Bayesian modeling exercises

Johannes Karreth

Applied Introduction to Bayesian Data Analysis

All exercises in this tutorial can be completed using the code shown during the workshop. Other code examples can also be found on my github page at `https://github.com/jkarreth/Bayes`.

If you have any questions working through these exercises, please don't hesitate to contact me.

1 Linear model

Write code for and run a linear model in JAGS using the data file `http://www.jkarreth.net/files/exercise1.csv`.

The outcome variable is `vote`, the vote share for the incumbent party in 15 U.S. presidential elections. The explanatory variables are `gnp`, the change in GNP from the previous year, and `approval`, the approval rate of the incumbent in July of the election year. (You can consider the data points independent of each other.)

Play around with different priors for the betas and compare/contrast the different results. Specifically, fit 10 different models; in each model, vary the (a) information you insert via the prior and (b) the degree of certainty around your prior. If differing priors lead to different results, interpret and discuss why this is the case. Provide a brief summary of the findings from your model, including the degree of certainty with which you make your conclusions about the model parameters (and what informs your un/certainty).

2 Debugging model code

Debug code for and fit a linear model in JAGS using the data file `http://www.jkarreth.net/files/exercise2.dta` and the BUGS model `http://www.jkarreth.net/files/exercise2.mod.txt`.

This dataset is a random sample of a public health database of young males. The outcome variable is `wgt`, the weight of each individual in the database. The explanatory variables are `age`, the individual's age, `hgt`, the individual's height, and `hc`, the individual's head circumference. (You can consider the data points independent of each other.)

The data are in Stata format. For your homework, perform the following steps:

- Prepare the data for use with JAGS.
- Inspect the data to check for potential problems you may need to fix before fitting a Bayesian model.
- Inspect the model code in the text file `exercise2.mod.txt` to check for potential problems you may need to fix before fitting a Bayesian model.
- Fit a Bayesian model, present the debugged model code, and report one single table summarizing posterior estimates—demonstrating that you successfully debugged the provided code and were able to fit the model.

3 Logistic regression model

Using the file `http://www.jkarreth.net/files/exercise3.csv`, write code and fit a Bayesian logit model. Be sure to assess convergence and supply a brief interpretation of the model results. Present your results as you would in a scholarly article: make a results table with the important quantities of interest (posterior estimates, credible intervals), and also convert the logit coefficients into "something more interpretable" of your choice.

The dataset is a cleaned up and modified version of the 1996 American National Election studies as used in Hainmueller and Hiscox (2006). This study investigated the determinants of individuals' support for free trade or protectionist policies. The outcome variable is a dummy `protectionist`—coded as 1 if a respondent expressed a preference for more protectionist policies, and coded as 0 if a respondent favored free trade. The explanatory variables are (see Table A2 in Hainmueller and Hiscox (2006) for more details):

- `age`: the incumbent's age.
- `female`: a binary indicator for female respondents.
- `TUmember`: a binary indicator for trade union members.
- `partyid`: the respondent's party identification: coded from 0 "strong Democrat" to 6 "strong Republican".
- `ideology`: the respondent's ideology: coded 0 if conservative, 1 if moderate, and 2 if liberal.
- `schooling`: years of full-time education completed.

You will need to import the data into R. You may also need to recode or transform some variables.

4 Factor analysis

Estimate a Bayesian factor model on a dataset with missing observations on some of the observed variables. You will be using an example dataset from UCLA's Stata tutorial on factor analysis.[1] You can read the dataset into R or Stata using the following commands:

- Stata: `use http://www.ats.ucla.edu/stat/stata/output/m255`
- R: `dat <- read.dta("http://www.ats.ucla.edu/stat/stata/output/m255.dta")`

The dataset contains information on students' teaching evaluations of college instructors. We focus only on a subset of items; this subset covers the following 12 items:[2]

- `item13`: Was the instructor well prepared?
- `item14`: Instructor's scholarly grasp
- `item15`: Instructor's confidence
- `item16`: Instructor is focused in lectures
- `item17`: Instructor uses clear relevant examples
- `item18`: Instructor is sensitive to students
- `item19`: Instructor allows me to ask questions
- `item20`: Instructor is accessible to students outside class
- `item21`: Instructor is aware of students understanding the material
- `item22`: I am satisfied with instructor's student performance evaluation

[1] See `http://www.ats.ucla.edu/stat/stata/output/fa_output.htm`.

[2] In this dataset you will be using variables that are on a scale from 1 to 5, so you won't need to rescale any of the variables.

- `item23`: Compared to other instructors, this instructor is good/better
- `item24`: Compared to other courses this course was good/better

Do these items all capture one latent concept? Briefly conduct a (frequentist) factor analysis in your preferred software (e.g., Stata: `help factor`; R: `fa.parallel` in the `psych` package) to check whether one or more factors are best used to cover the latent variable/s underlying these items. If you don't have much time, a simple correlation matrix of these 12 variables might also provide sufficient information to answer this question.

Next, write a Bayesian factor model to estimate a posterior distribution for the factor/s you would like to extract from these items. You can use the example from the workshop as a foundation. Be sure to use the information from your frequentist factor model to identify the factor model (hint: should the priors for the coefficients be truncated to be positive or negative?).

Start with few iterations (no more than 10 or 100), as iterations in this model will take longer than those in the previous assignments.

After estimating the latent variable, bring the mean of the posterior distribution of the latent variable for each observation back into your original dataset and report the correlation between the mean posterior latent variable and the factor you estimated in your frequentist model. You can also make a scatterplot of the "frequentist" against the "Bayesian" factor.

Finally, write one or two sentences about what advantages you might have gained from estimating a Bayesian vs. frequentist factor model, and what you could use this "Bayesian" latent factor for.

5 Item Response model

In this assignment, you will fit a simple 2-parameter Bayesian IRT model to estimate the latent ability of 500 students who took the TIMSS test. The TIMSS (Trends in International Mathematics and Science Study) tests students' ability in mathematics. This dataset is an excerpt and contains 34 questions that were answered by 500 students. Each question was either answered correctly (1) or incorrectly (0). The data you will use are at `http://www.jkarreth.net/files/exercise5.dta`.

Using the technique you learned in the workshop, fit a 2-parameter IRT model on these data. Use this model for two tasks:

1. Describe how useful each of the 34 items is in evaluating students' mathematical ability. If you're short on time, pick 5 random items and write how you evaluated their "usefulness."
2. List which students (identified by their row number in the original dataset) are in the top 5% of this sample in terms of their mathematical ability.

6 Ordered logit model

Use the file `http://www.jkarreth.net/files/exercise6.dta`.[3] Be sure to convert string variables to numeric variables and to delete cases with missing values (R hint: `na.omit()`).

Your choice of initial values will be important.

Fit the model, assess convergence and supply a brief interpretation of the results. *Be careful*: Most of these model specifications might take some time to converge, so start with a small number of iterations (only 10 or 20). If you have time, please present predicted probabilities or other useful quantities of interest.

Conventional wisdom among scholars of interest groups in American politics states that a primary goal of groups is to develop and maintain access to policy makers. While much of this work has

[3] Taken from a homework assignment in Chris Zorn's MLE course at a previous ICPSR summer program.

focused on groups' ties to members of Congress, sometimes equally important is the extent to which groups cultivate connections within executive and regulatory agencies.

Here, you will examine the causes of group access to federal agencies. The data are from a 1985 survey by the late Jack Walker of interest groups and associations listed in Congressional Quarterly's Washington Information Directory ($N = 892$). A screening question identified 787 groups who reported at least one contact with a cabinet department or independent agency during the year prior to the study. These groups were then asked:

> "For the federal agency with which this association communicates, consults or interacts the most, does this association interact with it frequently, occasionally, seldom, or almost never?"

The dependent variable `interact` captures the groups' responses, with observations coded 1 for "almost never," 2 for "seldom," 3 for "occasionally" and 4 for "frequently."

The data also contain three general types of variables. `age` is the age of the group, in years (i.e., 1985 - the year the association was founded). `taxexmpt` is an indicator of tax-exempt status. Two other variables tap the nature of the group's membership: `indmembs` is coded 1 for groups whose members consist of individual persons, and 0 otherwise; `orgmembs` is coded 1 for associations where members are themselves associations (e.g "peak associations") and 0 otherwise (groups coded 0 on both variables are "mixed," having members of both types).

7 Multinomial logit model

Use the file `http://www.jkarreth.net/files/exercise7.dta`.[4] Be sure to convert string variables to numeric variables and to delete cases with missing values (R hint: `na.omit()`). feel free to use any/all of the variables in the data set.

Fit the model, assess convergence and supply a brief interpretation of the results. *Be careful*: Most of these model specifications might take some time to converge, so start with a small number of iterations (only 10 or 20). If you have time, please present predicted probabilities or other useful quantities of interest.
The time was January, 2005. Condoleeza Rice was sworn in as the first African-American Secretary of State, Mahmoud Abbas was declared the winner of the Palestinian election, and (perhaps most important) Texas light sweet crude was selling for the princely sum of $45 a barrel. It was during those innocent, happier days that ABC News and the Washington Post commissioned a poll about public opinion on traffic. Among other things, pollsters asked 1,204 lucky, randomly-selected Americans:

> "What kind of vehicle do you usually drive – a car, an SUV, a pickup truck, or what?"

What does this have to do with political science? The answer ought to be obvious. We'll explore the political dynamics of car ownership, using the data from the 2005 ABC/WP poll. The main variable of interest is `cartype`, coded one for cars, two for SUVs, and three for pickup trucks. Covariates include dummy variables for `urban` residence, being `married`, having `kids`, and being `black` and/or `female`, as well as a naturally coded variable for `age` and an ordinal variable for level of `education`. Best of all, we also have two dichotomous variables for political party (`democrat` and `GOP`, with independents as our baseline) and a four-point ordinal scale indicating each respondent's approval or disapproval for President Bush.

8 Multilevel model

The data, a classic example for data where multilevel modeling is appropriate, are from the "High School and Beyond" Study from the 1980s, described in more detail in Raudenbush et al. (2004), *HLM6: Hierarchical Linear and Nonlinear Modeling*. The portion of the data you will use provides

[4] Taken from a homework assignment in Chris Zorn's MLE course at a previous ICPSR summer program.

information at the level of students and at the level of their schools.

You can use the dataset `http://www.jkarreth.net/files/exercise8.small.dta`, which is a smaller version of the original HSB dataset and limited to only 22 schools. You may also use the full dataset with 160 schools, `http://www.jkarreth.net/files/exercise8.full.dta` - but be prepared for the time needed to fit complex models on larger datasets.

Outcome variable:

- `mathach`: student's mathematical ability (continuous)

Student-level predictors:

- `ses`: student's socioeconomic status (continuous)
- `minority`: binary variable for students from minorities
- `female`: binary variable for female students

School-level predictors:

- `size`: number of students attending that school
- `sector`: whether the school is public sector (0) or private (1)
- `disclim`: disciplinary climate of the school (continuous)
- `meanses`: average socioeconomic status of all students at the school

Assignment

Fit a multilevel model of student's mathematical ability, with one or two school-level predictors, and three student-level predictors. Present the results in a regression table, and briefly interpret them.

Compare your estimates to the results from an equivalent frequentist model (if you have easy access to software for this estimation).

Helpful hints

Group IDs. Remember that when working with unbalanced panels (or using nested indexing), JAGS code requires a group-level indicator that starts at 1. In `exercise8.small.dta` or `exercise8.full.dta`, replace the `school` variable with such an indicator before you fit your model. In R, you can do this using the `unique()` and `match()` commands:

```
> uniqschool <- unique(hsb.small$school)
> hsb.small$schoolid <- match(hsb.small$school, uniqschool)
```

Alternatively, you can use one line of code to achieve the same:

```
> hsb.small$schoolid <- as.numeric(as.ordered(hsb.small$school))
```

In Stata, you may use:

```
. egen schoolid = group(school)
```

Level-2 variables Level-2 variables need to be in your data as vectors of the same length as the number of level-2 units. For instance, if you want to fit a model with the level-2 covariate `size`, ensure that your data contain `size` as a vector of the length 22 (if you use the `hsb.small.dta` dataset). Note that you should not use the `unique()` command for covariates: you may end up with a vector shorter than the number of level-2 units if several level-2 units have the same value for that particular variable. Therefore, in R, you can construct this variable like this:

```
> l2size <- as.vector(by(hsb.small$size, hsb.small$schoolid, mean))
```

Note that a good check for whether you've "gotten this right" is whether binary variables remain binary variables, or whether all of a sudden the `sector` dummy gets a value of, for instance, 0.75. Of course you can also use other methods to collapse these level-2 variables, such as the dplyr package, or the solutions described at `http://www.ats.ucla.edu/stat/r/faq/collapse.htm`. What's important is that you generate a correctly sorted vector with a length of the number of your level-2 units.

In Stata, you would create a second dataset by collapsing the level-2 variables of interest, and use `wbvector` separately for the level-1 and level-2 datasets.

```
. collapse (mean) size sector disclim meanses, by(schoolid)
. rename (size sector disclim meanses) l2=
```

References

Hainmueller, J. and Hiscox, M. J. (2006). Learning to Love Globalization: Education and Individual Attitudes Toward International Trade. *International Organization*, 60(2):469–498.