Statistica Sinica **5**(1995), 689-708

GRAPHICS FOR STUDYING NET EFFECTS OF REGRESSION PREDICTORS

R. Dennis Cook

University of Minnesota

Abstract. Graphical methods are proposed for studying the contributions of selected predictors to regression problems. By developing low dimensional distributional index functions based on sliced inverse regression, problems with many predictors can be addressed. It is shown that added variable plots can fill this role under certain conditions, but that they may generally overestimate predictor contributions. Scatterplot brushing plays a basic role in the methodology.

Key words and phrases: Added variable plots, adjusted variable plots, conditional response plots, scatterplot brushing, scatterplot matrices, sliced inverse regression.

1. Prelude

Graphical displays of data have always played an important role in statistical analyses. Recent innovations in computer graphics like rotation, linking and brushing have greatly increased the potential to understand data graphically, particularly when combined with a modern computing environment like *Lisp-Stat* (Tierney (1990)). Most of the recent innovations involve methods for displaying data, with relatively little attention devoted to supporting statistical theory. As a consequence, many people seem to regard graphics as a collection of ad hoc techniques that lack the foundations associated with most statistical methodology. The power, applicability and acceptance of statistical graphics might be increased by establishing statistical foundations for existing techniques and developing statistical theory to guide the construction and interpretation of new displays.

In this article I discuss how linking and brushing might be used to study the effects of regression predictors. Specifically, how might graphics be used to study the effects of including a selected vector of p_2 predictors x_2 after fully accounting for the contribution of the remaining vector x_1 of p_1 predictors? Plots that facilitate such understanding can be collectively thought of as *net effect plots* since it is the net effect of x_2 on the response that is of primary interest. Net effect plots can differ depending on the effects of interest and on the structure of the regression problem. Added variable plots (Cook and Weisberg (1982, p. 44)),

for example, can be instances of net effect plots that do not require brushing or linking.

An expository discussion on the use of brushing and linking to study net effects is given in Section 2. For background on those techniques see, Becker and Cleveland (1987) and Becker, Cleveland and Wilks (1987). One idea of Section 2 is that net effect plots can be obtained by brushing cells in a scatterplot matrix when the number of predictors in the regression is sufficiently small. However, dimension reduction becomes necessary when the number of predictors is larger than 2 or 3.

In Section 3 I discuss possibilities for dimension reduction via distributional index functions. This section makes use of recent dimension-reduction methodology developed by Li (1991, 1992). New results and ideas are presented to establish links between dimension-reduction methodology and the construction of net effect plots to guide model development or to study contributions of selected predictors after an adequate model has been selected. These and other instances of net effect plots are intended as constructions to allow predictor effects to be studied directly. They are not intended as diagnostics for detecting failures in a target model.

The ideas of Sections 2 and 3 are used in Section 4 to indicate when added variable plots can and cannot be used as net effect plots, thus establishing guidelines that may help avoid over-interpretation of added variable plots. In particular, net effect plots are not generally devices for providing visual information about the estimation of regression coefficients, although they may serve that role in special cases. Section 5 contains two brief examples.

A $(q+1)$-dimensional scatterplot will be denoted by $\{a, b\}$ where the first argument, which will always be a scalar, is allocated to the vertical axis and the coordinates of the vector b are allocated to the "horizontal" axis or axes in any convenient way. Following Dawid (1979), I use the notation $u \perp v$ to indicate that the random variables u and v are independent. Similarly, $u \perp v | z$ means that u and v are independent given any value for the random variable z.

2. Net Effect Plots

Let y_i denote the ith observation on the univariate response y, and let x_i denote the ith observation on the $p \times 1$ vector of predictors x. Partition $x^T = (x_1^T, x_2^T)$ so that $p = p_1 + p_2$. The data (y_i, x_i^T), $i = 1, \ldots, n$, are assumed to be iid observations on the random vector (y, x^T). The ultimate goal of a regression analysis is to characterize the behavior of the conditional distribution of y given x, with cdf denoted by $F(y|x)$, as the value of x varies in the relevant sample space. Here, however, I concentrate on graphical methods for studying the contribution

of x_2 after accounting for x_1.

Perhaps the most elusive task in developing a general paradigm for constructing a net effect plot is understanding how to deal with x_1 prior to considering x_2. Historically, various phrases have been used in the literature to convey this general idea, including "the regression on x_2 after accounting or adjusting for x_1". In the construction of added variable plots, for example, x_1 is first taken into account by using the residuals $e_{y|1}$ and $e_{2|1}$ from the ordinary least square (OLS) regression of y on x_1 and the OLS regression x_2 on x_1, respectively. An added variable plot is then the two-dimensional plot $\{e_{y|1}, e_{2|1}\}$ when $p_2 = 1$. Chambers, Cleveland, Kleiner and Tukey (1983, p. 268) refer to added variable plots as adjusted variable plots. The virtues of these adjustments in terms of the underlying distributions do not seem to be considered explicitly in the literature, and this apparently has lead to some confusion about the role of added variable plots (adjusted variable plots) in regression analysis, a topic that is explored in Section 4. Nevertheless, one way to account for x_1 is to condition on a specific value for x_1. In this article I will use conditioning as the operational version of these ideas. Conditioning can be used with or without a model, as discussed in Section 3. The coplots discussed in Chambers and Hastie (1992) are a very nice implementation of conditioning to study regression predictors.

2.1. Conditional response plots

Ideally, a net effect plot for x_2 at a particular value for x_1 is a graphical display of y versus x_2 where the data are a sample from the joint conditional distribution of (y, x_2) given x_1, with cdf denoted by $F(y, x_2|x_1)$. There may rarely be enough data to satisfy this requirement exactly, but useful results can often be obtained by conditioning approximately on x_1. Let J denote a subset of the case indices $(1, 2, \ldots, n)$ so that x_1 is relatively constant for $i \in J$. The plot $\{y_i, x_{i2}|i \in J\}$ shows the relationship between y and x_2 near the selected value of x_1. Since x_1 is relatively constant, only x_2 is left to explain the remaining variation in y. I refer to plots of the form $\{y_i, x_{i2}|i \in J\}$ as *conditional response* (CORE) plots since they display the regression relationship between y and x_2 after approximately conditioning on x_1. CORE plots are a basic form of net effect plots because they show the effect of x_2 after accounting for (conditioning on) x_1.

CORE plots are useful for studying the net effect of x_2 on y at a particular value of x_1. Brushing, interactively modifying J and updating the CORE plot, can be used to visualize how $F(y, x_2|x_1)$ changes with the value of x_1. Direct construction of CORE plots is limited to regression problems in which the dimensions of x_1 and x_2 are small since practical limitations are encountered oth-

erwise. When x_2 is a single predictor, $p_2 = 1$, a CORE plot is two dimensional and can be constructed with graphics programs that allow linking and brushing. Three-dimension CORE plots are possible, but interpretation may be more difficult.

The dimension of x_1 is a second limitation because brushing is mostly confined to problems in which $p_1 = 1$ or 2. This is a much more serious dimension restriction than that associated with x_2 since it limits application to regression problems with at most 3 or 4 predictors.

2.2. Reducing brushing dimensions

Even if a plot of several predictors could be brushed, the sparseness usually encountered in high dimensional plots may make it difficult to capture a subset of the data that reasonably approximates a sample from $F(y, x_2|x_1)$. Interpretation may be difficult as well. Imagine that we do have a method of brushing a scatterplot of arbitrary dimension and that sparseness is not an issue. Let $p_2 = 1$ and assume that $F(y, x_2|x_1)$ depends on x_1 only through the two linear combinations $\alpha^T x_1$ and $\gamma^T x_1$ so that $(y, x_2) \perp x_1 | (\alpha^T x_1, \gamma^T x_1)$. Consider brushing a p_1-dimensional plot of x_1 that is linked to the plot $\{y, x_2\}$ without knowledge of the problem structure. What we see in the CORE plots $\{y, x_2|J\}$ obtained while brushing depends on the movement of the brush in R^{p_1} relative to the subspace $S = \text{span}\{\alpha, \gamma\}$. Imagine moving the brush to highlight the points that are near a line L (a one-dimensional affine subspace) that is orthogonal to S. The projection of any point in L onto S will yield the same value. Thus, while moving the brush along L the distribution $F(y, x_2|x_1)$ will not change since the key linear combinations $(\alpha^T x_1, \gamma^T x_1)$ remain essentially constant. If we brush along a line that is parallel to S it may appear that the net effect of x_2 depends strongly on x_1 since the value of $(\alpha^T x_1, \gamma^T x_1)$ will change with every brush movement. Interpreting a series of CORE plots when S is unknown may be difficult since brushing will probably be neither parallel nor orthogonal to S. On the other hand, if S were known then we could replace the p_1-dimensional plot of x_1 with the two-dimensional plot $\{\alpha^T x_1, \gamma^T x_1\}$ without loss of information since $F(y, x_2|x_1) = F(y, x_2|\alpha^T x_1, \gamma^T x_1)$ for all values of x_1. These ideas along with some that will be introduced subsequently are illustrated in the following example.

Example 1. Let w_1 and w_2 be independent uniform random variables on $(-1, 1)$, and let $w_3|(w_1, w_2)$ be a normal random variable with mean $(w_1 + w_2)^2$ and variance 0.2. These three variables are the predictors for the example, $w^T = (w_1, w_2, w_3)$. The distribution of $y|w$ is described by the linear model,

$$y|w = 1.5(w_1 + w_2) + w_3 + 0.5\varepsilon, \tag{2.1}$$

where ε is a standard normal random variable and $\varepsilon \perp w$. In the previous notation, set $x_1^T = (w_1, w_2)$ and $x_2 = w_3$ so it is the net effect of w_3 that we wish to understand.

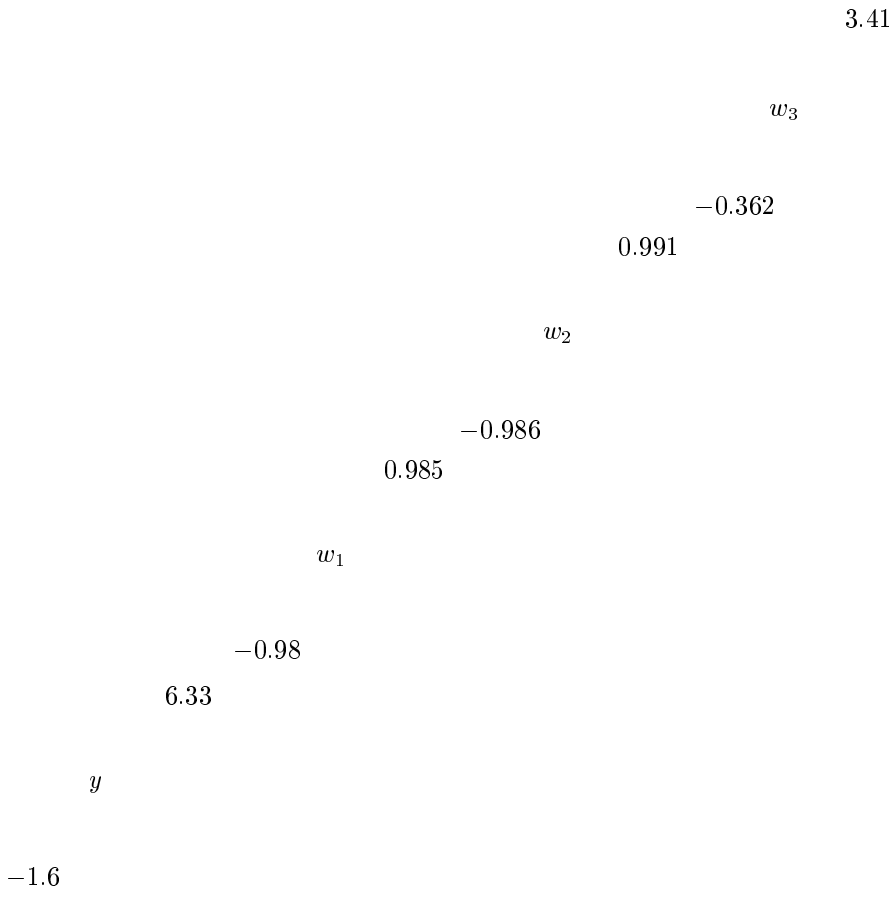
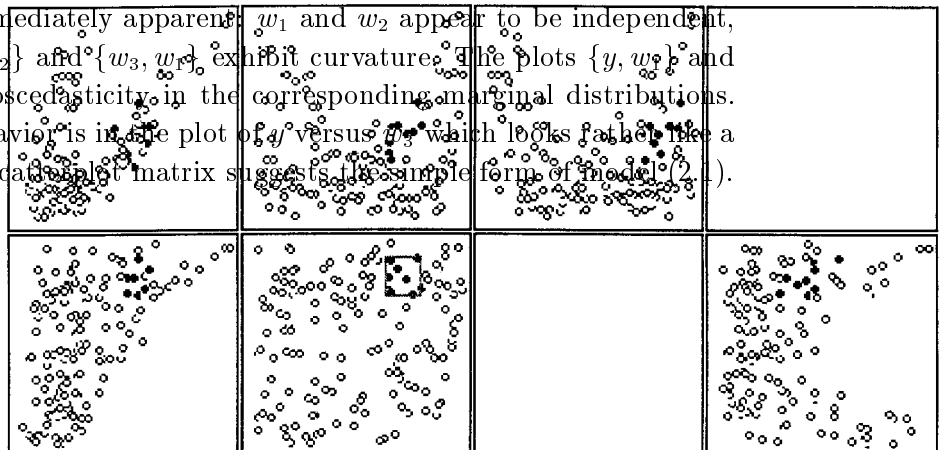


Figure 1. Scatterplot matrix of the data for Example 1.

A scatterplot matrix of 150 observations generated according to this model is given as Figure 1. The highlighted points will be discussed shortly. Imagine inspecting this scatterplot matrix without knowledge of the model. Several characteristics are immediately apparent: w_1 and w_2 appear to be independent, while the plots $\{w_3, w_2\}$ and $\{w_3, w_1\}$ exhibit curvature. The plots $\{y, w_1\}$ and $\{y, w_2\}$ suggest heteroscedasticity in the corresponding marginal distributions. The most curious behavior is in the plot of y versus w_3 which looks rather like a "C". Nothing in the scatterplot matrix suggests the simple form of model (2.1).



The variable x_1 is relatively constant for the highlighted points in the cell $\{w_1, w_2\}$ that are enclosed in a rectangular "brush", so the linked and highlighted points in $\{y, w_3\}$ form a CORE plot and they correspond approximately to a sample from $F(y, w_3|x_1)$. The variation in this CORE plot is substantially smaller than that in the marginal distribution $F(y, w_3)$. Because this happens regardless of brushing position in the cell $\{w_1, w_2\}$, there is considerable covariation between $\{y, w_3\}$ and $\{w_1, w_2\}$. Consequently, the bulk of the variation in $\{y, w_3\}$ can be associated with variation in w_1 and w_2, implying that the net effect of w_3 is relatively small.

While brushing the cell $\{w_1, w_2\}$, the corresponding points in $\{y, w_3\}$ may remain in roughly the same position or may move smoothly around the "C", depending on the movement of the brush. A closer look at the structure of the example can explain such movement in the context of the discussion at the beginning of this section. The distribution of $(y, w_3)|x_1$ is bivariate normal with mean

$$\mu(x_1) = (E(y|x_1), E(w_3|x_1))^T = (1.5(w_1 + w_2) + (w_1 + w_2)^2, (w_1 + w_2)^2)^T \quad (2.2)$$

and constant covariance matrix. Because this distribution depends only on $(w_1 + w_2)$, appropriate CORE plots for studying the net effects of w_3 could be constructed by brushing the variable $(w_1 + w_2)$ while observing $\{y, w_3\}$. The computer programs with which I am familiar allow only rectangular brushes with sides parallel to the coordinate axes, so it does not seem practically possible to use $\{w_1, w_2\}$ as a control plot for simultaneously brushing all points with about the same value of $(w_1 + w_2)$. It is possible to use small square brushes to highlight points falling near lines. When brushing along lines in $\{w_1, w_2\}$ that are orthogonal to $S((1, 1)^T)$, the highlighted points in $\{y, w_3\}$ remain in roughly the same position. On the other hand, when brushing along lines in $\{w_1, w_2\}$ that are parallel to $S((1, 1)^T)$, the highlighted points in $\{y, w_3\}$ move smoothly along sections of the "C" of points.

The essential structure of this example is reflected by the statement

$$(y, w_3) \perp (w_1, w_2)|(w_1 + w_2). \quad (2.3)$$

Finding similar structure in practice may have two important benefits. First, (2.3) means that (w_1, w_2) can be replaced with $(w_1 + w_2)$ in a study of $F(y, w_3|x_1)$, allowing the brushing dimension to be reduced by 1. This has the beneficial effect of increasing the number of observations in each CORE plot and potentially overcoming the sparseness encountered when brushing in high dimensions. Second, (2.3) implies that $y \perp w|(w_1 + w_2, w_3)$ and thus that (w_1, w_2) can be replaced with $(w_1 + w_2)$ in a study of $F(y|w)$ without loss of information.

Several ways of uncovering structure like that in (2.3) will be discussed in the next section. I illustrate one method to conclude this example, relying on the partial knowledge that $F(y, w_3|x_1)$ is a function of x_1 through only its unknown conditional mean $\mu(x_1)$. Figure 2 gives a plot of the estimates $\hat{\mu}(x_{i1})$, $i = 1, \ldots, 150$, obtained by using full second-order quadratic response models in $x_1^T = (w_1, w_2)$ to construct OLS estimates of $E(y|x_1)$ and $E(w_3|x_1)$. The pattern of points closely matches the plane curve traced by $\mu(x_1)$ given in (2.2) as x_1 varies in the square $(-1, 1)^2$. All of the distributions $F(y, w_3|x_1)$ have their centers on the plane curve $\mu(x_1)$ but are otherwise identical.

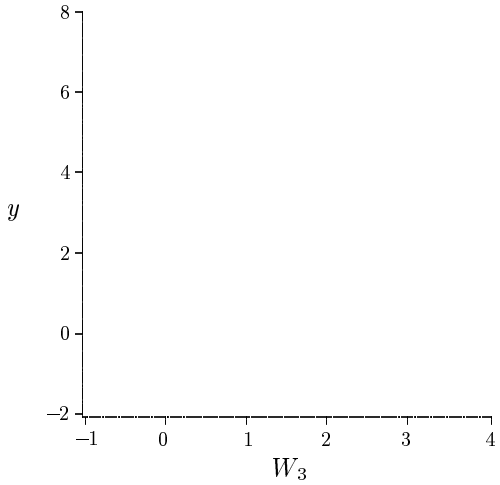


Figure 2. Estimated values of the conditional expectation curve $\hat{\mu}(x_{i1})$ for Example 1.

The scatterplot in Figure 2 can be used as a control plot for brushing just as the cell $\{w_1, w_2\}$ is used in the scatterplot matrix of Figure 1. Brushing along the curve in Figure 2 is essentially equivalent to brushing $(w_1 + w_2)$. Brushing around the plane curve while observing the corresponding linked points in $\{w_1, w_2\}$ and $\{y, w_3\}$ indicates that $F(y, w_3|x_1)$ depends the values of w_1 and w_2 only through their sum, and that the net effect of w_3 is relatively small.

The general conclusion of this example might be reached by brushing around the "C" in $\{y, w_3\}$ without constructing the plot of Figure 2. But this will not be possible when $p_1 > 2$. Moreover, by allowing a little more variation the two arms of the "C" in $\{y, w_3\}$ will merge making informative brushing difficult, while the plane curve of Figure 2 may still be apparent.

3. Distributional Indices

The practical difficulties in brushing a scatterplot of x_1 when $p_1 > 2$ might be overcome by replacing x_1 with a low dimension *distributional index function,* say $\tau(x_1)$. A distributional index function $\tau(x_1)$ serves to index the individual conditional distributions $F(y, x_2|x_1)$ just as x_1 itself does. Indices used in Example 1 are $\tau(x_1) = (w_1 + w_2)$ and $\tau(x_1) = \hat{\mu}(x_1)$. The basic idea is to brush a plot of $\tau(x_1)$ while observing the corresponding CORE plots $\{y, x_2|J\}$ arising in the linked plot $\{y, x_2\}$. In what follows, the case indices of the points at a particular brush location will be indicated by J_τ and the corresponding CORE plot will be indicated by $\{y, x_2|J_\tau\}$.

Distributional indices $\tau(x_1)$ should partition the values of x_1 into equivalence classes with the values in a class corresponding to identical or nearly identical distributions $F(y, x_2|x_1)$. Ideally, $\tau(x_1)$ will be at most three dimensional and have the property that

$$(y, x_2) \perp x_1 | \tau(x_1), \tag{3.1}$$

which is equivalent to $F(y, x_2|x_1) = F(y, x_2|\tau(x_1))$ for all values of x_1. Requiring that $\dim(\tau(x_1)) \leq 3$ is simply a matter of practical necessity. Condition (3.1) insures that no information will be lost by using the distributional index $\tau(x_1)$ in place of x_1. Of course (3.1) is trivially true when $\tau(x_1) = x_1$. Although in practice it may be difficult to satisfy (3.1) exactly, an index plot should be a useful tool for understanding the net effect of x_2 as long as (3.1) is a reasonable approximation. The following lemma may be helpful for constructing low dimensional indices. The justification can be established from Lemmas 4.1-4.3 in Dawid (1979).

Lemma 1. *Let* $\tau^T(x_1) = (\tau_1^T(x_1), \tau_2^T(x_1))$. *If* (a) $x \perp y | (\tau_1(x_1), x_2)$ *and* (b) $x_1 \perp x_2 | \tau_2(x_1)$ *then* (3.1) *holds.*

According to Lemma 1, if we can substitute τ_1 for x_1 in the regression of y on x without loss of information (Condition (a)) and similarly substitute τ_2 for x_1 in the regression of x_2 on x_1 (Condition (b)) then (3.1) holds and τ is a valid distributional index. In the following sections I discuss some possibilities for determining an index $\tau(x_1)$.

Example 1.1. Lemma 1 allows for an easy characterization of Example 1. From the distribution of $w_3|(w_1, w_2)$ it is easy to see that $w_3 \perp x_1 | (w_1 + w_2)$ and from (2.1) it follows that $y \perp w | ((w_1 + w_2), w_3)$. Lemma 1 now applies with $\tau = \tau_1 = \tau_2 = (w_1 + w_2)$. Similarly, $\tau(x_1) = ((w_1 + w_2), (w_1 + w_2)^2)$ and $\mu(x_1)$ are also valid distributional indices.

3.1. Location dependence

In some problems it may be reasonable to assume that $F(y, x_2|x_1)$ depends

on x_1 only through its conditional mean $\mu^T(x_1) = (E(y|x_1), E^T(x_2|x_1))$ so that

$$(y, x_2) \perp x_1 | \mu(x_1), \tag{3.2}$$

which is a special case of (3.1) with $\tau(x_1) = \mu(x_1)$. In practice it will be necessary to estimate $\mu(x_1)$ and this requires estimating two regression functions $E(y|x_1)$ and $E(x_2|x_1)$. This is essentially the structure illustrated in Example 1.

Condition (3.2) requires that $F(y, x_2|x_1)$ depend only on $\mu(x_1)$, but this is less restrictive than requiring $[y - E(y|x_1), x_2 - E(x_2|x_1)] \perp x_1$, so $F(y, x_2|x_1)$ need not be a location distribution. The covariance $\mathrm{Cov}(y, x_2|x_1)$ or any higher order moment may depend on $\mu(x_1)$ under (3.2) and this allows $y|x$ to be a binomial or Possion random variable, for example.

Suppose that analyses of the regressions of y on x_1 and x_2 on x_1 support the conclusions that $y \perp x_1 | E(y|x_1)$ and $x_2 \perp x_1 | E(x_2|x_1)$. This information is not generally sufficient to conclude that (3.2) holds since marginal location dependence need not imply joint location dependence. However, failing to find information in the data to contradict Condition (a) of Lemma 1 with $\tau_1 = E(y|x_1)$ may provide support for (3.2).

3.2. Postanalysis

A net effect plot may be relatively easy to construct after analysis has produced a useful characterization of the regression of y on x. Having a full analysis available need not necessarily mitigate our interest in studying the contribution of x_2 after x_1. To construct an index function in this case, extract from the results of the analysis the lowest dimensional function τ_1 so that Condition (a) of Lemma 1 is satisfied. The second component τ_2 can be obtained from an analysis of the regression of x_2 on x_1. For example, suppose that $p_2 = 1$ and that $F(y|x)$ has been characterized by a homoscedastic generalized additive model (Hastie and Tibshirani (1990)): $\hat{E}(y|x) = a + \Sigma_j g_{1j}(x_{1j}) + g_2(x_2)$, where g_{1j} is the estimated function for the jth coordinate x_{1j} of x_1 and g_2 is the function for x_2. Then set $\tau_1(x_1) = \Sigma_j g_{1j}(x_{1j})$. Assuming next that an estimated linear model $\hat{E}(x_2|x_1) = b_0 + b_1^T x_1$ is found to characterize the regression of x_2 on x_1, the distributional index is just $\tau^T = (\tau_1, \tau_2) = (\Sigma_j g_{1j}(x_{1j}), b_1^T x_1)$. The net effect of x_2 could then be studied by brushing $\{\Sigma_j g_{1j}(x_{1j}), b_1^T x_1\}$ while observing the CORE plots arising in the linked plot $\{y, x_2\}$. Hopefully, τ will generally have dimension 2.

The τ_1 coordinate of τ can be extracted from a full analysis of the regression of y on x, but a second analysis of the regression of x_2 on x_1 may be needed to obtain the second coordinate τ_2 of τ. This could become tiresome when we wish to study net effects for several different predictors x_2. A quick way of constructing

useful τ_2's can be obtained by constraining $\tau_2(x_1)$ to be a linear function of x_1 and using dimension-reduction subspaces.

Let u and v denote generic response and predictor vectors. A subspace S is called a *dimension-reduction subspace for the regression of u on v* if $u \perp v | b^T v$ where b is any basis for S. A dimension-reduction subspace with the smallest dimension is called a *minimum dimension-reduction subspace* for the regression of u on v and is denoted by $S_{u|v}$. Throughout this report I will assume that $S_{u|v}$ is contained in all dimension-reduction subspaces, which implies that $S_{u|v}$ is unique. For further discussion of dimension-reduction subspaces, see Cook (1994) and Li (1991).

Let γ denote a basis for $S_{x_2|x_1}$. Then Condition (b) of Lemma 1 holds with $\tau_2(x_1) = \gamma^T x_1$. The essential problem is now to estimate the subspace $S_{x_2|x_1}$. Once this is done we can choose a basis $\hat{\gamma}$ for the estimate and set $\tau_2(x_1) = \hat{\gamma}^T x_1$ for use in practice. Any basis will do theoretically, but practically it may be worthwhile to avoid colinearity in the index plot by insuring that the coordinates of $\tau_2(x_1)$ are uncorrelated or approximately so.

Recently regression methods have become available to estimate $S_{x_2|x_1}$ when $p_2 = 1$. Sliced inverse regression (aka SIR and slicing regression) as recently suggested by Li (1991) and Duan and Li (1991) is perhaps the first choice. SIR requires that $E(x_1|\gamma^T x_1)$ be linear in $\gamma^T x_1$. Although SIR is not very sensitive to modest violations of that condition, the re-weighting method described in Cook and Nachtsheim (1994) may be used to extend applicability. Other methods include SAVE (Cook and Weisberg (1991)) and pHd (Li (1992)).

3.3. Exploration: Univariate SIR

In some problems we may wish net effect plots in the exploratory stage of an analysis where sufficient knowledge of the regression of y on x is not yet available for τ_1 to be determined as described in Section 3.2. This is, perhaps, the most difficult situation because we must know how to construct a distributional index that can replace x_1 in the regression of y on x without a complete analysis in the first place. Nevertheless, it is possible to use inverse regression methods in this case.

One possibility is to apply SIR twice, once to the regression of y on x and once to the regression of x_2 on x_1. The latter application will provide $\tau_2(x_1)$ as discussed in the previous section. The former application will provide an estimate of $S_{y|x}$ from which $\tau_1(x_1)$ can be extracted according to Condition (a) of Lemma 1: Let $\beta = (\beta_1^T, \beta_2^T)^T$ be a partitioned basis for $S_{y|x}$ so that $\beta^T x = (\beta_1^T x_1 + \beta_2^T x_2)$. Then set $\tau_1(x_1) = \eta_1^T x_1$ where η_1 is any basis for $S(\beta_1)$. Determining a basis for

$S(\beta_1)$ is a necessary step since β_1 need not be of full rank. It is also possible to use inverse regression methods to estimate $S(\eta_1) = S(\beta_1)$ directly, where $\eta_1^T x_1$ comprises the fewest linear combinations of x_1 so that $y \perp x | (\eta_1^T x_1, x_2)$, as discussed in Cook (1994). Let $r_{1|2} = x_1 - E(x_1|x_2)$. If

$$y \perp r_{1|2} | \eta_1^T r_{1|2} \tag{3.3}$$

then $S(\eta_1)$ might be estimated by applying SIR to the regression of y on sample residuals $\hat{r}_{1|2} = x_1 - \hat{E}(x_1|x_2)$. Condition (3.3) holds for normally distributed predictors and is often a useful approximation in practice. For further discussion, see Cook (1992, 1994) and Cook and Wetzel (1993).

These separate inverse regression procedures can result in distributional indices with dimension larger than necessary when $S(\eta_1) \cap S_{x_2|x_1} \neq S$ (origin). An adaptation of inverse regression methods to the bivariate regression of (y, x_2) on x_1 may help avoid this possibility.

3.4. Exploration: Bivariate SIR

Following the ideas in Section 3.2 that lead to SIR, we restrict the full index $\tau(x_1)$ to be a linear function of x_1. Let η be a basis for $S_{(y,x_2)|x_1}$ so that

$$(y, x_2) \perp x_1 | \eta^T x_1. \tag{3.4}$$

As before we need an estimated basis $\hat{\eta}$ for $S_{(y,x_2)|x_1}$. Once found, we can set $\tau(x_1) = \hat{\eta}^T x_1$ for use in practice. To avoid colinearity in the index plot, we could again choose $\hat{\eta}$ to insure that the correlation between the elements of τ is small.

Although Li (1991) describes SIR in the context of regression problems with a univariate response variable, the same theory applies when the response variable is bivariate. The methodological change entails double slicing (Li (1991, p. 339)) the observations on (y, x_2) in the plane rather than slicing a univariate response. In effect, the bivariate response (y, x_2) is replaced by a discretized bivariate response $(\tilde{y}, \tilde{x}_2)$ say, assuming that $S_{(y,x_2)|x_1} = S_{(\tilde{y},\tilde{x}_2)|x_1}$. Once the bivariate slices are constructed, the methodology follows the steps for SIR.

Example 1.2. SIR was applied to the bivariate regression in Example 1 by first partitioning the 150 observations on y into 5 slices each containing 30 observations. The 30 values of w_3 in each slice were then partitioned into 5 slices of 6 observations each. In this way the data were partitioned into 25 slices of 6 observations each. The results from SIR strongly indicate that the dimension of $S_{(y,x_2)|x_1}$ is 1, with estimated basis $\hat{\eta}^T = (0.707, 0.708)$ which is nearly identical to the population basis $\eta^T = (1, 1)$.

4. Combining CORE Plots

The action of brushing a plot of the distributional index $\tau(x_1)$ while observing the corresponding CORE plots $\{y, x_2|J_\tau\}$ provides two kinds of information. First, each CORE plot corresponds approximately to a sample from $F(y, x_2|x_1)$. Studying a fixed CORE plot provides information about the conditional regression structure of y on x_2 at the selected value for x_1. It can be interpreted as any regression scatterplot, understanding that x_1 is essentially fixed. Second, brushing while observing the corresponding changes in the CORE plots provides information on how the conditional regression structure changes with the values of x_1 in the sample. In some problems key characteristics of the conditional regression structure may be constant, allowing individual CORE plots to be combined into a single net effect plot that gives a better impression of the constant aspects of the conditional regression structure.

The problem of combining CORE plots translates into the problem of combining the conditional random variables $(y, x_2)|x_1$ over the values of x_1. Informative combinations of these variables need not be easy to construct since there is no reason that $(y, x_2)|(x_1 = c_1)$ should be at all similar to $(y, x_2)|(x_1 = c_2)$, particularly if there are interactions. Nevertheless, a first method is to shift the individual CORE plots so that they have the same mean. The plot $\{r_{y|1}, r_{2|1}\} = \{y - E(y|x_1), x_2 - E(x_2|x_1)\}$ does just that since the expectation of $(r_{y|1}, r_{2|1})$ is at the origin for all values of x_1. The translation to $\{r_{y|1}, r_{2|1}\}$ shifts the individual conditional distributions to coincide at the origin and leaves the conditional covariance structure unchanged. In Example 1 this corresponds to sliding the individual distributions $F(y, x_2|x_1)$ along the conditional mean curve corresponding to Figure 2 until they coincide at the origin. If

$$(r_{y|1}, r_{2|1}) \perp x_1 \tag{4.1}$$

then the translation has the desired effect of combining the conditional distributions over x_1 while leaving the conditional regression structure intact, except for the predictable shift in location. Otherwise, combining CORE plots in this fashion may be undesirable.

If $E(y|x_1)$ and $E(x_2|x_1)$ are estimated by using OLS linear regressions of y on x_1 and x_2 on x_1, respectively, then the resulting combined CORE plot is just a standard added variable plot for x_2 after x_1. Thus an added variable plot is an instance of a net effect plot when (4.1) holds, and $E(y|x_1)$ and $E(x_2|x_1)$ are both linear in x_1. There is no restriction that $E(y|x)$ be linear in x, which is a bit of freedom that allows added variable plots to be used as diagnostics. Using the notation described at the beginning of Section 2, $\{e_{y|1}, e_{2|1}\}$ is a generic representation of an added variable plot for x_2 after x_1.

Example 1.3. Condition (4.1) holds in Example 1 so it is possible to construct a single net effect plot while preserving the conditional regression structure. Figure 3 gives the sample plot $\{y_i - \hat{E}(y|x_{i1}), w_{i3} - \hat{E}(w_3|x_{i1})\}$, where the conditional means were estimated by using the quadratic regressions described near the end of Example 1. This figure is interpreted as showing the overall net effect of $x_2 = w_3$; that is, the regression of y on w_3 with x_1 fixed, the particular value of x_1 being important only for determining the location of the point cloud. Again the net effect of w_3 does not seem very strong relative to the variation in $\{y, w_3\}$.

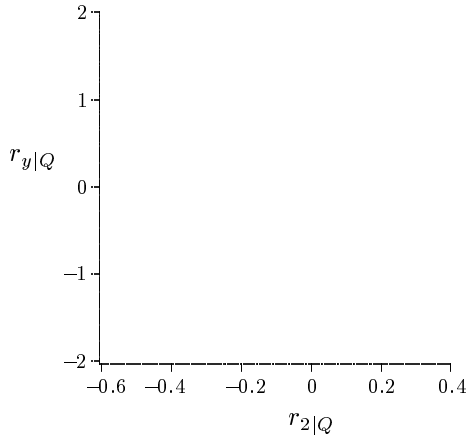


Figure 3. Scatterplot of the OLS residuals $r_{y|Q}$ from the regression of y on the full quadratic predictor Q in (w_1, w_2) versus the OLS the residuals $r_{y|Q}$ from the regression of w_3 on Q.

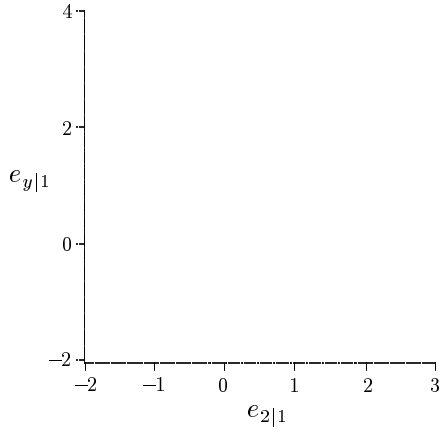


Figure 4. Added variable plot for w_3 in Example 1: Scatterplot of the OLS residuals $e_{y|1}$ from the regression of y on $x_1 = (w_1, w_2)^T$ versus the OLS residuals $e_{2|1}$ from the regression of $x_2 = w_3$ on x_1.

The regression functions $E(y|x_1)$ and $E(w_3|x_1)$ in this example were estimated by using quadratics in x_1, although the full regression function $E(y|x)$ is linear in x. If first-order linear regression models had been used to estimate $E(y|x_1)$ and $E(w_3|x_1)$, resulting in an added variable plot, the plot would display more than the net effect of w_3 since we would not have fully accounted for (conditioned on) x_1. The usual added variable plot for w_3 is shown as Figure 4. Clearly the impressions left by Figures 3 and 4 are quite different, with the plot in Figure 4 suggesting the stronger net effect.

The plots of Figures 3 and 4 illustrate the general conclusion that added variable plots tend to over-estimate net effects, unless $E(y|x_1)$ and $E(x_2|x_1)$ are both linear functions of x_1.

Estimated versions of the plot $\{r_{y|1}, r_{2|1}\}$ will provide visual information on the net effect of x_2 after x_1 when only the location of $(y, x_2)|x_1$ varies with the value of x_1 since then $(r_{y|1}, r_{2|1}) \perp x_1$. The regression functions $E(y|x_1)$ and $E(x_2|x_1)$ could be estimated by using OLS as in Example 1, a robust estimation method or generalized additive models, for example. When $(r_{y|1}, r_{2|1}) \perp x_1$ fails the plot $\{r_{y|1}, r_{2|1}\}$ can be viewed as an average over plots of the form $\{r_{y|1}, r_{2|1}|J_\tau\}$ where the averaging is with respect to the marginal distribution of $\tau(x_1)$. Whether this average interpretation is useful depends on the nature of the effects involved. It is certainly possible to have an extreme interaction where the CORE plots $\{r_{y|1}, r_{2|1}|J_\tau\}$ show systematic trends in x_2 and yet no clear relations are evident in the combined plot $\{r_{y|1}, r_{2|1}\}$ so that x_2 is unimportant on the average.

5. Additional Examples

5.1. Functionally related predictors

The discussion of the previous section shows that there are important differences between added variable plots and net effect plots, although an added variable plot can serve as a net effect plot when $E(y|x_1)$ and $E(x_2|x_1)$ are both linear functions of x_1. Condition (4.1) requires that sample versions of the net effect plot $\{r_{y|1}, r_{2|1}\}$ be constructed by using estimates of the conditional expectations $E(y|x_1)$ and $E(x_2|x_1)$. Conditional expectations are not the basis for added variable plots, however, and this can have important implications when considering functionally related predictors.

In their discussion of adjusted variable plots (added variable plots) Chambers et al.(1983) use a data set relating the tar content (Tar) of a gas to the

temperature (T) of a chemical process and the speed (S) of a rotor. Because speed was expected to have a nonlinear effect on tar content, S^2 was used as an additional predictor to form the initial model $\text{Tar} = \beta_0 + \beta_1 T + \beta_2 S + \beta_{22} S^2 + \varepsilon$. To study the contribution of S^2, Chambers et al. (1983, p. 273) use the adjusted variable plot for S^2, as shown in Figure 5.

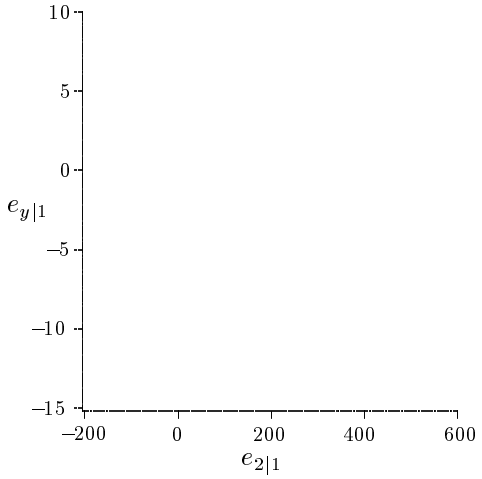


Figure 5. Adjusted (Added) variable plot for S^2: Scatterplot of the OLS residuals $e_{y|1}$ from the regression of $y = \text{Tar}$ on $x_1 = (S, T)$ versus the OLS residuals $e_{2|1}$ from the regression of $x_2 = S^2$ on x_1.

Regardless of the adequacy of the initial model, the adjusted variable plot of Figure 5 provides visual information on the numerical calculation of the coefficient of S^2 from an OLS fit (Draper and Smith (1966, Section 4.1)). But the adjusted variable plot for S^2 cannot be interpreted as a net effect plot as defined here. Because $S^2 - E(S^2|S,T) = 0$, the horizontal coordinate of the plot $\{r_{y|1}, r_{2|1}\}$ for S^2 is identically zero, which is just a reflection of the fact that the second coordinate of the conditional variable $(\text{Tar}, S^2)|(S,T)$ is degenerate. Thus we again see that there are important differences between net effect plots and adjusted variable plots.

The basic issue in this example is to understand the contribution of S to the conditional distributions $F(\text{Tar}|S,T)$, and this can be done by studying the regression of Tar on S at fixed values of T. The CORE plots $\{\text{Tar}, S|J\}$ obtained by brushing T are not very helpful because there are only 31 data points. However, the net effect plot $\{r_{\text{Tar}|T}, r_{S|T}\}$ may be useful. Inspection of the scatterplots $\{\text{Tar}, T\}$ and $\{S, T\}$ suggest that $E(\text{Tar}|T)$ and $E(S|T)$ are both reasonably lin-

ear in T, and the 3D scatterplot $\{T, (e_{\text{Tar}|T}, e_{S|T})\}$ supports the notion that T and $(e_{\text{Tar}|T}, e_{S|T})$ are independent. The net effect plot $\{e_{\text{Tar}|T}, e_{S|T}\}$, which is also an added variable plot in this case, is shown in Figure 6 along with a quadratic fit.

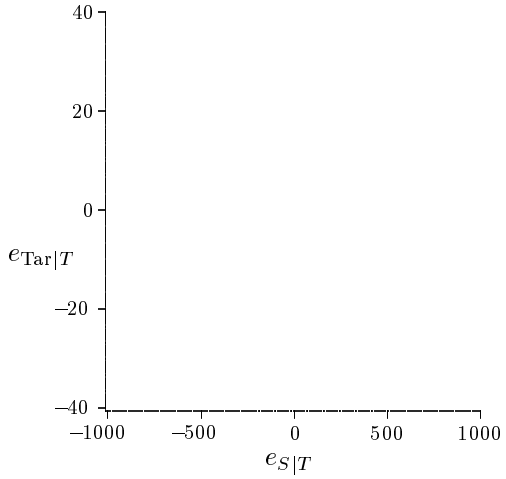


Figure 6. Net effect plot for S with a quadratic fit.

As in Figure 3, the net effect plot in Figure 6 provides visual information on the conditional regression structure between Tar and S at a fixed value of T, the particular value of T serving only to determine the location of the point cloud. The dominant conditional relationship between Tar and S is linear, but the quadratic fit indicates that there may be curvature as well. The curvature is largely a consequence of a few points at the extremes. Removing the linear trend from the plot in Figure 6 provides additional visual support for this conclusion.

5.2. Exploration with SIR

For this example I use the mussel data as reported by Cook and Weisberg (1994). The data consist of observations on the length L, width W, height H, shell mass S and muscle mass M for a sample of 82 Horse Mussels collected in the Marlborough Sounds located off the Northeast coast of New Zealand's South Island. The response variable is muscle mass M, the edible portion of the mussel. The purpose of this example, which focuses on the net effect of shell mass M, is to illustrate a few ideas for constructing net effect plots in the model development stage of an analysis. For notational convenience, let $x_1 = (L, W, H)^T$.

An inspection of a 3D scatterplot of x_1 indicates that all conditional expectations of the form $E(x_1|a^T x_1)$ are strongly dominated by linear trends, supporting

the application of bivariate SIR for reducing the dimension of x_1. Bivariate SIR was applied with 25 slices of 3-4 observations each, following the description in Example 1. There was only one significant linear combination of the predictors, $b^T x_1 = (.15L + .85W + .50H)$, indicating that condition (3.4) – $(M, S) \perp x_1 | b^T x_1$ – may hold to a useful approximation. Thus the net effect of S might be studied by brushing the single variable $b^T x_1$ while observing the corresponding CORE plots arising in the linked plot $\{M, S\}$.

A scatterplot matrix of M, S, and $b^T x_1$ is shown in Figure 7. The CORE plot corresponding to the slice on $b^T x_1$ suggests that much of the variation in the plot $\{M, S\}$ can be attributed to variation in $b^T x_1$, although an effect of S is still evident within the CORE plot. While brushing the scatterplot matrix in Figure 7 provides useful visual information, a single net effect plot constructed using the ideas of Section 4 may still be worthwhile.

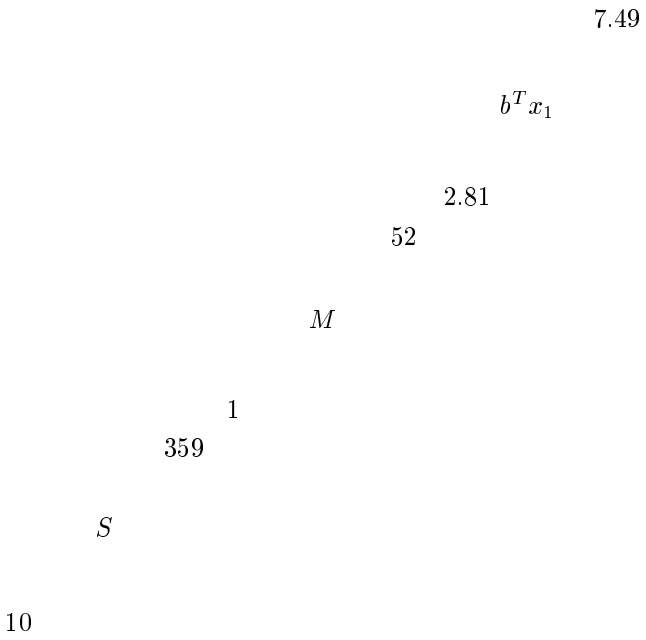


Figure 7. Scatterplot matrix for the mussel data.

Condition (4.1) provides one possibility for construction a combined net effect plot to gain visual information on $F(M, S|b^T x_1)$. The scatterplot matrix of Figure 7 indicates that $E(M|b^T x_1)$ and $E(S|b^T x_1)$ are nonlinear functions of the value of $b^T x_1$ so a standard added variable plot would likely over-estimate the net effect

of S. The conditional expectations do seem to be well-approximated by fitting quadratic polynomials in $b^T x_1$. Nevertheless, (4.1) might still not be sufficiently approximated because of the nonconstant variance that is evident in the plots $\{S, b^T x_1\}$ and $\{M, b^T x_1\}$.

Alternatively, power transformations of S and M might be used to force linear, homoscedastic relationships with $b^T x_1$. Because SIR is in invariant under strictly monotonic transformations of the response, transforming S and M at this stage will not change the previous results. Transforming both S and M to the 0.2 power seems to do a reasonable job, although two observations are highlighted as potential outliers on the transformed scale. Because the conditional expectations $E(M^{0.2}|b^T x_1)$ and $E(S^{0.2}|b^T x_1)$ are essentially linear, the added variable plot for $S^{0.2}$ shown in Figure 8 is reasonable for studying the net effect of $S^{0.2}$. The highlighted points correspond to the potential outliers noted earlier. The net effect plot in Figure 8 provides a visualization of the regression of $M^{0.2}$ on $S^{0.2}$ at a fixed value of x_1. The particular value of x_1 serves to determine the location of the point cloud through the linear combination $b^T x_1$ determined by using a bivariate SIR procedure.

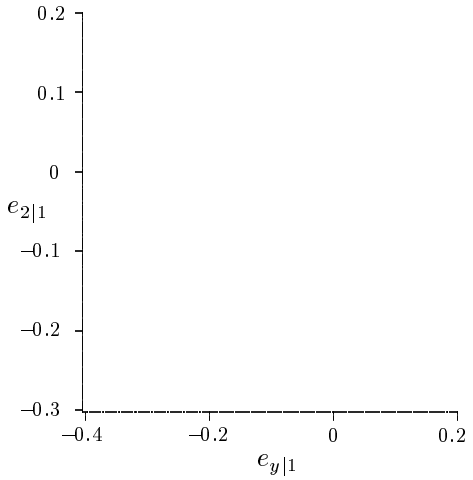


Figure 8. Added variable plot showing the net effect of $S^{0.2}$ in the regression of $M^{0.2}$ on $S^{0.2}$ and $b^T x_1$, along with the OLS fitted line for reference.

6. Postlude

Scatterplot brushing is a useful graphical tool for understanding the net effects of predictors in regression problems. Direct application is limited to regression problems with at most three or four predictors, however. When there

are more predictors we can always display all the data in a scatterplot matrix and then brush a single cell while observing the changing CORE plots in other cells of the matrix. But this technique still only allows investigation of marginal regression problems with at most three predictors since we cannot condition on more than two predictors simultaneously.

The essential proposal in this paper is to reduce the number of brushing dimensions by replacing the conditioning predictors x_1 with a low dimension distributional index function $\tau(x_1)$ that discards extraneous information from x_1 and this has the potential to allow net effects to be studied via brushing in regression problems with many predictors. The basic ideas do not depend on the nature of the response and they apply to generalized linear models as well as models with additive errors.

Acknowledgements

Research for this article was supported in part by grant DMS-9212413 from the National Science Foundation. The author thanks Ray Carroll, Ker-Chau Li and Sandy Weisberg for helpful comments on an earlier version of the manuscript. Nate Wetzel provided programming assistance.

References

Becker, R. A. and Cleveland, W. S. (1987). Brushing scatterplots. Technometrics **29**, 127-142.

Becker, R. A., Cleveland, W. S. and Wilks, A. R. (1987). Dynamic graphics for data analysis (with discussion). Statist. Sci. **2**, 355-395.

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

Chambers, J. M. and Hastie, T. J. (1992). Statistical Models in S. Pacific Grove, Wadsworth & Brooks/Cole, CA.

Cook, R. D. (1992). Graphical regression. In Computational Statistics, Vol 1 (Edited by Y. Dodge and J. Whittaker), 11-22, Physica-Verlag, Heidelberg.

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

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 Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, New York.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by K. C. Li. J. Amer. Statist. Assoc. **86**, 328-332.

Cook, R. D. and Weisberg, S. (1994). An Introduction to Regression Graphics. John Wiley, New York.

Cook, R. D. and Wetzel, N. (1993). Exploring regression structure with graphics (with discussion). TEST **2**, 1-57.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). J. Roy. Statist. Soc. Ser.B **41**, 1-31.

Draper, N. R. and Smith, H. (1966). Applied Regression Analysis. John Wiley, New York.
Duan, N. and Li. K. C. (1991). Slicing regression: A link-free regression method. Ann. Statist. **19**, 505-530.
Hastie, T. J. and Tibshirani, R. J. (1990). Generalized Additive Models. Chapman and Hall, London.
Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). 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**, 1025-1039.
Tierney, L. (1990). LISP-STAT. John Wiley, New York.

Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108, U.S.A.

(Received July 1993; accepted November 1994)