Confidence intervals for posterior intercepts, with application to the PIAAC literacy survey

Jochen Einbeck[1], Elizabeth Gray[12], Nick Sofroniou[3], Antonio Hermes Marques da Silva Junior[14], Jacob Gledhill[15]

[1] Durham University, UK
[2] University of Bath, UK
[3] Institute for Employment Research, University of Warwick, UK
[4] Universidade Federal do Rio Grande do Norte, Natal, Brazil
[5] Department for Communities and Local Government, UK

E-mail for correspondence: jochen.einbeck@durham.ac.uk

Abstract: For variance component models, it is often the posterior estimate of the random effect ('posterior intercept') rather than the estimate of the fixed effect parameters, which is of main interest. This is the case, for instance, when ranking region–wise mortality rates (where the crude, regional rates are unreliable due to small observed counts) or for the construction of educational league tables from complex sample surveys. However, in order to be able to decide whether two cluster–level units can *actually be distinguished*, it is clear that one needs a measure of variability of these posterior intercepts. We present an exploration of methods to address this issue which appears to be still undeveloped in the context of the model class considered.

Keywords: Nonparametric maximum likelihood; Empirical Bayes shrinkage; Bootstrap

1 Posterior intercepts

Consider variance component models of type

$$\mu_{ij} = E(y_{ij}|z_i) = h(x_{ij}^{\mathrm{T}}\beta + z_i), \tag{1}$$

where μ_{ij} is the expected response for unit j in cluster i, x_{ij} are the fixed effect covariates which may depend on i, j, or both, and z_i is the random

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

effect operating at the cluster level. If no assumption on the random effect distribution is made, then estimation can be carried out via 'nonparametric maximum likelihood' (Aitkin, 1999). Briefly, the marginal likelihood is approximated by a discrete mixture, the parameters of which are estimated alongside with the fixed effect parameters via the EM algorithm, yielding estimates $\hat{\beta}$, $\hat{z}_1, \ldots, \hat{z}_k$ with masses $\hat{\pi}_1, \ldots, \hat{\pi}_K$. Denote by $\hat{\theta}$ the collection of these estimates, and by $y_i = (y_{i\cdot})$ the set of response values for cluster i. Aitkin (1999) suggested to estimate the mean of the posterior distribution $z_i|y_i$ via 'Empirical Bayes Predictions'

$$\tilde{z}_i = \sum_{k=1}^{K} w_{ik} \hat{z}_k, \tag{2}$$

where $w_{ik} = \hat{P}(k|\hat{\theta}, y_i)$ are the posterior probabilities ('responsibilities') that observation i stems from component k, which can be computed via Bayes' theorem from the parameter estimates $\hat{\theta}$ of the last M step. The quantity of interest are these posterior intercepts, $\tilde{z}_i$.

2 PIAAC data

The PIAAC survey of adult skills was carried out from 01/08/2011 to 31/03/2012 by the OECD in 24 countries (or sub–country entities), and was designed to assess the proficiency of adults in the key competencies of literacy, numeracy, and problem–solving in technology–rich environments. We focus here on the 'literacy' output variable, with six possible outcomes for an assessed individual. We dichotomized this variable as 'people reaching level 3 or above', with 'level 2 and below' being considered as low–skilled, which corresponds to the key European Commission policy marker used to demarcate poor basic skills in the complementary PISA survey carried out at 15–years of age (Eurostat, 2016). As covariates we will use gender, as well as a factor for age (covering the intervals 16–24, 25–34, 35–44, 45–54, and 55+), though we have also explored more complex models using employment status and reading habits which are not reported here. This leads to a (rescaled) logistic regression model $y_{ij} \sim \text{Bin}(n_{ij}, \mu_{ij})/n_{ij}$ where n_{ij} is the (effective) sample size of the jth subpopulation (defined by the covariate combination of interest) for country i, and function $h(\cdot)$ in (1) is the logistic function. Data were extracted using the PIAAC explorer. A model with age*gender interaction and $K = 5$ turned out to capture the upper-level heterogeneity well (Table 1 right).

3 Uncertainty of posterior intercepts

3.1 Analytic approximation

We initially approach the problem analytically, considering the weights w_{ik} in (2) as constants (which, strictly, they are not, since they depend on the

parameters estimated in the last M-step). It follows then from (2) that

$$\mathrm{Var}(\tilde{z}_i) = \sum_{k=1}^{K} w_{ik}^2 \mathrm{Var}(\hat{z}_k) + \sum_{j \neq k} w_{ij} w_{ik} \mathrm{Cov}(\hat{z}_j, \hat{z}_k) \tag{3}$$

where the variances and covariances are available from the fitted model according to standard GLM theory. Clearly, the covariance terms cannot be naively omitted since the positions of the $\hat{z}_k's$ are strongly correlated. However, as $\sum_{k=1}^{K} w_{ik} = 1$ for all i, it is clear that $0 \leq w_{ij} w_{ik} \leq 1/4$ for all pairs $j \neq k$. In addition, it is often (but not always) the case that after EM convergence observations are classified to one of the components with probability equal or close to 1, in which case $w_{ij} w_{ik} \approx 0$. [To exemplify this point, Table 1 gives an excerpt of the matrix $W = (w_{ik})$ for the gender*age model with $K = 5$ components.] In either case, it is clear that the product $w_{ij} w_{ik}$ will be very small for all (or almost all) i, j, k with $j \neq k$, so that the 'naive' approximation

$$\mathrm{Var}(\tilde{z}_i) \approx \sum_{k=1}^{K} w_{ik}^2 \mathrm{Var}(\hat{z}_k) \tag{4}$$

will usually be a good one. Confidence intervals for the posterior intercepts are then obtained from either (3) or (4) via $\tilde{z}_i \pm q\sqrt{\mathrm{Var}(\tilde{z}_i)}$ where q is an appropriate quantile for which we use the 97.5% Gaussian quantile, 1.96.

3.2 NPML–Bootstrap

In order to assess the variability in a potentially more realistic way, we also developed a bootstrap routine which proceeds in two layers. Specifically, for $i = 1, \ldots, n$,

(i) from the set of mass points $\hat{z}_1, \ldots, \hat{z}_k$ draw a masspoint $\check{z}_i$ with probability w_{ik};

(ii) generate new $\check{y}_{ij} \sim \mathrm{Bin}(n_{ij}, \check{\mu}_{ij})/n_{ij}$, where $\check{\mu}_{ij}$ is defined in the natural way via (1), using $\hat{\beta}$ and $\check{z}_i$.

Having $\check{y}_{ij}$, we refit the model, yielding a new set of n posterior intercepts. Repeating these steps M times we have a bootstrap sample of estimates for posterior intercepts. Therefore, by taking the standard deviation of these we have an estimate for their variability.

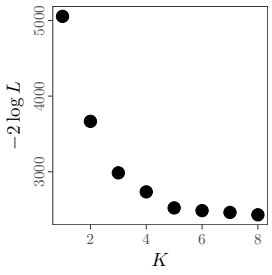
3.3 Results

Figure 1 (left) gives the $\tilde{z}_i$ along with 'full' confidence intervals (3) and bootstrapped confidence intervals (using $M = 9999$). The analytic and

simulation–based intervals are very similar, with minor differences only recognizable for a small subset of countries. The naive intervals (4) cannot be visually distinguished from the full intervals. Therefore, we provide in Figure 1 (right) the ratio of the widths of the naive and full intervals, as well as the bootstrapped and full intervals. All ratios are very close to 1, with slightly larger deviations for the bootstrapped intervals. We also see that five groups of countries can be robustly distinguished (since the corresponding intervals do not overlap), with Japan being the sole best–performing country.

TABLE 1. Left: Excerpt of matrix $W = (w_{ik})$ (4.d.p.) for age*gender model with $K = 5$; right: $-2 \log L$ as a function of K.

k	1	2	3	4	5
Australia	0	0	0	1	0
Austria	0	0.0023	0.9977	0	0
Canada	0	0	1	0	0
...					
Japan	0	0	0	0	1
Netherlands	0	0	0	1	0
...					
$\hat{z}_k$	-0.490	0.011	0.273	0.622	1.307



4 Uncertainty of posterior probabilities

4.1 Sampling from posterior likelihood

A potential issue with the methodology discussed so far is that by plugging the ML parameter estimates $\hat{\theta}$ into the expression for w_{ik}, the uncertainty in those estimates is ignored. Hence, the 'certainty' of mass point allocation when taking the w_{ik} at face value can be considered as overstated. To address this problem, Aitkin et al. (2014) suggested the following procedure based on the concept of posterior likelihood (Aitkin, 2010):

a) Assuming flat priors for θ, the posterior distribution $p(\theta|y_1, \ldots, y_n)$ is proportional to the likelihood, $L(\theta)$. Hence, one can take M random draws $\hat{\theta}^{[m]}$, $m = 1, \ldots, M$, from $L(\theta)$.

b) Compute $w_{ik}^{[m]} = P(k|\hat{\theta}^{[m]}, y_i)$, $m = 1, \ldots, M$.

For our purposes, one would then proceed further,

c) Apply step (i) in the algorithm in subsection 3.2 using $w_{ik}^{[m]}$ instead of w_{ik} in the m–th bootstrap repetition.

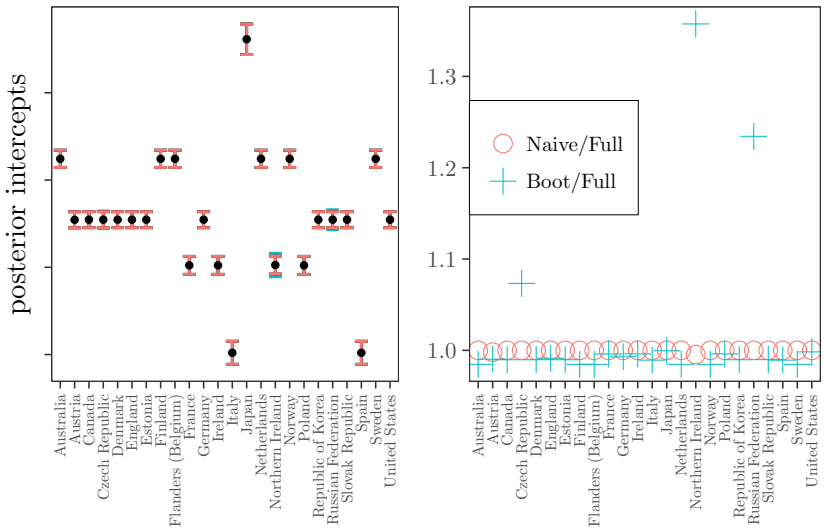


FIGURE 1. Left: Posterior intercepts [black dots] with analytic ('full') [inner interval; red in the online version] and bootstrapped intervals [outer; turquoise]; right: relative width of intervals.

However, the implementation is non–straightforward and will require a computationally expensive MCMC analysis, involving Gibbs samplers for each k, alternate draws between different types of model parameters, and ad–hoc solutions to the starting value and the label switching problems. Also, having already carried out a full EM procedure, a full–blown MCMC analysis only for the sake of analyzing the posterior intercepts feels rather out of scale. Hence, a simpler alternative idea is considered below.

4.2 Sensitivity assessment via EM process trail

As stated above, an EM algorithm has already been executed. As such, in this process, a series of 'draws' from the full likelihood $L(\theta)$ has been obtained. Assume that, in EM iteration $s = 1, \ldots, S$, we have obtained parameter estimates $\hat{\theta}^{[s]}$ with associated weight matrices $W^{[s]}$ and likelihoods $L^{[s]} \equiv L(\hat{\theta}^{[s]})$. Hence, we possess S draws from $L(\theta)$, including the final iteration, which corresponds to the MLE $\hat{\theta}^{[s]} \equiv \hat{\theta}$. While it is clear that these S draws in no way represent the correct shape of $L(\theta)$, the matrices $W^{[s]}$ can still be used to assess the *sensitivity* of the NPML–Bootstrap to imprecision in the $w_{ik}{}'s$, especially as some of the estimates along the EM process trail correspond to really 'bad' likelihoods (that is, estimates which have a likelihood of effectively 0 to be sampled in part a) of the Aitkin routine).

4.3 Results

For the data and model at hand, the number of required EM iterations turned out to be $S = 6$, and Figure 2 (left) shows $L(\hat{\theta}^{[s]})$ as a function of s. Figure 2 (right) shows the interval length of the NPML–bootstrapped confidence intervals when using $w_{ik}^{[s]}$, $s = 1, \ldots, 5$ relative to that using $w_{ik} \equiv w_{ik}^{[6]}$. It is clear that for all s corresponding to appreciable likelihoods the difference is less than 10%, and even for posterior weights corresponding to really poor likelihoods the increase is generally not more than 50%, indicating robust upper bounds for the uncertainty in this process.

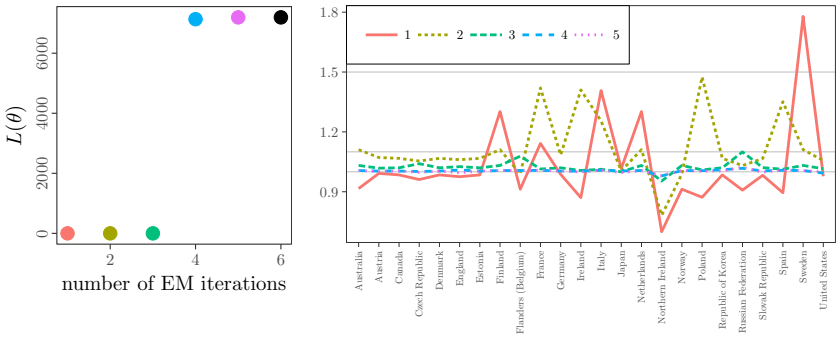


FIGURE 2. Left: Maximum Likelihood $L(\hat{\theta}^{[s]})$ versus EM iteration s; right: interval length using $w_{ik}^{[s]}$, $s = 1, \ldots, 5$ relative to using $w_{ik} \equiv w_{ik}^{[6]}$. The value s is given in the legend. [In the printed conference proceedings, there was a problem with the labelling of this graph. This has been corrected in this version.]

The present paper has demonstrated the utility of bootstrap methods to characterize the sometimes substantial uncertainty in cluster–level estimates which commonly arises in league-table comparisons. While no claim is made that the relative magnitudes of the different intervals will in general behave in the manner of this particular case study, the tools proposed to arrive at this judgement are applicable for arbitrary two–level problems.

References

Aitkin, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics*, **55**, 117 – 128.

Aitkin, M. (2010). *Statistical Inference — an Integrated Bayesian/Likelihood Approach.* Chapman & Hall/CRC:Boca Raton.

Aitkin, M., Duy, V. and Francis, B. (2014). Statistical Modelling of the group structure for social networks. *Social Networks*, **38**, 74–87.

Eurostat (2016). Smarter, greener, more inclusive? Indicators to support the Europe 2020 strategy. *Publications Office of the European Union*, European Commission, Luxembourg.