Red State Tutorial

Introduction

> I never said all Democrats are saloon-keepers. What I said is that all saloonkeepers are Democrats. —Horace Greeley, 1860

> Pat doesn't have a mink coat. But she does have a respectable Republican cloth coat. —Richard Nixon, 1952

> Like upscale areas everywhere, from Silicon Valley to Chicago's North Shore to suburban Connecticut, Montgomery County supported the Democratic ticket in last year's presidential election, by a margin of 63 percent to 34 percent. —David Brooks, 2001

> There is, for example, this large class of affluent professionals who are solidly Democratic. DataQuick Information Systems recently put out a list of 100 ZIP code areas where the median home price was above $500,000. By my count, at least 90 of these places — from the Upper West Side to Santa Monica — elect liberal Democrats. —David Brooks, 2004

> A lot of Bush's red zones can be traced to wealthy enclaves or sun-belt suburbs where tax cuts are king. —Matt Bai, 2001

> But in the Ipsos-Reid surveys, 38% of voters in "strong Bush" counties said that they had household incomes below $30,000, while 7% said that their families earned at least $100,000. In "strong Gore" counties, by contrast, only 29% of voters pegged their household income below $30,000, while 14% said that it was above $100,000. —James Barnes, 2002

For decades, the Democrats have been viewed as the party of the poor, with the Republicans representing the rich. Recent presidential elections, however, have shown a reverse pattern, with Democrats performing well in the richer blue states in the northeast and coasts, and Republicans dominating in the red states in the middle of the country and the south.

The purpose of this homework assignment, then, is to use Stan and hierarchical models to understand this seeming paradox. For our purposes, we will use data from the 2004 election. If you enjoy doing this analysis and would like to see this analysis done on other elections, you can read Andrew Gelman, Boris Shor, Joseph Bafumi, and David Park's *Rich State, Poor State, Red State, Blue State: What's the Matter with Connecticut?* (from which this tutorial is derived).

Preliminaries

If you haven't already, please install RStan. Instructions can be found here: https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started.

Then, please download the folder with the data to your desktop. In your R session, you'll need to set the working directory to this folder.

For Windows, the command to do so will be:

```
setwd("C:/Users/Public/Desktop/Red State Tutorial")
```

For Macs, the command will be:

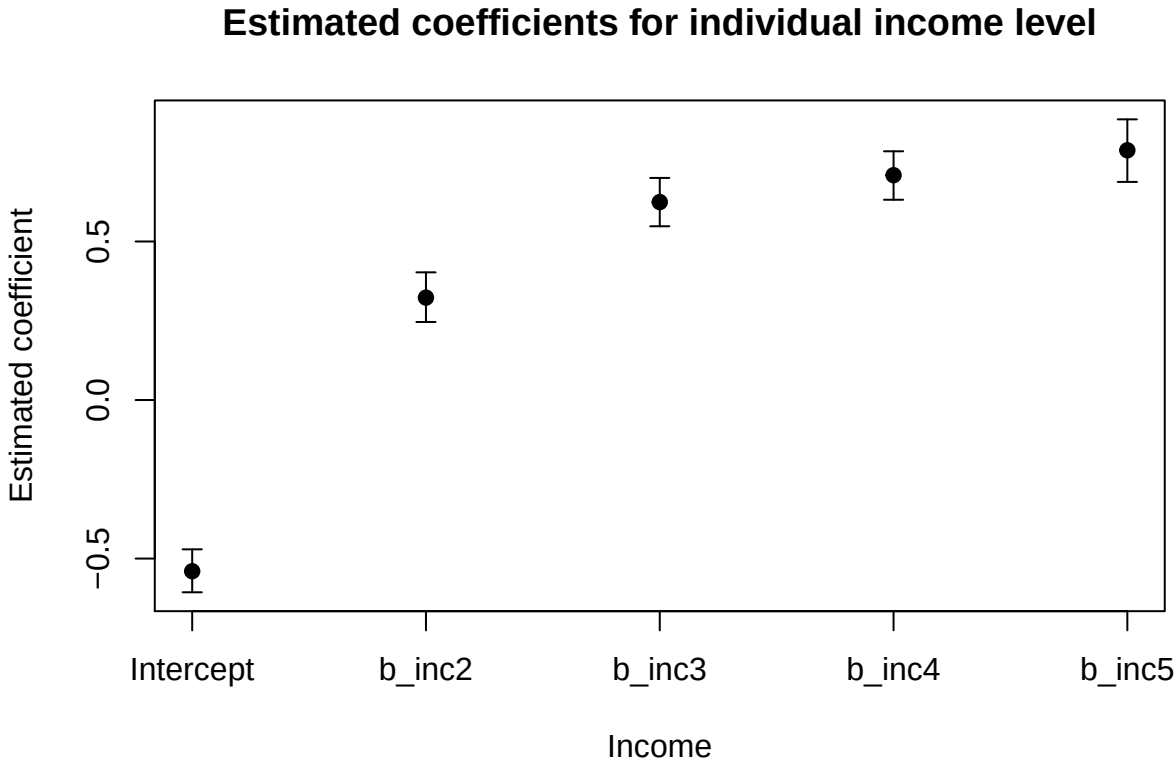
```
setwd("~/Desktop/Red State Tutorial")
```

Visualizing the "Paradox"

First, let us try to visualize this seeming paradox. Using the National Annenberg 2004 Election Survey data, we will look at the likelihood of voting Republican based on an individual's income. Note that in this tutorial, we'll categorize an individual's income in the following manner: $0 - $20,000, $20,000 - $40,000, $40,000 - $75,000, $75,000 - $150,000, and $150,000+.

First, the Stan analysis:

```
# Get the data and the R package we'll be using for
# the analysis
library(rstan)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
load("Data/stan_inc_regression_data.Rdata")

# Use Stan to run the regression model
stan_2004_ind_inc_fit_obj <- stan(file = "Stan/regression_by_one_var.stan",
    data = stan_2004_ind_inc_data, model_name = "Republican vote by individual income")

# Pull out the regression coefficient and the
# information needed for a confidence interval
interested_regression_rows <- c("Intercept", "b_inc2",
    "b_inc3", "b_inc4", "b_inc5")
interested_regression_cols <- c("mean", "2.5%", "97.5%")
ind_inc_regression_data <- summary(stan_2004_ind_inc_fit_obj)$summary[interested_regression_rows,
    interested_regression_cols]

# Plot the regression coefficients with bars
plot(ind_inc_regression_data[, "mean"], ylim = range(c(ind_inc_regression_data[,
    "2.5%"], ind_inc_regression_data[, "97.5%"])),
    pch = 19, xaxt = "n", xlab = "Income", ylab = "Estimated coefficient",
    main = "Estimated coefficients for individual income level")
axis(1, 1:5, interested_regression_rows)
arrows(1:5, ind_inc_regression_data[, "2.5%"], 1:5,
    ind_inc_regression_data[, "97.5%"], length = 0.05,
    angle = 90, code = 3)
```

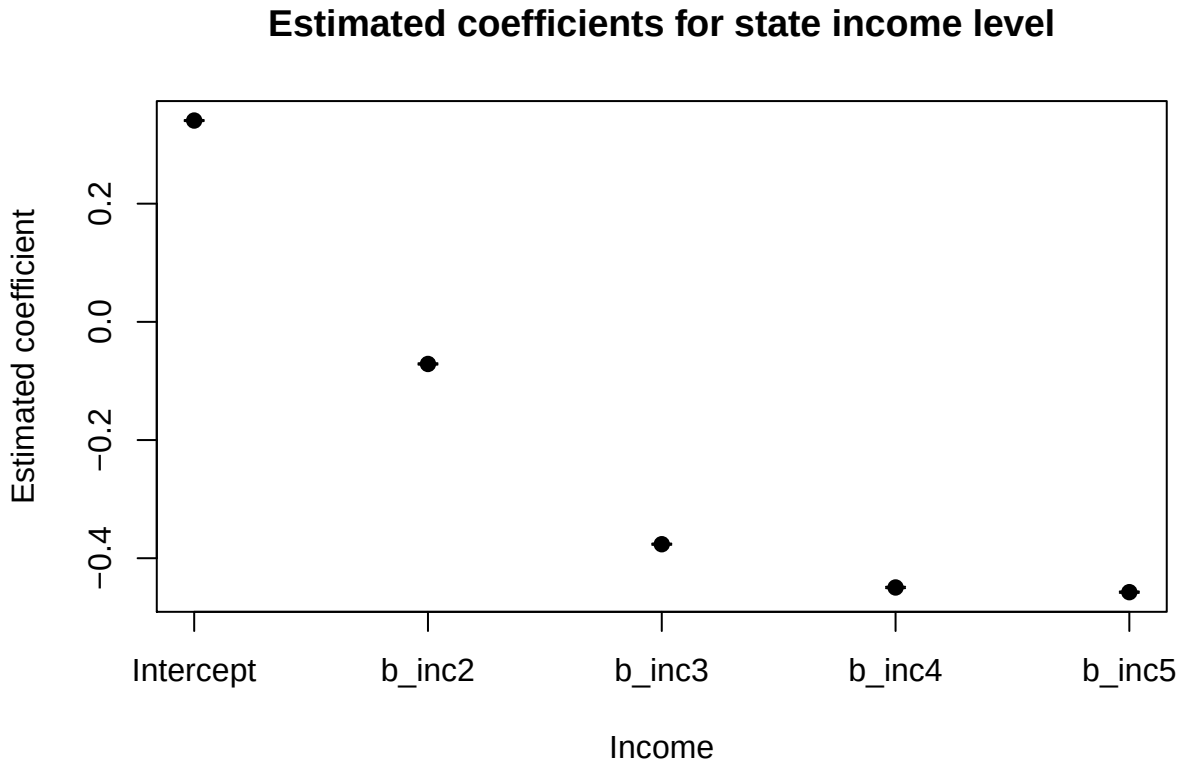


Note that we're plotting the regression coefficients in the plots. We can do so because in our model, it's a simple transform to get the probability.

Next, let us look at the likelihood of voting Republican based on the state's average income using the Census data. We'll divide the state average income into five categories so that there are the same number of states in each category. The categories are: $0 - $39,000, $39,000 - $43,003, $43,003 - $44,476, $44,476 - $50,614 and $50,614+.

```
# Get the data and the R package we'll be using for
# the analysis
library(rstan)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
load("Data/stan_state_regression_data.Rdata")

# Use Stan to run the regression model
stan_2004_state_inc_fit_obj <- stan(file = "Stan/regression_by_one_var.stan",
    data = stan_2004_state_inc_data, model_name = "Republican vote by state average income")

# Pull out the regression coefficient and the
# information needed for a confidence interval
interested_regression_rows <- c("Intercept", "b_inc2",
    "b_inc3", "b_inc4", "b_inc5")
interested_regression_cols <- c("mean", "2.5%", "97.5%")
state_inc_regression_data <- summary(stan_2004_state_inc_fit_obj)$summary[interested_regression_rows,
    interested_regression_cols]

# Plot the regression coefficients with bars
plot(state_inc_regression_data[, "mean"], ylim = range(c(state_inc_regression_data[,
    "2.5%"], state_inc_regression_data[, "97.5%"])),
```

```
    pch = 19, xaxt = "n", xlab = "Income", ylab = "Estimated coefficient",
    main = "Estimated coefficients for state income level")
axis(1, 1:5, interested_regression_rows)
arrows(1:5, state_inc_regression_data[, "2.5%"], 1:5,
    state_inc_regression_data[, "97.5%"], length = 0.05,
    angle = 90, code = 3)
```



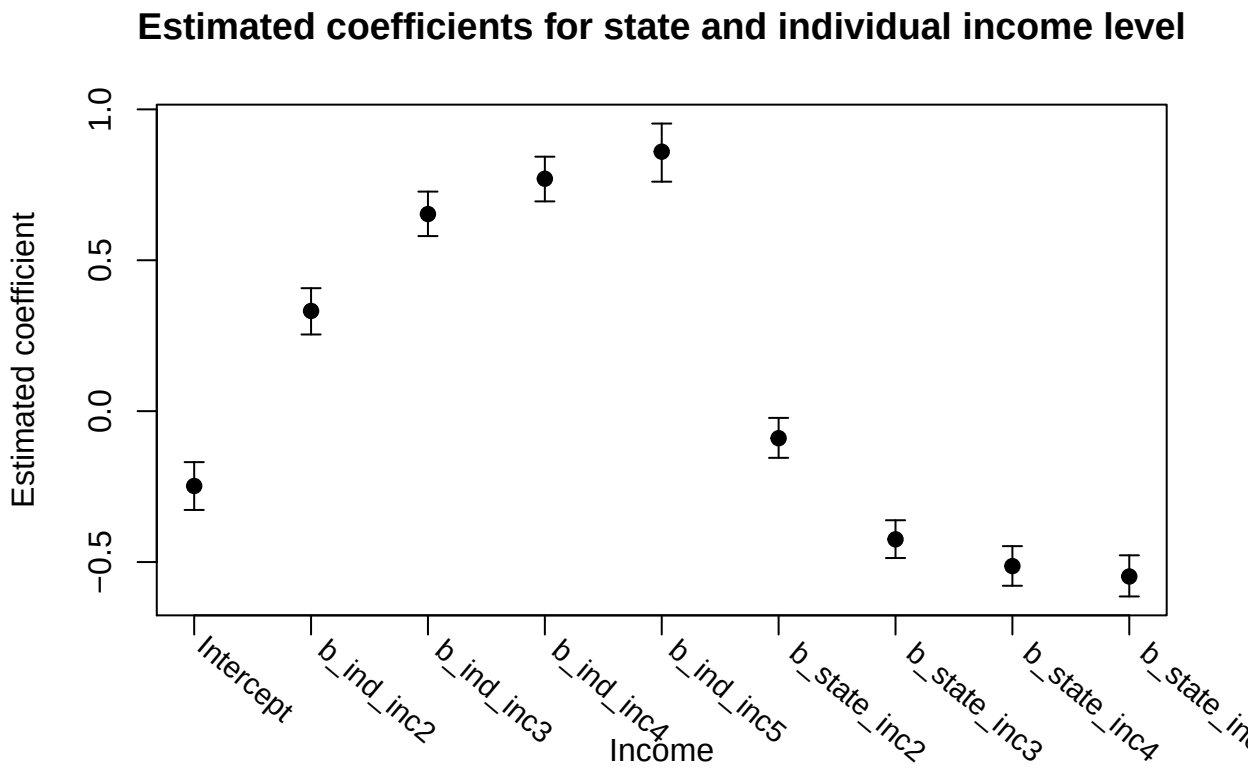
We can see that as an individual's income increases, the likelihood of voting Republican does so as well, but as a state's average income increases, the likelihood of voting Republican goes down.

Understanding this "paradox"

For our first attempt to understand this paradox, what if we treated the state's average income and individual's income as predictor variables?

```
# Get the data and the R package we'll be using for
# the analysis
library(rstan)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
load("Data/stan_state_inc_and_ind_inc_regression_data.Rdata")

# Use Stan to run the regression model
stan_2004_state_and_ind_inc_fit_obj <- stan(file = "Stan/regression_by_state_inc_and_ind_inc.stan",
    data = stan_2004_state_and_ind_inc_data, model_name = "Republican vote by state and ind income")

# Pull out the regression coefficient and the
```

```r
# information needed for a confidence interval
interested_regression_rows <- c("Intercept", "b_ind_inc2",
    "b_ind_inc3", "b_ind_inc4", "b_ind_inc5", "b_state_inc2",
    "b_state_inc3", "b_state_inc4", "b_state_inc5")
interested_regression_cols <- c("mean", "2.5%", "97.5%")
state_and_ind_inc_regression_data <- summary(stan_2004_state_and_ind_inc_fit_obj)$summary[interested_reg
    interested_regression_cols]

# Plot the regression coefficients with bars
plot(state_and_ind_inc_regression_data[, "mean"], ylim = range(c(state_and_ind_inc_regression_data[,
    "2.5%"], state_and_ind_inc_regression_data[, "97.5%"])),
    pch = 19, xaxt = "n", xlab = "Income", ylab = "Estimated coefficient",
    main = "Estimated coefficients for state and individual income level")
text(axis(1, 1:9, labels = F), par("usr")[3], labels = interested_regression_rows,
    srt = 320, xpd = T, adj = c(-0.2, 1.2), cex = 0.9)
arrows(1:9, state_and_ind_inc_regression_data[, "2.5%"],
    1:9, state_and_ind_inc_regression_data[, "97.5%"],
    length = 0.05, angle = 90, code = 3)
```



As we can see, that was not helpful. We still see the likelihood of voting Republican increase as one's income increases and decrease as one's state average income increases. So, let's take a step back and think about this problem some more. What do we know?

1. We know that some states are more likely to vote Republican.
2. We know that individuals with higher income are more likely to vote Republican.
3. We know that states with higher average income are less likely to vote Republican.

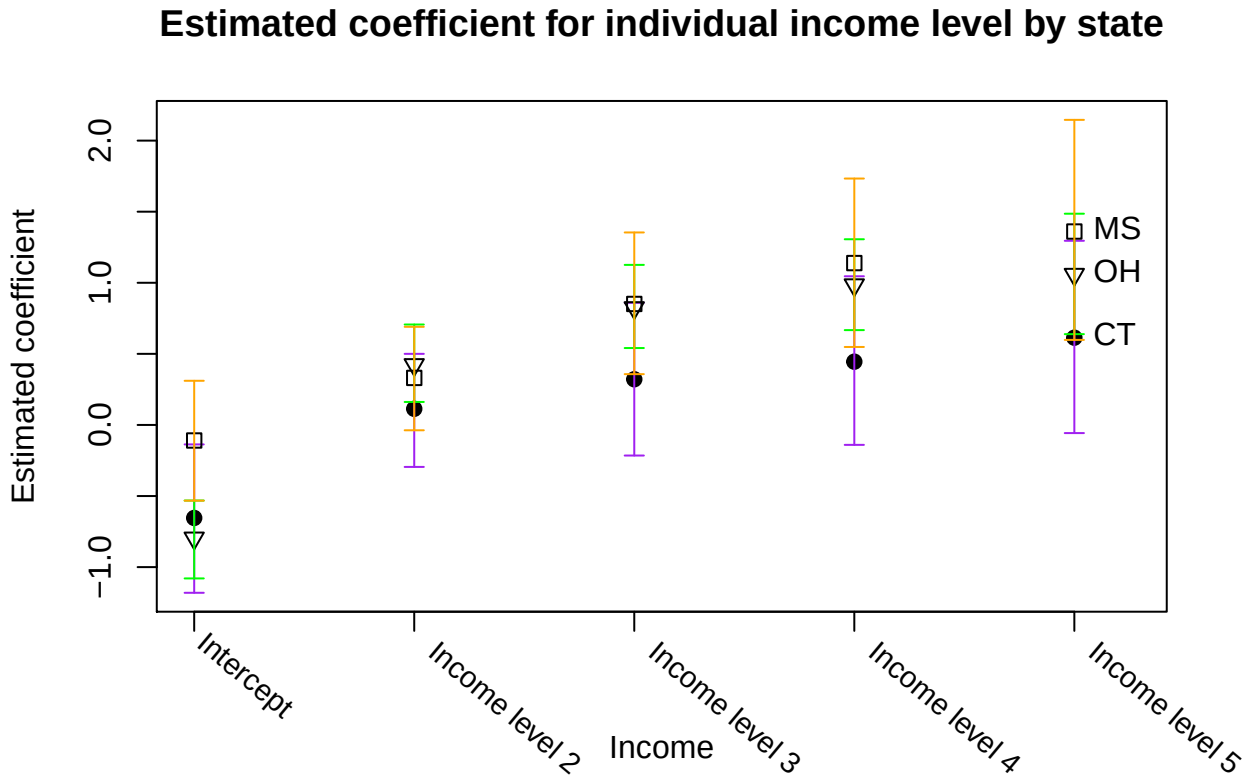
We can very easily express the first point in a hierarchical or mixed effects model. After all, to say that some states are more likely to vote Republican, we're also essentially saying that we want the intercept in our regression to vary by state.

So, let's build on this idea. If we can vary the intercept by state, why not also vary the slope associated with an individual's income by state? (It doesn't really make sense to vary the slope associated with a state's average income by state.) In other words, are all individuals with higher income more likely to vote Republican or does this vary by state?

Let us conduct this analysis. For our plot, let us focus on Mississippi, Ohio, and Conneticut because they represent a poor, middle-income, and rich state respectively.

```
# Get the data and the R package we'll be using for
# the analysis
library(rstan)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
load("Data/stan_state_and_inc_regression_data.Rdata")

# Use Stan to run the regression model
stan_2004_state_and_inc_fit_obj <- stan(file = "Stan/mixed_effects_regression_for_state_and_inc.stan",
    data = stan_2004_state_and_income_data, model_name = "Republican vote by income and state")

# Pull out the regression coefficient and the
# information needed for a confidence interval
ct_regression_rows <- sapply(1:5, function(i) {
    paste("m_Intercept_inc2_inc3_inc4_inc5_by_state[6,",
        i, "]", sep = "")
}, simplify = T)

oh_regression_rows <- sapply(1:5, function(i) {
    paste("m_Intercept_inc2_inc3_inc4_inc5_by_state[34,",
        i, "]", sep = "")
}, simplify = T)

ms_regression_rows <- sapply(1:5, function(i) {
    paste("m_Intercept_inc2_inc3_inc4_inc5_by_state[24,",
        i, "]", sep = "")
}, simplify = T)

interested_regression_cols <- c("mean", "2.5%", "97.5%")
state_and_inc_regression_data <- summary(stan_2004_state_and_inc_fit_obj)$summary[c(ct_regression_rows,
    oh_regression_rows, ms_regression_rows), interested_regression_cols]

# Plot the regression coefficients with bars Plot
# Conneticut first
plot(state_and_inc_regression_data[ct_regression_rows,
    "mean"], xlim = c(1, 5.25), ylim = c(min(state_and_inc_regression_data[,
    "2.5%"]), max(state_and_inc_regression_data[, "97.5%"])),
    pch = 19, xaxt = "n", xlab = "Income", ylab = "Estimated coefficient",
    main = "Estimated coefficient for individual income level by state")
text(axis(1, labels = F), par("usr")[3], labels = c("Intercept",
    "Income level 2", "Income level 3", "Income level 4",
    "Income level 5"), srt = 320, xpd = T, adj = c(-0.2,
    1.2), cex = 0.9)
```

```
arrows(1:5, state_and_inc_regression_data[ct_regression_rows,
    "2.5%"], 1:5, state_and_inc_regression_data[ct_regression_rows,
    "97.5%"], length = 0.05, angle = 90, code = 3,
    col = "purple")
text(5, state_and_inc_regression_data[ct_regression_rows[5],
    "mean"], labels = c("CT"), pos = 4)

# Plot Ohio
points(state_and_inc_regression_data[oh_regression_rows,
    "mean"], pch = 6)
arrows(1:5, state_and_inc_regression_data[oh_regression_rows,
    "2.5%"], 1:5, state_and_inc_regression_data[oh_regression_rows,
    "97.5%"], length = 0.05, angle = 90, code = 3,
    col = "green")
text(5, state_and_inc_regression_data[oh_regression_rows[5],
    "mean"], labels = c("OH"), pos = 4)

# Plot Mississippi
points(state_and_inc_regression_data[ms_regression_rows,
    "mean"], pch = 0)
arrows(1:5, state_and_inc_regression_data[ms_regression_rows,
    "2.5%"], 1:5, state_and_inc_regression_data[ms_regression_rows,
    "97.5%"], length = 0.05, angle = 90, code = 3,
    col = "orange")
text(5, state_and_inc_regression_data[ms_regression_rows[5],
    "mean"], labels = c("MS"), pos = 4)
```



From this analysis, we see that the effect of income on voting Republican seems to vary by state. Specifically, it seems that while individuals with higher income are more likely to vote Republican, this trend is smaller in

richer states and so these states are more likely to vote Democrat.