**STAT 578 – Advanced Bayesian Modeling – Spring 2020**

# Data Analysis Report

**DUE: May 10, 2020**

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 `EUCOVIDdeaths.csv` contains data from the European Center for Disease Prevention and Control (ECDC) on daily deaths related to COVID-19 in the 27 European Union (EU) member states for the second half of March 2020.[1] Each row represents an EU country, and the columns are as follows:

| | |
|---|---|
| `Country` | name of the country (member state) |
| `PopulationM` | 2018 population of the country, in millions |
| `Mar16`–`Mar31` | number of COVID-19 deaths recorded in the country for that date |

Use JAGS and R software, and use only the data in `EUCOVIDdeaths.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 COVID-19 and its spread, especially in the EU during the second half of March. (Use footnotes to acknowledge all sources you consult, including web sites.) *Do not plagiarize!*

2. **Data** Briefly describe the variables in `EUCOVIDdeaths.csv`. Create two segmented line time-series plots: The first will show number of deaths versus day (16 to 31), with a separate segmented line for each country. The second will be the same, except plotting number of deaths per million inhabitants of the country (i.e., death rate per capita). Clearly and accurately label the axes of both plots. Discuss the results, answering the following questions:

   - Which country has the most deaths over this period?
   - Which country has the highest death rate per capita over this period?
   - Which countries have *no* deaths over this period?

3. **First Model** You will use the JAGS model in the file named `firstmodel.bug`. The data-related nodes are as follows:

[1] Accessed April 22, 2020 from `https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide`

- `deaths[i,j]`: number of COVID-19 deaths recorded in country $i$ ($i = 1, \ldots, 27$) on day $j$ ($j = 1$ is March 16, $j = 2$ is March 17, etc.)
- `logpopulation[i]`: *natural logarithm* of the 2018 population of country $i$, where population is expressed in millions
- `daycent[j]`: day index *centered* so that this variable has an average of zero (but is *not* standardized); the difference between `daycent[j+1]` and `daycent[j]` should equal 1

Carefully set up the data structure that you will pass to JAGS. 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?
   - Is it a rate model? What would be the rate (exposure) variable?
   - What are the parameters and hyperparameters?
   - For the linear portion of the model, does the intercept vary by country? Does the slope (multiplying the centered day variable) vary by country?

(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 all 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) Approximate the posterior mean, posterior standard deviation, and 95% central posterior interval for each top-level (hyper)parameter.

(e) Which country has the *highest* posterior *median* value for its intercept? Which country has the *lowest* posterior *median* value for its intercept? (Remark: These are the EU countries with the highest and lowest median per capita death rates for the second half of March, according to this model.)

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

4. **Second Model** Starting with the JAGS model in `firstmodel.bug`, create an extended JAGS model that allows each country to have a separate slope (multiplying the centered day variable):

   - Each country should have its own slope parameter.
   - Under the prior, the slope parameters should be (conditionally) independent, and from a common *normal* distribution with a mean $\mu_{\text{slope}}$ and a *standard deviation* $\sigma_{\text{slope}}$.
   - Let the hyperprior for $\mu_{\text{slope}}$ be $\text{N}(0, 100^2)$, let the hyperprior for $\sigma_{\text{slope}}$ be $\text{U}(0, 100)$, and let these be independent under the prior.

- Do not change anything related to the model for the intercepts. The hyperparameters for the slopes and the hyperparameters for the intercepts should be independent under the prior.

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) Graph an approximate posterior density (*not* a histogram) for $\sigma_{\text{slope}}$. Does it suggest that there are actual differences among the slopes of different countries?

(d) For each country, compute an approximate posterior expected intercept and an approximate posterior expected slope. Plot these pairs on a scatterplot of slope (vertical axis) versus intercept (horizontal axis). Instead of the usual plotting symbol, use the *name of the country* to plot the point for each country. For example, for Malta, plot the name "Malta" centered at the coordinates for its posterior expected intercept and slope. (See NOTES for some tips.)

(e) 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`).
- In R, one way to plot using names rather than symbols is to use the `plot` function with argument `type="n"`, and then use the `text` function. For example, consider this:

```
plot(1:3, c(1,3,2), type="n")
text(1:3, c(1,3,2), c("one","two","three"), cex=0.8)
```

You may adjust `cex` to reduce overlap between names.

## 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 | separate plots |
| | 3 | discussion of results |
| First Model | | |
| | 4 | (a) |
| | 1 | (b) |
| | 4 | (c) |
| | 3 | (d) |
| | 2 | (e) |
| | 3 | (f) |
| Second Model | | |
| | 4 | (a) |
| | 4 | (b) |
| | 2 | (c) |
| | 3 | (d) |
| | 3 | (e) |
| Conclusions | | |
| | 3 | brief, clearly stated, appropriate summary of results |
| Appendix | | |
| | 2 | all R code present |
| | 2 | comments for different blocks of code |
| Total: | 55 | |