STAT 578 – Advanced Bayesian Modeling – Fall 2019

Data Analysis Report

DUE: December 15, 2019

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: The data file `marijuanause.csv` contains survey data on reported marijuana use during adolescence in a sample of 236 people of approximately the same age.[1] Each row represents a person in the sample, and the columns are as follows:

`female`	an indicator that the person is female (1 if so, 0 if not) (All people in the survey were either female or male.)
`use1976`	an indicator that the person used marijuana in the year 1976 (1 if so, 0 if not)
`use1977`	an indicator that the person used marijuana in the year 1977 (1 if so, 0 if not)
`use1978`	an indicator that the person used marijuana in the year 1978 (1 if so, 0 if not)
`use1979`	an indicator that the person used marijuana in the year 1979 (1 if so, 0 if not)
`use1980`	an indicator that the person used marijuana in the year 1980 (1 if so, 0 if not)

Use JAGS and R software, and use only the data in `marijuanause.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 marijuana and its use, especially during adolescence. (Use footnotes to acknowledge any sources you consult, including web sites.) *Do not plagiarize!*

2. **Data** Briefly describe and *statistically* summarize the variables in `marijuanause.csv`. Make two plots of percentage of use versus year: one plot for females and the other for males. Clearly and accurately label the axes of both plots.

3. **First Model** You will fit a (univariate) Bayesian logistic regression model to explain marijuana use based on the year and on whether or not the person is female:

    - The response variable will be Bernoulli: 1 if marijuana was used, 0 if not. Note that there is one response value for each combination of person and year, for a total of 1180 observations. (Rows of the data set correspond to individual people, not individual observations.)

[1] Based on a data set from Julian Faraway (2016). *faraway: Functions and Datasets for Books by Julian Faraway.* R package version 1.0.7. `https://CRAN.R-project.org/package=faraway`

- The model will be a logistic regression.
- The linear portion of the model will be almost like a linear regression: There will be an ordinary "intercept" term, a coefficient multiplying the (centered and rescaled) year number, and a coefficient multiplying the (centered) indicator of being female. (Of course, there is no "error" term.) None of the parameters depends on person, i.e., the same "intercept" and the same coefficients apply to every person.
- The independent variables are:
  - a centered and rescaled version of the year number: centered to have sample mean of zero, and rescaled to have sample standard deviation of 0.5 (*not* 1), as recommended in BDA3.
  - a centered (but *not* rescaled) version of the indicator variable for being female
- As recommended in BDA3, the prior for the "intercept" should be $t_1(0, 10^2)$, the priors for the other coefficients should be $t_1(0, 2.5^2)$, and these should all be independent. Note: These distributions are expressed in BDA3 notation. Be careful when converting to JAGS code.

This first model will *not* use any random effect terms. All responses will be regarded as (conditionally) independent under the sampling model, even if the responses come from the same person.

- Since the response values are stored in a two-way format (by person and by year), consider using nested `for` structures in your JAGS model.
- 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 code.

(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. You should use plots to check convergence, but do *not* include them in your report.

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

(d) Approximate the posterior probability that the coefficient for the (centered) indicator of being female exceeds zero. Interpret this result. (For a given year, does it appear that females are more likely to use marijuana than males? Less likely? About as likely?)

(e) For each year (1976 through 1980), approximate the *posterior expected probability* (according to the model) that a newly sampled *female* person would use marijuana in that year. Do the same for a newly sampled *male* person. Plot these probabilities versus year, separately for females and males. (Make sure to include clear and accurate labels.)

(f) 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 person to have a separate additive random effect:

   - Starting with the first model (as described previously), add to the linear portion of the model a random effect term that varies by person, i.e., is the same for all responses from the same person but is different for different people.
   - 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 common *standard deviation* σ_{person}.
   - Let the hyperprior for σ_{person} be approximately flat on $(0, \infty)$. (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 code.

   (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) Approximate the posterior mean, posterior standard deviation, and 95% central posterior interval for the "intercept," for the coefficient of the (centered and rescaled) year number, for the coefficient of the (centered) indicator of being female, and for σ_{person}.

   (d) 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 and summary of variables
	2	separate plots
First Model		
	6	(a)
	4	(b)
	3	(c)
	2	(d)
	4	(e)
	3	(f)
Second Model		
	5	(a)
	4	(b)
	4	(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
Total:	58	