SOME NEW ESTIMATION METHODS FOR WEIGHTED REGRESSION WHEN THERE ARE POSSIBLE OUTLIERS

David M. Giltinan
Merck Sharp and Dohme Research Laboratories
P.O. Box 2000, Rahway, NJ 07065

Raymond J. Carroll David Ruppert
University of North Carolina
Chapel Hill, NC 27514 USA

ABSTRACT

The problem of estimating the variance parameter robustly in a heteroscedatic linear model is considered. The situation where the variance is a function of the explanatory variables is treated. To estimate the variance robustly in this case, it is necessary to guard against the influence of outliers in the design as well as outliers in the response. By analogy with the homoscedastic regression case, two estimators are proposed which do this. Their performance is evaluated on a number of data sets. We had considerable success with estimators that bound the "self-influence", that is, the influence an observation has on its own fitted value. We conjecture that in other situations, for example, homoscedastic regression, bounding the self-influence will lead the estimators with good robustness properties.

Acknowledgement

The research of Professor Carroll and Dr. Giltinan was supported by the Air Force of Scientific Research AFOSR F 49620-82-C-0009. Professor Ruppert's research was supported by the National Science Foundation MCS-8100748.

1. Introduction

Heteroscedastic linear models occur frequently in practice, in such fields as radioimmunoassay (Finney (1976) and Rodbard and Frazier (1975)), chemical kinetics (Box and Hill (1974)), and econometrics (Hildreth and Houck (1968)). A number of recent papers are concerned with modeling the variance as a parametric function either of the mean response or the independent variables; see, for example, Box and Hill (1974), Carroll and Ruppert (1982), and Dent and Hildreth (1977).

Modeling the heteroscedasticity not only allows the statistician to understand better the nature of the statistical variability in the data, but also to estimate the mean response more efficiently by using estimated variances as weights.

In this paper we introduce outlier-resistant (or "bounded-influence") estimation methods for such heteroscedastic linear models. Outlier-resistant methods are, of course, useful when the error distribution is non-normal, but they are of benefit as well for other types of deviations from the ideal model. For example, the model for the mean response or for the variance may be adequate only over a restricted region of the design space. At the extreme points of the design these parametric models may break down; this is a particularly dangerous situation since these so-called high-leverage points are extremely influential in determining the values of the estimators unless "bounded-influence" estimators are used. Also, an occasional value of an independent variable may be grossly in error, for example, due to a recording error, though otherwise the x-variables are assumed to be without error - we are *not* thinking of the errors-in-variables problem where the independent variables are measured with errors. Here, again, the use of traditional, unbounded-influence estimators is risky since a disastrously wrong x-value is likely to have high leverage.

The preceeding motivation for using bounded-influence methods is as compelling for homoscedastic models as for heteroscedastic models, and Mallows (1975), Hampel (1978), and Krasker and Welsch (1982) have already begun the development of outlier-resistant methodologies for the homoscedastic case.

Here we extend the work of Mallows (1975) and Krasker and Welsch (1982) to heteroscedastic models. We begin by considering an example, reported by Leurgans in the *Biostatistics Casebook* (1980). The data are concerned with the comparison of a test method and a reference method for measuring glucose concentration in blood, and consist of 46 pairs of measurements (x_i, y_i), x being the reference and y the test method. A scatter plot of y against x for these data reveals a pronounced linear trend. However, a plot of the least squares residuals against the independent variable, x, gives a clear indication of heteroscedasticity, the variances tending to increase with the value of x.

On encountering such heterogeneity of variance in the data, there are two traditional approaches; to transform the data, or to perform a weighted least squares type of analysis. Leurgans chose to perform a log transformation, fitting the model

$$\ell n\ y_i = \alpha' + \beta' \ \ell n\ x_i + e_i \tag{1.1}$$

While Leurgans' choice of transformation is certainly reasonable for these data, in many applied situations one might be reluctant to transform, since transformation can make it difficult to make inference in the original scale. In such cases, a weighted analysis may be preferable. We shall consider weighted analyses, assuming an underlying model of the form

$$y_i = \alpha + \beta\ x_i + \sigma_i \varepsilon_i, \tag{1.2}$$

for the glucose data, where the $\{\varepsilon_i\}$ are independent and identically distributed disturbances and the $\{\sigma_i\}$ reflect the heterogeneity of variance in the model. The optimal normal theory weighted analysis of the data would involve basing inference on the transformed variables $y_i^* = y_i/\sigma_i$, $x_i^* = x_i/\sigma_i$. Since the σ_i's are typically unknown, in practice one will work instead with $y_i^* = y_i/\hat{\sigma}_i$, $x_i^* = x_i/\hat{\sigma}_i$, where $\hat{\sigma}_i$ is an estimate of σ_i. Clearly, for efficient inference, one would like to use a good estimate of the σ_i's. In the absence of replication, the usual approach is to assume some parametric model for the σ_i's; see Carroll (1982) for an alternative. For the Leurgans data we modelled the variance in the following way:

$$\sigma_i = \sigma |x_i|^\lambda, \tag{1.3}$$

somewhat analogous to a model used by Box and Hill (1974), although in their model, the variance is a function of the mean.

A number of estimation procedures have been proposed for models similar to that described by (1.3). Typically, a preliminary estimate of β is obtained, and the residuals from this preliminary fit are used to gain information about the variances. This information is then used to obtain a more efficient estimate of β. For example, Box and Hill (1974) and Jobson and Fuller (1980) both suggest a form of generalized least squares, hereafter denoted GLS. If one assumes a normal error distribution, one can proceed in the following three stages:

(i) Obtain a preliminary unweighted least squares estimate of β, say $\hat{\beta}_p$,

(ii) Obtain a maximum likelihood estimate of the variance parameters (σ,λ) pretending that $\beta = \hat{\beta}_p$, and treating the residuals as if they were the true errors, and form estimated standard deviations $\hat{\sigma}_i = \hat{\sigma}|x_i|^{\hat{\lambda}}$.

(iii) Repeat stage (i) with the transformed variables $y_i^* = y_i/\hat{\sigma}_i$, $x_i^* = x_i/\hat{\sigma}_i$.

When applied to the data set in our example, this generalized least squares estimation procedure yielded a value of .52 for the estimate of λ. In her analysis of the log-transformed data, Leurgans identified observation #31 (the point with the lowest x- and y-value) as a massive outlier, and deleted it from her analysis. On deletion of this point from the data set, the GLS estimation procedure above gave a value of .86 for the estimate of λ. A change of this size in the parameter estimate, on deletion of a single point, is disturbing and indicates an undesirable sensitivity of the maximum likelihood method to a small fraction of the data.

In the homoscedastic regression case, two general approaches to this kind of problem have been considered:

(i) the influence approach, where the focus is on identifying those points in the data set which have a large influence on inferences drawn from the data; see Belsley, Kuh and Welsch (1980) and Cook and Weisberg (1983).

(ii) developing new estimation techniques which automatically limit the influence of any small subset of the data on parameter estimates.

In our experience, the two approaches complement rather than compete with one another. Surprisingly little has been done to make the two approaches applicable for heteroscedastic regression. We shall pursue the second approach by developing bounded influence estimators for the variance parameter θ in heteroscedastic models. We emphasize that we are *not* interested in developing a "black-box" estimation procedure - we would like our estimators to provide diagnostic information about influential points, and to add to our understanding of the structure of the data.

In Section 2 we show, by considering the influence function, why the MLE for the variance parameter is so sensitive to outliers, and we describe an estimator which partially alleviates this sensitivity. We see that, analogously to the homoscedastic regression case, it is not just outliers in the response which exert a large influence, but also outliers in the design, or high leverage points.

In Sections 3 and 4 we consider two classes of estimators for the variance parameter which alleviate the sensitivity of the estimation procedure to extreme data points. In Section 5 the performance of these estimators is evaluated on a number of data sets, and their computation is discussed in the appendix.

2. The model

In what follows we consider the heteroscedastic linear model

$$y_i = x_i'\beta + \sigma_i\varepsilon_i; \quad i=1,\dots,n \tag{2.1}$$

where β is a p×1 vector of regression parameters to be estimated, x_i is a p×1 vector, and the $\{\varepsilon_i\}$ are independent with a distribution F which is assumed to be symmetric with mean 0 and variance 1. The quantities σ_i^2 are the variances of the y_i and reflect the heterogeneity of variance in the model.

A common situation in practice is that σ_i is assumed to be some known function of the explanatory variables. A very flexible model for this situation is

$$\sigma_i = \exp[h'(x_i)\theta], \qquad (2.2)$$

where h is a known $\mathbb{R}^q$-valued function and θ is an unknown q×1 vector of variance parameters. Note that (2.2) includes the power model $\sigma_i = \sigma|x_i|^\lambda$ of the previous section, by taking

$$h(x) = \begin{bmatrix} 1 \\ \log|x| \end{bmatrix} \quad \text{and} \quad \theta = \begin{bmatrix} \log\sigma \\ \lambda \end{bmatrix} .$$

In many cases, σ_i may reasonably be modelled as a function of the mean response, τ_i, which we might express as

$$\sigma_i = \exp[h'(\tau_i)\theta] \qquad (2.3)$$

analogously to (2.2). The techniques developed below for the model in (2.2) extend readily to that given by (2.3). Details may be found in Giltinan (1983). For brevity, we shall discuss only the first model in this article. We shall be concerned with obtaining bounded influence estimates of the variance parameter θ.

To do this, it proves helpful to investigate why the maximum likelihood estimator of θ described above is so sensitive to outliers. Under a normal error assumption, if β were known, the MLE of θ would solve the equation

$$\sum_{i=1}^{n}\left\{\left(\frac{y_i-x_i'\beta}{\exp[h'(x_i)\hat{\theta}_{MLE}]}\right)^2 - 1\right\} h(x_i) = 0. \tag{2.4}$$

The influence function (Hampel; 1968, 1974) for the maximum likelihood estimator is proportional to $(\varepsilon^2-1)h(x)$, where $\varepsilon = (y-x'\beta)\exp[-h'(x)\theta]$, and is thus unbounded. Since the influence function for the MLE is quadratic in the residual ε, in theory a point with a sufficiently large residual can have an arbitrarily large effect on the maximum likelihood estimate of θ. Carroll and Ruppert (1982) suggest guarding against this by replacing the term in braces in (2.4) by a bounded function $\chi(\varepsilon)$. In practice, one would also replace $x_i'\beta$ by $x_i'\hat{\beta}_p$ in (2.4); that is, they suggest solving

$$\sum_{i=1}^{n} \chi\left[\frac{y_i-x_i'\hat{\beta}_p}{\exp[h'(x_i)\hat{\theta}]}\right] h(x_i) = 0 \tag{2.5}$$

to obtain $\hat{\theta}$. They suggest a choice of χ which generalizes the classical Huber Proposal 2 for the homoscedsatic case. Implementation of this technique with the glucose data gave estimates of λ as 0.72 and 0.87 based on the full and reduced data sets respectively. Bounding the effect of a large residual certainly narrows the gap, but not by as much as one might like. Examination of the influence function for this estimation method reveals why this is the case - it is proportional to $\chi(\varepsilon)h(x)$. Choice of a bounded χ-function limits the effect of large ε on the estimate, but the factor of $h(x)$ may still lead to unbounded influence. A moderate residual occurring in conjunction with a large value of $|h(x)|$ may still have quite an effect on $\hat{\theta}$. Clearly, if one wishes to protect against this, the estimator for θ must limit not only the effect of large residuals, but also of large values of $|h(x)|$.

This is analogous to the situation in the usual homoscedastic regression case. In that situation, the least squares estimate of β is the solution to

$$\sum_{i=1}^{n} (y_i - x_i'\beta) x_i = 0 \tag{2.6}$$

and has influence function proportional to $r x$, where $r = y - x'\beta$. The influence function will be large in absolute value if either $|r|$ is large, or if $\|x\|$ is large - i.e. if a point has a large residual or high leverage. Various suggestions have been made in the recent literature on how to modify (2.6) to obtain efficient bounded influence estimates of β (e.g. Mallows, 1975; Maronna, Bustos and Yohai, 1979; Krasker, 1980; Krasker and Welsch, 1982). Extension of these methods to the problem of estimating θ is discussed in Giltinan (1983). In this article we focus on estimates of the type proposed by Mallows (1975) and by Krasker and Welsch (1982); the paper by Krasker and Welsch also discusses Mallows estimators.

One approach to bounded influence estimation proceeds from the view that one is interested chiefly in limiting the sensitivity of the estimate to anomalous data, that is, in bounding $\gamma_1 := \sup_{(x,y)} |IF(x,y)|$, where $IF(x,y)$ is the influence function at a point (x,y). This definition is not invariant to a change of coordinate system, and Krasker and Welsch propose two alternative definitions of sensitivity which circumvent this lack of invariance:

$$\gamma_2 := \sup_{(x,y)} \sup_{\lambda \neq 0} \frac{|\lambda' IF(x,y)|}{(\lambda' V \lambda)^{\frac{1}{2}}} = \sup_{(x,y)} [IF(x,y) V^{-1} IF(x,y)]^{\frac{1}{2}},$$

and

$$\gamma_3 := \sup_{(x,y)} |x' IF(x,y)|,$$

if the maximum influence a point can have on its own fitted value is of interest. Hampel (1978) refers to γ_3 as the "self-influence". Stahel (1981) calls γ_2 the "self-standardized sensitivity", because the estimator's influence function is normed by the covariance matrix of the estimator itself.

γ_2 is appropriate if one is chiefly concerned with inferences about β; $\lambda' IF(x,y)$ is the influence function of the linear function $\lambda'\hat{\beta}$ of the estimate and $(\lambda' V \lambda)^{\frac{1}{2}}$ is the (asymptotic) standard deviation of $\lambda'\hat{\beta}$. Therefore, bounding γ_2 insures that the standardized influence of $\lambda'\hat{\beta}$ is below a certain bound for all λ.

γ_3 is appropriate when one is more concerned with estimating the response surface at the observed x's. Using γ_3 rather than γ_2 leads to more extreme downweighting of high leverage points; this can be understood intuitively by noticing that for least-squares estimation where $IF(x,y)$ is proportional to rx, $\lambda' IF(x,y)$ and $x' IF(x,y)$ are, respectively, linear and quadratic in x. Thus, the amount of downweighting needed to keep γ_2 or γ_3 bounded is of order $\|x\|^{-1}$, respectively, $\|x\|^{-2}$.

In practice, we had most success when working with γ_3. We judged success by the stability of the overall inference, that is, how little inferences were changed by the deletion of one, or a few, influential points. In what follows, therefore, we present estimators and results developed with a view to bounding $\gamma_3 = \sup_{(x,y)} |x' IF(x,y;\beta)|$ in the regression case, and the analogous quantity $\gamma_3 = \sup_{(\varepsilon,x)} |h'(x) IF(\varepsilon,x;\theta)|$ in the variance estimation case. We remark that the methods generalize easily to cover the case of bounding γ_1 or γ_2, see e.g. Giltinan (1983).

Our success when estimating θ with methods that bound γ_3 lead us to conjecture that bounding γ_3 should be tried in other problems as well, e.g. homoscedastic regression. This may possibly lead to progress on the "masking problem" where two or more highly influential points mask each other.

3. Joint influence estimates

For technical convenience, throughout this and the next section it will be useful to treat the design points $\{x_i\}$ as if they were a sample from a distribution function H and independent of the disturbances $\{\varepsilon_i\}$. In practice, our estimators are equally appropriate for fixed and random $\{x_i\}$, but theoretical properties are more easily discussed for random design vectors.

In the homoscedastic case, Krasker and Welsch (1982) proposed estimating β by solving an equation of the type

$$\sum_{i=1}^{n} w(y_i,x_i)x_i(y_i-x_i'\hat{\beta}_{KW}) = 0 \tag{3.1}$$

where w(y,x) is a non-negative bounded continuous weight function which depends on y only through the absolute residual, $|r| = |y-x'\beta|/\sigma$, and therefore for symmetrically distributed ε gives Fisher consistency at the normal model for each x in $\mathbb{R}^p$:

$$E_\Phi\ w(\varepsilon,x;\beta)\varepsilon = 0. \tag{3.2}$$

The influence function for the estimator solving (3.1) is given by

$$IF_w(y,x;\beta) = w(y,x)\ M_w^{-1}x(y-x'\beta),$$

where

$$M_w = E[w(y,x)\ (\frac{y-x'\beta}{\sigma})^2\ x\ x'].$$

For this class of estimators of β,

$$\gamma_3 = \sup_{(x,y)} |x'IF_w(y,x;\beta)| = \sup_{(x,y)} w(y,x)\ x'M_w^{-1}x\ |y-x'\beta|.$$

Clearly, the quantity γ_3 may be bounded by suitable choice of w. Krasker and Welsch address the question of choosing w to bound γ_3 in an efficient manner. The weight function which they suggest has the intuitively appealing form: downweight an observation only if its influence would otherwise exceed the maximum allowable influence, else give the observation a full weight of one. That is, if γ_3 is not to exceed the bound a_3, then the appropriate weight function satisfies

$$w(y,x;\beta) = \text{Min}\left[1, \frac{a_3}{\left|\frac{y-x'\beta}{\sigma}\right| x' M_w^{-1} x}\right] \tag{3.3}$$

Such a weight function will downweight an observation which has a large residual, or high leverage. As discussed in the last section, for fixed $\left|\frac{y-x'\beta}{\sigma}\right|$ $w(y,x;\beta)$ is of order $||x||^{-2}$ as $||x|| \to \infty$.

We may extend the method of Krasker and Welsch to cover the estimation of the variance parameter θ as follows. Consider the class of estimators of θ obtained by solving an equation of the form

$$\sum_{i=1}^{n} w(y_i,\tau_i;\hat{\theta})[(\varepsilon_i(\hat{\theta}))^2-1]h(x_i) = 0 \tag{3.4}$$

where w is a bounded, continuous non-negative weight function, depending on y only through the residual $\varepsilon(\theta) = (y-\tau)\exp[-h'(x)\theta]$. The influence function for $\hat{\theta}$ is given by

$$IF_w(y,\tau;\theta) = w(y,\tau;\theta)\; M_w^{-1}h(x)(\varepsilon^2-1) \tag{3.5}$$

where

$$M_w = E\{w(\varepsilon,\tau)(\varepsilon^2-1)^2\; h(x)h'(x)\}.$$

Thus, for this class of estimators of θ,

$$\gamma_3 = \sup_{(\varepsilon,x)} |h'(x)IF_w(\varepsilon,x;\theta)| = \sup_{(\varepsilon,x)} h'(x)M_w^{-1}h(x)w(\varepsilon,x)|\varepsilon^2-1|.$$

By analogy to the homoscedastic regression case, if one wishes to choose w subject to a bound a_3 on γ_3, then a reasonable choice of w is as follows:

$$w(\varepsilon,\tau;\theta) = \text{Min}\left[1, \frac{a_3}{|\varepsilon^2-1|h'(x)M_w^{-1}h(x)}\right]. \tag{3.6}$$

The similarity with (3.3) is clear. It may be shown that under normality of ε,

$$M_w = E\left[g\left(\frac{a_3}{h'(x)M_w^{-1}h(x)}\right) h(x)h'(x)\right] \tag{3.7}$$

where $g(u) := E_\Phi[\mathrm{Min}|z^2-1|,u]\ |z^2-1|$.

The methods suggested above may be combined into a single estimation procedure, the details for which are given in the appendix. Since the goal has been to bound influence simultaneously over the design and the residuals, we call this a joint influence estimate. For the power model (1.1) applied to the glucose data, the joint influence estimate of λ changed only from 0.84 to 0.87 with the deletion of observation 31. If observation 31 is included, then its weight (3.6) equals 0.07, strongly supporting Leurgans' deletion of the point.

4. Separate influence estimates

A second approach which we have found useful in estimating θ is to handle high leverage points and outliers separately. We do this by adapting an idea of Mallows, see Krasker and Welsch (1982). Specifically, if the mean $\tau_i = x_i'\beta$ were known, then a Mallows-type estimator for θ would solve

$$\sum_{i=1}^{n} h_i w(h_i) \chi\left(\frac{(y_i - \tau_i)}{\exp(h_i'\theta)}\right) = 0, \tag{4.1}$$

where $h_i = h(x_i)$, $0 \le w(h_i) \le 1$ is a weight function and $\chi(\cdot)$ is an even function with mean zero. If $w(h_i) = 1$ and $\chi(v) = v^2 - 1$, the solution to (4.1) is the maximum likelihood estimate. The influence function for $\hat{\theta}$ solving (4.1) is

$$IF(\varepsilon,x,\theta) = M_w^{-1} h\, w(h)\ \chi(\varepsilon)/E(\varepsilon\dot{\chi}(\varepsilon)),$$

where $\quad M_w = E[w(h)h\,h']$.

The asymptotic covariance of $\hat{\theta}$ is

$$V(w,\chi) = M_w^{-1} N_w M_w^{-1} E\chi^2(\varepsilon)/\{E\,\varepsilon\,\dot{\chi}(\varepsilon)\}^2,$$

where

$$N_w = [E\,w^2(h)h\,h'],$$

so that the self-influence is

$$\gamma_3 = \sup|h'\,IF(\varepsilon,x,\theta)| = \sup_x h'M_w^{-1}h \sup_\varepsilon \chi(\varepsilon)/E\{\varepsilon\,\dot{\chi}(\varepsilon)\} .$$

Note that both the covariance matrix and the self-influence are products of functions of (h,w) and (ε,χ). To bound the self-influence efficiently, we follow the approach of Maronna, Bustos and Yohai (1979) in the homoscedastic case and split the problem into two parts, minimizing

$$M_w^{-1} N_w M_w^{-1} \quad \text{and} \quad E\chi^2(\varepsilon)/\{E\varepsilon\dot{\chi}(\varepsilon)\}^2$$

subject to bounds on

$$\sup_x h(x)'M_w^{-1}h(x) \quad \text{and} \quad \sup_\varepsilon \chi(\varepsilon)/E\{\varepsilon\dot{\chi}(\varepsilon)\},$$

respectively. By analogy with the previous section, we suggest choosing

$$w(x) = \text{Min}[1,a/x'M_w^{-1}x],$$

where

$$M_w = E\{\text{Min}[1,a/x'M_w^{-1}x]\ x\,x'\}$$

and in practice we estimate M_w by solving

$$\hat{M}_w = n^{-1}\sum_{i=1}^{N} \text{Min}[1,a/x_i'M_w^{-1}x_i]\ x_i x_i' .$$

The estimated design weights are

$$w_i = w(x_i) = \text{Min}[1,a/x_i'\hat{M}_w^{-1}x_i]. \qquad (4.2)$$

The effect of large residuals may be controlled by choice of the function χ - one possibility is to take $\chi(\cdot) = \psi_k^2(\cdot)-\xi$, where ψ_k is a Huber psi-function and ξ is a constant chosen to give consistency at the normal model.

The techniques described in this section may be combined to obtain a bounded influence estimate for (β,θ); see the appendix for details. A descriptive term for the process is separate influence estimates. Since the weights (4.2) depend only on the design, they serve mostly as an indicator of leverage in the values $\{h(x_i)\}$. We have had moderate success with the following indicator of large residuals:

$$\psi(r_i)/r_i, \; r_i = (y_i - x_i' \hat{\beta}_p)/\exp(h'(x_i)\hat{\theta}), \tag{4.3}$$

where $\psi(\cdot)$ is the Huber function

$$\psi(x) = \max(-c, \min(x,c)), \tag{4.4}$$

and $\hat{\beta}_p$ is a preliminary estimate of β as in the appendix. When applied to the glucose data, the separate influence estimate of λ in (1.1) changed only from 0.83 to 0.90 with the deletion of observation 31. The design weight (4.2) was 0.46, with residual weight (4.3) given by .364.

The choice of tuning constants "a" in (4.2) and "c" in (4.4) is up to the user. We generally vary these constants, using smaller values for diagnostic purposes and larger values for inference. In this paper, to save space, we have chosen fixed values which represent a compromise between the two goals.

5. Examples

(1) Glucose data: Table 5.1 summarizes the parameter estimates for the Leurgans data set, obtained by each of the four methods described above. The gain in stability of the parameter estimates by using bounded influence methods is evident from the table. The price exacted is that robust estimators are asymptotically less efficient at the normal model. While in this example the robust methods do seem to have higher estimated standard errors for the full data, with the outlier included, they are probably better than the optimistic GLS estimates.

Since predicting Y from X is a primary concern for these data, the widths of prediction intervals are of interest. We therefore computed such prediction intervals for a mean at various values in the range of X, using each of the four methods, and investigated the change upon deleting observation #31. The results are summarized in Table 5.2, which gives ratios of confidence interval lengths for different values of X for the respective estimation methods.

Again, the gain in stability by using bounded influence methods is obvious from Table 5.2 The stability is obtained at a cost - the prediction interval lengths for the joint and separate influence methods show an increase typically of about 10-16% over those obtained by maximum likelihood methods, when operating on the reduced data. This is, of course, in agreement with what is known to happen exactly at the normal error model.

Both the joint and separate influence procedures downweighted observation #31 considerably in the full data set. In this example, our attention had already been drawn to the point; however, sometimes the weights provided by the bounded influence methods are a useful diagnostic tool in drawing our attention to previously unsuspected influential points. This is illustrated by our next example.

(2) Gas vapor data: Our second example is taken from Weisberg (1980, page 146). The data are concerned with the amount of gas vapor emitted into the atmosphere when gasoline is pumped into a tank. In a laboratory experiment to investigate this, a sequence of 32 experimental fills was carried out. Four variables were thought to be relevant for predicting Y, the quantity of emitted hydrocarbons:

x_1 = the initial tank temperature, in °F

x_2 = the temperature of the dispensed gasoline, in °F

x_3 = the initial vapor pressure in the tank, in psi

x_4 = the vapor pressure of the dispensed gasoline, in psi

This data set has been used by Cook and Weisberg (1983) to illustrate their proposed score test for heteroscedasticity. They found definite evidence of heteroscedasticity, with the variance being a function of x_1 and x_4, and obtained this empirical estimate of the direction in which the variance is increasing:

$$x_5 = 0.778 + .110\, x_1 - 1.432\, x_4 .$$

A plot of the residuals from fitting the ordinary least squares model against x_5 shows obvious heteroscedasticity. We used a power model for the variance:

$$\sigma_i = \sigma |x_{5i} + 0.5|^{\lambda} , \qquad (5.1)$$

adding 0.5 to make $x_{5i} + 0.5 \geq 0$ in all observations, and analyzed the data using our three-stage estimation procedure. We employed the same four estimation techniques as in the previous example. All the discussion will focus on estimation of λ, although the estimation of β after appropriate weighting is interesting and nontrivial.

Table 5.3 summarizes the estimates of λ and associated standard errors, provided by the different estimation methods. For these data, Cook and Weisberg note that for unweighted least squares there are no unduly influential cases. Upon performing the initial analyses on the full data, we thus were surprised at the disparity in the estimate of λ provided by the different techniques. It turns out that least squares estimation of β and generalized least squares estimation of λ operate entirely differently. Closer examination reveals that the bounded influence methods downweighted observations #1 and #2 considerably in the estimation of λ, while giving close to full weight to all other data (see Table 5.3). This suggests that these two points, observation #1 in particular, exert considerable influence in determining the estimate of λ. This is borne out by Table 5.3; the generalized least squares estimate, which makes no attempt to control the influence of observations #1 and #2 in the full data set, changes considerably on their deletion. The bounded influence techniques, on the other hand, control their influence in the full data set and are relatively insensitive to their deletion.

Again, we find that the bounded influence methods provide credible, stable estimates at the price of some loss of efficiency. This example illustrates their use also as a diagnostic method for locating influential points. We stress that an influential observation is not necessarily 'bad' - it may, in fact, be highly informative. We do feel, however, that it is of value to be able to identify points which are highly influential. The methods proposed in this article provide one way of doing this, while at the same time providing the possibility of inference based on the well-known ideas of M-estimation.

Upon reflection, we realized that much of the difficulty with observations #1 and #2 in terms of estimating λ is due to the model (5.1). For the first two observations, $x_{5i} + 0.5$ is very nearly zero so that there is a huge relative disparity among the values of $\{x_{5i}+0.5\}$ in the data. Thus the strong influence of the first two observations on estimating λ is somewhat artificial and can be weakened by using $x_{5i}+1.0$.

This example is a good illustration of the power of our techniques to provide stable estimation and inference. However, from the view of modelling the variances it is somewhat unsatisfactory mainly because we used the linear combination x_5 as given. For example, one might consider an alternative variance model

$$\sigma_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i}. \qquad (5.2)$$

It is not clear how best to adapt our techniques to the model (5.2), and some care will be necessary. This data set is actually quite a difficult one for model (5.2). A generalized least squares estimate gives very large weight to the first two observations, while the likelihood appears to be unbounded.

6. Conclusion

We have introduced new methods of estimation for heteroscedastic linear models when the variances are a power of the mean or a single predictor. These methods are simple adaptations of the ideas of bounded influence regression. The estimation methods should serve as a complement to influence techniques such as the graphs of Cook and Weisberg (1983) or the deletion diagnostic ideas of Cook (1985).

Table 5.1 Glucose data estimates of λ for the model (1.2), with estimated standard errors in parentheses

	Complete Data	Reduced Data without #31
Generalized Least Squares	0.52 (0.18)	0.86 (0.17)
Carroll-Ruppert	0.72 (0.19)	0.87 (0.21)
Joint Influence Method	0.84 (0.20)	0.87 (0.19)
#31 Weight for θ	.07	
#31 Weight for β	.04	
Separate Influence Method	0.83 (0.21)	0.90 (0.20)
#31 Weight for θ	.46	
#31 Weight for β	.16	

Table 5.2 Ratio of length of mean prediction confidence interval for the reduced glucose data versus the complete glucose data.

Value of x	Generalized Least Squares	Carroll-Ruppert	Joint Influence Method	Separate Influence Method
60	0.92	1.05	1.00	0.98
100	0.87	0.95	0.96	0.94
180	1.14	1.05	0.99	0.99
260	1.20	1.10	1.00	1.00
380	1.21	1.12	1.01	1.01

Table 5.3 The Gas Vapor data estimates for λ in model (5.1). Estimated standard errors are in parentheses.

	Full Data	Without Observ. #1	Without Observ. #2
Generalized Least Squares	0.28 (0.09)	0.50 (0.11)	0.81 (0.18)
Carroll-Ruppert	0.28 (0.14)	0.40 (0.18)	0.90 (0.24)
Joint Influence Method	1.01 (0.29)	1.03 (0.25)	1.04 (0.29)
Weight for #1 in θ	.00		
Weight for #2 in θ	.00	.00	
Separate Influence Method	0.76 (0.22)	0.81 (0.23)	0.84 (0.31)
Weight for #1 in θ	.02		
Weight for #2 in θ	.03	.03	

Appendix Computation of the bounded influence estimates

For completeness, we include here the estimating equations for the three-stage bounded influence estimates discussed in Sections 3 and 4. We also outline the algorithm used to compute the three-stage separate influence estimator of Section 4. A similar algorithm may be used to compute the joint influence estimator of Section 3. Details are given in Giltinan (1983).

The three-stage joint influence estimator is obtained as follows:

Stage 1:

Solve

$$\sum_{i=1}^{n} \mathrm{Min}\left[1, \frac{a_1}{\left|\frac{y_i - x_i'\hat{\beta}_p}{\hat{\sigma}_1}\right| x_i'\hat{M}_1^{-1}x_i}\right](y_i - x_i'\hat{\beta}_p)x_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} f\left(\frac{a_1}{x_i'\hat{M}_1^{-1}x_i}\right) x_i x_i' - \hat{M}_1 = 0$$

$$\sum_{i=1}^{n} \chi\left[\frac{y_i - x_i'\hat{\beta}_p}{\hat{\sigma}}\right] = 0$$

simultaneously for $\hat{M}_1$, $\hat{\sigma}_1$ and $\hat{\beta}_p$, where $f(u) := E_\Phi(\mathrm{Min}|z|,u)|z$ and χ is an even function satisfying $E\chi(\varepsilon) = 0$.

We used χ of the form $\chi(\cdot) = \psi_c^2(\cdot) - E_\Phi\psi_c^2(z)$, where ψ_c is a Huber psi-function.

Stage 2:

Next solve

$$\sum_{i=1}^{n} \mathrm{Min}\left[1, \frac{a_2}{\left|\left[\frac{y_i - t_i}{\exp[h'(x_i)\hat{\theta}]}\right]^2 - 1\right| h'(x_i)\hat{M}_2^{-1}h(x_i)}\right]\left\{\left(\frac{y_i - t_i}{\exp[h'(x_i)\hat{\theta}]}\right)^2 - 1\right\}h(x_i) = 0$$

and

$$\frac{1}{n} \sum_{i=1}^{n} g\left[\frac{a_2}{h'(x_i)\hat{M}_2^{-1}h(x_i)}\right] h(x_i)h'(x_i) - \hat{M}_2 = 0$$

for $\hat{M}_2$ and $\hat{\theta}$, where $t_i = x_i'\hat{\beta}_p$ and g is as above.

Stage 3:

Form estimated standard deviations, $\hat{\sigma}_i = \exp[h'(x_i)\hat{\theta}]$ and repeat Stage 1, using the transformed variables $x_i^* = x_i/\hat{\sigma}_i$, $y_i^* = y_i/\hat{\sigma}_i$.

For the three-stage separate influence procedure, the estimating equations are as follows:

Stage 1: Solve

$$\sum_{i=1}^{n} x_i \operatorname{Min}\left[1, \frac{a_1}{x_i'\hat{M}_1^{-1}x_i}\right] \psi\left[\frac{y_i - x_i'\hat{\beta}_p}{\hat{\sigma}_1}\right] = 0 \qquad (A.1)$$

$$\hat{M}_1 = \frac{1}{n} \sum_{i=1}^{n} \operatorname{Min}\left[1, \frac{a_1}{x_i'\hat{M}_1^{-1}x_i}\right] x_i x_i' \qquad (A.2)$$

$$\sum_{i=1}^{n} \chi\left[\frac{y_i - x_i'\hat{\beta}_p}{\hat{\sigma}_1}\right] = 0 \qquad (A.3)$$

simultaneously for $\hat{\beta}_p$, $\hat{M}_1$ and $\hat{\sigma}_1$, where ψ is a nondecreasing odd function and χ is an even function $E\chi(\varepsilon) = 0$. In our applications we used a Huber psi-function in (A.1) and χ as in Huber's proposal 2 in (A.3)

Stage 2: Next solve

$$\sum_{i=1}^{n} h(x_i) \operatorname{Min}\left[1, \frac{a_2}{[h'(x_i)\hat{M}_2^{-1}h(x_i)]}\right] \chi\left[\frac{y_i - t_i}{\exp[h'(x_i)\hat{\theta})}\right] = 0 \qquad (A.4)$$

$$\hat{M}_2 = \frac{1}{n} \sum_{i=1}^{n} \operatorname{Min}\left[1, \frac{a_2}{[h'(x_i)\hat{M}_2^{-1}h(x_i)]}\right] h(x_i)h'(x_i) \qquad (A.5)$$

for $\hat{\theta}$ and $\hat{M}_2$, where $t_i = x_i'\hat{\beta}_p$.

Stage 3:

Form estimated standard deviations, $\hat{\sigma}_i = \exp[h'(x_i)\hat{\theta}]$ and repeat Stage 1, using the transformed variables $x_i^* = x_i/\hat{\sigma}_i$ and $y_i^* = y_i/\hat{\sigma}_1$. In solving the system (A.1)-(A.3) we first solved (A.2) iteratively to obtain $\hat{M}_1$, using $\frac{1}{n}\sum_{i=1}^{n} x_i x_i'$ as starting value. We then formed estimated weights, $w_{1i} = \text{Min}[1, a_1/(x_i'\hat{M}_1^{-1}x_i)]$, and proceeded to solve (A.1) and (A.3) by means of an iteratively reweighted least squares algorithm similar to that described in Huber (1981, pages 183-186). The equation (A.5) may be solved in a manner similar to (A.2), and estimated weights for the second stage computed. In solving (A.4) we used subroutine ZXGSN in the IMSL library and assumed that λ was in the interval (-2,2).

The truncation values a_1, a_2 and a_3 in the estimating equations still need to be specified. The matrix M_1 satisfies

$$M_1 = E\,\text{Min}\left[1, \frac{a_1}{x'M_1^{-1}x}\right] x\,x',$$

whence it follows that

$$I_{p\times p} = E\,\text{Min}\left[1, \frac{a_1}{x'M_1^{-1}x}\right] x\,x'\,M_1^{-1}.$$

Taking traces across this equation:

$$p = E\,\text{Min}\left[1, \frac{a_1}{x'M_1^{-1}x}\right] x'M_1^{-1}x = E\,\text{Min}[x'M_1^{-1}x, a_1],$$

so that, if (A.2) is to have a solution, a_1 must exceed p. Similarly, q is a lower bound for a_2 We have had good success in setting the truncation values at one-and-a-half times the lower bound, and the results reported in

Section 5 above used this cutoff value.

For the separate inference estimation procedure, formal influence calculations, given in detail in Giltinan (1983), yield the following expression for the influence function of the final estimate of β:

$$IF(x,y;\beta) = M_3^{-1} \, x \, \mathrm{Min}\left[1, \frac{a_3}{x'M_3^{-1}x}\right] \frac{\psi(y-x'\beta)}{E\,\dot{\psi}(\varepsilon)} .$$

The asymptotic covariance matrix of $\hat{\beta}$ is given by

$$\mathrm{Var}[\sqrt{n}\,(\hat{\beta}-\beta)] = M_3^{-1}N_3M_3^{-1} \, \frac{E\,\psi^2(\varepsilon)}{E^2\dot{\psi}(\varepsilon)}$$

where $N_3 = E\,w_3^2(x)\,x\,x'$.

In estimating the covariance matrix, we estimated M_3 by $\hat{M}_3$, N_3 by $\hat{N}_3 = \frac{1}{n}\sum w_{3i}^2 x_i x_i'$, and the scalar quantity $E\psi^2(\varepsilon)/E^2\dot{\psi}(\varepsilon)$ by $\frac{1}{n}\sum \psi^2(e_i)/[\frac{1}{n}\sum\dot{\psi}(e_i)]^2$, where $e_i = y_i^* - x_i^{*\prime}\hat{\beta}$, multiplying in practice by a finite-sample correction factor:

$$c_1 = \left[1 + \left(\frac{p+2}{n}\right) \frac{1 - \frac{1}{n}\sum\dot{\psi}(e_i)}{\sum\dot{\psi}(e_i)}\right]^2$$

- compare with Carroll and Ruppert (1982), or Huber (1981, page 173).

Another approach to computing standard errors might be to use the bootstrap. While this is a reasonable possibility, at this point it is not clear how best to implement bootstrapping in the presence of outliers and high leverage points. Future work in this area is clearly needed.

The influence function for $\hat{\theta}$ is given by

$$IF(x,y;\theta) = M_2^{-1}h(x) \, \mathrm{Min}\left[1, \frac{a_2}{h'(x)M_2^{-1}h(x)}\right] \frac{\chi\left[\frac{y-x'\beta}{\exp[h'(x)\theta]}\right]}{E(\varepsilon\dot{\chi}(\varepsilon))}$$

and thus $\hat{\theta}$ has asymptotic covariance matrix given by

$$\mathrm{Var}[\sqrt{n}\,(\hat{\theta}-\theta)] = E[IF(x,y;\theta)\; IF'(x,y;\theta)]$$

$$= M_2^{-1} N_2 M_2^{-1} \frac{E\chi^2(\varepsilon)}{E^2[\varepsilon\dot{\chi}(\varepsilon)]} \text{, where } N_2 = E w_2^2(h(x))h(x)h'(x).$$

This may be estimated in a fashion quite analogous to that used to estimate the asymptotic covariance matrix of β. Full details may be found in Giltinan (1983).

REFERENCES

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: John Wiley and Sons.

Box, G. E. P. and Hill, W. J. (1974) "Correcting inhomogeneity of variance with power transformation weighting", *Technometrics, 16,* 385-389.

Carroll, R. J. (1982) "Adapting for heteroscedasticity in linear models", *Annals of Statistics, 10,* 1224-1233.

Carroll, R. J. and Ruppert, D. (1982) "A comparison between maximum likelihood and generalized least squares in a heteroscedastic model", *Journal of the American Statistical Association, 77,* 878-882.

Carroll, R. J. and Ruppert, D. (1982b) "Robust estimation in heteroscedastic linear models", *Annals of Statistics, 10,* 429-441.

Cook, R. D. (1985) A general class of diagnostics. Preprint.

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

Cook, R. D. and Weisberg, S. (1983) "Diagnostics for heteroscedasticity in regression", *Biometrika, 70,* 1-10.

Dent, W. T. and Hildreth, C. (1977) "Maximum likelihood estimation in random coefficient models", *Journal of the American Statistical Association 72,* 69-72.

Finney, D. J. (1976) "Radioligand assay", *Biometrics 32,* 721-740.

Giltinan, D. M. (1983) "Bounded influence estimation in heteroscedastic linear models", Ph.D. thesis, University of North Carolina.

Hampel, F. R. (1968) "Contributions to the theory of robust estimation", unpublished PH.D. thesis, University of California, Berkeley.

Hampel, F. R. (1974) "The influence curve and its role in robust estimation", *Journal of the American Statistical Association, 69,* 383-394.

Hampel, F. R. (1978) "Optimality bounding the gross-error-sensitivity and the influence of position in factor space", *1978 Proceedings of the ASA Statistical Computing Section,* American Statistical Association, Washington, DC, 59-64.

Hildreth, C. and Houck, J. P. (1968). "Some estimators for a linear model with random coefficients,"*Journal of the American Statistical Association, 63,* 584-595.

Huber, P. J. (1981) Robust Statistics, New York: John Wiley and Sons.

Jobson, J. D. and Fuller, W. A. (1980) "Least squares estimation when the covariance matrix and parameter vector are functionally related", *Journal of the American Statistical Association*, *75*, 176-181.

Krasker, W. S. (1980) "Estimation in linear regression models with disparate data points", *Econometrica*, *48*, 1333-1346.

Krasker, W. S. and Welsch, R. E. (1982) "Efficient bounded-influence regression estimation using alternative definitions of sensitivity", *Journal of the American Statistical Association*, *77*, 595-604.

Leurgans, S. (1980) "Evaluating laboratory measurement techniques", in Biostatistics Casebook, John Wiley and Sons. (Ed. R. G. Miller, B. Efron, B. W. Brown, L. E. Moses)

Mallows, C. L. (1975) "On some topics in robustness", unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Maronna, R. A., Bustos, O. H., and Yohai, V. J. (1979), "Bias-and-efficiency-robustness of general M-estimators for regression with random carriers", in Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics, No. 757, Berlin-Heidelberg-New York: Springer-Verlag, 91-116.

Rodbard, D. and Frazier, G. R. (1975) Statistical analysis of radioligand assay data, *Methods of Enzymology*, *37*, 3-22.

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