VARIANCE FUNCTION ESTIMATION

M. Davidian and R. J. Carroll

University of North Carolina at Chape: Hill
Department of Statistics
Phillips Hall 039A
Chapel Hill, NC 27514

Research supported by the Air Force Office of Scientific Research AFOSR-F-49620-85-C-0144.

Key Words and Phrases : Regression, heteroscedasticity, asymptotic efficiency, quality control, variance estimation.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A174961

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Mimeo Series #1700

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 86-2145

6a. NAME OF PERFORMING ORGANIZATION
UNC-CH

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Air Force Office of Scientific Research

6c. ADDRESS *(City, State and ZIP Code)*
Univ. of North Carolina, Statistics Dept.
Phillips Hall, Chapel Hill, NC 27514

7b. ADDRESS *(City, State and ZIP Code)*
same as 8c

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL *(If applicable)*
nm

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
~~AFOSR~~-F-49620-85-C-0144

8c. ADDRESS *(City, State and ZIP Code)*
Bolling Air Force Base
Washington, DC 20332

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2304	A5	

11. TITLE *(Include Security Classification)*
"Variance Function Estimation"

12. PERSONAL AUTHOR(S)
Davidian, Marie and Carroll, Raymond

13a. TYPE OF REPORT
technical

13b. TIME COVERED
FROM 8/85 TO 8/86

14. DATE OF REPORT *(Yr., Mo., Day)*
July 1986

15. PAGE COUNT
31

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Regression, heteroscedasticity, asymptotic efficiency, quality control, variance estimation

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

We develop a general theory for variance function estimation in regression. Most methods in common use are included in our development. The general qualitative conclusions are these. First, most variance function estimation procedures can be looked upon as regressions with "responses" being transformations of absolute residuals from a preliminary fit or sample standard deviations from replicates at a design point. Our conclusion is that the former is typically more efficient, but not uniformly so. Secondly, for variance function estimates based on transformations of absolute residuals, we show that efficiency is a monotone function of the efficiency of the fit from which the residuals are formed, at least for symmetric errors. Our conclusion is that one should iterate so that the residuals are based on generalized least squares. Finally, robustness issues are of even more importance here than in estimation of a regression function for the mean. The loss of efficiency of the standard method away from the normal distribution is much more rapid than in the regression problem.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
~~Lisa Brooks~~ Capt Thomas

22b. TELEPHONE NUMBER *(Include Area Code)*
767-5025
~~(919) 962-2307~~

22c. OFFICE SYMBOL
nm

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

ABSTRACT

This document ~~We~~ develops a general theory for variance function estimation in regression. Most methods in common use are included in our development. The general qualitative conclusions are these. First, most variance function estimation procedures can be looked upon as regressions with "responses" being transformations of absolute residuals from a preliminary fit or sample standard deviations from replicates at a design point. Our conclusion is that the former is typically more efficient, but not uniformly so. Secondly, for variance function estimates based on transformations of absolute residuals, we show that efficiency is a monotone function of the efficiency of the fit from which the residuals are formed, at least for symmetric errors. Our conclusion is that one should iterate so that the residuals are based on generalized least squares. Finally, robustness issues are of even more importance here than in estimation of a regression function for the mean. The loss of efficiency of the standard method away from the normal distribution is much more rapid than in the regression problem. [illegible] Heteroscedasticity

1. INTRODUCTION

Consider a heteroscedastic regression model for observable data Y:

$$(1.1) \qquad EY_i = \mu_i = f(x_i,\beta); \qquad \mathrm{Var}\,(Y_i) = \sigma^2 g^2(z_i,\beta,\theta).$$

Here, $\{x_i(k \times 1)\}$ are the design vectors, $\beta(p \times 1)$ is the regression parameter, f is the mean response function, and the variance function g expresses the heteroscedasticity, where $\{z_i(\ell \times 1)\}$ are known vectors, possibly the $\{x_i\}$, σ is an unknown scale parameter, and $\theta(r \times 1)$ is an unknown parameter. Models which may be regarded as special cases of (1.1) are used in diverse fields, including radioimmunoassay, econometrics, pharmokinetic modeling, enzyme kinetics and chemical kinetics among others. The usual emphasis is on estimation of β with estimation of the variances as an adjunct.

The most common method for estimating β is generalized least squares, in which one estimates $g(z_i,\beta,\theta)$ by using an estimate of θ and a preliminary estimate of β and then performs weighted least squares; see, for example, Carroll and Ruppert (1982a) and Box and Hill (1974). This might be iterated, with the preliminary estimate replaced by the current estimate of β, a new estimate of θ obtained and the process repeated. Standard asymptotic theory as in Carroll and Ruppert (1982a) or Jobson and Fuller (1980) shows that as long as the preliminary estimators for the parameters of the variance function are consistent, all estimators of β obtained in this way will be asymptotically equivalent to the weighted least squares estimator with known weights.

There is evidence that for finite samples, the better one's estimate of θ, the better one's final estimate of β. Williams (1975) states that "both analytic and empirical studies...indicate that...the ordering of efficiency (of estimates of β)...in small samples is in accordance with the ordering by efficiency (of estimates of θ)." Rothenberg (1984) shows via second order calculations that if g does not depend on β, when the data are normally distributed the covariance matrix of the generalized least squares estimator of β is an increasing function of the covariance matrix of the estimator of θ.

Second order asymptotics provide only a weak justification for studying the properties of variance function estimates. Instead, our thesis is that estimation of the structural variance parameter θ is of independent interest. In many engineering applications, an important goal is to estimate the error made in predicting a new observation; this can be obtained from the variance function once a suitable estimate of θ is available. In chemical and biological assay problems, issues of prediction and calibration arise. In such problems, the estimator of θ plays a central role; the statistical properties of prediction intervals and calibration constructs such as the minimal detectable concentration will be highly dependent on how one estimates θ; see Carroll, Davidian and Smith (1986). In off-line quality control, the emphasis is not only on the mean response but also on its variability; Box and Meyer (1986) state that "one distinctive feature of Japanese quality control improvement techniques is the use of statistical experimental design to study the effect of a number of factors on variance as well as the mean.". Effective estimation of variance functions could play a major role in this application. It should be evident from this brief review that far from being only a nuisance parameter, the structural variance parameter θ can be an important part of a statistical analysis.

The above discussion suggests the need for a unified investigation of estimation of variance functions, in particular, estimation of the structural parameter θ. Previous work in the literature tends to treat various special cases of (1.1) as different models with their own estimation methods. The intent of this paper is to study parametric variance function estimation in a unified way. Nonparametric variance function estimation has also been studied, see for example Carroll (1982); we will confine our study to the parametric setting.

Parametric variance function estimation may be thought of as a type of regression problem in which we try to understand variance as a function of known or estimable quantities, and in which θ plays the part of a "regression" parameter. The major insight which allows for a unified study is that the absolute residuals from the current fit to the mean or the sample standard deviations from replicates are basic building blocks for analysis. At the graphical level, this means that transformations of the absolute residuals and sample standard deviations can be used to gain insight into the structure of the variability and to suggest parametric models. For estimation, a major contribution is to point out that most of the methods proposed in the literature are (possibly weighted) regressions of transformations of the basic building blocks on their expected values. Many exceptions to this are dealt with in this article as well.

Our study yields these major qualitative conclusions. As stated here, they apply strictly only to symmetric error distributions, but they are fairly definitive and one is unlikely to be too successful ignoring them in practice. Our first conclusion is that robustness plays a great role in the efficiency of variance function estimation, probably even greater than in estimation of a mean function. For example, if the variance does not depend on the mean response, the standard method will be normal theory maximum likelihood as in Box & Meyer (1986). A weighted analysis of absolute residuals yields an estimator only 12% less efficient at the normal model, and with a large slope of improvement for heavy tailed distributions. This slope of improvement is much larger than is typical in regression on means. For a standard contaminated normal model for which the best robust estimators have efficiency

125% with respect to least squares, the absolute residual estimator of the variance function has efficiency 200%.

Our second conclusion concerns the fit to the means upon which the residuals are based. It has been our experience that unweighted least squares residuals yield unstable estimates of the variance function when the variances depend on the mean. This is confirmed in our study, in the sense that the asymptotic efficiency of the variance function estimators is an increasing function of the variability of the current fit to the means. Thus, we suggest the use of iterative weighted fitting, so that the variance function estimate is based on generalized least squares residuals. As far as we can tell, this part of our paper is one of the first formal justifications for iteration in a generalized least squares context.

It is standard in many applied fields to take m replicates at each design point, where usually $m \leq 4$. Rather than using (transformations of) absolute residuals for estimating variance function parameters, one might use the sample standard deviations. Our third conclusion involves the efficiency of this substitution, for which we develop an asymptotic theory. The effect is typically, although not always, a loss of efficiency, at least when there are $m \leq 4$ replicates. The clearest results occur when the variance does not depend on the mean. Normal theory maximum likelihood is a weighted regression of squared residuals; the corresponding method would be a weighted regression based on sample variances. Using the latter entails a loss of efficiency, no matter what the underlying distribution. For normally distributed data, the efficiency is (m-1)/m, thus being only 50% for duplicates. For other methods, using the replicate standard deviations can be more efficient. This is particularly true of a method due to Harvey (1976), which is based on the logarithm of absolute residuals. A small absolute residual, which seems to always occur in practice, can wreak havoc with this method. This is consistent with our influence function calculations, so that we suggest some trimming of the smallest absolute residuals before applying Harvey's method.

In Section 2 we describe a number of methods for estimation of θ. We confine our attention to methods which are in common use; in particular, we do not discuss robust methods, see Giltinan, Carroll and Ruppert (1986). In Section 3 we present an asymptotic theory for a general estimator of θ whose construction encompasses the methods of Section 2. Section 4 contains examples of specific applications of our theory and a discussion of the implications of our formulation. Sketches of proofs are presented in Appendix A.

2. ESTIMATION OF θ

We now discuss the form and motivation for several estimators of θ in (1.1). In what follows, let $\hat{\beta}_*$ be a preliminary estimator for β. This could be unweighted least squares or the current estimate in an iterative reweighted least squares calculation. Let the errors be given by $\epsilon_i = \{Y_i - f(x_i,\beta)\}/\{\sigma g(z_i,\beta,\theta)\}$ and denote the residuals by $r_i = Y_i - f(x_i,\hat{\beta}_*)$.

2.1 Regression Methods

Pseudo-likelihood. Given $\hat{\beta}_*$, the pseudo-likelihood estimator maximizes the normal log-likelihood $\ell(\hat{\beta}_*,\theta,\sigma)$, where

$$(2.1)\qquad \ell(\beta,\theta,\sigma) = -N \log \sigma - \Sigma_{i=1}^{N}\log\{g(z_i,\beta,\theta)\} - (2\sigma^2)^{-1}\Sigma_{i=1}^{N} \{Y_i-f(x_i,\beta)\}^2/g^2(z_i,\beta,\theta),$$

see Carroll and Ruppert (1982a). Generalizations of pseudo-likelihood for robust estimation have been studied by Carroll and Ruppert (1982a) and Giltinan, Carroll and Ruppert (1986).

Least squares on squared residuals. Besides pseudo-likelihood, other methods using squared residuals have been proposed. The motivation for these methods is that the squared residuals have approximate expectation $\sigma^2 g^2(z_i,\beta,\theta)$, see Jobson and Fuller (1980) and Amemiya (1977). This suggests a nonlinear regression problem in which the "responses" are $\{r_i^2\}$ and the "regression function" is $\sigma^2 g^2(z_i,\hat{\beta}_*,\theta)$. The estimator $\hat{\theta}_{SR}$ minimizes in θ and σ

$$\Sigma_{i=1}^{N} \{r_i^2 - \sigma^2 g^2(z_i,\hat{\beta}_*,\theta)\}^2.$$

For normal data the squared residuals have approximate variance $\sigma^4 g^4(z_i,\beta.\theta)$; in the spirit of generalized least squares, this suggests the weighted estimator which minimizes in θ and σ

$$\Sigma_{i=1}^{N} \{r_i^2 - \sigma^2 g^2(z_i,\hat{\beta}_*,\theta)\}^2/g^4(z_i,\hat{\beta}_*,\hat{\theta}_*), \tag{2.2}$$

where $\hat{\theta}_*$ is a preliminary estimator for θ, $\hat{\theta}_{SR}$ for example. Full iteration, when it converges, would be equivalent to pseudo-likelihood.

Accounting for the effect of leverage. One objection to methods such as pseudo-likelihood and least squares based on squared residuals is that no compensation is made for the loss of degrees of freedom associated with preliminary estimation of β. For example, the effect of applying pseudo-likelihood directly seems to be a bias depending on p/N. For settings such as fractional factorials where p is large relative to N this bias could be substantial.

Bayesian ideas have been used to account for loss of degrees of freedom; see Harville (1977) and Patterson and Thompson (1974). When g does not depend on β, the restricted maximum likelihood approach of the latter authors suggests in our setting one estimate θ from the mode of the marginal posterior density for θ assuming normal data and a prior for the parameters proportional to σ^{-1}. When g depends on β, one may extend the Bayesian arguments and use a linear approximation as in Box and Hill (1974) and Beal and Sheiner (1986) to define a restricted maximum likelihood estimator.

Let Q be the N × p matrix with ith row $f_\beta(x_i,\beta)^t/g(z_i,\beta,\theta)$, where $f_\beta(x_i,\beta) = \partial/\partial\beta \{f(x_i,\beta)\}$, and let $H = Q(Q^tQ)^{-1}Q^t$ be the "hat" matrix with diagonal element $h_{ii} = h_{ii}(\beta,\theta)$; the values $\{h_{ii}\}$ are the leverage values. It turns out that the restricted maximum likelihood estimator is equivalent to an estimator obtained by modifying pseudo-likelihood to account for the effect of leverage. This characterization, while not unexpected, is new; we derive this estimator and its equivalence to a modification of pseudo-likelihood in Appendix B.

The least squares approach using squared residuals can also be modified to show the effect of leverage. Jobson and Fuller (1980) essentially note that for nearly normally distributed data we have the approximations

$$E\ r_i^2 \approx \sigma^2(1-h_{ii})g^2(z_i,\beta,\theta),$$
$$\text{var}\ r_i^2 \approx \sigma^4(1-h_{ii})^2g^4(z_i,\beta,\theta).$$

To exploit these approximations modify (2.2) to minimize in θ and σ

$$(2.3) \qquad \Sigma_{i=1}^N \{r_i^2 - \sigma^2(1-\hat{h}_{ii})g^2(z_i,\hat{\beta}_*,\theta)\}^2/\{(1-\hat{h}_{ii})^2g^4(z_i,\hat{\beta}_*,\hat{\theta}_*)\},$$

where $\hat{h}_{ii} = h_{ii}(\hat{\beta}_*,\hat{\theta}_*)$ and $\hat{\theta}_*$ is a preliminary estimator for θ. An asymptotically equivalent variation of this estimator in which one sets the derivatives of (2.3) with respect to θ and σ equal to 0 and then replaces $\hat{\theta}_*$ by θ can be seen to be equivalent to pseudo-likelihood in which one replaces standardized residuals by studentized residuals. While this estimator also takes into account the effect of leverage, it is different from restricted maximum likelihood.

Least squares on absolute residuals. Squared residuals are skewed and long-tailed, which has lead many authors to propose using absolute residuals to estimate θ; see Glejser (1969) and Theil (1971). Assume that

$$E|Y_i - f(x_i,\beta)| = \eta g(z_i,\beta,\theta),$$

which is satisfied if the errors $\{\epsilon_i\}$ are independent and identically distributed. Mimicking the least squares approach based on squared residuals, one obtains the estimator $\hat{\theta}_{AR}$ by minimizing in η and θ

$$\Sigma_{i=1}^{N} \{|r_i| - \eta g(z_{i,}\hat{\beta}_*,\theta)\}^2.$$

In analogy to (2.2), the weighted version is obtained by mimimizing

$$\Sigma_{i=1}^{N} \{|r_i| - \eta g(z_i,\hat{\beta}_*,\theta)\}^2/g^2(z_i,\hat{\beta}_*,\hat{\theta}_*),$$

where $\hat{\theta}_*$ is a preliminary estimator for θ, probably $\hat{\theta}_{AR}$. As for least squares estimation based on squared residuals, one could presumably modify this approach to account for the effect of leverage.

Logarithm method. The suggestion of Harvey (1976) is to exploit the fact that the logarithm of the absolute residuals has approximate expectation log $\{\sigma g(z_i,\beta,\theta)\}$. Estimate θ by ordinary least squares regression of log $|r_i|$ on log $\{\sigma g(z_i,\hat{\beta}_*,\theta)\}$, since if the errors are independent and identically distributed, the regression should be approximately homoscedastic. If one of the residuals is near zero the regression could be adversely affected by a large "outlier," hence in practice one might wish to delete a few of the smallest absolute residuals, perhaps trimming the smallest few percent.

2.2 Other methods

Besides squares and logarithms of absolute residuals, other transformations could be used. For example, the square root and 2/3 root would typically be more normally distributed than the absolute residuals themselves. Such transformations appear to be useful, although they have not been used much to our knowledge. Our asymptotic theory applies to such transformations.

In a parametric model such as (1.1), joint maximum likelihood estimation is possible, where we use the term maximum likelihood to mean normal theory maximum likelihood. When the variance function does not depend on β, it can be easily shown that maximum likelihood is asymptotically equivalent to weighted least squares methods based on squared residuals. In the situation in which the variance function depends on β this is not the case. In this setting, it has been observed by Carroll and Ruppert (1982b) and McCullagh (1983) that while maximum likelihood estimators enjoy asymptotic optimality when the model and distributional assumptions are correct, the maximum likelihood estimator of β can suffer problems under departures from these assumptions. This suggests that joint maximum likelihood estimation should not be applied blindly.

Methods requiring $m_i \geq 2$ replicates at each x_i have been proposed in the assay literature; for simplicity, we will consider only the case of equi-replication $m_i \equiv m$ and write in obvious fashion $\{Y_{ij}\}$, $j = 1, \ldots m$, to denote the m observations at x_i where appropriate. These methods do not depend on the postulated form of the regression function; one reason that this may be advantageous is that in many assays along with observed pairs (Y_{ij}, x_i) there will also be pairs in which only Y_{ij} is observed. A popular and widely used method is that of Rodbard and Frazier (1975). If we assume

$$g(z_i, \beta, \theta) = g(\mu_i, z_i, \theta), \tag{2.4}$$

the method is identical to the logarithm method previously discussed except that one replaces $|r_i|$ by the sample standard deviation s_i and $f(x_i,\hat{\beta}_*)$ in the "regression" function by the sample mean $\overline{Y}_{i\cdot}$. As an alternative, under the assumption of independence and (2.4), the modified maximum likelihood method of Raab (1981) estimates θ by joint maximization in the (N+r+1) parameters $\sigma^2,\theta,\mu_1,\ldots,\mu_N$ of the "modified" normal likelihood

$$(2.5)\quad \Pi_{i=1}^{N}\{2\pi\sigma^2 g^2(\mu_i,z_i,\theta)\}^{(m-1)/2}\exp[-\Sigma_{j=1}^{m}(Y_{ij}-\mu_i)^2/\{2\sigma^2 g^2(\mu_i,z_i,\theta)\}]$$

3. AN ASYMPTOTIC THEORY OF VARIANCE FUNCTION ESTIMATION

In this section we construct an asymptotic theory for a general class of regression-type estimators for θ. Since our major interest lies in obtaining general insights, we do not state technical assumptions or details.

3.1 Methods based on transformations of absolute residuals

Write $d_i(\beta) = |Y_i - f(x_i,\beta)|$. Let H_1 be a smooth function and define $H_{2,i}$ by

$$H_{2,i} = H_{2,i}(\eta,\theta,\beta) = E\ [\ H_1\{d_i(\beta)\}\ ],$$

where η is a scale parameter which is usually a function of σ only. If $\hat{\eta}_*$, $\hat{\theta}_*$ and $\hat{\beta}_*$ are any preliminary estimators for η, θ, and β, define $\hat{\eta}$ and $\hat{\theta}$ to be the solutions of

$$(3.1)\quad N^{-1/2}\,\Sigma_{i=1}^{N}\ H_{4,i}(\eta,\theta,\hat{\beta}_*)\ \{H_1\{d_i(\hat{\beta}_*)\} - H_{2,i}(\eta,\theta,\hat{\beta}_*)\}\ /\ H_{3,i}(\hat{\eta},\hat{\theta}_*,\hat{\beta}_*),$$

where $H_{3,i}(\eta,\theta,\beta)$ is a smooth function and $H_{4,i}$ is usually the partial derivative of $H_{2,i}$ with respect to (η,θ).

The class of estimators solving (3.1) includes directly or includes an asymptotically equivalent version of the estimators of Section 2.1. For methods which account for the effect of leverage, $H_{2,i}$, $H_{3,i}$ and $H_{4,i}$ will depend on the h_{ii}. In this case we need the additional assumption that if $h = \max\{h_{ii}\}$, then $N^{1/2}h$ converges to zero.

Theorem 3.1. Let $\hat{\eta}_*$, $\hat{\theta}_*$ and $\hat{\beta}_*$ be $N^{1/2}$ consistent for estimating η, θ and β. Let $\dot{H}_1$ be the derivative of H_1 and define

$$C_i = H_{4,i}\,[H_1\{d_i(\beta)\} - H_{2,i}] \,/\, H_{3,i};$$
$$B_{1,N} = N^{-1}\Sigma_{i=1}^{N}\, H_{4,i}H_{4,i}^{t}/H_{3,i};$$
$$B_{2,N} = -N^{-1}\Sigma_{i=1}^{N}\,(H_{4,i}/H_{3,i})\;\partial/\partial\beta\;\{H_{2,i}(\eta,\theta,\beta)\};$$
$$B_{3,N} = -N^{-1}\Sigma_{i=1}^{N}\,(H_{4,i}/H_{3,i})f_{\beta}(x_i,\beta)\;E\;[\dot{H}_1\{d_i(\beta)\}\text{sign}(\epsilon_i)].$$

Then, under regularity conditions as N → ∞,

$$\text{(3.2)}\quad B_{1,N}\,N^{1/2}\begin{bmatrix}\hat{\eta}-\eta\\ \hat{\theta}-\theta\end{bmatrix} = N^{-1/2}\Sigma_{i=1}^{N}C_i + (B_{2,N}+B_{3,N})\,N^{1/2}(\hat{\beta}-\beta) + o_p(1).$$

We may immediately make some general observations about the estimator $\hat{\theta}$ solving (3.1). Note that if the variance function does not depend on β, then $H_{2,i}$ does not depend on β and hence $B_{2,N} \equiv 0$. For the estimators of Section 2.1, $\dot{H}_1$ is an odd function. Thus, if the errors $\{\epsilon_i\}$ are symmetrically distributed, $E[\,\dot{H}_i\{d_i(\beta)\}\text{sign}(\epsilon_i)\,] = 0$ and hence $B_{3,N} \equiv 0$.

Corollary 3.1(a). Suppose that the variance function does not depend on β and the errors are symmetrically distributed. Then the asymptotic distributions of the regression estimators of Section 2.1 do not depend on the method used to obtain $\hat{\beta}_*$. If both of these conditions do not hold simultaneously, then the asymptotic distributions will depend in general on the method of estimating β.

□

The implication is that in the situation for which the variance function does not depend on β and the data are approximately symmetrically distributed, for large sample sizes the preliminary estimator for β will play little role in determining the properties of $\hat{\theta}$. Note also from (3.2) that for weighted methods, the effect of the preliminary estimator of θ is asymptotically negligible regardless of the underlying distributions.

The preliminary estimator $\hat{\beta}_*$ might be the unweighted least squares estimator, a generalized least squares estimator or some robust estimator. See, for example, Huber (1981) and Giltinan, Carroll and Ruppert (1986) for examples of robust estimators for β. For some vectors $\{v_{N,i}\}$, these estimators admit an asymptotic expansion of the form

$$N^{1/2}(\hat{\beta}_* - \beta) = N^{-1/2}\Sigma_{i=1}^{N} \Psi(v_{N,i},\epsilon_i) + o_p(1). \tag{3.3}$$

Here Ψ is odd in the argument ϵ. In case the variance function depends on β, $B_{2,N} \neq 0$ in general; however, if the errors are symmetrically distributed and $\hat{\beta}_*$ has expansion of form (3.3), then the two terms on the right-hand side of (3.2) are asymptotically independent. The following is then immediate.

Corollary 3.1(b). Suppose that the errors are symmetrically distributed and that $\hat{\beta}_*$ has an asymptotic expansion of the form (3.3). Then for the estimators of Section 2.1, the asymptotic covariance matrix of $\hat{\theta}$ is a monotone nondecreasing function of the asymptotic covariance matrix of $\hat{\beta}_*$.

□

By the Gauss-Markov theorem and the results of Jobson and Fuller (1980) and Carroll and Ruppert (1982a), the implication of Corollary 3.1(b) is that using unweighted least squares estimates of β will result in inefficient estimates of θ. This phenomenon is exhibited in small samples in a Monte Carlo study of Carroll, Davidian and Smith (1986). If one starts from the unweighted least squares estimate, one ought to iterate the process of estimating θ -- use the current value $\hat{\beta}_*$ to estimate θ from (3.1), use these $\hat{\beta}_*$ and $\hat{\theta}$ to obtain an updated $\hat{\beta}_*$ by generalized least squares and repeat the process $\mathcal{C} - 1$ more times. It is clear that the asymptotic distribution of $\hat{\theta}$ will be the same for $\mathcal{C} \geq 2$ with larger asymptotic covariance for $\mathcal{C} = 1$, so in principle one ought to iterate this process at least twice. See Carroll, Ruppert and Wu (1986) for more on iterating generalized least squares.

3.2 Methods based on sample standard deviations

Assume at each of M design points we have $m \geq 2$ replicate observations so that $N = Mm$ represents the total number of observations. Let $\{s_i\}$ be the sample standard deviations, which themselves have been proposed as estimators of the variance in generalized least squares estimation of β. This can be disasterous, see Jacquez, Mather and Crawford (1968). When replication exists, however, practitioners feel comfortable with the notion that the $\{s_i\}$ may be used as a basis for estimating variances; thus, one might reasonably seek to estimate θ by replacing $d_i(\hat{\beta}_*)$ by s_i in (3.1).

The following result is almost immediate from the proof of Theorem 3.1 in Appendix A. Here we let $N \rightarrow \infty$ such that m remains fixed.

Theorem 3.2. If $d_i(\hat{\beta}_*)$ is replaced by s_i in (3.1), then under the conditions of Theorem 3.1 the resulting estimator for θ satisfies (3.2) with $B_{3,N} \equiv 0$ and the redefinitions

(3.4a) $$C_i = (H_{4,i}/H_{3,i})\{H_1(s_i) - H_{2,i}\};$$

(3.4b) $$H_{2,i} = E\{H_1(s_i)\} = H_{2,i}(\eta,\theta,\beta). \qquad \square$$

If the errors are symmetrically distributed, then from (3.2) and Theorem 3.2, whether one is better off using absolute residuals or sample standard deviations in the methods of Section 2.1 depends only on the differences between the expected values and variances of $H_1\{d_i(\beta)\}$ and $H_1(s_i)$. In Section 4 we exhibit such comparisons explicitly and show that absolute residuals can be preferred to sample standard deviations in situations of practical importance.

3.3 Methods not depending on the regression function

We assume throughout this discussion that the variance function has form (2.4) and that $m \geq 2$ replicates are available at each x_i. From Section 2.1 we see that the "regression function" part of the estimating equations depends on $f(x_i,\hat{\beta}_*)$, so that in the general equation (3.1) $H_{2,i}$, $H_{3,i}$ and $H_{4,i}$ all depend on $f(x_i,\hat{\beta}_*)$. In some settings, one may not postulate a form for the μ_i for estimating θ; the method of Rodbard and Frazier (1975), for example, uses s_i in place of $d_i(\hat{\beta}_*)$ as in Section 3.2 and replaces $f(x_i,\hat{\beta}_*)$ by the sample mean $\overline{Y}_{i.}$. We now consider the effect of replacing predicted values by sample means for the general class (3.1).

The presence of the sample means in the variance function in (3.1) requires more complicated and restrictive assumptions than the usual large sample asymptotics applied heretofore. The method of Rodbard and Frazier and the general method (3.1) with sample means are functional nonlinear errors in variables problems as studied by Wolter and Fuller (1982) and Stefanski and Carroll (1985). Standard asymptotics for these problems correspond to letting σ go to zero at rate $N^{-1/2}$. In Section 3.4 we discuss the practical implications of σ being small; for now, we state the following result.

Theorem 3.3.. Suppose that we replace $f(x_i,\hat{\beta}_*)$ by $\overline{Y}_{i.}$ in $H_{2,i}$, $H_{3,i}$ and $H_{4,i}$ in Theorems 3.1 and 3.2 and adopt the assumptions of those theorems. Further, suppose that as $N \rightarrow \infty$, $\sigma \rightarrow 0$ simultaneously and

(i) $N^{1/2}\sigma \rightarrow \lambda,\ 0 \leq \lambda < \infty$;

(ii) $N^{1/2}\Sigma_{i=1}^{N} C_i$ has a nontrivial asymptotic normal limit distribution;

(iii) The $\{\epsilon_i\}$ are symmetric and i.i.d ;

(iv) $\{|\overline{Y}_{i.} - \mu_i| \,/\, \sigma\}^2$ has uniformly bounded k moments, some $k > 2$.

Then the results of Theorems 3.1 and 3.2 hold with $B_{2,N} = B_{3,N} \equiv 0$.

□

This result shows that under certain restrictive assumptions, one may replace predicted values by sample means under replication; however, it is important to realize that the assumption of small σ is not generally valid and hence the use of sample means may be disadvantageous in situations where these asymptotics do not apply.

The estimator of Raab (1981) discussed in Section 2.2 is also a functional nonlinear errors in variables estimator, complicated by a parameter space with size of order N. Sadler and Smith (1985) have observed that the Raab estimator is often indistinguishable from the same estimator with μ_i replaced by $\overline{Y}_{i.}$ in (2.5); such an estimator is contained in the general class (3.1). Davidian (1986) has shown that under the asymptotics of Theorem 3.3 and additional regularity conditions that the two estimators are asymptotically equivalent in an important special case. We may thus consider the result of Theorem 3.3

relevant to this estimator.

3.4 Small σ asymptotics

In Section 3.3 technical considerations forced us to pursue an asymptotic theory in which σ is small. It turns out that in some situations of practical importance these asymptotics are relevant. In particular, in assay data we have observed values for σ which are quite small relative to the means. Such asymptotics are used in the study of data transformations in regression. It is thus worthwhile to consider the effect of small σ on the results of Sections 3.1 and 3.2 and to comment on some other implications of letting $\sigma \to 0$.

In the situation of Theorem 3.1, if the errors are symmetrically distributed, then for the estimators of Section 2.1, if $\sigma \to 0$ as $N \to \infty$, then there is no effect for estimating the regression parameter β. In the situation of Theorem 3.2, the errors need not even be symmetrically distributed. The major insight provided by these results is that in certain practical situations in which σ is small, the choice of $\hat{\beta}_*$ may not be too important even if the variance function depends on β.

Small σ asymptotics may be used also to provide insight into the behavior of other estimators for θ which do not fit into the general framework of (3.1). Davidian (1986) has shown that for fixed σ the extended quasi-likelihood estimator of θ of Nelder and Pregibon (1986) and McCullagh and Nelder (1983) need not be consistent. If one adopts the asymptotics of the previous section, however, it is easily shown that the extended quasi-likelihood estimator is asymptotically equivalent to regression estimators based on squared residuals.

4. APPLICATIONS AND FURTHER RESULTS

In Section 3 we constructed an asymptotic theory which and stateed some general characteristics of regression-type estimators of θ. In this section we use the theory to exhibit the specific forms for the various estimators of Section 2 and compare and contrast their properties. Throughout, define

$$\nu(i,\beta,\theta) = \log g(z_i,\beta,\theta),$$

and let $\nu_\theta(i,\beta,\theta)$ and $\nu_\beta(i,\beta,\theta)$ be the column vectors of partial derivatives of ν with respect to θ and β. Further, let $\xi(\beta,\theta)$ be the covariance matrix of $v_\theta(i,\beta,\theta)$. For simplicity, assume that the errors $\{\epsilon_i\}$ are independent and identically distributed with kurtosis κ; $\kappa = 0$ for normality.

4.1 Pseudo-likelihood, restricted maximum likelihood and weighted squared residuals.

If when accounting for the effect of leverage we let $h \rightarrow 0$ such that $N^{1/2}h \rightarrow 0$, then these methods are asymptotically equivalent. Writing $\eta = \log \sigma$, we have $H_1(x) = x^2$, $H_{2,i} = \exp(2\eta)\ g^2(z_i,\beta,\theta)$, $H_{3,i} = H^2_{2,i}$ and $E\ [\ \dot{H}_1\{d_i(\beta)\}\ \text{sign}(\epsilon_i)\ ] = 2\ E\ [\ Y_i - f(x_i,\beta)] = 0$ so that $B_{3,N} \equiv 0$ regardless of the underlying distributions. If g does not depend on β, or $\sigma \rightarrow 0$, then as long as $\hat{\beta}_* - \beta = O_p(\sigma N^{-1/2})$, $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

$$(2 + \kappa)\ \{4N\ \xi(\beta,\theta)\}^{-1}. \tag{4.1}$$

As mentioned in Section 3, under the small σ asymptotics of Theorem 3.3, the extended quasi-likelihood estimator of θ is asymptotically equivalent to the estimators here with asymptotic covariance matrix (4.1). It has been shown by Davidian (1986) that these methods are asymptotically equivalent to maximum likelihood for general underlying distributions, so that pseudo-likelihood, weighted squared residuals, restricted maximum likelihood, maximum likelihood and, if $\sigma \rightarrow 0$, extended quasi-likelihood, are all asymptotically equivalent. In addition, all of these estimators have influence functions which are linear in the squared errors, indicating substantial nonrobustness.

We may also observe that these methods are preferable to unweighted regression on squared residuals. Write (4.1) as

$$(4.2) \qquad (1/2 + \kappa/4)\ (VW^{-1}V)^{-1},$$

where W is the N × N diagonal matrix with elements $H_{3,i}$ and V is the N × p matrix with i^{th} row $H^t_{4,i}$. For the unweighted estimator based on squared residuals, calculations similar to those above show that the asymptotic covariance matrix when either g does not depend on β or $\sigma \rightarrow 0$ is given by

$$(4.3) \qquad (1/2 + \kappa/4)\ (V^tV)^{-1}(V^tWV)(V^tV)^{-1}.$$

The comparison between (4.2) and (4.3) is simply that of the Gauss-Markov theorem, so that (4.2) is no larger than (4.3).

4.2 Logarithms of absolute residuals and the effect of inliers

We do not consider deletion of the few smallest absolute residuals. Here $H_1(x) = \log x$ so that $\dot{H}_1(x) = x^{-1}$. Letting $\eta = \log \sigma$ and assuming independent and identically distributed errors we have $H_{2,i} = \eta + \nu(i,\beta,\theta) + E \log |\epsilon|$, $H_{3,i} \equiv 1$, and $H_{4,i} = \tau(i,\beta,\theta)$. Under the assumption of symmetry of the errors, with g not depending on β or $\sigma \rightarrow 0$, tedious algebra shows that $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

(4.4) $$\text{var}\ \{\log |\epsilon|^2\}\ \{4N\ \varsigma(\beta,\theta)\}^{-1}.$$

The influence function for this estimator is linear in the logarithm of the absolute errors. This indicates nonrobustness more for inliers than for ourliers, which at the very least is an unusual phenomenon. If the errors are not symmetric then there will be an additional effect due to estimating β not present for the methods of Section 4.1, even if g does not depend on β.

4.3 Weighted Absolute Residuals

Assume that the errors are independent and identically distributed and let $\exp(\eta) = \sigma E|\epsilon|$. Consider the weighted estimator. We have $H_1(x) = x$, $\dot{H}_1(x) = 1$, $H_{2,i} = \exp(\eta)\ g(z_i,\beta,\theta)$ and $H_{3,i} = H_{2,i}^2$. Thus, if the errors are symmetrically distributed and either g does not depend on β or $\sigma \to 0$, $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

(4.5) $$\{\delta/(1 - \delta)\}\ \{N\ \varsigma(\beta,\theta)\}^{-1},$$

where $\delta = \text{var}\ |\epsilon|$. The influence function for this estimator is linear in the absolu.e errors. By an argument similar to that at the end of Section 4.1, we may conclude that when the effect of $\hat{\beta}_*$ is negligible one should use a weighted estimator and iterate the method.

4.4 Comparison of methods based on residuals

We assume that the errors are symmetric and independent and identically distributed and that either g does not depend on β or σ is small. Using (4.1), (4.4) and (4.5), the asymptotic relative efficiency (ARE) of the three methods depends only on the distribution of the errors. The ARE of the weighted absolute residual method to pleudo-likelihood is the same as the asymptotic relative efficiency of the mean absolute deviation with respect ot the sample variance for a single sample, see Huber (1981, page 3). For normal errors, using absolute residuals results in a 12% loss in efficiency while for standard double exponential errors there is a 25% gain in efficiency for using absolute residuals. For normal errors, the logarithm method represents a 59% loss of efficiency with respect to pseudo-likelihood.

In Table 1 we present ARE's for various contaminated normal distributions. The table shows that while at normality neither the absolute residuals nor the logarithm methods are efficient, a very slight fraction of "bad" observations is enough to offset the superiority of squared residuals in a dramatic fashion. For example, just *two bad observations* in 1000 negate the superiority of squared residuals. If 1% or 5% of the data are "bad," absolute residuals and the logarithm method, respectively, show substantial gains over squared residuals. The implication is that while it is commonly perceived that methods based on squared residuals are to be preferred in general, these methods can be highly non-robust. Our formulation includes this result for maximum likelihood, showing its inadequacy under slight departures from the assumed distributional structure.

4.5 Methods based on sample standard deviations

Assume that $m \geq 2$ replicate observations are available at each design point. In practice, m is usually small , see Raab (1981). We compare using absolute residuals to using sample standard deviations in the estimators of Section 2.1. For simplicity, assume that the errors are independent and identically and symmetrically distributed and that either g does not depend on β or σ is small. If the errors are not symmetric and σ is not small or the variance depends on β, using sample standard deviations presumably will be more efficient than in the discussion below. This issue deserves further attention.

Let s_m^2 be the sample variance of m errors $\{\epsilon_1, \ldots, \epsilon_m\}$. It is easily shown by calculations analagous to those of section 4.1 that replacing absolute residuals by sample standard deviations has the effect of changing the asymptotic covariance matrices (4.1), (4.4) and (4.5) to

(4.6) Pseudo-likelihood : $\{(2 + \kappa) + 2/(m - 1)\} \, \{4N \varsigma(\beta,\theta)\}^{-1}$;

(4.7) Logarithm method : $m \operatorname{var} \{ \log (s_m^2) \} \, \{4N \varsigma(\beta,\theta)\}^{-1}$;

(4.8) Weighted absolute residuals : $\{m \delta_* \,/\, (1 - \delta_*)\} \, \{N \varsigma(\beta,\theta)\}^{-1}$,

where $\delta_* = \operatorname{var}(s_m)$. Table 2 contains the asymptotic relative efficiencies of using sample standard deviations to using transformations of absolute residuals for various values of m when the errors are standard normal. The values in the table for $H_1(x) = x^2$ and x indicate that if the data are approximately normally distributed, using sample standard deviations can entail a loss in efficiency with respect to using residuals if m is small. For substantial replication (m $\geq$ 10), using sample standard deviations produces a slight edge in efficiency with respect to weighted absolute residuals for $H_1 = x$.

The second column of Table 2 shows that for the logarithm method, using sample standard deviations surpasses using residuals in terms of efficiency except when m = 2 and is more than twice as efficient for large m. In its raw form, log $|r_i|$ is very unstable because, at least occasionally, $|r_j| \approx$ producing a wild "outlier" in the regression. The effect of using sample standard deviations is to decrease the possibility of such inliers; the sample standard deviations will be likely more uniform, especially as m increases.

The implication is that the logarithm method should not be based on residuals unless remedial measures are taken. The suggestion to trim a few of the smallest absolute residuals before using this method is clearly supported by the theory; presumably, such trimming would reduce or negate the theoretical superiority of using sample standard deviations.

Table 3 contains the asymptotic relative efficiencies of weighted squared sample standard deviations and logarithms of these to weighted squared residuals under normality of the errors. The first column is the efficiency of Raab's method to pseudo-likelihood, and the second column is the efficiency of the Rodbard and Frazier method to pseudo-likelihood. The results of the table imply that using the Raab and Rodbard and Frazier methods, which are popular in the analysis of radioimmunoassay data, can entail a loss of efficiency when compared to methods based on weighted squared residuals. Davidian (1986) has shown that the Rodbard and Frazier estimator can have a slight edge in efficiency over the weighted squared residuals methods for some highly contaminated normal distributions. Using (4.6), the squared residual methods will be more efficient than Raab's method in the limit. Table 3 also addresses the open question as to whether Raab's method is asymptotically more efficient that the Rodbard and Frazier method for normally distributed data. The answer is a general yes, thus agreeing with the Monte-Carlo evidence available when the variance is a power of the mean.

5. DISCUSSION

In Section 3 we constructed a general theory of regression-type estimation for θ in the heteroscedastic model (1.1). This theory includes as special cases common methods described in Section 2 and allows for the regression to be based on absolute residuals from the current regression fit as well as sample standard deviations in the event of replication at each design point. Under various restrictions such as symmetry or small σ, when the variance function g does not depend on β, we showed in Sections 3 and 4 that we can draw general conclusions about this class of estimators as well as make comparisons among the various methods.

When employing methods based on residuals, one should weight the residuals appropriately and iterate the process. There can be large relative differences among the methods in terms of efficiency. Under symmetry of the errors, squared residuals are preferable for approximately normally distributed data, but this preference is tenuous, these can be highly non-robust under only slight departures from normality; methods based on logarithms or the absolute residuals themselves exhibit relatively more robust behavior. For the small amount of replication found in practice, using sample standard deviations rather than residuals can entail a loss in efficiency if estimation is based on the squares of these quantities or the quantities themselves. For the logarithm method based on residuals, trimming the smallest few absolute residuals is essential, since for normal data using sample standard deviations is almost always more efficient than using residuals, even for a small number of replicates. Popular methods applications such as radioimmunoassay based on sample means and sample standard deviations can be less efficient than methods based on weighted squared residuals.

Efficient variance function estimation in heteroscedastic regression analysis is an important problem in its own right. There are important differences in estimators for variance when it is modeled parametrically.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1972). Handbook of Mathematical Functions. Dover Publications, New York.

Amemiya, T. (1977). A note on a heteroscedastic model. Journal of Econometrics 6, 365-370 and corrigenda 8, 265.

Beal, S. L. and Sheiner, L. B. (1985). Heteroscedastic nonlinear regression with pharmokinetic type data. Preprint.

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

Box, G. E. P. and Meyer, R. D. (1986). Dispersion effects from fractional designs. Technometrics 28, 19-28.

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

Carroll, R. J., Davidian, M. and Smith, W. (1986) Variance functions and the minimum detectable concentration in radioimmunoassay. Preprint.

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

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

Carroll, R. J., and Ruppert, D. (1985). Power transformations when fitting theoretical models to data. Journal of the American Statistical Association 79, 321-328.

Carroll, R. J., Ruppert, D., and Wu, C. F. J. (1986). Variance expansion and the bootstrap in generalized least squares. Preprint.

Davidian, M. (1986). Variance function estimation in heteroscedastic regression models. Unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill.

Giltinan, D. M., Carroll, R. J. and Ruppert, D. (1986). Some new methods for weighted regression when there are possible outliers. Technometrics 28, 000-000.

Glejser, H. (1969). A new test for heteroscedasticity. Journal of the American Statistical Association 64, 316-323.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 79, 302-308.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. Econometrics 44, 461-465.

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

Jacquez, J. A., Mather, F. J. and Crawford, C. R. (1968). Linear regression with non-constant, unknown error variances: sampling experiments with least squares and maximum likelihood estimators. Biometrics 24, 607-626.

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.

McCullagh, P. (1983). Quasi-likelihood functions. Annals of Statistics 11, 59-67.

McCullagh, P. and Nelder, J. A. (1983). Generalized Linear Models. Chapman & Hall, New York.

Nel, D. G. (1980). On matrix differentiation in statistics. South African Statistical Journal 14,87-101.

Nelder, J. A. and Pregibon, D. (1986). An extended quasi-likelihood function. Preprint.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545-554.

Raab, G. M. (1981a). Estimation of a variance function, with application to radioimmunoassay. Applied Statistics 30, 32-40.

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

Rothenberg, T. J. (1984). Approximate normality of generalized least squares estimates. Econometrica 52, 811-825.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. Journal of the American Statistical Association 77, 828-838.

Sadler, W. A. and Smith, M. H. (1985). Estimation of the response-error relationship in immunoassay. Clinical Chemistry 31/11, 1802-1805

Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. Annals of Statistics 13, 1335-1351.

Theil, H. (1971). Principles of Econometrics, New York: John Wiley and Sons.

Williams, J. S. (1975). Lower bounds on convergence rates of weighted least squares to best linear unbiased estimators. In A Survey of Statistical Design and Linear Models, J. N. Srivastava, editor. Amsterdam, North Holland.

Wolter, K. M., and Fuller, W. A. (1982). Estimation of nonlinear errors-in-variables models. Annals of Statistics 10, 539-548.

APPENDIX A. PROOFS OF MAJOR RESULTS

We now present sketches of the proofs of the theorems of Section 3. Our exposition is brief and nonrigorous as our goal is to provide general insights. In what follows, we assume that

$$N^{1/2}\begin{bmatrix}\hat{\eta}-\eta\\ \hat{\theta}-\theta\end{bmatrix} = \mathcal{O}_p(1); \tag{A.1}$$

under sufficient regularity conditions it is possible to prove (A.1). Such a proof would be long, detailed and essentially noninformative; see Carroll and Ruppert (1982a) for a proof of $N^{1/2}$ consistency in a special case.

Sketch of proof of Theorem 3.1: From (3.1), a Taylor series, the fact that E [$H_1\{d_i(\beta)\}$] = $H_{2,i}$ and laws of large numbers, we have

$$0 = N^{-1/2}\Sigma_{i=1}^{N}\ (H_{4,i}/H_{3,i})[H_1\{d_i(\hat{\beta}_*)\} - H_{2,i}(\hat{\eta},\hat{\theta},\hat{\beta}_*)] + o_p(1) \tag{A.2}$$

By the arguments of Ruppert and Carroll (1980) or Carroll and Ruppert (1982a),

$$N^{-1/2}\Sigma_{i=1}^{N}\ (H_{4,i}/H_{3,i})[H_1\{d_i(\hat{\beta}_*)\} - H_1\{d_i(\beta)\}] \tag{A.3}$$

$$= N^{-1/2}\Sigma_{i=1}^{N} (H_{4,i}/H_{3,i})\ \dot{H}_i(d_i(\beta))\{d_i(\hat{\beta}_*) - d_i(\beta)\} + o_p(1)$$

$$= B_{3,N}\ N^{1/2}(\hat{\beta}_* - \beta) + o_p(1).$$

Applying this result to (A.2) along with a Taylor series in $H_{2,i}$ gives

$$0 = N^{-1/2}\Sigma_{i=1}^{N} C_i + (B_{2,N} + B_{3,N})\ N^{1/2}(\hat{\beta}_* - \beta)$$

$$- B_{1,N}\ N^{1/2}\begin{bmatrix} \hat{\eta} - \eta \\ \hat{\theta} - \theta \end{bmatrix} + o_p(1),$$

which is (3.2). □

Theorem 3.2 follows by a similar argument; in this case the representation (A.3) is unnecessary.

<u>Sketch of proof of Theorem 3.3</u>: We consider Theorem 3.2; the proof for Theorem 3.1 is similar. Recall here that (2.4) holds. In the following, all derivatives are with respect to the mean μ_i and the definitions of C_i and $H_{2,i}$ are as in (3.4).

Assumption (iv) implies that $N^{1/2} \max_{1 \le i \le N} |\overline{Y}_{i.} - \mu_i| \xrightarrow{p} 0$ so that a Taylor series in η, θ and $\overline{Y}_{i.}$ gives

$$B_{1,N}\ N^{1/2}\begin{bmatrix} \hat{\eta} - \eta \\ \hat{\theta} - \theta \end{bmatrix} = N^{-1/2}\Sigma_{i=1}^{N} C_i - N^{-1/2}\Sigma_{i=1}^{N}(\dot{H}_{2,i}H_{4,i}/H_{3,i})(\overline{Y}_{i.} - \mu_i)$$
$$+ N^{-1/2}\Sigma_{i=1}^{N} \{(\dot{H}_{4,i}/H_{3,i}) - (\dot{H}_{3,i}/H_{3,i})\}(\overline{Y}_{i.} - \mu_i) + o_p(1).$$

Since $\overline{Y}_{i.} - \mu_i = \sigma\ g(\mu_i, z_i, \theta)\ \overline{\epsilon}_{i.} \approx \lambda N^{-1/2} g(\mu_i, z_i, \theta)\ \overline{\epsilon}_{i.}$, where $\overline{\epsilon}_{i.}$ is the mean of the errors at x_i, we can write the last two terms on the right-hand side of (A.4) as

$$\lambda N^{-1}\Sigma_{i=1}^{N} \bar{\epsilon}_{i.}(q_{i,1} + q_{i,2}C_i)$$

for constants $\{q_{i,j}\}$. By assumption (v), since $\bar{\epsilon}_{i.}$ has mean zero, (A.5) converges in probability to zero if $E(\bar{\epsilon}_{i.}c_i) = 0$, which holds under the assumption of symmetry. Thus, (A.5) converges to zero which from (A.4) completes the proof. Note that if we drop the assumption of symmetry, from (A.5) the asymptotic normal distribution of $N^{1/2}(\hat{\theta} - \theta)$ will have mean $\text{p-lim}_{N\to\infty} \{ \lambda B_{1,N}^{-1} N^{-1}\Sigma_{i=1}^{N} (\bar{\epsilon}_{i.}C_i q_{i,2})\}$. □

APPENDIX B. CHARACTERIZATION OF RESTRICTED MAXIMUM LIKELIHOOD

Let $\hat{\beta}_*$ be a generalized least squares estimator for β. Assume first that g does not depend on β. Let the prior distribution for the parameters $\pi(\beta,\theta,\sigma)$ be proportional to σ^{-1}. The marginal posterior for θ is hard to compute in closed form for nonlinear regression. Following Box and Hill (1974) and Beal and Sheiner (1986), we have the linear approximation

$$f(x_i,\beta) \approx f(x_i,\hat{\beta}_*) + f_\beta(x_i,\hat{\beta}_*)^t(\beta-\hat{\beta}_*).$$

Replacing $f(x_i,\beta)$ by its linear expansion, the marginal posterior for θ is proportional to

(B.1) $$p(\theta) = \frac{\{\Pi_{i=1}^{N} g_i^2(\theta)\}^{-1/2}}{\hat{\sigma}_G^{(N-p)}(\theta)\ \{\text{Det } S_G(\theta)\}^{1/2}},$$ where

$$\hat{\sigma}_G^2(\theta) = (N-p)^{-1}\Sigma_{i=1}^N r_i^2 \,/\, g^2(z_i,\hat{\beta}_*,\theta),$$

$$S_G(\theta) = N^{-1}\Sigma_{i=1}^N f_\beta(x_i,\hat{\beta}_*) f_\beta(x_i,\hat{\beta}_*)^t \,/\, g^2(z_i,\hat{\beta}_*,\theta),$$

and where Det A = determinant of A. If the variances depend on β, we extend the Bayesian arguments by replacing $g_i(\theta)$ by $g(z_i,\hat{\beta}_*,\theta)$.

Let H be the hat matrix H evaluated at $\hat{\beta}_*$ and let $h_{ii} = h_{ii}(\hat{\beta}_*,\theta)$. From (2.1), pseudo-likelihood solves in (θ,σ)

$$\text{(B.2)} \qquad \Sigma_{i=1}^N \; [r_i^2/\{\sigma^2 g^2(z_i,\hat{\beta}_*,\theta))\}] \begin{bmatrix} 1 \\ \nu_\theta(z_i,\hat{\beta}_*,\theta) \end{bmatrix} = \Sigma_{i=1}^N \begin{bmatrix} 1 \\ \nu_\theta(z_i,\hat{\beta}_*,\theta) \end{bmatrix} .$$

Since H is idempotent, the left hand side of (B.2) has approximate expectation

$$\text{(B.3)} \qquad \Sigma_{i=1}^N \begin{bmatrix} 1 - p/N \\ \nu_\theta(z_i,\hat{\beta}_*,\theta)\;(1-h_{ii}) \end{bmatrix}$$

To modify pseudo-likelihood to account for loss of degrees of freedom, equate the left hand side of (B.2) to (B.3). From matrix computations as in Nel (1980), this can be shown to be equivalent to restricted maximum likelihood.

Table 1

Asymptotic relative efficiency with respect to weighted squared residuals for contaminated normal distributions with distribution function $F(x) = (1 - \alpha)\Phi(x) + \alpha\Phi(x/3)$.

contamination fraction α	weighted absolute residuals	logarithms of absolute residuals
0.000	0.876	0.405
0.001	0.948	0.440
0.002	1.016	0.480
0.010	1.439	0.720
0.050	2.035	1.220

Table 2

Asymptotic relative efficiency of using sample standard deviations to using absolute residuals under normality for $H_1(x)$ (weighted methods).

	$H_1(x)$		
m	x^2	log x	x
2	0.500	0.500	0.500
3	0.667	1.000	0.696
4	0.750	1.320	0.801
⋮	⋮	⋮	⋮
9	0.889	1.932	0.986
10	0.900	1.984	1.001
∞	1.000	2.467	1.142

Table 3

Asymptotic relative efficiency of using sample standard deviations to weighted squared residuals under normal errors for $H_1(x)$.

	$H_1(x)$	
m	x^2	log x
2	0.500	0.203
3	0.667	0.405
4	0.750	0.535
5	0.800	0.620
6	0.833	0.680
7	0.857	0.723
8	0.875	0.757
9	0.889	0.783
10	0.900	0.804