Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. http://www.jstatsoft.org/

Rgbp: Bayesian Hierarchical Modeling and Frequentist Method Check

Joseph Kelly
Harvard University

Carl Morris
Harvard University

Hyungsuk Tak
Harvard University

Abstract

Bayesian-frequentist reconciliation via Bayesian hierarchical modeling for Gaussian, Binomial, and Poisson data and frequentist method check for good coverage probability.

Keywords: hierarchical model, multi-level model, random effects mixed model, method check, coverage probability, normal, binomial, poisson, shrinkage, R.

1. Introduction

Rgbp uses Bayesian machinery to estimate a two-level model (a random-effects mixed model) and allows for a check of its frequentist properties via a repeated sampling procedure (which we call a "method check"). It is found that even in small samples our procedure yields good frequency properties. Also, this package will be useful for Bayesians who want to see a non-informative reference point before and after constructing their full-Bayesian hierarchical model. For frequentists, it will provide confidence intervals of a random-effect mixed model with good repeated sampling properties.

2. Three Feasible Types of Data

This package is intended to fit a multi-level model on the group-level (or unit-level) data in which each group-level (or unit-level) observation is believed to have the Normal, Poisson, or Binomial distribution. In this section, we will introduce three specific types of feasible datasets.

2.1. Normal: 8 School Data

Education Testing Service conducted randomized experiments in eight separate schools and obtained this dataset. It contains the coaching effects on SAT scores ($y_j, j = 1, \ldots, 8$) and standard errors ($se_j, j = 1, \ldots, 8$) of eight schools obtained after an analysis of covariance adjustment (Rubin, 1981).

```
R> y  <- c(28,  8, -3,  7, -1,  1, 18, 12)
R> se <- c(15, 10, 16, 11,  9, 11, 10, 18)
```

In the original paper, each school's coaching effect has approximately Normal sampling distribution with known sampling variance, *i.e.*, standard error of each school is assumed to be known or to be accurately estimated. So, it is reasonable to think that each (group-level) coaching effect is distributed as independent Normal distribution given the unknown mean μ_j and known standard error: $y_j|\mu_j \overset{ind}{\sim} \text{Normal}(\mu_j, se_j^2),\ j = 1, \ldots, 8$. **Rgbp** includes this dataset and can be called by typing '`R> data(schools)`' on R.

2.2. Poisson: 31 Hospital Data

This dataset is about the medical profiling evaluations for Coronary Artery Bypass Graft (CABG) surgeries of 31 New York hospitals conducted in 1992 (Morris and Lysy, 2012). It comprises of the number of deaths within a month of CABG surgeries in each hospital ($z_j,\ j = 1, \ldots, 31$) and total number of patients receiving CABG surgeries (case load) in each hospital ($n_j,\ j = 1, \ldots, 31$). The below code is an example of input based on the last ten hospital data.

```
R> z <- c( 14,   9,  15,  13,  35,  26,  25,  20,   35,   27)
R> n <- c(593, 602, 629, 636, 729, 849, 914, 940, 1193, 1340)
```

Considering the type of data, it makes sense to assume the number of deaths in each hospital has independent Poisson distribution given the unknown parameter λ_j: $z_j|\lambda_j \overset{ind}{\sim} \text{Poisson}(n_j\lambda_j)$, $j = 1, \ldots, 31$, where n_j can be interpreted as an exposure (not necessarily an interger). This dataset is also included in the package and can be called by '`R> data(hospital)`' on R.

2.3. Binomial: 18 Baseball Data

This dataset contains information about batting averages of 18 major league players through their first 45 official at-bats of the 1970 season (Efron and Morris, 1975). Also, it has two covariates, League and Position, showing in which league and in which position each player was playing. In this paper, we will use Position for a tutorial purpose. For convenience, we transform this variable into a binary indicator, which is 1 if a player was a outfielder and 0 otherwise. The code below shows a way to make inputs. If we have more than one covariate, for example, `x1` and `x2`, then '`R> x <- cbind(x1, x2)`' will be the right input of the `gbp` function.

```
R> z <- c(18, 17, 16, 15, 14, 14, 13, 12, 11, 11, 10, 10, 10, 10, 10,  9,  8,  7)
R> n <- c(45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45)
R> x <- c( 1,  1,  1,  1,  1,  0,  0,  0,  0,  1,  0,  0,  0,  1,  1,  0,  0,  0)
```

The data indicate that independent Binomial distribution is appropriate for each player's number of hits among 45 at-bats conditioning on the unknown parameter p_j: $z_j|p_j \overset{ind}{\sim} \text{Binomial}(n_j, p_j),\ j = 1, \ldots, 18$. This dataset is also a part of the package and can be called on R by '`R> data(baseball)`'.

3. Multi-level Structure

Our multi-level model, also called a conditionally independent hierarchical model (Kass and Steffey, 1989), is a very powerful tool for exploring the hierarchical sturucture in data. For example, we can think about a district-level hierarchy (bigger population) for 8 schools, the state-level hierarchy for 31 hospitals, and the position-level hierarchy for 18 baseball players. `gbp`, one of functions in **Rgbp**, fits such a hierarchical model whose first-level hierarchy has a distribution of observed data and second-level (bigger population hierarchy) has a conjugate prior distribution on the first-level parameter. Users can determine one of three types of multi-level models, such as Normal-Normal, Poisson-Gamma, and Binomial-Beta, based on their datasets.

3.1. Normal-Normal

`gbp` can construct a two-level Normal-Normal hierarchical model on the 8 school data. For reference, σ_j^2 below is assumed to be known or to be accurately estimated, and subscript j indicates j-th school in the dataset.

$$y_j|\mu_j \overset{ind}{\sim} \text{Normal}(\mu_j, \sigma_j^2), \tag{1}$$

$$\mu_j|\beta, A \overset{ind}{\sim} \text{Normal}(\mu_{0j}, A), \tag{2}$$

where $\mu_{0j} = x_j^T\beta,\ j = 1, \ldots, 8$, x_j is j-th school's covariate vector $(m \times 1)$, and m is the number of regression coefficients. Note that if there is no covariate then m is 1 for an intercept term, making $\mu_{0j} = \mu_0 = \beta_0$ for all j. For reference, a paramter with a zero subscript, μ_{0j}, represents a mean parameter of the prior (second-level) distribution. Also, Based on this conjugate prior distribution, it is easy to derive corresponding posterior distribution.

$$\mu_j|\mathbf{y}, \beta, A \overset{ind}{\sim} \text{Normal}(\ (1 - B_j)y_j + B_j\mu_{0j},\ (1 - B_j)\sigma_j^2), \tag{3}$$

where $B_j \equiv \frac{\sigma_j^2}{\sigma_j^2 + A},\ j = 1, \ldots, 8$, is called a shrinkages.

3.2. Poisson-Gamma

`gbp` is also able to build a Poisson-Gamma multi-level model on the 31 hospital data. Note that a constant multiplied to the notation representing Gamma distribution below is a scale and a square bracket below indicates [mean, variance] of distribution. And for notational consistency, let's define $y_j \equiv \frac{z_j}{n_j}$ for all j.

$$z_j|\lambda_j \overset{ind}{\sim} \text{Poisson}(n_j\lambda_j), \tag{4}$$

$$\lambda_j|\beta, r \overset{ind}{\sim} \frac{1}{r}\text{Gamma}(\lambda_{0j}r) \sim \text{Gamma}[\lambda_{0j},\ \frac{\lambda_{0j}}{r}], \tag{5}$$

where $\log(\lambda_{0j}) = x_j'\beta$, and $j = 1, \ldots, 31$. Immediate posterior distribution of this Poisson-Gamma model is

$$\lambda_j|\mathbf{z}, \beta, r \overset{ind}{\sim} \frac{1}{r + n_j}\text{Gamma}(r\lambda_{0j} + n_j y_j) \sim \text{Gamma}[\lambda_j^*,\ \frac{\lambda_j^*}{r + n_j}], \tag{6}$$

where $\lambda_j^* \equiv (1 - B_j)y_j + B_j\lambda_{0j}$, $B_j \equiv \frac{r}{r+n_j}$, and $y_j \equiv \frac{z_j}{n_j}$, $j = 1, \ldots, 31$.

3.3. Binomial-Beta

Binomial-Beta hierarchical model is the last model that `gbp` can fit. Again, a square bracket below indicates [mean, variance] of distribution.

$$z_j|p_j \overset{ind}{\sim} \text{Binomial}(n_j,\ p_j), \tag{7}$$

$$p_j|\beta, r \overset{ind}{\sim} \text{Beta}(rp_{0j},\ r(1 - p_{0j})) \sim \text{Beta}[p_{0j},\ \frac{p_{0j}(1 - p_{0j})}{r + 1}], \tag{8}$$

where $\log(\frac{p_{0j}}{1-p_{0j}}) = x_j'\beta$ and $j = 1, \ldots, 18$. Then posterior distribution is

$$p_j|\mathbf{z}, \beta, r \overset{ind}{\sim} \text{Beta}(rp_{0j} + n_j y_j,\ r(1 - p_{0j}) + n_j(1 - y_j) \sim \text{Beta}\left[p_j^*,\ \frac{p_j^*(1 - p_j^*)}{r + n_j + 1}\right], \tag{9}$$

where $p_j^* \equiv (1 - B_j)y_j + B_j p_{0j}$, $B_j \equiv \frac{r}{r+n_j}$, and $y_j \equiv \frac{z_j}{n_j}$, $j = 1, \ldots, 18$.

3.4. Hyper-prior Distribution

Hyper-prior distribution indicates a distribution of the second-level parameters, which plays an important role in deriving a full posterior distribution of all the parameters. `gbp` sets a non-informative distribution on second-level parameters to let the data speak more about their estimation.

$$\beta \sim \text{Uniform on } \mathbf{R}^m, \quad A \text{ (or } \frac{1}{r}) \sim \text{Uniform}(0, \infty), \tag{10}$$

where m is the number of regression coefficients. For β, it is reasonable choice to take flat (non-informative) distribution because information about the location gets informative as number of groups increases. Another flat prior on the hyper-prameter A (or $1/r$) that we suggest here will guarantee a posterior propriety under the moderate conditions and will give users a good repeated sampling property.

4. Estimation

4.1. Shrinkage Estimation

Estimating shrinkage is a key part of our work because as we can see the posterior means in (3), (6), and (9) are a linear function of shrinkage and the posterior variances are also a

linear (Gaussian), quadratic (Poisson), or cubic (Binomial) function of shrinkage. Once we estimate it, we can approximate the posterior moments given only data by Adam's law, for example, $E(\mu_j|\mathbf{y}) = E(E(\mu_j|\mathbf{y}, r, A)|\mathbf{y})$ for Gaussian model and by Eve's law, for instance, $Var(\mu_j|\mathbf{y}) = E(Var(\mu_j|\mathbf{y}, r, A)|\mathbf{y}) + Var(E(\mu_j|\mathbf{y}, r, A)|\mathbf{y})$

4.2. Adjustment for Density Maximization

When it comes to estimating a shrinkage, we can notice that it is a function of the second-level variance component, *i.e.*, $B_j \equiv \frac{\sigma^2}{\sigma^2+A} = B_j(A)$ for Gaussian and $B_j \equiv \frac{r}{r+n_j} = B_j(r)$ for Poisson and Binomial models.

In this case, one way is to obtain an MLE of the variance component with its asymptotic Normal distribution and then to use Delta method for asymptotic Normal distribution of shrinkage. But is the Normal approximation a good approximation for shrinkage that takes on a value between 0 and 1?

Here the adjustment for density maximization (Morris and Tang, 2011), called ADM, comes. It assumes the Beta distribution for shrinkage, which the authors believe is a much better approximation for it, and estimates its posterior moments, *i.e.*, $E(B_j|\text{data})$ and $Var(B_j|\text{data})$.

4.3. Approximation to Posterior Distribution via Matching Moments

After estimating two posterior moments, `gbp` reasonably approximates posterior distribution given only data, *i.e.*, $(\mu_j|\text{data})$, $(\lambda_j|\text{data})$, or $(p_j|\text{data})$, by matching two moments with its parameters. The reason we said 'reasonablely approximates' is that it is hard to find a closed form of posterior distribution given only data unless it is a Gaussian model. So, we assumed $(p_j|\text{data})$ had also Beta(a_1, a_0) distribution and matched two estimated moments, $E(p_j|\text{data})$ and $Var(p_j|\text{data})$, with two parameters, a_1 and a_0, of this Beta distribution.

5. Method Check

Like the two sides of the same coin, checking a statistical model always comes with fitting a model. If a fitted model cannot pass a checking process, we usually go back to the fitting process and come back to checking process iteratively. In this sense, checking a fitted model is an interactive procedure for the model justification.

There are two kinds of model justification process; one is a model check and the other is a method check. The model check is for the justification of a hierarchical modeling on a specific dataset. One possible question is, "Can this dataset benefit from such a modeling?" Christiansen and Morris (1996) answered this question by using a mixture model, $(z_j|\beta, r) \sim$ Negative-Binomial, for Poisson data to justify the second-level hierarchy. They found that their data had more variation than expected of the first-level Poisson distribution and Poisson hierarchical model could successfully account for such additional variation.

Once we are sure that the hierarchical modeling can be appropriate for our data, the following quetsion will be about the validity of interval estimates. "Does the confidence interval obtained

via this Bayesian model-fitting process achieve pre-specified $100(1-\alpha)\%$ confidence level for any true parameter values?" Our answer is "yes" and **Rgbp** has a function to assure this point. From now on, all the explanations will be based on the Binomial model.

5.1. Pseudo-data Generation Process

Figure 1 will be helpful to understand this process. As we can see in (8), the distribution of each true parameter ($p_j,\ j = 1, \ldots, 18$) depends on two hyper-parameters, r and $\beta_{(m\times 1)}$, where m is the number of regression coefficients. Once we fix these hyper-parameters at specific values, we can generate true parameters. Suppose we sampled 500 $\mathbf{p}_{(18\times 1)}$'s, *i.e.*, $\{\mathbf{p}^{(i)}_{(18\times 1)},\ i = 1, \ldots, 500\}$. Then, we can also generate $\{\mathbf{z}^{(i)}_{(18\times 1)},\ i = 1, \ldots, 500\}$ given each p_{ij}, where i indicates i-th pseudo-dataset and j does j-th player. Next, `coverage` fits the Binomial hierarchical model 500 times on $\{(\mathbf{z}^{(i)}_{(18\times 1)}, \mathbf{n}_{(18\times 1)}),\ i = 1, \ldots, 500\}$ to obtain 500*18 interval estimates. Taking the first player as an example, we have 500 pairs of (z_{i1}, p_{i1}), $i = 1, \ldots, 500$, and 500 interval estimates.

$$r, \boldsymbol{\beta}_{(m\times 1)} \rightarrow \begin{array}{ccc} \boldsymbol{p}^{(1)}_{(18\times 1)} & \rightarrow & \boldsymbol{z}^{(1)}_{(18\times 1)} \\ \vdots & \rightarrow & \vdots \\ \boldsymbol{p}^{(500)}_{(18\times 1)} & \rightarrow & \boldsymbol{z}^{(500)}_{(18\times 1)} \end{array}$$

Figure 1: Pseudo-data generating process

5.2. Unbiased Coverage Probability

Based on the generated pseudo-datasets, let's define an indicator variable, I_{ij}, which is 1 if j-th player's interval estimate from i-th pseudo-dataset includes p_{ij} and 0 otherwise. One way to estimate the coverage probability is taking average over these indicator variables. We call it an unbiased coverage probability estimate. For example, $\bar{I}_1 = \sum_{i=1}^{500} I_{i1}/500$ is the estimated coverage probability for the first player.

5.3. Rao-Blackwellized Coverage Probability

Based on the definition of Rao-Blackwellization, we define $E(I_{ij}|z_{ij}, r, \beta)$, where r and β were fixed at first (see 5.1) and z_{ij} is a sufficient statistic. This expectation is the same as $\Pr(\hat{p}_{ij,low} \leq p_{ij} \leq \hat{p}_{ij,upp}|z_{ij}, r, \beta)$, where $(\hat{p}_{ij,low},\ \hat{p}_{ij,upp})$ is j-th player's interval estimates (numeric values) on the i-th dataset. We can calculated this probability exactly because we know the distribution of $(p_{ij}|z_{ij}, r, \beta)$ in (9). Then, we can estimate the first player's coverage probability, for instance, by $\sum_{i=1}^{500} E(I_{i1}|z_{i1}, r, \beta)/500$.

6. Example

6.1. Known Prior Mean

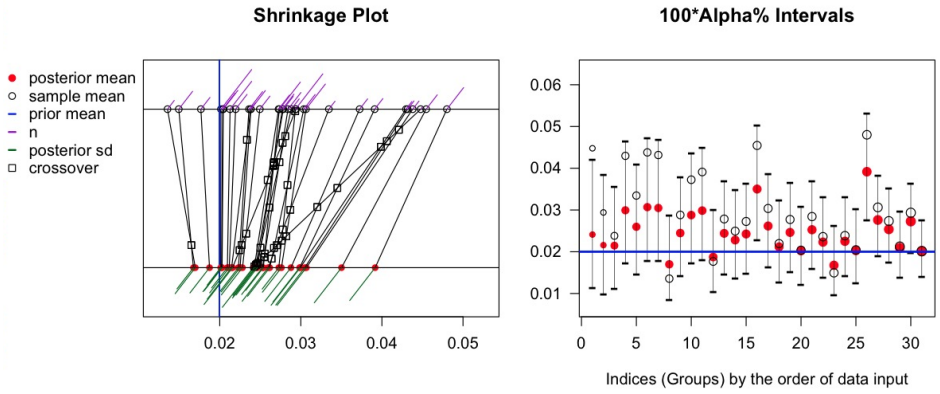

Figure 2: Coverage and Interval Plots of `R> plot(p)`

6.2. Unknown Prior Mean and No Covariate

Using 8 school data with GR

6.3. Unknown Prior Mean and One Covariate

Using 18 baseball data with BR

7. Discussion

8. Acknowledgments

9. Reference

1. Christiansen, C. and Morris, C. (1996). "Fitting and Checking a Two-Level Poisson Model: Modeling Patient Mortality Rates in Heart Transplant Patients," in *Bayesian Biostatistics*, eds. D. Berry and D. Stangl, New York: Marcel Dekker, pp. 467-561.

2. Efron, B. and Morris, C. (1975). "Data Analysis Using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association.* **70**. 311-319.

3. Kass, R. and Steffey, D. (1989). "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)." *Journal of the American Statistical Association.* **84**. 717-726.

4. Morris, C. and Tang, R. (2011). "Estimating Random Effects via Adjustment for Density Maximization." *Statistical Science.* **26**. 271-287.

5. Morris, C. and Lysy, M. (2012). "Shrinkage Estimation in Multilevel Normal Models." *Statistical Science*. **27**. 115-134.

6. Rubin, D. B. (1981). "Estimation in parallel randomized experiments." *Journal of Educational Statistics*. **6**. 377-401.

Affiliation:

Joseph Kelly
Department of Statistics
Harvard University
1 Oxford Street, Cambridge, MA
E-mail: kelly2@fas.harvard.edu
URL: http://www.people.fas.harvard.edu/~kelly2/

Carl Morris
Department of Statistics
Harvard University
1 Oxford Street, Cambridge, MA
E-mail: morris@fas.harvard.edu

Hyungsuk Tak
Department of Statistics
Harvard University
1 Oxford Street, Cambridge, MA
E-mail: hyungsuk.tak@gmail.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd