Analysis of batch variation in yields of dyestuff

Wyara Vanesa Moura e Silva
AMS 206 Quiz 1

Abstract

A data from batch variation in yields of dyestuff, which was first presented by Davies (1967), was analyze using random effects model. And it was found that the variance between the batches got close value with the error variance. However, the joint empirical distribution of these variances was asymmetric.

KEY WORDS: batch, random, effects, variance.

1. Introduction

The data analyzed represents batch to batch variation in yields of dyestuff. The data arise from a balanced experiment whereby the total product yield was determined for 5 samples from each of 6 randomly chosen batches of raw material. This data was analyzed by Bao & Tiao (1973). The data can be seen in the following table.

Table 1: Batch variation in yields of dyestuff.

Batch					
1	1545	1440	1440	1520	1580
2	1540	1555	1490	1560	1495
3	1595	1550	1605	1510	1560
4	1445	1440	1595	1465	1545
5	1595	1630	1515	1635	1625
6	1520	1455	1450	1480	1445

One objective of the experiment can be to learn to what extent batch to batch variation in a certain raw material was responsible for variation in the final product yield (Bao & Tiao 1973). For the pooled sample mean we have $\bar{y}_{..} = 1527.50$ and for the pooled sample standard deviation, $\sigma_y = 63.02$. In this paper, we are going to analyze this data using random effects modeling. First, in Section 2 it will be shown the model and methods used in the analysis. Then, the results of the analysis in Section 3.1.

2. Methods

2.1 Hierarchical Model

Denote y_{ij} the j-th observation in the i-th batch. In which $i = 1, ..., N$, $N = 6$ and $j = 1, ..., n$, $n = 5$. Assuming a model with the following likelihood for the y_{ij},

$$y_{ij}|b_i, \mu, \sigma_y^2, \sigma_b^2 \quad \sim \quad \mathcal{N}(\mu + b_i, \sigma_y^2)$$

and considering the random effects

$$b_i \quad \sim \quad \mathcal{N}(0, \sigma_b^2)$$

And in order to complete the model, it was assumed uniform priors for μ, σ_y^2 and σ_b^2. Thus, in this model, it is assumed that we have N independents normally distributed data points with each N having its own mean, but the same variance. Now we can evaluate the full conditionals for the three parameters.

The full conditionals distributions, for $\boldsymbol{\theta} = (b_i, \mu, \sigma_y, \sigma_b)$ are given by

$$\pi(b_i|\cdot) \quad \sim \quad \mathcal{N}\left(\frac{\sigma_b^2}{n\sigma_b^2 + \sigma_y^2}\sum_{j=1}^n (y_{ij} - \mu), \frac{\sigma_y^2 \sigma_b^2}{n\sigma_b^2 + \sigma_y^2}\right)$$

$$\pi(\mu|\cdot) \quad \sim \quad \mathcal{N}\left(\frac{1}{Nn}\sum_{i=1}^N \sum_{j=1}^n (y_{ij} - b_i), \frac{\sigma_y^2}{Nn}\right)$$

$$\pi(\sigma_y^2|\cdot) \quad \sim \quad \mathcal{IG}\left(\frac{Nn-2}{2}, \sum_{i=1}^N \sum_{j=1}^n \frac{[y_{ij} - (\mu + b_i)]^2}{2}\right)$$

$$\pi(\sigma_b^2|\cdot) \quad \sim \quad \mathcal{IG}\left(\frac{N-2}{2}, \sum_{i=1}^N \frac{b_i^2}{2}\right)$$

in which the $(\cdot)$ represents the data and all the other variables, $\mathcal{N}$ the normal density, and $\mathcal{IG}$ the inverse gamma density. The steps used to find those full conditionals are given in the Appendix. Now we can use Gibbs sampling to explore the posterior distribution for all the parameters. The idea is to generate posterior samples by sweeping through each variable to sample from its conditional distribution with the remaining variables fixed to their current values.

One thing that can be done it is to evaluate the joint marginal posterior distribution of μ, σ_y^2 and σ_b^2. This can be done by integrating out the random effects $(b_1, ..., b_N)$, then the marginal posterior density will be given by

$$\pi(\mu, \sigma_y, \sigma_b) \quad \sim \quad \prod_{i=1}^{N} \mathcal{N}\left(\bar{y}_{i.}|\mu, \sqrt{\sigma_y^2/n + \sigma_b^2}\right) \times$$

$$\times \quad \mathcal{G}\left(S_i|(n-1)/2, 1/(2\sigma_y^2)\right)$$

in which, $\mathcal{N}$ is the normal density, and $\mathcal{G}$ is the gamma density, and S_i is the sum of squares for the i-th batch. The steps used to find marginal posterior density are given in the Appendix.

3. Results

In the following results, it was obtained $M = 4000$ posterior samples after the burn in period. The results for the inferences about posterior mean and quantiles it is given in Table 2.

Table 2: Summary of inference for the batch variation in yields, posterior mean and quantiles, for all parameters estimated.

	mean	25%	median	75%
b_1	-18.667	-103.119	-18.500	67.711
b_2	1.085	-82.213	0.507	91.239
b_3	31.726	-48.045	29.909	121.375
b_4	-24.784	-112.126	-23.745	59.203
b_5	61.418	-15.952	58.979	152.498
b_6	-48.123	-136.857	-46.618	33.018
μ	1526.87	1447.781	1527.822	1602.309
σ_y	53.586	39.745	52.557	72.723
σ_b	77.518	24.608	64.556	210.996

Figure 1 shows the curves for the posterior densities of the random effects $(b_1, ..., b_N)$. We can see that the densities are slightly centered around zero, as expected.

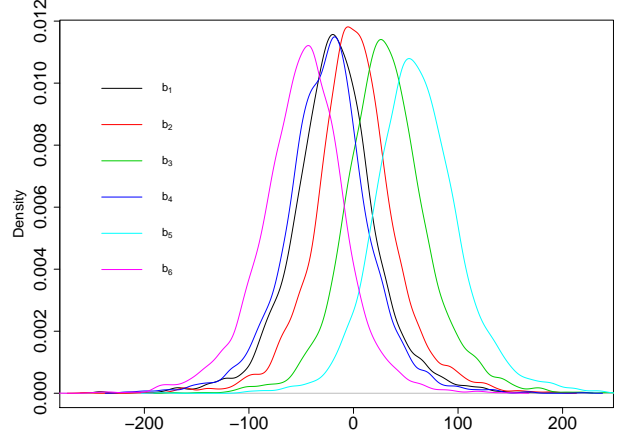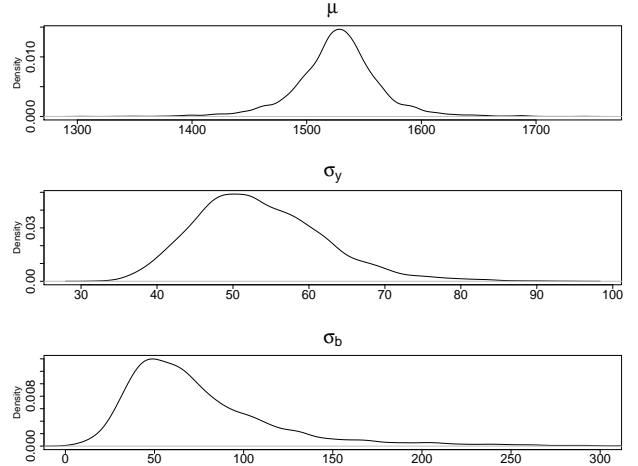Figure 1: Posterior densities for the b_i parameters.

Figure 2 displays the curves for the posterior densities regarding the parameters μ, σ_y and σ_b. We can see the asymmetry of the posterior distribution for both variances.

Figure 2: Posterior densities for the three parameters.

3.1   Normal approximation

Using a large n and given that the $\pi(\mu, \log \sigma_y, \log \sigma_b)|\mathbf{y})$ satisfy some regularity conditions, we can do the normal approximation for this posterior by: First, using the second order Taylor series expansion evaluated under $\hat{\mu}$, $\log \hat{\sigma_y}$ and $\log \hat{\sigma_b}$, in which these are the posterior mode, the maximum likelihood estimators. Then, setting $\boldsymbol{\theta} = (\mu, \log \sigma_y, \log \sigma_b)^T$ and $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \log \hat{\sigma_y}, \log \hat{\sigma_b})^T$,

2

we get the approximation:

$$\log(\pi(\boldsymbol{\theta}|\mathbf{y})) \quad \approx \quad \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathcal{H}_{\log(\pi(\hat{\boldsymbol{\theta}}|\mathbf{y}))}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

where, $\mathcal{H}_{\log(\pi(\hat{\boldsymbol{\theta}}|\mathbf{y}))}$ is the Hessian matrix. Then, using this expansion, the approximation for the posterior density will have as distribution: Normal $\left(\hat{\boldsymbol{\theta}}, [\mathrm{I}_f(\hat{\boldsymbol{\theta}})]^{-1}\right)$. Which we can find the Fisher information by using the hessian matrix. It was used the `optim` function in `R` to find the values for the parameters, this function use the BFGS method which is an optimization algorithm in the family of quasi-Newton methods (Gelman et al., 2014), and the results were:

$$\hat{\boldsymbol{\theta}} \quad = \quad (1527.3107, 3.9176, 3.7688)$$

$$\left[\mathrm{I}_f(\hat{\boldsymbol{\theta}})\right]^{-1} \quad = \quad \begin{pmatrix} 397.2024 & -0.0005 & -0.0473 \\ -0.0005 & 0.0212 & -0.0039 \\ -0.0473 & -0.0039 & 0.1596 \end{pmatrix}$$

$$(1)$$

Figure 3 displays the contour plot of the exact density overlaid on the normal distribution for the parameters $\log \sigma_y$ and $\log \sigma_b$, with μ fixed in its maximum likelihood estimator. We can see a asymmetrical shape of the distribution in the contour plots. In addition, the tails of the exact density seems to be heavier than the normal approximation for the posterior.

Figure 3: Contour plot of the exact overlaid on the normal approximation of the parameters $\log \sigma_y$ and $\log \sigma_b$.
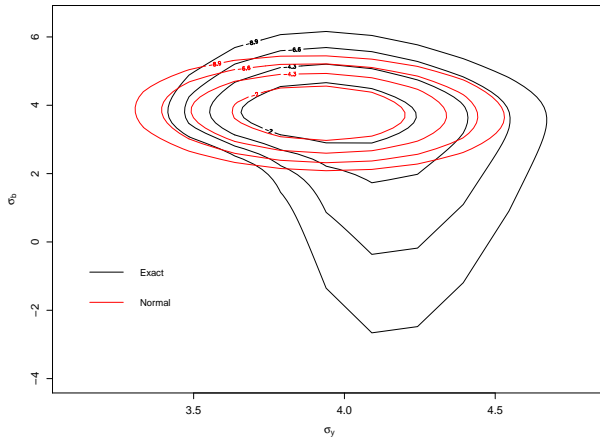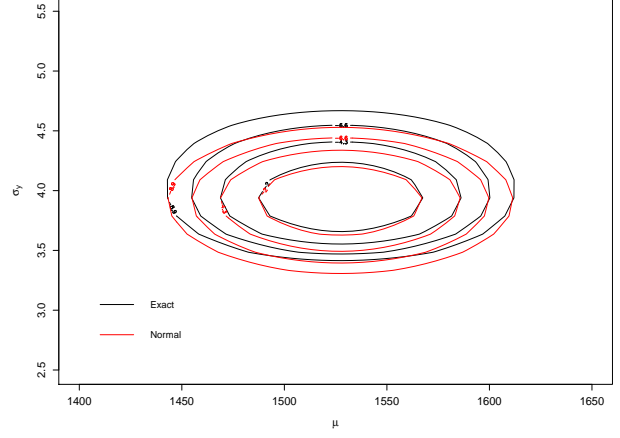
Figure 4: Contour plot of the exact overlaid on the normal approximation of the transformed parameters μ and σ_y.

In Figures 4 and 5 we can see the contour plots of the exact density overlaid on the normal distribution for the parameters μ by $\log \sigma_y$ and μ by $\log \sigma_b$, with σ_b and σ_y fixed in its maximum likelihood estimator, respectively.

For the normal approximation in Figure 4 the posterior distribution seem to fit well with the exact density. In contrast, the approximation in Figure 5 the modeling appears to be the worst, because the exact density has heavier tails than the normal density.

Figure 5: Contour plot of the exact overlaid on the normal approximation of the transformed parameters μ and σ_b.
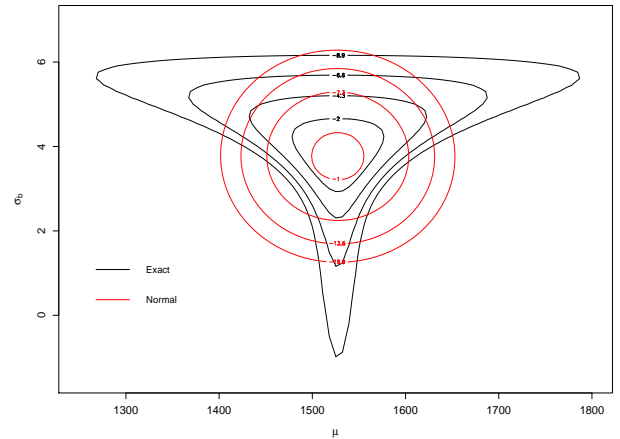
## 3.2 Rejection Sampling and SIR method

In order to obtain a random draws from the joint marginal distribution for the parameters $\boldsymbol{\theta} = (\mu, \log \sigma_y, \log \sigma_b)^T$, we need to define some features.

First, it was chosen a multivariate t distribution as proposal distribution $\pi(\boldsymbol{\theta})$ that resembles the posterior density, and another reason is that we need a function that has thicker tails.

The parameters for the proposal distribution are given by: $p(\boldsymbol{\theta}) \sim MVT_3(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}})$. In which, the degree of freedom is $\nu = 3$, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ found in Section 3.1 was used as the location parameter, and covariance matrix $\hat{\boldsymbol{\Sigma}}$ found in Section 3.1 was used as scale matrix. These values are in (1).

The value found for the identifiable constant the was $M = -78.0820$. And it was found that the maximum value M occurs at the value of $\boldsymbol{\theta} = (1444.3140, 3.9177, 4.7733)$. In which we can notice that this values of $\boldsymbol{\theta}$ are not at the extreme part of the parametric space. This gives an indication that the value found for M is, in fact, an approximate maximum. However, the original sample size was $55,000$ and we got the acceptance rate 0.1048 being a small number.

Figure 6: Contour plot of transformed parameters $\log \sigma_y$ and $\log \sigma_b$ with simulated draws from the rejection algorithm.
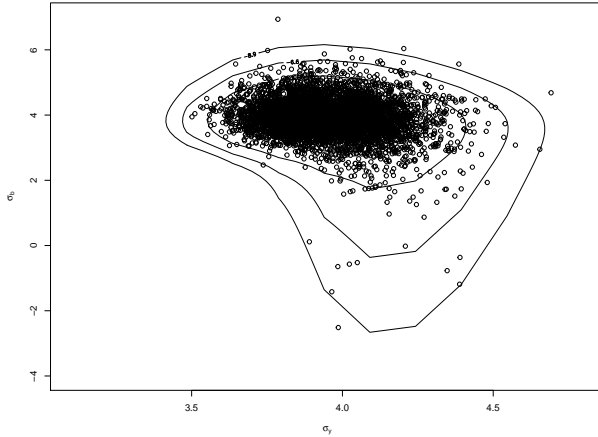
For the sampling importance resampling (SIR), as in the rejection method it was chose a multivariate t distribution as proposal distribution $p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta}) \sim MVT_3(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}})$. Then, it was computed the weights. The sample size was chose as the size of the accepted sample in the rejection method.

Figure 7: Contour plot of transformed parameters $\log \sigma_y$ and $\log \sigma_b$ with simulated draws from the SIR algorithm.
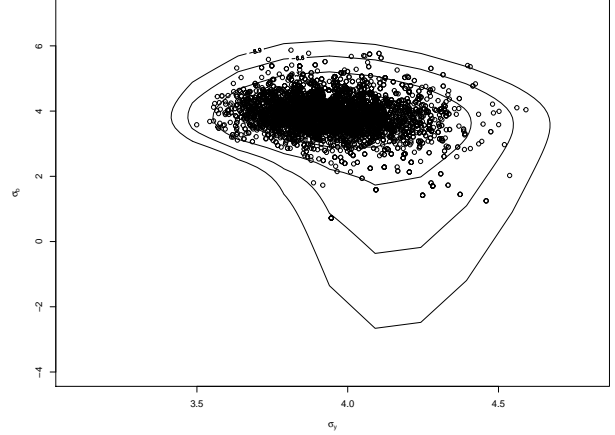
By looking the scatted plots in Figure 6 and 7 of the simulated points using the two methods, for the transformed parameters $\log \sigma_y$ and $\log \sigma_b$, seems that the rejection method gets more accurate results for the simulated points because the simulated draws had covered better the exact density of the posterior distribution.

In Figure 8, 9 and 10 we can see the density of the parameters μ, $\log \sigma_y$, $\log \sigma_b$, respectively, with curves from the three methods used to sample draws for the posterior distribution of the given parameters.

Figure 8: Posterior densities for μ using the three methods of sampling.

4
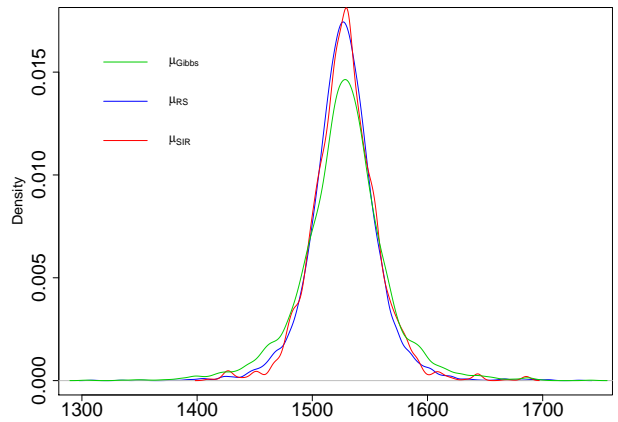
We can notice by looking at these graphs that the three methods got close values for the parameters, except for σ_b (Figure 10) that the location of the distribution using Gibbs sampling had a different value when compared to the method of rejection and SIR.

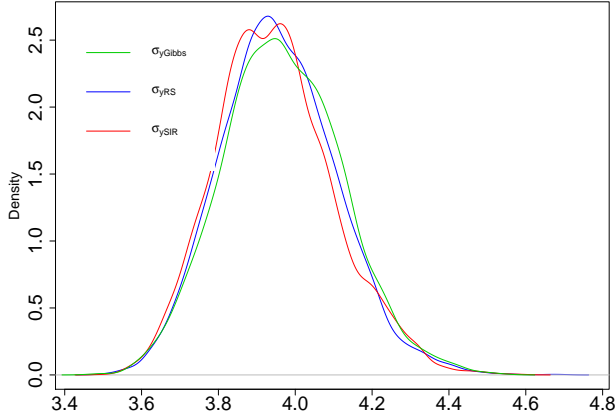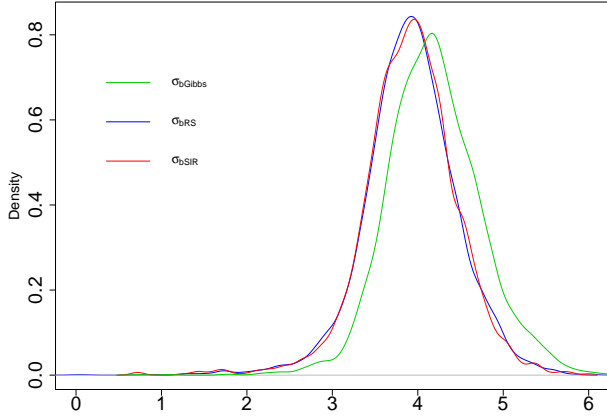Figure 9: Posterior densities for σ_y using the three methods of sampling.

Figure 10: Posterior densities for σ_b using the three methods of sampling.

4. Remarks

It was modeled the data from batch variation in yields of dyestuff using random effects model. The inference was made by using Gibbs sampling, normal approximation, rejection sampling and SIR method, in which the results obtained were similar in all three methods. Regarding the variance between the batch, it had close value with the variance from the whole data. When it is compared the μ parameters in the random effect and the variance for the batch we could notice that the distribution was very asymmetric if compared with the behavior between μ and the variance for the data.

5. Appendix

5.1 Full Conditionals Distribution

In this section, it will be showing some steps to find the full conditional distribution of the parameters, and then some steps to obtain the marginal posterior density for $\boldsymbol{\theta} = \mathbf{b}_i, \mu, \sigma_y, \sigma_b$.

First, we need the joint posterior distribution, which is given by:

$$
\begin{aligned}
\pi(\boldsymbol{\theta}|\mathbf{y}) \quad \propto \quad & \pi(\mathbf{b}_i, \mu, \sigma_y, \sigma_b)\pi(\mathbf{y}|\mathbf{b}_i, \mu, \sigma_y, \sigma_b) \\[6pt]
\pi(\boldsymbol{\theta}|\mathbf{y}) \quad \propto \quad & \prod_{i=1}^{N} \left(\frac{1}{\sigma_b^2}\right)^{-\frac{1}{2}} \exp\left\{\frac{b_i^2}{2\sigma_b^2}\right\} \times \\
& \times \prod_{i=1}^{N}\prod_{j=1}^{n} (\sigma_y^2)^{-\frac{1}{2}} \exp\left\{-\frac{(y_{ij} - (\mu - b_i))^2}{2\sigma_y^2}\right\} \\[6pt]
\propto \quad & \left(\frac{1}{\sigma_b^2}\right)^{-\frac{N}{2}} \exp\left\{\frac{1}{2\sigma_b^2}\sum_{i=1}^{N} b_i^2\right\} \times \\
& \times \left(\frac{1}{\sigma_y^2}\right)^{-\frac{Nn}{2}} \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^{N}\sum_{j=1}^{n} (y_{ij} - (\mu - b_i))^2\right\}
\end{aligned}
$$

Then, we can find all the full conditionals distribution, in which the ($\cdot$) represents the data and all the other variables, $\mathcal{N}$ the normal density, $\mathcal{G}$ the gamma density, and $\mathcal{IG}$ the inverse gamma density, the results are following:

- For $\boldsymbol{b}_i$

5

$$\pi(b_i|\cdot) \propto \exp\left\{\frac{1}{2\sigma_b^2}\sum_{i=1}^N b_i^2\right\} \times$$

$$\times\exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^N\sum_{j=1}^n (y_{ij}-(\mu-b_i))^2\right\}$$

$$\propto \exp\left\{\frac{1}{2\sigma_b^2}\sum_{i=1}^N b_i^2\right\} \times$$

$$\times\exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^N\sum_{j=1}^n (-2y_{ij}b_i + 2\mu b_i + b_i^2)\right\}$$

$$\propto \exp\left\{\frac{1}{2\sigma_y^2\sigma_b^2}\sum_{i=1}^N b_i^2(\sigma_y^2+n\sigma_b^2)\right\} \times$$

$$\times\exp\left\{-\frac{1}{2\sigma_y^2\sigma_b^2}\left[-2\sum_{i=1}^N b_i\sigma_b^2\sum_{j=1}^n(y_{ij}-\mu)\right]\right\}$$

$$\sim \mathcal{N}\left(\frac{\sigma_b^2}{n\sigma_b^2+\sigma_y^2}\sum_{j=1}^n(y_{ij}-\mu),\ \frac{\sigma_y^2\sigma_b^2}{n\sigma_b^2+\sigma_y^2}\right)$$

• For μ

$$\pi(\mu|\cdot) \propto \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^N\sum_{j=1}^n(y_{ij}-(\mu-b_i))^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^N\sum_{j=1}^n(\mu^2-2\mu y_{ij}-2\mu b_i)\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma_y^2}\left(Nn\mu^2-2\mu\sum_{i=1}^N\sum_{j=1}^n(y_{ij}-b_i)\right)\right\}$$

$$\propto \exp\left\{-\frac{Nn}{2\sigma_y^2}\left(\mu^2-2\mu\frac{1}{Nn}\sum_{i=1}^N\sum_{j=1}^n(y_{ij}-b_i)\right)\right\}$$

$$\sim \mathcal{N}\left(\frac{1}{Nn}\sum_{i=1}^N\sum_{j=1}^n(y_{ij}-b_i),\ \frac{\sigma_y^2}{Nn}\right)$$

• For σ_y^2

$$\pi(\sigma_y^2|\cdot) \propto (\sigma_y^2)^{-\frac{Nn}{2}}\exp\left\{\frac{1}{2\sigma_y^2}\sum_{i=1}^N\sum_{j=1}^n(y_{ij}-(\mu-b_i))^2\right\}$$

$$\sim \mathcal{IG}\left(\frac{Nn-2}{2},\ \sum_{i=1}^N\sum_{j=1}^n\frac{[y_{ij}-(\mu+b_i)]^2}{2}\right)$$

• For σ_b^2

$$\pi(\sigma_b^2|\cdot) \propto (\sigma_b^2)^{-\frac{N}{2}}\exp\left\{\frac{1}{2\sigma_b^2}\sum_{i=1}^N b_i^2\right\}$$

$$\sim \mathcal{IG}\left(\frac{N-2}{2},\ \sum_{i=1}^N\frac{b_i^2}{2}\right)$$

Now, by integrating out the random effects $\boldsymbol{b}_i$, we can obtain the following marginal posterior distribution, $\boldsymbol{\theta}' = \mu,\ \sigma_y,\ \sigma_b$.:

$$\pi(\boldsymbol{\theta}'|\mathbf{y}) \propto \int_{\boldsymbol{b}_i}\pi(\mathbf{b}_i,\mu,\sigma_y,\sigma_b|\mathbf{y})d\boldsymbol{b}_i$$

$$\propto \prod_{i=1}^N\int_{\boldsymbol{b}_i}\left(\frac{1}{\sigma_b^2}\right)^{-\frac{N}{2}}\exp\left\{\frac{1}{2\sigma_b^2}b_i^2\right\} \times$$
$$\times\left(\frac{1}{\sigma_y^2}\right)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma_y^2}\sum_{j=1}^n(y_{ij}-(\mu-b_i))^2\right\}d\boldsymbol{b}_i$$

$$\propto \prod_{i=1}^N\left(\frac{1}{\sigma_y^2}\right)^{-\frac{n}{2}}\left(\frac{1}{\sigma_b^2}\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_y^2}\sum_{j=1}^n(y_{ij}-\mu)^2\right\}$$

$$\times\int_{\boldsymbol{b}_i}\exp\left\{\frac{1}{2\sigma_b^2}b_i^2\right\} \times$$
$$\times\exp\left\{-\frac{1}{2\sigma_y^2}\sum_{j=1}^n(y_{ij}-(\mu-b_i))^2\right\}d\boldsymbol{b}_i$$

$$\propto \prod_{i=1}^N\left(\frac{1}{\sigma_y^2}\right)^{-\frac{n}{2}}\left(\frac{1}{\sigma_b^2}\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_y^2}\sum_{j=1}^n(y_{ij}-\mu)^2\right\}$$

$$\times\int_{\boldsymbol{b}_i}\exp\left\{-\frac{1}{2}\left[\frac{b_i^2-2\sum_{j=1}^n(y_{ij}-\mu)b_i\frac{\sigma_b^2}{n\sigma_b^2+\sigma_y^2}}{\frac{\sigma_b^2\sigma_y^2}{n\sigma_b^2+\sigma_y^2}}\right]\right\}d\boldsymbol{b}_i$$

$$\propto \quad \prod_{i=1}^{N} \left(\frac{1}{\sigma_y^2}\right)^{\frac{n}{2}} \left(\frac{1}{\sigma_b^2}\right)^{\frac{1}{2}} \left(\frac{\sigma_b^2 \sigma_y^2}{n\sigma_b^2 + \sigma_y^2}\right)^{-\frac{1}{2}} \times$$

$$\times \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{j=1}^{n}(y_{ij}-\mu)^2\right\}$$

$$\times \exp\left\{\left[\sum_{j=1}^{n}(y_{ij}-\mu)\left(\frac{\sigma_b^2}{n\sigma_b^2+\sigma_y^2}\right)\right]^2 \frac{n\sigma_b^2+\sigma_y^2}{2\sigma_b^2\sigma_y^2}\right\}$$

$$\propto \quad \prod_{i=1}^{N} \left(\frac{1}{\sigma_y^2}\right)^{\frac{n}{2}} \left(\frac{1}{\sigma_b^2}\right)^{\frac{1}{2}} \left(\left[\sigma_b^2+\frac{\sigma_y^2}{n}\right]\frac{n}{\sigma_b^2\sigma_y^2}\right)^{-\frac{1}{2}} \times$$

$$\times \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{j=1}^{n}(y_{ij}-\mu)^2\right\}$$

$$\times \exp\left\{\left[\sum_{j=1}^{n}(y_{ij}-\mu)\left(\frac{\sigma_b^2}{n\sigma_b^2+\sigma_y^2}\right)\right]^2 \frac{n\sigma_b^2+\sigma_y^2}{2\sigma_b^2\sigma_y^2}\right\}$$

$$\propto \quad \prod_{i=1}^{N} \left(\frac{1}{\sigma_y^2}\right)^{\frac{n-1}{2}} \left(\sigma_b^2+\frac{\sigma_y^2}{n}\right)^{-\frac{1}{2}} \times$$

$$\times \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2 + n(\bar{y}_{i\cdot}-\mu)^2\right\}$$

$$\times \exp\left\{\frac{\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot}+\bar{y}_{i\cdot}-\mu)^2\left(\frac{\sigma_b^2}{n\sigma_b^2+\sigma_y^2}\right)\right]^2}{2\left(\sigma_b^2+\frac{\sigma_y^2}{n}\right)}\right\}$$

$$\propto \quad \prod_{i=1}^{N} \left(\frac{1}{\sigma_y^2}\right)^{\frac{n-1}{2}} \left(\sigma_b^2+\frac{\sigma_y^2}{n}\right)^{-\frac{1}{2}} \times$$

$$\times \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{j=1}^{n}S_i\right\} \exp\left\{\frac{(\bar{y}_{i\cdot}-\mu)^2}{2\left(\sigma_b^2+\frac{\sigma_y^2}{n}\right)}\right\}$$

$$\sim \quad \prod_{i=1}^{N} \mathcal{G}\left(S_i\Big|\frac{n-1}{2},\frac{1}{2\sigma_y^2}\right) \mathcal{N}\left(\bar{y}_{i\cdot}\Big|\mu,\sqrt{\left(\sigma_b^2+\frac{\sigma_y^2}{n}\right)}\right)$$

Thus, the marginal posterior distribution it is given by the joint distribution of the sample mean between groups having a normal distribution, and the sum of squares a gamma density.

REFERENCES

Box, G. E. P., & Tiao, G. C. (1973). Bayesian inference in statistical inference. Adison-Wesley, Reading, Mass.

Davies, O. L. (1967), Statistical Methods in Research and Production, third edit ion, London: Oliver and Boyd.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 3). Boca Raton, FL: CRC press.