Statistica Sinica **6**(1996), 433-454

REGRESSION TRANSFORMATION DIAGNOSTICS FOR EXPLANATORY VARIABLES

Bo-Cheng Wei and Fred J. Hickernell

Southeast University and Hong Kong Baptist University

Abstract. Two types of diagnostics are presented for the transformation of explanatory variables in regression. One is based on the likelihood displacement proposed by Cook and Weisberg (1982) for assessing the influence of individual cases on the maximum likelihood estimate of a transformation parameter. The other is based on the local influence theory proposed by Cook (1986) for assessing the influence of small perturbations on the parameter estimates. Computations are performed on two data sets to illustrate the usefulness of these diagnostics.

Key words and phrases: Case deletions, likelihood displacement, local influence, perturbations.

1. Introduction

Transformations are commonly used in regression analysis. When a parametric transformation family, such as the Box-Cox power transformation, is used, the maximum likelihood estimate of the parameter is usually sensitive to perturbations of the data. Many diagnostic methods have been proposed in the literature to estimate this sensitivity (see Cook and Wang (1983), Atkinson (1983, 1986), Carroll and Ruppert (1987), Hinkley and Wang (1988), Lawrance (1988) and Tsai and Wu (1990)). Most authors have been concerned about transformations of the response.

In contrast, transformation diagnostics for explanatory variables have been studied to a lesser degree. Box and Tidwell (1962) proposed using constructed variables and added variable plots to guide the selection of transformations of explanatory variables. Their procedure is discussed by Cook and Weisberg (1982). Cook (1987) used the local influence approach to derive diagnostics for partially linear models, which include transformation of a single explanatory variable as a special case.

This paper proposes two types of diagnostics for assessing the influence of individual cases on the maximum likelihood estimates of transformation parameters for explanatory variables. Section 2 presents a diagnostic based on an approximation to the likelihood displacement (Cook and Weisberg (1982)) when one or more observations is deleted. Section 3 gives a number of diagnostics based on the local influence approach (Cook (1986, 1987)) for assessing the influence of small perturbations of the data. Specifically, we consider perturbations of case-weights, explanatory variables, and transformed explanatory variables. In Section 4, two numerical examples are presented to illustrate the usefulness of the diagnostics derived.

The starting point for this study is the linear model,

$$Y = X\beta + \varepsilon, \tag{1.1}$$

where $Y = (y_1, \ldots, y_n)^T$ denotes the observed response, X is a known full rank $n \times p$ data matrix with columns X_j $(j = 1, \ldots, p)$, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the unknown parameter vector, and ε is the random error.

In many cases model (1.1) may be improved by transforming one or more explanatory variables. For example, one might replace the first column of X by $(h(x_{11}, \lambda_1), \ldots, h(x_{n1}, \lambda_1))^T$, where $h(u, \nu)$ is a family of transformations indexed by a real parameter ν. Without loss of generality we may suppose that the first q explanatory variables are transformed. The new model can be written as

$$Y = X(\lambda)\beta + \varepsilon, \tag{1.2}$$

where $\varepsilon \sim N(0, \sigma^2 I)$, λ is a $q \times 1$ vector parameter, $X(\lambda) = (X_1(\lambda_1), \ldots, X_q(\lambda_q), X_{q+1}, \ldots, X_p)$, and $X_j(\lambda_j) = (h(x_{1j}, \lambda_j), \ldots, h(x_{nj}, \lambda_j))^T$ $(j = 1, \ldots, q)$. We assume that $X(\lambda)$ has full rank and that $\beta^{(1)} = (\beta_1, \ldots, \beta_q)^T \neq 0$.

For any fixed λ the maximum likelihood estimates of the regression coefficient and the variance of the noise are

$$\hat{\beta}(\lambda) = [X^T(\lambda)X(\lambda)]^{-1}X^T(\lambda)Y, \quad \hat{\sigma}^2_{\text{MLE}}(\lambda) = \frac{1}{n}s(\lambda), \tag{1.3}$$

respectively, where $Q(\lambda) = I - P(\lambda)$, $P(\lambda)$ is the projection matrix for $X(\lambda)$, and $s(\lambda)$ is the residual sum of squares,

$$s(\lambda) = Y^T Q(\lambda) Y. \tag{1.4}$$

The profile log-likelihood for λ is

$$L(\lambda) = -\frac{n}{2}\log[s(\lambda)], \tag{1.5}$$

omitting constant terms. So $\hat{\lambda}$, the maximum likelihood estimate of λ, must minimize the residual sum of squares. The corresponding estimates of β and σ^2 are $\hat{\beta} = \hat{\beta}(\hat{\lambda})$ and $\hat{\sigma}^2_{\text{MLE}} = \hat{\sigma}^2_{\text{MLE}}(\hat{\lambda})$ respectively.

2. Likelihood Displacement

To assess the influence of individual cases on the transformation parameter $\hat{\lambda}$ one may consider the case-deletion model,

$$Y_{(i)} = X_{(i)}(\lambda)\beta + \varepsilon. \tag{2.1}$$

The subscript (i) denotes the quantities with case i deleted. An alternative formulation of (2.1), usually more convenient for calculations, is the mean-shift outlier model,

$$Y = X(\lambda)\beta + d_i\gamma + \varepsilon, \tag{2.2}$$

where d_i is an n-vector having 1 in the ith position and 0 elsewhere, and γ is a scalar parameter.

Model (2.2) is equivalent to model (2.1) in the following sense. The residual sums of squares for (2.1) and (2.2) are $s_{(i)}(\lambda) = Y_{(i)}^T(I - P_{(i)}(\lambda))Y_{(i)}$ and $\tilde{s}(\lambda) = Y^T(I - \tilde{P}(\lambda))Y$ respectively, where $P_{(i)}(\lambda)$ and $\tilde{P}(\lambda)$ are the respective projection matrices for $X_{(i)}(\lambda)$ and $\tilde{X}(\lambda) = (X(\lambda), d_i)$. It can be shown that $s_{(i)}(\lambda) = \tilde{s}(\lambda)$ for all λ. Therefore, the maximum likelihood estimates of λ for models (2.1) and (2.2) are the same and can be denoted by $\hat{\lambda}_{(i)}$.

Note that this situation differs from that of transformations of the response. In the latter case the maximum likelihood estimates of λ for the case-deletion model and mean-shift outlier model are not equal since the Jacobian is involved. This was pointed out by Cook and Wang (1983) and Tsai and Wu (1990).

The maximum likelihood estimates for the parameters in the mean-shift outlier model (2.2) satisfy equations analogous to (1.3) and (1.5). For fixed λ let

$$\tilde{Q}(\lambda) = I - \tilde{P}(\lambda) = Q(\lambda) - Q(\lambda)d_i[d_i^TQ(\lambda)d_i]^{-1}d_i^TQ(\lambda), \tag{2.3a}$$

$$\begin{pmatrix}\tilde{\beta}(\lambda)\\ \tilde{\gamma}(\lambda)\end{pmatrix} = [\tilde{X}^T(\lambda)\tilde{X}(\lambda)]^{-1}\tilde{X}^T(\lambda)Y, \quad \tilde{\sigma}^2_{\mathrm{MLE}}(\lambda) = \frac{1}{n}Y^T\tilde{Q}(\lambda)Y, \tag{2.3b}$$

$$\tilde{e}(\lambda) = Y - X(\lambda)\tilde{\beta}(\lambda) - d_i\tilde{\gamma}(\lambda) = \tilde{Q}(\lambda)Y. \tag{2.3c}$$

Then $\hat{\lambda}_{(i)}$ must minimize the residual sum of squares $\tilde{s}(\lambda)=\tilde{e}^T(\lambda)\tilde{e}(\lambda)=Y^T\tilde{Q}(\lambda)Y$.

The difference between $\hat{\lambda}$ and $\hat{\lambda}_{(i)}$ can be measured by the likelihood displacement, $LD(\hat{\lambda}_{(i)})$, proposed by Cook and Weisberg (1982). This likelihood displacement may be defined as $LD(\lambda) = 2[L(\hat{\lambda}) - L(\lambda)]$. A large value of $LD(\hat{\lambda}_{(i)})$ indicates that $\hat{\lambda}$ is highly dependent on case i, which suggests that this case is influential and may be an outlier.

Computing $LD(\hat{\lambda}_{(i)})$ exactly involves nonlinear optimization to find $\hat{\lambda}_{(i)}$ plus some matrix operations involving $X(\hat{\lambda}_{(i)})$. The computational time can be considerable if there are many different case deletions to consider. In practice only a rough estimate of the size of $LD(\hat{\lambda}_{(i)})$ is needed to determine whether case i is influential. Therefore, we derive an approximation to $LD(\lambda)$ for small $\lambda - \hat{\lambda}$ and an approximation to $\hat{\lambda}_{(i)}$ for small $\hat{\lambda}_{(i)} - \hat{\lambda}$. Combining these two approximations gives a diagnostic that can be computed more quickly than the exact value of $LD(\hat{\lambda}_{(i)})$.

For simplicity of presentation consider first the transformation of a single explanatory variable, i.e. $q = 1$, $\lambda = \lambda_1$. The results for transformation of several variables are essentially the same and are given at the end of this section. The likelihood displacement can be rewritten as $LD(\lambda) = n\log[s(\lambda)/s(\hat{\lambda})]$, using (1.5). Since the residual sum of squares, $s(\lambda)$, has a local minimum at $\hat{\lambda}$, the first two nontrivial terms in its Taylor expansion are the constant and quadratic terms:

$$LD(\lambda) \approx n\log\left[1 + \frac{\ddot{s}(\hat{\lambda})(\lambda - \hat{\lambda})^2}{2s(\hat{\lambda})}\right] \approx \frac{n\ddot{s}(\hat{\lambda})(\lambda - \hat{\lambda})^2}{2s(\hat{\lambda})} = -\ddot{L}(\hat{\lambda})(\lambda - \hat{\lambda})^2. \quad (2.4)$$

Here and below $\dot{}$ denotes differentiation with respect to λ. Note that

$$\ddot{L}(\hat{\lambda}) = -\frac{n}{2}\,\frac{\ddot{s}(\hat{\lambda})}{s(\hat{\lambda})}. \quad (2.5)$$

Formulas for $\dot{s}(\lambda)$ and $\ddot{s}(\hat{\lambda})$ are derived with the help of the Lemma in the Appendix:

$$\dot{Q} = -Q\dot{X}(X^TX)^{-1}X^T - X(X^TX)^{-1}\dot{X}^TQ. \quad (2.6)$$

Letting $W = \dot{X}_1$ it follows from (1.4) and (2.6) that

$$\dot{s}(\lambda) = -2\hat{\beta}_1(\lambda)W^T(\lambda)Q(\lambda)Y, \quad (2.7)$$

where $\hat{\beta}_1(\lambda)$ is the first element of $\hat{\beta}(\lambda)$ defined in (1.3). Since $\hat{\lambda}$ is the maximum likelihood estimate, it must satisfy $\dot{s}(\hat{\lambda}) = 0$, which implies either

$$W^T(\hat{\lambda})Q(\hat{\lambda})Y = 0, \quad (2.8)$$

or $\hat{\beta}_1(\hat{\lambda}) = 0$. In the latter case the transformed variable does not enter the regression model. More importantly, straightforward calculations show that $\hat{\beta}_1(\hat{\lambda}) = 0$ implies $\ddot{s}(\hat{\lambda}) = -2\dot{\hat{\beta}}_1(\hat{\lambda})W^T(\hat{\lambda})Q(\hat{\lambda})Y \leq 0$, i.e. $\hat{\lambda}$ does not minimize the residual sum of squares. Thus, we may assume that $\hat{\beta}_1(\hat{\lambda}) \neq 0$. Differentiating (2.7) and using (2.6) and (2.8) leads to a formula for $\ddot{s}(\hat{\lambda})$.

$$\ddot{s}(\hat{\lambda}) = 2(\hat{e}_w^T\hat{e}_w - \hat{e}_v^T\hat{e}), \quad (2.9)$$

where $\hat{\beta}_1 = \hat{\beta}_1(\hat{\lambda})$, $\hat{e} = Q(\hat{\lambda})Y$, $\hat{e}_w = Q(\hat{\lambda})W(\hat{\lambda})\hat{\beta}_1$, $\hat{e}_v = Q(\hat{\lambda})V(\hat{\lambda})\hat{\beta}_1$ and $V = \ddot{X}_1$. The vector $\hat{e}$ contains the residuals for the original regression problem $Y =$

$X(\hat{\lambda})\beta + \varepsilon$. The vectors $\hat{e}_w$ and $\hat{e}_v$ are the residuals for the related problems of regressing $W(\hat{\lambda})\hat{\beta}_1$ and $V(\hat{\lambda})\hat{\beta}_1$ on $X(\hat{\lambda})$, respectively. Finally, substituting (2.9) into (2.4) and (2.5) yields

$$\ddot{L}(\hat{\lambda}) = \frac{-n[\hat{e}_w^T\hat{e}_w - \hat{e}_v^T\hat{e}]}{\hat{e}^T\hat{e}} = -n\frac{\hat{\sigma}_w^2 - \hat{\sigma}_{ve}}{\hat{\sigma}^2}, \tag{2.10}$$

$$LD(\lambda) \approx n\frac{\hat{\sigma}_w^2 - \hat{\sigma}_{ve}}{\hat{\sigma}^2}(\lambda - \hat{\lambda})^2, \tag{2.11}$$

where $\hat{\sigma}^2 = \|\hat{e}\|^2/(n-p)$, $\hat{\sigma}_w^2 = \|\hat{e}_w\|^2/(n-p)$, and $\hat{\sigma}_{ve} = \hat{e}_v^T\hat{e}/(n-p)$ are variance and covariance estimates.

Having derived an approximation to the likelihood displacement, the next step is to compute an approximation to $\hat{\lambda}_{(i)}$. Since $\hat{\lambda}_{(i)}$ is the maximum likelihood estimate of λ for model (2.2) it must satisfy an equation analogous to (2.8):

$$W^T(\hat{\lambda}_{(i)})\tilde{e}(\hat{\lambda}_{(i)}) = W^T(\hat{\lambda}_{(i)})\tilde{Q}(\hat{\lambda}_{(i)})Y = 0. \tag{2.12}$$

By following Atkinson's (1983) approach $\tilde{e}(\lambda)$ is approximated by means of a linear Taylor polynomial approximation to $X(\lambda)$:

$$\tilde{e}(\lambda) = Y - X(\lambda)\tilde{\beta}(\lambda) - d_i\tilde{\gamma}(\lambda) \approx Y - X(\hat{\lambda})\tilde{\beta}(\lambda) - d_i\tilde{\gamma}(\lambda) - W(\hat{\lambda})\hat{\beta}_1(\lambda - \hat{\lambda}).$$

Choosing $\tilde{\beta}(\lambda)$ and $\tilde{\gamma}(\lambda)$ to minimize the approximation to $\tilde{e}^T(\lambda)\tilde{e}(\lambda)$ yields an expression similar to (2.3):

$$\begin{aligned} \begin{pmatrix} \tilde{\beta}(\lambda) \\ \tilde{\gamma}(\lambda) \end{pmatrix} &\approx [\tilde{X}^T(\hat{\lambda})\tilde{X}(\hat{\lambda})]^{-1}\tilde{X}^T(\hat{\lambda})[Y - W(\hat{\lambda})\hat{\beta}_1(\lambda - \hat{\lambda})], \\ \tilde{e}(\lambda) &\approx \tilde{Q}(\hat{\lambda})[Y - W(\hat{\lambda})\hat{\beta}_1(\lambda - \hat{\lambda})], \\ \tilde{e}(\hat{\lambda}_{(i)}) &\approx \tilde{Q}(\hat{\lambda})[Y - W(\hat{\lambda})\hat{\beta}_1(\hat{\lambda}_{(i)} - \hat{\lambda})]. \end{aligned} \tag{2.13}$$

The term $W(\hat{\lambda}_{(i)})$ in (2.12) can be approximated in two ways. An approach analogous to Cook and Wang (1983) assumes

$$W(\hat{\lambda}_{(i)}) \approx W(\hat{\lambda}). \tag{2.14}$$

Substituting (2.13) and (2.14) into (2.12) then leads to the following approximation of $\hat{\lambda}_{(i)}$:

$$\begin{aligned} \hat{\lambda}_{(i)} - \hat{\lambda} &\approx [W^T(\hat{\lambda})\tilde{Q}(\hat{\lambda})W(\hat{\lambda})\hat{\beta}_1]^{-1}W^T(\hat{\lambda})\tilde{Q}(\hat{\lambda})Y \\ &= -\left(\hat{e}_w^T\hat{e}_w - \frac{\hat{e}_{wi}^2}{1-p_{ii}}\right)^{-1}\frac{\hat{e}_{wi}\hat{e}_i}{1-p_{ii}} = \frac{-\hat{\sigma}\ \hat{r}_{wi}\hat{r}_i}{\hat{\sigma}_w[(n-p) - \hat{r}_{wi}^2]}, \end{aligned} \tag{2.15}$$

where p_{ii} is the ith diagonal element of the projection matrix $P(\hat{\lambda})$, $\hat{e}_i$ is the ith element of $\hat{e}$, $\hat{e}_{wi}$ is the ith element of $\hat{e}_w$, and $\hat{r}_i$ and $\hat{r}_{wi}$ are scaled residuals: $\hat{r}_i = \hat{e}_i/[(1-p_{ii})^{1/2}\hat{\sigma}_w]$, $\hat{r}_{wi} = \hat{e}_{wi}/[(1-p_{ii})^{1/2}\hat{\sigma}_w]$.

Approximation (2.15) above neglects the $O(\hat{\lambda}_{(i)} - \hat{\lambda})$ contribution to $W(\hat{\lambda}_{(i)})$ in (2.12) while keeping the $O(\hat{\lambda}_{(i)} - \hat{\lambda})$ contribution to $\tilde{e}(\hat{\lambda}_{(i)})$. An asymptotically more accurate approach is to assume

$$W(\hat{\lambda}_{(i)}) \approx W(\hat{\lambda}) + V(\hat{\lambda})(\hat{\lambda}_{(i)} - \hat{\lambda}). \tag{2.16}$$

Substituting (2.13) and (2.16) into (2.12) and keeping terms up to $O(\hat{\lambda}_{(i)} - \hat{\lambda})$ leads to

$$\begin{aligned}
\hat{\lambda}_{(i)} - \hat{\lambda} &\approx [W^T(\hat{\lambda})\tilde{Q}(\hat{\lambda})W(\hat{\lambda})\hat{\beta}_1 - V^T(\hat{\lambda})\tilde{Q}(\hat{\lambda})Y]^{-1}W^T(\hat{\lambda})\tilde{Q}(\hat{\lambda})Y \\
&= -\left[\hat{e}_w^T\hat{e}_w - \frac{\hat{e}_{wi}^2}{1-p_{ii}} - \hat{e}_v^T\hat{e} + \frac{\hat{e}_{vi}\hat{e}_i}{1-p_{ii}}\right]^{-1}\frac{\hat{e}_{wi}\hat{e}_i}{1-p_{ii}} \\
&= \frac{-\hat{\sigma}\hat{r}_{wi}\hat{r}_i}{\hat{\sigma}_w\left[(n-p)\left(1-\frac{\hat{\sigma}_{ve}}{\hat{\sigma}_w^2}\right) - \hat{r}_{wi}^2 + \frac{\hat{\sigma}_v\hat{\sigma}\hat{r}_{vi}\hat{r}_i}{\hat{\sigma}_w^2}\right]},
\end{aligned} \tag{2.17}$$

where $\hat{e}_{vi}$ is the ith element of $\hat{e}_v$, $\hat{\sigma}_v^2 = \|\hat{e}_v\|^2/(n-p)$, and $\hat{r}_{vi} = \hat{e}_{vi}/[(1-p_{ii})^{1/2}\hat{\sigma}_v]$. This formula is analogous to that derived by Hinkley and Wang (1988), and it has an error of $o(\hat{\lambda}_{(i)} - \hat{\lambda})$.

The approximation of the likelihood displacement in (2.11) can now be combined with either of the two approximations for $\hat{\lambda}_{(i)}$. The following diagnostics are based on (2.15) and (2.17) respectively.

$$\begin{aligned}
LD(\hat{\lambda}_{(i)}) &\approx n\log\left\{1 + \left(1-\frac{\hat{\sigma}_{ve}}{\hat{\sigma}_w^2}\right)\left[\frac{\hat{r}_{wi}\hat{r}_i}{(n-p)-\hat{r}_{wi}^2}\right]^2\right\} \\
&\approx n\left(1-\frac{\hat{\sigma}_{ve}}{\hat{\sigma}_w^2}\right)\left[\frac{\hat{r}_{wi}\hat{r}_i}{(n-p)-\hat{r}_{wi}^2}\right]^2,
\end{aligned} \tag{2.18a}$$

$$\begin{aligned}
LD(\hat{\lambda}_{(i)}) &\approx n\log\left\{1 + \left(1-\frac{\hat{\sigma}_{ve}}{\hat{\sigma}_w^2}\right)\left[\frac{\hat{r}_{wi}\hat{r}_i}{(n-p)\left(1-\frac{\hat{\sigma}_{ve}}{\hat{\sigma}_w^2}\right) - \hat{r}_{wi}^2 + \frac{\hat{\sigma}_v\hat{\sigma}\hat{r}_{vi}\hat{r}_i}{\hat{\sigma}_w^2}}\right]^2\right\} \\
&\approx n\left(1-\frac{\hat{\sigma}_{ve}}{\hat{\sigma}_w^2}\right)\left[\frac{\hat{r}_{wi}\hat{r}_i}{(n-p)\left(1-\frac{\hat{\sigma}_{ve}}{\hat{\sigma}_w^2}\right) - \hat{r}_{wi}^2 + \frac{\hat{\sigma}_v\hat{\sigma}\hat{r}_{vi}\hat{r}_i}{\hat{\sigma}_w^2}}\right]^2.
\end{aligned} \tag{2.18b}$$

In Section 4 formulas (2.18) are applied two real data sets and their relative accuracies are compared.

The previous derivation considered the transformation of a single explanatory variable and deletion of a single case. It is straightforward to generalize (2.18) to the transformation of several explanatory variables and the deletion of multiple cases. Consider model (1.2) where q, the number of transformed variables, is now arbitrary. To allow for deletion of multiple cases the mean-shift outlier model, (2.2), is generalized to $Y = X(\lambda)\beta + D\gamma + \varepsilon$, where $J = \{i_1, \ldots, i_k\}$ is the set of indices of deleted cases, $D = (d_{i_1}, \ldots, d_{i_k})$, and γ is a $k \times 1$ vector.

The definition of the likelihood displacement LD remains unchanged. Its approximation by a quadratic Taylor polynomial is similar to (2.4), but now $\ddot{L}$ is a $q \times q$ matrix:

$$\ddot{L} = \frac{\partial^2 L}{\partial\lambda\partial\lambda^T}, \quad LD(\lambda) \approx -(\lambda - \hat{\lambda})^T \ddot{L}(\hat{\lambda})(\lambda - \hat{\lambda}). \tag{2.19}$$

The formula for $\ddot{L}(\hat{\lambda})$ corresponding to (2.10) is derived using $W = (\partial X_1/\partial\lambda_1, \ldots, \partial X_q/\partial\lambda_q)$, $V = (\partial X_1^2/\partial\lambda_1^2, \ldots, \partial X_q^2/\partial\lambda_q^2)$, $\hat{B} = \operatorname{diag}(\hat{\beta}_1(\hat{\lambda}), \ldots, \hat{\beta}_q(\hat{\lambda}))$, $\hat{E}_w = Q(\hat{\lambda})W(\hat{\lambda})\hat{B}$, and $\hat{E}_v = Q(\hat{\lambda})V(\hat{\lambda})\hat{B}$. The matrices $\hat{B}$, $\hat{E}_w$ and $\hat{E}_v$ are generalizations of $\hat{\beta}_1$, $\hat{e}_w$ and $\hat{e}_v$, respectively. The generalization of (2.10) is

$$\ddot{L}(\hat{\lambda}) = \frac{-n}{\hat{\sigma}^2}\left(\hat{\Sigma}_{ww} - \hat{\Sigma}_{ve}\right), \tag{2.20}$$

where $\hat{\Sigma}_{ww} = \hat{E}_w^T \hat{E}_w/(n-p)$ and $\hat{\Sigma}_{ve} = \operatorname{diag}(\hat{E}_v^T \hat{e})/(n-p)$.

The approximation to $\hat{\lambda}_{(J)}$, the estimated transformation parameter when cases J are deleted, is derived using the same argument as above. This results in an approximation corresponding to (2.15),

$$\hat{\lambda}_{(J)} - \hat{\lambda} \approx -[\hat{\Sigma}_{ww} - \hat{\Sigma}_{wJwJ}]^{-1}\hat{\Sigma}_{wJeJ}, \tag{2.21a}$$

and another corresponding to (2.17),

$$\hat{\lambda}_{(J)} - \hat{\lambda} \approx [(\hat{\Sigma}_{ve} - \hat{\Sigma}_{vJeJ}) - (\hat{\Sigma}_{ww} - \hat{\Sigma}_{wJwJ})]^{-1}\hat{\Sigma}_{wJeJ}, \tag{2.21b}$$

where $P_J = D^T P(\hat{\lambda}) D$, $\hat{e}_J = D^T \hat{e}$, $\hat{E}_{wJ} = D^T \hat{E}_w$, $\hat{E}_{vJ} = D^T \hat{E}_v$, $\hat{\Sigma}_{wJwJ} = \hat{E}_{wJ}^T (I - P_J)^{-1} \hat{E}_{wJ}/(n-p)$, $\hat{\Sigma}_{wJeJ} = \hat{E}_{wJ}^T (I - P_J)^{-1} \hat{e}_J/(n-p)$, and $\hat{\Sigma}_{vJeJ} = \operatorname{diag}(\hat{E}_{vJ}^T (I - P_J)^{-1} \hat{e}_J)/(n-p)$.

These approximations are combined with (2.20) to give the following generalizations of the diagnostics (2.18):

$$LD(\hat{\lambda}_{(J)}) \approx \frac{n}{\hat{\sigma}^2}\hat{\Sigma}_{wJeJ}^T(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wJwJ})^{-1}(\hat{\Sigma}_{ww} - \hat{\Sigma}_{ve})(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wJwJ})^{-1}\hat{\Sigma}_{wJeJ}, \tag{2.22a}$$

$$LD(\hat{\lambda}_{(J)}) \approx \frac{n}{\hat{\sigma}^2}\hat{\Sigma}_{wJeJ}^T[(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wJwJ}) - (\hat{\Sigma}_{ve} - \hat{\Sigma}_{vJeJ})]^{-1}(\hat{\Sigma}_{ww} - \hat{\Sigma}_{ve})$$
$$\times [(\hat{\Sigma}_{ww} - \hat{\Sigma}_{wJwJ}) - (\hat{\Sigma}_{ve} - \hat{\Sigma}_{vJeJ})]^{-1}\hat{\Sigma}_{wJeJ}. \quad (2.22b)$$

Although $\hat{\lambda}_{(J)}$ is a vector, LD is a scalar. Detailed information concerning the influence of one or more cases on the estimate of a single parameter $\hat{\lambda}_j$ is contained in the components of $\hat{\lambda}_{(J)} - \hat{\lambda}$ in (2.21). On the other hand, the likelihood displacement, (2.22), provides a convenient overall measure of influence of one or more cases.

Other diagnostic methods based on the above approach can be obtained. In particular, the method given by Atkinson (1986) and the method given by del Rio (1988) can be easily extended to the case of transformation of explanatory variables.

3. Local Influence

The local influence approach has been successfully applied to many statistical models (see, for example, Thomas (1990) and Weissfeld (1990)). Cook (1987) used this approach to derive regression diagnostics for partially nonlinear models, which include transformations of a single explanatory variable as a special case. Lawrance (1988) studied regression diagnostics for transformations of the response using the local influence method. This section extends the local influence method to transformations of several explanatory variables and also covers more perturbations than those considered by Cook (1987).

Again we consider model (1.2) where λ is the parameter of interest and the corresponding profile log-likelihood $L(\lambda)$ is given by (1.5). Now suppose that there are perturbations to the model (1.2) through an m-vector ω and that the maximum likelihood estimate of λ for the perturbed model is $\hat{\lambda}(\omega)$. The unperturbed state is denoted ω_0 which means that $\hat{\lambda} = \hat{\lambda}(\omega_0)$.

The likelihood displacement may be used to indicate the difference between the estimates $\hat{\lambda}$ and $\hat{\lambda}(\omega)$. Note that $LD(\hat{\lambda}(\omega)) = 2[L(\hat{\lambda}) - L(\hat{\lambda}(\omega))]$. To measure the sensitivity of the estimate $\hat{\lambda}$ to small perturbations one can compute the second derivative of $LD(\hat{\lambda}(\omega))$ with respect to ω at ω_0.

Following Cook (1986), $z = LD(\hat{\lambda}(\omega))$ is an m-dimensional surface in $\mathbb{R}^{m+1}$, and the normal curvature along the direction d at ω_0 is denoted by C_d. The direction $d_{\max}$ which corresponds to the maximum curvature $C_{\max} = \max_{\|d\|=1} C_d$ is the main diagnostic quantity. Cook (1986) showed that $C_d = |d^T G^T \ddot{L}(\hat{\lambda})Gd|$ ($\|d\| = 1$), where $G = [\partial\hat{\lambda}(\omega)/\partial\omega^T]_{\omega_0}$, and $\ddot{L}$ is the second derivative of the profile log-likelihood as before. It follows that $C_{\max}$ is the maximum absolute eigenvalue of the $m \times m$ matrix $F = G^T \ddot{L}(\hat{\lambda})G$, and $d_{\max}$ is the corresponding eigenvector.

This result can be derived in an alternative way using Taylor expansions as in the previous section. First we approximate $\hat\lambda(\omega)-\hat\lambda$ by differentials and then use the approximation for LD in (2.19):

$$\begin{aligned}\hat\lambda(\omega)-\hat\lambda &\approx G(\omega-\omega_0),\\ LD(\hat\lambda(\omega)) &\approx -(\hat\lambda(\omega)-\hat\lambda)^T\ddot L(\hat\lambda)(\hat\lambda(\omega)-\hat\lambda)\\ &\approx -(\omega-\omega_0)^T[G^T\ddot L(\hat\lambda)G](\omega-\omega_0)=-(\omega-\omega_0)^TF(\omega-\omega_0).\end{aligned}$$

From this last equation it is clear that for small perturbations the greatest influence on $LD(\hat\lambda(\omega))$ arises when $\omega-\omega_0$ is parallel to $d_{\max}$ as defined above.

Since formulas for $\ddot L(\hat\lambda)$ have been derived in the previous section, the only missing information is G, which depends on the form of the perturbation. We consider perturbations of case-weights, explanatory variables and transformed explanatory variables. The perturbed models all take the form $Y=X(\lambda,\omega)\beta+\varepsilon$, $\varepsilon\sim N(0,\sigma^2\Omega^{-1}(\omega))$, where $X(\lambda,\omega)=(X_1(\lambda,\omega),\dots,X_q(\lambda,\omega),X_{q+1}(\omega),\dots,X_p(\omega))$. Therefore, the profile log-likelihood for λ takes the form analogous to (1.5): $L(\lambda,\omega)=-(n/2)\log(s(\lambda,\omega))$, where $s(\lambda,\omega)=Y^TQ(\lambda,\omega)Y$ is the residual sum of squares, and

$$Q(\lambda,\omega)=\Omega(\omega)-\Omega(\omega)X(\lambda,\omega)[X^T(\lambda,\omega)\Omega(\omega)X(\lambda,\omega)]^{-1}X^T(\lambda,\omega)\Omega(\omega).$$

Since the unperturbed state is $\omega=\omega_0$, it follows that $X(\lambda,\omega_0)=X(\lambda)$, $\Omega(\omega_0)=I$, $Q(\lambda,\omega_0)=Q(\lambda)$ and $s(\lambda,\omega_0)=s(\lambda)$.

The maximum likelihood estimate $\hat\lambda(\omega)$ satisfies the equation $[\partial s(\lambda,\omega)/\partial\lambda]_{(\hat\lambda(\omega),\omega)}=0$. The argument leading to Equation (2.8) implies that $\hat\lambda(\omega)$ satisfies the equation $W^T(\hat\lambda(\omega),\omega)Q(\hat\lambda(\omega),\omega)Y=0$, where $W(\lambda,\omega)=(\partial X_1/\partial\lambda_1,\dots,\partial X_q/\partial\lambda_q)$. Differentiating this equation with respect to ω yields $G=[\partial\hat\lambda(\omega)/\partial\omega^T]_{\omega_0}=-\{[\partial(W^TQY)/\partial\lambda)]^{-1}[\partial(W^TQY)/\partial\omega^T)]\}_{(\hat\lambda,\omega_0)}$. The first term in the right-hand product is independent of the form of the perturbation and can be computed using the arguments leading to (2.9): $[\partial(W^TQY)/\partial\lambda^T]_{(\hat\lambda,\omega_0)}=-\hat B^{-1}(\Sigma_{ww}-\Sigma_{ve})$. Therefore $G=(\Sigma_{ww}-\Sigma_{ve})^{-1}A$, where $A=\hat B[\partial(W^TQY)/\partial\omega^T]_{(\hat\lambda,\omega_0)}$.

To compute the diagnostic $d_{\max}$ one must compute the $q\times m$ matrix A for the perturbation of interest. Then $d_{\max}$ is the eigenvector corresponding to the largest eigenvalue of the matrix $-G^T\ddot L(\hat\lambda)G=(n/\hat\sigma^2)A^T(\Sigma_{ww}-\Sigma_{ve})^{-1}A$. For transformations of a single explanatory variable ($q=1$) $(\Sigma_{ww}-\Sigma_{ve})^{-1}$ is a scalar, and $d_{\max}$ is proportional to A^T.

In the subsections below $W(\lambda,\omega)$ and $Q(\lambda,\omega)$ are identified for several interesting types of perturbations, and formulas are derived for A. Some of these perturbations were considered by Cook (1987), and we recover his results for $q=1$. Note that for all the perturbations considered below ω is an n-vector.

3.1. Case-weights perturbations

Consider a model where the case-weights are perturbed:

$$\Omega = \mathrm{diag}(\omega), \quad X(\lambda, \omega) = X(\lambda). \tag{3.1}$$

The unperturbed state is $\omega_0 = (1, \ldots, 1)^T$, for which $\Omega = I$. The Lemma in the Appendix implies that $[\partial Q(\lambda, \omega)/\partial \omega_i]_{(\hat{\lambda}, \omega_0)} = Q(\hat{\lambda}) d_i^T d_i Q(\hat{\lambda})$. Since $W(\lambda, \omega) = W(\lambda)$, it follows that A is

$$\begin{aligned} A_c = \hat{B}[\partial(W^T QY)/\partial \omega^T]_{(\hat{\lambda}, \omega_0)} &= \hat{B} W^T(\hat{\lambda})[\partial(Q(\lambda, \omega)Y)/\partial \omega^T]_{(\hat{\lambda}, \omega_0)} \\ &= \hat{E}_w^T \, \mathrm{diag}(\hat{e}). \end{aligned} \tag{3.2a}$$

For transformations of a single explanatory variable

$$d_{\max}^T \propto A_c = \hat{e}_w^T \, \mathrm{diag}(\hat{e}), \tag{3.2b}$$

which is equivalent to Equation (39) of Cook (1987).

3.2. Perturbations of explanatory variables

Without loss of generality, suppose that the $(q+1)$st column of the data matrix X is modified by adding a vector ω of perturbations. In this case

$$\Omega = I, \quad X(\lambda, \omega) = (X_1(\lambda_1), \ldots, X_q(\lambda_q), X_{q+1}(\omega), X_{q+2}, \ldots, X_p), \tag{3.3a}$$

$$X_{q+1}(\omega) = (x_{1,q+1} + \omega_1, \ldots, x_{n,q+1} + \omega_n)^T, \tag{3.3b}$$

and $\omega_0 = (0, \ldots, 0)^T$ represents the unperturbed state.

Again the Lemma in the Appendix plus the fact that $W(\lambda, \omega) = W(\lambda)$ are used. Straightforward calculations yield A:

$$\begin{aligned} A_{e1} = \hat{B}[\partial(W^T QY)/\partial \omega^T]_{(\hat{\lambda}, \omega_0)} &= \hat{B} W^T(\hat{\lambda})[\partial(Q(\lambda, \omega)Y)/\partial \omega^T]_{(\hat{\lambda}, \omega_0)} \\ &= -\,(\hat{E}_w^T \hat{\beta}_{q+1} + \hat{B}_{w,q+1} \hat{e}^T), \end{aligned} \tag{3.4a}$$

where $\hat{\beta}_i$ is the ith element of $\hat{\beta}$. The vector $\hat{B}_{wi}^T$ is the ith row of $\hat{B}_w = [X^T(\hat{\lambda})X(\hat{\lambda})]^{-1} X^T(\hat{\lambda}) W(\hat{\lambda}) \hat{B}$, the coefficient obtained in regressing $W(\hat{\lambda})\hat{B}$ on $X(\hat{\lambda})$. For $q = 1$ the diagnostic is

$$d_{\max}^T \propto A_{e1} = -(\hat{e}_w^T \hat{\beta}^2 + \hat{\beta}_{w2} \hat{e}^T), \tag{3.4b}$$

where $\hat{\beta}_{wi}$ is the ith element of $\hat{\beta}_w = [X^T(\hat{\lambda})X(\hat{\lambda})]^{-1} X^T(\hat{\lambda}) W(\hat{\lambda}) \hat{\beta}$.

Perturbation (3.3b) is additive. It also possible to consider proportional perturbations,

$$X_{q+1}(\omega) = (x_{1,q+1}\omega_1, \ldots, x_{n,q+1}\omega_n)^T, \tag{3.5}$$

where $\omega_0 = (1, \ldots, 1)^T$ represents the unperturbed problem. The effect of this modification is to right-multiply the value of A in (3.4) by $\operatorname{diag}(X_{q+1})$.

$$A_{e2} = -\,(\hat{E}_w^T \hat{\beta}_{q+1} + \hat{B}_{w,q+1}\hat{e}^T)\operatorname{diag}(X_{q+1}), \tag{3.6a}$$

$$d_{\max}^T \propto A_{e2} = -\,(\hat{e}_w^T \hat{\beta}_2 + \hat{\beta}_{w2}\hat{e}^T)\operatorname{diag}(X_2) \quad (q = 1). \tag{3.6b}$$

3.3. Perturbations of transformed variables

Now suppose that one of the transformed explanatory variables is perturbed by adding a vector ω of perturbations, and $\omega_0 = (0, \ldots, 0)^T$ represents the unperturbed state. Without loss of generality the first variable is perturbed. Then

$$\Omega = I, \quad X(\lambda, \omega) = (X_1(\lambda_1, \omega), X_2(\lambda_2), \ldots, X_q(\lambda_q), X_{q+1}, \ldots, X_p), \tag{3.7a}$$

$$X_1(\lambda_1, \omega) = (h(x_{11} + \omega_1, \lambda_1), \ldots, h(x_{n1} + \omega_n, \lambda_1))^T. \tag{3.7b}$$

Define the vectors $\hat{t}_1 = (h'(x_{11}, \nu)), \ldots, h'(x_{n1}, \nu))^T|_{\nu=\hat{\lambda}_1}$ and $\hat{u}_1 = (\partial h'(x_{11}, \nu)/\partial\nu, \ldots, h'(x_{n1}, \nu)/\partial\nu)^T|_{\nu=\hat{\lambda}_1}$, where $'$ denotes differentiation of h with respect to its first argument. Calculations similar to the case of perturbed explanatory variables yield the following value for A:

$$A_{t1} = \hat{\beta}_1(1, 0, \ldots, 0)^T \hat{e}^T \operatorname{diag}(\hat{u}_1) - (\hat{E}_w^T \hat{\beta}_1 + \hat{B}_{w1}\hat{e}^T)\operatorname{diag}(\hat{t}_1), \tag{3.8a}$$

$$d_{\max}^T \propto A_{t1} = \hat{\beta}_1 \hat{e}^T \operatorname{diag}(\hat{u}_1) - (\hat{e}_w^T \hat{\beta}_1 + \hat{\beta}_{w1}\hat{e}^T)\operatorname{diag}(\hat{t}_1) \quad (q = 1). \tag{3.8b}$$

Equation (3.8b) is equivalent to equation (42) of Cook (1987).

A proportional perturbation is given by

$$X_1(\lambda_1, \omega) = (h(x_{11}\omega_1, \lambda_1), \ldots, h(x_{n1}\omega_n, \lambda_1))^T, \tag{3.9}$$

where $\omega_0 = (1, \ldots, 1)^T$ represents the unperturbed problem. The corresponding value of A is then

$$A_{t2} = [\hat{\beta}_1(1, 0, \ldots, 0)^T \hat{e}^T \operatorname{diag}(\hat{u}_1) - (\hat{\beta}_1 \hat{E}_w^T + \hat{B}_{w1}\hat{e}^T)\operatorname{diag}(\hat{t}_1)]\operatorname{diag}(X_1), \tag{3.10a}$$

$$d_{\max}^T \propto A_{t2} = [\hat{\beta}_1 \hat{e}^T \operatorname{diag}(\hat{u}_1) - (e_w^T \hat{\beta}_1 + \hat{\beta}_{w1}\hat{e}^T)\operatorname{diag}(\hat{t}_1)]\operatorname{diag}(X_1) \;\; (q = 1). \tag{3.10b}$$

For the first data set considered in the next section the proportional perturbation is found to give a more appropriate diagnostic than the additive perturbation (3.7b).

4. Numerical Examples

4.1. Snow geese data

These data were given by Weisberg (1980) and discussed by Cook (1986). The response, y, is the true flock size, and the explanatory variable, x, is the visually estimated flock size for a sample of $n = 45$ flocks of snow geese. The proposed model is

$$y_i = \beta_1 + h(x_i, \lambda)\beta_2 + \varepsilon_i, \quad (i = 1, \ldots, n), \tag{4.1}$$

using the power transformation

$$h(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(x), & \lambda = 0. \end{cases} \tag{4.2}$$

The parameter estimates for this data are given in Table 1. The regression diagnostics discussed below indicate that case 29 is an outlier. Therefore, the parameters are also estimated for the data with case 29 deleted. The fitted curves with and without case 29 are plotted in Figure 1.

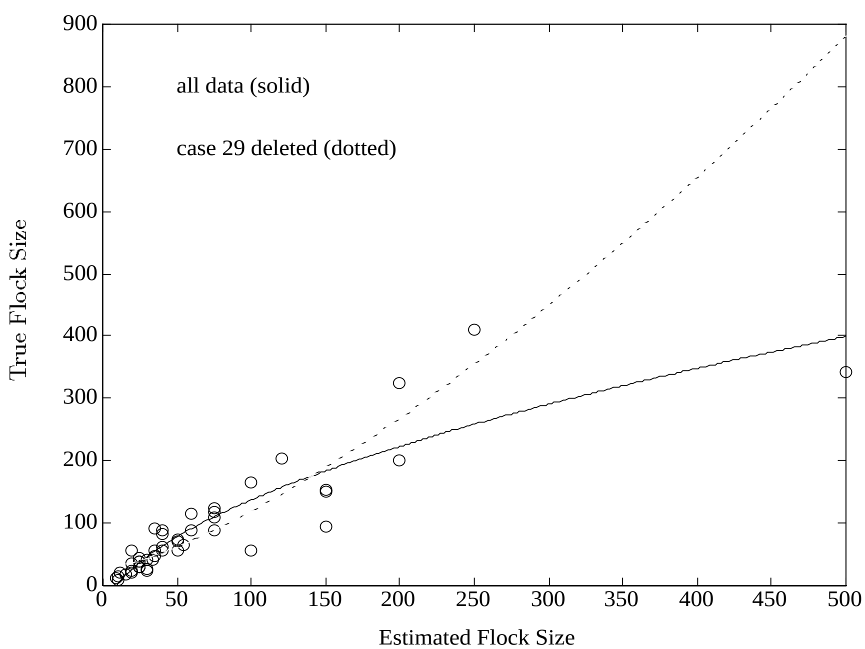

Figure 1. Snow geese data and fitted regression curves.

Table 1. Regression parameter estimates for snow geese data.

	All data	Case 29 deleted
$\hat{\lambda}$	0.53761	1.3772
$\hat{\beta}_1$	−35.759	27.405
$\hat{\beta}_2$	8.6038	0.22591
$\hat{\sigma}$	38.546	32.099

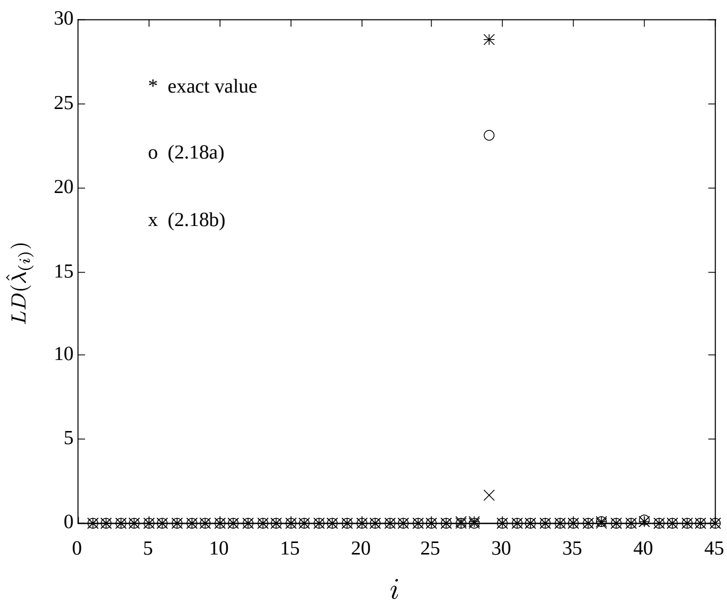

Figure 2. Index plot of likelihood displacement for snow geese data.

Figure 2 shows the likelihood displacement computed exactly and by (2.18). The exact value and both approximate values of $LD(\hat{\lambda}_{(29)})$ are far above the corresponding values for the other cases. Therefore, case 29 is the most influential, which is consistent with the scatter plot in Figure 1.

For the local influence approach $q = 1$ so $d_{\max}$ is scalar multiple of A^T. The three kinds of perturbations that can be considered are models (3.1), (3.7) and (3.9). Figure 3 gives the index plots of the vectors $d_{\max}$ for case-weights perturbations (3.1) and proportional perturbations of the transformed explanatory variable (3.9). $d_{\max}$ has been normalized so that $\|d_{\max}\|^2 = n$. For both kinds of perturbations case 29 is the most influential. This is consistent with the strong evidence of heteroscedasticity for these data as pointed out by Cook (1986).

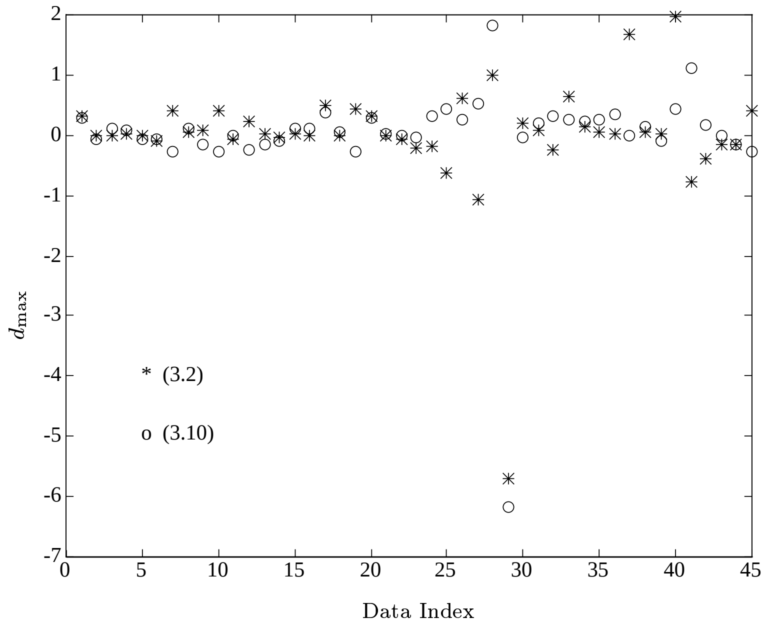

Figure 3. Index plot of $d_{\max}$ from (3.2) and (3.10) for snow geese data.

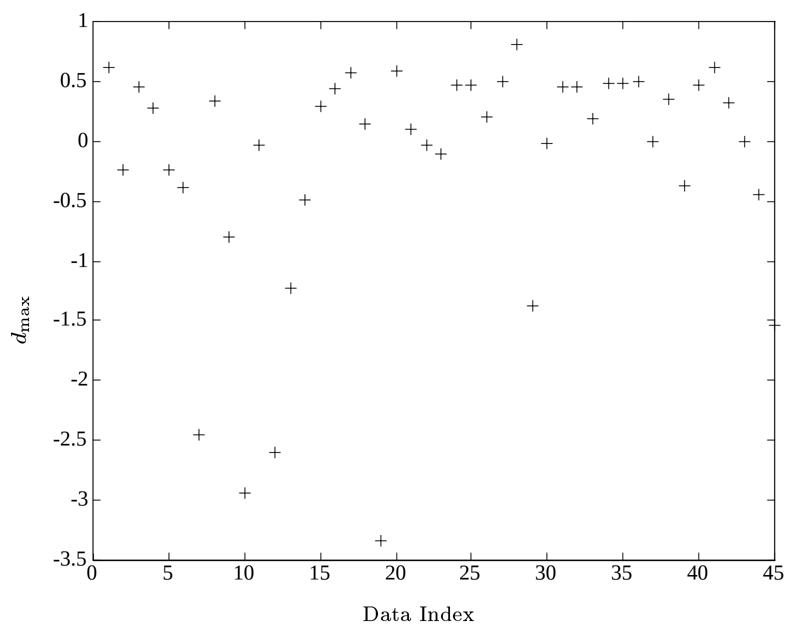

Figure 4. Index plot of $d_{\max}$ from (3.8) for snow geese data.

Figure 4 gives the index plot of $d_{\max}$ normalized as above for additive perturbations of the transformed explanatory variables according to model (3.7b). In contrast to the previous diagnostics case 29 is not the most influential. However, note that the four most influential cases (19, 10, 12 and 7 respectively) have small x values (9, 10, 10 and 12 respectively). The ith component of the vector $\hat{t}_1$ that appears in formula (3.8) is $\hat{t}_{1i} = h'(x_i, \hat{\lambda}) = x_i^{\hat{\lambda}-1} = x_i^{-0.46}$. Since $\hat{t}_{1i}$ is large when x_i is small, the diagnostic based on the additive perturbation (3.7) accentuates the influence of observations with small x in contrast to the diagnostic based on proportional perturbation (3.9). On the other hand the diagnostic for the proportional perturbation, (3.10), depends on $\hat{t}_{1i}x_i = x_i^{\hat{\lambda}} = x_i^{0.54}$, which gives more weight to observations with large x.

Box and Tidwell (1962) and Cook and Weisberg (1982) suggest using added variable plots to determine the need for transforming explanatory variables and whether one or more observations are influential. Using the notation introduced in (2.9) and (2.19) this corresponds to plotting $\hat{e}$ versus $\hat{e}_w$ (or the columns of $\hat{E}_w$ for $q \geq 1$). As Cook (1987) pointed out, a particular case i with $\hat{e}_i$ and $\hat{e}_{wi}$ large simultaneously may be identified as influential both from the added variable plot and from the case-weights perturbations diagnostic (3.2). The added variable plot for the geese data is shown in Figure 5 with case 29 labeled. Although case 29 is clearly the most influential point from the index plots in Figures 2 and 3, it is not so obviously the most influential from the added variable plot because $\hat{e}_{29}$ is moderate. Therefore, the new diagnostics proposed here have some advantages over added variable plots in identifying influential points.

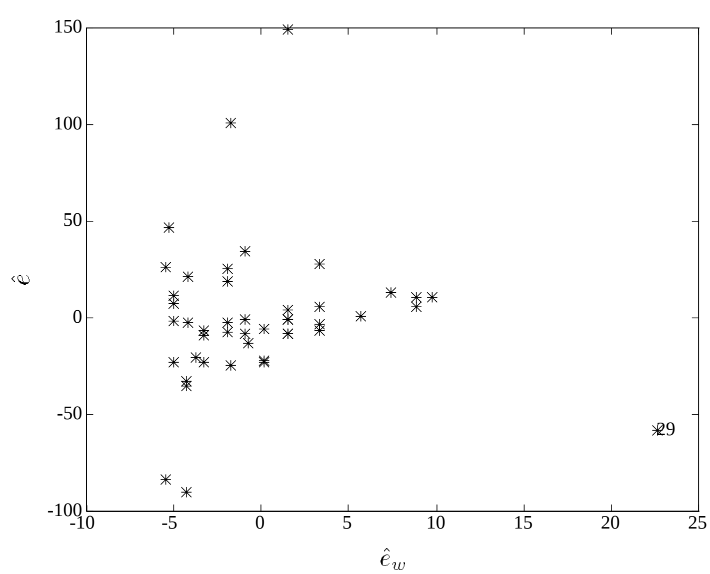

Figure 5. Added variable plot for snow geese data.

4.2. Tree data

These data were given by Ryan, Joiner and Ryan (1976) and discussed by Cook and Weisberg (1982), Cook and Wang (1983) and Tsai and Wu (1990). The response, y, is tree volume and the explanatory variables, x_2 and x_3, are tree diameter and tree height respectively. Two models are considered. In the first model only the tree diameter is transformed.

$$y_i = \beta_1 + h(x_{i2}, \lambda_2)\beta_2 + x_{i3}\beta_3 + \varepsilon_i, \quad (i = 1, \ldots, n). \tag{4.3}$$

In the second model both tree diameter and tree height are transformed.

$$y_i = \beta_1 + h(x_{i2}, \lambda_2)\beta_2 + h(x_{i3}, \lambda_3)\beta_3 + \varepsilon_i, \quad (i = 1, \ldots, n). \tag{4.4}$$

Again, h is the power transformation given by (4.2). Table 2. lists the parameter estimates for this data.

Table 2. Regression parameter estimates for tree data.

	Model (4.3)	Model (4.4)
$\hat{\lambda}_2$	2.6039	2.5831
$\hat{\lambda}_3$	—–	1.7375
$\hat{\beta}_1$	−21.249	−9.9487
$\hat{\beta}_2$	0.066366	0.070142
$\hat{\beta}_3$	0.36424	0.015134
$\hat{\sigma}$	2.5943	2.5929

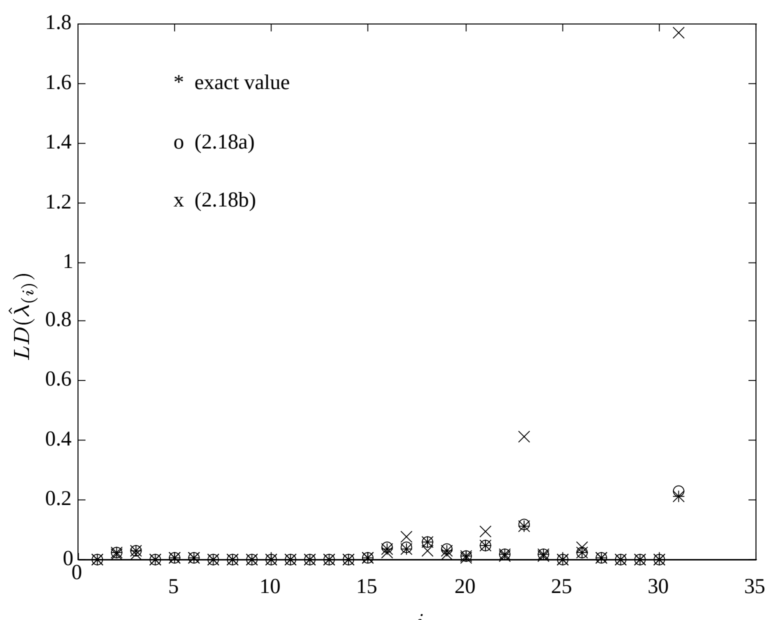

Figure 6. Index plot of likelihood displacement for tree data model (4.3).

First consider model (4.3). Figure 6 shows the likelihood displacement computed exactly and computed approximately by (2.18). This plot shows that cases 23 and 31 are influential. The local influence under all five perturbed models considered in Section 3 is shown by the index plots of the vectors $d_{\max}$ in Figure 7. Case 23 is influential for case-weights perturbations, while case 31 is influential for perturbations of either the explanatory variable or the transformed explanatory variable.

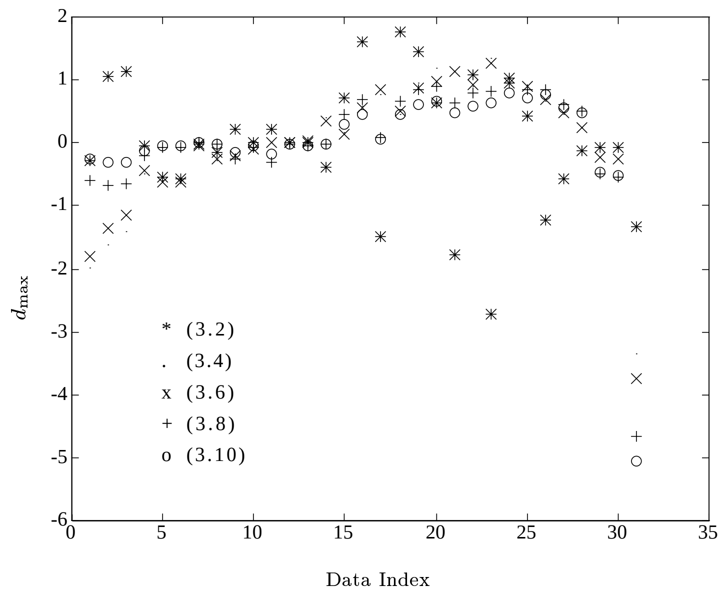

Figure 7. Index plot of $d_{\max}$ for tree data model (4.3).

A somewhat different situation arises for model (4.4). Figure 8 shows the likelihood displacement. While cases 23 and 31 are influential as before, cases 17 and 18 are now also influential. The index plots of the vectors $d_{\max}$ computed using (3.11) are shown in Figure 9. This plot is similar to that of Figure 7. Case 23 is still influential for case-weights perturbations, whereas case 31 is still influential for perturbations of the transformed explanatory variable.

The similar values of $\hat{\sigma}$ for models (4.3) and (4.4) in Table 2 suggest that the transformation of the tree height gives negligible improvement of the model. For model (4.4)

$$\ddot{L}(\hat{\lambda}) = -\begin{pmatrix} 13.420 & 0.38321 \\ 0.38321 & 0.073043 \end{pmatrix},$$

which means that changes in the transformation parameter for tree height have very little influence on the likelihood compared to changes in the transformation parameter for tree diameter. Consequently, λ_3 is estimated with much less accuracy than λ_2. Indeed, it was observed that the deletion of one case can make a great change in the estimate of λ_3. For the tree data $\hat{\lambda}_{(17)3} = -1.1370$ and

$\hat{\lambda}_{(18)3} = 6.2497.$

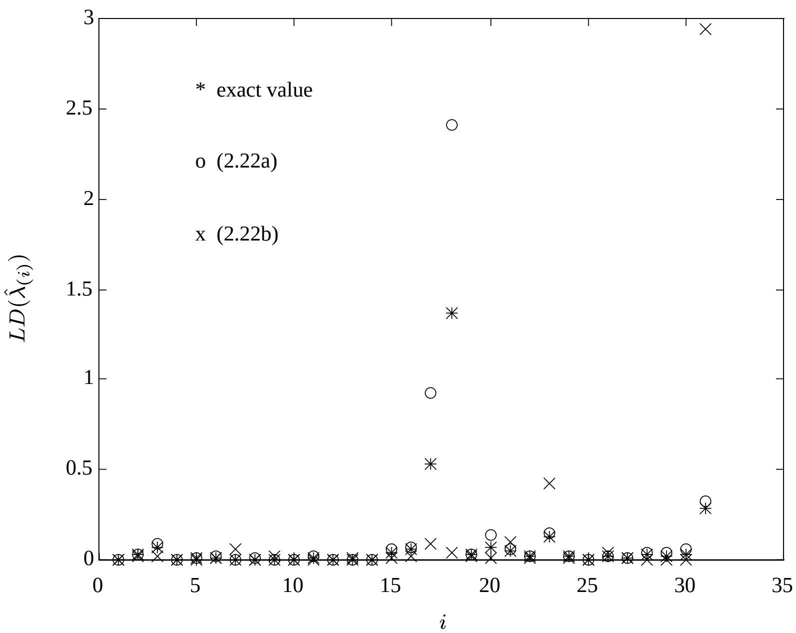

Figure 8. Index plot of likelihood displacement for tree data model (4.4).

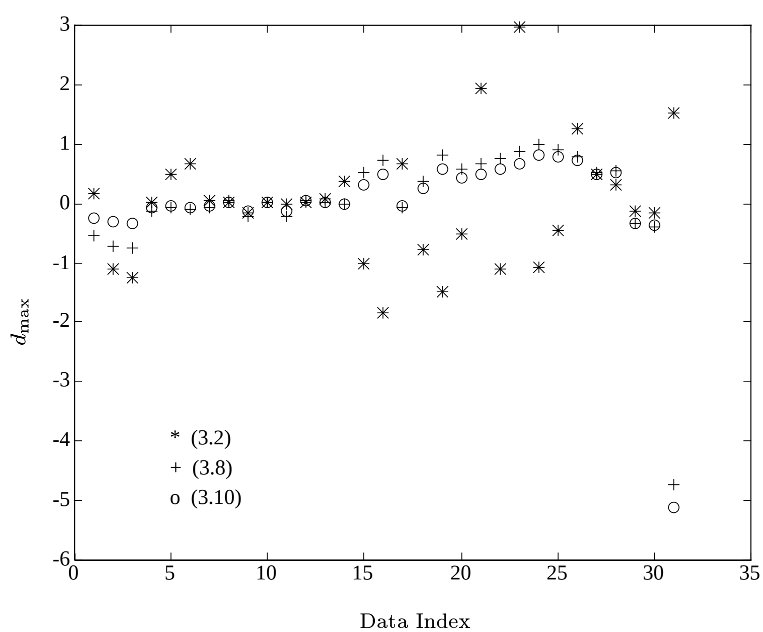

Figure 9. Index plot of $d_{\max}$ for tree data model (4.4).

Cook and Weisberg (1982, Section 2.4.4) discuss transformations of explanatory variables for the tree data using the method of Box and Tidwell (1962). Their estimate of the transformation parameter for the tree diameter, $\hat{\lambda}_2 \approx 2.53$, is nearly identical to maximum likelihood estimate. However, since this estimate, using the constructed variable $x_2 \log(x_2)$ is based on the assumption that $\hat{\lambda}_2 - 1 = 1.53$ is small, it is possible that this agreement is fortuitous. It is not apparent how the slopes and intercepts found by Cook and Weisberg correspond to those in Table 2. Cook and Weisberg conclude that the tree height should not be transformed, which is consistent with our analysis. From the added variable plots given by Cook and Weisberg (Figure 2.4.11) one might conclude that case 31 is influential because $\hat{e}_{31}$ and $\hat{e}_{w31}$ are simultaneously large for the constructed variable of $x_2 \log(x_2)$. However, an alternative conclusion is that $\hat{e}_{31}$ and $\hat{e}_{w31}$ are part of a trend suggesting that the tree diameter variable should be transformed. The added variable plot for model (4.3) is given in Figure 10 with cases 23 and 31 labeled. Here it is much less clear that cases 23 and 31 are influential than in the index plots in Figures 6 and 7. Moreover, the diagnostics proposed here are also able to identify cases 17, 18 as influential for model (4.4).

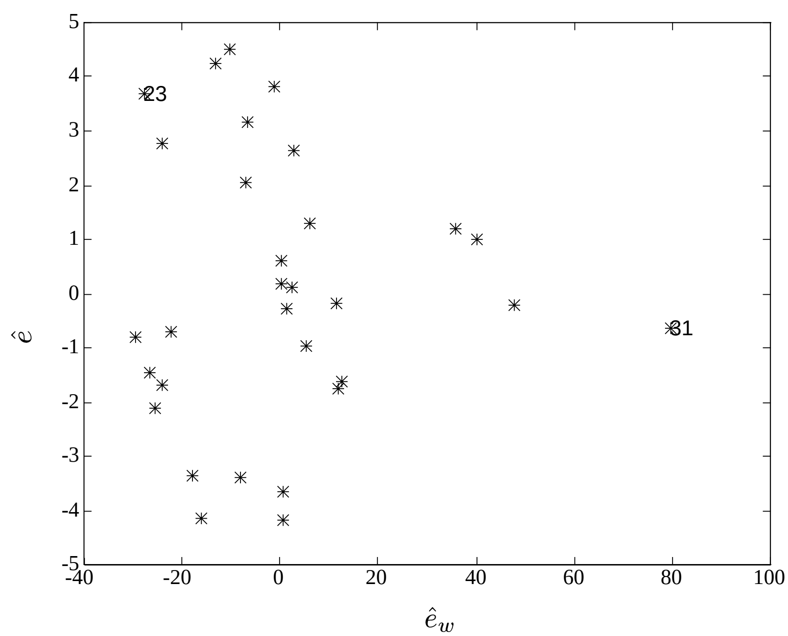

Figure 10. Added variable plot for tree diameter variable for model (4.3).

Figures 2, 6 and 8 compare the two approximations to the likelihood displacement derived in Section 2 with the exact values. Approximation (2.18b) and its generalization (2.22b) are more accurate for small $\hat{\lambda}_{(i)} - \hat{\lambda}$ in the above examples. (This cannot be seen in the figures since $LD(\hat{\lambda}_{(i)})$ is plotted on a linear, rather than logarithmic, scale.) However, the influential points which we wish to identify have relatively large $\hat{\lambda}_{(i)} - \hat{\lambda}$. From the above examples it is seen that both approximations may overestimate or underestimate $LD(\hat{\lambda}_{(i)})$, depending on the data and the model. However, the first approximation, (2.18a) and (2.22a), appears to give a somewhat better indication of the true value of $LD(\hat{\lambda}_{(i)})$ when it is large.

5. Discussion and Conclusion

The diagnostics proposed in this paper provide two approaches to identifying influential cases for the linear model with transformed explanatory variables (1.2). The first approach is through an approximation to the likelihood displacement for one or more case deletions. The second approach is by determining which cases have the greatest local influence under a variety of perturbations. The diagnostics derived using the local influence approach are related to those of Cook (1987); however, we consider a greater variety of perturbations than those he studied.

The diagnostic quantities derived are relatively simple to calculate since they do not require further time-consuming numerical optimizations once the original model has been fitted. At most, they involve straightforward matrix calculations. It is true that formulas for $LD(\hat{\lambda}_{(i)})$ and $d_{\max}$ are more complicated than those for $\hat{e}$ and $\hat{e}_w$ used in the added variable plots of Box and Tidwell (1962). However, the advantage of the diagnostics proposed here is that it is much easier to identify influential points from the index plots of $LD(\hat{\lambda}_{(i)})$ or $d_{\max}$ than from added variable plots.

The numerical results for the two data sets analyzed demonstrate that the diagnostics are useful in identifying influential cases and potential outliers. However, these two examples also demonstrate that the diagnostics need to be interpreted carefully. For some data sets a case may be influential according to virtually all diagnostics. On the other hand, one may discover that some cases are influential according to some diagnostics but not influential according to others. This would suggest that several different diagnostics should be computed, so that all potential outliers and influential points can be identified.

The local influence of a case depends strongly on the form of the perturbation. This was demonstrated by the snow geese data where a suspected outlier is not influential under additive perturbations of the transformed variable but was influential under proportional perturbations. For this data set proportional perturbations seem more appropriate because the errors in flock size appear to be proportional rather than additive. However, for other data sets an additive perturbation may be more appropriate.

Acknowledgement

This research was supported in part by a grant from the Hong Kong Research Grants Council. The authors would like to thank the referees for several valuable suggestions.

Appendix

In this appendix a lemma is derived that is used in Sections 2 and 3.

Lemma. *Suppose that* $Q = \Omega - \Omega X(X^T\Omega X)^{-1}X^T\Omega$ *and* $Q_0 = I - X(X^TX)^{-1}X^T$. *If the* $n \times p$ *matrix* X *is a function of a single variable* u *and* $\dot{X} = dX/du$, *then*

$$\frac{\partial Q}{\partial u} = -Q\dot{X}(X^T\Omega X)^{-1}X^T\Omega - \Omega X(X^T\Omega X)^{-1}\dot{X}^TQ, \tag{A.1}$$

$$\frac{\partial Q_0}{\partial u} = -Q_0\dot{X}(X^TX)^{-1}X^T - X(X^TX)^{-1}\dot{X}^TQ_0. \tag{A.2}$$

If the $n \times n$ *matrix* Ω *is a function of a single variable* v *and* $\Omega(v_0) = I$, $(\partial\Omega/\partial v)_{v_0} = E$, *then*

$$\left.\frac{\partial Q}{\partial v}\right|_{v_0} = Q_0EQ_0. \tag{A.3}$$

Proof. Using the formula for the derivatives of a matrix inverse,

$$\begin{aligned}\frac{\partial Q}{\partial u} =& -\Omega\dot{X}(X^T\Omega X)^{-1}X^T\Omega - \Omega X(X^T\Omega X)^{-1}\dot{X}^T\Omega \\ &+\Omega X\left[(X^T\Omega X)^{-1}\frac{\partial(X^T\Omega X)}{\partial u}(X^T\Omega X)^{-1}\right]X\Omega \\ =& -\Omega\dot{X}(X^T\Omega X)^{-1}X^T\Omega + \Omega X(X^T\Omega X)^{-1}(X^T\Omega\dot{X})(X^T\Omega X)^{-1}X\Omega \\ &-\Omega X(X^T\Omega X)^{-1}\dot{X}^T\Omega + \Omega X(X^T\Omega X)^{-1}(\dot{X}^T\Omega X)(X^T\Omega X)^{-1}X\Omega \\ =& -Q\dot{X}(X^T\Omega X)^{-1}X^T\Omega - \Omega X(X^T\Omega X)^{-1}\dot{X}^TQ\end{aligned}$$

which gives (A.1). If $\Omega = I$, then (A.1) reduces to (A.2). A similar derivation yields (A.3) (see also Lawrance (1988)).

References

Atkinson, A. C. (1983). Diagnostic regression analysis and shifted power transformations. Technometrics **25**, 23-33.

Atkinson, A. C. (1986). Diagnostic tests for transformations. Technometrics **28**, 29-37.

Box, G. E. P. and Tidwell, P. W. (1962). Transformation of the independent variables. Technometrics **4**, 531-550.

Carroll, R. J. and Ruppert, D. (1987). Diagnostics and robust estimation when transforming the regression model and the response. Technometrics **29**, 287-299.

Cook, R. D. (1986). Assessment of local influence (with discussion). J. Roy. Statist. Soc. Ser.B **48**, 133-155.

Cook, R. D. (1987). Influence assessment. J. Appl. Statist. **14**, 117-131.

Cook, R. D. and Wang, P. C. (1983). Transformations and influential cases in regression. Technometrics **25**, 337-343.

Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression, Chapman and Hall, London.

Hinkley, D. V. and Wang, S. (1988). More about transformations and influential cases in regression. Technometrics **30**, 435-440.

Lawrance, A. J. (1988). Regression transformation diagnostics using local influence. J. Amer. Statist. Assoc. **83**, 1067-1072.

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976). Minitab Student Handbook. Duxbury Press, North Scituate, MA.

Thomas, W. (1990). Influence on confidence regions for regression coefficients in generalized linear models. J. Amer. Statist. Assoc. **85**, 393-397.

Tsai, C.-L. and Wu, X. (1990). Diagnostics in transformation and weighted regression. Technometrics **32**, 315-322.

Weisberg, S. (1980). Applied Linear Regression. John Wiley, New York.

Weissfeld, L. A. (1990). Influence diagnostics for the proportional hazards model. Statist. Probab. Lett. **10**, 411-417.

Department of Mathematics, Southeast University, Nanjing 210018.

Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong.

(Received April 1993; accepted March 1995)