

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

An Analysis

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

2025-04-10

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 understand the possibility for inconclusive findings with the focus on lead exposure.

1. [The association between lead exposure and crime: A systematic review](#)

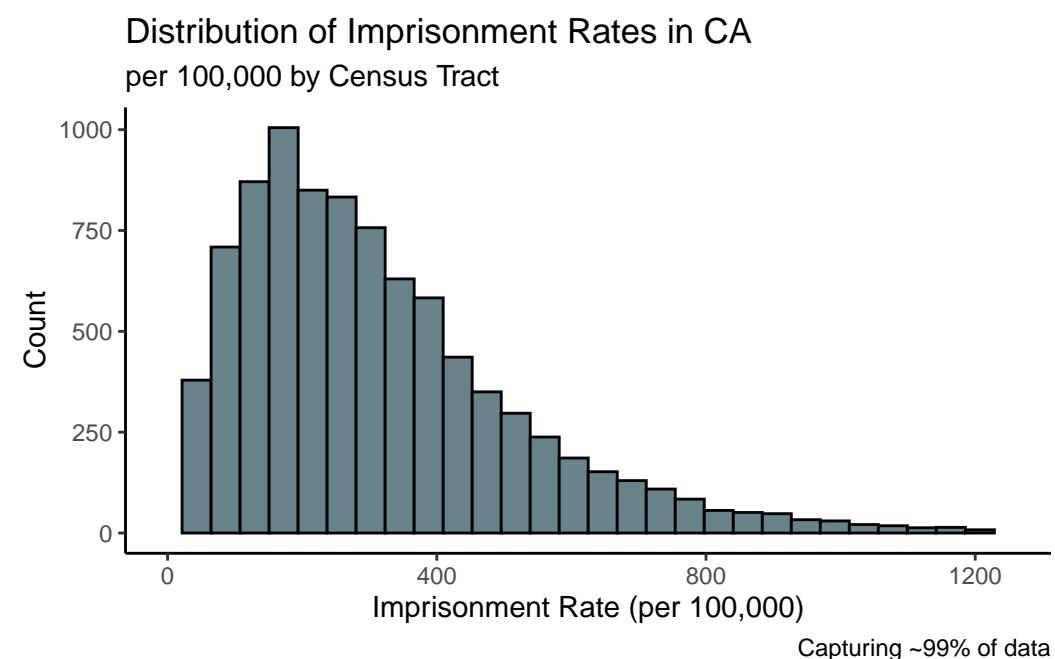
Exploratory Data Analysis

Our Data

We've chosen to create our dataframe 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`.

Data Exploration

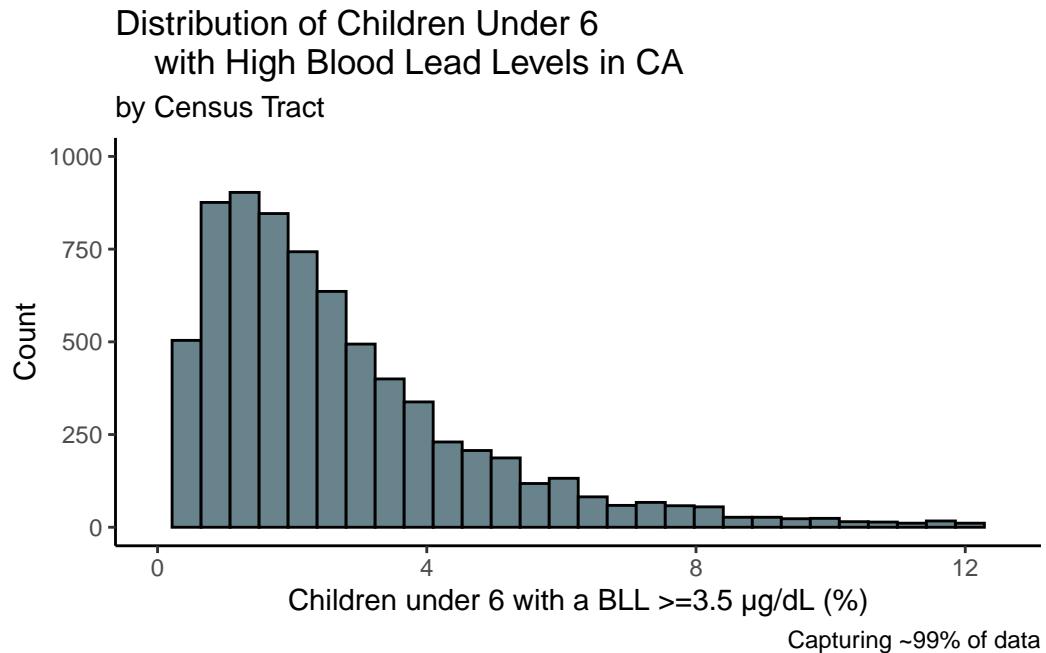
Incarceration Rate



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.

Blood Lead Levels

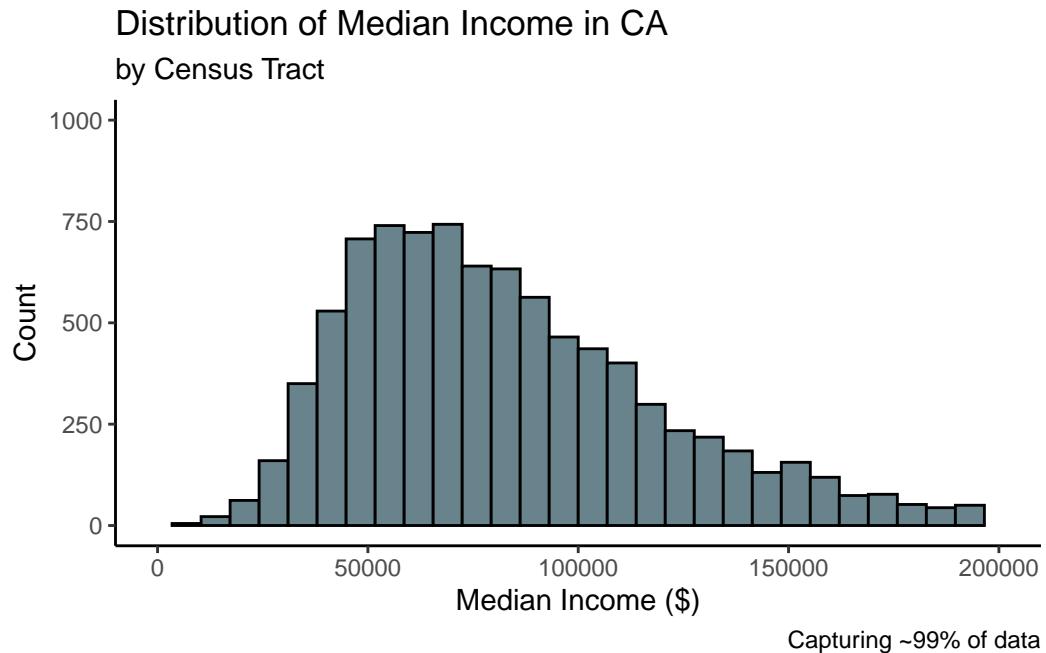
99%
12.57605



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 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 NA.

Income

99%
208825.1



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 `r_inc_iqr`. This doesn't appear surprising - we'd expect median incomes of tracts to be concentrated towards the left.

Spatial Distribution

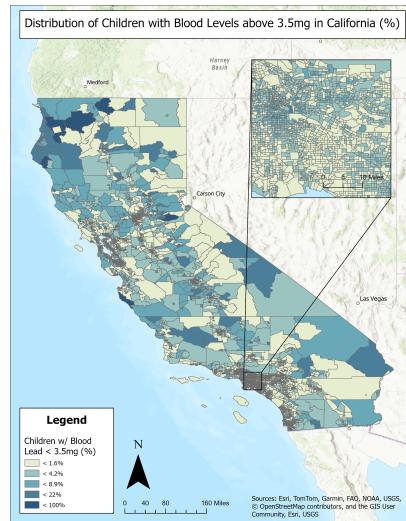
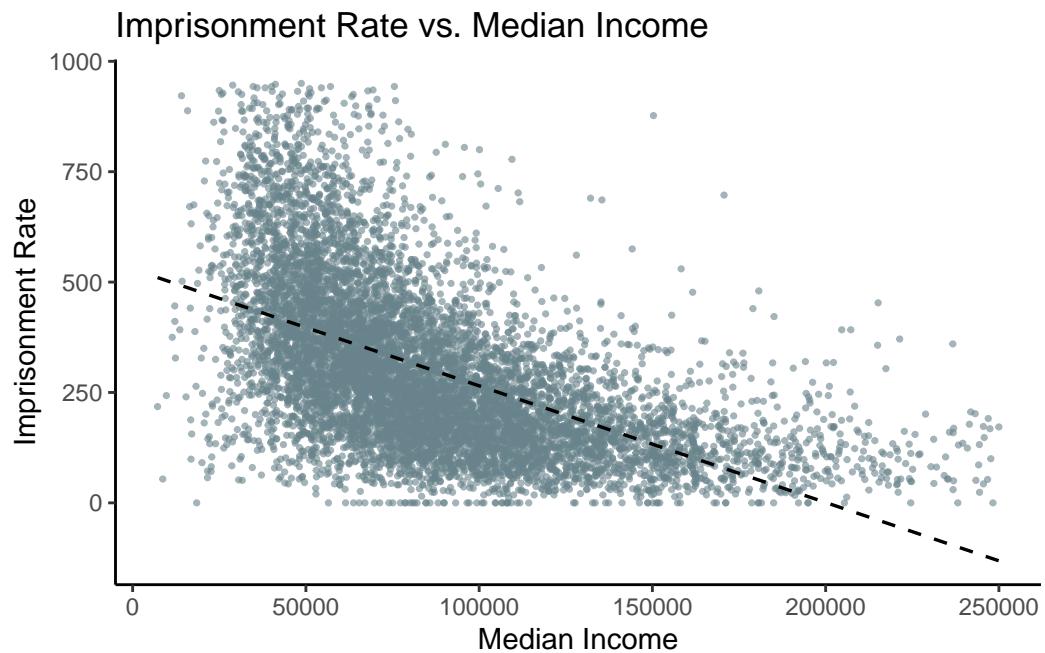
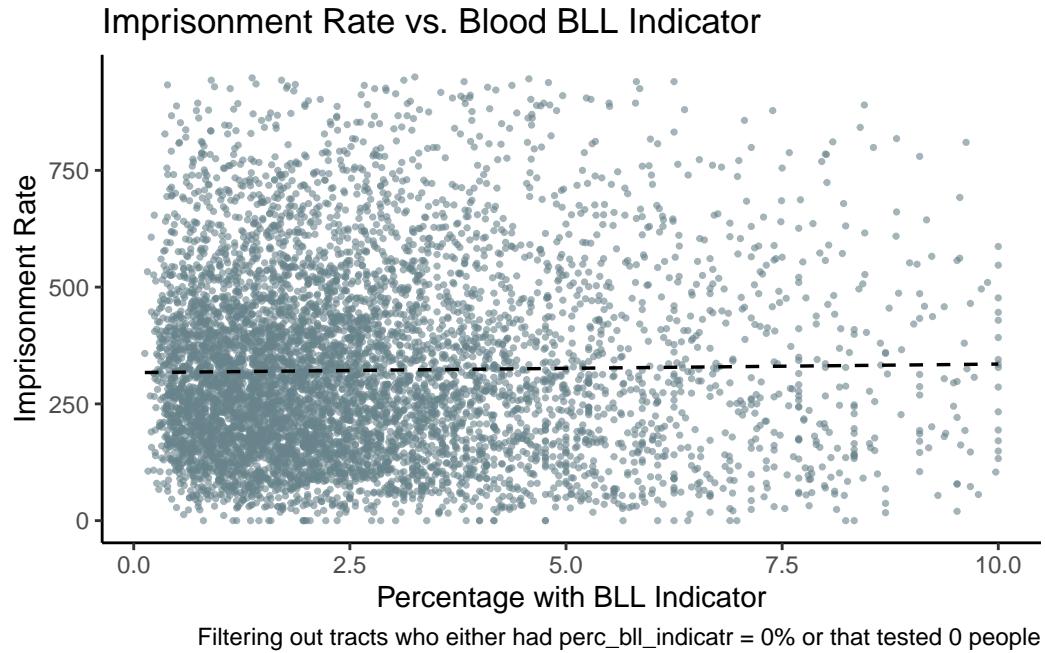
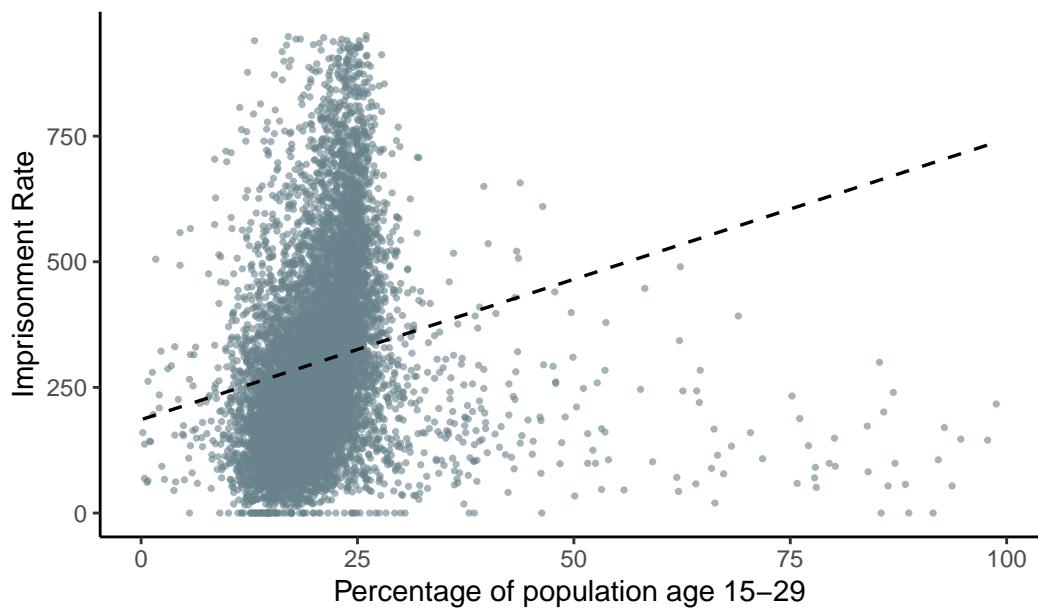


Figure 1: Spatial Distribution of BLL in California

Especially since census tracts can be extremely small, we'd have to account for spatial dependence in our analysis.



Imprisonment Rate vs. percentage age 15–29

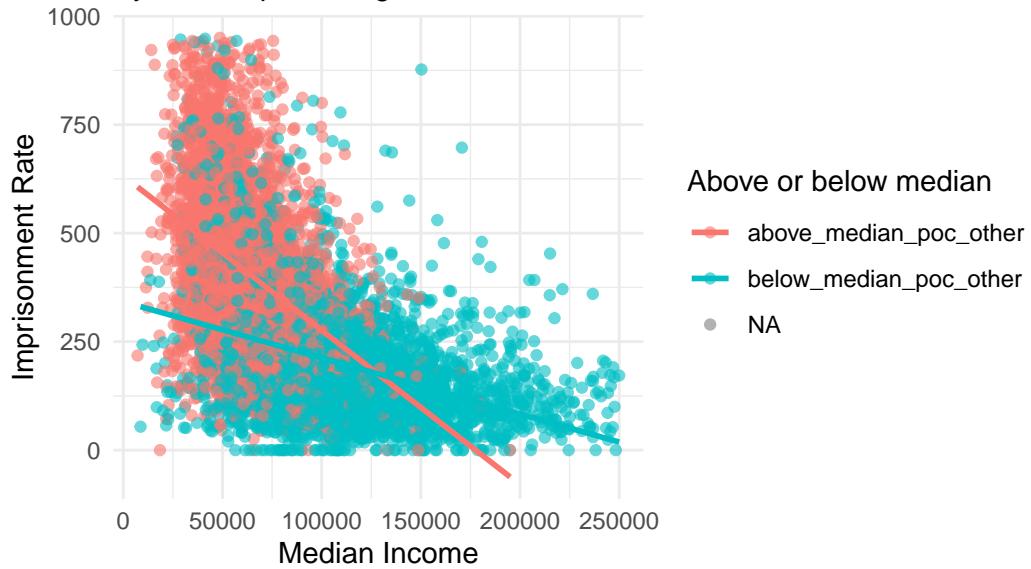


Imprisonment rate and the percentage of a census tract with a high blll has a generally positive correlation. Imprisonment rate and median income has a generally negative correlation. Imprisonment rate and percentage of population aged 15-29 in the census tract has a generally positive correlation.

Potential Interactions

Relationship between income and imprisonment

By median percentages of Black, Native American, and other races



When comparing the relationship between median income and imprisonment rate, it appears that generally they have a negative correlation. This graph suggests there could be an interaction effect between race and income, as the relationship between median income and imprisonment rate differs by race. We created a categorical variable for the percentage of the census tract population that is black, native american, or “other race” that is categorically above or below the median in the data. The relationship between imprisonment rate and median income appears more negatively correlated when categorically above the median census tract population percentage of black, native american, and other_race. This supports that there could be an interaction effect between race and income.

Splitting the Data

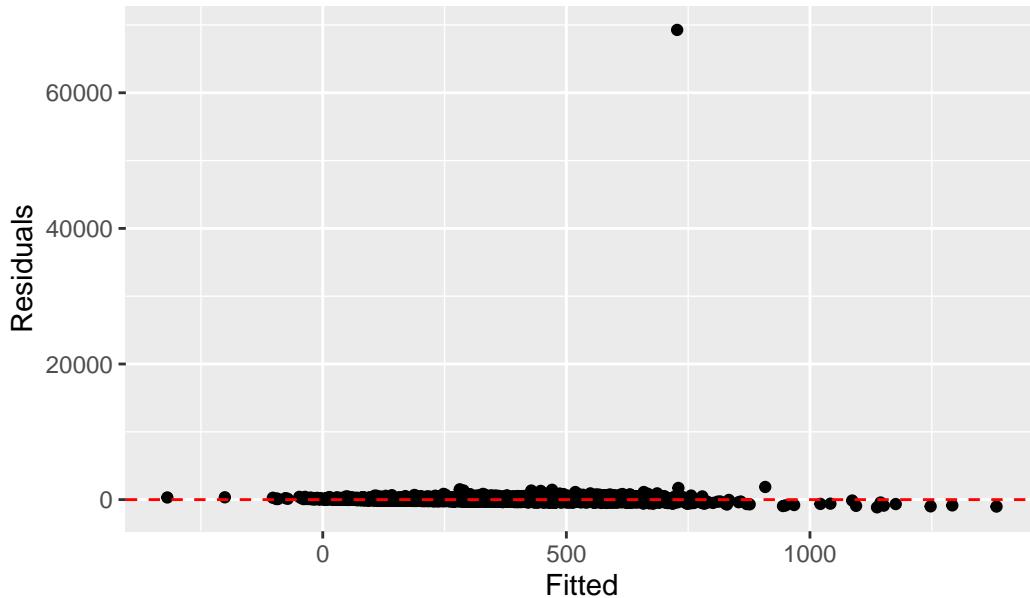
We are concerned about spatial correlation, so we will split the data 75/25 to minimize this effect.

Checking Initial Modelling Conditions

term	estimate	std.error	statistic	p.value
(Intercept)	-451.480	168.155	-2.685	0.007
perc_bll_indicator	-2.188	2.858	-0.766	0.444
POC_other	6.355	0.995	6.389	0.000
med_income	-0.001	0.000	-4.522	0.000
median_age	-3.192	1.407	-2.269	0.023
perc_male	19.283	2.999	6.431	0.000

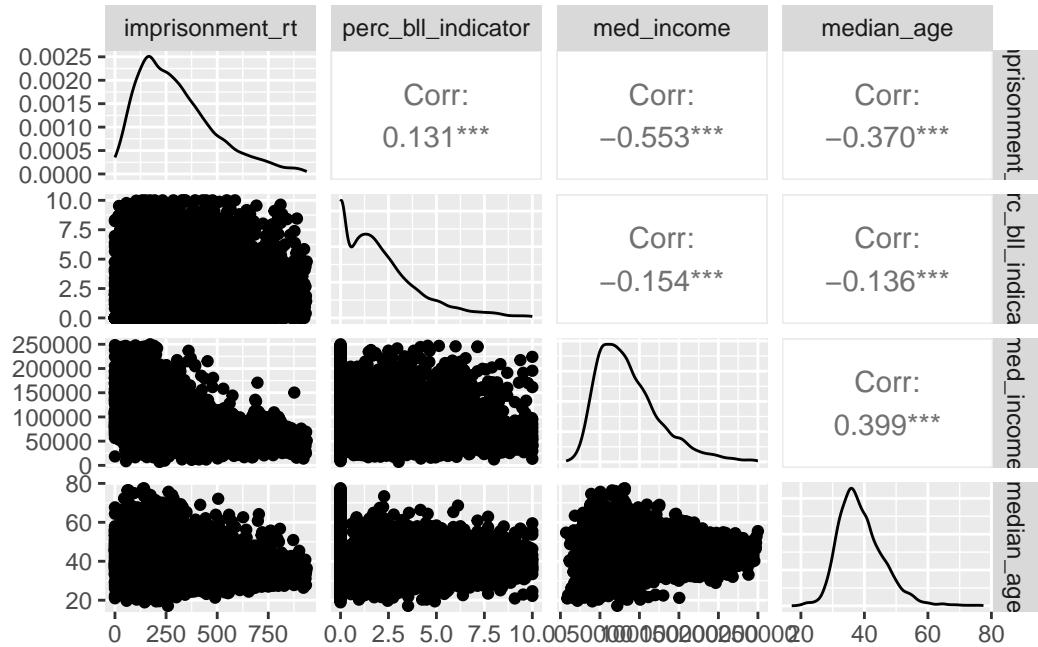
term	estimate	std.error	statistic	p.value
POC_other:med_income	0.000	0.000	-4.103	0.000

Residuals vs Fitted Values

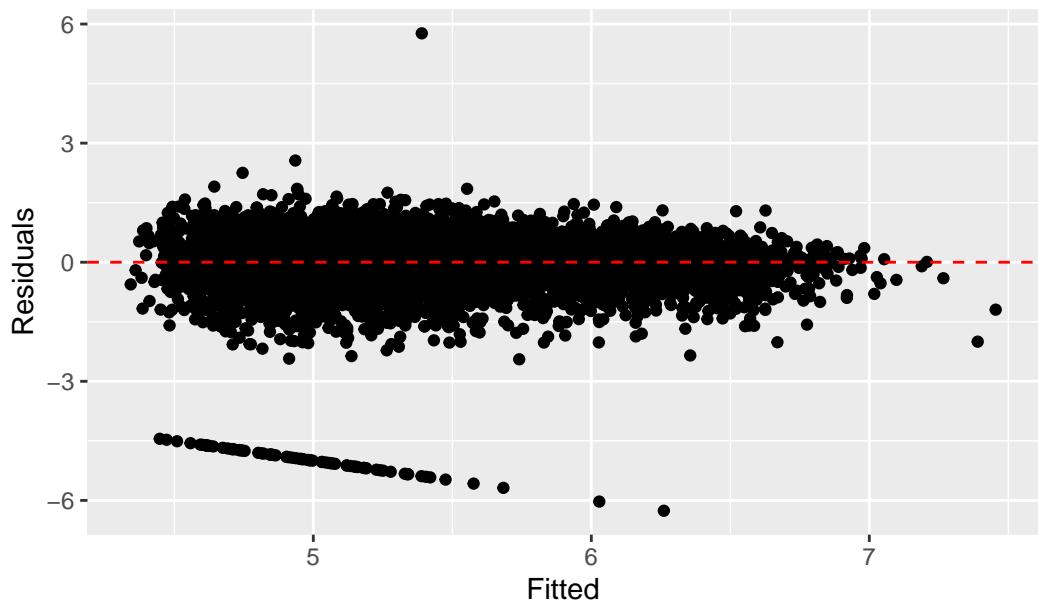


perc_bll_indicator	POC_other	med_income
1.049830	6.427185	2.321604
median_age	perc_male_POC_other:med_income	
1.637838	1.027161	3.972570

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.



Residuals vs Fitted Values



log_med_income	perc_bll_indicator	POC_other	median_age
1.679085	1.060100	2.176073	1.637213
perc_male			
1.023596			

The residual plot looks way better except for that one stripe of residuals. I am not sure what is causing that. The VIF values are awesome and super low too. Ignoring the stripe that would violate our other conditions, I am going to try a stepwise model selection.

The stripe could be due to census tracts next to each other influencing each other? This could be a good graph to show in class.

Start: AIC=118340.7

```
imprisonment_rt ~ perc_bll_indicator + POC_other + med_income +
    median_age + perc_male + med_income * POC_other
```

	Df	Sum of Sq	RSS	AIC
- perc_bll_indicator	1	333326	5075597167	118339
<none>			5075263840	118341
- median_age	1	2928906	5078192746	118344
- POC_other:med_income	1	9572938	5084836778	118356
- perc_male	1	23520298	5098784138	118380

Step: AIC=118339.3

```
imprisonment_rt ~ POC_other + med_income + median_age + perc_male +
    POC_other:med_income
```

	Df	Sum of Sq	RSS	AIC
<none>			5075597167	118339
+ perc_bll_indicator	1	333326	5075263840	118341
- median_age	1	2892561	5078489727	118342
- POC_other:med_income	1	9298152	5084895318	118354
- perc_male	1	23228936	5098826102	118378

Call:

```
lm(formula = imprisonment_rt ~ POC_other + med_income + median_age +
    perc_male + POC_other:med_income, data = bll_data_model)
```

Residuals:

Min	1Q	Median	3Q	Max
-1125	-108	-24	74	69283

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.513e+02	1.682e+02	-2.684	0.00729 **
POC_other	6.315e+00	9.933e-01	6.358	2.14e-10 ***
med_income	-1.367e-03	3.054e-04	-4.476	7.69e-06 ***

```

median_age      -3.172e+00  1.406e+00  -2.255  0.02415 *
perc_male       1.911e+01  2.990e+00   6.391  1.73e-10 ***
POC_other:med_income -5.480e-05  1.355e-05  -4.043  5.32e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 754.2 on 8924 degrees of freedom

(90 observations deleted due to missingness)

Multiple R-squared: 0.04212, Adjusted R-squared: 0.04159

F-statistic: 78.49 on 5 and 8924 DF, p-value: < 2.2e-16

term	estimate	std.error	statistic	p.value
(Intercept)	-451.301	168.151	-2.684	0.007
POC_other	6.315	0.993	6.358	0.000
med_income	-0.001	0.000	-4.476	0.000
median_age	-3.172	1.406	-2.255	0.024
perc_male	19.106	2.990	6.391	0.000
POC_other:med_income	0.000	0.000	-4.043	0.000

Interestingly, the lead predictor was removed through stepwise selection, perhaps indicating that it is not useful here.