J. R. Statist. Soc. B (1993)
55, *No.* 2, *pp.* 493-508

Modelling Variance Heterogeneity: Residual Maximum Likelihood and Diagnostics

By A. P. VERBYLA†

University of Adelaide, Australia

[Received March 1991. Final revision January 1992]

SUMMARY

The assumption of equal variance in the normal regression model is not always appropriate. To attempt to eliminate unequal variance a transformation is often used but if the transformation is not successful, or the variances are of intrinsic interest, it may be necessary to model the variances in some way. We consider the normal regression model when log-linear dependence of the variances on explanatory variables is suspected. Detection of the dependence, estimation and tests of homogeneity based on full and residual maximum likelihood are discussed as are regression diagnostic methods based on case deletion and log-likelihood displacement. Whereas the behaviour of full and residual maximum likelihood is similar under case deletion, changes in residual maximum likelihood estimates and log-likelihood displacements tend to be smaller than maximum likelihood.

Keywords: DIAGNOSTICS; LOG-LINEAR MODEL; MAXIMUM LIKELIHOOD; REGRESSION; RESIDUAL MAXIMUM LIKELIHOOD; SCORE TEST; VARIANCE HETEROGENEITY; VARIANCE MODEL

1. INTRODUCTION

In many situations, e.g. in economics, the assumption of variance homogeneity in the linear model is untenable and corrective action is required to ensure an efficient analysis. A standard approach is to transform the response variable (Box and Cox, 1964), which in addition addresses the questions of additivity and normality. However, it may be the case that modelling the variances is itself of interest (Box and Meyer, 1986) or a simple transformation is inadequate to correct for the variance heterogeneity.

The modelling of variances has been discussed extensively in econometrics publications. Park (1966) proposed a two-stage estimation process for a simple log-linear model for the variances while Harvey (1976) discusses maximum likelihood (ML) estimation and the likelihood ratio test for the more general model discussed here. Aitkin (1987) provides GLIM (Baker and Nelder, 1978) macros for ML estimation. Smyth (1989) extends some of the ideas to a wider class of error distributions and also to quasi-likelihood. A non-linear model for the mean in conjunction with variance modelling has been discussed by Stirling (1985). Carroll and Ruppert (1988) present an account of the theory for non-linear models, robust estimation and diagnostic methods for variance parameters using local influence (Cook, 1986) and the influence function.

The detection and tests of variance heterogeneity have been considered by many researchers. Cook and Weisberg (1983) consider graphical methods together with the

†*Address for correspondence*: Department of Statistics, University of Adelaide, GPO Box 498, Adelaide 5000, Australia.

 0035-9246/93/55493 $2.00

score test of variance homogeneity, a test previously considered by Breusch and Pagan (1979). Other references in this area can be found in Cook and Weisberg (1983) and Evans and King (1988).

In Section 2, we introduce the model to be assumed throughout the paper, a linear model for the mean and a log-linear model for the variances. Using a derived response, the added variable plot (Atkinson (1985), section 5.2) is shown to be useful for detecting heterogeneity, the qualification being that a log-linear form for the variances is appropriate. In Section 3 both ML and residual maximum likelihood (REML) estimation (Patterson and Thompson, 1971) are discussed. With a variance model, these researchers (see also Cooper and Thompson (1977)) recommend the use of REML because REML takes into account the loss of degrees of freedom in estimating the mean. The main aim of the present paper is to examine properties of REML as opposed to ML. Tests of homogeneity are discussed in Section 4.

For the normal regression model, diagnostic methods are now well established (Belsley *et al.*, 1980; Cook and Weisberg, 1982; Atkinson, 1985). We consider diagnostic methods based on case weights in Section 5 and again highlight the differences between ML and REML. Discussion concludes the paper in Section 6.

The MINITAB cherry tree data (Ryan *et al.*, 1976) are used to illustrate the results. The response is the volume of usable wood in 31 cherry trees, the explanatory variables being the height and diameter of the trees. Many models relating volume to height and diameter can be proposed and these are discussed by Cook and Weisberg (1983), Atkinson (1985) and Aitkin (1987). We consider the cube root of volume model used by Aitkin (1987). Although we consider only the cherry tree data, the results obtained are similar for other data sets and details are available from the author.

2. MODEL AND DETECTION OF HETEROSCEDASTICITY

Suppose that (Y_i, x_i, z_i) is the observation on the ith unit, $i = 1, 2, \ldots, n$, where Y_i is the response variable and x_i and z_i are $p \times 1$ and $q \times 1$ vectors of explanatory variables. Although we write x_i and z_i separately, they may and often do have common components. We consider the linear model

$$Y_i = x_i'\beta + e_i \tag{1}$$

where $e_i \sim N(0, \sigma_i^2)$, the e_i are assumed independent and β is a vector of unknown parameters. The variance model considered in this paper is the log-linear form

$$\log \sigma_i^2 = z_i'\lambda \quad \text{or} \quad \sigma_i^2 = \exp(z_i'\lambda), \tag{2}$$

where λ is a vector of unknown parameters. The first component of each z_i satisfies $z_{i1} = 1$, so that if $\lambda_2 = \ldots = \lambda_q = 0$ we have constant variance $\sigma^2 = \exp \lambda_1$. We do not discuss the case when the variance is a function of the mean in this paper.

Let Y be the $n \times 1$ vector of responses. The log-likelihood under models (1) and (2) is

$$\log L(\beta, \lambda; Y) = -\frac{1}{2}\left\{ \sum_{i=1}^{n} \log \sigma_i^2 + \sum_{i=1}^{n} \frac{(Y_i - x_i'\beta)^2}{\sigma_i^2} \right\}$$

$$= -\frac{1}{2}\left(\sum_{i=1}^{n} z_i'\lambda + \sum_{i=1}^{n} \frac{d_i}{\sigma_i^2} \right) \tag{3}$$

where $d_i = (Y_i - x_i'\beta)^2$.

We turn to the problem of determining which explanatory variables affect the variance. As β is unknown, we use the ordinary least squares estimator $\hat{\beta} = (X'X)^{-1}X'Y$ and define $d_i = (Y_i - x_i'\hat{\beta})^2$. Then $d_i \sim v_i^2\chi^2(1)$; Cook and Weisberg (1983), p. 6, show that

$$v_i^2 = (1-h_{ii})^2\sigma_i^2 + \sum_{j\neq i} h_{ij}^2\sigma_j^2$$

where h_{ij} are elements of the hat matrix $X(X'X)^{-1}X'$. Cook and Weisberg formulate a general model for the variances, $\sigma_i^2 = \sigma^2 w(z_i, \lambda)$, and consider a Taylor series expansion of $w(z_i, \lambda)$ about $\lambda = \lambda^*$ where λ^* is such that $w(z_i, \lambda^*) = 1$, for all i. If $w'(z_i, \lambda^*)$ denotes the derivative of $w(z_i, \lambda)$ with respect to λ evaluated at λ^*, Cook and Weisberg (1983) show that approximately

$$E\left\{\frac{d_i}{\sigma^2(1-h_{ii})}\right\} = 1 + (\lambda-\lambda^*)\left\{(1-h_{ii})\,w'(z_i,\lambda^*) + \sum_{j\neq i}\frac{h_{ij}^2}{1-h_{ii}}w'(z_j,\lambda^*)\right\}. \tag{4}$$

If s^2 is the usual unbiased estimate of σ^2 for the ordinary regression model, on the basis of approximation (4), Cook and Weisberg (1983) recommend that we plot the squared Studentized residuals $d_i/s^2(1-h_{ii})$ against $(1-h_{ii})\,w'(z_i,\lambda^*)$ to detect heterogeneity, the terms in h_{ij}^2 often being small. By the properties of the χ^2-distribution, as the mean of the Studentized residuals depends on explanatory variables so does the variance and the plots need to be interpreted carefully. In particular polynomial dependence can be missed; for the tree data Cook and Weisberg (1983) mention the linear effect in height on the log-variance but make no statement regarding the dependence on diameter. The Cook and Weisberg plots for both height and diameter are given in Figs 1(a) and 1(b). These plots are discussed later.

The main difficulty with the Cook and Weisberg plots is the non-constant variance of the Studentized residuals. If the variance model is log-linear as in equation (2), we can consider an approximate variance stabilizing transformation as follows. From Harvey (1976) or Smyth (1989) we have the results

$$\begin{aligned} E\{\log d_i + \log\tfrac{1}{2} - \psi(\tfrac{1}{2})\} &= \log v_i^2, \\ \operatorname{var}\{\log d_i + \log\tfrac{1}{2} - \psi(\tfrac{1}{2})\} &= \pi^2/2 \end{aligned} \tag{5}$$

where $\psi(\ )$ is the digamma function. Noting that $\log\frac{1}{2} - \psi(\frac{1}{2}) = 1.27036$ and following the development of Cook and Weisberg (1983), where the terms in h_{ij}^2 are assumed small, we find that $v_i^2 \approx (1-h_{ii})^2\sigma_i^2$ and hence under model (2)

$$E\left[\log\left\{\frac{d_i}{(1-h_{ii})^2}\right\} + 1.27036\right] = z_i'\lambda. \tag{6}$$

Thus, under equations (5) and (6), $\log\{d_i/(1-h_{ii})^2\} + 1.27036$ has a mean which is linear in explanatory variables and has a constant known variance. Ignoring higher order moments, we can use standard results for the normal linear model, at least in an explanatory manner.

Harvey (1976) and Smyth (1989) indicate that efficient and consistent estimators of β and λ can be obtained after two steps of an iterative procedure, with starting values constructed as above. Our aim, however, is purely exploratory.

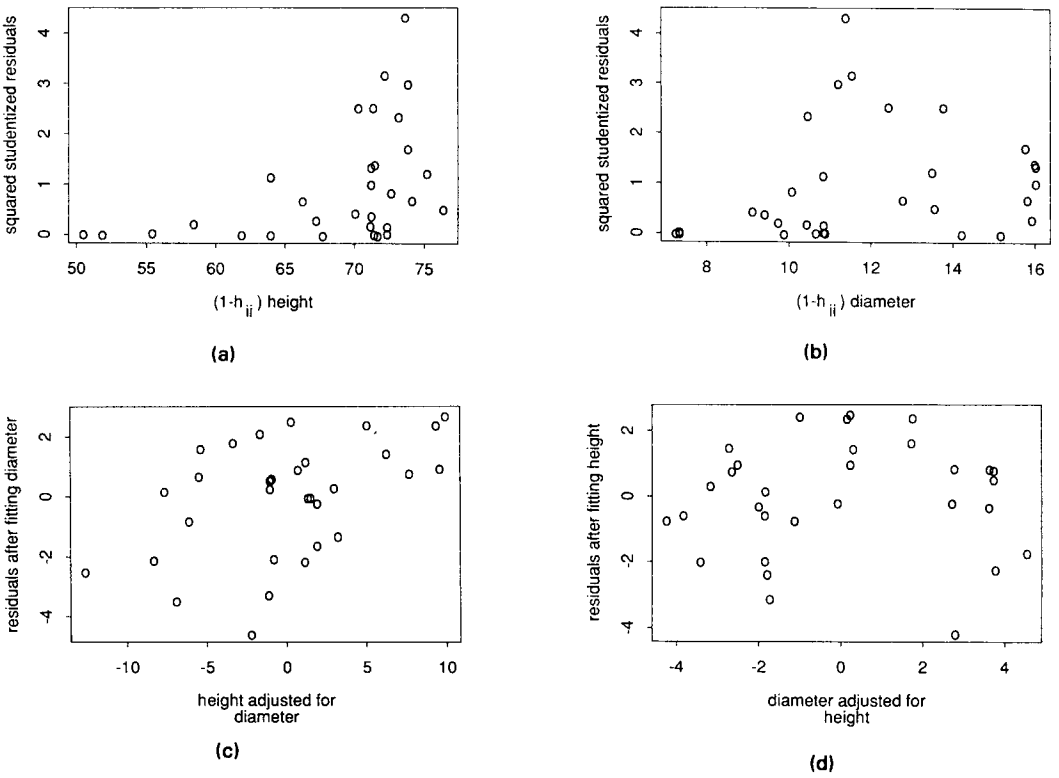

Fig. 1. Detection of heterogeneity—cherry tree data: (a) Cook and Weisberg plot for diameter; (b) Cook and Weisberg plot for height; (c) added variable plot for height; (d) added variable plot for diameter

One practical difficulty arises if $d_i = 0$. This case is fitted perfectly by the mean model and does not contribute to estimation of variance effects. Such points can be omitted in the procedure.

The approach is illustrated for the tree data. We fit a mean model, with response the cube root of volume and explanatory variables height and diameter (and including an intercept), by ordinary least squares and form the squared residuals, our d_i. The derived response $\log\{d_i/(1-h_{ii})^2\} + 1.27036$ is calculated and the added variable plot of Atkinson (1985), section 5.2, is used to determine whether each variable suspected of affecting the variance does have an effect over and above the other explanatory variables. The added variable plots for the cherry tree data are given in Figs 1(c) and 1(d). The residuals for the derived response after fitting the constant and appropriate explanatory variable appear to depend weakly on height (when adjusted for diameter) in a linear fashion and more strongly on diameter (when adjusted for height) in a quadratic manner. A tentative initial variance model on the log-scale is therefore linear in height and quadratic in diameter. These results are also apparent for the Cook and Weisberg plots in Figs 1(a) and 1(b). We see a wedge-shaped plot for height, suggesting a linear effect, and a triangular plot for diameter, suggesting a possible quadratic effect. A referee has suggested orthogonalizing height and diameter and then plotting; this seems to produce similar results.

The added variable plot is a standard graphical procedure usually used for determining the importance of explanatory variables on the mean. Provided that the log-linear form is appropriate, we can use the added variable plot to determine the importance of the explanatory variables on the variance. To the author this approach is a useful alternative to the approach of Cook and Weisberg (1983), which allows each explanatory variable to be examined by eliminating the other explanatory variables. Of course the Cook and Weisberg (1983) procedure extends beyond the log-linear form.

3. ESTIMATION

3.1. *Maximum Likelihood Estimation*

ML estimation is discussed by Harvey (1976) and Aitkin (1987); an outline is presented here. Let d be the vector whose ith element is d_i, X and Z be $n \times p$ and $n \times q$ matrices of the explanatory variables and Σ be the diagonal variance matrix with ith diagonal element σ_i^2. If $\mathbf{1}_n$ denotes the $n \times 1$ vector of unit elements, using equation (3), the score vector and the Fisher expected information are

$$U(\beta, \lambda; Y) = \begin{pmatrix} X'\Sigma^{-1}(Y - X\beta) \\ \frac{1}{2}Z'(\Sigma^{-1}d - \mathbf{1}_n) \end{pmatrix} \quad \text{and} \quad I(\beta, \lambda) = \begin{pmatrix} X'\Sigma^{-1}X & 0 \\ 0 & \frac{1}{2}Z'Z \end{pmatrix}. \tag{7}$$

The method of scoring yields

$$\hat{\beta}_{(m+1)} = (X'\Sigma_m^{-1}X)^{-1}X'\Sigma_m^{-1}Y, \tag{8}$$

$$\begin{aligned} \hat{\lambda}_{(m+1)} &= \lambda_m + (Z'Z)^{-1}Z'(\Sigma_m^{-1}d - \mathbf{1}_n) \\ &= (Z'Z)^{-1}Z'(\Sigma_m^{-1}d - \mathbf{1}_n + Z\lambda_m) \end{aligned} \tag{9}$$

where m indicates the mth iterate. Equations (8) and (9) are used iteratively to estimate β and λ. As Aitkin (1987) points out (see also Smyth (1989)) the estimation procedure can be viewed as a two-stage process involving two models, the mean and variance models. For given λ, equation (3) defines the mean model, a weighted normal regression with weights $1/\sigma_i^2$, whereas, for given β, equation (3) is viewed as a variance model with d_i the response. This is a generalized linear model with gamma errors and known scale parameter 2. Thus GLIM (or any program with a generalized linear modelling facility) can be used to carry out the estimation. I have used New S (Becker *et al.*, 1988) in all computations and the functions are available on request.

3.2. *Residual Maximum Likelihood*

The REML estimate of λ is found by using the marginal likelihood (Patterson and Thompson, 1971; Harville, 1974; Cooper and Thompson, 1977; Verbyla, 1990)

$$\begin{aligned} \log L_R(\lambda; Y) &= -\frac{1}{2}[\log(\det \Sigma) + \log\{\det(X'\Sigma^{-1}X)\} + Y'PY] \\ &= -\frac{1}{2}\left[\sum_{i=1}^{n} z_i'\lambda + \log\{\det(X'\Sigma^{-1}X)\} + Y'PY\right] \end{aligned} \tag{10}$$

where $P = \Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}$. The score vector for λ is given by

$$U_R(\lambda; Y) = \tfrac{1}{2}Z'(\Sigma^{-1}d - \mathbf{1}_n + h) \tag{11}$$

where $h = (h_{11}\, h_{22} \ldots h_{nn})'$ is the vector of diagonal elements of

$$H = \Sigma^{-1/2}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1/2}, \tag{12}$$

the hat matrix in a weighted regression. The expected information matrix is

$$I_R(\lambda) = \tfrac{1}{2}Z'VZ, \tag{13}$$

where $V = \text{var}(\Sigma^{-1}d)/2$ has diagonal elements $(1 - h_{ii})^2$ and off-diagonal elements h_{ij}^2. The method of scoring yields

$$\bar{\lambda}_{(m+1)} = \bar{\lambda}_m + (Z'VZ)^{-1}Z'\{\Sigma_m^{-1}d - \mathbf{1}_n + h\}, \tag{14}$$

where we use a bar to distinguish the REML estimate from the ML estimate and where the subscript is again the iteration count; d as well as all terms involving H will also change at each iteration but the subscript is omitted to avoid confusion with individual elements. A by-product of equation (14) is the REML estimate $\bar{\beta}$ of β which has the same form as equation (8), as we can write

$$Y'PY = (Y - X\bar{\beta})'\Sigma^{-1}(Y - X\bar{\beta})$$

in equation (10).

The off-diagonal elements of V are the terms h_{ij}^2 and following the discussion in Section 2 and Cook and Weisberg (1983) many of these terms may be small. If this is true, V is approximately a diagonal matrix with elements $(1 - h_{ii})^2$. The final estimates will be REML estimates as the score has not changed with this approximation. There are advantages in using the diagonal approximation of the weight matrix in computations and also for diagnostic methods to be discussed in Section 5.

A comparison of ML and REML can now be made. The difference between the score functions given by equations (7) and (11) is the vector h. Substituting in d_i for β by $\bar{\beta}$ which is given by equation (8) as a function of σ_i^2, we have

$$E\left(\frac{d_i}{\sigma_i^2}\right) = E\left\{\frac{(Y_i - x_i'\bar{\beta})^2}{\sigma_i^2}\right\} = 1 - h_{ii}$$

and hence $E(\Sigma^{-1}d) = (\mathbf{1}_n - h)$, so that

$$E\{U_R(\lambda; Y)\} = 0,$$

whereas the expected value of the ML score function for λ is 0 at β but not at $\bar{\beta}$. In this sense the REML estimator is less biased than the ML estimator as the score is an unbiased estimating equation at $\bar{\beta}$. Furthermore under REML the variances and covariances between the squared residuals d_i are taken into account through V in equation (13); this is ignored under ML.

3.3. *MINITAB Tree Data*

For the cherry tree data, Table 1 provides the full maximized log-likelihoods for various variance models, the mean being linear in height and diameter. The parameter estimates of the final variance model, quadratic in diameter, are given in Table 2. The added variable plots of Section 2 are confirmed by the tests and the weak dependence of log-variance on height is not significant. The parameter estimates under ML and

TABLE 1
2 log L for variance models

Model†	*ML*	*REML*
1	126.60	126.44
H	131.71	131.48
D	127.49	127.33
D, H	131.77	131.54
D, D^2	142.46	140.35
D, H, D^2	144.60	143.19
D, H, D^2, H^2	145.33	143.99
D, H, D^2, HD, H^2	147.13	146.15

†H, height; D, diameter.

TABLE 2
Modelling heterogeneity—tree data, cube root of volume

Method	$-2 \log L$	*Results for variance model*			*Results for mean model*		
		1	*D*	D^2	*1*	*D*	*H*
ML	−142.46	−41.15	5.1357	−0.1755	0.0942	0.1527	0.0117
Standard error		4.76	0.6986	0.0247	0.0537	0.0017	0.0010
REML	−140.35	−29.43	3.4194	−0.1151	0.0303	0.1513	0.0128
Standard error		7.16	1.0584	0.0378	0.0889	0.0031	0.0016
Approximate REML	−140.35	−29.43	3.4194	−0.1151	0.0303	0.1513	0.0128
Standard error		7.95	1.1654	0.0414	0.0889	0.0031	0.0016

REML differ, but as expected it is the standard errors where the pronounced differences appear. Standard errors found by using the approximate weight matrix are overestimated and hence are more conservative.

4. TESTS OF HOMOGENEITY

There are many tests available for homogeneity. Here we focus on the score test originally proposed by Breusch and Pagan (1979) and Cook and Weisberg (1983). The statistic is given by

$$S = \frac{1}{2}\left(\frac{d}{\hat{\sigma}_0^2} - \mathbf{1}_n\right)' Z(Z'Z)^{-1}Z'\left(\frac{d}{\hat{\sigma}_0^2} - \mathbf{1}_n\right), \tag{15}$$

where $\hat{\sigma}_0^2 = (1/n)(Y - X\hat{\beta}_0)'(Y - X\hat{\beta}_0)$ and $\hat{\beta}_0 = (X'X)^{-1}X'Y$. S is half the regression sum of squares in the regression of $d/\hat{\sigma}_0^2$ on Z.

Under REML equation (15) becomes

$$S_R = \frac{1}{2}\left(\frac{d}{\bar{\sigma}_0^2} - \mathbf{1}_n + h\right)' Z(Z'V_0Z)^{-1}Z'\left(\frac{d}{\bar{\sigma}_0^2} - \mathbf{1}_n + h\right), \tag{16}$$

where $\bar{\sigma}_0^2 = n\hat{\sigma}_0^2/(n-p)$ is the unbiased estimator of σ^2, V_0 is defined in terms of H as before but now $H = X(X'X)^{-1}X'$, the form under variance homogeneity. The inter-

TABLE 3
Score test for homogeneity

Method	*ML*	*REML*	*Approximate REML*
H	3.24	3.38	3.61
D	0.47	0.51	0.55
D, H	3.32	3.45	3.68
D, D^2	3.70	3.53	3.63
D, H, D^2	6.14	5.92	6.28
D, H, D^2, H^2	6.87	6.92	7.27
D, H, D^2, HD, H^2	8.32	8.20	8.44

pretation as half the regression sum of squares is maintained for equation (16) but it is a generalized regression of $V_0^{-1}d/\bar{\sigma}_0^2$ on Z with inner product matrix V_0. Replacing V_0 by the diagonal matrix with elements $(1-h_{ii})^2$ leads to an approximate test statistic. Asymptotically, the null distribution of all test statistics is χ^2 on $q-1$ degrees of freedom.

Table 3 provides tests of heteroscedasticity based on the score test for the cherry tree data. The table indicates that the score tests based on ML, REML and approximate REML are very similar in their magnitude and so it is unlikely to matter which is used. The tests indicate that squared diameter is important and that height is marginally important as far as the variance is concerned.

The ML score test is known to be conservative (Breusch and Pagan, 1979) and this is probably true of the test based on the REML score. Honda (1988) uses the general results of Harris (1985) for the test based on equation (15) to provide a test with a size that is closer to the nominal level. It may be possible to adjust equation (16) by using the results of Harris (1985) and this is the subject of current research.

5. DIAGNOSTICS

5.1. *Introduction*

We consider leverage and influence and in particular the use of case weights as a diagnostic tool. For the standard normal regression model the use of diagnostics is discussed by several researchers (Belsley *et al.*, 1980; Cook and Weisberg, 1982; Atkinson, 1985). For logistic regression, Pregibon (1981) discusses the use of one-step approximations to exact deletion statistics. Pregibon (1981) points out, and Williams (1987) shows, that the one-step approximation can also be used for generalized linear models. Williams (1987) uses this idea to examine the effect of deletion on various attributes of interest in an analysis. Cook (1986) considers local measures of influence on Pettitt and bin Daud (1989) use the ideas for the proportional hazards model.

The use of REML in diagnostic methods is implicit in most normal regression situations in that the common variance is estimated unbiasedly, which is the REML estimate. Cook *et al.* (1982) examine an alternative outlier model, where it is the variance of the ith observation which differs from the remainder. ML does not result in the observation with largest standardized residual being deemed the outlier, but REML does exactly that, as is shown by Thompson (1985).

5.2. *Leverage*

In the ordinary linear model, leverage refers to a measure of the extremity of the vector of explanatory variables, or the design space, for each case, i.e. x_i, $i = 1, 2, \ldots, n$. Leverage is measured by the diagonal elements of the hat matrix $X(X'X)^{-1}X'$. At points of high leverage, the fitted value of the model is mainly determined by the observed data at that point.

In a weighted regression with weights given by the diagonal matrix W, the hat matrix becomes $W^{1/2}X(X'WX)^{-1}X'W^{1/2}$. The diagonal elements of this hat matrix again indicate how extreme each x_i is in the design space, but the extremity is tempered by the magnitude of w_i, the weight for the ith observation. In this case, as in ordinary regression, we can consider a point to be of high leverage, if its fitted value is mainly determined by the observed data at that point. Of course, cases of high leverage are not necessarily influential cases in the fit of the model, in either the unweighted or the weighted regression.

Leverage measures can be constructed for both mean and variance model. We examine the appropriate hat matrix for each model. Under ML and REML, the hat matrix for the mean model is given by equation (12), where the ML and REML estimators are substituted for the unknown σ_i^2. Thus under both ML and REML the leverages involve estimated weights (reciprocals of variances) and as such are random. The hat matrix is an artefact of the scoring procedure, as Pregibon (1981), p. 712, points out. As in the logistic regression case discussed by Pregibon (1981), the weights are determined by the fit of the model, unlike true weighted regression where the weights determine the fit. None-the-less, for the mean model, the diagonal elements of equation (12) will indicate those cases for which the fitted values are largely determined by the observed value for that case. Thus a plot of the diagonal elements of H is a valuable diagnostic.

For the variance model, the hat matrices under ML and REML are

$$\begin{aligned} K_{\mathrm{M}} &= Z(Z'Z)^{-1}Z', \\ K_{\mathrm{R}} &= V^{1/2}Z(Z'VZ)^{-1}Z'V^{1/2} \end{aligned} \tag{17}$$

respectively. In the REML form we would prefer to use the diagonal approximation for V because of the simplifications that this entails. Then the ith diagonal element of K_{R} is

$$k_{\mathrm{R},ii} = (1 - h_{ii,\mathrm{R}})^2 z_i' \left\{ \sum_{i=1}^{n} (1 - h_{ii,\mathrm{R}})^2 z_i z_i' \right\}^{-1} z_i.$$

In what follows, we avoid subscripts M and R in h_{ii} and k_{ii} because the context will make it clear which is appropriate. The interpretation that points of high leverage largely determine the fitted value at that point is also applicable to the variance model. If $k_{ii} = 1$, under ML this implies that $d_i/\hat{\sigma}_i^2 - 1 = 0$ whereas under REML this becomes $d_i/\bar{\sigma}_i^2 - 1 + h_{ii} = 0$.

Under REML, the leverage of observations in the variance model is dependent on the leverage in the mean model. An observation, case i say, with high leverage in the mean model results in a small residual. If z_i is also extreme, under ML and using K_{M}, the variance model is forced to fit such a point well; this is not the case under REML.

Plotting the leverages of the variance model will be a useful diagnostic under either method of estimation.

For the tree data, plots of the leverages for the mean and variance models, the diagonal elements of equations (12) and (17), are given in Fig. 2. The plots for the mean model are similar, with REML exhibiting a decrease in leverage for the high leverage points and a slight increase for those points with low leverage. Equalization of leverage has been discussed by Dorfman (1991) and appears to be desirable in some contexts.

For the variance model, the points remote in the design space as indicated by ML are downweighted in REML, because they are also points of high leverage in the mean model. In particular, cases 1 and 31 have high leverage for the mean model and under ML they are also cases of high leverage for the variance model. Hence the fitted values are determined largely by the squared residuals for those cases. Under REML these cases have small leverage for the variance model fit, and the fit does not depend largely on the squared residual at each of those points.

5.3. *Influence*

We consider case weight perturbations in each of the mean and variance models and investigate the effect of such perturbations on parameter estimates and the likelihood. For the ordinary regression model, case weights often form the basis for the study of influence and

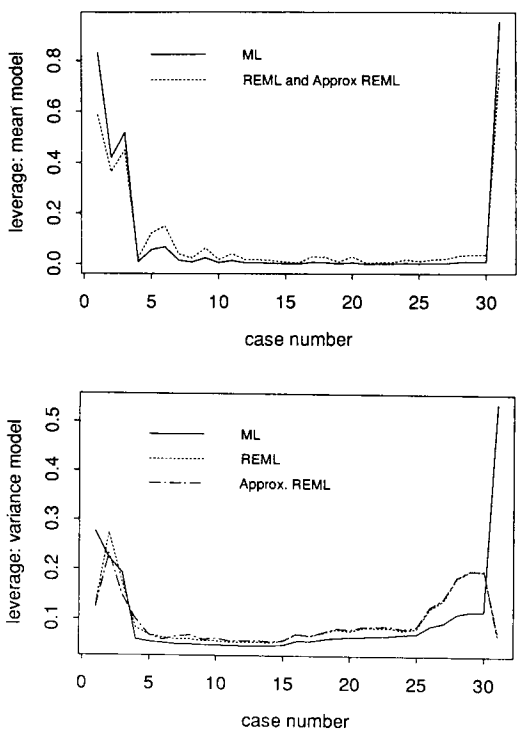

Fig. 2. Leverage plots for the mean and variance models: tree data

$$\sum_{i=1}^{n} \frac{w_i(y_i - x_i'\beta)^2}{\sigma^2}$$

forms the basis for a perturbation scheme. Case weights can therefore be interpreted as a modification of the variance of that case. In generalized linear models a case weight is incorporated in the log-likelihood. We could do precisely that here but because of the joint modelling situation we consider the two case weight schemes

$$l_m(\beta, \lambda, w; Y) = -\frac{1}{2}\left\{ \sum_{i=1}^{n} \log\left(\frac{\sigma_i^2}{w_i}\right) + \frac{w_i(y_i - x_i'\beta)^2}{\sigma_i^2}\right\}, \tag{18}$$

$$l_v(\beta, \lambda, w; Y) = -\frac{1}{2}\left\{ \sum_{i=1}^{n} \log \sigma_i^{2w_i} + \frac{(y_i - x_i'\beta)^2}{\sigma_i^{2w_i}}\right\}, \tag{19}$$

the first appropriate for mean model case weights, the second for variance model case weights.

In scheme (18), the case weights can be considered as changes in the variances as in ordinary regression, even though the variances are not necessarily equal. As $w_i \to 0$ for case i, the variance tends to infinity and there is no information available from case i. This corresponds to deletion of case i. Under scheme (19), the variances are again changed but via a power. If $w_i = 0$ and $w_j = 1, j \neq i$, case i has constant variance 1, whereas the remaining cases have variances depending on the explanatory variables. This perturbation scheme attempts to examine the influence of cases on the fit of the variance model. It is clearly artificial to impose a variance of 1 on a response but the advantage is the simple nature of the diagnostics produced. The relative changes in parameter estimates and log-likelihoods across observations and not the absolute changes are then important. It is more sensible to write $\sigma_i^2 = \sigma^2 s_i$, where s_i depends on the explanatory variables, and to consider $\sigma^2 s_i^{w_i}$. Unfortunately this leads to complicated expressions for diagnostics for the estimation of λ.

The REML equivalents to schemes (18) and (19) both involve a modification of the variances. The REML likelihood (10) is modified under equation (18) by replacing Σ^{-1} by $\Sigma^{-1}W$ where W is a diagonal matrix of case weights, whereas under equation (19) Σ^{-1} is replaced by a diagonal matrix whose elements are powers of the elements of Σ^{-1}. In both cases $\bar{\beta}$ is replaced by $\bar{\beta}(w)$, equation (8) with Σ^{-1} replaced appropriately.

Under schemes (18) and (19) we can investigate changes in parameter estimates and other attributes, in particular the log-likelihood displacement. Using the notation of Cook (1986), we consider the log-likelihood displacement

$$\mathrm{LD}(w) = 2\{l(\hat{\beta}, \hat{\lambda}; Y) - l(\hat{\beta}_w, \hat{\lambda}_w; Y)\}. \tag{20}$$

In this paper we focus on case deletion; thus we take $w_i = 0$ for each i in turn, the remaining weights being equal to 1. Local influence will be discussed elsewhere.

For case deletion under scheme (18), we can use the results of Pregibon (1981) to find a one-step approximation for the change in the ML estimates of β, namely for deletion of the ith observation, where the subscript (i) denotes deletion:

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X'\hat{\Sigma}^{-1}X)^{-1}x_i r_i}{\hat{\sigma}_i^2(1 - h_{ii})}. \tag{21}$$

The REML result is identical with ML estimates on the right-hand side of equation (21) replaced by their REML counterparts. Equation (21) usually understates the change in parameter estimates but is a useful approximation as it is in ordinary regression.

The modified score and information functions for λ under scheme (18) are easily found. Unfortunately the one-step approximation similar to equation (21) tends to underestimate the actual change. There is some improvement if we use $d_{j(i)}$, the residuals from the fit using equation (21), but the large changes still tend to be underestimated. Under ML, with initial estimate of λ equal to $\hat{\lambda}$, our one-step approximation for the modified λ under scheme (18) is

$$\hat{\lambda}_{(i)} = \hat{\lambda} + \hat{\delta} + \frac{(Z'Z)^{-1}z_i}{1-k_{ii}}\left(z_i'\hat{\delta} - \frac{d_{i(i)}}{\hat{\sigma}_i^2} + 1\right), \tag{22}$$

where $\hat{\delta} = (Z'Z)^{-1}Z'(\hat{\Sigma}^{-1}d_{(i)} - 1)$. Under REML, we find the one-step approximation

$$\bar{\lambda}_{(i)} = \bar{\lambda} + \bar{\delta} + \frac{(Z'VZ)^{-1}z_i(1-h_{ii})}{1-k_{ii}}\left\{(1-h_{ii})z_i'\bar{\delta} - \frac{d_{i(i)}}{\bar{\sigma}_i^2(1-h_{ii})} + 1\right\}, \tag{23}$$

where k_{ii} is given in equations (17) and $\bar{\delta} = (Z'VZ)^{-1}Z'(\hat{\Sigma}^{-1}d_{(i)} - \mathbf{1}_n + h)$. If we use the diagonal approximation for V, k_{ii} is given below equations (17). Having evaluated a one-step approximation, it is easy to calculate equation (20) by using ML or REML.

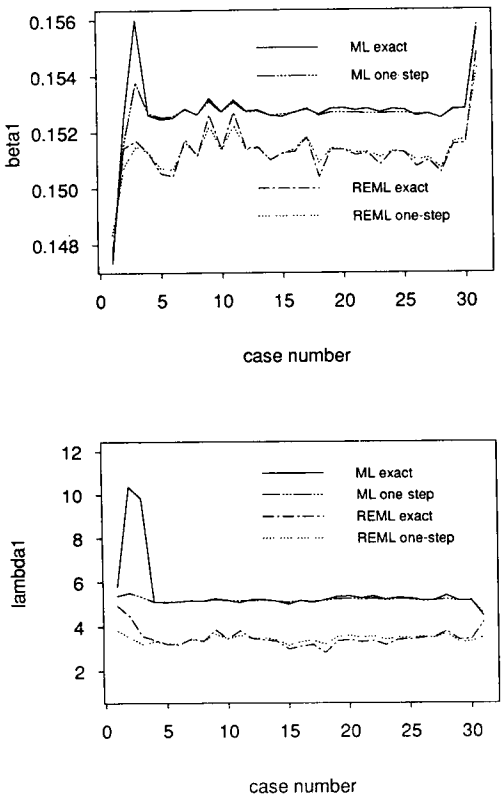

Fig. 3. Parameter estimates under deletion in the mean model

For the scheme defined by equation (19) with deletion and under ML the one-step approximation for β is

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X'\hat{\Sigma}^{-1}X)^{-1}x_i(1-\hat{\sigma}_i^2)r_i}{\hat{\sigma}_i^2\{1-(1-\hat{\sigma}_i^2)h_{ii}\}}, \tag{24}$$

which is Pregibon's (1981) result with w replaced by $\hat{\sigma}_i^2$. Again this form is also appropriate under REML. The one-step approximations for λ under scheme (19) turn out to be equations (22) and (23) for ML and REML respectively. The REML form requires an approximation which corresponds to using the diagonal form of V.

These ideas were applied to the MINITAB tree data. Both exact and one-step approximations, using equations (21)–(24), to the parameter estimates and the log-likelihood displacement (20) under case deletion were calculated under schemes (18) and (19) and for ML and REML.

Figs 3 and 4 contain plots of selected parameters and their estimates under schemes (18) and (19). The behaviour of the other parameters is very similar. There are several conclusions evident from Figs 3 and 4. The one-step approximations for the mean parameters appear quite reasonable. The corresponding one-step approximations for the variance parameters understate the important changes substantially under ML. The changes in the variance parameters under scheme (18) are more pronounced than under scheme (19). Perhaps the most significant result lies in the comparison of the

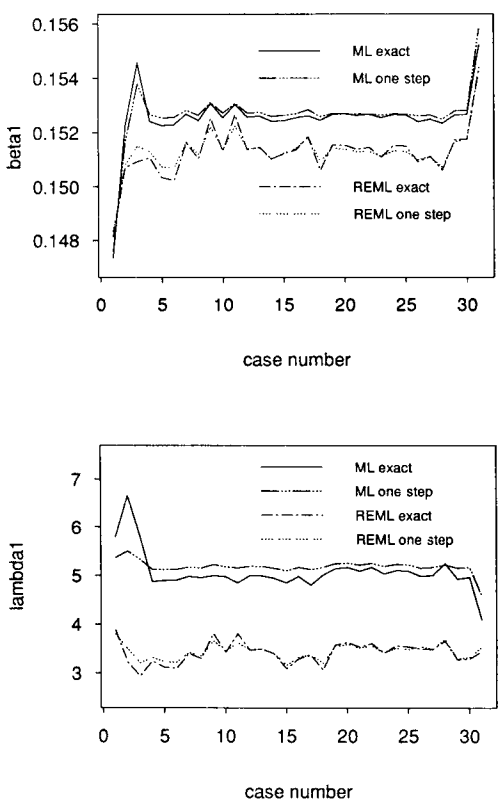

Fig. 4. Parameter estimates under deletion in the variance model

REML and ML estimates under deletion. Under ML we have some large changes in variance parameter estimates whereas for REML these changes are much less severe. These smaller changes in REML are accompanied by marginally larger changes for non-influential cases, a little like the leverage situation. In addition, the parameter estimates differ under ML and REML systematically and it appears that both mean and variance parameters under ML are biased, which is in line with previous work. Although not presented here, the use of the diagonal approximation for the matrix V in the one-step approximations (its use in the exact calculations provides the same results as full REML) was almost identical with the use of the full V and hence can be used to simplify the calculations.

The log-likelihood displacements, for both deletion in the mean (18) and the variance (19), are given in Fig. 5. Under ML, several log-likelihood displacements were too large to include formally on the plot and they are indicated on the plot by an arrow. The one-step approximations for the log-likelihood displacement are not always very good; they appear more appropriate under scheme (19). The insensitivity of changes under scheme (19) is again apparent when compared with the changes under scheme (18). If the exact changes under both deletion schemes are examined, REML is seen to have more stable log-likelihood displacements than does ML and in

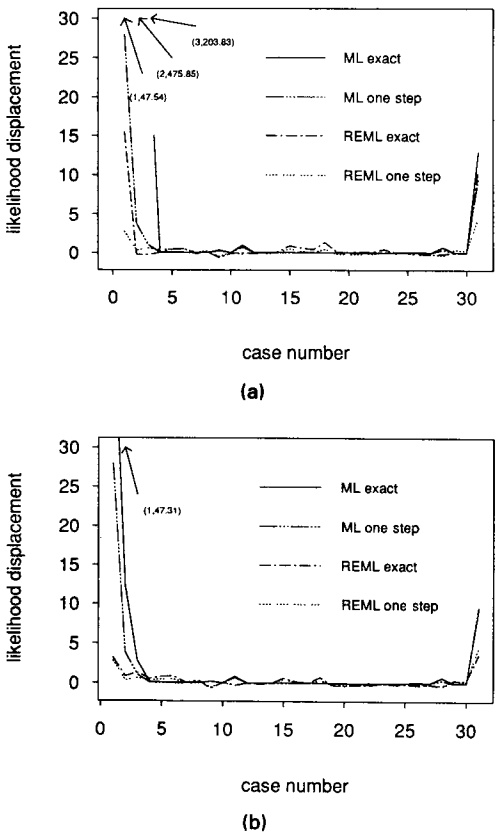

Fig. 5. Likelihood displacement under deletion: (a) perturbations in the mean model; (b) perturbations in the variance model

particular the very large likelihood displacements evident under ML in Fig. 5 are much smaller under REML. This suggests that REML estimation is affected less by individual cases than is ML.

A closer examination of the failure of the one-step approximations reveals that equations (21) and (24) are crucial for the approximations to be accurate. These two approximations indicate which cases are potentially important, but because they understate the important changes the approximations for λ and hence for the log-likelihood displacement are inaccurate. Williams (1987) suggests that once cases are recognized as potentially influential, here indicated by changes in the estimate of β by using the one-step approximations, the full iteration should be carried out for those cases, rather than relying on the one-step approximations. This seems advisable in the current situation.

6. DISCUSSION

There are theoretical reasons for preferring REML to ML. The reduction in the bias of variance parameter estimators is frequently cited, but parameter transformations in general disturb this property. More importantly, REML provides an unbiased score for the estimation of variance parameters and this unbiasedness is invariant under parameter transformations. Work on inference with nuisance parameters indicates that REML or the use of marginal and conditional likelihoods is very useful; see for example Cruddas *et al.* (1989) and references therein.

The aim of the present study was the comparison of ML and REML in a simple setting. In particular, how individual data points impact on ML and REML estimation was of interest. The leverage measures for variance estimation under REML and ML can differ substantially. Points of high leverage in the mean model result in small residuals. Under ML, the mean model leverage is ignored and the variance model may be forced close to such points if the leverage is also high for the variance model. Under REML, such points are downweighted in fitting the variance model and so other points will dictate the fit at these points. This difference in leverage can lead to some cases being highly influential under ML but not under REML. Changes in the mean parameters under deletion are qualitatively similar under ML and REML, whereas the changes in variance parameter estimates and log-likelihood displacements are more stable under REML than under ML. This suggests that fitting these models by using REML may be less sensitive with respect to individual cases than by ML.

ACKNOWLEDGEMENT

The author thanks a referee for valuable comments which led to improvements in the paper.

REFERENCES

Aitkin, M. (1987) Modelling variance heterogeneity in normal regression using GLIM. *Appl. Statist.*, **36**, 332–339.

Atkinson, A. C. (1985) *Plots, Transformations and Regressions*. Oxford: Clarendon.

Baker, R. J. and Nelder, J. A. (1978) *Generalized Linear Interactive Modelling: Release 3*. Oxford: Numerical Algorithms Group.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language: a Programming Environment for Data Analysis and Graphics*. Pacific Grove: Wadsworth and Brooks/Cole.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics*. New York: Wiley.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Box, G. E. P. and Meyer, R. D. (1986) Dispersion effects from fractional designs. *Technometrics*, **28**, 19–27.

Breusch, T. S. and Pagan, A. R. (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, **47**, 1287–1294.

Carroll, R. J. and Ruppert, D. (1988) *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Cook, R. D. (1986) Assessment of local influence (with discussion). *J. R. Statist. Soc.* B, **48**, 133–169.

Cook, R. D., Holschuh, N. and Weisberg, S. (1982) A note on an alternative outlier model. *J. R. Statist. Soc.* B, **44**, 370–376.

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

——— (1983) Diagnostics for heteroscedasticity in regression. *Biometrika*, **70**, 1–10.

Cooper, D. M. and Thompson, R. (1977) A note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika*, **64**, 625–628.

Cruddas, A. M., Reid, N. and Cox, D. R. (1989) A time series illustration of approximate conditional likelihood. *Biometrika*, **76**, 231–237.

Dorfman, A. H. (1991) Sound confidence intervals in the heteroscedastic linear model through releveraging. *J. R. Statist. Soc.* B, **53**, 441–452.

Evans, M. A. and King, M. L. (1988) A further class of tests for heteroscedasticity. *J. Econometr.*, **37**, 265–276.

Harris, P. (1985) An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika*, **72**, 653–659.

Harvey, A. C. (1976) Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, **41**, 461–465.

Harville, D. A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

Honda, Y. (1988) A size correction to the Lagrange multiplier test for heteroskedasticity. *J. Econometr.*, **38**, 375–386.

Park, R. E. (1966) Estimation with heteroscedastic error terms. *Econometrica*, **34**, 888.

Patterson, H. D. and Thompson, R. (1971) Recovery of interblock information when block sizes are unequal. *Biometrika*, **54**, 545–554.

Pettitt, A. N. and bin Daud, I. (1989) Case-weighted measures of influence for proportional hazards regression. *Appl. Statist.*, **38**, 51–67.

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.*, **9**, 705–724.

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

Smyth, G. K. (1989) Generalized linear models with varying dispersion. *J. R. Statist. Soc.* B, **51**, 47–60.

Stirling, W. D. (1985) Heteroscedastic models and an application to block designs. *Appl. Statist.*, **34**, 33–41.

Thompson, R. (1985) A note on restricted maximum likelihood estimation with an alternative outlier model. *J. R. Statist. Soc.* B, **47**, 53–55.

Verbyla, A. P. (1990) A conditional derivation of residual maximum likelihood. *Aust. J. Statist.*, **32**, 221–224.

Williams, D. A. (1987) Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statist.*, **36**, 181–191.