Is There a Linear Association Between Potential Risk of Lead Exposure and Poverty Levels?*

Megan Joseph

September 25, 2025

Many Americans live below the poverty line. As such, one might be interested in the indicators of poverty. Using data from the Calenviroscreen 4.0 tool, I show how the potential risk of lead exposure in homes may impact the poverty levels in a census tract. I fit a simple linear regression model with the percent of the population living two times below the federal poverty line as the response and the potential risk of lead exposure to children in low income homes as the predictor. The output is the model Y = 9.29 + 0.45X where Y is the response and X is the predictor.

1 Introduction

Can the potential risk of lead exposure in homes indicate poverty levels in an area? Many Americans live below the federal poverty line, especially in California where the cost of living is high. Identifying indicators of poverty may bring us closer to ending it. It has already been shown that older housing, which often has lead-based paint, and poverty are associated with high blood lead levels (August 2021).

[address knowledge gap, analysis, analysis, findings]

The remainder of the paper is structured as follows: Section 2 discusses the data, Section 3 discusses the model, and Section 4 presents the results.
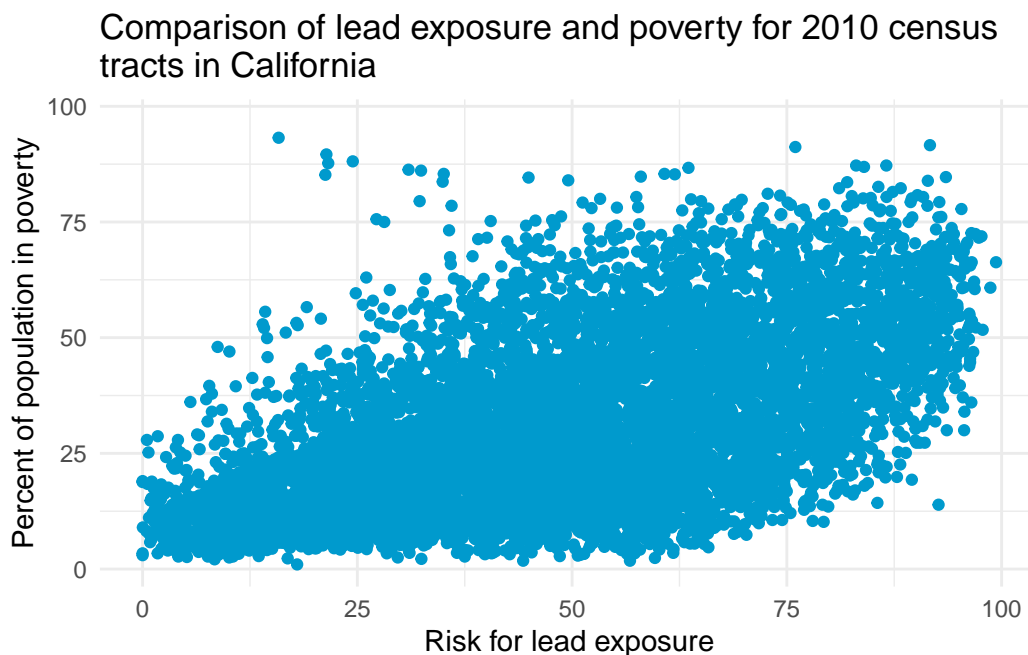
Figure 1: Scatter plot of the risk of lead exposure (x-axis) and percent of population in poverty (y-axis).

2 Data

The data is from CalEnviroScreen 4.0 which is a tool created by the Office of Environmental Health Hazard Assessment in the California Environmental Protection Agency used to help identify communities in California that are most affected by pollution (Wieland 2021). It was last updated in 2021. The data used for the tool come from national and state sources.

Each of the 8035 rows represents a census tract from 2010. The key variables are Lead, which quantifies the potential risk for lead exposure in children in low-income communities, and poverty, which represents the percent of population living below two times the federal poverty level (August 2021). Lead is transformed to a numeric variable because it is originally a character. Many other variables are present in the dataset but will not be used for this paper.

Poverty was taken from the American Community Survey, an ongoing survey of a sample of US population by the US Census Bureau, from 2015-2019. While the survey uses the federal poverty level to determine if people live below the federal poverty line, this dataset uses twice the federal poverty line because of California's high cost of living. Poverty is calculated by dividing the total population with a poverty status with the number of individuals below 200% of the poverty level for each census tract in California (August 2021).

---

*Project repository available at: https://github.com/meganajoseph/math261a_paper1.

One issue is that the American Community Survey come from a sample of the population rather than the entire population, so the Poverty variable may not be fully representative of the California population. To increase reliability, census tracts with estimates that either had a relative standard error less then 50 or a standard error less than the mean standard error of all California census tract estimates for poverty were included and the rest received no score (August 2021).

Another potential issue with the dataset is that there isn't direct information on exposures of lead, so the values used are based on pollution sources, releases, and environmental concentrations for potential exposures. To calculate this value, they used the percentage of households in each census tract with a likelihood of lead-based paint hazards based on housing age and the percentage of low-income households with children. Households with a likelihood of containing lead-based paint were determined by the construction period for each housing unit. The final lead risk for housing score is the weighted sum of the percent of home with likelihood of lead-based paint hazards percentile and the low-income household with children percentile. Here is the full equation: $0.6L + 0.4H = S$ where L is the percentile for lead-based paint hazards, H is the percentile of low-income household with children, and S is the final score (August 2021). I will assume that the indicators created are indicative of the potential for lead exposure.

Alternative sources of data include finding data directly from the US Census so as to use data from the entire population rather than a sample. Using a dataset that had actual measures of lead exposure in housing would also lead to more reliable answers.

?@fig-calenviron-scatter indicates a weak positive association between potential risk for lead exposure and poverty.

3 Methods

To answer the research question, I will fit a simple linear regression model to the data with Poverty as the response and Lead as the predictor. Let Y_i represent the percent of the population in the census tract living below two times the federal poverty level and X_i represent the potential risk for lead exposure in children living in low-income communities with older housing. The model can we written as

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

where β_0 represents the mean of the probability distribution of the percent of the population in poverty when risk for lead exposure is 0, β_1 represents the average increase in percent in poverty per unit increase in risk for lead exposure, and ε_i represents random error with mean 0 and variance σ^2.

The assumptions for linear regression are:

1. The variables used accurately reflect the quantities of interest, the model should include all predictors, and the model should be applicable for the research question.
2. The sample data is representative of the population of interest.
3. The mean function must be correct.
4. Errors must be independent.
5. The variance of the errors must be equal.
6. Errors must be normal.

[address assumptions]

I use the `lm()` function from the R programming language (R Core Team 2025) to fit the linear regression model.

4 Results

Comparison of lead exposure and poverty for 2010 census tracts in California with fitted simple linear regression model
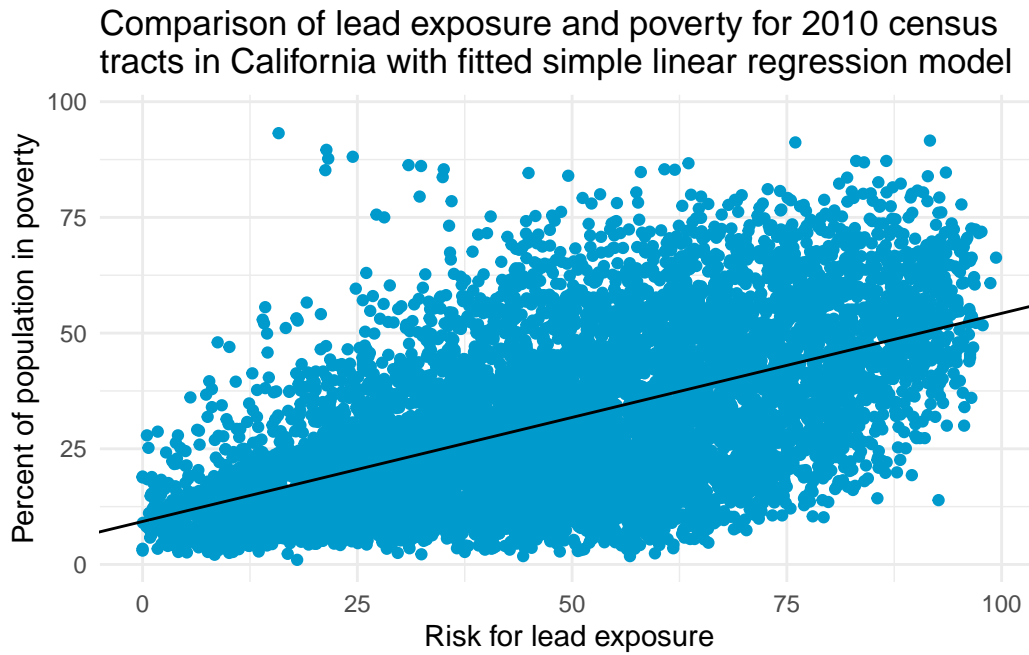
Figure 2: Scatter plot of the risk of lead exposure (x-axis) and percent of population in poverty (y-axis) with fitted linear regression model $Y_i = b_0 + b_0 X_i$.

The estimated slope parameter is $b_1 = 0.45$. This means that for each unit increase in potential risk for lead exposure, the percent in poverty increases by 0.45 on average. The estimated intercept parameter is $b_0 = 9.287$. This means that when the potential risk for lead exposure is 0, the average percent in poverty is 9.287.

[add hypothesis test]

[add discussion]

References

August, Laura et al. 2021. *Calenviroscreen 4.0*. Sacramento, CA: Office of Environmental Health Hazard Assessment. https://oehha.ca.gov/sites/default/files/media/downloads/cal enviroscreen/report/calenviroscreen40reportf2021.pdf.

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

Wieland, Walker. 2021. *Calenviroscreen 4.0 Excel Spreadsheet and Data Dictionary*. Sacramento, CA: Office of Environmental Health Hazard Assessment. https://calenviroscreen-oehha.hub.arcgis.com/documents/be09f14bef6244e8af4da6aeed89ec03/about.