2004

Visualizing 1D Regression

David J. Olive
Southern Illinois University Carbondale, dolive@math.siu.edu

Published in *Theory and Applications of Recent Robust Methods*, edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Series: Statistics for Industry and Technology, Birkhauser: Basel, 221-233. The original publication is available at www.springerlink.com.

Visualizing 1D Regression

David J. Olive

Abstract. Regression is the study of the conditional distribution of the response y given the predictors $\boldsymbol{x}$. In a 1D regression, y is independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors. Special cases of 1D regression include multiple linear regression, binary regression and generalized linear models. If a good estimate $\hat{\boldsymbol{b}}$ of some non-zero multiple $c\boldsymbol{\beta}$ of $\boldsymbol{\beta}$ can be constructed, then the 1D regression can be visualized with a scatterplot of $\hat{\boldsymbol{b}}^T\boldsymbol{x}$ versus y. A resistant method for estimating $c\boldsymbol{\beta}$ is presented along with applications.

1. INTRODUCTION

Regression is the study of the conditional distribution $y|\boldsymbol{x}$ of the response y given the $(p-1) \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. In a *1D regression* (or regression with 1–dimensional structure), y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}. \tag{1.1}$$

A *1D regression model* has the form

$$y = g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, e) \tag{1.2}$$

where g is a bivariate (inverse link) function and e is a zero mean error that is independent of $\boldsymbol{x}$. See [20] and [14, p. 414].

The above class of models is very rich. A *single index model* uses

$$y = g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, e) \equiv m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e, \tag{1.3}$$

and the *multiple linear regression model* is an important special case where m is the identity function: $m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Another important special case of 1D regression is the *response transformation model* where

$$g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e) \tag{1.4}$$

Received by the editors September 2, 2003.
1991 *Mathematics Subject Classification.* Primary 62G35; Secondary 62F35.
Key words and phrases. Diagnostics, Generalized Linear Models, Minimum Covariance Determinant, Response Transformations, Single Index Models.
This research was supported by NSF grant DMS 0202922.

and t^{-1} is a one to one (typically monotone) function so that $t(y) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e$. Generalized linear models (GLM's) are also a special case of 1D regression.

Some notation from the regression graphics literature will be useful. *Dimension reduction* can greatly simplify our understanding of the conditional distribution $y|\boldsymbol{x}$. If a 1D regression model is appropriate, then the $(p-1)$–dimensional vector $\boldsymbol{x}$ can be replaced by the 1–dimensional scalar $\boldsymbol{\beta}^T \boldsymbol{x}$ with no loss of information. A *sufficient summary plot* (SSP) is a plot that contains all the sample regression information about the conditional distribution of $y|\boldsymbol{x}$. For 1D regression, if $y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T \boldsymbol{x}$ then $y \perp\!\!\!\perp \boldsymbol{x}|c\boldsymbol{\beta}^T \boldsymbol{x}$ for any constant $c \neq 0$. The quantity $c\boldsymbol{\beta}^T \boldsymbol{x}$ is called a *sufficient predictor* (SP), and a plot of the SP versus y is a SSP. If a consistent estimator $\hat{\boldsymbol{b}}$ of $c\boldsymbol{\beta}$ can be found for some nonzero c, then an *estimated sufficient summary plot* (ESSP) is a plot of the *estimated sufficient predictor* (ESP) $\hat{\boldsymbol{b}}^T \boldsymbol{x}$ versus y.

Additional notation is needed before giving theoretical results. Let $\boldsymbol{x}$, $\boldsymbol{a}$, $\boldsymbol{t}$, and $\boldsymbol{\beta}$ be $(p-1) \times 1$ vectors where only $\boldsymbol{x}$ is random. The predictors $\boldsymbol{x}$ satisfy the condition of *linearly related predictors* with 1D structure ([14, p. 431]) if

$$E[\boldsymbol{x}|\boldsymbol{\beta}^T \boldsymbol{x}] = \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T \boldsymbol{x}. \tag{1.5}$$

Notice that $\boldsymbol{\beta}$ is a fixed $(p-1) \times 1$ vector. If $\boldsymbol{x}$ is elliptically contoured (EC) with 1st moments, then the assumption of linearly related predictors holds. See [6, p. 130].

Following [6, pp. 143-144], assume that there is an objective function

$$L_n(a, \boldsymbol{b}) = \frac{1}{n} \sum_{i=1}^{n} L(a + \boldsymbol{b}^T \boldsymbol{x}_i, y_i) \tag{1.6}$$

where $L(u, v)$ is a bivariate function that is a convex function of the first argument u. Assume that the estimate $(\hat{a}, \hat{\boldsymbol{b}})$ of $(a, \boldsymbol{b})$ satisfies

$$(\hat{a}, \hat{\boldsymbol{b}}) = \arg \min_{a, \boldsymbol{b}} L_n(a, \boldsymbol{b}). \tag{1.7}$$

For example, the ordinary least squares (OLS) estimator uses

$$L(a + \boldsymbol{b}^T \boldsymbol{x}, y) = (y - a - \boldsymbol{b}^T \boldsymbol{x})^2.$$

Maximum likelihood type estimators such as those used to compute GLM's and Huber's M–estimator also work, as does the Wilcoxon rank estimator. Assume that the population analog $(\alpha, \boldsymbol{\eta})$ is the unique minimizer of $E[L(a + \boldsymbol{b}^T \boldsymbol{x}, y)]$ where the expectation exists and is with respect to the joint distribution of $(y, \boldsymbol{x}^T)^T$. For example, $(\alpha, \boldsymbol{\eta})$ is unique if $L(u, v)$ is strictly convex in its first argument. The following result is a useful extension of [2, 3].

Theorem 1. ([20, p. 1016]): *Assume that the $\boldsymbol{x}$ are linearly related predictors, that $(y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some joint distribution and that* $\mathrm{Cov}(\boldsymbol{x}_{\mathrm{i}})$ *exists and is positive definite. Assume $L(u, v)$ is convex in its first argument and that $\boldsymbol{\eta}$ is unique. Assume that $y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T \boldsymbol{x}$. Then $\boldsymbol{\eta} = c\boldsymbol{\beta}$ for some scalar c.*

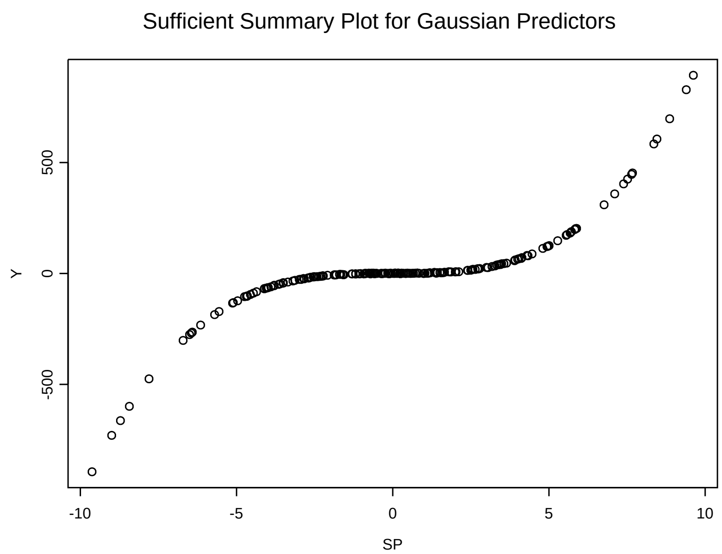

FIGURE 1. SSP for $m(u) = u^3$

Remark 1. If $\hat{\boldsymbol{b}}$ is a consistent estimator of $\boldsymbol{\eta} \equiv \boldsymbol{\beta}_b$, then certainly

$$\boldsymbol{\beta}_b = c\boldsymbol{\beta} + \boldsymbol{u}_g$$

where $\boldsymbol{u}_g = \boldsymbol{\beta}_b - c\boldsymbol{\beta}$ is the bias vector. If the conditions of Theorem 1 hold, then $\boldsymbol{u}_g = \boldsymbol{0}$. Under additional conditions, $(\hat{a}, \hat{\boldsymbol{b}}^T)^T$ is asymptotically normal (see [20, p. 1031]). In particular, the OLS estimator frequently has a $\sqrt{n}$ convergence rate. Often if no strong nonlinearities are present among the predictors, the bias vector is small enough so that $\hat{\boldsymbol{b}} \approx c\boldsymbol{\beta}$. See [14, pp. 431-441] for checking whether the predictors are linearly related and whether a 1D regression model is appropriate.

A very useful result is that if $y = m(x)$ for some function m, then m can be visualized with both a plot of x versus y and a plot of $c\,x$ versus y if $c \neq 0$. In fact, there are only three possibilities: if $c > 0$, then the two plots are nearly identical. If $c < 0$, then the plot appears to be flipped about the vertical axis. If $c = 0$, then the plot is a dot plot. Similarly, if $y_i = g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i, e_i)$, then the plot of $\boldsymbol{\beta}^T\boldsymbol{x}$ versus y and the plot of $c\boldsymbol{\beta}^T\boldsymbol{x}$ versus y will be nearly identical in overall shape if $c > 0$.

Example 1. Suppose that $\boldsymbol{x}_i \sim N_3(\boldsymbol{0}, \boldsymbol{I}_3)$ where $\boldsymbol{I}_3$ is the 3×3 identity matrix, and

$$y = m(\boldsymbol{\beta}^T\boldsymbol{x}) + e = (x_1 + 2x_2 + 3x_3)^3 + e \quad \text{with} \quad \text{n} = 100.$$

Then a 1D regression model holds with $\boldsymbol{\beta} = (1, 2, 3)^T$. Figure 1 shows the sufficient summary plot of $\boldsymbol{\beta}^T\boldsymbol{x}_i$ versus y_i, and Figure 2 shows the sufficient summary plot

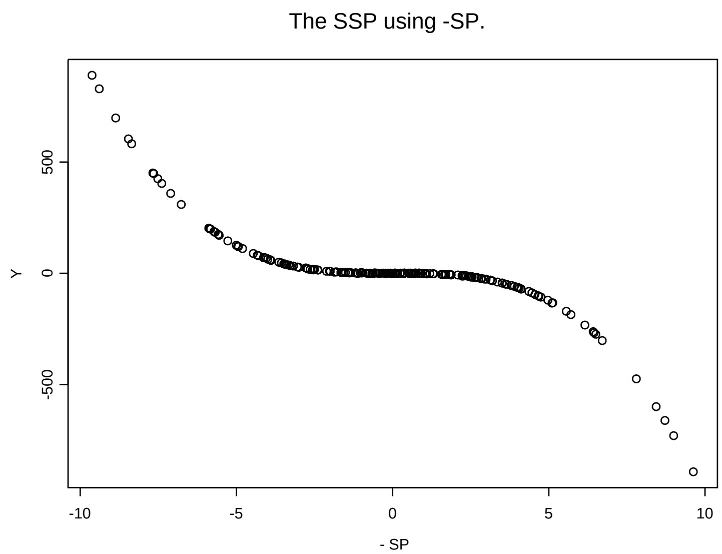

FIGURE 2. Another SSP for $m(u) = u^3$

of $-\boldsymbol{\beta}^T \boldsymbol{x}_i$ versus y_i. Notice that the functional form m appears to be cubic in both plots and that both plots can be smoothed by eye or with a scatterplot smoother.

Remark 2. The OLS estimator $(\hat{a}_o, \hat{\boldsymbol{b}}_o)$ is obtained from the usual multiple linear regression of y_i on $\boldsymbol{x}_i$, but *we are not assuming that the multiple linear regression model holds*; however, we are hoping that the 1D regression model $y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$ is a useful approximation to the data and that $\hat{\boldsymbol{b}}_o \approx c\boldsymbol{\beta}$ for some nonzero constant c. Theorem 1 provides some conditions for the above approximation to hold. Notice that if the multiple linear regression model does hold and if the errors e_i are such that OLS is a consistent estimator, then $c = 1$, and $\boldsymbol{u}_g = \boldsymbol{0}$.

The following result, perhaps first noted by [2, 3], is called the *1D Estimation Result* by [14, p. 432]: let $(\hat{a}_o, \hat{\boldsymbol{b}}_o)$ denote the OLS estimate obtained from the OLS multiple linear regression of y on $\boldsymbol{x}$. The *OLS view* is a plot of $\hat{\boldsymbol{b}}_o^T \boldsymbol{x}$ versus y. If the 1D regression model is appropriate and if no strong nonlinearities are present among the predictors, then *the OLS view will frequently be a useful estimated sufficient summary plot.* Hence the OLS predictor $\hat{\boldsymbol{b}}_o^T \boldsymbol{x}$ is a useful ESP.

Three additional methods considered in this paper that have proven useful for estimating the ESP are sliced inverse regression (SIR), principal Hessian directions (PHD), and sliced average variance estimation (SAVE). See [10] for a discussion of when these methods can fail. These methods frequently perform well if there are no strong nonlinearities present in the predictors. All of these methods fail if $c = 0$

or if the bias vector $\boldsymbol{u}_g$ is "large" compared to $c\boldsymbol{\beta}$. For example, the OLS view can fail if the sufficient summary plot of $\boldsymbol{\beta}^T\boldsymbol{x}$ versus y is approximately symmetric, and all of these methods can perform poorly if outliers are present ([17] shows that a single outlier can cause SIR to fail). Some useful references for SIR and related methods include [6, 7, 8], [14, 15], [16], [19] and [24].

Ellipsoidal trimming is a method for estimating the ESP that can reduce the bias $\boldsymbol{u}_g$. See [11] and [6, p. 152]. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed where T is a $(p-1)\times 1$ multivariate location estimator and $\boldsymbol{C}$ is a $(p-1)\times(p-1)$ symmetric positive definite dispersion estimator. Then the ith squared Mahalanobis distance is the scalar

$$D_i^2 = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{1.8}$$

for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \le D_{(j)}^2\}. \tag{1.9}$$

The ith case $(y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $c\boldsymbol{\beta}$ is computed from the untrimmed cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS could be used on the remaining cases.

The following procedure was suggested by [21]. First compute $(T, \boldsymbol{C})$ using the *Splus* function `cov.mcd` (see [23]). Trim the K% of the cases with the largest Mahalanobis distances, and then compute the OLS estimator $(\hat{\alpha}_K, \hat{\boldsymbol{\beta}}_K)$ from the untrimmed cases. Use $K = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten plots of $\hat{\boldsymbol{\beta}}_K^T\boldsymbol{x}$ versus y using all n cases. These plots will be called "OLS trimmed views." Notice that $K = 0$ corresponds to the OLS view. The *best OLS trimmed view* is the trimmed view with a smooth mean function and the smallest variance function and is the estimated sufficient summary plot. If $K^* = E$ is the percentage of cases trimmed that corresponds to the best trimmed view, then $\hat{\boldsymbol{\beta}}_E^T\boldsymbol{x}$ is the estimated sufficient predictor.

Example 2. For the data in Example 1, the OLS view is similar to Figure 1 except the plot is not quite as smooth and the horizontal scale is multiplied by $c \approx 42$. The best trimmed view appears to be identical to Figure 1 except that the horizontal scale is multiplied by $c \approx 12.5$. The OLS view used $\hat{\boldsymbol{\beta}}_0 = (41.68, 87.40, 120.83)^T \approx 42\boldsymbol{\beta}$ while the best trimmed view used $\hat{\boldsymbol{\beta}}_{50} = (12.61, 25.07, 37.26)^T \approx 12.5\boldsymbol{\beta}$.

This section has reviewed the existing literature on 1D regression. Section 2 shows that 1D regression can provide useful diagnostics when g is known. Section 3 considers estimating g when $g = g_\lambda$ and $\lambda \in \{\lambda_1, ..., \lambda_k\}$ where k is a small integer. Section 4 suggests using ellipsoidal trimming with methods other than OLS. This technique gives a resistant version of SIR and shows that the *Splus* function `lmsreg` can be very useful for finding certain types of curvature.

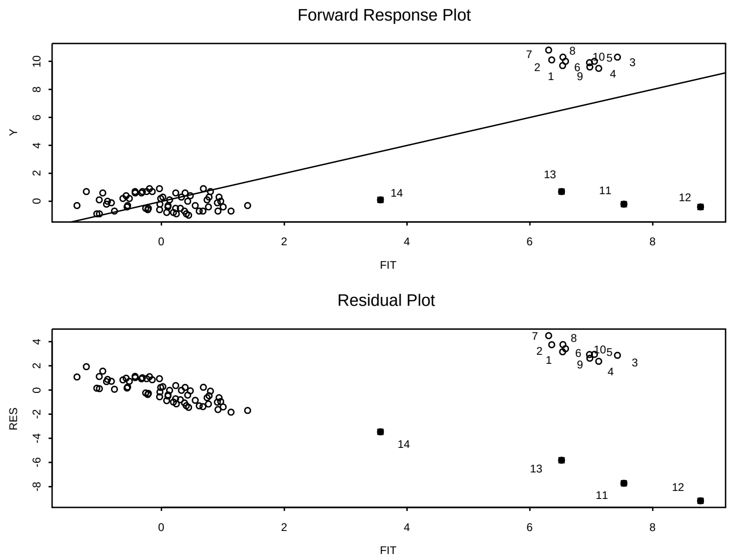

FIGURE 3. Plots for HBK Data

2. 1D REGRESSION DIAGNOSTICS

In this section, we suggest that when g is known, an estimated sufficient summary plot should be used in addition to the usual diagnostics for checking the model. Assume that the 1D regression model is $y_i = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i, e_i)$ for $i = 1, ..., n$ where the e_i are iid with zero mean and variance $V(e_i) = \sigma^2$.

Remark 3. If $y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$ then $y \perp\!\!\!\perp \boldsymbol{x} | a + c\boldsymbol{\beta}^T \boldsymbol{x}$ for any constants a and $c \neq 0$. Hence if $\hat{\boldsymbol{b}}^T \boldsymbol{x}$ is an ESP, so is $a + \hat{\boldsymbol{b}}^T \boldsymbol{x}$.

A good example is the multiple linear regression (MLR) model $y_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$. Let $(\hat{a}, \hat{\boldsymbol{b}})$ be a MLR estimator of $(\alpha, \boldsymbol{\beta})$. Then the fitted values are $\hat{y}_i = \hat{a} + \hat{\boldsymbol{b}}^T \boldsymbol{x}_i$, and the residuals are $r_i = y_i - \hat{y}_i$. The most used residual plot is a plot of $\hat{y}_i$ versus r_i, and the *forward response plot* is a plot of the fitted values $\hat{y}_i$ versus the response y_i.

Remark 3 shows that the forward response plot is an ESSP. Let the scalars $w_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$. Ignoring the errors gives $y_i = w_i$ which is the equation of the *identity line* that has unit slope and zero intercept. Hence if the MLR model is appropriate and if $(\hat{a}, \hat{\boldsymbol{b}})$ is a good estimator of $(\alpha, \boldsymbol{\beta})$, then the plotted points in the forward response plot should scatter about the identity line. The vertical deviations from the identity line are the residuals r_i since these deviations are $y_i - \hat{y}_i$. When the OLS estimator is used, the coefficient of determination R^2 is equal to the squared correlation of y_i and $\hat{y}_i$. See [4, p. 280].

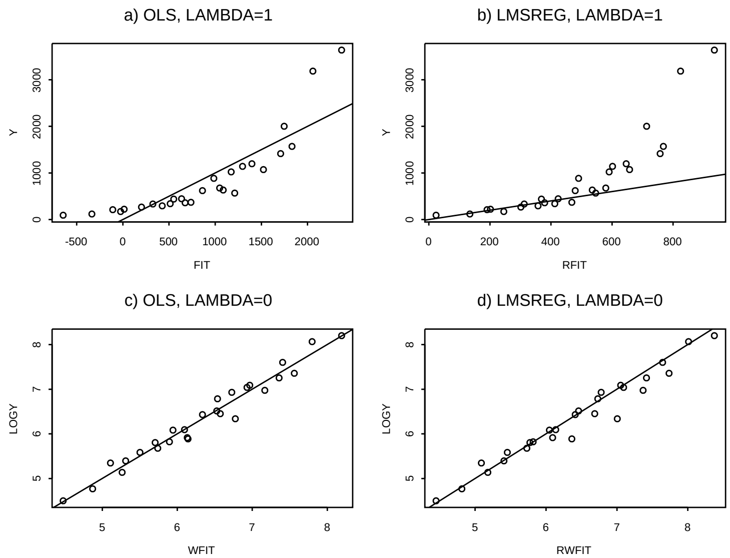

FIGURE 4. OLS and LMSREG Suggest Using log(y) for the Textile Data

High leverage outliers challenge conventional numerical MLR diagnostics such as Cook's distance ([5]), but, as shown in Example 3 below, can often be detected using the forward response and residual plots. Using trimmed views (see $\oint$ 4) is also effective for detecting outliers and other departures from the MLR model.

Example 3. In the well known artificial HBK data set ([18]), the first 10 cases are outliers while cases 11-14 are good leverage points. This data set has $n = 75$ cases and $p - 1 = 3$ nontrivial predictors. Figure 3 shows the residual and forward response plots based on the OLS estimator. The highlighted cases have Cook's distance $> \min(0.5, 2p/n)$, and the identity line is shown in the ESSP.

Now suppose that model (1.2) holds where g is known. If the estimator $(\hat{a}, \hat{\boldsymbol{b}})$ satisfies $\hat{\boldsymbol{b}} \approx c\boldsymbol{\beta}$, then the plot of $\hat{a} + \hat{\boldsymbol{b}}^T \boldsymbol{x}$ or $\hat{\boldsymbol{b}}^T \boldsymbol{x}$ versus y can be used to visualize g provided that $c \neq 0$. Since g is known, the classical (e.g. maximum likelihood) estimator for $\boldsymbol{\beta}$ should be used since then $c = 1$ and the bias vector should be small for large sample size n. Often adding a parametric fit and a lowess smooth to the plot will be useful.

Plots are also useful for additive error models

$$y_i = m(x_{i,1}, ..., x_{i,p-1}) + e_i = m(\boldsymbol{x}_i^T) + e_i = m_i + e_i.$$

Many anova, categorical, nonlinear regression, nonparametric regression, and multi-index models have this form. For these models $\hat{y}_i = \hat{m}_i$ and the residuals $r_i =$

$y_i - \hat{y}_i$. In the fit–response plot (FY plot) of $\hat{y}_i$ versus y_i, the plotted points should scatter about the identity line, and the vertical deviations from the identity line are equal to the residuals.

3. 1D REGRESSION MODEL SELECTION

In this section, we assume that a 1D regression model holds with $g = g_{\lambda_o}$ where $\lambda_o \in \Lambda = \{\lambda_1, ..., \lambda_k\}$ and $k > 1$ is a small integer. To estimate λ_o, make an ESSP for each of the k possible models. Examples include choosing a frequentist or a Bayesian model; a proportional hazards model or one of several competing 1D survival models; a logistic, probit or complementary log–log model in binary regression; a full or sub model in variable selection.

A good example for illustration is the response transformation model

$$t_{\lambda_o}(y_i) \equiv y_i^{(\lambda_o)} = \alpha_0 + \boldsymbol{\beta}_0^T \boldsymbol{x}_i + e_i \tag{3.1}$$

where the response variable $y_i > 0$ and the *power transformation family*

$$t_\lambda(y) \equiv y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda} \tag{3.2}$$

for $\lambda \neq 0$ and $y^{(0)} = \log(y)$. Assume $\lambda_o \in \Lambda = \{0, \pm 1/4, \pm 1/3, \pm 1/2, \pm 2/3, \pm 1\}$.

The literature for estimating λ_o is enormous, and at least two papers using results from 1D regression have appeared. Let the OLS estimator $(\hat{a}_o, \hat{\boldsymbol{b}}_o)$ be computed from the multiple linear regression of y_i on $\boldsymbol{x}_i$. Then [13] suggests that the inverse response plot of y versus $\hat{y}_{OLS}$ will often show t_{λ_o}. Hence the forward response plot of $\hat{y}_{OLS}$ versus y will show $t_{\lambda_o}^{-1}$. If t_λ is the appropriate transformation, [12] suggests that a plot of $\hat{y}_{OLS}$ versus $t_\lambda(y)$ will follow the identity line.

These ideas suggest a graphical method for selecting response transformations that can be used with any good MLR estimator. Let $w_i = t_\lambda(y_i)$ for $\lambda \neq 1$, and let $w_i = y_i$ if $\lambda = 1$. Next, perform the multiple linear regression of w_i on $\boldsymbol{x}_i$ and make the forward response plot of $\hat{w}_i$ versus w_i. If the plotted points follow the identity line, then take $\lambda_o = \lambda$. One plot is made for each of the eleven values of $\lambda \in \Lambda$, and if more than one value of λ works, contact subject matter experts and use the simplest or most reasonable transformation. (Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator by adding the estimate $\hat{\lambda}$ of λ_o to Λ.)

Example 4. A textile data set is given in [1] where samples of worsted yarn with different levels of the three factors were given a cyclic load until the sample failed. The goal was to understand how y = *the number of cycles to failure* was related to the predictor variables *length, amplitude* and *load*. Figure 4 shows the forward response plots for two MLR estimators: OLS and the *Splus* function `lmsreg`. Figures 4a and 4b show that a response transformation is needed while 4c and 4d both suggest that $\log(y)$ is the appropriate response transformation. Using OLS and a resistant estimator as in Figure 4 may be very useful if outliers are present.

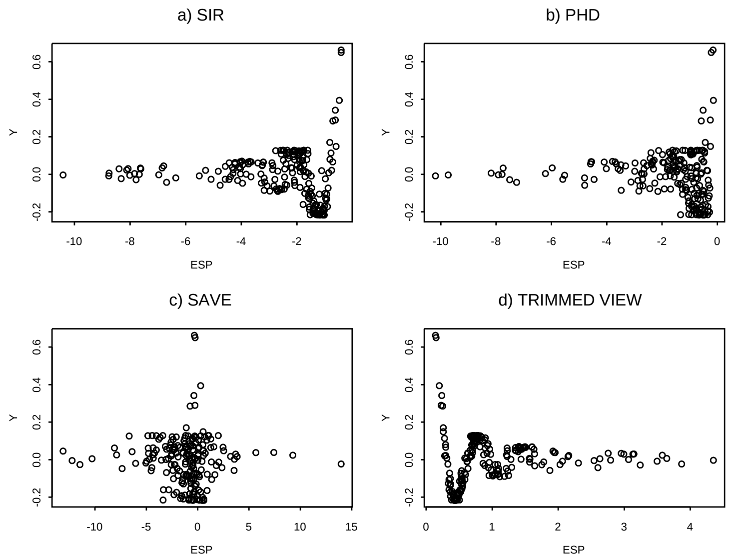

FIGURE 5. Estimated Sufficient Summary Plots

4. IMPROVING 1D ESTIMATORS

Assume that the 1D regression model (1.2) holds but both g and $\boldsymbol{\beta}$ are unknown. If a good estimator $\hat{\boldsymbol{b}} \approx c\boldsymbol{\beta}$ can be found where $c \neq 0$, then the ESSP can be used to visualize g. The next step might be to fit a tentative parametric or nonparametric model in the ESP.

Since methods for estimating the sufficient predictor such as OLS, PHD, SAVE and SIR can fail if strong nonlinearities such as outliers are present in the predictors or if $\hat{\boldsymbol{b}} \approx 0\ \boldsymbol{\beta}$, techniques for improving these methods are needed. The basic tool will be to use $(\hat{\boldsymbol{b}}, T, \boldsymbol{C})$ trimmed views where $\hat{\boldsymbol{b}}$ is an estimator of $c\boldsymbol{\beta}$ and $(T, \boldsymbol{C})$ is an estimator of multivariate location and dispersion. Two good choices for $(T, \boldsymbol{C})$ are the classical estimator ([6, p. 152]) or a robust estimator such as `cov.mcd`. Then, for example, SIR trimmed views generalize the SIR ESSP in the same way that OLS trimmed views generalize the OLS view: use ellipsoidal trimming to delete the K% of the cases with the largest Mahalanobis distances. Next, plot $\hat{\boldsymbol{b}}_K^T \boldsymbol{x}_i$ versus y_i where $\hat{\boldsymbol{b}}_K$ is the first SIR direction computed from the untrimmed cases. Again use 10 values of K where $K = 0$ corresponds to the usual SIR ESSP, and the best SIR trimmed view is the trimmed view with a smooth mean function and the smallest variance function.

Using trimmed views seems to work for several reasons. The ellipsoidal trimming divides the data into two groups: the *trimmed cases* and the *untrimmed cases*. Trimming often removes strong nonlinearities from the predictors, and the

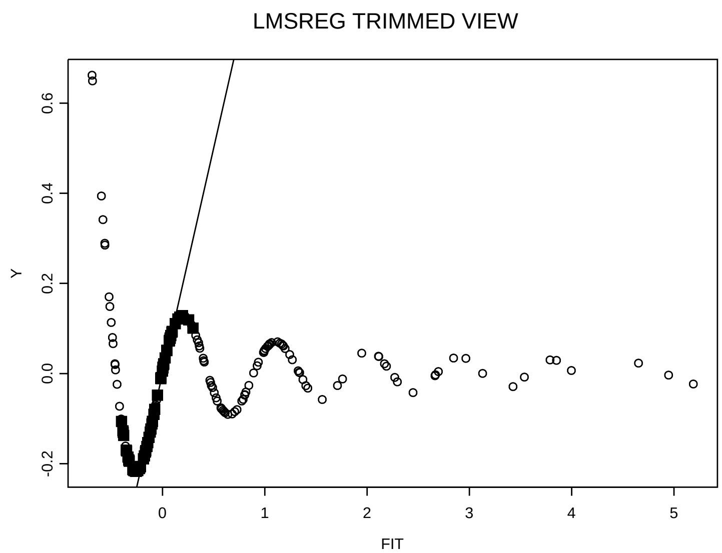

FIGURE 6. The Weighted `lmsreg` Fitted Values vs. Y

untrimmed predictor distribution is often more nearly elliptical contoured than the predictor distribution of the entire data set (recall Winsor's principle: "data are roughly Gaussian in the middle"). Secondly, under heavy trimming, the mean function of the untrimmed cases may be more linear than the mean function of the entire data set. Thirdly, if $|c|$ is very large, then the bias vector $\boldsymbol{u}_g$ may be small relative to $c\boldsymbol{\beta}$. From Remarks 1 and 2, any of these three reasons could produce a better estimated sufficient predictor. Also notice that trimmed views are resistant to y–outliers since the y values are plotted, and trimmed views are resistant to $\boldsymbol{x}$–outliers if $(T, \boldsymbol{C})$ is a resistant estimator.

Example 5. To illustrate the above discussion, an artificial data set with 200 trivariate vectors $\boldsymbol{x}_i$ was generated. The marginal distributions of $x_{i,j}$ are iid lognormal for $j = 1, 2,$ and 3. Since the response $y_i = \sin(\boldsymbol{\beta}^T\boldsymbol{x}_i)/\boldsymbol{\beta}^T\boldsymbol{x}_i$ where $\boldsymbol{\beta} = (1, 2, 3)^T$, the random vector $\boldsymbol{x}_i$ is not elliptically contoured and the function g is strongly nonlinear. The `cov.mcd` estimator was used for trimming, and Figure 5 shows the estimated sufficient summary plots for SIR, PHD, SAVE (using 8 slices), and the OLS 90% trimmed view. The OLS trimmed view is the best, while SAVE completely fails. Figure 6 shows that for this data, the `lmsreg` trimmed view is very useful for visualizing g. The `lmsreg` estimator attempts to make the median squared MLR residual small and implements the PROGRESS algorithm described in [22, pp. 197-204]. Table 1 shows the estimated sufficient predictor coefficients $\hat{\boldsymbol{b}}$ when the sufficient predictor coefficients are $c(1, 2, 3)^T$. Only the OLS and `lmsreg`

TABLE 1. Estimated Sufficient Predictors Estimating $c(1, 2, 3)^T$

method	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$
OLS View	0.0032	0.0011	0.0047
SIR	-0.394	-0.361	-0.845
PHD	-0.072	-0.029	-0.0097
SAVE	-1.09	0.870	-0.480
OLS 90% Trimmed View	0.086	0.182	0.338
LMSREG 70% Trimmed View	0.143	0.287	0.428

trimmed views produce estimated sufficient predictors that are highly correlated with the sufficient predictor. (The SAVE 40% trimmed view was also very good.)

Figure 6 helps illustrate why the best `lmsreg` trimmed view worked. This view used 70% trimming, and the open circles denote cases that were trimmed while the highlighted squares are the untrimmed cases. Note that the highlighted cases are far more linear than the data set as a whole. Also `lmsreg` will give about half of the highlighted cases zero weight, further linearizing the function. In Figure 6 the weighted `lmsreg` constant $\hat{\alpha}_{70}$ is included, and the plot is simply the forward response plot of the weighted `lmsreg` fitted values versus y. The vertical deviations from the identity line are the "MLR residuals" $y_i - \hat{\alpha}_{70} - \hat{\boldsymbol{\beta}}_{70}^T \boldsymbol{x}_i$ and at least half of the highlighted cases have small MLR residuals. There exist data sets where OLS is better than `lmsreg` for showing curvature (see [9]), but, as illustrated by Example 5, `lmsreg` often performed better for single index models when m was smooth.

A great deal of work remains to be done in the area of resistant dimension reduction. When a 1D regression model holds, the trimmed views work at least as well as the untrimmed view since the untrimmed view corresponds to 0% trimming. The webpage (http://www.math.siu.edu/olive/) contains programs and a good introduction to 1D regression models.

References

[1] G. E. P. Box, and D. R. Cox, *An Analysis of Transformations.* J. of the Royal Statistical Society, B **26** (1964), 211-246.

[2] D. R. Brillinger, *The Identification of a Particular Nonlinear Time Series.* Biometrika **64** (1977), 509-515.

[3] D. R. Brillinger, *A Generalized Linear Model with "Gaussian" Regressor Variables.* In *A Festschrift for Erich L. Lehmann,* eds. P.J. Bickel, K.A. Doksum, and J.L. Hodges, Wadsworth, Pacific Grove, CA, (1983), 97-114.

[4] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. Tukey, *Graphical Methods for Data Analysis.* Duxbury Press, Boston, 1983.

[5] R. D. Cook, *Deletion of Influential Observations in Linear Regression.* Technometrics **19** (1977), 15-18.

[6] R. D. Cook, *Regression Graphics: Ideas for Studying Regression Through Graphics.* John Wiley and Sons, Inc., NY. 1998.

[7] R. D. Cook, *Principal Hessian Directions Revisited.* J. of the American Statistical Association **93** (1998), 84-100.

[8] R. D. Cook, *SAVE: A Method for Dimension Reduction and Graphics in Regression.* Communications in Statistics Theory and Methods **29** (2000), 2109-2121.

[9] R. D. Cook, D. M. Hawkins, and S. Weisberg, *Comparison of Model Misspecification Diagnostics Using Residuals from Least Mean of Squares and Least Median of Squares.* J. of the American Statistical Association **87** (1992), 419-424.

[10] R. D. Cook, and B. Li, *Dimension Reduction for Conditional Mean in Regression.* The Annals of Statistics **30** (2002), 455-474.

[11] R. D. Cook, and C. J. Nachtsheim, *Reweighting to Achieve Elliptically Contoured Covariates in Regression.* J. of the American Statistical Association **89** (1994), 592-599.

[12] R. D. Cook, and D. J. Olive, *A Note on Visualizing Response Transformations in Regression.* Technometrics **43** (2001), 443-449.

[13] R. D. Cook, and S. Weisberg, *Transforming a Response Variable for Linearity.* Biometrika **81** (1994), 731-737.

[14] R. D. Cook, and S. Weisberg, *Applied Regression Including Computing and Graphics.* John Wiley and Sons, Inc., NY. 1999.

[15] R. D. Cook, and S. Weisberg, *Graphs in Statistical Analysis: is the Medium the Message?* The American Statistician **53** (1999), 29-37.

[16] W. K. Fung, X. He, L. Liu, and P. D. Shi, *Dimension Reduction Based on Canonical Correlation.* Statistica Sinica **12** (2002), 1093-1114.

[17] U. Gather, T. Hilker, and C. Becker, *A Note on Outlier Sensitivity of Sliced Inverse Regression.* Statistics **36** (2002), 271-281.

[18] D. M. Hawkins, D. Bradu, and G. V. Kass, *Location of Several Outliers in Multiple Regression Data Using Elemental Sets.* Technometrics **26** (1984), 197-208.

[19] K. C. Li, *Sliced Inverse Regression for Dimension Reduction.* J. of the American Statistical Association **86** (1991), 316-342.

[20] K. C. Li, and N. Duan, *Regression Analysis Under Link Violation.* The Annals of Statistics **17** (1989), 1009-1052.

[21] D. J. Olive, *Applications of Robust Distances for Regression.* Technometrics **44** (2002), 64-71.

[22] P. J. Rousseeuw, and A. M. Leroy, *Robust Regression and Outlier Detection.* John Wiley and Sons, Inc., NY. 1987.

[23] P. J. Rousseeuw, and K. Van Driessen, *A Fast Algorithm for the Minimum Covariance Determinant Estimator.* Technometrics **41** (1999), 212-223.

[24] T. M. Stoker, *Consistent Estimation of Scaled Coefficients.* Econometrica **54** (1986), 1461-1481.

Department of Mathematics, Mailcode 4408, Southern Illinois University, Carbondale, IL 62901-4408, USA
E-mail address: dolive@math.siu.edu