Graphical Detection of Regression Outliers and Mixtures.

R. Dennis Cook  
Department of Applied Statistics  
University of Minnesota  
1994 Buford Avenue  
St. Paul, MN 55108, USA

Regressions in practice can include outliers and other unknown subpopulation structure. For example, mixtures of regressions occur if there is an omitted categorical predictor like gender, species or location and different regressions occur within each category. A lurking variable that has an important effect but is not present among the predictors under consideration (Box 1966) can seriously complicate a regression analyses. Regression structure with lurking variables is illustrated in Figure 1a which is a stylized representation of subpopulation structures in a regression with response $Y$ predictors $X_k$. The contours A, C and E represent different subpopulation regressions. Point B represents an isolated outlier while the circular contours D represent an outlying cluster. The regression illustrated in the figure consists of a mixture of five distinct regressions, one for each of the four subpopulations and one for the isolated outlier.

Figure 1. (a) Stylized Representation of Regression Mixtures. (b) Summary Plot from SAVE Analysis of the Bank Note Data: × Authentic, ◦ Counterfeit.

In this article we discuss graphical methods for diagnosing the kind of structure illustrated in Figure 1a without prior knowledge of it and without requiring a model. Our approach rests on the theory of regression graphics based on central dimension-reduction subspaces.

1. Central Subspaces

Dimension reduction without loss of information is a dominant theme of regression graphics. The goal of a regression study is to infer about the conditional distribution of the response $Y$ given the $p \times 1$ vector of predictors $X$: How does the conditional distribution of $Y \mid X$ change with the value assumed by $X$? Recently developed dimension-reduction methods approach this
question through a population super parameter called the \textit{central subspace} (Cook 1994, 1998) and denoted by \( S_{Y|X} \). Letting the columns of the matrix \( \eta \) be a basis for \( S_{Y|X} \), the central subspace is the smallest subspace of \( \mathbb{R}^p \) such that \( Y \perp \!\!\!\!\perp X|\eta^T X \), where \( \perp \!\!\!\!\perp \) indicates independence. The statement is thus that \( Y \) is independent of \( X \) given \( \eta^T X \). It is equivalent to saying that the distribution of \( Y|X \) is the same as that of \( Y|\eta^T X \) for all values of \( X \) in its marginal sample space.

Knowledge of the central subspace is useful for parsimoniously characterizing how the distribution of \( Y|X \) changes with the value of \( X \) and, in particular, for identifying the kind structure illustrated in Figure 1a. If \( S_{Y|X} \) was known, the \textit{minimal sufficient summary plot} of \( Y \) versus \( \eta^T X \) could then be used to guide subsequent analysis and to identify subpopulation structure. If an estimated basis \( \hat{\eta} \) of \( S_{Y|X} \) was available then the summary plot of \( y \) versus \( \eta^T x \) could be used similarly.

Summary plots based on estimates of \( S_{Y|X} \) can be of significant value in many phases of a regression analysis, particularly during the initial phases when an adequate parsimoniously parameterized model is not yet available, and during the residual-based model criticism phase. Methods of estimating the central subspace or portions thereof include some standard methods like ordinary least squares, graphical regression (Cook 1998), principal Hessian directions (Li 1992), sliced average variance estimation (SAVE, Cook and Weisberg 1991), and sliced inverse regression (Li 1991). Cook and Weisberg (1999) gave an introductory account of studying regressions via central subspaces. A comprehensive discussion is available in Cook (1998).

2. Subpopulation Regressions

Without loss of generality, we work mostly in terms of the standardized predictor \( Z = \{\mathrm{Var}(X)\}^{-1/2}(X - E(X)) \), where \( \mathrm{Var}(X) \) is assumed to be positive definite. Then \( S_{Y|X} = \{\mathrm{Var}(X)\}^{-1/2}S_{Y|Z} \). Thus, any basis for \( S_{Y|Z} \) can be back-transformed to a basis for \( S_{Y|X} \). Replacing the population mean and covariance matrix by their usual estimates yields the corresponding sample version \( Z_i \).

To introduce subpopulation regressions, we assume that the outcome of the experiment depends on three random variables: the scalar response \( Y \), the \( p \times 1 \) vector of predictors \( Z \) and a binary indicator \( I \) that identifies the subpopulation, with \( I = 1 \) or 2. Considering only two subpopulations is intended to focus the discussion, and is not restrictive since either of the two subpopulations can itself be composed of multiple subpopulations. Although the structure of the population depends on three random variables, we assume that only \( Y \) and \( Z \) are observable. Our goal is to investigate how a graphical analysis of observations on \((Y,Z)\) can be used to uncover the regression structure in the two subpopulations. Thinking of one subpopulation as outlying the other, this goal can be rephrased as finding how a graphical analysis can be used to find outliers. The binary predictor \( I \) could correspond to a lurking variable, or it could indicate multiple subpopulations characterized by outliers and lurking variables, as illustrated in Figure 1a.

Averaging over \( I \), the regression of \( Y \) on \( Z \) can be represented as a mixture of the two subpopulation regressions, \( \mathcal{P}(Y \leq y|Z) = \sum_{i=1}^2 \mathcal{P}(I = i|Z) \mathcal{P}(Y \leq y|Z, I = i) \). This suggests that we need three central subspaces to characterize \( S_{Y|Z} \) in terms of the subpopulation regressions (Cook and Critchley 1998). Let \( S_{Y|Z}^i \) denote the central subspace for the regression of \( Y|(I = i) \) on \( Z|(I = i) \), \( i = 1, 2 \), and let \( S_{Z|Z} \) denote the central subspace for the regression of the binary indicator \( I \) on \( Z \). We use the columns of the matrices \( \beta_0, \beta_1 \) and \( \beta_2 \) to denote bases for \( S_{Z|Z}, S_{Y|Z}^1 \) and \( S_{Y|Z}^2 \).
Then Cook and Critchley (1998) show that \( S_{Y\mid Z} \subseteq S_{\mathcal{I}\mid Z} + S^1_{Y\mid Z} + S^2_{Y\mid Z} \). Thus, \( S_{Y\mid Z} \) is always contained in the direct sum of the three component central subspaces. Without constraints it is possible that it is in fact a proper subset, and thus that it loses information on the component subspaces. However, under weak technical requirements and the condition \( S(\beta_i) \cap S(\beta_j) \subseteq S(\beta) \) it can be shown that (Cook and Critchley 1998)

\[
S_{Y\mid Z} = S_{\mathcal{I}\mid Z} + S^1_{Y\mid Z} + S^2_{Y\mid Z}
\]  

(1)

Result (1) indicates that the methods mentioned in Section 1 for estimating \( S_{Y\mid Z} \) can be expected to find subpopulation structure. In effect, \( S_{Y\mid Z} \) expands automatically to incorporate outliers and regression mixtures. Thus, methods of estimating \( S_{Y\mid Z} \) can be used to identify these phenomena, without specifying a model. Experience has shown that methods for estimating the central subspace are sensitive to outliers. This sensitivity might be viewed as a disadvantage following traditional reasoning. However, Cook and Critchley (1998) argue that it can be an advantage, enabling the analyst to construct low-dimensional summary plots that incorporate mixtures and outliers without the need for a model. Sequential outlier deletion seems to fit nicely into the structure developed in this article because the removal of data from a specific subpopulation does not change the definitions of the remaining subpopulations.

3. Response Outliers

In this section we consider the implications of the previous results for regressions with only response outliers which are characterized by the conditions \( \mathcal{I} \perp Z \) and \( Y \perp Z \mid (\mathcal{I} = 2) \). It follows that \( S_{\mathcal{I}\mid Z} = S^2_{Y\mid Z} = S(0) \) and \( S_{Y\mid Z} = S^1_{Y\mid Z} \). Thus, response outliers have no effect on \( S_{Y\mid Z} \). But they can certainly change other aspects of the regression like the mean or variance function. For example, consider a linear regression in which with probability \( \pi \) we observe

\[
Y\mid(Z, \mathcal{I} = 1) = \alpha + \beta^T Z + \varepsilon
\]

and with probability \( 1 - \pi \) we obtain a response outlier \( Y\mid(Z, \mathcal{I} = 2) = W \), where \( W \perp (Z, \varepsilon) \), \( \varepsilon \perp Z \) and \( \varepsilon E(\varepsilon) = 0 \). Now straightforward application of ordinary least squares results in an unbiased estimate \( b \) of \( \pi \beta \), which of course is a biased estimate of \( \beta \) but nevertheless spans the same subspace as \( \beta \). Robust estimation methods may also be useful in this linear regression setting.

These results can be important in practice. For example, suppose that \( d = 1 \), and that \( E(Z|\beta^T Z) \) is linear or approximately so. Let \( b \) denote the \( p \times 1 \) vector of coefficients from the ordinary least squares regression of \( y_j \) on \( z_j \). Then \( S(b) \) is a consistent estimate of \( S_{Y\mid Z} \) (Cook 1998, Chapter 8), which implies that we may be able to infer usefully about both the regression structure and the response outliers from a summary plot of \( y_j \) versus \( b^T z_j \).

4. Swiss Bank Notes

Flury and Riedwyl (1988, p. 5) gave a data set on counterfeit Swiss bank notes. The response variable is a note’s authenticity, \( Y = 0 \) for genuine notes and \( Y = 1 \) for counterfeit notes. There are 6 predictors, each giving a different aspect of the size of a note: length of bottom edge, diagonal length, left edge length, length at center, right edge length and top edge length.

Application of \textit{SAVE} (Cook and Weisberg 1991) to the bank note data gave a clear indication that the central subspace has dimension 2. The summary plot of the first two \textit{SAVE}
predictors is shown in Figure 1b, with points marked according to a note’s authenticity. There are two striking features to Figure 1b: First, there seems to be an outlying authentic note. This may be a mislabeled counterfeit note, or an indication of a low-frequency mode for authentic notes. In view of the results discussed previously, we should perhaps not be surprised to find such outlying points in summary plots. Second, the counterfeit notes were apparently drawn from two distinct subpopulations, suggesting the presence of a lurking variable. The counterfeit subpopulations could reflect notes from two different sources, or a change in operational settings by a single counterfeiter.

REFERENCES


RÉSUMÉ