STATS 551 Homework 2

Multi-Parameter Models & Hierarchical Models

Due date: 6:00 pm (EST) Feb. 19, 2018

Analysis of proportions. A survey was done of bicycle and other vehicular traffic in the neighborhood of the campus of the University of California, Berkeley, in the spring of 1993. Sixty city blocks were selected at random; each block was observed for one hour, and the numbers of bicycles and other vehicles traveling along that block were recorded. The sampling was stratified into six types of city blocks: busy, fairly busy, and residential streets, with and without bike routes, with ten blocks measured in each stratum. Table 3.3 displays the number of bicycles and other vehicles recorded in the study.

Type of street	Bike route?	Counts of bicycles/other vehicles
Residential	yes	16/58, 9/90, 10/48, 13/57, 19/103, 20/57, 18/86, 17/112, 35/273, 55/64
Residential	no	12/113, 1/18, 2/14, 4/44, 9/208, 7/67, 9/29, 8/154
Fairly busy	yes	8/29, 35/415, 31/425, 19/42, 38/180, 47/675, 44/620, 44/437, 29/47, 18/462
Fairly busy	no	10/557, 43/1258, 5/499, 14/601, 58/1163, 15/700, 0/90, 47/1093, 51/1459, 32/1086
Busy	yes	60/1545, 51/1499, 58/1598, 59/503, 53/407, 68/1494, 68/1558, 60/1706, 71/476, 63/752
Busy	no	8/1248, 9/1246, 6/1596, 9/1765, 19/1290, 61/2498, 31/2346, 75/3101, 14/1918, 25/2318

Table 3.3 *Counts of bicycles and other vehicles in one hour in each of 10 city blocks in each of six categories. (The data for two of the residential blocks were lost.) For example, the first block had 16 bicycles and 58 other vehicles, the second had 9 bicycles and 90 other vehicles, and so on. Streets were classified as 'residential,' 'fairly busy,' or 'busy' before the data were gathered.*

1. For this part, restrict your attention to the first four rows of the table: the data on residential streets. (4×10 points)

   (a) Let $y_1, \ldots, y_{10}$ and $z_1, \ldots, z_8$ be the observed proportion of traffic that was on bicycles in the residential streets with bike lanes and with no bike lanes, respectively (so $y_1 = 16/(16 + 58)$ and $z_1 = 12/(12 + 13)$ for example). Set up a model so that the y_i's are independent and identically distributed given parameters θ_y and the z_i's are independent and identically distributed given parameters θ_z.

   (b) Set up a prior distribution that is independent in θ_y and θ_z.

   (c) Determine the posterior distribution for the parameters in your model and draw 1000 simulations from the posterior distribution. (Hint: θ_y and θ_z are independent in the posterior distribution, so they can be simulated independently.)

   (d) Let $\mu_y = E(y_i|\theta_y)$ be the mean of the distribution of the y_i's; μ_y will be a function of θ_y. Similarly, define μ_z. Using your posterior simulations from 1c, plot a histogram of the posterior simulations of $\mu_y - \mu_z$, the expected difference in proportions in bicycle traffic on residential streets with and without bike lanes.

2. For this problem, restrict your attention to the first two rows of the table: residential streets labeled as 'bike routes,' which we will use to illustrate this computational exercise. (6×10 points)

   (a) Set up a model for the data in Table 3.3 so that, for $j = 1, \ldots, 10$, the observed number of bicycles at location j is binomial with unknown probability θ_j and sample size equal to the total number of vehicles (bicycles included) in that block. The parameter θ_j can be interpreted as the underlying or 'true' proportion of traffic at location j that is bicycles. Assign a beta population distribution for the parameters θ_j and a noninformative hyperprior distribution as in the rat tumor example of Section 5.3 of BDA. Write down the joint posterior distribution.

   (b) Compute the marginal posterior density of the hyperparameters and draw simulations from the joint posterior distribution of the parameters and hyperparameters.

   (c) Compare the posterior distributions of the parameters θ_j to the raw proportions, (number of bicycles / total number of vehicles) in location j.

How do the inferences from the posterior distribution differ from the raw proportions?

(d) Give a 95% posterior interval for the average underlying proportion of traffic that is bicycles.

(e) A new city block is sampled at random and is a residential street with a bike route. In an hour of observation, 100 vehicles of all kinds go by. Give a 95% posterior interval for the number of those vehicles that are bicycles. Discuss how much you trust this interval in application.

(f) Was the beta distribution for the θ_j's reasonable?

Guideline for Submission: Submit R markdown (or jupyter notebook) with annotated code followed by results. Discussions about the results should follow the results.

Optional Reading. Read one of the following papers and post your summary and thoughts on Canvas. Bonus points up to 5 will be rewarded.

1. The selection of prior distributions by formal rules, Kass, R. E, and Wasserman, L. (1996), Journal of the American Statistical Association 91, 1343-1370.

2. Parameterization and Bayesian modeling, Gelman, A. (2004), Journal of the American Statistical Association 99, 537-545.