Lead in the Water: The Effects of Blood Lead Levels on Incareration Rates

A Socio-Spatial Analysis

staRstistions - Will Lieber, Wania Iftikhar Khan, AJ Tenser, Kami Akala

2025-04-30

Introduction

Could exposure to lead increase one's likelihood of going to prison? After our team's systematic review of studies, we have explored the potential effects of lead exposure on brain development in children and adults. Various existing studies highlight the detrimental effects of lead on different brain regions, noticeable in a decrease in executive control and cognitive control, thereby affecting memory, mood, behavior and comprehension skills. Such exposure to lead during the developmental years of children causes irreversible damage, the effects of which can be seen later on in life.

Studies in the past, such as one conducted by Talayero et al. (2023)[1], have highlighted a strong association between lead exposure during childhood and criminal tendencies during adulthood. One can be exposed to lead through different means, including water, which is the medium which we've chosen to investigate. Our research topic inquires about whether a relationship exists between a specified area's water lead levels and its incarceration rates, while also considering potential confounding effects of other demographic factors.

This research topic has important societal implications, namely the complicated intersection of crime, environmental racism, and more. It's an ever relevant question today and we hope to come to meaningful conclusions by the end of our analysis. Our initial hypothesis is that there is a positive relationship between water lead levels and the rate of incarceration with the existence of other interaction effects from things such as race and income. However, we acknowledge the intricate combination of social and institutional factors that increases one's likelihood of incarceration and are understand the possibility for inconclusive findings with the focus on lead exposure.

1. The association between lead exposure and crime: A systematic review

Our Data

We've chosen to create our data frame from a variety of census data relating to California in 2020. Our data looks at different California census tracts and their respective statistics relating to blood lead levels, income, incarceration rates, and racial demographics. For our analysis, we are particularly focused on `perc_bll_indicator`, `med_income`, our age and race variables and how well they can predict `imprisonment_rt`. The demographic data was collected through the Census data collection process which involves online surveys, in-person questionnaires, and is self-reported. The blood lead level data was compiled using the mandatory reported data entered by laboratories and healthcare workers. The data was sorted into census tracts based on reported street address and observations deemed false positive or false negative by the California Department of Public Health (CDPH) have been omitted. To clean the data, we joined each dataset to each other based on census tract and renamed columns for clarity. In addition, we calculated new columns to simplify categories (i.e. age ranges, ethnicity) and omitted fields which would not be relevant for this investigation.

Note: Figures referenced throughout the report can be found in the Appendix section.

Univariate Data Exploration

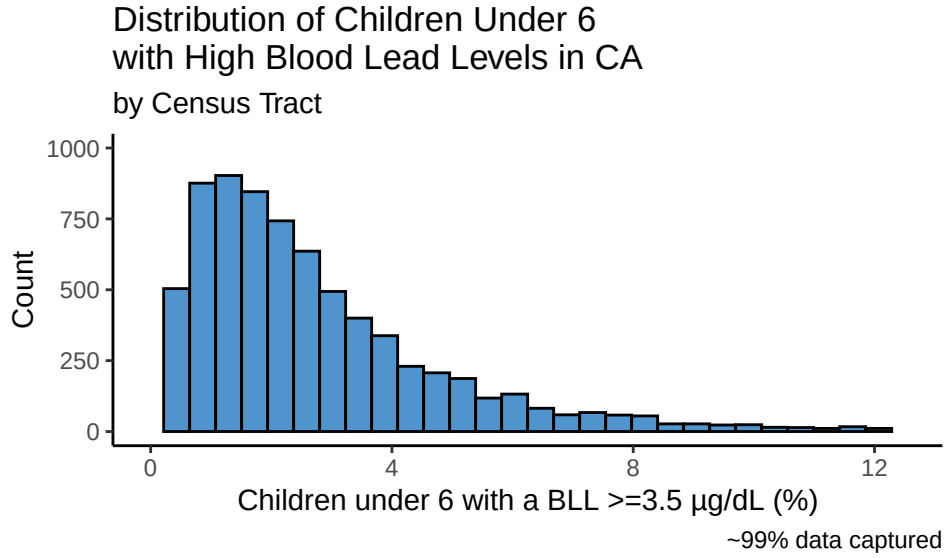

Figure 1: Distribution of lead in children

Blood Lead Levels (Figure 1): There are 94 census tracts that didn't test blood lead levels in this data. These will likely need to be removed because these observations are not useful in the analysis. Additionally, there is a large concentration of census tracts that have a percent

blood level indicator (bll) of 0 (1899 observations). While these should probably be included in the final analysis, removing these for data visualization produces a unimodal distribution that is heavily right skewed. Overall, the median blood level is 1.6891892 and the IQR is 2.599239.

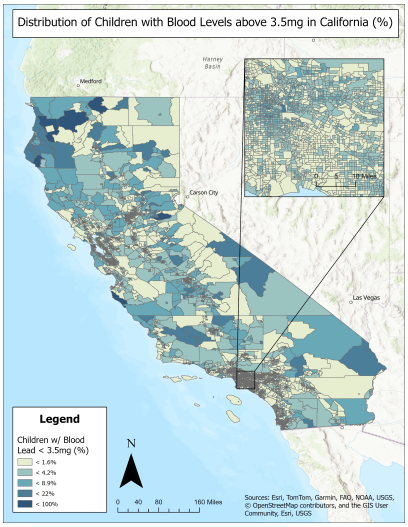

Figure 2: Spatial Distribution of BLL in California

Spatial Distribution (Figure 2): Given the spatial nature of this investigation, it is important to understand how the reported lead levels vary across the state. This map shows that there are contiguous series of tracts, especially in the Los Angeles region, where there may be spatial correlation. This is not surprising, as we expect water conditions across tracts to be similar.

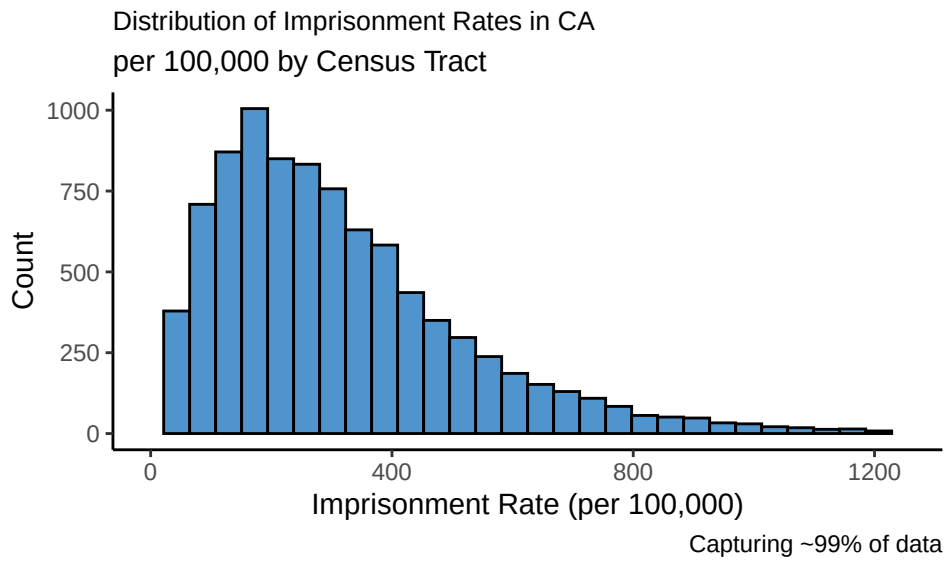

Figure 3: Distribution of Incarceration Rates in CA Census Tracts

Incarceration Rate (Figure 3): After removing extreme outliers (the top 1% incarceration rates - some may ultimately be removed because the census tract has an extremely low population ex. ~3 people), the shape of the distribution is unimodal and right skewed. The median incarceration rate is 267 out of 100,000 with an IQR of 257. This doesn't appear surprising - there are fewer census tracts with particularly high imprisonment rates.

Income (Figure 7): The shape of the income distribution is also unimodal with a less extreme right skew and a median value of 7.7225×10^4 and iqr of 5.044×10^4. This doesn't appear surprising - we'd expect median incomes of tracts to be concentrated towards the left.

Bivariate Data Exploration

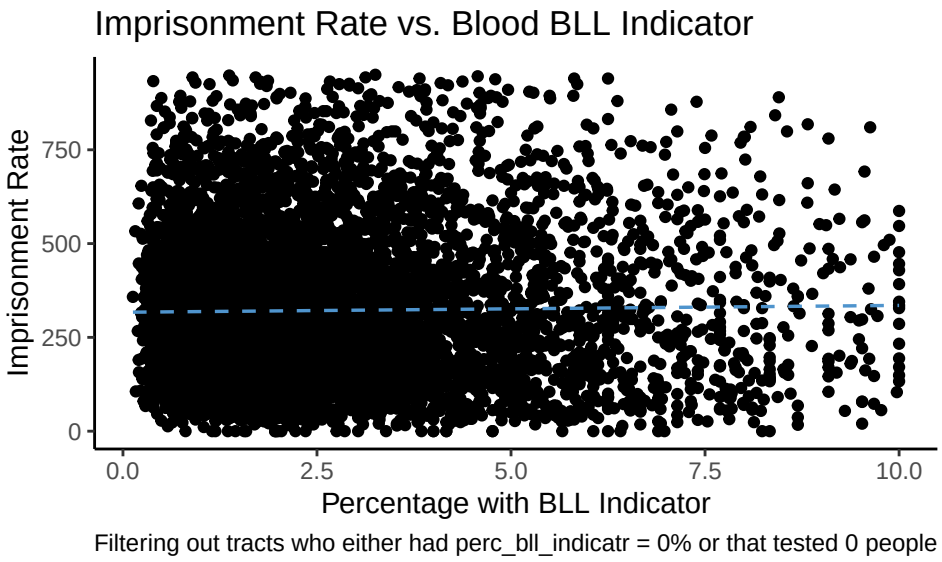

Figure 4: Imprisonment vs BLL Indicator

Imprisonment vs BLL Indicator (Figure 4): Imprisonment rate and the percentage of a census tract with a high bll has a weak positive correlation - at first glance, the bll indicator doesn't seem to have significant explanatory power for the variability in a census tract's imprisonment rate.

Imprisonment vs Median Income (Figure 8): Imprisonment rate and median income has a generally negative correlation. This isn't surprising as it is generally understood that a lack of resources, including financial, can be a driving force towards engaging in illegal activities.

Imprisonment vs Population Aged 15-29 (Figure 9): Imprisonment rate and percentage of population aged 15-29 in the census tract has a generally positive correlation. This isn't surprising as it's generally understood that one's likelihood to engage in illegal activities is higher as an adolescent.

Potential Interactions (Figure 10): There appears to be an interaction effect between race and income, as the relationship between median income and imprisonment rate differs by race. We created a variable, POC_other, indicating the percentage of a census tract population that is non-White. In this figure we categorized tracts with POC_other values that were above and below the median in the data of this variable to create a categorical split. The relationship between imprisonment rate and income appears more negatively correlated when the percentage of non-white people in the census tract is above the median.

Methodology and Results

For our analysis process, we iterated through different models until we were satisfied with its conditions and performance. As an initial model, we did simple linear regression to predict a census tract's imprisonment rate using bll indicator variable, racial makeup, median income, median age, the percentage of males, and the interaction between income and race. The model output and conditions are below:

Checking Initial Modelling Conditions

term	estimate	std.error	statistic	p.value
(Intercept)	-357.488	162.353	-2.202	0.028
perc_bll_indicator	-1.802	2.848	-0.633	0.527
POC_other	6.274	0.989	6.345	0.000
med_income	-0.001	0.000	-4.522	0.000
median_age	-3.145	1.379	-2.281	0.023
perc_male	17.270	2.859	6.041	0.000
POC_other:med_income	0.000	0.000	-3.974	0.000

Initial model output

Using the residual vs fitted plot **(Figure 11),** we can assess the model conditions. From this graph, we find

1. Normality condition can be relaxed.
2. Linearity condition not satisfied. Points not randomly scattered around the x = 0 line.
3. Constant variance condition not satisfied. Points fan out across the x = 0 line.
4. Independence condition also may not be satisfied. Census tracts next to each other could be more likely to have similar incarceration rates or blood lead levels.

Based on these initial conditions, we decided it would make sense to log transform some of our variables, specifically our median income and imprisonment rate variables.

Splitting the Data

We were also concerned about spatial correlation among census tracts especially since there can be multiple within a single county meaning they would share a lot of features. So we originally split the data with 20% in training and 80% testing to minimize this effect. We received abysmal R squared values when doing this. Using ChatGPT for suggestions, it recommended spatial cross validation, which we researched further to ensure it made sense for our analysis[2]. It appeared to be a feasible method to address model over-fitting and independence issues within our data. Because of that, we were then able to do a 80% training, 20% testing split.

2. Geocomputation with R

Figure 5: Spatial Correlation in CA

I DON'T KNOW WHAT THIS MAP SHOWS BUT WE SHOULD HAVE SOME DESCRIPTION TO GO WITH THIS MAP!

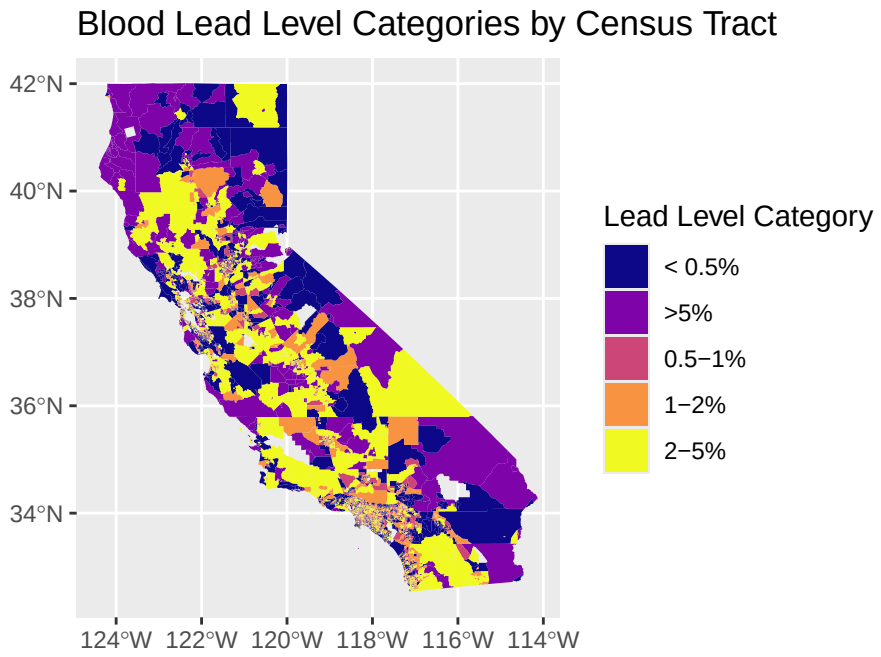

Figure 6: Spatial lead level category distribution

Final Model Selection

With our transformed variables, we got a better residuals plot **(Figure 12)** with only a single outlier. From there, we arrived at our final model using step-wise selection on our 80/20 split data to determine what predictor variables were most significant. Through this method, we found that indicators of bll in a census tract isn't statistically significant in predicting their imprisonment rate. However, median income, age, an the interaction between income and the racial makeup of a census tract were the most statistically significant in this model.

term	estimate	std.error	statistic	p.value
(Intercept)	11.971	0.330	36.315	0.000
log_med_income	-0.663	0.026	-25.624	0.000
perc_bll_indicator	-0.003	0.003	-1.045	0.296
POC_other	-0.027	0.009	-2.929	0.003
median_age	0.005	0.001	4.476	0.000
perc_male	0.008	0.003	2.915	0.004
log_med_income:POC_other	0.004	0.001	4.522	0.000

Table 3: Model Performance (RMSE and R-squared)

.metric	.estimator	.estimate
rmse	standard	184.971
rsq	standard	0.495

Discussion

Our ultimate research goal was to determine how well blood lead levels can explain rates of imprisonment, with a focus on California census tracts. We hypothesized that there would be a significant relationship based on existing research linking lead exposure to increased criminal tendencies. We also wanted to factor in other various social factors that are well understood to have effects on one's likelihood to be imprisoned, such as racial and financial demographic data. Based on our final chosen model, it appears that the blood lead levels of a census tract aren't statistically significant in predicting imprisonment rate while, income, age, and the interaction between income and race came out as our most significant predictors. We understand that the circumstances surrounding once likelihood of being imprisoned are complex and based on our analysis it appears that lead exposure isn't a particularly useful metric. General discourse on this topic seems to reveal that differing results in studies may be due to the overstated effect of lead exposure on incarceration rate or at the very least, it's not a useful metric on its own.

Interesting Insights

We used spatial cross validation to create training and testing sets that account for the fact that census tracts next to each other in our dataset likely have similar values. Inclusion of some census tracts in the training data that are right next to census tracts in the testing data could therefore bias model performance. So a model trained on spatially cross validated data evaluated using data that is geographically distant from the training set should in theory be more robust. We would expect the R squared values to be lower and the RMSE values to be higher for the spatially cross validated model. While the RMSE values were higher for the spatially cross validated model, in line with our prediction, interestingly, this was not the case for the R squared values–compared to a model that was training on randomly split data (see appendix: 80/20 train test- roughly the same proportion of the total census tracts as the Spatial CV train test), the spatial CV model has a higher R squared. This could potentially be caused by the random split model over-fitting to regions that generalize poorly to the new regions. Even though the RMSE value is worse, the spatial CV model likely provides a more realistic prediction on new regions given the way it was trained, making it, in our opinion, the superior model.

Limitations and Future Work

In terms of potential limitations or future changes of our analysis methods, it would be interesting to incorporate even more geographical regions as there are many states that currently

have significant levels of lead exposure. Perhaps there are broader patterns we're not seeing with a focus on California alone. Alternatively, it would be interesting to see the same analysis except on California counties instead of tracts to easily relate county-specific traits which might be more stable to our conclusions. A potential limitation of our current analysis is the way that we measure blood lead levels of a county. We used the test levels of children to relate to imprisonment rates of individuals that are likely older, making this indicator a bit unreliable in terms of measuring the exposure rate for individuals currently imprisoned.

Appendix

Univariate EDA

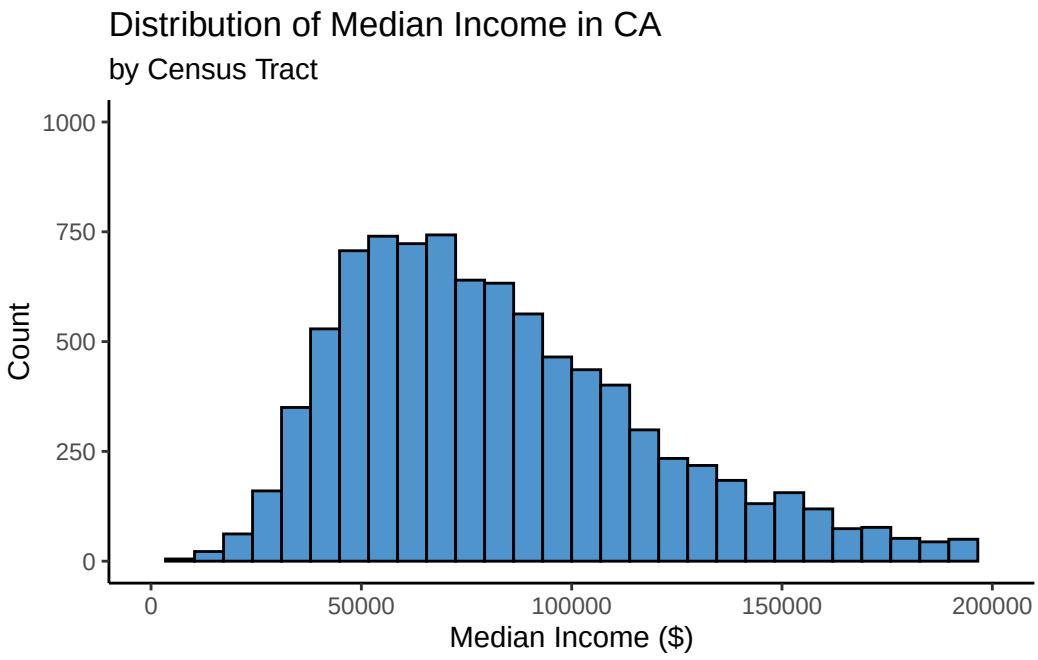

Figure 7: Income distribution in CA Census Tracts

Bivariate EDA

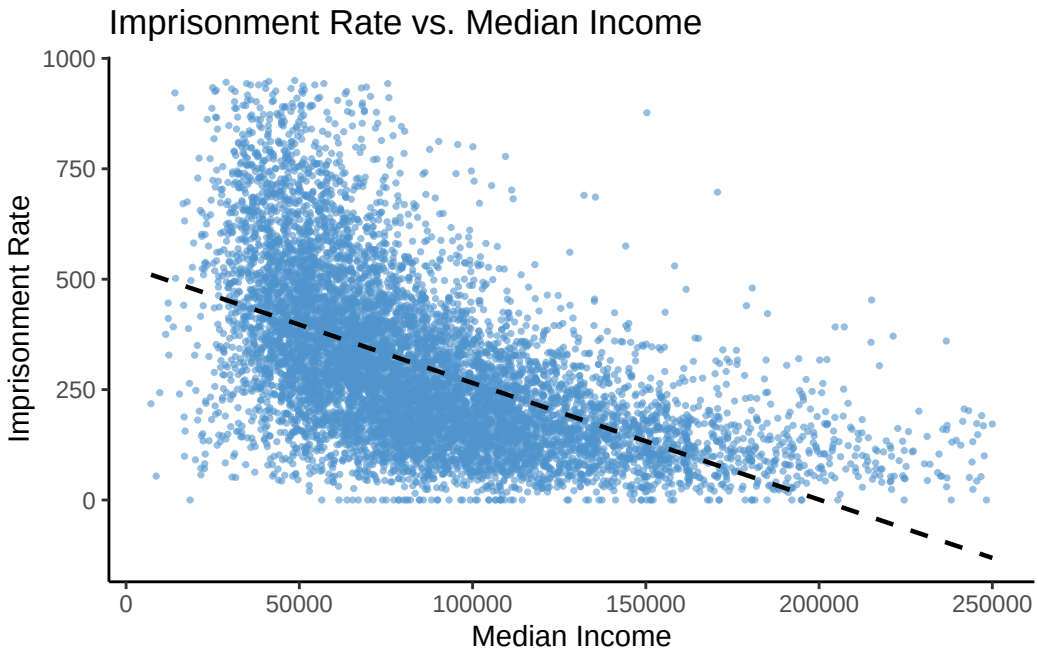

Figure 8: Imprisonment vs Median Income

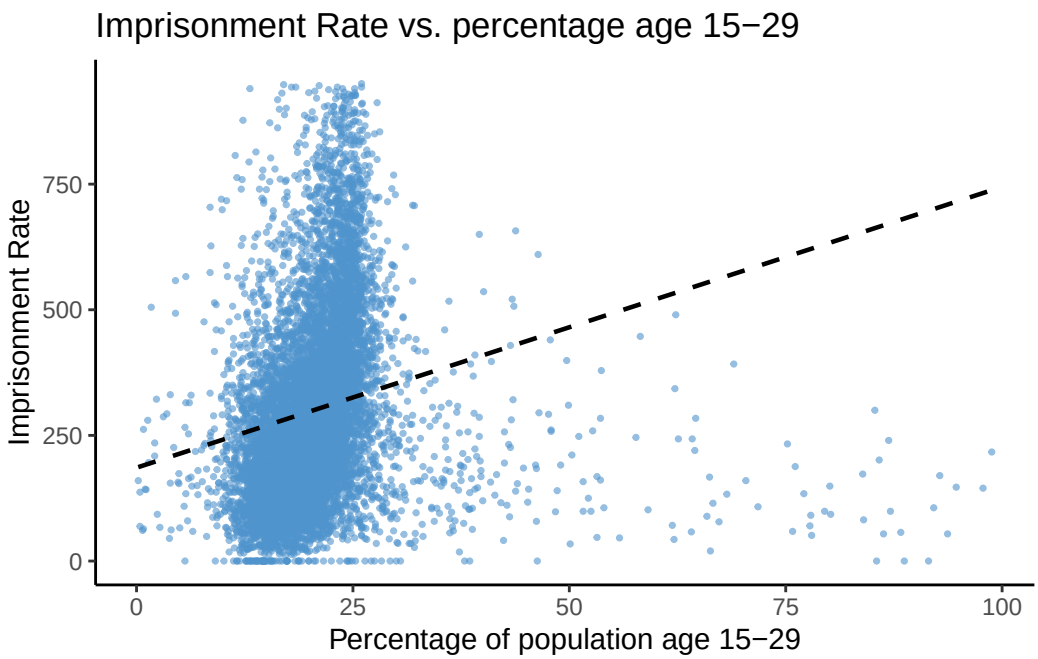

Figure 9: Imprisonment Rate vs Adolescent Population

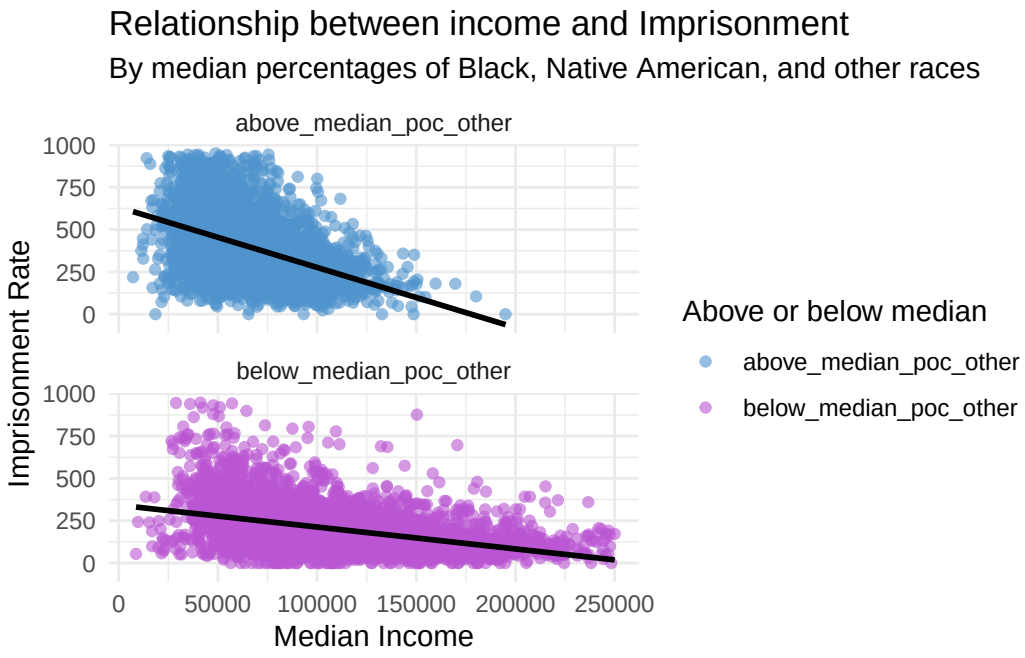

Figure 10: Relationship between race, income, and imprisonment

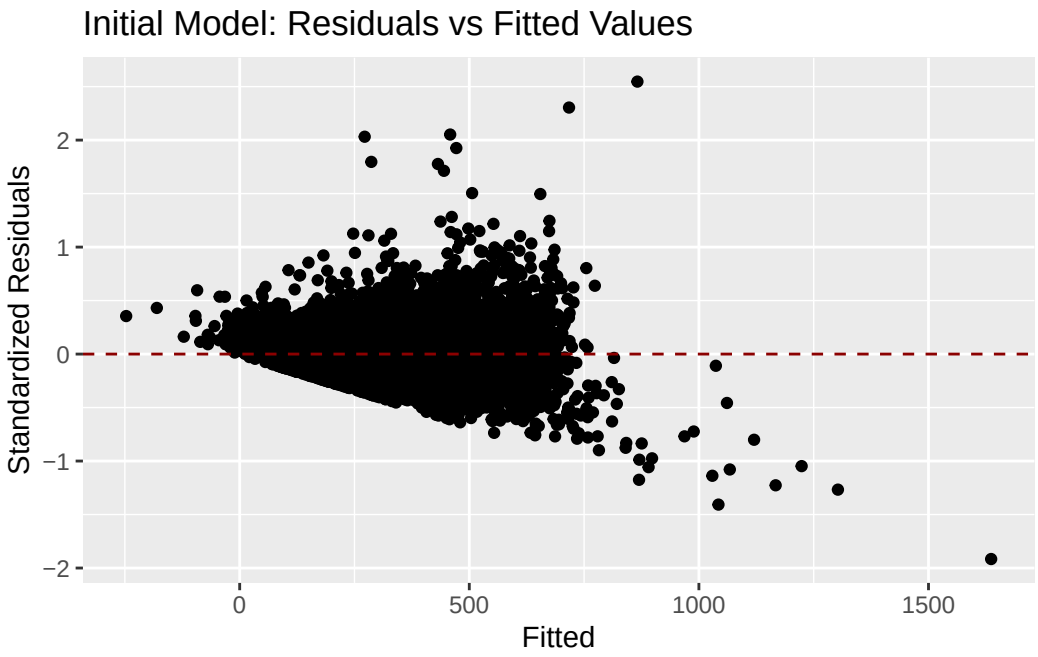

Figure 11: Residual plot of initial modeling output

```
perc_bll_indicator            POC_other           med_income
          1.047454             6.401602             2.310705
        median_age            perc_male POC_other:med_income
          1.631162             1.030406             3.966140
```

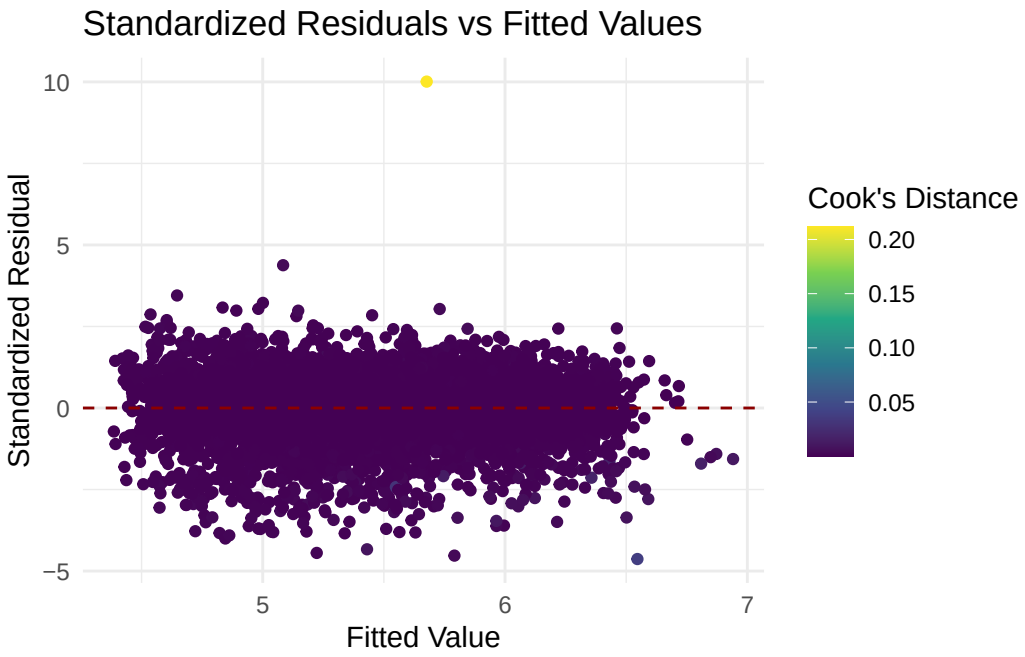

Figure 12: Residual plot of final modeling output

term	estimate	std.error	statistic	p.value
(Intercept)	11.971	0.330	36.315	0.000
log_med_income	-0.663	0.026	-25.624	0.000
perc_bll_indicator	-0.003	0.003	-1.045	0.296
POC_other	-0.027	0.009	-2.929	0.003
median_age	0.005	0.001	4.476	0.000
perc_male	0.008	0.003	2.915	0.004
log_med_income:POC_other	0.004	0.001	4.522	0.000

Residual plot of final modeling output