DATA ANALYTIC TOOLS FOR CHOOSING TRANSFORMATIONS IN SIMPLE LINEAR REGRESSION

Richard D. De Veaux and J. Michael Steele, Princeton University

Program in Statistics and Operations Research, School of Engineering and Applied Science, Princeton, NJ 08544

ABSTRACT

Transformations of the regressor and/or the response in simple regression are often sought to increase linear association and to make residuals appear more nearly normally distributed with constant variance. The ACE (alternating conditional expectation) algorithm of Breiman and Friedman (1985) finds the transformations maximizing the correlation between the regressor and response, while the AVAS (additivity and variance stabilization) algorithm of Tibshirani (1988) uses a variance-stabilizing transformation of the response. An exploratory data tool, the bulging rule of Mosteller and Tukey (1977) is used to find specific functional forms for the relationships suggested by the ACE and AVAS algorithms. Data on the water content of soil are used to illustrate the procedure.

1. Introduction

Recently, two powerful methods for estimating optimal transformations for regression and correlation have been proposed. For data (x_i, y_i), $1 \le i \le n$, the ACE algorithm of Breiman and Friedman finds transformations f and g such that the empirical correlation of the transformed data $(f(x_i), g(y_i))$, $1 \le i \le n$, is approximately maximized. The term ACE is an acronym for alternating condition expectation. If (X, Y) is a pair of jointly distributed random variables, one can define f and g as the limits of the functions f_n and g_n determined by taking $f_0(X) = X$, $g_0(Y) = Y$ and applying the recursions

$$f_{n+1}(X) = E(g_n(Y) \mid X)$$

and

$$g_{n+1}(Y) = E(f_n(X) \mid Y) / [\mathrm{var}(E(f_n(X) / Y))]^{1/2}$$

The f and g determined by this process can be shown to maximize $\mathrm{Corr}(f(X), g(Y))$ (subject to $\mathrm{var}(g(Y)) = 1$). Naturally, if the joint distribution of X and Y is not known, one cannot find f and g precisely by this method, but one can derive an empirically based algorithm as a natural modification of the theoretical algorithm, by replacing the conditional expectations by scatterplot smoothers. In their implementation, Breiman and Friedman used a refined version of the supersmoother of Friedman and Stuetzle (1982).

One problem of ACE is that the transformations, while maximizing linear association, may introduce heteroscedasticity in the response. The AVAS algorithm of Tibshirani (1988) is designed to alleviate this problem. It is similar to the ACE algorithm except that instead of using $g_{n+1}(y) = E(f_n(x) \mid y) / [\mathrm{var}(E(f_n(x) \mid y))]^{1/2}$ it uses the asymptotic variance-stabilizing transformation. (The details of the algorithm can be found in Tibshirani (1988)).

The ACE and AVAS algorithms can be useful as stand-alone tools for descriptive purposes. The result of each algorithm is two estimated functions $\hat{f}(x_i)$ and $\hat{g}(y_i)$, $1 \le i \le n$. However, it is often desirable or even necessary to obtain specific functional forms for f and g, that is, to find functions which approximate f and g and retain the desirable properties of the ACE or AVAS transformations.

The purpose of this paper is to illustrate the use of the bulging rule of Mosteller and Tukey (1977) as an aid in finding an explicit functional form approximating the ACE and AVAS transformations. Additionally, we compare the differences in the transformations suggested in these algorithms. Data from an experiment on soil water diffusivity are used as an example. The parameter of interest, the diffusion coefficient is a product of two functionals (a derivative and an integral) of X and Y. By finding an explicit additive functional form $F(Y) = a + bG(X)$ we are able to calculate the functionals explicitly.

This article is organized as follows. The diffusion problem is more extensively described in Section 2. The procedure for finding the transformations is outlined in Section 3. In Section 4 we carry out the procedure on the experimental data in detail. Section 5 contains discussion and some concluding remarks.

2. Soil Water Diffusion Problem

The movement of water in a horizontal column of unsaturated soil is commonly modeled by means of the one-dimensional diffusion equation

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial x}\left[D(\theta)\frac{\partial \theta}{\partial x}\right] \qquad 0 < x < \infty, \quad 0 < t < \infty, \qquad (2.1)$$

where θ is the water content of the soil, t is time, x is the position in the horizontal column, and $D(\theta)$ is the coefficient of soil water diffusivity at the moisture level θ. Any two variable function $\theta(x, t)$ that satisfies (2.1) can be shown (*c.f.* Jost (1960, p. 31)) to be a function of the single

variable $\lambda = x/t^{1/2}$, which is often called the Boltzman variable. After writing $\theta(\lambda)$ for the new function of one variable, one can check that $\theta(\lambda)$ satisfies the ordinary differential equation:

$$-\frac{\lambda}{2}\frac{d\theta}{d\lambda} = \frac{d}{d\lambda}\left[D(\theta)\frac{d\theta}{d\lambda}\right]. \qquad (2.2)$$

From this equation one can then easily obtain the expression for $D(\theta)$ which underlies our approach to its estimation:

$$D(\theta) = -\frac{1}{2}\frac{d\lambda}{d\theta}\int_{\theta_0}^{\theta}\lambda(u)du, \qquad (2.3)$$

where θ_0 is the initial water content of the soil. The simultaneous appearance of both derivative and integral terms in this expression for $D(\theta)$ provides one of the most intriguing features of its estimation.

The process that has been most widely used to estimate $D(\theta)$ is the transient-flow experiment of Bruce and Klute (1956). In that experiment, water is held at a constant head and permitted to infiltrate into a horizontal column containing air-dry soil. After a fixed time interval, the column is sectioned, and the water content of the individual sections is determined either by weighing, or by other methods. The data of Clothier and Scotter (1982) on Manawatu sandy loam plotted in Figure 1 are typical of those obtained through horizontal infiltration experiments. They also give an indication of some of the inherent difficulties in estimating $D(\theta)$. For instance, many smoothing methods when applied to the data of Figure 1 would lead to a virtually useless estimate of the derivative of λ with respect to θ. We will return to the problem of estimating $D(\theta)$ in Section 4. For further details on the experiment and historical background the reader is referred to De Veaux and Steele (1989) and Clothier and Scotter (1982).

3. Estimation Process

The details of the method we propose are possibly best explained in the context of an example such as the analysis of $D(\theta)$ of Manawatu sandy loam. Moreover, one almost has to have an honest example in hand in order to detail the role of the tools we have used to assist our transformation choice: the bulging rule and the ACE and AVAS algorithms. With that said, it seems useful to have a top-down view of the method of the proposed method. The four basic steps are the following:

Step 1. Find estimated transformations $\hat{f}(\theta)$ and $\hat{g}(\lambda)$ from the ACE and AVAS algorithms. (We use λ for the regressor and θ for the response.) The transformed data values $(\hat{f}(\theta_i), \hat{g}(\lambda_i))$ in both cases will exhibit a strong linear association. The R^2 values from the regressions of $\hat{f}(\theta_i)$ on $\hat{g}(\lambda_i)$ will be used as benchmarks against which we will compare the R^2 from more analytically tractable transformations.

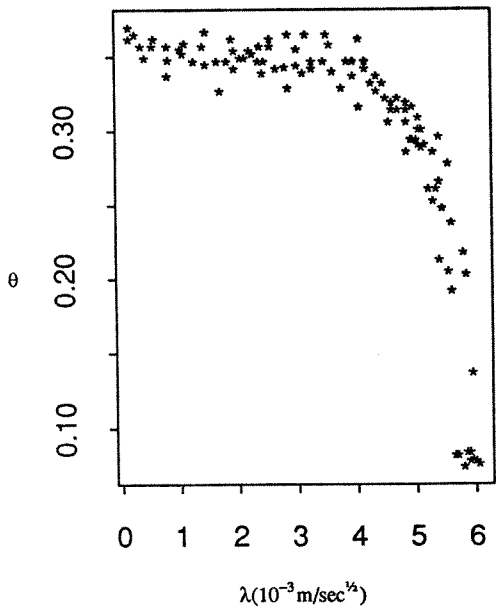

Figure 1. *Scatterplot of the volumetric water content, θ, versus the Boltzman variable λ.*

Step 2. Use the so-called bulging rule of Mosteller and Tukey (1977) to suggest analytically tractable functions $F(\theta)$ and $G(\lambda)$ which retain the desirable properties of the functions found in Step 1.

Step 3. Perform the regression of $F(\theta)$ on $G(\lambda)$, and use diagnostic tools to assess the appropriateness of the linear model:

$$F(\theta) = a + bG(\lambda) + \varepsilon. \qquad (3.1)$$

Step 4. If functionals of θ or λ need to be estimated, one can use (3.1) directly to obtain estimates.

As an example of step 4, consider the case of the diffusion equation (2.1). After performing steps 1-3, one would use the chain rule to extract from (3.1) an expression for $\frac{d\lambda}{d\theta}$ in terms of θ. Then either analytic or numerical integration is used to determine the values of the definite integrals:

$$I(\theta) = \int_{\theta_0}^{\theta}\lambda(u)\,du \qquad (3.2)$$

for all $\theta_0 \le \theta \le \theta_s$. Finally the diffusion coefficient $D(\theta)$ is estimated by the expression

$$D(\theta) = -\frac{1}{2}\frac{d\lambda}{d\theta}\int_{\theta_0}^{\theta} \lambda(u)\,du \qquad (3.3)$$

where the indicated derivative and integral are those determined previously.

4. An Example: Manawatu Sandy Loam

To understand the extent of the linear association between λ and θ that can be achieved by marginal transformations, we examine the results of applying the ACE and AVAS algorithm. Even for the best choices of f and g, the linear association between λ and θ is imperfect. Still, the ACE transformed variables plotted in Figure 2 and the AVAS transformed variables plotted in Figure 3 exhibit substantially greater linear association than the plot of the untransformed variables given in Figure 1. (We have used the implementation of the empirical ACE algorithm due to L. Brieman which is incorporated in *The Statistics Store* (J.M. Schilling (1985)), and the implementation of AVAS obtained from R. Tibshirani (see Tibshirani (1988)). When we measure the linear association of the transformed variables in terms of R^2, we find respectable values of $R^2 = .93$ for ACE and $R^2 = .92$ for AVAS. These values provide us with a benchmark, and, in fact, one of the principal benefits of the ACE and AVAS algorithms is that they provide a standard against which more analytically appealing transformations can be judged.

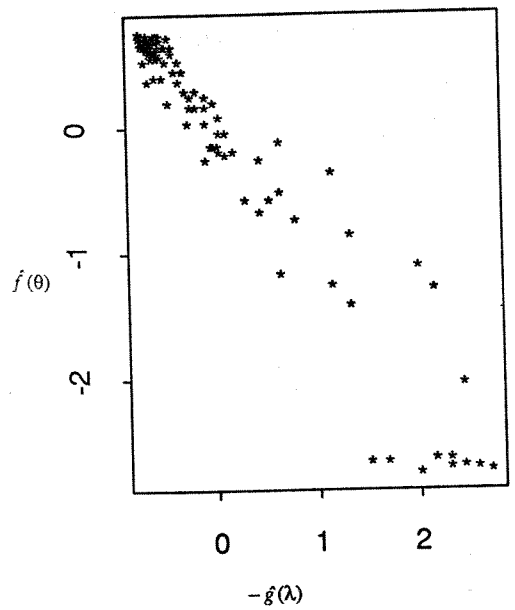

Figure 2. *Scatterplot of the ACE transformed θ_i versus the ACE transformed λ_i. This plot is used to assess linearity of the ACE transformation.*

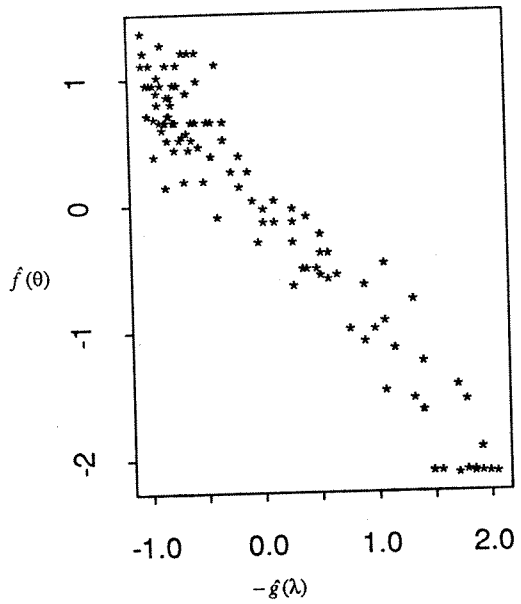

Figure 3. *Scatterplot of the AVAS transformed θ_i versus the AVAS transformed λ_i. This plot is used to assess linearity of the AVAS transformation.*

To aid the search for such surrogates for f and g, the ACE and AVAS transformed variables are plotted against the untransformed variables to see if simpler functional forms might suffice. Figures 4 and 5 show the plots of $(\lambda_i, \hat{g}(\lambda_i))$, $1 \le i \le n$ and $(\theta_i, \hat{f}(\theta_i))$, $1 \le i \le n$, respectively, for both ACE and AVAS.

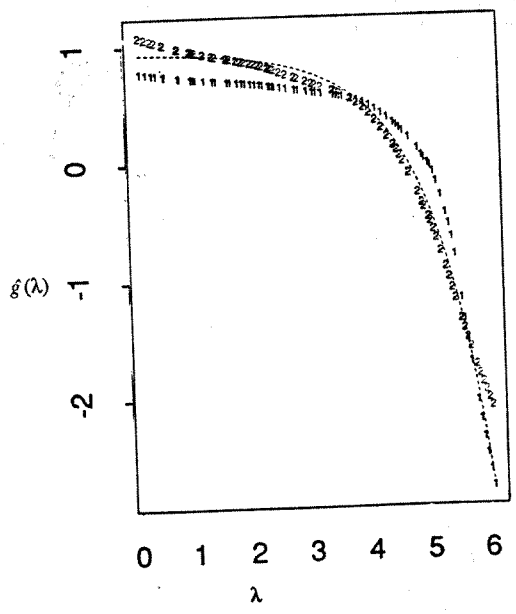

Figure 4. *Scatterplot of the ACE and AVAS transformed λ_i versus λ_i that is used to suggest a power transformation approximating $g(\lambda)$.*
1 = ACE 2 = AVAS ----- = standardized $(-e^{\lambda})$

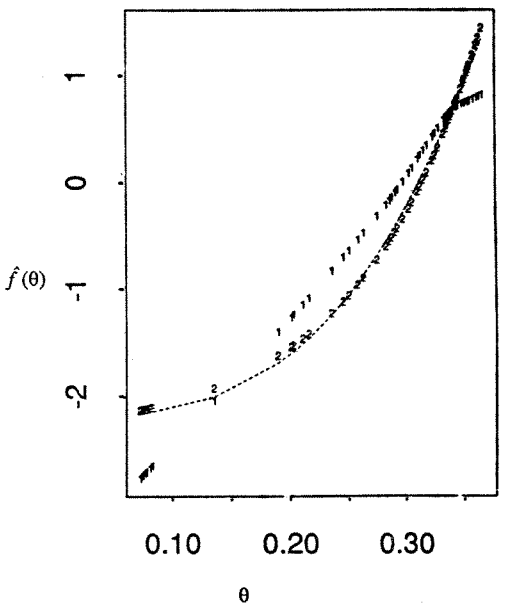

Figure 5. *Scatterplot of the ACE and AVAS transformed θ_i versus θ_i that is used to suggest a power transformation approximating $f(\theta)$. Notice that the bulge rule may not be directly applicable here.*
1 = ACE 2 = AVAS ----- = standardized (θ^3)

The hunt for analytically tractable replacements for f and g is further guided by the so-called bulging rule of Mosteller and Tukey (1977). Loosely speaking, the bulging rule suggests finding an outward normal to a smoothed plot of the data and using the signs of the normal components to guide one's choice of transfomation. For example, Figure 4 exhibits a bulge where both the x and y components of the outward normal are positive. The bulging rule then suggests that both the variables plotted on the horizontal axis should be transformed by moving up the scale of powers. In fact, the successive examination of plots of $(\lambda_i^{\alpha}, g(\lambda_i))$, $1 \le i \le n$, for larger values of α continues to suggest moving up the scale of powers, and we are thus led to consider the exponential transformation. For comparison $-e^{\lambda}$ (standardized to have mean 0 and variance 1) is also shown in Figure 4. The exponential appears to be a compromise between the transformations suggested by ACE and AVAS. An alternative approach to this exploratory search for an appropriate transformation would be to use the method of Box and Cox (1964).

When we begin a similar examination of the plot of $(\theta_i, \hat{f}(\theta_i))$, $1 \le i \le n$ given in Figure 5, the bulging rule for re-expression diverges for ACE and AVAS. For the ACE transformation, there may be a modest indication that we might wish to send θ down the scale of powers, but the indication is not supported when tried. Fortunately, we have recourse to a second approach that does suggest an appropriate transformation, and we can consider the plot of θ_i versus e^{λ_i} which is shown in Figure 6. After all, since we have having settled on e^{λ} as the surrogate for $f(\lambda)$, the principal remaining task is to determine a surrogate F for f such that the scatterplot $(F(\theta_i), e^{\lambda_i})$ is approximately linearized. The bulging rule applied to Figure 6 initially suggests that we consider a transformation F that moves θ up the ladder of powers, and successive applications of the bulging rule eventually lead us to the choice of $F(\theta) = \theta^3$.

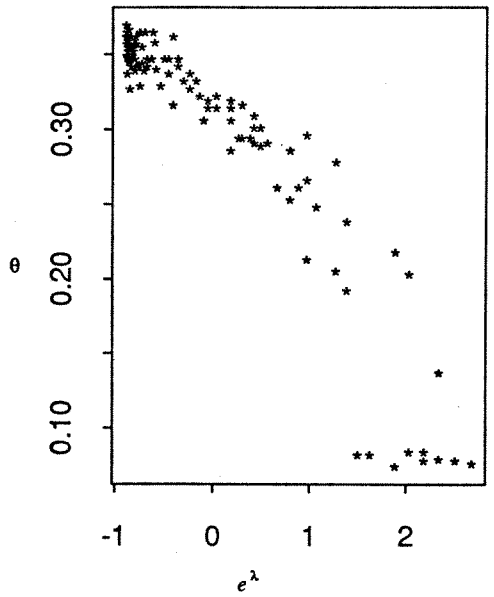

Figure 6. *Scatterplot of θ_i versus e^{λ_i}. The bulge rule suggests going up the ladder for either θ or λ.*

For the AVAS transformation in Figure 4, the bulging rule is directly applicable and suggests using $F(\theta) = \theta^3$ again. (The correlation between the AVAS $\hat{f}(\theta_i)$ and θ_i^3 is .999.) Thus, for this data set, we are led to the same transformation from both algorithms. Notice that $G(\lambda) = e^{\lambda}$ and $F(\theta) = \theta^3$ preserve the homoscedasticity of the AVAS transformations and the linearity of both ACE and AVAS. Strikingly, using $F(\theta) = \theta^3$ and $G(\lambda) = e^{\lambda}$ achieves an R^2 of .93 that meets the level of the optimal $R^2 = .93$ achieved by the ACE transformations. Moreover, when we consider the plot of θ_i^3 versus e^{λ_i} (both standarized) given in Figure 7, the visual impact of the linear association exhibited by this figure seems to compare well with that exhibited by the ACE transformed variables of Figure 2 and the AVAS transformed variables of Figure 3. On the basis of the quantitative evidence provided by comparing R^2's, the subjective evidence provided by comparison of the scatterplots of Figures 2, 3, and 7, and the fact that both the ACE and AVAS algorithms suggested the same transformations, it seems appropriate to settle on the transformation choices of Figure 7.

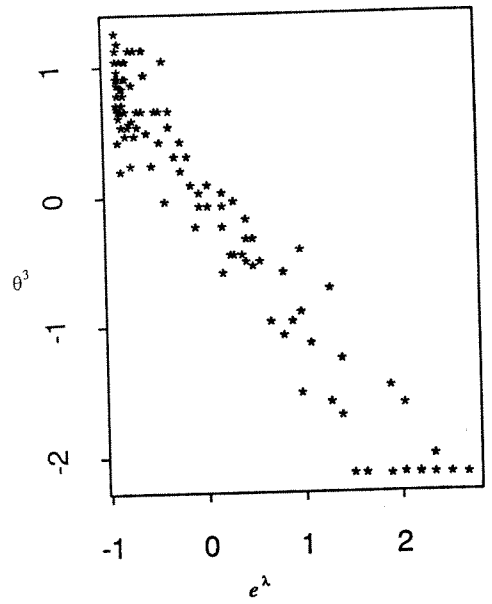

Figure 7. *Scatterplot of θ_i^3 versus e^{λ_i} (standardized). Approximate linearity is achieved with this transformation which should be compared with the ACE transformations of Figure 2.*

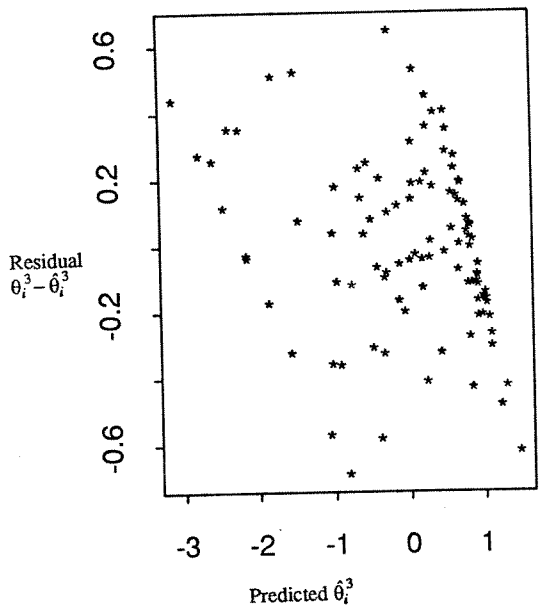

Figure 8. *Scatterplot of residuals versus predicted values for the model $\hat{\theta}_i^3 = a + be^{\lambda_i}$. Residuals appear to be approximately homoscedastic.*

For the Manawatu sandy loam data our exploratory analysis has led us to an approximate relationship of the form

$$F(\theta) = a + bG(\lambda), \tag{4.1}$$

where $F(\theta) = \theta^3$, $G(\lambda) = e^{\lambda}$. The coefficients in (4.1) can now be estimated by ordinary least squares from which we obtain $a = 4.48 \times 10^{-2}$, and $b = -1.20 \times 10^{-4}$ with nominal standard errors of 5.30×10^{-4} and 3.30×10^{-6}, respectively. In Figure 8, we show a plot of the predicted values $\hat{\theta}_i^3$ versus the residuals that are obtained from fitting the model (4.1) by ordinary least squares. The residuals appear approximately homoscedastic, and we have no reason to be discontent with the estimates obtained by ordinary least squares. If the scatterplot of Figure 8 had exhibited a greater heteroscedasticity, we would have probably elected to apply iteratively reweighted least squares, or a similarly directed technique.

As a final check on the reasonability of the fitted model, one should consider the fit in terms of the untransformed variables as exhibited in Figure 8. This plot has no flagrant defects; indeed it suggests that the procedure has been a reasonable one.

By differentiating (4.1) we find the general relationship

$$\frac{d\lambda}{d\theta} = \frac{F'(\theta)}{bG'(\lambda)} \tag{4.2}$$

$$= b^{-1}F'(\theta) / \{G'(G^{-1}((F(\theta) - a)/b))\},$$

and, for the choices that were made by means of the exploratory analysis of the Manawatu sandy loam data, one finds a particularly simple net result:

$$\frac{d\lambda}{d\theta} = 3\theta^2(\theta^3 - a)^{-1}. \tag{4.3}$$

In order to obtain $D(\theta)$ it remains only to determine the integral of $\lambda(\theta) = G^{-1}((F(\theta) - a)/b)$. For the Manawatu sandy loam data we find $\lambda(\theta) = \log((\theta^3 - a)/b)$, and the integral of $\lambda(\theta)$ can be determined analytically. For more details including a discussion of interval estimates of $D(\theta)$, the reader is referred to De Veaux and Steele (1989).

5. Discussion

Both the ACE and AVAS algorithms were used to suggest analytic forms for a transformation of a regressor and response which would exhibit linearity and homoscedasticity. To aid the search for such functional forms, the bulging rule of Mosteller and Tukey (1977) was used when appropriate. The ACE transformation, while displaying a high degree of linearity ($R^2 = .93$) also showed nonconstant variance in the response. The AVAS transformations did nearly as well in terms of linearity ($R^2 = .92$) and the transformed response had nearly constant variances. For our data set, both the ACE and AVAS algorithms led to the same function forms $G(\lambda) = e^{\lambda}$ and $F(\theta) = \theta^3$ for the regressor and response respectively. Functionals of the curves were directly attainable from the linear model $F(\theta) = a + bG(\lambda)$ which through residual analysis seemed

plausible. The success of the procedure in this case suggests that both the ACE and AVAS transformations should be considered as exploratory tools by the data analyst to suggest appropriate functional forms for transformation.

References

Box, G.E.P. and Cox, D.R. (1964). "An analysis of transformations," *J. Royal Statist. Soc.*, **B26**, 211-243, discussion 244-252.

Breiman, L. and Friedman, J.H. (1985). "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Statist. Soc.*, **80**, 580-597.

Bruce, R.R. and Klute, A. (1956). "The measurement of soil-moisture diffusivity," *Soil Sci. Soc. Am. Proc.*, **20**, 458-462.

Clothier, B.E. and Scotter, D.R. (1982). "Constant-flux infiltration from a hemispherical cavity," *Soil Sci. Soc. Am. J.*, **46**, 696-700.

De Veaux, R.D. and Steele, J.M. (1989). "ACE guided transformation method for estimation of the coefficient of soil water diffusivity," *Technometrics*, (In press).

DuChateau, P.C., Nofziger, D.L., Ahuja, L.R., and Schwartzendruber, D. (1972). "Experimental curves and rates of change from piecewise parabolic fits," *Agron. J.*, **64**, 538-542.

Friedman, J.H. and Stuetzle, W. (1982). "Smoothing of scatterplots," Technical Report Orion 3, Department of Statistics, Stanford University.

Jost, W. (1960). *Diffusion in Solids, Liquids, and Gases* 3rd ed., Academic Press, New York.

Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression*, Addison-Wesley, Reading, MA.

Schilling, J.M. (1985). *The Statistics Store*, AT&T Bell Laboratories, Murray Hill, NJ.

Tibshirani, R. (1988). "Estimating transformations for regression via additivity and variance stabilization," *J. Amer. Statist. Soc.*, **83**, 394-405.