Abdulhakim Alhowaish
January 20, 2024

Bayesian Hierarchical Statistics: Case Study using R
GASTST proposal

This is a to show how we can use Bayesian Hierarchical model to infer about social data. The case will review the data chosen from a common data repository. It begins with an exploration of the chosen dataset, histogram plots along with predictive analysis are shown. The Model section will represent the model choice. After that, I will justify my choice of model using convergence and autocorrelation plots in the Analysis section. Then, I will fit the model using the JAGS package. The posterior analysis and results are shown below. Finally, I will comment and conclude with observations and future prospective improvements.

1 Data

Using the library **COUNT** I used data from Fair 1987 [1]. Fair used a tobit model with the data. The target vector is the count of the number of affairs of every man in the dataset. There are 601 responses in it. Columns are 18 in the original dataset. I modified them to be only 5 columns including the target vector. Modifications are as follows:

- target vector is taken as itself **naffairs**.
- **Kids** column is taken as itself.
- Years married column is converted into one column having the number of years instead of six categories. Stored as **yrsmar**
- religious background five-columns are converted into single column with integer entry ranging between -2 as lowest, and 2 as highest. Stored as **relig**
- happiness index five-columns are converted into single column with integer entry ranging between -2 as lowest, and 2 as highest. Stored as **happidx**

To explore the nature of the data, Figure 1 shows histograms of all four columns (features) of men who have zero affairs versus men who have any number of affairs. It can be clearly seen that there are almost no paternal differences between men with recorded affairs and those who have had one.

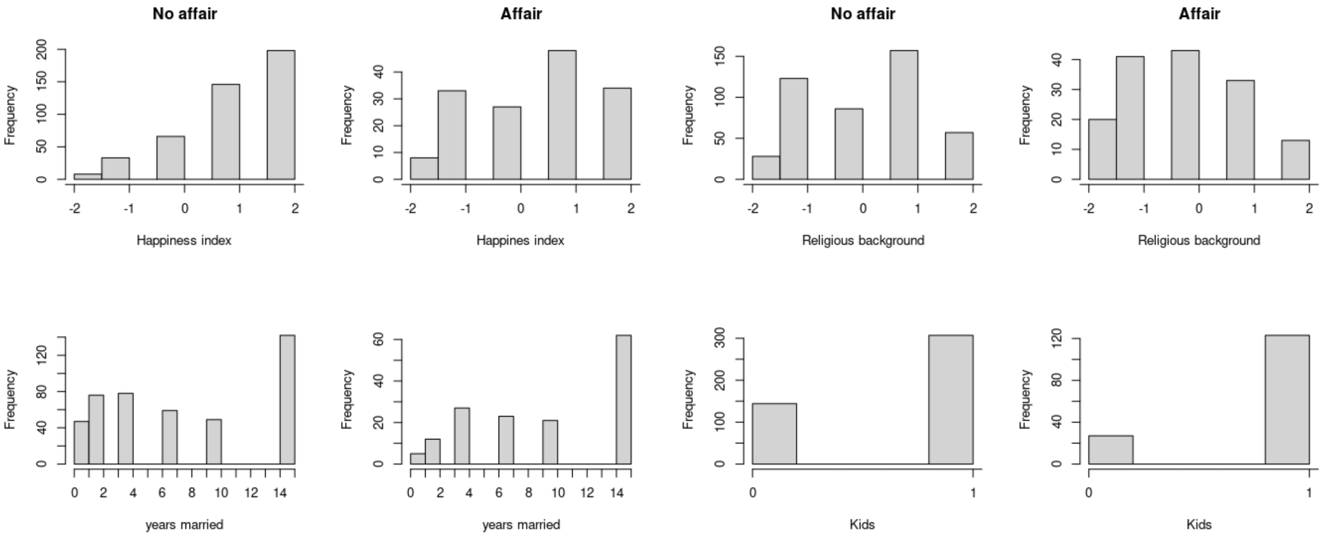

Figure 1: Distribution of categories throughout the features.

2 Model

The goal of this research can be either to predict if the man will have any affair in future or not, or to predict the number of affairs given some set up of features. These two contribute to two distinct models. The first can be modeled by having a Logit Beta-Bernoulli model. The other can be modeled as Poisson-Gamma model.

2.1 Logit Beta-Bernoulli Model

Using **rjags** library in R, this model can be illustrated as:

$$p(p_i|x_i, b_j) \propto p(x_i|p_i, w_j)p(p_i|b_j)p(b_j|\mu_0, \sigma_0^2) \tag{1}$$

$$p(x_i|p_i, b_j) \sim \mathbf{Bernoulli}(p_i, n_i) \tag{2}$$

$$logit(p_i|b_j) = \exp(\sum_j b_j * F_j) \tag{3}$$

$$p(b_j|\mu_0, \sigma_0^2) \sim \mathbf{Beta}(0.5, 0.5) \tag{4}$$

where F_j is the jth feature.

2.2 Poisson-Gamma Model

Using **rjags** library in R, this model can be illustrated as:

$$p(\lambda_i|x_i, b_j) \propto p(x_i|\lambda_i, b_j)p(\lambda_i|b_j)p(b_j|\alpha_0, \beta_0) \tag{5}$$

$$p(x_i|\lambda_i, b_j) \sim \mathbf{Poisson}(\lambda_i) \tag{6}$$

$$\lambda_i|b_j = \exp(\sum_j b_j * F_j) \tag{7}$$

$$p(b_j|\alpha_0, \beta_0) \sim \mathbf{Gamma}(\alpha_0, \beta_0) \tag{8}$$

where F_j is the jth feature.

3 Analysis

In this section, I will show mainly three tests:

- Convergence test showing the weight graph of the MC samples.
- Auto-correlation test by showing the autocorrelation plot of the weights.
- Deviance information criterion (DIC) to select the appropriate model.

3.1 Logit Beta-Bernoulli Model

using the model specified in equations (1 - 4) Figure 2 shows the convergence and autocorrelation tests. All weights are doing fine. Furthermore, the DIC result is relatively good. $pen.deviance = 837$.

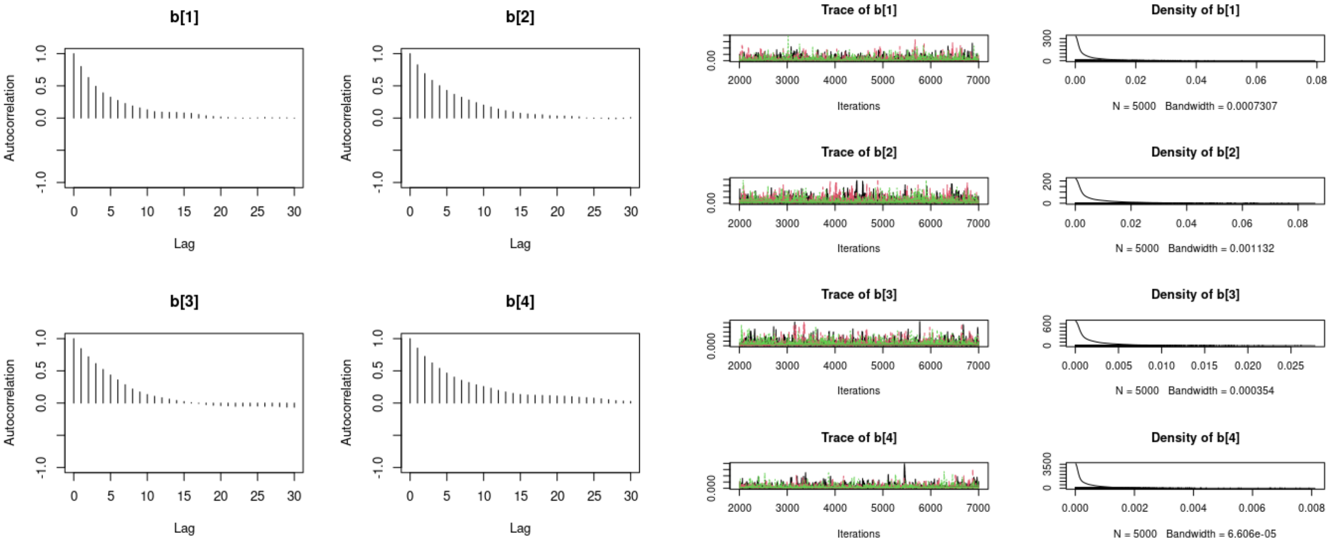

Figure 2: Convergence test and auto-correlation plot of the weights for all gamma prior model.

3.2 Poisson-Gamma Model

using model specified in equations (5-8) Figure 3 shows the convergence and autocorrelation tests. It can be clearly seen that only b_4 is doing good. All other three weights are not doing well. Furthermore, the DIC result is relatively high $pen.deviance = 3289$.

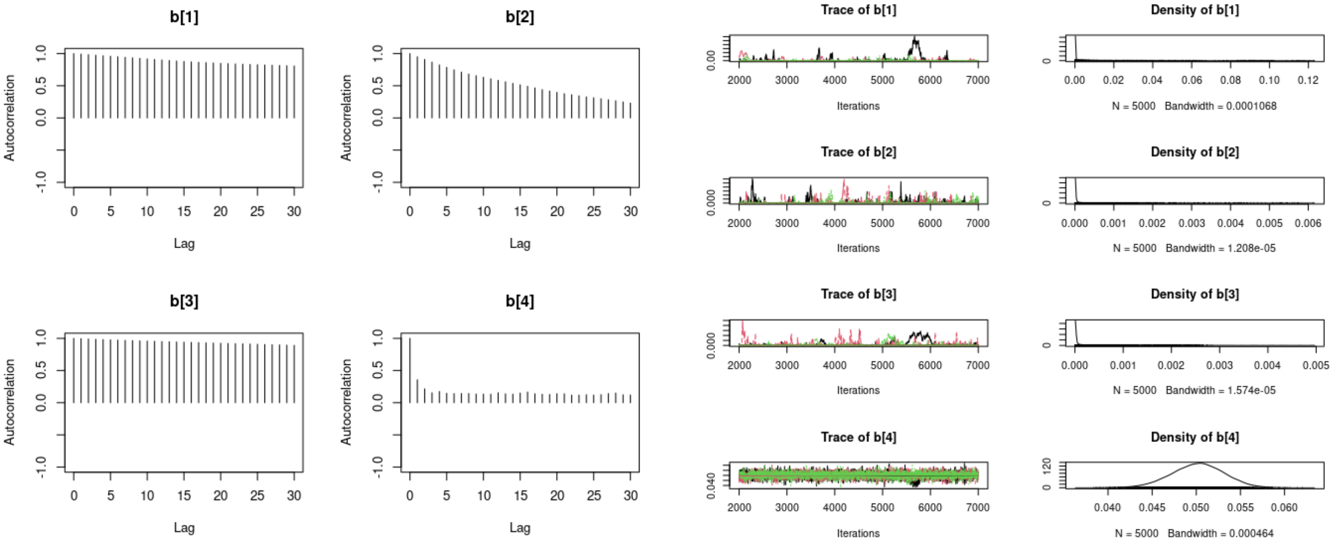

Figure 3: Convergence test and auto-correlation plot of the weights for all beta prior model.

So instead of using gamma prior in all weights, I replaced the three weights with prior to be normal. So I updated equation (8) to be:

$$p(b_i|\mu_0, \sigma_0^2) \sim \textbf{Normal}(0, 16) \textbf{ where } i = 1, 2, 3 \tag{9}$$

$$p(b_4|\alpha_0, \beta_0) \sim \textbf{Gamma}(\alpha_0, \beta_0) \tag{10}$$

Figure 4 shows the previous tests. A definite improvement in both graphs. Also, the DIC result is relatively better $pen.deviance = 2906$

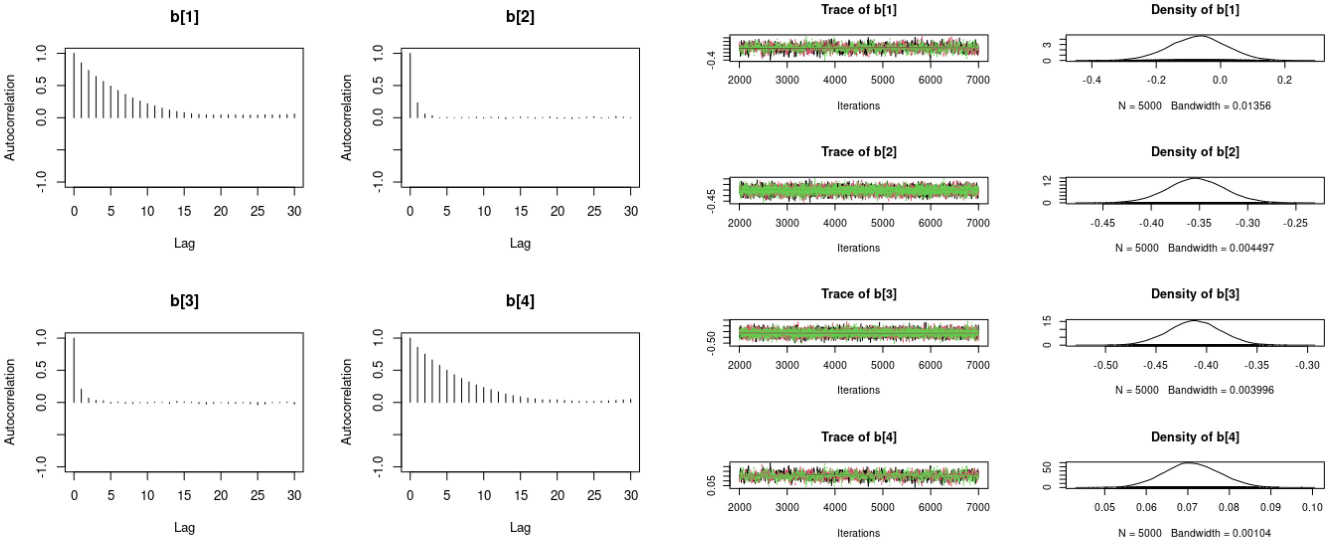

Figure 4: Convergence test and auto-correlation plot of the weights for mixture prior model.

4 Results

In this section, I will take the cases of randomly generated data, $n_{sim} = 10^5$, using emergent weights. Table () shows the probability of having an affair that each man will have given the setup.

Years Married (no affair)	Poisson-Gamma prob	Beta-Bernoulli prob	Real-data percentage
0.75	0.82	0.6	0.9
1.5	0.72	0.62	0.86
4	0.65	0.57	0.74
7	0.58	0.68	0.72
10	0.55	0.61	0.7
15	0.47	0.49	0.69

Conclusions

In this case study, I have used a data set from the R repository to apply to hierarchical models. The first model was to find whether a man is likely to have an affair given some features using **Logit Beta-Bernoulli Model** model. The other is to find how many affairs would he have. The other model is implemented using **Poisson-Gamma Model**. The results of both models were compared with the general linear model (glm) function in R. The results show some consistency with the original data. In the future, an extra step hierarchical model can be used. For example, in the Logit Beta-Bernoulli Model, we can have a non-informative beta prior to $\alpha, \beta = 0.5$. The course gives me a powerful statistical tool to understand the problem. In addition, I can create more data of the same nature and predict future trends.

References

[1] Fair, R. (1978). *A Theory of Extramarital Affairs*, Journal of Political Economy, 86: 45-61.