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

Ethan Kusnadi

October 4, 2025

This paper explores the relationship between high school educational attainment and toxic release in California census tracts using data from CalEnviroScreen 4.0. Educational 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 lower educational attainment and higher toxic release levels, but data issues and violations of regression assumptions cast doubt on the reliability of the results. Simple linear regression is likely unsuitable for this research question, and future analyses could apply data transformations or alternative models to better capture the relationship.

1 Introduction

This paper investigates the relationship between high school educational attainment and toxic release in a region. Educational 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 fully explored. Because these dynamics can influence where individuals and communities reside, this study evaluates whether lower educational attainment is associated with greater exposure to toxic waste levels.

Educational 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

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

correlation between lower educational attainment percentages and higher toxic release index levels in California, underlying issues in the data limit the strength of this 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.

2 Data

Data for this project was obtained 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 is available for download 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 educational attainment and toxic release index levels from facilities.

The educational 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). These values are based on 5-year ACS estimates from 2015 to 2019. Each census tract estimate was evaluated for reliability using both the standard error (SE) and relative standard error (RSE). Estimates were included in the dataset only if they met at least one reliability criterion: an RSE less than 50 or an SE less than the mean SE of all California census tract estimates for education. Because educational attainment is a percentage, the bounded scale may complicate model fit and interpretation.

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 (TRI) and Risk Screening Environmental Indicators (RSEI) maintained by the U.S. Environmental Protection Agency, as well as the Mexico Registry of Emissions and Transfer of Contaminants (RETC) (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). Figure 1 shows the frequency distributions of the observed values, both of which are heavily right-skewed.

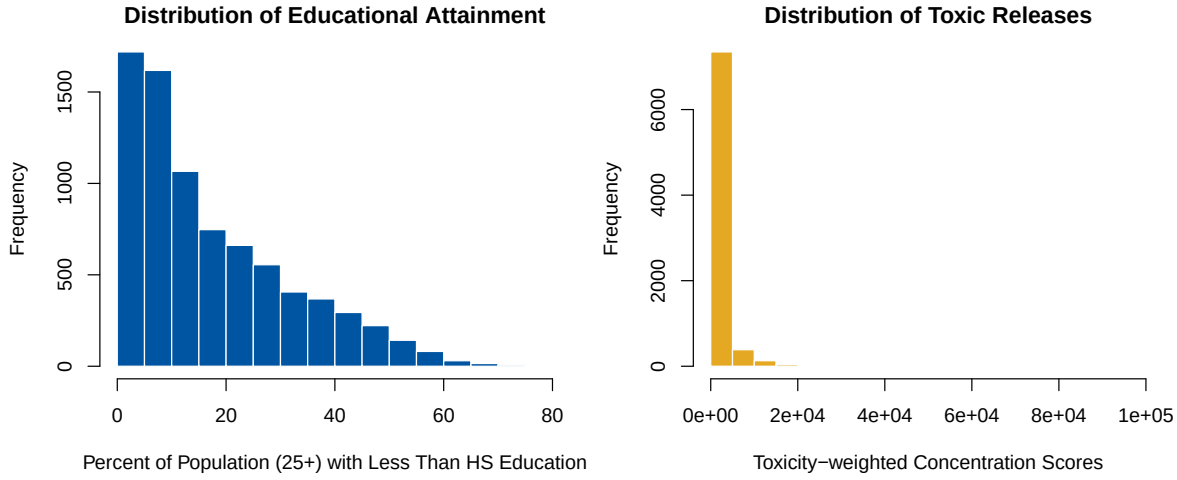

Figure 1: Distributions of Educational Attainment (left) and Toxic Releases (right).

3 Methods

A simple linear regression model was fitted to analyze the relationship between educational 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 educational 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 that the model includes all relevant predictors. In this dataset, both educational 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 research 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 with 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 implies that for a one percentage point increase in the proportion of residents without a high school education, the estimated toxic release index increases by about 42 units. The model explains only 2.9% of the variation in toxic release ($R^2 = 0.029$), indicating that educational attainment alone accounts for little of the variability in toxic release levels. Although the regression shows a positive correlation between a lower educational attainment of residents and higher toxic waste levels across census tracts in California, the limited explanatory power of the model makes this association uncertain. If this relationship were confirmed, it would suggest that communities with lower educational attainment experience disproportionate exposure to environmental hazards, highlighting environmental inequality.

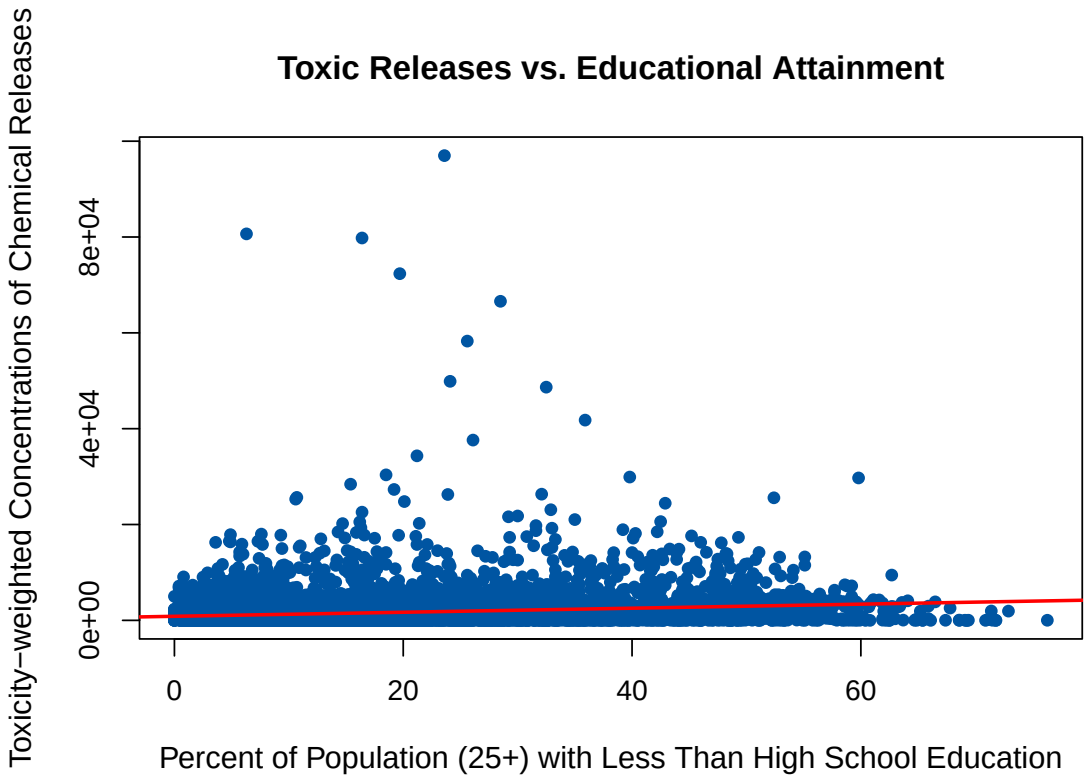

Figure 2: Scatter plot of Toxic Releases vs. Educational Attainment with fitted regression line shown in red.

These results are further illustrated in Figure 2. The scatter plot shows that most of the data is clustered near a toxic release variable level of 0. However, there are several massive outliers that may be affecting the regression line. The data is strongly right-skewed, and because the outcome spans several orders of magnitudes, the fitted regression line appears nearly flat.

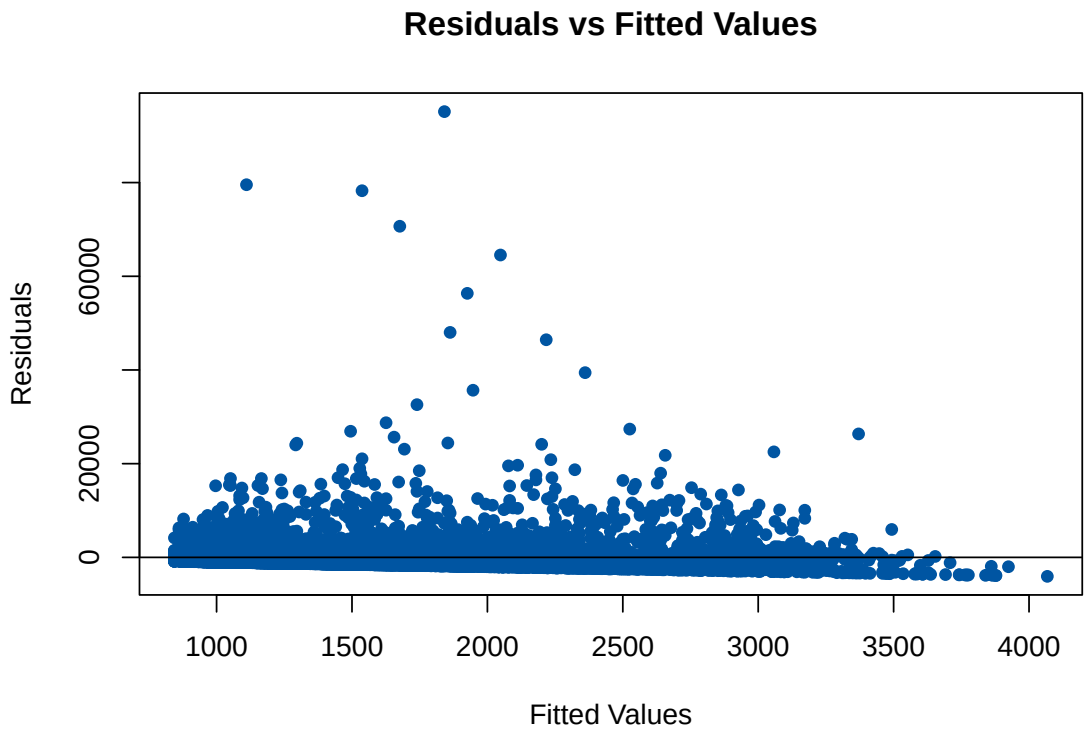

Figure 3: Residuals versus fitted values from the simple linear regression.

Plotting the residuals against the fitted values provides further insight into the validity of the regression assumptions. The residuals are not randomly scattered around the zero line and the spread of residuals decreases as fitted values increase. This pattern indicates heteroscedasticity, or unequal variance of the error terms, which violates the equal variance assumption of linear regression.

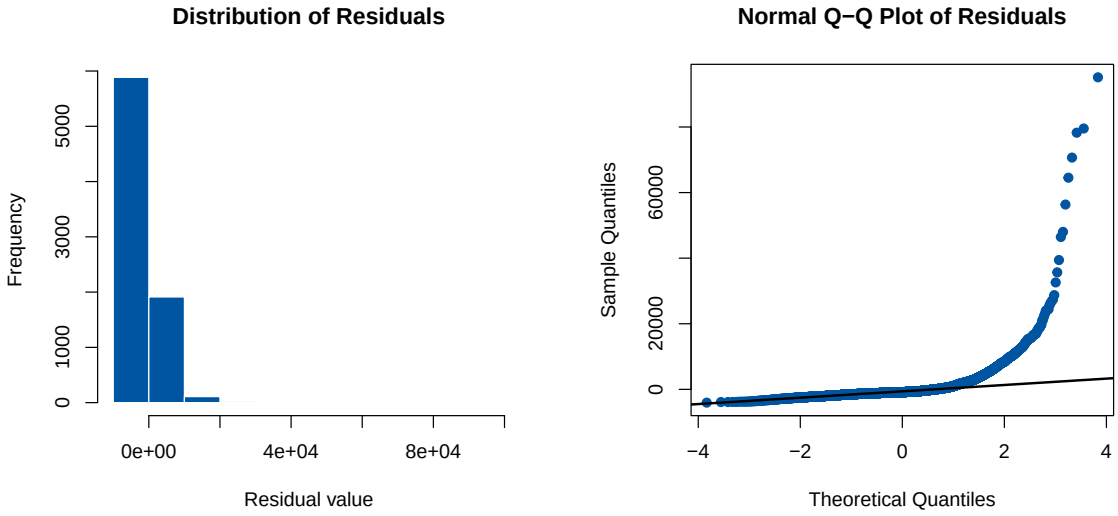

Figure 4: Histogram and Q-Q plot of residuals from the simple linear regression.

The histogram of residuals demonstrates a heavy right skew with extreme outliers, while the Q-Q plot shows the right tail deviating sharply from the reference line. Together, these diagnostics indicate a violation of the normality assumption for residuals in linear regression.

Therefore, although the positive slope suggests a positive relationship between lower educational attainment and higher toxic release levels, the diagnostic plots reveal significant issues with the analysis. As a result, the findings are deemed unreliable, and simple linear regression is likely not well suited for this research question.

Future analyses could apply simple linear regression to log-transformed toxic release data to account for the right-skewed distribution. Alternatively, nonlinear or multiple regression models incorporating additional socioeconomic variables could be tested to better isolate the effect of education.

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.