OLS for 1D Regression Models

Jing Chang
Department of Mathematics
College of Saint Mary
Omaha, Nebraska 68106-2377
jchang@CSM.edu

David J. Olive
Department of Mathematics
Southern Illinois University
Carbondale, Illinois 62901-4408
olive@math.siu.edu

Keywords Diagnostics; Outliers; Robust regression; Single index models.

Mathematics Subject Classification Primary 62G08; Secondary 62G35.

Abstract

In a 1D regression, the response variable is independent of the predictors given a single linear combination of the predictors. Theory for ordinary least squares (OLS) is reviewed, and it is shown that much of the OLS output originally meant for multiple linear regression is still relevant for a much wider class of regression models including single index models. Ellipsoidal trimming can be combined with OLS to create outlier resistant methods.

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 model*, Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x} \;\; \text{or} \;\; \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}. \tag{1}$$

Many of the most commonly used regression models are 1D regression models, and the additive error *single index model* has the form

$$Y = m(\boldsymbol{\beta}^T\boldsymbol{x}) + e \tag{2}$$

where e is a zero mean error that is independent of $\boldsymbol{x}$. The multiple linear regression model is the special case where $m(\boldsymbol{\beta}^T\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$:

$$Y = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e. \tag{3}$$

Important theoretical results for the additive error single index model were given by Brillinger (1977, 1983), and Li and Duan (1989) extended these results to single index models of the form $Y = g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, e)$ where g is a bivariate inverse link function.

A key condition for several of the theoretical results is the condition of *linearly related predictors* which holds if $E(\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x})$ is a linear function of $\boldsymbol{\beta}^T\boldsymbol{x}$. This condition holds if $\boldsymbol{x}$ is elliptically contoured (EC) with a nonsingular covariance matrix. Hall and Li (1993) show that the linearity condition often approximately holds in large dimensions even if $\boldsymbol{x}$ is not EC.

Section 2 reviews OLS theory for 1D regression. Section 3 shows ellipsoidal trimming can be used to create outlier resistant 1D methods that can give useful results when the assumption of linearly related predictors is violated, while Section 4 gives examples and simulations.

2. Some OLS Theory

This section reviews OLS theory for 1D regression. Let $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. The population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are $\alpha_{OLS} = E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x})$ and $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$.

Let the data be $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let the $p \times 1$ vector $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$, let $\boldsymbol{X}$ be the $n \times p$ OLS design matrix with ith row $(1, \boldsymbol{x}_i^T)$, and let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$. Then the OLS estimator $\hat{\boldsymbol{\eta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The sample covariance of $\boldsymbol{x}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \quad \text{where the sample mean} \quad \overline{\boldsymbol{x}} = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}} \boldsymbol{x}_{\text{i}}.$$

Similarly, define the sample covariance of $\boldsymbol{x}$ and Y to be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\,\overline{Y}.$$

The first result shows that $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

i) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then

$$\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T \overline{\boldsymbol{x}} \xrightarrow{D} \alpha_{OLS}$$

and

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \ \text{ as } \ \mathrm{n} \to \infty.$$

The following notation will be useful. Many 1D regression models have an error e with $\sigma^2 = \text{Var}(e) = E(e^2)$. Let the population OLS residual $v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T \boldsymbol{x}$ with

$$\tau^2 = E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T \boldsymbol{x})^2] = E(v^2), \tag{4}$$

and let the OLS residual be $r = Y - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}$. Typically the OLS residual r is not estimating the error e and $\tau^2 \neq \sigma^2$, but the following results show that the OLS residual is of great interest for 1D regression models. Assume that a 1D model (1) holds. Then under regularity conditions, results ii) – v) below hold.

ii) Li and Duan (1989): $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$ for some constant c.

iii) Li and Duan (1989) and Chen and Li (1998):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \boldsymbol{C}_{OLS}) \tag{5}$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T \boldsymbol{x})^2 (\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}. \tag{6}$$

iv) Chen and Li (1998): Let $\boldsymbol{A}$ be a known full rank constant $k \times (p-1)$ matrix. If the null hypothesis H_0: $\boldsymbol{A\beta} = \boldsymbol{0}$ is true, then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{A\beta}) = \sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T)$$

and

$$\boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T = \tau^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T. \tag{7}$$

If the multiple linear regression (MLR) model holds or if $E[v^2(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = E(v^2)\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then $\boldsymbol{C}_{OLS} = \tau^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$. If the MLR model holds, $\tau^2 = \sigma^2$. To create test statistics, the estimator

$$\hat{\tau}^2 = \mathrm{MSE} = \frac{1}{\mathrm{n} - \mathrm{p}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{r}_{\mathrm{i}}^2 = \frac{1}{\mathrm{n} - \mathrm{p}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{Y}_{\mathrm{i}} - \hat{\alpha}_{\mathrm{OLS}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}})^2$$

will be useful. Notice that MSE estimates τ^2, not σ^2. The estimator

$$\hat{\boldsymbol{C}}_{OLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} [(Y_i - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}_i)^2 (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T] \right] \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tag{8}$$

can also be useful.

v) Chen and Li (1998): A test statistic for $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ is

$$W_{OLS} = n \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{A}^T [\boldsymbol{A} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \boldsymbol{A}^T]^{-1} \boldsymbol{A} \hat{\boldsymbol{\beta}}_{OLS} / \hat{\tau}^2 \xrightarrow{D} \chi^2_k, \tag{9}$$

the chi–square distribution with k degrees of freedom.

Before presenting the next result, some results from OLS MLR theory are needed. Let the known $k \times p$ constant matrix $\tilde{\boldsymbol{A}} = [\boldsymbol{a} \ \boldsymbol{A}]$ where $\boldsymbol{a}$ is a $k \times 1$ vector, and let $\boldsymbol{c}$ be a known $k \times 1$ constant vector. Following Seber and Lee (2003, pp. 99–106), the usual F statistic for testing $H_0 : \tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{c}$ is

$$F_0 = \frac{(SSE(H_0) - SSE)/k}{SSE/(n-p)} = (\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})^T [\tilde{\boldsymbol{A}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \tilde{\boldsymbol{A}}^T]^{-1} (\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c}) / (k \hat{\tau}^2) \tag{10}$$

where $MSE = \hat{\tau}^2 = SSE/(n-p)$, $SSE = \sum_{i=1}^n r_i^2$ and $SSE(H_0) = \sum_{i=1}^n r_i^2(H_0)$ is the minimum sum of squared residuals subject to the constraint $H_0 : \tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{c}$. Recall that if H_0 is true, the MLR model holds and the errors e_i are iid $N(0, \sigma^2)$, then $F_0 \sim F_{k,n-p}$, the F distribution with k and $n-p$ degrees of freedom. Also recall that if a random variable $Z_n \sim F_{k,n-p}$, then as $n \to \infty$

$$Z_n \xrightarrow{D} \chi^2_k / k. \tag{11}$$

Theorem 2.1 below and (11) suggest that OLS output, originally meant for testing with the MLR model, can also be used for testing with many 1D regression data sets. Let the 1D model $Y \perp\!\!\!\perp \boldsymbol{x} | \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ be written as $Y \perp\!\!\!\perp \boldsymbol{x} | \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$ where the reduced model is $Y \perp\!\!\!\perp \boldsymbol{x} | \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R$ and $\boldsymbol{x}_O$ denotes the terms outside of the reduced model. Notice that OLS ANOVA F test corresponds to H_0: $\boldsymbol{\beta} = \boldsymbol{0}$ and uses $\boldsymbol{A} = \boldsymbol{I}_{p-1}$ where $\boldsymbol{I}_j$ is the $j \times j$ identity matrix. The tests for H_0: $\beta_i = 0$ use $\boldsymbol{A} = (0, ..., 0, 1, 0, ..., 0)$ where the 1 is in the ith position and are equivalent to the OLS t tests. The test H_0: $\boldsymbol{\beta}_O = \boldsymbol{0}$ uses $\boldsymbol{A} = [\boldsymbol{0} \ \boldsymbol{I}_j]$ if $\boldsymbol{\beta}_O$ is a $j \times 1$ vector, and the test statistic (10) can be computed by running OLS on the full model to obtain SSE and on the reduced model to obtain $SSE(R) \equiv SSE(H_0)$. In the theorem below, it is crucial that H_0: $\boldsymbol{A\beta} = \boldsymbol{0}$. Tests for H_0: $\boldsymbol{A\beta} = \boldsymbol{1}$, say, may not be valid.

Theorem 2.1. *Assume that a 1D regression model (1) holds and that Equation (9) holds when $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ is true. Then as $n \to \infty$, the test statistic (10) satisfies*

$$F_0 = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi^2_k/k.$$

To see this, notice that by (9), the result follows if $F_0 = (n-1)W_{OLS}/(kn)$. Let $\tilde{\boldsymbol{A}} = [\boldsymbol{0}\ \boldsymbol{A}]$ so that H_0:$\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{0}$ is equivalent to H_0:$\boldsymbol{A\beta} = \boldsymbol{0}$. Following Seber and Lee (2003, p. 106),

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \tag{12}$$

where the $(p-1) \times (p-1)$ matrix

$$\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1). \tag{13}$$

Using $\tilde{\boldsymbol{A}}$ and (12) in (10) shows that

$$F_0 = (\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS})^T \left[[\boldsymbol{0}\ \boldsymbol{A}] \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{A}^T \end{pmatrix} \right]^{-1} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/(k\hat{\tau}^2),$$

and the result follows from (13) after algebra.

Li and Duan (1989) suggest that OLS F tests are asymptotically valid if $\boldsymbol{x}$ is multivariate normal and if $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$. Freedman (1981), Brillinger (1983) and Chen and Li (1998) also discuss $\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS})$.

The above sufficient conditions are very restrictive, but the following remarks suggest that the OLS F tests for $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ are useful for exploratory purposes under very mild conditions. First some notation will be useful. If $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$, then $Y \perp\!\!\!\perp \boldsymbol{x}|a + c\boldsymbol{\beta}^T\boldsymbol{x}$ where a and $c \neq 0$ are constants. A sufficient predictor $SP = a + c\boldsymbol{\beta}^T\boldsymbol{x}$ and an estimated sufficient predictor $ESP = \hat{a} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. Let a candidate model $\boldsymbol{x}_R$ contain k terms including a constant. Let the full model $\boldsymbol{x}$ have design matrix $\boldsymbol{X}$ and residual sum of squares SSE. Let the corresponding terms for the candidate model be the $n \times k$ design matrix $\boldsymbol{X}_R$ and

SSE(R). If F_R is the test statistic for testing whether the $p-k$ predictor variables $\boldsymbol{x}_O$ not in $\boldsymbol{x}_R$ can be deleted, then the statistic $C_p(R) = (p-k)(F_R - 1) + k$.

Olive and Hawkins (2005), in the context of variable selection, note that if the full 1D model using $\boldsymbol{x}$ is good and $\hat{\boldsymbol{\beta}}$ is good estimator of $d\boldsymbol{\beta}$ for $d \neq 0$, then a candidate model using $\boldsymbol{x}_R$ is worth considering if the correlation $corr(\hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_R, \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ is high. For example, the maximum likelihood estimator would be used as $\hat{\boldsymbol{\beta}}$ for generalized linear models. Suppose the OLS estimator is such that $|corr(\hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}, \hat{\boldsymbol{\beta}}^T \boldsymbol{x})| = |corr(\text{OLS ESP}, \text{ESP})| \geq 0.95$. Then $corr(\hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_R, \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ should be high if $corr(\hat{\boldsymbol{\beta}}_{OLS,R}^T \boldsymbol{x}_R, \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x})$ is high.

To force high correlation, Olive and Hawkins (2005) showed that the following results are properties of OLS and hold even if the data does not follow a 1D model. Let OLS ESP(R) $= \hat{\alpha}_{OLS,R} + \hat{\boldsymbol{\beta}}_{OLS,R}^T \boldsymbol{x}_R$ be the ESP for the submodel and let $r_{R,i} = Y_i - \hat{\alpha}_{OLS,R} - \hat{\boldsymbol{\beta}}_{OLS,R}^T \boldsymbol{x}_{R,i}$. Let OLS ESP and r denote the corresponding quantities for the full model. Then

$$corr(r, r_R) = \sqrt{\frac{n-p}{C_p(R) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_R + n - p}},$$

and $C_p(R) \leq 2k$ corresponds to

$$corr(r, r_R) \geq \sqrt{1 - \frac{p}{n}}.$$

If $C_p(R) \leq 2k$ and $n \geq 10p$, then $0.9 \leq corr(r, r_R)$, and both $corr(r, r_R) \to 1.0$ and $corr$(OLS ESP, OLS ESP(R)) $\to 1.0$ as $n \to \infty$.

If $|corr(\text{OLS ESP}, \text{ESP})| \geq 0.95$, then often OLS variable selection can be used for the 1D data. If the F_R statistic is large, then OLS ESP for the full and reduced model will not have high absolute correlation, suggesting that the reduced model is not good. A screen for "good" candidate submodels $\boldsymbol{x}_R$ is $C_p(R) \leq \min(2k, p)$. The p-values from OLS output are often a useful benchmark. To see this, suppose that $n > 5p$ and first consider the model R_j that deletes the predictor x_j. Then the model has $k = p - 1$ predictors including the constant, and the test statistic is t_j where $t_j^2 = F_{R_j}$. It can be shown that $C_p(R_j) = C_p(R_{full}) + (t_j^2 - 2) = p + (t_j^2 - 2)$ where R_{full} is the full model. Using the C_p screen suggests that the predictor x_j can probably be deleted if $|t_j| < \sqrt{2} \approx 1.414$.

More generally, it can be shown that $C_p(R) \leq 2k$ iff $F_R \leq p/(p-k)$. Now k is the number of terms in the model including a constant while $p-k$ is the number of terms set to 0. As $k \rightarrow 0$, the OLS F test will reject H_0: $\boldsymbol{A\beta} = \boldsymbol{0}$ (ie, say that the full model should be used instead of the submodel R) unless F_R is not much larger than 1.

3. Resistant 1D methods

Olive (2002) showed that ellipsoidal trimming can be used to create outlier resistant 1D methods that can give useful results when the assumption of linearly related predictors is violated. To perform ellipsoidal trimming, a robust estimator of multivariate location and dispersion $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances $D_i^2 \equiv D_i^2(T, \boldsymbol{C}) = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x}_i - T)$ 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 $\{\boldsymbol{x} : (\boldsymbol{x}-T)^T \boldsymbol{C}^{-1}(\boldsymbol{x}-T) \leq D_{(j)}^2\}$. The ith case $(Y_i, \boldsymbol{x}_i)$ is trimmed if $D_i > D_{(j)}$. For example, if $j \approx 0.9n$, then about M% = 10% of the cases are trimmed, and a 1D method can be computed from the cases $(Y_M, \boldsymbol{x}_M)$ that remain. We used the Olive (2004) MBA estimator $(T_{MBA}, \boldsymbol{C}_{MBA})$ for $(T, \boldsymbol{C})$.

Several authors have noted that applying 1D methods to a subset $(Y_M, \boldsymbol{x}_M)$ of the data with the $\boldsymbol{x}_M$ distribution closer to being elliptically contoured is an effective method for making 1D methods such as OLS resistant to the presence of strong nonlinearities. See Li and Duan (1989, p. 1011), Brillinger (1991), Cook (1994, p. 188; 1998, p. 152), Cook and Nachtsheim (1994) and Li, Cook and Nachtsheim (2004).

The choice of M is important, and the Rousseeuw and Van Driessen (1999) DD plot of classical Mahalanobis distances MD_i vs MBA distances RD_i can be used to choose M. The MD_i use $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})$. Olive (2002) shows that the plotted points in the DD plot will follow the identity line with zero intercept and unit slope if the predictor distribution is multivariate normal (MVN), and will follow a line with zero intercept but non–unit slope if the distribution is EC (with nonsingular covariance matrix) but not MVN. Delete M% of the cases with the largest MBA distances so that the remaining cases follow the identity line (or some line through the origin) closely. Let $(Y_{Mi}, \boldsymbol{x}_{Mi})$ denote the data remaining after trimming where $i = 1, ..., n_M$. Then apply OLS on these n_M cases.

As long as M is chosen only using the predictors, OLS theory will apply if the data $(Y_M, \boldsymbol{x}_M)$ satisfies the regularity conditions. Let $\phi_M = \lim_{n\to\infty} n/n_M$, let c_M be a constant and let $\hat{\boldsymbol{\beta}}_M$ denote the OLS estimator applied to $(Y_{Mi}, \boldsymbol{x}_{Mi})$ with

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_M - c_M\boldsymbol{\beta}) = \frac{\sqrt{n}}{\sqrt{n_M}}\sqrt{n_M}(\hat{\boldsymbol{\beta}}_M - c_M\boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \phi_M \boldsymbol{C}_M). \tag{14}$$

If $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ is true and $\hat{\boldsymbol{C}}_M$ is a consistent estimator of $\boldsymbol{C}_M$, then from (9)

$$W_M = n_M \hat{\boldsymbol{\beta}}_M^T \boldsymbol{A}^T[\boldsymbol{A}\hat{\boldsymbol{C}}_M\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_M/\hat{\tau}_M^2 \xrightarrow{D} \chi_k^2.$$

For example, if the MLR model holds and the errors are iid $N(0, \sigma^2)$, then the OLS estimator

$$\hat{\boldsymbol{\eta}}_M = (\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T\boldsymbol{Y}_M \sim N_p(\boldsymbol{\eta}, \sigma^2(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}).$$

A tradeoff is that low amounts of trimming may not work while large amounts of trimming may be inefficient. For 1D models, Olive (2002) suggested plotting $\hat{\boldsymbol{\beta}}_M^T\boldsymbol{x}$ versus Y for $M = 0, 10, ..., 90$ and choosing M_{TV} such that the plot (called a trimmed view) has a smooth mean function and the smallest variance function. Notice that all n cases are used in the plot. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}}_M - c_M\boldsymbol{\beta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{C}_M)$ for $M = 0, 10, ..., 90$. Then $\hat{\boldsymbol{\beta}}_{M,TV}$ is $\sqrt{n}$ consistent if $c_M \equiv c_0$, e.g., for MLR $c_M \equiv 1$. But if $\hat{\boldsymbol{\beta}}_{M,TV}$ oscillates between $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_{10}$, then $\hat{\boldsymbol{\beta}}_{M,TV}$ need not be asymptotically normal. If there is oscillation and the c_M are not equal, then $\hat{\boldsymbol{\beta}}_{M,TV}$ is inconsistent.

Adaptive trimming can be used to obtain an asymptotically normal estimator that may avoid large efficiency losses. First, choose an initial amount of trimming M_I by using, e.g., the DD plot or trimmed views. Next compute $|corr(\hat{\boldsymbol{\beta}}_M^T\boldsymbol{x}, \hat{\boldsymbol{\beta}}_{M_I}^T\boldsymbol{x})|$ for $M = 0, 10, ..., 90$ and find the smallest value $M_A \le M_I$ such that the absolute correlation is greater than 0.95. If no such value exists, then use $M_A = M_I$. The resulting adaptive trimming estimator is asymptotically equivalent to the estimator that uses 0% trimming if $\hat{\boldsymbol{\beta}}_0$ is a consistent estimator of $c_0\boldsymbol{\beta}$ and if $\hat{\boldsymbol{\beta}}_{M_I}$ is a consistent estimator of $c_{M_I}\boldsymbol{\beta}$ for $c_0 \neq 0$ and $c_{M_I} \neq 0$.

4. Examples and Simulations

Example 4.1. The Buxton (1920) data consists of measurements taken on 87 men. Let *height* be the response. Figure 1a shows the DD plot made from the four predictors *head* *length, nasal height, bigonal breadth,* and *cephalic index.* The five massive outliers correspond to head lengths that were recorded to be around 5 feet. Figure 1b shows that after deleting these points, the predictor distribution is much closer to a multivariate normal distribution. Now 1D methods can be used to investigate the regression.

An estimated sufficient summary plot or response plot is a plot of an ESP versus the response Y, often with the estimated mean function added as a visual aid. A trimmed view is also a response plot. If $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $c\boldsymbol{\beta}$ for $c \neq 0$, then the response plot is often useful for visualizing the 1D regression. See Cook (1998, p. 10).

The following example and Tables 1 and 2 show that ellipsoidal trimming can be useful for 1D regression when $\boldsymbol{x}$ is not EC. There is a myth that transforming predictors is free, but using a log transformation for the example below will destroy the 1D structure.

Example 4.2. An artificial data set was generated with $Y = (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})^3 + e$ where $n = 100$, $\alpha = 0, \boldsymbol{\beta} = (1, 2, 3)^T, e \sim N(0, 1)$ and $x_i \sim lognormal(0, 1)$ for $i = 1, 2, 3$ where the x_i are iid. Figure 2 shows the trimmed views for M = 0, 10, 30 and 90. Table 1 shows the values of $\hat{\boldsymbol{\beta}}_M$. Notice that the 30% and 90% trimmed views capture the cubic function much better then the OLS = 0% trimmed view. Notice that $\hat{\boldsymbol{\beta}}_{30} \approx 205\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_{90} \approx 86\boldsymbol{\beta}$.

In a small simulation, the clean data $Y = (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})^3 + e$ where $n = 1000$, $\alpha = 1$, $\boldsymbol{\beta} = (1, 0, 0, 0)^T$, $e \sim N(0, 1)$ and $\boldsymbol{x} \sim N_4(\boldsymbol{0}, \boldsymbol{I}_4)$. The outlier percentage γ was either 0% or 49%. The 2 clusters of outliers were about the same size with $Y \sim N(0, 1)$ and $\boldsymbol{x} \sim N_4(\pm 10(1, 1, 1, 1)^T, \boldsymbol{I}_4)$. Table 2 records the averages of $\hat{\beta}_i$ over 100 runs where OLS used $M = 0$ or $M = 50\%$ trimming. When outliers were present, the average of $\hat{\boldsymbol{\beta}}_{50} \approx c(1, 0, 0, 0)^T$.

The following simulation study is extracted from Chang (2006) who used eight types of predictor distributions: d1) $\boldsymbol{x} \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1})$, d2) $\boldsymbol{x} \sim 0.6N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1}) + 0.4N_{p-1}(\boldsymbol{0}, 25\boldsymbol{I}_{p-1})$, d3) $\boldsymbol{x} \sim 0.4N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1}) + 0.6N_{p-1}(\boldsymbol{0}, 25\boldsymbol{I}_{p-1})$, d4) $\boldsymbol{x} \sim 0.9N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1}) + 0.1N_{p-1}(\boldsymbol{0}, 25\boldsymbol{I}_{p-1})$, d5) $\boldsymbol{x} \sim LN(\boldsymbol{0}, \boldsymbol{I})$ where the marginals are iid lognormal(0,1), d6) $\boldsymbol{x} \sim MVT_{p-1}(3)$, d7) $\boldsymbol{x} \sim MVT_{p-1}(5)$ and d8) $\boldsymbol{x} \sim MVT_{p-1}(19)$. Here $\boldsymbol{x}$ has a multivariate t distribution $\boldsymbol{x}_i \sim MVT_{p-1}(\nu)$ if $\boldsymbol{x}_i = \boldsymbol{z}_i / \sqrt{W_i/\nu}$ where $\boldsymbol{z}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1})$ is independent of the chi–square random variable $W_i \sim \chi^2_\nu$. Of the eight distributions, only d5) is not elliptically contoured. The MVT distribution gets closer to the MVN distribution d1) as $\nu \to \infty$. The MVT distribution has first moments for $\nu \geq 3$ and second moments for $\nu \geq 5$. See Johnson and Kotz (1972, pp. 134-135). All simulations used 1000 runs.

The simulations for single index models used $\alpha = 1$. Let the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Then the seven models considered were m1) $Y = SP + e$, m2) $Y = (SP)^2 + e$, m3) $Y = \exp(SP) + e$, m4) $Y = (SP)^3 + e$, m5) $Y = \sin(SP)/SP + 0.01e$, m6) $Y = SP + \sin(SP) + 0.1e$ and m7) $Y = \sqrt{|SP|} + 0.1e$ where $e \sim N(0, 1)$. Models m2), m3) and m4) can result in large $|Y|$ values which can cause numerical difficulties for OLS if $\boldsymbol{x}$ is heavy tailed.

For single index models with EC $\boldsymbol{x}$, OLS can fail if m is symmetric about the median θ of the distribution of $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. If m is symmetric about a, then OLS may become effective as $|\theta - a|$ gets large. This fact is often overlooked in the literature and is demonstrated by models m7), m5) and m2) where $Y = (SP)^2 + e$ with $\theta = \alpha = 1$. OLS has trouble with $Y = (SP - a)^2 + e$ as a gets close to $\theta = 1$ for the EC distributions. The type of symmetry where OLS fails is easily simulated, but may not occur often in practice.

First, coefficient estimation was examined with $\boldsymbol{\beta} = (1, 1, 1, 1)^T$, and the sample standard deviation (SD) of each entry $\hat{\beta}_{Mi,j}$ of $\hat{\boldsymbol{\beta}}_{M,j}$ was computed for $i = 1, 2, 3, 4$ with $j = 1, ..., 1000$. For each of the 1000 runs, the Chen and Li (1998) formula $SE_{cl}(\hat{\beta}_{Mi}) = \sqrt{n_M^{-1} (\hat{\boldsymbol{C}}_M)_{ii}}$ was computed where $\hat{\boldsymbol{C}}_M$ is the estimate (8) applied to $(Y_M, \boldsymbol{x}_M)$. The average of $\hat{\boldsymbol{\beta}}_M$ and of $\sqrt{n} SE_{cl}$ were recorded as well as $\sqrt{n} SD$ of $\hat{\beta}_{Mi,j}$ under the labels $\overline{\boldsymbol{\beta}}_M$, $\sqrt{n}\ \overline{SE}_{cl}$ and $\sqrt{n} SD$. Under regularity,

$$\sqrt{n}\ \overline{SE}_{cl} \approx \sqrt{n} SD \approx \sqrt{\frac{1}{1 - \frac{M}{100}} \ \mathrm{diag}(\boldsymbol{C}_{\mathrm{M}})}$$

where $\boldsymbol{C}_M$ is (6) applied to $(Y_M, \boldsymbol{x}_M)$.

For MVN $\boldsymbol{x}$, MLR and 0% trimming, all three recorded quantities were near (1,1,1,1) for $n = 60, 500$, and 1000. For 90% trimming and $n = 1000$, the results were $\overline{\boldsymbol{\beta}}_{90} = (1.00, 1.00, 1.01, 0.99)$, $\sqrt{n}\ \overline{SE}_{cl} = (7.56, 7.61, 7.60, 7.54)$ and $\sqrt{n} SD = (7.81, 8.02, 7.76, 7.59)$, suggesting that $\hat{\boldsymbol{\beta}}_{90}$ is asymptotically normal but inefficient.

For other distributions, Chang (2006) recorded results for 0 and 10% trimming as well as a "good" trimming value M_B. Results are "good" if all of the entries of both $\overline{\boldsymbol{\beta}}_{M_B}$ and $\sqrt{n}\ \overline{SE}_{cl}$ were approximately equal and if $\sqrt{n}\ \overline{SE}_{cl}$ was close to $\sqrt{n}SD$. The results were good for MVN $\boldsymbol{x}$ and all seven models, and the results were similar for $n = 500$ and $n = 1000$. The results were good for models m1 and m5 for all eight distributions. Model m6 was good for 0% trimming except for distribution d5, and model m7 was good for 0% trimming except for distributions d5, d6 and d7. Trimming usually helped for models m2, m3 and m4 for distributions d5 - d8. For $n = 500$, Table 3 shows that $\hat{\boldsymbol{\beta}}_M$ estimates $c_M\boldsymbol{\beta}$ and the average of the Chen and Li (1998) SE is often close to the simulated SD.

Next testing with nominal level 0.05 was considered. Let F_M denote the OLS statistic (10) applied to the n_M cases $(Y_M, \boldsymbol{x}_M)$ that remained after trimming. H_0 was rejected if $F_M > F_{k,n_M-p}(0.95)$. Let $\hat{p}$ be the proportion of runs where H_0 was rejected. Since 1000 runs were used, the count $1000\hat{p} \sim$ binomial$(1000,\ 1-\delta_n)$ where $1-\delta_n$ converges to the true large sample level $1-\delta$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/1000} \approx 0.0069$ for $p = 0.05$. An observed coverage $\hat{p} \in (0.03, 0.07)$ suggests that there is no reason to doubt that the true level is 0.05.

Let $Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e$. If $Y \perp\!\!\!\perp \boldsymbol{x}$, this reduces to $Y = m(\alpha) + e = c_\alpha + e$. For the corresponding test $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{0}$, the OLS F statistic (10) is invariant with respect to a constant. Hence this test is interesting because the results do not depend on the model (2), but only on the distribution of $\boldsymbol{x}$ and the distribution of e. Since $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$, power can be good if $c \neq 0$. The OLS test is equivalent to the ANOVA F test from MLR of Y on $\boldsymbol{x}$. Under H_0, the test should perform well provided that the design matrix is nonsingular and the error distribution and sample size are such that the central limit theorem holds. For the simulated data with $\boldsymbol{\beta} = \boldsymbol{0}$, the model is linear and normal, and the exact OLS level is 0.05 for $n > p$. Table 4 illustrates this claim for $n = 100$ and $n = 500$.

Next the test $H_0 : \beta_2 = 0$ was considered. The OLS test is equivalent to the t test from MLR of Y on $\boldsymbol{x}$. The true model used $\alpha = 1$ and $\boldsymbol{\beta} = (1, 0, 1, 1)^T$. To simulate adaptive trimming, $|corr(\hat{\boldsymbol{\beta}}_M^T\boldsymbol{x}, \boldsymbol{\beta}^T\boldsymbol{x})|$ was computed for $M = 0, 10, ..., 90$ and the initial trimming proportion M_I maximized this correlation. This process should be similar to choosing the best trimmed view by examining 10 plots. The rejection proportions were recorded for $M = 0, ..., 90$ and for adaptive trimming. Chang (2006) used the seven models, eight distributions and sample sizes $n = 60, 150$, and 500.

The test that used adaptive trimming had proportions ≤ 0.072 except for model m4 with distributions d2, d3, d4, d6, d7 and d8; m2 with d4, d6 and d7 for n = 500 and d6 with n = 150; m6 with d4 and n = 60, 150; m5 with d7 and n = 500 and m7 with d7 and n = 500. With the exception of m4, if the adaptive $\hat{p} > 0.072$, then 0% trimming had a rejection proportion near 0.1. Occasionally adaptive trimming was conservative with $\hat{p} < 0.03$. The 0% trimming worked well for m1 and m6 for all eight distributions and for d1 and d5 for all seven models. Models m2 and m3 usually benefited from adaptive trimming. For distribution d1, the adaptive and 0% trimming methods had identical $\hat{p}$ for $n = 500$ except for m3 where the values were 0.038 and 0.042. Table 5 used $n = 150$ and supports the claim that the adaptive trimming estimator can be asymptotically equivalent to OLS (0% trimming) and that trimming can greatly improve the type I error.

There are many competitors to OLS for 1D regression including sliced inverse regression (SIR) Li (1991), (residual based) principal Hessian directions (PHD) Li (1992), and sliced average variance estimation (SAVE) Cook and Weisberg (1991). These three methods can be computed in `R` using the Weisberg (2002) `dr` library. Chang (2006) applied ellipsoidal trimming and adaptive trimming to SIR, and includes much more extensive simulation results. The collection of functions `rpack`, available from (www.math.siu.edu/olive/rpack.txt) contains `R` functions. The function *drsim5* can be used to simulate OLS tests while *drsim6* can be used to simulate tests based on adaptive and 0% trimming. Power of the OLS tests can be examined by modifying the data so that the null hypothesis does not hold.

5. Conclusions

For 1D regression models, suppose that $|corr(\hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}, \hat{\boldsymbol{\beta}}^T \boldsymbol{x})| \geq 0.95$ where $\hat{\boldsymbol{\beta}}$ is a good estimator of $d\boldsymbol{\beta}$ for $d \neq 0$, or that the 1D regression can be visualized with the OLS response plot. For example, the plotted points cluster tightly about the estimated mean function. Then OLS should be a useful 1D estimator and output originally meant for MLR is also often useful for 1D regression (1DR) data. In particular, i) $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $\boldsymbol{\beta}$ for MLR and $c\boldsymbol{\beta}$ for 1DR. ii) The F test statistics tend to have a χ^2_k/k limiting distribution for MLR, and the $F_{k,n-p}$ cutoffs tend to be useful for exploratory purposes for 1DR. iii) Variable selection with the C_p statistic is effective. iv) The MSE estimates σ^2 for MLR and τ^2 for 1DR. v) The OLS response plot is a very effective tool for visualizing the regression and outlier detection. The estimated mean function for MLR is the unit slope line through the origin, but tends to be nonlinear for 1DR. vi) Resistant $\sqrt{n}$ consistent estimators based on OLS and ellipsoidal trimming exist for both MLR and 1DR. vii) Cook's distance is a useful influence diagnostic.

To see vii) for 1DR, notice that the ith Cook's distance

$$\text{CD}_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} = \frac{\|ESP(i) - ESP\|^2}{(p+1)MSE}$$

where $ESP(i) = \boldsymbol{X}^T\hat{\boldsymbol{\eta}}_{(i)}$ and $\hat{\boldsymbol{\eta}}_{(i)}$ is computed without the ith case, and the estimated sufficient predictor $ESP = \boldsymbol{X}^T\hat{\boldsymbol{\eta}}$ estimates $\alpha_{OLS} + c\,\boldsymbol{\beta}^T\boldsymbol{x}_j$ for some constant c and $j = 1, ..., n$. Thus Cook's distances give useful information on cases that influence the OLS ESP.

Fast exploratory analysis with OLS can be used to complement alternative 1D methods, especially if tests and variable selection for the 1D method are slow or unavailable from the software. Tests developed for parametric models such as the deviance tests for generalized linear models will often have more power than the "model free" OLS tests. Simonoff and Tsai (2002) suggest tests for single index models.

References for methods for single index models can be found in Patilea (2007), Kong and Xia (2007) and Hristache, Juditsky and Spokoiny (2001). Xia, Tong, Li, and Zhu (2002) describe additional methods, including the the minimum average conditional variance and refined minimum average conditional variance estimators (MAVE and rMAVE). `Matlab` implementations of several of these methods, including rMAVE, are available from (www.stat.nus.edu.sg/∼staxyc/). Xia (2006) tailors rMAVE to single index models (2), and shows with simulations and theory that rMAVE is an attractive method for obtaining an estimated sufficient predictor.

The DD plot should be used to detect outliers and influential cases for regressions with continuous predictors. The DD plot is also a diagnostic for the linearity condition since for

EC data the plotted points will follow a line through the origin, and for MVN data the plotted points will follow the identity line. In the case of no outliers, power transformations may be used to remove nonlinearities from the predictors and to transform the predictor distribution towards a MVN distribution. A DD plot and scatterplot matrix (when p is not too large) may then be useful tools for determining the success of such transformations.

Ellipsoidal trimming can be used to make many 1D methods resistant to $\boldsymbol{x}$ outliers. The response plot of $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus Y is crucial for visualizing the regression $Y|\boldsymbol{x}$ (where $\hat{\alpha} \equiv 0$ may be used). Trimming combined with the response plot makes the 1D methods resistant to Y and $\boldsymbol{x}$ outliers.

Acknowledgments: The authors thank the editor and referee for their comments that improved this article. This research was partially supported by the National Science Foundation grant DMS 0600933.

References

Brillinger, D.R. (1977). The identification of a particular nonlinear time series. *Biometrika* 64:509-515.

Brillinger, D.R. (1983). A generalized linear model with "Gaussian" regressor variables. In: Bickel, P.J., Doksum, K.A., Hodges, J.L. eds., *A Festschrift for Erich L. Lehmann.* Pacific Grove, CA: Wadsworth, 97-114.

Brillinger, D.R. (1991). Comment on 'Sliced inverse regression for dimension reduction' by K.C. Li. *J. Amer. Statist. Assoc.* 86:333.

Buxton, L.H.D. (1920). The anthropology of Cyprus. *J. Royal Anthropological Institute Great Britain Ireland* 50:183-235.

Chang, J. (2006). *Resistant Dimension Reduction.* Ph.D. Thesis, Southern Illinois University, online at (www.math.siu.edu/olive/sjingth.pdf).

Chen, C.H., Li, K.C. (1998). Can SIR be as popular as multiple linear regression? *Statistica*

Sinica 8:289-316.

Cook, R.D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* 89:177-189.

Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regression Through Graphics.* New York, Wiley.

Cook, R.D., Weisberg, S. (1991). Comment on 'Sliced inverse regression for dimension reduction' by K.C. Li. *J. Amer. Statist. Assoc.* 86:328-332.

Cook, R.D., Nachtsheim, C.J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *J. Amer. Statist. Assoc.* 89:592-599.

Freedman, D.A. (1981). Bootstrapping regression models. *Ann. Statist.* 9:1218-1228.

Hall, P., Li, K.C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* 21:867-889.

Hristache, M., Juditsky, A., Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* 29:595-623.

Johnson, N.L., Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* New York: Wiley.

Kong, E., Xia, Y. (2007). Variable selection for the single-index model. *Biometrika* 94:217-229.

Li, K.C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* 86:316-342.

Li, K.C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.* 87:1025-1040.

Li, K.C., Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* 17:1009-

1052.

Li, L., Cook, R.D., Nachtsheim, C.J. (2004). Cluster-based estimation for sufficient dimension reduction. *Computat. Statist. Data Analys.* 47:175-193.

Olive, D.J. (2002). Applications of robust distances for regression. *Technometrics* 44:64-71.

Olive, D.J. (2004). A resistant estimator of multivariate location and dispersion. *Computat. Statist. Data Analys.* 46:99-102.

Olive, D.J., Hawkins, D.M. (2005). Variable selection for 1D regression models. *Technometrics* 47:43-50.

Patilea, V. (2007). Semiparametric regression models with applications to scoring: a review. *Commun. Statist. Theor. Meth.* 36:2641-2653.

Rousseeuw, P.J., Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41:212-223.

Seber, G.A.F., Lee, A.J. (2003). *Linear Regression Analysis.* second ed., New York: Wiley.

Simonoff, J.S., Tsai, C. (2002). Score tests for the single index model. *Technometrics* 44:142-151.

Weisberg, S. (2002). Dimension reduction regression in R. *J. Statist. Software* 7: online at (www.jstatsoft.org/v017/i01).

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* 22:1112-1137.

Xia, Y., Tong, H., Li, W.K., Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. Royal Statist. Soc. B* 64:363-410.

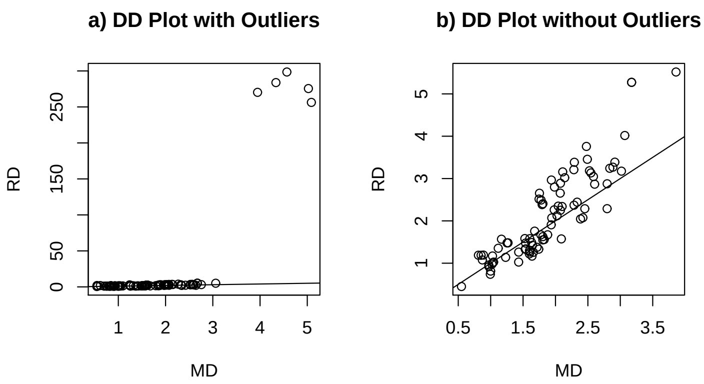

Figure 1: DD Plots for Buxton Data

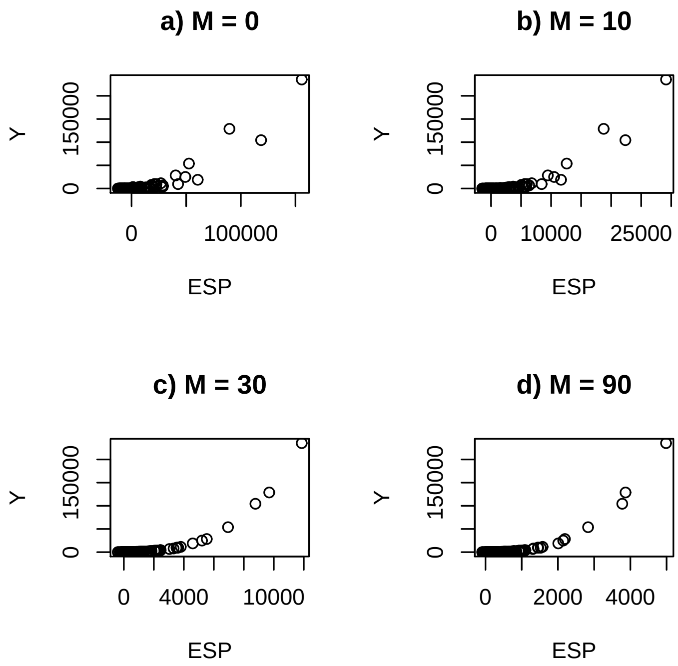

Figure 2: Trimmed Views

Table 1: Trimming with Non-EC Predictors, $\boldsymbol{\beta} = c(1, 2, 3)^T$

M	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0	346.034	3394.260	9000.226
10	292.575	731.751	1616.625
30	191.516	421.577	616.201
90	86.024	160.877	258.987

Table 2: Trimming with Outlier Percentage = γ, $\boldsymbol{\beta} = c(1,0,0,0)^T$

γ	M	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
0	0	5.974	.0083	$-.0221$	.0008
0	50	4.098	.0166	.0017	$-.0016$
49	0	2.269	$-.7509$	$-.7390$	$-.7625$
49	50	5.647	.0305	.0011	.0053

Table 3: OLS Coefficient Estimation with Trimming

m	$\boldsymbol{x}$	M	$\overline{\boldsymbol{\beta}}_M$	$\sqrt{n}\,\overline{SE}_{cl}$	$\sqrt{n}\,SD$
m2	d1	0	2.00,2.01,2.00,2.00	7.81,7.79,7.76,7.80	7.87,8.00,8.02,7.88
m3	d2	50	9.06,9.05,9.04,9.08	37.56,37.00,37.31,37.41	55.35,54.02,53.35,55.03
m4	d3	0	291.9,294.0,293.7,292.1	859.7,866.6,877.9,850.8	933.0,957.9,964.9,957.2
m5	d4	0	$-.03, -.03, -.03, -.03$	.30,.30,.30,.30	.31,.32,.33,.31
m6	d5	0	1.04,1.04,1.04,1.04	.36,.36,.37,.37	.41,.42,.42,.40
m7	d6	10	.11,.11,.11,.11	.58,.57,.57,.57	.60,.58,.62,.61

Table 4: Rejection Proportions for H_0: $\boldsymbol{\beta} = \mathbf{0}$

$\boldsymbol{x}$	n	$\hat{p}$	n	$\hat{p}$
d1	100	0.041	500	0.050
d2	100	0.050	500	0.045
d3	100	0.047	500	0.050
d4	100	0.045	500	0.048
d5	100	0.055	500	0.061
d6	100	0.042	500	0.036
d7	100	0.054	500	0.047
d8	100	0.044	500	0.060

Table 5: Rejection Proportions for H_0: $\beta_2 = 0$

m	$\boldsymbol{x}$	90	80	70	60	50	40	30	20	10	0	ADAP
1	1	.065	.073	.061	.056	.062	.051	.046	.050	.044	.043	.043
5	1	.025	.017	.019	.023	.019	.019	.020	.022	.027	.037	.029
2	2	.045	.033	.023	.024	.026	.070	.183	.182	.142	.166	.040
4	3	.044	.032	.027	.058	.096	.081	.071	.057	.062	.123	.120
6	4	.040	.023	.026	.024	.030	.032	.028	.044	.051	.088	.088
7	5	.056	.053	.058	.058	.053	.054	.046	.044	.051	.037	.037
3	6	.041	.030	.021	.024	.019	.025	.025	.034	.080	.374	.036
6	7	.041	.032	.027	.032	.023	.041	.047	.053	.052	.055	.055