115th Congress Analysis Using Bayesian Methods

Adrian Bandolon

December 7, 2018

Introduction

- The 115^{th} United States Congress is a meeting of the legislative branch of the United States of America. This congress convened on January 3, 2017 until January 3, 2019[1].
- This is the current Congress, and will be replaced by new representatives that were elected on November, 2018. Every two years, half of the representatives in the Congress is up for election[2].
- The United States House of Representatives consist one of the two chambers in the United States Congress. The US Senate is the upper chamber and the US House of Representatives is the lower chamber. Together these two chambers make up the legislative branch of the US government. Representatives in the House sit in congressional districts[3].
- There are 435 congressional districts in the United States House of Representatives. This number has been fixed by law. A Congressional District is based on population. A census is conducted every 10 years to determine the number of Representatives that each state sends to Congress and the number of congressional districts in each state[4].
- The political system in the United States of America has been dominated by two parties, the Republican and Democratic Parties. This has led a two-party system politicla system. There are "third" parties (Libertarian, Green and Constitution Parties) and Independents (those with no party affiliations). However, the Republican or Democratic Parties has won the presidential election since 1852, and has controlled the Congress to some extent since 1856. Typically, one of the two parties hold a majority in Congress[5].

Data

- Data on all congressional districts[6] represented in the *U.S. House of Representatives* during the *115th U.S. Congress* was supplied in `party115cong.csv`. Each row represented a district. Columns represented variables for each district:

Variable Name	Variable Description
state	The state where the district is found or the District of Columbia
district	Congressional district identifier within its State(eg. Texas's 1st Cong. District)
electedrep	Name of the Elected Representative for the District
party	Party affliation of the elected representaive (*D-Democrat; R-Republican*)
medHouseIncome	Median Household Income (*in Dollars*)

Table 1. Variable names and their corresponding descriptions, found in *party115cong.csv*.

- Below is an overview of some features of `party115cong.csv`.

[1] https://en.wikipedia.org/wiki/115th_United_States_Congress
[2] Pulled from memory of AP Government.
[3] https://en.wikipedia.org/wiki/United_States_House_of_Representatives
[4] https://en.wikipedia.org/wiki/Congressional_district
[5] https://en.wikipedia.org/wiki/Political_parties_in_the_United_States
[6] There are 435 congressional districts, but there are 436 considered in this dataset because the federal district of Washington, D.C. was regarded as a congressional district.

– We can see from *Figure 1* that there seems to be good separation in Median Household Income between districts with Republican vs. Democratic representatives.

– There seems to be no discernible difference in Median Household Income between states (*Figure 2*). However, similar to what we see in *Figure 1*, there seems to be differences in Median Household Income between districts with Republican vs. Democratic representatives, within each State.

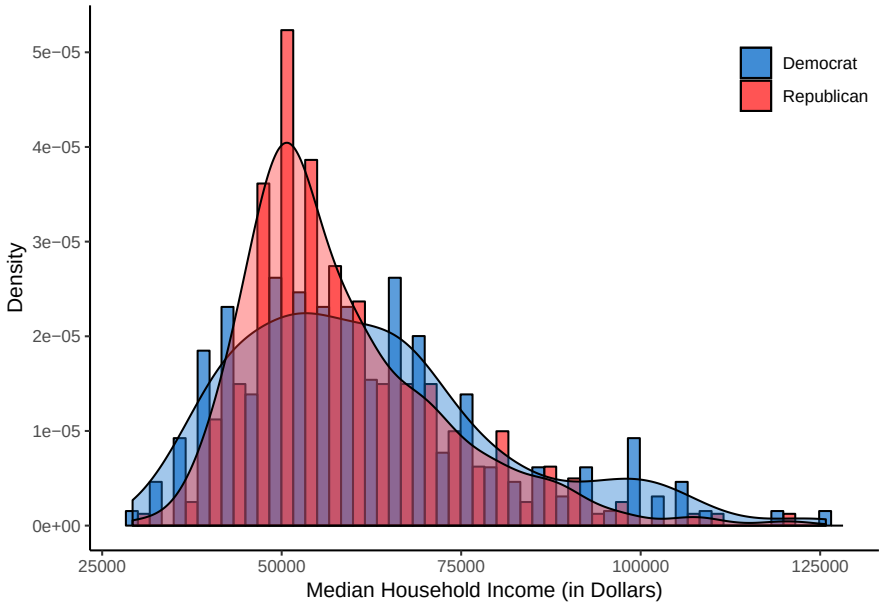

Figure 1. Histogram of Median Household Income (in Dollars) by Party Affiliation

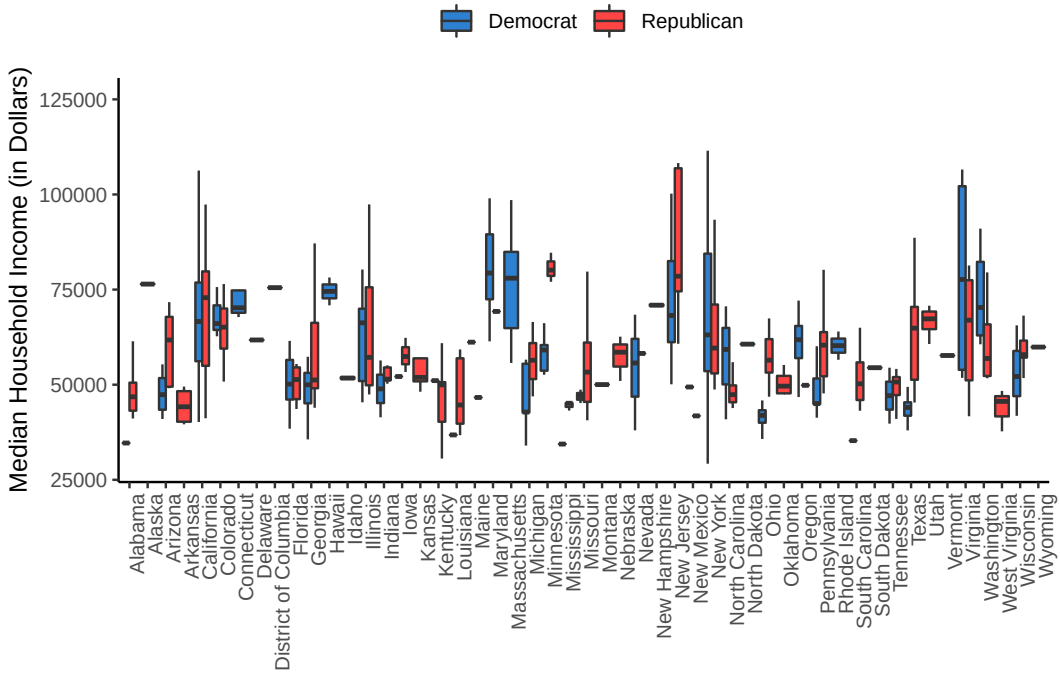

Figure 2. Median Household Income (in Dollars) by State and Party Affliation

First Model

- This is a Bayesian logistic regression model that attempts to explain `party` based on the *natural logarithm* of `medHouseIncome`.

a. JAGS model:

```
model {
  for (i in 1:length(party)) {
    party[i] ~ dbern(prob[i])
    logit(prob[i]) <- beta0 + betaIncome * IncomeStdized[i]

    party_rep[i] ~ dbern(prob[i])
    }

  beta0 ~ dt(0, 0.01, 1) # intercept
  betaIncome ~ dt(0, 0.16, 1) # slope
  }
```

Code 1. JAGS model (***.bug** file) specification for the *First* Model.

b. Model Summary:

Burn-In Length	1000
Iterations	2001 : 12000
Thinning Interval	1
Number of Chains	4

Table 2. Computation Summary for the *First* Model.

	beta0	betaIncome
Effective Sample Size	25395.56	25071.64

Table 3. Effective sample sizes for monitored coefficients of the *First* Model.

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

	Mean	SD	2.5%	97.5%
beta0	-0.21	0.10	-0.40	-0.02
betaIncome	0.36	0.19	-0.02	0.74

Table 4. Approximate posterior mean, posterior standard deviation, and 95% central posterior interval for each parameter of the *First* Model.

d. The probability of electing a Democrat as median household income increases evidently ***increases*** based on the posterior probability that the "slope" (`betaIncome`) is greater than zero.

```
mean(post.x1[,"betaIncome"] > 0)
```

```
[1] 0.9684
```

e. The effective number approximated by Plummer's DIC is around 1.99. This about the same as the actual number of parameters, which is 2.

```
Mean deviance:  598
penalty 1.987
Penalized deviance: 600
```

Second Model

a. JAGS model:

```
model {
  for (i in 1:length(party)) {
    party[i] ~ dbern(prob[i])
    logit(prob[i]) <- beta0 + betaState[state[i]] + betaIncome * IncomeStdized[i]

    party_rep[i] ~ dbern(prob[i])
    }
  # beta state prior
  for (j in 1:max(state)) {
    betaState[j] ~ dnorm(0, sigmaState)
    }
  # other priors
  beta0 ~ dt(0, 0.01, 1) # intercept
  betaIncome ~ dt(0, 0.16, 1) # slope
  # hyperprior
  sigmaState ~ dunif(0,100)
  }
```

Code 2. JAGS model (***.bug** file) specification for the *Second* Model.

b. Model Summary:

Burn-In Length	1000
Iterations	2001 : 22000
Thinning Interval	1
Number of Chains	4

Table 5. Computation Summary for the *Second* Model.

	beta0	betaIncome
Effective Sample Size	5424.83	8504.24

Table 6. Effective sample sizes for monitored coefficients of the *Second* Model.

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

	Mean	SD	2.5%	97.5%
beta0	-0.45	0.23	-0.92	-0.01
betaIncome	-0.46	0.28	-1.02	0.08

Table 7. Approximate posterior mean, posterior standard deviation, and 95% central posterior interval for each parameter of the *Second* Model.

d. After adjusting for state, and using the significance level of $p = 0.05$, the slope (`betaIncome`) is not significantly greater than zero. This means that the probability of electing a democrat does not increase with increasing Median Household Income, when adjusted for `state`.

```
mean(post.x2[,"betaIncome"] > 0)
```

```
[1] 0.046925
```

e. Which state has the largest (in the positive direction) posterior mean random effect? Which state has the smallest (in the negative direction) posterior mean random effect?

	Mean	SD	2.5%	97.5%
Massachusetts	2.34	0.92	0.77	4.39
Oklahoma	-1.37	0.96	-3.51	0.28

Table 8. States with the highest and lowest posterior mean random effect.

- To better understand these results, a closer look at the specific states is needed.

State	District	Elected Representative	Party
Oklahoma	1	Jim Bridenstine	Rep
Oklahoma	2	Markwayne Mullin	Rep
Oklahoma	3	Frank Lucas	Rep
Oklahoma	4	Tom Cole	Rep
Oklahoma	5	Steve Russell	Rep
Massachusetts	1	Richard Neal	Dem
Massachusetts	2	Jim McGovern	Dem
Massachusetts	3	Niki Tsongas	Dem
Massachusetts	4	Joseph P. Kennedy III	Dem
Massachusetts	5	Katherine Clark	Dem
Massachusetts	6	Seth Moulton	Dem
Massachusetts	7	Mike Capuano	Dem
Massachusetts	8	Stephen F. Lynch	Dem
Massachusetts	9	Bill Keating	Dem

Table 9. Representatives from states with the highest and lowest posterior mean random effect.

- From *Table 9* we can see that all Representatives from Oklahoma are Republicans and all Representatives from Massachussetts are Democrats. It is then no surprise that Massachussetts have the highest posterior mean random effect, while Oklahoma has the lowest.
  - For a similar effect, *Table 10* looks at the states with the second highest and second lowest posterior mean random effect.

State	District	Elected Representative	Party
Arkansas	1	Rick Crawford	Rep
Arkansas	2	French Hill	Rep
Arkansas	3	Steve Womack	Rep
Arkansas	4	Bruce Westerman	Rep
Connecticut	1	John B. Larson	Dem
Connecticut	2	Joe Courtney	Dem
Connecticut	3	Rosa DeLauro	Dem
Connecticut	4	Jim Himes	Dem
Connecticut	5	Elizabeth Esty	Dem

Table 10. Representatives from states with the second highest and second lowest posterior mean random effect.

f. The effective number of parameters approximated by (Plumber's) DIC is around 31. This actual number of parameters are 51 (for 50 states plus 1 for the District of Columbia) plus 1 each for the intercept and slope parameters.

- Model 2 (≈ 543) has a smaller DIC than Model 1 (≈ 600). Therefore, based on DIC **alone** Model 2 is better. However, what Model 1 tries to describe is slightly different from what Model 2 is describing. Model 1 describes how median household income affects voting at the national level. In this sense, Model 1 is "better" than Model 2. While Model 2 is "better" at describing how median household income influences voters at the state level.

```
Mean deviance:  512.1
penalty 31.14
Penalized deviance: 543.3
```

Conclusion:

Based on the results from the *First Model*, if we look at the whole country, for those elected in the 115^{th} Congress, the probability of electing a Democrat to the House Of Representatives increases with an increase in median household income. This can also be seen in *Figure 1* where there seems to be more Democratic representatives from districts with higher median household income.

If, however, we adjust for the state where each Representative is elected from, as in the *Second Model*, median household income is no longer as strong a predictor as it was in the *First Model*. This means that, adjusting for state, the probability of electing a Democrat is not significantly different from the probability of electing a Republican, based on an increase in median household income. This pattern is apparent in *Figure 2* where there is no discernible pattern as to which party gets elected based on median household income. In states like California, Florida, Georgia and Colorado for example, median household income are very similar to each other in districts held by each party. In states like Texas, Minnesota, Arizona and North Carolina, median household incomes are disparate between districts held by each party. Based on the results from the *Second Model*, and what *Figure 2* shows us, it seems that median household income is not a good predictor of which party gets elected to a district at the state level.

Appendix

```
# set working directory for easier knitting--LOCALLY!!
setwd("/media/adrian/SchoolFiles/AdvancedBayesianModelling_CS598/finalAnalysis")

# did not set seed as setting seed here does not affect rjag results
library(xtable); options(xtable.comment = FALSE) # for neater tables on pdf
library(knitr); library(rjags); library(ggplot2)

# load the data
congress <- read.csv("party115cong.csv", header = TRUE)
attach(congress) # to save some typing
```

Code 3. Report and analysis set-up and preparation.

```
### Code for Figure 1--a histogram with density plot overlay
ggplot(congress, aes(x = medHouseIncome,  fill = party)) +
  geom_histogram(aes(y=..density..), color = "black", position = "dodge", alpha=0.7) +
  geom_density(alpha=0.4)+
  labs(x = "Median Household Income (in Dollars)", y = "Density") +
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=16,face="bold")) +
  scale_fill_manual(values = c("dodgerblue3", "firebrick1"),
    name="",
    breaks=c("D", "R"),
    labels=c("Democrat", "Republican")) +
  theme_minimal()+theme_classic() +
  theme(legend.justification=c(1,1), legend.position=c(1,1))

### Code for Figure 2 -- a boxplot
ggplot(congress, aes(x=state, y=medHouseIncome, fill=party)) +
  geom_boxplot(outlier.shape = NA, position = "dodge", alpha = 0.9) +
  scale_fill_manual(values = c("dodgerblue3", "firebrick1"),
    name="",
    breaks=c("D", "R"),
    labels=c("Democrat", "Republican")) +
  theme_classic() +
  labs(x = "", y = "Median Household Income (in Dollars)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(legend.position="top")
```

Code 4. Data exploration/visualization.

```
### bug file for the first model
model {
  for (i in 1:length(party)) {
    party[i] ~ dbern(prob[i])
    logit(prob[i]) <- beta0 + betaIncome * IncomeStdized[i]

    party_rep[i] ~ dbern(prob[i])
    }

  beta0 ~ dt(0, 0.01, 1) # intercept
  betaIncome ~ dt(0, 0.16, 1) # slope
  }
```

```r
# recode the parties, 0=Republican, 1=Democrat
congress$party <- ifelse(congress$party=="R", 0, 1)
congress$medHouseIncome <- log(congress$medHouseIncome)

# set-up dataset for first model.
# center to mean 1 and scale to 0.5 sd's
d1 <- list(party = congress$party,
           IncomeStdized = as.vector(scale(
           congress$medHouseIncome,
           center = TRUE,
           scale = 2 * sd(congress$medHouseIncome)
           )))

# initial values for the first model coefficients
inits1 <- list(list(betaIncome=10, beta0= -10),
               list(betaIncome=-10, beta0= 10),
               list(betaIncome=10, beta0= 10),
               list(betaIncome=-10, beta0= -10))

# adaptation run for first model
m1 <- jags.model("model1.bug", d1, inits1, n.chains=4, n.adapt=1000)

# burn-in
update(m1, 1000)

# load dic module for Plummer's DIC later
load.module("dic")

# start
x1 <- coda.samples(m1, c("beta0","betaIncome"), n.iter=10000)

# convergence checks
gel1 <- gelman.diag(x1, autoburnin = F)
eff1 <- effectiveSize(x1)

gel1 <- as.data.frame(t(gel1$psrf))

# extract summaries for first model (x1)
sum1 <- summary(x1)

# break up the summaries into individual tables for improved aesthetics
sum1.1 <- rbind(1000,paste(sum1$start,":",sum1$end), sum1$thin, sum1$nchain)
sum1.1<- as.data.frame(sum1.1)

rownames(sum1.1) <-
  c("Burn-In Length",
  "Iterations",
  "Thinning Interval",
  "Number of Chains")
sum1.tab <- xtable(sum1.1)
print(sum1.tab, include.colnames = FALSE)

# effective sample size for model 1
```

```r
eff1 <- as.data.frame(effectiveSize(x1))
colnames(eff1) <- "Effective Sample Size"

eff1.tab <- xtable(t(eff1))
print(eff1.tab)

# extracts results into a matrix
post.x1 <- as.matrix(x1)
sum1.q <- as.data.frame(sum1$quantiles[,c(1,5)]) # for quantiles
sum1.m <- as.data.frame(sum1$statistics[,1:2]) # for mean and sd
sum1.t <- cbind(sum1.m, sum1.q) # combine both tables, again for aesthetics
xtable(sum1.t)

# check prob. of "slope" being greater than zero
mean(post.x1[,"betaIncome"] > 0)

# get Plummer's DIC, 100k iters.
dic.samples(m1, 100000)
```

Code 5. For *First Model* analysis

```r
# bug model for second model
model {
  for (i in 1:length(party)) {
    party[i] ~ dbern(prob[i])
    logit(prob[i]) <- beta0 + betaState[state[i]] + betaIncome * IncomeStdized[i]

    party_rep[i] ~ dbern(prob[i])
    }
  # beta state prior
  for (j in 1:max(state)) {
    betaState[j] ~ dnorm(0, sigmaState)
    }
  # other priors
  beta0 ~ dt(0, 0.01, 1) # intercept
  betaIncome ~ dt(0, 0.16, 1) # slope
  # hyperprior
  sigmaState ~ dunif(0,100)
  }

# model 2 dataset
d2 <- list(party = congress$party,
           state = unclass(congress$state),
           IncomeStdized = as.vector(scale( # standization same as in  model 1
           congress$medHouseIncome,
           center = TRUE,
           scale = 2 * sd(congress$medHouseIncome)
           )))

# initial values for model 2
inits2 <- list(list(betaIncome=10, beta0= -10, sigmaState = 90),
               list(betaIncome=-10, beta0= 10, sigmaState = 0.01),
               list(betaIncome=10, beta0= 10, sigmaState = 90),
               list(betaIncome=-10, beta0= -10, sigmaState = 0.01))
```

```r
# adaptation for model 2
m2 <- jags.model("model2.bug", d2, inits2, n.chains=4, n.adapt=1000)

# burn-in
update(m2, 1000)

# start
x2 <- coda.samples(m2, c("beta0","betaIncome", "betaState"), n.iter=20000)


# convergence checks
gel2 <- gelman.diag(x2, autoburnin = F)
eff2 <- effectiveSize(x2)

gel2 <- as.data.frame(t(gel2$psrf))

# extract summary for model 2
sum2 <- summary(x2)

# break the summary into different tables for improved aesthetics
sum2.1 <- rbind(1000,paste(sum2$start,":",sum2$end), sum2$thin, sum2$nchain)
sum2.1<- as.data.frame(sum2.1)

rownames(sum2.1) <-
  c("Burn-In Length",
  "Iterations",
  "Thinning Interval",
  "Number of Chains")
sum2.tab <- xtable(sum2.1)
print(sum2.tab, include.colnames = FALSE)

# effective sample size
eff2 <- as.data.frame(effectiveSize(x2[,1:2]))
colnames(eff2) <- "Effective Sample Size"

eff2.tab <- xtable(t(eff2))
print(eff2.tab)

# extracts results into a matrix
post.x2 <- as.matrix(x2[,1:2])
sum2.q <- as.data.frame(sum2$quantiles[,c(1,5)])
sum2.m <- as.data.frame(sum2$statistics[,1:2])
sum2.t <- cbind(sum2.m, sum2.q)
xtable(sum2.t[1:2,])

# check prob. of slope being greater than 0
mean(post.x2[,"betaIncome"] > 0)

# extract just the state coefficients (n=51)
# this was done so that which.min and
# which.max indexing is a little bit more intuitive, for me at least.
sum2.s <- sum2.t[3:length(sum2.t[,2]),]
```

```r
# returns index (row number) state with highest/lowest posterior mean
max.state <- which.max(sum2.s[, 1])
min.state <- which.min(sum2.s[, 1])
minMax.states <- sum2.s[c(max.state,min.state),]

#state 22=Massachusetts; state 37=Oklahoma
rownames(minMax.states) <- c("Massachusetts", "Oklahoma")
min.max.tab1 <- xtable(minMax.states, align = "ccccc")
print(min.max.tab1)

okla <- congress[congress$state=="Oklahoma",]
mass <- congress[congress$state=="Massachusetts",]

minMax.states <- rbind(okla,mass)

# recode to 0=rep, 1=dem for improved table readability
minMax.states$party <- ifelse(minMax.states$party==0, "Rep", "Dem")
minMax.states$medHouseIncome <- exp(minMax.states$medHouseIncome)
colnames(minMax.states) <- c("State", "District", "Elected Representative", "Party")
minMax.tab <- xtable(minMax.states[,1:4], align = "ccccc")
print(minMax.tab, include.rownames=FALSE)

max.state1 <- which.max(sum2.s[-max.state, 1])
min.state1 <- which.min(sum2.s[-min.state, 1])
minMax.states1 <- sum2.s[c(max.state1,min.state1),]

# state 4=arkansas; 7=connecticut
arka <- congress[congress$state=="Arkansas",]
conn <- congress[congress$state=="Connecticut",]

minMax.states1 <- rbind(arka, conn)

minMax.states1$party <- ifelse(minMax.states1$party==0, "Rep", "Dem")
minMax.states1$medHouseIncome <- exp(minMax.states1$medHouseIncome)
colnames(minMax.states1) <- c("State", "District", "Elected Representative", "Party")
minMax.tab1 <- xtable(minMax.states1[,1:4], align = "ccccc")
print(minMax.tab1, include.rownames=FALSE)

# Plummer's DIC
dic.samples(m2, 100000)
```

Code 6. For *Second Model* analysis