251 Project

Anna Leman, Joshua Christensen, Justin Woffinden

12/12/2020

Introduction

For our research question we wanted to evaluate whether or not the proportion of money spent on teacher salary had any effect on the proportion of dropouts for high school aged students and whether this effect differs from red states to blue states. Teacher salaries have been a topic of debate for years. If teacher salaries had a significant impact on the amount of dropouts then this would be good evidence that teachers should be making more money for the benefit of students everywhere. It allows us to see whether the relationship between teacher salary and dropout rates differs between red states and blue states.

Methods

The data for this analysis were gathered from the website Data-Planet, the education department of Wyoming and Wisconsin (see links below). The likelihoods for the red and blue states are both normal distributions with β_0, β_1 and σ^2 ($N \sim (\beta_0 + \beta_1 x_i, \sigma^2)$). This is a linear regression analysis so for our prior distributions we chose a normal distribution for β_0 and β_1 and an inverse gamma distribution for σ^2. The parameters we chose for these distributions were $m_0 = m_1 = 0$, $v_0 = v_1 = 100$ and $a = b = 1$. We chose these because we had very little knowledge as to what the prior distribution parameters were. We plan on using the posterior to see if there is a difference between the relationship of teacher salary and dropouts between red and blue states. The posteriors will give us more information on the actual relationship between teacher salary and dropouts and then we can compare the two. The assumptions of a linear regression analysis include a linear relationship between our variables, independence, normality of the residuals and equal variance about the regression line. For Bayesian Regression we also include the assumptions built into our prior distributions. The residual plots in the assumptions section of the appendix indicate approximate normality of the residuals. The scatterplot of the data shows no obvious curvature so the asssumption of linearity is reasonable. We see no "megaphone" pattern in the scatterplot so the assumption of equal variance is reasonable.

Links: Finance, Dropouts, Enrollment

Results

Exploratory Data Analysis

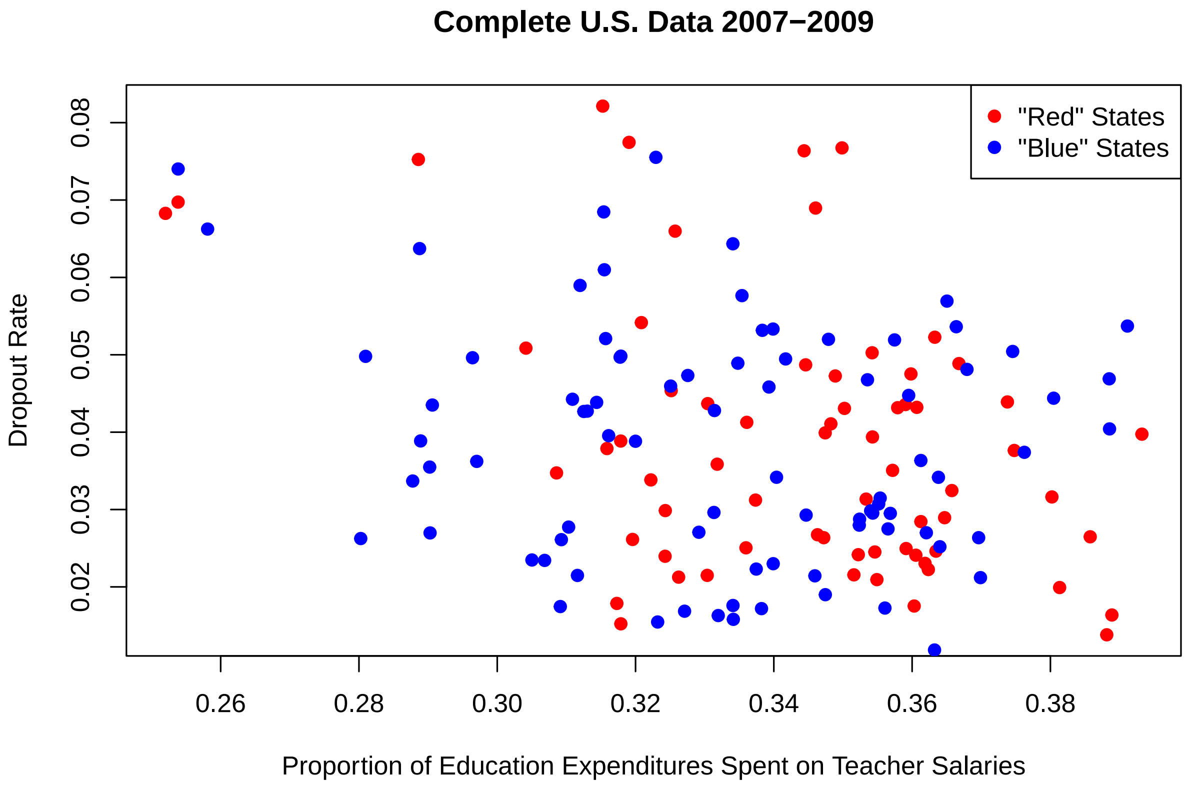
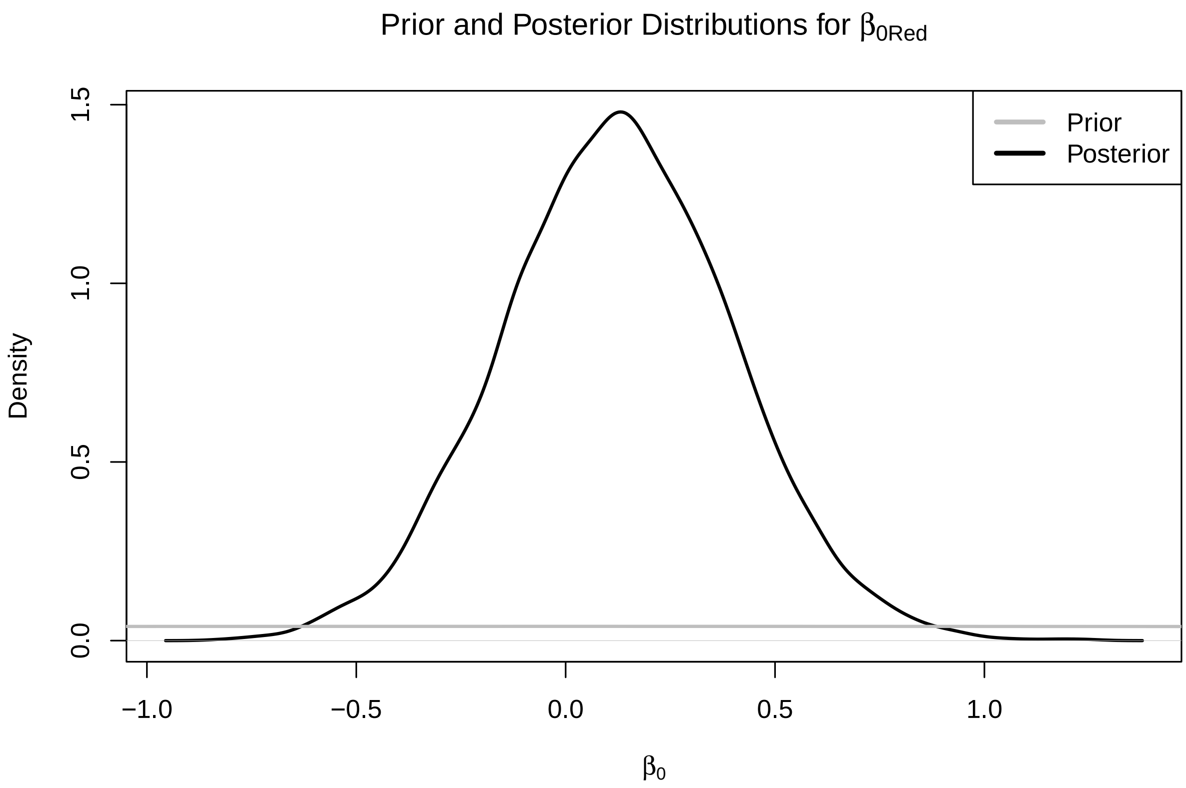


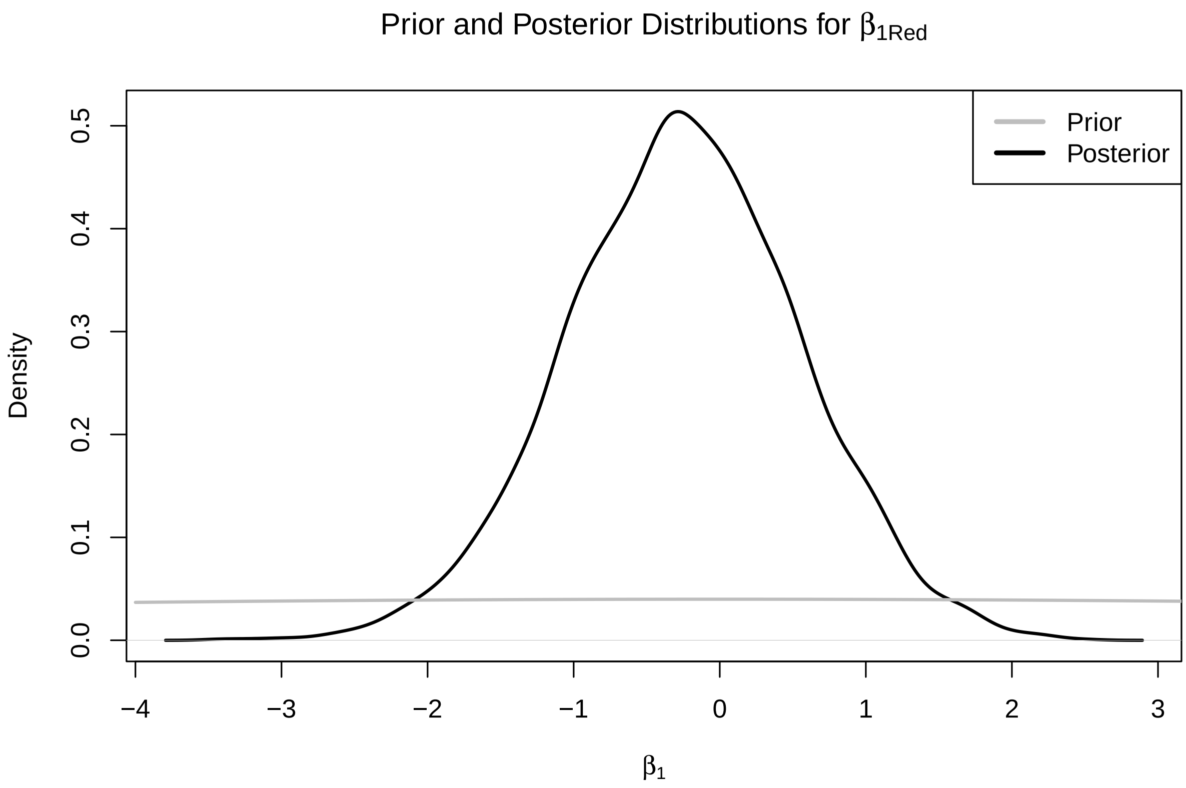
Table 1: Summary Statistics

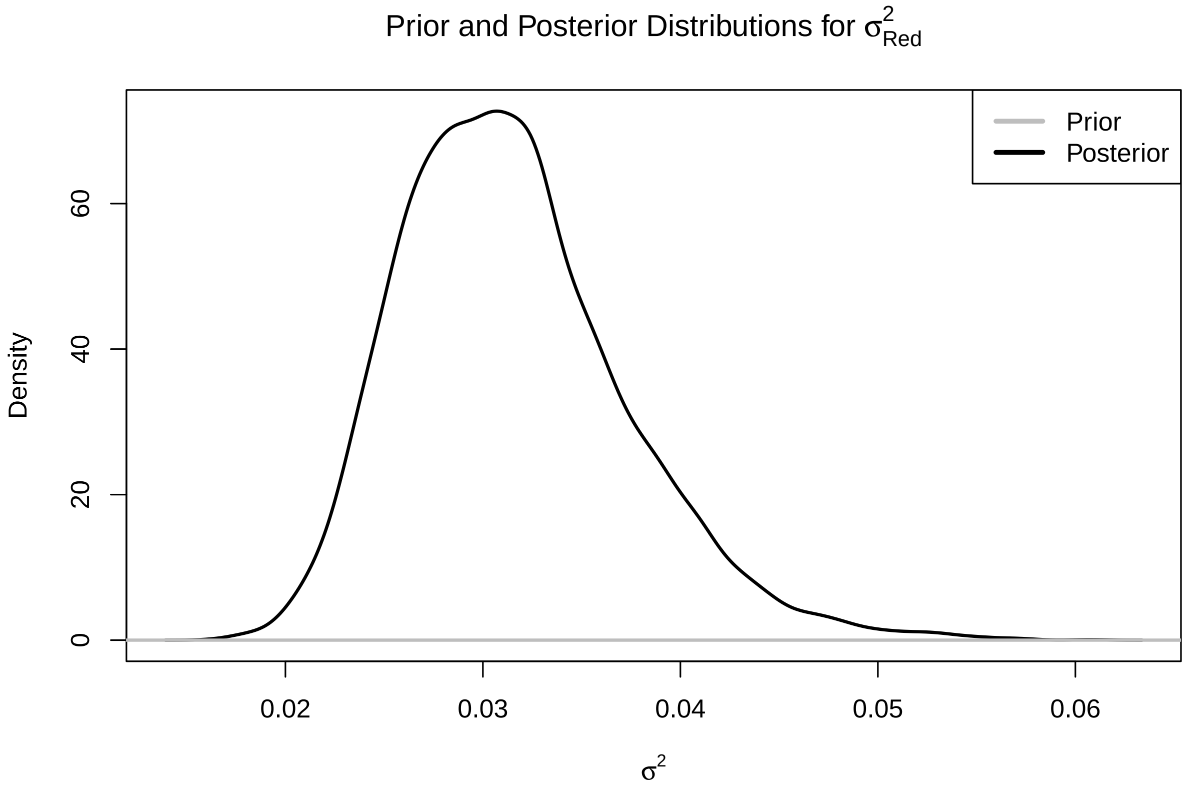
Region	Correlation	R-sq	Mean Dropout Rate	Mean Proportion of Budget for Teacher Salaries
Red States	-0.430	0.185	0.039	0.344
Blue States	-0.120	0.014	0.040	0.334
Overall	-0.249	0.062	0.039	0.338

Convergence and mixing is good in the draws that we kept. We took a burn-in of 100 draws and thinned by taking every 250th draw to achieve the best convergence and mixing in our final draws. See trace and acf plots in the assumptions section of the appendix.

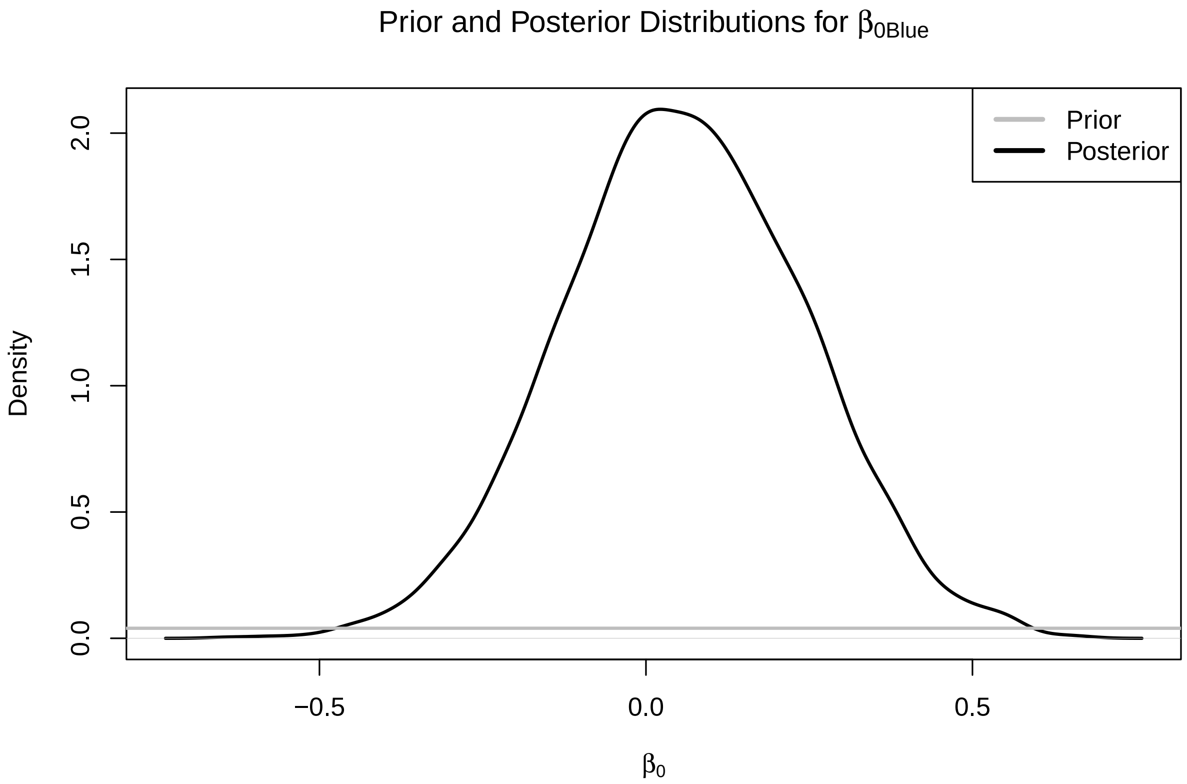
Individual Prior and Posterior Distributions



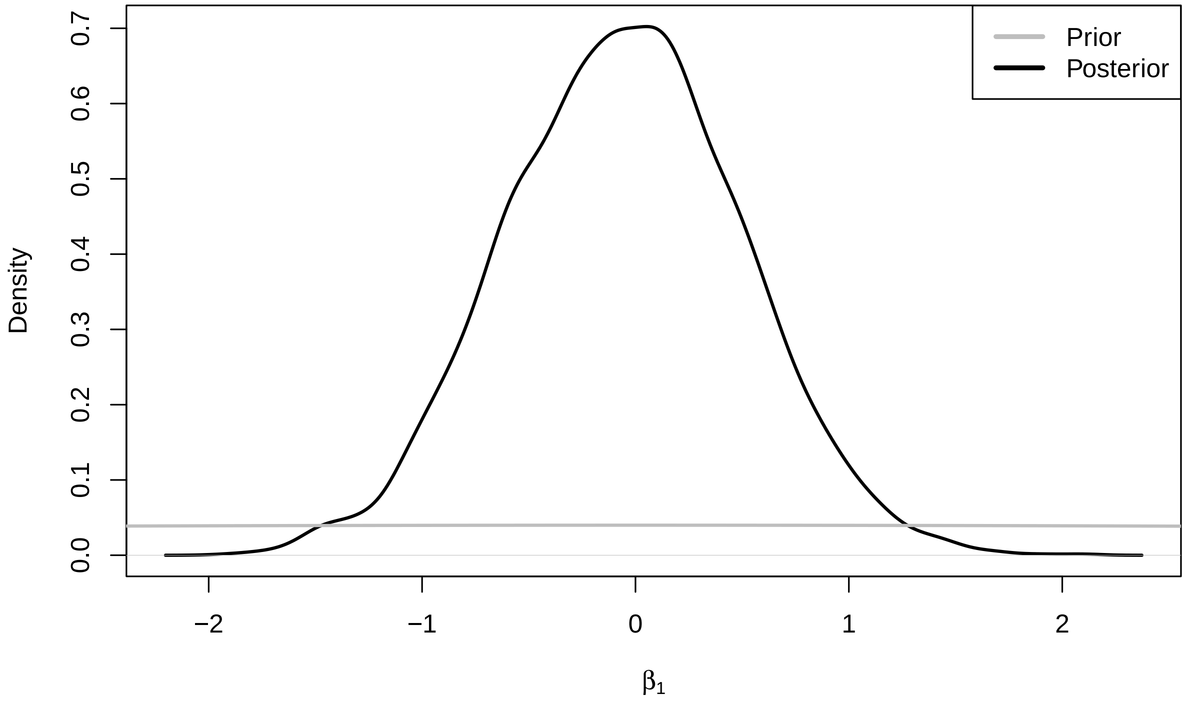




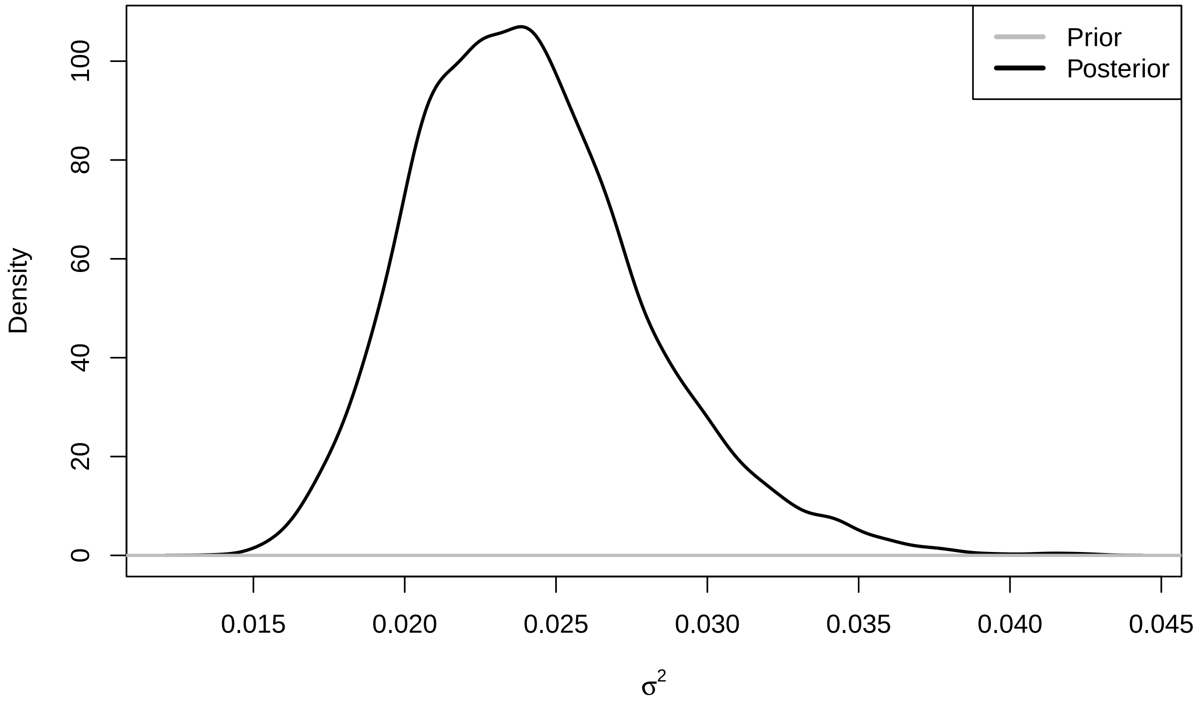
Prior and Posterior Distributions for σ^2_{Red}



Prior and Posterior Distributions for β_{0Blue}



Prior and Posterior Distributions for β_{1Blue}



Prior and Posterior Distributions for σ^2_{Blue}

Comparison of Posterior Distributions for β_{1Red} and β_{1Blue}

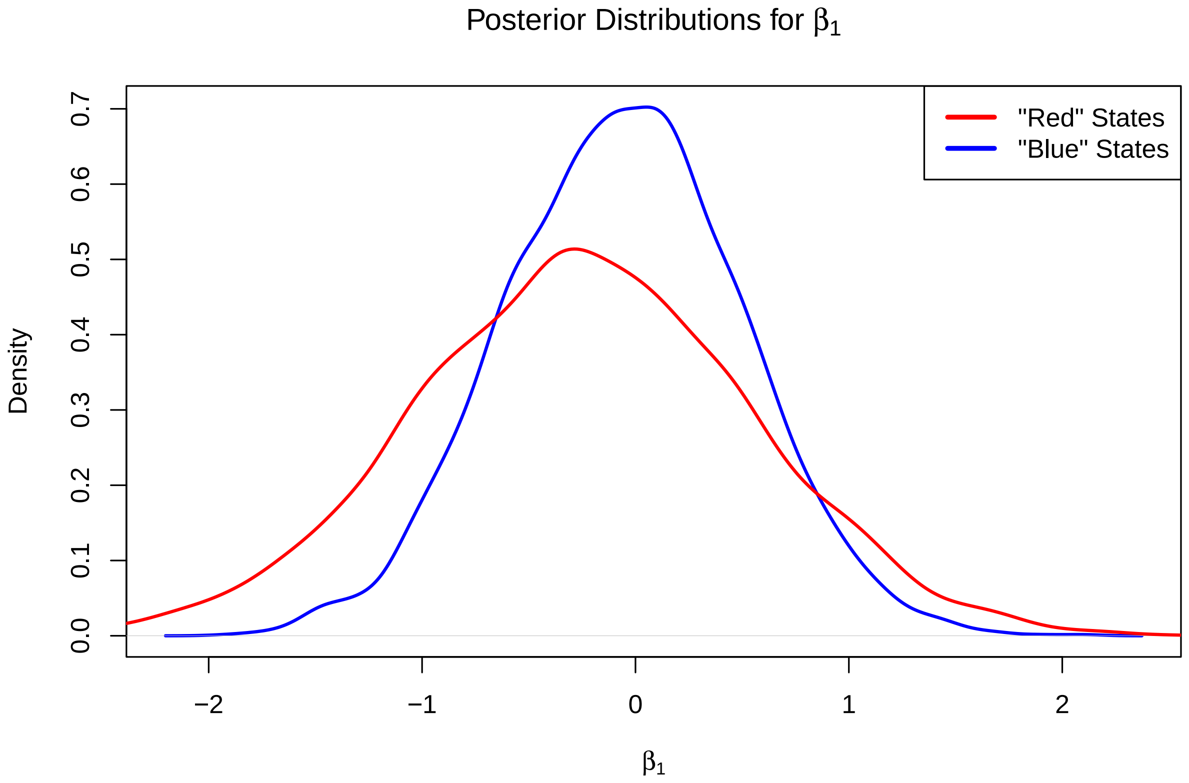
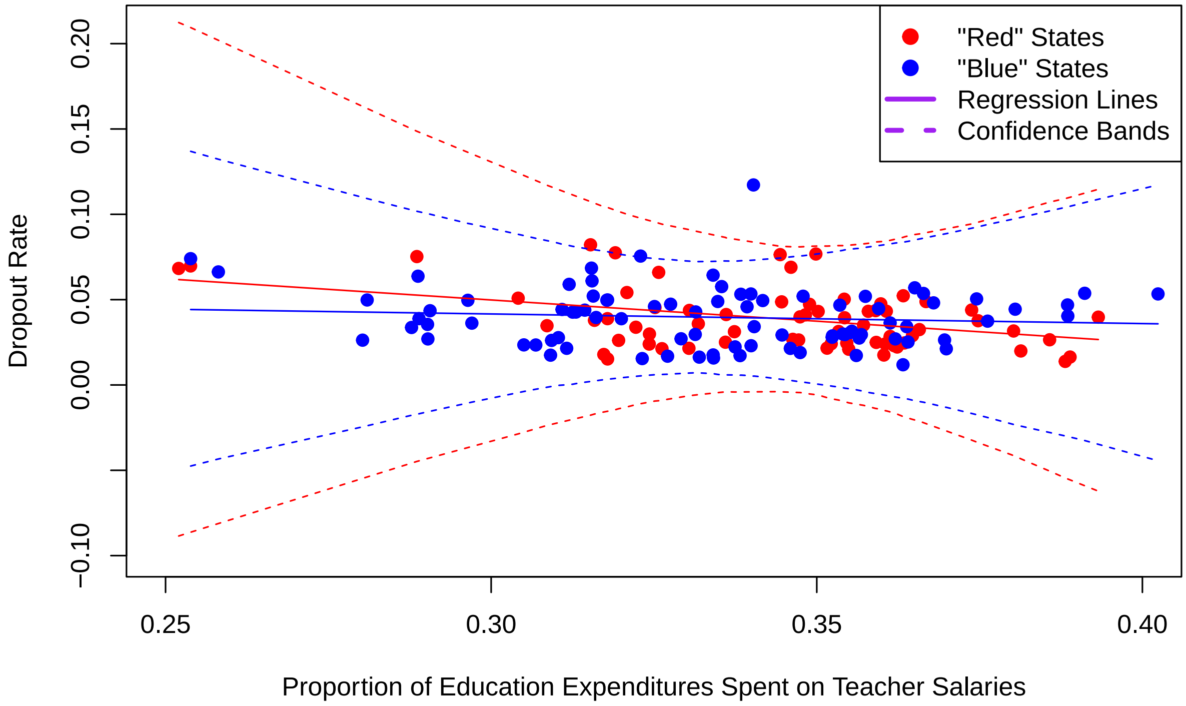


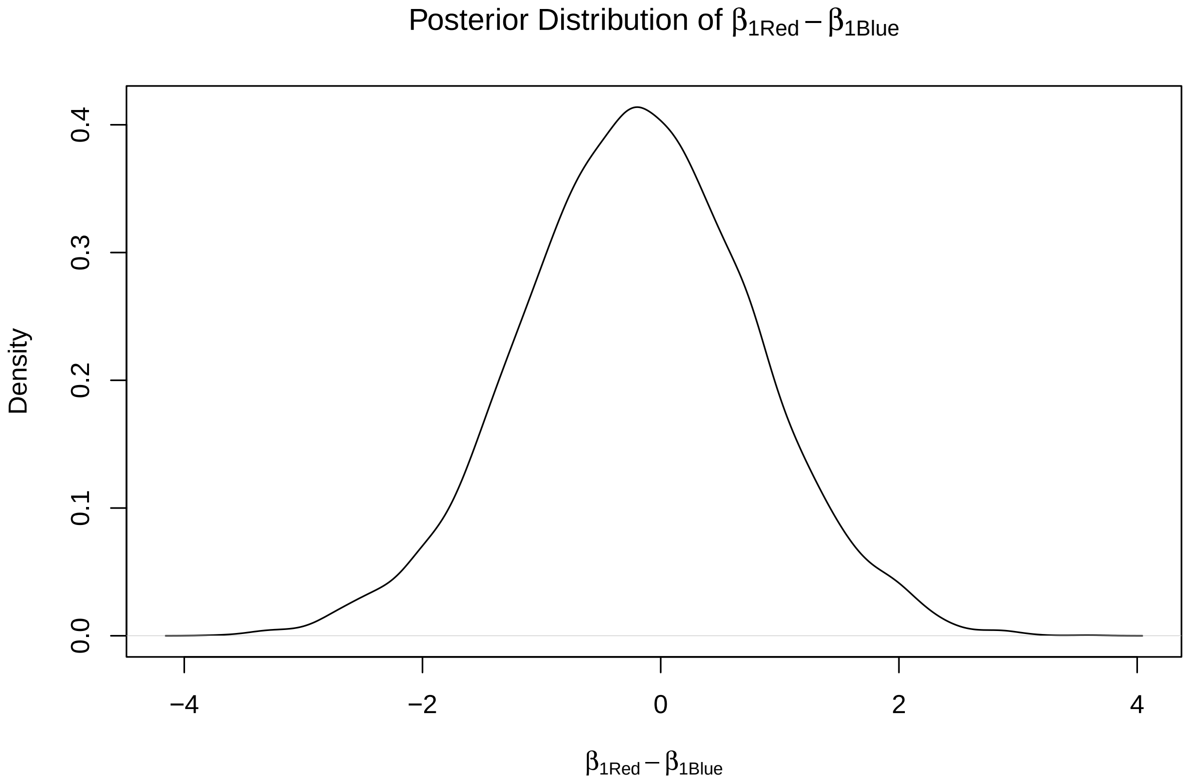
Table 2: Credible Intervals for Population Slopes

Population	Lower Bound	Upper Bound
Red States	-1.838	1.301
Blue States	-1.124	1.019

The confidence interval for β_{1Red} indicates that there is a 95% probability that the true slope of the regression line for red states is captured by the interval (-1.838,1.301).
The confidence interval for β_{1Blue} indicates that there is a 95% probability that the true slope of the regression line for blue states is captured by the interval (-1.124,1.019).


"Red" States
"Blue" States
Regression Lines
Confidence Bands
Dropout Rate
−0.10
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Proportion of Education Expenditures Spent on Teacher Salaries

Inference for the Posterior Distribution of $\beta_{1Red} - \beta_{1Blue}$.



According to our credible interval there is a 95% probability that the difference between the two population slopes ($\beta_{1Red} - \beta_{1Blue}$) is captured by the interval (-2.136, 1.762).

Discussion

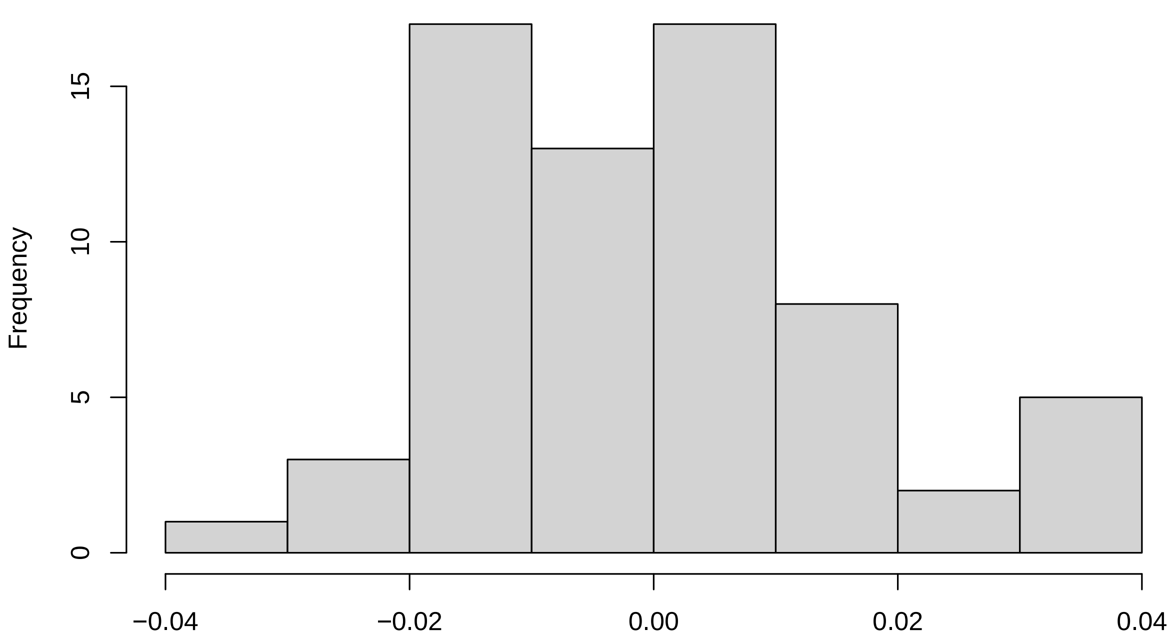
This project shows that teacher pay and dropout rates do not have a significant relationship, and there is no significant difference in the change of dropout rates by teacher pay between states that voted a certain way during the 2008 election. However, there are so many other factors that impact dropout rates that we may have restricted ourselves in our choice of a linear model. Such factors include textbook/student ratios, time spent on homework, and classroom size, but also societal demographics such as child homelessness rates, family layouts (e.g. the presence of a father figure in the home), and the size of school districts, both in area and population. Rather than doing a simple linear regression model, it may be more effective to utilize more of these variables in a multiple linear regression setting that accounts for these variables all at once. It is also important to recognize that dropout rates are not the only indicator of the quality of a high school student's education. We could have easily switched this specific response variable with a variety of other topics, such as student learning aptitude, grades, and college/job acceptance rates. Repeating our same procedure with one of these new response variables or as part of a multiple regression model is worth future exploration.

Another lurking variable that shouldn't be ignored is the existence of swing states in our country's political landscape. We chose to use how each state voted in the 2008 election since that fit the time frame of our dataset. However, there were several states who voted Democratic in 2008 that had voted Republican in 2004, such as Florida, Ohio, and Indiana. Those three states in particular have voted Republican in presidential elections since then. That means the way we separated our data between "red" and "blue" states will change during every presidential election, and hence the results can change. It is also worth noting that how a state votes in a presidential election does not tell the whole story about a state's general political affiliation. For example, states like Indiana and Vermont both voted Democrat in the 2008 presidential election, but both states voted in Republican governors. The opposite can be said for states like Missouri and West Virginia. Similar trends can also be found between how a state votes for the President in relation to their votes for Congressional members. There is great fluidity in the voting and timing of "red" vs "blue" states, so it can be challenging to create a firm definition that separates them.

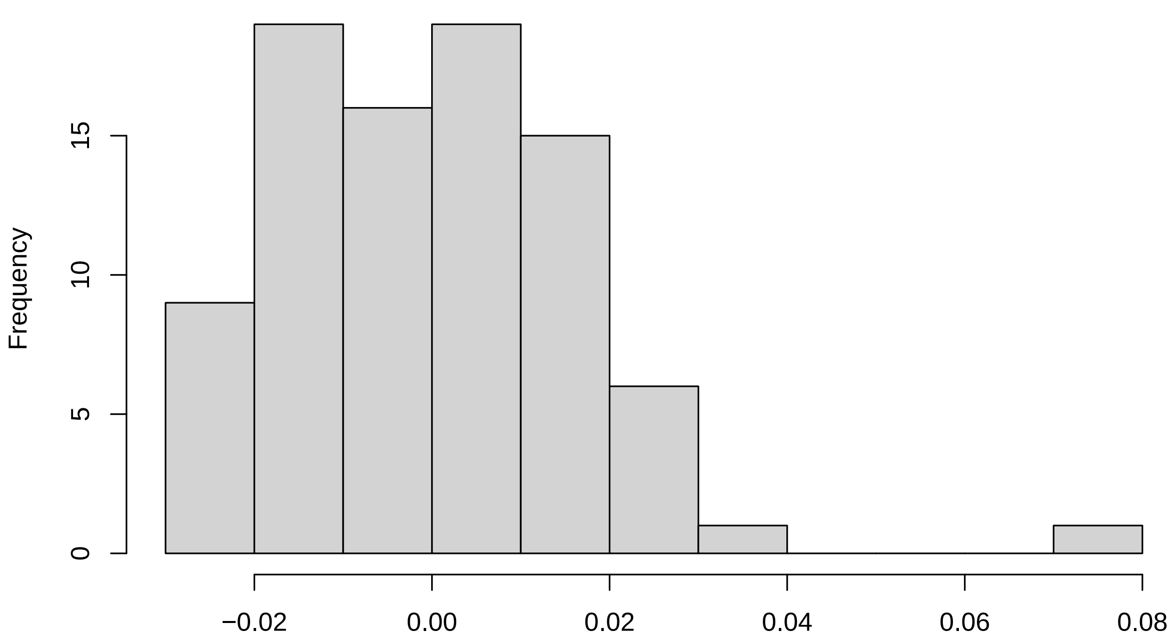
Appendix

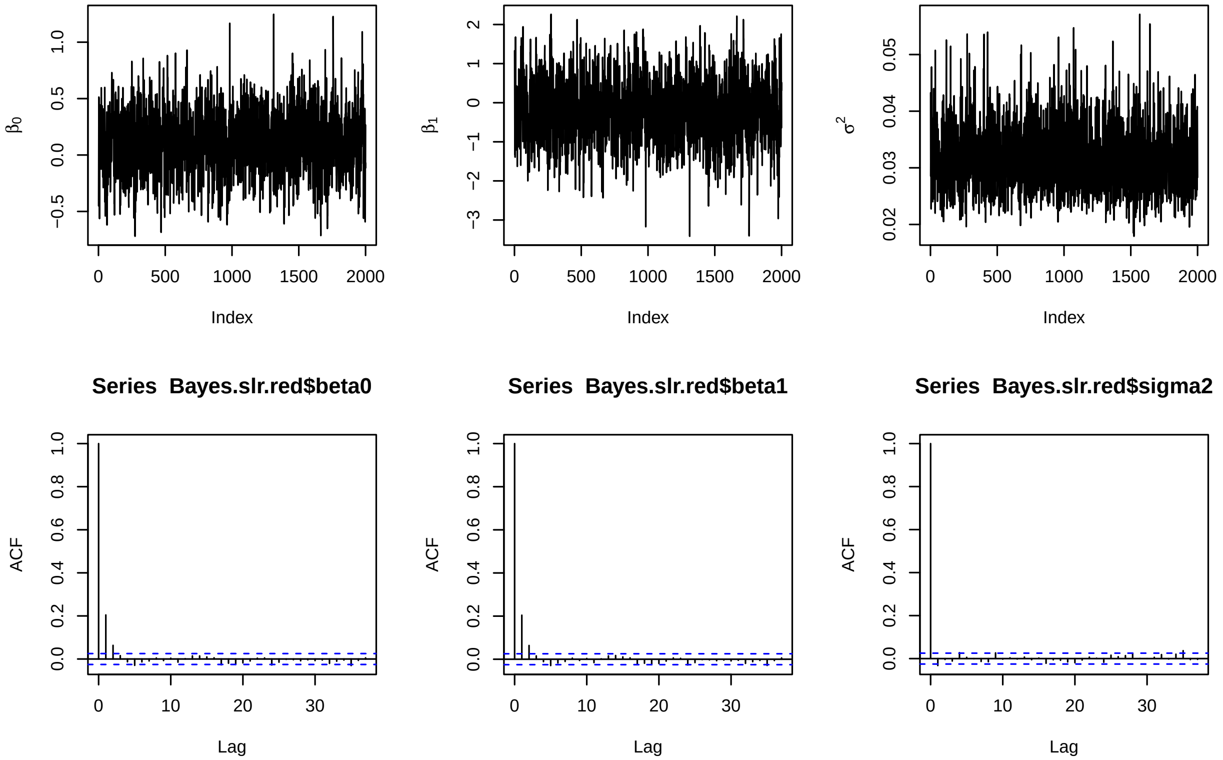
Assumptions

Residual Plot for "Red" States



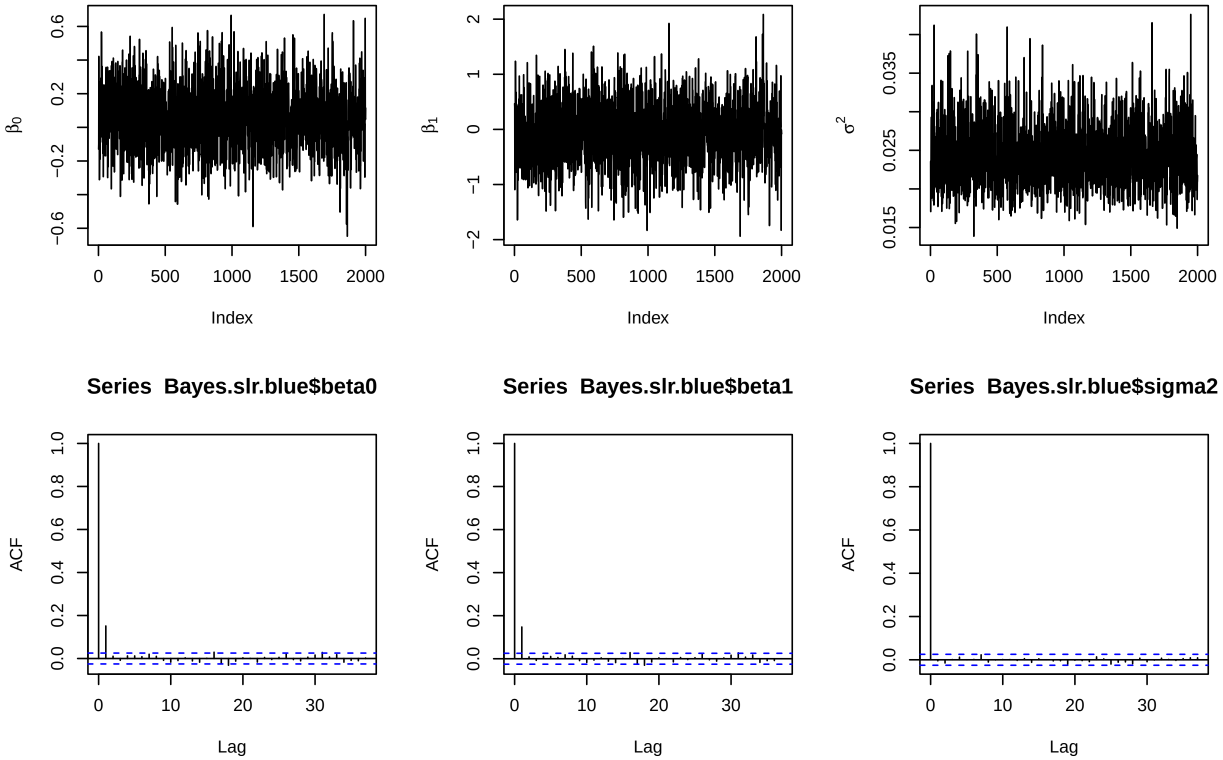
Residual Plot for "Blue" States




Series Bayes.slr.red$beta0

Series Bayes.slr.red$beta1

Series Bayes.slr.red$sigma2


Series Bayes.slr.blue$beta0

Series Bayes.slr.blue$beta1

Series Bayes.slr.blue$sigma2

Code

```r
# Residual plot
hist(resid(lm(droprate ~ salprop, data = red_states)),
     main = "Residual Plot for \"Red\" States",
     xlab = "")
hist(resid(lm(droprate ~ salprop, data = blue_states)),
     main = "Residual Plot for \"Blue\" States",
     xlab = "")

#Check convergence and mixing for each parameter for each population
# Red states
par(mfrow = c(2,3))
plot(Bayes.slr.red$beta0[1:2000], type = "l",ylab = expression(beta[0]))
plot(Bayes.slr.red$beta1[1:2000], type = "l", ylab = expression(beta[1]))
plot(Bayes.slr.red$sigma2[1:2000], type = "l", ylab = expression(sigma^2))
acf(Bayes.slr.red$beta0)
acf(Bayes.slr.red$beta1)
acf(Bayes.slr.red$sigma2)

# Blue states
plot(Bayes.slr.blue$beta0[1:2000], type = "l",ylab = expression(beta[0]))
plot(Bayes.slr.blue$beta1[1:2000], type = "l", ylab = expression(beta[1]))
plot(Bayes.slr.blue$sigma2[1:2000], type = "l", ylab = expression(sigma^2))
acf(Bayes.slr.blue$beta0)
acf(Bayes.slr.blue$beta1)
acf(Bayes.slr.blue$sigma2)
par(mfrow = c(1,1))
```

```r
library(readxl)
library(knitr)
library(invgamma)
# Read in the data
exps <- read.csv("Total_Expenditures_07_09.csv")
teach <- read.csv("Instruction_Salaries_07_09.csv")
all_1 <- read_excel("All_Students.xlsx", skip = 2)
all_2 <- read_excel("All_Students_2.xlsx", skip = 2)
drop_1 <- read_excel("Total_Dropouts.xlsx", skip = 2)
drop_2 <- read_excel("Total_Dropouts_2.xlsx", skip = 2)

# Combine into data frames by like variables
enroll <- cbind(all_2,all_1[,-1])
drops <- cbind(drop_1,drop_2[,-1])
```

```r
# Clean up environment
rm(all_1,all_2,drop_1,drop_2)

# Create function that converts to long format
cleaner <- function(x){
  clean <- data.frame(year=numeric(),var=numeric(),source=character())
  for(i in 1:3){
    for(j in 1:(ncol(x)-1)){
      clean[nrow(clean)+1,] <- c(i+2006,x[i,j+1],colnames(x)[j+1])
    }
  }
  clean
}

# Convert to long format
enroll <- cleaner(enroll)
drops <- cleaner(drops)
exps <- cleaner(exps)
teach <- cleaner(teach)

# Create function to clean state names
namecleaner <- function(x,trim1,trim2){
  for(i in 1:nrow(x)){
    x[i,3] <- substr(x[i,3],trim1,nchar(x[i,3]))
    x[i,3] <- substr(x[i,3],1,nchar(x[i,3])-trim2)
  }
  x
}

# Clean state and column names from enrollment
enroll <- namecleaner(enroll,30,36)
enroll$source <- trimws(enroll$source)
colnames(enroll) <- c("Year","Enrollment","Region")
enroll$Enrollment <- as.numeric(enroll$Enrollment)

# Combine enrollment across years
enroll <- aggregate(Enrollment ~ Region + Year, data = enroll, FUN = sum)

# Add Wisconsin and Wyoming enrollments
enroll <- rbind(enroll,data.frame(Year = c(2007,2008,2009,2007,2008,2009),
           Region = c("Wisconsin","Wisconsin","Wisconsin","Wyoming","Wyoming","Wyoming"),
           Enrollment = c(290146,285591,280188,26839,26397,26146)))
```

```
# Clean state and column names from dropouts
drops <- namecleaner(drops,29,42)
colnames(drops) <- c("Year","Dropouts","Region")

# Add Wisconsin and Wyoming dropouts
drops <- rbind(drops,data.frame(Year = c(2007,2008,2009,2007,2008,2009),
          Region = c("Wisconsin","Wisconsin","Wisconsin","Wyoming","Wyoming","Wyoming"),
          Dropouts = c(6673,6370,6005,1365,1000,1416)))

# Clean state and column names from expenses
exps <- namecleaner(exps,21,24)
colnames(exps) <- c("Year","Expenditures","Region")

# Clean state and column names from teacher salaries
teach <- namecleaner(teach,21,28)
colnames(teach) <- c("Year","Salaries","Region")

# Reorder exp and teach to match enroll and drops
exps <- exps[c(1:49,52:100,103:151,50,101,152,51,102,153),]
teach <- teach[c(1:49,52:100,103:151,50,101,152,51,102,153),]

# Combine all relevant variables
allvar <- data.frame(Year = enroll$Year,
           Region = enroll$Region,
           Enrollment = enroll$Enrollment,
           Dropouts = as.numeric(drops$Dropouts),
           Expenditures = as.numeric(exps$Expenditures),
           Salaries = as.numeric(teach$Salaries))

# Calculate columns of relevant percentages
allvar$droprate <- allvar$Dropouts/allvar$Enrollment
allvar$salprop <- allvar$Salaries/allvar$Expenditures

# Create final data set
final_data <- data.frame(year = allvar$Year,
                         region = allvar$Region,
                         droprate = allvar$droprate,
                         salprop = allvar$salprop)

# Create red and blue vectors to be used in the analysis
reds <- unique(final_data$region)[c(1:4,10,12,16:18,24:27,34,36,40:44,49,51)]
blues <- unique(final_data$region)[c(5:9,11,13:15,19:23,28:33,35,37:39,45:48,50)]
```

```
# Add election column
final_data$election <- character(nrow(final_data))
final_data[final_data$region %in% reds,5] <- "red"
final_data[final_data$region %in% blues,5] <- "blue"

# Remove NAs
final_data <- na.omit(final_data)

# Split data by 2008 election results
red_states <- final_data[final_data$election == "red",]
blue_states <- final_data[final_data$election == "blue",]
```

```
# View red and blue scatterplots
plot(red_states$salprop,red_states$droprate, col = "red", pch = 19,
     xlab = "Proportion of Education Expenditures Spent on Teacher Salaries",
     ylab = "Dropout Rate",
     main = "Complete U.S. Data 2007-2009")
points(blue_states$salprop,blue_states$droprate, col = "blue", pch = 19)
legend("topright", c("\"Red\" States", "\"Blue\" States"),
col=c("red", "blue"), pch = 19)

# Table of summary statistics
titles <- c("Red States","Blue States","Overall")
correlations <- c(cor(red_states$salprop,red_states$droprate),
                  cor(blue_states$salprop,blue_states$droprate),
                  cor(final_data$salprop,final_data$droprate))
R2 <- correlations^2
ymeans <-c(mean(red_states$droprate),
           mean(blue_states$droprate),
           mean(final_data$droprate))
xmeans <-c(mean(red_states$salprop),
           mean(blue_states$salprop),
           mean(final_data$salprop))
tab <- data.frame(titles,correlations,R2,ymeans,xmeans)
kable(tab, col.names = c("Region",
                         "Correlation",
                         "R-sq",
                         "Mean Dropout Rate",
                         "Mean Proportion of Budget for Teacher Salaries"),
      digits = 3, caption = "Summary Statistics")
```

```
# Create funciton for MCMC algorithm
slr.mcmc <- function(y, x, m0, v0, m1, v1, a, b, J=100, nburn=0, nthin=1){
```

```r
# empty vectors to save MCMC iterates
beta0 <- beta1 <- sigma2 <- numeric()


# starting values
beta0[1] <- 0
beta1[1] <- 0
sigma2[1] <- 1


n <- length(x)
sumx2 <- sum(x^2)

# empty object for fitted line for
# simple linear regression line
slrfit <- matrix(NA, nrow=J, ncol=n)



for(j in 2:J){

  # start updating beta0
  vstar <- 1/(n/sigma2[j-1] + 1/v0)

  sumyxb <- sum(y - x*beta1[j-1])
  mstar <- ((1/sigma2[j-1])*sumyxb + m0/v0)*vstar

  beta0[j] <- rnorm(1, mstar, sqrt(vstar))

  # update beta1
  vstar <- 1/(sumx2/sigma2[j-1] + 1/v1)
  mstar <- ((1/sigma2[j-1])*sum(x*(y - beta0[j])) + m1/v1)*vstar
  beta1[j] <- rnorm(1, mstar, sqrt(vstar))

  # update sigma2
  astar <- 0.5*n + a
  bstar <- 0.5*sum((y - (beta0[j] + beta1[j]*x))^2) + b
  sigma2[j] <- rinvgamma(1, shape=astar, rate=bstar)

  # update the simple linear regression line fit
  slrfit[j,] <- beta0[j] + beta1[j]*x
  }

  keep <- seq(nburn+1, J, by = nthin)

  list(beta0 = beta0[keep],
       beta1 = beta1[keep],
       sigma2 = sigma2[keep],
       slrfit = slrfit[keep,])

}
```

```
# Run the algorithm for red states and blue states
Bayes.slr.red <- slr.mcmc(y=red_states$droprate, x=red_states$salprop,
                      m0=0, v0=100,
                      m1=0, v1=100,
                      a=1, b=1,
                      J = 1500000, nburn=100, nthin=250)
Bayes.slr.blue <- slr.mcmc(y=blue_states$droprate, x=blue_states$salprop,
                      m0=0, v0=100,
                      m1=0, v1=100,
                      a=1, b=1,
                      J = 1500000, nburn=100, nthin=250)
```

```
# Plot prior vs posterior for all parameters
# First for reds
# Prior and posterior for red beta0
x <- seq(-20,20,length.out = 1001)
prior <- dnorm(x,0,10)
posterior <- density(Bayes.slr.red$beta0)
plot(posterior,type = "l", xlab = expression(beta[0]), ylab = "Density",
     main = expression(paste("Prior and Posterior Distributions for ", beta[0]["Red"])), lwd = 2)
lines(x,prior, col = "grey", lwd = 2)
legend("topright", c("Prior", "Posterior"),
col=c("grey", "black"), lwd=3)

# Prior and posterior for red beta1
x <- seq(-4,5,length.out = 1001)
prior <- dnorm(x,0,10)
posterior <- density(Bayes.slr.red$beta1)
plot(posterior,type = "l", xlab = expression(beta[1]), ylab = "Density",
     main = expression(paste("Prior and Posterior Distributions for ", beta[1]["Red"])),
     lwd = 2)
lines(x,prior, col = "grey", lwd = 2)
legend("topright", c("Prior", "Posterior"),
col=c("grey", "black"), lwd=3)

# Prior and posterior for red sigma2
x <- seq(0,2,length.out = 1001)
prior <- dinvgamma(x,1,1)
posterior <- density(Bayes.slr.red$sigma2)
plot(posterior,type = "l", xlab = expression(sigma^2), ylab = "Density",
     main = expression(paste("Prior and Posterior Distributions for ", sigma["Red"]^2)),
     lwd = 2)
lines(x,prior, col = "grey", lwd = 2)
legend("topright", c("Prior", "Posterior"),
col=c("grey", "black"), lwd=3)
# Then for blues
# Prior and posterior for blue beta0
x <- seq(-20,20,length.out = 1001)
prior <- dnorm(x,0,10)
posterior <- density(Bayes.slr.blue$beta0)
plot(posterior,type = "l", xlab = expression(beta[0]), ylab = "Density",
     main = expression(paste("Prior and Posterior Distributions for ", beta[0]["Blue"])),
     lwd = 2)
lines(x,prior, col = "grey", lwd = 2)
legend("topright", c("Prior", "Posterior"),
col=c("grey", "black"), lwd=3)

# Prior and posterior for blue beta1
x <- seq(-4,5,length.out = 1001)
prior <- dnorm(x,0,10)
posterior <- density(Bayes.slr.blue$beta1)
plot(posterior,type = "l", xlab = expression(beta[1]), ylab = "Density",
     main = expression(paste("Prior and Posterior Distributions for ", beta[1]["Blue"])),
     lwd = 2)
lines(x,prior, col = "grey", lwd = 2)
legend("topright", c("Prior", "Posterior"),
col=c("grey", "black"), lwd=3)

# Prior and posterior for blue sigma2
x <- seq(0,2,length.out = 1001)
prior <- dinvgamma(x,1,1)
posterior <- density(Bayes.slr.blue$sigma2)
plot(posterior,type = "l", xlab = expression(sigma^2), ylab = "Density",
     main = expression(paste("Prior and Posterior Distributions for ", sigma["Blue"]^2)),
     lwd = 2)
lines(x,prior, col = "grey", lwd = 2)
```

```r
legend("topright", c("Prior", "Posterior"),
col=c("grey", "black"), lwd=3)
```

```r
# Posterior graphic of both beta1s
x <- seq(-4,5,length.out = 1001)
posterior <- density(Bayes.slr.blue$beta1)
plot(posterior,type = "l", xlab = expression(beta[1]), ylab = "Density",
     main = expression(paste("Posterior Distributions for ", beta[1])),
     lwd = 2, col = "blue")

posterior <- density(Bayes.slr.red$beta1)
lines(posterior,type = "l", lwd = 2, col = "red")
legend("topright", c("\"Red\" States", "\"Blue\" States"),
col=c("red", "blue"), lwd=3)

# Credible intervals for the individual slopes
conred <- quantile(Bayes.slr.red$beta1, c(0.025, 0.975))
conblue <- quantile(Bayes.slr.blue$beta1, c(0.025, 0.975))

# Table of confidence intervals
titles <- c("Red States","Blue States")
lower <- c(conred[1],conblue[1])
upper <- c(conred[2],conblue[2])
tab <- data.frame(titles,lower,upper)
kable(tab, col.names = c("Population","Lower Bound","Upper Bound"),
      digits = 3, caption = "Credible Intervals for Population Slopes")
```

```r
# order
sal_ord_red <- order(red_states$salprop)
sal_ord_blue <- order(blue_states$salprop)

# Scatterplot with lines and credible intervals
plot(red_states$salprop,red_states$droprate, col = "red", pch = 19,
     xlim = c(0.25,.4), ylim = c(-.1,.21),
     xlab = "Proportion of Education Expenditures Spent on Teacher Salaries",
     ylab = "Dropout Rate")
points(blue_states$salprop,blue_states$droprate, col = "blue", pch = 19)
lines(red_states$salprop[sal_ord_red],
      apply(Bayes.slr.red$slrfit,2,mean)[sal_ord_red],
      col = "red", type = "l")
lines(blue_states$salprop[sal_ord_blue],
      apply(Bayes.slr.blue$slrfit,2,mean)[sal_ord_blue],
      col = "blue")
```

```r
# red 95% credible bands
lines(red_states$salprop[sal_ord_red],
      apply(Bayes.slr.red$slrfit,2,quantile, p = .025)[sal_ord_red],
      lty = 2, col = "red")
lines(red_states$salprop[sal_ord_red],
      apply(Bayes.slr.red$slrfit,2,quantile, p = .975)[sal_ord_red],
      lty = 2, col = "red")

# blue 95% credible bands
lines(blue_states$salprop[sal_ord_blue],
      apply(Bayes.slr.blue$slrfit,2,quantile, p = .025)[sal_ord_blue],
      lty = 2, col = "blue")
lines(blue_states$salprop[sal_ord_blue],
      apply(Bayes.slr.blue$slrfit,2,quantile, p = .975)[sal_ord_blue],
      lty = 2, col = "blue")

legend("topright",
       c("\"Red\" States", "\"Blue\" States", "Regression Lines","Confidence Bands"),
       col=c("red", "blue","purple","purple"),
       lwd=3, pch = c(19,19,NA,NA), lty = c(NA,NA,1,2))
```

```r
# Plot of beta1red-beta1blue
plot(density(Bayes.slr.red$beta1-Bayes.slr.blue$beta1),
     type = "l",
     xlab = expression(beta[1]["Red"]-beta[1]["Blue"]),
     ylab = "Density",
     main = expression(paste("Posterior Distribution of ",
                             beta[1]["Red"]-beta[1]["Blue"])))

#Credible interval for the difference between the two slopes
diffcred<- quantile(Bayes.slr.red$beta1-Bayes.slr.blue$beta1, c(.025,.975))
```