# Modeling the Relationship Between Education Attainment and Toxic Releases in California Using Simple Linear Regression

**Author:** Ethan Kusnadi

**Date:** September 24, 2025

**Github Link:** https://github.com/ekusnadi/Math-261A-Project-1

## Abstract

This paper aims to explore the relationship between high school education attainment and toxic release in a region. Education attainment is an important socioeconomic factor, as it can influence income or social mobility, and so the intuition is that this may have an impact on where people live in relation to areas with high toxic release. A simple linear regression model is trained on education attainment and toxic release data originating from the CalEnviroScreen 4.0 data dashboard. Outcomes of the analysis show a slight positive correlation between education attainment and toxic waste levels, but issues with the input data cast some skepticism on the results. Assumptions of the simple linear model such as unbounded input data and constant error variance are violated, and so the results of this model are unreliable. Simple linear regression is probably not appropriate for this question.

## Introduction

In this paper, I intend to investigate the relationship between high school education attainment and toxic release in a region. As education attainment can be an important

factor in influencing where a person can afford to live or have power in politics, I wanted to evaluate whether educational disadvantages would be correlated to toxic pollution in a living region.

To conduct this experiment, I used education attainment and toxic release data from the CalEnviroScreen 4.0 data dashboard to train a simple linear regression model using R code. While my findings indicate a slight positive correlation between education attainment and toxic release in California, underlying issues in the data prevent me from drawing a strong conclusion.

The structure of the remainder of this paper will be as follows: Section 2 gives an overview of the data, Section 3 explains the methods used in the analysis, and Section 4 describes the results and possible further explorations.

## Data

The data used for this project comes from the CalEnviroScreen 4.0 data dashboard. Created by the California Office of Environmental Health Hazard Assessment (OEHHA), CalEnviroScreen is a mapping tool for ranking and categorizing California communities based on socioeconomic, environmental, and health data. This data is sourced from various state and federal government organizations, including California Air Resources Board (CARB), California Department of Pesticide Regulation (CDPR), Department of Toxic Substances Control (DTSC), and more.

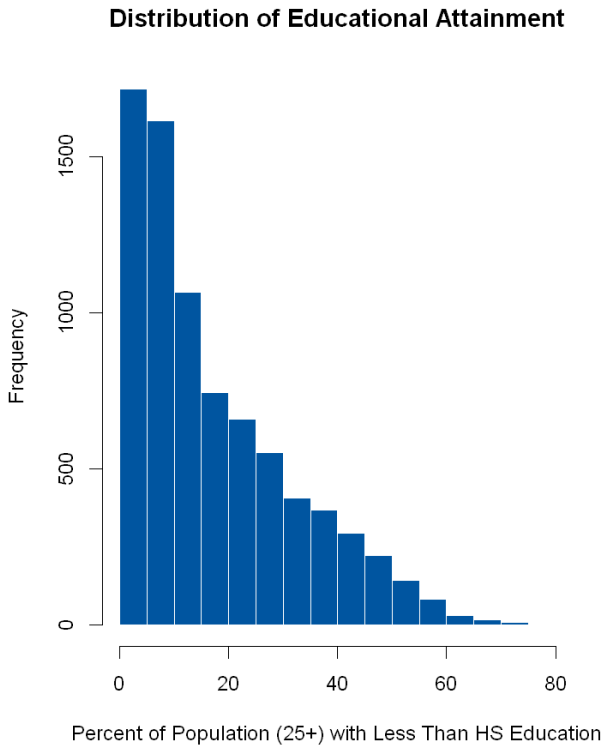
The raw data can be downloaded as an Excel spreadsheet, with metrics for each individual census tract in California. For my analysis, the relevant columns in this dataset are ‘Education’ and ‘Tox. Release’, which corresponds to education attainment and toxic releases from facilities.

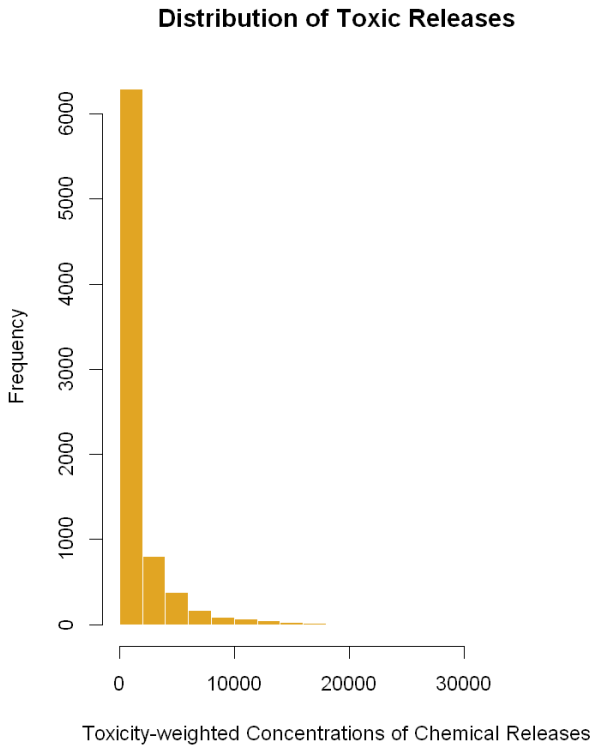
The education attainment data represents the percent of the population within a particular census tract that is over age 25 and has less than a high school education. This data was originally sourced from the American Community Survey (ACS), which is an ongoing survey of the US population conducted by the US Census Bureau. It is a result of 5-year estimates for 2015-2019, where estimates originate from a sample of the population but are evaluated based on the standard error and relative standard error and only included in the dataset if they meet a reliability criteria of either RSE less than 50 or

SE less than the mean SE of all California census tract estimates for education. As this is a percentage, it is bounded between 0 and 100, which may cause issues with our regression later on, which normally assumes an unbounded dependent variable from negative to positive infinity.

The toxic releases from facilities data contains the toxicity-weighted concentrations of modeled chemical releases to air from facility emissions and off-site incineration, averaged over 2017 to 2019 and including releases from Mexican facilities averaged over 2014 to 2016. This is a combination of data collected in the Toxics Release Inventory (TRI) and Risk Screening Environmental Indicators (RSEI) by the US Environmental Protection Agency (US EPA) as well as the Mexico Registry of Emissions and Transfer of Contaminants (RETC). Southern California borders Mexico, so the RETC data accounts for this cross-border pollution. Annual release of listed chemicals (in lbs/yr) in the TRI dataset is self-reported by facilities annually if they meet a threshold of at least 10 employees and at least 25,000 lbs manufactured or 10,000 lbs used of certain chemicals. RSEI then performs a data transformation on the TRI and RETC data using chemical-specific weights and atmospheric dispersion models to create an estimated pollution concentration score. This score is meant to be ranked against other RSEI scores and does not have meaning otherwise, which may lead to issues in the linear regression later on. This index has a lower bound of 0, which may also cause issues.

I removed every row in the dataset that had missing values (“NA”) for either education or toxic release. This resulted in removing 103 rows, leaving 7932 points in the dataset. There are several outliers in terms of the toxic release, but I have decided to include them in the model for now and try to verify if these numbers are correct in a future analysis. Below are histograms illustrating the frequency distributions of the observed values.




# Methods

For analysis of our data, we will be employing a simple linear regression model with the education variable as the predictor and the traffic variable as the response. The generalized equation for simple linear regression is presented below.

$$\mathbf{Y}_i = \beta_0 + \beta_1 \mathbf{X}_i + \varepsilon_i \text{ for } i = 1, \dots, n$$

In our scenario, the response $\mathbf{Y}_i$ represents the toxic release variable and the predictor $\mathbf{X}_i$ represents the education variable. The intercept $\beta_0$ represents the average education attainment of a census tract in California when the toxic release index is 0. The slope parameter $\beta_1$ represents the change in the mean of the probability distribution of the toxic release variable for a single percentage point increase in education attainment.

The linear regression model assumes validity, which is that the variables accurately reflect the quantities of interest, and the model includes all relevant predictors. In this case, we have education attainment percentages and toxic release index levels, which may not lead to perfect outcomes. As mentioned before, both are bounded variables, and the toxic release variable is a result of prior data processing.

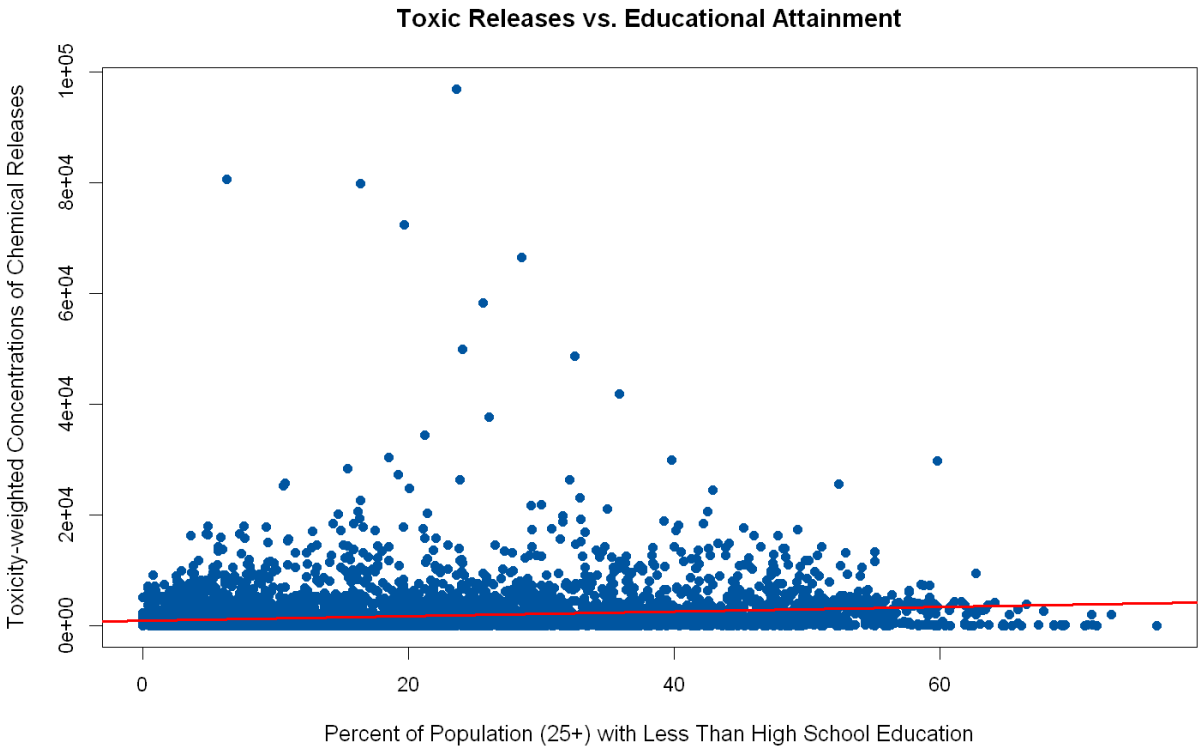
Another assumption of the linear regression model is that the data is representative of the population of interest. Representativeness will be true in our case if we only care about regions in California. If we want to extend this research outside of California, representativeness may not hold.

For the error terms, the linear regression model assumes independence and equal variance. We may run into issues here, as census tracts are arbitrary and there may in fact not be independence for either education attainment or toxic release.

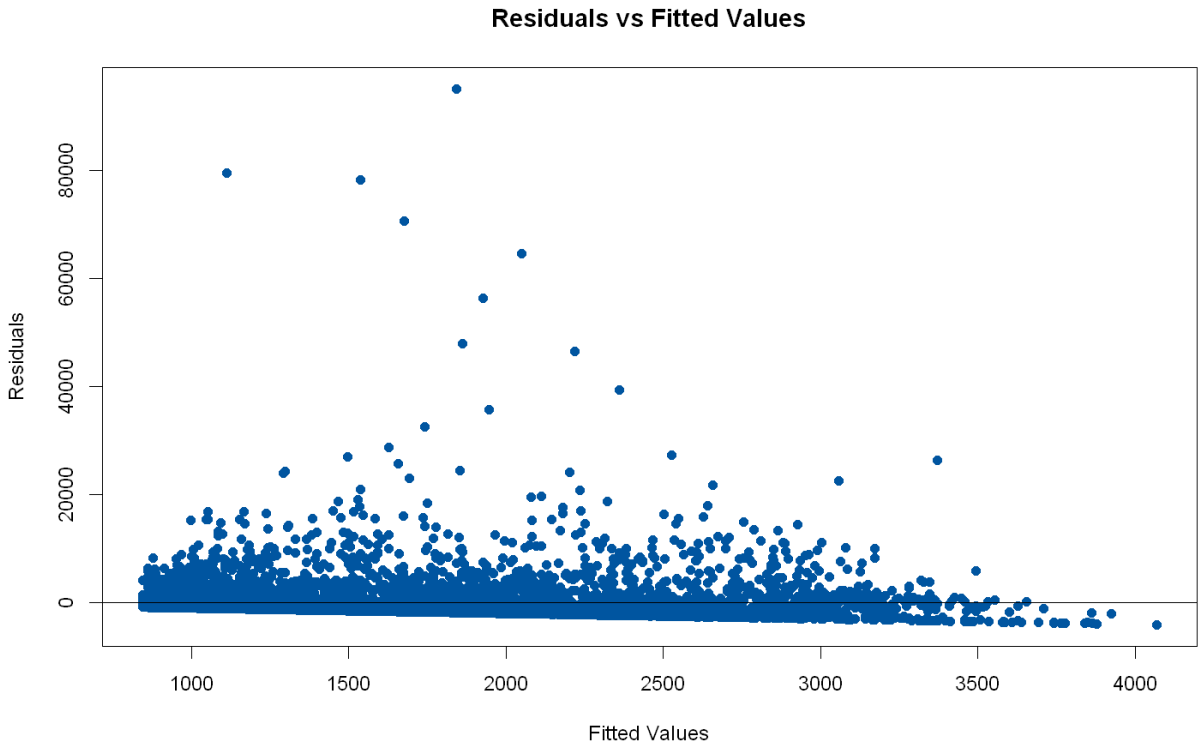
To perform the linear regression, I used the lm() function from the R programming language.

## Results

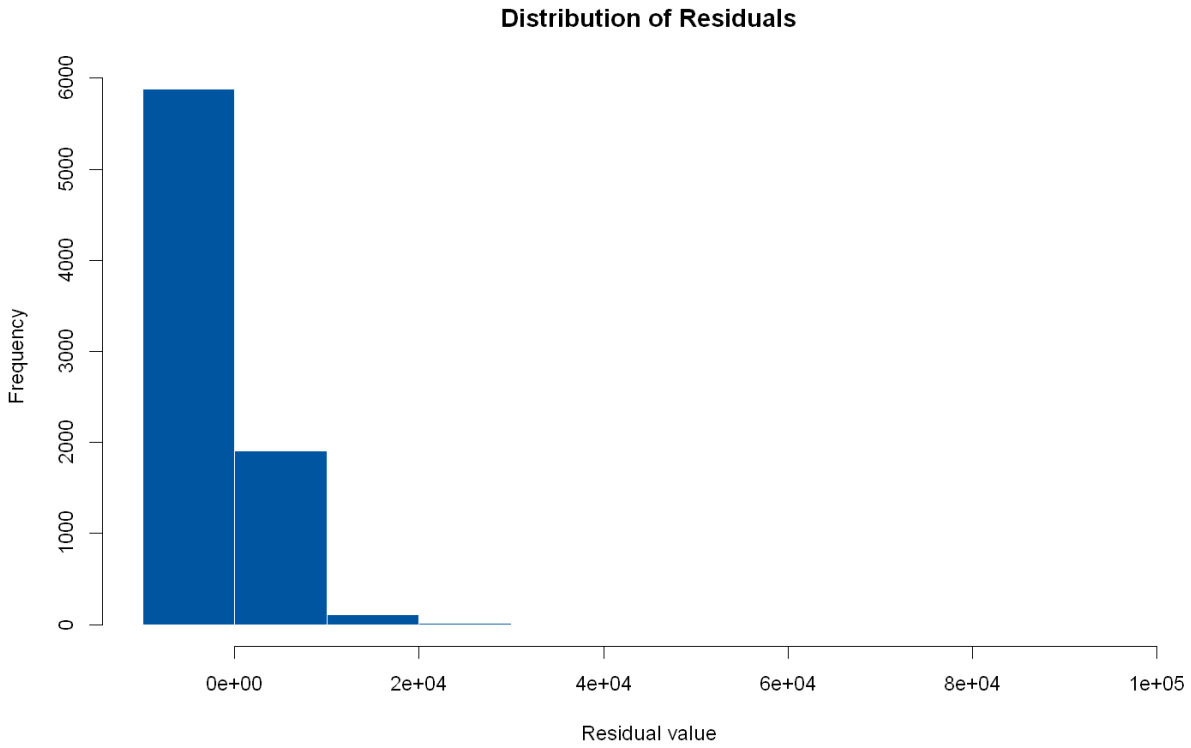
After training the simple linear regression model, I attained model parameters of intercept $\beta_0 = 844.3$ and slope $\beta_1 = 42.2$. This would indicate a positive correlation between the education attainment of residents and toxic release levels within a census tract in California. However, a quick look at the scatter plot with the linear regression line suggests that we should not jump to immediate conclusions.




In this plot, we see that most of the data is clustered around a toxic release variable level of 0. However, there are several massive outliers that may be affecting the regression line. The data is very right-skewed, and with the outcome varying over many orders of magnitudes, the linear regression line appears to be almost flat.




Plotting the residuals against the fitted values gives further insight on the validity of our linear regression. Clearly, the residuals are not randomly distributed around the 0 line. We see in this plot that the spread of the residuals increases as fitted values increase. This implies heteroscedasticity, or unequal variance of error terms.




From the histogram of residuals, we can see that the residuals are decidedly not normally distributed.

Therefore, although we found a positive slope, which would lead us to believe that there is a positive relationship between education attainment and toxic release, all of our resulting plots show that there are significant issues with this experiment. Thus, our results are unreliable. Simple linear regression is probably not best suited for this research question.

## References

Office of Environmental Health Hazard Assessment. (2021). *CalEnviroScreen 4.0* [Data set]. California Environmental Protection Agency. https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40

Office of Environmental Health Hazard Assessment. (2021). *CalEnviroScreen 4.0: Update to the California Communities Environmental Health Screening Tool* (Report). California Environmental Protection Agency. https://oehha.ca.gov/sites/default/files/media/downloads/calenviroscreen/report/calenviroscreen40reportf2021.pdf

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/