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

Megan Joseph

October 3, 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).

[include bg info on poverty and lead]

[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.

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

2 Data

2.1 Data Overview

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 (August 2021). The data used for the tool come from national and state sources and was last updated in 2021 (Wieland 2021).

Each of the 7960 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. Additionally, there are 24 missing rows whicih will be dropped from the dataset. Many other variables are present in the dataset but will not be used for this paper. Figure 1 shows a scatter plot of Poverty on the y-axis and Lead on the x-axis. As the risk for lead exposure increases, the percent in poverty increases a bit as well.

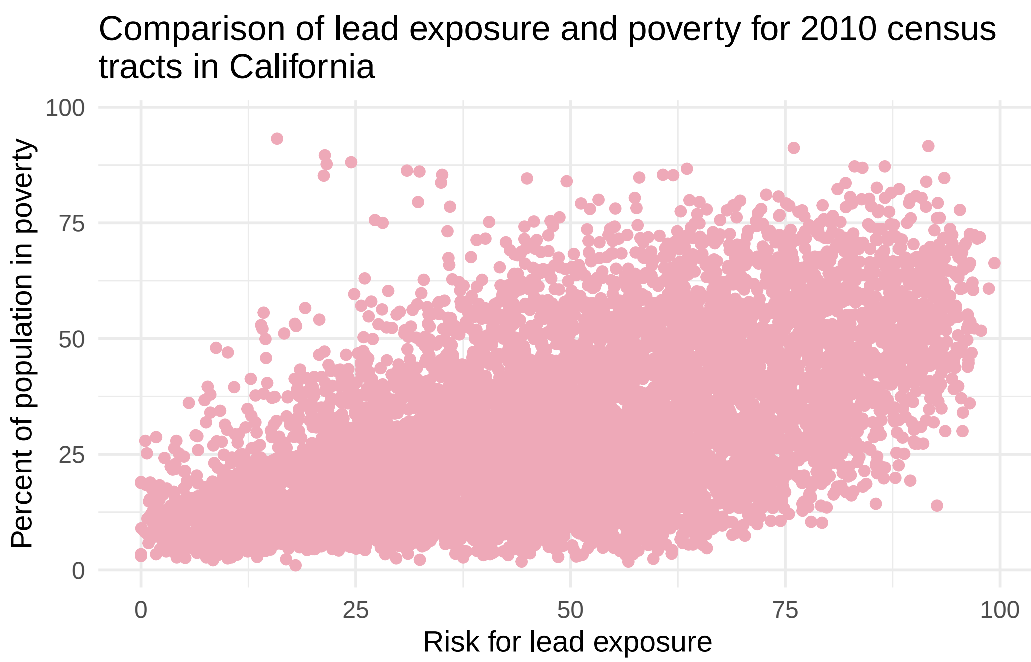

Figure 1: Scatter plot of the risk of lead exposure (x-axis) and percent of population in poverty (y-axis).

2.2 Poverty

The poverty column was taken from the American Community Survey (ACS), an ongoing survey of a sample of US population by the US Census Bureau, from 2015-2019 (August 2021). 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 (August 2021). 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). There are 0 missing rows which will be dropped from the data.

One issue with the dataset is that the ACS data 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).

2.3 Lead

Another potential issue 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 (August 2021). 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 (August 2021). The percentage of home with likelihood of lead-based paint hazards was calculated in a series of steps. First, California homes were grouped into categories based on age (August 2021). The, the number of housing units in each category was multiplpied by the reported percentage of homes with lead-based paint hazards (August 2021). Finally, the number of housing units in each category were added and divided by the total number of housing units in the census tract (August 2021). The percentage of low income households with children was calculated by estimating the number of households with incomes less than 80% of the county mean with one or more children under the age of six (August 2021). The final lead risk for housing score is the weighted sum of the percent of homes with likelihood of lead-based paint hazards percentile and the low-income household with children percentile: $0.6L + 0.4H = S$ 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.

2.4 Alternative Sources

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.

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 change in percent in poverty per unit increase in risk for lead exposure, and ε_i represents independent random error assumed to have 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.

3.1 Addressing Assumptions

1. The variables used here are Poverty and Lead from the dataset. They reflect the quantities of interest: the percent living under two times the federal poverty line and the potential risk of lead exposure. The model includes Lead as a predictor.

2. The population of interest is households in California. The sample data comes from each census tract in California which is representative of our population of interest. One caveat, though, is that the data is not per household but rather per census tract. Additionally, certain areas may not be accounted for in the Lead column due to ACS taking data from only a sample of the US population.

3. There is a slight departure from linearity as seen in Figure 2 (a) where the residuals are not equally dispersed around the line y=0. A curvilinear line may be better suited, but I will still fit a simple linear regression.

4. I plotted the residuals against zip codes because we need to check them against some series, usually time. This dataset only has geographic series, which still works. Figure 2 (a) shows that the errors seem independent. Most of the data are from zip codes greater than 87500, so we can ignore the few points around 0. There doesn't seem to be any strange patterns, so I will assume that the errors are independent.

5. Figure 2 (b) shows that the variance of the errors may not be equal. It shows a slight cone shape, especially below the y=0 line. Nonconstancy of error variance is very common and does not mean that the regression is invalid.

6. A histogram of residuals is the best way to check for normality because we can see if the graph resembles a normal distribution. Figure 2 (c) shows that the errors appear to be close to normal with a slight right skew. I will assume that the errors are normal.

Figure 2: Plots addressing the last four assumptions of linear regression.

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

4 Results

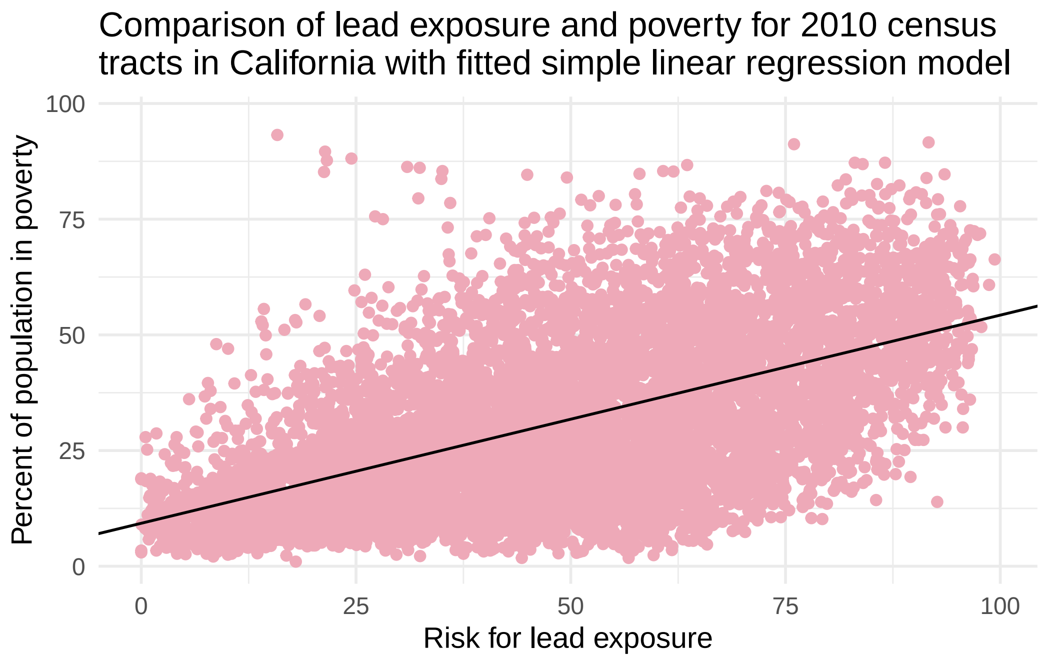

Figure 3: Scatter plot of the risk of lead exposure (x-axis) and percent of population in poverty (y-axis) with fitted linear regression model $Y_i = b_0 + b_0 X_i$.

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/calenviroscreen/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.