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

Rgbp: An R Package for Hierarchical Modeling and Method Checking

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, multilevel model, random effects mixed model, method check, coverage probability, normal, binomial, poisson, shrinkage, R.

1. Introduction

Rgbp is an R package for estimating and validating a two-level model (a random-effects mixed model). The estimation procedure utilizes Bayesian machinery and the validation involves checking frequency properties of the procedure via repeated sampling (which we call "method checking"). It is found that even in small samples our procedure yields good frequency properties in comparison to other methods such as Maximum Likelihood Estimation (MLE). This package will be useful for frequentists and Bayesians alike. Bayesians are able to use the package to see a non-informative reference point before and after constructing their full-Bayesian hierarchical model and frequentists, now have a procedure that will provide confidence intervals of a random-effect mixed model with good repeated sampling properties.

2. Three Feasible Types of Data

Rgbp is intended to fit a model where two levels of structure are assumed (unit and group levels). The model can be characterized by the distributional family assumed for the group level data and in the case of **Rgbp** this will be either the Normal, Poisson, or Binomial distribution. In this section, we will introduce three specific types of feasible data sets that illustrate this fact.

2.1. Normal: 8 Schools

The Education Testing Service (ETS) conducted randomized experiments in eight separate schools (group) to test whether students (unit) SAT scores are effected by coaching. The dataset contains the estimated coaching effects on SAT scores ($y_j, j = 1, \ldots, 8$) and standard errors ($se_j, j = 1, \ldots, 8$) of the eight schools Rubin (1981).

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

Due to the nature of the test each school's coaching effect has an 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 school-level coaching effect is distributed as an 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 data set and can be called by the command '`data(schools)`' on R.

2.2. Poisson: 31 Hospitals

The following data are from 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 the total number of patients receiving CABG surgeries (case load) in each hospital ($n_j,\ j = 1, \ldots, 31$). The following 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 is reasonable to assume the number of deaths in each hospital follow independent Poisson distributions given an unknown true rate parameter λ_j: $z_j|\lambda_j \overset{ind}{\sim} \text{Poisson}(n_j\lambda_j),\ j = 1, \ldots, 31$, where, for each j, n_j can be interpreted as an exposure (not necessarily an integer) and λ_j as the true death rate per exposure. This data set is also included in the package and can be called by '`data(hospital)`' on R.

2.3. Binomial: 18 Baseball Hitters

The following dataset contains information on the batting averages of 18 major league baseball players through their first 45 official at-bats of the 1970 season Efron and Morris (1975). In addition one covariate relating to the position each player was playing was observed and for illustrative purposes, we transform this variable into a binary indicator (1 if a player was a outfielder and 0 otherwise).

```
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 true batting average p_j:

$z_j|p_j \overset{ind}{\sim} \text{Binomial}(n_j, p_j),\ j = 1, \ldots, 18$. This data set is also a part of the package and can be called on R by '`data(baseball)`'.

3. Multilevel Structure

A two-level or multilevel model, also called a conditionally independent hierarchical model Kass and Steffey (1989), is a very powerful tool for exploring the hierarchical structure in data. For example, in the 8 schools dataset, we can imagine that there exists 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 the 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. The `gbp` function allows users to choose one of three types of multilevel models, such as Normal-Normal, Poisson-Gamma, and Binomial-Beta, based on their data sets.

3.1. Normal-Normal

The following is a general Normal-Normal hierarchical model. For reference, $V_j\ (\equiv \sigma^2/n_j)$ below is assumed to be known or to be accurately estimated, and subscript j indicates j-th group in the data set.

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

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

where $j = 1, \ldots k$, $\mu_{0j} = \mathbf{x}_j^T\beta$, x_j is j-th group's covariate vector $(m \times 1)$, m is the number of regression coefficients and β and A are unknown. Note that if there is no covariate then $x_j = 1$ for an intercept term $(m = 1)$ and so $\mu_{0j} = \mu_0 = \beta_0$ for all j, resulting in an exchangeable prior distribution. For reference, a parameter with a zero subscript, such as μ_{0j}, represents a mean parameter of the prior (second-level) distribution, *i.e.*, a prior mean. Based on this conjugate prior distribution it is easy to derive the corresponding posterior distribution

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

where $B_j \equiv V_j/(V_j + A),\ j = 1, \ldots, k$, are called shrinkages.

3.2. Poisson-Gamma

`gbp` is also able to build a Poisson-Gamma multilevel model. Note that a constant, $1/r$, multiplied to the Gamma distribution below is a scale and a square bracket below indicates [mean, variance] of distribution. And for notational consistency, let's define $y_j \equiv z_j/n_j$ for all j.

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

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

where $\log(\lambda_{0j}) = x_j'\beta,\ j = 1, \ldots, k$, with two unknown hyper-parameters, r and β. The posterior distribution of this Poisson-Gamma model is

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

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

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.

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

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

where $\log(\frac{p_{0j}}{1-p_{0j}}) = x_j'\beta,\ j = 1, \ldots, k$. Note that r and β are two unknown hyper-parameters. Then, corresponding 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 $p_j^* \equiv (1 - B_j)y_j + B_j p_{0j}$, $B_j \equiv r/(r + n_j)$, and $y_j \equiv z_j/n_j$, $j = 1, \ldots, k$.

3.4. Hyper-prior Distribution

In our examples the hyper-prior distribution is an assumed distribution on the second-level parameters and with the goal of objectivity in mind `gbp` assumes non-informative hyper-prior distributions.

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

where m is the number of regression coefficients. As for β, it is a reasonable choice to assume a flat (non-informative) distribution because information about the location gets plentiful as the number of groups increases. The next flat prior distribution of the second-level variance component A (or $1/r$) was chosen for its good repeated sampling properties and for posterior propriety under moderate conditions.

4. Estimation

4.1. Shrinkage Estimation

Estimating the shrinkage factors $(B_1, \ldots B_k)$ is the key estimation problem with the hierarchical models `gbp` assumes. As we can see in 3, 6, and 9, the posterior means are a linear function of the shrinkage factors and the posterior variances are also linear (Gaussian), quadratic

(Poisson), or cubic (Binomial) functions. A natural method then to estimate $E(\mu_j|\mathbf{y})$ and $Var(\mu_j|\mathbf{y})$ is to first estimate the shrinkage factors.

4.2. Adjustment for Density Maximization

It is noted that the shrinkage factors $(B_1, \ldots B_k)$ are a function of the second-level variance component, *i.e.*, $B_j \equiv V_j/(V_j + A) = B_j(A)$ for Gaussian and $B_j \equiv r/(r + n_j) = B_j(r)$ for Poisson and Binomial models. Current popular methods of estimation via maximization (MLE or MAP) of these unknown shrinkage factors can result in estimates lying on the boundary of the parameter space. In the Normal-Normal model this situation typically arises when the variation within schools is greater than the variation between resulting in a parameter estimate of $\hat{A} = 0$.

To continue with a maximization based estimation procedure but to steer clear of aforementioned issues we make us of adjustment for density maximization (ADM) Morris and Tang (2011). In general ADM is a procedure to obtain estimates of posterior moments via maximization if the underlying posterior can be approximated by a Pearson family. For our purposes we can approximate the posterior distribution of a shrinkage factor with the Beta distribution allowing us to finally obtain estimates the posterior moments, *i.e.*, $E(B_j|\text{data})$ and $Var(B_j|\text{data})$, without any trouble that MLE can cause. Please refer to Morris and Tang (2011) for additional advantages of ADM.

Once we estimate these two moments of shrinkage, we can also estimate the posterior moments given only the data. Taking the Normal-Normal model as an example we note the following identities

$$E(\mu_j|\mathbf{y}) = E(E(\mu_j|\mathbf{y}, r, A)|\mathbf{y}) \tag{11}$$

$$Var(\mu_j|\mathbf{y}) = E(Var(\mu_j|\mathbf{y}, r, A)|\mathbf{y}) + Var(E(\mu_j|\mathbf{y}, r, A)|\mathbf{y}). \tag{12}$$

Note that both $E(\mu_j|\mathbf{y}, r, A)$ and $Var(\mu_j|\mathbf{y}, r, A)$ are functions of the shrinkage factors (3) and since we have already can estimate $E(B_j|\mathbf{y})$ and $Var(B_j|\mathbf{y})$ via ADM, we can finally estimate $E(\mu_j|\mathbf{y})$ and $Var(\mu_j|\mathbf{y})$.

4.3. Approximation to Posterior Distribution via Matching Moments

After estimating the two posterior moments, $E(p_j|\mathbf{z})$ and $Var(p_j|\mathbf{z})$, `gbp` reasonably approximates a posterior distribution of the mean effects given the data by assuming a reasonable distribution and matching moments. The Binomial-Beta model we approximate $p_j|\mathbf{z}$ with a Beta distribution and for the Poisson-Gamma model we approximate $\lambda_j|\mathbf{z}$ with a Gamma distribution. Finally for the Normal-Normal model we actually estimate the first three moments and approximate $\mu_j|\mathbf{y}$ with a skewed-t distribution.

5. Method Checking

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 validation or checking process, we usually go back and forth from estimation and checking steps iteratively. In this sense, checking a fitted model is an interactive procedure for the model justification.

There are two kinds of model justification processes; one is a model check and the other is a method check. The model check is for the justification of the model 3 on a specific data set. One possible question is, "Can this data set benefit from such a multilevel modeling?" Christiansen and Morris (1997) answered this question by using a mixture model, $z_j|\beta, r \sim$ Negative-Binomial, on 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 question will be about the validity of interval estimates, the final product of this multilevel modeling. "Does the 95% (can be specified differently) confidence interval obtained via this Bayesian model-fitting process achieve 95% confidence level for any true parameter values?" Our answer is "yes" and **Rgbp** has a function to check this point. From now on, all the explanations will be based on the Binomial model because the idea is the same and can be easily applied to the other two models.

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 batting average ($p_j,\ j = 1, \ldots, 18$) depends on two hyper-parameters, r and $\beta_{(m\times 1)}$, where m is the number of regression coefficients. So, once we fix these hyper-parameters at specific values, we can generate true batting averages. Suppose we sampled 500 $\mathbf{p}_{(18\times 1)}$'s, *i.e.*, $\{\mathbf{p}^{(i)}_{(18\times 1)},\ i = 1, \ldots, 500\}$ from the prior distribution in (8), where r and β are given. 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-data set 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. After this generating process, as for the first player, we will have 500 pairs of (z_{i1}, p_{i1}) and 500 interval estimates, $\{(\hat{p}_{i1,low},\ \hat{p}_{i1,upp}),\ i = 1, \ldots, 500\}$.

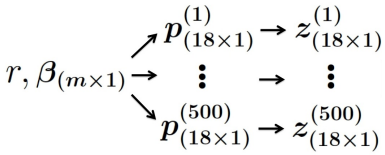

Figure 1: Pseudo-data generating process

5.2. Simple Unbiased Estimates of Coverage Probabilities

Based on the generated pseudo-data sets, let's define an indicator variable, I_{ij}, which is 1 if j-th player's interval estimate from i-th pseudo-data set includes p_{ij} and 0 otherwise. One way to estimate the coverage probability is taking average over these indicator variables. We call it a simple 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 Unbiased estimates of Coverage Probabilities

Rao-Blackwellization improves accuracy of an unbiased estimator by taking conditional expectation given a sufficient statistic. Based on this idea, we define $E(I_{ij}|z_{ij}, r, \beta)$, where r and β were given 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-Ch player's interval estimates on the i-th data set. We can calculate this probability because we know the distribution of $p_{ij}|z_{ij}, r, \beta$ in (9). Note that conditioning on z_{ij} is equivalent to conditioning on $\mathbf{z}$ as long as r and β are known. Then, we can estimate the first player's coverage probability by $\sum_{i=1}^{500} E(I_{i1}|z_{i1}, r, \beta)/500$, which is also unbiased but more accurate than the previous simple estimator.

6. Examples

6.1. 31 Hospitals: Known Second-level Mean

Suppose you live in the New York state (NY) and have been suffering from severe coronary heart disease (hopefully not). If you are supposed to receive the coronary artery bypass graft (CABG) surgery soon, you might want to find the most famous hospital for dealing with such a surgery. On top of that, if you can figure out each hospital's ability to handle this surgery, it will be useful for your decision in choosing a hospital.

For this purpose, you gathered data of 31 hospitals in NY composed of the number of deaths within a month of CABG surgeries and total number of patients receiving CABG surgeries in each hospital. In addition, while one looks for such information, suppose one knows that the state-level death rate per exposure of this surgery is 0.020.

The multilevel modeling that assumes a bigger population-level hierarchy will be insightful in this problem. Here, we presume a state-level (NY) hierarchy governing the true death rates of CABG surgery of all the hospitals in NY. This perspective is to view the true death rates of those 31 hospitals as sampled from the state-level population distribution whose mean is 0.020. For reference, a model check can be useful for evaluating the validity of such a view point (Christiansen and Morris, 1996).

Assuming an additional hierarchy is reasonable, a model-fitting process begins. Since the true death rate per exposure after CABG surgery might be small and the caseload (n_j) is much bigger than the number of deaths (z_j) in each hospital, the Poisson distribution would be the first choice to describe the uncertainty in our data. Next, `gbp` will help us fit the Poisson multilevel model with the Gamma conjugate prior distribution on the true death rate λ_j whose mean is 0.020 ($\lambda_0 = 0.020$) as described in 3.2 . For reference, the number of regression coefficients (m) is 0 because we do not need to estimate the prior mean via log-linear regression (see section 3.2).

```
R> p <- gbp(z, n, mean.PriorDist = 0.02, model = "pr")
R> p

Summary for each unit (sorted by n):

         obs.mean    n prior.mean shrinkage low.intv post.mean upp.intv post.sd
```

1	0.045	67	0.02	0.834	0.011	0.024	0.042	0.008
2	0.029	68	0.02	0.832	0.010	0.022	0.038	0.007
3	0.024	210	0.02	0.616	0.011	0.021	0.035	0.006
4	0.043	256	0.02	0.568	0.017	0.030	0.046	0.008
5	0.033	269	0.02	0.556	0.015	0.026	0.041	0.007
6	0.044	274	0.02	0.551	0.018	0.031	0.047	0.008
7	0.043	278	0.02	0.548	0.018	0.030	0.047	0.007
8	0.014	295	0.02	0.533	0.008	0.017	0.029	0.005
9	0.029	347	0.02	0.492	0.014	0.024	0.038	0.006
10	0.037	349	0.02	0.491	0.017	0.029	0.043	0.007
11	0.039	358	0.02	0.484	0.018	0.030	0.045	0.007
12	0.018	396	0.02	0.459	0.010	0.019	0.030	0.005
13	0.028	431	0.02	0.438	0.015	0.024	0.037	0.006
14	0.025	441	0.02	0.433	0.013	0.023	0.035	0.005
15	0.027	477	0.02	0.414	0.015	0.024	0.036	0.006
16	0.045	484	0.02	0.410	0.023	0.035	0.050	0.007
17	0.030	494	0.02	0.405	0.016	0.026	0.039	0.006
18	0.022	501	0.02	0.402	0.012	0.021	0.032	0.005
19	0.028	505	0.02	0.400	0.015	0.025	0.036	0.005
20	0.020	540	0.02	0.384	0.012	0.020	0.031	0.005
21	0.028	563	0.02	0.374	0.016	0.025	0.037	0.005
22	0.024	593	0.02	0.362	0.014	0.022	0.033	0.005
23	0.015	602	0.02	0.358	0.010	0.017	0.026	0.004
24	0.024	629	0.02	0.348	0.014	0.023	0.033	0.005
25	0.020	636	0.02	0.346	0.012	0.020	0.030	0.005
26	0.048	729	0.02	0.316	0.027	0.039	0.053	0.007
27	0.031	849	0.02	0.284	0.019	0.028	0.038	0.005
28	0.027	914	0.02	0.269	0.017	0.025	0.035	0.005
29	0.021	940	0.02	0.264	0.014	0.021	0.030	0.004
30	0.029	1193	0.02	0.220	0.020	0.027	0.036	0.004
31	0.020	1340	0.02	0.201	0.014	0.020	0.027	0.003
colMeans	0.029	517	0.02	0.438	0.015	0.025	0.037	0.006

For reference, we need to type '`R> print(p, sort = FALSE)`' instead of '`R> p`' in order to list hospitals by the order of data input in the above output. '`R> p`' automatically sorts the output by the increasing order of caseload (n_j).

The output contains information about sample mean (`obs.mean`), caseload (`n`), known prior mean (λ_0), shrinkage estimate ($\hat{B}_j$), lower interval, posterior mean ($E(\lambda_j|\mathbf{z})$), upper interval, and standard deviation of posterior distribution ($sd(\lambda_j|\mathbf{z})$).

As we can see in (6), the posterior mean, $(1 - B_j)y_j + B_j\lambda_0$, is a linear function of shrinkage, $B_j \equiv r/(r + n_j)$, locating between the sample mean and prior mean ($\lambda_0 = 0.02$). It makes sense because r can be interpreted as the amount of prior information and n_j as the amount of observed information. If the second level has more information than the first level, then the sample mean shrinks towards the prior mean more than 50%. This point is clear in the above output because, as caseload increases, shrinkage decreases, depending less on the prior distribution.

A function "summary" shows selective information on hospitals and more detailed estimation result as below. To be specific, it displays some hospitals (not all as above) with minimum, median, and maximum caseloads (n_j). On top of that, more specific estimation results, such as the estimation result of $\alpha \equiv \log(1/r)$, follow. Note that when we do not know the prior mean in advance unlike this hospital problem, gbp fits a linear regression for the Normal model, a log-linear regression for the Poisson model, or a logistic regression for the Binomial model and a summary of regression fit will appear.

```
R> summary(p)

Main summary:

                  obs.mean    n prior.mean shrinkage low.intv post.mean
Unit w/ min(n)       0.045   67       0.02     0.834    0.011     0.024
Unit w/ median(n)    0.045  484       0.02     0.410    0.023     0.035
Unit w/ max(n)       0.020 1340       0.02     0.201    0.014     0.020
Overall Mean         0.029  517       0.02     0.438    0.015     0.025

                  upp.intv post.sd
                     0.042   0.008
                     0.050   0.007
                     0.027   0.003
                     0.037   0.006

Second-level Variance Component Estimation Summary:
alpha = log(A) for Gaussian and log(1/r) for Binomial and Poisson data:

  post.mode.alpha post.sd.alpha
1          -5.818         0.411
```

Since estimated α is -5.818, we can easily calculate $\hat{r} = \exp(5.818) = 336$, which is a good indicator of how valuable and informative the hypothetical second-level hierarchy is. As mentioned before, r can be interpreted as the amount of prior information, although we do not know its true value. So, gbp asked data how much credit they wanted to give to this hypothetical second-level hierarchy. $\hat{r} = 336$ is the answer from the data. It means that observed sample means of hospitals whose caseloads are less than 336 will shrink toward the prior mean (0.020) more than 50%.

For example, the shrinkage estimate of the first hospital ($\hat{B}_1 = 0.834$) was calculated by 336 / (336 + 67), where 67 is its caseload (n_1), and its posterior mean is $(1 - 0.834) * 0.045 + 0.834 * 0.020 = 0.024$. As for this hospital, using more information from the prior distribution is an appropriate choice because its observed amount of information (67) is far less than the amount of state-level information (336).

We also need a graphical summary that can give us valuable insight buried in pile of numbers and a function 'plot' is exactly for this purpose.

```
R> plot(p)
```

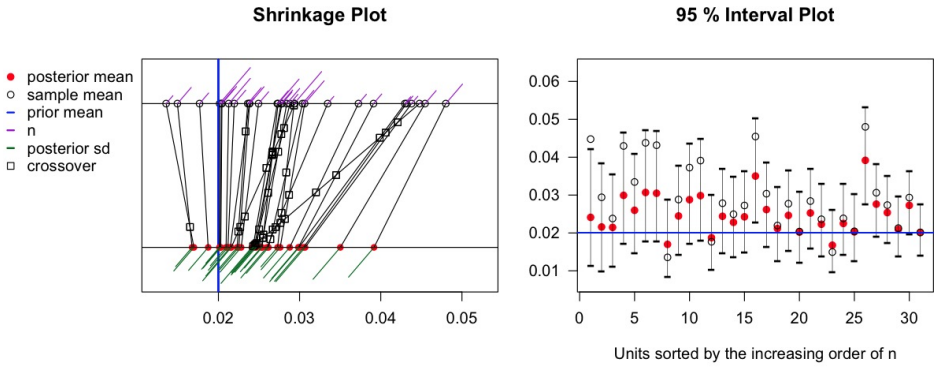

Figure 2: Shrinkage plot and 95% interval plot

The regression towards the mean (RTTM) is obvious in the left-side graph; the observed sample means, empty dots on the upper line, are shrinking towards the known second-level mean (a blue vertical line at 0.02) to the different extents. Note that some hospitals' ranks have changed by shrinking much harder towards 0.02 than others. For example, the empty square symbol at the crossing point of the two left-most lines (8th and 23rd hospitals on the list above) indicates that seemingly safest hospital among 31 hospitals in terms of the observed sample mean was not actually safer than the second safest hospital. Without this hierarchical modeling, we might have made a wrong decision in choosing a hospital.

Intuitively, the result of multilevel modeling makes more sense than that of naive sample mean. For example, suppose there are two hospitals, whose sample means (z_j/n_j) are 0 and 0.001 and caseloads (n_j) are 1 and 1000 each. Do you believe that the former hospital, whose observed death rate is 0, is better than the latter and are you going to choose the former hospital? Borrowing information from state-level hierarchy seems reasonable for the former hospital because it is hard to judge its true death rate per exposure with just one caseload. Though somewhat extreme, this is what happened to the two left-most hospitals in the first plot and this is why hierarchical modeling is a reasonable choice for this data set.

The estimated 95% intervals are displayed on the right-side plot. We can clearly see that all the posterior means (red dots) are between sample mean (empty dots) and second-level mean (a blue horizontal line). For reference, if we want to draw this plot by the order of data input, `plot(p, sort = FALSE)` is a right adjustment. Overall, as the caseload increases towards the right and as the posterior mean gets closer to 0, the length of interval gets shorter, which the formula of posterior variance in (6) implies.

This interval plot can give us one more valuable insight, which we could not have noticed. Let's look at the 8th and the 31st hospitals on the graph (or on the outcome for numeric reference). The point estimate of the true death rate per exposure of the 31st hospital is higher than the one of the 8th. But, the upper bound of interval estimate of this 31st hospital is lower than that of the 8th. This interval plot makes the 31st hospital emerge as one of your candidates. (Could you regard this 31st hospital as a possible candidate before you observe this plot?)

Also it reveals that the 23rd hospital, whose estimated true rate was the smallest, has also the smallest upper bound. If you are a risk-adverse, this hospital will attract you most strongly.

And if you already chose this 23rd hospital compared to the 8th from the shrinkage plot, your decision might become stronger at this point, excluding the 8th hospital with more certainty.

Then, how reliable are these intervals? Does our procedure to generate interval estimates have good repeated sampling property? The following method checking will answer this question. The `coverage` function below generates 10,000 pseudo-data sets regarding the estimated r (= 336.37) as a given true value. For reference, we can try any other value of r, for example $r = 200$, by replacing below code with '`R> pcv <- coverage(p, A.or.r = 200, mean.PriorDist = 0.02, nsim = 10000)`.

Also, `gbp` can give us interval estimates with different confidence level, for example 90%, and the appropriate code adjustment is `R> p <- gbp(z, n, Alpha = 0.9, mean.PriorDist = 0.02, model = "pr")`. Then, the function `coverage` will evaluate whether interval estimates achieve 90% confidence level.

```
R> pcv <- coverage(p, nsim = 10000)
```

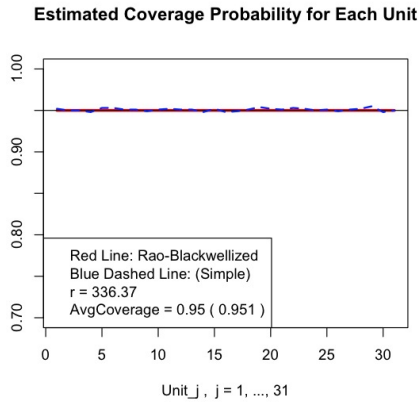

Figure 3: Coverage plot via method checking

It estimated coverage probabilities by simple unbiased estimates (a blue dashed line) and Rao-Blackwellized unbiased estimates (a red line). Both lines are indistinguishable from the horizontal line at 0.95 and the estimated overall average coverage rate is 0.950 from the Rao-Blackwellized estimate and 0.951 from the simple unbiased estimate. Considering this result, we finally conclude that the interval estimates from the suggested multilevel modeling on this particular data set has nice repeated sampling property.

Morris and Christiansen (1995) also looked into a similar ranking problem in hierarchical modeling, taking shrinkage into account.

6.2. 8 Schools: Unknown Second-level Mean and No Covariate

As stated in **??** the Education Testing Service (ETS) conducted randomized experiments in eight separate schools (group) to test whether students (unit) SAT scores are effected by coaching. The dataset contains the estimated coaching effects on SAT scores ($y_j, j = 1, \ldots, 8$) and standard errors ($se_j, j = 1, \ldots, 8$) of the eight schools Rubin (1981).

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

Due to the nature of the test each school's coaching effect has an 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. Hence, we can use `gbp` to fit this hierarchical model:

```
R> data(schools)
R> attach(schools)
R> g <- gbp(y, se, model = "gr")
R> g

Summary for each unit (sorted by n):

          obs.mean   se prior.mean shrinkage low.intv post.mean upp.intv post.sd
5            -1.00  9.0      8.168     0.408  -13.297     2.737   16.692   7.634
2             8.00 10.0      8.168     0.459   -7.255     8.077   23.361   7.810
7            18.00 10.0      8.168     0.459   -1.289    13.484   30.821   8.176
4             7.00 11.0      8.168     0.507   -8.780     7.592   23.602   8.257
6             1.00 11.0      8.168     0.507  -13.027     4.633   20.131   8.441
1            28.00 15.0      8.168     0.657   -2.315    14.979   38.763  10.560
3            -3.00 16.0      8.168     0.685  -17.130     4.650   22.477  10.096
8            12.00 18.0      8.168     0.734  -10.208     9.189   29.939  10.227
colMeans      8.75 12.5      8.168     0.552   -9.163     8.168   25.723   8.900
```

The output from `gbp` provides an easy to read summary of the results of the estimation. Here we see that the posterior mean estimates are simply a linear combination

```
R> plot(g)
```

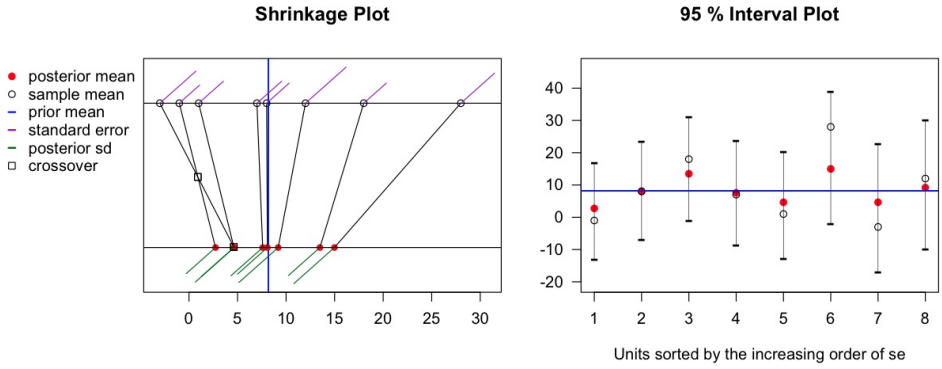

Figure 4: Shrinkage plot and 95% interval plot for 8 schools

```
R> summary(g)
```

```
Main summary:

                    obs.mean   se prior.mean shrinkage low.intv post.mean
Unit w/ min(se)        -1.00  9.0      8.168     0.408  -13.297     2.737
Unit w/ median(se)1     1.00 11.0      8.168     0.507  -13.027     4.633
Unit w/ median(se)2     7.00 11.0      8.168     0.507   -8.780     7.592
Unit w/ max(se)        12.00 18.0      8.168     0.734  -10.208     9.189
Overall Mean            8.75 12.5      8.168     0.552   -9.163     8.168

                     upp.intv post.sd
                       16.692   7.634
                       20.131   8.441
                       23.602   8.257
                       29.939  10.227
                       25.723   8.900

Second-level Variance Component Estimation Summary:
alpha = log(A) for Gaussian and log(1/r) for Binomial and Poisson data:

  post.mode.alpha post.sd.alpha
1           4.768         1.139


Regression Summary:

      estimate   se z.val p.val
beta0    8.168 5.73 1.425 0.154


R> gcv <- coverage(g, nsim = 10000)
```

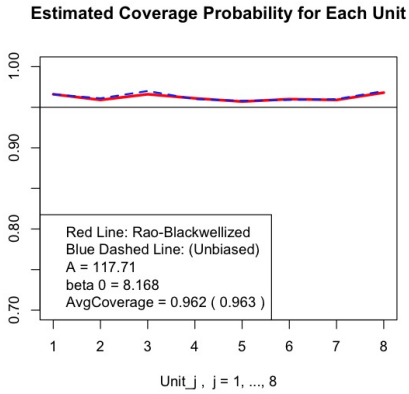

Figure 5: Coverage plot via method checking for 8 schools

6.3. 18 Baseball Players: Unknown Prior Mean and One Covariate

In this example, suppose you read the New York Times published on 26 April 1970. One article in this magazine covered 18 baseball players' batting averages through their first official 45 at-bats of the 1970 season. And then, you became interested in their batting averages over the remaining season. How could you predict? For reference, the data set `baseball` has each player's batting average over the remaining season.

First of all, the independent Binomial distribution given a true batting average of each player (see (7)) seems an appropriate probability distribution describing the uncertainty in this data set. And we know that the sample mean (MLE) under this distribution is an uniformly minimum variance unbiased estimator (UMVUE) with nice asymptotic properties. It is also the simplest way to estimate their true batting averages. But can we say this is the best way to predict their batting averages?

Efron and Morris (1975) said "no," and suggested James-Stein estimator, making each players' sample mean shrink towards an estimated mean of the second-level prior distribution. They showed that this estimator was more than three times efficient than the sample mean under squared error loss. On top of that, a few years later Morris (1983) introduced interval estimates that could successfully catch the batting averages over the remaining season, which MLE and its interval estimate could not.

Our hierarchical modeling is a sequel to these researches, which can add one more unique value. Suppose you strongly believe that the outfielder's batting averages are different from other positions' ones, not giving a credit to the assumption that all players' batting averages shrink towards the same mean. Instead, you want to assume two different second-level hierarchies for outfielders and for other positions, within each of which sharing information and regressing toward the mean (RTTM) occur. This perspective considers true batting averages of outfielders as sampled from the one and those of other positions from the other. Our multilevel modeling provides a way to incorporate such information (covariate) seamlessly into the second-level hierarchy.

```
R> b <- gbp(z, n, x, model = "br")
R> b

Summary for each unit (sorted by n):

          obs.mean  n   X1 prior.mean shrinkage low.intv post.mean upp.intv post.sd
1            0.400 45 1.00      0.310     0.715    0.248     0.335    0.429   0.046
2            0.378 45 1.00      0.310     0.715    0.244     0.329    0.420   0.045
3            0.356 45 1.00      0.310     0.715    0.240     0.323    0.411   0.044
4            0.333 45 1.00      0.310     0.715    0.236     0.316    0.403   0.043
5            0.311 45 1.00      0.310     0.715    0.230     0.310    0.396   0.042
6            0.311 45 0.00      0.233     0.715    0.179     0.256    0.341   0.041
7            0.289 45 0.00      0.233     0.715    0.175     0.249    0.331   0.040
8            0.267 45 0.00      0.233     0.715    0.171     0.243    0.323   0.039
9            0.244 45 0.00      0.233     0.715    0.166     0.237    0.315   0.038
10           0.244 45 1.00      0.310     0.715    0.210     0.291    0.379   0.043
11           0.222 45 0.00      0.233     0.715    0.161     0.230    0.308   0.038
12           0.222 45 0.00      0.233     0.715    0.161     0.230    0.308   0.038
```

```
13          0.222 45 0.00      0.233     0.715     0.161     0.230     0.308    0.038
14          0.222 45 1.00      0.310     0.715     0.202     0.285     0.375    0.044
15          0.222 45 1.00      0.310     0.715     0.202     0.285     0.375    0.044
16          0.200 45 0.00      0.233     0.715     0.155     0.224     0.302    0.038
17          0.178 45 0.00      0.233     0.715     0.148     0.218     0.297    0.038
18          0.156 45 0.00      0.233     0.715     0.140     0.211     0.292    0.039
colMeans    0.265 45 0.44      0.267     0.715     0.191     0.267     0.351    0.041
```

Our model reflects on the additional information, *i.e.*, indicator covariate (1 for outfielder and 0 for other positions), estimating two different prior means, 0.310 and 0.233. Also note that shrinkage estimates are the same for all players. It makes sense because shrinkage ($B_j \equiv r/(r+n_j)$) is determined by the relative amount of information between the first-level and the second-level hierarchies. The fact that all players have the same amount of observed information ($n_j = 45$) and $\hat{r} = \exp(4.727) = 113$ from below summary tell why the estimated shrinkage is 0.715 (=113 / (113 + 45)).

For reference, we can run the Normal hierarchical model using $y_j = z_j/n_j$ and $V_j = \bar{y}(1-\bar{y})/n$, where $\bar{y} = \sum_j z_j / \sum_j n_j$, $j = 1, \ldots, 18$. For comparison, we calculated squared error losses, $\sum_{j=1}^{18}(p_{j,remaining} - \hat{p}_j)^2$, where $p_{j,remaining}$ is each player's batting average over the remaining season and $\hat{p}_j$ is estimate from the Binomial and Normal multilevel models and from the sample mean (MLE) using the first 45 at-bats. These values were 0.042, 0.043, and 0.086 each, indicating that estimates from the Normal and Binomial models are twice better than MLE in terms of squared error loss.

One more thing to note is that the posterior standard deviation tends to be bigger among outfielders than among others. Let's see its formula in (9). We can notice that the estimated posterior mean is a critical factor that makes posterior variances different from each player because every player has the same at-bats (n_j) and r. As the posterior mean gets closer to 0.5, the posterior variance gets bigger and has the biggest value when the posterior mean is 0.5. Intuitively, it makes sense. If a true batting average is 0.1, we can easily expect less hits. But if it is 0.5, we are less certain whether an at-bat would be more likely a hit or not, like a coin tossing. Our model is automatically reflecting on such uncertainty.

```
R> summary(b)

Main summary:

                          obs.mean  n   X1 prior.mean shrinkage low.intv
Unit w/ min(obs.mean)        0.156 45 0.00      0.233     0.715    0.140
Unit w/ median(obs.mean)1    0.244 45 0.00      0.233     0.715    0.166
Unit w/ median(obs.mean)2    0.244 45 1.00      0.310     0.715    0.210
Unit w/ max(obs.mean)        0.400 45 1.00      0.310     0.715    0.248
Overall Mean                 0.265 45 0.44      0.267     0.715    0.191

                         post.mean upp.intv post.sd
                             0.211    0.292   0.039
                             0.237    0.315   0.038
                             0.291    0.379   0.043
```

```
                    0.335     0.429    0.046
                    0.267     0.351    0.041

Second-level Variance Component Estimation Summary:
alpha = log(A) for Gaussian and log(1/r) for Binomial and Poisson data:

  post.mode.alpha post.sd.alpha
1          -4.727         0.957


Regression Summary:

      estimate    se  z.val p.val
beta0   -1.194 0.131 -9.129 0.000
beta1    0.389 0.187  2.074 0.038
```

This summary also includes the result of logistic regression fit because we did not know prior means in advance. We had to estimate them via this logistic regression model. Let's look at the p-value of the indicator covariate (`beta1`). It shows that the two prior means distinguishing outfielders from other positions were significantly different, supporting your initial belief. The positive sign of $\hat{\beta}_1$ reflects on the first five players, outfielders, whose observed batting averages were higher than others.

```
R> plot(b)
```

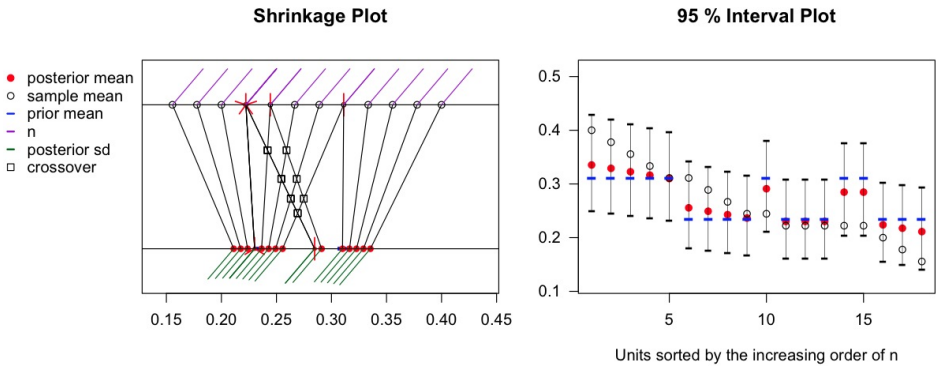

Figure 6: Shrinkage plot and 95% interval plot

Let's see shrinkage and 95% interval plots for more intuition. It is obvious that sample means (empty dots) on the upper line shrink towards two different means, 0.310 and 0.233. For reference, the red line symbols on dots are for when two or more points have the same mean and are plotted over each other. For example, five players (from the 11th player to the 15th) have the same sample mean (0.222) and at this point on the upper line, there are short red lines toward five directions. This symbol also means that the seemingly two lines crossing over others were actually three lines for those who were outfielders (10th, 14th, and 15th hitters).

For reference, the observed sample mean among outfielders is 0.308 and that among other positions is 0.231. But why are the estimated two prior means from this Binomial model 0.310 and 0.233 each? This can be attributed to the logistic regression which estimates prior mean by a non-linear logit function of $x'\beta$, causing a small bias (see 3.3). The Normal model can avoid this small bias because it uses a linear regression to estimate these prior means.

The 95% interval plot shows us range of true batting average for each player, which clarifies the regression toward the mean (RTTM) within two groups. Let's see the 10th, 14th, and 15th players on the graph for example. They are outfielders but their batting averages (sample mean) are far lower than the first five outfielders. RTTM means that it can happen by their bad luck though in the long run their batting averages will come back to their expected ones as outfielders (blue horizontal line). The fact that their sample means are located at the lower part of their 95% intervals supports this argument. So, their posterior means (red dots) can also be interpreted as their RTTM.

One more thing to emphasize is that observed sample means are all within the interval estimates. And these interval estimates also covered the batting averages over the remaining season successfully, although the result was left out. Then, how much can we trust these interval estimates? The following method check will answer it.

```
R> bcv <- coverage(b, nsim = 10000)
```

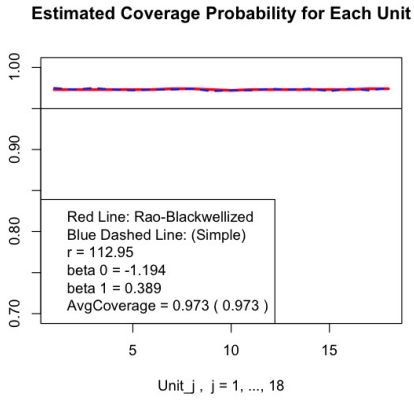

Figure 7: Coverage plot via method checking

The meaning of putting the `gbp` object, `b`, without any additional argument in the above `coverage` function is to make this function regard the estimated values, such as $\hat{r}$ (=112.95) and $\hat{\beta}$ ($= (-1.194,\ 0.389)^T$), as true values when it generates pseudo-data sets.

Finally, the estimated average coverage probability is 0.973 based on the Rao-Blackwellization, conservatively satisfying the definition of the 95% confidence interval.

7. Discussion

Bayes vs Freq, post.mean vs estimate, post.sd vs estimated standard error.

8. Acknowledgments

References

Christiansen CL, Morris CN (1997). "Hierarchical Poisson Regression Modeling." *Journal of the American Statistical Association*, **92**(438), pp. 618–632. ISSN 01621459. URL http://www.jstor.org/stable/2965709.

Efron B, Morris C (1975). "Data Analysis Using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association*, **70**(350), pp. 311–319. ISSN 01621459. URL http://www.jstor.org/stable/2285814.

Kass RE, Steffey D (1989). "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)." *Journal of the American Statistical Association*, **84**(407), pp. 717–726. ISSN 01621459. URL http://www.jstor.org/stable/2289653.

Morris C, Tang R (2011). "Estimating Random Effects via Adjustment for Density Maximization." *Statistical Science*, **26**(2), pp. 271–287. ISSN 08834237. URL http://www.jstor.org/stable/23059992.

Morris CN (1983). "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, **78**(381), pp. 47–55. ISSN 01621459. URL http://www.jstor.org/stable/2287098.

Morris CN, Christiansen C (1995). "Hierarchical Models for Ranking and for Identifying Extremes, With Application." In J Bernardo, J Berger, A Dawid, A Smith (eds.), *Bayesian Statistics 5*, pp. 227–296. New York: Oxford University Press.

Morris CN, Lysy M (2012). "Shrinkage Estimation in Multilevel Normal Models." *Statistical Science*, **27**(1), 115–134.

Rubin DB (1981). "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics*, **6**(4), pp. 377–401. ISSN 03629791. URL http://www.jstor.org/stable/1164617.

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

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