Mind the Education Gap: Mapping Poverty Predictions Across California's Communities*

A Regression Analysis Using CalEnviroScreen Data

Isis Martinez

October 29, 2025

We investigate the relationship between educational attainment and poverty rates in California using CalEnviroScreen 4.0 census-tract data compiled by the Office of Environmental Health Hazard Assessment (OEHHA). A simple linear regression of poverty rates on the percentage of adults without a high school diploma shows that each one-percent increase in low educational attainment is associated with an average 0.98-percentage point poverty level increase. A supplemental model including unemployment shows only a minimal improvement in explanatory power, with education remaining the highest predictor explored. Diagnostic tests indicate that several OLS assumptions are violated, so results should be viewed as descriptive associations rather than causal effects. Despite these limitations, the analysis suggests that educational attainment is strongly correlated with socioeconomic vulnerability across California communities.

1 Introduction

The link between educational attainment to socioeconomic well-being has been widely documented in prior research (Cutler and Lleras-Muney 2006; Zajacova and Lawrence 2018).

The persistent geographic inequities California faces highlights the importance of examining how educational outcomes relate to poverty across communities. Using the CalEnviroScreen 4.0 dataset (OEHHA 2021b), we explore the following question in this project: **To what extent is tract-level poverty associated with low levels of educational attainment in California?**

The CalEnviroScreen dataset was originally designed to identify communities facing disproportionate environmental burdens. However, it also provides tract-level socioeconomic and

*Project repository available at: https://github.com/iterrall/MATH261A-project-1-martinez.

environmental data suitable for exploring correlates of poverty across communities (OEHHA 2021b). These relationships can offer descriptive insights into how socioeconomic disadvantage aligns with environmental and health risks (OEHHA 2021a; Morello-Frosch and Shenassa 2006).

We investigate this question using a simple linear regression of poverty on the percentage of adults without a high school diploma. A supplemental model including unemployment was fitted for descriptive comparison, but its model assumption criteria was not fully assessed.

The remainder of this paper is structured as follows: Section 2 discusses the data, Section 3 describes the modeling approach, Section 4 presents the results, and Section 5 discusses the conclusions and limitations of the model.

2 Data

We use **California census tracts** as our observational units. These are small, relatively stable geographic areas defined by the U.S. Bureau (U.S. Census Bureau 2025a). We used the data in the **CalEnviroScreen 4.0** (OEHHA 2021a), which is a statewide screening tool developed by the California Office of Environmental Health Hazard Assessment (OEHHA). CalEnviroScreen compiles socioeconomic, environmental, and health indicators to support data-driven policy and business decisions.

From the CalEnviroScreen 4.0 that is based on 2015–2019 American Community Survey 5-year estimates (U.S. Census Bureau 2025a; OEHHA 2021b, 2021a), we focus on the following socioeconomic measures in our analyses:

- **Poverty** is defined in this project as the percent of the total tract population living below twice the federal poverty level (FPL). Using a 200% threshold adjusts for California's relatively high cost of living (OEHHA 2021a; U.S. Census Bureau 2025b).
- **Education** represents the percentage of adults age 25 years and older without a high school diploma. This is calculated as 100 minus the share of adults who have completed high school or higher education (U.S. Census Bureau 2025a, 2025b).
- **Unemployment** is the percentage of the labor force that is unemployed in the tract (OEHHA 2021a; U.S. Census Bureau 2025b).

All three variables are percentages in between 0 and 100, so we interpret coefficients as percentage-point changes. This means they represent expected changes in poverty rate per one-point percentage change. For example, a one-point increase in the share of adults without a high school diploma corresponds to an expected change in the poverty rate by the estimated coefficient (Wickham, Hester, and François 2023).

We imported data with the readr package (Wickham, Hester, and François 2023; R Core Team 2024) and cleaned it using dplyr (Wickham et al. 2025; R Core Team 2024) to retain complete records. Tracts flagged by OEHHA for high sampling uncertainty based on ACS standard error thresholds, as well as those missing socioeconomic variables, were excluded. Using the dplyr package (Wickham et al. 2025; R Core Team 2024), all variables were converted to numeric form and missing values were removed for consistency. Two datasets were created: a primary sample with 7906 tracts containing valid data for poverty and education, and a supplemental sample with 7658 tracts that also include unemployment. These preprocessing steps aimed to maximize data completeness and reliability, though residual measurement error from ACS estimates likely remains.

We note the **descriptive summaries** in the table Table 1 that we created using the dplyr, kableExtra, and knitr packages (Wickham et al. 2025; Xie 2015). Across 7658 tracts, the average poverty rate is 31.3%, ranging from 1.0% to 93.2%. The average share of adults (25+) without a high school diploma is 17.6%, with some tracts as high as 76.3%. The unemployment rate averages 6.3% and can reach 41.1%. These wide ranges highlight high variability across communities.

Table 1: Descriptive statistics for 7658 census tracts from the supplemental dataset including unemployment. Distributions represent tract-level rates (percentages) for each variable and summarize the mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values.

	Distribution (percent)						
Variable	Mean	SD	Min	25th	Median	75th	Max
Poverty	31.3%	18.2%	1.0%	16.3%	27.8%	44.3%	93.2%
Low Education	17.6%	14.6%	0.0%	5.8%	12.7%	26.1%	76.3%
Unemployment	6.3%	3.8%	0.0%	3.6%	5.5%	8.0%	41.1%

We also calculated correlation coefficients among the key variables using `stats::cor()` (R Core Team 2024). Poverty is strongly correlated with low educational attainment (r $\approx$ 0.79) and moderately correlated with unemployment (r $\approx$ 0.55). The relationship between education and unemployment is weaker (r $\approx$ 0.39). Because education shows the strongest association with poverty, it serves as a reasonable starting point for modeling tract-level differences in poverty rates.

That said, relying on simple pairwise correlations to choose predictors has clear limitations: correlations show association, not causation, and they don't account for confounding or overlap among socioeconomic factors. Here, the goal is descriptive, to highlight broad tract-level patterns rather than make causal claims. Despite these limitations, education was chosen as the main explanatory variable because it is both theoretically relevant and empirically linked to poverty, while keeping the model straightforward and easy to interpret.

Although education shows the strongest correlation with poverty, we recognize that selecting explanatory variables solely based on bivariate correlations is not ideal statistical practice.

Correlation does not imply causation, and such measures do not account for potential confounding or multicollinearity among predictors. However, our aim here is exploratory: to illustrate tract-level socioeconomic associations using a simple, interpretable model. Education was therefore selected as the primary explanatory variable because it provides a theoretically grounded and empirically strong relationship with poverty, while keeping the model parsimonious and transparent.

In our analysis, we include visualizations including a scatterplot of poverty versus education (Figure 1) that shows a positive linear trend. Additionally, we include a second plot coloring points by unemployment (Figure 5) that shows unemployment is also positively correlated with poverty, but with a weaker association than education (Figure 5) (Wickham 2016).

Finally, we note the **underlying data limitations** of our simple regression analysis. As with any ACS-derived data, estimates include sampling error, especially in smaller tracts. The variables have a bounded range between 0% and 100%, which could be introducing non-constant variance in regression models, and data clustering for percentage rounding. Additionally, the education measure applies only to adults 25+, while poverty covers all residents, and unemployment covers the workforce, so there is a variable denominator mismatch. Another limitation is the likely presence of geographic dependence because neighboring tracts could share similar socioeconomic conditions. This possible environmental clustering could bias standard errors and inference, which could suggest a need for spatial models or robust standard errors in future work.

3 Methods

To investigate the relationship between educational attainment and poverty, we adopt a simple linear regression model with poverty as the response and low-educational rate as the predictor. Let

- Y_i denote the percentage of the population living below 200% of the federal poverty line in tract i (poverty),
- X_i denote the percentage of adults age 25+ without a high-school diploma in tract i (education).

We fit the following model:

$$\mathbf{Y_i = {}_0 + {}_1 X_i + {}_i} \text{ for } \mathbf{i = 1, ..., n}$$

where β_0 is the intercept (expected poverty rate if no adults lack a diploma), β_1 is the slope (expected change in poverty for one-percentage point increase in X_i), and ε_i is the error term that encapsulates unobserved factors that impact poverty rates not explained by low-educational attainment percentage (R Core Team 2024; Gelman, Hill, and Vehtari 2021; Kutner et al. 2005). As a robustness check of this model, we fit a supplemental multiple linear regression that adds tract unemployment rate U_i as a second explanatory variable:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 U_i + \varepsilon_i \text{ for } i = 1, ..., n$$

where β_2 is the slope on unemployment. This makes it possible for us to test whether unemployment explains variation in poverty beyond education.

We estimate the parameters of the model (β_0 and β_1) using ordinary least squares (OLS) in R with the `lm()` function (R Core Team 2024; Kutner et al. 2005). In our case, lm(pov ~ edu, data = to_analyze_df) regresses tract-level poverty rates in the percentage of adults without a high school diploma. The function outputs estimated coefficients, residuals, fitted values, and summary statistics that we accessed using functions like summary() and coef()(R Core Team 2024).

When choosing **explanatory variables**, we sought to identify which socioeconomic variables in CalEnviroScreen best explain variation in tract-level poverty. To guide this process, we initially ran simple regressions of poverty on each socioeconomic indicator and compared their R^2 values. Education (percent of adults without a high school diploma) exhibited the strongest association, with unemployment showing a moderate relationship. We used this step as an exploratory tool to understand which indicators are most strongly associated with poverty. However, it is important to note that selecting variables based on bivariate relationships is not a best-practice approach to model building. Regardless for this project, based on this testing in addition to prior research linking low educational attainment to poverty in addition to poverty, we chose education as the primary explanatory variable and added unemployment in a secondary model to assess whether it improves explanatory power.

As discussed in Section 2, the variables are percentages (0–100%), so we interpret coefficients as percentage-point changes. We did not complete any **transformations** on them to preserve clarity of interpretations, though we note that bounded outcomes could produce non-normal error distribution.

Model validation involved evaluating overall model fit using the R^2 value, statistical significance through p-values, t-tests, and confidence intervals, as well as examining diagnostic plots generated in R using `ggplot2` (R Core Team 2024; Wickham 2016). Inference for regression parameters relies on linear model assumptions, so we evaluated whether their conditions were satisfied. The assumptions:

1. **Linearity**: For the primary model in this project, the conditional mean of the response is a linear function of the predictor(s):

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_{1i} \text{ for the primary model, and}$$

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \text{ For the supplemental model with unemployment,}$$

   where Y_i represents poverty, X_{1i} education, and X_{2i} unemployment. We assessed this using a scatterplot of poverty versus education with an overlaid fitted line. A straight line supports linearity, and curvature could indicate non-linearity and model misspecification.

2. **Independence of errors**: The model error terms are independent across observations:

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ for all } i \neq j.$$

We evaluated independence using a residuals vs. fitted values plot. Random point scatter around zero indicates independence, while clustering could indicate correlated errors (for example, spatial dependence among tracts).

3. **Equal variance of errors (homoscedasticity)**: For of the predictors, error variance is constant:

$$\forall X_i, \quad Var(\varepsilon_i \mid X_i) = \sigma^2.$$

We also used the residuals vs. fitted values plot to assess this. A consistent vertical spread suggests constant variance, and a funnel shape or clustering indicates unequal variance (heteroskedasticity). Because the data are percentages bounded between 0 and 100, it is likely we will see heteroskedasticity.

4. **Normality of errors**: The error terms are assumed to be normally independent and identically distributed with mean zero:

$$\varepsilon_i \sim N(0, \sigma^2).$$

We assessed normality using both a residual histogram and a Q–Q plot. A roughly bell-shaped histogram and points that follow the diagonal in the Q–Q plot suggest that the residuals are approximately normal, while noticeable skew or curved tails indicate departures from this assumption. These diagnostics help evaluate whether the OLS assumptions hold closely enough for reliable inference. The plots and their results are presented in Section 4.

The usual t-test of $H_0 : \beta_1 = 0$ is exact only when the model errors are independent, have homoskedasticity, and are normally distributed. When the normality assumption is relaxed, the test remains approximately valid for large samples under independence, due to the Central Limit Theorem. However, if the errors exhibit heteroskedasticity, the usual OLS standard errors become unreliable. In that case, heteroskedasticity-robust or spatially robust standard errors should be used to obtain valid large-sample inference.

However, possible pitfalls and **methodological limitations** should be noted. Spatial clustering of tracts could violate error independence assumption, ACS sampling variability introduces measurement error, and the bounded nature of percentage variables likely contributes to heteroskedasticity. Additionally, the mismatch in the sample populations for the different variables could bias coefficient estimates or inflate residual variance, since the predictor and response are not drawn from the same reference group. The difference could partially impact heteroskedasticity and weaken the precision of estimated relationships (Wickham 2016).

Future work could address listed challenges with robust standard errors, variance-stabilizing transformations, or spatial models that explicitly account for geographic dependence. Additionally, like we inspect another factor on poverty rates in the dataset such as unemployment

(Figure 5), we could explore other possible predictors of poverty by exploring more robust multiple linear regression models (Wickham 2016).

Together, these methods supported data cleaning and model validation to assess the relationship between education and poverty in a reproducible way

4 Results

The simple linear regression of poverty level on education gives the following fitted model:

$$\widehat{pov} = \hat{\beta}_0 + \hat{\beta}_1 \times (\text{Low Education}) = 14.255 + 0.979 \times (\text{Low Education})$$

where $\hat{\beta}_0$ is the estimated intercept parameter and $\hat{\beta}_1$ is the estimated slope parameter. In other words, for a tract with 0% of its adults lacking at least a high school diploma, this model would predict an average tract poverty rate of 14.26%. The estimated slope $\hat{\beta}_1 \approx 0.979$, indicating a predicted average 0.98% increase in tract poverty rate for one-percent increase in adults without a high-school diploma. The model explains approximately 61.59% of the variation in poverty rates across census tracts ($R^2 \approx 0.616$).

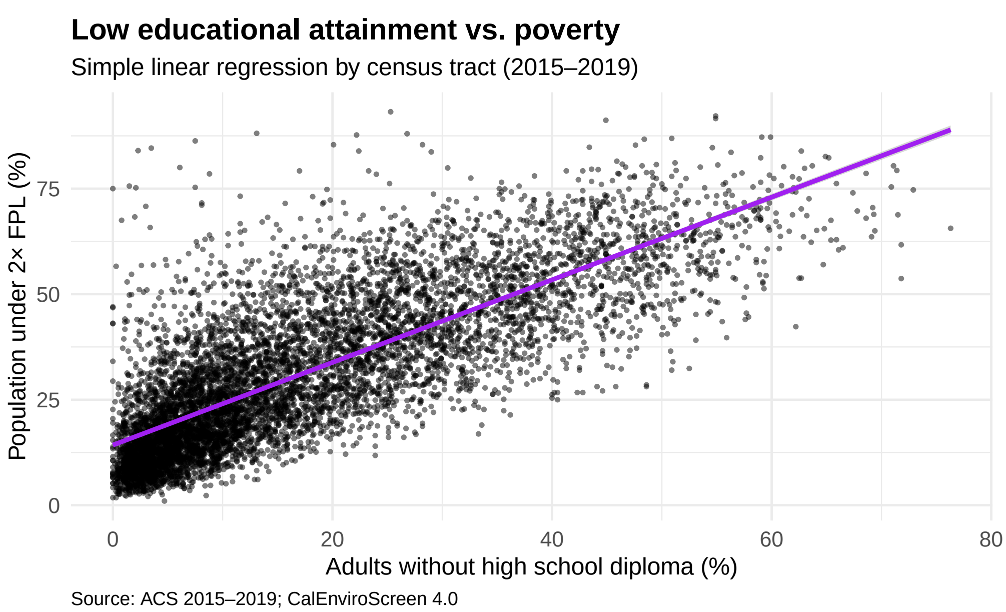

Figure 1: California census tracts (2015–2019): each +1 pp in adults without HS diploma is associated with 0.98 pp higher poverty (OLS). R^2 = 0.62.

To formally evaluate whether education is associated with poverty we conduct a t-test for the slope coefficient. Let our type I error rate be $\alpha = 0.05$. Let us test the following hypotheses:

H_0: $\beta_1 = 0$ (no relationship between education and poverty)

H_a: $\beta_1 \neq 0$ (nonzero association between education and poverty).

Under the usual linear model assumptions for t-inference (approximately linear mean structure, independent and homoscedastic errors, and near-normal residuals in large samples), the t-test of H_0 for the education coefficient yields $t \approx 113$ with a p-value <2e-16 (< 0.001) (Robinson 2014). Therefore, we see there is a positive association between low educational attainment and poverty. In other words, tracts with higher shares of adults lacking a high school diploma tend to have higher poverty rates on average. The 95% confidence interval for the slope, [0.96166, 0.99574], indicates that under repeated sampling, about 95% of such intervals would capture the true slope. Based on this interval, each additional percentage point of adults without a high school diploma is associated with an estimated 0.9617 – 0.9957 percentage point increase in tract poverty.

As detailed in Section 3, our diagnostic plots provide visual evidence about how well the data meet linear regression **assumptions**. The scatterplot (Figure 1) shows evidence for an approximately linear relationship between education and poverty. The residuals–fitted plot (Figure 2) shows clustering and a funneling of points, which implies heteroskedasticity and possible spatial dependence among nearby tracts. The Q–Q plot (Figure 3) displays a heavy right tail and a light left tail and the residual histogram (Figure 4) shows a right skew. Both of these plots show evidence of non-normality of this model's error terms.

These results show that the error independence, constancy, and normality assumptions are violated. These deviations primarily affect the precision of estimated standard errors, meaning inferences should be interpreted with caution. Applying more robust or spatially adjusted standard errors could be an extension for future work to make inferences with more reliable precision.

Additionally, we share a supplemental model with unemployment included to measure if another factor changes our results, which shows a similar effect ($\beta_1 = 0.838$) and a slightly higher $R^2 = 0.688$. Figure 5 visualizes this relationship by showing education remains a predictor of poverty even when controlling for unemployment. This suggests that differences in unemployment rates across tracts do not account for most of the variation in poverty once educational attainment is considered.

5 Discussion

Summary: Our regression analysis (Section 4) demonstrates a positive association between lower educational attainment and higher poverty levels. Tracts with a greater number of adults who have not received a high school diploma tend to exhibit higher average poverty rates. Similarly, unemployment has a positive (although weaker) correlation with poverty after accounting for education. These results suggest that education explains more of the variation in tract-level poverty across the CalEnviroScreen data set (OEHHA 2021a).

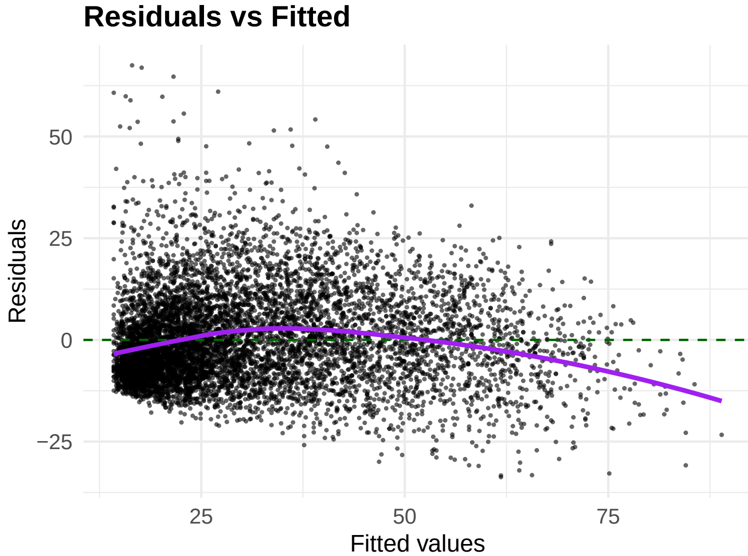

Figure 2: Residuals vs. fitted values for the simple OLS (poverty ∼ education). Each point shows how much the model's prediction differs from the actual poverty rate for each census tract, and the purple line shows the general trend in these residuals. The point clustering and trend line curvature suggests the model may suggest dependence of errors and heteroscedasticity.

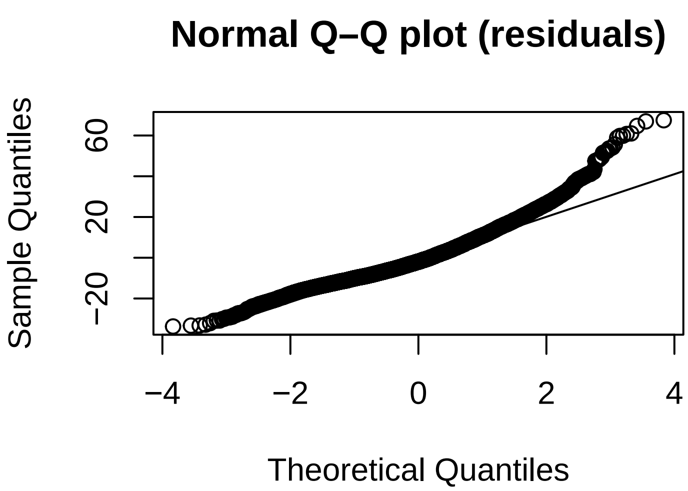

Figure 3: Normal Q–Q plot of residuals. Points near the line indicate approximate normality; curvature indicates deviations from normality.

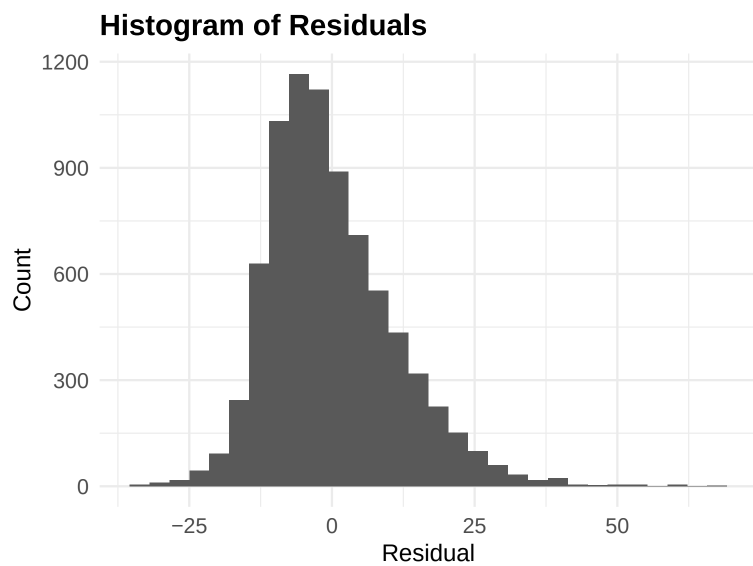

Figure 4: Histogram of residuals. A roughly bell-shaped, symmetric distribution supports the normal-errors assumption, skew indicates deviations from normality.

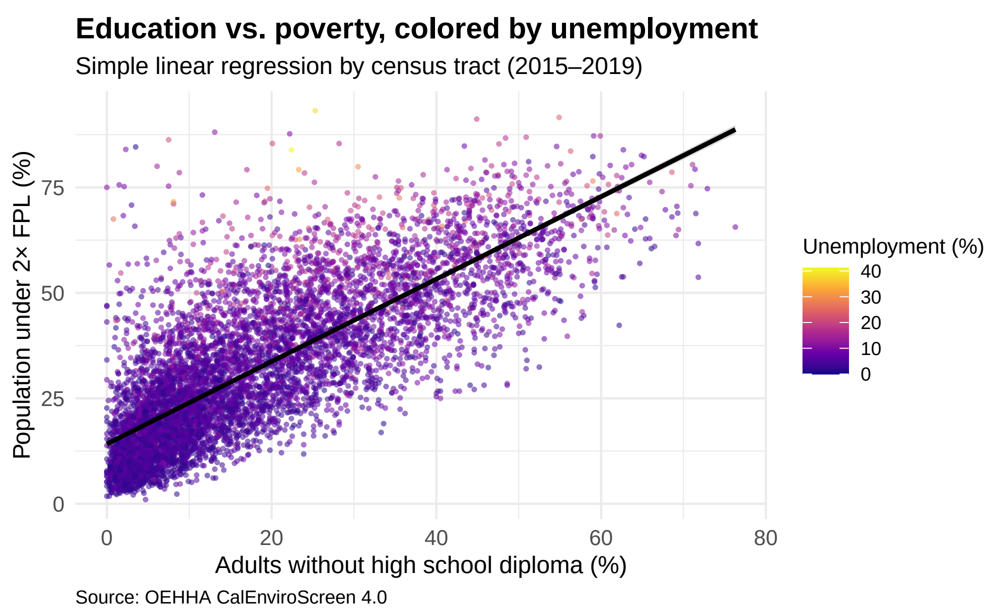

Figure 5: California census tracts (2015–2019). Poverty is higher where educational attainment is lower: each +1 % point in adults without a HS diploma is associated with 0.84 pp higher poverty (OLS), R^2 = 0.69.

Model assumptions: As discussed in Section 4, our diagnostic checks show that the linear model captures the overall relationship between education and poverty, but still show evidence for assumption violations. The heteroskedasticity, potential spatial dependence, and non-normal residuals are all evidence that OLS estimates will understate uncertainty and overstate statistical significance. These issues do not overturn the observed positive association but limit confidence in its estimated precision. Future work could re-estimate with robust or spatial standard errors to verify the strength of the education–poverty link.

Comparing education and unemployment: Including unemployment as an additional explanatory variable produced only minor changes in the estimated effect of education. Tracts with high poverty also tend to have high unemployment rates, as depicted in Figure 5. However, it fails to explain all the variation in poverty across tracts. The above findings are descriptive and should not be taken to imply causation, as unmeasured factors could influence both education and poverty (OEHHA 2021a).

Our simple linear regression model provides a descriptive summary of the association but fails to account for all the factors influencing poverty. The positive slope along with a moderate R^2, suggests that low educational attainment accounts for some of the variation in observed poverty rates. However, the analysis is still correlational and should be interpreted as evidence of association rather than causation.

Limitations: We used cross-sectional, observational data, which limits our causal inferences. The nature of tract-level geographical dependence of the data likely violates the independence assumption. Additionally, the bounded percentage outcomes produce unequal variances. Finally the ACS sampling error introduces measurement error.

On a broader level, a limitation of our analysis is the definition of poverty. CalEnviroScreen uses 200% of the federal poverty level (FPL) to account for California's high cost of living. This is a more appropriate benchmark than the standard FPL; it does not capture wide regional differences within the state. For example, housing costs in the Bay Area vs rural areas of California have a large range. Consequently, the same income threshold may reflect very different levels of economic hardship depending on location. This limitation implies that the poverty measure could overestimate poverty in some low-cost neighborhoods and underestimate it in high-cost neighborhoods. This could add more variation explained outside of education or unemployment.

Additionally, the education measure is only for adults aged 25 and older, but the poverty measures encompass the entire population. This denominator discrepancy means that our predictor and outcome are not measured on exactly the same group, thus introducing another limitation. For example, tracts with many children in poverty but relatively well-educated adults could weaken the observed association of our model. On the other hand, tracts with low adult education may experience higher poverty rates even among children and elderly residents who are not part of the education measure. This difference in denominators introduces another possible measurement error into our regression.

Implications: For the scope of this analysis, our regression model predicts that tracts with lower educational attainment tend to exhibit higher observed poverty rates. While the findings are correlational, they underscore education's possible relevance as a socioeconomic indicator in understanding community-level disadvantage. Because CalEnviroScreen informs environmental and equity-focused resource allocation, applying education as a contextual factor could enhance poverty vulnerability analyses. Future research could evaluate whether changes in educational access or attainment are associated with following changes in poverty using additional model designs.

6 References

Cutler, David M., and Adriana Lleras-Muney. 2006. "Education and Health: Evaluating Theories and Evidence." *NBER Working Paper Series*, no. 12352. https://www.nber.org/papers/w12352.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2021. *Regression and Other Stories*. Cambridge University Press.

Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models*. 5th ed. New York: McGraw-Hill/Irwin.

Morello-Frosch, Rachel, and Edmond D. Shenassa. 2006. "Environmental Justice and the Distribution of Pollution: The Case of California's Central Valley." *Environmental Health Perspectives* 114 (12): 1810–17. https://doi.org/10.1289/ehp.9310.

OEHHA. 2021a. "CalEnviroScreen 4.0: Updated Analysis." California Environmental Protection Agency. https://oehha.ca.gov/sites/default/files/media/downloads/calenviroscreen/report/calenviroscreen40reportf2021.pdf.

———. 2021b. "CalEnviroScreen Data Hub." https://calenviroscreen-oehha.hub.arcgis.com/#Data.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David. 2014. "Broom: An r Package for Converting Statistical Analysis Objects into Tidy Data Frames." *arXiv Preprint arXiv:1412.3565*. https://arxiv.org/abs/1412.3565.

U.S. Census Bureau. 2025a. "American Community Survey Design and Methodology Report." https://www.census.gov/programs-surveys/acs/.

———. 2025b. "Data.census.gov." https://data.census.gov/cedsci/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2025. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Romain François. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2015. *Dynamic Documents with r and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

Zajacova, Anna, and Elizabeth M. Lawrence. 2018. "Education and Health: The Casual Association and Challenges." *Annual Review of Public Health* 39: 273–89. https://doi.org/10.1146/annurev-publhealth-031816-044628.