STAT/CSSS 564: Assignment 3

Jeff Arnold, Sheridan Grant

May 2nd, 2017

Instructions

1. Fork this repository to your account
2. Edit the file `solutions.Rmd` with your solutions to the problems.
3. Submit a pull request to have it graded. Include either or both a HTML and PDF file.

For updates and questions follow the Slack channel: #assignment3

This assignment will require the following R packages:

```
library("rstan")
library("bayesplot")
library("tidyverse")
```

1 Poisson and Negative Binomial Models

Mosteller and Wallace (1963) analysis of the authorship of disputed letters in the *The Federalist Papers* records the the word frequency in selected articles by Alexander Hamilton and James Madison. The articles were divided into blocks of approximately 200 words each, and the number of instances of various words in each block were recorded. For authorship attribution, they are concerned with the frequency of frequency words. The file `data/federalist.csv`contains data on the occurrence of the word "may" in these papers.[^federalist]

[federalist]: This problem is based on Bayesian Data Analysis 3 Gelman et al. (2013 Ch. 17, Ex. 2 and 3).

```
col_types <- cols(
   author = col_character(),
   count = col_integer(),
   number = col_integer()
)
may <- read_csv('data/federalist.csv', col_types = col_types)
may
```

```
## # A tibble: 14 × 3
##      author count number
##       <chr> <int>  <int>
## 1  Hamilton     0    128
## 2  Hamilton     1     67
## 3  Hamilton     2     32
## 4  Hamilton     3     14
## 5  Hamilton     4      4
## 6  Hamilton     5      1
## 7  Hamilton     6      1
## 8   Madison     0    156
## 9   Madison     1     63
## 10  Madison     2     29
## 11  Madison     3      8
## 12  Madison     4      4
```

```
## 13  Madison     5      1
## 14  Madison     6      1
```

Expand the dataset to have one observation per occurrence. This data will be easier to work with given the Stan models seen thus far. However, it would be more efficient to work with the original data (though the model may seem more confusing). The following code creates on observation per block, per author, with a single variable indicating the occurrences:

```
may_long <- may %>%
  mutate(occur = map2(count, number, rep)) %>%
  unnest(occur) %>%
  select(author, occur) %>%
  mutate(hamilton = as.integer(author == "Hamilton"),
         hamilton_scaled = hamilton - mean(hamilton))

glimpse(may_long)
```

```
## Observations: 509
## Variables: 4
## $ author          <chr> "Hamilton", "Hamilton", "Hamilton", "Hamilton"...
## $ occur           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ hamilton        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ hamilton_scaled <dbl> 0.5147348, 0.5147348, 0.5147348, 0.5147348, 0....
```

1. Fit a Poisson model (`stan/poisson.stan`) to these data with an intercept and a single indicator variable for `hamilton`:

$$
\begin{aligned}
y_i &\sim \texttt{Poisson}(\lambda_i) \\
\lambda_i &= \exp(a + b \times \texttt{hamilton}_i) \\
a &\sim N(0, 10) \\
b &\sim N(0, 2.5)
\end{aligned}
$$

   Remember to use the mean 0 scaled version of the Hamilton indicator variable.

2. Fit a Negative Binomial model (`stan/neg_binomial.stan`) to the data with different parameters for each author and a non-informative prior distribution.

$$
\begin{aligned}
y_i &\sim \texttt{Poisson}(\mu) \\
\lambda_i &= \exp(a + b \times \texttt{hamilton}_i) \\
a &\sim N(0, 10) \\
b &\sim N(0, 2.5)
\end{aligned}
$$

   Remember to use the mean 0 scaled version of the Hamilton indicator variable.

3. In these models, what is the support (domain) of the response variable? Of the linear predictors $a + b \times \texttt{hamilton}$, and the expected value (λ or μ)?

4. Stan provides two parameterizations of the negative binomial distribution: `neg_binomial` and `neg_binomial_2`. Describe the parameters used in each, and how these parameters relate to each other. The model provided uses `neg_binomial_2`. What makes that a more convenient parameterization for our purposes.

5. What does the ϕ parameter in the Negative Binomial model do? The Poisson model does not have an equivalent additional parameter; what constraint on the shape of the Poisson distribution does that impose?

6. There are separate parameters for each author in these models. Is this any different than fitting a separate model for each author. Consider both the negative binomial and Poisson models.

7. Both the Poisson and Negative Binomial Models are for "unbounded" counts, and thus allow for potentially infinite values. However, the description notes that each of these chunks has approximately 200 words, so the number of uses of the word "may" cannot be infinite in any particular chunk. This suggests a third model, in which the response is modeled with a Binomial distribution, `y ~ binomial(200, p)` where `p` is the proportion of words that are "may". Without actually running the models (though nothing is stopping you), would you expect that using a binomial model would have much different results? What are some advantages or disadvantages of the Poisson or Negative Binomial models?

8. For each model, calculate the probability that Hamilton used the word "may" more often. This can be done outside the Stan model, by extracting the parameters from the `stanfit` object, or within the Stan model, by adding a new variable (the best place to put it would be the `generated quantities` block).

9. The `generated quantities` blocks of these models simulate samples from the posterior predictive distribution, and store them in the `y_rep` variable. Plot the predictive posterior distributions for each model and compare them using the visual methods. See this bayesplot vignette.

10. Consider a reasonable test statistic for comparing the model to the data and calculate its posterior predictive p-value for each model. See this bayesplot vignette.

11. Compare the models using PSIS-LOO statistic. The LOO-PSIS is an estimator of a model's the *expected* log-likelihood of a new (out-of-sample) observation. Read the loo vignette for more information.[^loo]. WAIC, discussed in McElreath (2016), is also an estimator of the expected out-of-sample error. However, Vehtari, Gelman, and Gabry (2016) show that the more recently developed PSIS-LOO is not much more computationally burdensome and generally has better performance than WAIC.

    1. Calculate the PSIS-LOO for each observation and interpret it. Use the `loo` function.
    2. What does the value of `p_loo` mean? How does it compare to the actual number of parameters in the model? Why might it be different? Ex-ante, why would you expect the negative binomial to have a larger value of `p`?
    3. Compare the models using the `loo::compare` function. Which model has the better predictive fit?

12. The previous analysis could potentially be used for model selection. However, model choice effectively puts zero probability on some types of models, violating Cromwell's rule. Another approach is *continuous model expansion*, which is to build a larger model that incorporates special cases and thus allows for model uncertainty; see Gelman et al. (2013 Ch 6). How is the Poisson model a specific case of the negative binomial model? If the Poisson model were "true", what parameter values would you expect to see in the negative binomial model?

2 Statistical Simulation

Read King, Tomz, and Wittenberg (2000). They propose a statistical simulation approach for interpreting statistical analysis; see section "Simulation-Based Approaches to Interpretation". Compare and contrast this to a full Bayesian approach.

3 Student-t Prior

The robust regression with Student-t error example uses the following prior on the degrees of freedom parameter, a Gamma distribution with shape 2 and inverse-scale (rate) of 0.1,

$$\nu \sim \text{Gamma}(2, 0.1).$$

The Student-t distribution is used because it has wider tails and thus is less sensitive to outliers than a normal distribution. However, the researcher generally has no information about the value of the degrees of freedom. So why was this distribution chosen?

1. Plot this prior distribution, and the values of the 5th and 95th quantiles. You can use `dgamma(x, 2, rate = 0.1)` and `qgamma(x, 2, scale = 0.1)`. What is
2. Additionally, the prior is truncated at 2. Why? Hint: What moments of the Student-t distribution are not-defined for values between 2.

4 Student-t as a Mixture of Normals

The Student-t distribution is a scale mixture of normal distributions.[1] This means that a Student-t distribution can be represented as normal distributions in which the variances are drawn from different distributions. Suppose X is distributed Student-t with degrees of freedom ν, location μ, and scale σ,

$$X \sim t_\nu(\mu, \sigma).$$

Samples from Y can be drawn by

$$x_i \sim N(\mu, \lambda_i^2 \gamma^2)$$

If the local variance parameters are distributed inverse-gamma

$$1/\lambda^2 \sim \text{Gamma}(\nu/2, \nu/2).$$

Many distributions used in regression shrinkage: Double Exponential (Laplace), and Hierarchical Shrinkage (Horseshoe), have this representation.

You can draw a sample from this:

```
df <- 10
n <- 1000
sigma2 <- rgamma(n, 0.5 * df, 0.5 * df)
x <- rnorm(n, sd = sqrt(1 / sigma2))
```

Plot samples drawn in this way against either samples or theoretical values of the Student-t distribution. Try a few values of the degrees of freedom. Try something small (3) and large (100).

You can draw samples directly from a Student-t with `rt`. A quantile-quantile plot (geom_qq) or a density plot with the function (geom_density and stat_function).

Note: there isn't a right answer to this. Well, actually, there is, and you know it—they are equivalent, a proof is in the link. So for credit, do a little work, and show it. This pattern appears often, so wrap your head around it.

5 Transformations of Coefficients

Rainey (2016) notes that unbiased estimators of parameters does not imply that transformations of those parameters are unbiased estimators. See Carpenter (2016) for a more in depth discussion of this.

1. One property of frequentist estimators is their bias. Let β be the true value of a parameter, and $E(\beta)$ be the expected value of sampling distribution of a statistic, the bias of that statistic is,

   $$\text{bias}(\hat{\beta}) = E(\hat{\beta}) - \beta$$

   Often there is great concern whether an estimator is unbiased on not. How does the property of "bias" apply to Bayesian estimators?

[1] mix

References

Carpenter, Bob. 2016. "The Impact of Reparameterization on Point Estimates." *Stan Case Studies*, April. http://mc-stan.org/documentation/case-studies/mle-params.html.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, and Aki Vehtari. 2013. *Bayesian Data Analysis*. Taylor & Francis Ltd. http://www.ebook.de/de/product/15022612/andrew_gelman_john_b_carlin_hal_s_stern_david_b_dunson_aki_vehtari_bayesian_data_analysis.html.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2). JSTOR: 347. doi:10.2307/2669316.

McElreath, Richard. 2016. *Statistical Rethinking*. Apple Academic Press Inc. http://www.ebook.de/de/product/24465987/richard_mcelreath_statistical_rethinking.html.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58 (302). Informa UK Limited: 275–309. doi:10.1080/01621459.1963.10500849.

Rainey, Carlisle. 2016. "Transformation-Induced Bias: Unbiased Coefficients Do Not Imply Unbiased Quantities of Interest." http://www.carlislerainey.com/papers/bias.pdf.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2016. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing*, August. Springer Nature. doi:10.1007/s11222-016-9696-4.