Variable Selection for 1D Regression Models

David J. Olive

Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408,

and Douglas M. Hawkins

School of Statistics, University of Minnesota, Minneapolis, MN 55455-0493.

Variable selection, the search for j relevant predictor variables from a group of p candidates, is a standard problem in regression analysis. The class of 1D regression models is a broad class that includes generalized linear models. We show that existing variable selection algorithms, originally meant for multiple linear regression and based on ordinary least squares and Mallows' C_p, can also be used for 1D models. Graphical aids for variable selection are also provided.

KEY WORDS: C_p; Cook's Distance; Generalized Linear Models; Outliers; Regression Graphics; Single Index Models.

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

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

Many important regression models, including *generalized linear models* (GLM's), satisfy (1.1). Another example is the *response transformation model*,

$$y = t^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e), \tag{1.2}$$

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 an observed survival time, then many *survival models* including the Cox (1972) *proportional hazards model* are response transformation models. Yet another example satisfying (1.1) is the *single index model* which has the form

$$y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e.$$

The *multiple linear regression model* is an important special case of this model with $m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

The class of 1D models also includes many other special cases. Li and Duan (1989, p. 1014) list binary regression, censored regression, and projection pursuit models, while Stoker (1986), Horowitz (1998) and Cook and Weisberg (1999) also provide applications.

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). 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.

A standard problem in 1D regression is variable (or subset) selection. Assume that model (1.1) holds and that $\boldsymbol{x} = (x_1, ..., x_{p-1})^T$ are the $p-1$ nontrivial predictors. Then variable selection is a search for a subset of variables that can be deleted without important loss of information.

To clarify ideas, 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

$$SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_E^T\boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S. \tag{1.3}$$

The extraneous terms that can be eliminated given that the subset S is in the model have zero coefficients.

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

$$\mathrm{corr}(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{i}}, \alpha + \boldsymbol{\beta}_{\mathrm{I}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{I,i}}) = 1. \tag{1.4}$$

This observation, which is true regardless of the explanatory power of the model, 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 corr(ESP,ESP(I)). Any such subset for which the correlation is high is worth closer investigation and consideration. To make this advice more specific, use the rule of thumb that a candidate subset of predictors I is worth considering if the sample correlation of ESP and ESP(I) satisfies

$$\text{corr}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \tilde{\alpha}_{\mathrm{I}} + \tilde{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = \text{corr}(\tilde{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \tilde{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) > 0.95. \tag{1.5}$$

The difficulty in using this approach for general 1D problems is a computational one; with even modest numbers of predictors, there is a huge number of possible subsets I, and in general, fitting each of these subset models involves substantial computation. For this reason, proposals for subset selection in 1D problems have tended to use methods such as forward selection and backward elimination, despite their known inferiority – see for example Naik and Tsai (2001), Fan and Li (2002), Agresti (2002, pp. 211-217) or Cook and Weisberg (1999, pp. 485, 536-538).

The exception to this general difficulty is OLS, where there are computationally highly efficient algorithms (notably the Furnival-Wilson (1974) 'leaps and bounds' algorithm) for exploring all possible subsets.

This observation ties in with another. As shown by Li and Duan (1989), it is frequently found that fitting the full model as an ordinary least squares (OLS) regression gives a coefficient vector which is consistent for some non-zero multiple of the true ESP, even if the 1D model is not a linear regression. Pairing these observations leads to an approach in which the computational ease of OLS can be applied to the more general 1D subsetting problem:

- Fit a full model using the methods appropriate to that 1D problem to find the ESP

$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$.

- Find the OLS ESP $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}$.

- If the 1D ESP and the OLS ESP have 'a strong linear relationship' – for example $|\text{corr}(\text{ESP}, \text{OLS ESP})| > 0.95$ – then infer that the 1D problem is one in which OLS may serve as an adequate surrogate for the correct 1D model fitting procedure.

- Use computationally fast OLS subsetting procedures such as the leaps and bounds algorithm to identify predictor subsets that are effectively equivalent to the full set of predictions (as measured by such metrics as C_p, see Mallows 1973 and Jones 1946).

- Perform a final check on interesting-looking subsets identified in this way by using them to fit the 1D model.

This strategy allows us to use computationally efficient OLS procedures to perform the computationally intensive portion of subset investigation for some problems, restricting the potentially much heavier computations of the 1D fitting to just the final verification stages. Section 3 will show that if the model I contains k predictors including a constant and if $C_p(I) \leq 2k$, then

$$\text{corr}(\hat{\alpha}_{\text{OLS}} + \hat{\boldsymbol{\beta}}_{\text{OLS}}^{\text{T}} \boldsymbol{x}_{\text{i}}, \hat{\alpha}_{\text{OLS,I}} + \hat{\boldsymbol{\beta}}_{\text{OLS,I}}^{\text{T}} \boldsymbol{x}_{\text{I,i}}) \to 1$$

as the sample size $n \to \infty$.

Section 4 examines the impact of influential cases on variable selection for the multiple linear regression model. We show how to use an RC plot of residuals versus Cook's

distances to detect the influential cases. This graphical technique could be used to complement robust numerical variable selection methods. See Burman and Nolan (1995), Ronchetti and Staudte (1994) and Sommer and Huggins (1996).

2 Some Plotting Aids for Subset Selection

After performing a variable selection procedure, there will often be several subsets that look competitive (e.g., to subject matter experts). A large number of numerical and graphical quantities can be produced to compare the models. The ESP, the response y, and the difference $y - ESP$ can always be generated for a 1D regression, and sometimes Wald p-values are available for the coefficients $\hat{\beta}_i$. Often fitted values, residuals, diagnostics such as Cook's distances (Cook, 1977), and goodness of fit quantities such as the deviance and AIC are also available.

We use the following notation for naming plots.

F is the fitted value.

E is the ESP.

R is the fitted residual.

V is the difference $V = y - ESP$.

C is the Cook's distance.

In OLS, $E = F$ and $V = R$, but in other 1D problems this correspondence falls away. The term 'wz' plot refers to a plot with w on the horizontal axis and z on the vertical axis. So for example an EY plot has the ESP on the horizontal axis and the response

y on the vertical axis. FY, FR, ER and RC plots are all commonly useful. Equation 1.5 leads to an EE plot using ESP(I) on the horizontal axis and the full model ESP on the vertical. If several submodels $I_1, ..., I_d$ are under consideration, let I_0 denote the full model. Then a scatterplot matrix of y and ESP(I_j) for $j = 0, ..., d$ provides a compact comparison of all the subsets; those showing high correlation with the full ESP can be retained for closer study.

For multiple linear regression, the EY plot has been called a forward response plot and is a familiar model checking plot (Cook and Weisberg, 1997, 1999). It has been suggested for more general 1D model diagnostics also. Brillinger (1983) suggested using the OLS EY plot to visualize m for single index models. Li and Duan (1989) showed that under fairly reasonable conditions, the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is a $\sqrt{n}$ consistent asymptotically normal estimator of $c\boldsymbol{\beta}$, showing that the EY plot can be used to diagnose a general nonlinear 1D relationship.

The key to understanding which plots are the most useful is the observation that a wz plot is used to visualize the conditional distribution of z given w. Since a 1D regression is the study of the conditional distribution of y given $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, the EY plot is used to visualize this conditional distribution and should be made for any 1D regression analysis. Adding visual aids such as the estimated parametric mean function $m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x})$ for 1D models such as the binary logistic regression model can be useful. If an estimated nonparametric mean function $\hat{m}(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x})$ such as lowess follows the parametric curve closely, then often numerical goodness of fit tests will suggest that the model is good.

Similarly, an ER residual plot is used to visualize the conditional distribution of the residuals given the ESP. The EE plot can be used to quickly check that the correlation

is high due to linearity (not due to outliers), that the plotted points fall about some line, and that the line is the identity line (with unit slope and zero intercept). In the EY plot, the vertical discrepancies from the identity line are $V_{I,i} = y_i - \tilde{\alpha}_I - \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$. Section 3 will show that the VV plot is important for understanding how to use the OLS ESP for variable selection.

Efficient use of the graphical and numerical quantities is very important for variable selection. Experience suggests that the EY plot should be made for both the full model and the final submodel. If $\mathrm{corr}(\mathrm{ESP}, \mathrm{ESP(I)}) > 0.95$, then the two EY plots look nearly identical. For correlations less than 0.85, sometimes the two plots look very different. If a lack of fit plot such as a residual plot is available, then it should also be made for both models. If several competing submodels are available, an EE scatterplot matrix may be used to compare them compactly. In a binary regression, marking the "successes" and "failures" with different plotting symbols or colors adds considerable insight without any chart clutter.

The following rules of thumb may be useful for multiple linear, logistic, and loglinear regression. The submodel should have a small number of predictors subject to the constraint that $\mathrm{SSE}(I)$ or the deviance $G^2(I)$ is close to that of the full model in that the partial F test or change in deviance test should conclude that the submodel is good. Also the submodel I should not have many variables with large Wald p–values.

3 Using OLS and C_p for 1D Variable Selection

This section provides theoretical results for the OLS ESP, and the following notation will be useful. Assume that all models include a constant and that $\boldsymbol{X}$ is the $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively. In multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ plus a constant has k terms, then the F_I statistic for testing whether the $p - k$ predictor variables can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k) - (n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}[\frac{SSE(I)}{SSE} - 1]$$

where SSE is the error sum of squares from the full model and SSE(I) is the error sum of squares from the candidate submodel. Also recall that

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model. Notice that $C_p(I) \leq k$ if and only if $F_I \leq 1$.

When the 1D model is not a multiple linear regression model, the OLS ESP is equal to the OLS fit and the OLS vertical discrepancies $V_{I,i}$ are equal to the OLS residuals $r_{I,i}$. Hence the FY, FF and RR plots should be called EY, EE and VV plots, respectively. For a plot having w on the horizontal axis and z on the vertical axis, denote the OLS line by $\hat{z} = a + bw$. The following proposition is a property of OLS and holds even if the

data does not follow a 1D regression model.

Proposition 3.1. Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.

EY or FY Plot:

i) If $w = \hat{y}_I$ and $z = y$, then the OLS line is the identity line.

ii) If $w = y$ and $z = \hat{y}_I$, then the OLS line has slope $b = [\mathrm{corr}(y, \hat{y}_I)]^2 = R_I^2$ and intercept $a = \overline{y}(1 - R_I^2)$ where $\overline{y} = \sum_{i=1}^n y_i/n$ and R_I^2 is the coefficient of multiple determination from the candidate model.

EE or FF Plot:

iii) If $w = \hat{y}_I$ and $z = \hat{y}$, then the OLS line is the identity line.

iv) If $w = \hat{y}$ and $z = \hat{y}_I$, then the OLS line has slope $b = [\mathrm{corr}(\hat{y}, \hat{y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.

VV or RR Plot:

v) If $w = r$ and $z = r_I$, then the OLS line is the identity line.

vi) If $w = r_I$ and $z = r$, then $a = 0$ and the OLS slope $b = [\mathrm{corr}(r, r_I)]^2$ and

$$\mathrm{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

Proof: See appendix.

In many settings (not all of which meet the quite strict Li-Duan sufficient conditions), the full model OLS ESP is a good estimator of the sufficient predictor. When this is the case, if p is fixed and $C_p(I) \leq k$ or $F_I \leq 1$, then in the VV plot the plotted points will cluster about the identity line and the correlation of the plotted points will be large. Then the same result will hold for the plotted points in the EE plot: OLS ESP $\approx$ OLS

ESP(I), and the EY plots based on the full and submodel ESP can both be used to visualize the conditional distribution of y. (The correlations of the plotted points in the two EY plots will be nearly the same since $\sqrt{R^2} \approx \sqrt{R_I^2}$.)

If a 1D model holds, a common assumption made for variable selection is that the fitted full model ESP is a good estimator of the sufficient predictor, and the usual numerical and graphical checks on this assumption should be made. To see that this assumption is weaker than the assumption that the OLS ESP is good, notice that if a 1D model holds but $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $c\boldsymbol{\beta}$ where $c = 0$, then the $C_p(I)$ criterion could wrongly suggest that all subsets I have $C_p(I) \leq k$. Hence we also need to check that $c \neq 0$.

There are several methods for checking the OLS ESP, including: a) if an ESP from an alternative fitting method is believed to be useful, check that the ESP and the OLS ESP have a strong linear relationship – for example that |corr(ESP, OLS ESP)| > 0.95. b) Often examining the EY plot shows that a 1D model is reasonable. For example, if the data are tightly clustered about a smooth curve, then a single index model may be appropriate. c) Verify that $\boldsymbol{x}$ has an elliptically contoured distribution with 2nd moments and that the mean function $m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ is not symmetric about the median of the distribution of $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Then results from Li and Duan (1989) suggest that $c \neq 0$.

Condition a) is both the most useful (being a direct performance check) and the easiest to check. A standard fitting method should be used when available (e.g., for parametric 1D models or the proportional hazards model). Condition c) needs $\boldsymbol{x}$ to have a continuous multivariate distribution while the predictors can be factors for a) and b). Olive (2002) gives a graphical procedure for checking that a distribution is elliptically contoured and gives a weighted ESP that can sometimes cause condition b) to hold when

c) is violated.

Daniel and Wood (1971, p. 87) suggest using Mallows' graphical method for screening subsets by plotting k versus $C_p(I)$ for models close to or under the $C_p = k$ line. Proposition 3.1 vi) implies that if $C_p(I) \leq k$ then $\text{corr}(V, V(I))$ and $\text{corr}(ESP, ESP(I))$ both go to 1.0 as $n \to \infty$. Hence models I that satisfy the $C_p(I) \leq k$ screen will contain the true model S with high probability when n is large. This result does not guarantee that the true model S will satisfy the screen, hence overfit is likely (see Shao 1993). Let d be a lower bound on $\text{corr}(V, V(I))$. Proposition 3.1 vi) implies that if

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(V, V(I)) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

$$d_n \equiv \sqrt{1 - \frac{p}{n}}.$$

To reduce the chance of overfitting, use the $C_p = k$ line for large values of k, but also consider models close to or under the $C_p = 2k$ line for small values of k. A referee noted that the true simulated logistic regression model S satisfied $C_p(S) \leq k$ for about 60% of the simulated data sets. We simulated multiple linear regression and single index model data sets with $p = 8$ and $n = 50, 100, 1000$ and 10000. Again 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. The following example helps illustrate the above discussion.

Example 1. Li (1997) showed that the Boston housing data of Harrison and Rubinfeld (1978) grossly violates the Li and Duan (1989) conditions. One model for the data is a response transformation with $t(y) = \log(y)$ where the response $y = \text{CRIM}$, the per capita crime rate by town. The predictors were $x_1 =$ proportion of residential land zoned for

lots over 25,000 sq.ft., x_2 = log(proportion of non-retail business acres per town), x_3 = Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), $x_4 = NOX$ = nitric oxides concentration (parts per 10 million), x_5 = average number of rooms per dwelling, x_6 = proportion of owner-occupied units built prior to 1940, x_7 = log(weighted distances to five Boston employment centers), $x_8 = RAD$ = index of accessibility to radial highways, x_9 = log(full-value property-tax rate per \$10,000), x_{10} = pupil-teacher ratio by town, $x_{11} = 1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town, and x_{12} = log(% lower status of the population).

To illustrate the potential of the OLS ESP, consider the full model with the response y untransformed (that is, on the natural, and not the logarithmic scale) and predictors x_2, x_3, x_4, x_5, x_7, x_8, x_9 and x_{12}. If a multiple linear regression of $\log(y)$ on $\boldsymbol{x}$ is appropriate, then this model is a nonlinear 1D model. (As pointed out by a referee, some readers may disagree that the multiple linear regression model is appropriate, but our method can still produce interesting subsets since Proposition 3.1 holds even if the data does not follow a 1D regression model.)

The essentially unique "interesting" C_p value (searching all subsets with the branch and bound algorithm) is the value 5.7 obtained using x_2, x_4, x_7, x_8 and x_{12} as predictors. Figure 1 shows the VV and EE plots for this minimum C_p submodel. Notice the similarity of the EY plots for the full model and submodel. Since $C_p(I) = 5.7 < k = 6$, the correlation of the plotted points in the VV plot is high, as expected.

Despite the nonlinearity in the model, using fast OLS subsetting technology leads to a good model of the relationship. Further exploration of this data suggests that NOX and RAD are the most important predictors. A plot of NOX vs. RAD reveals two clusters

of locales with high NOX and high RAD that correspond to the cases with the highest per capita crime rate.

4 A Graphical Aid for Multiple Linear Regression

In this section we assume that the multiple linear regression model holds and that the full model uses all $p-1$ predictor variables plus a constant. Cases that have atypical leverage and/or deviation often have substantial impact on numerical variable selection methods, and the subsets identified from the "cleaned data" that excludes these cases may be very different from those using the full data set, a situation that should cause concern. This result suggests running the numerical variable selection procedure on the entire data set and on the cleaned data set, keeping track of interesting models from both data sets. For a candidate submodel I, let $C_p(I, c)$ denote the value of the C_p statistic for the cleaned data.

The RC plot of the residuals r_i versus the Cook's distances CD_i is useful for finding the influential cases. Recall that

$$\mathrm{CD}_i = \frac{r_i^2}{p\hat{\sigma}^2} \frac{h_i}{(1-h_i)^2}, \tag{3.1}$$

where h_i is the leverage and $\hat{\sigma}^2$ is the usual estimate of the error variance.

Though two-dimensional, the RC plot is attractive because it shows three case diagnostics, giving the cases' residuals, leverage, and influence together. Cases with the same leverage define a parabola in the RC plot; this parabola is steep if the leverage is large, and flat if it is small. In an ideal setting with no outliers or undue case leverage, this plot should be an evenly-populated parabola. This leads to a graphical approach of mak-

ing the RC plot, temporarily deleting cases that depart from this ideal shape (through extreme lateral or radial location), refitting the model and regenerating the plot to see whether it now conforms to the desired shape. The following example illustrates the approach.

Example 2. Gladstone (1905-1906) attempts to estimate $y = weight$ of the human brain (in grams, measured after death) using simple linear regression with a variety of predictors including $x_1 = age$ in years, $x_2 = height$ in inches, $x_3 =$ *head height* in mm, $x_4 =$ *head length* in mm, $x_5 =$ *head breadth* in mm, $x_6 =$ *head circumference* in mm, and $x_7 =$ *cephalic index*. The predictor $x_8 = sex$ (coded as 0 for females and 1 for males) of each subject was also included. Head *size*, the product of the *head length*, *head breadth*, and *head height*, is a volume measurement. Hence $x_9 = (size)^{1/3}$ was also used as a predictor with the same physical dimensions as the other lengths. Thus there are 9 nontrivial predictors and one response, and all models will also contain a constant. Of the original 276 cases, nine were deleted because of missing values, leaving 267 cases.

Table 1 shows the summary statistics of the more interesting subset regressions. The smallest C_p value came from the subset x_1, x_5, x_8, x_9, and in this regression x_5 has a t value of 1.76. Deleting a single predictor from an adequate regression changes the C_p by approximately $t^2 - 2$, where t stands for that predictor's Student's t in the regression – as illustrated by the increase in C_p from 4.4 to 6.3 following deletion of x_5. Analysts must choose between the larger regression with its smaller C_p but a predictor that does not pass the conventional screens for statistical significance, and the smaller, more parsimonious, regression using only apparently statistically significant predictors, but (as assessed by

C_p) possibly less accurate predictive ability.

Figure 2 shows a sequence of RC plots used to identify cases 118, 234, 248 and 258 as atypical; deleting them leads to an RC plot that is a reasonably evenly-populated parabolic band. There is nothing particularly striking about these four atypical cases other than their incompatibility with the main sweep of the data, but data capture errors are a possible factor.

One of the biggest advantages of using a sequence of RC plots to detect influential cases is that the sequence tends to be small and there is a stopping criterion. Another advantage of the RC plot is that there could be a point with a residual near zero but the Cook's distance does not stick out (is the 5th largest, for example). This case is likely to be influential on numerical variable selection methods but can't be found with an FR residual plot or an FC plot of fitted values versus Cook's distances.

Figure 3 shows the FY plots and FR residual plots for the full model and the more parsimonious choice for a final submodel I – that using a constant, $x_1 = age$, $x_8 = sex$ and $x_9 = size^{1/3}$. A further five cases (230, 254, 255, 256 and 257) are well separated from the bulk of the data in each of the four plots. These correspond to five infants. They reflect the age gap between the handful of infants and the bulk of the data. By definition they must have higher leverage than average, and so good exploratory practice would be to remove them also to see the effect on the model fitting. The right columns of Table 1 reflect making all 9 deletions. As in the full data set, the subset x_1, x_5, x_8, x_9 gives the smallest C_p, but x_5 is of only modest statistical significance and might reasonably be deleted to get a more parsimonious regression. What is striking after comparing the left and right columns of Table 1 is that the adequate C_p values for the cleaned data set seem

substantially smaller than their full-sample counterparts: 1.2 versus 4.4, and 2.3 versus 6.3. Since these C_p for the same p are dimensionless and comparable, this suggests the otherwise non-obvious fact that these 9 cases are primarily responsible for any additional explanatory ability in the 6 unused predictors, and so are influential to variable selection.

Multiple linear regression data sets with cases that influence numerical variable selection methods are common, and subsets selected using both the entire data set and the clean data set should be examined. Two data archives for the *Arc* software (Cook and Weisberg 1999) were examined, and Table 2 shows results for seven of the more interesting data sets. The first five data sets are available from the website (http://www.math.siu.edu /olive) while the final two data sets come with the *Arc* software available from the website (http://www.stat.umn.edu/arc/). The first 4 rows of Table 2 correspond to the Gladstone data of Example 2, with and without the 5 infants.

The full model used p predictors, including a constant. The final submodel I also included a constant, and the nontrivial predictors are listed in the third column of Table 2. The fourth column lists p, $C_p(I)$ and $C_p(I, c)$ while the second column gives the set of influential cases. Two rows are presented for each data set. The second row gives the response variable and any predictor transformations. For example, for the Gladstone data $p = 10$ since there were 9 nontrivial predictors plus a constant. Only the predictor *size* was transformed, and the final submodel is the one given in Example 2. For the rat data, the final submodel used a constant but did not use any of the 3 nontrivial predictors. The major and ais data sets show that deleting the influential cases may increase the C_p statistic.

5 CONCLUSIONS

To summarize, if the fitted full 1D model $y \perp\!\!\!\perp \boldsymbol{x} | \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ is a useful approximation to the data and if $\hat{\boldsymbol{\beta}}_{OLS}$ is a good estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, then a subset I will produce an EY plot similar to the EY plot of the full model if corr(OLS ESP, OLS ESP(I)) > 0.95. Assume that subset I uses k predictors including the intercept, that $C_p(I) \leq 2k$ and $n \geq 10p$. Then $0.9 \leq \text{corr}(V, V(I))$, and both $\text{corr}(V, V(I)) \rightarrow 1.0$ and corr(OLS ESP, OLS ESP(I)) $\rightarrow 1.0$ as $n \rightarrow \infty$. For a fixed value of k, the model I with the smallest value of $C_p(I)$ maximizes $\text{corr}(V, V(I))$. Notice that within the (large) subclass of 1D models where the OLS ESP is useful, the OLS partial F test is robust (asymptotically) to model misspecifications in that $F_I \leq 1$ correctly suggests that submodel I is good.

A framework for variable selection for models that produce fitted values $\hat{y}$ for the response variable y can also be developed. Such models include single and multi–index, nonlinear regression, nonparametric regression and time series models. For these models the ESP may not exist, but a subset I is "interesting" if the correlation $\text{corr}(\hat{y}, \hat{y}_I)$ of the fitted values from the full and submodel is higher than 0.95.

For 1D regression models, the OLS ESP variable selection method can often be used to examine all subsets. The Furnival-Wilson OLS branch and bound algorithm permits an exhaustive study of up to some 30 predictors and arbitrarily many cases on standard desktop computers. This problem size far exceeds what can be accommodated in direct fitting of 1D models in most non-OLS settings.

All of the plots discussed in the paper are easy to produce with good general purpose regression software since they involve conventional OLS diagnostics. Object-linking soft-

ware that supports brushing and temporary case deletion with automatic plot updates is particularly suitable for exploring the interplay between cases and subset selection criteria. The plots used in this paper were produced using both *Splus* and *Arc* (Cook and Weisberg 1999), a public-domain regression system on an *Xlisp-Stat* base.

Section 4 showed how to use the RC plot for multiple linear regression. In principle, this same approach can be used in other 1D modeling settings, with the substitution of model-appropriate definitions of residuals and Cook's distance. For example, an RC plot for logistic regression can be made using the standardized Pearson's residual and the Cook type distance suggested by Collett (1991, p. 151).

The literature on numerical methods for variable selection in the OLS multiple linear regression model is enormous, and the literature for other given 1D regression models is also growing. If the variable selection techniques in these papers are successful, then the estimated sufficient predictors from the full and candidate model should be highly correlated. Influential cases will often appear in the VV and EE plots, and the EY plot is useful for detecting clusters of outliers and for visualizing the conditional distribution of y. Influential cases may also appear in residual and added variable plots.

The Boston housing data can be obtained from the STATLIB website (http://lib.stat.cmu.edu/datasets/boston).

Acknowledgments

The authors are grateful to the editor, William Notz, associate editor and referees for a number of helpful suggestions for improvement in the article. Stan Young and Dennis Cook read an earlier version of the manuscript. This work was supported by the National

Science Foundation under grants DMS 0202922, DMS 0306304, DMS 9803622 and ACI 9619020.

APPENDIX

Proof of Proposition 3.1: Several authors (e.g., Draper and Smith 1981, p. 140; Chambers, Cleveland, Kleiner, and Tukey 1983, p. 280) have suggested using the FY plot to visualize R^2. Hence the proofs of i) and ii) are straightforward modifications of known full model results.

Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing z on w is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

$$b = \frac{\sum(w_i - \overline{w})(z_i - \overline{z})}{\sum(w_i - \overline{w})^2} = \frac{SD(z)}{SD(w)}\text{corr}(z, w).$$

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing z on w is equal to one if and only if the OLS slope from regressing w on z is equal to $[\text{corr}(z, w)]^2$.

The proofs of ii), iv) and vi) follow from (*) and the proofs of i), iii) and v).

i) The slope $b = 1$ if $\sum \hat{y}_{I,i} y_i = \sum \hat{y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{y} - \overline{y} = 0$.

iii) The slope $b = 1$ if $\sum \hat{y}_{I,i} \hat{y}_i = \sum \hat{y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T \boldsymbol{H}\boldsymbol{H}_I \boldsymbol{Y}$ $= \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{y} - \overline{y} = 0$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T \boldsymbol{r}_I / \boldsymbol{r}^T \boldsymbol{r}$. Since $\boldsymbol{r}^T \boldsymbol{r}_I = \boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) = \boldsymbol{I} - \boldsymbol{H}$, the numerator $\boldsymbol{r}^T \boldsymbol{r}_I = \boldsymbol{r}^T \boldsymbol{r}$ and $b = 1$.

6 References

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

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.

Burman, P., and Nolan D. (1995), "A General Akaike-Type Criterion for Model Selection in Robust Regression," *Biometrika*, 82, 877-886.

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

Collett, D. (1991), *Modelling Binary Data,* Chapman & Hall/CRC, Boca Raton, Florida.

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

Cook, R.D., and Weisberg, S. (1997), "Graphs for Assessing the Adequacy of Regression Models," *Journal of the American Statistical Association,* 92, 490-499.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* John Wiley and Sons, NY.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Daniel, C., and Wood, F.S. (1971), *Fitting Equations to Data,* John Wiley and Sons, NY.

Draper, N.R., and Smith, H. (1981), *Applied Regression Analysis,* 2nd Ed., John Wiley

and Sons, NY.

Fan, J., and Li, R. (2002), "Variable Selection for Cox's Proportional Hazard Model and Frailty Model," *The Annals of Statistics*, 30, 74-99.

Furnival, G., and Wilson, R. (1974), "Regression by Leaps and Bounds," *Technometrics,* 16, 499-511.

Gladstone, R.J. (1905-1906), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

Harrison, D. and Rubinfeld, D.L. (1978), "Hedonic Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management,* 5, 81-102.

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

Jones, H.L. (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

Koenker, R., and Geling, O. (2001), "Reappraising Medfly Longevity: a Quantile Regression Survival Analysis," *Journal of the American Statistical Association,* 96, 458-468.

Li, K.C. (1997), "Nonlinear Confounding in High-Dimensional Regression," *The Annals of Statistics,* 25, 577-612.

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

Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

Naik, P.A., and Tsai, C. (2001), "Single-Index Model Selections," *Biometrika,* 88, 821-832.

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

Ronchetti, E., and Staudte, R.G. (1994), "A Robust Version of Mallows's C_p", *Journal of the American Statistical Association,* 89, 550-559.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association,* 88, 486-494.

Sommer, S., and Huggins, R.M. (1996), "Variables Selection Using the Wald Test and a Robust C_p," *Applied Statistics*, 45, 15-29.

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

Table 1: Some Subsets – Brain Data

		All cases		Cleaned data	
Subset I	k	RSS $\times 10^3$	$C_p(I)$	RSS$\times 10^3$	$C_p(I,c)$
x_1, x_9	3	1486	12.6	1352	10.8
x_8, x_9	3	1655	43.5	1516	42.8
x_1, x_8, x_9	4	1442	6.3	1298	2.3
x_1, x_5, x_9	4	1463	10.1	1331	8.7
x_1, x_5, x_8, x_9	5	1420	4.4	1282	1.2
All	10	1397	10.0	1276	10.0

Table 2: Summaries for Seven Data Sets

file	influential cases	submodel I	p, $C_p(I)$, $C_p(I,c)$
file	response	transformed predictors	
cbrain	118, 234, 248, 258	$(size)^{1/3}$, age, sex	10, 6.337, 3.044
cbrain	brnweight	$(size)^{1/3}$	
cbrain-5	118, 234, 248, 258	$(size)^{1/3}$, age, sex	10, 5.603, 2.271
cbrain-5	brnweight	$(size)^{1/3}$	
pop	14, 55	$\log(x_2)$	4, 12.665, 0.679
pop	log(y)	$\log(x_1)$, $\log(x_2)$, $\log(x_3)$	
cyp	11, 16, 56	sternal height	7, 4.456, 2.151
cyp	height	none	
major	3, 44	x_2, x_5	6, 0.793, 7.501
major	height	none	
ais	11, 53, 56, 166	log(LBM), log(Wt), sex	12, -1.701, 0.463
ais	%Bfat	log(Ferr), log(LBM), log(Wt), $\sqrt{Ht}$	
rat	3	no predictors	4, 6.580, -1.700
rat	y	none	

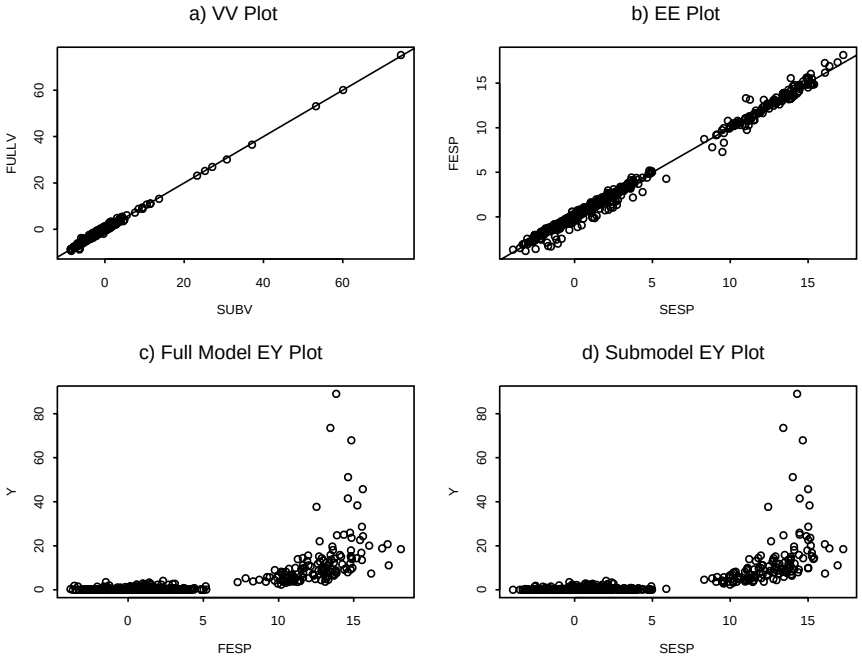

Figure 1: Boston Housing Data: Nonlinear 1D Regression Model

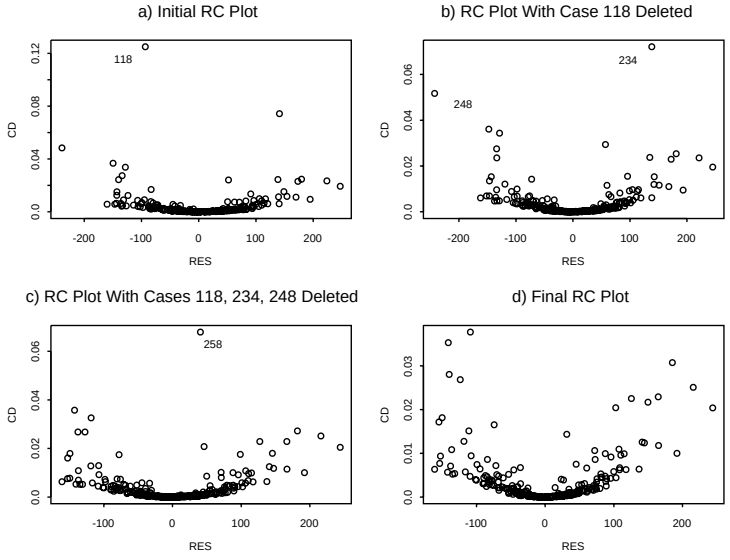

Figure 2: RC Plots for the Gladstone Brain Data

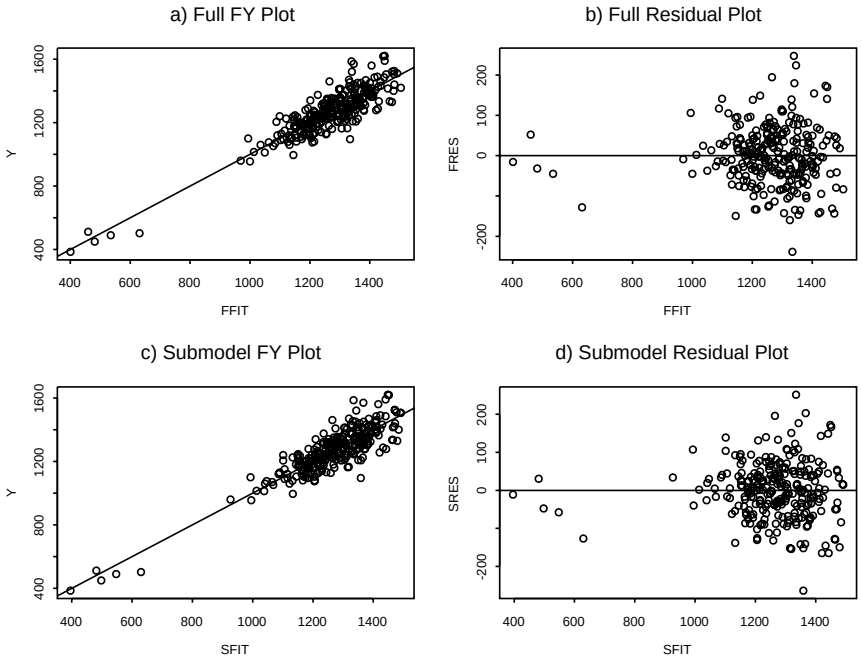

Figure 3: Gladstone data: comparison of the full model and the submodel.