic is minimized. The following transformation on drainage area was selected after exploratory analysis:

$$A' = A^{\lambda}, \tag{4}$$

where A' is the transformed value for the regression, A is drainage area, and λ is a real number. The transformation is referred to in this report as the power transformation.

Two computer programs were written in R to loop through successive non-integer values of λ and record the value that yields a minimum PRESS for each of the six recurrence intervals. Tens of thousands of regressions were done in the process of exploratory data analysis and for the final minimization reported here. The first program implemented the watershed characteristics drainage area, mean annual precipitation, and main-channel slope as predictor variables; the second program implemented only drainage area as a predictor variable.

The programs and incremental output are provided in the appendix. The purpose of including the programs and output in the report is to provide an archive of the PRESS minimization algorithm and the regression analysis results summarized in tables 5 and 6.

The results of the power transformation of drainage area using the three predictor variables are listed in table 5. The value of λ is the exponent on A in the equations. The values of λ increase in absolute magnitude with increasing recurrence interval; the larger the absolute value of λ, the larger the amount of concavity in the trend line of residuals (systematic bias) that is removed relative to the $\log_{10}$-exclusive equations.

In all six equations, the p-values for the coefficients on the watershed characteristics are less than .0001. The diagnostic statistics of adjusted R-squared and residual standard error in the table are greater and less than, respectively, those in table 3. Therefore, the equations using the power transformation have less uncertainty. However, the PRESS statistic is the more important statistic to compare. The PRESS statistic for a given recurrence interval is less when the power transformation on drainage area is used instead of the $\log_{10}$ transformation. The percentage changes in the PRESS statistics associated with the power transformation (table 5) from those associated with the $\log_{10}$ transformation (table 3) show that, as recurrence interval increases, the power transformation produces an increasingly more valid regression.

Residual standard errors of the PRESS-minimized equations in table 5 are similar to those of the equations reported in AS1997. For example, the 100-year residual standard error is about 0.33 and the AS1997 weighted value is 0.27 for the 11 regions collectively.

Residuals for the 100-year peak-streamflow equation using the three predictor variables (table 5) are shown in figure 4. A LOWESS trend line through the data is superimposed on the figure. Downward concavity of the LOWESS trend line is not present, unlike the LOWESS trend line in figure 1. In fact, the LOWESS trend line is essentially flat, which indicates that systematic bias in the equation has been removed through the use of the specified power transformation. The power transformation on drainage area effectively linearizes the relation between 100-year peak streamflow and drainage area. Minimization of the PRESS statistic effectively removes systematic bias. Similar results (not reported here) for the five remaining recurrence intervals were obtained.

Table 5. Regression equations based on power transformation of drainage area using three predictor variables.

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; PRESS, PRediction Error Sum of Squares; A, drainage area in square miles; P, mean annual precipitation in inches; S, main-channel slope in feet per mile. The exponent of A is the power λ.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic	Percent change from PRESS in table 3
$Q_2 = 10^{35.60 - 36.09A^{-0.008}} P^{1.448} S^{0.3472}$	0.8286	0.2863	54.27	-0.17
$Q_5 = 10^{11.16 - 11.28A^{-0.0299}} P^{1.279} S^{0.4640}$	.8461	.2646	46.37	-2.9
$Q_{10} = 10^{9.047 - 8.950A^{-0.0400}} P^{1.188} S^{0.5172}$	.8396	.2707	48.57	-5.0
$Q_{25} = 10^{7.949 - 7.628A^{-0.0497}} P^{1.096} S^{0.5699}$	.8217	.2889	55.32	-6.8
$Q_{50} = 10^{7.554 - 7.090A^{-0.0553}} P^{1.039} S^{0.6021}$	.8048	.3062	62.18	-7.5
$Q_{100} = 10^{7.307 - 6.714A^{-0.0601}} P^{0.9883} S^{0.6295}$	.7862	.3253	70.19	-7.9

The results of the power transformation of drainage area using only drainage area as a predictor variable are listed in table 6. Again, the value of λ is the exponent on A in the equations. In all six equations, the p-values for the coefficients on the watershed characteristics are less than .0001. Adjusted R-squared and residual standard error for regression based on power transformation are greater and less than, respectively, those for regression based exclusively on $\log_{10}$ transformation (table 4). Therefore, the equations using the power transformation have less uncertainty. The PRESS statistic for a given recurrence interval is less when the power transformation on drainage area is used instead of the $\log_{10}$ transformation. The percentage changes in the PRESS statistic associated with the power transformation (table 6) from those associated with the $\log_{10}$ transformation (table 4) show that, as recurrence interval increases, the power transformation produces an increasingly more valid regression.

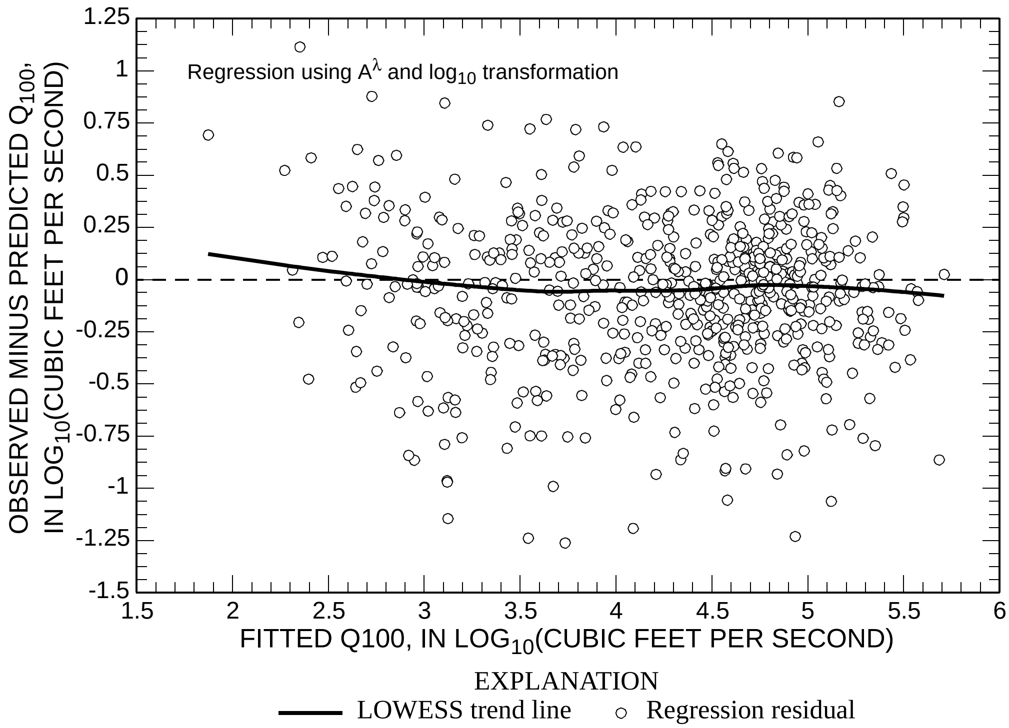

Figure 4. Residual plot of regression of 100-year peak streamflow using power transformation of drainage area using three predictor variables.

Table 6. Regression equations based on power transformation of drainage area using drainage area as the only predictor variable.

[Q_T, peak streamflow for *T*-year recurrence interval in cubic feet per second; PRESS, PRediction Error Sum of Squares; *A*, drainage area in square miles. The exponent of *A* is the power λ.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic	Percent change from PRESS in table 4
$Q_2 = 10^{8.280 - 6.031A^{-0.0465}}$	0.7710	0.3309	72.06	-2.9
$Q_5 = 10^{7.194 - 4.614A^{-0.0658}}$	.8030	.2994	59.00	-6.8
$Q_{10} = 10^{6.961 - 4.212A^{-0.0749}}$	.8002	.3021	60.10	-8.4
$Q_{25} = 10^{6.840 - 3.914A^{-0.0837}}$	.7834	.3184	66.77	-9.4
$Q_{50} = 10^{6.806 - 3.766A^{-0.0890}}$	.7659	.3354	74.08	-9.5
$Q_{100} = 10^{6.800 - 3.659A^{-0.0934}}$	.7462	.3545	82.78	-9.4

Residuals for the 100-year peak-streamflow equation in table 6 are graphed in figure 5. A LOWESS trend line through the data is superimposed. The concave-down shape of the LOWESS trend line in the residuals graph from the $\log_{10}$ transformation (fig. 2) is not present in the graph derived from the power transformation. In fact, the LOWESS trend line is essentially flat, which indicates that systematic bias in the equation has been removed. The authors conclude that the power transformation on drainage area effectively linearizes the relation between 100-year peak streamflow and drainage area. Minimization of PRESS effectively removes systematic bias. Similar results (not reported here) for the five remaining recurrence intervals were obtained.

A graph of the four PRESS statistics by recurrence interval is shown in figure 6. From the figure it is clear that the power transformation with PRESS minimization produces PRESS statistics less than those from the $\log_{10}$-exclusive equations. PRESS minimization becomes increasingly important as recurrence interval increases because the $\log_{10}$ transformation does not sufficiently linearize the relation between peak streamflow and drainage area for the larger recurrence-interval events. The smallest PRESS statistics occur for the 5-year recurrence interval. The PRESS statistics for the 2-year recurrence interval are not exceeded until the 25-year and larger recurrence intervals are reached.

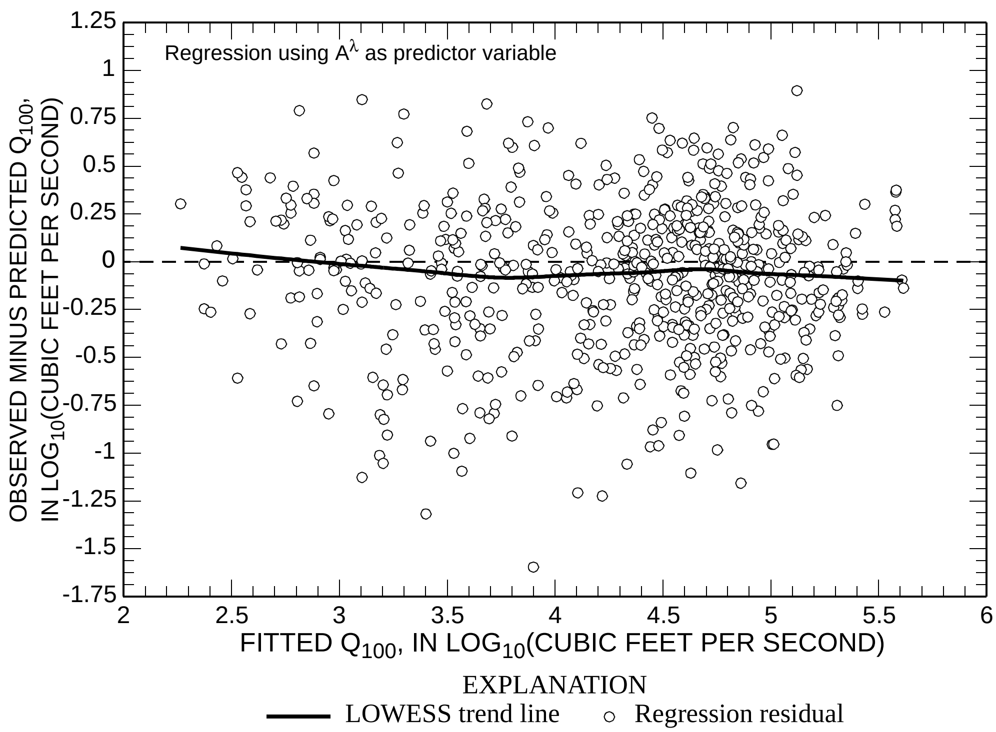

Figure 5. Residual plot of regression of 100-year peak streamflow using power transformation of drainage area using drainage area as the only predictor variable.

An interpretation of the PRESS statistic is that estimation of the 2-year peak streamflow using watershed characteristics is more difficult than estimation of the 5-year and 10-year peak streamflow. This observation is consistent with residual standard errors reported in AS1997.

Finally, the magnitude and extent of the bias between the $\log_{10}$-exclusive regression and the PRESS-minimized regression is informative. The magnitude of the bias can be expressed as the ratio (the bias ratio) of the $\log_{10}$ equations (tables 3 or 4) to the PRESS-minimized equations (tables 5 or 6). For example, the bias ratio for the 100-year peak streamflow for the drainage-area-only equations is

$$\frac{Q_{100}^{\log 10}}{Q_{100}^{PRESS}} = \frac{10^{3.318} A^{0.5094}}{10^{6.8000 - 3.659 A^{-0.0934}}} . \tag{5}$$

When the ratio is greater than 1, the $\log_{10}$-exclusive regression overestimates peak streamflow relative to the PRESS-minimized regression. Similar equations of the bias ratio for other recurrence intervals are easily defined. Together, the six equations defining the bias ratio document the inherent differences between the $\log_{10}$-exclusive peak-streamflow equations and the PRESS-minimized equations.

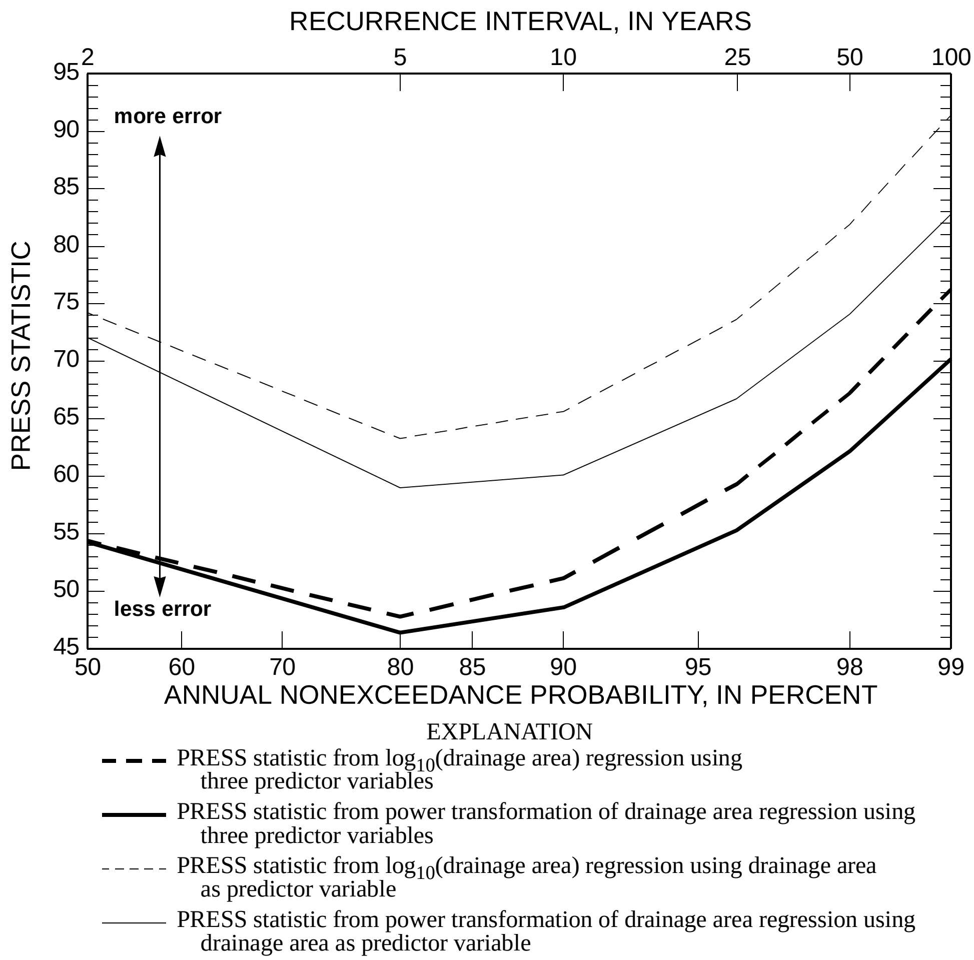

Figure 6. Comparison of PRESS statistics from regression analysis.

The extent of the bias ratio is shown by the ratio as a function of drainage area. An example for each of the recurrence intervals for the regressions using drainage area as the only predictor variable is shown in figure 7. An interpretation of the figure is that the $\log_{10}$-exclusive regressions overestimate peak streamflow frequency for drainage areas less than about 8 square miles and drainage areas greater than about 2,000 square miles. The overestimation for drainage areas less than about 2 square miles is substantial. The overestimation for drainage areas less than about 0.5 square mile is in excess of 100 percent for all but the 2-year peak streamflow. Alternatively, the $\log_{10}$-exclusive regressions slightly underestimate peak streamflow frequency for drainage areas between about 8 and 2,000 square miles.

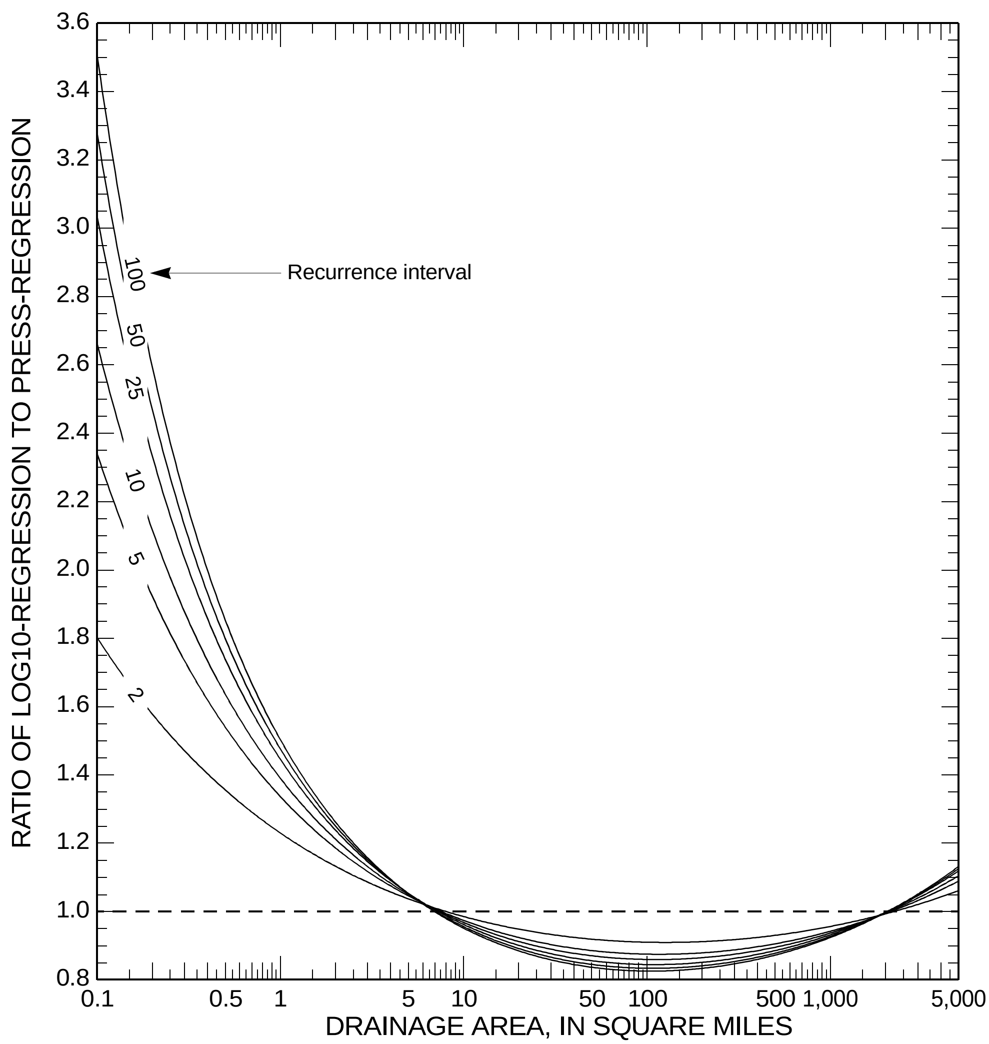

Figure 7. Relation between bias ratio and drainage area by recurrence interval for the regressions using drainage area as the only predictor variable.

SUMMARY

Peak-streamflow frequency estimates are needed for flood-plain management; for objective assessment of flood risk; for cost-effective design of dams, levees, and other flood-control structures; and for design of roads, bridges, and culverts. Peak-streamflow frequency represents the collective peak streamflow for recurrence intervals of 2, 5, 10, 25, 50, and 100 years. A common model for estimation of peak-streamflow frequency is based on the regional regression method, which relates peak-streamflow frequency to watershed characteristics. As part of TxDOT research project 0–4405, the USGS and Texas Tech University evaluated an alternative statistical framework for developing regression equations with enhanced ability to predict peak streamflow for small undeveloped watersheds in Texas.

The current (2005) 96 regional regression equations for 11 geographic regions of Texas are based on $\log_{10}$ transformations on all regression variables (the peak-streamflow values and the watershed characteristics of drainage area, main-channel slope, basin shape factor). The $\log_{10}$ transformation does not fully linearize the relations between the variables, which is a major assumption in linear regression analysis. As a result, some systematic bias remains in the current equations. The primary source of the bias is the discernible curvilinear relation between peak streamflow and drainage area in $\log_{10}$ space. The bias results in overestimation of peak streamflow for both the smallest and largest watersheds, and the bias increases with increasing recurrence interval.

To demonstrate the bias, equations using $\log_{10}$(drainage area) for the study area are reported. These equations are independent of geographic region. Inclusion of mean annual precipitation provides a surrogate spatial location variable that takes the place of geographic region designation associated with the current equations, which reduces the number of equations from 96 to six—one for each of the six recurrence intervals.

To address the bias, a statistical framework based on minimization of the PRESS statistic through power transformation on drainage area is described. The PRESS statistic is an important diagnostic of regression performance. It is a validation-type statistic, and small values are desirable. Minimization of PRESS is appropriate for peak-streamflow frequency analysis because the equations are used in hydrologic engineering practice to predict new data.

Compared to $\log_{10}$(drainage area) equations, the equations derived from PRESS minimization have PRESS statistics and residual standard errors less than those of the $\log_{10}$(drainage area) equations. Selected residual plots for the PRESS-minimized equations demonstrate that the systematic bias in regional regression equations for peak-streamflow frequency estimation in Texas can be removed. Because the overall error is similar to the overall error associated with the equations currently in use and bias is removed, the PRESS-minimized equations provide an alternative technique for peak-streamflow frequency estimation. Therefore, the regression equations developed by PRESS minimization are potential alternatives to the current equations. A promising line of research into peak-streamflow frequency estimation through the regional regression method thus is demonstrated.

REFERENCES

Asquith, W.H., and Slade, R.M., Jr., 1997, Regional equations for estimation of peak-streamflow frequency for natural basins in Texas: U.S. Geological Survey Water-Resources Investigations Report 96–4307, 68 p.

Asquith, W.H., and Slade, R.M., Jr., 1999, Site-specific estimation of peak-streamflow frequency using generalized least-squares regression for natural basins in Texas: U.S. Geological Survey Water-Resources Investigations Report 99–4172, 19 p. [URL http://water.usgs.gov/pubs/wri/wri994172]

Cleveland, W.W., 1979, Robust locally weighted regression and smoothing scatterplots: Journal of the American Statistical Association, v. 80, p. 829–836.

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources—Studies in environmental science 49: New York, Elsevier, 522 p.

Maindonald, J., and Braun, J., 2003, Data analysis and graphics using R—An example-based approach: Cambridge, United Kingdom, Cambridge University Press, 362 p.

Montgomery, D.C., Peck, E.A., and Vining, G.G., 2001, Introduction to linear regression analysis (3rd ed.): New York, Wiley, 641 p.

R Development Core Team, 2004, R—A language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria, ISBN 3–900051–07–0.

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou, E., 1992, Frequency analysis of extreme events, *in* Maidment, D.A., ed., Handbook of hydrology: New York, McGraw-Hill [variously paged].

This page intentionally left blank.

APPENDIX

Computational Script Using PRESS Minimization and Drainage Area, Mean Annual Precipitation, and Main-Channel Slope

```
R : Copyright 2004, The R Foundation for Statistical Computing
Version 2.0.1  (2004-11-15), ISBN 3-900051-07-0
> MLRweights <- function(vector) {
+   tmp = length(vector)/sum(vector)
+   return (tmp*vector)
+ }
>
> PRESS <- function(model) {
+   if(is.null(model$terms)) stop("invalid 'lm' object: no terms")
+   sum( (weighted.residuals(model)/(1-hatvalues(model)))^2 )
+ }
> DATA <- read.csv("tx664.csv",header=T)
> attach(DATA)
> names(DATA)
 [1] "Station" "LatD"    "LatM"    "LatS"    "LonD"    "LonM"    "LonS"
 [8] "EqYrs"   "CDA"     "MAP"     "P224"    "Slope"   "Shape"   "Q2"
[15] "Q5"      "Q10"     "Q25"     "Q50"     "Q100"    "C2"      "C25"
[22] "C100"
> outliers <- c(212,323,358,602,614,620,628,637)
> CDA  <- CDA[-outliers]
> Q2   <- Q2[-outliers]
> Q5   <- Q5[-outliers]
> Q10  <- Q10[-outliers]
> Q25  <- Q25[-outliers]
> Q50  <- Q50[-outliers]
> Q100 <- Q100[-outliers]
> MAP  <- MAP[-outliers]
> Slope <- Slope[-outliers]
> WEIGHTS <- MLRweights(EqYrs[-outliers])
>
> > sum.lowess.slope <- function(lowss) {
+   x <- lowss$x
+   y <- lowss$y
+   n <- length(x)
+   sum <- 0
+   for(i in (seq(2,n))) {
+     delx <- x[i] - x[i-1]
+     if(delx == 0) next
+     sum <- sum + (y[i] - y[i-1])/delx
+   }
+   return(sum)
+ }
>
> doQt <- function(Q,type) {
+   smallpress <- 10000
+   smallpower <- 10000
+   smallsl    <- 1e45
+   smallslope <- 1e45
+   for(power in seq(0.007,.08,by=0.0001)) {
+     if(power == 0) next
+     power <- -1 * power
+     CDA1 <- 10^CDA
+     CDA1 <- CDA1^power
+     WLS.OUT <- lm(Q~CDA1+MAP+Slope, weights=WEIGHTS)
+     sm <- lowess(fitted(WLS.OUT),y=residuals(WLS.OUT))
+     sl <- sum(abs(sm$x - residuals(WLS.OUT)))
+     sslope <- sum.lowess.slope(sm)
+     press <- PRESS(WLS.OUT)
+     if(press < smallpress) {
+       smallpress <- press
+       smallpower <- power
+     }
+     if(sl < smallsl) {
+       smallsl <- sl
+       smallslpower <- power
```

```
+     }
+     if(sslope < smallslope) {
+       smallslope <- sslope
+       smallslopepower <- power
+     }
+   }
+   rsq <- summary(WLS.OUT)$r.squared
+   rme <- summary(WLS.OUT)$sigma
+   print(c(type,smallpower,smallpress,smallsl,smallslpower,smallslope,smallslopepower,rsq,rme))
+   return(c(type,smallpower,smallpress,smallsl,smallslpower,smallslope,smallslopepower,rsq,rme))
+ }
>
>
> vals2   <- doQt(Q2,2)
[1]    2.0000000   -0.0082000   54.2650236 2180.5800371   -0.0574000
[6]    7.7574682   -0.0073000    0.8094416    0.3025241
> vals5   <- doQt(Q5,5)
[1]    5.0000000   -0.0299000   46.3680582 2422.5205699   -0.0676000
[6]    6.8520003   -0.0136000    0.8361210    0.2736545
> vals10  <- doQt(Q10,10)
[1]   10.0000000   -0.0400000   48.5688087 2546.4914507   -0.0723000
[6]    2.1089430   -0.0193000    0.8333676    0.2765250
> vals25  <- doQt(Q25,25)
[1]   25.0000000   -0.0497000   55.3159612 2676.3858864   -0.0768000
[6]   -4.6458696   -0.0256000    0.8186019    0.2920446
> vals50  <- doQt(Q50,50)
[1]   50.0000000   -0.0553000   62.1782942 2759.1174546   -0.0795000
[6]   -8.4980157   -0.0246000    0.8031427    0.3082154
> vals100 <- doQt(Q100,100)
[1]  100.0000000   -0.0601000   70.1874647 2832.8009817   -0.0800000
[6]  -12.1432954   -0.0368000    0.7855992    0.3265377
>
>
> finalQt <- function(Q,power,type) {
+   CDA1 <- 10^CDA
+   CDA1 <- CDA1^power
+   WLS.OUT <- lm(Q~CDA1+MAP+Slope, weights=WEIGHTS)
+   print(summary(WLS.OUT))
+   W <- diag(WEIGHTS)
+   X = model.matrix(WLS.OUT)
+   Xt = t(X)
+   tmp <- chol2inv( chol( Xt %*% W %*% X ) )
+   wlshat1 <- X %*% tmp %*% Xt
+   print(tmp)
+   print(max(diag(wlshat1)))
+   m.wls.out <- tmp %*% Xt %*% W %*% Q
+   print(m.wls.out)
+   PRESS(WLS.OUT)
+   return(WLS.OUT)
+ }
>
>
> F2.OUT   <- finalQt(Q2,vals2[2],2)

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-1.134364 -0.169019 -0.008506  0.190237  1.097330

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.59867    0.78972  45.078  < 2e-16 ***
CDA1        -36.09370    0.96760 -37.302  < 2e-16 ***
MAP           1.44788    0.09866  14.676  < 2e-16 ***
Slope         0.34723    0.04963   6.996 6.53e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2863 on 652 degrees of freedom
Multiple R-Squared: 0.8293,Adjusted R-squared: 0.8286
```

```
F-statistic:  1056 on 3 and 652 DF,  p-value: < 2.2e-16
```

```
           [,1]       [,2]        [,3]        [,4]
[1,]  7.6093394 -9.2195944  0.55243205  0.36807256
[2,] -9.2195944 11.4232854 -0.80660576 -0.48359566
[3,]  0.5524321 -0.8066058  0.11875864  0.04243762
[4,]  0.3680726 -0.4835957  0.04243762  0.03005193
[1] 0.1053207

              [,1]
  [1,]  35.5986711
  [2,] -36.0936961
  [3,]   1.4478828
  [4,]   0.3472258
> F5.OUT   <- finalQt(Q5,vals5[2],5)

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-0.933747 -0.170028 -0.005652  0.142126  0.836051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.16386    0.14688   76.01   <2e-16 ***
CDA1        -11.27975    0.26878  -41.97   <2e-16 ***
MAP           1.27941    0.08982   14.24   <2e-16 ***
Slope         0.46404    0.04544   10.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2646 on 652 degrees of freedom
Multiple R-Squared: 0.8468,Adjusted R-squared: 0.8461
F-statistic:  1201 on 3 and 652 DF,  p-value: < 2.2e-16
```

```
            [,1]       [,2]        [,3]        [,4]
[1,]  0.30819904 -0.3799637 -0.01210666  0.03055440
[2,] -0.37996370  1.0321015 -0.23486075 -0.14338346
[3,] -0.01210666 -0.2348608  0.11524761  0.04091835
[4,]  0.03055440 -0.1433835  0.04091835  0.02949867
[1] 0.1047414

              [,1]
  [1,]  11.1638574
  [2,] -11.2797452
  [3,]   1.2794083
  [4,]   0.4640428
> F10.OUT  <- finalQt(Q10,vals10[2],10)

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-0.949383 -0.169673 -0.007528  0.145502  0.782993

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.04660    0.11863   76.26   <2e-16 ***
CDA1        -8.94992    0.21392  -41.84   <2e-16 ***
MAP          1.18822    0.09121   13.03   <2e-16 ***
Slope        0.51724    0.04625   11.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2707 on 652 degrees of freedom
Multiple R-Squared: 0.8403,Adjusted R-squared: 0.8396
F-statistic:  1143 on 3 and 652 DF,  p-value: < 2.2e-16
```

```
             [,1]       [,2]        [,3]         [,4]
[1,]  0.192009986 -0.1216301 -0.06356396 -0.000679897
```

```
[2,] -0.121630080  0.6244004 -0.17970515 -0.110637868
[3,] -0.063563963 -0.1797051  0.11352360  0.040132667
[4,] -0.000679897 -0.1106379  0.04013267  0.029183281
[1] 0.1041505

             [,1]
  [1,]  9.0465969
  [2,] -8.9499203
  [3,]  1.1882238
  [4,]  0.5172355
> F25.OUT  <- finalQt(Q25,vals25[2],25)

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.01532 -0.18941 -0.01757  0.14541  0.94158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.94932    0.11881   66.91   <2e-16 ***
CDA1        -7.62804    0.19053  -40.03   <2e-16 ***
MAP          1.09555    0.09660   11.34   <2e-16 ***
Slope        0.56992    0.04906   11.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2889 on 652 degrees of freedom
Multiple R-Squared: 0.8225,Adjusted R-squared: 0.8217
F-statistic:  1007 on 3 and 652 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]        [,4]
[1,]  0.16917355 -0.01930396 -0.09202358 -0.01816880
[2,] -0.01930396  0.43507942 -0.14753723 -0.09156878
[3,] -0.09202358 -0.14753723  0.11183416  0.03934196
[4,] -0.01816880 -0.09156878  0.03934196  0.02885127
[1] 0.1034083

             [,1]
  [1,]  7.9493152
  [2,] -7.6280350
  [3,]  1.0955540
  [4,]  0.5699208
> F50.OUT  <- finalQt(Q50,vals50[2],50)

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10808 -0.20353 -0.02085  0.14901  1.00273

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.55371    0.12581   60.04   <2e-16 ***
CDA1        -7.08958    0.18527  -38.27   <2e-16 ***
MAP          1.03865    0.10195   10.19   <2e-16 ***
Slope        0.60210    0.05183   11.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3062 on 652 degrees of freedom
Multiple R-Squared: 0.8057,Adjusted R-squared: 0.8048
F-statistic: 901.1 on 3 and 652 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]        [,4]
[1,]  0.16878887  0.01314134 -0.10338009 -0.02523111
[2,]  0.01314134  0.36601926 -0.13398330 -0.08354382
[3,] -0.10338009 -0.13398330  0.11084897  0.03887227
[4,] -0.02523111 -0.08354382  0.03887227  0.02864815
[1] 0.1029075
```

```
           [,1]
  [1,]  7.5537081
  [2,] -7.0895811
  [3,]  1.0386544
  [4,]  0.6021032
> F100.OUT <- finalQt(Q100,vals100[2],100)

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21129 -0.21871 -0.03113  0.15723  1.08687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.30661    0.13476  54.218   <2e-16 ***
CDA1        -6.71364    0.18421 -36.445   <2e-16 ***
MAP          0.98827    0.10790   9.159   <2e-16 ***
Slope        0.62951    0.05489  11.468   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3253 on 652 degrees of freedom
Multiple R-Squared: 0.7872,Adjusted R-squared: 0.7862
F-statistic: 803.9 on 3 and 652 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]        [,4]
[1,]  0.17159632  0.03242555 -0.11114151 -0.03010189
[2,]  0.03242555  0.32062551 -0.12431136 -0.07782179
[3,] -0.11114151 -0.12431136  0.11000107  0.03846329
[4,] -0.03010189 -0.07782179  0.03846329  0.02846809
[1] 0.1024386

           [,1]
  [1,]  7.3066130
  [2,] -6.7136406
  [3,]  0.9882687
  [4,]  0.6295080
```

Computational Script using PRESS Minimization and Drainage Area

```
R : Copyright 2004, The R Foundation for Statistical Computing
Version 2.0.1  (2004-11-15), ISBN 3-900051-07-0
> MLRweights <- function(vector) {
+   tmp = length(vector)/sum(vector)
+   return (tmp*vector)
+ }
>
> PRESS <- function(model) {
+   if(is.null(model$terms)) stop(“invalid ‘lm’ object: no terms”)
+   sum( (weighted.residuals(model)/(1-hatvalues(model)))^2 )
+ }
>
> DATA <- read.csv(“tx664.csv”,header=T)
> attach(DATA)
> names(DATA)
 [1] “Station” “LatD”    “LatM”    “LatS”    “LonD”    “LonM”    “LonS”
 [8] “EqYrs”   “CDA”     “MAP”     “P224”    “Slope”   “Shape”   “Q2”
[15] “Q5”      “Q10”     “Q25”     “Q50”     “Q100”    “C2”      “C25”
[22] “C100”
>
> outliers <- c(212,323,358,602,614,620,628,637)
> CDA  <- CDA[-outliers]
> Q2   <- Q2[-outliers]
> Q5   <- Q5[-outliers]
> Q10  <- Q10[-outliers]
> Q25  <- Q25[-outliers]
> Q50  <- Q50[-outliers]
```

```
> Q100 <- Q100[-outliers]
> MAP  <- MAP[-outliers]
> Slope <- Slope[-outliers]
> WEIGHTS <- MLRweights(EqYrs[-outliers])
>
> sum.lowess.slope <- function(lowss) {
+   x <- lowss$x
+   y <- lowss$y
+   n <- length(x)
+   sum <- 0
+   for(i in (seq(2,n))) {
+     delx <- x[i] - x[i-1]
+     if(delx == 0) next
+     sum <- sum + (y[i] - y[i-1])/delx
+   }
+   return(sum)
+ }
>
> doQt <- function(Q,type) {
+   smallpress <- 10000
+   smallpower <- 10000
+   smallsl    <- 1e45
+   smallslope <- 1e45
+   for(power in seq(0.04,.2,by=0.0001)) {
+     if(power == 0) next
+     power <- -1 * power
+     CDA1 <- 10^CDA
+     CDA1 <- CDA1^power
+     WLS.OUT <- lm(Q~CDA1, weights=WEIGHTS)
+     sm <- lowess(fitted(WLS.OUT),y=residuals(WLS.OUT))
+     sl <- sum(abs(sm$x - residuals(WLS.OUT)))
+     sslope <- sum.lowess.slope(sm)
+     press <- PRESS(WLS.OUT)
+     if(press < smallpress) {
+       smallpress <- press
+       smallpower <- power
+     }
+     if(sl < smallsl) {
+       smallsl <- sl
+       smallslpower <- power
+     }
+     if(sslope < smallslope) {
+       smallslope <- sslope
+       smallslopepower <- power
+     }
+   }
+   rsq <- summary(WLS.OUT)$r.squared
+   rme <- summary(WLS.OUT)$sigma
+   print(c(type,smallpower,smallpress,smallsl,smallslpower,smallslope,smallslopepower,rsq,rme))
+   return(c(type,smallpower,smallpress,smallsl,smallslpower,smallslope,smallslopepower,rsq,rme))
+ }
>
> vals2   <- doQt(Q2,2)
[1]    2.0000000   -0.0465000   72.0647133 2197.2123196   -0.0997000
[6]    3.2248956   -0.0434000    0.7120751    0.3712962
> vals5   <- doQt(Q5,5)
[1]    5.0000000   -0.0658000   59.0006834 2442.2796806   -0.1078000
[6]   -2.3722788   -0.0530000    0.7557700    0.3335610
> vals10  <- doQt(Q10,10)
[1]   10.0000000   -0.0749000   60.0973064 2567.6641814   -0.1116000
[6]   -7.5542365   -0.0567000    0.7592680    0.3318613
> vals25  <- doQt(Q25,25)
[1]   25.0000000   -0.0837000   66.7667046 2698.9820344   -0.1154000
[6]  -15.1968702   -0.0597000    0.7488384    0.3431186
> vals50  <- doQt(Q50,50)
[1]   50.0000000   -0.0890000   74.0838966 2782.5850449   -0.1176000
[6]  -19.1690278   -0.0642000    0.7351320    0.3569672
> vals100 <- doQt(Q100,100)
[1]  100.0000000   -0.0934000   82.7837507 2857.0060272   -0.1195000
[6]  -22.4427987   -0.0720000    0.7186556    0.3734860
>
> finalQt <- function(Q,power,type) {
```

```
+   CDA1 <- 10^CDA
+   CDA1 <- CDA1^power
+   WLS.OUT <- lm(Q~CDA1, weights=WEIGHTS)
+   print(summary(WLS.OUT))
+   W <- diag(WEIGHTS)
+   X = model.matrix(WLS.OUT)
+   Xt = t(X)
+   tmp <- chol2inv( chol( Xt %*% W %*% X ) )
+   wlshat1 <- X %*% tmp %*% Xt
+   print(tmp)
+   print(max(diag(wlshat1)))
+   m.wls.out <- tmp %*% Xt %*% W %*% Q
+   print(m.wls.out)
+   PRESS(WLS.OUT)
+   return(WLS.OUT)
+ }
>
> F2.OUT   <- finalQt(Q2,vals2[2],2)

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43150 -0.18074  0.01207  0.20405  1.28629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.2799     0.1002   82.63   <2e-16 ***
CDA1         -6.0308     0.1284  -46.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3309 on 654 degrees of freedom
Multiple R-Squared: 0.7713,Adjusted R-squared: 0.771
F-statistic:  2206 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]       [,2]
[1,]  0.09169293 -0.1165254
[2,] -0.11652540  0.1505865
[1] 0.01885060

            [,1]
  [1,]  8.279869
  [2,] -6.030849
> F5.OUT   <- finalQt(Q5,vals5[2],5)

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-1.064167 -0.195349  0.003766  0.160790  0.957834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.19379    0.06350  113.29   <2e-16 ***
CDA1        -4.61402    0.08928  -51.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2994 on 654 degrees of freedom
Multiple R-Squared: 0.8033,Adjusted R-squared: 0.803
F-statistic:  2671 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.04499556 -0.06218392
[2,] -0.06218392  0.08895185
[1] 0.02071799

            [,1]
  [1,]  7.193790
```

```
  [2,] -4.614025
> F10.OUT  <- finalQt(Q10,vals10[2],10)

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-1.069176 -0.212532 -0.002709  0.158296  0.889868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.96115    0.05610  124.08   <2e-16 ***
CDA1        -4.21241    0.08222  -51.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3021 on 654 degrees of freedom
Multiple R-Squared: 0.8005,Adjusted R-squared: 0.8002
F-statistic:  2625 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.03449016 -0.04941901
[2,] -0.04941901  0.07408408
[1] 0.02164636

            [,1]
  [1,]  6.961146
  [2,] -4.212410
> F25.OUT  <- finalQt(Q25,vals25[2],25)

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.14066 -0.23002 -0.01563  0.16378  1.06665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.84023    0.05276  129.64   <2e-16 ***
CDA1        -3.91357    0.08038  -48.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3184 on 654 degrees of freedom
Multiple R-Squared: 0.7838,Adjusted R-squared: 0.7834
F-statistic:  2370 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.02746451 -0.04066401
[2,] -0.04066401  0.06374535
[1] 0.02257363

            [,1]
  [1,]  6.840234
  [2,] -3.913570
> F50.OUT  <- finalQt(Q50,vals50[2],50)

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18085 -0.24244 -0.02337  0.16170  1.17321

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.80600    0.05219   130.4   <2e-16 ***
CDA1        -3.76582    0.08133   -46.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3354 on 654 degrees of freedom
Multiple R-Squared: 0.7662,Adjusted R-squared: 0.7659
F-statistic:  2144 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.02422098 -0.03653889
[2,] -0.03653889  0.05882340
[1] 0.02314619

            [,1]
  [1,]  6.806003
  [2,] -3.765824
> F100.OUT <- finalQt(Q100,vals100[2],100)

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21299 -0.25175 -0.03704  0.15891  1.26325

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.80019    0.05252  129.49   <2e-16 ***
CDA1        -3.65879    0.08336  -43.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3545 on 654 degrees of freedom
Multiple R-Squared: 0.7465,Adjusted R-squared: 0.7462
F-statistic:  1926 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.02194667 -0.03360600
[2,] -0.03360600  0.05530056
[1] 0.0236296

            [,1]
  [1,]  6.800189
  [2,] -3.658793
```