Final Project

Incumbency in congressional elections

Haiwei Zhou, Zijian Han

MA 578 Bayesian Statistics

December 18, 2017

Contents

Abstract

This project analyzed the U.S. House of Representatives Electoral Data (Gelman, 1994), which contains data for every US House of Representatives election from 1896 to 1992. In the report, we demonstrated three methods to make analysis in the incumbency advantage. First, we collected, cleaned and explored the data from the website. Then, we performed model selection using various methods including linear regression via OLS, Bayesian linear regression and Hierarchical Model to select the best fitted model and make analysis. We used some popular MCMC diagnostics methods to evaluated the performance of the model. The results showed that our model fitted well and the incumbency has positive influence on the election.

Keywords: Linear Regression, Bayesian, Hierarchical Model

Introduction

The United States House of Representatives is the lower chamber of the United States Congress, the Senate being the upper chamber. Together they compose the legislature of the United States. Elections for representatives are held in every even-numbered year, on Election Day the first Tuesday after the first Monday in November. By law, Representatives must be elected from single-member districts. After a census is taken (in a year ending in 0), the year ending in 2 is the first year in which elections for U.S. House districts are based on that census (with the Congress based on those districts starting its term on the following Jan. 3).

In most states, major party candidates for each district are nominated in partisan primary elections, typically held in spring to late summer. In some states, the Republican and Democratic parties choose their respective candidates for each district in their political conventions in spring or early summer, which often use unanimous voice votes to reflect either confidence in the incumbent or the result of bargaining in earlier private discussions. Exceptions can result in so-called floor fight—convention votes by delegates, with outcomes that can be hard to predict. Especially if a convention is closely divided, a losing candidate may contend further by meeting the conditions for a primary election.

Observers of legislative election in the united states have often noted that incumbency—that is, being the current representative in a district—is an advantage for candidates. We are interested in the magnitude of the effect. We shall use linear regression to study the advantage of incumbency in elections for the U.S. House of Representatives in the past century.

We use the data on congressional elections to analyze the incumbency advantage with the link http://www.stat.columbia.edu/~gelman/book/data/incumbency/, from which we download the data. The incumbency advantage dataset is a longitudinal dataset containing incumbency information for 51 states and their sub-districts along with the Democratic and Republican votes starting from year 1896 to year 1992. Each asc file contains data for a single year incumbency and voting information. Each file is formatted identically with blanks between columns and with the following fields:

(1) State Index, (2) District, (3) Incumbency Code,

(4) Democratic votes, (5) Republican Votes.

Our primary research goal is analyzing the incumbency effect on the vote received by the incumbent party's candidate based on the known Incumbency code for current and previous election, incumbency party's votes received, State and District Index.

At the beginning, we aim to use all the provided asc files to implement our analysis, however, according to data description file, there are a lot of missing values for incumbency in year 1986, we decide to only use 20 century data from year 1900 to year 1992. Additionally, in our original data, there are some missing values recorded with "-9", we need to figure out those missing values from an exception file attached with the dataset before further analysis. In the following three conditions, we could reply on the exception file and impute back the missing values, (1) Both parties have vote totals = -9, (2) One party has vote = 0 and the other party has vote = -9, (3) One part has vote > 0 and the other party has vote = -9. In usual, if it is condition (2) or condition (3), we could some useful information from the exception file, however, if it is condition (1) and there is a third-part winner, we will simply discard those data because we only interest in the Democratic and Republican incumbency condition. After cleaning and imputing those missing values, we create some new variables, namely, "Incumbency Votes Proportion", "Incumbency", "Incumbency Party", based on those provided information from original dataset. Because our primary interest is effect of incumbency, the incumbency advantage can be defined theoretically as:

$$Incumbency\ advantage_i = y^I_{complete\ i} - y^O_{complete\ i}\ ,$$

where:

$y^I_{complete\ i}$ = proportion of vote in district *i* of current incumbent legislator

$y^O_{complete\ i}$ = proportion of vote in district *i* of incumbency party if two parties compete for open seat

To estimate the incumbency advantage, we need to calculate "Incumbency Votes Proportion" as the proportion votes for the current incumbent legislator. If there is an open seat, we just choose the higher proportion of votes between two parties. "Incumbency" is a binary indicator with 0 means open seat and 1 means incumbency. We include another variable called "Incumbency Party": it is 1 if the Democrats control the seat and -1 if the Republicans control the seat, it is 0 otherwise.

Methods and Analysis

Linear Regression via OLS

The first model is an ordinary least square linear regression. We aim to model the incumbency's effect for each year from 1900 to 1992 and we have 36 simple linear models in total. For those consecutive elections with different number of districts with different number of districts, we just simply discard them as is mentioned before. In this model, we include variables with incumbency information about the current and last elections with current incumbency votes proportion as the response variable. We are primarily interested in what a frequentist model will tell us about the incumbency effect. To be more specific, our model is built as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \,,$$

Where

y is the response variable of the current incumbent party's vote proportion?

x_1 is an binary indicator of incumbent situation (0: open seat, 1 : incumbent).

x_2 is past election's votes proportion ranges continuously from 0 to 1.

x_3 is also an indicator of past election's incumbency situation (0: open seat, 1: Democratic, -1: Republican).

The result in table 1 is for the model of 1988 with information from 1986:

Variables	*OLS result*				
	estimates	*2.5%*	*97.5%*	*Standard error*	*p-value*
β_0	*0.315*	*0.236*	*0.394*	*0.040*	0.000
β_1	*0.078*	*0.028*	*0.128*	*0.026*	0.002
β_2	*0.467*	*0.381*	*0.551*	*0.026*	0.000
β_3	*0.010*	*-0.003*	*0.023*	*0.007*	0.144

Table 1: Linear regression for 1988

In 1988's congressional election, the effect of incumbency is huge on the votes proportion. The incumbent situation in 1988 and incumbent proportion in 1986 are all significant as is shown in table 1. However, it only shows us a single year's result.

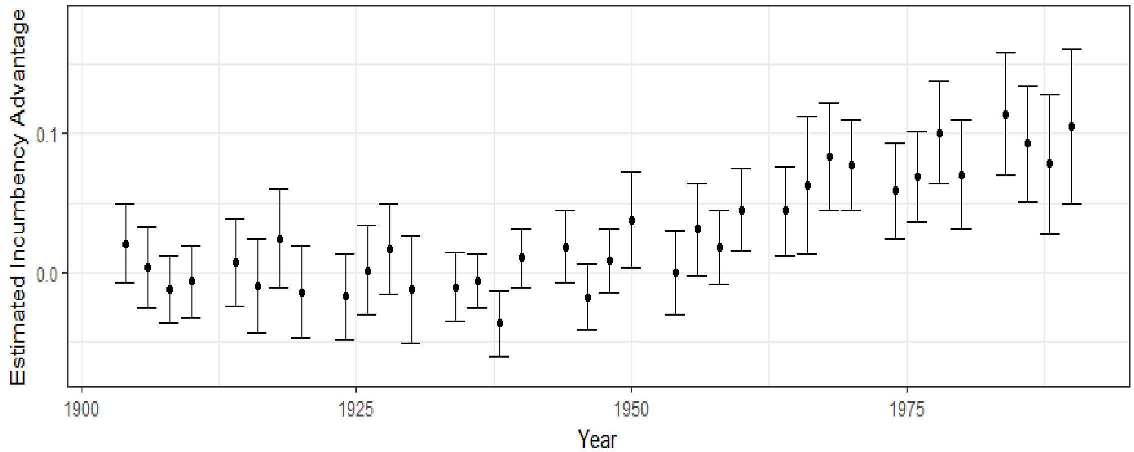

Figure 1: Linear regression incumbency effect for all years

In Figure1, we could have a complete view of the incumbency's effect from linear model among every elections ranges from 1900 to 1992. There is an increasing trend on the incumbency's effect. Before 1960, the 95% confidence intervals contain 0 so it is hard to say current incumbency parties have any advantage on current elections. However, after 1960, an incumbency party is more like to have a higher election vote proportion.

Bayesian with g-prior

The second model we tried is Bayesian Linear Regression model. In order to assess changes over time, we run a separate regression for each election year in our study. As an initial analysis, we estimate separate regressions for each of the election years in the twentieth century, excluding election years immediately following redrawing of the district boundaries, for it is difficult to define incumbency in those years.

A Bayesian analysis of a regression model requires specification of the prior parameters. Finding values of these parameters that represent actual prior information can be difficult. In this case, it's easily to find that the analysis must be done in the absence of precise prior information or information that is easily converted into the parameters of a conjugate prior distribution. We could stick to least square estimation, with the drawback that probability statements about beta would be unavailable. Alternatively, it is sometimes possible to justify a prior distribution with other criteria. Our idea is that, if the prior distribution is not going to represent real prior information about the parameters, then it should be as minimally informative as possible. The resulting posterior distribution would then represent the posterior information of someone who began with little knowledge of the population being studied. One type of weakly informative prior is the unit informative prior. A unit informative prior is one that contains the same amount of information as that would be contained in only a single observation. Another principle for constructing a prior distribution for beta is based on the idea that the parameter estimation should be invariant to changes in the scale of the regresses. Thus we decide to use the g-prior for this model fitting.

For g-prior in Bayesian analysis, we consider a regression model

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$

With a Jeffery's prior on σ^2

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}$$

And a prior called a g-prior on the vector of regression coefficients β

$$\beta|\sigma^2 \sim MVN(0, g\sigma^2(X^T X)^{-1})$$

We can get the priors

$$\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$$

$$\beta|\sigma^2 \sim MVN(0, g\sigma^2(X^T X)^{-1})$$

The sampling model is

$$y \sim MVN(X\beta, \sigma^2 I)$$

The Posterior is

$$\sigma^2|y \sim Inv - \chi^2(\nu_0 + n, \frac{\nu_0\sigma_0^2 + SSR_g}{\nu_0 + n})$$

Where

$$SSR_g = \mathrm{y}^{\mathrm{T}}(I - \frac{g}{g+1} X(X^T X)^{-1} X^T)y$$

With this assumption, we can acquire the posterior mean and variance as following:

$$\mathrm{E}[\beta|y, X, \sigma^2] = \frac{g}{g+1}(X^T X)^{-1} X^T y$$

$$\mathrm{Var}[\beta|y, X, \sigma^2] = \frac{g}{g+1}\sigma^2(X^T X)^{-1}$$

In the simulation, we can demonstrate in two steps:

First simulate σ^2 from $p(\sigma^2|y)$

Then use the σ^2 to simulate β from $p(\beta|y, \sigma^2)$

Similar to the OLS method, we use the data of 1986 and 1988 to fit a Bayesian regression model with the g-prior that $\nu_0 = 1$ and $g = n = 435$. We get the similar graphs and posterior distribution of β as following:

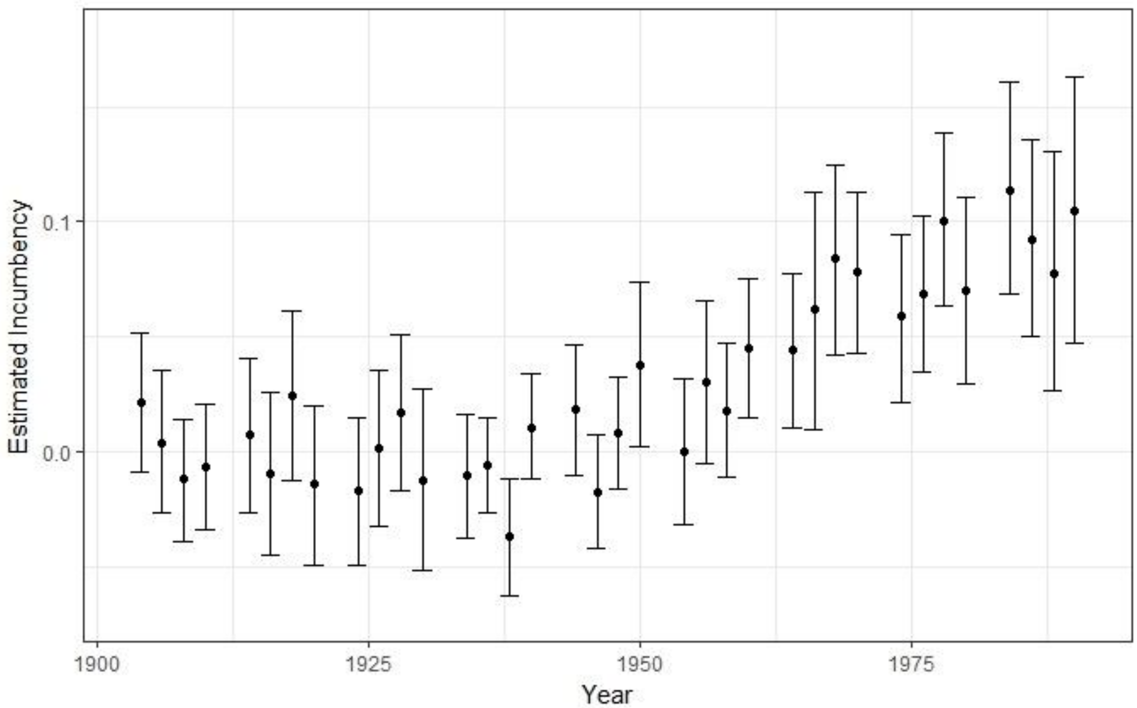

Figure 2: Bayesian Linear Regression for 1988

Variables	*Bayesian Result*		
	Median	*2.5%*	*97.5%*
β_0	*0.314*	*0.233*	*0.394*
β_1	*0.078*	*0.026*	*0.129*
β_2	*0.465*	*0.375*	*0.551*
β_3	*0.009*	*-0.005*	*0.025*

Table 2: Bayesian Linear regression Posterior Estimation for 1988

With a large g (over 400), we get a similar result to traditional linear regression model with a larger 95% confidence interval for the posterior distribution. Posterior standard deviations of these parameters are (0.212 0.162 0.203 0.08), based on 5000 independent Monte Carlo samples generated using R-code. We can conclude that the Bayesian model fit well and the incumbency advantage exists with increasing influence as time goes on.

Hierarchical Model

The third model we have fit is a Bayesian hierarchical model which considers the year as a grouping factor. This model is actually a combination of two separate models which are a regression model to describe within-group variation, and a multivariate normal model to describe heterogeneity among regression coefficients across the groups. In the third model, we will use exactly the same predictors as the ordinary least square model with the year as a grouping parameter. Mathematically, our model can be written as:

$$\mathrm{Y}_{\mathrm{i,j}} = \theta^T x_{i,j} + \gamma_j^T x_{i,j} + \epsilon_{i,j}$$

where

θ is the fixed effect as same as the previous linear model,

γ_j is the random effect, as in our case, the year.

Let $\beta_j = \theta + \gamma_j$, and we use semi-conjugate priors distribution. We could use Gibbs sampling to approximate posterior distribution with full conditional distributions. The priors and full conditional posteriors are:

$$\epsilon_{i,j} \sim N(0, \sigma^2), \beta_1 \ldots \beta_m \sim MVN(\theta, \Sigma)$$

$$\theta \sim MVN(\mu_0, \Lambda_0)$$

$$\Sigma \sim inv - Wishart(\eta_0, S_0^{-1})$$

$$\sigma^2 \sim inv - gamma(\nu_0/2, \nu_0\sigma_0^2/2)$$

$$\theta|\beta_1..\beta_m, \Sigma \sim MVN((\Lambda_0^{-1} + m\Sigma^{-1})^{-1}, (\Lambda_0^{-1} + m\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + m\Sigma^{-1}\bar{\beta})$$

$$\Sigma|\theta, \beta_1 \ldots \beta_m \sim inv - Wishart(\eta_0 + m, \left[S_0 + \sum(\beta_j - \theta)(\beta_j - \theta)^T\right]^{-1})$$

$$\sigma^2 \sim inv - gamma((\nu_0 + \sum n_j)/2, [\nu_0\sigma_0^2 + \sum\sum(y_{i,j} - \beta_j^T x_{i,j})^2]/2)$$

Here, we choose to use an unbiased but weak prior. We take μ_0 to be equal to the average of the ordinary least square regression estimates and the prior variance Λ_0 to be their sample covariance. Similarly, we let the prior sum of squares matrix S_0 to be equal to the covariance of the least square estimates and prior degree of freedom η_0 to be $p + 2 = 6$, so the prior distribution Σ has an expectation equal to the sample covariance of the ordinary least squares estimates. Additional, we let σ_0^2 to be the average of the within-group sample variance and $\nu_0 = 1$. We want to compare the difference of incumbency effect between using ordinary least square regression and Bayesian hierarchical regression with exactly the same predictors.

Variables	*Bayesian hierarchical result*				
	2.5%	*25%*	*50%*	*75%*	*97.5%*
β_0	*0.022*	*0.128*	*0.182*	*0.234*	*0.337*
β_1	*-0.042*	*0.006*	*0.029*	*0.054*	*0.101*
β_2	*0.395*	*0.595*	*0.700*	*0.803*	*1.012*
β_3	*-0.053*	*-0.010*	*0.010*	*0.031*	*0.073*

Table 3: Bayesian hierarchical regression

The Bayesian hierarchical model result is shown in Table3 and it is the overall averaged performance of all groups. With $\beta_0, \beta_1, \beta_2, \beta_3$ standing for the parameters of the same predictors with ordinary least square regression in Table1, we can see the results in Table3 are similar to the results in Table1.

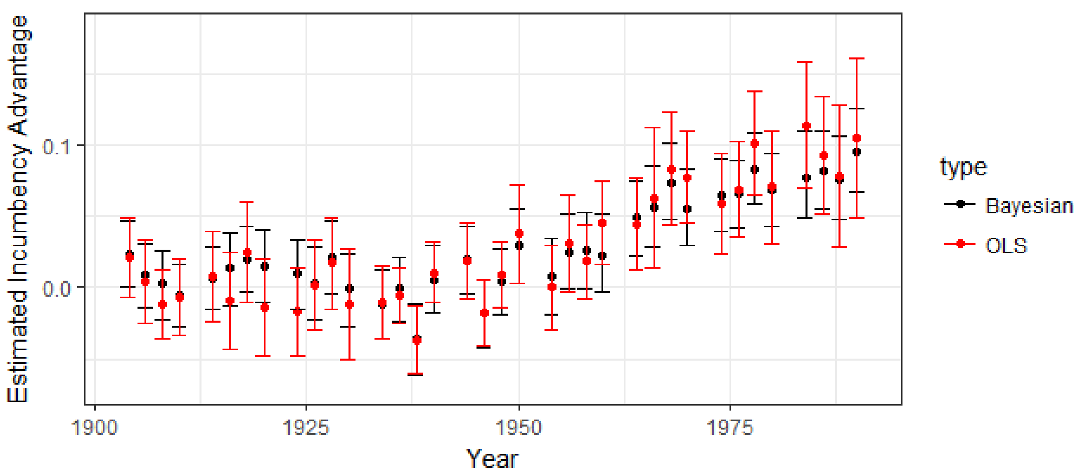

Figure 3: Estimated incumbency advantage from Bayesian model

By imputing the year grouping factor back to the results, we can show the result for difference years. In Fugure3, we also compare the incumbency effect by years from the two different models, and we see the Bayesian hierarchical model gives us a similar main trend of incumbency effect from 1990 to 1992. Additionally, Bayesian model provides us wider posterior 95% confidence intervals compared to the ordinary least square result. From Monte Carlo approximation, we can calculate the posterior probability that the incumbency effect is less or equal to 0 is around 80%. It is also a strong proof that the incumbency effect will have a positive effect to the current incumbency party's votes proportion. With a closer look to Figure3, we can find the posterior estimates of incumbency effect for difference years tend to move closed to the overall posterior mean of incumbency effect, in other words, the Bayesian hierarchical model is able to share information across groups, shrinking extreme regression lines towards the across-group average. The result estimates are more convincible compared to the ordinary least square model.

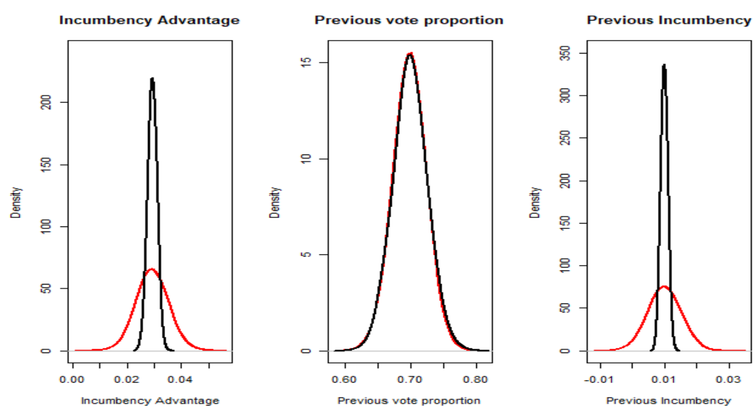

Figure 4: Comparison of prior(black) and posterior(red) distributions

From a comparison between the prior and posterior distributions for the mean of βs in Figure4. We can see the posterior distributions have a strong alteration in our prior information about βs.

Diagnostics

In this section, we implement some popular MCMC diagnostics methods to evaluate our result precision. The ideal sampling result will have a low autocorrelation with a high effective sample size as to obtain a stationary Markov chain. We calculate the effective sample size for the simulated βs very close to the simulation iteration. The autocorrelation function plot and trace plot for βs are shown below.

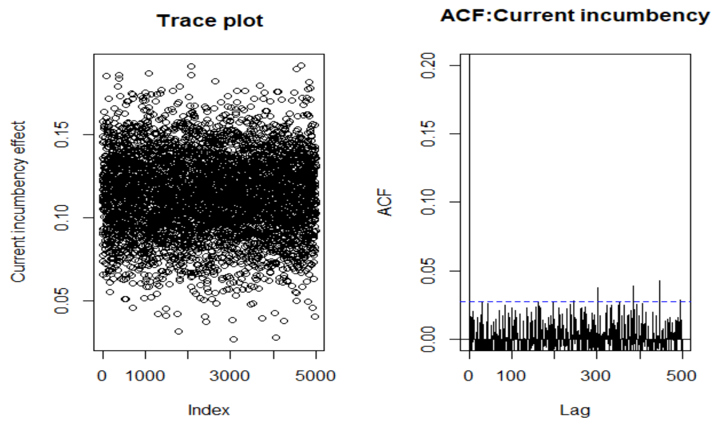

Figure 5: Diagnostic plots for the Bayesian model for year 1984

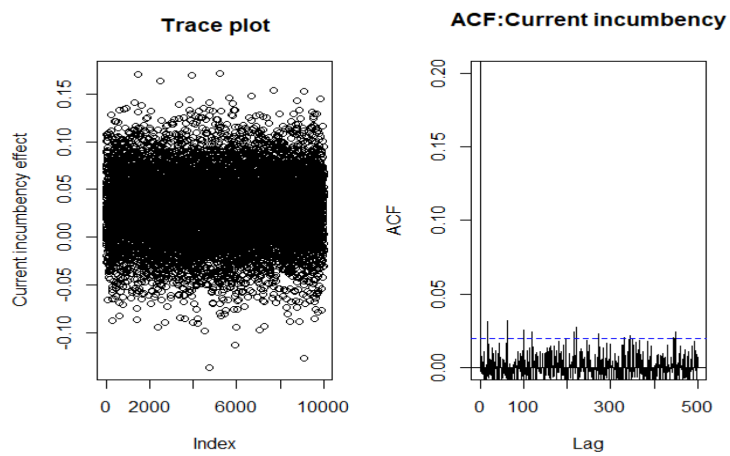

Figure 6: Diagnostic plots for the Bayesian Hierarchical Model for all years

The above two figures both show us an immediate convergence and a low degree of autocorrelation. For the first 500 lags, the autocorrelation function values are essentially equal to zero for approximation purposes. These plots indicate an approximately independently sampled sequence for incumbency effect. The Gibbs sampler for both plain Bayesian model and Bayesian hierarchical model performs quite well.

Conclusion

This project studied the incumbency advantage from the data set of congressional elections. As an initial analysis, we estimate separate regressions for each of the election years in the twentieth century, excluding election years immediately following redrawing of the district boundaries, for it is difficult to define incumbency in those years. Posterior means and 95% posterior intervals for the coefficient for the incumbency are displayed for each election year. As usual, we can use posterior simulations of the regression coefficients to compute any quantity of interest. For example, the increase from the average incumbency advantage in the 1950s to the average advantage in the 1980s has a posterior mean of 0.05.

These results are based using incumbent party and previous election result as control variables (in addition to the constant term). Including more control variables to account for earlier incumbency and election results did not substantially change the inference about the coefficient of the treatment variable, and in addition make the analysis more difficult because of complications such as previous elections that were uncontested. We demonstrated three different methods to acquire results in different version.

Comparing the three models we use, we can find that they all work well and give similar results with some difference in details. The whole trends of incumbency's influence increased in the last century with the median coefficient grew from nearly 0.02 to 0.1. The Bayesian regression model performs larger quantile than the original linear regression model while the hierarchical model is the most smoothie one. We can conclude from all three models that the incumbency has positive influence on the election results and the influence is more significant in recent years.

Appendix

Contribution

Haiwei Zhou: Finished the Bayesian Linear Model part, wrote the introduction, Bayesian parts and summary of the report.

Zijian Han: Collected and cleaned the data, finished the simulation of OLS, Hierarchical Model and Diagnostic and wrote the report of relative parts.

We discussed all the methods and analyze the simulation results together.

R code

````
```{r}
library(tidyr)
library(dplyr)
library(mvtnorm)
library(ggplot2)
```
```{r}
# read data
data_file_path <-  "./datafolder/"
file_list.asc <- list.files(data_file_path)
file_list_dir <- lapply(file_list.asc, function(x) paste(data_file_path,x,sep = ""))
file_list_dir <- unlist(file_list_dir)

## Exception file
data_excp <- read.table('./datafolder/excepth.asc',header = TRUE)

length(file_list_dir) #49 files
```

```{r}
# merge data together with year indicator
years <- seq(1896,1992,2)

data_all <- NULL
for(i in 1:length(years))
{
  data_sub <- read.table(file_list_dir[i])
  data_sub$year <- years[i]
````
````

```
  data_all <- rbind(data_all,data_sub)
}



dim(data_all)
# change name
colnames(data_all) <- c("State","District","Incumbency_Code","Democratic_votes","Republican_votes","Year")

Minus9_to_NA <- function(col)
{
  col[which(col == -9)] <- NA
  return(col)
}

data_all <- apply(data_all,2,Minus9_to_NA)
data_all <- as.data.frame(data_all)
```

Linear model
```{r}
data_fill <- data_all

data_fill$Incumbency <- NA
data_fill$Incumbency_vote <- NA
### Get Incumbency Votes
Rep_inc <- which(data_fill$Incumbency_Code == -1)
Dem_inc <- which(data_fill$Incumbency_Code == 1)
Open_seat <- which(data_fill$Incumbency_Code == 0)

data_fill$Incumbency_vote[Rep_inc] <- data_fill$Republican_votes[Rep_inc]/(data_fill$Republican_votes[Rep_inc]+data_fill$Democratic_votes[Rep_inc])
data_fill$Incumbency[Rep_inc] <- 1


data_fill$Incumbency_vote[Dem_inc] <- data_fill$Democratic_votes[Dem_inc]/(data_fill$Republican_votes[Dem_inc]+data_fill$Democratic_votes[Dem_inc])
data_fill$Incumbency[Dem_inc] <- 1

for(i in Open_seat)
{
data_fill$Incumbency_vote[i] <- max(data_fill$Democratic_votes[i]/(data_fill$Republican_votes[i]+data_fill$Democratic_votes[i]),data_fill$Republican_votes[i]/(data_fill$Republican_votes[i]+data_fill$Democratic_votes[i]))
}
data_fill$Incumbency[Open_seat] <- 0
```

```
lm_model <- function(t)
{
  ref_year <- t-2
  data_ols <- data.frame(ref_prop_vote = data_fill[which(data_fill$Year ==
ref_year),'Incumbency_vote'], y =data_fill[which(data_fill$Year ==
t),'Incumbency_vote'],incumbency = data_fill[which(data_fill$Year ==
t),'Incumbency'],incumbency_part = data_fill[which(data_fill$Year ==
ref_year),'Incumbency_Code'])

  model <- lm(y~. , data = data_ols)
  return(as.numeric(confint(model, 'incumbency', level=0.95)))
}

r = NULL

year.period1 <-
c(1904,1906,1908,1910,1914,1916,1918,1920,1924,1926,1928,1930,1934,1936,1938,194
0,1944,1946,1948,1950,1954,1956,1958,1960,1964,1966,1968,1970,1974,1976,1978,1980
,1984,1986,1988,1990)
for(year in year.period1)
{
  ci <- lm_model(year)
  r <- rbind(r,ci)
}

colnames(r) <- c("L","U")
r <- as.data.frame(r)
r$point <- apply(r,1,mean)
r$year <- year.period1


ggplot(r, aes(x = year, y = point)) + geom_point(size = 1.5) + ylab("Estimated Incumbency
Advantage") + xlab("Year") + geom_errorbar(aes(ymax = U,ymin = L)) + theme_bw() +
ylim(-0.07,0.18)
```

```{r}
Yr <- seq(1900,1990,2)
Yr <- Yr[-seq(1,46,5)]

result <- matrix(NA,nrow = 36,ncol = 13)
for(i in 1:36){
  y <- subset.data.frame(data_fill,data_fill$Year==Yr[i]+2,select = c('Incumbency_vote'))
  X <- cbind(y,data_fill[which(data_fill$Year ==
Yr[i]),"Incumbency_vote"],data_fill[which(data_fill$Year ==
```

```
Yr[i]+2),"Incumbency"],rep(1,length(y)),data_fill[which(data_fill$Year ==
Yr[i]),"Incumbency_Code"])
 X <- na.omit(X)
 y <- as.vector(X[,1])
 X <- as.matrix(X[,-1])

 #  Read data dimensions
 n <- dim(X)[1] ; p <- dim(X)[2]
 # Set prior parameters
 nu0 <- 1; g <- n
 s20 <- summary(lm(y ~ X ))$sigma^2
 mXX <- solve(t(X)%*%X)
 # Posterior calculations
 Hg <- (g/(g+1)) * X%*%mXX%*%t(X)
 SSRg <- t(y)%*%( diag(1,nrow=n)-Hg )%*%y
 Vbeta <- (g/(g+1))*mXX
 Ebeta <- Vbeta%*%t(X)%*%y
 # Simulate sigma^2 and then beta|sigma^2
 s2.post <- beta.post<-NULL
 for(s in 1:5000)
 {
  s2.post<-c(s2.post, (nu0*s20+SSRg)/rchisq(1,nu0+n))
  beta.post<-rbind(beta.post, rmvnorm(1,Ebeta,s2.post[s]*Vbeta))
 }

 temp <- apply(beta.post,2,quantile,probs=c(.025,.5,.975))
 result[i,] <- c(Yr[i]+2,temp[,1],temp[,2],temp[,3],temp[,4])
}
View(result)
colnames(result) <- c('Year','Prop 2.5%','Prop 50%','Prop 97.5%',
                  'Incumbency 2.5%','Incumbency 50%','Incumbency 97.5%',
                  'Constant 2.5%','Constant 50%','Constant 97.5%','Incumbency code
2.5%','Incumbency code 50%','Incumbency code 97.5%')

result <- data.frame(result)
attach(result)
ggplot(data = result, aes(x = Year, y = Incumbency.50.))+geom_point(size = 1.5) +
geom_errorbar(aes(ymax = Incumbency.97.5.,ymin = Incumbency.2.5.))+ ylab('Estimated
Incumbency') + theme_bw()+ylim(-0.07,0.18)
```

Bayesian Hierarchical model

```{r}
## Manipulate data
year.period1 <-
c(1904,1906,1908,1910,1914,1916,1918,1920,1924,1926,1928,1930,1934,1936,1938,194
0,1944,1946,1948,1950,1954,1956,1958,1960,1964,1966,1968,1970,1974,1976,1978,1980
,1984,1986,1988,1990)
```

```
## Ad-hoc estimates from OLS

lm_model <- function(t)
{
  ref_year <- t-2
  data_ols <- data.frame(y =data_fill[which(data_fill$Year ==
t),'Incumbency_vote'],previous_vote = data_fill[which(data_fill$Year ==
ref_year),'Incumbency_vote'],previous_incum = data_fill[which(data_fill$Year ==
ref_year),'Incumbency_Code'], current_incum = data_fill[which(data_fill$Year ==
t),'Incumbency'])

  model <- lm(y~. , data = data_ols)
  return(as.numeric(model$coefficients))
}

OLS.Beta <- NULL
for(t in year.period1)
{
  r <- lm_model(t)
  OLS.Beta <- rbind(OLS.Beta,r)
}

lm_model_red <- function(t)
{
  ref_year <- t-2
  data_ols <- data.frame(y =data_fill[which(data_fill$Year ==
t),'Incumbency_vote'],previous_vote = data_fill[which(data_fill$Year ==
ref_year),'Incumbency_vote'],previous_incum = data_fill[which(data_fill$Year ==
ref_year),'Incumbency_Code'], current_incum = data_fill[which(data_fill$Year ==
t),'Incumbency'])

  model <- lm(y~. , data = data_ols)
  return(residuals(model))
}

RESI <- NULL

for(t in year.period1)
{
  resi_var <- var(lm_model_red(t))
  RESI <- rbind(RESI,resi_var)
}

#sigma2
s20 <- mean(RESI)

# Theta
mu0 <- apply(OLS.Beta,2,mean) # Take the means for each column

# Sigma
```

```
L0<-var(OLS.Beta) # Variance as variance of prior
#L0

merge_data <- NULL

for(year in year.period1)
{
  data_merge <- data_frame(current_vote = data_fill[which(data_fill$Year ==
year),"Incumbency_vote"],previous_vote = data_fill[which(data_fill$Year == year-
2),"Incumbency_vote"],current_incum = data_fill[which(data_fill$Year ==
year),"Incumbency"], previous_incum = data_fill[which(data_fill$Year == year-
2),"Incumbency_Code"],State = data_fill[which(data_fill$Year == year),"State"],District =
data_fill[which(data_fill$Year == year),"District"],Year = data_fill[which(data_fill$Year ==
year),"Year"])
  merge_data <- rbind(merge_data,data_merge)
}

# Eliminate the "at-large" districts
merge_data <- merge_data[-which(merge_data$District > 50),]
# Eliminate NA values
merge_data <- na.omit(merge_data)
# Construct of Y and X
Y <- NULL
for(i in 1:length(year.period1))
{
  Y[[i]] <- merge_data$current_vote[which(merge_data$Year == year.period1[i])]
}

X.l <- NULL
for(i in 1:length(year.period1))
{
  X.l[[i]] <- cbind(rep(1,length(which(merge_data$Year ==
year.period1[i]))),merge_data$previous_vote[which(merge_data$Year ==
year.period1[i])],merge_data$previous_incum[which(merge_data$Year ==
year.period1[i])],merge_data$current_incum[which(merge_data$Year == year.period1[i])])
}

# Define two functions
rmv<-function(n,mu,Sigma){ # samples Y~MVN(mu,Sigma)
cm<-chol(Sigma);d<-dim(Sigma)[1]
Y0<-matrix(rnorm(n*d),nrow=d)
t(cm)%*%Y0 + mu
}
riw<-function(n,nu0,Sigma){ # Sigma~IW(nu0,Sigma^(-1)); requires rmv
m<- solve(Sigma)
sapply(1:n,function(i)
solve(crossprod(t(rmv(nu0*n,0,m))[(i-1)*nu0+1:nu0,])), simplify = 'array')
}

year.count <- merge_data %>% group_by(Year) %>% summarise(n = n())
```

```
n <- year.count$n
m <- length(year.period1)
eta0 <- 6; S0 <- L0
nu0 <- 1; s02 <- s20
# Gibbs Sampler initial values
beta<-mu0
S2<-S0
s2<-s02
Beta<-list(); Beta.t<-TH<-Sigma<-sigma2<-NULL
S<-10000



set.seed(1)
for (s in 1:S)
{
# update theta
Lm<-solve(solve(L0)+m*solve(S2)); Bm<-mu0
mum <- Lm%*%(solve(L0)%*%mu0+m*solve(S2)%*%Bm)
theta <- c(rmv(1,mum,Lm))
# Update beta for each group
mS2<- solve(S2)
mS <- lapply(1:m,function(j) solve(mS2+t(X.l[[j]])%*%X.l[[j]]/s2))
beta<-sapply(1:m,function(j)
rmv(1,mS[[j]]%*%(mS2%*%theta+t(X.l[[j]])%*%Y[[j]]/s2),mS[[j]]))
# Update sigma^2
s2 <- (nu0*s02+do.call('sum',lapply(1:m,function(j) crossprod(Y[[j]]-
X.l[[j]]%*%beta[,j]))))/rchisq(1,nu0+sum(n))
# Update Sigma
St<-crossprod(t(beta-theta))
S2<-riw(1,eta0+m,S0+St)[,,1]
# Predictive distribution of beta
beta.t <- rmv(1,theta,S2)
Beta[[s]]<-beta; Beta.t<-rbind(Beta.t,c(beta.t))
sigma2<-c(sigma2,s2);Sigma<-rbind(Sigma,c(S2));TH<-rbind(TH,theta)
}

Beta.p.array <- array(0,dim = c(10000,m,4))

apply(Beta.t,2,function(x)quantile(x,c(0.025,0.25,0.5,0.75,0.975)))


# Obtain posterior expectations for Beta
for(i in 1:S)
{
group.coef <- Beta[[i]]
 for(j in 1:m)
 {
  Beta.p.array[i,j,] <- group.coef[,j]
```

```
 }
}

M <- NULL
L <- NULL
U <- NULL

for(j in 1:36)
{
  M[j] <- as.numeric(quantile(Beta.p.array[,j,4],0.5))
  L[j] <- as.numeric(quantile(Beta.p.array[,j,4],0.025))
  U[j] <- as.numeric(quantile(Beta.p.array[,j,4],0.975))
}

hierar_incum_p <- data.frame(L = L,U = U,point= M ,year = year.period1)

hierar_incum_p$type <- "Bayesian"
r$type <- 'OLS'

data_to_plot <- rbind(hierar_incum_p,r)


ggplot(data_to_plot, aes(x = year, y = point,color = type)) + geom_point(size = 1.5) +
ylab("Estimated Incumbency Advantage") + xlab("Year") + geom_errorbar(aes(ymax =
U,ymin = L,color = type)) + theme_bw() + ylim(-0.07,0.18) + scale_color_manual(values =
c(1,2))

par(mfrow=c(1,3))
plot(density(TH[,4],adjust = 2),xlab = "Incumbency Advantage",main = "Incumbency
Advantage",col = 2,lwd = 2,ylim=c(0,240))
lines(density(rnorm(1000,mu0[4],S0[4,4]),adjust = 2),col = 1,lwd = 2)

plot(density(TH[,2],adjust = 2),xlab = "Previous vote proportion",main = "Previous vote
proportion",col = 2,lwd = 2)
lines(density(rnorm(1000,mu0[2],S0[2,2]),adjust = 2),col = 1,lwd = 2)

plot(density(TH[,3],adjust = 2),xlab = "Previous Incumbency",main = "Previous
Incumbency",col = 2,lwd = 2,ylim = c(0,350))
lines(density(rnorm(1000,mu0[3],S0[3,3]),adjust = 2),col = 1,lwd = 2)
```

Gibbs sampling diagnostics

```{r}
# previous_vote, current_incumbency, constant, previous incumb
model2_beta <- beta.post
library(coda)
effectiveSize(beta.post[,1])
effectiveSize(beta.post[,4])
acf(beta.post[,1],lag = 1000)
acf(beta.post[,2],lag = 1000)
```

````
par(mfrow = c(2,2))
plot(beta.post[,1],ylab=("Previous vote effect"))
plot(beta.post[,4],ylab=("Previous incumbency effect"))
plot(beta.post[,2],ylab=("Current incumbency effect"))
plot(beta.post[,3],ylab=("Contant term"))
acf(beta.post[,1],lag = 500,main=("ACF:Previous vote"),ylim=c(0,0.2))
acf(beta.post[,4],lag = 500,main=("ACF:Previous incumbency"),ylim=c(0,0.2))
acf(beta.post[,2],lag = 500,main=("ACF:Current incumbency"),ylim=c(0,0.2))
acf(beta.post[,3],lag = 500,main=("ACF:Contant term"),ylim=c(0,0.2))
```


```{r}
y <- subset.data.frame(data_fill,data_fill$Year==Yr[33]+2,select = c('Incumbency_vote'))
  X <- cbind(y,data_fill[which(data_fill$Year ==
Yr[33]),"Incumbency_vote"],data_fill[which(data_fill$Year ==
Yr[33]+2),"Incumbency"],rep(1,length(y)),data_fill[which(data_fill$Year ==
Yr[33]),"Incumbency_Code"])
  X <- na.omit(X)
  y <- as.vector(X[,1])
  X <- as.matrix(X[,-1])

  #  Read data dimensions
  n <- dim(X)[1] ; p <- dim(X)[2]
  # Set prior parameters
  nu0 <- 1; g <- n
  s20 <- summary(lm(y ~ X ))$sigma^2
  mXX <- solve(t(X)%*%X)
  # Posterior calculations
  Hg <- (g/(g+1)) * X%*%mXX%*%t(X)
  SSRg <- t(y)%*%( diag(1,nrow=n)-Hg )%*%y
  Vbeta <- (g/(g+1))*mXX
  Ebeta <- Vbeta%*%t(X)%*%y
  # Simulate sigma^2 and then beta|sigma^2
  s2.post <- beta.post<-NULL
  for(s in 1:5000)
  {
    s2.post<-c(s2.post, (nu0*s20+SSRg)/rchisq(1,nu0+n))
    beta.post<-rbind(beta.post, rmvnorm(1,Ebeta,s2.post[s]*Vbeta))
  }

par(mfrow=c(1,2))

plot(beta.post[,2],ylab=("Current incumbency effect"),main=("Trace plot"))
acf(beta.post[,2],lag = 500,main=("ACF:Current incumbency"),ylim=c(0,0.2))

plot(Beta.t[,4],ylab = ("Current incumbency effect"),main=("Trace plot"))
acf(Beta.t[,4],lag = 500,main=("ACF:Current incumbency"),ylim=c(0,0.2))
```
````
````