On h-likelihood, random effects, and penalised likelihood

J.K. Lindsey

Biostatistics, Limburgs Universitair Centrum, Diepenbeek

Email: jlindsey@luc.ac.be

www.luc.ac.be/∼jlindsey

Abstract

The h-likelihood approach of Lee and Nelder (1996) is here interpreted as a penalised likelihood for estimation of doubly-constrained fixed effects that are shrinkage estimates similar to those provided by random effect models. In this way, it can be extended to arbitrary distributions, censored data, and nonlinear regression functions. The estimates of both the random effect parameters and the variance components are directly available.

The new procedure is illustrated by application to a standard split-plot design and to the nonlinear parameters of a pharmacokinetic one-compartment model with left-censored data.

KEYWORDS: censoring, compartment model, fixed effects, h-likelihood, model selection, nonlinear regression, random effects, shrinkage estimates.

1 Introduction

When a cluster of multiple measurements is made on each of several individuals, random effects (RE) models provide a useful approach to handling the heterogeneity among the individuals that is not accounted for by the available inter-cluster covariates. Responses $(y_{ij}, j = 1, \ldots, n_i)$ in a cluster i are assumed independent given one or more unobserved 'variables' or random effects. (For simplicity in what follows, I shall restrict attention to one such random effect, although the discussion extends immediately to the more general case.) Then, models can be constructed based on a mixture distribution whereby these random effects are integrated out yielding a multivariate distribution for the observations on each individual (of t):

$$\prod_i f(y_{i1}, \ldots, y_{in_i}; \boldsymbol{\theta}, \phi, \alpha) = \prod_i \int \left[\prod_j f_0(y_{ij}|u_i; \boldsymbol{\theta}, \phi) \right] f_1(u_i; \alpha) \mathrm{d}u_i \tag{1}$$

where $\boldsymbol{\theta}$ is a vector of regression parameters, and ϕ and α are dispersion (or shape) parameters. (Dispersion parameters will be defined so that they increase with their corresponding variances.)

Here, $f_1(\cdot)$ is a density, whereas $f_0(\cdot)$ may be a density or a discrete probability distribution depending on the nature of y_{ij}. The unobservable $\mathbf{u}$ disappear in the construction of the multivariate distribution and need play no further role in the analysis. The model resulting from this procedure then yields a likelihood function upon which inferences can be made.

A major drawback of such models is the integration which most often must be done numerically, in several dimensions if there is more than one random effects. However, when $f_0(\cdot)$ and $f_1(\cdot)$ are normal distributions, or the latter is the conjugate distribution of the former, the integration can sometimes be performed analytically. Even in these cases, important exceptions occur when the regression function is nonlinear in the unobserved random effects to be integrated out and/or the observations are censored, both of which occur, for example, in applications of pharmacokinetic compartment models. Thus, it would be valuable to have available a procedure which is easily applicable for any arbitrary suitable distributions $f_0(\cdot)$, and its cumulative distribution function (cdf), and $f_1(\cdot)$, with regression functions nonlinear in the unobserved random effects.

Let us call

$$\prod_i f(y_{i1}, \ldots, y_{in_i}; \boldsymbol{\theta}, \phi, u_i) = \prod_i \prod_j f_0(y_{ij}|u_i; \boldsymbol{\theta}, \phi) \tag{2}$$

the corresponding fixed effects (FE) model, where the u_is, suitably constrained, are the $t-1$ unknown FE parameters to be estimated along with $\boldsymbol{\theta}$ and ϕ. In addition, call the model based on $f_0(y_{ij}; \boldsymbol{\theta}, \phi)$, conditioning on available covariates, but not on the u_is, the null model. If an appropriate mixing distribution has been selected for Equation (1), one might expect it to fit better than the corresponding FE model of Equation (2) by model selection criteria such as the AIC, because of the penalty for estimation of a much larger number of parameters in the latter.

Lee and Nelder (1996) introduced a different approach which, unlike the RE model, does not require integration when $f_0(\cdot)$ is a generalised linear model (GLM). Subsequently, they (2001a & b) have somewhat widened the scope. They propose to base inferences on what they call an h-likelihood:

$$\prod_i f(y_{i1}, \ldots, y_{in_i}; \boldsymbol{\theta}, \phi, u_i, \alpha) = \prod_i \prod_j f_0(y_{ij}|u_i; \boldsymbol{\theta}, \phi) \prod_i f_1(u_i; \alpha) \tag{3}$$

where $\mathbf{u}$ is a vector of unknown parameters to be estimated, as in the FE model. However, this h-likelihood is puzzling: it appears to be based on a model that contains $t-1$ more parameters to estimate than the standard RE model of Equation (1) and one more (α) than the corresponding standard FE model in Equation (2). Is the latter parameter identifiable and can it be estimated? On the other hand, arguments in favour of this approach include that integration is not required and that shrinkage estimates are directly provided in place of the usual FE estimates.

The presentation of Lee and Nelder only covers GLMs which allow a rather restricted class of distributions and *linear* regression functions, without censoring. Their estimation procedures depend heavily on the special characteristics of the exponential dispersion family upon which GLMs are based. The question, then, is if Equation (3) can be adapted in some way so that it

can be treated as an ordinary likelihood for models based on distributions outside the exponential dispersion family in the presence of censoring and having regression functions nonlinear in the unobserved random effects.

2 h-likelihoods

In a way similar to that used for ordinary GLMs, Lee and Nelder (1996, 2001a & b) do not use the h-likelihood based on Equation (3) to estimate the dispersion parameters, but only $\boldsymbol{\theta}$ and $\mathbf{u}$ for fixed values of ϕ and α. Their procedure for doing this depends heavily on the special characteristics of the exponential dispersion family and hence cannot be used in the more general case considered here.

2.1 h-likelihoods correspond to no model

For arbitrary distributions $f_0(\cdot)$ and $f_1(\cdot)$, possibly with a nonlinear regression function and/or censoring, we must be able to maximise the h-likelihood globally over all parameters, as can be done with any standard likelihood. However, an h-likelihood cannot be treated as an ordinary likelihood for at least two reasons.

1. Equation (3) appears to correspond to a model that can yield the probability of the data for fixed values of the parameters, including $\mathbf{u}$. However, the sum or integral of the first factor on the right over all possible values of y_{ij} is unity so that multiplying by the second factor will give a value less than or equal to one. Thus, the h-likelihood is does not correspond to a true probability model. Normalising simply eliminates the second factor.

2. The h-likelihood based on Equation (3), when optimised over all parameters, has an infinite maximum with all u_i identical and $\alpha = 0$ (that is, zero variance for the mixing distribution), the null model. However, this problem does not occur in the Lee and Nelder approach because they fix α at some finite nonzero value calculated by other procedures. An infinite likelihood is characteristic of inappropriate use of a density instead of a probability in a likelihood function; it can be eliminated by replacing the densities by differences of cdfs for some finite unit of precision about the random variables so that the maximum of the h-likelihood is then less than or equal to unity. Unfortunately, that does not solve the problem here.

I shall now outline ways in which these two problems can be overcome. In constructing an h-likelihood, let $L_0(\boldsymbol{\theta}, \phi, u_i; y_{ij})$ represent the first factor on the right of Equation (3), $L_1(\alpha, u_i)$ the second factor, and $L(\boldsymbol{\theta}, \phi, \alpha, u_i; y_{ij})$ their product.

2.2 h-likelihood estimates

The problematic parameter in this h-likelihood is α. Inspection of Equation (3) shows that this parameter must be a function of the u_is. For example, in a random intercept model with $f_0(\cdot)$ and

$f_1(\cdot)$ both normal distributions, the estimate of the mixing variance is $\hat{\alpha} = \sum \hat{u}_i^2/t$. The special case of the null model with all $u_i = 0$ has $\hat{\alpha} = 0$. Thus, α is not, in fact, an additional parameter; it is not identifiable. Let us look at this more closely.

Consider first optimising the second factor of the h-likelihood alone. For *one fixed* u_i, the maximum is given by

$$\lim_{\alpha \to 0} L_1(\alpha, u_i) = 1$$

(if based on probabilities; if based on densities, the maximum is infinite). However, in the complete h-likelihood, there are several u_is and they are not fixed, but are unknown parameters. Thus, consider next the simultaneous maximum of the product, $\prod_i L_1(\alpha, u_i)$. This will still be unity when $\alpha = 0$, implying that all of the u_is are identical, corresponding to the null model.

Consider now the unconstrained maximum of the first factor in the h-likelihood, alone. This corresponds to the FE model of Equation (2), yielding a value much smaller than one. Generally, the values of the u_is estimated in this way will all be different unless all clusters are very similar. Thus, this conflicts with optimisation of the second factor.

Now let us look at the two factors together. If the u_is in the second factor were set to the FE values, this function would have a maximum when α is the dispersion of these values (for example, the variance of the u_is when $f_1(\cdot)$ is a normal distribution). This would also be very considerably less than its unconstrained maximum of unity. Thus, the maximum of the complete h-likelihood, $\prod_{ij} L(\boldsymbol{\theta}, \phi, \alpha, u_i; y_{ij})$, should yield a compromise between the null and FE models, depending on the value of α. The structure of the h-likelihood should act to place a constraint on the FE u_i parameter values. However, this does not occur when α is estimated simultaneously with the u_is, instead of being fixed, because the second factor can reach one, greatly outweighing the first factor, so that the overall maximum of the h-likelihood corresponds to the null model.

Thus, the problem is to obtain a value for α; the h-likelihood has a maximum at $\alpha = 0$, pointing to the null model. Lee and Nelder (1996) overcome this by using an *adjusted* h-likelihood in which Equation (3) is multiplied by a multivariate normal distribution $f_2(\cdot)$:

$$\prod_i f(y_{i1}, \ldots, y_{in_i}; \boldsymbol{\theta}, \phi, \mathbf{u}, \alpha) = \prod_i \prod_j f_0(y_{ij}|u_i; \boldsymbol{\theta}, \phi) \prod_i f_1(u_i; \alpha) f_2(0, \phi \mathbf{H}^{-1}) \tag{4}$$

where $\mathbf{H}$ is obtained from the GLM estimation procedure and involves α. They use this to iterate between estimating $\boldsymbol{\theta}, \mathbf{u}$ and ϕ, α so that this is in fact the 'likelihood' they are using, not the h-likelihood $\prod_{ij} L(\boldsymbol{\theta}, \phi, \alpha, u_i; y_{ij})$. Again, for the first reason given above for h-likelihoods, this adjusted h-likelihood also does not correspond to a probability model.

The effect of including this additional factor is that α is no longer estimated only from $\sum \hat{u}_i^2/t$. Thus, for example, in a linear regression model where $f_0(\cdot)$ and $f_1(\cdot)$ are both normal distributions, with variances respectively ϕ and α, the estimate of α obtained from Equation (4) is a weighted sum of the two variance estimates, $\hat{\phi}$ and $\sum \hat{u}_i^2/t$. See the derivation by Lee and Nelder (1996) following their Equation (4.8). The question is how to interpret this in a useful general way.

2.3 Doubly-constrained fixed effects

Consider first the classical normal-normal model. In the FE model with a different intercept in each cluster,

$$\widehat{u_i} = \bar{y}_{i+} - \mu_i(\widehat{\boldsymbol{\theta}})$$

where $\mu_i(\cdot)$ is the regression function fitted, not including u_i. In the corresponding RE model, Y_{ij} given u_i has conditional distribution $f_0(\cdot)$, with variance ϕ and, from Equation (3), α is the variance of the marginal distribution of the u_is. Y_{ij} has marginal distribution given by Equation (1) with variance $\phi + \alpha$, which can be estimated by $\hat{\phi} + \sum \hat{u}_i^2/t$, essentially the estimate of α obtained from the adjusted h-likelihood based on Equation (4).

Now let us look at arbitrary conditional and marginal distributions $f_0(\cdot)$ and $f_1(\cdot)$. Although, in general, the variances will not be explicit parameters in the model, let σ_0^2 represent the conditional variance of Y_{ij} given u_i and σ_1^2 the marginal variance of the u_is parametrised with some constraint, such as sum zero or product one. Then, the former can be estimated by

$$\hat{\sigma}_0^2 = \sum_i \sum_j [y_{ij} - \mu_i(\widehat{\boldsymbol{\theta}}, \hat{\mathbf{u}})]^2 / n_+$$

where $\mu_i(\cdot)$ is now some regression function possibly nonlinear in $\boldsymbol{\theta}, \mathbf{u}$. ($n_+$ might be adjusted by the number of estimated parameters.)

Let us assume that the marginal distribution $f_1(\cdot)$ is such that its dispersion or shape parameter α is some function of σ_1^2: $\alpha(\sigma_1^2)$. Examples of such distributions include the normal, gamma, inverse Gauss, Weibull, and beta, and modifications of them such as the log normal and inverse gamma; distributions with infinite variance, such as the Cauchy, must be excluded. Then, I propose to estimate this parameter, not by $\hat{\alpha} = \alpha(\hat{\sigma}_1^2)$ as when Equation (3) is optimised, but by $\hat{\alpha} = \alpha(\hat{\sigma}_0^2 + \hat{\sigma}_1^2)$. For example, in the classical normal-normal linear model, α will be estimated by the marginal variance of Y_{ij}, $\hat{\sigma}_0^2 + \sum \hat{u}_i^2/t$, as described above. Using this relationship, we can maximise the h-likelihood in Equation (3) directly in one step using a nonlinear optimiser. Because the estimate of α is constrained away from zero by the inclusion of σ_0^2, this generalised h-likelihood function is prevented from going to infinity.

This procedure can be interpreted in the following way. Consider the FE model of Equation (2). The FE parameters u_i have some constraint such as sum zero or product one, so that $t-1$ independent estimates are obtained from the corresponding likelihood. Now introduce a second constraint. Not only will the mean of these parameters (or their logarithm) be zero but they will have some chosen distribution $f_1(\cdot)$ with variance $\sigma_0^2 + \sigma_1^2$. With these two constraints incorporated into the model, $t-2$ independent estimates need to be estimated from the h-likelihood, one less than for the corresponding FE model. Let us call this the doubly-constrained FE model.

The first constraint can be introduced into the standard likelihood based on Equation (2) directly in the usual way. However, point estimation using the second requires that this likelihood be multiplied by $f_1(\cdot)$, yielding the h-likelihood of Equation (3), a form of penalised likelihood.

Then, this function must be maximised using $\hat{\alpha} = \alpha(\hat{\sigma}_0^2 + \hat{\sigma}_1^2)$. On the other hand, the likelihood for making inferences about the doubly-constrained FE model, such as model selection, remains that based on Equation (2). The h-likelihood is an *estimation* procedure, a form of penalised likelihood for introducing an additional constraint on the u_is, not an ordinary likelihood function.

Note that the choice of the way in which α depends on the other parameters, here via $\alpha = \alpha(\sigma_0^2 + \sum u_i^2/t)$, only determines the constraint placed on the u_i. It does not involve any approximate inference procedure, as do the derivation of the adjusted h-likelihood of Equation (4) or quasi-likelihood.

2.4 Goodness of fit

In complex models outside the exponential family, standard procedures for goodness of fit, such as examination of residuals, are generally of little use for indicating problems with a model and for suggesting alternatives. The interpretation of h-likelihood as an ordinary FE likelihood penalised by an additional constraint on the FE parameters allows us to use the likelihood based on Equation (2), not the h-likelihood, in model selection criteria, such as the AIC or BIC.

One example used by Lee and Nelder (1996) is the seed germination data of Crowder (1978). As shown by Lindsey (1999), there is little evidence of overdispersion in these data. Because Lee and Nelder do not have available an objective model selection criterion, they do not realise that the model based on their h-likelihood fits no better than the ordinary binomial model. A standard binomial GLM has an AIC (negative log likelihood plus the number of estimated parameters) of 58.9 as compared to 58.8 for the corresponding beta-binomial model which has been penalised for one extra parameter. A binomial generalised linear mixed model (GLMM) for overdispersion, with normal mixing distribution, has an AIC of 58.6, again with one more parameter than the standard binomial model. According to Lee and Nelder (1996), the latter analysis is essentially similar to their h-likelihood.

Thus, some objective model selection criterion is essential when doing any model fitting, both to obtain reasonably fitting models and to avoid overfitting as in this example.

3 Examples

3.1 Chocolate cakes

A second example used by Lee and Nelder (1996) involves the breaking angle of chocolate cakes, from Cochran and Cox (1957, p. 300). For these data, there are 15 clusters (the replications) with 18 observations in each (three recipes at each of six temperatures).

All approaches point to a model with differences among replications and recipes and a linear trend in temperature, so that I shall only discuss this. Consider first the classical analysis of Cochran and Cox using a normal distribution. Applying maximum likelihood to the RE model, we

Table 1: Estimates of replication effects from various models fitted to the chocolate cake data with their standard errors (s.e.). The last replicate has effect equal to minus the sum of the others. Those for the gamma-inverse gamma are on the log scale.

	Normal-normal		Gamma-inverse gamma	
	Fixed	Doubly-constrained FE	Fixed	Doubly-constrained FE
1	14.71	14.30	0.397	0.410
2	13.43	13.06	0.365	0.379
3	4.82	4.69	0.155	0.169
4	1.16	1.13	0.057	0.070
5	0.27	0.27	0.020	0.035
6	−3.40	−3.30	−0.094	−0.081
7	−4.79	−4.66	−0.152	−0.136
8	−4.73	−4.60	−0.146	−0.132
9	−4.79	−4.66	−0.144	−0.131
10	−3.29	−3.20	−0.094	−0.080
11	−2.23	−2.17	−0.053	−0.040
12	−1.18	−1.15	−0.023	−0.009
13	−0.23	−0.23	0.014	0.030
14	−4.23	−4.12	−0.131	−0.116
15	−5.51	−5.36	−0.172	−0.158
s.e.	1.06	1.04	0.033	0.031

find the estimated variance components to be 22.89 and 36.52. (In fact, recipes are nested within replications, giving variance estimates 19.11, 35.57, and 3.47, but, following both Cochran and Cox and Lee and Nelder, I shall ignore this in what follows.) The corresponding point estimates using the doubly-constrained FE model are 21.65 and 35.72. Thus, these penalised estimates, obtained without integration, are very close to those from the RE model. The FE and corresponding shrinkage estimates are shown in the first two columns of Table 1. Because of the large number of observations within each cluster (replication), little shrinkage occurs.

Lee and Nelder (1996), following Firth and Harris (1991), use a multiplicative model (log link) with constant coefficient of dispersion (gamma distribution) and the conjugate mixing distribution. To obtain this here, I shall modify the above model in three steps, changing sequentially the link from identity to log, $f_0(\cdot)$ from normal to gamma, and $f_1(\cdot)$ from normal to inverse gamma. As can be seen in Table 2, introducing a log link does not improve the fit. On the other hand, introducing the gamma conditional distribution for $f_0(\cdot)$ does improve it. Again, the doubly-constrained FE model provides a somewhat better fit than the FE model, with almost no shrinkage.

Lee and Nelder (1996) use a gamma distribution with log link and inverse gamma mixing

Table 2: AICs for various models fitted to the chocolate cake data.

$f_0(\cdot)$	Normal		Normal		Gamma	
Link	Identity		Log		Log	
Null model	939.5		939.6		921.3	
Fixed effect	817.0		817.0		808.2	
$f_1(\cdot)$	Normal	Inverse gamma	Normal	Inverse gamma	Normal	Inverse gamma
Random effect	837.2	834.6	837.5	835.7	826.0	826.5
Doubly-constrained FE	816.2	816.3	816.8	816.9	807.2	807.2

distribution; this is the third step. As seen in Table 2, the inverse gamma mixing distribution does not fit better than the normal. The FE and corresponding shrinkage estimates are shown in the last two columns in Table 1.

Lee and Nelder (1996) estimate the shape parameter of the inverse gamma mixing distribution to be 219.1. With my procedure, the estimate is 41.2, but the likelihood is almost constant over a wide range of values. With fixed mean, the variance is the reciprocal of these values. Lee and Nelder claim that their residual plots point to the inverse gamma mixing distribution in preference to the normal mixing distribution, but these plots can be misleading and are always subjective. A gamma or inverse gamma distribution with such a small variance is virtually identical to a normal distribution with the same variance. The RE model with normal mixing distribution has a slightly larger likelihood (with the same number of parameters estimated) and the two doubly-constrained FE models both show the same fit, slightly better than the FE model.

It is interesting to note that, although not the best model, the normal conditional distribution RE model, whether with identity or log link, fits better with the inverse gamma mixing distribution than with the more usual normal mixing distribution, as can also be seen in Table 2.

3.2 Pharmacokinetic compartment models

Lindsey *et al.* (2000) analyse Phase I pharmacokinetic data concerning concentrations of flosequinan and its metabolite in 18 healthy volunteers. They found that a gamma distribution was required with a first-order one-compartment model

$$\mu_t = \frac{k_a d}{V(k_a - k_e)} \left(\mathrm{e}^{-k_e t} - \mathrm{e}^{-k_a t}\right) \tag{5}$$

describing the mean concentration over time, where the absorption rate k_a, the elimination rate k_e, and the volume V are parameters to be estimated, d is the dose administered, and t is time. They also showed that it is important to model the left-censoring due to undetectably small values and to use an appropriate function for the change in dispersion over time. For the latter, they use Equation (5) raised to a power, an additional parameter, where all four parameters have different

Table 3: AICs for the FE models fitted to the flosequinan data.

	Log normal	Gamma
Null	563.9	548.9
Absorption	545.2	543.4
Elimination	471.4	426.6
Volume	550.2	491.6

Table 4: AICs for the FE, RE, and doubly-constrained FE models for the elimination rate with a gamma conditional distribution, fitted to the flosequinan data.

Mixing distribution	Normal	Gamma
Fixed effect	426.6	
Random effect	460.9	458.7
Doubly-constrained FE	425.7	425.7

values than in the regression function for the mean. However, the only dependencies that they considered were those accounted for by FE models or by covariates.

Here, I shall apply random effects and the doubly-constrained FE model to the three nonlinear parameters individually, using an appropriate regression function for the dispersion parameter and allowing for left censoring. I shall compare the results from the standard log normal distribution with those from the gamma distribution. For simplicity, I shall only consider the lowest dose level of the parent drug, flosequinan. For this subset of the data, there are 12 observations over time for each of the 18 subjects and 111 nondetectable values out of 216 observations. Two individuals have extreme curves (numbers 12 and 18), both higher than the others.

Inspection of the fits of the FE models in Table 3 shows that differences in the elimination rate explain the most variability among the individual curves. As for the complete data set, the gamma distribution fits much better than the log normal.

Let us now look at the RE and doubly-constrained FE models for the elimination rate, with the gamma conditional distribution and using normal and gamma mixing distributions. As can be seen in Table 4, the doubly-constrained FE models fit somewhat better than the FE model, and much better than the corresponding RE models. There is little difference among the two mixing distributions. The differences in log elimination rate for each subject are shown in Table 5 using the three models. Here, there is virtually no shrinkage. Individuals 12 and 18 clearly stand out as extreme, with lower elimination rates.

If the left-censored nondetectable values are set to one-half the detectable level and the density is used for them instead of the cdf, the FE and doubly-constrained FE models for the elimination parameter converge to the null model with an AIC of about 632, much worse than those that allow

Table 5: Estimates of individual effects for the log elimination rate from models fitted to the flosequinan data with their standard errors for the gamma conditional distribution with normal and gamma mixing distributions.

	Fixed	Doubly-constrained FE	
		Normal	Gamma
1	0.30	0.29	0.12
2	0.05	0.05	−0.12
3	0.57	0.56	0.38
4	0.60	0.59	0.41
5	0.93	0.92	0.75
6	−0.04	−0.04	−0.22
7	0.38	0.37	0.19
8	0.48	0.47	0.30
9	0.26	0.25	0.08
10	0.61	0.60	0.42
11	−0.32	−0.31	−0.49
12	−1.64	−1.62	−1.80
13	−0.26	−0.26	−0.43
14	0.15	0.15	−0.03
15	−0.42	−0.41	−0.59
16	−0.09	−0.08	−0.26
17	0.24	0.24	0.06
18	−1.79	−1.77	−1.95
s.e.	0.10	0.07	0.06

correctly for censoring. As with the models fitted to the complete data set by Lindsey *et al.* (2000), the compartment model curve is greatly distorted.

4 Discussion

Penalties on the likelihood have been used in two different ways here. The h-likelihood involves a penalty that places a constraint on the parameters in the statistical model, used in their point estimation. This penalty is not used in the model selection process once estimates with the constraint have been obtained. On the other hand, the AIC and other model selection criteria involve a penalty on the complexity of the model. These do not affect the point estimates of the parameters of a given model, but provide an objective procedure for comparing distinct models.

If the 'correct' mixing distribution $f_1(\cdot)$ were chosen, one might expect that the RE model would fit better than the corresponding FE model, whether with one or two constraints, because the latter has a large number of parameters estimated. This, however, will depend both on the amount of information available within each cluster and on the penalty imposed. As is well known, the BIC may indicate a simpler model than the AIC. For the cake data, the BIC gives a small advantage to the RE model whereas, for the pharmacokinetic data, it points strongly to the FE and doubly-constrained FE models over the RE one.

A likelihood function plays several roles in the analysis of data.

1. It corresponds to some probabilistic representation of the observed data, containing unknown parameters.

2. By its maximisation, it allows point estimates of these parameters to be calculated.

3. By study of its shape, the precision of these estimates can be determined.

The h-likelihood of Lee and Nelder (1996) fulfils none of these criteria, although their adjusted h-likelihood does meet point 2. On the other hand, the interpretation of the h-likelihood as applying one additional constraint to the standard FE model to bring it closer to the RE model allows Equation (2) to be interpreted as a standard likelihood. The estimation procedure proposed above allows a wide class of models to be fitted to clustered data involving censoring and nonlinear regression functions.

All of the models used in the examples above can be routinely fitted using my libraries for R available at `www.luc.ac.be/~jlindsey/rcode.html`. The random effects models are fitted by Romberg integration, except for certain with the normal mixing distribution for which the faster Gauss-Hermite integration may used.

References

[1] Cochran, W.G. and Cox, G.M. (1957) *Experimental Designs.* New York: John Wiley.

[2] Crowder, M.J. (1978) Beta-binomial anova for proportions. *Applied Statistics* **27**, 34–37.

[3] Firth, D. and Harris, I.R. (1991) Quasi-likelihood for multiplicative random effects. *Biometrika* **78**, 545–555.

[4] Lee, Y. and Nelder, J.A. (1996) Hierarchical generalised linear models. *Journal of the Royal Statistical Society B* **58**, 619–678.

[5] Lee, Y. and Nelder, J.A. (2001a) Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88**, 987–1006.

[6] Lee, Y. and Nelder, J.A. (2001b) Modelling and analysing correlated non-normal data. *Statistical Modelling* **1**, 3–16.

[7] Lindsey, J.K. (1999) On the use of corrections for overdispersion. *Journal of the Royal Statistical Society* **C48**, 553–561.

[8] Lindsey, J.K., Byrom, W.D., Wang, J., Jarvis, P., and Jones, B. (2000) Generalised nonlinear models for pharmacokinetic data. *Biometrics* **56**, 81–88.