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 from the U.S. South, or all counties in states where slavery was legal in 1860. 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 our paper, and is defined as residents per 100,000 population in a given county who are currently in prison.
- **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 "urbanicty" 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 our paper, and is defined as the proportion of population that were slaves. Because many states in the North had already abolished slavery by 1860, this limits the analysis to counties in the U.S. South.
- **Whether the county has a prison and what type of prison. (`prison_type`)** Could there be some sort of cultural effect that makes prisons themselves increase incarceration rate in a county? To study this, we include a categorical variable indicating whether a county has a prison or not, and what type of prison it is.

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 fourty years, but we use only data from 2015 to avoid complications of time-series data. Slave population rate was taken from the U.S. Census in 1860, provided 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(SAIPE) Program.
- **Rent burden,** from the Eviction Lab at Princeton. Rent burden is defined as the proportion of income spent on rent.
- **Gini coefficient,** taken from the County Health Rankings report and originally collected by the U.S. Census Bureau. 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.

An important note for the graders

Code for this project is extensive because data was integrated from several sources and each one required cleaning. To cut down on length and communicate our results clearly, we suppress nearly all code.

For readers that do wish to look at my code in the Rmarkdown file, please note that we use the `tidyverse` syntax previewed in SDS230's last class. The `tidyverse` is a collection of pacakges including dplyr and ggplot2 that make life easier in R. 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. We also provide extensive comments in the code.

Very importantly for graders, we provide all of the data that we reference in a Github repository. The most important file is labeled "full_dt.csv" and is in the "data" folder. This has all of our cleaned data, assembled from different sources and ready to use. It should be downloaded by the code below, but in case it is not, feel free to download it manually here.

Finally, sometimes Rmarkdown moved our plots from the area of discussion to a few pages forward or backward to best fit the page.

Data cleaning

Slavery data

We read in a current counties boundary shapefile. 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.

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. These time periods had different identifiers for each county, and we could not find a "crosswalk" online that could link these two together quickly. To address this problem, we took "centroids" in each 1860 county 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 counties from these data together. Figure 1 explains this process with Georgia as an example. Code for Figure 1 is suppressed but available in the accompanying Rmarkdown file.

This is not an optimal solution, and the best way to approach this would have been to take block-level population data to weight overlaps between 1860 and 2015 shapefiles. We considered this too far outside of the scope of the class to be worthwhile to pursue. Figure 1 suggests our method works for the majority of counties.

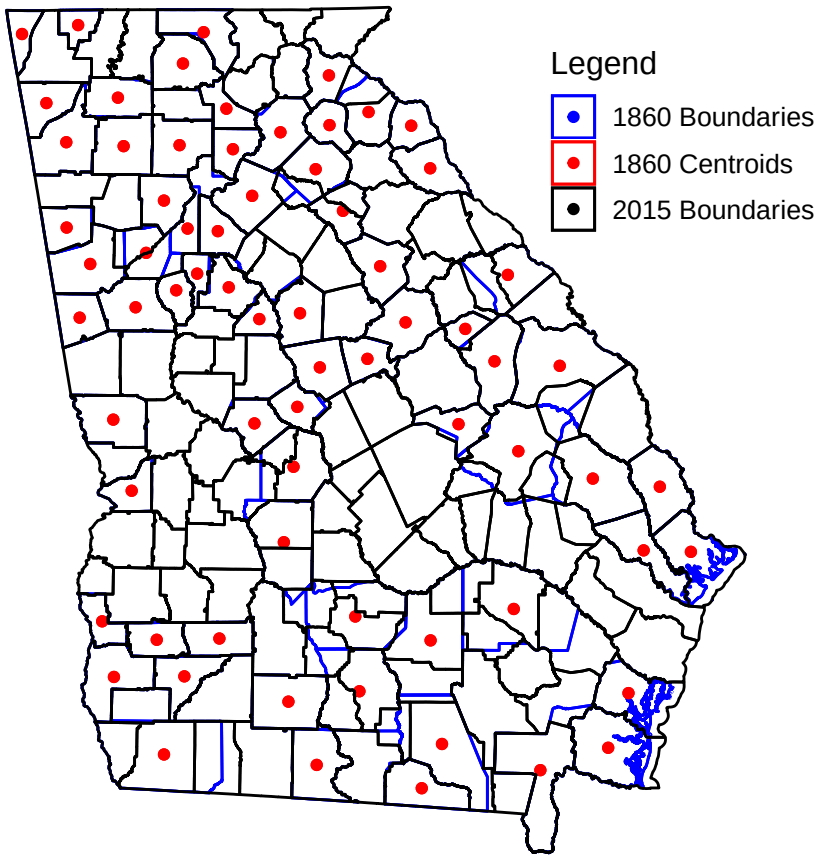


Figure 1: Diagram of County Overlaps and Changes for Georgia

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

To summarise, we read in the 1860 county shapefile with `sf::st_read()`, fit the shapefile with the correct projection (i.e. calibrated the longitude-latitute 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 we only filtered for 2015 data, created Black population rate, and selected columns of interest.

`prison_type`, or whether a county has a prison and what type of prison it has, was also fairly easy to obtain from the Prison Policy Initiative's Correctional Facility Locator 2010. We scraped this using the `rvest` package as we saw in class as we could not download it directly. They also did not have 5-digit county identifiers directly on the table, but were embedded in a set of links in each row. We used the `stringr` string manipulation package to pull these identifiers out. Finally, we filtered for only state and federal correctional facilities (otherwise known as "prisons") and aggregated the dataset to have only one row per county instead of one row per prison.

Data cleaning of control variables

The control variables were also straightforward to obtain. In general, we changed columns to numeric, removed missing values, joined state and county identifiers to one five-digit identifier, and joined them through the `*._join()` functions from the tidyverse. In one case, we had to write a loop to download a separate file for each state, then used the `bind_rows()` command to assemble them into one dataset. The EPA's data on air pollution (`pm` in my dataset for "particulate matter"), was provided at the tract level, which meant we had to create county-level estimates by combining the tract-level data.

Merging

The five-digit codes that uniquely identified counties allowed us to assign 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, 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?

[1]Note: It could definitely be argued that counties with high slave populations having high incarceratoin 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 dmeographic 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 plots

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. Both sets of boxplots do not reveal large differences in incarceration rates between counties with different prison types and urbanicity settings, but we cannot make any definitive conclusions without further analysis.

Matrix plot

To assess correlation between continuous, we present a matrix plot for all variables using the `pairsJDRS()` command. The labels were hard to read with nine variables, so I manually added text on the top of the

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

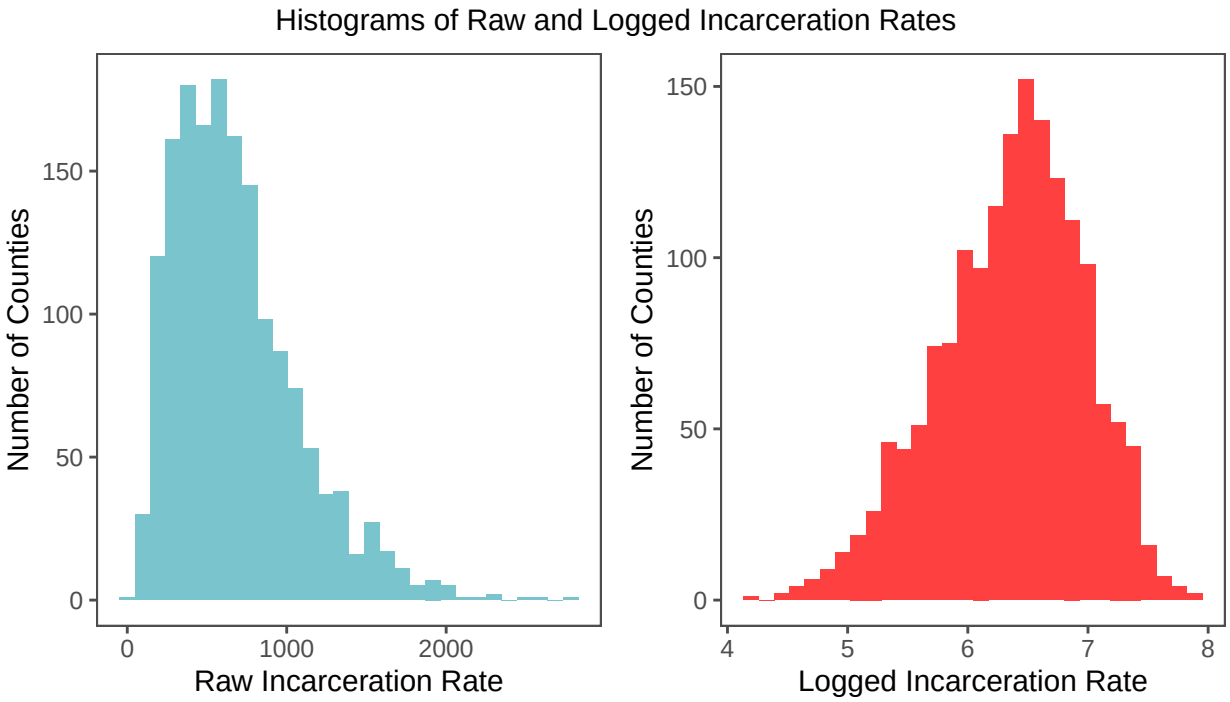


Figure 2: Histogram of Incarceration Rates

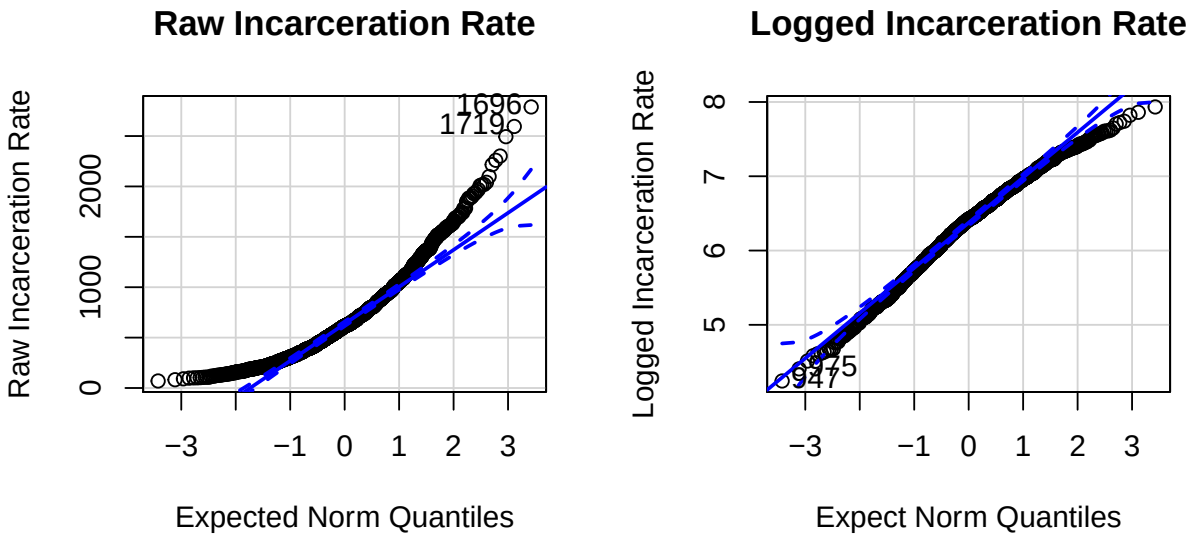


Figure 3: Normal Quantile Plots of Incarceration Rates

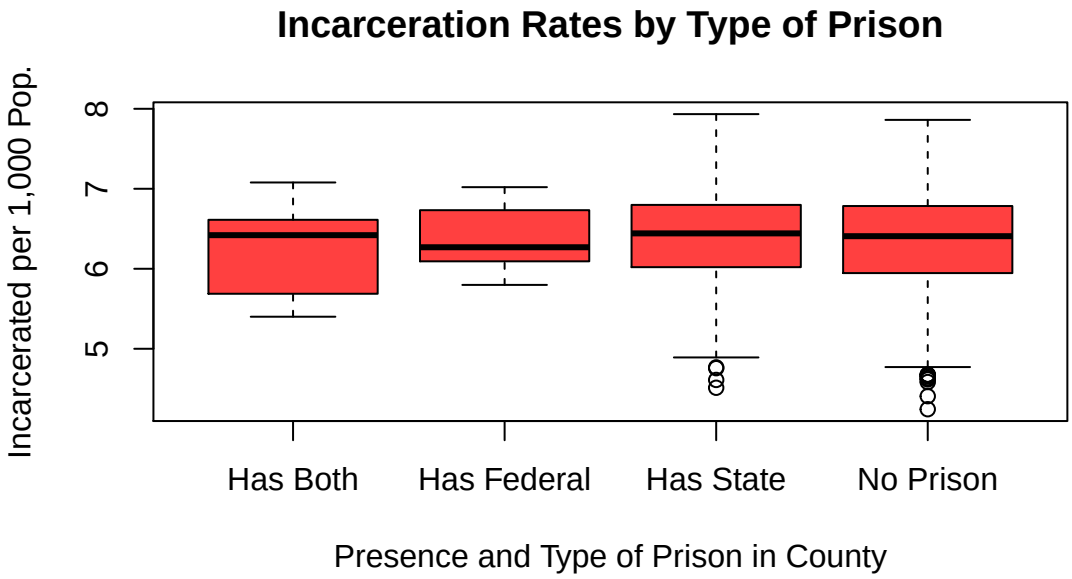


Figure 4: Boxplot by Type of Prison

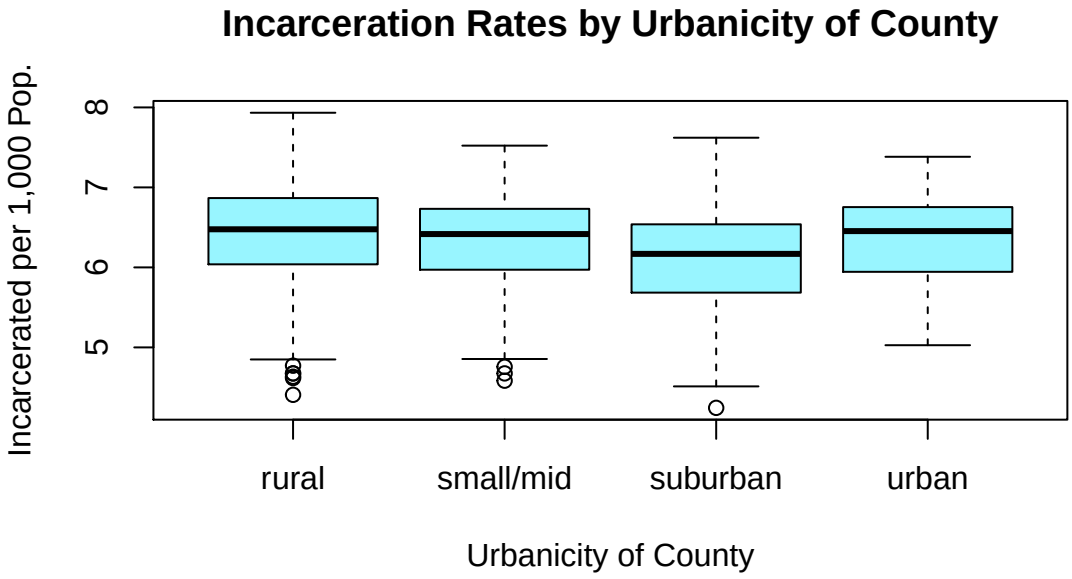


Figure 5: Boxplot by Urbanicity

graph. It is clear here that some degree of association is present between many of our predictors and the response variable, with some multicolinearity. We will address multicolinearity in the "Conclusions" section.

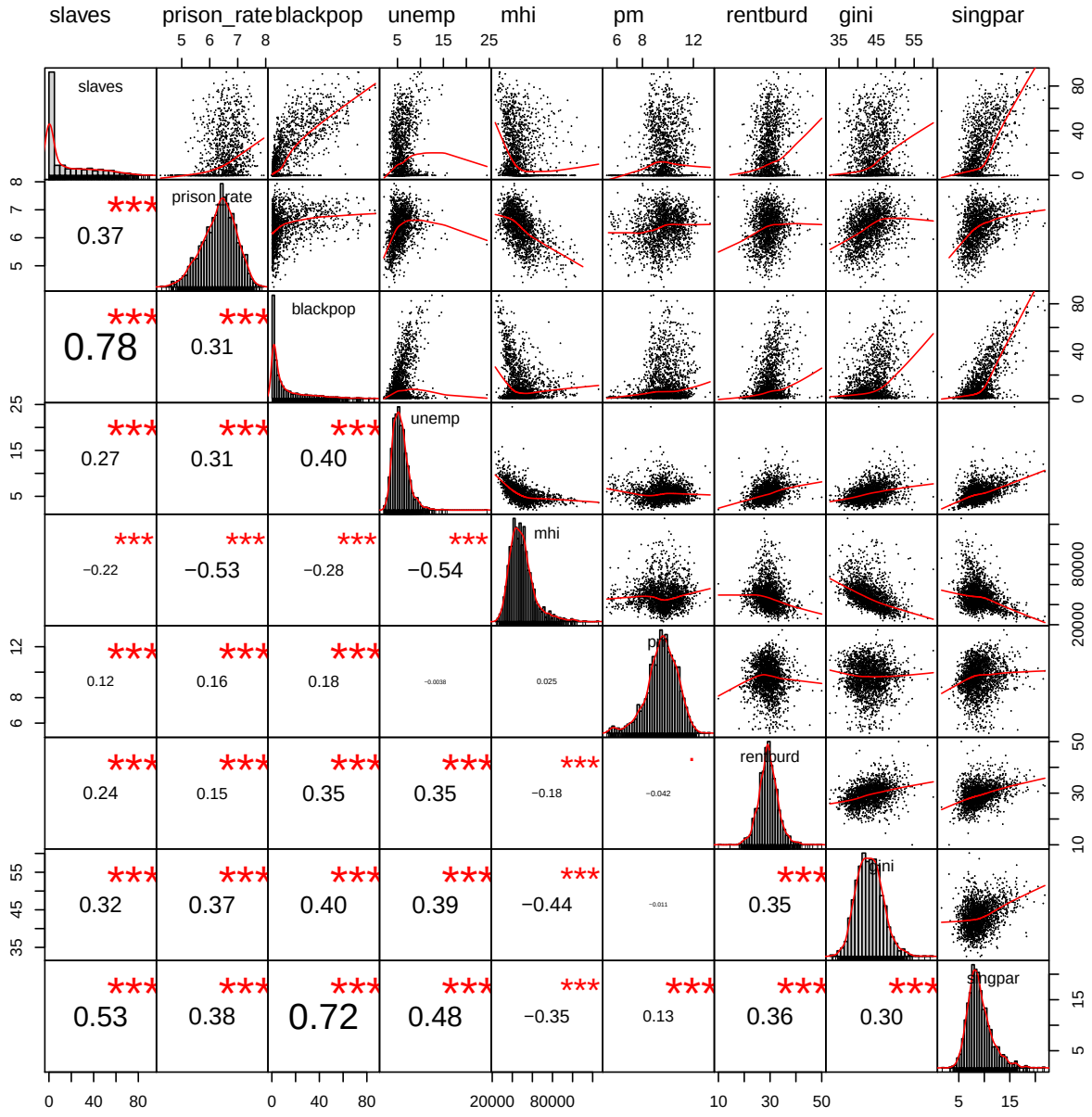


Figure 6: Matrix Plot of Continuous Variables

Correlation test with permuted p-values

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 ten million simulations returned a p-value of 0, again indicating that the true p-value is less than $1e-7$. 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 evidened with 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 are ambiguous 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.

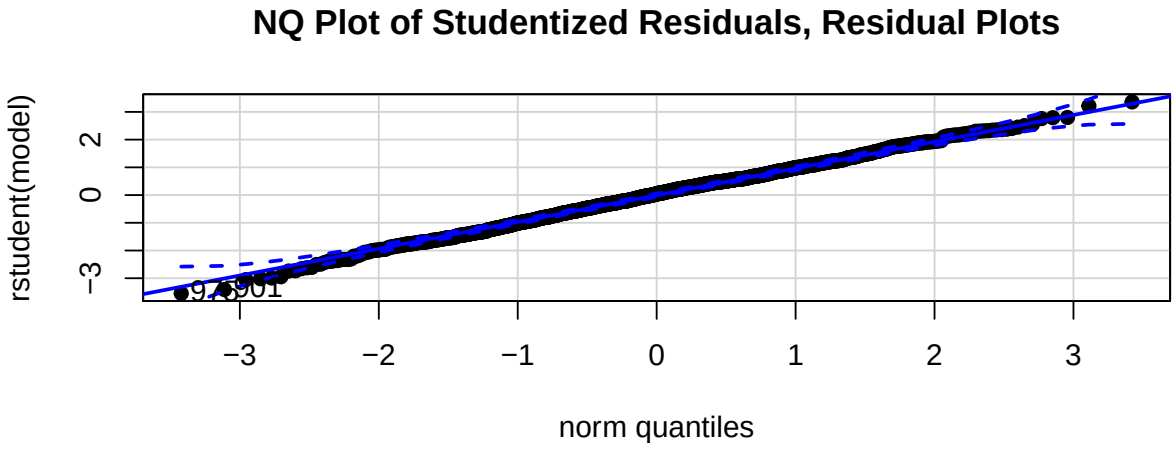
Best subsets regression

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 two models selected by BIC and adjusted R-squared after a `regsubsets()` regression. The response variable in all models is logged incarceration rate.

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 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 distributed[4]. Though there is some heteroskedasticity in the residuals, the extent is mild and within our personal standards of "acceptable heteroskedasticity."



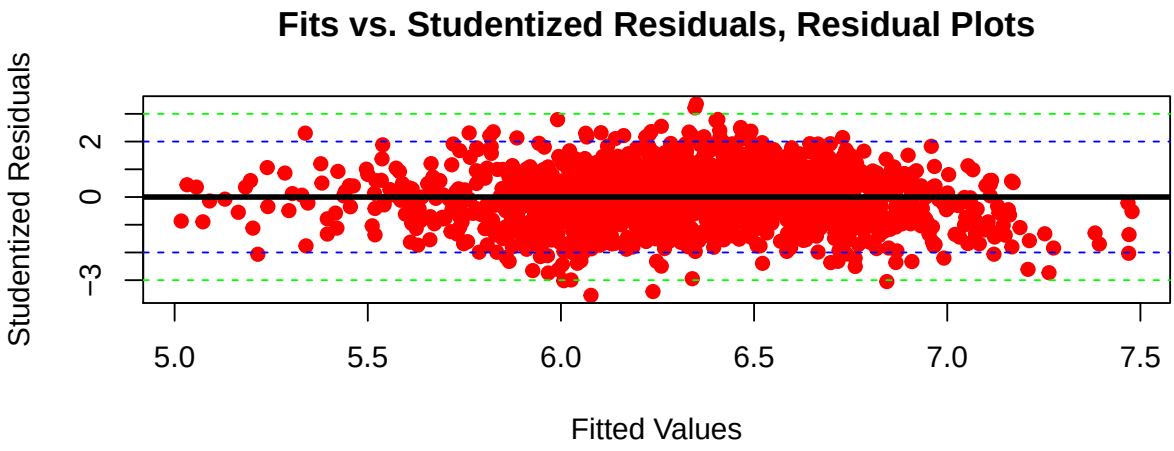
[3] $e^{7.212e-03} = 1.007239 = 0.7\%$ increase. See here for further explanation on logged dependent variables and their interpretation.

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

Table 2: Results of Best Subsets Regression

	Model chosen by:		
	Base Model	BIC	Adj. R2
slaves	0.011***	0.007***	0.007***
	(0.001)	(0.001)	(0.001)
	$p = 0.000$	$p = 0.000$	$p = 0.000$
urbanicity_rural			−0.150*
			(0.082)
			$p = 0.068$
‘urbanicity_small/mid‘			−0.144*
			(0.081)
			$p = 0.075$
urbanicity_suburban			−0.027
			(0.083)
			$p = 0.745$
urbanicity_urban			
blackpop		−0.010***	−0.010***
		(0.002)	(0.002)
		$p = 0.000$	$p = 0.000$
‘prison_type_Has Both‘			0.138
			(0.150)
			$p = 0.360$
‘prison_type_Has Federal‘			−0.083
			(0.124)
			$p = 0.503$
‘prison_type_Has State‘			0.050*
			(0.028)
			$p = 0.076$
‘prison_type_No Prison‘			
unemp			−0.021**
			(0.009)
			$p = 0.016$
mhi		−0.00002***	−0.00002***
		(0.00000)	(0.00000)
		$p = 0.000$	$p = 0.000$
pm		0.077***	0.068***
		(0.010)	(0.010)
		$p = 0.000$	$p = 0.000$
gini		0.023***	0.024***
		(0.004)	(0.004)
		$p = 0.000$	$p = 0.000$
singpar		0.057***	0.057***
		(0.007)	(0.008)
		$p = 0.000$	$p = 0.000$
Constant	6.193***	5.051***	5.503***
	(0.017)	(0.227)	(0.286)
	$p = 0.000$	$p = 0.000$	$p = 0.000$
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)	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



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 a plot of Tukey-corrected confidence intervals for an ANOVA analysis of all urbanicity types and their incarceration rates. 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.

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.

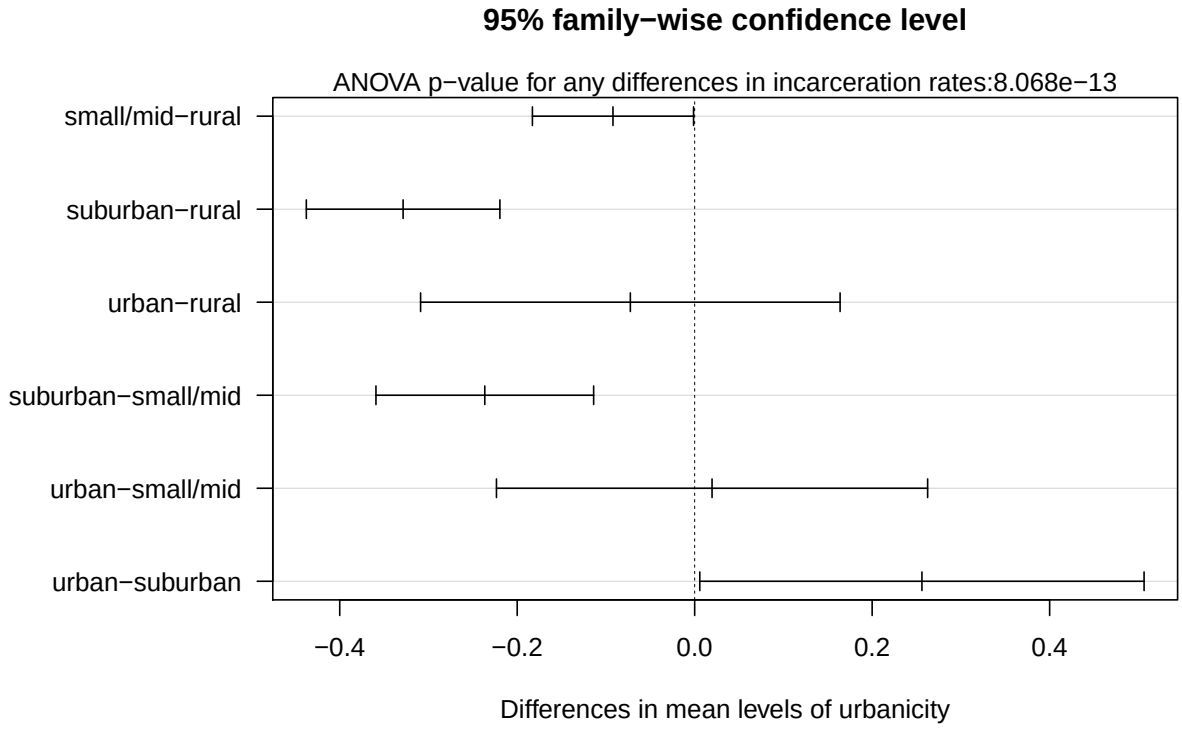


Figure 7: ANOVA of Incarceration Rate and Urbanicity, with Tukey Corrections for Confidence Intervals

Conclusions

Summary of main results

Our results indicate that slave population rate in 1860 positively predicts incarceration rate in 2015, even when accounting for potential mediating variables like Black population rate and median household income. This points towards a separate (although undeniably related) "memory" of slavery that operates through geography, whereby local townships' penal codes, police forces, and need for capitalist labor creates both slavery and mass incarceration. Wacquant's words in 2002 came before the Vera Institute released their dataset of prison incarceration, and his observation lends a disturbing insight much before numbers could.

We also find evidence to question the traditional narrative of "mass incarceration" as an "urban issue," given that rural counties do not have significantly different incarceration rates. It could perhaps be that two distinct mechanisms operate in urban environments and rural environments, given that suburban counties in the middle of the urban-rural spectrum do not have these same high incarceration rates; more work is needed to explore this idea.

Model assumptions and limitations

Our results have a number of assumptions and limitations.

- Independent observations
  - Traditionally, linear regression relies on an "independent observations assumption," or that any given member in a sample studied is not affected by another member of the sample. We acknowledge that this could very well be the case for slavery and mass incarceration, given that many

Southern states fought to continue slavery for other Southern states. A high slave population of one county could affect incarceration rates in an adjacent county as well, in a sort of "spillover effect." However, the methods to account for this (fixed-effects analysis and spatial lag models) are outside the scope of this class, and we end with these results not to offer definitive conclusions but encourage further study.

- Multicolinearity
  - Many of our predictors are colinear (i.e. counties with lower median household income may have higher proportions of single-parent households), and we were cautioned that this could distort our conclusions by inflating or deflating the standard errors. However, the fact that the trends of interest appear across all models, no matter how many variables we include or exclude, suggest that these trends exist even if we account for related effects caused by other variables. See this Piazza post for further inquiry.
- Normally-distributed residuals
  - As we saw above, the residuals for the BIC-selected linear model were perfectly normally distributed. The response variable itself is only "almost" normally distributed, which may cause problems with the ANOVA assumption; however, from class we know ANOVA is fairly robust to small violations of normality and unequal variances.

Nevertheless, our results are strong evidence that slavery echoes and has effects in the modern day. Despite being one hundred and fifty years after abolition, the age of "freedom" is yet to come.