Statistics
in Medicine

Special Issue Paper

Received 15 November 2010, Accepted 11 July 2011 Published online 8 September 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4356

Identifying influential observations in Bayesian models by using Markov chain Monte Carlo

Dan Jackson,[a*†] Ian R. White[a] and James Carpenter[b]

In statistical modelling, it is often important to know how much parameter estimates are influenced by particular observations. An attractive approach is to re-estimate the parameters with each observation deleted in turn, but this is computationally demanding when fitting models by using Markov chain Monte Carlo (MCMC), as obtaining complete sample estimates is often in itself a very time-consuming task. Here we propose two efficient ways to approximate the case-deleted estimates by using output from MCMC estimation. Our first proposal, which directly approximates the usual influence statistics in maximum likelihood analyses of generalised linear models (GLMs), is easy to implement and avoids any further evaluation of the likelihood. Hence, unlike the existing alternatives, it does not become more computationally intensive as the model complexity increases. Our second proposal, which utilises model perturbations, also has this advantage and does not require the form of the GLM to be specified. We show how our two proposed methods are related and evaluate them against the existing method of importance sampling and case deletion in a logistic regression analysis with missing covariates. We also provide practical advice for those implementing our procedures, so that they may be used in many situations where MCMC is used to fit statistical models. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: Bayesian methods; generalised linear models; influence; Markov chain Monte Carlo

1. Introduction

Measures of leverage, influence and residuals are well-established tools in data analysis. Influence refers to how sensitive inferences are to the presence of particular observations and has proved its usefulness in maximum likelihood and least squares analyses.

It is therefore of interest to explore efficient algorithms for measuring influence in Bayesian analyses. Our interest in this was motivated by examples such as the sudden unexpected death in infancy (SUDI) analysis described in detail in Section 2. Here we wish to fit a standard model and examine the usual diagnostics such as influence and residuals, but some observations have missing covariates. A natural way to incorporate these observations into the analysis is to posit a model for the missing data and use maximum likelihood, but this is computationally intensive in all but the simplest of situations. In particular, this type of approach becomes increasingly difficult as the dimension of the integrals with respect to the missing data and the parameter space over which the maximisation is performed become large.

In order to overcome these difficulties, Markov chain Monte Carlo (MCMC) [1] can be used. MCMC is a powerful but computationally intensive method for fitting Bayesian models. It has the nice feature of yielding a random sample from the posterior distribution of the parameters of interest, from which means and credible intervals can be readily calculated. Although relatively simple models present few difficulties, the iteration times for complex and hence more realistic models may be long. If the mixing is poor, many iterations can be required to obtain estimates to within an appropriate degree of Monte Carlo error. Numerical estimation of influence by directly removing observations is therefore extremely tedious at best.

[a] *MRC Biostatistics Unit, Cambridge, UK*
[b] *London School of Hygiene and Tropical Medicine, UK*
* *Correspondence to: Dan Jackson, MRC Biostatistics Unit, Cambridge, UK.*
† *E-mail: daniel.jackson@mrc-bsu.cam.ac.uk*

Here the simplest definition of influence will be taken as the difference between whole sample parameter estimates and case-deleted estimates, where each case-deleted estimate can be obtained by removing the observation in question and refitting the model [2]. Although alternative measures of influence are possible, such as distances between posterior distributions of parameters of interest and the corresponding case-deleted distributions [3], these are not developed further here.

An established way to obtain influence statistics by using MCMC is importance sampling [3, 4]. Here we take advantage of the relationship between the full data posterior distribution and case-deleted posterior distributions; their ratio is the likelihood of the removed observation. Hence, case-deleted posterior summaries can be obtained from weighted summaries of the sampled values by using the importance weight which is proportional to the reciprocal of the likelihood of the deleted observation evaluated at the simulated parameter values. Although this procedure is effective in many situations, it has its limitations. If the weights are very variable, then extremely large numbers of iterations may be needed; highly unusual and hence influential observations can provide surprisingly volatile weights. Hence, the problem is that, if deleting an observation causes a big change in an estimate, then the importance sampling influence statistic depends heavily on very few MCMC draws and sometimes only one. Perhaps the other main difficulty is that importance sampling requires the repeated computation of the observations' contributions to the likelihood. This becomes particularly computationally demanding as the model becomes more complex, and as a consequence importance sampling is unfeasible for very complicated models. These concerns provided additional motivation for our work which provides influence statistics that do not make any further likelihood-based computations and, in contrast to the alternatives, do not become more computationally intensive as the model complexity increases.

The routine evaluation of influence statistics was firmly established in the context of linear regression in Cook's seminal paper [5]. As emphasised by Cook, 'On the surface it might seem that any desirability this measure [a standardised measure of influence] has would be overshadowed by the computations necessary for the determination of $n+1$ regressions'. However, Cook shows how influence statistics can be obtained using only quantities calculated from the original regression involving the complete sample, and this also provides the foundation of approximate influence measures for the generalised linear model (GLM) introduced by Pregibon [6] and Williams [7]. Hence, approximate case-deleted estimates can be obtained for any GLM without the need to fit additional regressions. In particular, the case-deleted influence statistics for GLMs given by Williams provide the justification for the procedures for obtaining these using MCMC developed here.

Although the use of MCMC to fit models with missing data provides our motivation, our methods may be used much more generally. We do however require that the likelihood dominates the prior, so that Bayesian and likelihood-based analyses provide inferences that are in good numerical, but not necessarily philosophical, agreement. This is to ensure that the established likelihood-based influence statistics are similar to their Bayesian counterparts. Here we use models that either are a standard single-level GLM or comprise a collection of connected GLMs, so we can ignore the prior specification asymptotically. Many Bayesian models are specified in this way, and this type of approach is emphasised by the models typically fitted using WinBUGS [8] (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml). 'Vague' or 'uninformative' priors help to ensure that their role is negligible but, provided the sample size is sufficiently large and truly dogmatic priors are avoided, are not required by the methods and models that follow. We return to the additional issues presented by more complicated models, such as hierarchical and mixed effects models, in the discussion.

The paper is set out as follows. In Section 2, we analyse the SUDI data. In Section 3, we review existing classical/frequentist methods for obtaining influence statistics for GLMs that lead us directly to our first proposal for obtaining these. In Section 4, we develop a procedure for obtaining influences using perturbed regression parameters, which, unlike our first proposal, does not require the form of the GLM to be specified. In Section 5, we describe the results from a simulation study, and in Section 6, we explain why our methods apply to models that comprise a collection of connected GLMs. In Section 7, we apply our methods to our example and compare the results with importance sampling and direct case deletion. We conclude in Section 8 with a discussion.

2. Sudden unexpected death in infancy

We will examine an illustrative analysis of a case control study exploring risk factors for SUDI in families with a previous SUDI [9]. We have information for 137 Care of Next Infants (CONI) infants, of whom

33 are cases. This was analysed using logistic regression with key matching variables as covariates. WinBUGS [8], via the R package BRugs, was used to perform all MCMCs.

We found four variables to be especially good predictors of death in this sample by using a logistic model: Income Deprivation Affecting Children Index (DEP; a low score indicates high deprivation), mother's age in years (MAGE), birthweight in grams (BWT) and the number of previous terminations and miscarriages (TERMIN). All covariates were centred, and continuous variables were standardised, so we fit the model

$$\text{logit}(p(\text{death})) = \beta_1 + \beta_2(\text{DEP} - 9500)/9000 + \beta_3(\text{MAGE} - 29)/6 + \tag{1}$$

$$\beta_4(\text{BWT} - 3000)/700 + \beta_5(\text{TERMIN} - 0.9).$$

From a standard complete case maximum likelihood analysis, we obtain the following (standard errors in parentheses): $\hat{\beta}_1 = -2.51$ (0.55), $\hat{\beta}_2 = -1.22$ (0.47), $\hat{\beta}_3 = -0.69$ (0.29), $\hat{\beta}_4 = -1.09$ (0.33) and $\hat{\beta}_5 = -1.80$ (0.58).

This complete case analysis discards 15% of the data: 21 observations have missing TERMIN values, four of which also have missing MAGE; DEP and BWT are complete. To include the incomplete cases, we fit model (1) jointly with models for the incomplete covariates. Specifically, we use a linear model for MAGE conditional on DEP and BWT and a log linear Poisson model for TERMIN on DEP, BWT and MAGE. We also use WinBUGS' truncation command to truncate TERMIN at 20. Without this, a few of the simulated TERMIN values can be unrealistically large, which has unfortunate implications for the model fit.

We place uniform priors on all regression parameters and a Gamma (0.001,0.001) prior on the precision of the variance in the linear model for MAGE. We used a burn in of 10^3 iterations and a further 10^5 iterations to make inferences. This MCMC took around 10 min on a Windows Terminals Server and provided $\hat{\beta}_1 = -2.62$ (0.54), $\hat{\beta}_2 = -1.10$ (0.42), $\hat{\beta}_3 = -0.54$ (0.28), $\hat{\beta}_4 = -1.09$ (0.33) and $\hat{\beta}_5 = -1.91$ (0.60), where the estimates are given by the simulated sample means. This analysis adds further weight to the conclusion that the four variables considered are good predictors of death. We also obtained the correlation matrix of these estimates and therefore $\text{Var}(\hat{\beta})$. We use the notation $\hat{\beta}$ for both Bayesian and maximum likelihood estimates because we assume that they are in good numerical agreement. In particular, we use the hat notation to denote the posterior mean in all Bayesian analyses that follow, where we obtain these quantities as the sample mean of the simulated values from the MCMC.

This analysis has successfully incorporated all data but makes it difficult to assess issues such as influence. With the additional assumptions made about the missing data, coupled with the fact that there is a large amount of variation in the observed covariates, there is the natural concern that a few unusual observations might be driving the inferences. We now turn our attention to methods for deriving influences using MCMC output and begin by considering the simpler case where there is just a single GLM and no missing data.

3. Case-deleted estimates for generalised linear models

We condition inference on X, an n by p matrix of explanatory variables. The observations, denoted by y, are an n by 1 vector of conditionally independent responses. We assume for now that the data are complete but return to the possibility of missing data in Sections 6 and 7. Denote $\mu_i = E[Y_i|X_i]$, where μ_i depends on the ith row of X, X_i^T, through the linear predictor $\eta_i = g(\mu_i) = X_i^T\beta$, where $g(\cdot)$ denotes the link function and β is the p by 1 vector of regression parameters. We assume that Y_i is a member of the exponential family with variance v_i and dispersion parameter ϕ. Fitting the model in a Bayesian framework, and assuming that the likelihood dominates the prior, implies that the posterior means, medians and modes will all be close to the maximum likelihood point estimates, obtained using iterated weighted least squares [10, 11], with agreement as the sample size tends to infinity.

Williams [7] provides an approximation to the ith case-deleted maximum likelihood estimate, $\hat{\beta}_{(i)}$, obtained by using the complete sample estimate $\hat{\beta}$ as an initial value and one step of the weighted least squares algorithm:

$$\hat{\beta}_{(i)} \approx \hat{\beta} - \sqrt{\frac{w_i}{v_i}}(1-h_i)^{-1}(X^TWX)^{-1}X_ir_i \tag{2}$$

where $w_i = v_i^{-1}(\partial\mu_i/\partial\eta_i)^2$, $W = \mathrm{diag}(w_i)$, r_i denotes the residual $(y_i - \mu_i)$ and h_i is the ith diagonal element of the 'hat' matrix $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$; all terms on the right-hand side are evaluated at the complete sample estimates. Let θ_i denote the canonical parameter for the regression. The density of y_i is $\exp((b(\theta_i)+y_i\theta_i)/\phi + c(y_i,\phi))$, where $\mu_i = -b'(\theta_i)$ and $v_i = -\phi b''(\theta_i)$. Furthermore, the matrix $\mathrm{Var}(\hat\beta)$ is approximately $(X^TWX)^{-1}$.

We suggest omitting the term involving h_i in (2). The sum of the positive h_i equals the number of regression parameters p, and we assume that the sample size n is much greater than p. Hence, although the relative magnitudes of h_i play an important role in determining the observations' leverages, they are much less important when obtaining influence statistics using (2); $(1-h_i) \approx 1$ for all i in large samples with no grossly outlying values. With this simplification and substitution of $\partial\mu_i/\partial\eta_i = \partial\theta_i/\partial\eta_i \times \partial\mu_i/\partial\theta_i$, where $\partial\mu_i/\partial\theta_i = v_i/\phi$, Williams' formula (2) becomes

$$\hat\beta_{(i)} \approx \hat\beta - \mathrm{Var}(\hat\beta)\frac{\partial\theta_i}{\partial\eta_i}\frac{X_i r_i}{\phi}. \tag{3}$$

This further approximation is useful as we routinely obtain $\hat\beta$ and $\mathrm{Var}(\hat\beta)$ when fitting the model to the full dataset. In particular, for the SUDI data, we have already obtained $\hat\beta$ and $\mathrm{Var}(\hat\beta)$ as described and reported in Section 2.

Any procedure for obtaining the rest of the right-hand side of (3) may be used to provide approximate case-deleted estimates. In particular, we can obtain influence statistics, $\hat\beta - \hat\beta_{(i)}$, directly from (3). If the canonical link is used, $\partial\theta_i/\partial\eta_i = 1$ and (3) simplifies further.

3.1. Our first proposal

Assuming a particular GLM, and that the impact of the prior specification is negligible, we propose additionally defining and storing the simulated values of

$$\delta_i = \frac{\partial\theta_i}{\partial\eta_i}\frac{X_i r_i}{\phi}$$

when running the MCMC. The derivatives $\partial\theta_i/\partial\eta_i$ can be evaluated in terms of η_i and are unity if a canonical link has been adopted. Hence, the δ_i are easily computed in terms of the covariates, β and ϕ. Replacing δ_i with its estimate (the mean of the simulated values) from the MCMC in (3), and $\hat\beta$ and $\mathrm{Var}(\hat\beta)$ with theirs, immediately yields an approximate influence for the ith observation. All of the unobserved quantities that we define in this article are evaluated at every iterate of the MCMC, including derivatives, and so all of these vary between MCMC draws, but only the estimates $\hat\delta_i$, $\hat\beta$ and $\mathrm{Var}(\hat\beta)$ are used to compute influence statistics in (3) when using our first proposal. An assessment of the MC error for $\hat\delta_i$ is equally important, but just as straightforward, as for $\hat\beta$ when obtaining influence statistics in this way.

4. Obtaining influence statistics using perturbations

Our first proposal avoids any further evaluation of the likelihood and so can be expected to be more computationally efficient than importance sampling but requires the form of the GLM to be known. In this section, we propose a second method based on perturbing the regression parameters. Although we use the theory of GLMs to justify the procedure, ultimately we use just the variance/covariance matrix of the complete sample point estimates and the posterior means of the perturbations.

We will define an n by p random matrix ϵ of 'perturbations' and replace the jth regression coefficient in the model, for the ith observation, by $\beta_{ij} = \beta_j + \epsilon_{ij}$. The entries ϵ_{ij} are to be given independent (of each other, β and ϕ) normal priors with mean 0. Otherwise, the regression is to be fitted using MCMC in the usual way. This essentially introduces a random effect component (over observations) on the regression parameters. By making the prior precisions of the random perturbations extremely large, we obtain approximately the same estimates $\hat\beta$ and $\mathrm{Var}(\hat\beta)$ as in the usual perturbation-free regression model. We use the quantities σ_{ij} to denote the prior standard deviations of ϵ_{ij}.

We next show in the case of a GLM that the posterior distributions of the ϵ_{ij} relate to influential outcomes in the model. The intuition is that, if an observation is directly influential for a particular parameter, then the corresponding posterior ϵ_{ij} distribution will move further from zero than its less influential counterparts, reflecting the 'pull' or influence that this observation exerts on the fitted model.

4.1. *The mathematical consequences of introducing the perturbations*

Let ϵ_i^T denote the ith row of the perturbation matrix. With the assumption that each of the n observations, conditional on all ϵ_{ij}, β and ϕ, are independent (equivalent to assuming that they are exchangeable in the original model),

$$P(\epsilon_1, \epsilon_2, \cdots, \epsilon_n, \beta, \phi | y) \propto P(\beta, \phi) \prod_{i=1}^{n} \{P(\epsilon_i) P(y_i | \epsilon_i, \beta, \phi)\} \tag{4}$$

Note that we include the prior $P(\beta, \phi)$ for completeness, but we assume that its role is negligible in practice. We show in the Appendix that, provided the prior perturbation variances (denoted by σ_{ij}^2) are made sufficiently small, the posterior mean of ϵ_{ij} is given by $\hat{\epsilon}_{ij}$, where

$$\frac{\hat{\epsilon}_{ij}}{\sigma_{ij}^2} \approx \frac{\partial\theta_i}{\partial\eta_i} \frac{X_{ij}(r_i - s_i)}{\phi}, \tag{5}$$

$$s_i = \frac{\partial\theta_i}{\partial\eta_i} \frac{X_i^T \epsilon_i v_i}{\phi},$$

X_{ij} denotes the jth entry of X_i^T, and all quantities are on the right-hand side of (5) are evaluated at their posterior modes.

4.2. *An approximation*

We assume that the magnitude of the posterior mode of s_i is small in relation to that of r_i. We can check this, provided that the form of the GLM is known, as shown for our example in Section 7. As $\sigma_{ij}^2 \to 0$, $s_i \to 0$, and we can ignore s_i. Hence,

$$\frac{\hat{\epsilon}_{ij}}{\sigma_{ij}^2} \approx \delta_{ij} \tag{6}$$

where δ_{ij} is the jth entry in δ_i and is evaluated at full sample estimates in exactly the same way as in Williams' approximation for case-deleted estimates. For models where the variance v_i is unbounded, extra care must be taken to ensure that s_i is small in relation to the residuals because s_i is linear in v_i. For example, for datasets with large Poisson counts, very small σ_{ij}^2 may be required to ensure that we can ignore s_i in this way.

We can therefore obtain the term $\hat{\delta}_i \approx \{\partial\theta_i/\partial\eta_i\}(X_i r_i/\phi)$ using MCMC by adding the perturbations and monitoring the simulated ϵ_i or equivalently by defining the random variables $\delta_{ij} = \epsilon_{ij}/\sigma_{ij}^2$ and monitoring the δ_i. With the combination of (6) and (3), the case-deleted estimates are

$$\hat{\beta}_{(i)} \approx \hat{\beta} - \text{Var}(\hat{\beta})\hat{\delta}_i$$

from which we can obtain the influences, $\hat{\beta} - \hat{\beta}_{(i)}$.

4.3. *Implementation and Monte Carlo error*

A practical issue is that, although it is possible to implement this method in a single stage as presented above, this is computationally demanding and frequently results in 'trap errors' in WinBUGS. We suggest a two-stage procedure that avoids these difficulties in practice. In the first stage, the original model is fitted without perturbations in the usual way in order to obtain $\hat{\beta}$, $\hat{\phi}$ and Var$(\hat{\beta})$.

In the second stage, to simplify the MCMC algorithms, we suggest constraining β and ϕ to their estimates from the first stage and then introducing the perturbations. With β and ϕ fixed in this way, the posterior for the ϵ vectors becomes (4) with $P(\beta, \phi) = 1$ and $P(y_i|\epsilon_i, \beta, \phi)$ replaced by $P(y_i|\epsilon_i, \hat{\beta}, \hat{\phi})$, and a similar $\hat{\delta}_i$ is obtained. This simplification is made purely for computational convenience because suitable $\hat{\delta}_i$ are evaluated regardless of whether a one-stage or a two-stage procedure is adopted.

Now that β and ϕ are to be held fixed in this second stage, the only observation which is used to update the prior of ϵ_i, and therefore δ_i, is y_i. Hence, we can add the ϵ_i to a single observation (but to all regression coefficients) at a time, providing y_i as the only observation, and run n additional analyses. In practice, we found it convenient and computationally efficient to add the perturbations in this manner, so that $\sigma^2_{k,j} = 0$ for all $k \neq i$ when obtaining the ith influence statistic. It is then a very simple task to update the vector ϵ_i *en bloc* using a Gibbs' sampler, and, because there are no other random variables, we simulate directly from the posterior $\epsilon_i | y_i$ and the simulated ϵ_i, and hence δ_i are independent. The resulting MCMC mixes and converges extremely rapidly.

Monte Carlo error is inherent in the influences obtained using equation (6), which is a little more difficult to assess because of the indirect manner in which influence statistics are obtained when using the perturbations. The error in $\hat{\delta}_i$ is approximately normally distributed, centred at the origin with $\text{Var}(\hat{\delta}_i) \approx (m\text{diag}(\sigma_i^2))^{-1}$, where σ_i^2 is the vector of σ_{ij}^2 associated with observation i, and m is the number of iterations. This is because δ_i has a prior variance of $(\text{diag}(\sigma_i^2))^{-1}$, and in the two-stage approach, the simulated δ_i are independent as explained above; the posterior variance of δ_i is therefore very similar to its prior. Hence, the Monte Carlo error of the influence of the ith observation, obtained as $\text{Var}(\hat{\beta})\hat{\delta}_i$, is approximately $E_i = \text{Var}(\hat{\beta})(\text{diag}(\sigma_i^2)^{-1})\text{Var}(\hat{\beta})/m$, so that a good indication of this Monte Carlo error can easily be obtained. Small values of σ_{ij}^2, which improve the accuracy of the approximations, also increase Monte Carlo error of the influence statistics, and some kind of tradeoff is needed in practice. If the perturbations are made quite large in this Monte Carlo error/approximation tradeoff, then they do not affect the primary analysis because this is performed in the first 'perturbation free' stage. Large entries of $\text{Var}(\hat{\beta})$ also have an unfortunate implication for the Monte Carlo error, so we suggest centring data before fitting models to reduce the variance of intercept terms.

Although an advantage of this approach is that we merely require the model to be some (unspecified) GLM, the form of the GLM does have implications for how small the perturbations' variances have to be, which in turn has implications for m. In practice, it may be desirable to use a variety of small σ_{ij} and large m in order to determine if reliable influence statistics have been obtained. Alternatively, and if known, the properties of the GLM in question can be examined and suitable criterion chosen. Despite this, the difficulties associated with choosing appropriate values of σ_{ij} and m, so that both the approximations are accurate and the MC error is small, are a disadvantage of this method. Our first proposal may therefore be considered preferable in situations where the form of the GLM is known.

5. A simulation study

In order to assess the accuracy of the two proposed methods, we performed a simulation study. Here we simulated 100 datasets, each with $n = 100$ observations, by using the linear model $Y \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, 1)$, where $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_3 = 3$. Burn ins of 10^3 iterations were used. A further 10^5 iterations were used when implementing our first proposal, and $m = 10^6$ iterations were used when obtaining influences using the perturbations to ensure that these had stabilised.

By using normally distributed data, exact frequentist influence statistics [5] can be obtained which provide a 'gold standard' to compare our methods with. Because we used a linear model with a canonical link, our simulation study does not assess the accuracy of the approximations described in the Appendix; rather, it assesses how well our proposals work when these approximations are valid. Flat priors over very wide ranges were used for the β parameters, and a Gamma (0.001,0.001) prior was used for the precision of the variance.

Perturbations with $\sigma_{ij} = \min(0.05/|X_{ij}|, 1)$ were used. These values have been found to ensure as large perturbations as possible are used, without compromising the accuracy of any approximation, when a canonical link is used in either a linear (with variance of around 1 or greater) or logistic regression involving a moderate number of parameters. For a GLM with a canonical link, the approximations described in the Appendix only require that values of $X_i^T \epsilon_i$ are small and the posterior modes of the s_i are small in relation to the residuals r_i; $s_i = X_i^T \epsilon_i$ for a standard linear regression. For example, for $p = 6$ as in our main example, using $|X_{ij}|\sigma_{ij} = 0.05$ ensures that $X_i^T \epsilon_i \sim N(0, 0.015)$, so that probable values of $X_i^T \epsilon_i$ are small on the log odds scale. Numerical difficulties ('trap errors' in WinBUGS) can, however, be obtained if instead $\sigma_{ij} = 0.05/|X_{ij}|$ is used because this results in very large σ_{ij} if covariates are close to zero and hence potential numerical instability. We suggest considering the use of $\sigma_{ij} = \min(a/|X_{ij}|, b)$ in practice and have found $a = 0.05$ and $b = 1$ to be suitable for our examples.

By combining the influence statistics across all simulated observations and datasets, we have $100 \times 100 = 10\,000$ exact frequentist influence statistics for each regression parameter to compare our methods with. Both our proposed methods provided influence statistics that were in good agreement with these standard influence statistics. To give an indication of this, the coefficients of the least squares regression lines of the proposed influence statistics on the standard frequentist influence statistics are given for each parameter in Table I, where the correlations between the proposed and standard influence statistics are also shown. The correlations are given to five decimal places so that they are distinguishable. The gradients of the regression lines are all between 0.95 and 0.99 suggesting that, compared with the exact frequentist influences, the proposed methods have a slight tendency to understate the influence of observations in this simulation study, but the overall picture is that all three methods are in good agreement. This tendency can be explained by the omittance of the $(1 - h_i)$ terms in (2), so this will lessen in larger samples.

6. Obtaining influences for models comprising a collection of connected generalised linear models and in situations where there are missing data

The above analysis assumes a single GLM (with complete data), but as explained in the introduction, Bayesian models are commonly defined as a collection of connected GLMs. For example, the model for the SUDI data is such a collection of three GLMs, where the logistic regression of death on all the covariates is of real interest. There are also some missing data, which motivated our methods as we explained in the introduction.

Our methods are, however, immediately applicable to any GLM component within the full model. When implementing our first proposal (Section 3.1), we simply define $\delta_i = \{\partial\theta_i/\partial\eta_i\}(X_i r_i/\phi)$ for the GLM for which influences are desired and combine these with the resulting regression estimates, $\hat{\beta}$ and $\mathrm{Var}(\hat{\beta})$, for this same GLM. Neglecting any prior dependence between regression parameters that might have been specified, and if there is no missing data, the parameter estimates for each GLM are independent, and hence, the standard theory described in Section 3 ensures that influence statistics are obtained from our methods.

There is, however, the issue of missing covariates, which is ignored in the above argument. This is in fact of little concern provided the fraction of missing data is not too severe because our procedures average the influences over the posterior distributions of any missing data, and hence, suitable influences are obtained. This way of handling missing data may be considered an advantage of our methods. We may therefore handle missing data by entering them as 'NA' in the usual way when using WinBUGS, for example.

The perturbations are justified as they obtain δ_i in a different way and hence are also appropriate. Here we constrain all parameters to their estimated values after fitting the model comprising a collection of connected GLMs in the first stage and then add the perturbations to the GLM in question and monitor the simulated δ_i in the same manner as before.

7. Influence statistics in the sudden unexpected death in infancy analysis

We used the proposed methods, and the alternatives, to estimate the influences in the SUDI analysis. Our first proposal is suitable because the form of the model of interest (the logistic regression for the probability of death) is known, but our second proposal will also be used so that the results can be compared.

There are some missing covariates, but otherwise model (1) is a standard logistic regression. We therefore also monitored $s_i \approx p_i(\text{death})(1 - p_i(\text{death}))X_i^T \epsilon_i$ when adding the corresponding perturbation to

Table I. Some results from the simulation study: c_1 and m_1 denote the intercepts and gradients of the least squares regression lines of influence statistics from our first proposal (Section 3.1) on the frequentist influence statistics; ρ_1 denotes the correlation between these influence statistics. c_2, m_2 and ρ_2 denote these same quantities for influences obtained using perturbations.

Parameter	c_1	m_1	ρ_1	c_2	m_2	ρ_2
β_0	0.0000	0.9832	0.99979	0.0002	0.9761	0.99978
β_1	0.0000	0.9648	0.99964	0.0000	0.9578	0.99963
β_2	0.0000	0.9726	0.99962	−0.0001	0.9655	0.99957

ensure that the role of s_i is negligible in (5). The average absolute posterior mean of $\hat{s}_i$ was just 0.0006, and its maximum was 0.0024. These are small values in relation to the residuals resulting from logistic regression, and we are reassured that approximation (6) is accurate. A burn in of 10^3 iterations was used. A further 10^5 iterations were used to obtain influences using our first proposal, and the influences from the perturbations-based method appeared to stabilise (small MC error) using $m = 10^6$. Perturbations with $|X_{ij}|\sigma_{ij} = 0.05$ were initially used, but some numerical difficulties were encountered for the three data points where the mother's age was 29 years; this corresponds to a covariate of zero in the regression as written above and, hence, with the decision to use $|X_{ij}|\sigma_{ij} = 0.05$, an infinite σ_{ij}. In fact, covariates were centred at exact means rather than the approximate ones given above, but the three very large σ_{ij} arising from this strategy presented problems. Following the procedure used in the simulation study to avoid these problems, σ_{ij} was therefore reduced to unity when used in conjunction with the mother's age covariate for these three observations.

7.1. Comparison of methods

A gold standard influence analysis was performed by removing each of the 137 observations in turn. Only 20 000 iterations were used to obtain estimates for each case-deleted sample to reduce the time taken (to around 5 h). The resulting case-deleted influences were standardised (divided by the standard error of the corresponding complete sample parameter estimate) and are shown as solid points in Figure 1 for the first 20 observations, where standardised influences for β_2 and β_3 are shown in the top two panels and those for β_4 and β_5 are shown in the bottom two panels. The 95% intervals describing Monte Carlo uncertainty, from normal approximations, are only slightly wider than the solid points which can thus be taken to be accurate. Hollow circles show the corresponding estimates obtained by importance sampling [3, 4]; triangles show these from our first proposal (Section 3.1), and diamonds show influences obtained using perturbations (our second proposal).

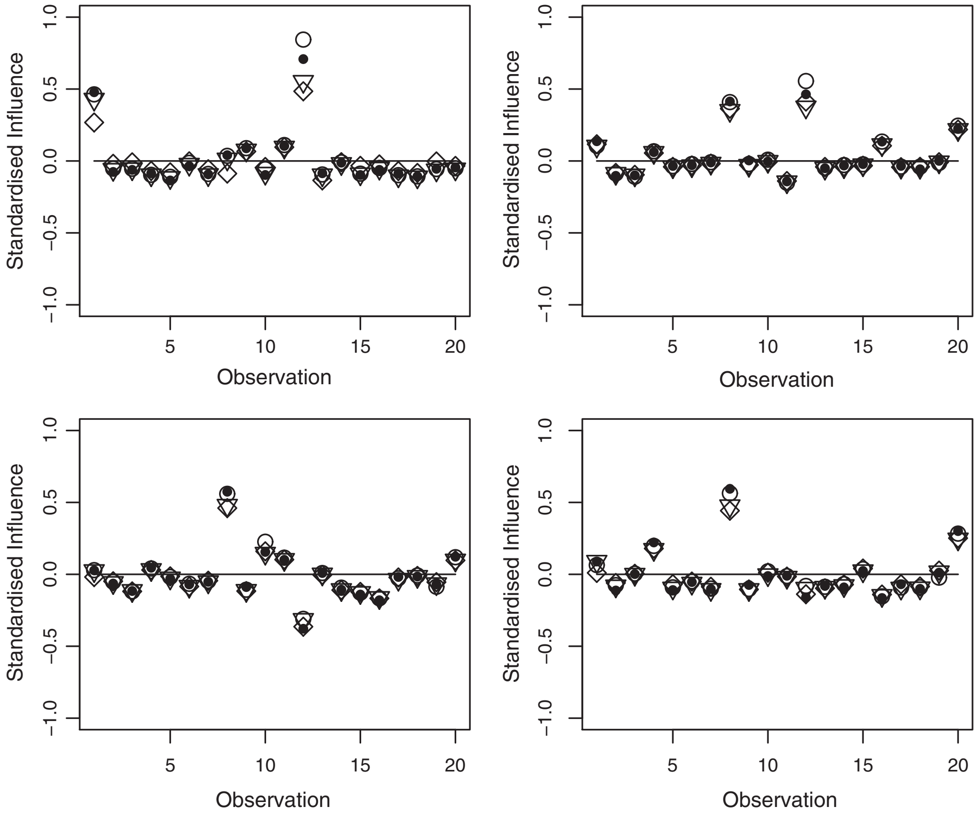

Figure 1. Standardised influence statistics for the first 20 observations. The top two panels show influences for β_2 and β_3, whilst the bottom two panels show these for β_4 and β_5. Solid points show the influence statistics obtained by case deletion; hollow circles show the corresponding estimates obtained by importance sampling; triangles show these using our first proposal (Section 3.1); and diamonds show influences obtained using perturbations.

All three approximations perform least well for the more influential observations but are generally effective in obtaining influences. It is to be expected that large influences are obtained with the least precision, as these provide the biggest challenge to Williams' approximation (which two of the methods are based on) and the most volatile weights for importance sampling. Very similar results were obtained for the remaining 117 observations. Note in particular that all methods successfully detected the influential observation 12 for β_2. It is not surprising that observation 12 is influential; this has the highest DEP score of the cases and one of the oldest MAGE values. Cook [5] suggests, in the context of linear regression, that if removing an observation results in an estimate at the edge of 50% of the whole sample confidence region, then this 'may be cause for concern'. Observation 12 can therefore be interpreted as being worryingly influential for β_2 but not excessively so.

Because all the procedures were successful, it is perhaps the additional computational burden required to obtain influences that determines which is to be recommended in practice. Importance sampling doubled the overall simulation time (from around 10 to 20 min). Using our first proposal only required an extra 5 min. Even for this relatively simple example, our first proposal was much faster than importance sampling. This relative efficiency will increase as the model complexity, and hence the difficulties associated with evaluating the observations' contributions to the likelihood, also increases.

The perturbations-based method required around 30 s for each observation, so that around an hour of computing time was needed to obtain influence statistics. Nevertheless, this method has the advantage that computation time does not increase with model complexity in the way it does with importance sampling and may prove useful for models where the precise form of the GLM cannot be specified. Indicative WinBUGS code for performing the analysis is available from the first author on request.

8. Discussion

This paper has developed novel ways of approximating established influence statistics, as used in classical methods, so that they can be used in the context of Bayesian analyses using MCMC. The two new approaches involve storing additional quantities and then use standard MCMC output to recover Williams' influence statistic. The type of analysis performed provided the motivation for this work, which can be applied much more generally. We have not, however, attempted to answer the much more difficult question of whether or not removing particular observations changes the conclusions qualitatively. Furthermore, we have not asked the even more difficult question of what we should do if we discover that an observation is very influential.

Direct case deletion, although easily implemented, is not a feasible option unless models can be fitted extremely quickly; even for the relatively simple model considered here, with just over a hundred observations, this took several hours. Importance sampling is a generally viable but fairly computationally intensive method for the type of models we have considered but becomes computationally prohibitive as the model becomes more complex and the repeated computation of the likelihood becomes less feasible. Even for our relatively simple example, the computational advantages of our first proposal are apparent. If any of the procedures that avoid direct case deletion highlight extremely influential observations, and an accurate indication of their influence is desired, it is necessary to remove the offending observations and refit the model. This is because all the approximate influences are less accurate for more influential observations. All that is typically needed in practice, however, is the identification of influential observations, and an indication of the extent of this influence and our procedures can be used for these purposes.

We have assumed that the model is a collection of connected single-level GLMs. Our methods might also be considered when the part of the model of interest is either known to be such a GLM or, at the very least, thought to behave similarly to one when using the perturbations. Bayesian Hierarchical models, where unobserved random effects present further issues [12], do not fit directly into our framework, however. Despite this, Hodges' [13] proposals for the assessment of case influence for hierarchical models, in his Section 4.4, bear some similarities to ours. Extensions or variations of our methods might be especially useful in this context and may form the subject of future work.

In particular, because the perturbations have the intuition that influential observations will exert more 'pull', it would be especially interesting to see if these are also useful for these and other types of statistical model. A difficulty in using the perturbations in practice is that we must ensure that all the various approximations used are appropriate. For models where the variance is unbounded, extra care must be taken to ensure that s_i is small in relation to the residuals because this is linear in v_i. Furthermore, if a canonical link has not been used, we have a further criterion that must be satisfied in terms of the derivatives of the canonical parameter and the linear predictor.

In conclusion, we have proposed two new methods for estimating the influence of observations when fitting models by using MCMC. We have shown how these are related and given practical guidance. Both proposals have been shown, in an example, to give accurate measures of influence. Both proposals avoid any further computation of the likelihood and so can be used in some situations where the complexity of the model renders case deletion and importance sampling unfeasible.

Acknowledgement

The authors thank Shaun Seaman for his helpful comments on the use and implementation of importance sampling. We also thank Bob Carpenter for comments and for providing the data. D. J. and I. R. W. were supported by MRC grant U.1052.00.006. J. C. was supported by grant RES 063-27-0257.

Appendix

We assume a GLM so that the observations y_i follow distributions from the natural exponential family with canonical parameter θ_i, so $P(y_i|\theta_i) \propto \exp((b(\theta_i) + y\theta_i)/\phi)$. With the perturbations and an arbitrary link function, we have $\theta_i = f(X_i^T\beta + X_i^T\epsilon_i)$, where $f(\cdot)$ is a strictly increasing function; if this is the identity function, then we have assumed the canonical link. We also have the standard results

$$\frac{\partial}{\partial\alpha}b(\alpha) = -E[Y_i|\theta_i = \alpha] = -\mu_i. \tag{7}$$

and

$$\frac{\partial^2}{\partial\alpha^2}b(\alpha) = -\frac{\partial}{\partial\alpha}E[Y_i|\theta_i = \alpha] = -\frac{1}{\phi}\mathrm{Var}[Y_i|\theta_i = \alpha] = -\frac{v_i}{\phi}. \tag{8}$$

Using independent normal priors for the ϵ_{ij}, and noting that all ϵ prior distributions are independent, gives

$$P(\epsilon_i) \propto \prod_{j=1}^{p} \exp(-\epsilon_{ij}^2/2\sigma_{ij}^2) \tag{9}$$

The distributional assumptions of the GLM, with the perturbations introduced, gives

$$P(y_i|\epsilon_i, \beta, \phi) \propto \exp((b(f(X_i^T\beta + X_i^T\epsilon_i)) + y_i f(X_i^T\beta + X_i^T\epsilon_i))/\phi). \tag{10}$$

The product of (9) and (10) gives the posterior density of ϵ_i to within a constant of proportionality. Taking the logarithm of the resulting density and differentiating with respect to ϵ_{ij} gives

$$\frac{\partial}{\partial\epsilon_{ij}}\log(P(\epsilon_{ij}|y)) = -\epsilon_{ij}/\sigma_{ij}^2 + X_{ij}f'(X_i^T\beta + X_i^T\epsilon_i)(y_i + b'(f(X_i^T\beta + X_i^T\epsilon_i)))/\phi \tag{11}$$

where X_{ij} denotes the jth entry of X_i^T; $f'(\cdot) = \partial\theta_i/\partial\eta_i$ and $b'(\cdot)$ are the derivatives of functions $f(\cdot)$ and $b(\cdot)$. In order to obtain the posterior distribution of ϵ_{ij}, a series of approximations are required.

Approximation 1

We use the linear approximation $f'(X_i^T\beta + X_i^T\epsilon_i) \approx f'(X_i^T\beta) + X_i^T\epsilon_i f''(X_i^T\beta)$ and note that this is exact if a canonical link is used and valid as $\sigma_{ij}^2 \to 0$ for all $j = 1, \cdots, p$. Hence, $f'(X_i^T\beta + X_i^T\epsilon_i) \approx \partial\theta_i/\partial\eta_i + X_i^T\epsilon_i\partial^2\theta_i/\partial\eta_i^2$ where the partial derivatives are evaluated at $\eta_i = X_i^T\beta$. We further assume $\partial\theta_i/\partial\eta_i >> \mathrm{abs}(X_i^T\epsilon_i\partial^2\theta_i/\partial\eta_i^2)$. Again as $\sigma_{ij}^2 \to 0$, this criterion is satisfied, and this is exact if a canonical link is used. Assuming sufficiently small perturbations, we approximate $f'(X_i^T\beta + X_i^T\epsilon_i) \approx \partial\theta_i/\partial\eta_i$ in (11).

Approximation 2

We use two linear approximations, firstly for the function $f(\cdot)$, taken around $X_i^T\beta$, and then for $b^{'}(\cdot)$, taken around $f(X_i^T\beta)$, to approximate $b^{'}(f(X_i^T\beta + X_i^T\epsilon_i)) \approx b^{'}(f(X_i^T\beta)) + X_i^T\epsilon_i f^{'}(X_i^T\beta)b^{''}(f(X_i^T\beta))$. From (7) and (8), this is equivalent to $-(\mu_i + (X_i^T\epsilon_i\{\partial\theta_i/\partial\eta_i\}v_i)/\phi)$, where the partial derivative and the moments of Y_i are evaluated at $\eta_i = X_i^T\beta$. The first of these linear approximations requires small $X_i^T\epsilon_i$, and the second requires small $X_i^T\epsilon_i\partial\theta_i/\partial\eta_i$. For a canonical link, these requirements are the same, and the first linear approximation is exact. The second linear approximation is exact for a linear model and canonical link. All these approximations are justified for any GLM as $\sigma_{ij}^2 \to 0$, however.

An approximate derivative

Upon making approximations 1 and 2, (11) becomes

$$\frac{\partial}{\partial\epsilon_{ij}}\log(P(\epsilon_{ij}|y)) \approx -\frac{\epsilon_{ij}}{\sigma_{ij}^2} + \frac{X_{ij}\partial\theta_i/\partial\eta_i(r_i - s_i)}{\phi} \qquad (12)$$

where r_i denotes the residual $(y_i - \mu_i)$, and $s_i = (X_i^T\epsilon_i\{\partial\theta_i/\partial\eta_i\}v_i)/\phi$. We obtain the posterior mode by setting derivatives (12), $j = 1, \cdots, p$, to zero. Hence, denoting (for the moment) the posterior mode of ϵ_{ij} using the 'hat' notation, we obtain (5). A linear approximation for $b(\cdot)$ in (3) further shows that the posterior of the ϵ_i are also approximately normal. Hence, the 'hat' notation in (5) can be taken to refer to any conventional measure of location, so we take this to indicate the posterior mean.

References

1. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London, 2006.
2. Bradlow ET, Zaslavsky AM. Case influence analysis in Bayesian influence. *Journal of Computational and Graphical Statistics* 1997; **6**:314–331.
3. Weiss RE, Cho M. Bayesian marginal influence assessment. *Journal of Statistical Planning and Inference* 1998; **71**:163–177.
4. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman and Hall: New York, 2004.
5. Cook RD. Detection of influential observation in linear regression. *Technometrics* 1977; **19**:15–18.
6. Pregibon D. Logistic regression diagnostics. *The Annals of Statistics* 1981; **9**:705–724.
7. Williams DA. Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* 1987; **36**:181–191.
8. Lunn DJ, Thomas A, Best N, Spiegelhalter D. Winbugs a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics in Computing* 2000; **10**:325–337.
9. Carpenter RG, Waite A, Coombs RC, Daman-Willems C, McKenzie A, Huber J, Emery JL. Repeat sudden unexpected and unexplained infant deaths: natural or unnatural?. *Lancet* 2005; **365**:29–35.
10. Lee Y, Nelder JA, Pawitan Y. *Generalised Linear Models with Random Effects*. Chapman and Hall: London, 2006.
11. McCullagh P, Nelder JA. *Generalised Linear Models*. Chapman and Hall: London, 1999.
12. Liu J, Hodges JS. Posterior bimodality in the balanced one-way random-effects model. *Journal of the Royal Statistical Society, Series B* 2003; **65**:247–255.
13. Hodges JS. Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). *Journal of the Royal Statistical Society, Series B* 1998; **60**:497–536.