1D Regression

David J. Olive [*]

Southern Illinois University

August 27, 2004

Abstract

Regression is the study of the conditional distribution of the response Y given the vector of predictors $\boldsymbol{x}$. In a 1D regression, Y is independent of $\boldsymbol{x}$ given a single linear combination $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ of the predictors. Special cases of 1D regression include multiple linear regression, logistic regression, generalized linear models and single index models. An estimated sufficient summary plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i can be used to study the conditional distribution of Y given $\boldsymbol{x}$, and should be made for any 1D regression analysis.

KEY WORDS: Generalized Linear Models; Outliers; Regression Graphics; Single Index Models; Sliced Inverse Regression.

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

$$(1.1) \qquad\qquad Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}.$$

See Cook and Weisberg (1999a, pp. 414-415).

    If the 1D regression model holds, then $Y \perp\!\!\!\perp \boldsymbol{x} | a + c\boldsymbol{\beta}^T \boldsymbol{x}$ for any constants a and $c \neq 0$. The quantity $a + c\boldsymbol{\beta}^T \boldsymbol{x}$ is called a *sufficient predictor* (SP), and an *estimated sufficient predictor* (ESP) is $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}$ where $\tilde{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant c. For semiparametric 1D models, the choice $a = 0$ is often used and sometimes the scaling is such that $\tilde{\boldsymbol{\beta}} = (1, \tilde{\beta}_2, ..., \tilde{\beta}_{p-1})^T$. See Horowitz (1998, pp. 14-16).

    Many important regression models satisfy (1.1). The *single index model* has the form

$$(1.2) \qquad\qquad Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e,$$

where e is zero mean error that is independent of $\boldsymbol{x}$. Important theoretical results for the single index model were given by Brillinger (1977, 1983) and Aldrin, Bølviken and Schweder (1993). Li and Duan (1989) extended these results to models of the form

$$(1.3) \qquad\qquad Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e)$$

where g is a bivariate inverse link function.

    Generalized linear models (GLM's), introduced by Nelder and Wedderburn (1972), are also 1D models, and the following three examples are important. *Multiple linear*

regression (MLR) is both a GLM and a single index model with $m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$.

Logistic regression (LR) is a special case of binomial regression, and the LR model states that $Y_1, ..., Y_n$ are independent random variables with

$$(1.4) \quad Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{x}_\text{i})) \quad \text{where} \quad \text{P}(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\text{T} \boldsymbol{x}_\text{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\text{T} \boldsymbol{x}_\text{i})},$$

and the binary logistic regression model has $m_i \equiv 1$ for $i = 1, ..., n$. *Loglinear regression (LLR)* is a special case of Poisson regression, and the LLR model states that $Y_1, ..., Y_n$ are independent random variables with

$$(1.5) \qquad Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i})) \quad \text{where} \quad \mu(\boldsymbol{x}_\text{i}) = \exp(\alpha + \boldsymbol{\beta}^\text{T} \boldsymbol{x}_\text{i}).$$

Another example of (1.1) is the *response transformation model,*

$$(1.6) \qquad Y = t^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e),$$

where t^{-1} is a one to one (typically monotone) function. Hence $t(Y) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e$. Koenker and Geling (2001) note that if Y is the survival time, then many *survival models* including the Cox (1972) *proportional hazards model* are response transformation models.

There are many ways to estimate 1D models, including maximum likelihood for parametric models. The literature for estimating $c\boldsymbol{\beta}$ when model (1.1) holds is growing, and Cook and Li (2002) summarize when competing methods such as ordinary least squares (OLS), sliced inverse regression (SIR), principal Hessian directions (PHD), and sliced average variance estimation (SAVE) can fail. All four methods frequently perform well if there are no strong nonlinearities present in the predictors. Further information about these and related methods can be found, for example, in Brillinger (1977, 1983), Bura and Cook (2001), Chen and Li (1998), Cook (1998ab, 2000, 2003, 2004), Cook and Critchley

3

(2000), Cook and Li (2002), Cook and Weisberg (1991, 1999ab), Fung, He, Liu and Shi (2002), Li (1991, 1992, 2000), Li and Duan (1989) and Yin and Cook (2002, 2003).

In addition to OLS, specialized methods for 1D models with an unknown inverse link function (e.g., models (1.2) and (1.3)) have been developed, and often the focus is on developing asymptotically efficient methods. See the references in Cavanagh and Sherman (1998), Delecroix, Härdle and Hristache (2003), Härdle, Hall and Ichimura (1993), Hristache, Juditsky, Polzehl, and Spokoiny (2001), Stoker (1986), Weisberg and Welsh (1994) and Xia, Tong, Li and Zhu (2002).

Section 2 discusses interpretation of the coefficients $\boldsymbol{\beta}$. Section 3 considers plots for the goodness of fit and lack of fit of the 1D model, while Section 4 considers variable selection. Section 5 suggests resistant methods and Section 6 gives conclusions.

2   Interpretation of Coefficients

The interpretation of the coefficients in a 1D model is nearly the same as the interpretation of the coefficients for multiple linear regression. Denote a model by $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$. Then β_i is the rate of change in the SP associated with a unit increase in x_i when all other predictor variables $x_1, ..., x_{i-1}, x_{i+1}, ..., x_{p-1}$ are held fixed:

$$\beta_i = \frac{\partial\, SP}{\partial x_i} \quad \text{for} \quad i = 1, ..., p-1.$$

The interpretation of β_i changes with the model in two ways. First, the interpretation changes as terms are added and deleted from the SP. Hence the interpretation of β_1 differs for models $SP = \alpha + \beta_1 x_1$ and $SP = \alpha + \beta_1 x_1 + \beta_2 x_2$. Secondly, the interpretation changes as the parametric or semiparametric form of the model changes. For multiple linear

regression, $E(Y|SP) = SP$ and an increase in one unit of x_i increases the conditional expectation by β_i. For binary logistic regression,

$$E(Y|SP) = \rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)},$$

and the change in the conditional expectation associated with a one unit increase in x_i is more complex.

Of course, holding all other variables fixed while changing x_i may not be possible. For example, if $SP = \alpha + \beta_1 x + \beta_2 x^2$, then

$$\frac{d\ SP}{dx} = \beta_1 + 2\beta_2 x.$$

The interpretation also changes if interactions and factors are present. Suppose a factor W is a qualitative random variable that takes on c categories $a_1, ..., a_c$. Then the 1D model will use $c - 1$ indicator variables $W_i = 1$ if $W = a_i$ and $W_i = 0$ otherwise, where one of the levels a_i is omitted, e.g., use $i = 2, ..., c$. Suppose X_1 is quantitative and X_2 is qualitative with 2 levels and $X_2 = 1$ for level a_2 and $X_2 = 0$ for level a_1. Then a first order model with interaction is $SP = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. This model yields two unrelated lines in the sufficient predictor depending on the value of x_2: $SP = \alpha + \beta_2 + (\beta_1 + \beta_3)x_1$ if $x_2 = 1$ and $SP = \alpha + \beta_1 x_1$ if $x_2 = 0$. If $\beta_3 = 0$, then there are two parallel lines: $SP = \alpha + \beta_2 + \beta_1 x_1$ if $x_2 = 1$ and $SP = \alpha + \beta_1 x_1$ if $x_2 = 0$. If $\beta_2 = \beta_3 = 0$, then the two lines are coincident: $SP = \alpha + \beta_1 x_1$ for both values of x_2. If $\beta_2 = 0$, then the two lines have the same intercept: $SP = \alpha + (\beta_1 + \beta_3)x_1$ if $x_2 = 1$ and $SP = \alpha + \beta_1 x_1$ if $x_2 = 0$. In general, as factors have more levels and interactions have more terms, e.g. $x_1 x_2 x_3 x_4$, the interpretation of the model rapidly becomes very complex.

3   Goodness of Fit and Lack of Fit Plots

There is an enormous literature for numerical and graphical diagnostics and tests for the goodness and lack of fit of regression models. See, for example, Agresti and Caffo (2002), Anderson-Sprecher (1994), Cheng and Wu (1994), Cook (1977, 1986), Heckman and Zamar (2000), Jiang (2001), Joglekar, Schuenemeyer and LaRiccia (1989), Kauermann and Tutz (2001), Menard (2000), Pardoe and Cook (2002), Pierce and Schafer (1986), Simonoff (1998), Simonoff and Tsai (2002), Tang (2001), Theus and Lauer (1999) and Xia, Li, Tong and Zhang (2004).

A 1D regression is the study of the conditional distribution $Y|SP$ of the response given the sufficient predictor, and the *estimated sufficient summary plot (ESSP or EY plot)* of the estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i can be used to visualize this conditional distribution. The EY plot can be used as a diagnostic for goodness of fit by adding the estimated parametric mean function and an estimated nonparametric mean function to the plot. The EY plot is a special case of a model checking plot, see Cook and Weisberg (1997, 1999a: ch. 17).

If there is only one predictor x, then a plot of x versus Y is an ESSP. Replacing x by ESP has two major advantages. First, the plot can be made for $p - 1 \geq 1$ and secondly, the possible shapes that the plot can take is greatly reduced. For example, in a plot of x_i versus Y_i, the plotted points will fall about some line with slope β and intercept α if the simple linear regression model holds, but in a plot of $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i = \hat{Y}_i$ versus Y_i, the plotted points will fall about the identity line with unit slope and zero intercept if the multiple linear regression model holds.

Example 3.1. Tremearne (1911) presents an MLR data set of measurements on 115 people of Hausa nationality. We deleted 3 cases because of missing values and used *height* as the response variable Y. The five predictor variables were *height when sitting, height when kneeling, head length, nasal breadth,* and *span*. Figure 1 presents the ESSP, also called a *forward response plot*. Notice that the estimated mean function is the identity line and that the vertical deviation of Y_i from the line is equal to the residual $r_i = Y_i - (\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$. See Chambers, Cleveland, Kleiner and Tukey (1983, p. 280). Points corresponding to cases with Cook's distance $> \min(0.5, 2 * p/n)$ are shown as highlighted squares. Figure 1 also shows the residual plot of the ESP versus the residuals, a widely used diagnostic for lack of fit.

Example 3.2. Figure 2 shows the ESSP for an artificial binary LR data set with the estimated mean function

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid. Divide the ESP into J "slices" each containing approximately n/J cases, and then compute the sample mean = sample proportion of the $Y's$ in each slice and add the resulting step function to the ESS plot. This is done in Figure 2 with $J = 10$ slices. This step function is a simple nonparametric estimator of the mean function $\rho(SP)$, and if the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (1980, 2000, pp. 147–156). For the binary logistic regression model, the residuals do not behave very well, but the Cook (1996) *binary response plot* (not shown) is a useful plot

for examining lack of fit.

Example 3.3. Figure 3 shows the ESSP for an artificial LLR data set with the estimated mean function

$$\hat{\mu}(ESP) = \exp(ESP)$$

added as a visual aid. The lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve) in Figure 3. If the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the LLR model fits the data well.

Simple diagnostic plots for the loglinear regression model can also be made using weighted least squares (WLS). Let $Z_i = Y_i$ if $Y_i > 0$, and let $Z_i = 0.5$ if $Y_i = 0$. Then the minimum chi–square estimator $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$ of the parameters $(\alpha, \boldsymbol{\beta})$ in a LLR model is found from the WLS regression of $\log(Z_i)$ on $\boldsymbol{x}_i$ with weights $w_i = Z_i$. Equivalently, use the OLS regression (without intercept) of $\sqrt{Z_i} \log(Z_i)$ on $\sqrt{Z_i}\boldsymbol{x}_i$. The minimum chi–square estimator tends to be consistent if n is fixed and all n counts Y_i increase to ∞ while the loglinear regression maximum likelihood estimator tends to be consistent if the sample size $n \to \infty$. See Agresti (2002, pp. 611-612) and Powers and Xie (2000, p. 284). However, the two estimators are often close for many data sets. This result and the equivalence of the minimum chi–square estimator to an OLS estimator suggest the following diagnostic plots. Let $(\tilde{\alpha}, \tilde{\boldsymbol{\beta}})$ be an estimator of $(\alpha, \boldsymbol{\beta})$.

For a loglinear regression model, a *weighted forward response plot* is a plot of $\sqrt{Z_i} ESP = \sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ versus $\sqrt{Z_i} \log(Z_i)$. The *weighted residual plot* is a plot of $\sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ versus the "WMLR" residuals $r_{Wi} = \sqrt{Z_i} \log(Z_i) - \sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$.

If the loglinear regression model is appropriate and if the estimators are reasonable, then the plotted points in the weighted forward response plot should follow the identity line. Cases with large WMLR residuals may not be fit very well by the model. Figure 4 shows the diagnostic plots for the artificial data using both the minimum chi–square estimator and the LLR MLE. Although the plots based on the MLE are attractive, more research is needed to determine when such plots are useful for contingency tables.

Example 3.4. Following Brillinger (1977, 1983) and Cook and Weisberg (1999a, 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, 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. For a single index model with unknown mean function m, assume that the lowess curve is a reasonable estimator for m and add both the lowess curve and the step function based on slices to the EY plot (perhaps using the OLS ESP). As a diagnostic for goodness of fit, check that both the plotted points and the step function follow the lowess curve.

Remark 3.1. The ESSP is also a useful visual aid for whether the predictors $\boldsymbol{x}$ are needed in the given model, e.g., for the ANOVA F or deviance test of $H_o : \boldsymbol{\beta} = \boldsymbol{0}$ versus $H_A : \boldsymbol{\beta} \neq \boldsymbol{0}$. For MLR, LLR and the binary LR models, if the predictors are not needed in the model, then $E(Y_i|\boldsymbol{x}_i)$ should be estimated by the sample mean $\overline{Y}$. If the predictors are needed, then $E(Y_i|\boldsymbol{x}_i)$ should be estimated by the appropriate function of the $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$. If it is clear that no horizontal line fits either the data or the estimated nonparametric mean function as well as the estimated mean function (as in Figures 1, 2 and 3), then the predictors are needed. For single index models, if the lowess

curve fits the data and the step function better than any horizontal line, then again the predictors are needed.

4 Variable Selection

Variable selection is the search for a subset of variables that can be deleted without important loss of information. Assume that there exists a subset S of predictor variables such that if $\boldsymbol{x}_S$ is in the 1D model, then none of the other predictors is needed in the model. Write E for these ('extraneous') variables not in S, partitioning $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$. Then

$$(4.1) \qquad SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S.$$

The extraneous terms that can be eliminated given that the subset S is in the model have zero coefficients: $\boldsymbol{\beta}_E = \boldsymbol{0}$.

Now suppose that I is a candidate subset of predictors and that $S \subseteq I$. Then

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I,$$

(if I includes predictors from E, these will have zero coefficients). For any subset I that includes all relevant predictors, the correlation

$$(4.2) \qquad \mathrm{corr}(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \alpha + \boldsymbol{\beta}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = 1.$$

This observation suggests that variable selection for 1D regression models is simple in principle. For each value of $j = 1, 2, ..., p - 1$ nontrivial predictors, keep track of subsets I that provide the largest values of $\mathrm{corr}(\mathrm{ESP}, \mathrm{ESP}(I))$. Any such subset for which the correlation is high is worth closer investigation and consideration. Experience suggests

10

that if corr(ESP,ESP(I)) > 0.95, then the EY plots based on the full model $\boldsymbol{x}$ and the submodel $\boldsymbol{x}_I$ will be nearly identical, visually.

Olive and Hawkins (2005) show that if the 1D ESP and the OLS ESP satisfy

(4.3) $$|\text{corr}(\text{ESP}, \text{OLS ESP})| > 0.95,$$

then existing variable selection algorithms, originally meant for multiple linear regression and based on OLS and the Jones (1946) and Mallows (1973) C_p criterion, can often be used for 1D models. In particular, the Furnival and Wilson (1974) procedure can be used to search all subsets if $p < 30$.

The method is very simple: check that Equation (4.3) holds, perform the OLS regression of Y on $\boldsymbol{x}$, and then perform the OLS variable selection procedure (e.g., forward selection or backward elimination). Assume that the submodel $\boldsymbol{x}_I$ plus a constant has k terms. Then keep track of submodels I with small k that satisfy $C_p(I)$ close to or less than $2k$. The basic idea is that if the correlation corr(r, r_I) of the OLS residuals from the full model and the submodel tends to one, then so does corr(OLS ESP, OLS ESP(I)), and the simple screen $C_p(I) \leq 2k$ corresponds to

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

In the literature, the screen $C_p(I) \leq k$ is often suggested, but in simulations for multiple linear regression, logistic regression and single index models, the true model S satisfied $C_p(S) \leq k$ for about 60% of the simulated data sets, but S satisfied $C_p(S) \leq 2k$ for about 97% of the data sets.

There is a massive literature for variable selection for multiple linear regression, and the literature for 1D models is growing. See, for example, Claeskins and Hjort (2003),

Efron, Hastie, Johnstone and Tibshirani (2004), Fan and Li (2001, 2002), Hastie (1987), Lawless and Singhai (1978), Naik and Tsai (2001), Nordberg (1982), Nott and Leonte (2004) and Tibshirani (1996). For generalized linear models, forward selection and backward elimination based on the AIC criterion are often used. See Agresti (2002, pp. 211-217) or Cook and Weisberg (1999a, pp. 485, 536-538).

5   Resistant Estimation

The presence of strong nonlinearities in the predictors or the presence of outliers can cause 1D regression methods to fail, but the 1D methods often work well if the predictors follow an elliptically contoured distribution. The literature on outlier resistant methods for multiple linear regression is enormous, and the literature for outlier resistant methods of other parametric 1D models is rapidly growing. These methods often use M-estimators, weighted likelihoods or trimmed likelihoods. See, for example, Cantoni and Ronchetti (2001), Choi, Hall and Presnell (2000), Croux and Haesbroeck (2003), Gervini (2005), Heritier and Ronchetti (2004), Luceño (1998), Markatou, Basu and Lindsay (1997), Morgenthaler (1992), Müller and Neykov (2003), Olive (2005) and Rousseeuw and Christmann (2001). Outlier resistant methods for general methods such as SIR are less common, but see Gather, Hilker and Becker (2001, 2002).

Several authors have noted that ellipsoidal trimming is an effective method for making regression graphics methods such as SIR resistant to the presence of strong nonlinearities. See Brillinger (1991), Cook (1998a, p. 152), Cook and Nachtsheim (1994), Heng-Hui (2001), Lexin, Cook and Nachtsheim (2004), and Olive (2002, 2004b).

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

$$(5.1) \qquad D_i^2 = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x}_i - T)$$

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

$$(5.2) \qquad \{\boldsymbol{x} : (\boldsymbol{x} - T)^T C^{-1}(\boldsymbol{x} - T) \le D_{(j)}^2\}.$$

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 resulting ESP is outlier resistant if a resistant estimator $(T, \boldsymbol{C})$ (such as the Olive 2004a estimator) is used.

The following procedure was suggested by Olive (2002, 2004b). First compute $(T, \boldsymbol{C})$ using the *Splus* function `cov.mcd` (see Rousseeuw and Van Driessen 1999). 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 5.1. 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 m is strongly nonlinear. The `cov.mcd` estimator was used for trimming, and Weisberg (2002) was used to produce the SIR, PHD and SAVE ESPs. Figure 5 shows the EY plots for SIR, PHD, SAVE, and OLS. Figure 6 shows that the EY plots based on trimming greatly improved the SIR, SAVE and OLS methods. Replacing the OLS trimmed views by alternative MLR estimators often produced good EY plots, and for single index models, the `lmsreg` estimator often worked the best. Table 1 shows the estimated sufficient predictor coefficients $\hat{\boldsymbol{b}}$ when the sufficient predictor coefficients are $c(1, 2, 3)^T$. Only the SIR, SAVE, OLS and `lmsreg` trimmed views produce estimated sufficient predictors that are highly correlated with the sufficient predictor. For this example, the `lmsreg` trimmed view (not shown) gave the best EY plot.

6 CONCLUSIONS

Heuristically, this paper has shown that OLS output still gives relevant results for the class of the 1D models. $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $c\boldsymbol{\beta}$, OLS variable selection is useful, the partial F test is useful in that $F_I \leq 1$ (which is equivalent to $C_p(I) \leq k$) suggests that the submodel I is good, and the OLS EY plot can be used to visualize the mean function.

The EY plot is used to visualize the conditional distribution of $Y|\boldsymbol{x}$, or, equivalently, of $Y|SP$. This plot should be made for any 1D analysis and emphasizes the goodness of

fit of the 1D model. Although a residual plot of W versus r can be very important, the plot emphasizes lack of fit and is used to visualize the conditional distribution $r|W$ of the residuals given W.

Another application of the EY plot is to choose between k 1D regression models where k is small. 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. Make an EY plot for each of the k competing models and choose the model corresponding to the best plot. Cook and Olive (2001) illustrate such a procedure for response transformations. Similar procedures may be effective for the 1D models given by Carroll and Ruppert (1984), Horowitz (1996), Sarker (1985) and Yeo and Johnson (2000).

7 References

Agresti, A. (2002). *Categorical Data Analysis.* 2nd ed., Wiley, Hoboken, NJ.

Agresti, A. and Caffo, B. (2002). Measures of relative model fit. *Computat. Statist. Data Analys.* **39** 127-136.

Aldrin, M., Bφlviken, E. and Schweder, T. (1993). Projection pursuit regression for moderate non-linearities. *Computat. Statist. Data Analys.* **16** 379-403.

Anderson-Sprecher, R. (1994). Model comparisons and R^2. *Amer. Statist.* **48** 113-117.

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 *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A. and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.

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

Bura, E. and Cook, R.D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *J. Roy. Statist. Soc. Ser. B* **63** 393-410.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96** 1022-1030.

Carroll, R.J. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *J. Amer. Statist. Assoc.* **79** 321-328.

Cavanagh, C. and Sherman, R.P. (1998). Rank estimators for monotonic index models. *J. Econometrics* **84** 351-381.

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

Chen, C.H. and Li, K.C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8** 289-316.

Cheng, K.F. and Wu, J.W. (1994). Testing goodness of fit for a parametric family of link functions. *J. Amer. Statist. Assoc.* **89** 657-664.

Choi, E., Hall, P. and Presnell, B. (2000). Rendering parametric procedures more robust by empirically tilting the model. *Biometrika* **87** 453-465.

Claeskins, G. and Hjort, N.L. (2003). The focused information criterion (with discussion). *J. Amer. Statist. Assoc.* **98** 900-916.

Cook, R.D. (1977). Deletion of influential observations in linear regression. *Technom.* **19** 15-18.

Cook, R.D. (1986). Assessment of local influence. *J. Roy. Statist. Soc. Ser. B* **48** 133-169.

Cook, R.D. (1996). Graphics for regressions with binary response. *J. Amer. Statist. Assoc.* **91** 983-992.

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

Cook, R.D. (1998b). Principal hessian directions revisited. *J. Amer. Statist. Assoc.* **93** 84-100.

Cook, R.D. (2000). SAVE: a method for dimension reduction and graphics in regression. *Commun. Statist. Th. Meth.* **29** 2109-2121.

Cook, R.D. (2003). Dimension reduction and graphical exploration in regression including survival analysis. *Statistics in Medicine* **2** 1399-1413.

Cook, R.D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32** 1062-1092.

Cook, R.D. and Critchley, F. (2000). Identifying outliers and regression mixtures graphically. *J. Amer. Statist. Assoc.* **95** 781-794.

Cook, R.D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** 455-474.

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

Cook, R.D. and Olive, D.J. (2001). A note on visualizing response transformations in

regression. *Technom.* **43** 443-449.

Cook, R.D. and 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. and Weisberg, S. (1997). Graphs for assessing the adequacy of regression models. *J. Amer. Statist. Assoc.* **92** 490-499.

Cook, R.D. and Weisberg, S. (1999a). *Applied Regression Including Computing and Graphics.* Wiley, New York.

Cook, R.D. and Weisberg, S. (1999b). Graphs in statistical analysis: is the medium the message? *Amer. Statist.* **53** 29-37.

Cox, D.R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187-220.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computat. Statist. Data Analys.* **44** 273-295.

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Multiv. Analys.* **86** 213-226.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407-451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazard model and frailty model. *Ann. Statist.* **30** 74-99.

Fung, W.K., He, X., Liu, L. and Shi, P.D. (2002). Dimension reduction based on

canonical correlation. *Statist. Sinica* **12** 1093-1114.

Furnival, G. and Wilson, R. (1974). Regression by leaps and bounds. *Technom.* **16** 499-511.

Gather, U., Hilker, T., and Becker, C. (2001). A robustified version of sliced inverse regression. In *Statistics in Genetics and in the Environmental Sciences*, eds. Fernholtz, T.L., Morgenthaler, S. and Stahel, W., Birkhäuser, Basel, 145-157.

Gather, U., Hilker, T. and Becker, C. (2002). A note on outlier sensitivity of sliced inverse regression. *Statistics* **36** 271-281.

Gervini, D. (2005). Robust adaptive estimators for binary regression models. *J. Statist. Plan. Infer.* to appear.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single index models. *Ann. Statist.* **21** 157-178.

Hastie, T., (1987). A closer look at the deviance. *Amer. Statist.* **41** 16-20.

Heckman, N.E. and Zamar, N.H. (2000). Comparing the shapes of regression functions. *Biometrika* **87** 135-144.

Heng-Hui, L. (2001). A study of sensitivity analysis on the method of principal hessian directions. *Computat. Statist.* **16** 109-130.

Heritier, S. and Ronchetti, E. (2004). Robust binary regression with continuous outcomes. *Canadian J. Statist.* **32** to appear.

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

Horowitz, J.L. (1998). *Semiparametric Methods in Econometrics.* Springer-Verlag, New York.

Hosmer, D.W. and Lemeshow, S. (1980). A goodness of fit test for the multiple logistic regression model. *Commun. Statist.* **A10** 1043-1069.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression.* 2nd ed., Wiley, New York.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29** 1537-1566.

Jiang, J. (2001). A nonstandard small χ^2-test with application to generalized linear model diagnostics. *Statist. Probab. Lett.* **53** 101-109.

Joglekar, G., Schuenemeyer, J.H. and LaRiccia, V. (1989). Lack-of-fit testing when replicates are not available. *Amer. Statist.* **43** 135-143.

Jones, H.L. (1946). Linear regression functions with neglected variables. *J. Amer. Statist. Assoc.* **41** 356-369.

Kauermann, G. and Tutz, G. (2001). Testing generalized linear and semiparametric models against smooth alternatives. *J. Roy. Statist. Soc. Ser. B* **63** 147-166.

Koenker, R. and Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *J. Amer. Statist. Assoc.* **96** 458-468.

Lawless, J.F. and Singhai, K. (1978). Efficient screening of nonnormal regression models. *Biometrics* **34** 318-327.

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

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. (2000). *High Dimensional Data Analysis via the SIR/PHD Approach.* Unpublished manuscript available from (http://www.stat.ucla.edu/ kcli/).

Li, K.C., and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009-1052.

Luceño, A. (1998). Multiple outliers detection through reweighted least deviances. *Computat. Statist. Data Analys.* **26** 313-326.

Mallows, C. (1973). Some comments on C_p. *Technom.* **15** 661-676.

Markatou, M., Basu, A. and Lindsay, B. (1997). Weighted likelihood estimating equations: the discrete case with applications to logistic regression. *J. Statist. Plan. Infer.* **57** 215-232.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *Amer. Statist.* **54** 17-24.

Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika* **79** 747-754.

Müller, C.H. and Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *J. Statist. Plan. Infer.* **116** 503-519.

Naik, P.A. and Tsai, C. (2001). Single-index model selections. *Biometrika* **88** 821-832.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy.*

Statist. Soc. Ser. A **135** 370-380.

Nordberg, L. (1982). On variable selection in generalized linear and related regression models. *Commun. Statist. Th. Meth.* **11** 2427-2449.

Nott, D.J. and Leonte, D. (2004). Sampling schemes for Bayesian variable selection in generalized linear models. *J. Computat. Graph. Statist.* **13** 362-382.

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

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

Olive, D.J. (2004b). Visualizing 1D regression. In *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Series: Statistics for Industry and Technology, Birkhäuser, Basel, Switzerland, 221-233.

Olive, D.J. (2005). Two simple resistant regression estimators. *Computat. Statist. Data Analys.* to appear.

Olive, D.J. and Hawkins, D.M. (2005). Variable selection for 1D regression models. *Technom.* to appear.

Pardoe, I. and Cook, R.D. (2002). A graphical method for assessing the fit of a logistic regression model. *Amer. Statist.* **56** 263-272.

Pierce, D.A. and Schafer, D.W. (1986). Residuals in generalized linear models. *J. Amer. Statist. Assoc.* **81** 977-986.

Powers, D.A. and Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis.* Academic Press, San Diego.

Rousseeuw, P.J. and Christmann, A. (2001). Measuring overlap in binary regression. *Computat. Statist. Data Analys.* **37** 65-75.

Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum co-variance determinant estimator. *Technom.* **41** 212-223.

Sarker, N. (1985). Box-Cox transformation and the problem of heteroscedasticity. *Commun. Statist. Th. Meth.* **14** 363-379.

Simonoff, J.S. (1998). Logistic regression, categorical predictors, and goodness-of-fit: it depends on who you ask. *Amer. Statist.* **52** 10-14.

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

Stoker, T.M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461-1481.

Tang, M.L. (2001). Exact goodness-of-fit test for binary logistic model. *Statist. Sinica* **11** 199-212.

Theus, M. and Lauer, S.R.W. (1999). Visualizing loglinear models. *J. Computat. Graph. Statist.* **8** 396-412.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.

Tremearne, A.J.N. (1911). Notes on some Nigerian tribal marks. *J. Roy. Anthropological Institute of Great Britain and Ireland* **41** 162-178.

Weisberg, S. (2002). Dimension reduction regression in R. *J. Statist. Software* **7** web-page (http://www.jstatsoft.org).

Weisberg, S. and Welsh, A.H. (1994). Adapting for the missing link. *Ann. Statist.* **22** 1674-1700.

Xia, Y.C., Li, W.K., Tong, H. and Zhang , D. (2004). A goodness-of-fit test for single-

index models. *Statist. Sinica* **14** 34-39.

Xia, Y., Tong, H., Li, W.K. and Zhu, L.-X. (2002). An adaptive estimation of dimension
   reduction space (with discussion and rejoinder). *J. Roy. Statist. Soc. Ser. B* **64** 363-
   410.

Yeo, I.K. and Johnson, R. (2000). A new family of power transformations to improve
   normality or symmetry. *Biometrika* **87** 954-959.

Yin X.R. and Cook, R.D. (2002). Dimension reduction for the conditional kth moment
   in regression. *J. Roy. Statist. Soc. Ser. B* **64** 159-175.

Yin, X. and Cook, R.D. (2003). Estimating central subspaces via inverse third moments.
   Biometrika, **90** 113-125.

Table 1: Estimated Sufficient Predictors Coefficients Estimating $c(1, 2, 3)^T$

method	b_1	b_2	b_3
OLS View	0.0032	0.0011	0.0047
90% Trimmed OLS View	0.086	0.182	0.338
SIR View	-0.394	-0.361	-0.845
10% Trimmed SIR VIEW	-0.284	-0.473	-0.834
SAVE View	-1.09	0.870	-0.480
40% Trimmed SAVE VIEW	0.256	0.591	0.765
PHD View	-0.072	-0.029	-0.0097
90% Trimmed PHD VIEW	-0.558	-0.499	-0.664
70% Trimmed LMSREG VIEW	0.143	0.287	0.428
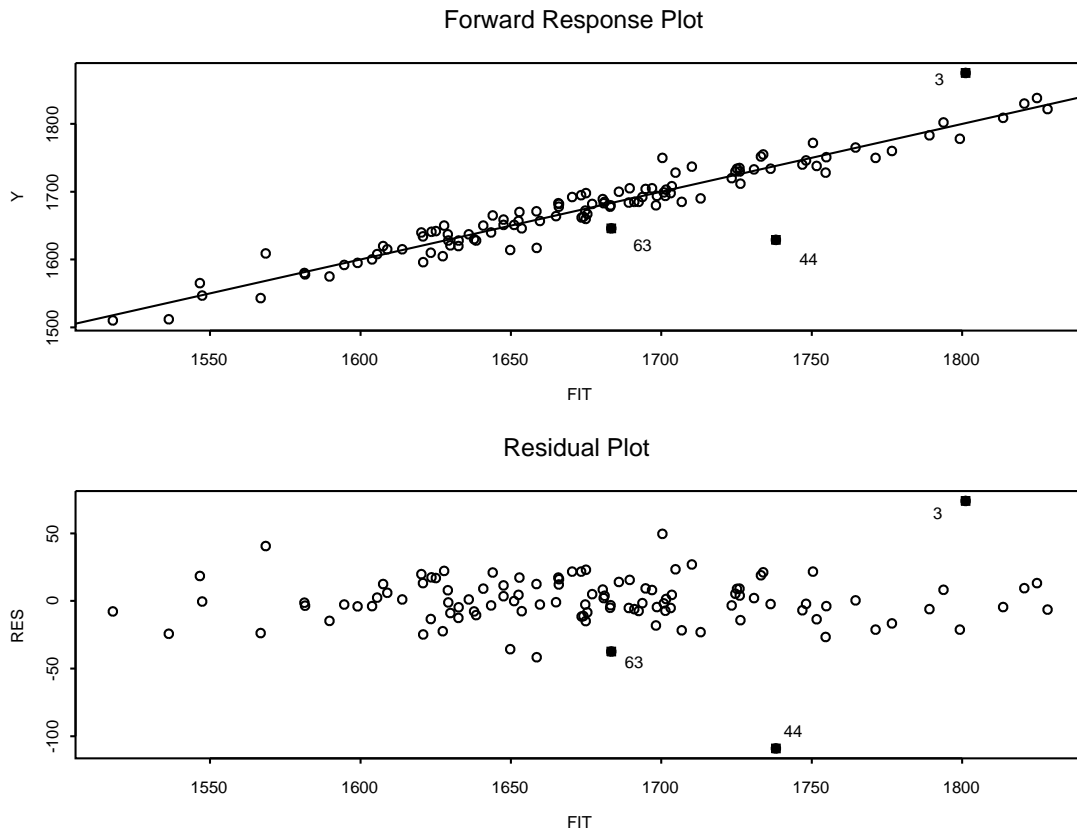
Figure 1: Residual and Forward Response Plots for the Tremearne Data. Highlighted Cases Have Large Cook's Distances.
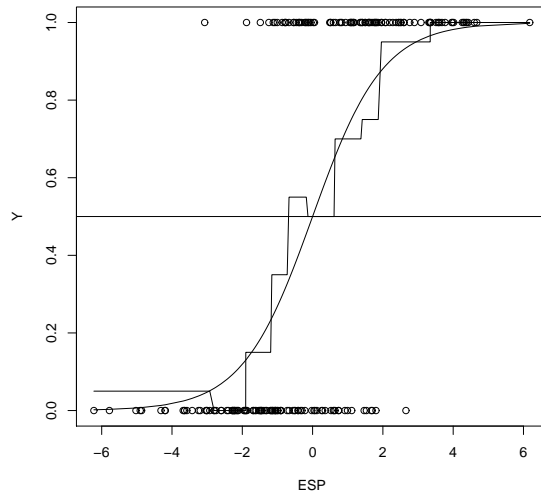
Figure 2: ESS Plot for LR Data
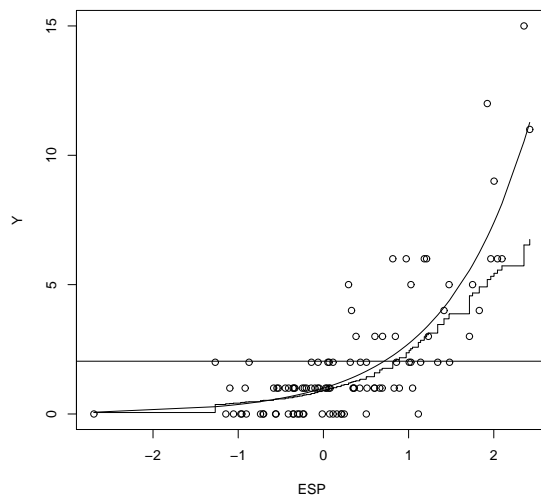
Figure 3: ESSP for Loglinear Regression

27
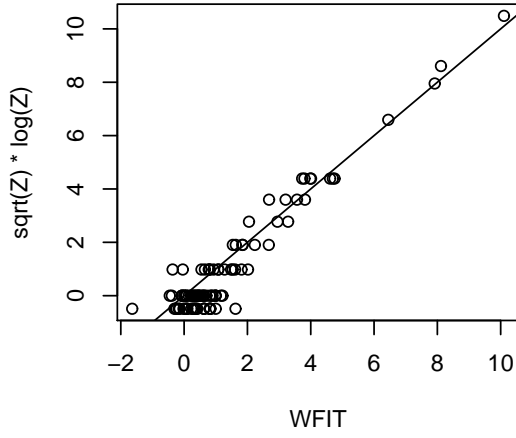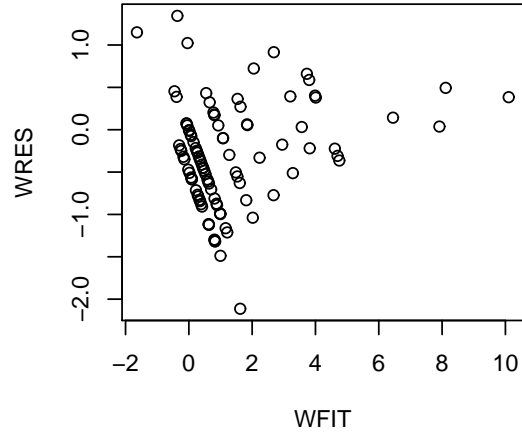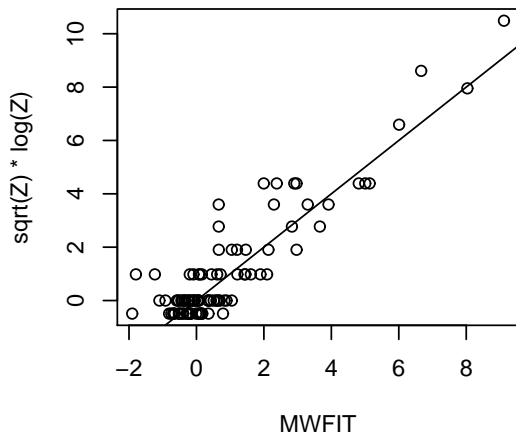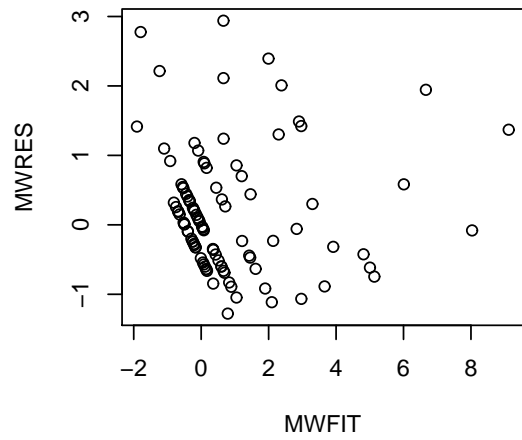
Figure 4: Diagnostic Plots for Loglinear Regression
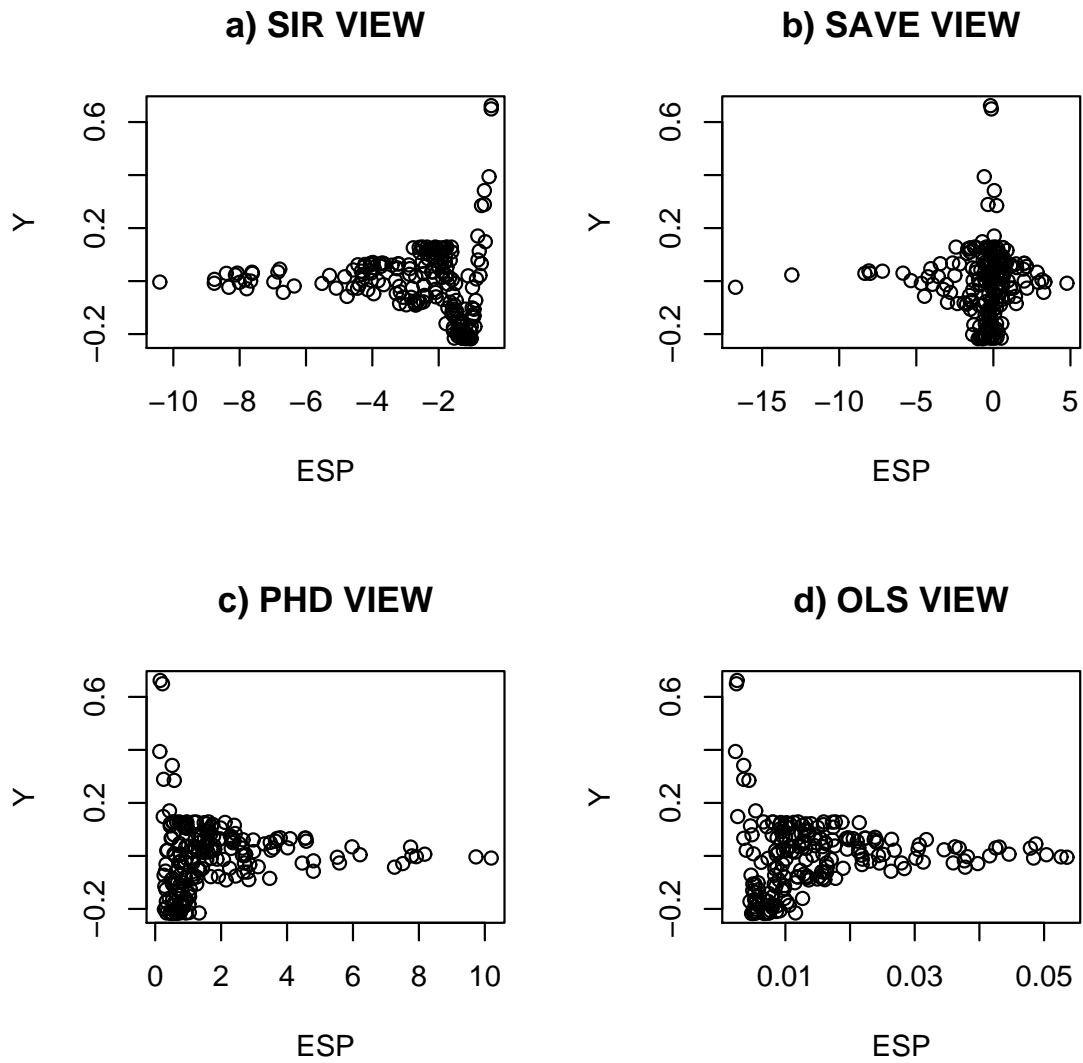
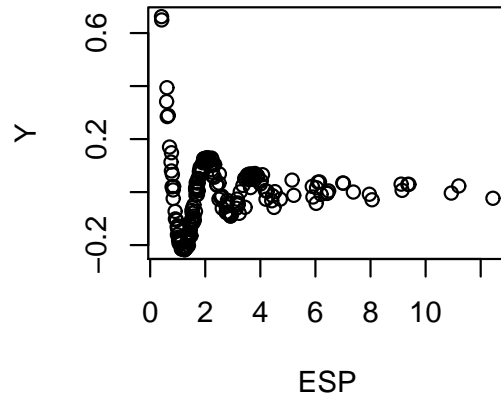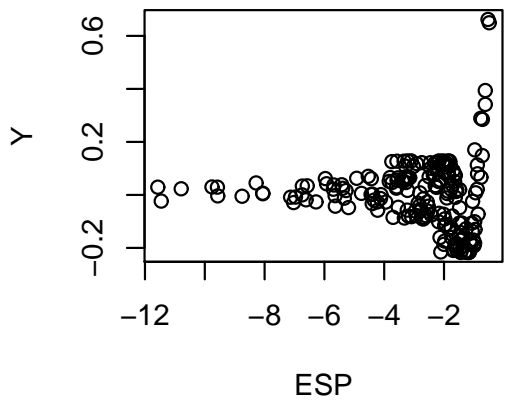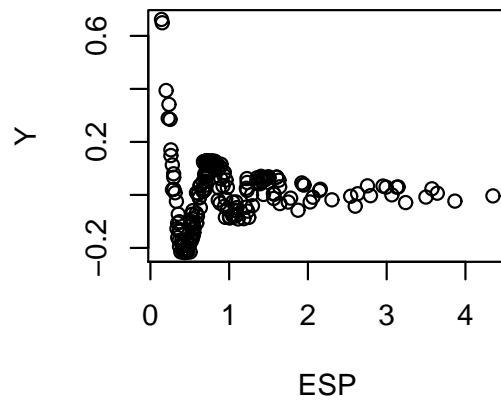Figure 5: Estimated Sufficient Summary Plots Without Trimming
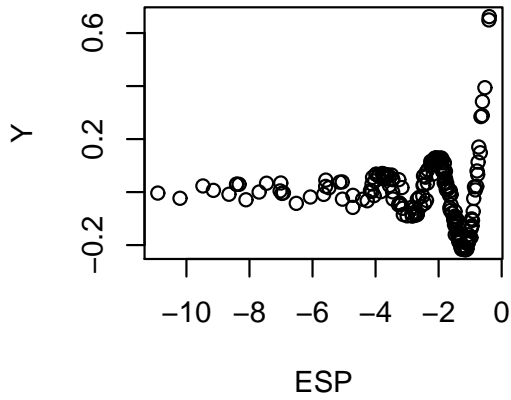
Figure 6: 1D Regression with Trimmed Views