Research Support
University of Michigan Business School

February 1997

Robust Estimation of Semiparametric Regression Models

Working Paper #9712-03

by

Martin R. Young
University of Michigan

and

Xuming He
University of Illinois

Robust Estimation of Semiparametric Regression Models

Martin R. Young[a*] Xuming He[b]

[a] University of Michigan Business School, Ann Arbor, MI 48109

[b] University of Illinois, Champaign, IL 61820.

February, 1997

Abstract

The goodness-of-fit of multiple linear regression models can often be improved by preliminary transformation of the dependent variable. In this paper, we consider the model $h(y) = \beta'\mathbf{x} + e$, where $h(\cdot)$ is a monotone function, and present a method for joint estimation of the transformation function $h(\cdot)$ and the regression coefficients β, via minimization of a sum of absolute deviations loss function. The resulting estimator is robust with respect to outliers, and has a simple and direct numerical solution, using linear programming. The paper presents consistency results, and applications of the technique to simulated and real economic data.

KEY WORDS: *index model, least absolute deviations, splines, transformation*

JEL CLASSIFICATION: *C14*

[*] Corresponding author. Phone: 313-936-1332, Fax: 313-936-0274, email: myoung@umich.edu.

The research of Xuming He was supported by NSA Grant MDA904-96-1-0011.

1 Introduction

The goal of linear regression modeling is the identification of the statistical equation relating predictor variables x to some response variable y. The standard linear regression model assumes that the predictor variables are linearly related to the dependent variate: the validity of this assumption is often improved by preliminarily transforming the dependent variable nonlinearly, and then applying the linear model to the transformed data. Box and Cox (1964) was an early and influential contribution to the understanding of transformations in regression; Carroll and Ruppert (1988) provides an overview of classical and modern methods for choosing transformations for regression models.

In this paper, we present a new method for automatically identifying a transformation which optimizes the goodness-of-fit between the transformed dependent variate and a linear combination of the predictor variables. In this approach, the transformation function is chosen nonparametrically, without restricting the function to belong to a class such as the power function class. Because the predictor is determined as a parametric, specifically linear, function of the independent variables, the overall procedure is termed semiparametric (Horowitz 1996). The method in this paper uses the least absolute deviations (LAD) goodness-of-fit criterion, and as a result is resistant to outliers in the dependent variable. Finally, the method is very computationally efficient: the transformation function and the regression coefficients are estimated jointly as the solution to a single linear programming problem. Throughout the paper, the proposed procedure will be referred to as the "ROSE" technique, short for "Robust Semiparametric Estimation".

Section 2 of this article describes the proposed method for robust semiparametric regression estimation, section 3 discusses inference issues, section 4 describes the application of the ROSE technique on some real and simulated data, section 5 presents results on the asymptotic properties of the estimator, and section 6 gives further discussion, including references to related work.

2 The Algorithm

The data for the problem under consideration are pairs $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ where the y_i are observations on a dependent variable, and $\mathbf{x}_i = (x_{i0}, \ldots, x_{ip})$ are predictor variables, with, typically, $x_{i0} = 1$ for all i. The usual model for relating dependent and predictor variables is the linear regression model:

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + e_i, \tag{1}$$

where the e_i are zero mean residuals. In this paper, we consider a generalization of model (1), namely:

$$h(y_i) = \boldsymbol{\beta}'\mathbf{x}_i + e_i, \tag{2}$$

where $h(\cdot)$ is a monotone, nonlinear transformation function. The objective of the analysis is to identify the function $h(\cdot)$ and the coefficient vector $\boldsymbol{\beta}$ to optimize the goodness-of-fit between $h(y_i)$ and $\boldsymbol{\beta}'\mathbf{x}_i$. In order to achieve robustness with respect to outliers in the dependent variable, we use the LAD criterion to assess goodness-of-fit; see Dielman and Pfaffenberger (1982) and Gonin and Money (1989) for overviews of LAD methods in robust regression.

The semiparametric LAD estimation problem can be formulated as:

$$\min \sum_{i=1}^{n} |h(y_i) - \boldsymbol{\beta}'\mathbf{x}_i| \tag{3}$$

$$\text{subject to}$$

$$h(y_1) = y_1 \tag{4}$$

$$h(y_n) = y_n, \tag{5}$$

where y_1 and y_n are, respectively, the smallest and largest values of y in the dataset. The optimization in (3) is over the product space $\mathbf{H} \times \mathbf{R}^{p+1}$, where $\mathbf{H}$ is the space of monotone functions $h(\cdot)$ satisfying boundary conditions (4) and (5). The boundary conditions are needed in order to prevent the trivial solution $h(y) = 0$, $\beta_0 = \beta_1 = \cdots = \beta_p = 0$, which would have a sum of absolute deviations equal to zero. Because the boundary conditions constrain both the scale and location of the transformation, it will typically be necessary to include the intercept term $x_{i0} = 1$, in order to achieve a good fit between the $h(y_i)$ and $\beta' \mathrm{x}_i$. The values of $h(y_1)$ and $h(y_n)$ can clearly be constrained to have arbitrary, distinct values; the choice represented in equations (4)-(5) leads to a location and scale for the transformed variable that match those of the original dependent variable, which correspondence may be helpful in interpreting the estimated model.

To implement the estimation, the function $h(\cdot)$ can be represented as a linear combination of spline basis functions:

$$h(y) = \sum_{k=0}^{q} \alpha_k \phi_k(y), \tag{6}$$

where the $\phi_k(y)$ are the known basis functions. In the examples described in this paper, the linear "power spline" basis (Smith 1979) is used. For a linear power spline with knots at locations $k_1, \ldots, k_r$, the series expansion has $r + 2$ basis functions, and is given by:

$$h(y) = \alpha_0 + \alpha_1 y + \sum_{j=2}^{q} \alpha_j \left(y - k_{j-1} \right)_+ . \tag{7}$$

where $(z)_+ = \max\{z, 0\}$, and $q = r + 1$. In this parameterization, $h(y)$ is monotone as long as the following linear restrictions on $\boldsymbol{\alpha}$ are maintained: $\alpha_1 \geq 0$, $\alpha_1 + \alpha_2 \geq 0$, $\ldots$, $\alpha_1 + \cdots + \alpha_q \geq 0$. Ramsay (1988) presents an alternative parameterization for splines, called I-splines, which have the convenient property that monotonicity of the spline function is ensured simply by enforcing positivity on all coefficients α_k in the basis expansion (6); the I-spline basis could also be used in the procedure described below.

Given the linear power spline representation for $h(y)$, the estimation problem can be written as:

$$\min \sum_{i=1}^{n} (\epsilon_i^+ + \epsilon_i^-) \tag{8a}$$

subject to

$$\sum_{k=0}^{q} \alpha_k \phi_k(y_i) - \sum_{k=0}^{p} \beta_k x_{ik} = \epsilon_i^+ - \epsilon_i^- \tag{8b}$$

$$\sum_{k=0}^{q} \alpha_k \phi_k(y_1) = y_1 \tag{8c}$$

$$\sum_{k=0}^{q} \alpha_k \phi_k(y_n) = y_n \tag{8d}$$

$$\alpha_1 \geq 0 \tag{8e}$$

$$\alpha_1 + \alpha_2 \geq 0 \tag{8f}$$

$$\vdots$$

$$\alpha_1 + \cdots + \alpha_q \geq 0 \tag{8g}$$

$$\epsilon_i^+ \geq 0, \quad \epsilon_i^- \geq 0, \quad i = 1, \ldots, n. \tag{8h}$$

The quantities ϵ_i^+ and ϵ_i^- are the positive and negative parts of the residual $h(y_i) - \beta' \mathrm{x}_i$. The optimization in (8) is over the variables $\{\epsilon_i^+, \epsilon_i^-, \quad i = 1, \ldots, n\}$, $\{\alpha_k, \quad k = 1, \ldots, q\}$, and $\{\beta_k, \quad k = 1, \ldots, p\}$. The objective function and constraints are linear in the decision variables: thus the problem can be solved as a single linear programming problem, using standard methods such as the simplex, dual simplex, or interior point algorithms. The robust semiparametric estimator (ROSE) is defined, then, as the minimizing solution to linear program (8).

Because the model (2) is estimated using linear programming, it is simple to incorporate additional linear constraints on the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, should such constraints be appropriate. For example, in certain applications, the signs of the linear predictor coefficients β_k will be known a-priori; adding constraints $\beta_k \geq 0$ or $\beta_k \leq 0$ will not change the linear programming character of the solution. Incorporating such prior information may be helpful in regularizing parameter estimates in cases of multicollinearity, or small sample size: see, e.g., the example in section 4.3 below. Also, the transformation $h(y)$ can be restricted to be concave or convex through linear restrictions on the α parameters. For example, with the power spline parameterization for $h(y)$ in equation (7), $h(y)$ will be guaranteed to be concave if $\alpha_k \leq 0$ for $k = 2, \ldots, q$.

Solving the linear program (8) is typically rapid; for example, for the data described in section 4.2, with 506 observations, 14 predictor variables, and with 6 basis function coefficients used to estimate the transformation function, the optimal parameter estimates are obtained in less than 1 second on a 66 Mhz PC with an Intel 486DX CPU, using the algorithm of Barrodale and Roberts (1978).

Because the transformation function $h(\cdot)$ is constrained to be monotone, oversmoothing is not as significant of a concern as it is in a typical nonparametric regression problem. However, it is useful to have a heuristic for choosing a model with as few knots as are necessary to achieve a good fit to the data. An approach which seems to provide reasonable model choice is to consider a small number of different values for q, the number of parameters in the spline transformation (e.g., q=2, 4, 6, and 8) and to select the model which minimizes the AIC-like measure AIC $= 2\ell - 2(p+q)$, where

$$\ell = n \log n - n \log \left(\sum_{i=1}^{n} |h(y_i) - \beta' \mathbf{x}_i| \right) - n + \sum_{i=1}^{n} h'(y_i) \tag{9}$$

is a surrogate for a concentrated log-likelihood function, assuming Laplace errors, and $p+q$ is the total number of free parameters ($p+1$ for the regression coefficients, $q+1$ for the spline transformation, minus 2 for the normalizing constraints). The quantity $h'(y_i)$ is the derivative of $h(y)$ at y_i, which enters the likelihood function via the Jacobian of the transformation, and which can be evaluated easily as $h'(y_i) = \sum_{k=0}^{q} \alpha_k \phi'_k(y_i)$; for the power spline basis, with $\phi_0(y) = 1$, $\phi_1(y) = y$, and $\phi_j(y) = (y - k_{j-1})_+$, $j = 2, \ldots, q$, the derivatives are $\phi'_0(y) = 0$, $\phi'_1(y) = 1$, and $\phi'_j(y) = I(y \geq k_{j-1})$, $j = 2, \ldots, q$. Hastie and Tibshirani (1990, chapter 7) describes a related approach to model selection in the context of transformation estimation.

LAD methods have previously been employed in nonparametric regression settings in Koenker, Ng, and Portnoy (1994), Young (1996), and He and Shi (1996). The model described in this paper differs from earlier work in that the nonlinear transformation being estimated is applied to the dependent variable, rather than the independent variates.

3 Inference on Regression Coefficients

The issue of how inference on regression coefficients should proceed in the face of uncertainty about the appropriate transformation for the dependent variable is not fully resolved. Bickel and Doksum (1981) showed that the sampling variability of the estimator $\hat{\beta}$ in a Box-Cox regression model may be considerably greater when the transformation parameter is estimated than when it is fixed. Box and Cox (1982) and Hinkley and Runger (1984) have argued, though, that the parameter β is not "physically meaningful" unless the scaling for the dependent variable is considered fixed, and therefore argue for making inference on β conditional upon the estimated transformation. If this latter view is taken, then inference on the coefficients β in the robust semiparametric model can be made, conditional upon the estimated transformation $h(y)$, using the sampling theory for LAD estimators developed in Bassett and Koenker (1978). This is so, because, if $h(y)$ is conditioned upon, or fixed, the estimate of β in the robust semiparametric regression is an LAD regression estimate.

Bassett and Koenker (1978) demonstrated that an LAD regression coefficient estimator $\hat{\beta}$ is asymptotically unbiased and normally distributed with covariance matrix $\tau^2(\mathrm{X}'\mathrm{X})^{-1}$, where $\mathbf{X}$ is the usual regression design matrix, and τ^2/n is the variance of the median of a sample of n observations from the error distribution. Thus, for example, a $(1-\alpha)100\%$ confidence interval for a linear compound $\mathrm{r}'\beta$ can be written as $\mathrm{r}'\hat{\beta} \pm z_{\alpha/2}\tau \left[\mathrm{r}' (\mathrm{X}'\mathrm{X})^{-1} \mathrm{r}\right]^{1/2}$ (Dielman and Pfaffenberger 1982). A consistent estimate of the quantity τ can be obtained based on the residuals from the regression. Let the ordered residuals be denoted by $e_{(i)}$, then an estimate of τ is given by:

$$\hat{\tau} = \frac{e_{(t)} - e_{(s)}}{2(t-s)/n}, \tag{10}$$

where t and s are symmetric around the median sample residual, and $t-s$ is small relative to n; Dielman and Pfaffenberger (1982) suggest the estimate is not sensitive to the choice of $t-s$. Other estimates of τ based on kernel density estimation or neighboring regression quantiles may also be used.

4 Examples

4.1 Seasonal Regression Modeling

A common model for time series forecasting is the seasonal regression model (Cryer 1986):

$$y_t = \beta_0 + \beta_1 t + \beta_2 Q_2 + \beta_3 Q_3 + \beta_4 Q_4 + e_t \tag{11}$$

where

$$\begin{aligned} y_t &= \text{The time series value at time period } t \\ Q_2 &= 1 \text{ if period } t \text{ is in Quarter 2; 0 otherwise} \\ Q_3 &= 1 \text{ if period } t \text{ is in Quarter 3; 0 otherwise} \\ Q_4 &= 1 \text{ if period } t \text{ is in Quarter 4; 0 otherwise.} \end{aligned}$$

The term $\beta_1 t$ in equation (11) captures the linear trend in the time series, while the terms $\beta_2 Q_2$, $\beta_3 Q_3$, and $\beta_4 Q_4$ account for seasonal variation. Often it is necessary to nonlinearly transform the data y_t before the linear trend model is appropriate. Here, we estimate the optimal transformation automatically from the data. through fitting the model:

$$h(y_t) = \beta_0 + \beta_1 t + \beta_2 Q_2 + \beta_3 Q_3 + \beta_4 Q_4 + e_t. \tag{12}$$

Figure 1 displays aggregate sales figures for the Ford Motor Company. in nominal dollars; note the nonlinearity of the trend. Also, the additivity of model (11) is questionable for these data, as the seasonal variation appears to increase in magnitude over time. Figure 2 shows the transformation function $h(y)$ estimated by solving (8) using 4 knots spaced at the quintiles of the empirical distribution for y. Figure 3 shows the transformed time series; here the linear trend assumption appears more valid. Fortuitously, this transformation also appears to render the model more nearly additive: on the transformed scale, the seasonal variation is fairly constant over time. The model in (12) provides a forecast equation for $h(y_t)$ for some future time t; it is elementary to invert the function $h(y)$, to obtain a forecast for y_t.

[Insert Figures 1-3 Here]

4.2 Boston Housing Data

Harrison and Rubinfeld (1978) present a hedonic price index model for estimating homeowners' marginal willingness-to-pay for various characteristics of a neighborhood, including crime rate, and environmental quality. The unit of observation in this study was a census tract in Boston, and the entire set of variables used in the analysis were:

y = MEDV = median housing value (in \$1000)

x_1 = CRIM = per capita crime rate

x_2 = ZN = proportion of land zoned for lots greater than 25,000 square feet

x_3 = INDUS = proportion of nonretail business acres

x_4 = CHAS = bounds Charles river (1=yes, 0=no)

x_5 = NOX = nitrous oxide concentration (parts per 10 million)

x_6 = RM = average number of rooms in home squared

x_7 = AGE = proportion of owner-occupied units built prior to 1940

x_8 = DIS = weighted distance to five employment centers in Boston region

x_9 = RAD = index of accessibility to radial highways

x_{10} = TAX = full value property tax rate (per $10,000)

x_{11} = PTRATIO = pupil teacher ratio

x_{12} = B = 1000(Bk-0.63)^2, where Bk is the proportion of blacks in the population

x_{13} = LSTAT = proportion of the population that is lower status

Belsley, Kuh, and Welsch (1980) states that the data for this study "possess much heavier tails than the Gaussian (normal) distribution"; they estimate the data using iterative robust regression procedures, using log-transformed median housing value as the y variate in the regression. Here, we use median value as the y variate, and automatically select an appropriate scaling $h(y)$ via the ROSE procedure.

With just 1 knot for the spline expansion of $h(y)$, the estimated transformation is very close to a log transformation, with a Pearson correlation between $h(y)$ and $\log(y)$ of 0.993. Table 1 displays the normalized coefficient estimates obtained via the ROSE procedure, via robust (LAD) regression on log(y), and via OLS estimation on log(y). it is seen that the robust semiparametric method produces coefficient estimates similar to those obtained via robust regression applied to the log-transformed data. A possible advantage of the semiparametric method is that it does not require the user to recognize the need for a log transformation prior to analysis of the data.

[Insert Table 1 Here]

4.3 Automobile Fuel Efficiency

Lock (1993) provides a dataset of specifications for new car models from the 1993 year; measures are provided for evaluating price, fuel economy, engine size, body size. and features. The cars included in the dataset were selected at random from among 1993 passenger car models. Here, regression models are estimated for predicting city fuel economy, as a function of various covariates. The analysis includes all complete cases listed in the dataset – a total of 82 observations – and incorporates the following variables:

y = CITYMPG = city fuel efficiency (mpg)
x_1 = CYL = # of cylinders
x_2 = ENGSIZE = engine size (liters)
x_3 = HP = horsepower (100's)
x_4 = LENGTH = length (feet)
x_5 = WIDTH = width (feet)
x_6 = WEIGHT = weight (1000 pounds)
x_7 = DOMESTIC = 1 if U.S. manufacturer, 0 otherwise

Table 2 presents the coefficient estimates obtained via OLS; the positive coefficients on LENTGH and WIDTH obtained from OLS seem counterintuitive, and are due to the extreme multicollinearity in the data. A further deficiency of the OLS model is revealed in the residual plot, shown in Figure 4, in which there appears to be a nonlinear pattern not captured by the linear model.

The data were also analyzed using the semiparametric regression model estimated via the ROSE procedure. The estimation was regularized by adding the restrictions that the coefficients on the β_k must all be non-positive, except for the coefficient for DOMESTIC. The coefficient estimates are presented in Table 2, and the estimated transformation function is plotted in Figure 5. The zero coefficient estimates are, if not completely satisfactory, more reasonable than negative coefficients. Figure 6 displays the residuals from the transformed regression model; these seem more nearly random than do the residuals from the untransformed model.

This example provides an illustration of the three potentially beneficial features of the ROSE procedure: the robustness with respect to outliers, the ability to nonlinearly transform the dependent variable, and the ability to easily constrain the parameter estimates to a space that appears "reasonable" a-priori. The modeling approach presented here does not represent the conclusive analysis for these data; in particular, the analysis might benefit from judicious use of case deletion (Anscombe 1960; Andrews and Pregibon 1978), variable selection (Mallows 1973), and/or such techniques as ridge regression (Hoerl and Kennard 1970) or principal components regression (Mansfield, Webster, and Gunst 1977). The ROSE approach, though, can be useful in general as a simple method for generating plausible candidate models in the face of multicollinearity, nonlinearity, and residual outliers.

[Insert Table 2 Here]

[Insert Figures 4-6 Here]

4.4 Monte Carlo Analysis

A modest Monte Carlo study was performed to evaluate the effectiveness of the joint estimation procedure. Data were generated from the model

$$\tilde{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i, \quad i = 1, \ldots, n, \tag{13}$$

with $\beta_0 = 1$, $\beta_1 = 11$, $\beta_2 = 21$, $\beta_3 = 31$, with x_{ij} all uniformly distributed on $[0, 1]$, and with the e_i coming from the Laplace (double-exponential) density with mean 0 and standard deviation σ. The $\tilde{y}$ were then nonlinearly transformed into observable quantities y by

$$y_i = \frac{a}{1 + \exp\left(-\left(\tilde{y} - 40\right)/5\right)} + b, \tag{14}$$

with a and b chosen so that $y_i = \tilde{y}_i$ for the smallest and largest values of $\tilde{y}$ in the sample. Figure 7 shows a plot of y vs. β'x for a sample dataset, using the true values of β.

Inverting equation (14) shows that the optimal transformation function from y back to $\tilde{y}$ is not a member of the Box–Cox class of functions.

The regression coefficients β were estimated three ways: via the ROSE procedure presented in section 2, via LAD regression of y vs. $\mathbf{x}$, and via OLS regression of y vs. $\mathbf{x}$. The latter two methods would only be appropriate for estimating the direction of β: thus, in each case, the estimates of β were normalized as $\hat{\beta}_n = \hat{\beta}/\sqrt{\hat{\beta}'\hat{\beta}}$. The goodness of fit for each simulation replication was measured as $\left(\hat{\beta}_n - \beta_n\right)' \left(\hat{\beta}_n - \beta_n\right)^{\frac{1}{2}}$ and the overall performance of the estimators was assessed as the mean of this quantity over all simulation replications. The statistical significance of the difference in performance was assessed by a binomial sign test, for which the test statistic is the proportion of simulation replications in which the ROSE estimator is more accurate than a competing estimator (LAD or OLS). Table 3 shows the performance of the three methods for different noise levels σ and different samples sizes n. In each case examined, the ROSE method substantially outperformed the methods which do not employ a preliminary transformation, with the differences statistically significant. These results suggest the value of applying the appropriate nonlinear transformation to the dependent variable, even when all that is desired is an estimate of the direction of the regression coefficient vector.

[Insert Table 3 Here]

5 Asymptotic Analysis: Consistency

This section is concerned with asymptotic consistency of the proposed estimates in Section 2. We argue that when the model (2) holds, the direction of β is being estimated consistently under rather general conditions on the predictor variables. If both the predictor and the error variables are normally distributed, the estimated transformation also approximates the true link function h up to a multiplicative constant.

To facilitate our arguments, we consider an equivalent form of the model (2) with the intercept term absorbed into the link function:

$$h_0(Y) = X'\beta_0 + e \tag{15}$$

where $\beta_0 \in S_p$, the p-dimensional unit ball, and $h_0 \in H$, the space of increasing functions that are Lipschitz on any compact interval $[A, B] \subset D_Y$, where D_Y is the interior of the support of Y. Note that D_Y could take the form of $(0, \infty)$ or $(-\infty, \infty)$ or a finite interval like $(0, 1)$. Since the Lipschitz condition is imposed on a compact set contained by D_Y, the function h_0 is allowed to have arbitrarily large derivatives at the "boundaries" of D_Y. Furthermore, we assume that $\mathrm{E}\,|X| < \infty$ and that the median of e is zero.

Let (h^*, β^*) be the minimizer of $E|h(Y) - X'\beta|$ over $\beta \in S_p$ and $h \in H$.

Lemma 1: If the distribution of X is elliptical, then $\beta^* = \beta_0$. In addition, if $X'\beta_0$ and e are independently and normally distributed, we have $h^*(y) = c^* h_0(y)$ for some constant c^*.

In general, there is no assurance that h^* would recover the true link function h_0.

Proof: Since β^* is the solution to the problem of minimizing $\mathrm{E}\,|h^*(Y) - X'\beta|$ over $\beta \in S_p$, Theorem 2.1 of Li and Duan (1989) implies that if the distribution of X is elliptical, the solution (h^*, β^*) is Fisher consistent in β, that is, $\beta^* = \beta_0$.

To prove the second part of Lemma 1, we assume without loss of generality that $\mathrm{E}(X'\beta_0) = 0$, $\mathrm{Var}(X'\beta_0) = \sigma_x^2$, and $\mathrm{E}\,e^2 = \sigma_e^2$. The conditional distribution of ϵ given Y (or equivalently $h_0(Y)$) is normal with mean $\sigma_e^2 h_0(y)/(\sigma_x^2 + \sigma_e^2)$. Consider

$$|h(Y) - X'\beta^*| = |\epsilon - (h_0(Y) - h(Y))|. \tag{16}$$

Note that h^* minimizes the expected value of (16) over $h \in H$. By a well-known property of population median, the conditional expectation of (16) given $Y = y$ is minimized by $h(y)$ satisfying $h_0(Y) - h(Y) = \mathrm{median}\,[e|y] = \sigma_e^2 h_0(y)/(\sigma_x^2 + \sigma_e^2)$. Thus, $h^*(y) = \sigma_x^2 h_0(y)/(\sigma_x^2 + \sigma_e^2)$. The proof is then complete.

In fact, we have also seen that the solution (h^*, β^*) is unique, but h^* is not equal to h_0. A scale multiplier however does not affect the goodness-of-fit through a linear equation.

Next, we consider the sampling property of the estimator (h_n, β_n) that minimizes $\sum_i |h(y_i) - x_i'\beta|$ over $h \in H$ and $\beta \in S_p$. The observations are assumed to be independent from the model (15).

It is easy to see that for every sequence h_n in H and any given compact interval $[A, B]$ there exists a subsequence h_{n_k} that converges to a limit h_∞ in H in the sense that $\sup_{A \le y \le B} |h_{n_k}(y) - h_\infty(y)| \to 0$.

Lemma 2: The estimator (h_n, β_n) is consistent in the sense that $\beta_n \to \beta^*$ and $h_n(y) \to h^*(y)$ for any $y \in D_Y$.

Proof: If the estimator (h_n, β_n) is not consistent for (h^*, β^*) as stated in Lemma 2, we consider, for some given pairs $A < B$ (to be chosen later), a convergent subsequence, still denoted by (h_n, β_n) for notational simplicity, such that $\beta_n \to \beta_1$ and $\max_{A \le y \le B} |h_n(y) - h_1(y)| \to 0$. In this case, we have either $\beta_1 \ne \beta^*$ or $\max_{A \le y \le B} |h_1(y) - h^*(y)| > 0$.

For any arbitrarily small $\delta > 0$, we can choose $A = A(\delta)$ and $B = B(\delta)$ to be sufficiently close to the boundaries of D_Y and then extend h_1 outside $[A, B]$ so that $h_1 \in H$ and $E\{|h_1(Y) - X'\beta_1| I(A \le Y \le B)\} \ge \mathrm{E}\,|h_1(Y) - X'\beta_1| - \delta$. Then, there exists sufficiently large $N = N(\delta)$ such that $n > N$ implies

$$\mathrm{E}\,|h^*(Y) - X'\beta^*| + \delta \ge n^{-1} \sum |h^*(y_i) - x_i'\beta^*|$$

$$\ge n^{-1} \sum |h_n(y_i) - x_i'\beta_n| \ge n^{-1} \sum_{A \le y_i \le B} |h_n(y_i) - x_i'\beta_n|$$

$$\ge n^{-1} \sum_{A \le y_i \le B} |h_1(y_i) - x_i'\beta_1| - \delta \ge \mathrm{E}\{|h_1(Y) - X'\beta_1| I(A \le Y \le B)\} - 2\delta$$

$$\ge \mathrm{E}\,|h_1(Y) - X'\beta_1| - 3\delta.$$

Letting $\delta \to 0$, we would arrive at $\mathrm{E}\,|h_1(Y) - X'\beta_1| \le \mathrm{E}\,|h^*(Y) - X'\beta^*|$, which contradicts the definition of (h^*, β^*). The proof is then complete.

Finally, we consider the estimator $(\hat{h}_n, \hat{\beta}_n)$ that minimizes $\sum |h(y_i) - X_i'\beta|$ over $h \in H_B$ and $\beta \in S_p$, where H_B is the space of "power splines" defined on the interval $[y_1, y_n]$ with the set of knots $k_1 < \cdots < k_r$, see (7).

Theorem 1: If the knots are quasi-uniform in the sense that

$$\frac{\min_i(k_{i+1} - k_i)}{\max_i(k_{i+1} - k_i)} > \gamma$$

for some constant $\gamma > 0$, and the number of knots in each finite internal tends to infinity with n. then the estimator $(\hat{h}_n, \hat{\beta}_n)$ is consistent in the sense of Lemma 2.

Proof: For any choice of $[A, B]$, we know from the theory of spline approximations that there exists $\check{h}_n \in H_B$ such that $\max_{A \leq y \leq B} |\check{h}_n(y) - h^*(y)| \to 0$ as $n \to \infty$. Without loss of generality, we can choose $\check{h}_n$ to be bounded outside $[A, B]$. Thus. for sufficiently large n and sufficiently large interval $[A, B]$,

$$n^{-1} \sum_{A \leq y_i \leq B} |h^*(y_i) - x_i'\beta^*| \geq n^{-1} \sum_{A \leq y_i \leq B} |\check{h}_n(y_i) - x_i'\beta^*| - \delta \geq n^{-1} \sum |\check{h}_n(y_i) - x_i'\beta^*| - 2\delta$$

$$\geq n^{-1} \sum |\hat{h}_n(y_i) - x_i'\hat{\beta}_n| - 2\delta \geq n^{-1} \sum_{A \leq y_i \leq B} |\hat{h}_n(y_i) - x_i'\hat{\beta}_n| - 2\delta.$$

Then the same arguments used in the proof of Lemma 2 for the consistency of (h_n, β_n) apply to that of $(\hat{h}_n, \hat{\beta}_n)$.

The consistency result we obtained for the function estimate is not uniform due to the fact that the true link function h_0 may not have a bounded derivative function on its support. If we start with a finite support such as $D_Y = (0, 1)$, the model (15) with Gaussian error has to be satisfied with a link function with unbounded derivatives at 0 and 1. In this sense, uniform consistency in compact intervals inside D_Y is the best result that can be achieved.

6 Conclusion

We consider the joint estimation of a vector of regression coefficients and a monotone transformation on the dependent variable of the regression. An estimation procedure is presented which has the desirable properties of robustness to outliers, and computational efficiency. The algorithm can be easily implemented using a standard linear program solver.

In recent decades, a number of useful extensions to the standard multiple linear regression model have been developed; these include ACE (Breiman and Friedman 1985), AVAS (Tibshirani 1988), generalized additive models (Hastie and Tibshirani 1990), projection pursuit regression (Friedman and Stuetzle 1981), ATS methods (Cleveland, Mallows, and McRae 1993) slicing regression (Li 1991), transform-both-sides methods (Nychka and Ruppert 1995), robust transformation estimation (Carroll 1980; Carroll and Ruppert 1985; Carroll and Ruppert 1987), and semiparametric regression (Powell and Stoker 1989; Ichimura 1993; Bonneu, Delecroix, and Malin 1993; Horowitz 1996; Wang and Ruppert 1996; He and Shen 1997). Bayesian contributions to flexible regression modeling include West, Müeller, and Escobar (1993), Mallick and Gelfand (1994), and Laud, Damien, and Smith (1996). The ROSE algorithm presented in the present paper offers a potentially useful addition to the currently available set of techniques. The ROSE technique may be of use in cases in which the data may contain outliers, and the response variate requires nonlinear transformation, but in which the modeler wishes to have explanatory variables enter the model in an easily understandable linear and additive fashion.

References

Andrews, D. and D. Pregibon, 1978, Finding the outliers that matter, Journal of the Royal Statistical Society, Series B 40, 85–93.

Anscombe, F., 1960, Rejection of outliers, Technometrics 2, 123–147.

Barrodale, I. and F. Roberts, 1978, An efficient algorithm for discrete l_1 linear approximation, SIAM Journal of Numerical Analysis 15, 603–611.

Bassett, G. and R. Koenker, 1978, Asymptotic theory of least absolute error regressions, Journal of the American Statistical Association 73, 618–622.

Belsley, D., E. Kuh, and R. E. Welsch, 1980, Regression diagnostics (John Wiley and Sons, New York, NY).

Bickel, P. and K. Doksum, 1981, An analysis of transformations revisited, Journal of the American Statistical Association 76, 296–311.

Bonneu, M., M. Delecroix, and E. Malin, 1993, Semiparametric versus nonparametric estimation in single index regression model: A computational approach, Computational Statistics Quarterly 8, 207–222.

Box, G. and D. Cox, 1964, An analysis of transformations, Journal of the Royal Statistical Society, Series B 26, 211–246.

Box, G. and D. Cox, 1982, An analysis of transformations revisited, Journal of the American Statistical Association 77, 209–210.

Breiman, L. and J. Friedman, 1985, Estimating optimal transformations for multiple regression and correlation (with discussion), Journal of the American Statistical Association 80, 580–619.

Carroll, R., 1980, A robust method for testing transformations to achieve approximate normality, Journal of the Royal Statistical Society, Series B 42, 71–78.

Carroll, R. and D. Ruppert, 1985, Transformations: A robust analysis, Technometrics 27, 1–12.

Carroll, R. and D. Ruppert, 1987, Diagnostics and robust estimation when transforming the regression model and the response, Technometrics 29, 287–299.

Carroll, R. and D. Ruppert, 1988, Transformation and weighting in regression (Chapman & Hall, London, UK).

Cleveland, W. S., C. L. Mallows, and J. E. McRae, 1993, ATS methods: Nonparametric regression for non-Gaussian data, Journal of the American Statistical Association 88, 821–835.

Cryer, J. D., 1986, Time series analysis (Duxbury Press, Boston, MA).

Dielman, T. and R. Pfaffenberger, 1982, LAV (least absolute value) estimation in linear regression: A review, TIMS/Studies in the Management Sciences 19, 31–52.

Friedman, J. and W. Stuetzle, 1981, Projection pursuit regression, Journal of the American Statistical Association 76, 817–823.

Gonin, R. and J. Money, 1989, Non-linear l_p–norm estimation (John Wiley and Sons, New York, NY).

Harrison, D. and D. Rubinfeld, 1978, Hedonic prices and the demand for clean air, Journal of Environmental Economics and Management 5, 81–102.

Hastie, T. and R. Tibshirani, 1990, Generalized additive models (Chapman & Hall, London, UK).

He, X. and L. Shen, 1997, Linear regression after spline transformation, Biometrika (to appear) .

He, X. and P. Shi, 1996, Monotone B–spline smoothing, Technical report, University of Illinois Department of Statistics.

Hinkley, D. and G. Runger, 1984, Analysis of transformed data, Journal of the American Statistical Association 79, 302–308.

Hoerl, A. and R. Kennard, 1970, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12, 55–67.

Horowitz. J. L., 1996. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable, Econometrica 64. 103–137.

Ichimura, H., 1993, Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, Journal of Econometrics 58, 71–120.

Koenker, R., P. Ng, and S. Portnoy, 1994, Quantile smoothing splines. Biometrika 81. 673–680.

Laud, P., P. Damien, and A. Smith, 1996, Bayesian nonparametric and covariate analysis of failure time data. Technical report. University of Michigan Business School.

Li. K.-C., 1991, Slicing inverse regression for dimension reduction. Journal of the American Statistical Association 86, 316–342.

Li, K.-C. and N. Duan, 1989, Regression analysis under link violation. Annals of Statistics 17, 1009–1052.

Lock. R., 1993, 1993 new car data, Journal of Statistics Education 1.

Mallick. B. and A. Gelfand. 1994, Generalized linear models with unknown link functions. Biometrika 81, 237–245.

Mallows, C., 1973, Some comments on Cp, Technometrics 15, 661–676.

Mansfield, E., J. Webster, and R. Gunst, 1977, An analytic variable selection technique for principal component regression, Applied Statistics 26, 34–40.

Nychka, D. and D. Ruppert, 1995, Nonparametric transformations for both sides of a regression model, Journal of the Royal Statistical Society, Series B 57, 519–532.

Powell, J. L. and T. M. Stoker, 1989, Semiparametric estimation of index coefficients, Econometrica 57, 1403–30.

Ramsay, J., 1988, Monotone regression splines (with discussion), Statistical Science 3. 425–461.

Smith, P., 1979, Splines as a useful and convenient statistical tool, American Statistician 33, 57–62.

Tibshirani, R., 1988, Estimating optimal transformations for regression via additivity and variance stabilization, Journal of the American Statistical Association 83, 394–405.

Wang, N. and D. Ruppert, 1996, Estimation of regression parameters in a semiparametric transformation model, Journal of Statistical Planning and Inference 52, 331–351.

West, M., P. Müeller, and M. Escobar, 1993, Hierarchical priors and mixture models with applications in regression and density estimation, Technical Report A02, Duke University, Institute of Statistics and Decision Sciences.

Young, M., 1996, Robust seasonal adjustment by Bayesian modeling, Journal of Forecasting 15, 355–367.

Table 1: Parameter estimates. Boston housing data.

Variable	ROSE	LAD	OLS
CRIM	-0.0198	-0.0239	-0.2693
ZN	0.0020	0.0031	0.0907
INDUS	0.0065	0.0079	0.0498
CHAS	0.1586	0.1596	0.0806
NOX	-0.8490	-0.8203	-0.2478
RM	0.4889	0.5296	0.1932
AGE	-0.0020	-0.0019	0.0067
DIS	-0.0799	-0.0940	-0.3002
RAD	0.0192	0.0208	0.3728
TAX	-0.0012	-0.0014	-0.3233
PTRATIO	-0.0752	-0.0900	-0.2574
B	0.0018	0.0018	0.1144
LSTAT	-0.0464	-0.0569	-0.6313

Table 2: Parameter estimates, 1993 car fuel economy data.

	OLS	ROSE
INTERCEPT	29.289	54.380
CYL	-0.241	-0.915
ENGSIZE	+0.888	-0.000
HP	+0.655	-0.595
LENGTH	+0.251	-0.085
WIDTH	+4.228	-0.000
WEIGHT	-23.72	-13.618
DOMESTIC	-1.804	-0.830

Table 3: Results of Monte Carlo evaluation of estimators.

		RMSE				
n	σ	ROSE	LAD	OLS	Signif. LAD	Signif. OLS
100	2.0	0.0263	0.0495	0.0360	*	*
100	4.0	0.0397	0.0666	0.0521	*	*
50	2.0	0.0402	0.0791	0.0621	*	*
50	4.0	0.0665	0.1010	0.0856	*	*

* = Semiparametric estimator more efficient. $p < .01$. binomial sign test.

Figure 1: Aggregate sales in millions of dollars, Ford Motor Company

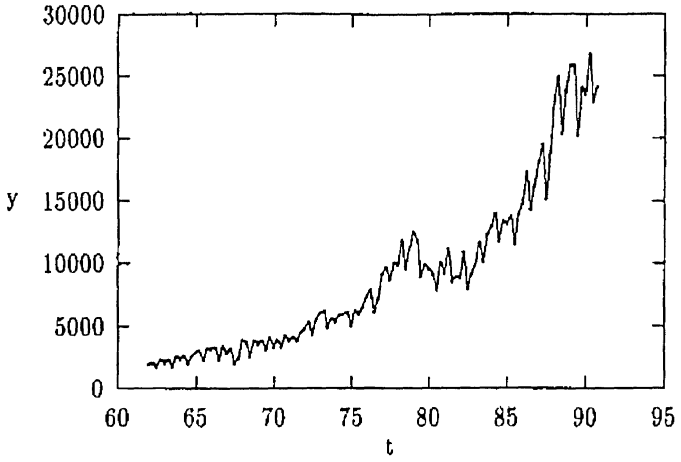


Figure 2: Estimated transformation of dependent variable in seasonal regression model of Ford Motor Company sales data

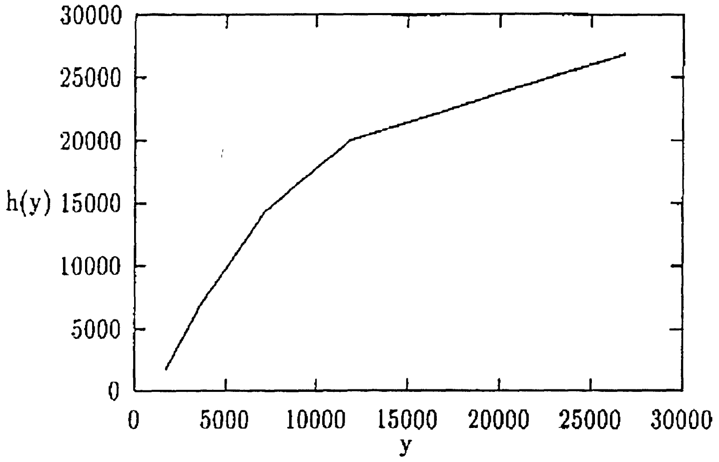


Figure 3: Transformed sales data, Ford Motor Company

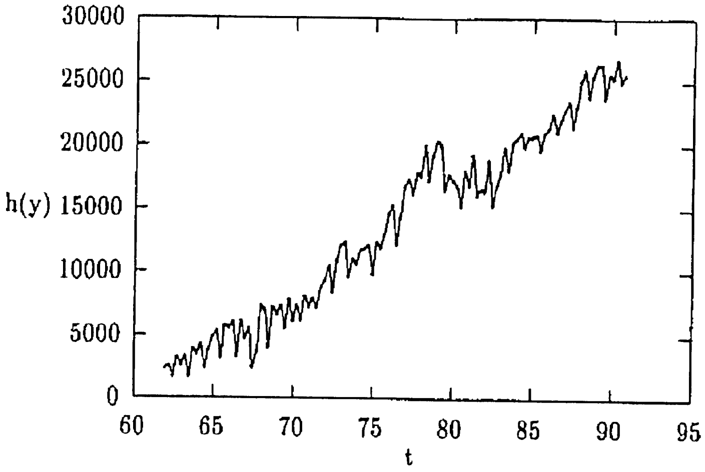


Figure 4: Residual plot, OLS regression of 1993 cars fuel economy data

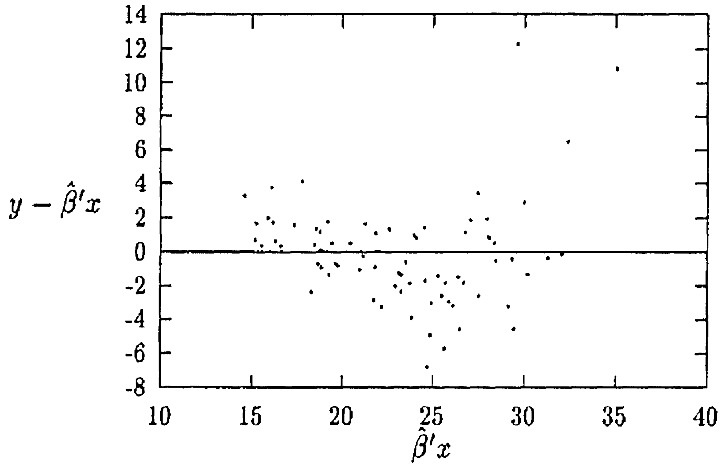


Figure 5: Estimated transformation of dependent variable in 1993 cars fuel economy regression model

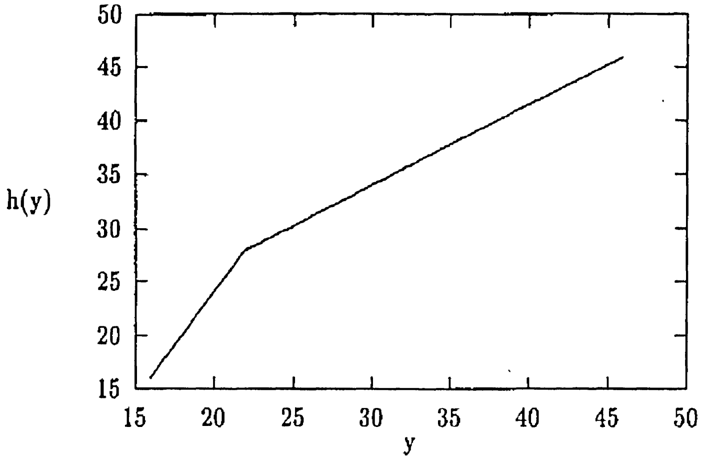


Figure 6: Residual plot, robust semiparametric regression of 1993 cars fuel economy data

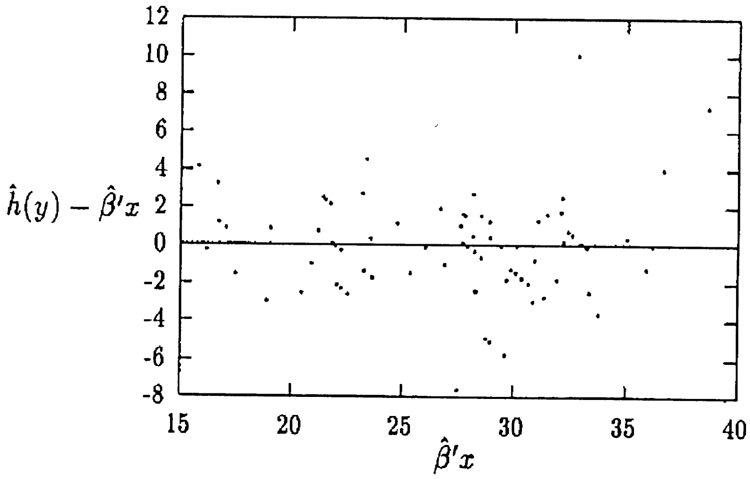
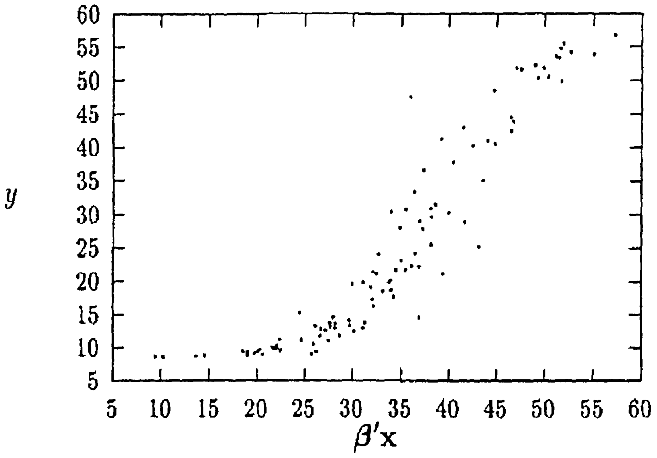


Figure 7: Sample dataset from Monte Carlo study: $n = 100$, $\sigma = 2.0$.


60
55
50
45
40
35
30
25
20
15
10
5
y
5
10
15
20
25
30
35
40
45
50
55
60
β'x