# STAT/CSSS 564: Assignment 4

*Jeff Arnold & Sheridan Grant*

*May 14, 2017*

## Instructions

This repository contains the assignment instructions. Submitted solutions will use a **separate** repository.

1. Fork UW-CSSS-564/assignment-2017-4-submissions repository to your account.
2. Edit the file `README.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: #assignment4.

This assignment will require the following R packages:

```r
library("rstan")
library("rstanarm")
library("haven")
library("tidyverse")
library("loo")
```

Set the following options for faster sampling sampling.[^rprofile] This option sets the default to save a compiled model to disk and reuse it if the code hasn't changed. This will avoid needless recompilation.

```r
rstan_options(auto_write = TRUE)
```

If you sample with multiple chains and your computer has multiple cores, this will run the chains in parallel.

```r
options(mc.cores = parallel::detectCores())
```

Fearon and Laitin (2003) is a famous paper in the civil war (intra-state) war literature. It analyzes the factors associated with the onset of civil (intra-state) war between 1945–99. They consider a variety of variables such as prior civil wars, per-capita income, population, non-contiguous state, oil-exporter, new-state, democracy, ethnic fractionalization.

Montgomery and Nyhan (2010) replicate this work using Bayesian Model Averaging. This assignment pursues a similar replication, but we will use regularization The replication data for the original paper is here.

The original code is con

```r
# variables we'll use later
keepvars <- c("onset", "warl", "gdpenl", "lpopl1", "lmtnest", "ncontig",
"Oil", "nwstate", "instab", "polity2l", "ethfrac", "relfrac",
"anocl", "deml", "nwarsl", "plural", "plurrel", "muslim", "loglang",
"colfra", "eeurop", "lamerica", "ssafrica", "asia", "nafrme",
"second")

# original Fearon & Laitin war
fl <- read_dta('https://github.com/UW-CSSS-564/assignment-2017-4/blob/master/data/fl.dta?raw=true') %>%
# remove a coding error
  filter(onset != 4) %>%
  # add the count of wars in neighboring countries
  inner_join(read_dta("https://github.com/UW-CSSS-564/assignment-2017-4/raw/master/data/nwarsl.dta"), by
  # log(number of languages)
  mutate(loglang = log(numlang)) %>%
  select(one_of(keepvars))
```

# 1 Replicating Fearon and Laitin

Let $y_{c,y} \in \{0,1\}$ be whether country $c$ in year $y$ has the onset of a civil war. We will model this as a logistic model in which the probability of civil war onset for a country-year, $\mu_{c,y}$, is a function of $K$ predictors, $x_{c,y}$.

$$
\begin{aligned}
y_{c,y} &\sim \mathsf{Bernoulli}(\mu_{c,y}) \\
\mu_{c,y} &= \mathrm{logit}^{-1}(\eta_{c,y}) = \frac{1}{1 + \exp(-\eta_{c,y})} \\
\eta_{c,y} &= x_{c,y}\beta
\end{aligned}
$$

We will consider various prior distributions of the coefficient parameters, $\beta$.

Estimate the two models that Montgomery and Nyhan (2010) uses in the paper using weakly informative priors,

$$
\begin{aligned}
\beta_0 &\sim N(0,5) \\
\beta_k &\sim N(0,2.5) \quad \text{for } k \in 1, \dots, K.
\end{aligned}
$$

and calculate the LOO performance of these methods. When replicating results from papers, you will often have to dig through some confusing code or files, perhaps in programming languages or file formats you're unfamiliar with (we had to do this to write this question!). The two logit models are the first and third used by Fearon and Laitin (2003). The original paper used Stata, and the code is contained in the file reference-code/f&l-rep.do. In Stata, the command `logit` is followed by the response variable and a lists of the predictors.

To estimate this (and the other models). You can directly use either a Stan model, as we have used in class, or use the **rstanarm** package. The function `stan_glm` can estimate the See the vignette Estimating Generalized Linear Models for Binary and Binomial Data with rstanarm describes

Here's a few examples which run similar logit models:

```
mod <- stan_glm(onset ~ loglang, family = binomial(), data = fl)
loo_mod <- loo(mod)
mod2 <- stan_glm(onset ~ loglang + Oil, family = binomial(), data = fl)
loo_mod2 <- loo(mod2)
compare(loo_mod, loo_mod2)
```

When estimating these models ensure that you scale the variables. The priors in **rstanarm** do this automatically when `autoscale = TRUE` (default). If you are using **rstan**, you will have to do this manually.

# 2 Regularization Priors

- Now estimate this model with all 25 predictor variables and the following priors
    - weakly informative priors
    - hierarchical shrinkage prior

  You can use either **rstan** or **rstanarm**.
- Plot the mean and 90% credible intervals for these models.
    - How do the coefficients differ between models?

  - Which coefficients have the largest effects?
  - How do these results compare with the two models chosen in Fearon and Laitin (2003)?

As before, be sure that the variables are scaled.

# 3 Variable Scaling

Rerun the weakly informative and hierarchical shrinkage models, but do not scale the variables. Set `autoscale = FALSE` if using `stan_glm` or do not scale the parameters if using **rstan**.

- What does this option do?

- Which coefficients changed the most?
- Compare the changes in the coefficients to the standard deviations of these coefficients?
- Explain how rescaling the variables affects the priors on the coefficients.

# 4 Model Comparison

- Calculate and compare the LOO-PSIS estimates of the elpd for each model. Do you get any warning messages? If so what do they mean - and how would you address it?

- Which model has the best fit?

- The LOO-PSIS approximates Leave-one-Out cross validation. LOO-CV estimates the out-of-sample model fit by fitting the model to $n-1$ observations and predicting the observation that was not included. Given the structure of the data, is this the out-of-sample quantity of interest? Provide another cross-validation example that may be more appropriate and discuss why. You do not need to implement it.

- For the best-fitting model, extract the observation level `elpd`. If `foo` is an `loo` object you can extract these as follows,

  ```
  foo$pointwise
  ```

  Plot the summaries of the observation level elpd values by year and country. For years and countrys does it work well or poorly?

# 5 Model Size

- Compare the model sizes given by `loo` using the results from the previous section. How does that compare to the actual number of parameters in the model?
- The HS prior more aggressively shrinks coefficients towards zero. Is the mean of any coefficient exactly zero? Can you think of a method to define a thresh-hold where coefficients of some variables could be treated as effectively zero? The solutions will provide some examples from the literature (and my Bayesian notes have references to some), but try to think it through on your own. The idea isn't to get it "right", but think about the problem prior to finding out how others have approached (and maybe solved?) the problem.

## 5.1 Posterior Predictive Checks

Thus far, we've only compared models using the log-posterior values. Using a statistic of your choice, assess the fit of data generated from the model to the actual data using posterior predictive checks.

# 6 Taking Time Seriously

One variable not in the previous models is the time since the last civil war.[1] Beck, Katz, and Tucker (1998) note that a duration model with time-varying covariates can be represented as a binary choice model that includes a function of the time at risk of the event. As such we could rewrite the model

$$\eta_{c,y} = x'_{c,y}\beta + f(d_{c,y})$$

where $d_{i,t}$ is the time since the last civil war or the first observation of that country in the data.

One issue is that we don't know the duration function, $f$. Since $f$ is unknown, and the analyst generally has few priors about it, generally a flexible functional form is used. Beck, Katz, and Tucker (1998) suggest using a cubic spline, while Carter and Signorino (2010) suggest a polynomial. In particular, Carter and Signorino (2010) suggest a cubic polynomial, meaning the linear predictors now becomes,

$$\eta_{c,y} = x'_{c,y}\beta + \gamma_1 d_{c,y} + \gamma_2 d^2_{c,y} + \gamma_3 d^3_{c,y}$$

- Carter and Signorino (2010) argue that a cubic polynomial is usually sufficient to capture the time-dependence in this sort of data. This is another sort of model choice. How would you solve the choice of the the order of the polynomial with regularization? Include this variable, and re-estimate a model.
- Box-Steffensmeier and Zorn (2001) discuss how including only duration function as above in the model is equivalent to a "proportional hazards" assumption. In this context, it would mean that all variables have the same effect (coefficient) on the probability of failure regardless of the duration. They suggest estimating a model that interacts all the variables with a function of the duration, and running an F-test that all the interactions were zero. How would you address this concern using Bayesian regularization?

# 7 Time Trends and Time-Varying Coefficients

The time since the last war is not the only way in which time can affect predictions and inference.

The baseline probability of civil-war may vary over time. Notably there was an increase in war after the civil war. We could model that as a time-trend, which is an unknown function of $y$ (in this case):

$$\eta_{c,y} = x'_{c,y}\beta + f(y)$$

In classical regression, special cases of time trends are considered for purposes of parsimony

- No trend
- Linear trend
- Time indicators

The linear trend is the most restrictive, and including an indicator variable for each unique value of time (e.g. year dummies) is the most restrictive.

With regularization it is possible to include and estimate flexible time trend while using the shrinkage prior to impose parsimony.

- Re-estimate the model with a flexible time trend.
- How would you extend the model to include time-varying coefficients on these variables? At least write it out it, if not try to estimate the model.

# Changelog

See this page for any differences between when the assignment was released and the current version.

[1] Though it is discussed in a footnote of Fearon and Laitin (2003 fn. 26).

## 2017-05-17

- Replicating Fearon and Laitin
  - add reminder and instructions to rescale variable
- Regularization Priors
  - add reminder and instructions to rescale variable
- Variable Scaling
  - clarify that the user is to run the weakly informative and hierarchical shrinkage models.
- Model Comparison
  - On last problem, edit the instructions for clarity.
  - Provide examples for extracting pointwise elpd values
- Other
  - Add CHANGELOG
  - Fix numbering of problems
  - Rename `index.pdf` to `assignment-2017-4.pdf`
  - Add `README.md` generated from `index.Rmd`

## References

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42 (4). [Midwest Political Science Association, Wiley]: 1260–88. http://www.jstor.org/stable/2991857.

Box-Steffensmeier, Janet M., and Christopher J. W. Zorn. 2001. "Duration Models and Proportional Hazards in Political Science." *American Journal of Political Science* 45 (4). [Midwest Political Science Association, Wiley]: 972–88. http://www.jstor.org/stable/2669335.

Carter, David B., and Curtis S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18 (03). Cambridge University Press (CUP): 271–92. doi:10.1093/pan/mpq013.

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97 (01). Cambridge University Press (CUP): 75–90. doi:10.1017/s0003055403000534.

Montgomery, Jacob M., and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18 (02). Cambridge University Press (CUP): 245–70. doi:10.1093/pan/mpq001.