

Advanced Bayesian Modeling

Data Analysis Report

You will submit a PDF file containing your data analysis report, which must follow the format described below.

Important: You may not collaborate or discuss your analysis with anyone else. Plagiarism from *any* source is an academic integrity infraction.

Scenario: Rivera & Rosenbaum (2020)¹ discuss data evidence for racial bias in police stops in two US cities. One assessment method they use is the *outcome test*, in which the proportion of police searches (after a stop) that find contraband is compared by race of the subject searched. If subjects of a US minority race are less likely to be found with contraband after a search, relative to white subjects, racial bias may have influenced the police decision to stop or to conduct the search.

Data file `policesearchSanFrancisco.csv` contains data aggregated from original data files used by Rivera & Rosenbaum (2020). It records frequencies and outcomes of police searches conducted in San Francisco from January 1, 2015, through June 30, 2016. Each row represents a combination of the searched subject's race and the reporting police district. The columns are as follows:

<code>SubjectRace</code>	racial grouping of searched subject (API=asian/pacific islander)
<code>District</code>	designation for police district reporting the search
<code>ContrabandFound</code>	number of searches (out of the total) in which contraband was found
<code>TotalSearches</code>	total number of searches conducted (subsequent to police stops)

Use JAGS and R software, and use only the data in `policesearchSanFrancisco.csv`. JAGS code should be included in the appropriate sections, but **all R code and any direct R text output listings you choose to include should be in the Appendix only.**

Your report must be neatly typed and can be at most **8 pages**, excluding the Appendix. It must follow this outline:

1. **Introduction** Provide brief background information about police stops and searches in the US, and the issue of racial bias in US policing. (Use footnotes to acknowledge all sources you consult, including web sites.) *Do not plagiarize!*
2. **Data** Briefly describe the variables in `policesearchSanFrancisco.csv`. For each subject racial group (including `Other`), produce a boxplot of the district-level raw proportions of searches that find contraband. (Omit any proportions that cannot be calculated.) Display all five boxplots side-by-side on the same graph (same axis), so they can be compared.²

Answer the following questions:

¹Rivera, R., & Rosenbaum, J. (2020, August). Racial disparities in police stops in US cities. *Significance*, 17(4), 04–05.

²Consider using the R function `boxplot`. Consult its R help file for assistance.

- In the data, which race/district combinations have no searches?
 - For which racial group are the proportions of searches that find contraband generally the largest?
 - For which racial group are the proportions of searches that find contraband generally the smallest?
3. **First Model** You will use the JAGS model in the file named `firstmodel.bug`. The data-related nodes are as follows:
- **found**: a node array containing the numbers of searches that resulted in finding contraband (for each race/district combination)
 - **searches**: a node array containing the total numbers of searches conducted (for each race/district combination)
 - **race**: a node array, each element containing an integer index from 1 to 5 indicating the race group of the subject (for each race/district combination)
 - **district**: a node array, each element containing an integer index from 1 to 12 indicating the reporting district (for each race/district combination)

Carefully set up the R data structure that you will pass to JAGS.³⁴ Then run your analysis (being careful to follow the usual procedures) and report as follows:

- (a) Describe the model in `firstmodel.bug`. Make sure that the following questions are answered by your description:
 - What type of (generalized) linear model is this? What does the response variable represent?
 - What are the parameters and hyperparameters?
 - Are the racial group parameters treated as if they are fixed effects or random effects? What about the district parameters?
- (b) List the JAGS code in `firstmodel.bug`.
- (c) Summarize the details of your computation, including number of chains, length of burn-in, number of iterations used per chain, any thinning (if used), and effective sample sizes of the top-level parameters. You should use plots to check convergence, but do *not* include them in your report.

Note: Use overdispersed starting values, but make them less extreme if you encounter convergence problems.
- (d) Graph an approximate posterior density (*not* a histogram) for `sigmadistrict`. Does your graph suggest that there are actual differences among the districts (in terms of the probability of a search finding contraband)?

³You can convert `SubjectRace` and `District` to factor variables in R, if they are not so already. Then you can convert each factor variable to an integer index by applying the `unclass` function. When interpreting results, it is up to you to figure out which factor level corresponds to which integer index.

⁴You may omit any row of the data set for which the total number of searches is zero, since such rows will not contribute to the likelihood function. While you can run the JAGS model even without omitting those rows, omitting them may make the rest of your analysis easier to perform.

- (e) Let β_B be the coefficient associated with the subject race being Black and β_W the coefficient associated with the subject race being White. Briefly explain why $\beta_B < \beta_W$ would indicate that contraband is less likely to be found in a search of a Black subject than of a White subject (within a given district). Then approximate the posterior probability that $\beta_B < \beta_W$. What do you conclude?
 - (f) Check the model for overdispersion: Approximate the posterior predictive p -value based on using the chi-square discrepancy. What do you conclude?
 - (g) Approximate the value of (Plummer's) DIC and its associated effective number of parameters. Compare the effective number of parameters with the actual (total) number of parameters (including hyperparameters).
4. **Second Model** Starting with the JAGS model in `firstmodel.bug`, create an extended JAGS model that can account for overdispersion:

- Add a random effect term ϵ_i to the linear portion of the model (where i ranges over the observations, i.e., the rows of the data set).
- Under the prior, let the random effects ϵ_i be (conditionally) independent and have the same normal distribution: one that has mean 0 and variance σ_ϵ^2 .
- Let the hyperprior for σ_ϵ (*not* σ_ϵ^2) be uniform from 0 to 10.
- Do not change anything related to the other aspects of the model.

Run your analysis (being careful to follow the usual procedures) and report as follows:

- (a) List all of the JAGS code for your extended model.
 - (b) Summarize the details of your computation, including number of chains, length of burn-in, number of iterations used per chain, any thinning (if used), and effective sample sizes of the top-level parameters. You should use plots to check convergence, but do *not* include them in your report.
Note: Use overdispersed starting values, but make them less extreme if you encounter convergence problems.
 - (c) As you did for the previous model, consider the proposition that contraband is less likely to be found in a search of a Black subject than of a White subject (within a given district). Approximate the posterior probability of this for your new model. Does your conclusion change?
 - (d) Approximate the value of (Plummer's) DIC and its associated effective number of parameters. Is your second model better than the first, according to DIC?
5. **Conclusions** Briefly summarize your results in a non-technical manner.
6. **Appendix** Provide the R code you used to conduct your analysis. Include comments that label the purpose of each block of code.

NOTES:

- Comma-separated variable (`.csv`) files can be read into R with `read.csv`.

- Effective sample sizes of at least 2000 are recommended for accuracy.
- If your computer runs out of memory, consider using thinning (e.g., the `thin` argument of `coda.samples`).

POINT ALLOCATIONS

Specifications	2	neatly typed
	2	no more than 8 pages (excluding Appendix)
Introduction	4	background given
	1	sources acknowledged
Data	1	description of variables
	2	boxplots
	3	questions
First Model	6	(a)
	1	(b)
	4	(c)
	2	(d)
	3	(e)
	3	(f)
	3	(g)
Second Model	4	(a)
	4	(b)
	2	(c)
	3	(d)
Conclusions	3	brief, clearly stated, appropriate summary of results
Appendix	2	all R code present
	2	comments for different blocks of code
<hr/>		
Total:	57	