Case Study: Examination of Bicycle Traffic at the University of California, Berkeley

Robert Espinoza

Abstract

In this case study we will examine the proportion of bicycles at the University of California, Berkeley with survey data of counts of bicycles and other vehicles in one hour for 10 different city blocks in the residential streets with bike routes. The goal of the study is to compare designs of hierarchical models of the data. The first model attempts to model the number of bicycles at location j, for j=1,...,10 as a binomial random variable with unknown probability θ, and sample size n_j, which is the total count of bicycles and other vehicles in the j^{th}. θ is interpreted as the true proportion of traffic of bicycles in all the blocks. In this model we assign a beta distribution to model θ with hyperparameters α and β, and use a non-informative hyperprior to model α and β. The second model we will be examining uses the same assumptions as the first hierarchical model, but with the modification that each location j has it's on proportion θ_j.

1. Model analysis of a single θ parameter for all j^{th} locations

The hierarchical model for the first model includes the following assumptions: $y_j \sim Binomial(\theta, n_j)$, $\theta \sim Beta(\alpha, \beta)$, and the hyperprior for the α and β hyperparameters will use the non-informative prior from Gelman's *Bayesian Data Analysis*, rat tumor example from chapter 5 of the textbook, which is :

$$p(\alpha, \beta) \propto (\alpha + \beta)^{\frac{5}{2}}$$

and the reparamterized hyperprior we will be using in the model is:

$$p(log(\frac{\alpha}{\beta}), log(\alpha + \beta)) \propto \alpha\beta((\alpha + \beta)^{\frac{5}{2}}$$

The joint posterior probability for all of the parameters is:

$$p(\theta, \alpha, \beta | y) \propto p(\alpha, \beta) p(\theta | \alpha, \beta) p(y | \theta, \alpha, \beta)$$

$$\propto p(\alpha, \beta) \prod_{j=1}^{10} \frac{\Gamma(\alpha + \beta)}{\Gamma(\beta)\Gamma(\beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \prod_{j=1}^{10} \theta^{y_j} (1 - \theta)^{n_j - y_j}$$

By using the joint posterior distribution we can derive the full conditional distribution of θ:

$$p(\theta | \alpha, \beta, y) \propto$$

$$\prod_{j=1}^{10} \frac{\Gamma(\alpha + \beta + n_j)}{\Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)} \theta^{\alpha + y_j - 1} (1 - \theta)^{\beta + n_j - y_j - 1}$$

The marginal posterior distribution of (α, β) can be found by using the following conditional probability formula:

$$p(\phi | y) = \frac{p(\theta, \phi | y)}{p(\theta | \phi, y)}$$

Thus the marginal posterior distribution of $(alpha, beta)$ is:

$$p(\alpha, \beta | y) \propto$$

$$p(\alpha, \beta) \prod_{j=1}^{10} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)}{\Gamma(\alpha + \beta + n_j)}$$

The reparameterized marginal posterior distribution of (α, β) by applying the appropriate change of variables and Jacobian is:

$$\alpha = \frac{\mathrm{e}^{log\frac{\alpha}{\beta}} \mathrm{e}^{log\alpha + \beta}}{\mathrm{e}^{log\alpha + \beta} + 1}$$

$$\beta = \frac{\mathrm{e}^{log\alpha + \beta}}{\mathrm{e}^{log\frac{\alpha}{\beta}} + 1}$$

$$p(log(\frac{\alpha}{\beta}), log(\alpha + \beta) | y) \propto log\Gamma(\beta) + log\Gamma(\alpha)$$

$$- \frac{5}{2} log\Gamma(\alpha + \beta) \sum_{j=1}^{10} [log\Gamma(\alpha + \beta) - log\Gamma(\alpha) - log\Gamma(\beta)$$

$$log\Gamma(\alpha + y_j) + log\Gamma(\beta + n_j - y_j) - log\Gamma(\alpha + \beta + n_j)]$$

1.1 Simulation for model 1

Taking a closer look at the data we find that the mean proportion of bicycles per city block is 0.1961 and the variance is 0.0111.

By using the mean and variance of the proportion of bicycles per block we can determine an initial starting point to simulate the α and β values, which would be an (α, β) near (2.47, 10.39). By using the transformation of α and β we should expect values of $(log(\frac{\alpha}{\beta}), log(\alpha + \beta))$ near (-1.435, 2.554).

The next step is to simulate from the reparamaterized posterior marginal distribution of (α, β) by using a grid approximation of the distribution since it is not in a closed form. First we will generate a grid of values for $(log(\frac{\alpha}{\beta}), log(\alpha + \beta))$ around the estimated values previously mentioned, and then evaluate the grid values at the reparamaterized posterior marginal distribution. Then we must generate a posterior sample from the probabilities generated from the grid values. And finally we transform back to our original (α, β) coordinate system. Once we have the sampled (α, β) values, we can use them to sample θ values, and determine an appropriate estimation for θ.

The means of the sampled $(log(\frac{\alpha}{\beta}), log(\alpha + \beta))$ are (-1.35, 2.72), which represents the means (α, β) as (3.53, 13.72). The normalized contour plot of the marginal posterior density, and a scatter plot of 500 draws from the numerically computed marginal posterior density of $(log(\frac{\alpha}{\beta}), log(\alpha+\beta))$ is displayed below:

By using the sampled (α, β) values we can determine a sample for θ. Since the full conditional distribution of θ has a closed form, which is a beta distribution, we can use a beta random number generator to generate a sample of theta. By using the parameters of the full conditional of θ described previously, our sample has a mean of 19.79 and has a 95% posterior interval of [0.08094, 0.4522].

1.1.1 *Posterior predictive using model 1*

Now that we have a numerically generated samples of theta for model 1, we will use them to determine a predictive posterior interval for a new city block. Let us assume we are trying to predict the bicycle count in a new city block within an hour of observing the block, the block is a residential area with a bike route. In the hour of observing this block there are a total of 100 vehicles including bicycles. In order to achieve a posterior predictive interval we must use the sampled thetas from the beta distribution, and use the sample to generate a new sample from the sampling distribution of y, which is a binomial distribution. By using a binomial random number generator with the parameter $n^* = 100$ and the vector of sampled θ values we can generate a sample of bicycle counts when 100 total vehicles are observed. The mean for the predicted sample of bicycle counts is 19.68 bicycles, and the 95% posterior interval is [7, 46.53].

2. Model analysis of a unique θ parameter for each j^{th} locations

The second model will observe is similar in nature as the first model, except for the fact that is has a unique $theta_j$ for each location, which means there will be 10 total θ_j This model will also have a hierarchical structure with the following assumptions: $y_j \sim Binomial(\theta_j, n_j)$, $\theta_j \sim Beta(\alpha, \beta)$, and the hyperprior for (α, β) is same non-informative prior used in model 1. The joint posterior of all of the parameters for this model is a as follows:

$$p(\theta, \alpha, \beta | y) \propto$$

$$p(\alpha, \beta) \prod_{j=1}^{10} \frac{\Gamma(\alpha + \beta)}{\Gamma(\beta)\Gamma(\beta)} \theta_j^{\alpha-1} (1-\theta_j)^{\beta-1} \prod_{j=1}^{10} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}$$

The full conditional distribution of θ is:

$$p(\theta | \alpha, \beta, y) \propto$$

$$\prod_{j=1}^{10} \frac{\Gamma(\alpha + \beta + n_j)}{\Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)} \theta_j^{\alpha + y_j - 1} (1 - \theta_j)^{\beta + n_j - y_j - 1}$$

The same simulation is used as in model 1 to find a sample for (α, β), and for this example we will use the same (α, β) sample to conduct the further analysis on model 2. The sampled θ_j's for this model will be composed of a matrix of dimension n by 10, where n is the number of samples, and for this example we will use 500 draws. The following table provides the mean and posterior intervals for each θ_j:

θ_j	Mean	95% Posterior Interval
1	0.213	[0.137, 0.304]
2	0.107	[0.056, 0.177]
3	0.179	[0.106, 0.275]
4	0.191	[0.118, 0.286]
5	0.161	[0.105, 0.220]
6	0.251	[0.169, 0.350]
7	0.178	[0.115, 0.244]
8	0.139	[0.084, 0.203]
9	0.119	[0.085, 0.155]
10	0.429	[0.343, 0.525]

As analyzed in the previous model, we compute the 95% predictive posterior interval for the new city block. Since The prediction is with respect to a new group j then our computation is the same as in model 1, by means of sampling a new $theta_j$ for the new city block observed, and then sampling the bicycle count from the binomial distribution using the parameters of $(n^* = 100, \theta_{j*})$.

3. Model Checking

The next step of the case study is to conduct modeling checking analysis on both of the models examined. We will use a posterior predictive check, specifically posterior predictive p-values. We first generate replication data for both models. The test used to compute the posterior predictive p-values will be a paired t-test. The graphs below show the comparison of the replicated data for each model vs the original data of bicycle counts.

Below are the posterior predictive p-values computed using a t-test for each of the models, which is comparing the similarity between the replicated data using model 1 versus the original count of bicycles observed, and similarly for model 2.

Paired t-test for model 1

```
data: y1.rep and y
t = -0.5098, df = 9, p-value = 0.6224
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: [-14.679944 , 9.279944]
mean of the differences: -2.7
```

Paired t-test for model 2

```
data: y2.rep and y
t = -0.3166, df = 9, p-value = 0.7588
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: [-4.072808, 3.072808]
mean of the differences: -0.5
```

From the model checking conducted above it seems that model 2 is preferable in this setting. Model 2 has a higher p-value compared to model 1, therefore meaning that the data that was replicated from model 2 is more similar to the original observed data. Also the mean of the differences was lower for model 2 versus model 1.

One possible modeling extension is to analyze a model with batched $theta_j$ that are batched on similar frequency of bicycle traffic, or batched according to spatial distance of the city blocks.

REFERENCES

Gelman A., Robert C, Chopin N., Rousseau J.(2013), "Bayesian Data Analysis"