Crime in the USA: Using SAS to Analyze Recidivism Rates

Philip Mayevskiy, San Francisco State University

.

ABSTRACT

This paper focuses on using the Annual Parole Survey (2014), produced by the Bureau of Justice, Statistics to analyze recidivism rates in the United States criminal justice system. The goal of this paper is to illustrate how to use SAS University Edition to analyze the Annual Parole Survey dataset and interpret the results to make accurate conclusions. Though this paper focuses on using statistical analysis to understand recidivism rates in the United States, the techniques applied are widely applicable to various other government statistics.

INTRODUCTION

The Bureau of Justice Statistics (BJS) has been releasing their Annual Parole Survey since 1980. The latest release being the 2016 data. However, the BJS regularly updates the information, therefore, more recent datasets are often less complete than the older versions.

The 2014 versions of the BJS's parole survey was released in late 2015 (BJS 2015). This dataset describes conditions and circumstances for parolees in state and federal juristictions. Because of the volume of factors that describe the status of parolee, as well as the descriptive statistics for the parolee, there are almost 170 variables that are recorded in the survey.

Some of the more notable variables include parolee demographics (race, sex), types of offenses, and number of people on parole who reoffend.

The Annual Parole Survey does not overtly include a variable on recidivism rates, however this piece of information can be extrapolated easily from the data provided by the survey. To do this, steps need to be taken to clean the data to increase readability and make it usable for statistical analysis. The 2014 data has many missing values and is not formatted properly for analysis; once the data has been properly formatted, then various statistical techniques in SAS can be used in interpreting which variables are important when analyzing recidivism rates in the United States.

CLEANING THE DATA

The BJS includes a codebook with the downloadable dataset. This deciphers what the variables in the dataset represent and what the values mean. For example, in the 2014 Parole Survey, missing values are presented by either a '-8' or '-9', this is important to note so that SAS reads these values in as the correct missing value rather than a numeric value.

To mark these values so SAS knows that it is not numeric and to ignore it in calculations, you use the following code:

```
DATA datain.Parole;
SET datain.da36320p1;

format _numeric_ 9.4;

   IF (EXINCNEW = -8 OR EXINCNEW = -9) THEN EXINCNEW = '.';
RUN ;
```

'EXINCNEW' being the variable name for the number parolees returned to incarceration with new sentence. This 'IF-THEN' statement ensures all '-8' and '-9' values will be replaced by a '.', which is the SAS default for missing values. The 'format' code ensures that the data remains in its numeric type. The next step is to do this for every variable in the data. Though this is tedious, it is vital to read the data in correctly so that spurious conclusions can be avoided.

Supplementary text which includes this code to remove missing values is dependent on where the dataset is obtained. Downloading it from the icpsr.umich.edu website (which offers all the parole surveys) provides a separate SAS file, "Supplemental_syntax.sas", that provides all of this code for you.

Since the goal is to analyze recidivism rates, obtaining the response variable is the next step. There are two variables that stand out, "Returned to incarceration with new sentence" (EXINCNEW) and "no new sentence" (EXINCREV). Because these values are unique for each state, dividing this sum of these two variables by the total parole population at the beginning of the year (TOTBEG) in each state yields the recidivism rate (RRATE) for that state.

The code for this would be:

```
RRATE = ( EXINCNEW + EXINCREV ) / TOTBEG ;
```

Because some of the states have missing values in the original variables used to create the RRATE variable, they need to be dropped from the rest of the analysis. Here is an example of the code to do this:

```
IF (STATEID = 06 or STATEID = 09 or STATEID = 11) then delete;
```

The code to delete certain observations references a State ID. Each State ID number corresponds to an individual state. Dropping all the states that had many missing values throughout the dataset, 8 states were deleted (CA, CT, NH, NM, MS, WA, WI, DC).

The next step in preparing the data for analysis is to make sure the variables are compatible with the response variable. Many of the variables are measured as populations, and to make them useful in the analysis, they must be changed to be proportional. In the same way we made the RRATE, we can make all the variables proportions:

```
WHITE = WHITE / TOTRACE;
BLACK = BLACK / TOTRACE;
HISP = HISP / TOTRACE;
```

In this case, the variables that correspond with the race population were changed to proportional. The denominator depends on the specific variable, therefore, understanding the Codebook is important. In this case the total race population was 'TOTRACE'. Other Variables that correspond to types of offender, type of sentence, etc. must also be made proportional.

The final aspect of the data is to set all the binary, or categorical, variables to proper format. In the parole data, the binary values are marked as either '1' or '2'. To set this to the right format it needs to be '0' and '1'. This is because to run a statistical test on categorical variables it needs to be in ANOVA format. The basic way to do this is the following:

```
IF GPS = 1 THEN GPS = 0;
IF GPS = 2 THEN GPS = 1;
```

For this example, GPS was used. But again, all the binary (or categorical) variables must be changed to this format.

ANALYZING THE DATA

Demographics

Now that the data is cleaned and prepared for analysis, initial regression tests can be run to see which variables are best for understanding recidivism. Starting with demographic data I used PROC REG to run forward, backward, and stepwise selection tests on the data.

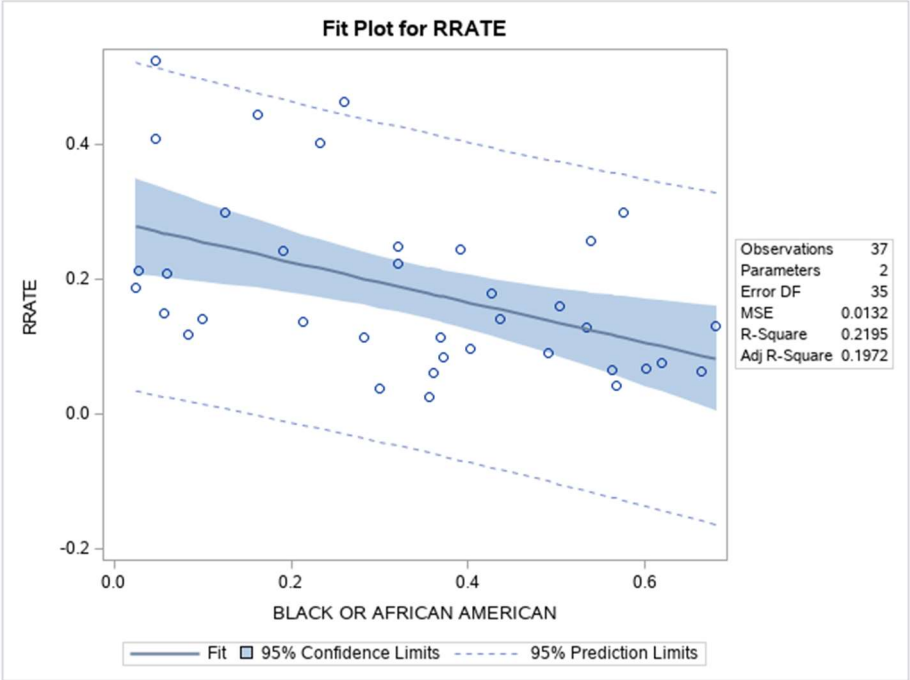
```
model rrate = WHITE BLACK HISP ASIAN / selection = forward;
model rrate = WHITE BLACK HISP ASIAN / selection = stepwise;
model rrate = WHITE BLACK HISP ASIAN / selection = backward;
```

The results were in order of significance as follows:

Forward: BLACK, WHITE, HISPANIC
Stepwise: BLACK
Backwards: BLACK

The p-value for the RRATE = BLACK regression was 0.0035 and it had a negative coefficient. What this means is that the higher concentration of Black or African American people in a parole population the lower the recidivism rate.

Using a Fit Plot for the equation displays the points, trend, and confidence interval. There are, however, 2 points outside of the confidence interval.



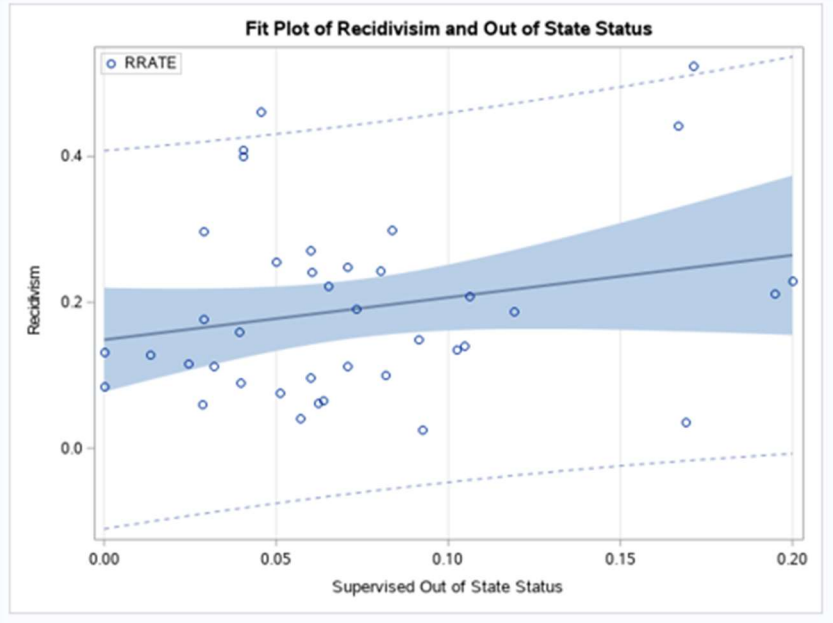
Removing those two points increases to p-value to 0.01 and tightens in the confidence bands so much so that 2 more points become outliers. Doing the same thing again we get a p-value of 0.02 and the confidence bands become narrower again. It may be that there are a handful of states that have a very large recidivism rate with a lower Black or African American parole population that may be affecting this trend. These would be states like Utah, Colorado, Minnesota, and North Dakota.

Doing a regression on RRATE = FEMALE we see that more Females as a percentage of the parole population increases likelihood to reoffend, however, the p-value for this regression is 0.07. Therefore, it is insignificant if we use a p-value = 0.05 as our standard.

Variables of Interest

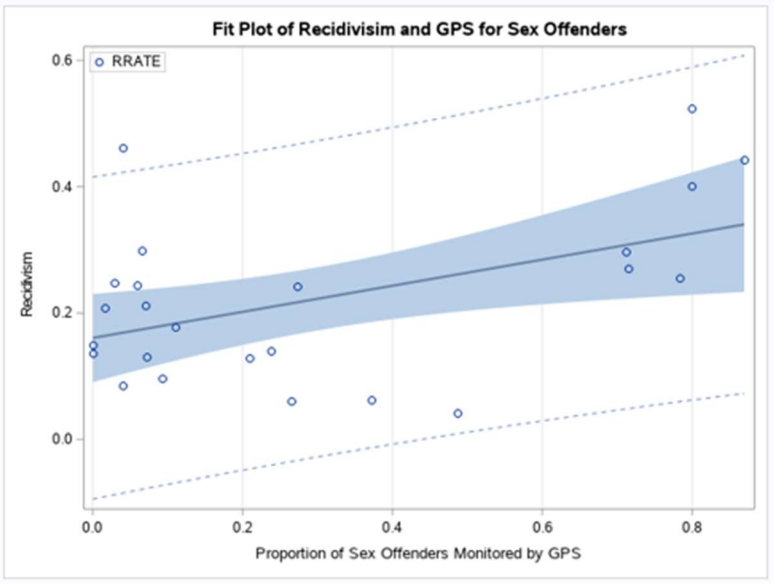
There are a variety of different variables in this dataset that can potentially be very helpful in understanding recidivism . Many of the variables could not be used because there were too many missing values or the 'significant' variable was not relevant. For example, variables like "Unknown Offense" are significant but most of the observations have a value of '0' and only a select few states recorded anything for this column, so dropping the variable would be the best option. But with the remaining variables we can do forward, stepwise, and backwards selection again to find the best subset of variables. Doing the analysis showed none of the variables were significant but the most significant variable seemed to be proportion of people supervised out of state. To display a custom Fit Plot we use:

```
proc sgplot data=datain.Parole;
 /*Create a title at the top*/
title 'Fit Plot of Recidivisim and Out of State Status';
 /*Use Regression to visualize the trends and confidence intevals*/
reg y=rrate x=outstate / cli clm nomarkers;
 /*Create a scatter plot to see where the points lie*/
scatter y=rrate x=outstate;
 /*Create a legend to easily identify what the points are*/
keylegend 'scatter' / location=inside across=1 position=topleft;
 /*Label the x-axis*/
xaxis grid label = "Supervised Out of State Status";
/*Label the y-axis*/
yaxis label = " Recidivism ";
run;
```

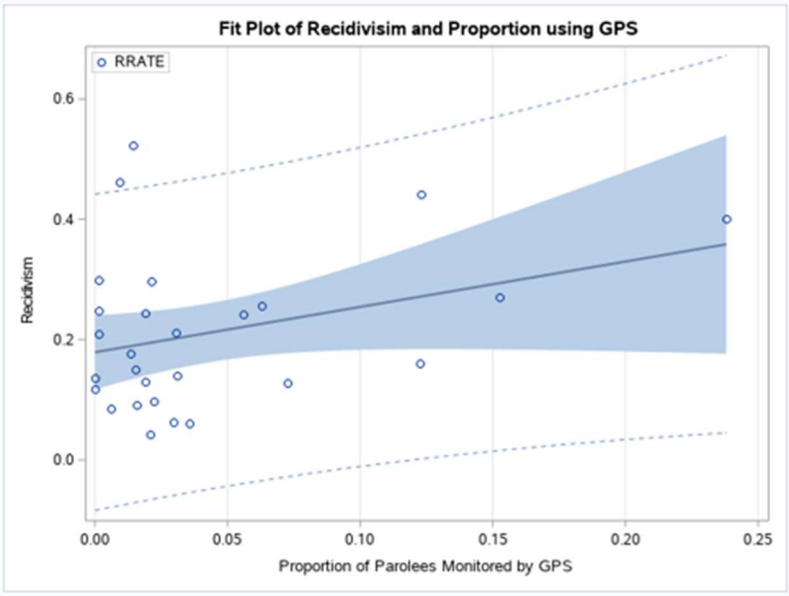


This regression RRATE = 'Supervised out of State' had a p-value of 0.10. The trend in this graph (one that shows more recidivism with more 'out of state' parolee's) seems to be highly influenced by two points in the top right of the plot. It does seem reasonable that this variable would affect recidivism in some way or another. But to make this conclusion we either need better data or to instead survey only parolees 'out of state' to see if they have a higher rate directly rather than knowing the proportion of the parolee's in each state that are in this category.

Another variable of interest is GPS Tracking. These variables have a many missing values but it still have enough observations enough look into. These are the GPS variables fit plots:



The regression for the Proportion of Sex Offenders Monitored by GPS with respect to Recidivism was significant. States with higher proportion of sex offenders with GPS devices have higher recidivism. It may in fact be the case that states with higher recidivism rates are more likely to put GPS tracking devices on their parolee sex offender population instead. As stated before, the only way to really know is to survey the sex offenders individually rather than analyzing them as a proportion of the population.



As for the Proportion of Parolees Monitored by GPS Devices with respect to recidivism, it is insignificant.

The categorical variables also came up short. None of them were significant for predicting recidivism.

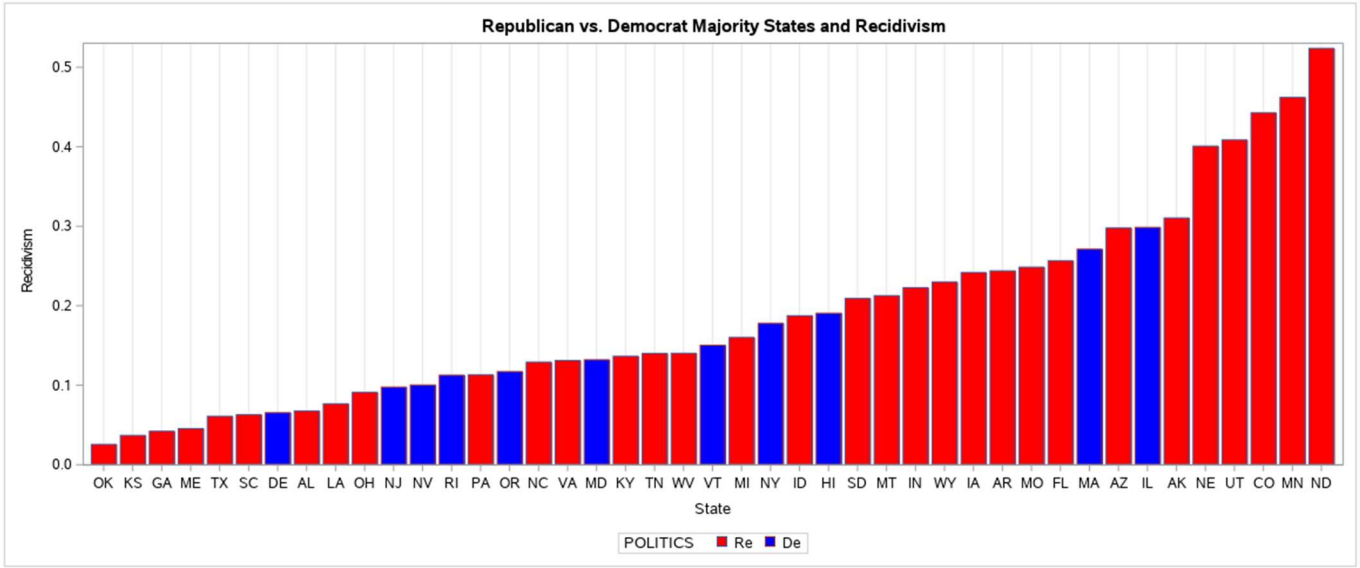
Politics of Recidivism

Kff.org (The Henry J. Kaiser Family Foundation 2018) offers data on political party's by state in 2018. Specifically, data on which states are majority Democrat or majority Republican in terms of the State House of Representatives and the State Senate. Using our STATE variable and the data on which states are affiliated with which party, we can create a new variable with the following code:

```
POLITICS = STATE;
IF STATE = 'AL' THEN POLITICS = 'Republican';
IF STATE = 'AK' THEN POLITICS = 'Republican';
```

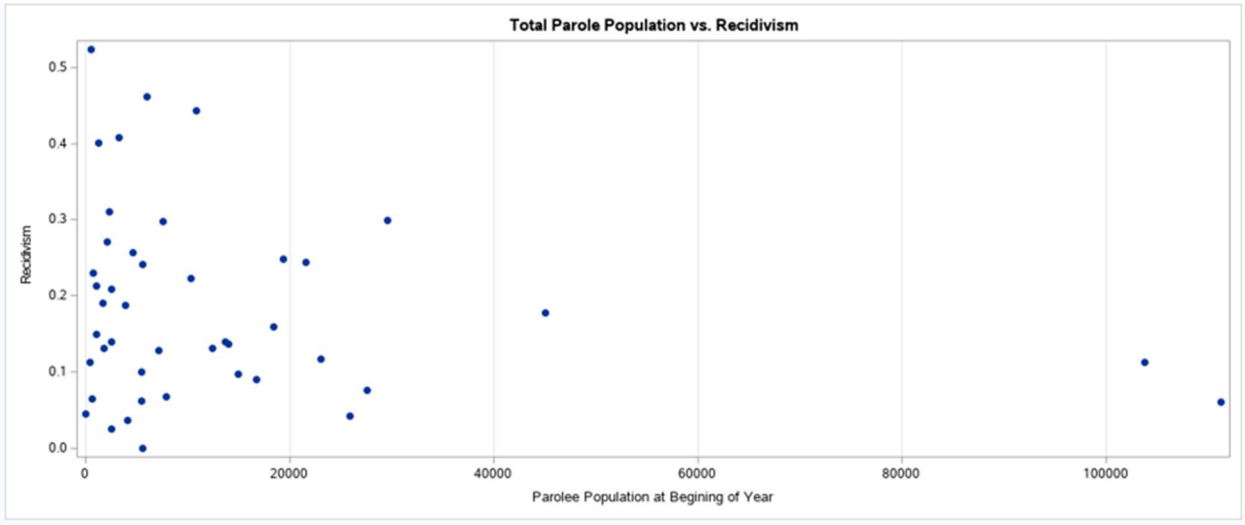
Once every state has been labeled with the corrects political party, we can create a graph to get a visual on the distribution of recidivism rate by political party. We can use SGPLOT for this:

```
proc sgplot data=datain.Parole;
   title 'Republican vs. Democrat Majority States and Recidivism';
   /*Set colors for Bar*/
   styleattrs datacolors =(blue red) ;
   /*Create Bar Graph and order it by RRATE*/
   vbar STATE / response = rrate group= POLITICS categoryorder= respasc;
   xaxis grid label = "State";
   yaxis label = " Recidivism ";
run;
```



From this graph we can surmise that republican led governments make up the extreme ends of the recidivism rates. Republicans make up the political majority in their states about twice as often as democrats do (The Henry J. Kaiser Family Foundation 2018). Just this fact may play a role. But it's also the case that states will a low population density tend to be more Republican.

To test if population plays a role we can make a scatterplot of the Parolee population size with respect to the recidivism rate. When we do that with the SGPLOT like before, we get:



The scatterplot makes it clear that there is a cone shape. There is lots of variability at the beginning where there is a very small population of parolees. And this is to be expected. With a small population, a small amount of variability could result in a big change in value whereas in a larger population, the mean stays relatively steady.

CONCLUSION

Analyzing raw data can begin to paint a better picture of the circumstances on the ground. Understanding the recidivism rates in your own state and knowing the environments in which they thrive can help people make more informed decisions on what problems to address.

However, national data could be so encompassing that few conclusions can be made to find the reason as to why something is the way it is. Because there are so many variables that affect the reason for recidivism, variables that try to generalize a large population often will not be enough for a good predictor. To explain something, you must single out what you are looking for. Say, for example, Using GPS Tracking devices on parolees, you cannot just get a population size and use that to find a correlation. An ideal situation is a randomized study across states and their populations. Differences in recidivism rates among states can say more about the state parole population itself than about the methods in which they treat parolees. There is an uncountable number of factors, whether social, economic, or something else, that affect recidivism. And most of these factors cannot be made into data.

REFERENCES

"Data Collection: Annual Probation Survey and Annual Parole Survey." *Bureau of Justice Statistics (BJS)*, www.bjs.gov/index.cfm?ty=dcdetail&iid=271.

"State Political Parties." *The Henry J. Kaiser Family Foundation*, The Henry J. Kaiser Family Foundation, 12 June 2018, https://www.kff.org/other/state-indicator/state-political-parties/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22desc%22%7D.

United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. Annual Parole Survey, 2014. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-12-02. https://doi.org/10.3886/ICPSR36320.v1

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Philip Mayevskiy
pmayevskiy@gmail.com