

Memories of Slavery in Mass Incarceration

A Quantitative Analysis of Carceral Systems in the U.S. South

April 20th, 2020

Slavery and mass imprisonment are genealogically linked and [...] one cannot understand the latter—its timing, composition, and smooth onset as well as the quiet ignorance or acceptance of its deleterious effects on those it affects—without returning to the former as its historic starting point and functional analogue.

— Loic Wacquant, *From Slavery to Mass Incarceration*

Introduction

Scholars and activists like Wacquant have argued for over a century against the strict dichotomization of “slavery” and “freedom.” After formal abolition in 1865, “freedom” was not universally granted to slaves; instead, many iterations of carceral systems have restricted freedom in ways America no longer formally recognizes as slavery. From the ghetto, to Jim Crow laws, to exclusion from institutions of power, to the modern prison, the practice of imprisonment and punishment has shifted and transformed. We live now in an age of mass incarceration, one where Black men are six times as likely to be incarcerated and laws are created as part of a racialized project of incarceration. This paper attempts to describe our modern prison as an echo of former plantations, that the age of slavery is the present and not wholly behind us. We use current and historical data to describe relationships between incarceration, slavery, and urbanicity. **We find that counties with high incarceration rates now also had high slave populations in 1860s, a significant result even when accounting for the accompanying high Black population rate. Additionally, prison population rate is highest in rural areas,** mirroring known geographies of plantations.

For full disclosure, this project is inspired by a feature story on The Pudding called [The Shape of Slavery](#), written by Yale Professor Bill Rankin and reporter Matt Daniels. We provide statistical analysis to supplement their beautiful, beautiful, maps (seriously, check them out). # Data

All observations in this project are counties. Variables used in this analysis are below, with abbreviated names in parentheses.

Key variables

- **Incarceration rate rate, 2015. (`prison_rate`)** This is the main response variable for my paper, and is defined as residents per 100,000 population in a given county who are currently in prison. Note that this *not* the number of prisoners in a county, but in fact the number of people who were sent to prison (possibly elsewhere). This is an important distinction when thinking about many urban residents who are sent to prison in rural areas.
- **Black population rate, 2015. (`blackpop`)** The proportion of a county’s residents between ages 16 and 64 who are Black.
- **Urbanicity, 2015. (`urbanicity`)** This is a factor variable that describes the “urbanicity” of a county, or how urban a county is. Its values are “rural”, “small/mid”, “suburban”, and “urban.”

- **Slave population rate, 1860.** (`slaves`) This is the main independent variable for my paper, and is defined as the proportion of population that were slaves.
- **Whether the county has a prison and what type of prison.** (`prison_type`) I show in an ANOVA that rural counties have higher prison population rates. Would this be because more rural counties are more likely to have prisons, and some sort of cultural effect makes prisons increase incarceration rate in a county? To study this, I include a categorical variable indicating whether a county has a prison or not.

The first four variables were taken from the [Vera Institute's repository of prisons and jails data](#). Their dataset contains granular data by race and gender at the county level over forty years, but I take data from 2015 to avoid time-series data. Slave population rate was taken from the U.S. Census in 1860, standardized by the team at [IPUMS NHGIS](#). `prison_type` was taken from the Prison Policy Initiative's [Correctional Facility Locator](#).

Additional control variables

To attenuate and contextualize the main results, additional control variables were added that center around socioeconomic status.

- **Unemployment rate**, from the Bureau of Labor Statistics' [Local Area Unemployment Statistics Program](#).
- **Particular matter less than 2.5 μm in the air**, from the Environmental Protection Agency's [EJSCREEN tool](#). This is a proxy for air quality.
- **Median household income**, from the Census Bureau's [Small Area Income and Poverty Estimates\(SAIME\) Program](#).
- **Rent burden**, from the [Eviction Lab](#) at Princeton. Rent burden is defined as the proportion of income spent on rent.
- **Gini coefficient**, originally taken from the U.S. Census Bureau and taken from the [County Health Rankings](#) report. The Gini coefficient is a measure of economic inequality on a continuous scale from 0 to 1, with higher values indicating greater economic inequality.
- **Proportion of single-parent households**, also taken from the County Health Rankings report.

Data cleaning

An important note on syntax and code

I spent some time learning R outside of class this semester, and use it here. I stick to a set of libraries known as the “tidyverse,” made for making life easier in R than the base R code. There are enough new functions that it would not make sense to describe each function as it appears, but many of these functions are named intuitively and can be guessed from their names. I also provide descriptions of each code block as summaries of the code inside.

One important symbol is `%>%`, known as the pipe operator. This lets users chain series of commands, saving writing and organizing code into readable “sentences.” Its formal definition is “take the output of the command on the left and insert it as the first argument to the command on the right,” and intuitively it can be read as “and then do.” For example, if I wrote `f(x) %>% g()`, this would be equivalent to `g(f(x))` or “do `f(x)` first and then run `g()` on that output.”

Importing and basic cleaning of data.

I read in a current counties boundary file. Data from NHGIS was aggregated to one observation per county, and counties in states that only had 0 for slave population (i.e. states that had outlawed slavery by 1860,

which were all Northern states) were replaced with NA values to be excluded from the analysis. Rates were created by dividing the slave population by the total population of the county.

Spatial manipulation

The most challenging part of data cleaning came from the issue of U.S. counties in 1860 being wildly different from U.S. counties in 2015. Codes for the two systems were not compatible, and borders for many counties over the past have changed over the past years. To my knowledge, no crosswalk exists to map these two systems together. To address this problem, I took centroids from 1860 counties and observed which 2015 counties they resided in. In coding terms, I used the `st_centroid()` and `st_join()` commands in the `sf` package to map data from these data together. Figure 1 explains this process with Virginia as an example. Code for Figure 1 is suppressed but available in the accompanying Rmarkdown file.

This is not an optimal solution because counties are heterogeneous, and the best way to approach this would have been to take block-level population data to weight overlaps between 1860 and 2015 shapefiles. I considered this outside the scope of this project as only a good approximation was needed. Figure 1 suggests that the criteria for “good” seems to have been met.

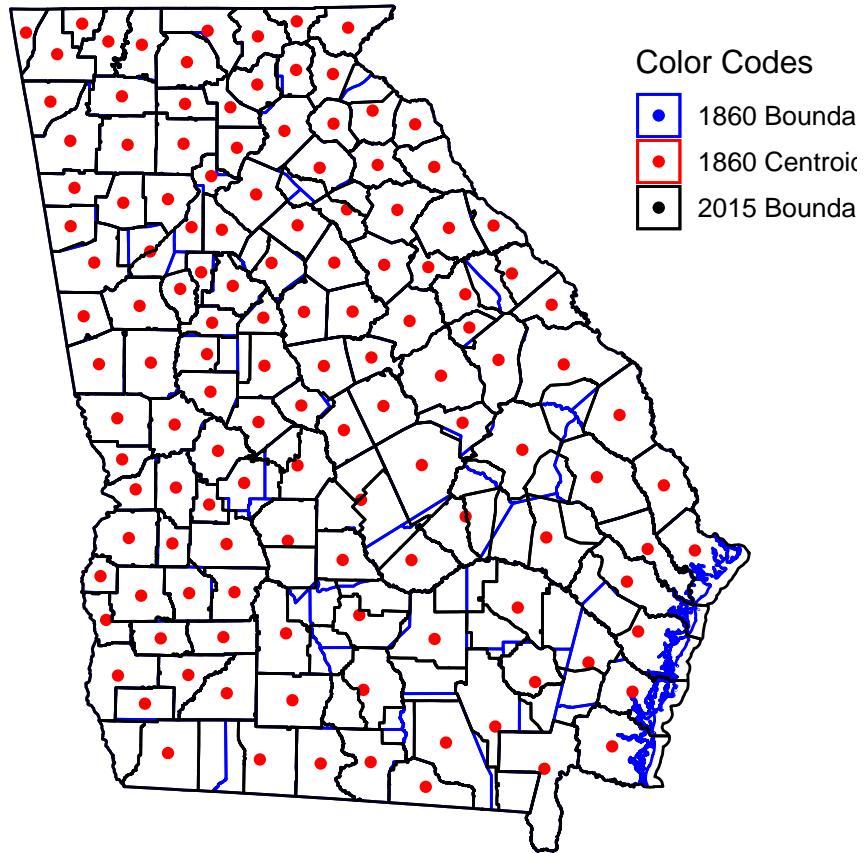


Figure 1: : Diagram of County Overlaps and Changes

I also aggregated the age- and sex-specific data on Black population rates to one observation per county.

To summarise, I read in the 1860 county shapefile with `sf::st_read()`, fit the shapefile with the correct projection (i.e. calibrated the longitude-latitude coordinates to rely on the same system as my 2015 counties

shapefile), transformed the data to 2015 counties as noted above, joined the shapefile with my dataset by the “GISJOIN” identifier, and aggregated 2015 counties where two or more 1860 counties had fallen into.

Other data cleaning of key variables

The Vera dataset is incredibly tidy, with one row for each county in a given year. Minimal data cleaning was needed, and I only filtered for 2015 data, created rates for population density and Black population rate, and selected columns of interest.

The prisons location data also was fairly easy to obtain, although not as clean as the Vera dataset. This information came from the Prison Policy Initiative’s Correctional Facility Locator 2010, which combines a few different tables from the U.S. Census and corrects errata for public presentation. I wasn’t able to download this directly, so I scraped this using rvest as we saw in class. The webpage also did not have county codes listed directly, but had them inside of links in several places. I pulled this out with the `stringr` string manipulation package and regex syntax. Almost every county has a local correctional facility (jail), but for this project I was only interested in prisons data, so I pulled only state and federal correctional facilities. It could be the case that these included jails, but a quick visual check scrolling through the website indicates that these were mostly prisons. Finally, I aggregated the dataset to only have one row per county, with different labels depending on if a county had a state prison, federal prison, or both.

Data cleaning of control variables

I changed columns to numeric, removed missing values, joined state and county FIPS codes to one five-digit identifier, and joined them through the `*.join()` functions from the tidyverse.

Code is suppressed here for clarity and length, but included in the Rmarkdown version of this file in case students are interested.

Merging

The five-digit FIPS codes that uniquely identify counties allowed us to merge variables from different datasets to the same counties. We dropped spatial attributes because they were no longer needed. We also changed NA values for `prison_type` to “No Prison” to be a more descriptive value (this was only possible once we merged the location data together).

Descriptive Plots and Summary Information

To restate our questions of interest clearly:

1. Is there an association between 2015 county incarceration rates and 1860 slave population rates?
2. If so, is this just because counties with higher slave populations in 1860 now have higher Black populations, and the prison system is more likely to incarcerate Black residents (i.e. not because of a direct “memory” between 1860 and 2015)?¹
3. If #1 holds, is it just because prisons are generally located in isolated areas like rural counties, and prisons have some sort of cultural effect that increases probability of being incarcerated in a given environment?

¹Note: It could definitely be argued that counties with high slave populations having high incarceration rates *only because of high Black populations* is still an effect of the “memory of slavery,” but my findings indicate a distinct, though related, “memory” effect.

4. (Playing with and off of #3) How does this trend relate to classifications of urbanicity, considering scholars have written about “the urban ghetto” as a carceral system? Are urban counties more likely to have higher incarceration rates?
5. What *is* this problem due to, if not completely the things above?

Descriptive statistics

To get started exploring these data, here is a glimpse of my continuous variables:²

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
slaves	2,058	15.33	21.53	0.00	0.00	27.29	92.50
prison_rate	1,629	681.49	392.88	8.27	389.23	884.19	2,788.86
blackpop	2,074	11.41	15.74	0.22	1.21	14.69	87.21
unemp	2,076	5.70	1.78	2.00	4.50	6.60	24.50
mhi	2,076	48,794.20	12,718.37	23,014.00	40,293.00	54,207.25	125,900.00
pm	2,076	9.58	1.22	5.21	8.88	10.44	13.32
rentburd	2,076	29.32	3.78	10.00	27.00	31.50	50.00
gini	2,076	43.13	3.66	32.60	40.50	45.50	60.10
singpar	2,076	9.02	2.42	1.62	7.44	10.25	21.87

Histogram

A histogram of incarceration rates reveals a unimodal right-skewed distribution. Given a Box-Cox transformation on a preliminary model suggested an exponent of .18 and many demographic variables operate on the log scale, I performed a log transformation and removed one outlier (on the lower end). The new logged variable is not completely normally distributed and is a bit left-skewed, but produces normally distributed residuals in the end model.

Normal quantile plot

To further emphasize that our raw response variable is not normally distributed and that our transformed response variable is near-but-not-quite normally distributed, we present two normal quantile plots of raw and logged incarceration rates.

Box plots

To provide a snapshot of our categorical variables, one box plot each for `prison_type` and `urbanicity` are presented with the dependent variable being 2015 incarceration rates.

Correlation test with permuted p-values

To assess correlation between continuous, we present a matrix plot for all variables. The labels were hard to read with nine variables, so I manually added text on the top of the graph. It is clear here that some degree of

²The `stargazer` package in R was used for formatting.

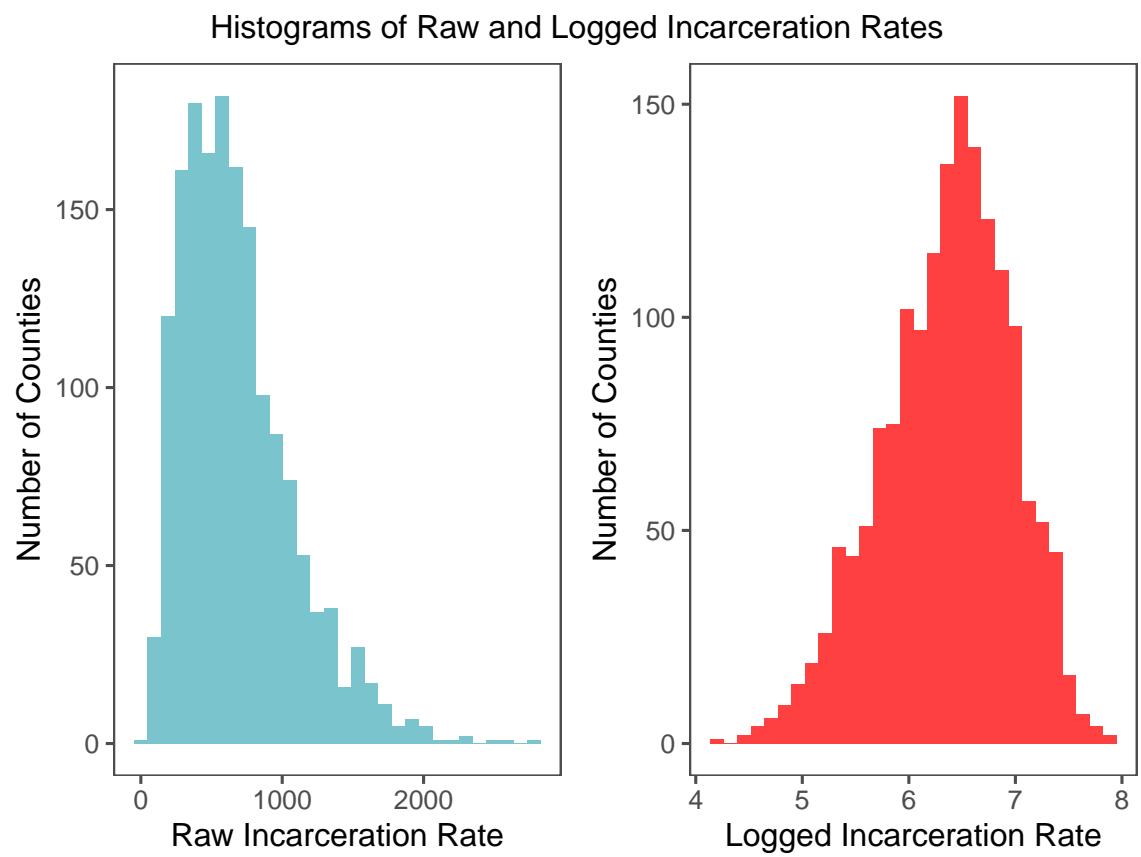


Figure 2: Histogram of Incarceration Rates

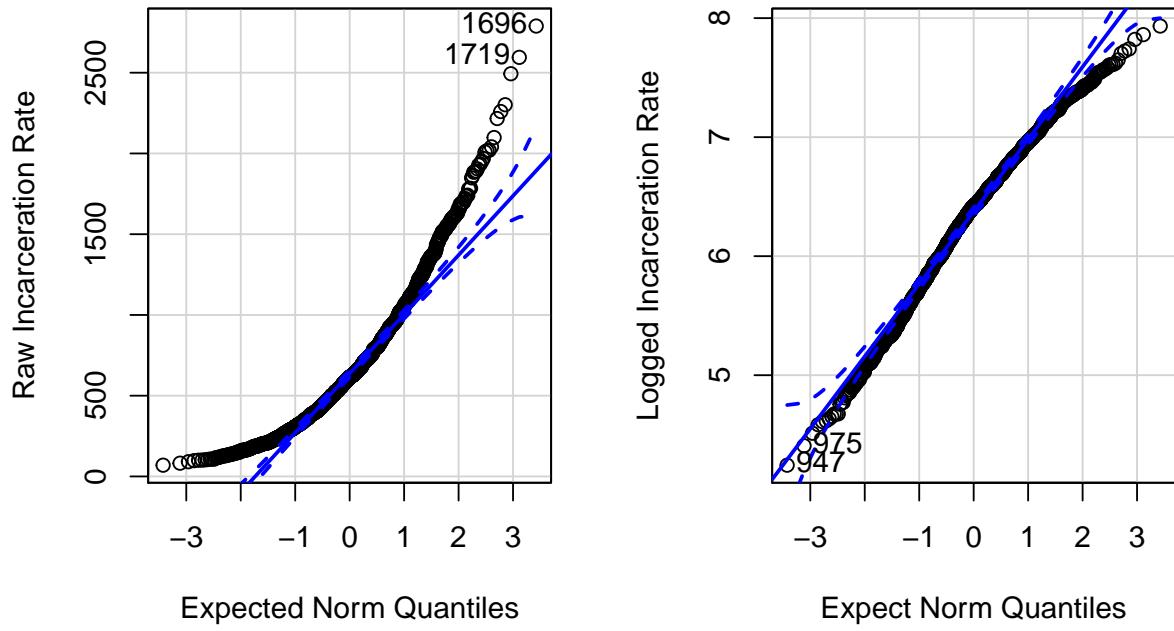


Figure 3: Normal Quantile Plots of Incarceration Rates

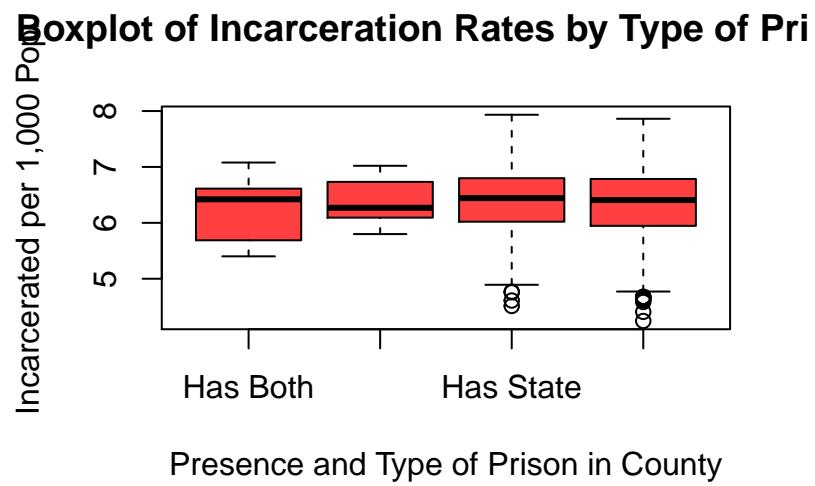


Figure 4: Boxplot by Type of Prison

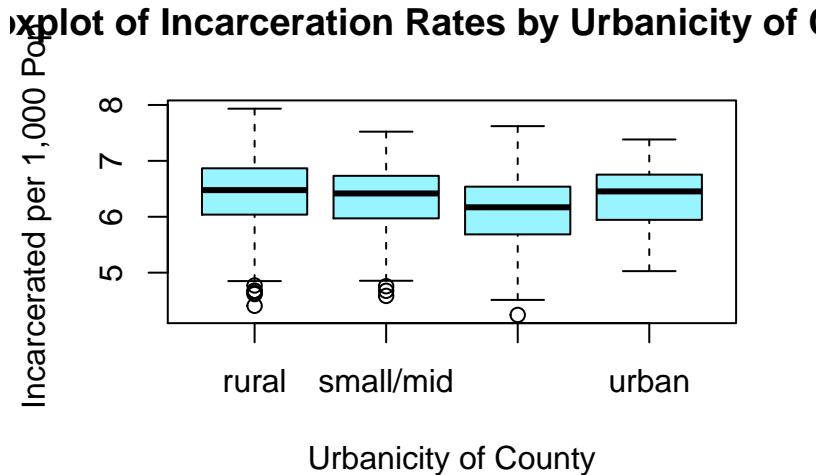


Figure 5: Boxplot by Urbanicity.

Focusing on our two main variables of interest, a traditional correlation test revealed that 2015 incarceration rates and 1860 slave population rates were positively correlated with a coefficient of .365, with an associated p-value of $6.65e - 52$ (in other words, extremely small), and a 95% confidence interval of .322 to .407. A **permutation correlation** test with a million simulations returned a p-value of 0, indicating that the true p-value is less than $1e - 6$. These preliminary results are encouraging pieces of evidence that incarceration rates and slave population rates are associated.

Analysis

To begin, we assert that counties with high rates of incarceration now had high slave population rates in 1860, as evidenced with a bootstrapped correlation plot and best subsets regression. We find this interesting given that mass incarceration is usually thought to hit urban environments hardest, but areas with high slave population rates were likely rural areas with large plantations. The boxplots above also indicate ambiguous results as to how this occurs. To explore this trend, we present an ANOVA result, a traditional t-test with bootstrapped confidence intervals, and a permutation test for urban-rural counties to show differences in incarceration rates across the urban-rural scale.

Best subsets regression

We show four models here to show robustness in our results – that is, seeing the trends of interest across multiple types of models shows that it is not just an artifact of our methods. We present a “base model” with 1860 slave population as the sole predictor, then show three models given by `regsubsets()` and selected by BIC and adjusted R-squared. Code is again suppressed for clarity and length, and can be found in our Rmarkdown file. The response variable in all models is logged incarceration rate.

Reordering variables and trying again:

The BIC-selected model shows a highly significant relationship between slave population rate and current prison incarceration rates, with a one-percentage-point increase in slave population rate associated with

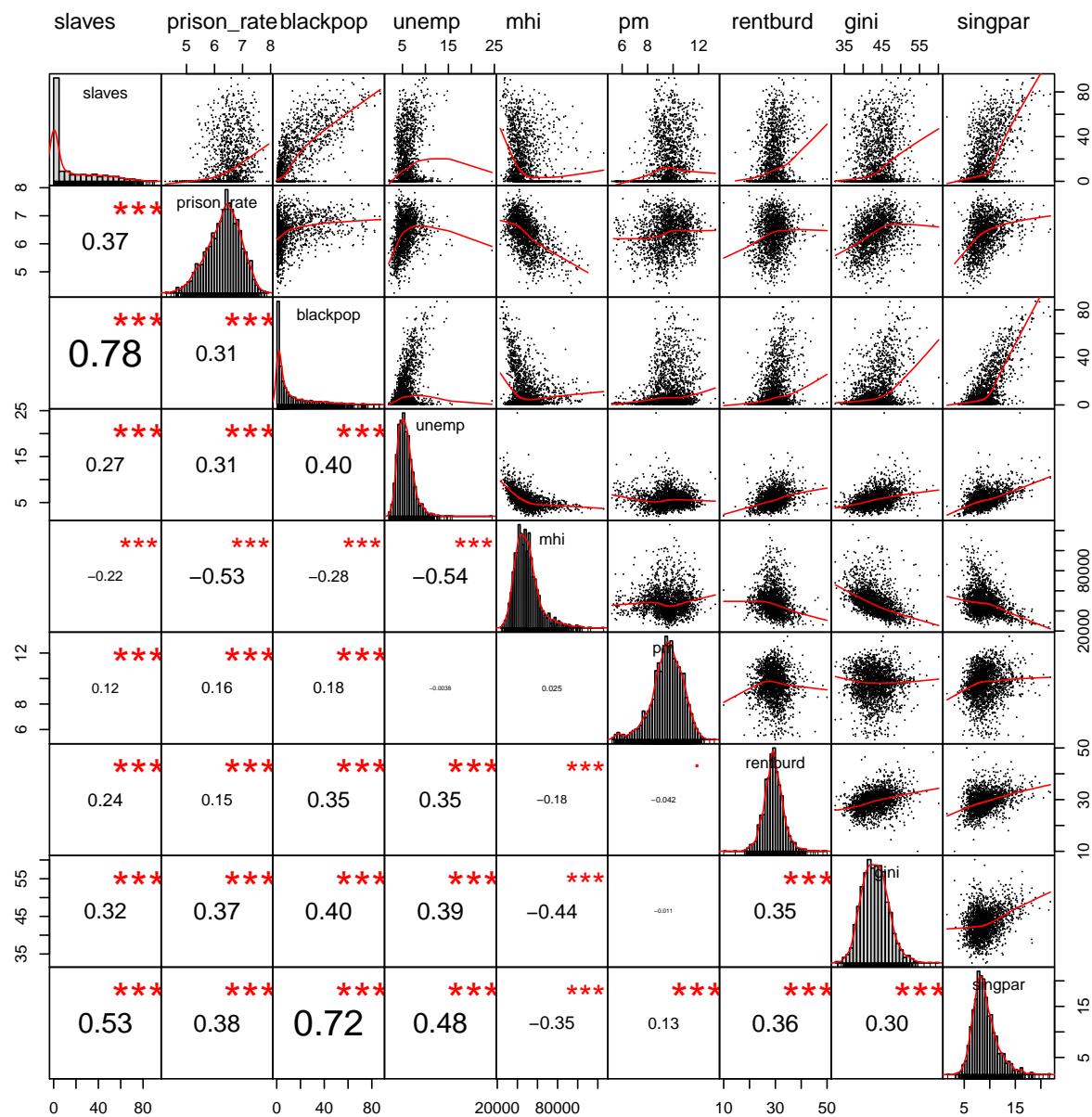


Figure 6: Matrix Plot of Continuous Variables

Table 2: Results of Best Subsets Regression

	Model chosen by:		
	Base Model	BIC	Adj. R2
slaves	0.011*** (0.001) <i>p</i> = 0.000	0.007*** (0.001) <i>p</i> = 0.000	0.007*** (0.001) <i>p</i> = 0.000
urbanicity_rural			-0.150* (0.082) <i>p</i> = 0.068
‘urbanicity_small/mid‘			-0.144* (0.081) <i>p</i> = 0.075
urbanicity_suburban			-0.027 (0.083) <i>p</i> = 0.745
urbanicity_urban			
blackpop		-0.010*** (0.002) <i>p</i> = 0.000	-0.010*** (0.002) <i>p</i> = 0.000
‘prison_type_Has Both‘			0.138 (0.150) <i>p</i> = 0.360
‘prison_type_Has Federal‘			-0.083 (0.124) <i>p</i> = 0.503
‘prison_type_Has State‘			0.050* (0.028) <i>p</i> = 0.076
‘prison_type_No Prison‘			
unemp			-0.021** (0.009) <i>p</i> = 0.016
mhi		-0.00002*** (0.00000) <i>p</i> = 0.000	-0.00002*** (0.00000) <i>p</i> = 0.000
pm		0.077*** (0.010) <i>p</i> = 0.000	0.068*** (0.010) <i>p</i> = 0.000
gini		0.023*** (0.004) <i>p</i> = 0.000	0.024*** (0.004) <i>p</i> = 0.000
singpar		0.057*** (0.007) <i>p</i> = 0.000	0.057*** (0.008) <i>p</i> = 0.000
Constant	6.193*** (0.017) <i>p</i> = 0.000	5.051*** (0.227) <i>p</i> = 0.000	5.503*** (0.286) <i>p</i> = 0.000
Observations	1,609	1,609	1,609
R ²	0.133	0.388	0.396
Adjusted R ²	0.133	0.386	0.391
Residual Std. Error	0.562 (df = 1607) ¹⁰	0.473 (df = 1602)	0.471 (df = 1595)
F Statistic	247.101*** (df = 1; 1607)	169.420*** (df = 6; 1602)	80.505*** (df = 13; 1595)

Note:

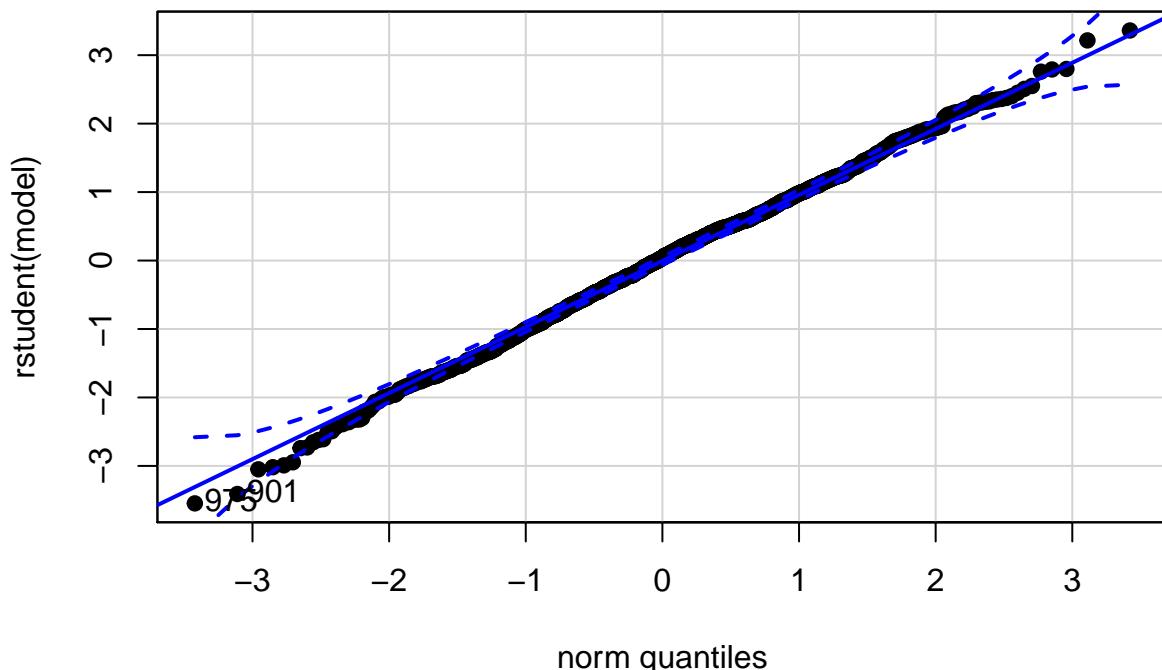
*p<0.1; **p<0.05; ***p<0.01

about a .00721 increase in logged incarceration rates. This corresponds to about a 0.7% multiplicative increase of incarceration rates for a one-percent increase in slave population rate.³

Notably, this trend is robust across multiple models – though the coefficient of “slaves” decreases as we add more variables, slave population rate is significant even as we control for potential mediating variables like the Black population rate, median household income, income inequality with the Gini coefficient, and so on. Though these coefficients are significant as well, seeing slave population rate positively predict incarceration rates in all models suggests a distinct effect between incarceration and slavery.

We verify the BIC-selected model with residual plots. The residuals appear almost perfectly normally distributed⁴. Though there is some heteroskedasticity in the residuals, the extent is mild and within our personal standards of “acceptable heteroskedasticity.”

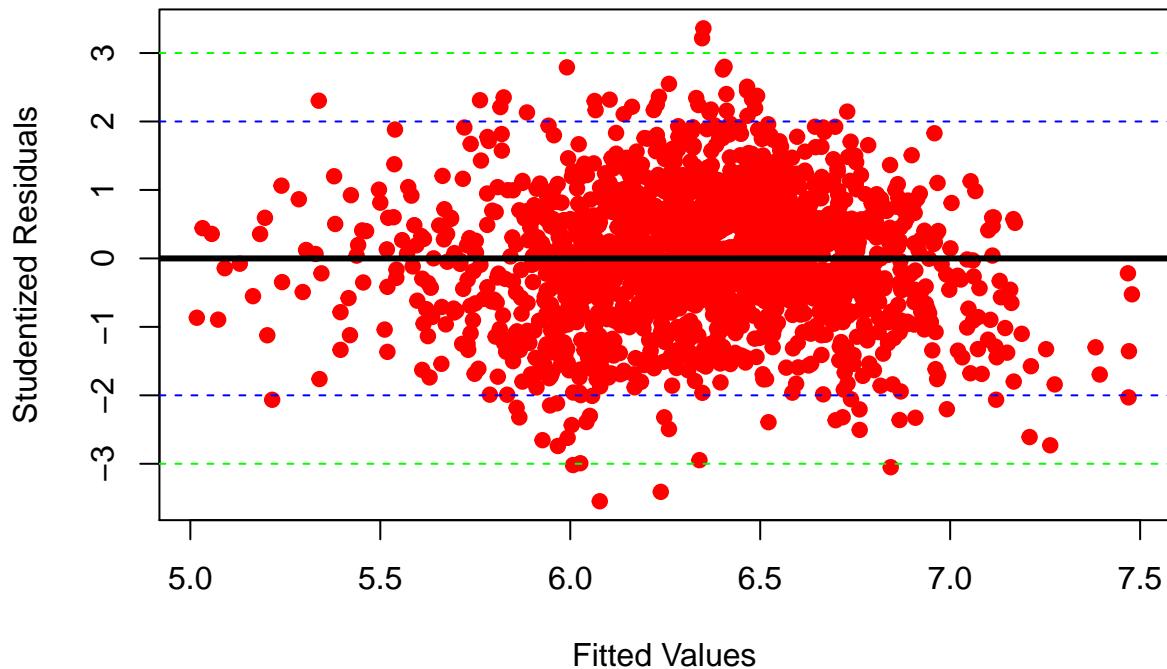
NQ Plot of Studentized Residuals, Residual Plots



³ $e^{7.212e-03} = 1.007239 = 0.7\%$ increase. See [here](#) for further explanation on logged dependent variables and their interpretation.

⁴a trend that I've actually never seen before with these kinds of demographic variables.

Fits vs. Studentized Residuals, Residual Plots



Urban-rural rates of incarceration

Our intuition before we began this analysis held that higher slave population rates are usually associated with rural counties (where plantations were) and high incarceration rates are usually associated with urban areas (at the epicenter of mass incarceration). Off of this idea alone, counties with high slave population rates in 1860 would have lower incarceration rates now. However, our results indicate the opposite. This is made even more intriguing seeing that the model lends suggestive evidence for rural counties having lower incarceration rates than urban counties – “urbanicity_rural” has a negative coefficient and a p-value of .068, indicating that its incarceration rate is lower than the baseline of “urbanicity_urban.”

To explore this further, we present an ANOVA analysis, a t-test with bootstrapped confidence intervals, and a permutation test.

ANOVA

We first present an ANOVA analysis of all urbanicity types. From smallest to largest, these are “rural”, “small/mid”, “suburban”, and “urban.”

This provides evidence that there is some difference in incarceration rates between different sizes of counties, and suburban counties appear to have the lowest incarceration rates. However, the key difference of interest (“urban-rural”) does not seem to reveal large differences.

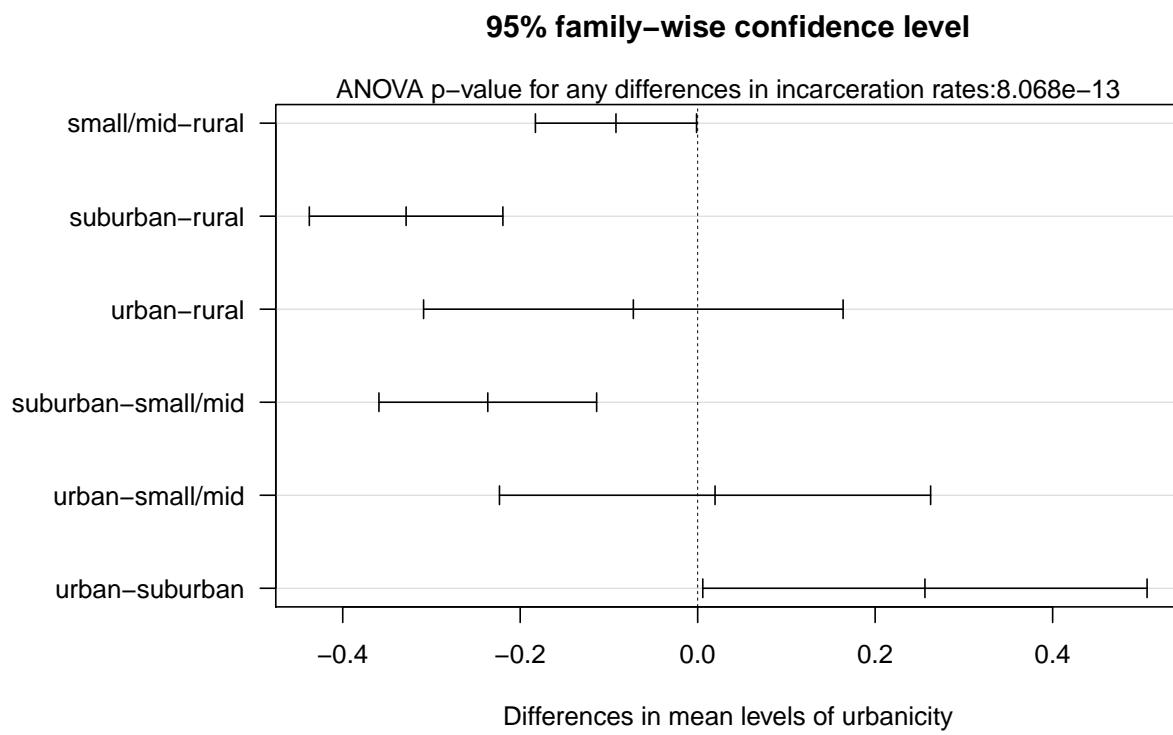


Figure 7: ANOVA of Incarceration Rate and Urbanicity, with Tukey Corrections for Confidence Intervals

t-test and bootstrapped confidence interval

To drive this point home, we performed a **t-test and created bootstrapped confidence intervals** on just the incarceration rates of rural counties and urban counties. The traditional t-test revealed a 95% confidence interval for the difference between -.0917 and .2368, and the bootstrapped confidence interval revealed a similar range between -.088 and .235.

Though of course, absence of evidence is not evidence of absence, these results suggest that whatever difference there is in incarceration rates between urban and rural counties exists is not a consistent one and one cannot say that either group has a higher incarceration rate than the other.

- at least one each of boxplot, scatterplot, qqplot, residual plot, histogram
- t-tests, correlation, and bootstrapped intervals for the two
- permutation test
- at least one of anova, ancova, logistic, multinomial

Conclusions and Summary

Summary of main results

Future implications and next steps

Model assumptions and limitations

- Independent observations

- Multicollinearity
- Normally-distributed residuals

Here, I present a descriptive scatter plot of 1860 slave population rate against black population rate. Two lines are presented, one line from model 1 (base model) and one line from model 2 (controlling for Black population rate). I use the ggplot2 package, one that gives much more flexibility on formatting than graphing functions in base R. Two outliers were removed here.