1351.0.55.029

Australian Bureau of Statistics

Research Paper

Small Area Estimation Using a Multinomial Logit Mixed Model with Category Specific Random Effects

www.abs.gov.au

Research Paper

Small Area Estimation Using a Multinomial Logit Mixed Model with Category Specific Random Effects

Janice Scealy

Australian Bureau of Statistics
and Australian National University

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 28 JAN 2010

ABS Catalogue no. 1351.0.55.029

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Janice Scealy, Analytical Services Branch on Canberra (02) 6252 5764 or email <analytical.services@abs.gov.au>.

CONTENTS

SMALL AREA ESTIMATION USING A MULTINOMIAL LOGIT MIXED MODEL WITH CATEGORY SPECIFIC RANDOM EFFECTS

Janice Scealy
Australian Bureau of Statistics
and Australian National University

ABSTRACT

This paper describes a model based approach to producing small area estimates of counts for different categories of the Australian labour force based on a multinomial logit mixed model with category specific random effects. By category specific we mean that within each small area there are two correlated random effects, one associated with the employed category and the other associated with the unemployed category. Estimates of the model parameters are produced using penalized quasi-likelihood combined with approximated restricted maximum likelihood estimation and using these, estimated counts are then produced for each small area. Mean squared error estimates of the estimated counts are approximated using two methods: 1) a parametric bootstrap and 2) analytical approximations and we compare the performance of both. Using a parametric bootstrap we also examine the properties of the combined penalized quasi-likelihood and restricted maximum likelihood estimators and discuss model goodness of fit measures and diagnostics.

Keywords: small area estimation, multinomial logit mixed model, parametric bootstrap, labour force survey.

1. INTRODUCTION

The Australian Bureau of Statistics (ABS) produces labour force estimates using direct survey estimators for regions with large enough sample sizes for these estimators to be reliable. In recent years there has been a growing demand for labour force estimates to be produced in smaller geographical regions. The direct estimates for these small regions are considered to be too unreliable because the standard errors are large due to small sample sizes. A way around this is to produce model based estimates which borrow strength from administrative and Census data and other types of auxiliary variables. The hope is that the model based estimators will produce estimates with mean squared error less than the direct survey estimators.

The model based approach relies on an appropriate choice of model and good auxiliary variables. The aim here is to produce estimates for each of three labour force statuses: employment, unemployment and not in the labour force for a set of small areas. Auxiliary data are available within age/sex classes for each small area and sample counts of the three labour force statuses are obtained from the Australian Labour Force Survey. The total numbers of people within each sex/age group are also assumed known and are obtained from the Estimated Resident Population (ERP) projections published by the ABS (for further details, see ABS, 2007). Molina *et al.* (2007) describe a methodology based on the application of the multinomial logit mixed model which can be used to produce estimates in this small area estimation situation. Random area effects are included in the models to account for potential correlations between the age/sex class counts in the small areas not explained by the auxiliary variables. The inclusion of random area effects in the model specifically accounts for the area level variation not explained by the auxiliary variables.

In the model described in Molina *et al.* (2007), only one random area effect is used within each small area and the random effect is therefore the same across the multinomial classes. In our situation this may not be appropriate. Some work carried out at the ABS on fitting three separate logistic mixed models to the data suggests that the variances of the random effects are not the same across each category. A more appropriate model would be to introduce category specific random effects. This allows for the variances of the random effects to differ between the categories and also allows for a potential correlation between them as well. In our case it does not make sense to assume that the category specific random effects are perfectly correlated. By allowing for a general arbitrary covariance matrix, Hartzel *et al.* (2001) make the point that the model will be structurally the same regardless of the choice of baseline category which is a good property.

In this paper we extend the model in Molina *et al.* (2007) to include category specific random effects. To estimate model parameters, we develop a similar penalized quasi-likelihood (PQL) estimation scheme with approximate maximum likelihood (ML) and/or restricted maximum likelihood (REML) for the variance components. These parameter estimates are then used to produce estimated labour force counts for each small area. Mean squared error estimates of the estimated counts are approximated using two methods:

1. a parametric bootstrap, and
2. analytical approximations,

and we compare the estimates produced using these two methods. Using a parametric bootstrap, we also examine the properties of the combined PQL and REML estimators and discuss model goodness of fit. Unlike Molina *et al.* (2007), we also consider estimation for out-of-sample small areas and briefly review some alternative estimation schemes. Note that the primary focus of this paper is to give technical details on how one might produce model based estimates for the Australian labour force. This is an experimental procedure and the ABS will not be publishing any model based estimates for the Australian labour force as part of the ABS product at this stage.

2. THE MODEL

Similar to Molina *et al.* (2007) let index i ($i = 1, 2, \ldots, I_d$) denote the sex/age groups and d ($d = 1, 2, \ldots, D$) denote the small areas. We have a slightly different set up because I_d is not constant ($I_d \leq 10$). The reason why we have unbalanced data is because some of the sex/age groups within a small area have no sample and will be excluded from model estimation since they do not contribute to the likelihood. The labour force sample counts are denoted by y_{di1}, y_{di2} and y_{di3} which represent respectively, employment, unemployment and not in the labour force counts in sex/age group i in the d-th small area. Let $m_{di} = y_{di1} + y_{di2} + y_{di3}$ denote the sample size and p_{di1}, p_{di2} and p_{di3} denote the respective probabilities of employed, unemployed and not in the labour force. Let u_{d1} and u_{d2} denote the category specific random effects. We assume that the vectors $(y_{di1}, y_{di2}, y_{di3})^t$ given m_{di} and $\boldsymbol{u}_d = (u_{d1}, u_{d2})^t$ are independent across d and i with multinomial distribution, that is with the probability density function

$$f(y_{di1}, y_{di2} \mid \boldsymbol{u}_d) = \frac{m_{di}!}{y_{di1}!\, y_{di2}!\, y_{di3}!}\, p_{di1}^{y_{di1}}\, p_{di2}^{y_{di2}}\, p_{di3}^{y_{di3}}. \tag{2.1}$$

It is also assumed that for $j = 1, 2$

$$\log \frac{p_{dij}}{p_{di3}} = \boldsymbol{x}_{dij}^t \boldsymbol{\beta}_j + u_{dj}, \tag{2.2}$$

where $\boldsymbol{\beta}_j$ is a vector of parameters and $\boldsymbol{x}_{dij}$ is a vector of explanatory variables associated with the j-th category. Note that in (2.1) above, technically we should also be conditioning on $\boldsymbol{x}_{dij}$, that is, f should be defined as $f(y_{di1}, y_{di2} \mid \boldsymbol{u}_d, \boldsymbol{x}_{dij})$. It will be assumed throughout this paper that whenever we condition on $\boldsymbol{u}_d$ we also condition on $\boldsymbol{x}_{dij}$ even when it is omitted from the notation. We also assume that $\boldsymbol{u}_d$ is independently and identically distributed as bivariate normal and its probability density function is

$$f(\boldsymbol{u}_d) = \frac{1}{2\pi \left| \boldsymbol{W}_d \right|^{\frac{1}{2}}} e^{-\frac{1}{2} \boldsymbol{u}_d^t \boldsymbol{W}_d^{-1} \boldsymbol{u}_d}$$

where

$$\boldsymbol{W}_d = \begin{pmatrix} \varphi_1 & \varphi_{12} \\ \varphi_{12} & \varphi_2 \end{pmatrix}.$$

Under this model we have a vector of variance components $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_{12})^t$ that will need to be estimated along with $\boldsymbol{\beta} = (\beta_1^t, \beta_2^t)^t$. This model is a GLMM (generalised linear mixed model) and to estimate the parameters a variety of different techniques can be used. In the next section we discuss some of these.

3. ESTIMATION OF β AND φ

Let $\boldsymbol{y}_{di} = (y_{di1}, y_{di2})^t$ for $i = 1, 2, \ldots, I_d$ and $d = 1, 2, \ldots, D$ and let $\boldsymbol{y}$ be the vector obtained by stacking the $\boldsymbol{y}_{di}$'s into a column. Using the definition of extended likelihood in Pawitan (2001) we may write the likelihood function for $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{u} = (\boldsymbol{u}_1^t, \boldsymbol{u}_2^t, \ldots, \boldsymbol{u}_D^t)^t$ as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{u}) &= f(\boldsymbol{y} \mid \boldsymbol{u}) f(\boldsymbol{u}) \\ &= \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} f(y_{di1}, y_{di2} \mid \boldsymbol{u}_d) \right) \left(\prod_{d=1}^{D} f(\boldsymbol{u}_d) \right). \end{aligned} \tag{3.1}$$

Pawitan also notes that this definition of the likelihood is called hierarchical likelihood by Lee and Nelder (1996).

For the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$, ideally likelihood based estimation should be based on maximising $L(\boldsymbol{\beta}, \boldsymbol{\varphi})$, where

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\varphi}) &= \int \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} f(y_{di1}, y_{di2} \mid \boldsymbol{u}_d) \right) \left(\prod_{d=1}^{D} f(\boldsymbol{u}_d) \right) d\boldsymbol{u} \\ &= \prod_{d=1}^{D} \left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\boldsymbol{u}_d) \left(\prod_{i=1}^{I_d} f(y_{di1}, y_{di2} \mid \boldsymbol{u}_d) \right) du_{d1} du_{d2} \right). \end{aligned} \tag{3.2}$$

To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ one could try to maximise the marginal likelihood defined at (3.2) using, for example, Monte-Carlo methods or numerical integration techniques to evaluate the integrals. A procedure like Newton-Raphson could then be used to solve the likelihood equations since the equations are non-linear. More specifically, Hartzel *et al.* (2001) describe an adaptive Gauss-Hermite, quasi-Newton algorithm which could be used in the multinomial logit mixed model case. This method is appropriate when the dimension of the integrals are small and the dataset size is not large. In our case, the dimension of the integrals are small but the dataset size is large. That is, there are a large number of double integrals to evaluate at each iteration and convergence will therefore be very slow.

Another method which could be used to maximise (3.2) is to use an automated Monte Carlo EM algorithm as described in Hartzel *et al.* (2001). Again this will be computationally intensive in our case. The method consists of implementing an EM algorithm which treats the random effects $\boldsymbol{u}$ as the missing data. In the E-step a conditional expectation needs to be evaluated and this is approximated by using Monte Carlo methods (actually, an independent sample from the conditional distribution is generated). The M-Step is then undertaken by maximising this

approximate expectation. At each EM iteration, the Monte Carlo sample size is increased in an automated way until convergence. In any case, for computation reasons we do not pursue exact marginal likelihood maximisation approaches further here. Instead we will now discuss approximate methods.

The approach taken in Molina *et al.* (2007) is to use the PQL (penalized quasi-likelihood) method introduced by Breslow and Clayton (1993) combined with either ML (maximum likelihood) or REML (restricted maximum likelihood) for the variance components. The PQL method of Breslow and Clayton (1993) obtains estimates of $\boldsymbol{\beta}$ given $\boldsymbol{\varphi}$ by maximising an approximation to the marginal likelihood $L(\boldsymbol{\beta}, \boldsymbol{\varphi})$. Part of this approximation involves a Laplace integral approximation. Estimates for $\boldsymbol{u}$ are also produced as a by-product of the approximation and hence the method produces joint estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ given $\boldsymbol{\varphi}$. As Jiang (2007) states, the PQL method is equivalent to the maximum hierarchical likelihood method of Lee and Nelder (1996) in the case of normality of the random effects. The maximum hierarchical likelihood method obtains joint estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ by maximising the hierarchical likelihood (the log of (3.1) in our case). Interestingly in the case of normal linear mixed models the estimate of $\boldsymbol{\beta}$ obtained by maximising $L(\boldsymbol{\beta}, \boldsymbol{\varphi})$ and $L(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{u})$ given $\boldsymbol{\varphi}$ are equivalent. But this is not the case in general for GLMMs.

As mentioned by Jiang (2007), Lee and Nelder (1996) showed that in general the maximum hierarchical likelihood estimates of the fixed effects are asymptotically equivalent to the marginal maximum likelihood estimates of the fixed effects. On the surface this suggests that when the sample sizes are large, then to obtain estimates of $\boldsymbol{\beta}$, maximising (3.2) and (3.1) are equivalent. However as Jiang (2007) states, the asymptotics here are in the sense of the cluster sample sizes approaching infinity, but the number of clusters remaining bounded. This is not satisfied in our small area estimation case since the number of clusters (small areas) is large but the cluster sample sizes are small and bounded (≤ 10, since the sampled units are the age/sex classes). Therefore estimators for $\boldsymbol{\beta}$ derived from maximising (3.1) will not be equivalent to maximising (3.2) even as we increase the number of small areas in the sample. As Jiang (2007) also states, there are a number of approximations involved in deriving the PQL and these approximations have introduced bias into the estimates and this bias does not vanish asymptotically (PQL estimators are known to be inconsistent). When the cluster sample sizes are small there is insufficient information to estimate both the random and fixed effects $\boldsymbol{\beta}$ simultaneously.

Hartzel *et al.* (2001) points out that PQL methods have been shown to be biased especially for highly non-normal cases such as Bernoulli response data and biases tend to increase as the variance components increase. Hartzel *et al.* (2001) suspect that a similar problem will exist for the multinomial logit random effects model when the multinomial sample sizes are small. In our case the variance components are

expected to be small and the average m_{di} is approximately 13. Also note that the number of multinomial observations per cluster is ≤ 10. In most cases these are exactly 10. Therefore the m_{di} and cluster sizes might be sufficiently large together and the variance components sufficiently small to allow the PQL estimators to work reasonably well. This will of course need to be confirmed via simulation which is undertaken later.

In any case, two issues have been highlighted so far. The first is the computational difficulty of maximising (3.2) directly due to the presence of the integrals which have no closed form solution and the second is the inconsistency of the PQL estimates associated with maximising (3.1). These two issues give some motivation for trying to come up with alternative estimators. Jiang (1998) proposes an alternative estimation method called the method of simulated moments which is both computationally attractive and results in consistent estimators. A set of estimating equations are obtained by equating sample moments of the sufficient statistics to their expectations. The expectations are then approximated by simulating sequences of normal random variables (i.e. the integrals in the expectations are approximated by Monte Carlo simulation). The equations are then solved by a Newton-Raphson procedure. However, there is one issue associated with this method. Jiang (1998) shows that the method of simulated moment estimators can be quite inefficient. For small samples the method of simulated moment estimators seem to have substantially larger variance than estimators based on PQL. So it appears there is a bias versus variance trade-off when choosing between such methods as PQL or the method of simulated moments. There are clearly issues associated with all estimators that have been discussed so far.

4. PQL ESTIMATION OF β AND u

To obtain the PQL estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ we need to maximise the log of the joint likelihood defined at (3.1). Assume for the moment that the variance components $\boldsymbol{\varphi}$ are known. The joint log likelihood is

$$l(\boldsymbol{\beta},\boldsymbol{u}) = c - \frac{1}{2}\sum_{d=1}^{D} \boldsymbol{u}_d^t \boldsymbol{W}_d^{-1} \boldsymbol{u}_d + \sum_{d=1}^{D}\sum_{i=1}^{I_d}\sum_{j=1}^{3} y_{dij} \log p_{dij}, \tag{4.1}$$

where c is a constant. The maximum likelihood estimators can be obtained by equating the first derivatives of (4.1) to zero and then solving this system of equations. Let $q = 1, 2, \ldots, Q_j$ index the components of the vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{x}_{dij}$ and denote the q-th component of each by $\beta_{j(q)}$ and $x_{dij(q)}$ respectively. By noting that

$$p_{di1} = \frac{e^{\boldsymbol{x}_{di1}^t \boldsymbol{\beta}_1 + \boldsymbol{u}_{d1}}}{1 + e^{\boldsymbol{x}_{di1}^t \boldsymbol{\beta}_1 + \boldsymbol{u}_{d1}} + e^{\boldsymbol{x}_{di2}^t \boldsymbol{\beta}_2 + \boldsymbol{u}_{d2}}}$$

$$p_{di2} = \frac{e^{\boldsymbol{x}_{di2}^t \boldsymbol{\beta}_2 + \boldsymbol{u}_{d2}}}{1 + e^{\boldsymbol{x}_{di1}^t \boldsymbol{\beta}_1 + \boldsymbol{u}_{d1}} + e^{\boldsymbol{x}_{di2}^t \boldsymbol{\beta}_2 + \boldsymbol{u}_{d2}}}$$

and

$$p_{di3} = \frac{1}{1 + e^{\boldsymbol{x}_{di1}^t \boldsymbol{\beta}_1 + \boldsymbol{u}_{d1}} + e^{\boldsymbol{x}_{di2}^t \boldsymbol{\beta}_2 + \boldsymbol{u}_{d2}}}$$

after some algebra it can be shown that for $j = 1, 2, 3; j' = 1, 2; q = 1, 2, \ldots, Q_{j'}$; $i = 1, 2, \ldots, I_d$ and $d = 1, 2, \ldots, D$,

$$\frac{\partial \log p_{dij}}{\partial \beta_{j'(q)}} = \begin{cases} x_{dij'(q)}\left(1 - p_{dij'}\right) & \text{if } j = j' \\ -x_{dij'(q)} p_{dij'} & \text{otherwise} \end{cases}$$

and for $j' = 1, 2$ and $q = 1, 2, \ldots, Q_{j'}$,

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial \beta_{j'(q)}} = \sum_{d=1}^{D}\sum_{i=1}^{I_d} x_{dij'(q)}\left(y_{dij'} - m_{di} p_{dij'}\right). \tag{4.2}$$

Now we need to find the first derivatives with respect to the random effects. Again after some algebra it can be shown that for $j = 1, 2, 3$; $j' = 1, 2$; $i = 1, 2, \ldots, I_d$; $d = 1, 2, \ldots, D$ and $d' = 1, 2, \ldots, D$,

$$\frac{\partial \log p_{dij}}{\partial u_{d'j'}} = \begin{cases} 1 - p_{d'ij'} & \text{if } j = j' \text{ and } d = d' \\ -p_{d'ij'} & \text{if } j \neq j' \text{ and } d = d' \\ 0 & \text{otherwise,} \end{cases}$$

and for $j' = 1, 2$ and $d' = 1, 2, \ldots, D$,

$$\frac{\partial\left(\sum_{d=1}^{D}\sum_{i=1}^{I_d}\sum_{j=1}^{3} y_{dij}\log p_{dij}\right)}{\partial u_{d'j'}} = \sum_{i=1}^{I_d}\left(y_{d'ij'} - m_{d'i}p_{d'ij'}\right).$$

Now we note that for $d = 1, 2, \ldots, D$,

$$\boldsymbol{W}_d^{-1} = \begin{pmatrix} \varphi_1 & \varphi_{12} \\ \varphi_{12} & \varphi_2 \end{pmatrix}^{-1} = \frac{1}{\varphi_1\varphi_2 - \varphi_{12}^2}\begin{pmatrix} \varphi_2 & -\varphi_{12} \\ -\varphi_{12} & \varphi_1 \end{pmatrix}$$

and $$\boldsymbol{u}_d^t\boldsymbol{W}_d^{-1}\boldsymbol{u}_d = \frac{1}{\varphi_1\varphi_2 - \varphi_{12}^2}\left(\varphi_2 u_{d1}^2 - 2\varphi_{12}u_{d1}u_{d2} + \varphi_1 u_{d2}^2\right).$$

Therefore for $d' = 1, 2, \ldots, D$ and $j' = 1, 2$,

$$\frac{\partial\left(-\frac{1}{2}\sum_{d=1}^{D}\boldsymbol{u}_d^t\boldsymbol{W}_d^{-1}\boldsymbol{u}_d\right)}{\partial u_{d'j'}} = \begin{cases} \dfrac{-1}{\varphi_1\varphi_2 - \varphi_{12}^2}(\varphi_2 u_{d'1} - \varphi_{12}u_{d'2}) & \text{if } j' = 1 \\ \dfrac{-1}{\varphi_1\varphi_2 - \varphi_{12}^2}(\varphi_1 u_{d'2} - \varphi_{12}u_{d'1}) & \text{if } j' = 2 \end{cases}$$

and hence for $d' = 1, 2, \ldots, D$ and $j' = 1, 2$,

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{u})}{\partial u_{d'j'}} = \begin{cases} \dfrac{-1}{\varphi_1\varphi_2 - \varphi_{12}^2}\left(\varphi_2 u_{d'1} - \varphi_{12} u_{d'2}\right) + \sum_{i=1}^{I_d}(y_{d'ij'} - m_{d'i}p_{d'ij'}) & \text{if } j' = 1 \\ \dfrac{-1}{\varphi_1\varphi_2 - \varphi_{12}^2}\left(\varphi_1 u_{d'2} - \varphi_{12} u_{d'1}\right) + \sum_{i=1}^{I_d}(y_{d'ij'} - m_{d'i}p_{d'ij'}) & \text{if } j' = 2 \end{cases} \tag{4.3}$$

Estimates for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are found by equating all the derivatives defined by (4.2) and (4.3) to 0 and solving the resulting system of equations. Because these equations are non-linear they cannot be solved directly. Instead they can be solved by using a Newton-Raphson algorithm. In order to use the Newton-Raphson method we will also need to work out all the second derivatives of the loglikelihood function.

For $j' = 1, 2$; $j'' = 1, 2$; $q = 1, 2, \ldots, Q_{j'}$ and $q' = 1, 2, \ldots, Q_{j''}$ it can be shown that

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{u})}{\partial\beta_{j''(q')}\partial\beta_{j'(q)}} = \begin{cases} \sum_{d=1}^{D}\sum_{i=1}^{I_d} -x_{dij'(q)}x_{dij'(q')}m_{di}p_{dij'}(1 - p_{dij'}) & \text{if } j' = j'' \\ \sum_{d=1}^{D}\sum_{i=1}^{I_d} x_{dij'(q)}x_{dij''(q')}m_{di}p_{dij'}p_{dij''} & \text{if } j' \neq j''. \end{cases}$$

Also for $d' = 1, 2, \ldots, D$; $d'' = 1, 2, \ldots, D$; $j' = 1, 2$ and $j'' = 1, 2$,

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{u})}{\partial u_{d''j''} \partial u_{d'j'}} = \begin{cases} \dfrac{-\varphi_2}{\varphi_1\varphi_2 - \varphi_{12}^2} - \sum_{i=1}^{I_d} m_{d'i} p_{d'i1}(1 - p_{d'i1}) & \text{if } d'' = d' \text{ and } j'' = j' = 1 \\ \dfrac{-\varphi_1}{\varphi_1\varphi_2 - \varphi_{12}^2} - \sum_{i=1}^{I_d} m_{d'i} p_{d'i2}(1 - p_{d'i2}) & \text{if } d'' = d' \text{ and } j'' = j' = 2 \\ \dfrac{\varphi_{12}}{\varphi_1\varphi_2 - \varphi_{12}^2} + \sum_{i=1}^{I_d} m_{d'i} p_{d'i1} p_{d'i2} & \text{if } d'' = d' \text{ and } j'' \neq j' \\ 0 & \text{if } d'' \neq d', \end{cases}$$

and for $d'' = 1, 2, \ldots, D$; $j'' = 1, 2$; $j' = 1, 2$ and $q = 1, 2, \ldots, Q_{j'}$,

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{u})}{\partial u_{d''j''} \partial \beta_{j'(q)}} = \begin{cases} -\sum_{i=1}^{I_d} x_{d''ij'(q)} m_{d''i} p_{d''ij'}(1 - p_{d''ij'}) & \text{if } j'' = j' \\ \sum_{i=1}^{I_d} x_{d''ij'(q)} m_{d''i} p_{d''ij'} p_{d''ij''} & \text{if } j'' \neq j'. \end{cases}$$

Similar to Molina *et al.* (2007), let $\boldsymbol{\theta}_{di} = (\theta_{di1}, \theta_{di2})^t$, where for $j = 1, 2$,

$$\theta_{dij} = \log \frac{p_{dij}}{p_{di3}}.$$

We can also write (2.2) as

$$\boldsymbol{\theta}_{di} = \boldsymbol{X}_{di}\boldsymbol{\beta} + \boldsymbol{Z}_{di}\boldsymbol{u},$$

where

$$\boldsymbol{X}_{di} = \begin{pmatrix} \boldsymbol{x}_{di1}^t & \mathbf{0}_{1\times Q_2} \\ \mathbf{0}_{1\times Q_1} & \boldsymbol{x}_{di2}^t \end{pmatrix},$$

$$\boldsymbol{Z}_{di} = \begin{pmatrix} \mathbf{0}_{1\times 2(d-1)} & 1 & 0 & \mathbf{0}_{1\times 2(D-d)} \\ \mathbf{0}_{1\times 2(d-1)} & 0 & 1 & \mathbf{0}_{1\times 2(D-d)} \end{pmatrix}$$

and $\mathbf{0}_{a^*\times b^*}$ denotes a matrix of zeros with dimension $a^* \times b^*$.

Denote the mean and covariance matrix of $\boldsymbol{y}_{di}$ given $\boldsymbol{u}$ as $\boldsymbol{\mu}_{di}$ and $\boldsymbol{\Sigma}_{di}$ and these are

$$\boldsymbol{\mu}_{di} = m_{di}(p_{di1}, p_{di2})^t$$

and

$$\boldsymbol{\Sigma}_{di} = m_{di}\begin{pmatrix} p_{di1}(1-p_{di1}) & -p_{di1}p_{di2} \\ -p_{di1}p_{di2} & p_{di2}(1-p_{di2}) \end{pmatrix}.$$

Again similar to Molina *et al.* (2007), let $\boldsymbol{s}_\beta$ and $\boldsymbol{s}_u$ be the vectors of first derivatives of the loglikelihood. That is,

$$\boldsymbol{s}_\beta = \left(\frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial \beta_{1(1)}}, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial \beta_{1(2)}}, \ldots, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial \beta_{1(Q_1)}}, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial \beta_{2(1)}}, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial \beta_{2(2)}}, \ldots, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial \beta_{2(Q_2)}} \right)^t$$

and

$$\boldsymbol{s}_u = \left(\frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial u_{11}}, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial u_{12}}, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial u_{21}}, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial u_{22}}, \ldots, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial u_{D1}}, \frac{\partial l(\boldsymbol{\beta},\boldsymbol{u})}{\partial u_{D2}} \right)^t.$$

In matrix notation we have

$$\boldsymbol{s}_\beta = \sum_{d=1}^{D} \sum_{i=1}^{I_d} \boldsymbol{X}_{di}^t \left(\boldsymbol{y}_{di} - \boldsymbol{\mu}_{di} \right)$$

and

$$\boldsymbol{s}_u = \sum_{d=1}^{D} \sum_{i=1}^{I_d} \boldsymbol{Z}_{di}^t \left(\boldsymbol{y}_{di} - \boldsymbol{\mu}_{di} \right) - \sum_{d=1}^{D} \boldsymbol{Z}_{d1}^t \boldsymbol{W}_d^{-1} \boldsymbol{Z}_{d1} \boldsymbol{u}$$

and the non-linear system of equations that we need to solve is described by

$$\boldsymbol{s} = \left(\boldsymbol{s}_\beta^t, \boldsymbol{s}_u^t \right)^t = \boldsymbol{0}_{(Q_1+Q_2+2D)\times 1}. \tag{4.4}$$

Let

$\boldsymbol{J}_\beta$ be a square symmetric matrix containing all the derivatives of $\boldsymbol{s}_\beta$ with respect to $\boldsymbol{\beta}$,

$\boldsymbol{J}_u$ be a square symmetric matrix containing all the derivatives of $\boldsymbol{s}_u$ with respect to $\boldsymbol{u}$,

$\boldsymbol{J}_{\beta u}$ be the matrix containing all the derivatives of $\boldsymbol{s}_u$ with respect to $\boldsymbol{\beta}$.

It can be shown that

$$J_\beta = -\sum_{d=1}^{D}\sum_{i=1}^{I_d} X_{di}^t \Sigma_{di} X_{di} \,,$$

$$J_{\beta u} = -\sum_{d=1}^{D}\sum_{i=1}^{I_d} X_{di}^t \Sigma_{di} Z_{di}$$

and

$$J_u = -\sum_{d=1}^{D}\sum_{i=1}^{I_d} Z_{di}^t \Sigma_{di} Z_{di} - \sum_{d=1}^{D} Z_{d1}^t W_d^{-1} Z_{d1} \,.$$

The Newton Raphson algorithm can now be applied to find the solution to (4.4). This iterative algorithm has updating equations as follows

$$\begin{pmatrix} \boldsymbol{\beta}^{k+1} \\ \boldsymbol{u}^{k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{k} \\ \boldsymbol{u}^{k} \end{pmatrix} - \begin{pmatrix} J_\beta^k & J_{\beta u}^k \\ (J_{\beta u}^k)^t & J_u^k \end{pmatrix}^{-1} S^k, \tag{4.5}$$

where the superscript k indicates the iteration number and the current values of all parameters are used to evaluate the functions within. Note that in its current form (4.5) contains a large matrix which needs to be inverted at each step. This term can be further simplified by noting the following partitioned matrix identity given in Henderson and Searle (1981)

$$\begin{pmatrix} J_\beta & J_{\beta u} \\ J_{\beta u}^t & J_u \end{pmatrix}^{-1} = \begin{pmatrix} \left(J_\beta - J_{\beta u} J_u^{-1} J_{\beta u}^t\right)^{-1} & -\left(J_\beta - J_{\beta u} J_u^{-1} J_{\beta u}^t\right)^{-1} J_{\beta u} J_u^{-1} \\ -J_u^{-1} J_{\beta u}^t \left(J_\beta - J_{\beta u} J_u^{-1} J_{\beta u}^t\right)^{-1} & J_u^{-1} + J_u^{-1} J_{\beta u}^t \left(J_\beta - J_{\beta u} J_u^{-1} J_{\beta u}^t\right)^{-1} J_{\beta u} J_u^{-1} \end{pmatrix}, \tag{4.6}$$

for any square matrices J_β and J_u with J_u nonsingular and J_β possibly singular. Molina *et al.* (2007) also make use of this identity. Note that this identity simplifies the inversion quite a lot since we no longer need to invert a square matrix of dimension $Q_1 + Q_2 + 2D$. Instead we need to invert $J_\beta - J_{\beta u} J_u^{-1} J_{\beta u}^t$ and J_u. The matrix $J_\beta - J_{\beta u} J_u^{-1} J_{\beta u}^t$ is a square matrix of dimension $Q_1 + Q_2$ which is the total number of explanatory variables in the model and is a lot smaller than $Q_1 + Q_2 + 2D$ since D is large. As for J_u, this is a a square matrix with dimension $2D$ which is still quite large. But note that J_u is a block diagonal matrix with block sizes of 2. So all as we need to do in this case is invert a series of 2 by 2 matrices which is trivial.

So to compute PQL estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ we use the iterative formula (4.5) until convergence (and using the identity given at (4.6)). An initial guess is needed for the parameters to start off the iterations. We suggest using a small value of $\boldsymbol{u}$ as its initial value and an initial value of $\boldsymbol{\beta}$ obtained by fitting a multinomial logit model without random effects.

The criterion for convergence we use is one given in Booth and Hobert (1999)

$$\max \left(\begin{array}{ll} \dfrac{\left|\beta_{j(q)}^{k+1} - \beta_{j(q)}^{k}\right|}{\left|\beta_{j(q)}^{k}\right| + \varepsilon_1} & , j = 1,2 \text{ and } q = 1,2,...,Q_j, \\ & \\ \dfrac{\left|u_{dj}^{k+1} - u_{dj}^{k}\right|}{\left|u_{dj}^{k}\right| + \varepsilon_1} & , d = 1,2,...,D \text{ and } j = 1,2 \end{array} \right) < \varepsilon_2, \qquad (4.7)$$

where ε_1 and ε_2 are both small positive numbers (we use $\varepsilon_1 = 0.01$ and $\varepsilon_2 = 0.001$).

In all of the above it was assumed that $\boldsymbol{\varphi}$ is known. In the next section we derive an approximate maximum likelihood estimator of $\boldsymbol{\varphi}$ given the other terms. To obtain joint estimates of $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{\varphi}$ we will need to iterate between updating each, where $\boldsymbol{\beta}$ and $\boldsymbol{u}$ will be updated using the algorithm in this section.

5. APPROXIMATE ML ESTIMATION OF $\boldsymbol{\varphi}$

It was mentioned in Section 3 that for the normal linear mixed model, given the variance components, estimates of $\boldsymbol{\beta}$ obtained by maximising $L(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{u})$ are equivalent to maximising $L(\boldsymbol{\beta}, \boldsymbol{\varphi})$. However for the normal linear mixed model, given $\boldsymbol{\beta}$, ML estimates for the variance components based on maximising $L(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{u})$ and $L(\boldsymbol{\beta}, \boldsymbol{\varphi})$ are not equivalent. This suggests that it is probably inappropriate to maximise $L(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{u})$ to get estimates of $\boldsymbol{\varphi}$ in our situation too.

To obtain an approximate marginal likelihood $L(\boldsymbol{\beta}, \boldsymbol{\varphi})$, Molina *et al.* (2007) adapted the ideas of Schall (1991) to their bivariate setting. We will follow this approach too.

Assume that $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are known.

Let $$g_j(\boldsymbol{y}_{di}) = \log \frac{y_{dij}}{m_{di} - y_{di1} - y_{di2}} = \log \frac{y_{dij}}{y_{di3}} \qquad \text{for } j = 1, 2.$$

A first order Taylor series expansion about the point $\boldsymbol{\mu}_{di}$ leads to

$$g_j(\boldsymbol{y}_{di}) = g_j(\boldsymbol{\mu}_{di}) + \left.\frac{\partial g_j}{\partial y_{di1}}\right|_{\boldsymbol{\mu}_{di}} (y_{di1} - \mu_{di1}) + \left.\frac{\partial g_j}{\partial y_{di2}}\right|_{\boldsymbol{\mu}_{di}} (y_{di2} - \mu_{di2}), \text{ for } j = 1, 2.$$

Let $\boldsymbol{\xi}_{di} = (g_1(\boldsymbol{y}_{di}), g_2(\boldsymbol{y}_{di}))^t$ and $\boldsymbol{e}_{di} = \boldsymbol{\Sigma}_{di}^{-1}(\boldsymbol{y}_{di} - \boldsymbol{\mu}_{di})$.

Calculating the expressions of the derivatives involved and using matrix notation, the above Taylor series expansion becomes

$$\boldsymbol{\xi}_{di} = \boldsymbol{X}_{di}\boldsymbol{\beta} + \boldsymbol{Z}_{di}\boldsymbol{u} + \boldsymbol{e}_{di},$$

where $Var(\boldsymbol{e}_{di} | \boldsymbol{u}_d) = \boldsymbol{\Sigma}_{di}^{-1}$ and $E(\boldsymbol{e}_{di} | \boldsymbol{u}_d) = \boldsymbol{0}_{2\times 2}$.

It is also clear that $E(\boldsymbol{\xi}_{di} | \boldsymbol{u}_d) = \boldsymbol{X}_{di}\boldsymbol{\beta} + \boldsymbol{Z}_{di}\boldsymbol{u}$ and $Var(\boldsymbol{\xi}_{di} | \boldsymbol{u}_d) = \boldsymbol{\Sigma}_{di}^{-1}$.

Although it is not clear in Molina *et al.* (2007), on correspondence with the author the following was established. The term $\boldsymbol{\xi}_{di} | \boldsymbol{u}_d$ is assumed to be approximately normal. Since $\boldsymbol{u}_d$ is normal, this also implies that the joint distribution of $\boldsymbol{\xi}_{di}$ and $\boldsymbol{u}_d$ is approximately normal and hence the marginal distribution of $\boldsymbol{\xi}_{di}$ is approximately normal.

Let $\boldsymbol{W} = \text{diag}(\boldsymbol{W}_d, d = 1, 2, \ldots, D)$. Now it is easily shown that $E(\boldsymbol{\xi}_{di}) = \boldsymbol{X}_{di}\boldsymbol{\beta}$ and

$$Var(\boldsymbol{\xi}_{di}) = E\left(Var(\boldsymbol{\xi}_{di} | \boldsymbol{u}_d)\right) + Var\left(E(\boldsymbol{\xi}_{di} | \boldsymbol{u}_d)\right) = E\left(\boldsymbol{\Sigma}_{di}^{-1}\right) + \boldsymbol{Z}_{di}\boldsymbol{W}\boldsymbol{Z}_{di}^t.$$

The matrix $\boldsymbol{\Sigma}_{di}$ is a function of the random effects and we now need to assume that this is approximately constant and hence $Var(\boldsymbol{\xi}_{di}) \approx \boldsymbol{\Sigma}_{di}^{-1} + \boldsymbol{Z}_{di}\boldsymbol{W}\boldsymbol{Z}_{di}^t$, where the random effects are replaced by their estimated values.

Let $\boldsymbol{\xi}$ denote the vector that is constructed by stacking the vectors $\boldsymbol{\xi}_{di}$ in one column and $\boldsymbol{V} = Var(\boldsymbol{\xi})$.

Then $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^t + \boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma} = diag(\boldsymbol{\Sigma}_{di}, i = 1, 2, \ldots, I_d, d = 1, 2, \ldots, D)$.

Now we can define the approximated normal log likelihood for $\boldsymbol{\varphi}$ as (ignoring constant terms and assuming $\boldsymbol{\beta}$ is known)

$$l(\boldsymbol{\varphi}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{X}\boldsymbol{\beta})^t \boldsymbol{V}^{-1}(\boldsymbol{\xi} - \boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{X}$ is obtained by stacking the matrices $\boldsymbol{X}_{di}$. Also let $\boldsymbol{Z}$ be the matrix obtained by stacking $\boldsymbol{Z}_{di}$. Given $\boldsymbol{\beta}$, we can now obtain approximate ML estimates of $\boldsymbol{\varphi}$ by maximising $l(\boldsymbol{\varphi})$ with respect to $\boldsymbol{\varphi}$. However, before we do this we need to simplify $l(\boldsymbol{\varphi})$.

To compute the PQL estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$, we use the updating equations (4.5). As noted by Jiang (2007), the following alternative iterative procedure originally proposed by Breslow and Clayton (1993) can also be used to produce the same PQL estimates. That is, for fixed $\boldsymbol{\varphi}$ compute

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{\xi} \tag{5.1}$$

and

$$\hat{\boldsymbol{u}} = \boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\left(\boldsymbol{\xi} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right), \tag{5.2}$$

where given $\boldsymbol{\xi}$ one may first use (5.1) to update $\hat{\boldsymbol{\beta}}$, then use (5.2) to update $\hat{\boldsymbol{u}}$ then update $\boldsymbol{\xi}$ and so on until convergence. Note that we do not use (5.1) and (5.2) to update $\boldsymbol{\beta}$ and $\boldsymbol{u}$ because (4.5) is computationally more convenient. However the form of (5.1) and (5.2) is useful here because we can use these to help simplify $l(\boldsymbol{\varphi})$.

From (5.2), (5.1) and the results on page 446 in Pawitan (2001),

$$\left(\boldsymbol{\xi} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t \boldsymbol{V}^{-1}\left(\boldsymbol{\xi} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{\xi} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{u}}\right)^t\boldsymbol{\Sigma}\left(\boldsymbol{\xi} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{u}}\right) + \hat{\boldsymbol{u}}^t\boldsymbol{W}^{-1}\hat{\boldsymbol{u}}$$

and

$$|\boldsymbol{V}| = \left|\boldsymbol{\Sigma}^{-1}\right|\left|\boldsymbol{W}\right|\left|\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z} + \boldsymbol{W}^{-1}\right|.$$

We can now write an approximated pseudo loglikelihood of $\boldsymbol{\varphi}$ as (ignoring terms that are not functions of $\boldsymbol{\varphi}$)

$$l(\boldsymbol{\varphi}) = -\frac{1}{2}\log|\boldsymbol{W}| - \frac{1}{2}\log\left|\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z} + \boldsymbol{W}^{-1}\right| - \frac{1}{2}\hat{\boldsymbol{u}}^t\boldsymbol{W}^{-1}\hat{\boldsymbol{u}}. \tag{5.3}$$

Given current estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$, to obtain approximate ML estimates of $\boldsymbol{\varphi}$ we need to differentiate (5.3) with respect to $\boldsymbol{\varphi}$, set the resulting three equations to 0 and solve them.

In the derivation of $l(\boldsymbol{\varphi})$ it was assumed that $\boldsymbol{\Sigma}$ was constant and does not depend on $\boldsymbol{u}$ or $\boldsymbol{\beta}$. When this is the case, joint estimation of $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{u}$ is equivalent to maximising the function

$$Q(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{u}) = l(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{u}) - \frac{1}{2}\log\left|\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z} + \boldsymbol{W}^{-1}\right| \tag{5.4}$$

since $$\frac{\partial Q}{\partial \boldsymbol{\beta}} = \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{u})}{\partial \boldsymbol{\beta}}, \frac{\partial Q}{\partial \boldsymbol{u}} = \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{u})}{\partial \boldsymbol{u}} \text{ and } \frac{\partial Q}{\partial \boldsymbol{\varphi}} = \frac{\partial l(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}}.$$

This justifies the following algorithm to obtain joint estimates of $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{u}$

1. Compute $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ given $\boldsymbol{\varphi}$ by using (4.5).
2. Fixing $\boldsymbol{\beta}$ and $\boldsymbol{u}$ at their current values $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$, update $\boldsymbol{\varphi}$ by maximising (5.3).
3. Iterate between 1 and 2 until convergence.

However this algorithm assumes that $\boldsymbol{\Sigma}$ is constant in (5.3). Pawitan (2001) notes that this algorithm is appropriate when $\boldsymbol{\Sigma}^{-1}$is a slowly varying function of $\boldsymbol{\mu}$ ($\boldsymbol{\mu}$ is the vector obtained by stacking all the $\boldsymbol{\mu}_{di}$). This means we can ignore the derivative of the second term of Q with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$, so the first step is justified.

Pawitan (2001) notes that for certain generalised linear mixed models studied by Breslow and Clayton (1993), estimates of the variance components based on maximising (5.4) are close to the exact marginal likelihood estimates provided that the variance component is not too large. The method tends to underestimate the variance component, and the problem can be severe for large values of the variance component. However in our application, the variance components are expected to be small, so hopefully this should not be too much of an issue.

We now need to maximise (5.3) and to do this we need to differentiate this function with respect to each component of $\boldsymbol{\varphi}$. For $a = 1, 2$ and 12,

$$\begin{aligned}\frac{\partial l(\boldsymbol{\varphi})}{\partial \varphi_a} &= -\frac{1}{2}\frac{\partial \log|\boldsymbol{W}|}{\partial \varphi_a} - \frac{1}{2}\frac{\partial \log\left|\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z} + \boldsymbol{W}^{-1}\right|}{\partial \varphi_a} - \frac{1}{2}\frac{\partial \boldsymbol{u}^t\boldsymbol{W}^{-1}\boldsymbol{u}}{\partial \varphi_a} \\ &= -\frac{1}{2}Tr\left[\boldsymbol{W}^{-1}\frac{\partial \boldsymbol{W}}{\partial \varphi_a}\right] - \frac{1}{2}Tr\left[\left(\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z} + \boldsymbol{W}^{-1}\right)^{-1}\frac{\partial \boldsymbol{W}^{-1}}{\partial \varphi_a}\right] - \frac{1}{2}\boldsymbol{u}^t\frac{\partial \boldsymbol{W}^{-1}}{\partial \varphi_a}\boldsymbol{u},\end{aligned} \tag{5.5}$$

where $Tr[\ ]$ denotes matrix trace.

After some algebra, it can be shown that

$$Tr\left[\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \varphi_1}\right]=\frac{D\varphi_2}{\varphi_1\varphi_2-\varphi_{12}^2}, \tag{5.6}$$

$$Tr\left[\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \varphi_2}\right]=\frac{D\varphi_1}{\varphi_1\varphi_2-\varphi_{12}^2} \tag{5.7}$$

and

$$Tr\left[\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \varphi_{12}}\right]=\frac{-2D\varphi_{12}}{\varphi_1\varphi_2-\varphi_{12}^2}\,. \tag{5.8}$$

Also,

$$\boldsymbol{u}^t\frac{\partial \mathbf{W}^{-1}}{\partial \varphi_1}\boldsymbol{u}=\frac{1}{\left(\varphi_1\varphi_2-\varphi_{12}^2\right)^2}\sum_{d=1}^{D}\boldsymbol{u}_d^t\begin{pmatrix}-\varphi_2^2 & \varphi_2\varphi_{12}\\ \varphi_2\varphi_{12} & -\varphi_{12}^2\end{pmatrix}\boldsymbol{u}_d, \tag{5.9}$$

$$\boldsymbol{u}^t\frac{\partial \mathbf{W}^{-1}}{\partial \varphi_2}\boldsymbol{u}=\frac{1}{\left(\varphi_1\varphi_2-\varphi_{12}^2\right)^2}\sum_{d=1}^{D}\boldsymbol{u}_d^t\begin{pmatrix}-\varphi_{12}^2 & \varphi_1\varphi_{12}\\ \varphi_1\varphi_{12} & -\varphi_1^2\end{pmatrix}\boldsymbol{u}_d \tag{5.10}$$

and

$$\boldsymbol{u}^t\frac{\partial \mathbf{W}^{-1}}{\partial \varphi_{12}}\boldsymbol{u}=\frac{1}{\left(\varphi_1\varphi_2-\varphi_{12}^2\right)^2}\sum_{d=1}^{D}\boldsymbol{u}_d^t\begin{pmatrix}2\varphi_{12}\varphi_2 & -\left(\varphi_{12}^2+\varphi_1\varphi_2\right)\\ -\left(\varphi_{12}^2+\varphi_1\varphi_2\right) & 2\varphi_1\varphi_{12}\end{pmatrix}\boldsymbol{u}_d. \tag{5.11}$$

Now for $d=1,2,\ldots,D$ let

$$\begin{aligned}q_{d1}&=\textstyle\sum_{i=1}^{I_d} m_{di}p_{di1}(1-p_{di1}),\\ q_{d2}&=\textstyle\sum_{i=1}^{I_d} m_{di}p_{di2}(1-p_{di2}) \text{ and}\\ q_{d3}&=\textstyle\sum_{i=1}^{I_d} -m_{di}p_{di1}p_{di2}\,.\end{aligned}$$

After some algebra it can be shown that

$$Tr\left[\left(\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z}+\mathbf{W}^{-1}\right)^{-1}\frac{\partial \mathbf{W}^{-1}}{\partial \varphi_1}\right]=$$

$$-\sum_{d=1}^{D}\left(\frac{q_{d1}\varphi_{12}^2+\varphi_2\left(1+2q_{d3}\varphi_{12}+q_{d2}\varphi_2\right)}{\left(-\varphi_{12}^2+\varphi_1\varphi_2\right)\left(\left(1+q_{d3}\varphi_{12}\right)^2+\left(q_{d2}-q_{d3}^2\varphi_1\right)\varphi_2+q_{d1}\left(\varphi_1-q_{d2}\varphi_{12}^2+q_{d2}\varphi_1\varphi_2\right)\right)}\right), \tag{5.12}$$

$$Tr\left[\left(\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z}+\boldsymbol{W}^{-1}\right)^{-1}\frac{\partial \boldsymbol{W}^{-1}}{\partial \varphi_2}\right]=$$

$$-\sum_{d=1}^{D}\left(\frac{\varphi_1+q_{d1}\varphi_1^2+2q_{d3}\varphi_1\varphi_{12}+q_{d2}\varphi_{12}^2}{\left(-\varphi_{12}^2+\varphi_1\varphi_2\right)\left(\left(1+q_{d3}\varphi_{12}\right)^2+\left(q_{d2}-q_{d3}^2\varphi_1\right)\varphi_2+q_{d1}\left(\varphi_1-q_{d2}\varphi_{12}^2+q_{d2}\varphi_1\varphi_2\right)\right)}\right) \tag{5.13}$$

and

$$Tr\left[\left(\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z}+\boldsymbol{W}^{-1}\right)^{-1}\frac{\partial \boldsymbol{W}^{-1}}{\partial \varphi_{12}}\right]=$$

$$-2\sum_{d=1}^{D}\left(\frac{\left(\varphi_{12}+q_{d1}\varphi_1\varphi_{12}+q_{d3}\varphi_{12}^2+q_{d3}\varphi_1\varphi_2+q_{d2}\varphi_{12}\varphi_2\right)}{\left(\varphi_{12}^2-\varphi_1\varphi_2\right)\left(\left(1+q_{d3}\varphi_{12}\right)^2+\left(q_{d2}-q_{d3}^2\varphi_1\right)\varphi_2+q_{d1}\left(\varphi_1-q_{d2}\varphi_{12}^2+q_{d2}\varphi_1\varphi_2\right)\right)}\right). \tag{5.14}$$

By substituting the relevant terms (5.6)–(5.14) into (5.5), we can now calculate the first derivatives of $l(\boldsymbol{\varphi})$ with respect to φ_1, φ_2 and φ_{12}.

Let $$\boldsymbol{S}_\varphi=\left(\frac{\partial l(\boldsymbol{\varphi})}{\partial \varphi_1},\frac{\partial l(\boldsymbol{\varphi})}{\partial \varphi_2},\frac{\partial l(\boldsymbol{\varphi})}{\partial \varphi_{12}}\right)^t.$$

To obtain an update for $\boldsymbol{\varphi}$ given $\boldsymbol{\beta}$ and $\boldsymbol{u}$ we need to solve $\boldsymbol{S}_\varphi=\boldsymbol{0}_{3\times 1}$. The multinomial model described in the paper Molina *et al.* (2007) has one variance component. In this case, an explicit updating formula is available for the variance component based on rearranging the single equation

$$\frac{dl(\varphi)}{d\varphi}=0.$$

An experiment was undertaken to determine whether simple updating equations could be obtained by rearranging the equations $\boldsymbol{S}_\varphi=\boldsymbol{0}_{3\times 1}$ but we could not get it to converge. Unfortunately we will need to use a single iteration of the Newton-Raphson algorithm to update $\boldsymbol{\varphi}$ instead, which means we will also need all the second derivatives of $l(\boldsymbol{\varphi})$.

After some algebra it can be shown that

$$\frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1^2} = -\frac{1}{\left(-\varphi_{12}^2+\varphi_1\varphi_2\right)^3}\sum_{d=1}^{D}\boldsymbol{u}_d^t\begin{pmatrix}\varphi_2^3 & -\varphi_{12}\varphi_2^2\\ -\varphi_{12}\varphi_2^2 & \varphi_{12}^2\varphi_2\end{pmatrix}\boldsymbol{u}_d$$
$$+\sum_{d=1}^{D}\frac{\left(q_{d1}+q_{d1}q_{d2}\varphi_2-q_{d3}^2\varphi_2\right)^2}{2\left(\left(1+q_{d3}\varphi_{12}\right)^2+\left(q_{d2}-q_{d3}^2\varphi_1\right)\varphi_2+q_{d1}\left(\varphi_1-q_{d2}\varphi_{12}^2+q_{d2}\varphi_1\varphi_2\right)\right)^2},$$

$$\frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1\partial \varphi_2} = -\frac{1}{2\left(-\varphi_{12}^2+\varphi_1\varphi_2\right)^3}\sum_{d=1}^{D}\boldsymbol{u}_d^t\begin{pmatrix}2\varphi_{12}^2\varphi_2 & -\left(\varphi_{12}^3+\varphi_1\varphi_{12}\varphi_2\right)\\ -\left(\varphi_{12}^3+\varphi_1\varphi_{12}\varphi_2\right) & 2\varphi_1\varphi_{12}^2\end{pmatrix}\boldsymbol{u}_d$$
$$+\sum_{d=1}^{D}\frac{\left(q_{d3}-q_{d1}q_{d2}\varphi_{12}+q_{d3}^2\varphi_{12}\right)^2}{2\left(\left(1+q_{d3}\varphi_{12}\right)^2+\left(q_{d2}-q_{d3}^2\varphi_1\right)\varphi_2+q_{d1}\left(\varphi_1-q_{d2}\varphi_{12}^2+q_{d2}\varphi_1\varphi_2\right)\right)^2},$$

$$\frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1\partial \varphi_{12}} = -\frac{1}{2\left(-\varphi_{12}^2+\varphi_1\varphi_2\right)^3}\sum_{d=1}^{D}\boldsymbol{u}_d^t\begin{pmatrix}-4\varphi_{12}\varphi_2^2 & \varphi_2\left(3\varphi_{12}^2+\varphi_1\varphi_2\right)\\ \varphi_2\left(3\varphi_{12}^2+\varphi_1\varphi_2\right) & -2\left(\varphi_{12}^3+\varphi_1\varphi_{12}\varphi_2\right)\end{pmatrix}\boldsymbol{u}_d$$
$$-\sum_{d=1}^{D}\frac{\left(q_{d3}-q_{d1}q_{d2}\varphi_{12}+q_{d3}^2\varphi_{12}\right)\left(q_{d3}^2\varphi_2-q_{d1}\left(1+q_{d2}\varphi_2\right)\right)}{\left(\left(1+q_{d3}\varphi_{12}\right)^2+\left(q_{d2}-q_{d3}^2\varphi_1\right)\varphi_2+q_{d1}\left(\varphi_1-q_{d2}\varphi_{12}^2+q_{d2}\varphi_1\varphi_2\right)\right)^2},$$

$$\frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_2^2} = -\frac{1}{\left(-\varphi_{12}^2+\varphi_1\varphi_2\right)^3}\sum_{d=1}^{D}\boldsymbol{u}_d^t\begin{pmatrix}\varphi_1\varphi_{12}^2 & -\varphi_1^2\varphi_{12}\\ -\varphi_1^2\varphi_{12} & \varphi_1^3\end{pmatrix}\boldsymbol{u}_d$$
$$+\sum_{d=1}^{D}\frac{\left(q_{d2}+q_{d1}q_{d2}\varphi_1-q_{d3}^2\varphi_1\right)^2}{2\left(\left(1+q_{d3}\varphi_{12}\right)^2+\left(q_{d2}-q_{d3}^2\varphi_1\right)\varphi_2+q_{d1}\left(\varphi_1-q_{d2}\varphi_{12}^2+q_{d2}\varphi_1\varphi_2\right)\right)^2},$$

$$\frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_2\partial \varphi_{12}} = -\frac{1}{2\left(-\varphi_{12}^2+\varphi_1\varphi_2\right)^3}\sum_{d=1}^{D}\boldsymbol{u}_d^t\begin{pmatrix}-2\left(\varphi_{12}^3+\varphi_1\varphi_{12}\varphi_2\right) & \varphi_1\left(3\varphi_{12}^2+\varphi_1\varphi_2\right)\\ \varphi_1\left(3\varphi_{12}^2+\varphi_1\varphi_2\right) & -4\varphi_1^2\varphi_{12}\end{pmatrix}\boldsymbol{u}_d$$
$$-\sum_{d=1}^{D}\frac{\left(q_{d2}+q_{d1}q_{d2}\varphi_1-q_{d3}^2\varphi_1\right)\left(q_{d1}q_{d2}\varphi_{12}-q_{d3}\left(1+q_{d3}\varphi_{12}\right)\right)}{\left(\left(1+q_{d3}\varphi_{12}\right)^2+\left(q_{d2}-q_{d3}^2\varphi_1\right)\varphi_2+q_{d1}\left(\varphi_1-q_{d2}\varphi_{12}^2+q_{d2}\varphi_1\varphi_2\right)\right)^2}$$

$$\frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_{12}^2} = -\frac{1}{\left(-\varphi_{12}^2 + \varphi_1\varphi_2\right)^3} \sum_{d=1}^{D} \boldsymbol{u}_d^t \begin{pmatrix} \varphi_2\left(3\varphi_{12}^2 + \varphi_1\varphi_2\right) & -\left(\varphi_{12}^3 + 3\varphi_1\varphi_{12}\varphi_2\right) \\ -\left(\varphi_{12}^3 + 3\varphi_1\varphi_{12}\varphi_2\right) & \varphi_1\left(3\varphi_{12}^2 + \varphi_1\varphi_2\right) \end{pmatrix} \boldsymbol{u}_d$$

$$+\sum_{d=1}^{D} \frac{2\left(q_{d3} - q_{d1}q_{d2}\varphi_{12} + q_{d3}^2\varphi_{12}\right)^2}{\left(\left(1 + q_{d3}\varphi_{12}\right)^2 + \left(q_{d2} - q_{d3}^2\varphi_1\right)\varphi_2 + q_{d1}\left(\varphi_1 - q_{d2}\varphi_{12}^2 + q_{d2}\varphi_1\varphi_2\right)\right)^2}$$

$$-\sum_{d=1}^{D} \frac{-\left(q_{d1}q_{d2}\right) + q_{d3}^2}{\left(1 + q_{d3}\varphi_{12}\right)^2 + \left(q_{d2} - q_{d3}^2\varphi_1\right)\varphi_2 + q_{d1}\left(\varphi_1 - q_{d2}\varphi_{12}^2 + q_{d2}\varphi_1\varphi_2\right)}.$$

Now let

$$\boldsymbol{J}_{\varphi} = \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial^2 \varphi_1} & \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1 \partial \varphi_2} & \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1 \partial \varphi_{12}} \\ \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1 \partial \varphi_2} & \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial^2 \varphi_2} & \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_2 \partial \varphi_{12}} \\ \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1 \partial \varphi_{12}} & \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_2 \partial \varphi_{12}} & \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial^2 \varphi_{12}} \end{pmatrix}.$$

To obtain an update of $\boldsymbol{\varphi}$ given $\boldsymbol{\beta}$ and $\boldsymbol{u}$ we simply compute

$$\boldsymbol{\varphi}^{\text{update}} = \boldsymbol{\varphi}^{\text{previous}} - \boldsymbol{J}_{\varphi}^{-1}\boldsymbol{S}_{\varphi},$$

where all terms on the right hand side are evaluated using the current values. Only one update is needed because after $\boldsymbol{\varphi}$ is updated once we then have to go back to updating $\boldsymbol{\beta}$ and $\boldsymbol{u}$ by using repeats of (4.5) until convergence given the current value of $\boldsymbol{\varphi}$.

The iterative algorithm for computing joint estimates of $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{\varphi}$ stops when both (4.7) and the following condition is satisfied for all three variance components

$$\frac{\left|\varphi_a^{\text{update}} - \varphi_a^{\text{previous}}\right|}{\left|\varphi_a^{\text{previous}}\right| + \varepsilon_1} < \varepsilon_2,$$

where subscript a denotes the particular component.

6. APPROXIMATE REML ESTIMATION OF φ

Approximate REML (Restricted Maximum Likelihood) estimation could also be used as an alternative to approximate ML estimation for estimating $\boldsymbol{\varphi}$. This approach was also undertaken by Molina *et al.* (2007).

For normal linear mixed models, it is well known that in some cases the ML estimator of the variance components can be downwardly biased (see page 235 of Rao and Kleffe (1988)). As mentioned in Harville (1977), one criticism of the ML approach to estimating variance components is that the ML estimator takes no account of the loss of degrees of freedom for estimating the fixed effects $\boldsymbol{\beta}$. For estimating the variance components, $\boldsymbol{\beta}$ can be considered as a nuisance parameter. A way of eliminating the influence of $\boldsymbol{\beta}$ is to construct a marginal (or approximated marginal) ML estimator for the variance components. In the normal linear mixed model case, this involves transforming the original observations such that the new data are independent of $\boldsymbol{\beta}$ and then maximising the likelihood for the variance components of this new data. The transformed data are called error contrasts as described in Harville (1977) and have smaller dimension than the original data.

In the normal linear mixed model case, the REML loglikelihood can be derived as a modified profile likelihood (see pages 286–292 in Pawitan (2001)). In our case the approximated REML loglikelihood is proportional to

$$l(\boldsymbol{\varphi}) - \frac{1}{2}\log\left|\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right|, \tag{6.1}$$

where $l(\boldsymbol{\varphi})$ is given by (5.3) and as Pawitan (2001) mentions, the second term can be interpreted as a penalty term, subtracting from the profile loglikelihood the 'undeserved' information on the nuisance parameter (which is $\boldsymbol{\beta}$). To obtain approximate REML estimates of $\boldsymbol{\varphi}$ in our case, we simply need to maximise (6.1) which can be done using similar methods to what is described in Section 5. That is, we need to implement the same Newton-Raphson algorithm except that now $\boldsymbol{S}_\varphi$ and $\boldsymbol{J}_\varphi$ contain extra terms. These extra terms can be obtained by finding all the first and second derivatives of the second term in (6.1) with respect to all the variance components.

Although we are potentially reducing bias by using the approximated REML estimator we should keep in mind that the variance may be larger than the variance of the approximated ML estimator. Therefore in some situations the MSE of the approximated ML estimator might be smaller. See Harville (1977) for further details for a discussion in relation to normal linear mixed models.

In any case we need all the first and second derivatives of $\log|\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}|$.

For $a = 1, 2$ and 12,

$$\frac{\partial \log\left|X^t V^{-1} X\right|}{\partial \varphi_a} = Tr\left[\left(X^t V^{-1} X\right)^{-1} X^t \frac{\partial V^{-1}}{\partial \varphi_a} X\right].$$

Now $$\frac{\partial V^{-1}}{\partial \varphi_a} = -V^{-1}\frac{\partial V}{\partial \varphi_a}V^{-1} = -V^{-1} Z \frac{\partial W}{\partial \varphi_a} Z^t V^{-1}$$

and therefore

$$\frac{\partial \log\left|X^t V^{-1} X\right|}{\partial \varphi_a} = -Tr\left[\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Z \frac{\partial W}{\partial \varphi_a} Z^t V^{-1} X\right], \tag{6.2}$$

where $$\frac{\partial W}{\partial \varphi_a} = \text{diag}\left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right) \text{ or } \text{diag}\left(\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\right) \text{ or } \text{diag}\left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right),$$

depending on which value of a is chosen.

On first glimpse (6.2) looks quite difficult to calculate, but note that the matrix within the trace is only a $Q_1 + Q_2$ dimensional matrix in total. Because V and W are block diagonal (V has $d = 1, 2, \ldots, D$ blocks each of size $2I_d$) simplifications can also be made within the matrix multiplications in the calculations.

Now we need to calculate the second derivatives. For $a = 1, 2, 12$ and $b = 1, 2, 12$,

$$\begin{aligned}
\frac{\partial^2 \log\left|X^t V^{-1} X\right|}{\partial \varphi_b \partial \varphi_a} &= -2\,Tr\left[\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Z \frac{\partial W}{\partial \varphi_a} Z^t \frac{\partial V^{-1}}{\partial \varphi_b} X\right] \\
&-Tr\left[\frac{\partial\left(X^t V^{-1} X\right)^{-1}}{\partial \varphi_b} X^t V^{-1} Z \frac{\partial W}{\partial \varphi_a} Z^t V^{-1} X\right] \\
&= 2\,Tr\left[\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Z \frac{\partial W}{\partial \varphi_a} Z^t V^{-1} Z \frac{\partial W}{\partial \varphi_b} Z^t V^{-1} X\right] \\
&+Tr\left[\left(X^t V^{-1} X\right)^{-1} X^t \frac{\partial V^{-1}}{\partial \varphi_b} X\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Z \frac{\partial W}{\partial \varphi_a} Z^t V^{-1} X\right] \\
&= 2\,Tr\left[\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Z \frac{\partial W}{\partial \varphi_a} Z^t V^{-1} Z \frac{\partial W}{\partial \varphi_b} Z^t V^{-1} X\right] \\
&-Tr\left[\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Z \frac{\partial W}{\partial \varphi_b} Z^t V^{-1} X\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Z \frac{\partial W}{\partial \varphi_a} Z^t V^{-1} X\right].
\end{aligned}$$

Now let

$$S_{\varphi}^{*} = S_{\varphi} - \frac{1}{2}\left(\frac{\partial \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{1}}, \frac{\partial \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{2}}, \frac{\partial \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{12}} \right)^{t}$$

and

$$J_{\varphi}^{*} = J_{\varphi} - \frac{1}{2}\begin{pmatrix} \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial^{2}\varphi_{1}} & \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{1}\partial \varphi_{2}} & \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{1}\partial \varphi_{12}} \\ \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{1}\partial \varphi_{2}} & \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial^{2}\varphi_{2}} & \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{2}\partial \varphi_{12}} \\ \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{1}\partial \varphi_{12}} & \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial \varphi_{2}\partial \varphi_{12}} & \frac{\partial^{2} \log\left|X^{t}V^{-1}X\right|}{\partial^{2}\varphi_{12}} \end{pmatrix},$$

where S_{φ} and J_{φ} are as defined in Section 5.

To obtain a REML update of $\boldsymbol{\varphi}$ given $\boldsymbol{\beta}$ and $\boldsymbol{u}$ we compute

$$\boldsymbol{\varphi}^{\text{update}} = \boldsymbol{\varphi}^{\text{previous}} - (J_{\varphi}^{*})^{-1} S_{\varphi}^{*},$$

where all terms on the right hand side are evaluated using the current values.

7. EMPIRICAL BEST PREDICTION, SMALL AREA ESTIMATION AND A NOTE ON MSE ESTIMATION

Now that we have derived the PQL and REML estimators of the model parameters, we would like to use these to help predict the small area totals of employed, unemployed and not in the labour force. These sets of totals are, for $j = 1, 2, 3$ and $d = 1, 2, \ldots, D$,

$$\delta_{dj} = \sum_{i=1}^{I_d} y_{dij} + \sum_{i=1}^{I} y^r_{dij}$$

where y^r_{dij} denotes the total non sample in small area d, age/sex group i and labour force class j. Note that $I = 10$ (total number of age/sex classes within a small area). In previous sections, summations were defined using I_d and classes $I_d + 1, \ldots, I$ were excluded because they did not contribute to the likelihood functions since there was no sample in these classes. However, we now need to include these.

Obviously we cannot use δ_{dj} directly because all the y^r_{dij} are unknown. If we knew the values of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ then we could estimate δ_{dj} using

$$\tilde{\delta}_{dj} = \sum_{i=1}^{I_d} y_{dij} + \sum_{i=1}^{I} \mu^r_{dij}$$

where

$$\mu^r_{dij} = \begin{cases} m^r_{di} \dfrac{e^{\boldsymbol{x}^t_{di1}\boldsymbol{\beta}_1 + \boldsymbol{u}_{d1}}}{1 + e^{\boldsymbol{x}^t_{di1}\boldsymbol{\beta}_1 + \boldsymbol{u}_{d1}} + e^{\boldsymbol{x}^t_{di2}\boldsymbol{\beta}_2 + \boldsymbol{u}_{d2}}} & \text{if } j = 1 \\ m^r_{di} \dfrac{e^{\boldsymbol{x}^t_{di2}\boldsymbol{\beta}_2 + \boldsymbol{u}_{d2}}}{1 + e^{\boldsymbol{x}^t_{di1}\boldsymbol{\beta}_1 + \boldsymbol{u}_{d1}} + e^{\boldsymbol{x}^t_{di2}\boldsymbol{\beta}_2 + \boldsymbol{u}_{d2}}} & \text{if } j = 2 \\ m^r_{di} \dfrac{1}{1 + e^{\boldsymbol{x}^t_{di1}\boldsymbol{\beta}_1 + \boldsymbol{u}_{d1}} + e^{\boldsymbol{x}^t_{di2}\boldsymbol{\beta}_2 + \boldsymbol{u}_{d2}}} & \text{if } j = 3, \end{cases}$$

and m^r_{di} is the total non-sample in sex/age group i, in small area d and these are assumed known. Again $\tilde{\delta}_{dj}$ cannot be used here directly since $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are unknown.

The prediction problem we now have is one where we need to predict

$$\tau_{dj} = \sum_{i=1}^{I} \mu^r_{dij},$$

for $j = 1, 2$ ($j = 3$ is not needed since it can be obtained via subtraction since m^r_{di} is known). There are two ways to predict τ_{dj}. The first is to simply replace $\boldsymbol{u}$ with the PQL estimate $\hat{\boldsymbol{u}}$ and replace $\boldsymbol{\beta}$ with the PQL estimate $\hat{\boldsymbol{\beta}}$.

That is,

$$\hat{\tau}_{dj} = \sum_{i=1}^{I} \hat{\mu}_{dij}^{r}, \qquad (7.1)$$

where the hat on μ_{dij}^{r} means that we are using estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ within this function. This is the approach taken by Molina *et al.* (2007). An alternative way of predicting τ_{dj} is to use an empirical best predictor (EBP) as defined by Jiang (2007, pp. 143–144). Assuming that $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ are known, the best predictor in the sense of minimum MSE of τ_{dj} would be

$$E\left(\tau_{dj} \middle| \boldsymbol{y}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tau_{dj} f\left(\boldsymbol{u}_d \middle| \boldsymbol{\beta}, \boldsymbol{\varphi}; \boldsymbol{y}_d\right) du_{d1} du_{d2} .$$

An empirical best predictor replaces $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ in the integral above with respectively the PQL and REML estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varphi}}$. The double integral above has no closed form solution but we could approximate it using Monte Carlo methods.

Because we need to approximate the double integrals in EBP using Monte Carlo methods, this adds to the computation time and MSE estimation becomes a problem. For instance, bootstrap MSEs (and those based on other resampling methods) are not computationally feasible for EBP. Alternatively Jiang (2007) on page 144 describes a method of approximating the MSEs of EBP based on a Taylor Series expansion that gives an estimate whose bias is corrected to the second order. However in the derivation it is assumed that the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ have certain properties which PQL estimators may not satisfy (see page 158). Therefore these MSE estimates may not be correct to second order in our context. Another problem with this MSE estimator is that one of its terms relates to the expected value of the EBP squared over the distribution of $\boldsymbol{y}$ and this calculation is not computationally feasible for our model.

One advantage of using the estimator at (7.1) is that MSE estimation is relatively straight forward. Bootstrap MSEs are computationally feasible or alternatively, estimators based on Taylor series approximations can be used. In particular the Taylor series approximation method in Molina *et al.* (2007) could be extended to our category specific multinomial mixed model case. In any case it is not even clear whether the MSEs will be smaller for the EBP than those for (7.1). Therefore at this stage EBP is not recommended.

8. ANALYTICAL APPROXIMATION OF THE MSE

For reasons discussed in the previous section we will be using the PQL estimators of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ along with (7.1) to predict the small area totals. To approximate the MSEs of these small area total estimates, we use a similar approach to the one described in Molina *et al.* (2007) based on an analytical approximation. We begin by briefly outlining this approach and then later in this section we describe the technical details of the MSE approximation.

As mentioned in Molina *et al.* (2007), under linear mixed models, Prasad and Rao (1990) obtained an analytical approximation of the MSE of an estimator of the type $t(\hat{\boldsymbol{v}}) = \boldsymbol{\lambda}^t\hat{\boldsymbol{\beta}} + \boldsymbol{m}^t\hat{\boldsymbol{u}}$ where $\boldsymbol{\lambda}$ and $\boldsymbol{m}$ are vectors of constants and $\boldsymbol{v}$ and $\hat{\boldsymbol{v}}$ are respectively the vector of variance components and the estimated vector of variance components. The Prasad and Rao (1990) approximation takes the form

$$MSE\left(t(\hat{\boldsymbol{v}})\right) = G_1(\boldsymbol{v}) + G_2(\boldsymbol{v}) + G_3(\boldsymbol{v})$$

and the estimator is given by

$$\widehat{MSE}\left(t(\hat{\boldsymbol{v}})\right) = G_1(\hat{\boldsymbol{v}}) + G_2(\hat{\boldsymbol{v}}) + 2G_3(\hat{\boldsymbol{v}})$$

which corrects for the bias in the $G_1(\hat{\boldsymbol{v}})$ term. In the derivation, Prasad and Rao used a result from Kackar and Harville (1984) who showed that under certain conditions

$$MSE\left(t(\hat{\boldsymbol{v}})\right) = MSE(t(\boldsymbol{v})) + E((t(\hat{\boldsymbol{v}}) - t(\boldsymbol{v}))^2).$$

In the Prasad and Rao context

$$MSE(t(\boldsymbol{v})) = G_1(\boldsymbol{v}) + G_2(\boldsymbol{v}) \text{ and } G_3(\boldsymbol{v}) = E((t(\hat{\boldsymbol{v}}) - t(\boldsymbol{v}))^2)$$

and Prasad and Rao (1990) proposed a new approximation to $G_3(\boldsymbol{v})$ relevant in a small area estimation context.

Prasad and Rao's formula was adapted by Baillo and Molina (2005) to a multivariate mixed linear model and a multidimensional parameter. Both the Prasad and Rao (1990) and the Baillo and Molina (2005) MSE estimators are for linear mixed models. These estimators can be adapted to our context by noting that our model can be written as an approximate bivariate linear mixed model (see Section 5 for further details). In our context we use the predictions (7.1) for the non-sample labour force counts and in order to apply the Prasad and Rao (1990) and Baillo and Molina (2005) MSE approximations, we need to linearise (7.1) using a first order Taylor series

approximation (Molina *et al.* (2007) use a similar approach). This then results in a MSE approximation of the form $\boldsymbol{G}_1(\boldsymbol{\varphi}) + \boldsymbol{G}_2(\boldsymbol{\varphi}) + \boldsymbol{G}_3(\boldsymbol{\varphi})$. There is also an additional term needed called $\boldsymbol{G}_4(\boldsymbol{\varphi})$ which is added to the above MSE approximation. This extra term comes from the fact that we are estimating the actual non-sample counts $\sum_i y^r_{dij}$ and not simply τ^r_{dj}which introduces additional variability (this extra term is also found in the MSE approximation in Molina *et al.* (2007)).

Note that the multivariate linear mixed model in Baillo and Molina (2005) is different from our approximate multivariate linear mixed model. In Baillo and Molina (2005) there is only one random effect associated with each small area and we have two correlated random effects in each small area. Therefore the formulas in Baillo and Molina (2005) cannot be immediately applied to our case. Nonetheless we are still able to produce a MSE estimator but note that it cannot be guaranteed to be accurate to a known order even if the multivariate linear mixed model was not an approximation (a more detailed and rigorous proof along similar lines to pages 7–17 in Baillo and Molina (2005) would be needed for that). In any case, we show in later sections that our MSE estimator performs well. We now derive this MSE estimator.

For $d = 1, 2, \ldots, D$, let $\boldsymbol{\delta}_d = (\delta_{d1}, \delta_{d2})^t$ be the vector of small area totals that we are interested in predicting (note that the third total within each small area can be obtained via subtraction). Now

$$\boldsymbol{\delta}_d = \sum_{i=1}^{I} \boldsymbol{y}_{di} + \sum_{i=1}^{I} \boldsymbol{\mu}^r_{di} + \sum_{i=1}^{I} \left(\boldsymbol{y}^r_{di} - \boldsymbol{\mu}^r_{di} \right),$$

where $\boldsymbol{y}_{di}$ and $\boldsymbol{y}^r_{di}$ are respectively the vectors of sample totals (known) and non-sample totals (unknown) and

$$\boldsymbol{\mu}^r_{di} = m^r_{di} \left(p_{di1}, p_{di2} \right)^t, \tag{8.1}$$

where m^r_{di} is the total non-sample which is assumed known.

Now define

$$\tilde{\boldsymbol{\delta}}_d = \sum_{i=1}^{I} \boldsymbol{y}_{di} + \sum_{i=1}^{I} \boldsymbol{\mu}^r_{di} = \sum_{i=1}^{I} \boldsymbol{y}_{di} + \boldsymbol{\tau}_d$$

and

$$\hat{\boldsymbol{\delta}}_d = \sum_{i=1}^{I} \boldsymbol{y}_{di} + \sum_{i=1}^{I} \hat{\boldsymbol{\mu}}^r_{di} = \sum_{i=1}^{I} \boldsymbol{y}_{di} + \hat{\boldsymbol{\tau}}_d, \tag{8.2}$$

where $\hat{\boldsymbol{\mu}}^r_{di}$ is (8.1) with the $\boldsymbol{u}$'s and $\boldsymbol{\beta}$'s in p_{di1} and p_{di2} replaced with the PQL estimates $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\beta}}$.

The estimator $\hat{\boldsymbol{\delta}}_d$ is used to estimate $\boldsymbol{\delta}_d$. We will define the MSE matrix of $\hat{\boldsymbol{\delta}}_d$ as

$$\begin{aligned} MSE\left(\hat{\boldsymbol{\delta}}_d\right) &= E\left(\left(\hat{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right) \\ &= E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d + \tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d + \tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right) \\ &= E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\right) + E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right). \end{aligned} \tag{8.3}$$

The MSE matrix contains $MSE\left(\hat{\delta}_{d1}\right)$ and $MSE\left(\hat{\delta}_{d2}\right)$ and the cross product term

$$E\left(\left(\hat{\delta}_{d1} - \delta_{d1}\right)\left(\hat{\delta}_{d2} - \delta_{d2}\right)\right).$$

Note that the terms

$$E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\right)$$

and

$$E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right)$$

are omitted from (8.3) because these are matrices containing all zeros. This is because given $\boldsymbol{u}_d$, $\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d$ and $\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d$ are independent.

Hence $$E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\right) = E\left(E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\middle|\boldsymbol{u}_d\right)E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\middle|\boldsymbol{u}_d\right)\right)$$

and note that $$E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\middle|\boldsymbol{u}_d\right) = 0_{2\times 1}.$$

By a similar argument,

$$E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right) = 0_{2\times 2}.$$

Now

$$E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\right) = E\left(\left(\hat{\boldsymbol{\tau}}_d - \boldsymbol{\tau}_d\right)\left(\hat{\boldsymbol{\tau}}_d - \boldsymbol{\tau}_d\right)^t\right) = MSE\left(\hat{\boldsymbol{\tau}}_d\right)$$

and

$$E\left(\left(\tilde{\boldsymbol{\delta}}_d-\boldsymbol{\delta}_d\right)\left(\tilde{\boldsymbol{\delta}}_d-\boldsymbol{\delta}_d\right)^t\right)=E\left(\left(\sum_{i=1}^{I}\left(\boldsymbol{y}_{di}^r-\boldsymbol{\mu}_{di}^r\right)\right)\left(\sum_{i=1}^{I}\left(\boldsymbol{y}_{di}^r-\boldsymbol{\mu}_{di}^r\right)\right)^t\right)$$

$$=E\left(E\left(\left(\sum_{i=1}^{I}\left(\boldsymbol{y}_{di}^r-\boldsymbol{\mu}_{di}^r\right)\right)\left(\sum_{i=1}^{I}\left(\boldsymbol{y}_{di}^r-\boldsymbol{\mu}_{di}^r\right)\right)^t\middle|\boldsymbol{u}_d\right)\right)$$

$$=E\left(\sum_{i=1}^{I}\boldsymbol{\Sigma}_{di}^r\right), \tag{8.4}$$

where $$\boldsymbol{\Sigma}_{di}^r=m_{di}^r\begin{pmatrix}p_{di1}(1-p_{di1}) & -p_{di1}p_{di2}\\ -p_{di1}p_{di2} & p_{di2}(1-p_{di2})\end{pmatrix}.$$

The term $E\left(\sum_{i=1}^{I}\boldsymbol{\Sigma}_{di}^r\right)$ cannot be further simplified because $\boldsymbol{\Sigma}_{di}^r$ is a non-linear function of the random effects and the expectation involves an integral with no closed form solution. To estimate this term we can use (as in Molina *et al.* (2007))

$$\mathbf{G}_4(\hat{\varphi})=\sum_{i=1}^{I}\hat{\boldsymbol{\Sigma}}_{di}^r\,,$$

where $\hat{\boldsymbol{\Sigma}}_{di}^r$ is $\boldsymbol{\Sigma}_{di}^r$ with the $\boldsymbol{u}$'s and $\boldsymbol{\beta}$'s within replaced by the PQL estimates $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\beta}}$. Now we need to further simplify and approximate the matrix $MSE(\hat{\boldsymbol{\tau}}_d)$.

By definition, $$\hat{\boldsymbol{\tau}}_d=\sum_{i=1}^{I}\hat{\boldsymbol{\mu}}_{di}^r$$

and each of the $\hat{\boldsymbol{\mu}}_{di}^r$ can be written as functions of

$$\hat{\boldsymbol{\theta}}_{di}=\left(\hat{\theta}_{di1},\hat{\theta}_{di2}\right)^t$$

where $$\hat{\boldsymbol{\theta}}_{di}=\boldsymbol{X}_{di}\hat{\boldsymbol{\beta}}+\boldsymbol{Z}_{di}\hat{\boldsymbol{u}}\,.$$

We can now calculate a first order Taylor series approximation for each of the vectors $\hat{\boldsymbol{\mu}}_{di}^r$ about the point $\boldsymbol{\theta}_{di}$. This approximation is

$$\hat{\boldsymbol{\mu}}_{di}^r\approx\boldsymbol{\mu}_{di}^r+\boldsymbol{\Sigma}_{di}^r\left(\hat{\boldsymbol{\theta}}_{di}-\boldsymbol{\theta}_{di}\right)$$

and hence

$$\begin{aligned}\hat{\boldsymbol{\tau}}_d - \boldsymbol{\tau}_d &\approx \sum_{i=1}^{I} \boldsymbol{\Sigma}_{di}^r \hat{\boldsymbol{\theta}}_{di} - \sum_{i=1}^{I} \boldsymbol{\Sigma}_{di}^r \boldsymbol{\theta}_{di} \\ &= \sum_{i=1}^{I} \boldsymbol{\Sigma}_{di}^r \left(\boldsymbol{X}_{di} \hat{\boldsymbol{\beta}} + \boldsymbol{Z}_{di} \hat{\boldsymbol{u}} \right) - \sum_{i=1}^{I} \boldsymbol{\Sigma}_{di}^r \left(\boldsymbol{X}_{di} \boldsymbol{\beta} + \boldsymbol{Z}_{di} \boldsymbol{u} \right) \\ &= \hat{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d,\end{aligned}$$

where $\boldsymbol{\tau}'_d = \boldsymbol{K}_d \boldsymbol{\beta} + \boldsymbol{M}_d \boldsymbol{u}$

and $\hat{\boldsymbol{\tau}}'_d = \boldsymbol{K}_d \hat{\boldsymbol{\beta}} + \boldsymbol{M}_d \hat{\boldsymbol{u}},$

where $\boldsymbol{K}_d = \sum_{i=1}^{I} \boldsymbol{\Sigma}_{di}^r \boldsymbol{X}_{di}$

and $\boldsymbol{M}_d = \sum_{i=1}^{I} \boldsymbol{\Sigma}_{di}^r \boldsymbol{Z}_{di}$.

Technically $\boldsymbol{K}_d$ and $\boldsymbol{M}_d$ are random variables since $\boldsymbol{\Sigma}_{di}$ is dependent on $\boldsymbol{u}_d$. However, from this point forward we need to assume that $\boldsymbol{K}_d$ and $\boldsymbol{M}_d$ are constant and do not depend on $\boldsymbol{u}_d$. Note that this will be a reasonable assumption if the variance components are small. Now, using the fact that $\hat{\boldsymbol{\tau}}_d - \boldsymbol{\tau}_d \approx \hat{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d$, it follows that $MSE(\hat{\boldsymbol{\tau}}_d) \approx MSE(\hat{\boldsymbol{\tau}}'_d)$ (Molina *et al.* (2007) also make this assumption) and therefore

$$\begin{aligned}MSE\left(\hat{\boldsymbol{\tau}}_d\right) &\approx E\left(\left(\hat{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)\left(\hat{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)^t\right) \\ &= E\left(\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d + \tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d + \tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)^t\right), \qquad (8.5)\end{aligned}$$

where $\tilde{\boldsymbol{\tau}}'_d = \boldsymbol{K}_d \tilde{\boldsymbol{\beta}} + \boldsymbol{M}_d \tilde{\boldsymbol{u}}$

and $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$ are the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ assuming that the variance components $\boldsymbol{\varphi}_1$, $\boldsymbol{\varphi}_2$ and $\boldsymbol{\varphi}_{12}$ are known.

As noted in Section 5, the PQL estimators can be obtained via equations (5.1) and (5.2). So we can write

$$\hat{\boldsymbol{\beta}} = \left(X^t \hat{V}^{-1} X\right)^{-1} X^t \hat{V}^{-1} \boldsymbol{\xi},$$

$$\hat{\boldsymbol{u}} = \hat{W} Z^t \hat{V}^{-1} \left(\boldsymbol{\xi} - X\hat{\boldsymbol{\beta}}\right),$$

$$\tilde{\boldsymbol{\beta}} = \left(X^t V^{-1} X\right)^{-1} X^t V^{-1} \boldsymbol{\xi} \tag{8.6}$$

and

$$\tilde{\boldsymbol{u}} = W Z^t V^{-1} \left(\boldsymbol{\xi} - X\tilde{\boldsymbol{\beta}}\right), \tag{8.7}$$

where $V = ZWZ^t + \Sigma^{-1}$ and $\hat{V} = Z\widehat{W}Z^t + \hat{\Sigma}^{-1}$. Note that $\widehat{W}$ is W but with φ_1, φ_2 and φ_{12} within replaced with either the approximated ML or REML estimates. Also, $\hat{\Sigma}^{-1}$ is just Σ^{-1} but with $\boldsymbol{\beta}$ and $\boldsymbol{u}$ within replaced with the PQL estimates (which are also technically functions of $\hat{\varphi}_1$, $\hat{\varphi}_2$ and $\hat{\varphi}_{12}$).

Baillo and Molina (2005) apply a result which was derived by Kackar and Harville (1984) in a linear mixed model setting. We will also apply this result and assume that the estimator of the variance components is translation invariant. This assumption means that we can further simplify (8.5) to

$$\begin{aligned} MSE\left(\hat{\tau}'_d\right) &= E\left(\left(\hat{\tau}'_d - \tilde{\tau}'_d\right)\left(\hat{\tau}'_d - \tilde{\tau}'_d\right)^t\right) + E\left(\left(\tilde{\tau}'_d - \tau'_d\right)\left(\tilde{\tau}'_d - \tau'_d\right)^t\right) \\ &= E\left(\left(\hat{\tau}'_d - \tilde{\tau}'_d\right)\left(\hat{\tau}'_d - \tilde{\tau}'_d\right)^t\right) + MSE\left(\tilde{\tau}'_d\right). \end{aligned} \tag{8.8}$$

To approximate the first term in (8.8) we note that $\hat{\tau}'_d = \tilde{\tau}'_d(\hat{\varphi})$ and $\tilde{\tau}'_d = \tilde{\tau}'_d(\varphi)$, where $\varphi = (\varphi_1, \varphi_2, \varphi_{12})^t$. This only works if we assume that $\hat{\Sigma}_{di}$ is not a function of $\hat{\varphi}$ (technically $\hat{\Sigma}_{di}$ is a function of $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\beta}}$ which are themselves functions of $\hat{\varphi}$). With these assumptions we can now approximate $\hat{\tau}'_d - \tilde{\tau}'_d$ using a first order Taylor series expansion like in Kackar and Harville (1984). Here though we are in a bivariate setting.

First let $\tilde{\boldsymbol{\tau}}'_d = (\tilde{\tau}'_{d1}, \tilde{\tau}'_{d2})^t$, then

$$\tilde{\boldsymbol{\tau}}'_d\left(\hat{\varphi}\right) \approx \tilde{\boldsymbol{\tau}}'_d\left(\varphi\right) + \begin{pmatrix} \dfrac{\partial \tilde{\tau}'_{d1}}{\partial \varphi_1} & \dfrac{\partial \tilde{\tau}'_{d1}}{\partial \varphi_2} & \dfrac{\partial \tilde{\tau}'_{d1}}{\partial \varphi_{12}} \\ \dfrac{\partial \tilde{\tau}'_{d2}}{\partial \varphi_1} & \dfrac{\partial \tilde{\tau}'_{d2}}{\partial \varphi_2} & \dfrac{\partial \tilde{\tau}'_{d2}}{\partial \varphi_{12}} \end{pmatrix} \left(\hat{\varphi} - \varphi\right).$$

This then implies that

$$E\left((\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d)(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d)^t\right) \approx E\begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix},$$

where for $j = 1, 2$ and $k = 1, 2$,

$$\begin{aligned} g_{jk} = {} & \frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_1}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_1}(\hat{\varphi}_1 - \varphi_1)^2 + \frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_2}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_2}(\hat{\varphi}_2 - \varphi_2)^2 + \frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_{12}}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_{12}}(\hat{\varphi}_{12} - \varphi_{12})^2 \\ & + \left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_1}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_2} + \frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_2}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_1}\right)(\hat{\varphi}_1 - \varphi_1)(\hat{\varphi}_2 - \varphi_2) \\ & + \left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_1}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_{12}} + \frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_{12}}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_1}\right)(\hat{\varphi}_1 - \varphi_1)(\hat{\varphi}_{12} - \varphi_{12}) \\ & + \left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_2}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_{12}} + \frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_{12}}\frac{\partial \tilde{\tau}'_{dk}}{\partial \varphi_2}\right)(\hat{\varphi}_2 - \varphi_2)(\hat{\varphi}_{12} - \varphi_{12}). \end{aligned}$$

Now let for $j = 1, 2$,

$$\frac{\partial \tilde{\tau}'_{dj}}{\partial \boldsymbol{\varphi}} = \left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_1}, \frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_2}, \frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_{12}}\right)^t,$$

then for $j = 1, 2$ and $k = 1, 2$,

$$g_{jk} = Tr\left[\left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \boldsymbol{\varphi}}\right)\left(\frac{\partial \tilde{\tau}'_{dk}}{\partial \boldsymbol{\varphi}}\right)^t (\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})^t\right].$$

Now assuming that the elements of each of $\left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \boldsymbol{\varphi}}\right)\left(\frac{\partial \tilde{\tau}'_{dk}}{\partial \boldsymbol{\varphi}}\right)^t$, for each combination of j and k, are approximately independent of the elements of $(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})^t$, then

$$E\left((\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d)(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d)^t\right) \approx \begin{pmatrix} g^*_{11} & g^*_{12} \\ g^*_{12} & g^*_{22} \end{pmatrix}, \tag{8.9}$$

where for $j = 1, 2$ and $k = 1, 2$,

$$g^*_{jk} = Tr\left[E\left(\left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \boldsymbol{\varphi}}\right)\left(\frac{\partial \tilde{\tau}'_{dk}}{\partial \boldsymbol{\varphi}}\right)^t\right) E\left((\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})^t\right)\right].$$

This idea is motivated from Kackar and Harville (1984) and is also applied in Prasad and Rao (1990) and Baillo and Molina (2005).

The term $$E\left((\hat{\boldsymbol{\varphi}}-\boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}}-\boldsymbol{\varphi})^t\right)$$

can be approximated by using the inverse of the approximated observed Fisher information matrix. That is, if approximated ML is being used for the variance components then we can use the final value of $-\boldsymbol{J}_{\boldsymbol{\varphi}}^{-1}$ from Section 5. Or if approximated REML is being used for the variance components then we can use the final value of $-\boldsymbol{J}_{\boldsymbol{\varphi}}^{-1}$ from Section 6.

As for the other terms $$E\left(\left(\frac{\partial\tilde{\tau}'_{dj}}{\partial\boldsymbol{\varphi}}\right)\left(\frac{\partial\tilde{\tau}'_{dk}}{\partial\boldsymbol{\varphi}}\right)^t\right)$$

for $j=1,2$ and $k=1,2$, these can be simplified further by applying some ideas from Prasad and Rao (1990) and Baillo and Molina (2005) with some adjustments. We do this now.

Note that $\tilde{\tau}'_d$ can be written as follows

$$\tilde{\boldsymbol{\tau}}'_d=\begin{pmatrix}\tilde{\tau}'_{d1}\\ \tilde{\tau}'_{d2}\end{pmatrix}=\begin{pmatrix}\boldsymbol{K}_{d1}\tilde{\boldsymbol{\beta}}+\boldsymbol{M}_{d1}\tilde{\boldsymbol{u}}\\ \boldsymbol{K}_{d2}\tilde{\boldsymbol{\beta}}+\boldsymbol{M}_{d2}\tilde{\boldsymbol{u}}\end{pmatrix},$$

where $\boldsymbol{K}_{d1}$ and $\boldsymbol{K}_{d2}$ are respectively the first and second rows of the matrix $\boldsymbol{K}_d$. Similarly, $\boldsymbol{M}_{d1}$ and $\boldsymbol{M}_{d2}$ are respectively the first and second rows of the matrix $\boldsymbol{M}_d$.

For $j=1,2$ we have

$$\begin{aligned}\tilde{\tau}'_{dj}&=\boldsymbol{K}_{dj}\tilde{\boldsymbol{\beta}}+\boldsymbol{M}_{dj}\tilde{\boldsymbol{u}}\\ &=\boldsymbol{K}_{dj}\left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{\xi}+\boldsymbol{M}_{dj}\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\left(\boldsymbol{\xi}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)\\ &=\boldsymbol{K}_{dj}\left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{\xi}+\boldsymbol{M}_{dj}\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\left(\boldsymbol{\xi}-\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{\xi}\right).\end{aligned}$$

Now substitute $\boldsymbol{\xi}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}+\boldsymbol{e}$ into the above equation and after some algebra and some simplifications we obtain

$$\begin{aligned}\tilde{\tau}'_{dj}&=\boldsymbol{K}_{dj}\boldsymbol{\beta}\\ &+\left(\boldsymbol{K}_{dj}\left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}+\boldsymbol{M}_{dj}\boldsymbol{W}\boldsymbol{Z}^t\left(\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\right)\right)(\boldsymbol{Z}\boldsymbol{u}+\boldsymbol{e}).\end{aligned}$$

Now for $a = 1, 2$ and 12 and $j = 1, 2$ we have

$$\frac{\partial \tilde{\tau}'_{dj}}{\partial \varphi_a} = \frac{\partial}{\partial \varphi_a}\left(\begin{array}{l} K_{dj}\left(X^t V^{-1} X\right)^{-1} X^t V^{-1} + \\ M_{dj} W Z^t \left(V^{-1} - V^{-1} X \left(X^t V^{-1} X\right)^{-1} X^t V^{-1}\right) \end{array}\right)(Zu + e)$$

$$\approx \frac{\partial\left(M_{dj} W Z^t V^{-1}\right)}{\partial \varphi_a}(Zu + e). \qquad (8.10)$$

The justification for the above approximation can be roughly explained as follows. The terms that we are ignoring contain the term $(X^t V^{-1} X)^{-1}$. This term is expected to be small and negligible if the number of small areas D is large (which is often the case in practise) compared to the number of parameters in $\boldsymbol{\beta}$. This is because $(X^t V^{-1} X)^{-1}$ is the approximate variance of $\hat{\boldsymbol{\beta}}$ which gets smaller as the sample size increases.

Using the approximation (8.10) we can now write for $j = 1, 2$ and $k = 1, 2$,

$$E\left(\left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \boldsymbol{\varphi}}\right)\left(\frac{\partial \tilde{\tau}'_{dj}}{\partial \boldsymbol{\varphi}}\right)^t\right)$$

$$\approx \left(\frac{\partial\left(M_{dj} W Z^t V^{-1}\right)}{\partial \boldsymbol{\varphi}}\right) E\left((Zu + e)(Zu + e)^t\right)\left(\frac{\partial\left(M_{dk} W Z^t V^{-1}\right)}{\partial \boldsymbol{\varphi}}\right)^t$$

$$= \left(\frac{\partial\left(M_{dj} W Z^t V^{-1}\right)}{\partial \boldsymbol{\varphi}}\right) V \left(\frac{\partial\left(M_{dk} W Z^t V^{-1}\right)}{\partial \boldsymbol{\varphi}}\right)^t, \qquad (8.11)$$

where

$$\frac{\partial\left(M_{dj} W Z^t V^{-1}\right)}{\partial \boldsymbol{\varphi}} = \left(\left(\frac{\partial\left(M_{dj} W Z^t V^{-1}\right)}{\partial \varphi_1}\right)^t, \left(\frac{\partial\left(M_{dj} W Z^t V^{-1}\right)}{\partial \varphi_2}\right)^t, \left(\frac{\partial\left(M_{dj} W Z^t V^{-1}\right)}{\partial \varphi_{12}}\right)^t\right).$$

To approximate each of the matrices given by (8.11) for $j = 1, 2$ and $k = 1, 2$, we simply calculate the derivatives and then replace the unknown parameters by their estimated values. To calculate the elements within the matrices we can use the following simplifications which we obtained after some algebra.

For $j = 1, 2$ and $k = 1, 2$ and $a = 1, 2, 12$ and $b = 1, 2, 12$

$$\left(\frac{\partial\left(\boldsymbol{M}_{dj}\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\right)}{\partial\varphi_a}\right)\boldsymbol{V}\left(\frac{\partial\left(\boldsymbol{M}_{dk}\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\right)}{\partial\varphi_b}\right)^t = \boldsymbol{M}_{dj}\frac{\partial\boldsymbol{W}}{\partial\varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{W}}{\partial\varphi_b}\boldsymbol{M}_{dk}^t$$

$$-\boldsymbol{M}_{dj}\frac{\partial\boldsymbol{W}}{\partial\varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{W}}{\partial\varphi_b}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{W}\boldsymbol{M}_{dk}^t - \boldsymbol{M}_{dj}\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{W}}{\partial\varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{W}}{\partial\varphi_b}\boldsymbol{M}_{dk}^t$$

$$+\boldsymbol{M}_{dj}\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{W}}{\partial\varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{W}}{\partial\varphi_b}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{W}\boldsymbol{M}_{dk}^t. \tag{8.12}$$

So we can now fully approximate $E\left(\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d\right)\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d\right)^t\right)$.

Next we need to approximate the matrix $MSE(\tilde{\boldsymbol{\tau}}'_d)$. To do this we follow almost directly the approach taken in Molina *et al.* (2007) and after some algebra we obtain terms very similar to those obtained in that paper.

By definition

$$\tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d = \boldsymbol{K}_d\tilde{\boldsymbol{\beta}} + \boldsymbol{M}_d\tilde{\boldsymbol{u}} - \boldsymbol{K}_d\boldsymbol{\beta} - \boldsymbol{M}_d\boldsymbol{u}. \tag{8.13}$$

Let $$\boldsymbol{\Pi} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}$$

and $$\boldsymbol{P} = \left(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}$$

and substitute (8.6) and (8.7) and $\boldsymbol{\xi} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$ into (8.13). We then obtain

$$\tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d = \left(\boldsymbol{K}_d\boldsymbol{P}\boldsymbol{X}^t\boldsymbol{V}^{-1} + \boldsymbol{M}_d\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{\Pi}\right)\left(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}\right) - \boldsymbol{M}_d\boldsymbol{u}$$

and hence

$$\begin{aligned}
&\left(\tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)\left(\tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)^t \\
&= \left(\boldsymbol{K}_d\boldsymbol{P}\boldsymbol{X}^t\boldsymbol{V}^{-1} + \boldsymbol{M}_d\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{\Pi}\right)\left(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}\right)\left(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}\right)^t\left(\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{P}\boldsymbol{K}_d^t + \boldsymbol{\Pi}\boldsymbol{Z}\boldsymbol{W}\boldsymbol{M}_d^t\right) \\
&+ \boldsymbol{M}_d\boldsymbol{u}\boldsymbol{u}^t\boldsymbol{M}_d^t - \boldsymbol{M}_d\boldsymbol{u}\left(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}\right)^t\left(\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{P}\boldsymbol{K}_d^t + \boldsymbol{\Pi}\boldsymbol{Z}\boldsymbol{W}\boldsymbol{M}_d^t\right) \\
&- \left(\boldsymbol{K}_d\boldsymbol{P}\boldsymbol{X}^t\boldsymbol{V}^{-1} + \boldsymbol{M}_d\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{\Pi}\right)\left(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}\right)\boldsymbol{u}^t\boldsymbol{M}_d^t.
\end{aligned} \tag{8.14}$$

Now $$E\left((\boldsymbol{Zu}+\boldsymbol{e})(\boldsymbol{Zu}+\boldsymbol{e})^t\right)=\boldsymbol{V},$$

$$E\left(\boldsymbol{uu}^t\right)=\boldsymbol{W},$$

$$E\left(\boldsymbol{ue}^t\right)=0$$

and so after the above substitutions and with some further simplifications we have

$$E\left((\tilde{\boldsymbol{\tau}}'_d-\boldsymbol{\tau}'_d)(\tilde{\boldsymbol{\tau}}'_d-\boldsymbol{\tau}'_d)^t\right)=\boldsymbol{K}_d\boldsymbol{P}\boldsymbol{K}_d^t-\boldsymbol{K}_d\boldsymbol{P}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{W}\boldsymbol{M}_d^t$$

$$-\boldsymbol{M}_d\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{P}\boldsymbol{K}_d^t+\boldsymbol{M}_d\boldsymbol{W}\boldsymbol{M}_d^t-\boldsymbol{M}_d\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{\Pi}\boldsymbol{Z}\boldsymbol{W}\boldsymbol{M}_d^t. \tag{8.15}$$

Now let $$\boldsymbol{T}=\left(\boldsymbol{W}^{-1}+\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{Z}\right)^{-1} \tag{8.16}$$

and we will need the following identity from Henderson and Searle (1981) (for example)

$$\boldsymbol{V}^{-1}=\boldsymbol{\Sigma}-\boldsymbol{\Sigma}\boldsymbol{Z}\boldsymbol{T}\boldsymbol{Z}^t\boldsymbol{\Sigma}. \tag{8.17}$$

After some algebra it can be proved that

$$\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{W}=\boldsymbol{\Sigma}\boldsymbol{Z}\boldsymbol{T} \tag{8.18}$$

and $$\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}=\boldsymbol{T}\boldsymbol{Z}^t\boldsymbol{\Sigma}. \tag{8.19}$$

Using (8.16)–(8.19) we can further simplify (8.15) to

$$E\left((\tilde{\boldsymbol{\tau}}'_d-\boldsymbol{\tau}'_d)(\tilde{\boldsymbol{\tau}}'_d-\boldsymbol{\tau}'_d)^t\right)=\boldsymbol{M}_d\boldsymbol{T}\boldsymbol{M}_d^t+\boldsymbol{K}_d\boldsymbol{P}\boldsymbol{K}_d^t+\boldsymbol{M}_d\boldsymbol{T}\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^t\boldsymbol{\Sigma}\boldsymbol{Z}\boldsymbol{T}\boldsymbol{M}_d^t$$

$$-\boldsymbol{K}_d\boldsymbol{P}\boldsymbol{X}^t\boldsymbol{\Sigma}\boldsymbol{Z}\boldsymbol{T}\boldsymbol{M}_d^t-\boldsymbol{M}_d\boldsymbol{T}\boldsymbol{Z}^t\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{P}\boldsymbol{K}_d^t$$

which on collecting terms is

$$E\left((\tilde{\boldsymbol{\tau}}'_d-\boldsymbol{\tau}'_d)(\tilde{\boldsymbol{\tau}}'_d-\boldsymbol{\tau}'_d)^t\right)=\mathbf{M}_d\mathbf{T}\mathbf{M}_d^t+\left(\mathbf{K}_d-\mathbf{M}_d\mathbf{T}\mathbf{Z}^t\boldsymbol{\Sigma}\mathbf{X}\right)\mathbf{P}\left(\mathbf{K}_d-\mathbf{M}_d\mathbf{T}\mathbf{Z}^t\boldsymbol{\Sigma}\mathbf{X}\right)^t.$$

To approximate the above equation we simply substitute in the estimates of $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{\varphi}$.

We now have a formula to approximate the matrices $MSE(\hat{\boldsymbol{\delta}}_d)$. This approximation is

$$\widehat{MSE}(\hat{\boldsymbol{\delta}}_d) = \boldsymbol{G}_1(\hat{\boldsymbol{\varphi}}) + \boldsymbol{G}_2(\hat{\boldsymbol{\varphi}}) + 2\boldsymbol{G}_3(\hat{\boldsymbol{\varphi}}) + \boldsymbol{G}_4(\hat{\boldsymbol{\varphi}}), \tag{8.20}$$

where

$$\boldsymbol{G}_1(\hat{\boldsymbol{\varphi}}) = \hat{\boldsymbol{M}}_d \hat{\boldsymbol{T}} \hat{\boldsymbol{M}}_d^t ,$$

$$\boldsymbol{G}_2(\hat{\boldsymbol{\varphi}}) = \left(\hat{\boldsymbol{K}}_d - \hat{\boldsymbol{M}}_d \hat{\boldsymbol{T}} \boldsymbol{Z}^t \hat{\boldsymbol{\Sigma}} \boldsymbol{X}\right) \hat{\boldsymbol{P}} \left(\hat{\boldsymbol{K}}_d - \hat{\boldsymbol{M}}_d \hat{\boldsymbol{T}} \boldsymbol{Z}^t \hat{\boldsymbol{\Sigma}} \boldsymbol{X}\right)^t ,$$

$$\boldsymbol{G}_4(\hat{\boldsymbol{\varphi}}) = \sum_{i=1}^{I} \hat{\boldsymbol{\Sigma}}_{di}^r$$

and $\boldsymbol{G}_3(\hat{\boldsymbol{\varphi}})$ is given by (8.9) with estimates substituted in and with some further simplifications such as (8.12). Note that $\boldsymbol{G}_3(\hat{\boldsymbol{\varphi}})$ is multiplied by two in the above approximation. This is because as in Molina *et al.* (2007), $E(G_1(\hat{\boldsymbol{\varphi}})) \approx G_1(\boldsymbol{\varphi}) - G_3(\boldsymbol{\varphi})$ and we therefore multiply $\boldsymbol{G}_3(\hat{\boldsymbol{\varphi}})$ by two to correct for the bias. The proof that $E\left(\boldsymbol{G}_1(\hat{\boldsymbol{\varphi}})\right) \approx \boldsymbol{G}_1(\boldsymbol{\varphi}) - \boldsymbol{G}_3(\boldsymbol{\varphi})$ is given in the Appendix.

9. OUT-OF-SAMPLE SMALL AREAS

All previous sections implicitly assumed that for $d = 1, 2, \ldots, D$, each small area had some sample in at least one of its age/sex classes. Now suppose that as before we have $d = 1, 2, \ldots, D$ in-sample small areas, but we now have an additional set of R small areas with no sample. We can still produce estimates and MSEs for these out-of-sample small areas, but a slightly different methodology needs to be applied.

In total we have $d = 1, 2, \ldots, D, D+1, \ldots, D+R$ small areas that we wish to produce estimates and MSEs for. For the $d = 1, 2, \ldots, D$ small areas in sample we simply follow the methodology in the previous sections (discard the out-of-sample small areas for estimation of parameters, small area totals and MSEs). The rest of this section discusses estimation for the out-of-sample small areas.

Given our model, the best prediction we have of the out-of-sample small area random effects $(u_{d1}, u_{d2})^t$ for $d = D+1, D+2, \ldots, D+R$ is $(0, 0)^t$. This is because in our model we are assuming that the random effects are independent between small areas and knowing what the observed in-sample data are does not give us any additional information about the out-of-sample $\boldsymbol{u}$'s. If we had a model with spatially correlated random effects, then we could adapt the method outlined in Saei and Chambers (2005) to produce estimates and MSEs. But of course, our model does not have spatially correlated random effects.

We therefore need to resort to producing synthetic estimates for the out-of-sample small areas. These are for $d = D+1, D+2, \ldots, D+R$,

$$\hat{\boldsymbol{\delta}}_d = \sum_{i=1}^{I} \hat{\boldsymbol{\mu}}_{di}^{*r} = \hat{\boldsymbol{\tau}}_d,$$

where

$$\hat{\boldsymbol{\mu}}_{di}^{*r} = m_{di}^r \begin{pmatrix} \dfrac{e^{\boldsymbol{x}_{di1}^t \hat{\boldsymbol{\beta}}_1}}{1 + e^{\boldsymbol{x}_{di1}^t \hat{\boldsymbol{\beta}}_1} + e^{\boldsymbol{x}_{di2}^t \hat{\boldsymbol{\beta}}_2}} \\ \dfrac{e^{\boldsymbol{x}_{di2}^t \hat{\boldsymbol{\beta}}_2}}{1 + e^{\boldsymbol{x}_{di1}^t \hat{\boldsymbol{\beta}}_1} + e^{\boldsymbol{x}_{di2}^t \hat{\boldsymbol{\beta}}_2}} \end{pmatrix} = m_{di}^r \begin{pmatrix} p_{di1}^{syn} \\ p_{di2}^{syn} \end{pmatrix},$$

where the estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}^t_1, \hat{\beta}^t_2)^t$ are the PQL estimates based on the in-sample data $d = 1, 2, \ldots, D$.

Most of the derivation for an approximation to the MSE matrices for the out-of-sample small area totals proceeds in a similar way as in the previous section.

First define

$$\boldsymbol{\mu}_{di}^{*r} = m_{di}^{r}\begin{pmatrix} \dfrac{e^{\boldsymbol{x}_{di1}^{t}\boldsymbol{\beta}_1}}{1+e^{\boldsymbol{x}_{di1}^{t}\boldsymbol{\beta}_1}+e^{\boldsymbol{x}_{di2}^{t}\boldsymbol{\beta}_2}} \\ \dfrac{e^{\boldsymbol{x}_{di2}^{t}\boldsymbol{\beta}_2}}{1+e^{\boldsymbol{x}_{di1}^{t}\boldsymbol{\beta}_1}+e^{\boldsymbol{x}_{di2}^{t}\boldsymbol{\beta}_2}} \end{pmatrix} = m_{di}^{r}\begin{pmatrix} p_{di1}^{*} \\ p_{di2}^{*} \end{pmatrix},$$

and let $$\boldsymbol{\delta}_d = \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{*r} + \sum_{i=1}^{I}\left(\boldsymbol{y}_{di}^{r} - \boldsymbol{\mu}_{di}^{r} + \boldsymbol{\mu}_{di}^{r} - \boldsymbol{\mu}_{di}^{*r}\right)$$

denote the real totals we are trying to estimate.

Now define $$\tilde{\boldsymbol{\delta}}_d = \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{*r} = \boldsymbol{\tau}_d.$$

The MSE matrix of $\hat{\boldsymbol{\delta}}_d$ for $d = D+1, D+2, \ldots, D+R$ can now be written as

$$MSE\left(\hat{\boldsymbol{\delta}}_d\right) = E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^{t}\right) + E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^{t}\right)$$
$$+ E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^{t}\right) + E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^{t}\right). \qquad (9.1)$$

We now go about further simplifying the terms in (9.1).

Firstly we note that

$$E\left(\left(\sum_{i=1}^{I}\boldsymbol{y}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r}\right)\left(\sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r*}\right)^{t}\right)$$

$$= E\left(\left(E\left(\sum_{i=1}^{I}\boldsymbol{y}_{di}^{r}\middle|\boldsymbol{u}_d\right) - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r}\right)\left(\sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r*}\right)^{t}\right) = 0$$

and therefore

$$E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^{t}\right) = E\left(\left(\sum_{i=1}^{I}\boldsymbol{y}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r}\right)\left(\sum_{i=1}^{I}\boldsymbol{y}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r}\right)^{t}\right)$$
$$+ E\left(\left(\sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{*r}\right)\left(\sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{*r}\right)^{t}\right). \qquad (9.2)$$

The first term in the above equation is simply (8.4) and to approximate this term we can use (assuming the variance components are small)

$$\sum_{i=1}^{I} \hat{\boldsymbol{\Sigma}}_{di}^{*r},$$

where

$$\hat{\boldsymbol{\Sigma}}_{di}^{*r} = m_{di}^{r} \begin{pmatrix} p_{di1}^{syn}\left(1-p_{di1}^{syn}\right) & -p_{di1}^{syn} p_{di2}^{syn} \\ -p_{di1}^{syn} p_{di2}^{syn} & p_{di2}^{syn}\left(1-p_{di2}^{syn}\right) \end{pmatrix}. \tag{9.3}$$

The second term in (9.2) can be approximated by taking a first order Taylor series expansion. Let $\boldsymbol{\theta} = \boldsymbol{X}_{di}\boldsymbol{\beta} + \boldsymbol{u}_d$ and $\boldsymbol{\theta}^* = \boldsymbol{X}_{di}\boldsymbol{\beta}$. Then

$$\begin{aligned} \boldsymbol{\mu}_{di}^{r} &\approx \boldsymbol{\mu}_{di}^{*r} + \boldsymbol{\Sigma}_{di}^{*r}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{*}\right), \\ &= \boldsymbol{\mu}_{di}^{*r} + \boldsymbol{\Sigma}_{di}^{*r}\boldsymbol{u}_d \end{aligned} \tag{9.4}$$

where $\boldsymbol{\Sigma}_{di}^{*r}$ has the same form as (9.3) but with p_{di1}^{syn} and p_{di2}^{syn} replaced respectively with p_{di1}^{*} and p_{di2}^{*}. Hence,

$$\begin{aligned} E\left(\left(\sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{*r}\right)\left(\sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r} - \sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{*r}\right)^{t}\right) &\approx \boldsymbol{M}_{d}^{*} E\left(\boldsymbol{u}_d \boldsymbol{u}_d^{t}\right) \boldsymbol{M}_{d}^{*t} \\ &= \boldsymbol{M}_{d}^{*} \boldsymbol{W}_d \boldsymbol{M}_{d}^{*t}, \end{aligned} \tag{9.5}$$

where

$$\boldsymbol{M}_{d}^{*} = \sum_{i=1}^{I} \boldsymbol{\Sigma}_{di}^{*r}.$$

Note that (9.5) can be approximated with $\hat{\boldsymbol{M}}_{d}^{*}\hat{\boldsymbol{W}}_d\hat{\boldsymbol{M}}_{d}^{*t}$,

where

$$\hat{\boldsymbol{M}}_{d}^{*} = \sum_{i=1}^{I} \hat{\boldsymbol{\Sigma}}_{di}^{*r}$$

and

$$\hat{\boldsymbol{W}}_d = \begin{pmatrix} \hat{\varphi}_1 & \hat{\varphi}_{12} \\ \hat{\varphi}_{12} & \hat{\varphi}_2 \end{pmatrix}.$$

Now we need to approximate

$$E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^{t}\right) = MSE\left(\hat{\boldsymbol{\tau}}_d\right).$$

Let $\hat{\theta}^{*}_{di} = \boldsymbol{X}_{di}\hat{\boldsymbol{\beta}}$ and $\theta^{*}_{di} = \boldsymbol{X}_{di}\boldsymbol{\beta}$. By a first order Taylor series expansion we have

$$\hat{\boldsymbol{\mu}}^{*r}_{di} \approx \boldsymbol{\mu}^{*r}_{di} + \boldsymbol{\Sigma}^{*r}_{di}\left(\hat{\boldsymbol{\theta}}^{*}_{di} - \boldsymbol{\theta}^{*}_{di}\right) \tag{9.6}$$

and hence

$$\hat{\boldsymbol{\tau}}_d - \boldsymbol{\tau}_d \approx \sum_{i=1}^{I} \boldsymbol{\Sigma}^{*r}_{di}\left(\boldsymbol{X}_{di}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_{di}\boldsymbol{\beta}\right). \tag{9.7}$$

Now let $\boldsymbol{\tau}'_d = \boldsymbol{K}^{*}_d\boldsymbol{\beta}$ and $\hat{\boldsymbol{\tau}}'_d = \boldsymbol{K}^{*}_d\hat{\boldsymbol{\beta}}$, where $\boldsymbol{K}^{*}_d = \sum_{i=1}^{I} \boldsymbol{\Sigma}^{*r}_{di}\boldsymbol{X}_{di}$. Now,

$$\begin{aligned}
MSE\left(\hat{t}_d\right) &\approx MSE\left(\hat{\boldsymbol{\tau}}'_d\right) \\
&= E\left(\left(\hat{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)\left(\hat{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)^t\right) \\
&= E\left(\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d + \tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d + \tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)^t\right) \\
&= E\left(\left(\tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)\left(\tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)^t\right) + E\left(\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d\right)\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d\right)^t\right)
\end{aligned} \tag{9.8}$$

where $$\tilde{\boldsymbol{\tau}}'_d = K^{*}_d\tilde{\boldsymbol{\beta}}$$

and to get to the last line we assume as in the previous section that the estimator of the variance components is translation invariant.

The second term in (9.8) is

$$E\left(\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d\right)\left(\hat{\boldsymbol{\tau}}'_d - \tilde{\boldsymbol{\tau}}'_d\right)^t\right) = E\left(\left(\boldsymbol{K}^{*}_d\hat{\boldsymbol{\beta}} - \boldsymbol{K}^{*}_d\tilde{\boldsymbol{\beta}}\right)\left(\boldsymbol{K}^{*}_d\hat{\boldsymbol{\beta}} - \boldsymbol{K}^{*}_d\tilde{\boldsymbol{\beta}}\right)^t\right) \approx \mathbf{0}.$$

To see why the above term is approximately the $\mathbf{0}$ matrix, see the argument just after (8.10) and note that we do not have an $\boldsymbol{M}_d$ type component here.

The first term in (9.8) is

$$E\left(\left(\tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)\left(\tilde{\boldsymbol{\tau}}'_d - \boldsymbol{\tau}'_d\right)^t\right) = E\left(\left(\boldsymbol{K}^{*}_d\tilde{\boldsymbol{\beta}} - \boldsymbol{K}^{*}_d\boldsymbol{\beta}\right)\left(\boldsymbol{K}^{*}_d\tilde{\boldsymbol{\beta}} - \boldsymbol{K}^{*}_d\boldsymbol{\beta}\right)^t\right),$$

where $$\tilde{\boldsymbol{\beta}} = \left(X^t\boldsymbol{V}^{-1}X\right)^{-1}X^t\boldsymbol{V}^{-1}\boldsymbol{\xi}$$

and $\boldsymbol{X}$ and $\boldsymbol{V}$ are as defined in the previous sections and are therefore based on the in-sample areas only. Again we need to assume that $\boldsymbol{\Sigma}^{-1}$ in $\boldsymbol{V}$ does not depend on the in-sample random effects $\boldsymbol{u}$ to do the following.

After some simplifications we have

$$K_d^*\tilde{\boldsymbol{\beta}} - K_d^*\boldsymbol{\beta} = K_d^*\left(X^tV^{-1}X\right)^{-1}X^tV^{-1}(Z\boldsymbol{u}+\boldsymbol{e}),$$

where Z, $\boldsymbol{u}$ and $\boldsymbol{e}$ are defined using the in-sample areas.

Hence we have

$$\begin{aligned}
&E\left((\tilde{\boldsymbol{\tau}}_d' - \boldsymbol{\tau}_d')(\tilde{\boldsymbol{\tau}}_d' - \boldsymbol{\tau}_d')^t\right) \\
&= K_d^*\left(X^tV^{-1}X\right)^{-1}X^tV^{-1}E\left((Z\boldsymbol{u}+\boldsymbol{e})(Z\boldsymbol{u}+\boldsymbol{e})^t\right)V^{-1}X\left(X^tV^{-1}X\right)^{-1}K_d^{*t} \\
&= K_d^*\left(X^tV^{-1}X\right)^{-1}X^tV^{-1}VV^{-1}X\left(X^tV^{-1}X\right)^{-1}K_d^{*t} \\
&= K_d^*\left(X^tV^{-1}X\right)^{-1}K_d^{*t} \qquad (9.9)
\end{aligned}$$

and we can approximate (9.9) with

$$\hat{K}_d^*\left(X^t\hat{V}^{-1}X\right)^{-1}\hat{K}_d^{*t}$$

where $$\hat{K}_d^* = \sum_{i=1}^{I}\hat{\Sigma}_{di}^{*r}X_{di}$$

and $\hat{V}$ is just V with the estimates $\hat{\boldsymbol{\varphi}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ substituted in as in previous sections.

Now we need to consider the terms

$$E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right)$$

and $$E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\right).$$

Note that $\hat{\boldsymbol{\delta}}_d - \tilde{\tilde{\boldsymbol{\delta}}}_d$ is a function of the in-sample data and is independent of $\tilde{\tilde{\boldsymbol{\delta}}}_d - \boldsymbol{\delta}_d$ which is a function of the out-of-sample data. Therefore

$$E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right) = E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\right)E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right)$$

and $$E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\right) = E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\right)E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\right).$$

Now using (9.6) and (9.7) we have

$$E\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right) \approx \sum_{i=1}^{I} \boldsymbol{\Sigma}_{di}^{*r}\left(\boldsymbol{X}_{di}E\left(\hat{\boldsymbol{\beta}}\right) - \boldsymbol{X}_{di}\boldsymbol{\beta}\right) = \mathbf{0}_{2\times 1}$$

since $\hat{\boldsymbol{\beta}}$ is unbiased in a linear mixed model framework.

Also using (9.4)

$$\begin{aligned}
E\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right) &= -E\left(\sum_{i=1}^{I}\left(\boldsymbol{y}_{di}^{r} - \boldsymbol{\mu}_{di}^{r} + \boldsymbol{\mu}_{di}^{r} - \boldsymbol{\mu}_{di}^{*r}\right)\right) \\
&= -E\left(E\left(\sum_{i=1}^{I}\left(\boldsymbol{y}_{di}^{r} - \boldsymbol{\mu}_{di}^{r} + \boldsymbol{\mu}_{di}^{r} - \boldsymbol{\mu}_{di}^{*r}\right)\middle|\boldsymbol{u}_d\right)\right) \\
&= -E\left(\sum_{i=1}^{I}\boldsymbol{\mu}_{di}^{r} - \boldsymbol{\mu}_{di}^{*r}\right) \\
&\approx -E\left(\sum_{i=1}^{I}\boldsymbol{\Sigma}_{di}^{*r}\boldsymbol{u}_d\right) \\
&= \mathbf{0}_{2\times 1}\,.
\end{aligned}$$

Hence $$E\left(\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^t\right) \approx \mathbf{0}_{2\times 2}$$

and $$E\left(\left(\tilde{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \tilde{\boldsymbol{\delta}}_d\right)^t\right) \approx \mathbf{0}_{2\times 2}\,.$$

We can now define an MSE estimator for the out-of-sample areas, $d = D+1, D+2, \ldots, D+R$ as

$$\widehat{MSE}\left(\hat{\boldsymbol{\delta}}_d\right) = \hat{\boldsymbol{K}}_d^{*}\left(\boldsymbol{X}^t\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\hat{\boldsymbol{K}}_d^{*t} + \hat{\boldsymbol{M}}_d^{*}\hat{\boldsymbol{W}}_d\hat{\boldsymbol{M}}_d^{*t} + \sum_{i=1}^{I}\hat{\boldsymbol{\Sigma}}_{di}^{*r}\,.$$

10. APPROXIMATE STANDARD ERRORS OF THE ELEMENTS OF $\hat{\beta}$

In Section 8 we derived the approximate MSE matrix of an estimator of the form $K\hat{\beta}+M\hat{u}$, where M and K are given matrices.

If we set K to be the identity matrix and M to be a matrix containing all 0's then we can apply (8.20) directly to obtain

$$\begin{aligned}\widehat{MSE}\left(\hat{\beta}\right) &\approx K\hat{P}K^t \\ &= \left(X^t\hat{V}^{-1}X\right)^{-1}\end{aligned} \tag{10.1}$$

by noting that the component $G_4(\hat{\varphi})$ is not relevant here. The approximation (10.1) can be used to check the significance of the covariates in the model. Note that it is not appropriate to use $-J_{\beta}^{-1}$ defined in Section 4 for this purpose (since it is not a function of W and does not account properly for the influence of the random effects). Note also that in this framework (of linear mixed models), $\hat{\beta}$ is unbiased, and so the approximate standard errors of the elements of $\hat{\beta}$ are the square root of the diagonal entries in (10.1).

11. PARAMETRIC BOOTSTRAP MEAN SQUARED ERRORS

The mean squared error estimators in Sections 8 and 9 were derived using various approximations. In this section we consider an alternative mean squared error estimation technique based on a parametric bootstrap. This approach is very similar to the bootstrap method described in Molina *et al.* (2007). The main difference here is that instead of simulating $\boldsymbol{u}_d$ from a univariate normal distribution, we need to simulate from a bivariate normal distribution. We also extend this bootstrap technique to incorporate out-of-sample small areas.

The bootstrap method is outlined as follows:

(a) Model fitting: fit the model to the original data (this will be for in-sample areas $d = 1, 2, \ldots, D$), obtaining parameter estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\varphi}}$.

(b) Generation of random effects: For $d = 1, 2, \ldots, D, D+1, D+2, \ldots, D+R$ (includes R out-of-sample areas), independently generate $\boldsymbol{u}_d^* = (u_{d1}^*, u_{d2}^*)^t$ from a bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\hat{\boldsymbol{W}}_d = \begin{pmatrix} \hat{\varphi}_1 & \hat{\varphi}_{12} \\ \hat{\varphi}_{12} & \hat{\varphi}_2 \end{pmatrix}.$$

(c) Generation of a bootstrap population: for $d = 1, 2, \ldots, D, D+1, D+2, \ldots, D+R$, calculate the probabilities

$$p_{di1}^* = \frac{e^{\boldsymbol{x}_{di1}^t \hat{\boldsymbol{\beta}}_1 + u_{d1}^*}}{1 + e^{\boldsymbol{x}_{di1}^t \hat{\boldsymbol{\beta}}_1 + u_{d1}^*} + e^{\boldsymbol{x}_{di2}^t \hat{\boldsymbol{\beta}}_2 + u_{d2}^*}}$$

and

$$p_{di2}^* = \frac{e^{\mathbf{x}_{di2}^t \hat{\boldsymbol{\beta}}_2 + u_{d2}^*}}{1 + e^{\mathbf{x}_{di1}^t \hat{\boldsymbol{\beta}}_1 + u_{d1}^*} + e^{\mathbf{x}_{di2}^t \hat{\boldsymbol{\beta}}_2 + u_{d2}^*}}$$

and generate the following sample and non-sample multinomial vectors

$$\boldsymbol{y}_{di}^* = \left(y_{di1}^*, y_{di2}^*\right)^t \sim \text{Multinomial}\left(m_{di}, p_{di1}^*, p_{di2}^*\right),$$

$$\boldsymbol{y}_{di}^{*r} = \left(y_{di1}^{*r}, y_{di2}^{*r}\right)^t \sim \text{Multinomial}\left(m_{di}^r, p_{di1}^*, p_{di2}^*\right).$$

Calculate the true area totals

$$\boldsymbol{\delta}_d^* = \left(\delta_{d1}^*, \delta_{d2}^*\right)^t = \sum_{i=1}^{10} \left(\boldsymbol{y}_{di}^* + \boldsymbol{y}_{di}^{*r}\right).$$

(d) Model fitting to the bootstrap sample and parameter estimation: fit model to the bootstrap sample data $\boldsymbol{y}_{di}^{*}$, for $d = 1, 2, \ldots, D$ only, obtaining estimates $\hat{\boldsymbol{\beta}}_1^{*}$, $\hat{\boldsymbol{\beta}}_2^{*}$, $\hat{\boldsymbol{\varphi}}^{*}$ and predicted values $\hat{\boldsymbol{u}}_d^{*}$. Note that the predicted values of $\boldsymbol{u}_d$ for the out-of-sample areas $d = D+1, D+2, \ldots, D+R$ are always $\hat{\boldsymbol{u}}_d^{*} = \mathbf{0}$. From these, calculate individual predicted values for $d = 1, 2, \ldots, D+R$:

$$\hat{y}_{di}^{*r} = m_{di}^{r}\left(\frac{e^{\boldsymbol{x}_{di1}^{t}\hat{\boldsymbol{\beta}}_1^{*}+\hat{u}_{d1}^{*}}}{1+e^{\boldsymbol{x}_{di1}^{t}\hat{\boldsymbol{\beta}}_1^{*}+\hat{u}_{d1}^{*}}+e^{\boldsymbol{x}_{di2}^{t}\hat{\boldsymbol{\beta}}_2^{*}+\hat{u}_{d2}^{*}}}, \frac{e^{\boldsymbol{x}_{di2}^{t}\hat{\boldsymbol{\beta}}_2^{*}+\hat{u}_{d2}^{*}}}{1+e^{\boldsymbol{x}_{di1}^{t}\hat{\boldsymbol{\beta}}_1^{*}+\hat{u}_{d1}^{*}}+e^{\boldsymbol{x}_{di2}^{t}\hat{\boldsymbol{\beta}}_2^{*}+\hat{u}_{d2}^{*}}}\right)^{t}.$$

Then calculate bootstrap estimates of totals by

$$\hat{\boldsymbol{\delta}}_d^{*} = \left(\hat{\delta}_{d1}^{*}, \hat{\delta}_{d2}^{*}\right)^{t} = \sum_{i=1}^{10}\left(\boldsymbol{y}_{di}^{*} + \hat{\boldsymbol{y}}_{di}^{*r}\right).$$

(e) Bootstrap replicates: repeat steps (b)–(d) B times. Let $\delta_{d1}^{*(b)}$ and $\delta_{d2}^{*(b)}$ denote the true values of the parameters and $\hat{\delta}_{d1}^{*(b)}$ and $\hat{\delta}_{d2}^{*(b)}$ the estimators that are obtained in the b-th repetition, $b = 1, 2, \ldots, B$. The bootstrap estimators of the mean squared error matrices $E\left((\hat{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d)(\hat{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d)^{t}\right)$, for $d = 1, 2, \ldots, D+R$ are

$$\hat{E}\left(\left(\hat{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)\left(\hat{\boldsymbol{\delta}}_d - \boldsymbol{\delta}_d\right)^{t}\right) =$$

$$\begin{pmatrix} B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_{d1}^{*(b)} - \delta_{d1}^{*(b)}\right)^2 & B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_{d1}^{*(b)} - \delta_{d1}^{*(b)}\right)\left(\hat{\delta}_{d2}^{*(b)} - \delta_{d2}^{*(b)}\right) \\ B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_{d1}^{*(b)} - \delta_{d1}^{*(b)}\right)\left(\hat{\delta}_{d2}^{*(b)} - \delta_{d2}^{*(b)}\right) & B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_{d2}^{*(b)} - \delta_{d2}^{*(b)}\right)^2 \end{pmatrix}.$$

(f) Other estimates: after step (e) it is also possible to estimate mean squared errors of other parameter estimates such as $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varphi}}$. For example, after implementing the repetitions $b = 1, 2, \ldots, B$ we can also approximate root relative mean squared errors of each element in $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. Let index k denote the k-th element of $\boldsymbol{\beta}_1$ (or $\boldsymbol{\beta}_2$). Let $\hat{\boldsymbol{\beta}}_{1(k)}$ denote the value from step (a) and $\hat{\boldsymbol{\beta}}_{1(k)}^{*(b)}$ denote the bootstrap estimate at iteration b. A bootstrap estimate of the $RMSE(\hat{\boldsymbol{\beta}}_{1(k)})$ is

$$\sqrt{B^{-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\beta}}_{1(k)}^{*(b)} - \hat{\boldsymbol{\beta}}_{1(k)}\right)^2}$$

and this can be compared directly with the standard error estimate from Section 10.

12. AUXILIARY DATA

The previous sections described the theory behind producing small area estimates of labour force counts and estimating mean squared errors. We now apply this theory to real data.

The small areas in our model represent LGAs (Local Government Areas) which have boundaries defined according to the 2001 ASGC (Australian Standard Geographical Classification). These are chosen because they provide sufficiently fine geographical areas, but still have adequate samples in the area for analysis. The total number of in-sample areas in August 2001 and August 2006 are respectively 424 and 413 and the total number of out-of-sample small areas in August 2001 and August 2006 are respectively 220 and 231.

The first step is to fit an appropriate model to the in-sample data and produce PQL estimates of $\boldsymbol{u}$ and $\boldsymbol{\beta}$ and approximated REML or ML estimates of $\boldsymbol{\varphi}$. One essential part of this first step is to choose an appropriate set of explanatory variables to include in the models. That is, we need to fully define $\boldsymbol{x}_{di1}$ and $\boldsymbol{x}_{di2}$.

Auxiliary data are available from a variety of different sources. The two main sources are administrative Centrelink benefit payment data from DEEWR (Department of Education, Employment and Workplace Relations) and the Australian Census of Population and Housing. These data are available for the time points August 2001 and August 2006 and so it will be possible to fit two separate models with the same explanatory variables for the two different time points as a comparison.

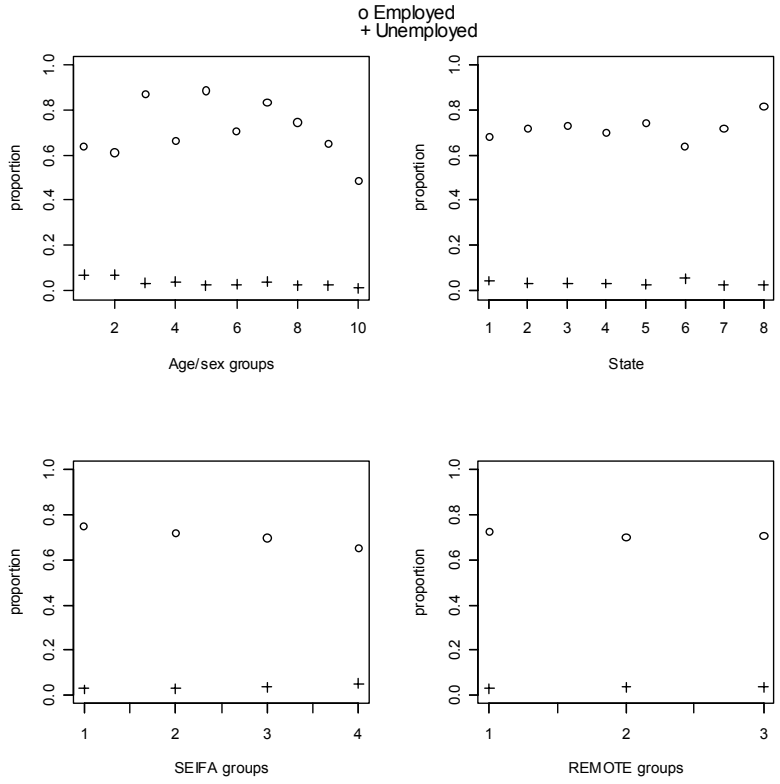
There are quite a large number of potential explanatory variables that could be used in the models. Ideally some kind of model variables selection process would need to be undertaken, however we do not consider this approach here. Instead we note that the ABS has already produced experimental estimates of small area labour force counts using three separate binomial logit mixed effects models (using the methodology outlined in Saei and Chambers, 2003). A careful model selection process was applied in this case to determine an appropriate set of explanatory variables for the binomial logit mixed models. For consistency and comparative purposes we will use the same set of explanatory variables that were used in these earlier models. Therefore each of our vectors $\boldsymbol{x}_{di1}$ and $\boldsymbol{x}_{di2}$ will contain the same set of 37 variables each. In summary, these variables are benefits payments variables, state indicators, age/sex indicators, remoteness indicators, socio-economic indexes for areas, and household type.

The following list outlines the explanatory variables in more detail.

- AS1–AS10: ten age/sex indicators. The five age groupings are 15–24 years, 25–34 years, 35–44 years, 45–54 years and 55–64 years. Note that we only consider the working age population for this analysis. AS1–AS10 are defined in order of increasing age and all odd AS groups are males. The base category is AS1.
- STATE1–STATE8: eight state indicators. These refer in order to New South Wales, Victoria, Queensland, South Australia, Western Australia, Tasmania, the Northern Territory and the Australian Capital Territory. The base category is New South Wales.
- NSA_YAO: Proportion of population in the class registered to receive full payment of Newstart Allowance (unemployment benefits) or Youth Allowance (other).
- ASPAY1–ASPAY10: ten age/sex indicator by PAY interactions. PAY is the proportion of the population in the class registered to receive full other benefit payments. For example, disability support pension, parenting payments, partner allowance, wife pension, etc.. Note that this set up implies that the effect that PAY has on the probability is different for each age/sex class.
- REMOTE1–REMOTE3: three remoteness indicators. These refer to major city (REMOTE1), non-remote area (REMOTE2) and remote area (REMOTE3) as per the ASGC (Australian Standard Geographical Classification) 2001 (see ABS (2001) for further details). The base category is REMOTE1.
- SEIFA1–SEIFA4: four socio-economic index of advantage-disadvantage indicators. SEIFA1 indicates advantaged areas (whether the area is in the top 25% of SEIFA scores), SEIFA2 indicates the next 25%, SEIFA3 the next 25% and SEIFA4 indicates the most disadvantaged areas (whether the area is in the bottom 25% of SEIFA scores). The base category is SEIFA1. For further details see ABS (2003).
- HH1: Proportion of Census population in class that lives in dwelling consisting of married couple only or married couple with at least one child aged 15 or over.
- HH2: Proportion of Census population in class that lives in dwelling consisting of married couple with children all aged 0 to 14 years.
- HH3: Proportion of Census population in class that lives in dwelling consisting of one person only or one person with at least one child aged 15 or over.
- HH4: Proportion of Census population in class that lives in dwelling consisting of one person with children all aged 0 to 14 years.

Before we fit the multinomial logit mixed models, we undertake a brief exploratory data analysis using the August 2006 LFS and Census data to examine the suitability of our above chosen explanatory variables. Figure 12.1 contains average proportions of employed and unemployed within the ten AS groups, the eight STATE groups, the three REMOTE groups and the four SEIFA groups. The labour force proportions (especially employment) depend strongly on age and sex since the mean proportions vary across these categories. The AS indicators should certainly be considered as explanatory variables. Figure 12.1 also shows that the relationships between the labour force average proportions and each of STATE, SEIFA and REMOTE are in general not as strong as age/sex but there is still some variation.

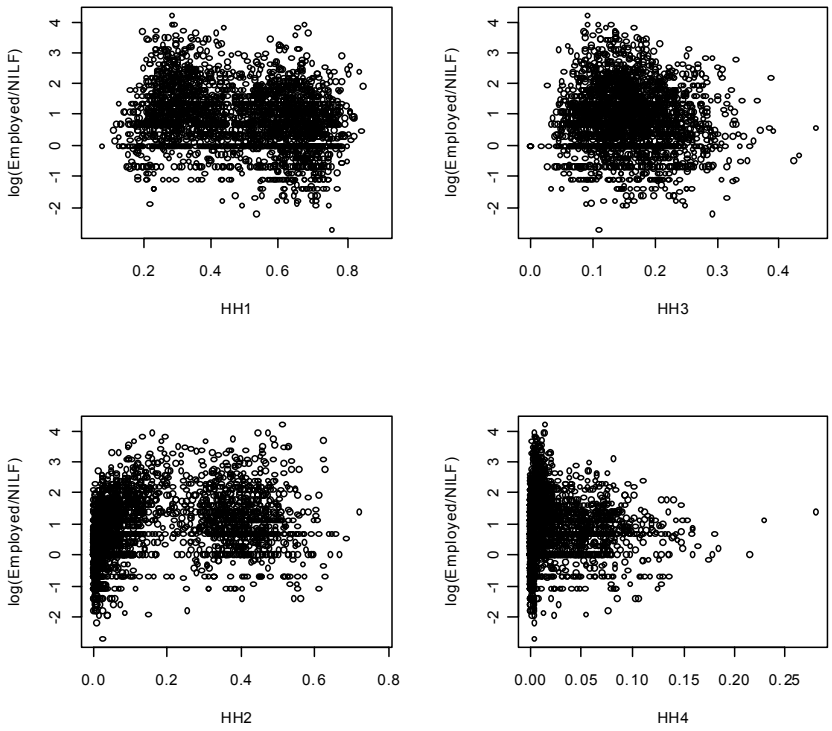
12.1 Plot of average labour force proportions within certain groups for 2006



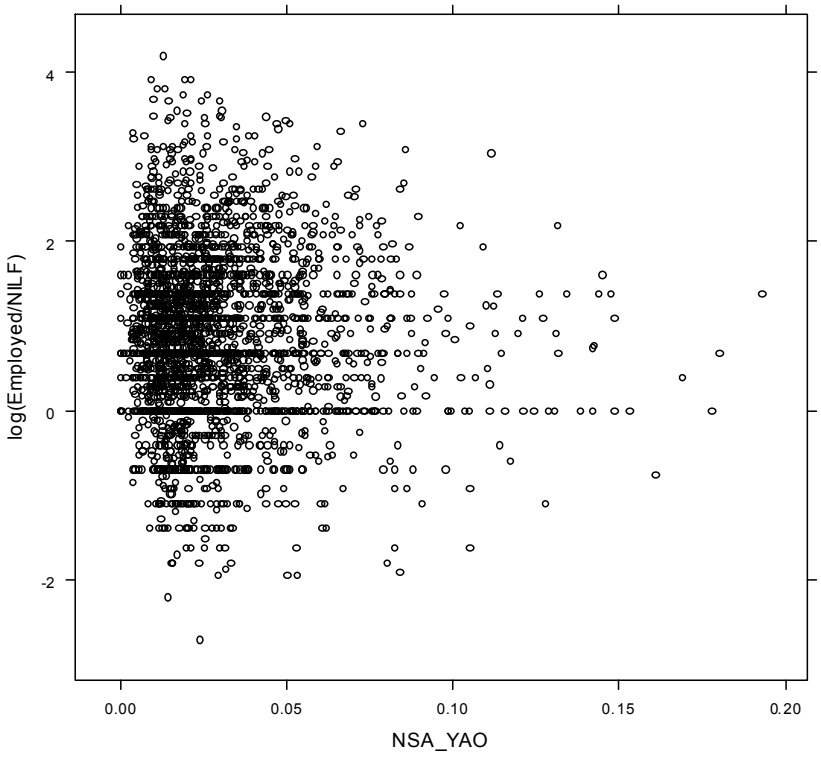
The variables we have examined so far are the categorical variables. We now examine relationships between $\log(y_{di1}/y_{di3})$ for the cases $y_{di3} > 0$ and $y_{di1} > 0$ and the continuous explanatory variables for 2006. Note that we do not do this for y_{di2} since there are a large number of zeros in this case.

Figure 12.2 contains plots of $\log(y_{di1}/y_{di3})$ versus each of HH1, HH2, HH3 and HH4. From these it appears there might be a weak association, especially for HH1 and HH3. A similar plot for the variable NSA_YAO is given in figure 12.3, although we note that the linear relationship in this case does not look very strong. However, note that irrespective of this observation, a relationship may still exist between $\log(y_{di2}/y_{di3})$ and NSA_YAO and this is the reason why this variable is still included.

12.2 Plot of $\log(y_{di1}/y_{di3})$ versus each of HH1, HH2, HH3 and HH4 for 2006



12.3 Plot of $\log(y_{di1}/y_{di3})$ versus NSA_YAO for 2006



12.4 Plot of log(y_{di1}/y_{di3}) versus PAY within AS groups for 2006

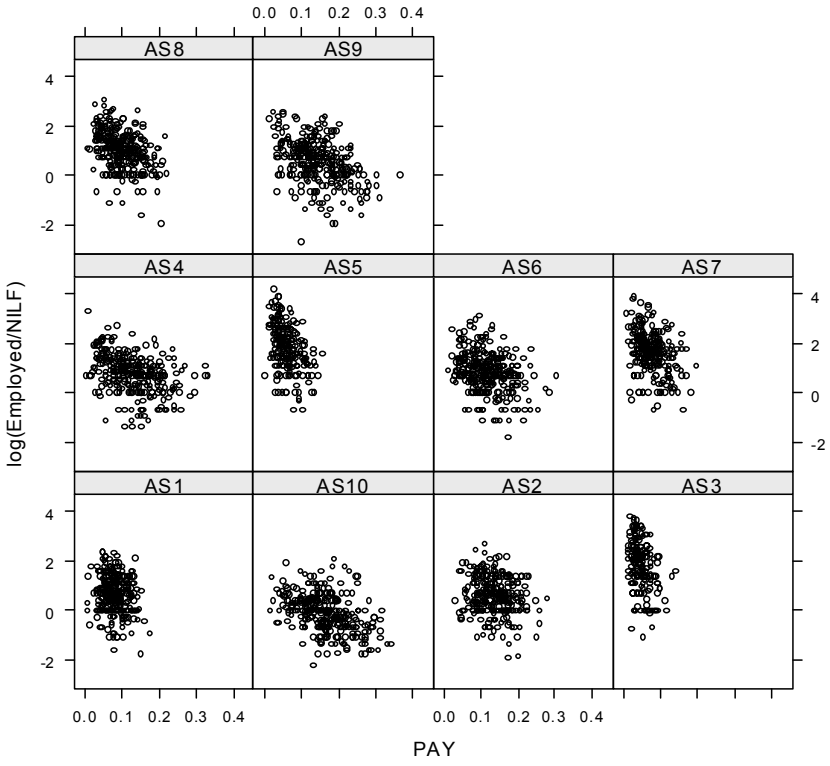


Figure 12.4 contains a plot of $\log(y_{di1}/y_{di3})$ versus PAY within AS groups. There appears to be a negative relationship between $\log(y_{di1}/y_{di3})$ and PAY. However, note that the slopes appear to differ between AS groups. This is the reason why we include the PAY variable in the model as an interaction between each of the AS indicators.

13. ESTIMATES OF MODEL PARAMETERS

Tables 13.1 and 13.2 contain the PQL estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for the years 2006 and 2001. Estimates of the standard errors of the parameter estimates and p-values are also given in these tables. The p-values are calculated assuming that the distribution of the parameter estimate is approximately normal. Estimates which are significant at the 0.05 level are marked with a *. Not all parameter estimates appear to be significant, but as we stated in the previous section, all variables are still included for consistency with the previously fitted binomial logit mixed models. As expected, for employment (and this holds for both years), the AS group indicators and ASPAY variables are highly significant. For both years, the variable NSA_YAO is highly significant for unemployment. This is what we would expect as NSA_YAO is related to unemployment.

The standard errors in tables 13.1 and 13.2 use the analytical approximation (10.1) which has many assumptions behind it. As a comparison, we also calculate parametric bootstrap RMSEs using the algorithm described in Section 11 (in particular see step (f)). We implement the parametric bootstrap using the 2006 data, with the simulation size set to $B = 1000$ and use the combined PQL-REML algorithm for parameter estimation. A comparison between the 74 estimated analytical standard errors and the parametric bootstrap estimated root mean squared errors is given in figure 13.3 (figure 13.4 contains the smaller standard errors in the range 0–0.5 only). From these plots we can see that the differences between the estimated SEs and RMSEs are very small. Also in all cases we find that the bias component of the bootstrap RMSEs is small (<1.5%). Therefore the analytical SE estimates of $\hat{\boldsymbol{\beta}}$ appear to be good approximations.

The PQL parameter estimates in tables 13.1 and 13.2 use the approximated REML method to estimate the variance components $\boldsymbol{\varphi}$. As an alternative to this we could also have used the approximated ML method. Table 13.5 contains estimates of the variance components for both August 2001 and August 2006, using the two different estimation methods. To estimate the RRMSEs of the approximated ML and REML estimators, the parametric bootstrap of Section 11 (with $B = 1000$) can be used with some slight modifications that we now briefly discuss:

- In step (a) of the algorithm, compute the PQL-REML estimates. These will be conditioned on in the rest of the simulation
- In step (d) we need to compute both the REML and ML estimates of the variance components.
- At the conclusion of the simulation we have 1000 REML and 1000 ML estimates of $\boldsymbol{\varphi}$. These can then be used to calculate estimates of bias and RMSEs.

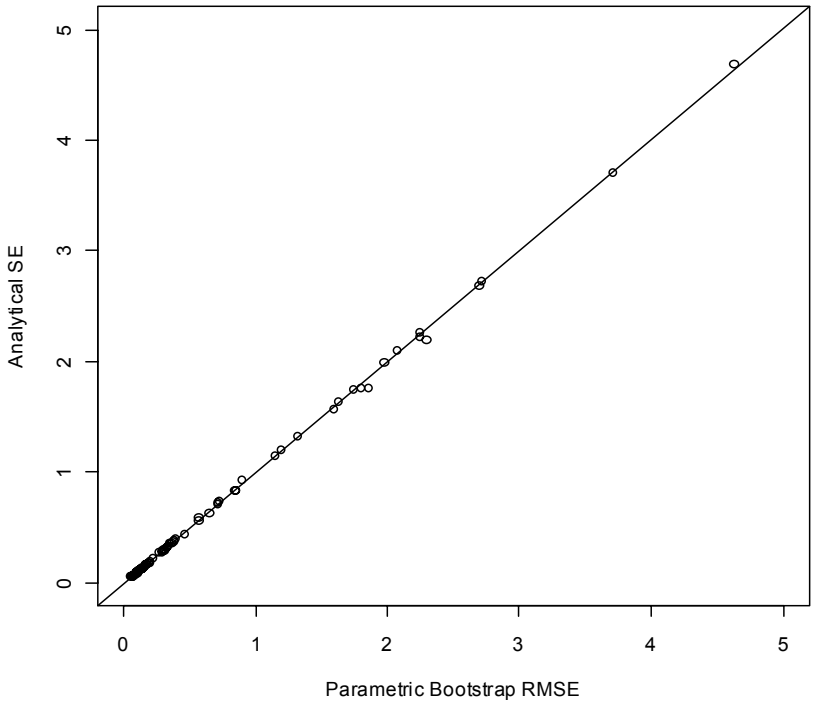
13.1 PQL Estimates of β_1 and β_2 for 2006 (REML used for variance components)

	Employed					*Unemployed*				
Variable	*Estimate*	*SE*	*Z*	*p-value*		*Estimate*	*SE*	*Z*	*p-value*	
Intercept	0.428	0.192	2.23	0.0257	*	–1.890	0.358	–5.27	< 0.0001	*
STATE2	0.092	0.059	1.55	0.1210		–0.016	0.096	–0.17	0.8690	
STATE3	0.175	0.062	2.81	0.0050	*	0.057	0.101	0.56	0.5740	
STATE4	0.096	0.069	1.39	0.1650		–0.095	0.115	–0.83	0.4090	
STATE5	0.096	0.065	1.47	0.1420		–0.162	0.114	–1.42	0.1560	
STATE6	0.002	0.091	0.02	0.9870		–0.013	0.144	–0.09	0.9260	
STATE7	–0.137	0.158	–0.87	0.3870		–0.730	0.276	–2.65	0.0081	*
STATE8	0.444	0.131	3.39	0.0007	*	0.155	0.199	0.78	0.4350	
AS2	0.094	0.138	0.68	0.4950		–0.154	0.269	–0.57	0.5660	
AS3	2.130	0.161	13.20	< 0.0001	*	0.986	0.297	3.33	0.0009	*
AS4	0.627	0.138	4.55	< 0.0001	*	–0.819	0.311	–2.63	0.0085	*
AS5	2.320	0.184	12.60	< 0.0001	*	1.210	0.382	3.18	0.0015	*
AS6	0.637	0.170	3.75	0.0002	*	–0.374	0.377	–0.99	0.3210	
AS7	1.600	0.145	11.10	< 0.0001	*	–0.048	0.297	–0.16	0.8720	
AS8	0.476	0.129	3.69	0.0002	*	–1.220	0.314	–3.88	0.0001	*
AS9	0.360	0.134	2.68	0.0074	*	–0.930	0.328	–2.84	0.0045	*
AS10	–0.647	0.132	–4.91	< 0.0001	*	–2.540	0.443	–5.73	< 0.0001	*
REMOTE2	0.101	0.078	1.30	0.1940		0.028	0.115	0.24	0.8090	
REMOTE3	0.215	0.055	3.92	< 0.0001	*	0.063	0.089	0.71	0.4790	
SEIFA2	–0.145	0.059	–2.48	0.0131	*	–0.041	0.094	–0.44	0.6630	
SEIFA3	–0.157	0.073	–2.14	0.0324	*	0.057	0.122	0.46	0.6430	
SEIFA4	–0.098	0.082	–1.19	0.2340		0.106	0.140	0.76	0.4500	
ASPAY1	–5.430	1.210	–4.50	< 0.0001	*	–4.700	2.260	–2.08	0.0375	*
ASPAY2	–4.060	0.843	–4.81	< 0.0001	*	–1.970	1.570	–1.25	0.2110	
ASPAY3	–17.100	2.720	–6.30	< 0.0001	*	–12.900	4.680	–2.76	0.0058	*
ASPAY4	–4.510	0.739	–6.10	< 0.0001	*	0.652	1.630	0.40	0.6890	
ASPAY5	–16.300	1.740	–9.39	< 0.0001	*	–17.400	3.700	–4.70	< 0.0001	*
ASPAY6	–4.790	0.735	–6.52	< 0.0001	*	–2.640	1.770	–1.49	0.1360	
ASPAY7	–10.500	1.330	–7.94	< 0.0001	*	–2.160	2.690	–0.80	0.4220	
ASPAY8	–5.410	0.844	–6.41	< 0.0001	*	0.325	2.220	0.15	0.8840	
ASPAY9	–6.390	0.638	–10.00	< 0.0001	*	–3.730	1.760	–2.12	0.0340	*
ASPAY10	–4.720	0.566	–8.34	< 0.0001	*	0.034	2.200	0.02	0.9880	
NSA_YA0	–0.030	1.160	–0.03	0.9800		8.900	2.100	4.23	< 0.0001	*
HH1	0.637	0.219	2.91	0.0036	*	0.444	0.400	1.11	0.2670	
HH2	–0.244	0.292	–0.84	0.4030		–0.582	0.591	–0.98	0.3250	
HH3	1.510	0.362	4.18	< 0.0001	*	0.814	0.725	1.12	0.2630	
HH4	1.090	0.924	1.18	0.2380		3.630	1.990	1.82	0.0688	

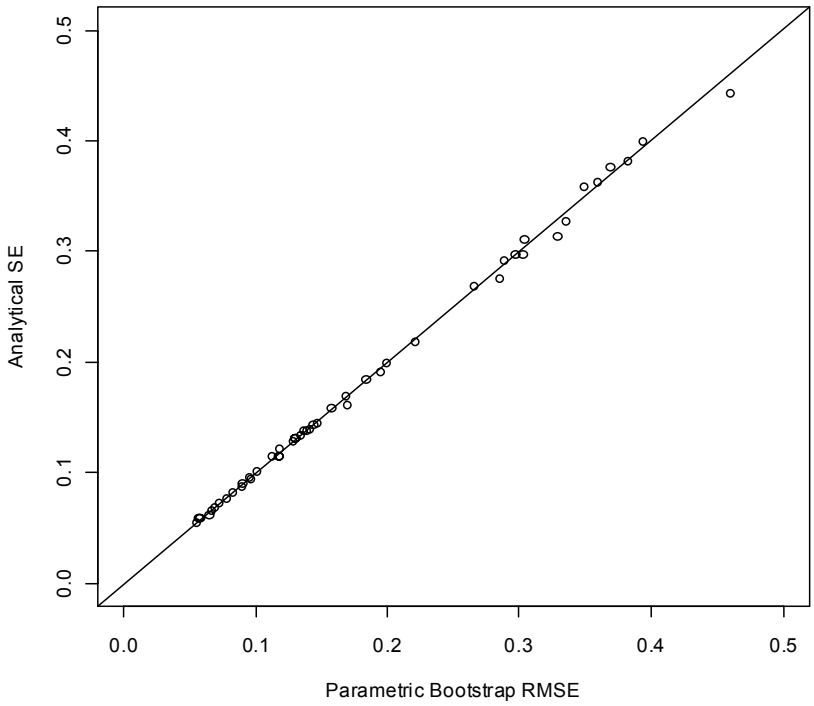
13.2 PQL Estimates of β_1 and β_2 for 2001 (REML used for variance components)

	Employed					*Unemployed*				
Variable	*Estimate*	*SE*	*Z*	*p-value*		*Estimate*	*SE*	*Z*	*p-value*	
Intercept	0.779	0.204	3.82	0.0001	*	–2.610	0.370	–7.05	< 0.0001	*
STATE2	0.024	0.045	0.53	0.5980		0.122	0.084	1.45	0.1470	
STATE3	0.131	0.048	2.74	0.0061	*	0.330	0.086	3.83	0.0001	*
STATE4	0.045	0.056	0.81	0.4170		0.119	0.104	1.15	0.2500	
STATE5	0.039	0.052	0.75	0.4520		0.146	0.095	1.53	0.1260	
STATE6	–0.018	0.071	–0.25	0.8020		0.220	0.124	1.77	0.0767	
STATE7	0.075	0.124	0.61	0.5450		0.135	0.219	0.62	0.5370	
STATE8	0.145	0.096	1.51	0.1310		0.101	0.177	0.57	0.5670	
AS2	–0.089	0.151	–0.59	0.5540		–0.148	0.267	–0.55	0.5800	
AS3	1.960	0.182	10.80	< 0.0001	*	1.650	0.287	5.77	< 0.0001	*
AS4	0.781	0.147	5.31	< 0.0001	*	0.200	0.282	0.71	0.4770	
AS5	2.460	0.200	12.30	< 0.0001	*	1.800	0.344	5.24	< 0.0001	*
AS6	0.715	0.180	3.98	< 0.0001	*	–0.133	0.358	–0.37	0.7110	
AS7	1.720	0.151	11.40	< 0.0001	*	0.966	0.265	3.64	0.0003	*
AS8	0.415	0.135	3.08	0.0021	*	–0.767	0.296	–2.59	0.0096	*
AS9	0.080	0.143	0.56	0.5770		–0.795	0.315	–2.52	0.0117	*
AS10	–0.766	0.151	–5.09	< 0.0001	*	–2.680	0.506	–5.30	< 0.0001	*
REMOTE2	0.036	0.058	0.62	0.5330		0.017	0.101	0.17	0.8690	
REMOTE3	0.138	0.042	3.27	0.0011	*	0.077	0.075	1.02	0.3080	
SEIFA2	–0.053	0.046	–1.15	0.2500		0.039	0.083	0.48	0.6350	
SEIFA3	–0.069	0.060	–1.16	0.2460		–0.083	0.108	–0.77	0.4400	
SEIFA4	–0.084	0.068	–1.24	0.2150		–0.141	0.122	–1.15	0.2500	
ASPAY1	–3.830	1.010	–3.79	0.0002	*	0.201	1.650	0.12	0.9030	
ASPAY2	–2.370	0.746	–3.18	0.0015	*	0.658	1.340	0.49	0.6230	
ASPAY3	–9.230	2.630	–3.51	0.0004	*	–7.110	3.780	–1.88	0.0601	
ASPAY4	–4.730	0.819	–5.78	< 0.0001	*	–1.030	1.680	–0.61	0.5410	
ASPAY5	–13.800	1.770	–7.78	< 0.0001	*	–9.390	2.850	–3.29	0.0010	*
ASPAY6	–4.400	0.746	–5.90	< 0.0001	*	–0.604	1.680	–0.36	0.7200	
ASPAY7	–11.200	1.300	–8.62	< 0.0001	*	–8.210	2.380	–3.45	0.0006	*
ASPAY8	–5.590	0.665	–8.41	< 0.0001	*	–1.780	1.760	–1.01	0.3120	
ASPAY9	–4.460	0.507	–8.81	< 0.0001	*	–1.430	1.260	–1.14	0.2540	
ASPAY10	–6.020	0.614	–9.82	< 0.0001	*	0.843	2.370	0.36	0.7230	
NSA_YA0	–0.697	0.897	–0.78	0.4370		7.510	1.450	5.16	< 0.0001	*
HH1	0.164	0.217	0.75	0.4510		0.741	0.394	1.88	0.0601	
HH2	–0.819	0.303	–2.70	0.0069	*	–0.911	0.557	–1.63	0.1030	
HH3	1.580	0.457	3.45	0.0006	*	2.580	0.938	2.75	0.0060	*
HH4	0.271	1.310	0.21	0.8360		3.610	2.660	1.36	0.1740	

13.3 Plot of analytical SEs versus bootstrap RMSEs of $\hat{\beta}$ for 2006



13.4 Plot of the small analytical SEs versus bootstrap RMSEs of $\hat{\beta}$ for 2006



13.5 Estimates of φ

	2001		2006	
Parameter	ML	REML	ML	REML
φ_1	0.0238	0.0280	0.0683	0.0758
φ_2	0.0468	0.0595	0.0691	0.0853
φ_{12}	0.0105	0.0125	0.0424	0.0466
ρ	0.316	0.307	0.617	0.580

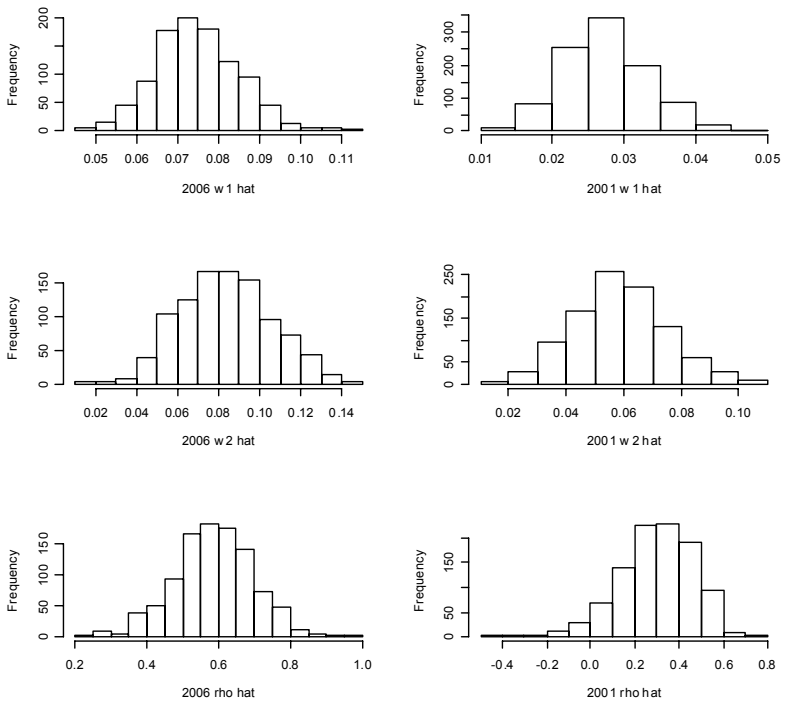
Table 13.6 contains the bootstrap results for $\boldsymbol{\varphi}$ for August 2006. In terms of bias, the REML estimator is performing reasonably well since all relative bias estimates are $<2\%$ in absolute value. From these results it is clear that the REML estimator is to be preferred over the ML estimator since not only are the estimated biases smaller (as one might expect), but the RRMSE estimates are also smaller for the REML estimator.

13.6 Comparison of the REML and ML estimators of φ for 2006

		Relative bias (%)		RRMSE (%)	
Parameter	Original estimate	ML	REML	ML	REML
φ_1	0.0758	–9.15	–1.56	15.56	13.22
φ_2	0.0853	–19.54	–1.85	31.87	27.08
ρ	0.580	6.71	0.80	22.68	20.15

It is also of interest to test the significance of the variance components $\boldsymbol{\varphi}$ (since if these are not significantly different from zero, then there is no point using a mixed effects model). In Molina *et al.* (2007), the authors test the significance of their single variance component using a likelihood ratio test (which is based on the approximate marginal likelihood). In theory we could apply a similar test here, but unfortunately the distribution of the likelihood ratio test statistic in our case will not be easy to derive under the null hypothesis. This is because under the null hypothesis $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_{12})^t = \mathbf{0}$ and $\boldsymbol{\varphi}$ is on the boundary of the parameter space which is a non standard condition. Molina *et al.* (2007) were able to apply this test because in the case of one variance component, the null distribution is known to be a mixture of two χ^2 distributions.

13.7 Histograms of $\hat{\varphi}_1$, $\hat{\varphi}_2$ and $\hat{\rho}$ for 2006 and 2001



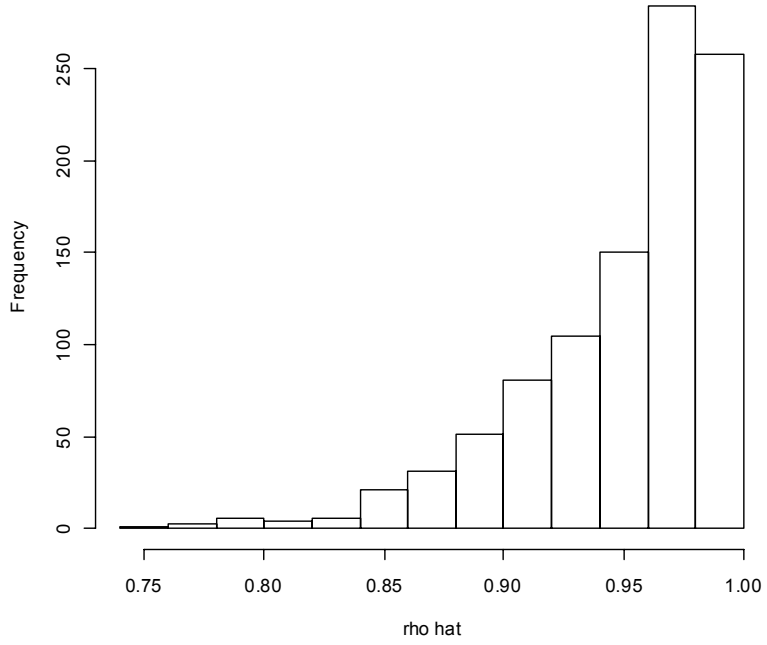
In any case, as an alternative, the previous parametric bootstrap used to produce table 13.6 will give some insight into the significance of the variance components. Using the 1000 REML estimates of each parameter, we build up empirical distributions under the fitted model of the REML estimators. Histograms of the 1000 sets of REML variance component estimates are given in figure 13.7. From these we can calculate approximate 95% parametric bootstrap confidence intervals using the percentile method. These confidence intervals are given in table 13.8. Notice that the confidence intervals for $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ for both years are not close to zero and the distributions look roughly symmetric, giving some evidence that the variance components are significant and a random effects model is appropriate in both years.

13.8 Approximate 95% confidence intervals for $\hat{\varphi}_1$, $\hat{\varphi}_2$ and $\hat{\rho}$ for 2006 and 2001

	2006		*2001*	
Parameter	*Lower*	*Upper*	*Lower*	*Upper*
φ_1	0.0556	0.0948	0.0169	0.0395
φ_2	0.0439	0.1280	0.0272	0.0937
ρ	0.371	0.788	–0.048	0.581

If we were to apply the model in Molina *et al.* (2007) to our LFS data, we would be imposing that $\boldsymbol{\varphi} = (\varphi_1, \varphi_1, \varphi_1)^t$. Tables 13.5 and 13.8 give some evidence that this more restrictive model may not be appropriate especially for the 2001 data. To test the null hypothesis of $\boldsymbol{\varphi} = (\varphi_1, \varphi_1, \varphi_1)^t$ for the 2006 data, a parametric bootstrap is used to generate empirical distributions under the null. The algorithm in Section 11 is used for this purpose, but note that we need to make one small change. Since we need to simulate under the null, in step (c) we replace all instances of u^*_{d2} with u^*_{d1} to ensure there is only one variance component. A histogram of the empirical distribution of $\hat{\rho}$ under the null hypothesis is given in figure 13.9 for the 2006 data. There is strong evidence to suggest that $\rho = 1$ does not hold since the observed $\hat{\rho}$ is 0.580 and is nowhere near any of the simulated $\hat{\rho}$ values under the null. Therefore our category specific random effects model appears to be more appropriate for our data than the more restrictive model used in Molina *et al.* (2007).

13.9 Histogram of $\hat{\rho}$ under the null for 2006



14. RESIDUAL PLOTS AND GOODNESS OF FIT TESTS

In the case of the multinomial logit model with no random effects or variance components, it is usual to assess model fit by computing various different summary statistics, measuring differences between the observed and fitted values. For example, as stated in Dobson (2001), one could use the Pearson χ^2 statistic, Deviance or the Likelihood ratio chi-squared statistic.

Suppose we have no random effects or variance components $\boldsymbol{\varphi}$ (the model is a GLM (generalised linear model)), then the standard Pearson χ^2 goodness of fit statistic is

$$\chi^2_p = \sum_{j=1}^{3} \sum_{d=1}^{D} \sum_{i=1}^{I_d} \frac{\left(y_{dij} - E_{dij}\right)^2}{E_{dij}},$$

where E_{dij} is the estimated expected value of y_{dij}. The Pearson χ^2 goodness of fit test is used to determine whether or not the model appears to hold. Given that the model holds and under appropriate asymptotic conditions (expected counts E_{dij} are large), the Pearson χ^2 statistic will follow an approximate $\chi^2_{2(N-p)}$ distribution, where $N = \Sigma_{d=1}^{D} I_d = 7,820$ is the total number of multinomial observations for the August 2006 data (similarly we could also compute this for August 2001) and $p = 37$ is the number of parameters in $\boldsymbol{\beta}_1$ (or $\boldsymbol{\beta}_2$). A large or small value of χ^2_p indicates that overall the model does not appear to hold (the null hypothesis is that the model holds).

Unfortunately we cannot apply the above test to our data since our model is not a GLM. The application of goodness of fit tests in a GLMM (generalised linear mixed model) situation is not straight forward theoretically. For instance, the observations are no longer independent and the Pearson statistic and other statistics are not guaranteed to have a χ^2 distribution under the null hypothesis even when the expected counts are large. Also, another problem is that the estimates E_{dij} are not easy to calculate as they involve integrals with no closed form solution.

There are two approaches that we consider to get around the above issues. The first is to use as E_{dij}, the conditional expectation estimates $m_{di}\hat{p}_{dij}$ predicted from the model and the second is to use an estimate of $E(y_{dij})$ from a Taylor series expansion about $\boldsymbol{u}_d = \mathbf{0}$, since in our case the variance components are small. The distribution of these χ^2 statistics can then be estimated by applying the parametric bootstrap of Section 11 using $B = 1000$.

For our multinomial logit mixed model sample data we have

$$E\left(y_{dij}\right) = E\left(E\left(y_{dij} \mid \boldsymbol{u}_d\right)\right) = m_{di} E\left(p_{dij}\right) \tag{14.1}$$

and

$$Var\left(y_{dij}\right) = E\left(Var\left(y_{dij} \mid \boldsymbol{u}_d\right)\right) + Var\left(E\left(y_{dij} \mid \boldsymbol{u}_d\right)\right)$$
$$= m_{di}E\left(p_{dij}\right)\left(1 - E\left(p_{dij}\right)\right) + m_{di}\left(m_{di} - 1\right)Var\left(p_{dij}\right). \qquad (14.2)$$

It can be shown using Taylor series expansions that when the variance components are small,

$$E\left(p_{di1}\right) \approx p^*_{di1} + \frac{1}{2}\begin{pmatrix} \varphi_1 p^*_{di1}\left(1 - 2p^*_{di1}\right)\left(1 - p^*_{di1}\right) \\ +\varphi_2\left(p^{*2}_{di2}p^*_{di1} - p^*_{di1}p^*_{di2}\left(1 - p^*_{di2}\right)\right) \\ +2\varphi_{12}\left(2p^{*2}_{di1}p^*_{di2} - p^*_{di1}p^*_{di2}\right) \end{pmatrix}, \qquad (14.3)$$

$$E\left(p_{di2}\right) \approx p^*_{di2} + \frac{1}{2}\begin{pmatrix} \varphi_1\left(p^{*2}_{di1}p^*_{di2} - p^*_{di2}p^*_{di1}\left(1 - p^*_{di1}\right)\right) \\ +\varphi_2 p^*_{di2}\left(1 - 2p^*_{di2}\right)\left(1 - p^*_{di2}\right) \\ +2\varphi_{12}\left(2p^{*2}_{di2}p^*_{di1} - p^*_{di1}p^*_{di2}\right) \end{pmatrix} \qquad (14.4)$$

and

$$E\left(p_{di3}\right) \approx p^*_{di3} + \frac{1}{2}\begin{pmatrix} \varphi_1\left(p^{*2}_{di1}p^*_{di3} - p^*_{di3}p^*_{di1}\left(1 - p^*_{di1}\right)\right) \\ +\varphi_2\left(p^{*2}_{di2}p^*_{di3} - p^*_{di3}p^*_{di2}\left(1 - p^*_{di2}\right)\right) \\ +4\varphi_{12}p^*_{di1}p^*_{di2}p^*_{di3} \end{pmatrix}, \qquad (14.5)$$

where p^*_{di1}, p^*_{di2} and p^*_{di3} are respectively p_{di1}, p_{di2} and p_{di3} with $\boldsymbol{u}_d$ replaced with the zero vector (the Taylor series expansions were taken about the point $\boldsymbol{u}_d = \boldsymbol{0}$). Also, we can show that

$$Var\left(p_{di1}\right) \approx \varphi_1 p^{*2}_{di1}\left(1 - p^*_{di1}\right)^2 + \varphi_2 p^{*2}_{di1}p^{*2}_{di2} - 2\varphi_{12}p^{*2}_{di1}\left(1 - p^*_{di1}\right)p^*_{di2}, \qquad (14.6)$$

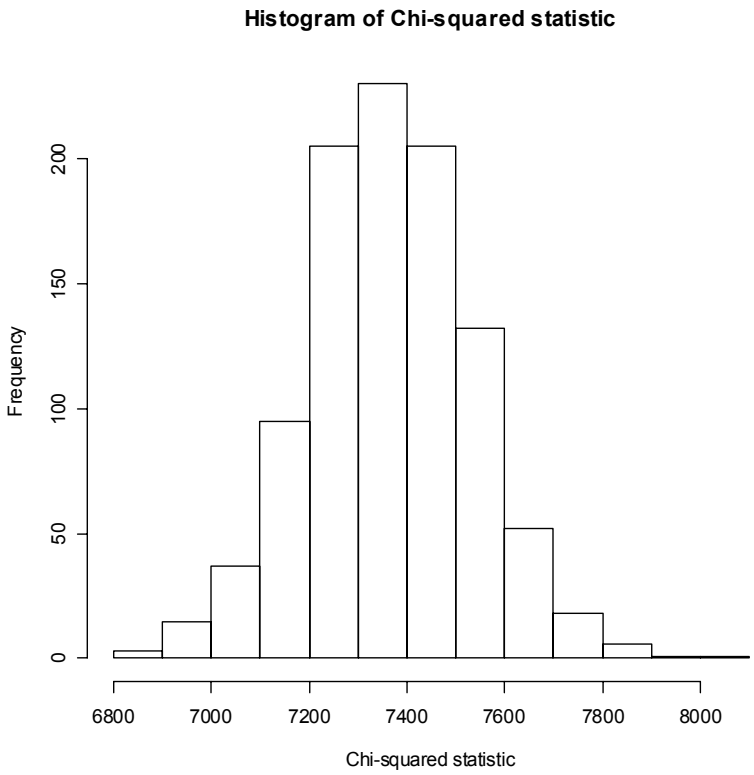
$$Var\left(p_{di2}\right) \approx \varphi_1 p^{*2}_{di1}p^{*2}_{di2} + \varphi_2 p^{*2}_{di2}\left(1 - p^*_{di2}\right)^2 - 2\varphi_{12}p^{*2}_{di2}p^*_{di1}\left(1 - p^*_{di2}\right) \qquad (14.7)$$

and

$$Var\left(p_{di3}\right) \approx \varphi_1 p^{*2}_{di1}p^{*2}_{di3} + \varphi_2 p^{*2}_{di2}p^{*2}_{di3} + 2\varphi_{12}p^*_{di1}p^*_{di2}p^{*2}_{di3}. \qquad (14.8)$$

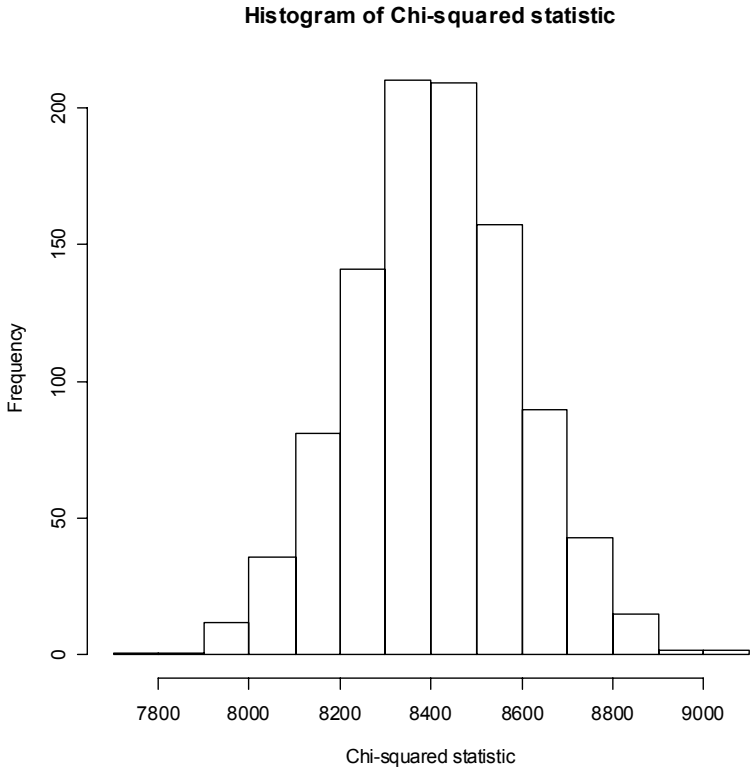
We are now at a point where we can approximate E_{dij} in the χ^2 statistic using estimates of $E(y_{dij})$ based on the Taylor series approximations (14.3)–(14.5) and noting the relationship (14.1). Note that within these approximations, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ are replaced by their estimates. Call this the unconditional approximation method. The other method as we mentioned earlier is to use the estimates of the conditional expectation $E(y_{dij}|\boldsymbol{u}_d)$ for E_{dij} in the χ^2 statistic. Call this method the conditional method.

14.1 Histogram of the χ^2 statistic using the conditional method for 2006



Using the parametric bootstrap of Section 11, we generate empirical distributions of the two χ^2 statistics under the fitted model based on a simulation of size $B = 1000$. The two distributions are summarised in figures 14.1 and 14.2 for the August 2006 data. An approximate 95% confidence interval for the conditional method is (7035, 7708) and an approximate 95% confidence interval for the unconditional approximation method is (8045, 8775). The values of the χ^2 statistic for the sample data are 7669 (conditional method) and 8703 (unconditional approximation method). Clearly these values are within the appropriate approximate confidence intervals, suggesting that overall both the models for y_{dij} and $y_{dij}|\boldsymbol{u}_d$ are adequate for the August 2006 sample data (similar conclusions can also be drawn for the August 2001 data, but details are not given here).

14.2 Histogram of the χ^2 statistic using the unconditional approximation method for 2006



One issue associated with goodness of fit tests is that they only produce one overall summary measure. It is of interest to also examine individual deviations and look for outliers/influential points. Residual plots are useful for this purpose. For each labour force status $j = 1, 2, 3$, define the following conditional standardised residuals for the in-sample data (Molina *et al.* (2007) call these Pearson residuals)

$$r_{dij}^{c} = \frac{y_{dij} - m_{di}\hat{p}_{dij}}{\sqrt{m_{di}\hat{p}_{dij}\left(1-\hat{p}_{dij}\right)}}.$$

We can also calculate approximate unconditional standardised residuals as follows

$$r_{dij}^{uc} = \frac{y_{dij} - \hat{E}\left(y_{dij}\right)}{\sqrt{\widehat{Var}\left(y_{dij}\right)}}$$

where $\hat{E}\left(y_{dij}\right)$ and $\widehat{Var}\left(y_{dij}\right)$ are estimates based on the earlier Taylor series approximations (14.3)–(14.8) and the relationships (14.1) and (14.2) with parameters $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ replaced with estimates.

Interestingly, we can also calculate the following summary statistics for each labour force category $j = 1, 2, 3$

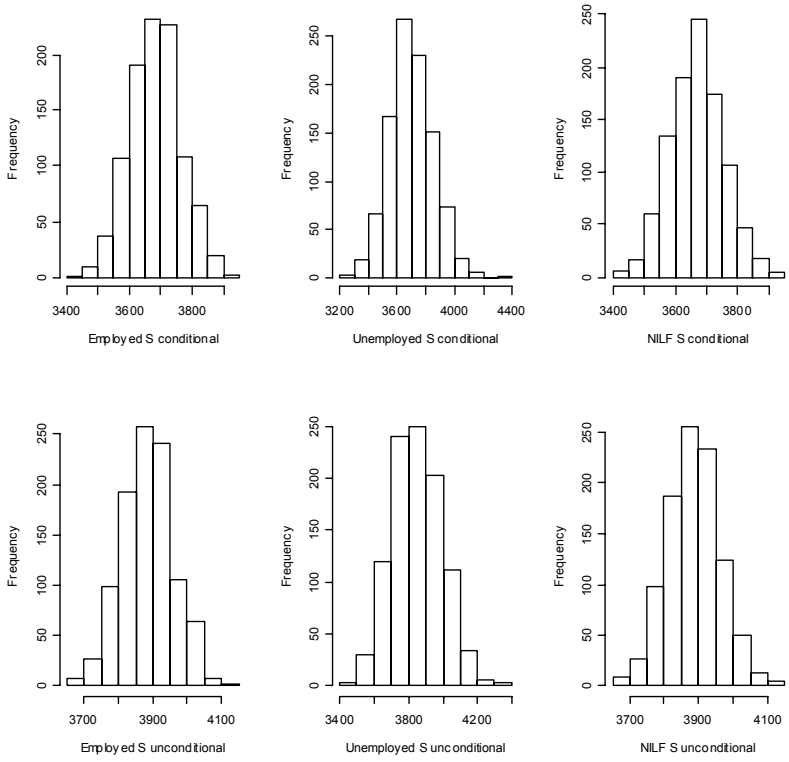
$$S_j^c = \sum_{d=1}^{D} \sum_{i=1}^{I_d} \left(r_{dij}^c \right)^2$$

and

$$S_j^{uc} = \sum_{d=1}^{D} \sum_{i=1}^{I_d} \left(r_{dij}^{uc} \right)^2$$

and compare these with appropriate empirical distributions generated from a parametric bootstrap under the fitted model with $B = 1000$ (again we use the bootstrap approach in Section 11). The empirical distributions are given in figure 14.3. Table 14.4 contains the S^c and S^{uc} values for the August 2006 sample data as well as approximate 95% parametric bootstrap confidence intervals using the percentile method. None of the S^c or S^{uc} values are within their corresponding confidence intervals. These statistics indicate that there may be underdispersion present for the Unemployed counts and overdispersion for both the Employed and NILF counts. Previously when we calculated the overall χ^2 statistics we did not obtain significant values. This is because the over and underdispersion in a way averaged themselves out overall.

14.3 Histogram of the S^c and S^{uc} statistics for 2006 from a parametric bootstrap



14.4 S^c and S^{uc} values and associated parametric bootstrap 95% confidence intervals

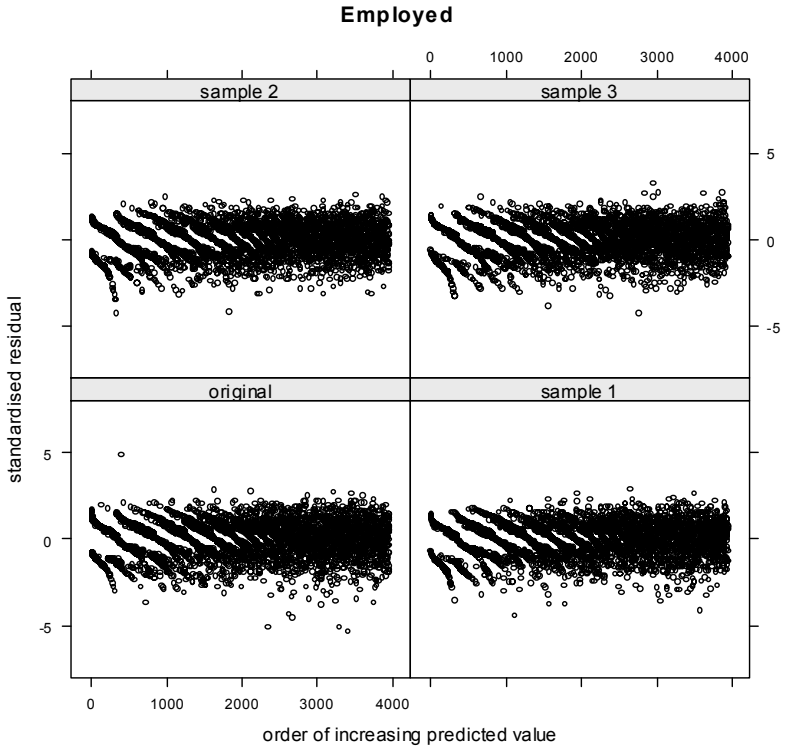
	Conditional			Unconditional		
LFS status	S^c	Lower	Upper	S^{uc}	Lower	Upper
Employed	4,224	3,522	3,848	4,306	3,741	4,030
Unemployed	3,393	3,407	4,000	3,565	3,578	4,137
Not in labour force	4,253	3,503	3,842	4,407	3,736	4,043

We now examine some residual plots to look more closely at this potential under and overdispersion issue. Figure 14.5 contains the standardised residuals r^c_{di1} versus the order of increasing predicted values $m_{di}\hat{p}_{di1}$ for the 2006 employed sample data. The panel labelled original contains the original sample and the other three panels contain values calculated from three parametric bootstrap samples. Similarly, figure 14.6 contains the unconditional standardised residuals r^{uc}_{di1} versus the order of increasing predicted values $\hat{E}(y_{di1})$. Similar plots for NILF and unemployed are given in figures 14.7–14.10.

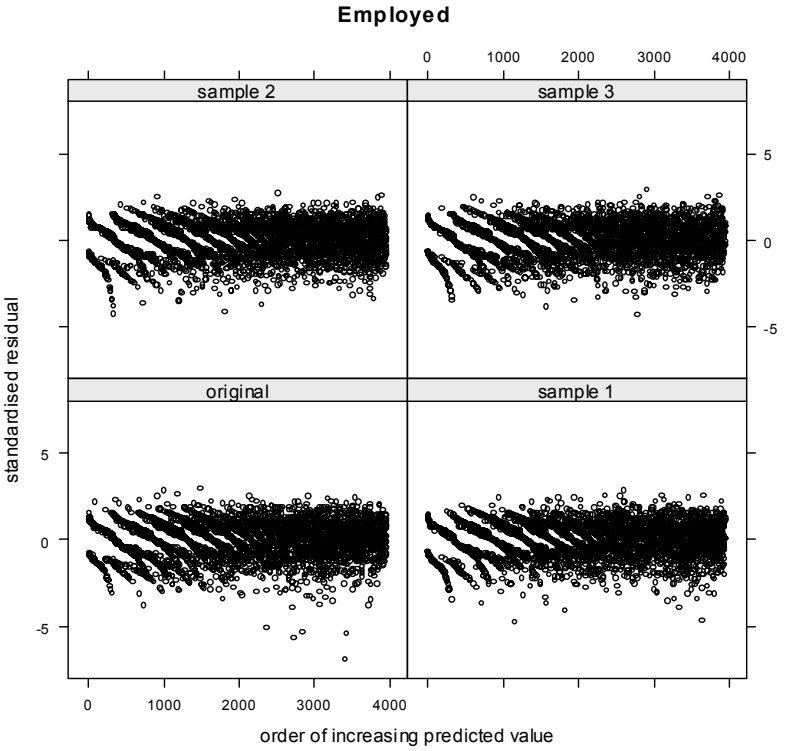
When comparing the bootstrap generated employed and NILF samples with the original data, it appears that the original data contain a small number of larger absolute residual values for both Employed and NILF. We actually recalculated the S^c and S^{uc} statistics by setting these few larger absolute residuals to zero for employed and NILF, but the resulting S^c and S^{uc} statistics were still significantly large. Hence these couple of larger absolute residuals are not solely responsible for the apparent overdispersion. In any case, when ignoring the small number of larger absolute residuals, the sample and original data distributions look roughly similar and hence we argue that the apparent under and overdispersion is not large enough for us to be overly concerned. In the significance tests we are clearly only picking up small significant differences and we suspect this is because our sample sizes are large.

We mentioned above that there were a small number of larger residuals in absolute value than one might expect when compared to the bootstrap samples. Most of these do not appear to be overly large and the NILF and employed ones mostly correspond to the same units. Upon further investigation, the only thing these outliering units have in common is that they are mostly all from remote areas (REMOTE3=1). Therefore the model may not be doing as well in the remote areas. Note that there is not much we can do about this issue because the sample sizes are small in the remote areas and we do not have any further covariates available.

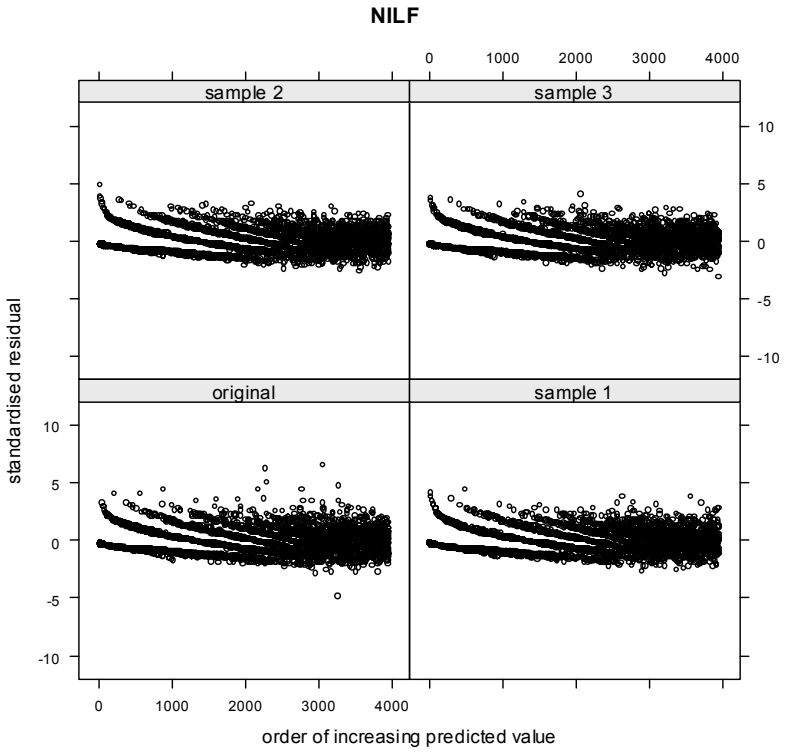
14.5 Plots of the conditional employed standardised residuals versus order of predicted values for original 2006 data and for three parametric bootstrap simulations



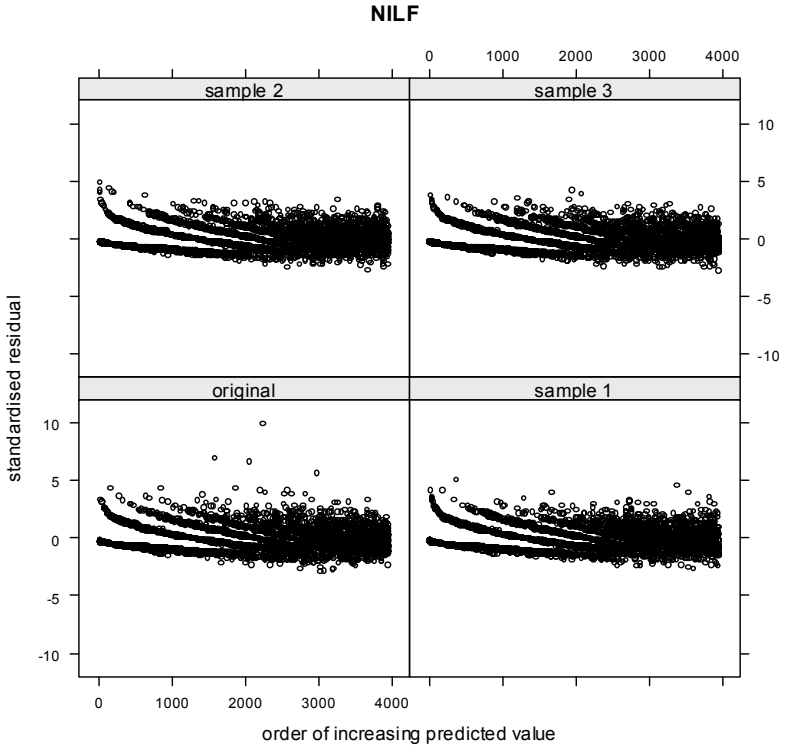
14.6 Plots of the unconditional employed standardised residuals versus predicted values for original 2006 data and for three parametric bootstrap simulations



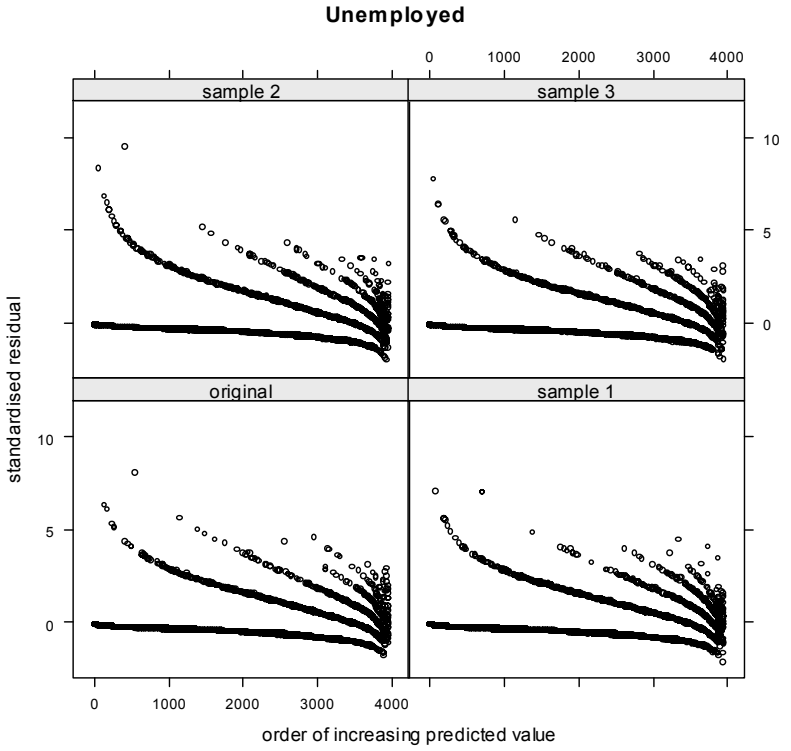
14.7 Plots of the conditional NILF standardised residuals versus order of predicted values for original 2006 data and for three parametric bootstrap simulations



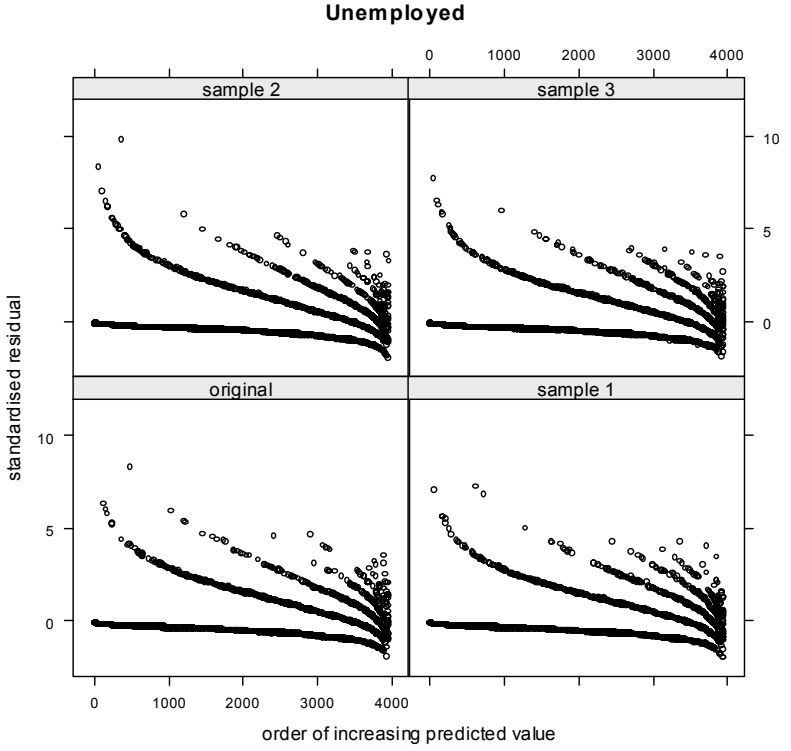
14.8 Plots of the unconditional NILF standardised residuals versus predicted values for original 2006 data and for three parametric bootstrap simulations



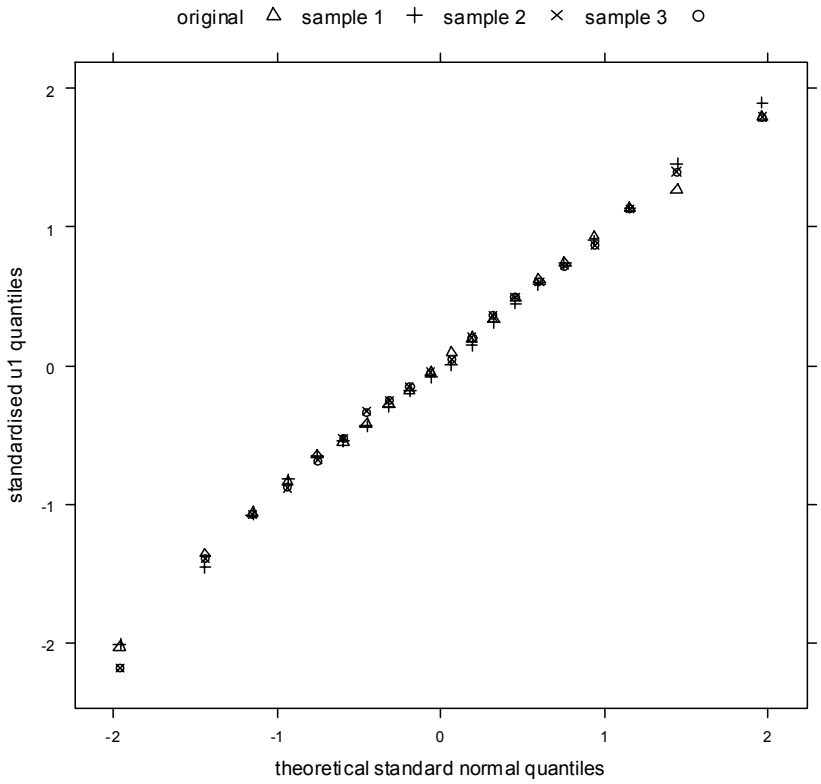
14.9 Plots of the conditional unemployed standardised residuals versus order of predicted values for original 2006 data and for three parametric bootstrap simulations



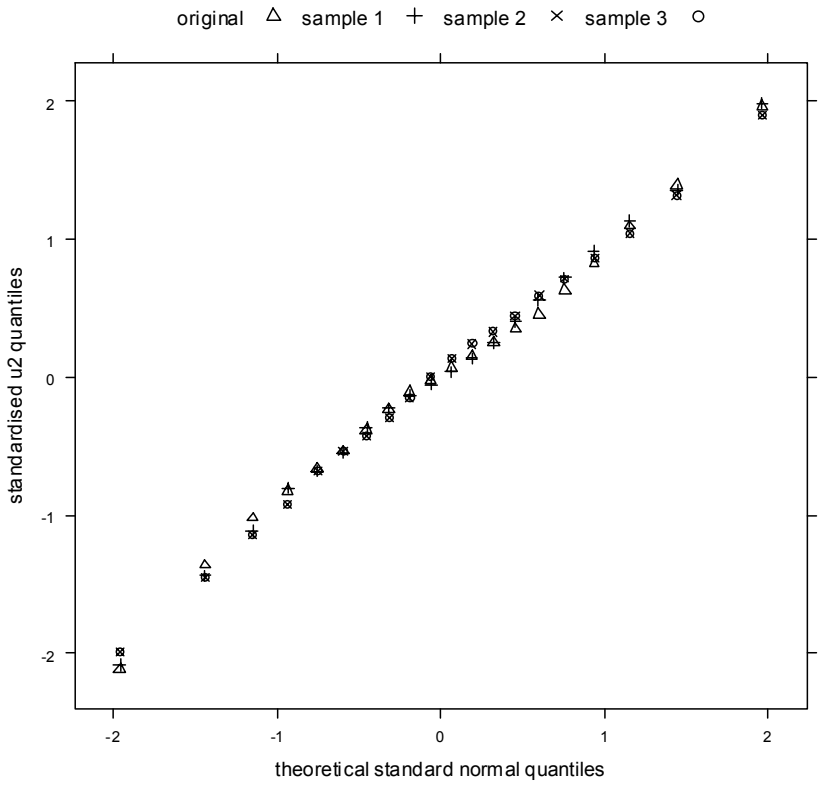
14.10 Plots of the unconditional unemployed standardised residuals versus predicted values for original 2006 data and for three parametric bootstrap simulations



14.11 Q–Q plots of $\hat{u}_{d1}$ for the original 2006 sample and three bootstrap samples

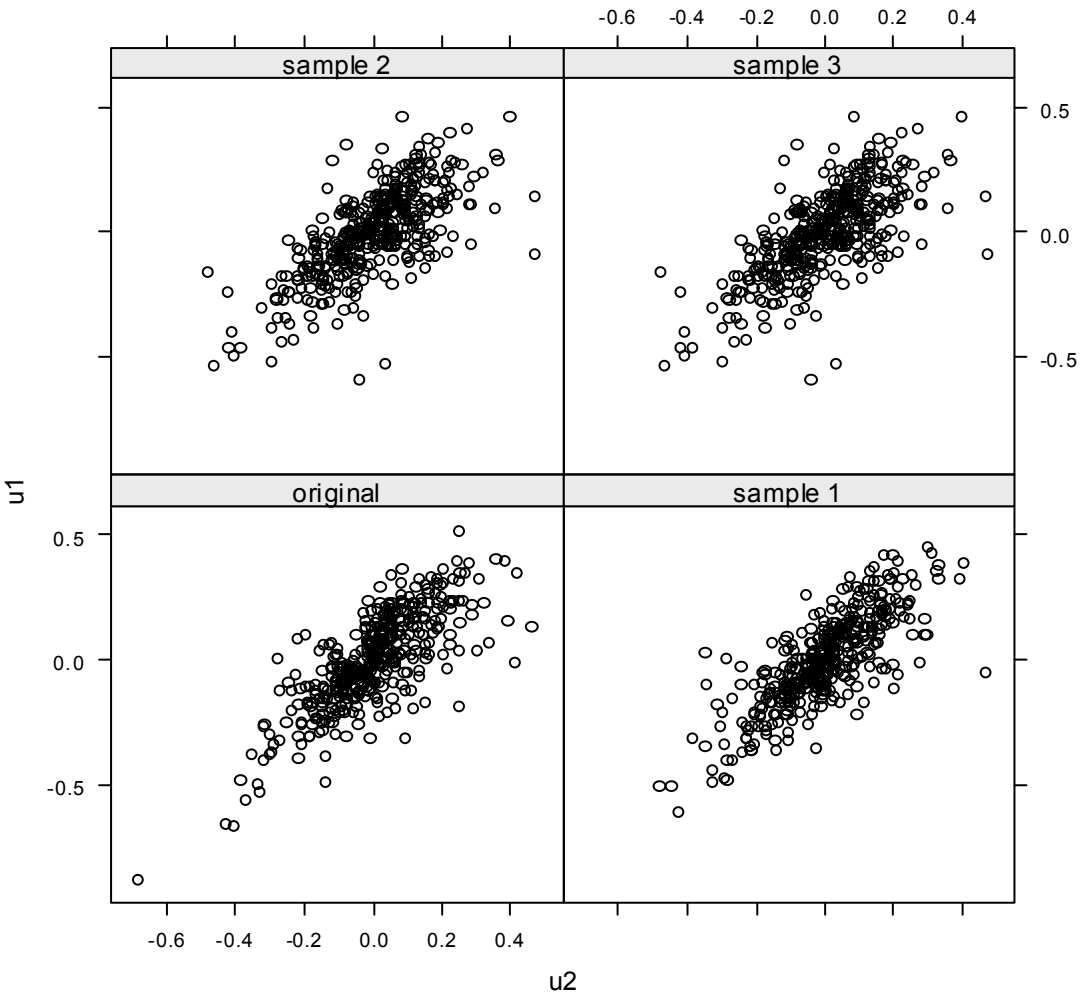


14.12 Q–Q plots of $\hat{u}_{d2}$ for the original 2006 sample and three bootstrap samples



So far we have said nothing about the predicted random effects $\hat{\boldsymbol{u}}$. Figure 14.11 contains a plot of the quantiles of standardised $\hat{u}_{d1}$ values versus theoretical standard normal quantiles. This is done for the original 2006 sample and three samples generated using the parametric bootstrap. Similarly, figure 14.12 contains a plot of the quantiles of standardised $\hat{u}_{d2}$ values versus theoretical standard normal quantiles. Figure 14.13 contains a plot of $\hat{u}_{d1}$ versus $\hat{u}_{d2}$. From these figures it appears that the original 2006 sample $\hat{\boldsymbol{u}}_d$ values behave similar to those obtained from 'typical' samples. The estimated $\hat{\boldsymbol{u}}_d$ values from the original sample therefore do not appear to give any indication of model departure.

14.13 Plots of $\hat{u}_{d1}$ versus $\hat{u}_{d2}$ for the original 2006 sample and three bootstrap samples



15. SMALL AREA ESTIMATES AND MSE ESTIMATES

Similar to figure 8 in Molina *et al.* (2007), figure 15.1 contains plots of the ratios of direct RSE estimates to model analytical RRMSE estimates versus sample sizes for 2006. The line $y = 1$ is also plotted. A ratio greater than 1 indicates we get gains by using the model based approach, whereas a ratio less than 1 indicates we get gains by using the direct survey estimation approach. Since all ratios are greater than 1 we are always getting gains by using the model based approach. The gains are quite large when the sample sizes are small and are small when the sample sizes are larger. Therefore when the sample size is small, the model based estimates have much small estimated MSEs than the direct survey estimates. Hence we have successfully reduced the MSEs by using a model based approach.

15.1 Plots of the ratios of direct RSE estimates to model analytical RRMSE estimates versus sample sizes for 2006

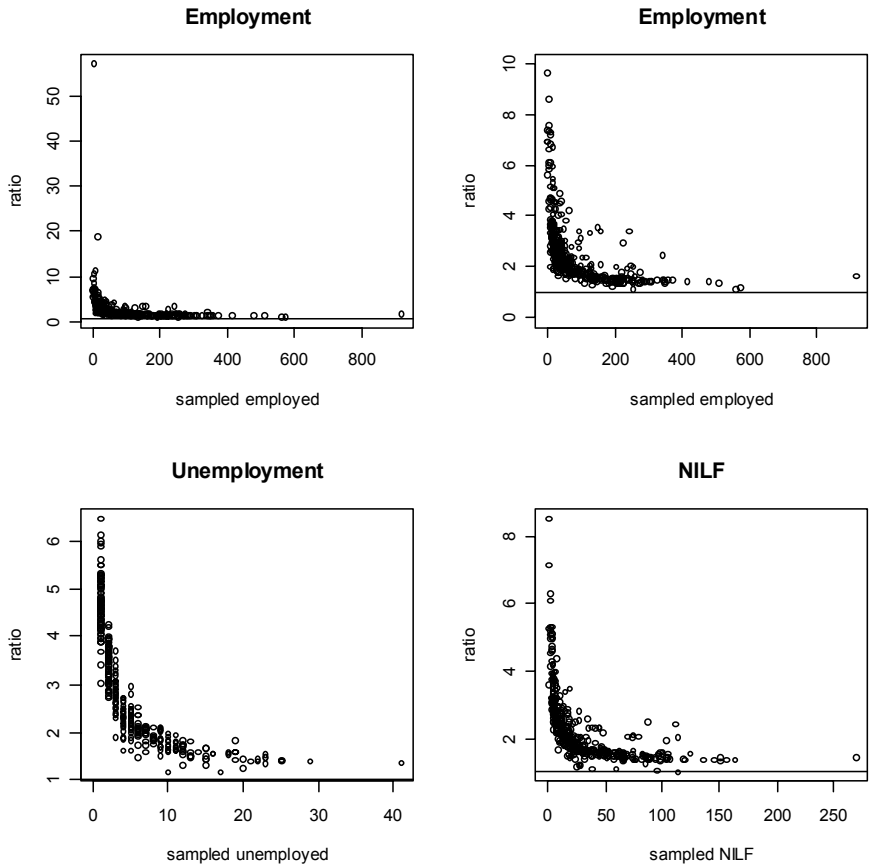
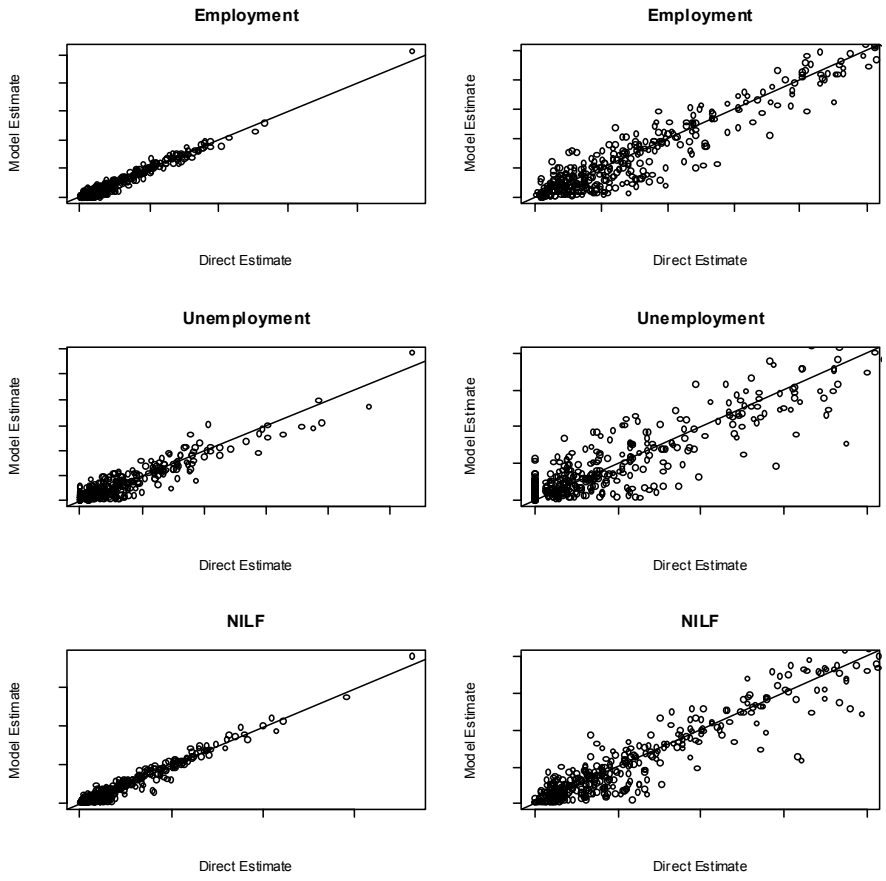


Figure 15.2 contains a comparison between the model based estimates (8.2) and the direct survey estimates for the August 2006 in-sample small areas. Note that the estimates of the NILF small area totals are obtained via subtraction. These plots are useful as a check for bias. For confidentiality reasons we have removed the actual numbers on the plots. The three plots on the left contain all of the estimates, whereas the plots on the right contain only the estimates closer to zero. The line $y = x$ is also drawn on these plots. Note that the direct survey estimates should be approximately design unbiased but with large standard errors. Figure 15.2 suggests that the model based estimates for Employed and NILF are roughly unbiased since although there is some variation, the estimates are distributed roughly about the line $y = x$. The unemployment model based estimates appear to be a little worse in some cases. For instance, when the direct estimates are large, the model based estimate tends to be smaller. However, for the most part, the model based estimators appear roughly unbiased or have a small bias.

15.2 Comparison between model based small area estimates and direct survey estimates for 2006



We now consider mean squared error estimation for the estimated small area totals. The mean squared error matrices can be estimated using two methods, either by an analytical approximation or a parametric bootstrap. For further details on the analytical approximation and the parametric bootstrap see Sections 8, 9 and 11. A comparison of the average percentage RRMSEs for 2006 derived from the analytical approximation and the parametric bootstrap with $B = 1000$ is given in table 15.3. On average the differences between the two methods are very small. The largest average absolute difference is for unemployment and this is only 0.83% and 0.62% for respectively the in-sample and out-of-sample small areas.

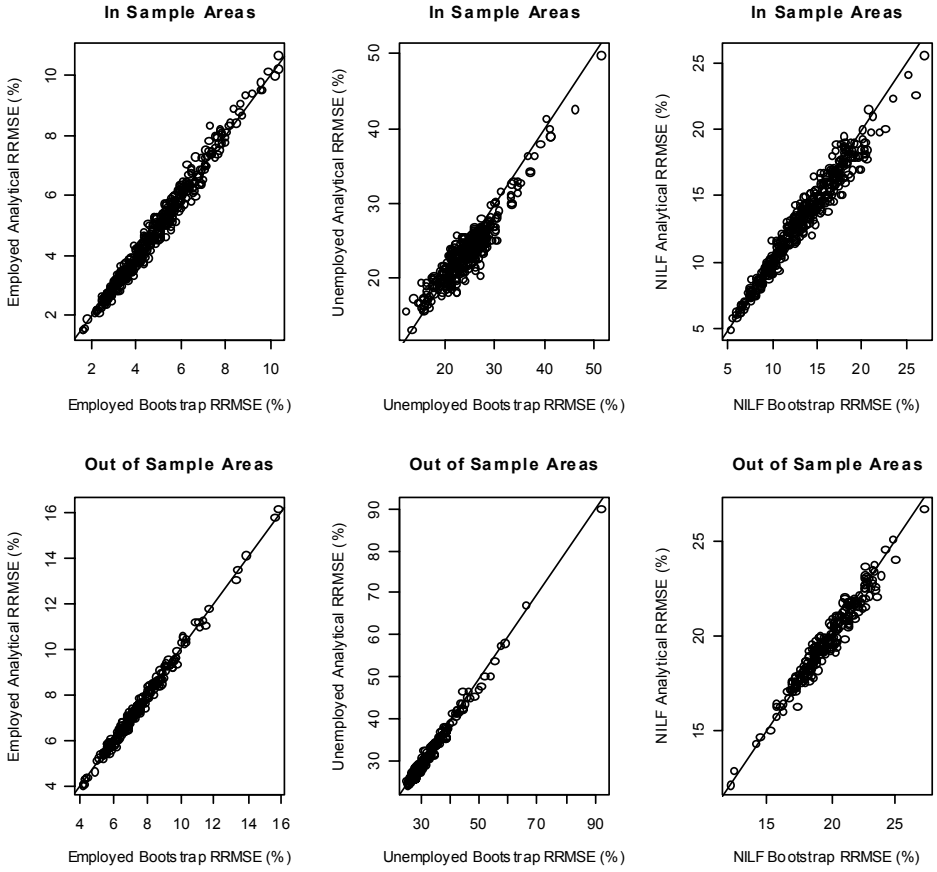
15.3 2006 Average RRMSE (%) estimates

	In-sample areas		*Out-of-sample areas*	
LFS status	*Bootstrap*	*Analytical*	*Bootstrap*	*Analytical*
Employed	4.84	4.86	7.44	7.47
Unemployed	24.52	23.69	32.30	31.68
Not in labour force	13.34	13.26	19.66	19.72

Figure 15.4 contains an overall comparison between the analytical and parametric bootstrap RRMSEs for August 2006. The line $y = x$ is also plotted. Figure 15.4 also confirms that on average the RRMSEs from both methods compare well since the values are roughly distributed about the line $y = x$. The in-sample unemployment analytical RRMSE estimates appear a little worse than the others on average since the analytical approximation looks to be on average slightly overestimating the smaller RRMSEs and underestimating the larger RRMSEs.

Table 15.5 contains the 2.5th and 97.5th percentiles of the distribution of the differences between the 2006 parametric bootstrap and the analytical RRMSE percentages. Figure 15.4 and table 15.5 shows that there is some variability in the differences between the parametric bootstrap and analytical RRMSEs. However, this variability is not overly large with the worse case being for the in-sample unemployed and the majority of these differences are $< 4\%$ in absolute value. Note also that some of these larger differences could also be due to extra variation resulting from the parametric bootstrap since $B = 1000$ is only a moderately large value.

15.4 Comparison between analytical and bootstrap RRMSEs



15.5 2.5th and 97.5th percentiles of the distribution of the differences between the 2006 parametric bootstrap and analytical RRMSE percentages

	In-sample areas		*Out-of-sample areas*	
LFS status	*2.5th percentile*	*97.5th percentile*	*2.5th percentile*	*97.5th percentile*
Employed	–0.50	0.50	–0.37	0.32
Unemployed	–2.54	4.24	–1.03	2.45
Not in labour force	–1.50	2.12	–0.95	1.00

16. CONCLUSION

This paper successfully adapted the model and estimators described in Molina *et al.* (2007) to include category specific random effects. We showed that the category specific multinomial logit mixed model is more appropriate for our dataset than the more restrictive one given by Molina *et al.* (2007). The PQL-REML estimation procedure worked very well in our context and we showed via a parametric bootstrap that the PQL-REML estimators had good statistical properties. For instance, the bias in the REML variance component estimates was found to be small.

Similar to Molina *et al.* (2007), we described and derived two different estimators of the mean squared errors of the small area estimated totals. These two different approaches are based on using an analytical approximation and a parametric bootstrap. In the paper by Molina *et al.* (2007), the authors undertook a simulation study and concluded that the bootstrap estimator performed better than the analytical approximation and recommended the bootstrap be used. However they noted that the differences were smaller for the actual UK unemployment data. We showed that for the Australian labour force data that the analytical approximation RRMSEs compared very well with the parametric bootstrap RRMSEs and the differences were all reasonably small. In our context we recommend that the analytical RRMSEs be used because our parametric bootstrap is much more computationally intensive than the one given in Molina *et al.* (2007). We believe the small gains in accuracy will not be worth the extra computational effort involved for the parametric bootstrap in our case.

In this paper we used residual plots and χ^2 goodness of fit tests to check model assumptions. We used a parametric bootstrap to generate the empirical distributions of the χ^2 statistic. These tests and plots indicated that the model assumptions appeared to hold approximately for the sample data. There was very slight under and overdispersion present and a couple of small outliers for remote areas. In a future study we might try to improve the model for remote areas. In any case, for the most part the multinomial logit mixed model appears to work reasonably well for modelling the Australian Labour Force count data. Interestingly, the multinomial model has quite a restrictive variance and correlation structure and the fact that the multinomial model works so well here is very convenient. This is because extending the model to account for under and overdispersion in a small area context would not be straight forward. This would certainly be an interesting topic for future research.

Another future research topic could be to try account better for the sample design and any design informativeness. Our estimators like those in the Molina *et al.* (2007) paper essentially assume that the Labour Force sample has been collected using SRSWOR.

ACKNOWLEDGEMENTS

This work was initially intended to form a small introductory part of my PhD thesis and has since evolved into a separate more detailed joint ABS and ANU research paper. As such, I would like to acknowledge my PhD supervisor Professor Alan Welsh for his technical advice and continual support throughout this work and for putting me onto this topic in the first place.

I would also like to thank my small area estimation colleagues, in particular Daniel Elazar for their support and help and for allowing me to work on this topic part time at the ABS. Also, I would especially like to thank those ABS staff who manipulated the data and got it into a useable form which I could then run the multinomial logit mixed models on.

I would also like to acknowledge DEEWR, the Department of Education, Employment and Workplace Relations, for making their administrative data available to use, and I very much appreciate their effort.

Finally, I would like to acknowledge two other ABS colleagues. Thanks especially to Peter Rossiter for his substantial help with formatting this document. Also, thanks very much to Frank Yu for proof reading this paper and providing some very useful and relevant comments.

REFERENCES

Australian Bureau of Statistics (2001) *Australian Standard Geographical Classification (ASGC), 2001*, cat. no. 1216.0, ABS, Canberra.

—— (2003) *Information Paper: Census of Population and Housing – Socio-Economic Indexes for Areas, Australia, 2001,* cat. no. 2039.0, ABS, Canberra.

—— (2007) *Regional Population Growth,* cat. no. 3218.0, ABS, Canberra.

Baillo, A. and Molina, I. (2005) *Mean Squared Errors of Small Area Estimators under a Unit-Level Multivariate Model*, Statistics and Econometrics Working Paper ws054007, Universidad Carlos III de Madrid, Madrid. (last viewed 13 January 2010)
< http://ideas.repec.org/p/cte/wsrepe/ws054007.html >

Booth, J.G. and Hobert, J.P. (1999) "Maximising Generalised Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm", *Journal of the Royal Statistical Society: Series B (Statistical Methodology*), 61(1), pp. 265–285.

Breslow, N.E. and Clayton, D.G. (1993) "Approximate Inference in Generalized Linear Mixed Models", *Journal of the American Statistical Association*, 88(421), pp. 9–25.

Dobson, A.J. (2001) *An Introduction to Generalized Linear Models*, Second Edition, Chapman and Hall/CRC Press, London.

Geweke, J. (1996) "Monte Carlo Simulation and Numerical Integration", in H.M. Amman, D.A. Kendrick and J. Rust (eds), *Handbook of Computational Economics, Volume I*, Elsevier Science Publishers B.V., Amsterdam.

Hartzel, J.; Agresti, A. and Caffo, B. (2001) "Multinomial Logit Random Effects Models", *Statistical Modelling*, 1, pp. 81–102.

Harville, D.A. (1977) "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems", *Journal of the American Statistical Association*, 72(358), pp. 320–340.

Henderson, H.V. and Searle, S.R. (1981) "On Deriving the Inverse of a Sum of Matrices", *SIAM Review*, 23(1), pp. 53–60.

Jiang, J. (1998) "Consistent Estimators in Generalised Linear Mixed Models", *Journal of the American Statistical Association*, 93(442), pp. 720–729.

Jiang, J. (2007) *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Science and Business Media LLC, New York.

Kackar, R.N. and Harville, D.A. (1984) "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models", *Journal of the American Statistical Association*, 79(388), pp. 853–862.

Lee, Y. and Nelder, J.A. (1996) "Hierarchical Generalized Linear Models (with Discussion)", *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4), pp. 619–678.

Molina, I.; Saei, A. and Lombardía, M.J. (2007) "Small Area Estimates of Labour Force Participation under a Multinomial Logit Mixed Model", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4), pp. 975–1000.

Pawitan, Y. (2001) *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press.

Prasad, N.G.N. and Rao, J.N.K. (1990) "The Estimation of the Mean Squared Error of Small-Area Estimators", *Journal of the American Statistical Association*, 85(409), pp. 163–171.

Rao, C.R. and Kleffe, J. (1988) *Estimation of Variance Components and Applications*, Elsevier Science Publishers B.V., Amsterdam.

Saei, A. and Chambers, R. (2003) *Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects*, S3RI Methodology Working Papers, M03/15, Southampton Statistical Sciences Research Institute. (last viewed 13 January 2010) < http://eprints.soton.ac.uk/8165/ >

Saei, A. and Chambers, R. (2005) *Empirical Best Linear Unbiased Prediction for Out of Sample Areas*, S3RI Methodology Working Papers, M05/03, Southampton Statistical Sciences Research Institute. (last viewed 13 January 2010) < http://eprints.soton.ac.uk/14073/ >

Schall, R. (1991) "Estimation in Generalized Linear Models with Random Effects", *Biometrika*, 78(4), pp. 719–727.

APPENDIX

This appendix contains the proof of $E\left(\boldsymbol{G}_1\left(\hat{\boldsymbol{\varphi}}\right)\right) \approx \boldsymbol{G}_1\left(\boldsymbol{\varphi}\right) - \boldsymbol{G}_3\left(\boldsymbol{\varphi}\right)$ (see Section 8).

First we take a second order Taylor series expansion (and assume that $\hat{\boldsymbol{\Sigma}}$ in $\hat{\boldsymbol{T}}$ does not depend on $\hat{\boldsymbol{\varphi}}$)

$$\begin{aligned}\hat{\boldsymbol{T}} = \boldsymbol{T}\left(\hat{\boldsymbol{\varphi}}\right) \approx \boldsymbol{T}\left(\boldsymbol{\varphi}\right) &+ \frac{\partial \boldsymbol{T}}{\partial \varphi_1}\left(\hat{\varphi}_1 - \varphi_1\right) + \frac{\partial \boldsymbol{T}}{\partial \varphi_2}\left(\hat{\varphi}_2 - \varphi_2\right) \\ &+ \frac{\partial \boldsymbol{T}}{\partial \varphi_{12}}\left(\hat{\varphi}_{12} - \varphi_{12}\right) + \frac{1}{2}\frac{\partial^2 \boldsymbol{T}}{\partial \varphi_1^2}\left(\hat{\varphi}_1 - \varphi_1\right)^2 + \frac{1}{2}\frac{\partial^2 \boldsymbol{T}}{\partial \varphi_2^2}\left(\hat{\varphi}_2 - \varphi_2\right)^2 \\ &+ \frac{1}{2}\frac{\partial^2 \boldsymbol{T}}{\partial \varphi_{12}^2}\left(\hat{\varphi}_{12} - \varphi_{12}\right)^2 + \frac{\partial^2 \boldsymbol{T}}{\partial \varphi_1 \partial \varphi_2}\left(\hat{\varphi}_1 - \varphi_1\right)\left(\hat{\varphi}_2 - \varphi_2\right) \\ &+ \frac{\partial^2 \boldsymbol{T}}{\partial \varphi_1 \partial \varphi_{12}}\left(\hat{\varphi}_1 - \varphi_1\right)\left(\hat{\varphi}_{12} - \varphi_{12}\right) + \frac{\partial^2 \boldsymbol{T}}{\partial \varphi_2 \partial \varphi_{12}}\left(\hat{\varphi}_2 - \varphi_2\right)\left(\hat{\varphi}_{12} - \varphi_{12}\right).\end{aligned}$$

Assuming that $\boldsymbol{T}$ is approximately constant and does not depend on $\boldsymbol{u}$ (technically $\boldsymbol{T}$ depends on $\boldsymbol{u}$) and $E(\hat{\boldsymbol{\varphi}}) \approx \boldsymbol{\varphi}$, then

$$\begin{aligned}E\left(\boldsymbol{T}\left(\hat{\boldsymbol{\varphi}}\right)\right) \;\approx\; \boldsymbol{T}\left(\boldsymbol{\varphi}\right) &+ \frac{1}{2}\frac{\partial^2 \boldsymbol{T}}{\partial \varphi_1^2} E\left(\left(\hat{\varphi}_1 - \varphi_1\right)^2\right) + \frac{1}{2}\frac{\partial^2 \boldsymbol{T}}{\partial \varphi_2^2} E\left(\left(\hat{\varphi}_2 - \varphi_2\right)^2\right) \\ &+ \frac{1}{2}\frac{\partial^2 \boldsymbol{T}}{\partial \varphi_{12}^2} E\left(\left(\hat{\varphi}_{12} - \varphi_{12}\right)^2\right) + \frac{\partial^2 \boldsymbol{T}}{\partial \varphi_1 \partial \varphi_2} E\left(\left(\hat{\varphi}_1 - \varphi_1\right)\left(\hat{\varphi}_2 - \varphi_2\right)\right) \\ &+ \frac{\partial^2 \boldsymbol{T}}{\partial \varphi_1 \partial \varphi_{12}} E\left(\left(\hat{\varphi}_1 - \varphi_1\right)\left(\hat{\varphi}_{12} - \varphi_{12}\right)\right) + \frac{\partial^2 \boldsymbol{T}}{\partial \varphi_2 \partial \varphi_{12}} E\left(\left(\hat{\varphi}_2 - \varphi_2\right)\left(\hat{\varphi}_{12} - \varphi_{12}\right)\right). \qquad \text{(A.1)}\end{aligned}$$

After some algebra and making use of (8.16), (8.18) and (8.19) it can be proved that for $a = 1, 2, 12$ and $b = 1, 2, 12$,

$$\begin{aligned}\frac{\partial^2 \boldsymbol{T}}{\partial \varphi_a \partial \varphi_b} = &-\frac{\partial \boldsymbol{W}}{\partial \varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_b} - \frac{\partial \boldsymbol{W}}{\partial \varphi_b}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_a} \\ &+ \frac{\partial \boldsymbol{W}}{\partial \varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_b}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{W} + \frac{\partial \boldsymbol{W}}{\partial \varphi_b}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{W} \\ &+ \boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_b} + \boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_b}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_a} \\ &- \boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_b}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{W} \\ &- \boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_b}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{W}}{\partial \varphi_a}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{W}. \qquad \text{(A.2)}\end{aligned}$$

Now, assuming that $\widehat{\boldsymbol{M}}_d \approx \boldsymbol{M}_d$ (i.e. $\widehat{\boldsymbol{M}}_d$ is approximately constant), then

$$E\left(\boldsymbol{G}_1\left(\hat{\boldsymbol{\varphi}}\right)\right) \approx \boldsymbol{M}_d E\left(\hat{\boldsymbol{T}}\right)\boldsymbol{M}_d^t = \begin{pmatrix} \boldsymbol{M}_{d1}E\left(\hat{\boldsymbol{T}}\right)\boldsymbol{M}_{d1}^t & \boldsymbol{M}_{d1}E\left(\hat{\boldsymbol{T}}\right)\boldsymbol{M}_{d2}^t \\ \boldsymbol{M}_{d2}E\left(\hat{\boldsymbol{T}}\right)\boldsymbol{M}_{d1}^t & \boldsymbol{M}_{d2}E\left(\hat{\boldsymbol{T}}\right)\boldsymbol{M}_{d2}^t \end{pmatrix}. \tag{A.3}$$

Now substitute (A.1) into (A.3) and using (8.12) and (A.2) we obtain

$$\begin{aligned} E\left(\boldsymbol{G}_1\left(\hat{\boldsymbol{\varphi}}\right)\right) \approx \boldsymbol{G}_1\left(\boldsymbol{\varphi}\right) &- \begin{pmatrix} h_{1,1}^{(1,1)} & h_{1,1}^{(1,2)} \\ h_{1,1}^{(1,2)} & h_{1,1}^{(2,2)} \end{pmatrix} - \begin{pmatrix} h_{2,2}^{(1,1)} & h_{2,2}^{(1,2)} \\ h_{2,2}^{(1,2)} & h_{2,2}^{(2,2)} \end{pmatrix} - \begin{pmatrix} h_{12,12}^{(1,1)} & h_{12,12}^{(1,2)} \\ h_{12,12}^{(1,2)} & h_{12,12}^{(2,2)} \end{pmatrix} \\ &- \begin{pmatrix} h_{1,2}^{(1,1)} & h_{1,2}^{(1,2)} \\ h_{1,2}^{(1,2)} & h_{1,2}^{(2,2)} \end{pmatrix} - \begin{pmatrix} h_{2,1}^{(1,1)} & h_{2,1}^{(1,2)} \\ h_{2,1}^{(1,2)} & h_{2,1}^{(2,2)} \end{pmatrix} - \begin{pmatrix} h_{1,12}^{(1,1)} & h_{1,12}^{(1,2)} \\ h_{1,12}^{(1,2)} & h_{1,12}^{(2,2)} \end{pmatrix} \\ &- \begin{pmatrix} h_{12,1}^{(1,1)} & h_{12,1}^{(1,2)} \\ h_{12,1}^{(1,2)} & h_{12,1}^{(2,2)} \end{pmatrix} - \begin{pmatrix} h_{2,12}^{(1,1)} & h_{2,12}^{(1,2)} \\ h_{2,12}^{(1,2)} & h_{2,12}^{(2,2)} \end{pmatrix} - \begin{pmatrix} h_{12,2}^{(1,1)} & h_{12,2}^{(1,2)} \\ h_{12,2}^{(1,2)} & h_{12,2}^{(2,2)} \end{pmatrix} \end{aligned}$$

where for $a = 1, 2, 12$, $b = 1, 2, 12$, $j = 1, 2$ and $k = 1, 2$,

$$h_{a,b}^{(j,k)} = \left(\frac{\partial\left(\boldsymbol{M}_{dj}\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\right)}{\partial\varphi_a}\right)\boldsymbol{V}\left(\frac{\partial\left(\boldsymbol{M}_{dk}\boldsymbol{W}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\right)}{\partial\varphi_b}\right)^t E\left(\left(\hat{\varphi}_a - \varphi_a\right)\left(\hat{\varphi}_b - \varphi_b\right)\right)$$

and therefore

$$\begin{aligned} E\left(\boldsymbol{G}_1\left(\hat{\boldsymbol{\varphi}}\right)\right) &\approx \boldsymbol{G}_1\left(\boldsymbol{\varphi}\right) - \begin{pmatrix} g_{11}^* & g_{12}^* \\ g_{12}^* & g_{22}^* \end{pmatrix} \\ &= \boldsymbol{G}_1\left(\boldsymbol{\varphi}\right) - \boldsymbol{G}_3\left(\boldsymbol{\varphi}\right). \end{aligned}$$

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au