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

Ethan Kusnadi

October 3, 2025

This paper explores the relationship between high school education attainment and toxic release in California census tracts using data from CalEnviroScreen 4.0. Education attainment is examined as a socioeconomic factor that may shape where residents live in relation to environmental hazards. A simple linear regression model shows a slight positive correlation between education attainment and toxic release, but data issues and violations of regression assumptions cast doubt on the reliability of the results. Simple linear regression is probably not appropriate for this question.

1 Introduction

This paper investigates the relationship between high school education attainment and toxic release in a region. Education attainment is an important socioeconomic factor, often linked to income, mobility, and access to safer environments. While previous studies have examined how socioeconomic factors such as race and wealth relate to residence in areas with higher toxic release, the specific role of education has not been explored. Because these dynamics can influence where individuals and communities reside, this study evaluates whether lower education attainment is associated with greater exposure to toxic pollution.

Education attainment and toxic release data from the CalEnviroScreen 4.0 data dashboard were employed to fit a simple linear regression model using R (Office of Environmental Health Hazard Assessment 2021a; R Core Team 2025). While the analysis indicates a slight positive correlation between education attainment and toxic release in California, underlying issues in the data limit the strength of this conclusion.

*Project repository available at: https://github.com/ekusnadi/Math-261A-Project-1.

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.

2 Data

The data used for this project comes from the Office of Environmental Health Hazard Assessment (OEHHA) CalEnviroScreen 4.0 data dashboard (Office of Environmental Health Hazard Assessment 2021b). 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 the California Air Resources Board (CARB), California Department of Pesticide Regulation (CDPR), Department of Toxic Substances Control (DTSC) (Office of Environmental Health Hazard Assessment 2021b).

The raw data can be downloaded as an Excel spreadsheet from OEHHA's CalEnviroScreen website, with metrics for every individual census tract in California (Office of Environmental Health Hazard Assessment 2021b). A census tract is a geographic subdivision of a county defined for the purpose of collecting and analyzing census data. For this 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), an ongoing survey of the US population conducted by the US Census Bureau (U.S. Census Bureau 2019). 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 independent 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 releases from Mexican facilities averaged over 2014 to 2016. This combines data from the Toxics Release Inventory and Risk Screening Environmental Indicators maintained by the U.S. Environmental Protection Agency, as well as the Mexico Registry of Emissions and Transfer of Contaminants (U.S. Environmental Protection Agency (EPA) 2021; Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) 2021). 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.

Every row in the dataset containing missing values ("NA") for either education or toxic release was removed. This resulted in removing 103 rows, leaving 7932 points in the dataset. Several outliers are present in the toxic release data, but they are retained in the current model, as their accuracy has not yet been verified (Office of Environmental Health Hazard Assessment 2021b). Below are histograms illustrating the frequency distributions of the observed values.

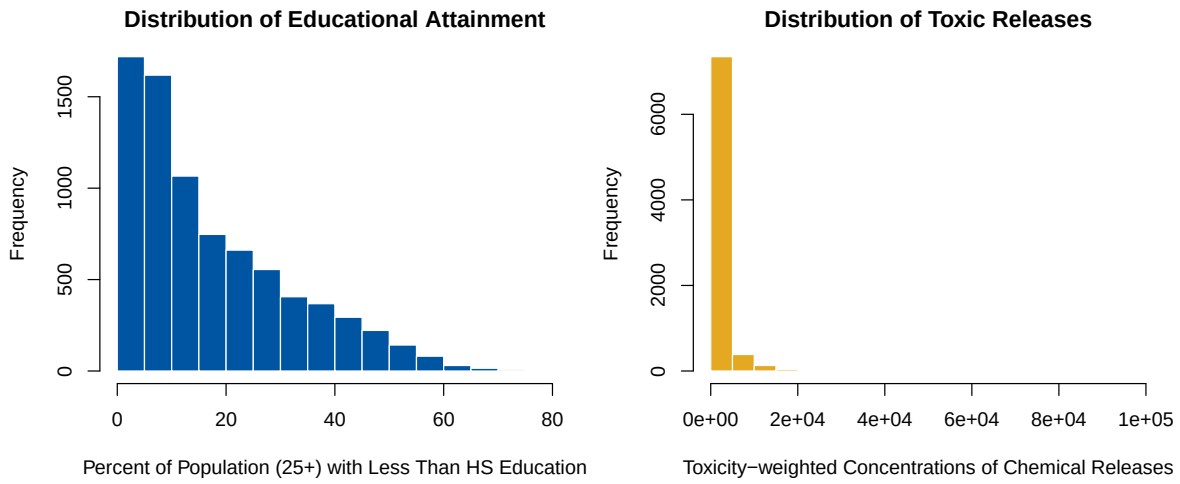

Figure 1: Distributions of Educational Attainment (left) and Toxic Releases (right).

3 Methods

A simple linear regression model was fitted to analyze the relationship between education attainment and toxic release. The generalized equation for simple linear regression is presented below:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{for } i = 1, \ldots, n$$

In this analysis, the response Y_i represents the toxic release variable and the predictor X_i represents the education variable. The intercept β_0 corresponds to the expected toxic release index in a census tract in California when the education variable equals zero. The slope parameter β_1 represents the expected change in toxic release for a one percentage point increase in education attainment, which would mean an increased percentage of people without a high school education. The random error term ε_i captures unexplained variance.

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 dataset, both education attainment percentages and toxic release index levels are bounded and result from prior data processing, which may limit the precision of estimates. Another assumption of the linear regression model is that the data is representative of the population of interest. Representativeness will hold if inference is restricted to regions in California, but may be violated if extended to other areas.

For the error terms, the linear regression model assumes independence, equal variance, a mean of zero, and normal distribution. In this problem, nearby census tracts may be spatially correlated, meaning that educational attainment or toxic releases may not be independent across tracts. As the linear regression model is estimated using the ordinary least squares method to minimize the sum of squared residuals, the residuals produced have a sample mean of zero by construction.

The linear regression was performed by employing the lm() function from the R programming language (R Core Team 2025).

4 Results

After training, the simple linear regression model produced 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 across census tracts in California. However, a quick look at the scatter plot with the linear regression line suggests that we should not jump to immediate conclusions.

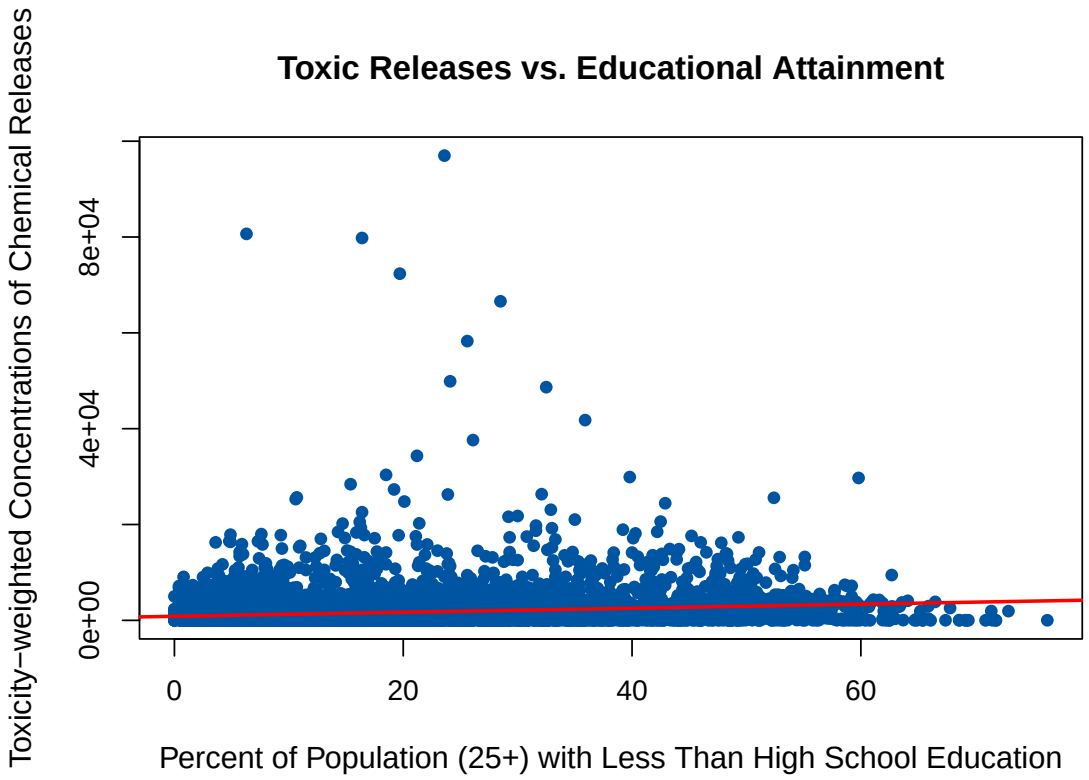

Figure 2: Scatter plot of Toxic Releases vs. Educational Attainment with fitted regression line.

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.

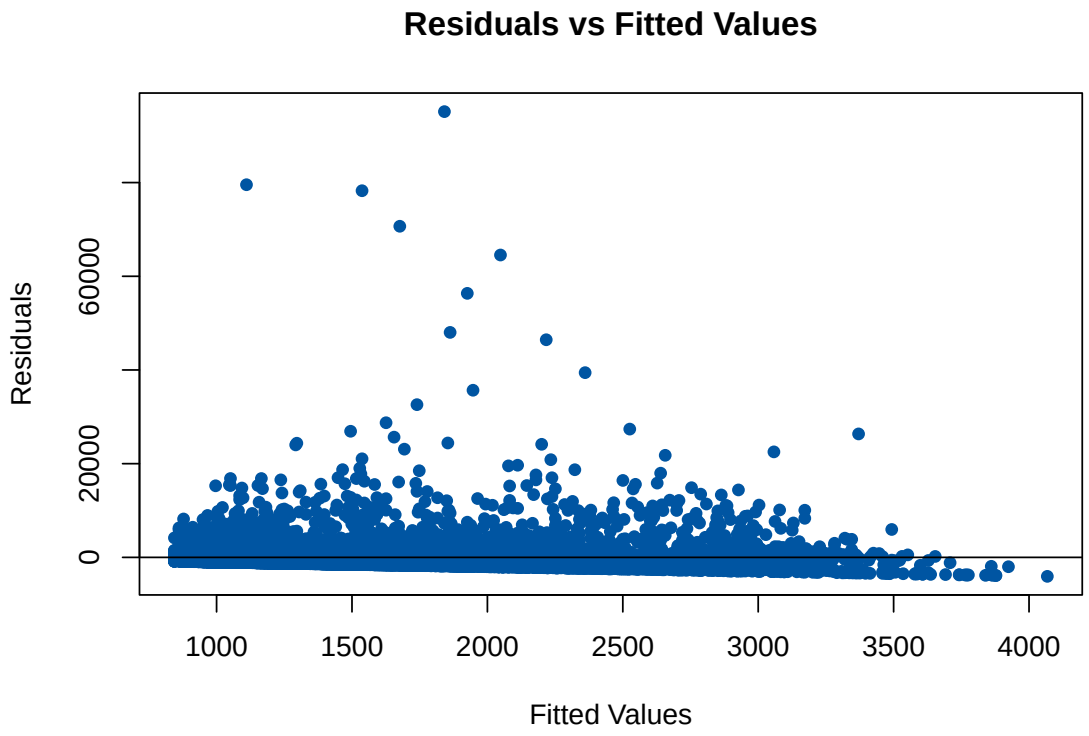

Figure 3: Residuals versus fitted values from the simple linear regression.

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 zero line. We see in this plot that the spread of the residuals decreases as fitted values increase. This implies heteroscedasticity, or unequal variance of error terms.

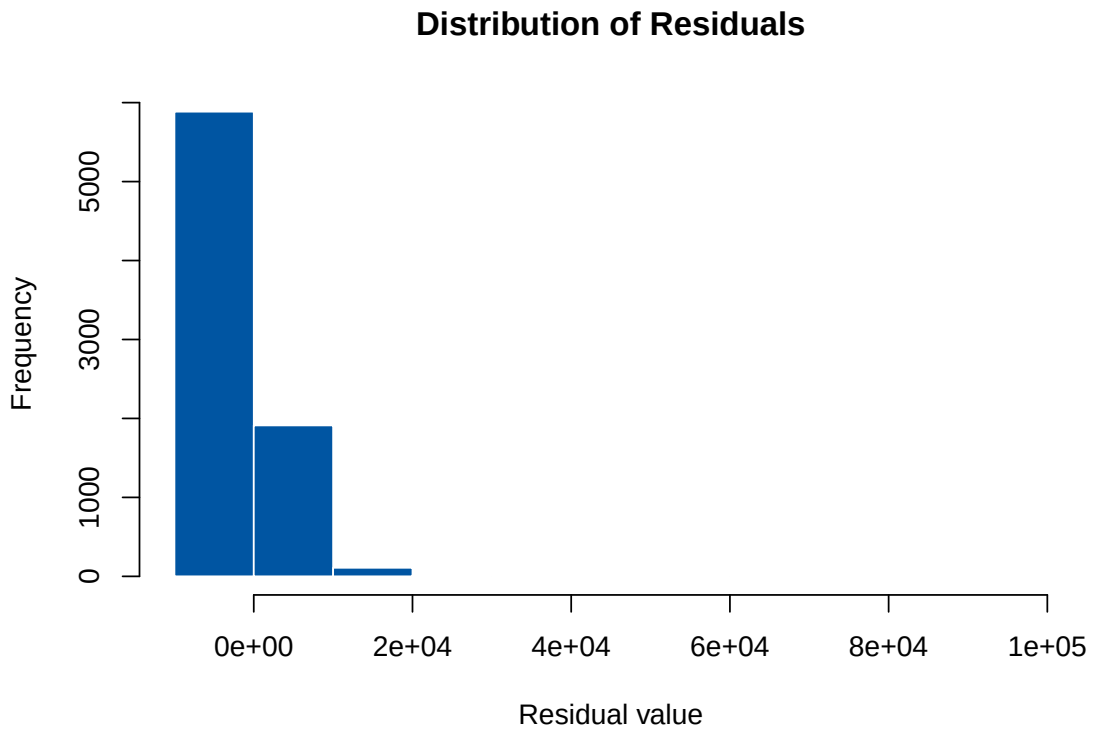

Figure 4: Histogram of residuals from the simple linear regression.

From the histogram of residuals, we can see that the residuals are right-skewed with a heavy tail, indicating the presence of large outliers. This violates the assumption of normally distributed residuals.

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. 2021a. "CalEnviroScreen 4.0 [Data Set]." https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40.

———. 2021b. "CalEnviroScreen 4.0: Update to the California Communities Environmental Health Screening Tool." California Environmental Protection Agency. https://oehha.ca.gov/media/downloads/calenviroscreen/report/calenviroscreen40reportf2021.pdf.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/.

Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT). 2021. "Registro de Emisiones y Transferencia de Contaminantes (RETC) [Data Set]." http://sinat.semarnat.gob.mx/retc.

U.S. Census Bureau. 2019. "American Community Survey (ACS) 5-Year Estimates (2015–2019) [Data Set]." https://www.census.gov/programs-surveys/acs.

U.S. Environmental Protection Agency (EPA). 2021. "Toxics Release Inventory (TRI) and Risk Screening Environmental Indicators (RSEI) [Data Set]." https://www.epa.gov/toxics-release-inventory-tri-program.