STAT 578 – Advanced Bayesian Modeling – Fall 2018

Data Analysis Report

DUE: December 16, 2018

You will submit a data analysis report following the description below.

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

Scenario: The data file `party115cong.csv` contains data on all congressional districts[1] represented in the U.S. House of Representatives during the 115th U.S. Congress. Each row represents a district, and the columns are as follows:

`state`	the U.S. state containing the district, or the District of Columbia
`district`	an identifier of Congressional district within `state`
`electedrep`	name of person elected to the 115th Congress
`party`	the party affiliation of the Representative during the 115th Congress: `D` for Democrat, `R` for Republican
`medHouseIncome`	median household income (dollars)

Use JAGS and R software, and use only the data in `party115cong.csv`. JAGS code should be included in the appropriate sections, but all R code and any direct R 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 the U.S. House of Representatives, congressional districts, the 115th U.S. Congress, and U.S. political parties. (Use footnotes to acknowledge any sources you consult, including web sites.)

2. **Data** Briefly describe and summarize the variables in `party115cong.csv`, as appropriate. Plot a histogram of `medHouseIncome`.

3. **First Model** You will fit a Bayesian logistic regression model to explain `party` based on the *natural logarithm* of `medHouseIncome`:

   - The response will be Bernoulli: 1 if Democrat (`D`), 0 if Republican (`R`). You will need to create a variable having this coding.
   - The model will be a logistic regression.
   - The linear portion of the model will be almost like a simple linear regression: There will be an ordinary "intercept" term and a "slope" multiplying the (centered and rescaled) `log(medHouseIncome)`, but, of course, no "error" term.

[1] For the purposes of this assignment, the federal district of Washington, D.C. is regarded as a congressional district.

- The independent variable is a centered and rescaled version of `log(medHouseIncome)`: centered to have sample mean of zero, and rescaled to have sample standard deviation of 0.5 (*not* 1), as recommended in BDA3. Note that the centering and rescaling should be *after* taking the natural logarithm.
- As recommended in BDA3, the prior for the "intercept" should be $t_1(0, 10^2)$, and the prior for the "slope" should be $t_1(0, 2.5^2)$ (and these should be independent). Note: These are expressed in BDA3 notation. Be careful when converting to JAGS code.

This first model will *not* use `state` (or any of the other variables).

- For the Bernoulli, use the `dbern` distribution specifier in JAGS.
- You may wish to consult the JAGS manual to make sure that you are correctly using the `dt` distribution specifier (for a t distribution).

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

(a) List your JAGS 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 all parameters. Do not include plots.

(c) Approximate the posterior mean, posterior standard deviation, and 95% central posterior interval for each parameter.

(d) Approximate the posterior probability that the "slope" exceeds zero. Interpret this result. (What apparently happens to the probability of electing a Democrat as median household income increases?)

(e) Approximate the value of (Plummer's) DIC and the associated effective number of parameters. Compare the effective number of parameters with the actual number of parameters.

4. **Second Model** Now extend your first model by allowing each state to have a separate additive random effect:

- Create an indexing variable in which the variable `state` is recoded with the integers 1 to 51. (Refer to the Ebola data example in Week 14.)
- Starting with the first model (as described above), add to the linear portion of the model a random effect term that varies by state. (For comparison, consider the term `betavirus[virus[i]]` in the JAGS model for the Ebola data example in Week 14.)
- Let the prior for these random effects be (conditionally) independent from a *normal* distribution with mean zero (since the model already has an intercept) and *standard deviation* σ_{state}.
- Let the prior for σ_{state} be approximately flat. (You need to determine how to implement this. It may require a preliminary run and some adjustment.)

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

(a) List your JAGS 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. Do not include plots.

(c) Approximate the posterior mean, posterior standard deviation, and 95% central posterior interval for the "intercept" and for the "slope" coefficient related to median household income.

(d) Approximate the posterior probability that the "slope" exceeds zero. Interpret this result. (What apparently happens to the probability of electing a Democrat as median household income increases, after adjustment for state?)

(e) Which state has the largest (in the positive direction) posterior mean random effect? Which state has the smallest (in the negative direction) posterior mean random effect? Interpret in terms of the apparent party preferences of the two states (after adjustment for median household income).

(f) Approximate the value of (Plummer's) DIC and the associated effective number of parameters. Is this second model better than the first?

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		
	2	description/summary of variables
	1	histogram
First Model		
	5	(a)
	4	(b)
	3	(c)
	2	(d)
	3	(e)
Second Model		
	5	(a)
	4	(b)
	3	(c)
	2	(d)
	3	(e)
	3	(f)
Conclusions		
	3	brief, clearly stated, appropriate summary of results
Appendix		
	2	all R code present
	2	comments for different blocks of code
Total:	56	