Estimate proportion of bicycle traffic

Garima Garg

Abstract

In this paper, our aim is to estimate θ, the underlying proportion of traffic that is bicycles. Two models are being used to estimate the proportion of bicycle traffic. First model takes θ as fixed - "true" proportion of traffic in all the blocks that is bicycles. The second model is hierarchical model in which θ_j vary with location - underlying proportion of traffic at location j that is bicycles. Model checking methods are used to compare between these two models.

KEY WORDS: Marginal Distribution, Posterior Predictive, Contour Plot

1. Introduction

A standard model assumes that the count of bicycles yi follows a Binomial distribution and the objective is to estimate the underlying proportion of traffic that is bicycles, θ. The fraction yi/ni is the count of bicycles per total number of vehicles and can be viewed as an estimate of proportion of traffic that is bicycles at location i. In Figure 1, I plot the ratios yi/ni against the total number of vehicles ni for all 10 city blocks in the residential streets with bike routes, where each point is labeled by the count of bicycles yi. The data shows that 16 out of 74 vehicles were in one of the residential street with bike route(Refer Table 1). It is natural to assume a binomial model for the count of bicycles. I used two models to estimate $/theta$, Model I and Model II. Both of them are defined in consecutive sections.

2. Model I - Analysis with equal proportion of bicycle traffic

The count of bicycles follow binomial distribution.

$$y_i|\theta \propto Bin(n_i, \theta)$$

From given data, I can see that $/theta$, the proportion of bicycle traffic follows beta distribution

$$\theta = Beta(\alpha, \beta)$$

.

Now I need to choose a non-informative hyperprior distribution on α, β. Before assigning a hyperprior distribution, I reparamaterize in terms of logit($\frac{\alpha}{\alpha+\beta}$)=log($\frac{\alpha}{\beta}$) and log($\alpha + \beta$),which are the logit of the mean and the logarithm of the 'sample size' in the beta distribution for θ. I use the logistic and logarithmic transformations to put each on a $(-\infty, \infty)$ scale.

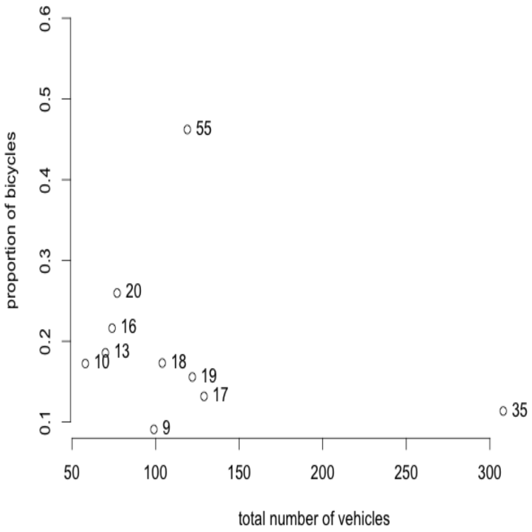

Figure 1: *Plot of proportion of bicycles against total number of vehicles for all the city blocks. Each point is labeled by the count of bicycles.*

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

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

Thus, assuming a Beta prior distribution for θ, and non-informative hyperprior distribution on α, β yields a Beta posterior distribution on θ.

$$p(\theta|\alpha, \beta, y) \sim Beta(\alpha + \sum_{i=1}^{10} y_i, \beta + \sum_{i=1}^{10} n_i - \sum_{i=1}^{10} y_i)$$

The Marginal posterior distribution on α, β takes this form

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \frac{B(\alpha + \sum_{i=1}^{10} y_i, \beta + \sum_{i=1}^{10} n_i - \sum_{i=1}^{10} y_i)}{B(\alpha, \beta)}$$

Now I have established a full probability model for data and parameters, and so I computed the marginal posterior distribution of the hyper parameters. I used grid approximation and set a grid in the range of log($\frac{\alpha}{\beta}$),log($\alpha + \beta$) $\epsilon[-3, 3] \times [-7, 12]$. To avoid computational overflows, I subtracted the maximum value of the log density from

Table 1: The table below gives the counts of bicycles and other vehicles in one hour in each of 10 city blocks in the residential streets with bike routes.

Type of Street	*Count of bicycles/other vehicles*
Residential	16/58, 9/90, 10/48, 13/57, 19/103, 20/57, 18/86, 17/112, 35/273, 55/64

each point on the grid and exponentiate, yielding values of the unnormalized marginal posterior density. Fig 2 shows a contour plot of the unnormalized marginal posterior density on a grid of values. Now I can sample from the numerically computed marginal posterior distribution using the sample function. Fig.3 shows the scatter plot of 10000 draws of $\log(\frac{\alpha}{\beta})$,$\log(\alpha+\beta)$. I can now transform the 10000 random draws to (α, β) scale. Now using these values, I can simulate 10000 draws of θ from its conditional posterior density. The 95% confidence interval of θ is [0.074,0.49].

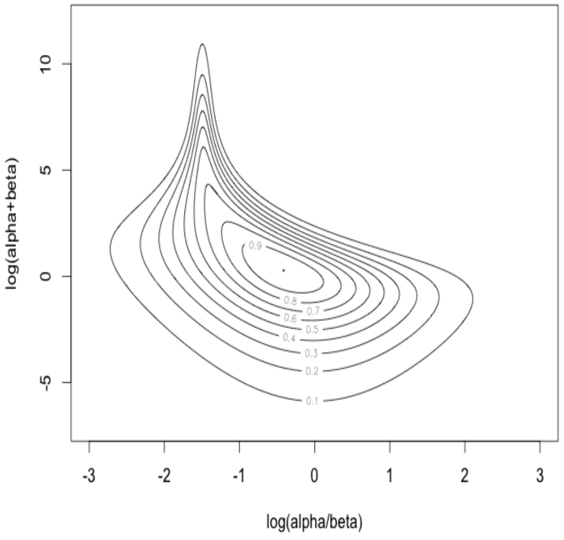

Figure 2: *Model I- Contour plot of the marginal posterior density of $\log(\frac{\alpha}{\beta})$,$\log(\alpha+\beta)$.*

2.0.1 *Model I-Posterior Predictive Model Checking*

For the posterior predictive check, consider y_8, and let's see if our model fits the data well or not. If the actual number of count of bicycles yi is in the middle of this predictive distribution, then I can say that our observation is consistent with our model fit. On the other hand, if the observed yi is in the extreme tails of the distribution ,then this observation indicates that the model is inadequate in fitting this observation. Fig.4 shows Histogram of simulated draws from the posterior predictive distribution of y8*. The actual number of count of bicycles is shown by a vertical line. If in an hour of observation, 100 vehicles of all kind go by, then 95% posterior predictive interval for the number of vehicles that are bicycles is [6,50].

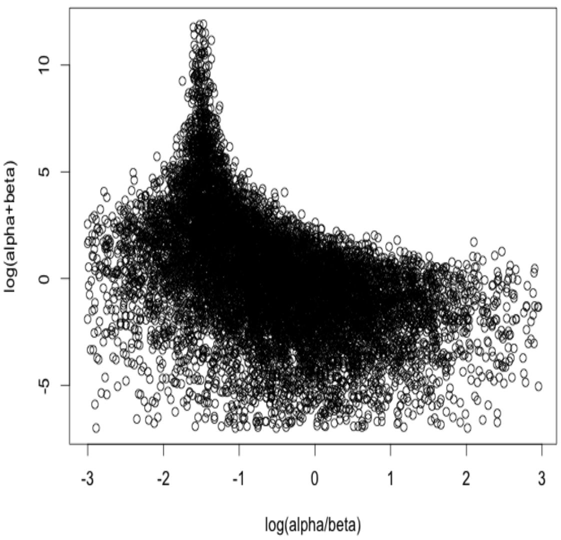

Figure 3: *Model II-Scatterplot of 10000 draws of $\log(\frac{\alpha}{\beta})$,$\log(\alpha+\beta)$ from the numerically computed marginal posterior density.*

3. Model II - Analysis using a Hierarchical Model

In the hierarchical model, I consider that θ, proportion of traffic that is bicycle vary because of different factors and this governs the bicycle traffic at any location j. This might be a possibility that one particular city block observes the minimum proportion of bicycle traffic at all hours due to variety of reasons, no school/college is in the vicinity of that particular city block etc. The count of bicycle follows the binomial distribution with mean $_j\theta_j$.

$$y_i|\theta \propto Bin(n_j, \theta_j)$$

Our prior distribution for θ_j in this model is also Beta Distribution. As given in the problem, I assumed same non-informative hyperprior in the previous part.

$$\theta_j = Beta(\alpha, \beta)$$

.

The joint posterior distribution of all parameters is

$$p(\theta, \alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}$$

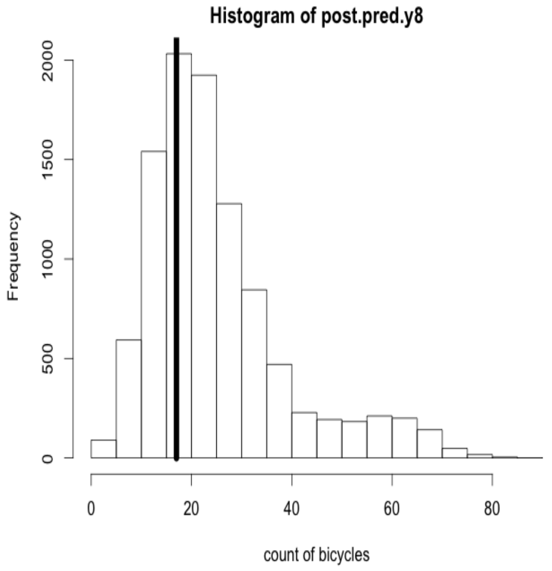

Figure 4: *Histogram of simulated draws from the posterior predictive distribution of y8*. The actual number of count of bicycles is shown by a vertical line..*

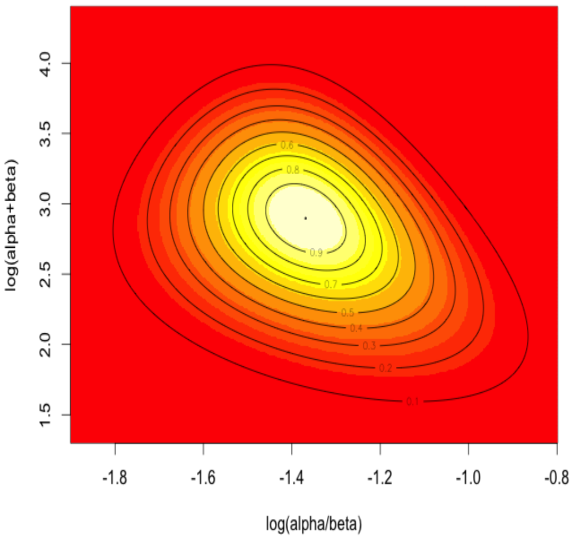

Figure 5: *Model II - Contour plot of the marginal posterior density of log($\frac{\alpha}{\beta}$),log($\alpha+\beta$) .*

$$\times \prod_{j=1}^{J} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}$$

The conditional posterior density of θ is

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

$$\times \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

The marginal posterior density of α, β can be determined by integrating out θ from joint posterior density

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

To put α, β on a $(-\infty,\infty)$ scale, I again transformed the hyperparameters to logit($\frac{\alpha}{\alpha+\beta}$)=log($\frac{\alpha}{\beta}$) and log($\alpha+\beta$). I used grid approximation and set a grid in the range of log($\frac{\alpha}{\beta}$), log($\alpha+\beta$) $\epsilon[-1.9,-0.8]\times$[1.3,4.4]. Figure 5 shows the contour plot of the unnormalized marginal posterior density. Now I can sample from the numerically computed marginal posterior distribution, and then using these random draws of hyperparameters, simulate the values of θ_j. I transformed the log($\frac{\alpha}{\beta}$), log($\alpha+\beta$) scale to (α,β) scale. The approximated value of (α,β)=(3.7,14.0). Figure 6 shows the scatterplot of 10000 draws of (α,β). Using the values of hyper parameters, I can make inferences about the θ posterior. Table 2 gives the 95% confidence interval for all of the θ. Fig 7 shows the posterior means against the values of y/n, to assess shrinkage.

Table 2: Model-II 95% posterior interval for θ.

θ	5%	95%
θ_1	0.14679911	0.2868546
θ_2	0.06319828	0.1591639
θ_3	0.11087149	0.2564617
θ_4	0.12352644	0.2611295
θ_5	0.11225471	0.2153299
θ_6	0.17786850	0.3261577
θ_7	0.12347541	0.2385467
θ_8	0.09523349	0.1890420
θ_9	0.09092846	0.1499693
θ_{10}	0.35624388	0.5045657

3.0.2 Model II-Posterior Predictive Model Checking

For the posterior predictive check, consider y_9, and let's see if our model fits the data well or not. If the actual number of count of bicycles yi is in the middle of this predictive distribution, then I can say that our observation is consistent with our model fit. On the other hand, if the observed yi is in the extreme tails of the distribution ,then this observation indicates that the model is inadequate in fitting this observation. Fig.8 shows Histogram of simulated draws from the posterior predictive distribution of y9*. The actual number of count of bicycles is shown by a vertical line.
If in an hour of observation, 100 vehicles of all kind go by, then 95% posterior predictive interval for the number of vehicles that are bicycles is [12, 28].

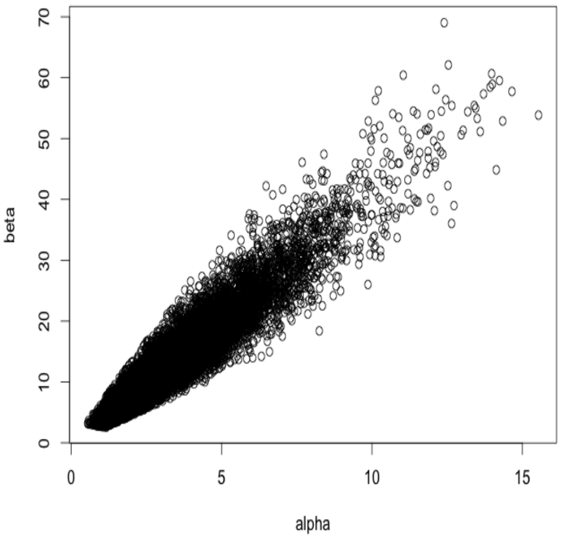

Figure 6: *Model II-Scatter plot of 10000 draws of alpha, beta from the numerically computed marginal posterior density.*

4. Model Comparison

Model I is definitely better than Model II in terms of posterior predictive checks. Model I better fits the percentages of estimated θ, future valyes of count of bicycles.

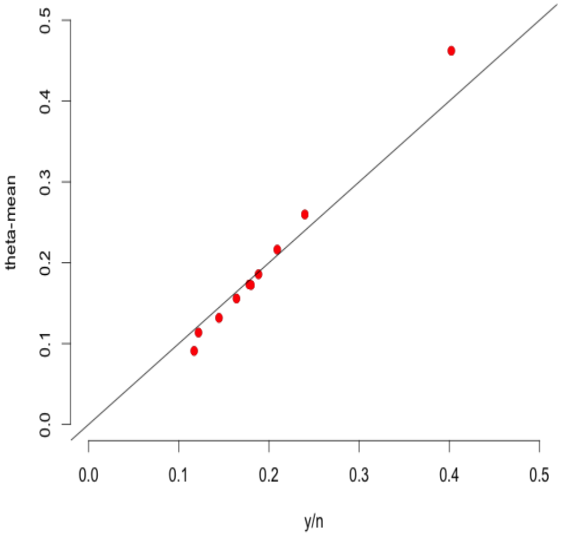

Figure 7: *Model II-Posterior means to assess shrinkage.*

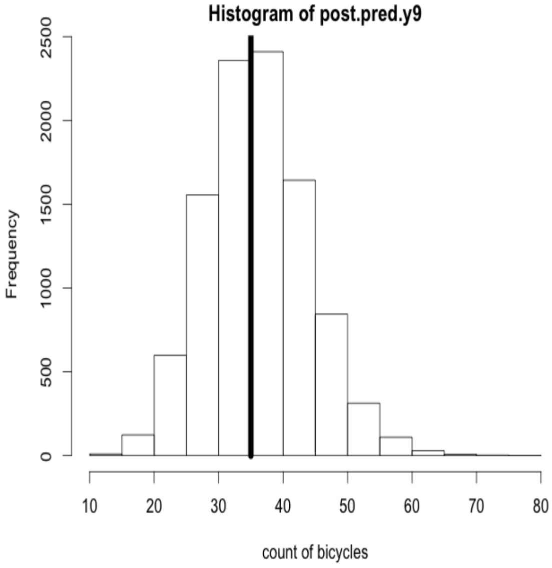

Figure 8: *Model II - Histogram of simulated draws from the posterior predictive distribution of y9*. The actual number of count of bicycles is shown by a vertical line*