A Use of the Contributions to lack-of-fit in Deletion Diagnostic for Generalized Least Squares

Sakyajit Bhattacharya[a,1,*], John Haslett[b], Thomas Brendan Murphy[a]

[a] *School of Mathematical Sciences, University College Dublin, Dublin 4, Ireland*
[b] *School of Computer Science & Statistics, Trinity College Dublin, Dublin 2, Ireland*

Abstract

The paper studies the use of the *contributions to lack of fit* in detecting outliers in linear models with Gaussian errors (see Haslett and Hayes (1998) for the concept of contributions). The paper shows that contributions are an important factor behind an observation's departure from the estimated model and from the dataset. The paper also shows by illustration that contributions perform better than marginal and conditional residuals in detecting outliers in a dataset. The demerits of this measure are that they can be negative and can lead to false identification of outliers.

1. Introduction

This paper examines the use of the *contributions to lack of fit* in a dataset, first introduced by Haslett and Hayes (1998), in detecting outlying observations for linear models with (possibly) correlated Gaussian errors. In the linear model $Y = X\beta + \varepsilon$, the contribution to lack of fit of the ith observation can be expressed as

$$T_i = d_i^{-1}\hat{e}_i\tilde{e}_i \tag{1}$$

where $\hat{e}_i$ is an estimate of the marginal residual of the ith observation which measures the deviation of a point from the fitted trend, $\tilde{e}_i$ is an estimate of the conditional residual of the ith observation which measures the deviation of a point from its neighbourhood, and $d_i = \mathrm{Var}(\tilde{e}_i)$ (see Haslett and Haslett (2007) for the introduction of these two residuals). Haslett and Hayes (1998) proposed that contributions can be used to detect abnormality in a dataset under a fitted model. The authors suggested that if T_i is of large magnitude, then the ith observation is an anomaly in the sense that it either lies far away from the fitted trend, or lies far away from its neighbouring points, or both. The present paper can be considered as a critical examination of the proposal of Haslett and Hayes (1998). The paper mathematically proves that T_i, as a joint effect of $\hat{e}_i$ and $\tilde{e}_i$, contributes significantly to an observation's departure from the dataset as well as from the fitted model. The paper also shows by illustration that T_is are more effective than $\hat{e}_i$ or $\tilde{e}_i$ in detecting abnormal observations. However, the demerits of T_is lie in the facts that they may be negative, and can be 0 even if one of the residuals is 0 and the other is not.

Contributions can be effectively defined for a block of more than observations, too. If the dataset Y is partitioned as (Y_a, Y_b), then contribution for the block of observations Y_a can be defined as

$$T_a = \hat{e}_a^T D_a^{-1}\tilde{e}_a \tag{2}$$

where $\hat{e}_a$ is the set of marginal residuals for the corresponding observations belonging to Y_a, $\tilde{e}_a$ is the set of conditional residuals for the corresponding observations belonging to Y_a, and $D_a = \mathrm{var}(\tilde{e}_a)$. Haslett and Hayes

*Corresponding author
Email addresses: sakyajit.bhattacharya@ucd.ie (Sakyajit Bhattacharya), john.haslett@tcd.ie (John Haslett), brendan.murphy@ucd.ie (Thomas Brendan Murphy)

[1] This work was supported by SFI Research Frontier Grants (2007/RFP/MATH228).

(1999) pointed out that contributions enjoy a unique property of additivity in this respect. They showed that the contribution for a block of observations is the sum of the contributions of the singletons belonging to that block. So, when there are some naturalized blocks in a dataset (for example, Linear Mixed Models (LMM)), the additivity of contributions can be effectively used to find the contribution of a block.

Haslett and Hayes (1998) introduced contributions in diagnostics of the model $Y = X\beta + \varepsilon$ where Y is an $n \times 1$ vector, X is an $n \times p$ matrix, β is a $p \times 1$ vector of parameters, ϵ is the $n \times 1$ vector of errors, and $\text{Var}(\varepsilon) = V$. The authors showed that, if $P = (a, b, c, ...)$ is a complete partition of the indexes of Y, and if $\tilde{e}_a$, $\tilde{e}_b$, $\tilde{e}_c$... denote the conditional residuals associated with this partition and $\tilde{e}_P$ is the stacked vector of these residuals, then

$$D_P^{-1}\tilde{e}_P = V^{-1}\hat{e} = QY \tag{3}$$

where $\hat{e}$ is the marginal residual, $Q = V^{-1}\left(I - X(X^TV^{-1}X)^{-1}X^TV^{-1}\right)$, and D_P is block-diagonal so that $D_a = \text{Var}(\tilde{e}_a) = Q_{aa}^{-1}$, $D_b = \text{Var}(\tilde{e}_b) = Q_{bb}^{-1}$ and so on.

The above argument validates the dual roles of marginal and conditional residuals in a general linear model with correlated errors. If $(T_a, T_b, T_c, ...)$ is the stacked vector of block contributions as defined in equation 2, then equation 3 shows that the lack of fit S can be expressed as

$$S = \sum_{m \in P} T_m,$$

where T_i is the weighted combination of two different types of residuals. Further, the authors showed that marginal residual measures a global deviation of a point as represented by its distance from its estimated mean, and conditional residual measures the local deviation of a point as represented by its distance from its estimated conditional mean. Hence the total lack of fit of a dataset under a fitted model is a joint effect of both global and local deviations. Since T_is take into account of both kinds of deviations, the authors suggested that T_is can be a useful tool in understanding the influence of global and local deviations on the abnormal behaviour of a dataset.

There were, however, two shortcomings in the method suggested by Haslett and Hayes (1998) which this paper addresses:

- The authors did not illustrate how the re-estimation of the variance matrix after deletion of data points can affect the contributions. Throughout their analysis they assumed that the variance matrix needs not be re-estimated after each stage of deletion, which in some sense can falsely represent an observation to be outlier, as shall be shown in our data analysis in section 4.

- The sampling distribution of T is that of a weighted distribution of two independent chi-squared variables. The cdf and probability interval of such a distribution is very complicated; hence the authors went to an approximate standardization of T which is useful in the upper tail, being based on the transformation of a single chi-squared variable. Such an approximation has two disadvantages. The standardized contribution loses its additive property, and we can not approximate the lower tail of T which can also be a source of adequate information.

The present paper deals with the first shortcoming by re-estimating V at each stage of deletion by methods suggested by Haslett and Dillane (2004). We have used two types of deletion diagnostics. The one by re-estimating V at each stage of deletion, the other by not re-estimating V. The method of re-estimation proposed by Haslett and Dillane (2004) is a computationally cheaper procedure(See the appendix for more details) compared with the other previous methods like Christensen et al. (1992). We then compared the two methods and studied how the change in the variance structure affects the dataset. Re-estimating the variance parameters is important in the sense that the abnormal behaviour of an observation might be caused due to large variance which can influence the neighbouring points, too, to behave abnormally. When the point is deleted, re-estimating the variance parameters would get rid of the influence, if any, of the point's large variance over other points.

To address the second shortcoming, we have used some close approximations of the cdf of T derived by Bhattacharya et al. (2011) which helps to derive the approximate quantiles of the probability distribution.

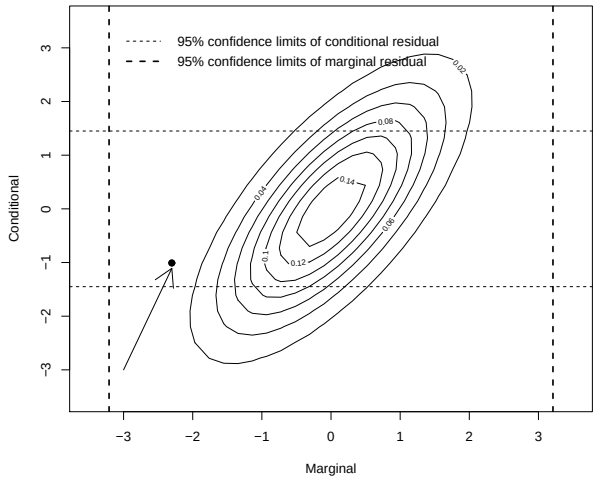


Figure 1: Point 9 from the simulated data has normal marginal and conditional residuals, but the corresponding $(\hat{e}_i, \tilde{e}_i)$ lies far from the 95% bivariate contour, and hence shows high contribution.

Such an approximation retains the additive property of T and takes care of the left tail of the distribution. Then we propose to plot T_is along with their corresponding 95% probability intervals, and detect the points that lie outside the corresponding intervals.

To study the relative merits and demerits of contributions as compared to marginal and conditional residuals, we aim to validate our argument at theoretical discussion and by simple illustrations. We have theoretically interpreted contributions as a joint effect of global and local residuals in section 2. We have also compared the 'sensitivity' of contributions with that of marginal and conditional residuals in detecting unusual observations. By sensitivity, here we meant the rate of change in the value of a measure with the change in the value of a given observation. It can be shown that even for moderate outliers the contributions are more sensitive in detecting them. The drawback arises when one of the conditional or marginal residuals is 0. Then contribution can be 0 but the other non-zero deviation can still be large.

For a quick example, we have studied the observations from a simulated time series data with AR(1) error. We have analysed the dataset in detail in section 3. For the time being, we only considered observation 9 which shows a high contribution. The observation has marginal and conditional residuals falling well inside the respective 95% probability intervals, but the corresponding $(\hat{e}_i, \tilde{e}_i)$ lies far away from the 95% bivariate contour of the marginal and conditional residuals, as shown in figure 1. So, if we look at the marginal and conditional residuals for point 9, they do not reflect any abnormality. But the pair of residuals jointly lie far away from the contour. So, the joint effect of the global and local deviations are large, even though their individual effects are not. The contribution is 4.35, larger than the upper limit of its 95% probability interval which is 3.90. This is an illustration to show that contributions can detect abnormalities which may not be reflected by marginal or conditional residuals by themselves.

To compare the use of the above measures in the context of data analysis, we simulated an AR(1) data in section 3 and studied the properties of contributions along with their rate of detection for additive outliers, innovative outliers, and group of additive outliers. We validated our arguments through comparison with marginal and conditional residuals.

In section 4 we examined two real life data sets where the observations are grouped into a number of blocks. These two data sets, combined with the simulated one, cover a vast area of examples. The simulated data is an example of a time-series, Corn data is an example of an LMM, and Ovary data is an example of repeated measures. Moreover, the simulated data looks at the singletons and the effect of their deletion

over the remaining dataset. Corn data has some naturalized blocks and so we can study the effect of block deletion. Ovary data has naturalized blocks where each block behaves as an AR(1) process. In ovary data, therefore, we can study the effect of deletion of singletons as well as of the blocks that behave as time series.

2. Mathematical properties of Contribution

Contribution for a block of observations Y_a is defined as

$$T_a = \hat{e_a}^T D_a^{-1} \tilde{e_a}.$$

For Gaussian errors, Haslett and Hayes (1998) showed that T_a can be expressed as

$$T_a = \frac{\gamma_a + \phi_a}{2\kappa_a}\chi^2_{1,\kappa_a} - \frac{\gamma_a - \phi_a}{2\kappa_a}\chi^2_{2,\kappa_a} \equiv \alpha\chi^2_{1,\kappa_a} - \beta\chi^2_{2,\kappa_a} \tag{4}$$

where κ_a is the length of the block a, χ^2_{1,κ_a} and χ^2_{2,κ_a} are independent χ^2 variables with degrees of freedom κ_a, $\gamma_a = \text{tr}\left[(G^{.5}_{aa} D_a^{-1} G^{.5}_{aa})^{.5}\right]$, $G = VQV$, G_{aa} is the ath block of the block diagonal matrix G_P which is created by taking the diagonal blocks of the matrix G, and $\phi_a = \text{tr}(VQ)_{aa}$.

Let us consider the case when the block of observations Y_a contains only a singleton, i.e. $a = \{i\}$, say. Then T_i will depend on three components, marginal residual of the ith observation $\hat{e}_i$, conditional residual of the i-th observation $\tilde{e}_i$, and members of the diagonal matrix D. This is a drawback of the measure in the sense that if $\hat{e}_i$ is 0 and $\tilde{e}_i$ is large (or vice versa), T_i will be 0, without taking into account of the large conditional residual. So, a risk of falsely identifying an unusual observation to be normal always lies in using contributions.

Another drawback appears when $\hat{e}_i$ and $\tilde{e}_i$ are of different signs. Then T_i will be negative. A negative contribution is hard to explain and it certainly is a departure from our general idea of measures of deviation. The other well-known measures like Cook's distance or Mahalanobis distance are all positive, and we can interpret them as some kind of metrics in the Euclidean space. Hence we can well describe the situation when these measures are 0. But even when contribution of a point is 0, we can not conclude that the point has no abnormality. Similarly, when the contribution is negative, we can not interpret what kind of deviation makes this happen.

d_i^{-1} is defined as the ith element of diag(Q), where Q can be defined as $V^{-1}(I - H)$ where H is the Hat matrix. If $V = I$, then $Q = I - H$. Also, the leverage is defined as diag(H). Hence, high-leverage will imply that the elements of D will be large. Since $T_i = \hat{e}_i^T d_i^{-1} \tilde{e}_i$, T_i will be small for high leverage. The result can similarly be generalized for a general covariance structure.

2.1. Approximate cdf of contribution

Equation 4 shows that the analytical cdf of T is of complex form and not invertible. We shall use some close approximation of the cdf of T by using some approximations of the Confluent Hypergeometric function of the second type (See, for example, Press (1966) and Bhattacharya et al. (2011)). The approximate cdf is given below:

$$P(T \le t) \simeq \begin{cases} 1 - A_1(\beta)\Gamma\left(\frac{\kappa}{2}, \frac{t}{2\alpha}\right) & \text{for large positive } t \text{ and } \kappa \le 2 \\ A_1(\alpha)\Gamma\left(\frac{\kappa}{2}, -\frac{t}{2\beta}\right) & \text{for large negative } t \text{ and } \kappa \le 2 \\ 1 - A_2(\beta,\alpha)\Gamma\left(\frac{\kappa-2}{2}, \frac{t}{2\alpha}\right) & \text{for large positive } t \text{ and } \kappa \ge 3 \\ A_2(\alpha,\alpha)\Gamma\left(\frac{\kappa-2}{2}, -\frac{t}{2\beta}\right) & \text{for large negative } t \text{ and } \kappa \ge 3 \\ A_3(\alpha) + A_4(\alpha)e^{-t} & \text{for small positive } t \text{ and } \kappa \ge 3 \\ 1 - A_3(\alpha) - A_4(\beta)e^{-t} & \text{for small negative } t \text{ and } \kappa \ge 3 \end{cases} \tag{5}$$

where

$$\lambda = \frac{\alpha + \beta}{2\alpha\beta}, A_1(x) = (2\lambda x)^{-\frac{\kappa}{2}}, A_2(x,y) = A_1(x)\left[\frac{\kappa}{2}\left(1 + \frac{1}{2y\lambda}\right) - 1\right],$$

$$A_3(x) = I_{\frac{\gamma-\phi}{2\gamma}}\left(\frac{\kappa}{2},\frac{\kappa}{2}\right) + \frac{c(\kappa)}{\Gamma(\frac{\kappa}{2})}\frac{\Gamma(\kappa-1)}{\Gamma(\frac{\kappa}{2})}2x\,(1+\lambda x)\,\lambda^{-(\kappa-1)},$$

$$A_4(x) = \frac{\Gamma(\kappa-1)}{\Gamma(\frac{\kappa}{2})}2x\,(1+2\lambda x)\,\lambda^{-(\kappa-1)},$$

where $\Gamma(s,t) = (\Gamma(s))^{-1}\int_t^\infty e^{-u}u^{s-1}du$ is the incomplete gamma function of order s and $I_a(\lambda,\mu) = (\text{Beta}\,(\lambda,\mu))^{-1}\int_0^a u^{\lambda-1}(1-u)^{\mu-1}du$ is the incomplete Beta function.

For $\kappa = 2$, no approximation is needed for $P(T \le t)$ with small t. We can directly calculate the value of ψ. But for $\kappa = 1$, there is no satisfactory approximation for $P(T \le t)$ with small t. In that case we can use the large value approximations which work moderately well for small values, too. Using the form of cdf, we can now easily find the quantiles and thus the probability intervals of T.

2.2. Comparison with marginal and conditional residuals

For Gaussian errors, the marginal residual for the block Y_a, denoted by $\hat{e}_a$, is distributed as $\text{MVN}\,(0,(VQV)_{aa})$ and conditional residual $\tilde{e}_a$ is distributed as $\text{MVN}\,(0,D_{aa})$ where $D_{aa} = Q_{aa}^{-1}$.

Also,

$$\text{Cov}(\hat{e}_a,\tilde{e}_a) = (VQ)_{aa}\,D_a.$$

So,

$$(\hat{e}_a,\tilde{e}_a) \sim \text{MVN}(0,S_a) \tag{6}$$

where

$$S_a = \begin{pmatrix} (VQV)_{aa} & (VQ)_{aa}\,D_a \\ D_a\,(QV)_{aa} & D_a \end{pmatrix}.$$

Let us consider Y_i and its joint deviation. By joint deviation, we mean the departure of Y_i both from its estimated marginal mean $\hat{Y}_i$ and estimated conditional mean $\tilde{Y}_i$. So, the Mahalanobis distance of the joint deviation can be expressed as

$$\begin{aligned} M_i &= (\hat{e}_i,\tilde{e}_i)^T S_i^{-1}(\hat{e}_i,\tilde{e}_i) \\ &= \lambda_1 M_{\hat{e}_i} + \lambda_2 M_{\tilde{e}_i} + \lambda_3 T_i \end{aligned}$$

where $M_{\hat{e}_i} = \hat{e}_i^2/\text{Var}(\hat{e}_i)$, $M_{\tilde{e}_i} = \tilde{e}_i^2/\text{Var}(\tilde{e}_i)$ and λ_is are constants depending on the variance parameters.

$M_{\hat{e}_i}$ can be interpreted as the Mahalanobis distance of Y_i from its estimated mean and $M_{\tilde{e}_i}$ can be interpreted as the Mahalanobis distance from its conditional mean. Thus $M_{\hat{e}_i}$ and $M_{\tilde{e}_i}$ can be interpreted as the measures of deviation caused by the global and local characteristics of the underlying error of estimation, while T_i is the measure of deviation caused by the interaction between these two characteristics.

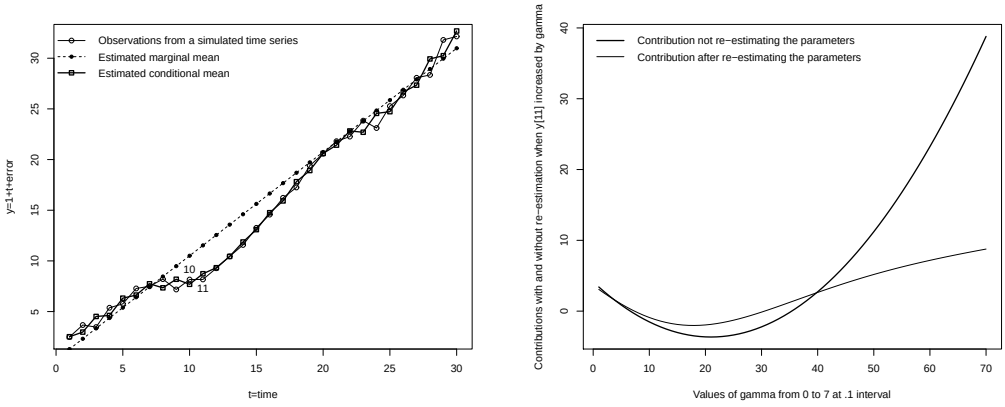
Hence, three factors contribute to an observation's unusual behaviour. Its departure from the fitted model, its departure from the neighbourhood, and a joint effect of the first two kinds of departures. T_i takes account of the last factor.

This can be similarly generalized for a block of more than one observations. For a block, the result comes in terms of weighted marginal and conditional residuals.

If a point's distance from its estimated marginal mean is expressed as l, then T_i can be expressed in a convex functional form $T_i = l(l-\alpha)/d_i$ where α is the difference between the marginal and the conditional mean, denoted as $\hat{Y}_a - \tilde{Y}_a$. Hence, $\partial T_i/\partial l = \sqrt{v_{ii}/d_i}(2l-\alpha)$. on the other hand, $\partial\hat{e}_i/\partial l = 1$.

That means, if a point's distance from its marginal mean is greater than $(\alpha+\sqrt{d_i/v_{ii}})/2$, its contribution is more sensitive than its marginal residual to detect additive outliers, because the rate of change in the value of T_i will be much more rapid than the rate of change in the value of $\hat{e}_i$ when the initial value of an observation changes. For most practical datasets, $(\alpha+\sqrt{d_i/v_{ii}})/2$ is a very small value. So, even for moderate additive outliers, T_is are practically more sensitive than marginal residuals in detecting them. The result is similar for conditional residuals also.

We discuss these features in details in the next section, where we have analysed the simulated data for different parametric values.



(a) (b)

Figure 2: Plot (a) shows observations along with estimated marginal and conditional means. Observations 10 and 11 are particularly far away from the estimated mean. Plot (b) shows the contributions for different values of γ when the additive outlier γ is put in the 11th place. Methods with and without re-estimation of parameters at each stage of the values of γ have been compared.

3. Analysis of a simulated data

We have considered a time series model $Y = 1 + t + \varepsilon$ where t is time and ε is an AR(1) process. Using this model, a series containing 30 observations are simulated with correlation parameter $\rho = 0.9$ and variance $\sigma^2 = 1$. Plot (a) of figure 2 shows the plot of the observations with their estimated marginal and conditional means. Observations 10 and 11 are indicated because they are far from the estimated mean. Observation 11 lies below the estimated mean as well as the neighbouring observations 10 and 12. We shall move observation 11 from its initial position to see how the contribution changes. In particular, we shall make the observation as an additive outlier and study the behaviour of contributions.

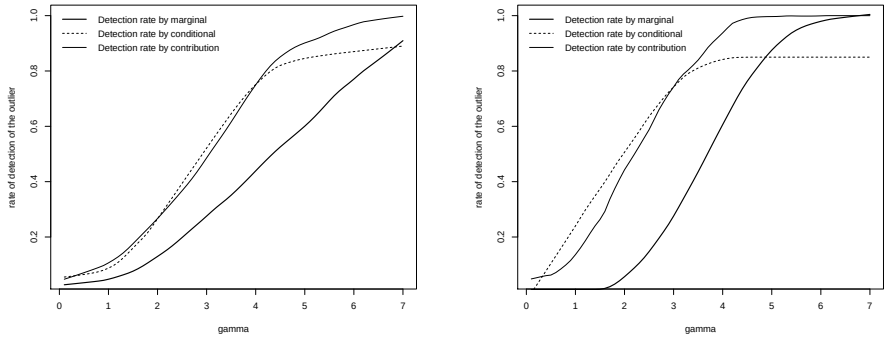
3.1. Additive outliers

For a set of observations $(y_1, y_2, ..., y_n)$ with an additive outlier at the kth position,

$$y_t = \begin{cases} u_t & \text{for } t \neq k \\ u_t + \gamma & \text{for } t = k \end{cases}$$

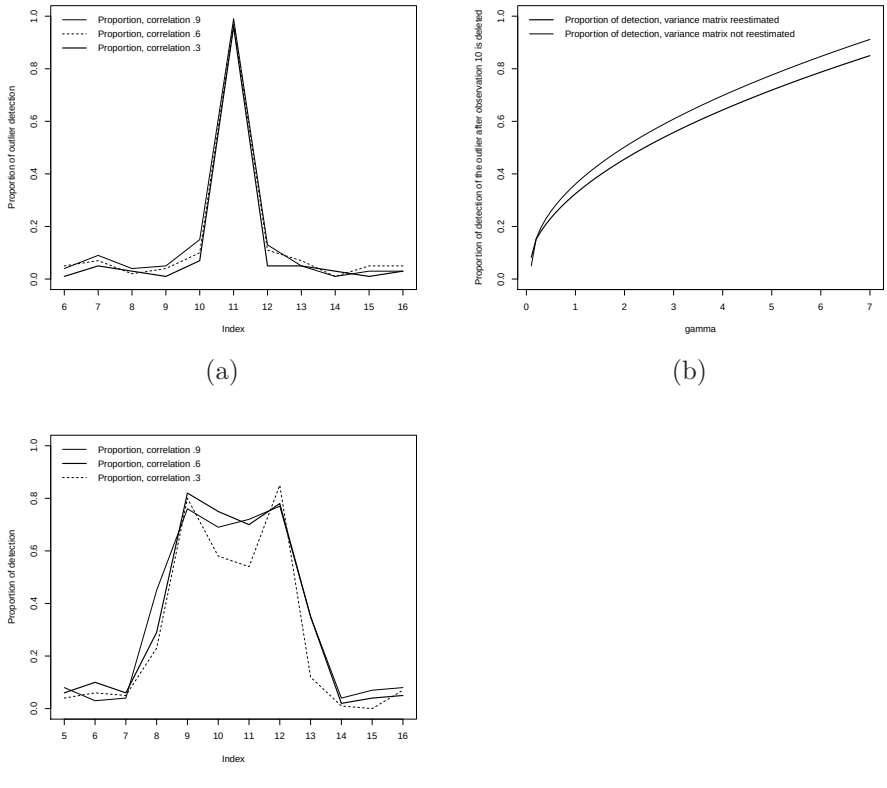
(for example, see Fox (1972)). To build an additive outlier, we put a value γ to the 11th observation, and looked at the behaviour of T_i for different values of γ, both with and without re-estimating the parameters. Plot (b) of figure 2 shows the two curves of the value of contributions, with and without re-estimation, for values of γ from 0 to 7 at 0.1 interval. The plot shows that when parameters are not re-estimated, the curve behaves like a parabola. However, the curve changes at a much slower rate when the parameters are re-estimated. So, re-estimation is recommended at each stage of the change in the additive outlier because it can get rid of false detection of outliers. Our analysis throughout has been based on re-estimation of parameters.

To compare the performance of T_i with $\hat{e}_i$ and $\tilde{e}_i$, we have run 100 simulations for $\rho = 0.9$ and $\sigma^2 = 1$. We placed γ at the 1st and the 11th position, one by one, and studied the rate of detection of the observation as an outlier for both these cases with the values of γ from 0 to 7 at 0.1 interval. Figure 3 shows the rate of detection by contributions, marginal residuals and conditional residuals for different values of γ. The figure shows that contributions behave much better than the marginal residuals in detecting outliers. However, the conditional residuals perform mildly better than contributions in detecting moderate outliers. For extreme cases, contribution is the winner.



(a) (b)

Figure 3: Plot (a) shows the rate of detection of the additive outlier when the outlier is placed in the 1st position. Plot (b) shows the rate of detection of the outlier when the outlier is placed in the 11th position. Values of γ are taken from 0 to 7 at 0.1 interval. Rate of detection of the outlying point by contribution has been compared with those of marginal and conditional residuals.



(a) (b)

(c)

Figure 4: Plot (a) shows the rate of detection of observations 6 to 16 as outliers when $\gamma = 4$ is placed in the 11th position. Three models with $\rho = 0.3$, 0.6 and 0.9 have been considered. Plot (b) shows the rate of detection of outlying observation 11 when observation 10 is deleted. Methods with and without re-estimating the variance matrix after deletion have been compared. Plot (c) shows the rate of detection of the observations 5 to 16 when $\gamma = 4$ is placed in the positions from 9 to 12. The proportions of detection for the cases $\rho = 0.3$, 0.6 and 0.9 have been compared.

To study the rate of falsely detecting an observation as an outlier, we have run 100 simulations each for $\rho = 0.3$, 0.6 and 0.9 and placed $\gamma = 4$ in the 11th position. We considered the neighbouring observations of the outlier and studied if they are detected by T_is. Plot (a) of figure 4 shows that for correlation 0.9, observations 9 and 10 have been detected by approximately 12% and 18% of times. The same is true for observations 12 and 13. Observation 11 has been detected 98% of times by contributions. The rate of false detection is significantly lower when the correlation is low. This happens due to the influence of the outlying observations over its neighbourhood. The influence has been studied and denoted as 'swamping' by Barnett and Lewis (1978) where in a group of observations the extreme outlying observation 'swamps' the other points, so that the other points also show high outlyingness. Swamping increases with high correlation. Hence contributions with low correlation structure in the model show lower rate of false detection of observations.

Lastly, to study the effect of estimating variance parameters after deletion of a point, we have deleted observation 10 and placed γ in the 11th position. Then the rates of detection of observation 11 with and without re-estimating the variance matrix after deletion were plotted in plot (b) of figure 4. The plot shows that when the variance matrix is not re-estimated, the rate of detection is more than 20% even for a small γ of value 2. For a moderate value 4, the rate of detection is almost 40% . The rate of detection is much lower when the variance matrix is re-estimated. So, we can conclude that re-estimating the variance matrix after deletion of a point gets rid of the undue influence of the variance of the deleted observation over the other points.

3.2. Group of additive outliers

If, instead of placing γ in a single position, we place γ in a block of positions, then we get a group of additive outliers. To study the performance of contributions in detecting the outlying group, we placed $\gamma = 4$ in t position 9 to 12. We have then run 100 simulations each for $\rho = 0.3$, 0.6 and 0.9 and plotted the rate of detection of observations 5 to 16 for each case in plot (c) of figure 4. The plot shows that for low correlation, the rate of detection of observations 10 and 11 is significantly lower than the rate of detection of observations 9 and 12. Also, the rate of false detection of observations 8 and 13 is much lower for low correlation. So, comparing plots (a) and (c) of figure 4, we can conclude that high correlation structure can affect the contribution's detection skill.

3.3. Innovative outliers

For a set of AR(1) observations $(y_1, y_2, ..., y_n)$ with an innovative outlier at the kth position, $y_t = \phi y_{t-1} + u_t$ and

$$u_t = \begin{cases} \epsilon_t & \text{for } t \neq k \\ \epsilon_t + \gamma & \text{for } t = k \end{cases}$$

(see Fox (1972)). To build an innovative outlier in our model, we placed $\gamma = 4$ in the 20th position and ran the simulation 100 times each for $\rho = 0.3$, 0.6 and 0.9. Plot (a) of figure 5 shows the rate of detection of observations 15 to 30 for three correlation parameters. The plot shows that for low correlation, the rate of detection of innovative outliers is very much similar to that of additive outliers, which is expected, because for low correlation the future observations behave unaffected for moderate values of γ. Also, the rate of false detection of observations 16, 17, 18 and 19 is much higher for high correlation. On the other hand, the rate of detection of observation 20 is lower for high correlation. That means, for innovative outliers, high correlation structure significantly affects the contribution's detection skill.

To compare contributions with marginal and conditional residuals, we varied γ from 0 to 7 at 0.1 interval with correlation 0.9 and ran 100 simulations. Plots (b) and (c) of figure 5 show that when γ is placed at the 20th position, then the rate of detection of the 20th and the 21st observation is more by contribution than by marginal or conditional residuals. Strikingly, the rate of detection of an innovative outlier is lower than the rate of detection of an additive outlier.

So far we have been mostly studying the singletons and the effect of the deletion of a single observation over the remaining data. How can we detect a block of unusual observations in a dataset and study the

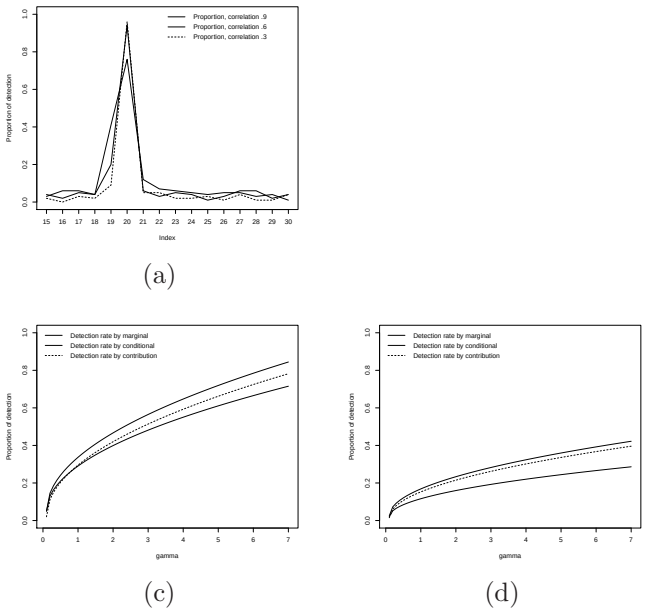


Figure 5: Plot (a) shows the rate of detection of the observations 15 to 30 when $\gamma = 4$ is placed as an innovative outlier in the 20th position. The proportions of detection for the cases $\rho = 0.3$, 0.6 and 0.9 have been compared.Plot (b) shows the rate of detection of observation 20 when the innovative outlier γ is placed in the 20th position. Plot (c) shows the rate of detection of observation 21 with observation 20 as the innovative outlier. Values of γ are taken from 0 to 7 at 0.1 interval. Rate of detection of the outlying point by contribution has been compared with those of marginal and conditional residuals.

effect of their deletion over other observations? This question is especially relevant when there are some naturalized blocks in the dataset. In this respect we move to the next section for some real life data analyses where the datasets contain naturalized blocks.

4. Data Analysis

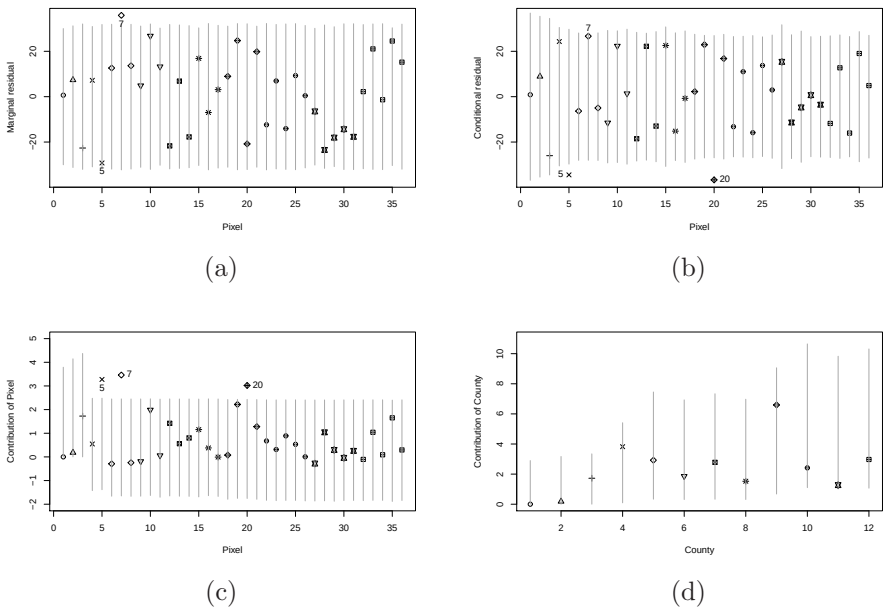
We shall now analyse two data sets to detect unusual observations using the properties and interval estimation of T.

4.1. Corn data

We consider an example of a linear mixed model. The dataset is a prediction of areas under corn and soy-bean for 12 counties in north-central Iowa, based on 1978 June Enumerative Survey and satellite data. There are 37 segments of those 12 counties. Battese et al. (1988) deleted the second segment of Hardin county from their analysis as the reported hectares of corn was identical for that of the first segment, and propose to fit the following model

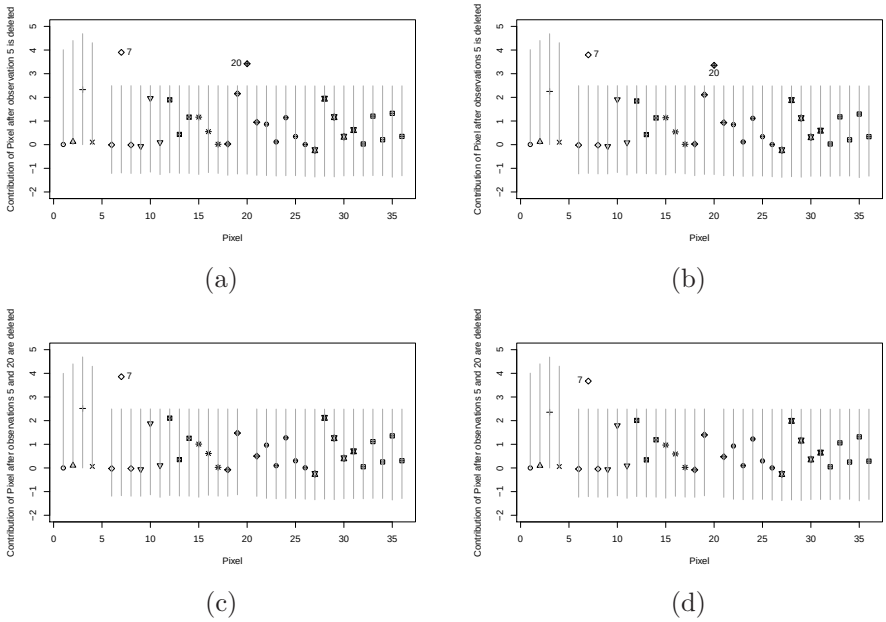
$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{ij}$$

where i is the subscript of the county, j is the subscript for a segment within a given county, y_{ij} is the number of hectares of corn (we have used the corn data as the y variable) in the jth segment of the ith county, x_{1ij} and x_{2ij} are number of pixels classified as corn and soy-beans, respectively, in the jth segment of the ith county. The random error u_{ij} can be expressed as $u_{ij} = v_i + e_{ij}$ where v_i is the ith county effect and e_{ij} is the random effect. v_i and e_{ij} are assumed to be iid Normal random variables with mean zero and variances σ_v^2 and σ_e^2 respectively. The reported crop hectares for a crop are positively correlated within



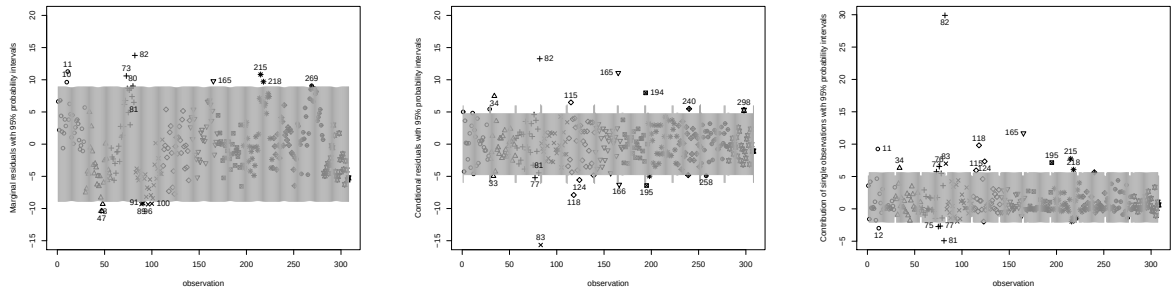
(a) (b) (c) (d)

Figure 6: An analysis of Corn data. (a) Shows the marginal residuals with 95% probability intervals, (b) Shows the conditional residuals with 95% probability intervals, (c) Shows the contributions of single observations with 95% probability intervals, (d) shows the contribution of the counties with 95% upper probability limit. Different symbols have been used for different counties.



(a) (b) (c) (d)

Figure 7: An analysis of Corn data. (a) shows the contributions after observation 5 is deleted and the variance matrix is re-estimated, (b) shows the contributions after observation 5 is deleted and the variance matrix matrix is not re-estimated,(c)shows the contributions after observations 5 and 20 are deleted and the variance matrix is re-estimated,(d)shows the contributions after observations 5 and 20 are deleted and the variance matrix is not re-estimated.



(a) (b) (c)

Figure 8: An analysis of Ovary data. (a) shows the marginal residuals with 95% probability intervals, (b) Shows the conditional residuals with 95% probability intervals, (c) Shows the contributions of single observations with 95% probability intervals.Different symbols have been used for different mares.

given counties but uncorrelated between different counties. The REML estimates of the variance parameters give $\hat{\sigma_v^2} = 103$ and $\hat{\sigma_e^2} = 194$.

A characteristic of the residuals should be noted here. Since the variance matrix is block diagonal with each block representing a county, the marginal residuals $\hat{e}_a$ for county a will be independent of the marginal residuals $\hat{e}_b$ for county b. So, a simple calculation shows that the block contributions will always be positive, distributed as an weighted chi-squared variable. Hence we have plotted the block contributions with 95% upper probability limit.

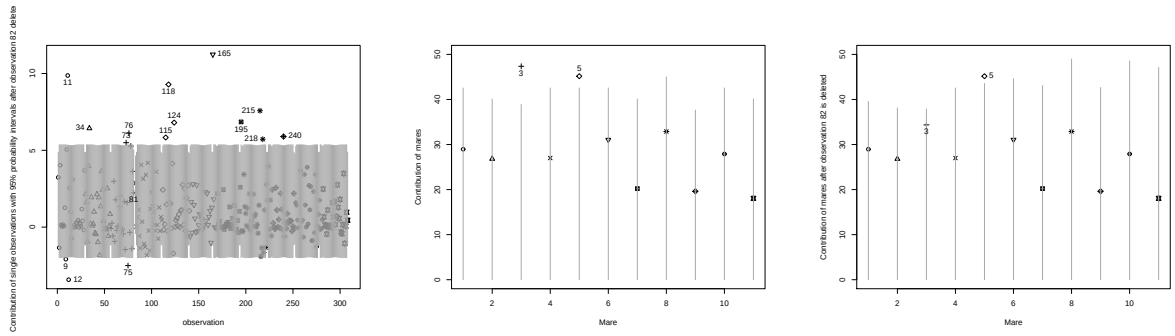
Figure 6 shows that observations 5, 7 and 20 have high contributions. But marginal residual detects observations 5 and 7 only. Conditional residual detects observations 5 and 20. So, if we study only with marginal or conditional residual, we miss at least one abnormal point. Plot (d) of the figure shows that contribition of each country falls well within the upper 95% probability limit.

Figure 7 shows how the unusual observations affect the dataset. We first deleted point 5, and then points 5 and 20 together, to look at their influence. Like the simulated data, here also we compared the deletion diagnostic by re-estimating the variance matrix with the diagnostic by not re-estimating the variance matrix, as shown in plots (a) to (d) of figure 7. They show that deleting unusual observations one by one does not affect other unusual observations. For example, deleting observation 5 still makes the observations 7 and 20 as unusual. That is because of the block diagonal structure of the variance matrix. Since the blocks of observations are independent of one another, and 5, 7 and 20 belong to three different blocks, deleting one leaves the other two unaffected. Also, we observe that re-estimating the variance matrix after deletion of an observation has the same effect of not re-estimating the variance matrix. So, for this LME model, deleting unusual observations would not lead to much changes in the variance matrix.

Corn data showed that no block in the dataset is unusual. One may naturally ask that, if some block has significantly high contribution, how to detect the observations within that block that are mainly responsible for this unusual behaviour? Also, if a block is an outlier, can we identify the points within that block that are really outliers, and the points that behave abnormally due to 'swamping'? A study of the Ovary data answers that question.

4.2. Ovary data

Pierson and Ginther (1987) reported on a study of the number of large ovarian follicles detected in 11 mares at several times in their estrus cycles. The data has three sets of observations. The first set is the mares with ordered factor. The second set is time in the estrus cycle. The data were recorded daily from 3 days before ovulation until 3 days after the next ovulation. The measurement times for each mare are scaled so that the ovulations for each mare occur at times 0 and 1. The third set represents the number of ovarian follicles greater than 10 mm in diameter. There are overall 308 observations for these 11 mares.



(a) (b) (c)

Figure 9: An analysis of Ovary data. (a) shows contribution of the mares, (b) shows contribution of single observations after observation 82 is deleted, (c) shows contribution of mares after observation 82 is deleted. From plot (a) it comes out that mares 3 and 5 are unusual, having high contribution. Probing into mare 3, it comes out that observation 82, the point with highest contribution in the mare, is the most influential behind the mare's unusual behaviour. When the point is deleted, the contribution of mare 3 comes down to a normal level, as apparent in plot (c).

The underlying model, as proposed by Pinheiro and Bates (2000) is an AR(1) process for each mare with the fitted trend as

$$Y = \beta_0 + \text{Sin}(2\pi X)\beta_1 + \text{Cos}(2\pi X)\beta_2 + \varepsilon$$

where Y is the number of follicles and X is time in the estrus cycle. The errors are Gaussian and are assumed to be uncorrelated between the mares.

In our analysis we took the mares as the blocks. Figure 8 presents an analysis of the 308 observations. It shows that there are 17 observations having high magnitude of contribution, falling outside the probability intervals. The numbers of observations having high marginal residuals and high conditional residuals are 14 and 15, respectively. So, here also contributions detect more observations than marginal and conditional residuals. For example, observation 81 has marginal and conditional residual falling well inside the respective probability intervals. But it has high negative contribution. Observations 82, 118 and 165 have significantly large positive contributions.

We do the next step of analysis which we call 'probing' into a block. Since mare 3 is unusual, we probe into the observations belonging to mare 3 in order to find out which observations influence the mare. Plot (a) of figure 8 shows that observation 82, belonging to Mare 3, has the highest contribution within that mare. Plot (a) of figure 9 shows that mares 3 and 5 are unusual, having high contribution. In order to look at how much the observation is responsible for the high contribution of mare 3, we delete point 82 and re-analyse the other points belonging to that mare. Plots (b) and (c) of figure 9 present the contributions of single observations as well as the blocks after deleting observation 82. It shows an interesting picture regarding point 81. Initially plot (a) of figure 8 showed that point 81 has high negative contribution, and apparently it was an unusual observation. But after deletion of point 82, observation 81 shows a nominal contribution, lying well inside the probability interval. That means the apparent abnormality of point 81 was caused by the extreme abnormal influence of the neighbouring point 82. This is an example of 'swamping' as discussed earlier.

Point 82 is observed to be the deciding factor behind the abnormal behaviour of mare 3. If the point is removed, contribution of mare 3 decreases by a large extent. Mare 5 also has high contribution. If we probe into mare 5, we shall observe that removing observation 118, the point with the highest contribution within mare 5, can make the mare behaving normally.

5. Conclusion

The proposition by Haslett and Hayes (1998) that contributions can be used for detecting unusual observations has merits and demerits. The authors decomposed the total lack of fit into the sum of contributions,

and showed that it involves two kinds of complimentary residuals that arise in the analysis of linear models. Their proposition is furthered by the present paper by showing that in a dataset, the joint deviation of an observation can be decomposed in three parts. Deviation due to global characteristic of the residual, deviation due to local characteristic of the residual, and deviation due to the joint effect of both these characteristics. Contribution measures the third part. In that sense, contribution is more important than marginal and conditional residuals because it takes into account of both kinds of deviations, and hence can detect more unusual observations in a dataset. We illustrated this feature for the detection of additive and innovative outliers and a patch of additive outliers in a time series model.

However, the significant demerit lies in the fact that contributions can be negative. A negative contribution is difficult to interpret as a measure of deviation. Secondly, when one of the marginal and conditional residuals is 0 and the other is large, contribution becomes 0, even when the observation shows one type of high residual. That way, the use of contributions can be misleading and should be carefully treated.

Re-estimation of the variance parameters after deletion of a number of observations is another issue related to the use of contributions. We have illustrated by data analysis that assuming the variance matrix fixed, or known, can lead to false identification of a normal observation as outlier. So, it is highly recommendable to estimate the variance parameters after each stage of deletion.

References

Barnett, V., Lewis, T., 1978. Outliers in Statistical Data. Chichester:Wiley.

Battese, G., R.M., H., Fuller, M., 1988. An error components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association 83, 28–36.

Bhattacharya, S., Murphy, T.B., Haslett, J., 2011. A Derivation of Quantiles of the Weighted Difference of Two Independent Chi Squared Random Variables using the Confluent Hypergeometric Function. Technical Report. University College Dublin. Dublin 4, Ireland.

Christensen, R., Pearson, L., Johnson, W., 1992. Case deletion diagnostics for mixed models. Technometrics 34, 38–45.

Dillane, D.M., 2004. Deletion Diagnostics for the Linear Mixed Model. Ph.D. thesis. Trinity College Dublin.

Fox, A., 1972. Outliers in time series. Journal of the Royal Statistical Society Series B 34, 350–363.

Haslett, J., Dillane, D., 2004. Application of delete = replace to deletion diagnostics for variance component estimation in the linear mixed model. Journal of the Royal Statistical Society Series B 66, 131–144.

Haslett, J., Haslett, S.J., 2007. The three basic types of residuals for a linear model. International Statistical Review 75, 1–24.

Haslett, J., Hayes, K., 1998. Residuals for the linear model with general covariance structure. Journal of the Royal Statistical Society Series B (Statistical Methodology) 60, 201–215.

Haslett, J., Hayes, K., 1999. Simplifying general least squares. American Statistician 53, 376–381.

Pierson, R.A., Ginther, O.J., 1987. Follicular population dynamics during the estrus cycle of the mare. Animal Reproduction Science 14, 219–231.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York.

Press, S.J., 1966. Linear combinations of non-central chi-square variates. Annals of Mathematical Statistics 37, 480–487.

Sullivan, C., 2001. A Diagnostic for the General Linear Model: an Application to Time Series. Ph.D. thesis. Trinity College Dublin.

Appendix

Subsets deletion for components of variance in the Linear Mixed Model

The Linear Mixed Model (LMM) may be defined as

$$Y \sim N(X\beta + Z\gamma, V)$$

where X and $Z = (Z_1, Z_2, ..., Z_r)$ are known matrices, β is a vector of fixed effects and γ is a vector of random effects with $E(\gamma) = 0, \mathrm{Cov}(\gamma) = D$ and $\mathrm{Cov}(\gamma, \varepsilon) = 0$. V will then be of the form $ZDZ^T + A$ where Z^T is the transpose of Z and $A = \mathrm{Var}(\varepsilon)$.

The above LMM can be equivalently expressed as

$$Y = W\mu + \varepsilon$$

where $V = Var(\varepsilon) = \sum_{j=0}^{r} \sigma_j^2 Z_j Z_j^T$. Let $\sigma = (\sigma_0^2, \sigma_1^2, ..., \sigma_r^2)^T$. Our purpose is to estimate $\tilde{\sigma}_{(a_i)}$ following deletion of subsets $Y_{(a_i)}(i = 1, ..., k)$. Christensen et al. (1992) proposed a method based on REML for the re-estimation, but Haslett and Dillane (2004) have provided an easier approximation method.

We drop the suffix i for notational simplicity. Haslett and Dillane (2004) proposed the approximation of $\tilde{\sigma}_{(a)}$ by the following recursive equation:

$$\tilde{\sigma}_{(a)} \approx T^{-1} \tilde{s}_{(a)}$$

where the (i, j)th element of the matrix T is defined as $t_{ij} = \mathrm{tr}(QZ_j Z_j^T QZ_i Z_i^T)$ and $\tilde{s}_{(a)}$ is a vector following the deletion of subset a which has the j-th element as $\tilde{s}_{j(a)} = \mathrm{tr}\left(QZ_j Z_j^T Q \mathrm{Var}_{\tilde{\sigma}_{(a)}}(Y)\right)$. The matrix T is available from the full fit and is thus already available. The above equation leads to an iterative solution of $\tilde{\sigma}_{(a)}$.

The same form of approximation holds for a time series model, but some details might get changed. For example, for an MA process, the exact method as described above can be applied, while for an AR process minor changes need to be made for the approximation of the correlation parameter. Dillane (2004) provides a detailed analysis of reestimating variance parameters for an AR process.