Analyzing Correlation Between Socioeconomic Factors and Incarceration

Identifying Domains for Impactful Policy

Abstract

This study delves into the intricate relationship between socioeconomic factors and incarceration rates, explicitly focusing on community demographics including race, urban-rural categorizations, proportions of teenagers in each community, unemployment rates, poverty rates, population density, and education rates. A large body of extant literature observes relationships between racial subjugation, economic subjugation, and the employment of carceral mechanisms (see Cantekin & Elgin,2019). We sought to bring this home– exploring the relationship between such indicators on incarceration rates in California. We aim to use methods in multiple linear regression to locate California communities especially impacted by the carceral web and pinpoint potential policies to reduce incarceration rates within vulnerable communities.

We utilized linear regression models to quantify the extent to which these variables explain and predict incarceration rates within each of California's census tracts. Preliminary findings suggest that, while some socio-economic metrics play a significant role in determining incarceration rates, the simple racial composition of each tract emerges as the strongest predictor. In line with our assumptions, unemployment rate, poverty rate, and some education metrics proved significant. Against our assumptions, urban-rural categorization and population size were not significant predictors of incarceration. The results of this study suggest that, while incarceration rates are likely to be marginally reduced by addressing socio economic imbalances, critical analysis centered on racial subjugation itself likely needs to be the primary focus of any effective decarceration policy.

Introduction

In the wake of ongoing civil rights movements and increasing awareness of social injustices, there is a pressing need to scrutinize the underlying factors that contribute to disparities in incarceration rates. The United States grapples with a prison system that is often criticized for its disproportionate impact on marginalized communities. This study is motivated by a desire to dissect and understand the complex interplay between civil rights issues and incarceration, with an emphasis on the roles of race, urban-rural categorizations, proportions of teenagers in each community, unemployment rates, poverty rates, population density, and education rates.

Our investigation is rooted in the hypothesis that specific socioeconomic factors and race are intricately woven into the fabric of incarceration rates. By employing linear regression models, we aim to unveil the existing quantitative relationships, offering statistically robust and socially enlightening insights. The objective is not merely to highlight correlations but to provide a data-driven foundation upon which more intentional, equitable, and effective policies can be crafted.

Dataset

Sources and Collection

Data for this study was sourced from The Prison Policy Initiative, a reputable organization dedicated to producing research that exposes the broader harm of mass criminalization (Prison Policy Initiative, 2023). We created a script to scrape information from The Prison Policy Initiatives website. This gave us 2020 incarceration rates (per 100,000 per) for each of California's more than 8,000 census tracts.

To enrich our analysis and provide a more comprehensive view of the relationship between socioeconomic factors and incarceration rates, we combined this dataset with data from the United States Census Bureau. Specifically, we downloaded tract-level socioeconomic information from 2020 and joined it with the incarceration rate dataset on census tract. This approach provided incarceration rates linked with granular data on factors including racial proportions, the plurality racial group (categorical), urban-rural categorizations (categorical), the proportion of teenagers in each community, unemployment rates, poverty rates, population density, and education rates.

Data Integration and Cleaning

Upon obtaining the datasets, initial data cleaning was performed to extract numerical information (i.e. census tract numbers) from columns with textual data and handle any inconsistencies in data types between merged datasets. The datasets were then merged based on shared identifiers, ensuring that the combined dataset was consistent and suitable for analysis.

Limitations

- The data is primarily focused on California, which might not be representative of the entire United States.
- Potential discrepancies or biases in the original data collection methods could influence the results.
- The datasets might not capture all relevant factors that influence incarceration rates, and the study's findings should be interpreted in this context.

Model Diagnostics

Before fitting a regression model, we tested for any potential data structure problems (i.e. multicollinearity and influential points) or model assumption problems (i.e. heteroscedasticity, normality, and linearity).

Multicollinearity

We first checked for multicollinearity. When not addressed, unacceptable levels of multicollinearity can distort the interpretation of regression coefficients and inflate the standard errors of predictors.

To diagnose multicollinearity, we first calculated the Variance Inflation Factor (VIF) for each predictor– keeping in mind that a VIF value greater than 10 suggests high multicollinearity.

After initial analysis, we found high VIF values indicating unacceptable multicollinearity for plurality race (categorical), white proportion, Asian proportion "Other" racial identity proportion, and education levels.

To address this, we first removed the "education_total" predictor. We reasoned that other education parameters are subsets of the total education numbers, which may

explain the perfect correlations. We additionally needed to remove proportions of the population that identified racially as “Other.” This was an unfortunate necessity to retain appropriate levels of multicollinearity in the model, but in future analysis we would employ more sophisticated dimension-reduction methods to not exclude information from a population that deserves to be included in analysis. Applying these changes, VIFs for all numerical predictors dropped below 10.

To address the VIF scores related to our categorical predictor “plurality_race” , we decided to compute the GVIF for race. GVIF is an extension of VIF generalized to more accurately assess the multicollinearity of categorical variables. After computing GVIF for “plurality_type”, our results still indicated unacceptable rates of multicollinearity. We reasoned that this was likely because we were including both racial proportions and the racial category that the plurality of populations identify as in the same model. With this in mind, we decided to only include only numerical predictors analyzing racial demographics in the model. This resulted in a pool of possible predictors with acceptable levels of multicollinearity.

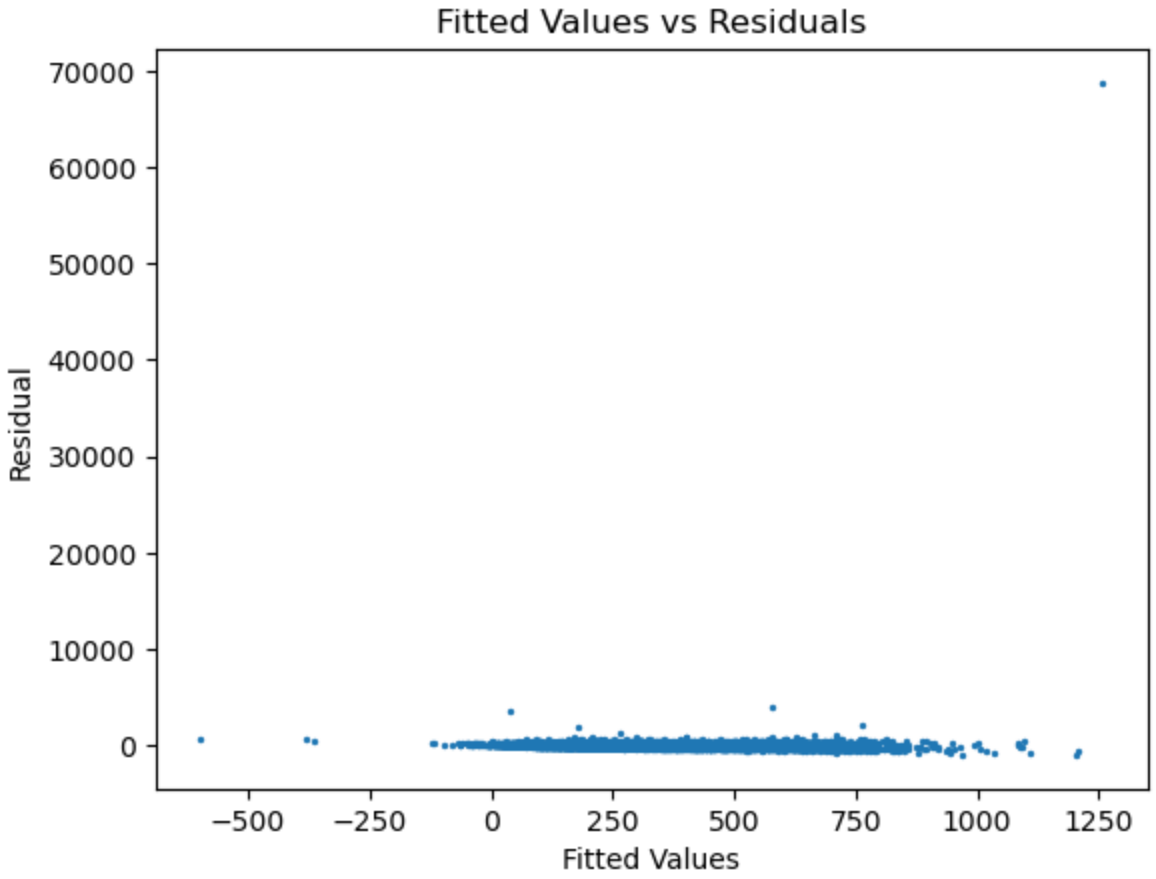
VIF/GVIF Factor	Features
6915.763476	Intercept
1.223349	C(Area_Type)[T.Urban]
684.216906	C(plurality_race)[T.Asian]
120.548228	C(plurality_race)[T.Black or African American]
1162.153641	C(plurality_race)[T.Some Other Race]
1448.874315	C(plurality_race)[T.White]
1.308042	unemployment_rate
6.042376	population
1.767698	percent_poverty
45.540319	white_proportion
7.068499	black_proportion
2.489335	native_american_proportion
25.802133	asian_proportion
1.182019	native_hawaiian_pac_islander_proportion
43.208835	other_race_proportion
inf	education_total
inf	education_less_than_highschool
inf	education_highschool
inf	education_some_college
inf	education_bachelors_or_higher
1.598651	total_population_16_to_19
40.42780856587475	GVIF for race

VIF Factor	Features
96.139142	Intercept
1.202601	C(Area_Type)[T.Urban]
1.298417	unemployment_rate
6.017656	population
1.703425	percent_poverty
4.596294	white_proportion
1.646150	black_proportion
1.435877	native_american_proportion
2.858378	asian_proportion
1.132574	native_hawaiian_pac_islander_proportion
3.547390	education_less_than_highschool
3.125224	education_highschool
2.773845	education_some_college
3.829052	education_bachelors_or_higher
1.593271	total_population_16_to_19

VIF and GVIF Factors before and after dropping correlated predictors

Influential Points

We also remained aware that influential points with high leverage could disproportionately affect the fitted model. To identify influential points, we calculated both external studentized residuals and Cook's distances– taking note of any points returning scores indicative of being influential. We narrowed in on five influential points that were identified using both external studentized residuals and Cook's distances: indices 374, 2233, 3613, 3615, and 3628.



Plot of fitted values versus residuals including outliers

Given the potential impact of these points on the regression results, we decided to analyze two models: one with these influential points and one without them.

Heteroscedasticity

Heteroscedasticity refers to the circumstance in which the variability of a variable is unequal across levels of another variable. This violates our assumption of constant variance and can bias eventual inferences.

To identify potential heteroscedasticity in our model, we ran a Breusch-Pagan (BP) test. Unfortunately, our initial model returned a BP Statistic of 70.73 and resulting p-value < .001. Even after trying to apply a log transformation and Box-Cox

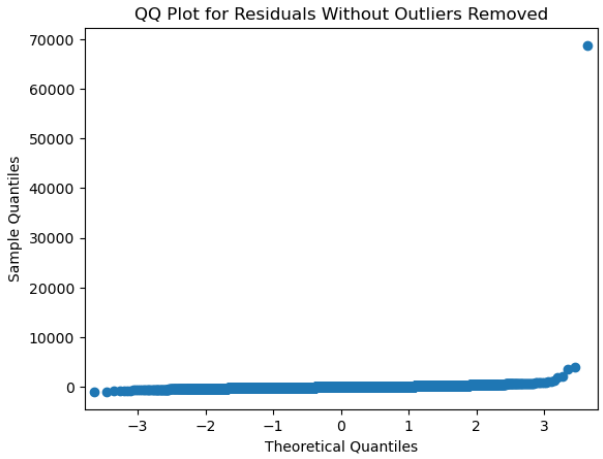
transformation on Y, the issue persisted. To address this, we decided to use robust standard error to analyze our regression models.

Normality

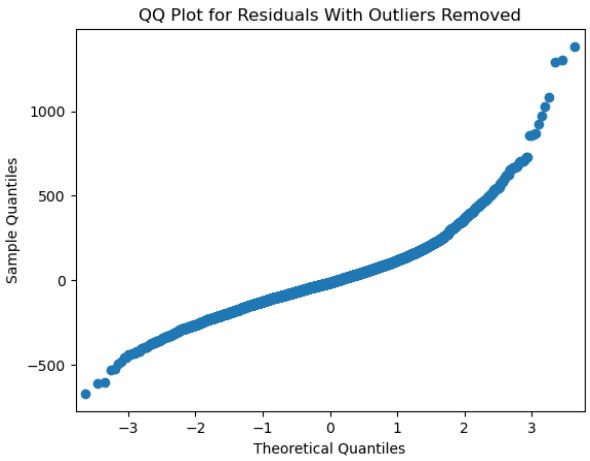
We additionally decided to check for violations of normality in the model. Acknowledging that only very extreme departures of the distribution of Y have the potential to yield spurious results, we decided it was better to be safe than sorry!

The residuals from both our regression models did not follow a normal distribution, as indicated by the Jarque-Bera (JB) test. The model with outliers *included* returned a JB score of 13349487520.526 (Prob(JB) = 0.00). The model with outliers *excluded* returned a JB score of 13232.190 (Prob(JB) = 0.00).

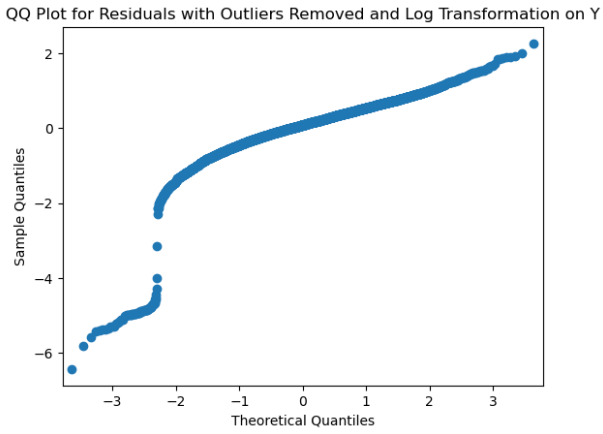
Our model without outliers removed produced a largely ineffectual QQ plot due to the y-axis needing to account for one especially large sample quantile:

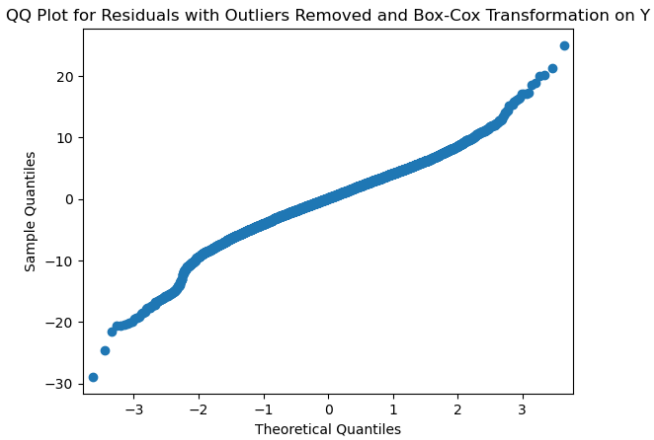


Removing previously-identified outliers, we observed a QQ plot that approaches normality, but shows problematic departures from normality at the tails:



We then compared QQ plots after applying either box-cox or natural log transformations on Y. QQ plots suggest that a box-cox transformation gets us closest to normality:

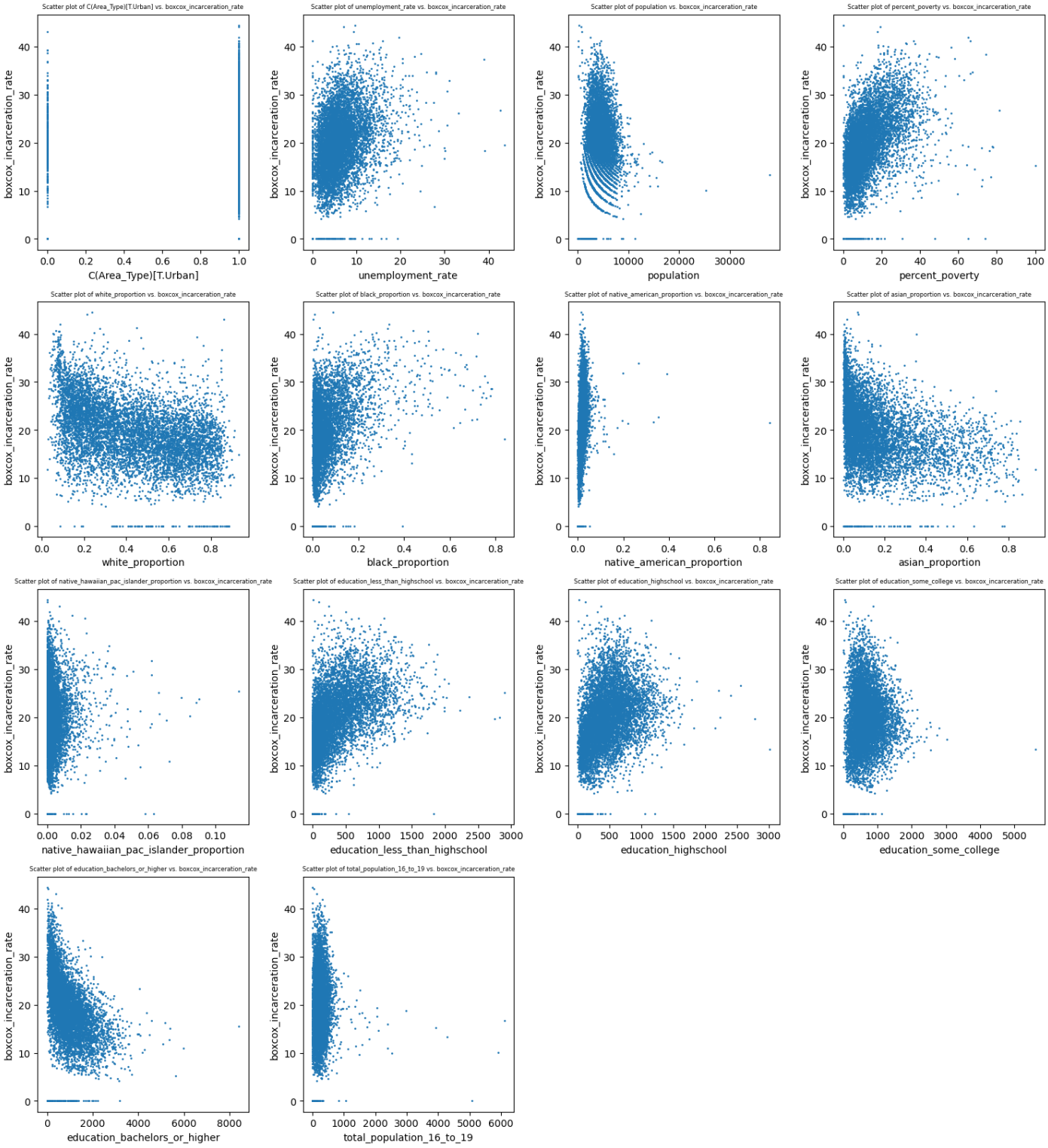




Although QQ plot analysis suggested that a box-cox transformation on incarceration rates gets data closest to a normal distribution, a Jarque-Bera test on our model– with outliers removed and after applying a box-cox transformation on incarceration rates– still indicated that the transformed data did not possess characteristics of a normal distribution (JB = 1620.84, Prob(JB) = 0.00). We moved forward with analysis utilizing the box-cox transformed model, but remained aware of this assumption violation.

Linearity

The relationship between predictors and the response variable appeared mostly linear, with some exceptions. For instance, there was a slight non-linear pattern with population. For this reason, we applied a log transformation on the population data.



Scatter plots demonstrating largely linear relationships between incarceration rates and predictors

Model Selection (Influential Points excluded)

Full model: `boxcox_incarceration_rate ~ unemployment_rate + C(Area_Type) + log_population + percent_poverty + white_proportion + black_proportion + native_american_proportion + asian_proportion + native_hawaiian_pac_islander_proportion + education_less_than_highschool + education_highschool + education_some_college + education_bachelors_or_higher + total_population_16_to_19`

Method 1 - t-test for individual predictors

Based on the p-value of t-test from the MLR summary, the predictors log_population, Area_type, education_less_than_high_school and education_some_college should be excluded from the model because their p-values for t-test results are less than 0.05.

Method 2 - Best Subset Selection

With the results from performing Best Subset Selection, given that including at least 4-5 predictors achieves the similar performance, the results of Mallows Cp should be incorporated into the evaluation. To have the lowest Mallows Cp and highest Adj-R^2, only white_proportion should be excluded from the selection.

Method 3 - Stepwise Selection

Forward stepwise selection : Area_Type, education_some_college , education_less_than_highschool are excluded. AIC = 22087.66. The selected model is similar to the model selected by t-test statistics.

Model Selection (Influential Points included)

Method 1 - t-test for individual predictors

Based on the p-value of t-test from the MLR summary, the predictors Area_type, education_less_than_high_school should be excluded from the model because their p-values for t-test results are less than 0.05.

Method 2 - Best Subset Selection

With the results from performing Best Subset Selection, given that all models including at least 4-5 predictors achieves the similar performance, the results of Mallows Cp should be incorporated into the evaluation. To have the lowest Mallows Cp and highest Adj-R^2, native_american_proportion, other_race_proportion, and education_less_than_high_school should be excluded from the model. In addition, compared to the model where outliers are excluded, the Adjusted R^2 decreases from 50 to 47 in best subset selection.

Method 3 - Stepwise Selection

Using the metric AIC, education_less_than_highschool and Area_Type should be excluded from the model. However, AIC decreases in this model where outliers are included.

Conclusion: Combining results from different model selection methods, we should remove area_type and education_less_than_high_school from the model.

Regression Results

Outliers excluded:

This is the model selected by Best Subset and outliers excluded, in this model white_proportion is removed, and it has a lower Adj-R^2.

Final model: *boxcox_incarceration_rate ~ unemployment_rate + C(Area_Type) + log_population + percent_poverty + black_proportion + native_american_proportion + asian_proportion + native_hawaiian_pac_islander_proportion + education_less_than_highschool + education_highschool + education_some_college + education_bachelors_or_higher + total_population_16_to_19*

OLS Regression Results

Dep. Variable:	boxcox_incarceration_rate	R-squared:	0.478
Model:	OLS	Adj. R-squared:	0.478
Method:	Least Squares	F-statistic:	515.4
Date:	Thu, 12 Oct 2023	Prob (F-statistic):	0.00
Time:	13:22:27	Log-Likelihood:	-21617.
No. Observations:	7317	AIC:	4.326e+04
Df Residuals:	7303	BIC:	4.336e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	14.1457	1.851	7.640	0.000	10.516	17.775
C(Area_Type)[T.Urban]	1.2934	0.276	4.678	0.000	0.751	1.835
unemployment_rate	0.0766	0.016	4.932	0.000	0.046	0.107
log_population	0.2812	0.252	1.118	0.264	-0.212	0.774
percent_poverty	0.0982	0.007	13.652	0.000	0.084	0.112
black_proportion	19.2926	0.709	27.207	0.000	17.903	20.683
native_american_proportion	32.0990	3.509	9.146	0.000	25.219	38.979
asian_proportion	-4.6220	0.382	-12.094	0.000	-5.371	-3.873
native_hawaiian_pac_islander_proportion	-43.4142	9.576	-4.534	0.000	-62.186	-24.643
education_less_than_highschool	0.0033	0.000	13.420	0.000	0.003	0.004
education_highschool	0.0021	0.000	6.952	0.000	0.002	0.003
education_some_college	-0.0003	0.000	-1.310	0.190	-0.001	0.000
education_bachelors_or_higher	-0.0021	0.000	-15.168	0.000	-0.002	-0.002
total_population_16_to_19	-0.0021	0.000	-7.109	0.000	-0.003	-0.002

Omnibus:	636.402	Durbin-Watson:	1.347
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2027.691
Skew:	-0.439	Prob(JB):	0.00
Kurtosis:	5.425	Cond. No.	2.43e+05

Outliers included:

Given most models suggest removing Area_type and education_less_than_high_school and outliers should not be easily removed, with the metric AIC, we are going to include outliers where AIC is smaller using stepwise selection.

Final model: *boxcox_incarceration_rate ~ unemployment_rate + log_population + percent_poverty + black_proportion + native_american_proportion + asian_proportion + native_hawaiian_pac_islander_proportion + education_highschool + education_some_college + education_bachelors_or_higher + total_population_16_to_19*

OLS Regression Results

Dep. Variable:	boxcox_incarceration_rate	R-squared:	0.471
Model:	OLS	Adj. R-squared:	0.470
Method:	Least Squares	F-statistic:	542.4
Date:	Thu, 12 Oct 2023	Prob (F-statistic):	0.00
Time:	13:29:57	Log-Likelihood:	-17792.
No. Observations:	7322	AIC:	3.561e+04
Df Residuals:	7309	BIC:	3.570e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	23.2250	1.018	22.812	0.000	21.229	25.221
unemployment_rate	0.0463	0.009	5.058	0.000	0.028	0.064
log_population	-0.8969	0.130	-6.891	0.000	-1.152	-0.642
percent_poverty	0.0542	0.004	13.024	0.000	0.046	0.062
white_proportion	-6.4641	0.246	-26.270	0.000	-6.947	-5.982
black_proportion	4.4880	0.472	9.499	0.000	3.562	5.414
native_american_proportion	4.6881	2.121	2.210	0.027	0.530	8.846
asian_proportion	-7.6376	0.300	-25.501	0.000	-8.225	-7.050
native_hawaiian_pac_islander_proportion	-31.1632	5.651	-5.514	0.000	-42.241	-20.085
education_highschool	0.0012	0.000	7.001	0.000	0.001	0.002
education_some_college	0.0007	0.000	5.140	0.000	0.000	0.001
education_bachelors_or_higher	-0.0005	7.97e-05	-5.952	0.000	-0.001	-0.000
total_population_16_to_19	-0.0010	0.000	-5.991	0.000	-0.001	-0.001

Omnibus:	2086.622	Durbin-Watson:	1.373
Prob(Omnibus):	0.000	Jarque-Bera (JB):	206108.769
Skew:	0.293	Prob(JB):	0.00
Kurtosis:	28.985	Cond. No.	2.37e+05

Key Findings

Final models with both outliers excluded and outliers included can be observed above. We decided to base final conclusion returned by the stepwise selection method:

1. Racial Proportions: The proportions of different races in the population were significant predictors of incarceration rates. Specifically, tracts with a higher proportion of Black and Native American residents returned positive association with incarceration rates. For example, for every one percent increase in the population being Black, incarcerations rates increased by 4.49. Moreover, beta values associated with racial proportions were notably higher than beta values observed for mere socioeconomic indicators.

2. Unemployment Rate: Unemployment rates were a significant predictor of incarceration rates. For every one percent rise in unemployment rate, we can expect incarceration rates to increase .05.

3. Poverty Rate: As expected, increases in poverty rate correspond with increases in unemployment rate. For every one percent increase in a community's poverty rate, we can expect incarceration rates to rise .05.

4. Population: Population demonstrated an inverse relationship with incarceration rates.

5. Education: Higher levels of educational attainment prior to attainment of a bachelor's degree was associated with higher incarceration rates. Higher levels of attainment of a bachelor's degree (or higher-level degree) was associated with lower incarceration rates.

6. Proportion of Teenages in Community: Contrary to our expectations, we observed a slight inverse relationship between the proportion of teenagers in a community and incarceration rates.

7. Adj-R^2: The final adjusted R^2 of our model is .47. This means that, as is, the model describes about 47% of the variability in California's incarceration rates.

Conclusion

Our regression analysis provides a comprehensive understanding of the factors influencing California incarceration rates. While some predictors, like racial proportions, had expected relationships with incarceration rates, others, like the proportion of teenagers in the community, did not.

It is important, however, to note assumption violations we ran into throughout the course of analysis. Outliers were detected, so results were presented with and without outliers. Heteroscedasticity remained a problem, so final interpretations utilized robust standard error. There was an additional violation of normality. While a Box-Cox transformation on Y helped data approach normality, we still observed notable skew and kurtosis after applying the transformation, but moved forward with analysis.

In general, these findings underscore the importance of addressing both socio-economic imbalances and race to reduce incarceration rates.

References

Cantekin, K., & Elgin, C. (2019). Incarceration and Labor Market Conditions of the Underclass in the United States: An Empirical Investigation. *European Journal on Criminal Policy and Research, 26*, 529 - 546.

Prison Policy Initiative. (2023). Prison Policy Initiative. Retrieved from https://www.prisonpolicy.org/