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 "paradox". For our purposes, we will use data from the 2004 election.

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:

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

For Macs, the command will be:

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

In addition, you should also run the following commands in your R Session to use RStan and take full advantage of your computer's processing power. It'll also add some helper functions that we'll use later.

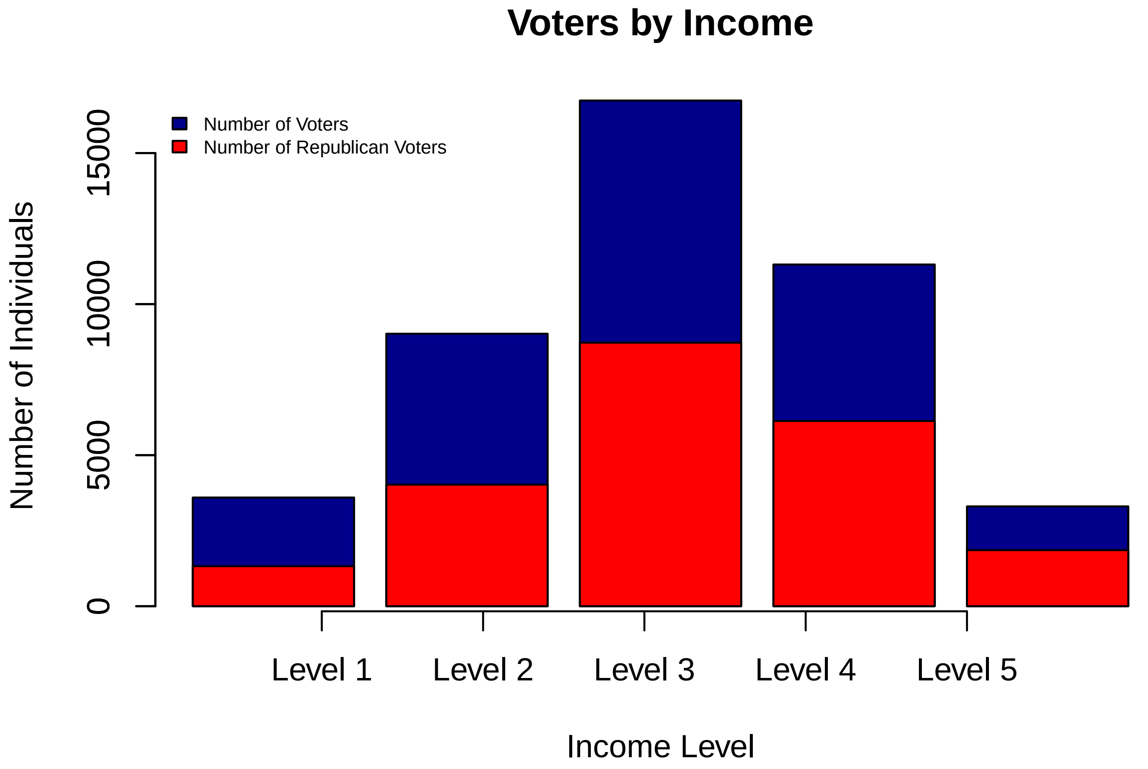
```
library(rstan)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
source("red_state_tutorial_helper_functions.R")
```

Finally, if you'd like to run the entire code for each scenario, check the appendix.

Visualizing the "Paradox"

Let's first try to analyze the data and see if this "paradox" exists. There are two parts. The first is that the Democrats are the "party of the poor". What do we mean when we say that? Well, we expect that people with lower income are more likely to vote Democrat and people with higher income are more likely to vote Republican.

Is this the case? Let's use the National Annenberg 2004 Election Survey data to explore whether income is correlated to the likelihood of voting Republican. In this data set, income is not a continuous variable, but divided into five categories: \$0 - \$20,000, \$20,000 - \$40,000, \$40,000 - \$75,000, \$75,000 - \$150,000, and \$150,000+. Looking at the data, we see:



While it looks like there are proportionally fewer Republican voters for those in the lowest income level and proportionally more Republican voters for those in the highest income level, the trend isn't fully clear.

So, let's analyze it. We're trying to predict the likelihood of voting Republican for a given income level based on the number of Republican voters for each income level. Because we'll ignore third party voters and think of the outcome as voting Republican or voting Democrat, we'll be using a binomial logistic regression model to do so. "Success" in our model will be voting Republican though we could have considered "success" as voting Democrat.

On the other side, income level is a categorical variable and not a continous variable. As a result, the regression coefficients for income level we're looking for will tell us the difference between one level and any other level. Because we're trying to see if the likelihood of voting Republican or the number of Republican voter increases when income increases, let's use the first or lowest income level as the level to compare all other levels to.

So, now that we know our model, we can write our Stan code and process our data for the Stan code. Fortunately, the Stan code and data for that model is provided for you so all you have to do is run it in your R session:

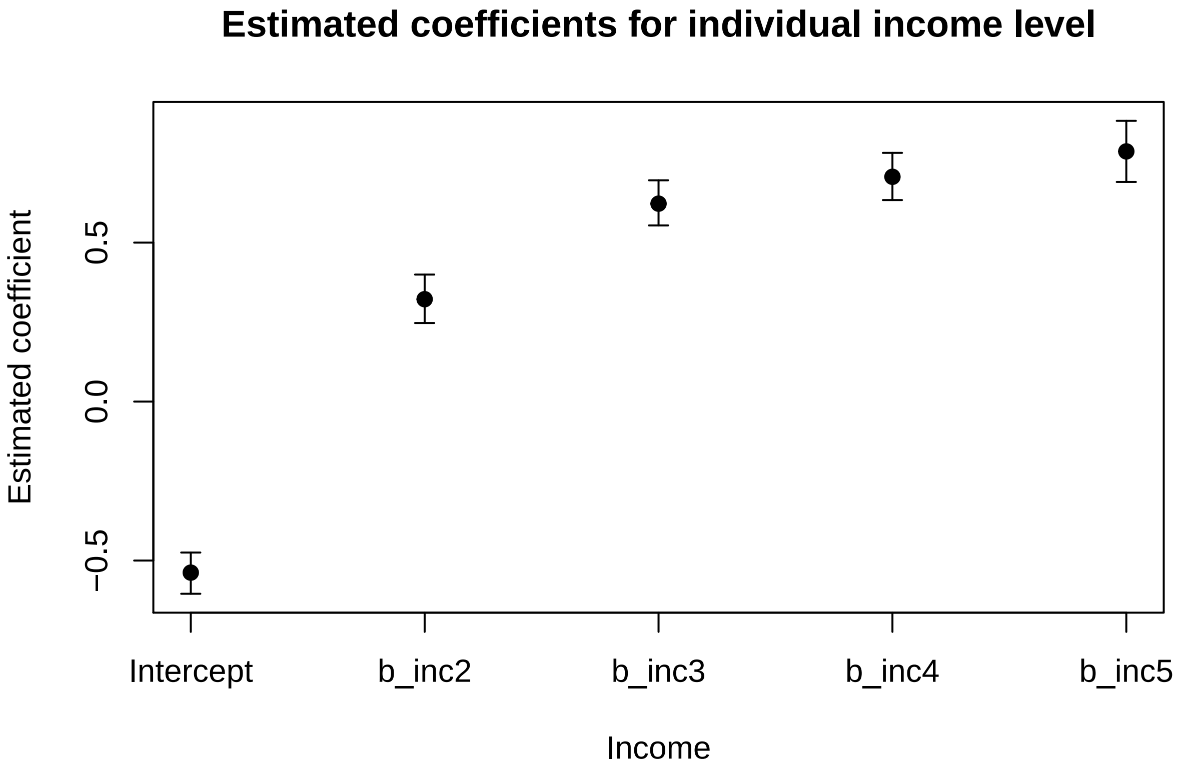
```r
load("Data/stan_inc_regression_data.Rdata")

# 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")
```

The stan function call returns a Stanfit object. To read more about it, you can visit http://mc-stan.org/interfaces/rstan.html and click on the RStan Manual and Vignette link. For us, we'll focus on the summary function provided because it'll allow us to access our regression coefficients. In the Stan code, the regression coefficients for income level are named "b_inc2", "b_inc3", "b_inc4", and "b_inc5", one for each level between two and five. There isn't a regression coefficient for the first income level because again, our other regression coefficients are showing us the difference in outcome between the first income level and them. However, for our simple model, we are in some sense capturing the first income level's effect in our Intercept term. So, let's pull out the mean of these coefficients and the 95% bounds for their credible intervals, which is a feature of Bayesian analysis. We'll plot these 95% credible intervals because if our model and its assumptions are correct, there's a 95% chance the credible intervals capture the emperical coefficients' means.

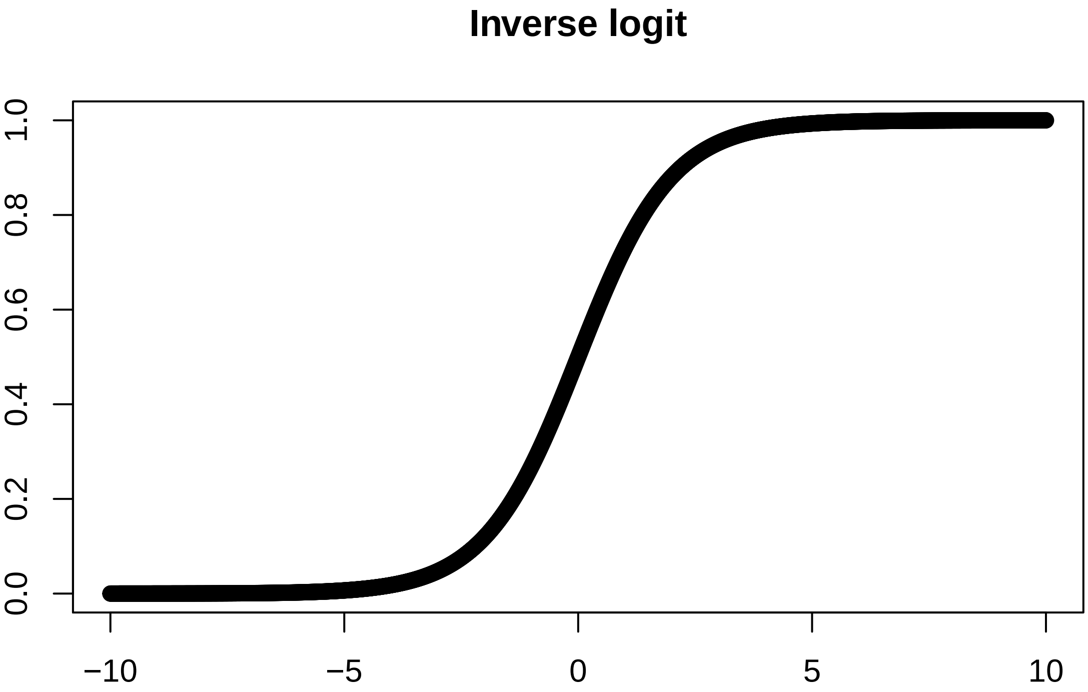
```r
# Pull out the regression coefficient and the
# information needed for a credible interval
interested_regression_rows <- c("Intercept", "b_inc2",
    "b_inc3", "b_inc4", "b_inc5")
interested_regression_cols <- c("mean", "2.5%", "97.5%")
ind_inc_regression_data <- summary(stan_2004_ind_inc_fit_obj)$summary[interested_regression_rows,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

We didn't have to save off the information in order to plot it, but we did so to make the plotting code cleaner. Also, we could pull out the median or any other point estimate of these coefficients and any other credible interval bounds, but the mean and the 95% bounds are provided to us in the summary call by default. However, if we had pulled point estimates of a certain type and the n% credible interval bounds for these estimates, our interpretation of the credible intervals would change to there being a n% chance the credible intervals capture the emperical coefficients of that type of point estimate.

```r
# 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)
```



You'll notice in the plot that the coefficents range from -0.5 to 0.5. Clearly, they aren't probabilities because probabilities are between 0 and 1. Instead, they're points on the inverse logit function curve, the untransformed version shown below.



In our plot, it is clear that going from a lower income category to a higher income category leads to higher coefficients. As we can see on the untransformed inverse logit plot, higher coefficients mean higher probabilities of voting Republican though how much of a difference depends on what two points you're comparing. Thus, when we look at the coefficients plot, it does look like an increase in an individual's income is correlated to an increase in voting Republican. So, there's support that the Democrats are the party of the poor.

However, the other part of the paradox is that the Democrats are the party of the rich states. In other words, are states with higher average income more likely to vote Democrat? We again use the binomial logistic regression model from before with the same definition of "success". However, instead of using the 2004 Annenberg Survey data, we'll be using the Census CPS data. Thus, while our state average income level variable will still have five categories, we'll divide it 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+.

Because our model is essentially the same as before, we can reuse our Stan code from before. However, we'll be feeding the code a different data set.

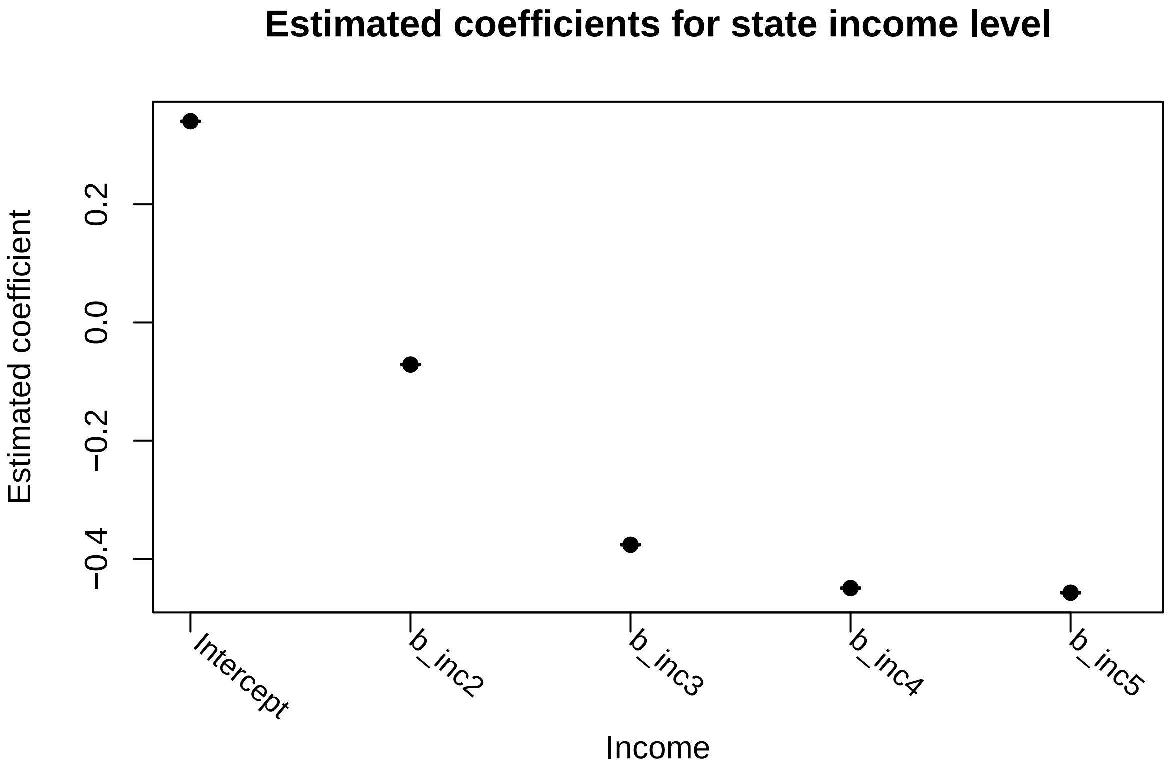
```r
# Get the data and the R package we'll be using for
# the analysis
load("Data/stan_state_regression_data.Rdata")

# 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")
```

We'll again summarize the data in the same way as before. Only this time, we'll be doing it using a function instead of looking at the Stanfit object itself.

```r
# 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")
state_inc_regression_data <- extract_results_from_stanfit_object(interested_regression_rows,
    stan_2004_state_inc_fit_obj)
```

And now, we'll plot it, but we'll again use a function that's wrapping the code called previously.

```r
# Plot the coefficients and their 95% credible
# interval bounds
plot_title = "Estimated coefficients for state income level"
plot_coefficients_with_credible_intervals(plot_title,
    interested_regression_rows, state_inc_regression_data)
```



In our plot, we see that as income level increases, the coefficients decrease. Based on our understanding of the inverse logit function, this means the probability of voting Republican decreases as well. So, as a state's average income increases, the likelihood of voting Republican goes down.

All in all, 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. Let's now try to understand the paradox.

Understanding this "paradox"

For our first attempt to understand this paradox, what if we treated the state's average income level and individual's income level as predictor variables and added their effects together? Our outcome is still the same as before so we'll use a binomial logistic regression as before. On the other hand, we now have two categorical variables. Still, for each variable, we'll compare the outcome of one level to those of all other levels. So, let's run the Stan code.

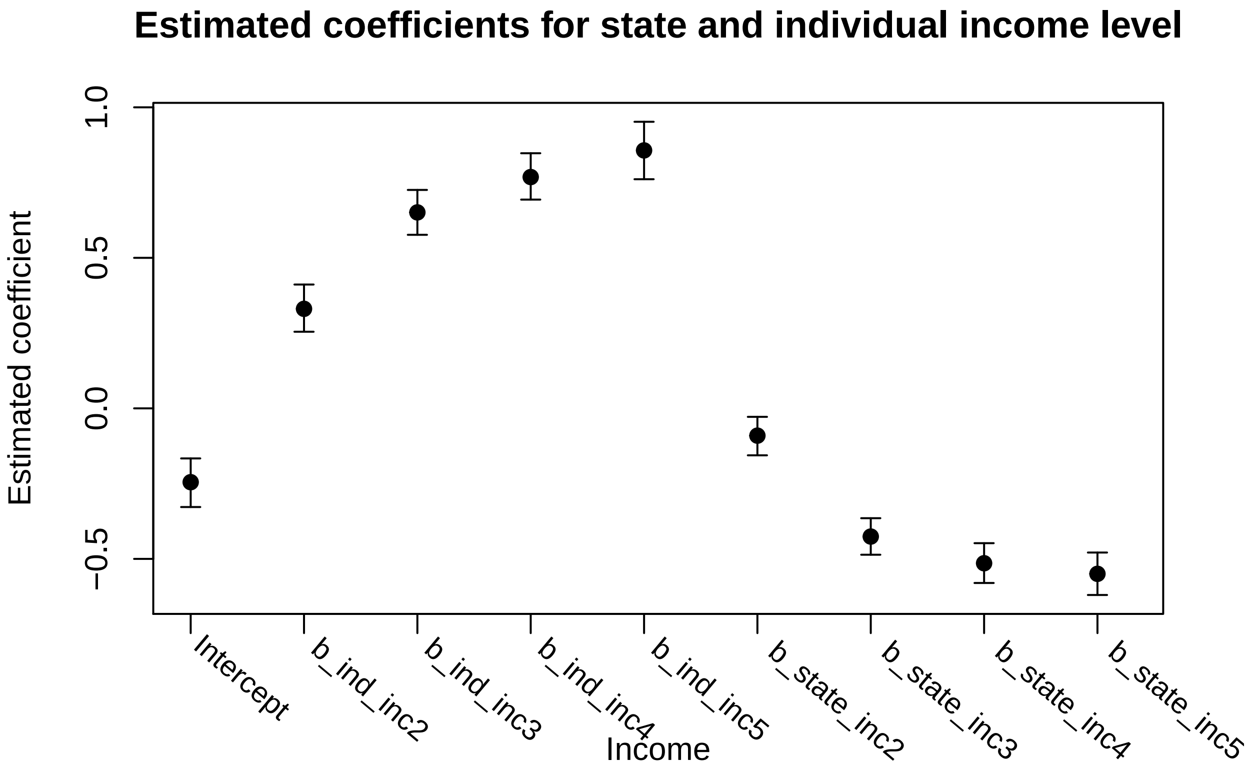
```r
# Get the data we'll be using for the analysis
load("Data/stan_state_inc_and_ind_inc_regression_data.Rdata")

# 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")
```

Again, let's pull the results. Our individual income level regression coefficients are "b_ind_inc2", "b_ind_inc3", "b_ind_inc4", and "b_ind_inc5" and our state average income level regression coefficients are "b_state_inc2", "b_state_inc3", "b_state_inc4", and "b_state_inc5". We don't have regression coefficients for the first or lowest levels because again, we've set our model to use the lowest level as the reference categories for comparisons. However, the intercept in some sense now represents the effect of the lowest individual income level and lowest state average income level on voting Republican.

```
# Pull out the regression coefficient and the
# information needed for a credible 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")
state_and_ind_inc_regression_data <- extract_results_from_stanfit_object(interested_regression_rows,
    stan_2004_state_and_ind_inc_fit_obj)
```

Now, let us plot these results.

```
# Plot the coefficients and their 95% credible
# interval bounds
plot_title = "Estimated coefficients for state and individual income level"
plot_coefficients_with_credible_intervals(plot_title,
    interested_regression_rows, state_and_ind_inc_regression_data)
```



From the plot, it looks like the regression coefficients for individual income level increase as income level increases whereas the regression coefficients for state income level decrease as income level increases. Because our model is linear on the logit scale, one way to interpret these regression coefficients is that:

1. For a given individual income level, the likelihood of voting Republican decreases as the state average income increases.
2. For a given state average income level, the likelihood of voting Republican increases as the state average income increases.

So, we haven't really gotten any more insight into the "paradox".

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's conduct this analysis. Our outcome will still be the same so we'll still use a binomial logistic regression model.

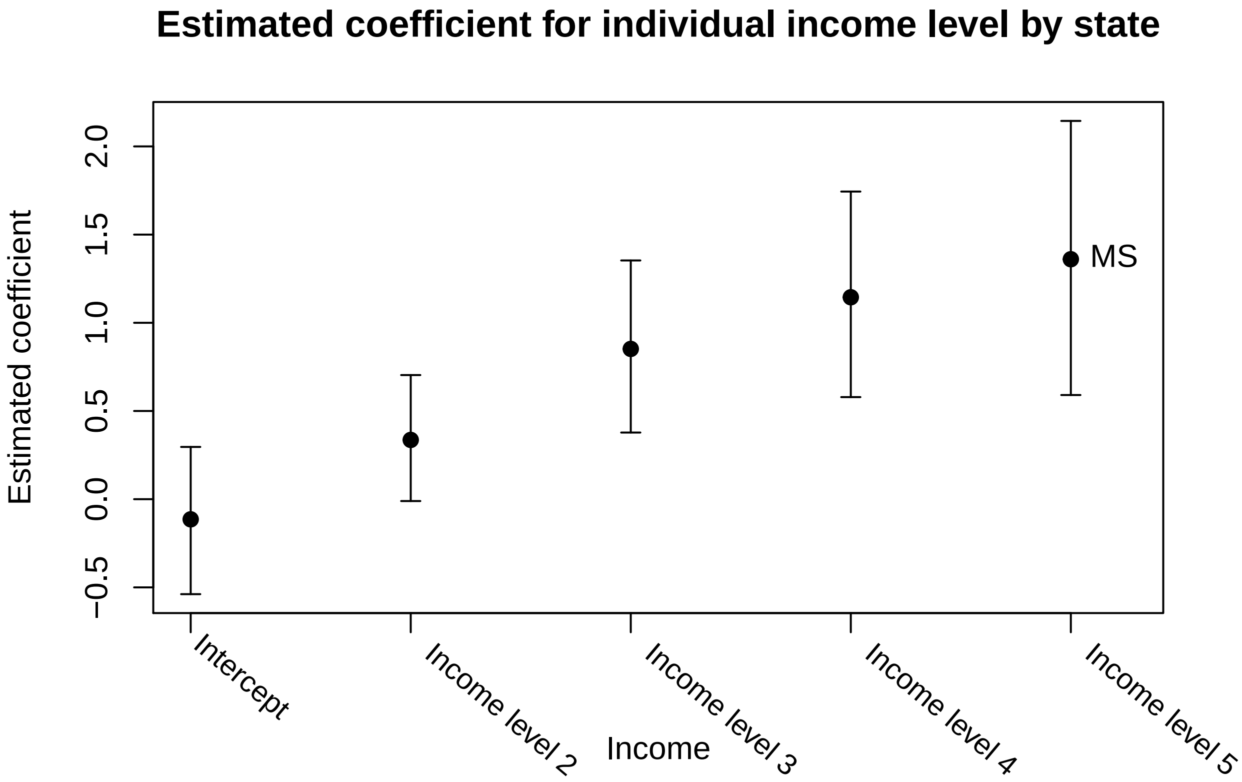
```
load("Data/stan_state_and_inc_regression_data.Rdata")

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

One way to interpret our results is to look at the effect of income for a given state. So, let's focus on Mississippi, Ohio, and Conneticut because they represent a poor, middle-income, and rich state respectively.

```
# Pull out the regression coefficient and the
# information needed for a confidence interval
ct_regression_rows <- generate_state_labels(6)
oh_regression_rows <- generate_state_labels(34)
ms_regression_rows <- generate_state_labels(24)
state_and_inc_regression_data <- extract_results_from_stanfit_object(c(ct_regression_rows,
    oh_regression_rows, ms_regression_rows), stan_2004_state_and_inc_fit_obj)
```

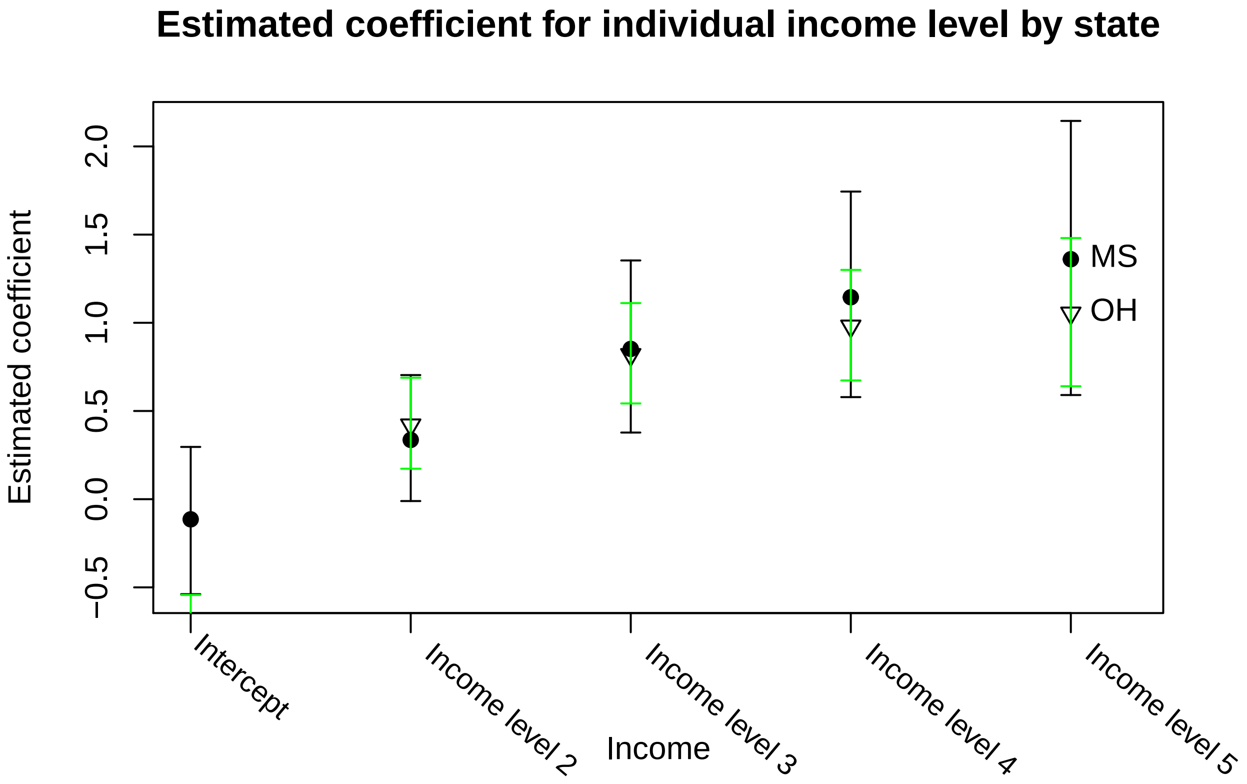
Let's plot the regression coefficients of Mississippi first.

```
# Plot the regression coefficients with bars Plot
# Mississippi first
plot_title = "Estimated coefficient for individual income level by state"
x_labels = c("Intercept", "Income level 2", "Income level 3",
    "Income level 4", "Income level 5")
plot_coefficients_with_credible_intervals(plot_title,
    x_labels, state_and_inc_regression_data[ms_regression_rows,
        ])
text(5, state_and_inc_regression_data[ms_regression_rows[5],
    "mean"], labels = c("MS"), pos = 4)
```



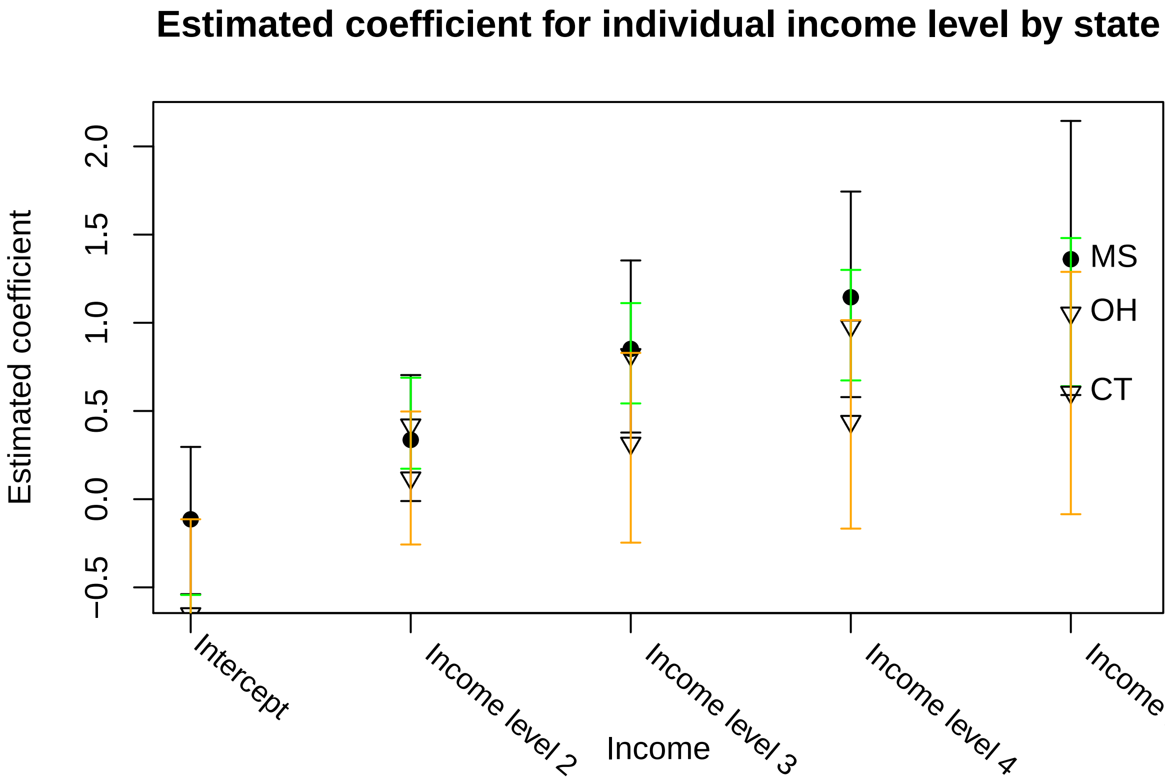
It still looks like the individual income graph from before. What about Ohio, the middle-income state?

```
# Plot Ohio
add_state_to_plot("OH", oh_regression_rows, color = "green",
    state_and_inc_regression_data[oh_regression_rows,
        ])
```



Huh. It looks like the change in coefficients isn't as much as before. What about Conneticut, the rich state?

```
# Plot CT
add_state_to_plot("CT", ct_regression_rows, color = "orange",
    state_and_inc_regression_data[ct_regression_rows,
        ])
```



So, it seems that while individuals with higher income are more likely to vote Republican, this trend is smaller in richer states. These states are more likely to vote Democrat because there's less support across income categories. There isn't a paradox.

Conclusion

Let's take a step back and see what we've learned. We sought to understand the paradox that even though the Democrats claim to be the party of the poor, states with higher average income tended to swing Democrat. When we analyzed the 2004 National Election Annenberg Survey data using a binomial logistic regression on individual income level and the number of Republican voters, we did indeed see on our coefficients box plots that the likelihood of voting Republican increases as income level increases. On the other hand, when we analyzed the Census CPS data using a binomial logistic regression on state average income level and the Republican vote share, we saw on our coefficients box plots that the likelihood of voting Republican decreases as average income level increases. So, we confirmed that the paradox seemed to exist.

When we sought to decipher the paradox, we saw that using individual income level and state average income level as predictors model didn't help. Instead, it was through a hierarchical regression model in which we allowed the intercept and individual income to vary by state. In our coefficient regression plots, we saw that while individuals with higher income are more likely to vote Republican, the effect is smaller in states with higher average income. This explained our paradox.

So, 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). Or, you could do it yourself and compare your results to their paper.

In any case, you've had an opportunity to go over a few simple regression models and see a hierarchical regression model.

Appendix

Code for Individual Income Model

```
# 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)
source("red_state_tutorial_helper_functions.R")
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")

# Plot the coefficients Extract the Stan
# coefficient estimates and their 95% credible
# interval bounds
interested_regression_rows <- c("Intercept", "b_inc2",
    "b_inc3", "b_inc4", "b_inc5")
ind_inc_regression_data <- extract_results_from_stanfit_object(interested_regression_rows,
    stan_2004_ind_inc_fit_obj)

## Plot the coefficients and their 95% credible
## interval bounds
plot_title = "Estimated coefficients for individual income level"
plot_coefficients_with_credible_intervals(plot_title,
    interested_regression_rows, ind_inc_regression_data)
```

Code for State Income Model

```
# 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)
source("red_state_tutorial_helper_functions.R")

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")

# Plot the coefficients Extract the Stan
# coefficient estimates and their 95% credible
# interval bounds
interested_regression_rows <- c("Intercept", "b_inc2",
    "b_inc3", "b_inc4", "b_inc5")
state_inc_regression_data <- extract_results_from_stanfit_object(interested_regression_rows,
    stan_2004_state_inc_fit_obj)

## Plot the coefficients and their 95% credible
## interval bounds
plot_title = "Estimated coefficients for state income level"
plot_coefficients_with_credible_intervals(plot_title,
    interested_regression_rows, state_inc_regression_data)
```

Code for State and Individual Income Model

```
# 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)
source("red_state_tutorial_helper_functions.R")
load("Data/stan_state_inc_and_ind_inc_regression_data.Rdata")

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

# Plot the coefficients Extract the Stan
# coefficient estimates and their 95% credible
# interval bounds
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")
state_and_ind_inc_regression_data <- extract_results_from_stanfit_object(interested_regression_rows,
    stan_2004_state_and_ind_inc_fit_obj)

## Plot the coefficients and their 95% credible
## interval bounds
plot_title = "Estimated coefficients for state and individual income level"
plot_coefficients_with_credible_intervals(plot_title,
    interested_regression_rows, state_and_ind_inc_regression_data)
```

Code for Hierarchical Model for Individual Income by State

```
# 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
model_file = "Stan/mixed_effects_regression_for_state_and_inc.stan"
stan_2004_state_and_inc_fit_obj <- stan(file = model_file,
```

```r
    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 <- generate_state_labels(6)
oh_regression_rows <- generate_state_labels(34)
ms_regression_rows <- generate_state_labels(24)
state_and_inc_regression_data <- extract_results_from_stanfit_object(c(ct_regression_rows,
    oh_regression_rows, ms_regression_rows), stan_2004_state_and_inc_fit_obj)

# Plot the regression coefficients with bars Plot
# Mississippi first
plot_title = "Estimated coefficient for individual income level by state"
x_labels = c("Intercept", "Income level 2", "Income level 3",
    "Income level 4", "Income level 5")
plot_coefficients_with_credible_intervals(plot_title,
    x_labels, state_and_inc_regression_data[ms_regression_rows,
        ])
text(5, state_and_inc_regression_data[ms_regression_rows[5],
    "mean"], labels = c("MS"), pos = 4)

# Plot Ohio
add_state_to_plot("OH", oh_regression_rows, color = "green",
    state_and_inc_regression_data[oh_regression_rows,
        ])

# Plot Mississippi
add_state_to_plot("CT", ct_regression_rows, color = "orange",
    state_and_inc_regression_data[ct_regression_rows,
        ])
```