# Monitoring-Federal-Criminal-Sentences

1996 - 2015

Project Topic and Relevance

This project is interesting from my perspective because of the past experiences I have had with the federal criminal justice system.

Since the establishment of the Sentencing Reform Act of 1987 which introduced the gridline system that was originally supposed to achieve a standardized procedure in sentencing, there has been research that disproportionate sentences have been given to black males in the federal system. The rhetoric used to instigate mass incarceration is perhaps being reused in the new administration to justify more unforgiving punishments. The dramatic increase in incarceration rates in the mid-70s has produced long-term social, economic, and institutional impact on American society. Two major concerns in contemporary literature surrounding mass incarceration include the causes and consequences of mass incarceration. The penal system has produced inequality by disproportionately reshaping the life course of a generation of young minority men and negatively influencing their economic opportunity and family life in disadvantaged communities and neighborhoods (Raphael and Stoll 2009; Travis, Western, and Redburn 2014; Western, Pettit, and Lopoo 2006).

Although violent and drug crimes have been one of the focal points of research in mass incarceration, little is known about the direct comparison of punishments between the different types of crimes. Furthermore, understand how the demographics of the prison population has shifted over the recent years. This project aims to evaluate the severity of different types of cases in relation to one another, while seeing if other legal and extralegal characteristics have played a role in contributing to the outcome of the sentence.

There have been many researches done on the outcomes of cases of the federal system, but not the internal processes that produces the results. Although this research does not shed light on the lack of research due to data constraints, the findings of this project does show sentencing disparity that will in turn attract social scientists’ into investigating the internal processes of the federal legal framework.

Finding Data

To investigate the possible correlation between sentencing outcomes with legal and extralegal characteristics, twenty years of federal sentencing data was downloaded from ICPSR. One advantaged this dataset offered over the other datasets is that it includes legal both extralegal characteristics of the case defendant, documents consistent variables over time, and has been cited by numerous academic citations. Due to my original research interests, I did not have to adopt this dataset for another. Another advantage that this data has over other types of data is that it is recorded on an individual basis. Utilizing standardized federal court documents such as detailed pre-sentence reports, the data was able to accurately capture demographic and factual data determines the outcome of the case. Because pre-sentence reports are constructed by probation, and the facts within the report are bargained by the prosecution and the defense based on the discovery of the case, the data will capture the outcomes of those legal process. Similarly, because the data series includes case information like the disposition of the case, it opens the possibility to examine how each type of legal process may also affect the outcomes of the case.

Data Structure and Data Munging

Despite being a mostly clean and "tidy" data series, there were some issues that this project had to hurdle prior to the data analysis process. The two major steps that this project encountered were importing data, lumping overly granular data, remove any observations that required special coding, and merging data.

Importing data was one major step that R users might encounter. Because the data prior to 2005 are in SPSS, STATA, or even ASCII format, this project had to utilize the “asciiReader” and “readr” package from CRAN with the corresponding ASCII helper. Furthermore, the SPSS and STATA dataset provided from 2001-2003 were not readable by R, therefore the data had to be loaded in SPSS and converted to csv. prior being used in R.

Having data that is overly-granular could sometimes be a disadvantage, as it is in the case for the purpose of this research project. Because the data had incrementally more variables each year, the researcher of this research had to go through the codebook each year and document the differences each variable. For example, the type of the offense from 1996-2000, “OFFTYPE”, included 40 crime categories, but the variable after that year was retired, and the data from after that year included 37 categories. For that reason, this project had to go through individual codebooks and ensure the accuracy of each categorization. Other extralegal characteristic variables such as race, education, and citizenship had to undergo a similar procedure.

In addition to selecting variables with consistent categories, this research project also converted all character and factor categorizations to numeric values manually due to some inconsistencies over the years. The process is completed in the “clean.r” script. The corresponding category of each variable can be found in the documented and descriptive codebook.

The dataset categories were also lumped prior to merging. After formatting the data to consistent numeric in the previous step, the data from individual years are also lumped into major categories. The main reasons for lumping categories into more generalizable categories are its ability to be visualized and modeled later in the project.

Despite the challenging but necessary steps that this project encountered, this project was able to produce a merged dataset that is consistently categorized and easy to be analyzed.

Reproducible Research

The three reasons why this research project is reproducible is because 1) it includes data that is published and easily accessible by anyone, and that 2) the scripts that this project has utilized are all included in the “scripts” folder, and 3) the codebook is explicit and explains the demographic makeup of the data series population.

The individual datasets are accessible for anyone under conditions specified by ICPSR, meaning that people can document and complete the same data munging process in any data program. Additionally, the data series provides codebooks from individual years to help researchers identify how variables might have changed from last year, if there are any new variables that are created from any old variables, or if there are any variables that are being discarded.

The scripts of this project were written to be reproducible. Each code leading up the data that was used in the data analysis has been well documented, and this research suggests that other researchers review the process prior to utilizing the code. There are also labs that are well commented and describes potential methods that novice R use when producing their own research.

Lastly, the codebook produced by this research project independent from ICPSR provides a good description of the data variables that this project has decided to select and utilize. Although some of the variables were not used in the data analysis, it informs future researchers of what variables to consider when preparing their dataset for analysis. The codebook also includes summary information regarding the variables that were selected in the data structure and munging process.

In general, this research is reproducible because the easiness to access data, organized and commented scripts and codebooks that comes with the data, and the inclusion of any work that was being used in the process of producing the research. Lastly, individuals who have an interest in contacting the author of this research may contact Johnathan Hsu at [hsujohnathan@gmail.com](mailto:hsujohnathan@gmail.com).