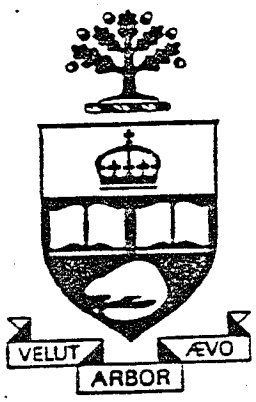

Estimating transformations for regression: a variation on ACE

by

Robert Tibshirani

Department of Preventative Medicine and Biostatistics and Department of Statistics
University of Toronto

Technical Report No. 4, February (1986)

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Estimating transformations for regression: a variation on ACE

Robert Tibshirani
Department of Preventive Medicine & Biostatistics
and
Department of Statistics
University of Toronto

ABSTRACT

We propose a variant of Breiman and Friedman's ACE (Alternating Conditional Expectation) algorithm (JASA 1985, vol 80, pg 580-619) for the estimation of optimal transformations in regression and correlation. ACE is a powerful, versatile procedure that produces nonparametric estimates of transformations of variables X and Y by alternately smoothing the current transformation for X on Y and the current transformation for Y on X. Our proposal is aimed specifically at regression problems and differs from ACE in one way: if Y is the response variable in the problem, we use the (approximate) variance stabilizing transformation based on the current model as the estimated function for Y. The estimates of transformations for the predictor variable (or variables) are obtained as in ACE. We call the resulting algorithm "RACE" for Regression by Alternating Conditional Expectation. In a number of examples, many of which were looked at by the discussants to the forementioned paper, RACE seems to alleviate some of the anomalies of ACE in regression problems. These include the failure of ACE to reproduce model transformations and sensitivity to the marginal distribution of the predictors.

1. Introduction.

L. Breiman and J. Friedman (1985) (hereafter BF) proposed a powerful method called "ACE" (Alternating Conditional Expectation) for the estimation of optimal transformations for regression and correlation. The idea of their algorithm is best described not for data but for random variables with a known distribution. Given random variables X and Y, ACE finds the transformations $\theta(Y)$ and $\phi(X)$ that maximize the correlation $corr(\theta(Y), \phi(X))$ subject to $Var\ \theta(Y) = 1$. Equivalently, it finds the transformations that minimize $E(\theta(Y) - \phi(X))^2$ subject to $Var\ \theta(Y) = 1$. ACE achieves this by repeated alternation of two conditional expectations: $\phi(X) = E(\theta(Y) \mid X)$, and $\theta(Y) = E(\phi(X) \mid Y)/Var(E(\phi(X) \mid Y))$ until convergence. Given data realizations of X and Y, the ACE algorithm replaces the conditional expectations by scatterplot smoothers. In their implementation, BF use a refined version of a running lines smoother, called the "supersmoother" (Friedman and Stuetzle 1982). The result of ACE is two estimated functions $\hat{\theta}(Y)$ and $\hat{\phi}(X)$, useful for descriptive purposes or to suggest transformations for the regression of Y on X or X on Y.

ACE is a powerful and useful tool, but for a number of reasons it seems more suited for correlation analysis than regression. In particular, note that ACE is essentially symmetric in X and Y (the standardizing of $\theta(Y)$ could be applied to $\phi(X)$ without essentially changing the solutions). More specifically, suppose Y is in fact a response variable and X a predictor variable, with

$$\theta^0(Y) = \phi^0(X) + \varepsilon \tag{1}$$

where $\theta^0(Y)$ is invertible, $E(\varepsilon \mid X) = 0$ and $Var(\varepsilon \mid X) = \sigma^2$, $0 < \sigma^2 < \infty$.

Now since $\theta^0(Y)$ is assumed to be invertible, specification of the distributions of ε and X determines a joint distribution of X and Y, say $\mathcal{L}(X,Y)$. Given $\mathcal{L}(X,Y)$, the transformations $\theta^*(Y)$ and $\phi^*(X)$ that maximize $corr(\theta^*(Y), \phi^*(X))$ are NOT in general $\theta^0(Y)$ and $\phi^0(X)$ (BF pg 581). Since $\theta^*(Y)$ and $\phi^*(X)$ are the solutions produced by ACE, we see that ACE does not reproduce model transformations. The discussants in BF pointed out other problems with ACE in the regression context. These include sensitivity to the shape of the marginal distribution of X and the collapsing of disjoint clusters. We believe that it is the smoothing

1. Introduction.

L. Breiman and J. Friedman (1985) (hereafter BF) proposed a powerful method called "ACE" (Alternating Conditional Expectation) for the estimation of optimal transformations for regression and correlation. The idea of their algorithm is best described not for data but for random variables with a known distribution. Given random variables X and Y, ACE finds the transformations $\theta(Y)$ and $\phi(X)$ that maximize the correlation $corr(\theta(Y), \phi(X))$ subject to $Var\ \theta(Y) = 1$. Equivalently, it finds the transformations that minimize $E(\theta(Y) - \phi(X))^2$ subject to $Var\ \theta(Y) = 1$. ACE achieves this by repeated alternation of two conditional expectations: $\phi(X) = E(\theta(Y) \mid X)$, and $\theta(Y) = E(\phi(X) \mid Y)/Var(E(\phi(X) \mid Y))$ until convergence. Given data realizations of X and Y, the ACE algorithm replaces the conditional expectations by scatterplot smoothers. In their implementation, BF use a refined version of a running lines smoother, called the "supersmoother" (Friedman and Stuetzle 1982). The result of ACE is two estimated functions $\hat{\theta}(Y)$ and $\hat{\phi}(X)$, useful for descriptive purposes or to suggest transformations for the regression of Y on X or X on Y.

ACE is a powerful and useful tool, but for a number of reasons it seems more suited for correlation analysis than regression. In particular, note that ACE is essentially symmetric in X and Y (the standardizing of $\theta(Y)$ could be applied to $\phi(X)$ without essentially changing the solutions). More specifically, suppose Y is in fact a response variable and X a predictor variable, with

$$\theta^0(Y) = \phi^0(X) + \varepsilon \tag{1}$$

where $\theta^0(Y)$ is invertible, $E(\varepsilon \mid X) = 0$ and $Var(\varepsilon \mid X) = \sigma^2$, $0 < \sigma^2 < \infty$.

Now since $\theta^0(Y)$ is assumed to be invertible, specification of the distributions of ε and X determines a joint distribution of X and Y, say $\mathcal{L}(X,Y)$. Given $\mathcal{L}(X,Y)$, the transformations $\theta^*(Y)$ and $\phi^*(X)$ that maximize $corr(\theta^*(Y), \phi^*(X))$ are NOT in general $\theta^0(Y)$ and $\phi^0(X)$ (BF pg 581). Since $\theta^*(Y)$ and $\phi^*(X)$ are the solutions produced by ACE, we see that ACE does not reproduce model transformations. The discussants in BF pointed out other problems with ACE in the regression context. These include sensitivity to the shape of the marginal distribution of X and the collapsing of disjoint clusters. We believe that it is the smoothing

on Y that causes much of the difficulty.

In this paper we propose an alternative to ACE that we believe is more suitable for the estimation of regression transformations It differs from ACE in one way: instead of using $E(\phi(X) \mid Y)$ as the current estimate of $\theta(Y)$, it uses the (estimated) variance stabilizing transformation. We call the resultant algorithm "RACE" for Regression by Alternating Conditional Expectation, or Regression ACE for short. RACE seems to solve the regression difficulties suffered by ACE. In particular, if model (1) holds, then $\theta^0(Y)$ and $\phi^0(X)$ are fixed points of the RACE procedure. Other problems with ACE, both in constructed and real data sets, seem to be alleviated by RACE (see Section 3).

We have yet to discuss multiple predictors. ACE incorporates these through the so-called backfitting algorithm; RACE also uses backfitting, so that again the only difference between ACE and RACE is the method used to estimate $\theta(Y)$.

A word of caution is appropriate. Although the results of RACE are very encouraging, we do not have substantial theoretical support for RACE like that provided by BF for ACE. Also, RACE has not yet been given the careful practical scrutiny that uncovered the difficulties with ACE. Indeed, such scrutiny could reveal anomalies of RACE of a different nature.

This paper is layed out as follows. In Section 2 we define the RACE procedure both for single and multiple predictor models. In Section 3 we discuss a number of examples, comparing the performance of ACE and RACE. In Section 4 we conclude with some discussion including a conjecture about the convergence of RACE.

2. The RACE algorithm.

2.1. Single predictor case

Suppose we have random variables X and Y satisfying (1). Assume $\mathcal{L}(Y \mid X = x)$ is absolutely continuous for each x, and F is not degenerate. For convenience we assume that $EX = EY = 0$, and also that all estimated functions have mean 0.

Our problem can be stated as follows: find functions $\theta(Y)$ and $\phi(X)$ such that $\phi(X) = E(\theta(Y) \mid X)$ and $Var(\theta(Y) \mid \phi(X)) = \textit{constant}$, subject to $E(\theta(Y) - \phi(X))^2 = 1$. This last

constraint rules out the solution $\theta(Y) = \phi(X) \equiv 0$ a.e. $(\mathcal{L}(X,Y))$ and also solutions for which $\theta(Y) = \phi(X)$ a.e. $(\mathcal{L}(X,Y))$. Note that $\theta^0(Y)$ and $\phi^0(X)$, suitably scaled, satisfy these conditions. Now given a function $\phi(X)$, the (approximate) variance stabilizing transformation for Y is $\theta(Y) = \int^Y [1/\sqrt{E(Y-\phi)^2 \mid \phi}]d\phi$. Alternating this with the step $\phi(X) = E(\theta(Y) \mid X)$ forms the basis of the RACE algorithm. Denoting the percentage variance explained by $r^2(\theta,\phi) = 1 - E(\theta(Y) - \phi(X))^2/E\theta^2(Y)$, the RACE algorithm is given below.

The RACE algorithm for a single predictor

Initialize: $\phi(X) = E(Y \mid X)$, $\theta(Y) = Y$.

Iterate until $r^2(\theta,\phi)$ fails to decrease

$$v(u) = E(\theta(Y) - \phi(X))^2 \mid \phi(X) = u)$$

$$\theta(Y) = \int_0^{\theta(Y)} \frac{1}{\sqrt{v(u)}} du$$

$$\theta(Y) = \theta(Y) - E\theta(Y)$$

$$\phi(X) = E(\theta(Y) \mid X)$$

$$e = \sqrt{E(\theta(Y) - \phi(X))^2}$$

$$\theta(Y) = \theta(Y)/e,\ \phi(X) = \phi(X)/e$$

End loop

In order for $\int_0^{\theta(Y)} \frac{1}{\sqrt{v(u)}} du$ to make sense in general, we define the integrand to be 0 where $v(u)$ is undefined and consider the integral to be a sum when X is a categorical variable.

If instead we have a sample from the joint distribution of X and Y, we replace the conditional expectations by a scatterplot smoother and e and $r^2(\theta,\phi)$ by their sample analogues, as in the ACE algorithm. Any scatterplot smoother might be used: for example a spline smoother (see Wahba and Wold 1979 or Silverman 1985) or a robust kernel smoother (Cleveland 1979). We have chosen to use the super-smoother employed by BF to facilitate comparisons with ACE. (As in the ACE algorithm, if X is categorical, the "smoother" would simply take averages in each category.) An estimate of the integral for $\theta(Y)$ is also needed: we use a trapezoid rule with linear interpolation between sample values of Y. We use a change in

constraint rules out the solution $\theta(Y) = \phi(X) \equiv 0$ a.e. $(\mathcal{L}(X,Y))$ and also solutions for which $\theta(Y) = \phi(X)$ a.e. $(\mathcal{L}(X,Y))$. Note that $\theta^0(Y)$ and $\phi^0(X)$, suitably scaled, satisfy these conditions. Now given a function $\phi(X)$, the (approximate) variance stabilizing transformation for Y is $\theta(Y) = \int^Y [1/\sqrt{E(Y-\phi)^2 \mid \phi}]d\phi$. Alternating this with the step $\phi(X) = E(\theta(Y) \mid X)$ forms the basis of the RACE algorithm. Denoting the percentage variance explained by $r^2(\theta,\phi) = 1 - E(\theta(Y) - \phi(X))^2/E\theta^2(Y)$, the RACE algorithm is given below.

The RACE algorithm for a single predictor

Initialize: $\phi(X) = E(Y \mid X)$, $\theta(Y) = Y$.

Iterate until $r^2(\theta,\phi)$ fails to decrease

$$v(u) = E(\theta(Y) - \phi(X))^2 \mid \phi(X) = u)$$

$$\theta(Y) = \int_0^{\theta(Y)} \frac{1}{\sqrt{v(u)}} du$$

$$\theta(Y) = \theta(Y) - E\theta(Y)$$

$$\phi(X) = E(\theta(Y) \mid X)$$

$$e = \sqrt{E(\theta(Y) - \phi(X))^2}$$

$$\theta(Y) = \theta(Y)/e,\ \phi(X) = \phi(X)/e$$

End loop

In order for $\int_0^{\theta(Y)} \frac{1}{\sqrt{v(u)}} du$ to make sense in general, we define the integrand to be 0 where $v(u)$ is undefined and consider the integral to be a sum when X is a categorical variable.

If instead we have a sample from the joint distribution of X and Y, we replace the conditional expectations by a scatterplot smoother and e and $r^2(\theta,\phi)$ by their sample analogues, as in the ACE algorithm. Any scatterplot smoother might be used: for example a spline smoother (see Wahba and Wold 1979 or Silverman 1985) or a robust kernel smoother (Cleveland 1979). We have chosen to use the super-smoother employed by BF to facilitate comparisons with ACE. (As in the ACE algorithm, if X is categorical, the "smoother" would simply take averages in each category.) An estimate of the integral for $\theta(Y)$ is also needed: we use a trapezoid rule with linear interpolation between sample values of Y. We use a change in

$r^2(\theta,\phi)$ of .001 as a convergence criterion, and convergence was typically obtained in ≤ 5 iterations in the examples discussed in Section 3.

2.2. RACE for multiple predictors

With multiple predictors $X_1, X_2, \ldots X_p$, our model takes the form $\theta^0(Y) = \sum_1^p \phi_j^0(X_j) + \varepsilon$, where $\theta^0(Y)$ is invertible, $E(\varepsilon \mid X_1, X_2, \ldots X_p) = 0$ and $Var(\varepsilon \mid X_1, X_2, \ldots X_p) = \sigma^2$, $0 < \sigma^2 < \infty$. In the multiple predictor version of RACE, given a function $\theta(Y)$, we need to find functions $\phi_1(X_1), \ldots \phi_p(X_p)$ such that $E(\theta(Y) \mid X_1, \ldots X_p) = \sum_1^p \phi_j(X_j)$, We achieve this through the backfitting algorithm used by Breiman and Friedman. It consists of cycling through $X_1, X_2, \ldots X_p$, updating $\phi_j(X_j)$ by $\phi(X_j) = E(\theta(Y) - \sum_{i \neq k} \phi_i(X_i))$ until the expected mean squared error of the model doesn't change much. BF show that backfitting does indeed convergence to functions $\phi_1(X_1), \ldots \phi_p(X_p)$ such that $E\theta(Y) = \sum_1^p \phi_j(X_j)$, under regularity conditions. Letting $r^2(\theta, \phi_1, \ldots \phi_p) = 1 - E(\theta(Y) - \sum_1^p \phi_j(X_j))^2 / E^2\theta(Y)$ we have the multiple predictor RACE algorithm:

The RACE algorithm for multiple predictors.

Initialize: $\theta(Y) = Y$; Backfit Y on $X_1, \ldots X_p$ to give $\phi_1(X_1), \ldots \phi_p(X_p)$

Iterate until $r^2(\theta, \phi_1, \phi_2, \ldots \phi_p)$ fails to decrease

$v(u) = E(\theta(Y) - \sum_1^p \phi_j(X_j))^2 \mid \sum_1^p \phi_j(X_j) = u)$

$\theta(Y) = \int_0^{\theta(Y)} \frac{1}{\sqrt{v(u)}} du$

$\theta(Y) = \theta(Y) - E\theta(Y)$

Backfit $\theta(Y)$ on $X_1, \ldots X_p$ to give $\phi_1(X_1), \ldots \phi_p(X_p)$

$e = \sqrt{E(\theta(Y) - \sum_1^p \phi_j(X_j))^2}$

$\theta(Y) = \theta(Y)/e$, $\phi_j(X) = \phi_j(X)/e$, $j = 1, \ldots p$

End loop

As in the single predictor RACE algorithm, the data version of the above algorithm replaces each conditional expectation by a scatterplot smoother, the integral by an estimate

based on the trapezoid rule, and e and $r^2(\theta, \phi_1, \phi_2, \ldots \phi_p)$ by their sample analogues.

Remark A. The ACE algorithm does not assume that $\theta(Y)$ is monotone (although mononiticity can be enforced in the data algorithm). This makes sense in the ACE setting because the optimal transformation for Y given the joint distribution of X and Y may not be monotone. In the present setting, however, we assume that X and Y are generated from a model of the form (1) and this is meaningful only if $\theta(Y)$ is monotone. Appropriately, the estimate of $\theta(Y)$ given by RACE is monotone, since it is the integral of a positive function.

Remark B. A strength of the RACE algorithm is that the solutions it produces are (essentially) independent of the marginal distribution of X. In particular, if X is transformed by a monotone function $f(X)$ before applying RACE, then ignoring scaling, $\theta(Y)$ is unchanged and $\phi(X)$ is mapped in the obvious way, i.e. $\phi^{new}(f(X)) = \phi^{old}(X)$. This is not the case for the ACE algorithm. Note that in the data version of RACE, this invariance will hold only approximately because a scatterplot smoother typically uses an implicit metric in the X space.

Remark C. One of the assumptions in model (1) is $Var(\theta(Y) \mid X) = \sigma^2$, but in the RACE algorithm we have used the quantity $Var(\theta(Y) \mid \phi(X))$. It is clear, however, that $Var(\theta(Y) \mid X) = \sigma^2$ implies $Var(\theta(Y) \mid \phi(X)) = \sigma^2$. We use the latter expression because the variance stabilizing transformation is a function of $E(\theta(Y) \mid X) = \phi(X)$.

Remark D. Note that in the RACE algorithm we begin by computing $\phi(X)$ instead of $\theta(Y)$: this is important in case X is a categorical variable, for computation of $v(u)$ requires $\phi(X)$ to be real-valued. Also, it might seem more natural to put the initial step $\phi(X) = E(Y \mid X)$ into the iteration part of the algorithm and remove the corresponding step at the end of the loop. We decided on the above form of the algorithm because it made the check of the $r^2(\theta, \phi)$ condition easier, due to the rescaling that occurs.

Remark E. It is interesting to note that the popular Box-Cox method for estimating

based on the trapezoid rule, and e and $r^2(\theta, \phi_1, \phi_2, \ldots \phi_p)$ by their sample analogues.

Remark A. The ACE algorithm does not assume that $\theta(Y)$ is monotone (although mononiticity can be enforced in the data algorithm). This makes sense in the ACE setting because the optimal transformation for Y given the joint distribution of X and Y may not be monotone. In the present setting, however, we assume that X and Y are generated from a model of the form (1) and this is meaningful only if $\theta(Y)$ is monotone. Appropriately, the estimate of $\theta(Y)$ given by RACE is monotone, since it is the integral of a positive function.

Remark B. A strength of the RACE algorithm is that the solutions it produces are (essentially) independent of the marginal distribution of X. In particular, if X is transformed by a monotone function $f(X)$ before applying RACE, then ignoring scaling, $\theta(Y)$ is unchanged and $\phi(X)$ is mapped in the obvious way, i.e. $\phi^{new}(f(X)) = \phi^{old}(X)$. This is not the case for the ACE algorithm. Note that in the data version of RACE, this invariance will hold only approximately because a scatterplot smoother typically uses an implicit metric in the X space.

Remark C. One of the assumptions in model (1) is $Var(\theta(Y) \mid X) = \sigma^2$, but in the RACE algorithm we have used the quantity $Var(\theta(Y) \mid \phi(X))$. It is clear, however, that $Var(\theta(Y) \mid X) = \sigma^2$ implies $Var(\theta(Y) \mid \phi(X)) = \sigma^2$. We use the latter expression because the variance stabilizing transformation is a function of $E(\theta(Y) \mid X) = \phi(X)$.

Remark D. Note that in the RACE algorithm we begin by computing $\phi(X)$ instead of $\theta(Y)$: this is important in case X is a categorical variable, for computation of $v(u)$ requires $\phi(X)$ to be real-valued. Also, it might seem more natural to put the initial step $\phi(X) = E(Y \mid X)$ into the iteration part of the algorithm and remove the corresponding step at the end of the loop. We decided on the above form of the algorithm because it made the check of the $r^2(\theta, \phi)$ condition easier, due to the rescaling that occurs.

Remark E. It is interesting to note that the popular Box-Cox method for estimating

transformations is a special case of ACE, not RACE. This is clear because the Box-Cox procedure minimizes the residual sum of squares, restricting the $\phi_j(X_j)$'s to be linear and $\theta(Y)$ to be of the form $(Y^\lambda - 1)/\lambda$. The Box-Cox method does not produce many of the anomalies of ACE because the transformation on Y is restricted. Of course, it is much less general than either ACE or RACE.

3. Examples.

Example 1. *Disjoint clusters.* The following situation was discussed by Buja and Kass (1985) and attributed to Charles Stone. Suppose the distribution of (X, Y) falls into 2 disjoint clusters in diagonally opposite quadrants, that is for some a and b, $P(X \leq a, Y \leq b)$ and $P(X > a, Y > b)$ are both non-zero and sum to 1. Then ACE produces a function $\theta(Y)$ that maps the sets $\{Y \leq b\}$ and $\{Y > b\}$ onto different constants and a function $\phi(X)$ that maps the sets $\{X \leq a\}$ and $\{X > a\}$ onto different constants. The resulting correlation between the transformed variables is 1. Buja and Kass say that this is perplexing and note that it occurs regardless of the distributions in each cluster.

We can see that RACE will not collapse the clusters, at least in the following special case. Assume that $E(Y \mid X) = c_1(X)$ for $X \leq a$, $E(Y \mid X) = c_2(X)$ for $X > a$, and $Var(Y \mid X)$ is constant. Then RACE will produce $\theta(Y) = Y$, and $\phi(X) = c_1(X)$ for $X \leq a$ and $\phi(X) = c_2(X)$ for $X > a$. (The actual functions will be centered and scaled versions of these). This is clear because the above functions are the initial ones used by RACE and the RACE iteration leaves them unchanged. ¿From a regression point of view, the RACE solution makes more sense: it is simply estimating $E(Y \mid X)$ in each cluster. Note that both Pregibon and Vardi (1985) and BF point out that with real data ACE does not collapse the clusters but produces transformations with large bends to try to account for them, because of the implicit smoothness forced by the supersmoother.

Example 2. *Data generated with skewed X's.* As mentioned earlier, ACE does not generally reproduce model transformations. Consider for example the case $Y = X + \varepsilon$ with $\varepsilon \sim$

$N(0,1)$. Then if X is also normal, ACE will produce the identity transformations. However, if X is highly skewed, ACE will produce a transformation of Y that is curved. Figure 1 shows the function $\theta(Y)$ produced by both RACE (x's) and ACE(o's) for 100 data points from the above model, with $X = exp(exp(U))$, U being uniform on $[0,1]$. (In all the figures, the functions have been scaled so that $E\phi(X) = E\theta(Y) = 0$ and $E\theta^2(Y) = 1$.) The transformations for X were both very nearly linear. The histogram of the X values is shown in Figure 2. Notice that ACE bends the transformation of Y where there is a concentration of X values.

Example 3. *Brain and body weight data.* Pregibon and Vardi (1985) look at ACE applied to these data from Weisberg (1980, pp128-129). Figure 3 shows brain weight (Y) plotted versus body weight (X) for 62 species of animals. As Pregibon and Vardi note, the sparsity of points in the northeast corner suggests taking logarithms, but ACE barely transforms the data at all. Figure 4 shows a plot of $\theta(Y)$ versus $\phi(X)$ from ACE. A plot of $\theta(Y)$ versus $\phi(X)$ from RACE is shown in Figure 5 and the transformations (Figures 6 and 7) are close to logarithmic (solid curves). BF in their rejoinder claim that outliers in the Y space are creating the problem for ACE, but the results here suggest that skewness in the marginal distribution of X is causing the problem.

Example 4. *Unequal variances.* The following example was suggested by David Andrews. We let $Y = X + \varepsilon$ where $\varepsilon \sim N(0, X^2)$ and X is uniform on $[0,3]$. Figure 8 shows 200 observations generated from this model and the results of RACE (x's) and ACE(o's) are shown in Figures 9 and 10. The ACE transformation are nearly linear, while RACE has found the square root transformations that stabilize the variance. A plot of $\theta(Y)$ versus $\phi(X)$ from RACE (Figure 11) indicates that RACE has achieved variance stabilization while maintaining a fairly linear relationship between the transformed values.

Example 5. *Missing group variable.* Pregibon and Vardi (1985) consider the following problem. Data consisting of 200 observations, 100 from $Y = 3 + 5X$ and 100 from $Y = -3 + 5X + \varepsilon$, are generated with X uniform on $[-5,5]$ and $\varepsilon \sim N(0,1)$. The available data consist only of X and Y and not the grouping variable. Pregibon and Vardi show that ACE

$N(0,1)$. Then if X is also normal, ACE will produce the identity transformations. However, if X is highly skewed, ACE will produce a transformation of Y that is curved. Figure 1 shows the function $\theta(Y)$ produced by both RACE (x's) and ACE(o's) for 100 data points from the above model, with $X = exp(exp(U))$, U being uniform on $[0,1]$. (In all the figures, the functions have been scaled so that $E\phi(X) = E\theta(Y) = 0$ and $E\theta^2(Y) = 1$.) The transformations for X were both very nearly linear. The histogram of the X values is shown in Figure 2. Notice that ACE bends the transformation of Y where there is a concentration of X values.

Example 3. *Brain and body weight data.* Pregibon and Vardi (1985) look at ACE applied to these data from Weisberg (1980, pp128-129). Figure 3 shows brain weight (Y) plotted versus body weight (X) for 62 species of animals. As Pregibon and Vardi note, the sparsity of points in the northeast corner suggests taking logarithms, but ACE barely transforms the data at all. Figure 4 shows a plot of $\theta(Y)$ versus $\phi(X)$ from ACE. A plot of $\theta(Y)$ versus $\phi(X)$ from RACE is shown in Figure 5 and the transformations (Figures 6 and 7) are close to logarithmic (solid curves). BF in their rejoinder claim that outliers in the Y space are creating the problem for ACE, but the results here suggest that skewness in the marginal distribution of X is causing the problem.

Example 4. *Unequal variances.* The following example was suggested by David Andrews. We let $Y = X + \varepsilon$ where $\varepsilon \sim N(0, X^2)$ and X is uniform on $[0,3]$. Figure 8 shows 200 observations generated from this model and the results of RACE (x's) and ACE(o's) are shown in Figures 9 and 10. The ACE transformation are nearly linear, while RACE has found the square root transformations that stabilize the variance. A plot of $\theta(Y)$ versus $\phi(X)$ from RACE (Figure 11) indicates that RACE has achieved variance stabilization while maintaining a fairly linear relationship between the transformed values.

Example 5. *Missing group variable.* Pregibon and Vardi (1985) consider the following problem. Data consisting of 200 observations, 100 from $Y = 3 + 5X$ and 100 from $Y = -3 + 5X + \varepsilon$, are generated with X uniform on $[-5,5]$ and $\varepsilon \sim N(0,1)$. The available data consist only of X and Y and not the grouping variable. Pregibon and Vardi show that ACE

produces a linear transformation for X but a cubic-looking transformation for Y. It is clear that in the case of known distributions, RACE will produce linear transformations for X and Y because after computing $\phi(X) = E(Y \mid X)$, the variance of $Y - \phi(X)$ is constant. We also ran RACE on a sample of size 200 and the transformations were very close to linear. It is not clear if ACE or RACE is preferable in this situation: ACE tries to account for the missing variable by bending the transformation of Y while RACE simply averages over the two groups.

Example 6. *Data sets from Breiman and Friedman.* We applied RACE to some of the examples given in BF, for comparison with ACE. We briefly summarize the results here. In the simulated example $Y = exp(X^3) + \varepsilon$ with X^3 and $\varepsilon \sim N(0,1)$ the transformations produced were nearly identical, and close to the correct ones. This is not surprising in light of the discussions of Section 2 and Example 2. The transformations were also very close for the case $Y = exp(sin(2\pi X) + \varepsilon/2)$ where X was uniform on $[0,1]$ and $\varepsilon \sim N(0,1)$. Finally, we tried RACE on the ozone concentration data analyzed by BF. Again, the results were very similar, both methods producing a very mild transformation for the response.

4. Discussion.

In this paper we have proposed a variant of the ACE algorithm that uses variance stabilization to estimate the transformation of the response. In a number of regression situations, RACE seems to alleviate some of the anomalies of ACE. However, much more work is needed to establish the relative merits of the two methods. On the theoretical side, BF were able to prove, under regularity conditions, that given a joint distribution for X and Y, transformations maximizing $corr(\theta(Y), \phi(X))$ exist and ACE converges to them (in the L^2 sense). In the RACE setup, if we begin by assuming model (1), then existence of the solutions is not an issue. One would like to show that RACE converges to the optimal transforms $\theta^0(Y)$ and $\phi^0(X)$. We have had difficulty establishing this, however, because the transformation used to update $\theta(Y)$ is non-linear, unlike in ACE. More generally, we can ask the following interesting question (suggested by Larry Wasserman): given random variables X and Y, do transformations exist so that the relationship (1) holds? We conjecture that under suitable regularity

conditions these transformations do exist and RACE converges to them. In the multiple predictor case, it is clear that such transformations do not necessarily exist. For example, if $Y = exp(X_1) + X_1X_2 + \epsilon$ then no transformation $\theta(Y)$ will make $E(\theta(Y) \mid X_1, X_2)$ an additive function of X_1 and X_2.

Acknowledgments

We would like to thank David Andrews and Larry Wasserman for helpful discussion and the Natural Sciences and Engineering Research Council of Canada for its support.

References

Breiman, L. and Friedman, J.H. (1985). *Estimating optimal transformations for multiple regression and correlation.* J. Amer. Statist. Assoc. Vol. 80, No. 391, pg. 580-597.

Buja, A. and Kass, R.E. (1985). Discussion of *Estimating optimal transformations for multiple regression and correlation.* J. Amer. Statist. Assoc. Vol. 80, No. 391, pg. 580-597.

Cleveland, W.S. (1979). *Robust locally weighted regression and smoothing scatterplots.* J. Amer. Statist. Assoc. 74, 829-836.

Fowlkes, E. and Kettenring, J.R. (1985). Discussion of *Estimating optimal transformations for multiple regression and correlation.* J. Amer. Statist. Assoc. Vol. 80, No. 391, pg. 580-597.

Friedman, J.H. and Stuetzle, W. (1982). *Smoothing of scatterplots,* Dept. of Statistics Tech. Rept. Orion 3, Stanford University.

Pregibon, D. and Vardi, Y. (1985). Discussion of *Estimating optimal transformations for multiple regression and correlation.* J. Amer. Statist. Assoc. Vol. 80, No. 391, pg. 580-597.

Silverman, B.W. (1985). *Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion).* J. Roy. Statist. Soc. B, 36, 111-147.

Wahba, G. and Wold, S. (1975). *A completely automatic French curve: fitting spline functions by cross-validation.* Comm. Stat 4,1-7.

Weisberg, S. (1980). *Applied linear regression.* New York, Wiley.

conditions these transformations do exist and RACE converges to them. In the multiple predictor case, it is clear that such transformations do not necessarily exist. For example, if $Y = exp(X_1) + X_1X_2 + \epsilon$ then no transformation $\theta(Y)$ will make $E(\theta(Y) \mid X_1, X_2)$ an additive function of X_1 and X_2.

Acknowledgments

We would like to thank David Andrews and Larry Wasserman for helpful discussion and the Natural Sciences and Engineering Research Council of Canada for its support.

References

Breiman, L. and Friedman, J.H. (1985). *Estimating optimal transformations for multiple regression and correlation.* J. Amer. Statist. Assoc. Vol. 80, No. 391, pg. 580-597.

Buja, A. and Kass, R.E. (1985). Discussion of *Estimating optimal transformations for multiple regression and correlation.* J. Amer. Statist. Assoc. Vol. 80, No. 391, pg. 580-597.

Cleveland, W.S. (1979). *Robust locally weighted regression and smoothing scatterplots.* J. Amer. Statist. Assoc. 74, 829-836.

Fowlkes, E. and Kettenring, J.R. (1985). Discussion of *Estimating optimal transformations for multiple regression and correlation.* J. Amer. Statist. Assoc. Vol. 80, No. 391, pg. 580-597.

Friedman, J.H. and Stuetzle, W. (1982). *Smoothing of scatterplots,* Dept. of Statistics Tech. Rept. Orion 3, Stanford University.

Pregibon, D. and Vardi, Y. (1985). Discussion of *Estimating optimal transformations for multiple regression and correlation.* J. Amer. Statist. Assoc. Vol. 80, No. 391, pg. 580-597.

Silverman, B.W. (1985). *Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion).* J. Roy. Statist. Soc. B, 36, 111-147.

Wahba, G. and Wold, S. (1975). *A completely automatic French curve: fitting spline functions by cross-validation.* Comm. Stat 4,1-7.

Weisberg, S. (1980). *Applied linear regression.* New York, Wiley.

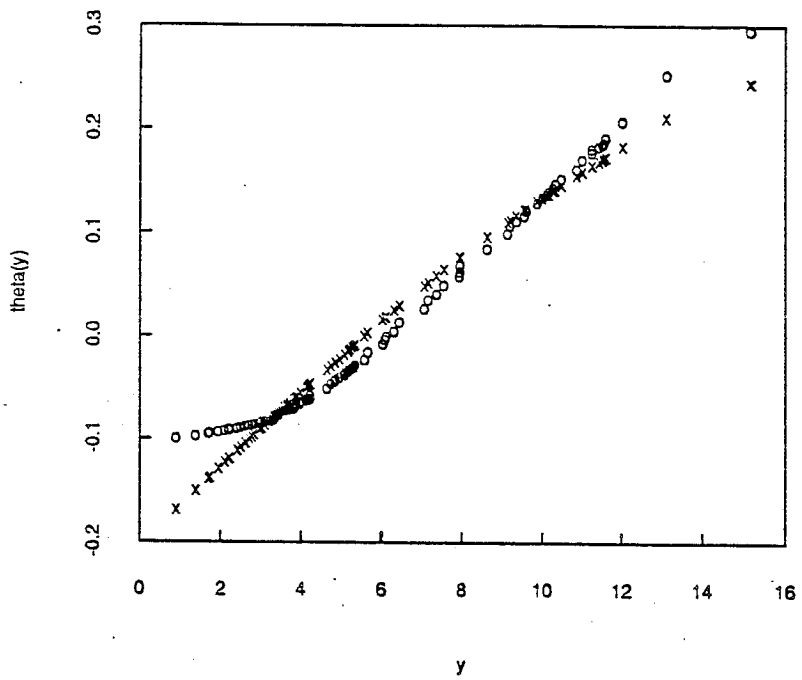

Figure 1. Response transformations for Example 2. The x's are from RACE, o's from ACE

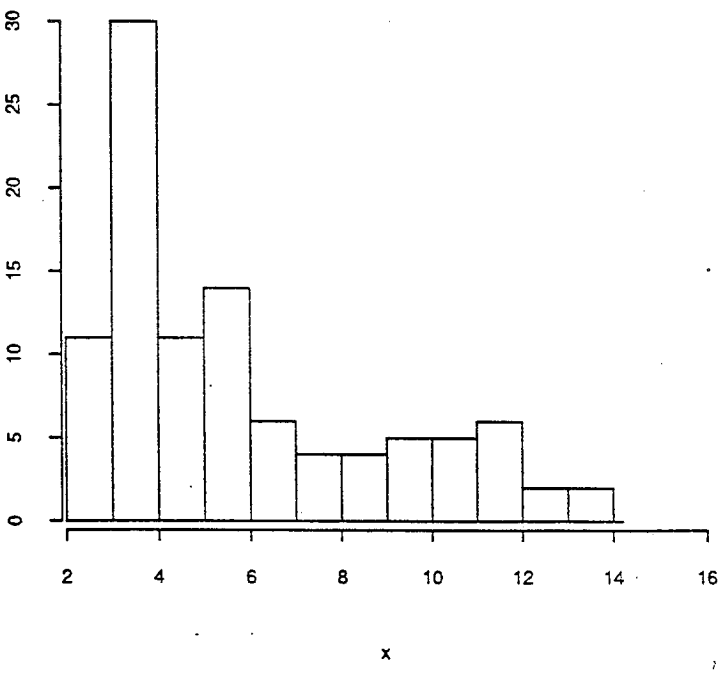

Figure 2. Histogram of X values for Example 2

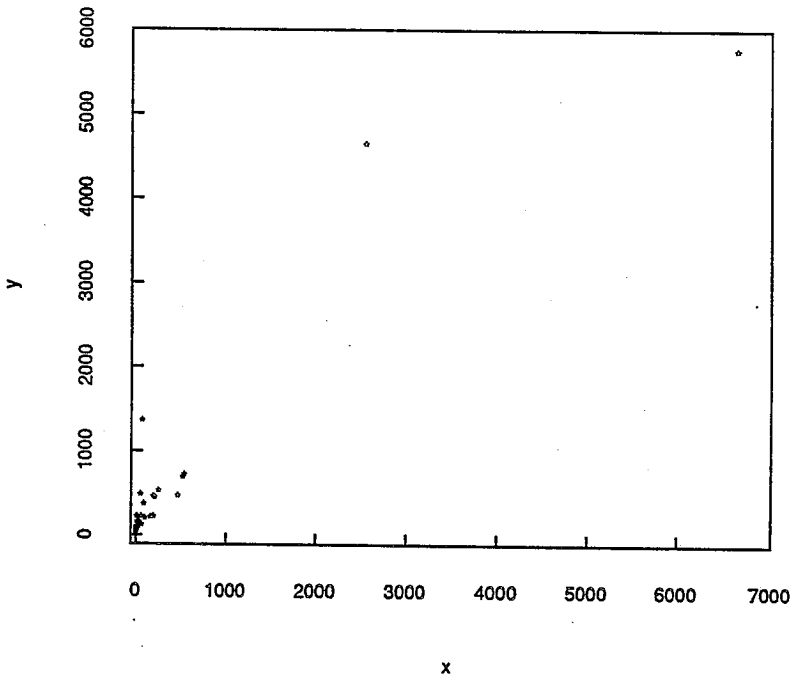

Figure 3. Brain weight (Y) versus body weight (X) for Example 3.

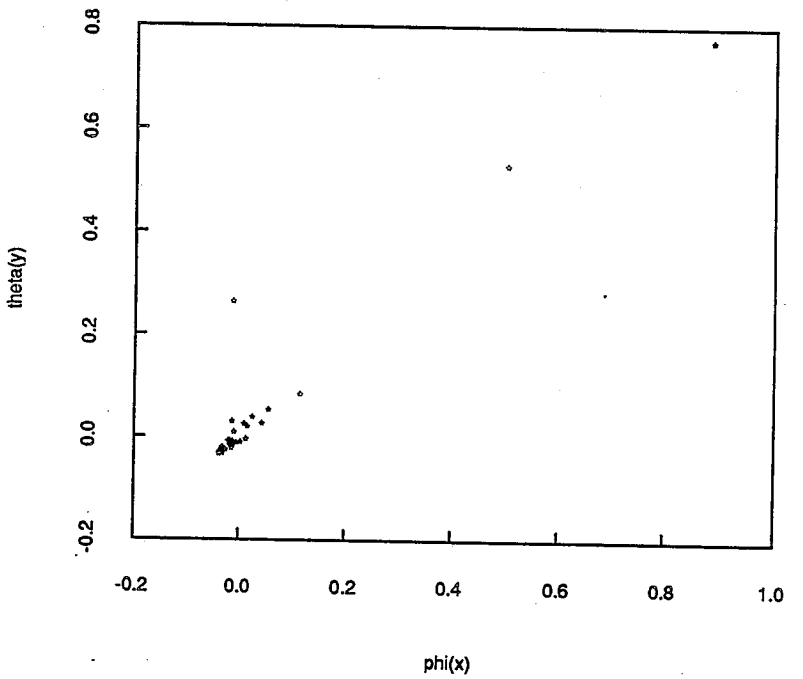

Figure 4. $\theta(Y)$ versus $\phi(X)$ from ACE, Example 3.

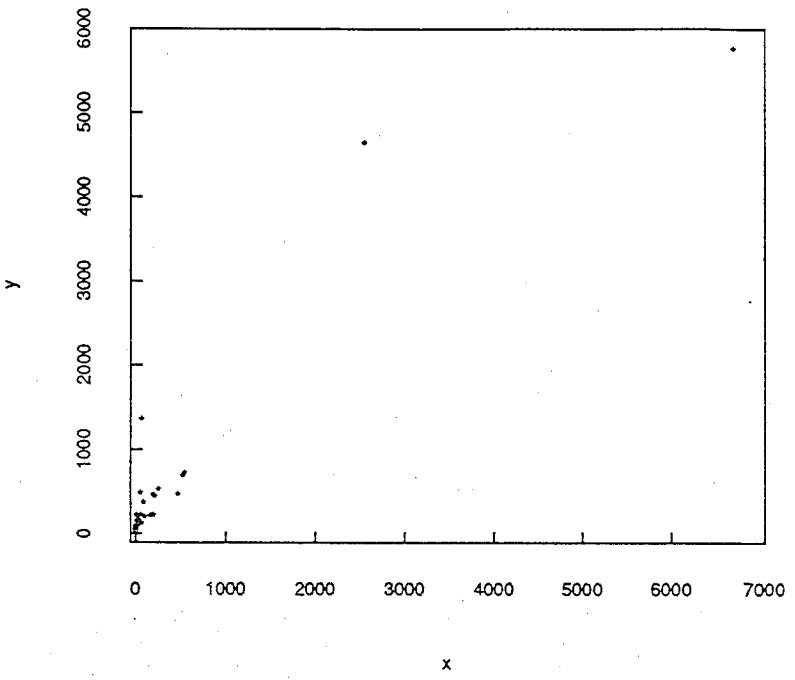

Figure 3. Brain weight (Y) versus body weight (X) for Example 3.

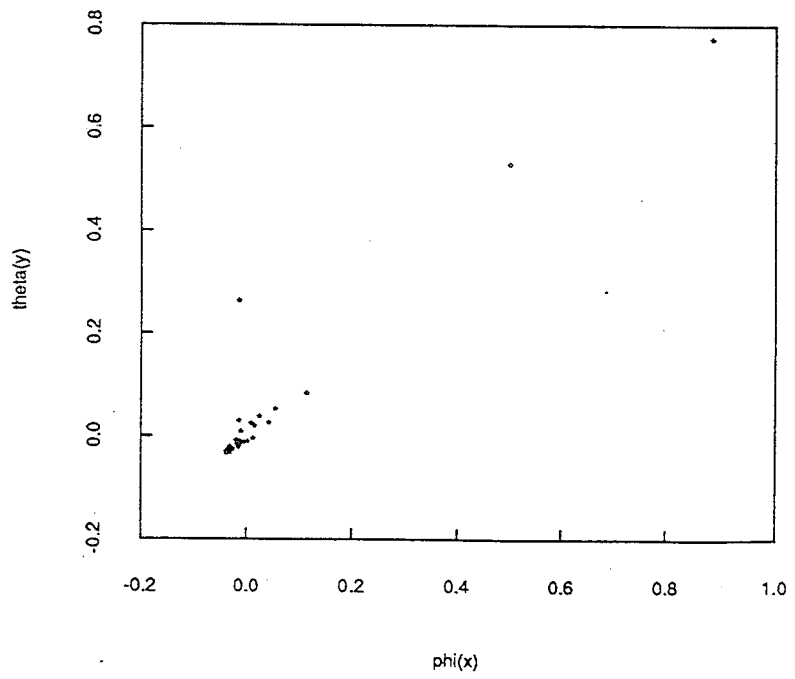

Figure 4. $\theta(Y)$ versus $\phi(X)$ from ACE, Example 3.

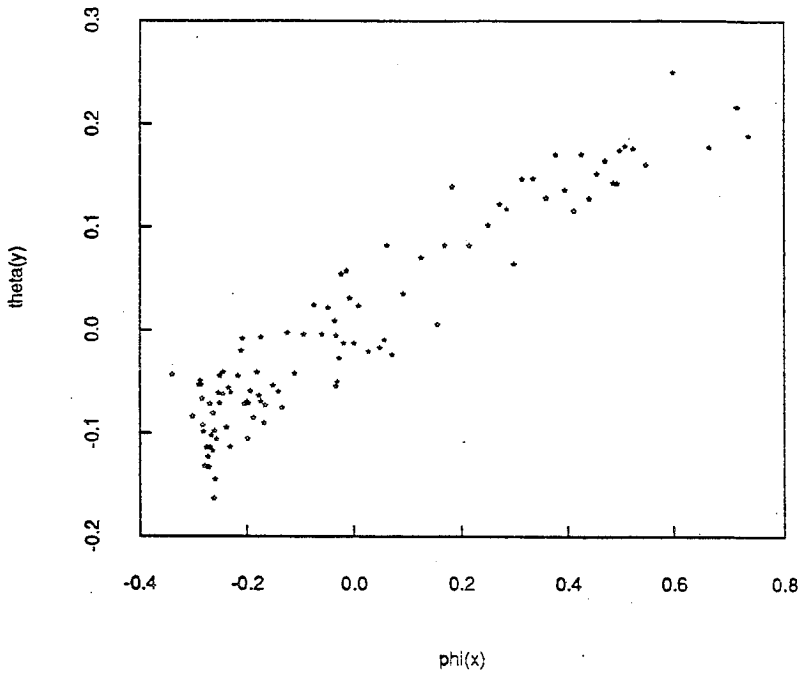

Figure 5. $\theta(Y)$ versus $\phi(X)$ from RACE, Example 3.

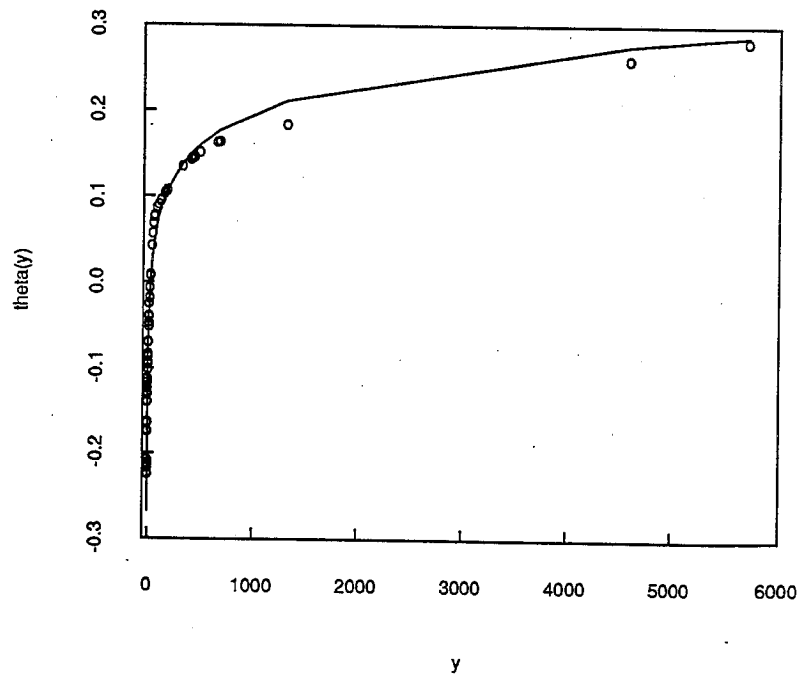

Figure 6. $\theta(Y)$ from RACE and log function (scaled) (solid curve)

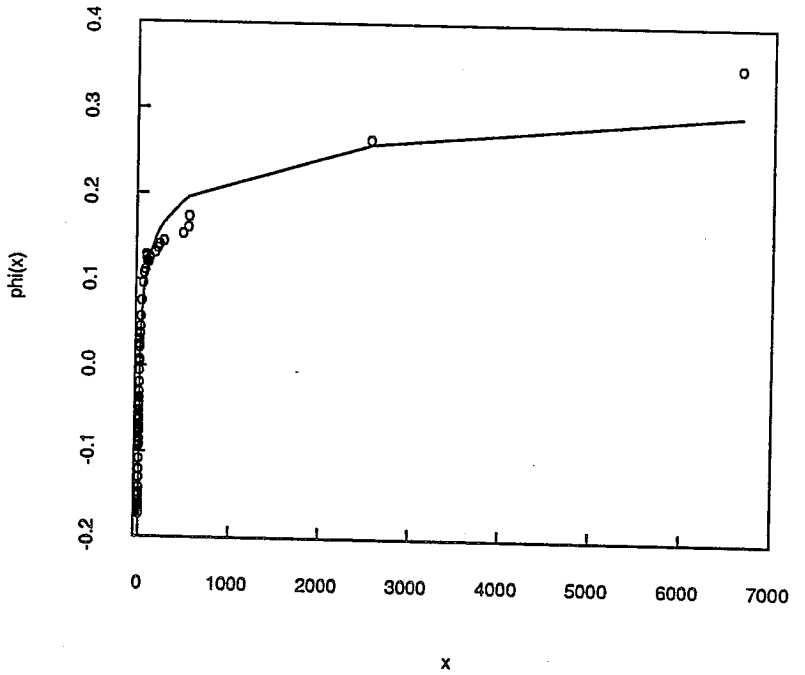

Figure 7. $\phi(X)$ from RACE and log function (scaled) (solid curve)

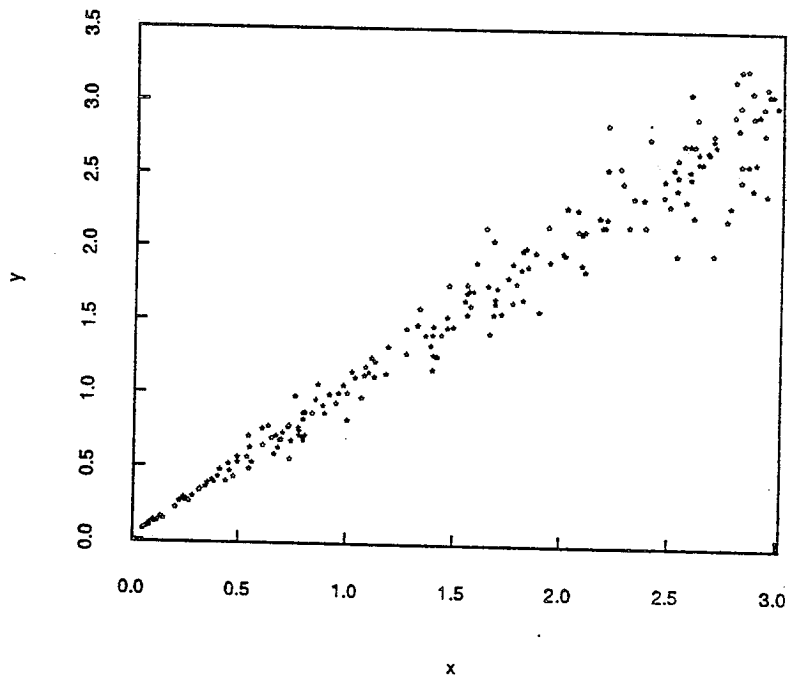

Figure 8. Data for Example 4

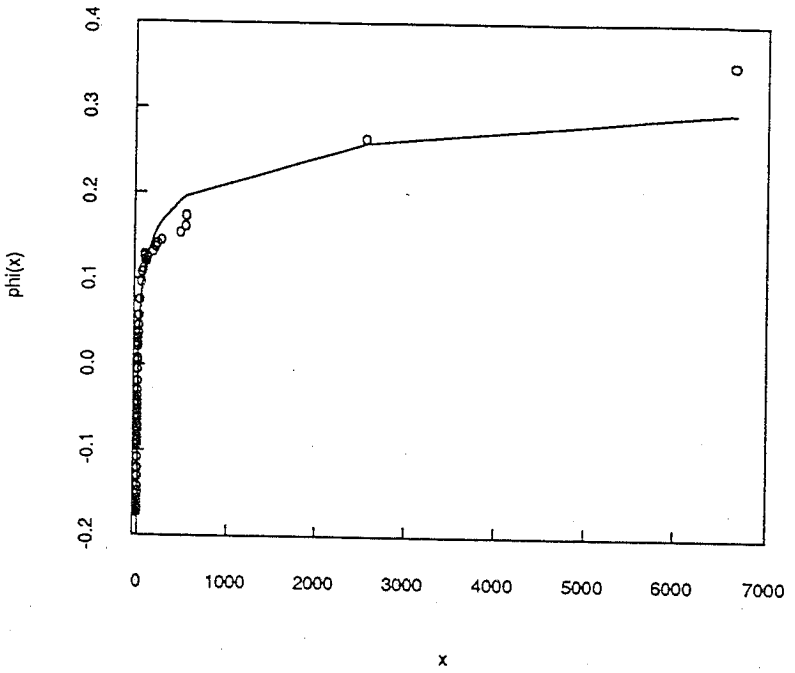

Figure 7. $\phi(X)$ from RACE and log function (scaled) (solid curve)

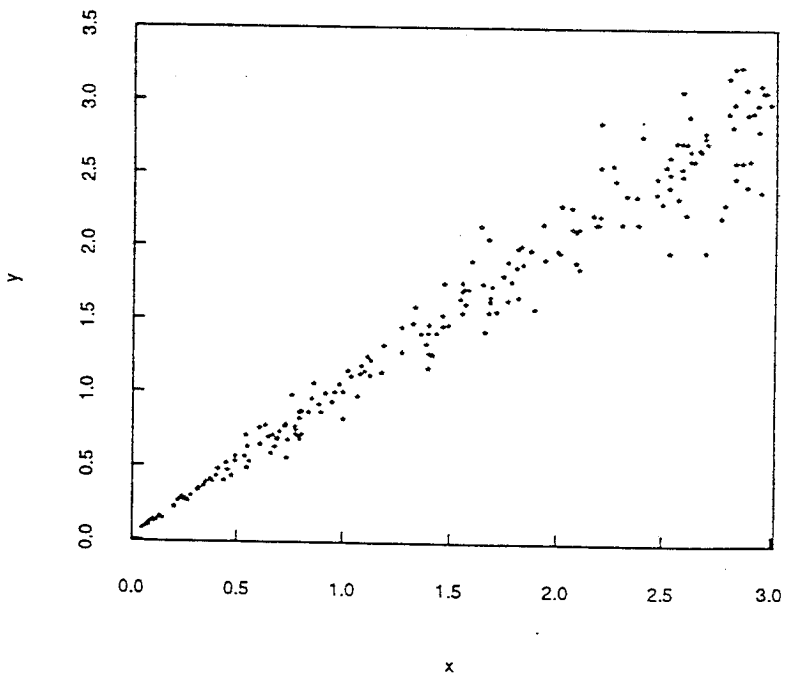

Figure 8. Data for Example 4

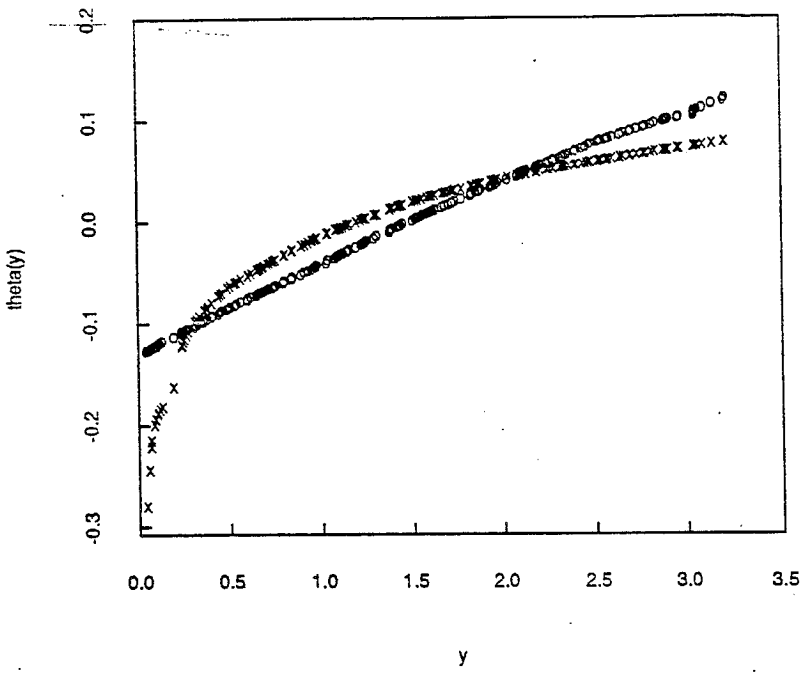

Figure 9. $\theta(Y)$ from RACE (x's) and ACE (o's), Example 4

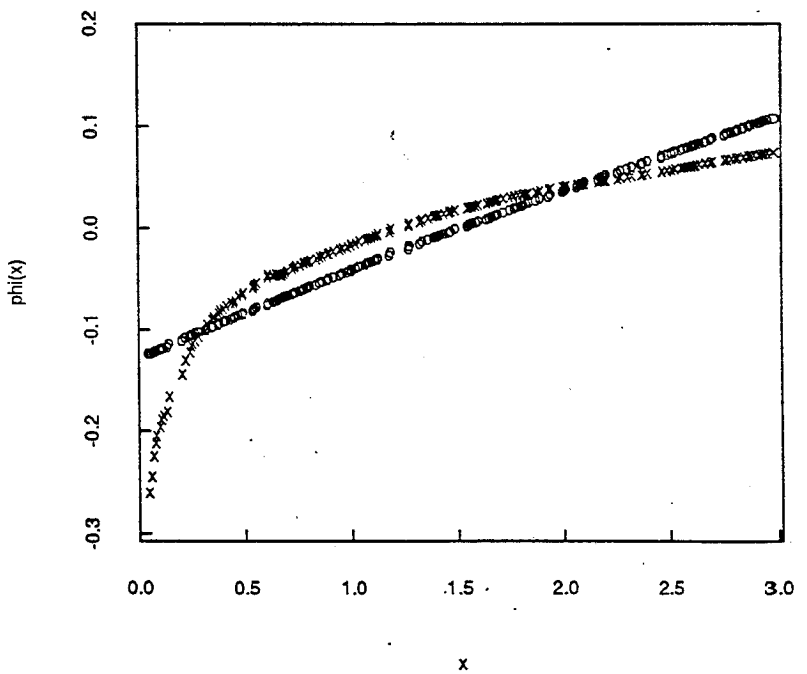

Figure 10. $\phi(X)$ from RACE (x's) and ACE (o's), Example 4