Linking Education, Unemployment, and Poverty*

Insights from California's CalEnviroScreen Data

Isis Martinez

October 6, 2025

This paper investigates the relationship between educational attainment and poverty in California using census-tract data from CalEnviroScreen 4.0. We fit a simple linear regression of poverty rates on the percentage of adults without a high school diploma, finding that each one-point increase in low educational attainment corresponds to about a 0.98-point increase in poverty rate. A supplemental model including unemployment shows only a minimal improvement in explanatory power, with education remaining the stronger predictor. While diagnostics reveal violations of normality and variance constancy, the results highlight the significant role of education in shaping socioeconomic vulnerability. These findings underscore the potential for targeted education policies to reduce poverty disparities across California communities.

1 Introduction

The link between educational attainment to socioeconomic well-being has been widely corroborated in research (for example, Cutler and Lleras-Muney (2006) and Zajacova and Lawrence (2018)). The persistent geographic inequities California faces drives the need for us to take a closer look at how education relates to poverty when related to location. Therefore, using the CalEnviroScreen 4.0 dataset, (Office of Environmental Health Hazard Assessment (OEHHA) 2021), we ask the following question in this project: **To what extent is tract-level poverty associated with low levels of education in California?**

While its primary focus is to better understand and address environmental concerns in California communities, we can learn a great deal from the raw data that can be found on the CalEnviroScreen Data Hub (Office of Environmental Health Hazard Assessment (OEHHA)

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

2021) about possible socioeconomic well-being predictors and outcomes. There is a wealth of information around geospatial data that can be used to drive more equitable decision-making by exploring possible factors that play a role in changes to poverty levels. Understanding this question can be an important way to learn poverty disparity mitigation techniques within communities to improve health, since low socioeconomic status has been tied to multiple environmental risks and health disparities Morello-Frosch and Shenassa (2006).

We address this question by fitting a simple linear regression by taking poverty rate as a function of percentage of adults lacking a high school diploma. We then extend the analysis by exploring the contribution of unemployment makes on the model and examining the robustness of linear regression model assumptions.

The remainder of this paper is structured as follows: Section 2 discusses the data, Section 3 the model and the methods we used, Section 4 presents the results, and Section 5 discusses the conclusions in addition to weaknesses with the conclusions from this model.

2 Data

We use California census tracts (locations) as our observational, as they serve as small, relatively stable geographic areas defined by the U.S. Census Bureau (2025a). We used this data that was compiled in the **CalEnviroScreen 4.0** (OEHHA 2021), which is a statewide screening tool that compiles socioeconomic, health, and environmental indicators to support policy and business decisions.

We focus on the following socioeconomic measures drawn from the American Community Survey (U.S. Census Bureau (2025a), 2015–2019 5-year estimates) for this analysis (OEHHA 2021) from the CalEnviroScreen that was published in 2021 (Office of Environmental Health Hazard Assessment (OEHHA) 2021):

- **Poverty (pov)**: Percent of the tract population living below twice the federal poverty level (FPL). Using 200% of the FPL helps account for California's high cost of living (Office of Environmental Health Hazard Assessment (OEHHA) 2021; U.S. Census Bureau 2025b).
- **Education (edu)**: Percent of adults age 25 years and older without a high school diploma, which was derived as 100 minus the share of adults with at least a diploma (U.S. Census Bureau 2025a, 2025b).
- **Unemployment (unemp)**: Percent of the labor force that is unemployed.

All 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 in education (and employment in our supplementary model).

Before analysis, we removed tracts with missing values that were flagged as unreliable by OEHHA based on ACS standard error screening. After cleaning the raw data with Wickham et al. (2025), the dataset contained 7906 tracts for the primary model and 7658 for the supplemental model including unemployment.

Table 1: Descriptive summaries for key variables (tract level).

variable	mean	sd	min	p25	median	p75	max	n
pov	31.3	18.2	1	16.3	27.8	44.3	93.2	7658
edu	17.6	14.7	0	5.8	12.7	26.1	76.3	7658
unemp	6.3	3.8	0	3.6	5.5	8.0	41.1	7658

We note the **descriptive summaries** in the table Table 1 that we created using Robinson (2014) and Xie (2015). Across 7906 tracts, the average poverty rate is 31.33 %, with a wide range of values from 1 % to 93.2 %. Educational attainment shows large variability as well with an average 17.56 of adults lacking a highschool diploma, but there are tracts where the rate is over 75%. Unemployment rate is lower on average (6.26), but can be as high as 41.1 These wide ranges highlight substantial variability across communities, which could impact the reliability of our regression model.

We also calculated simple Pearson correlations among the key variables using stats::cor() in R Core Team (2024). Poverty is strongly correlated with low educational attainment (r 0.79), and moderately correlated with unemployment (r 0.55). The correlation between education and unemployment is weaker (r 0.39). These results reinforce our choice of education as the primary explanatory variable for modeling tract-level poverty.

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 **limitations** of our simple regression analysis. As with any ACS-derived data, estimates include sampling error, especially in smaller tracts. Variables are bounded between 0% and 100%, which can introduce non-constant variance in regression models, clustering for percentage rounding in the data. Additionally, the education measure applies only to adults 25+, while poverty covers all residents, creating a mismatch in denominators. Another limitation is the likely presence of geographic dependence, since neighboring tracts could share similar socioeconomic conditions. This spacial clustering could bias standard errors and inference, which would 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:

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

  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 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). S 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 with functions like summary() and coef().

Model choice and considerations:

When choosing **explanatory variables**, we sought to identify which socioeconomic variables in CalEnviroScreen best explain variation in tract-level poverty. To guide variable selection, we initially ran simple regressions of poverty on each socioeconomic indicator and compared their R^2 values. Education (percent without a high school diploma) showed the strongest relationship, with unemployment also showing a moderate association. 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 can produce non-normal residuals.

We completed **model validation** and fit by evaluating R^2, p-values, and diagnostic plots. We include the diagnostics we used to assess the validity of our standard linear regression **assumptions** using Wickham (2016) and R Core Team (2024):

1. **Linearity**: a straight line Figure 1 and Figure 2 suggest linearity between the predictor educational attainment and the outcome poverty.

2. **Independence of errors**: Figure 2 suggests independence if there is an even spread of points, while any clustering or patterns of the points would indicate a lack of independence between errors across census tracts

3. **Constant variance**: consistent spread of points in Figure 2 provide evidence of constant variance of error terms across values of X_i, while any funneling of the points suggest a violation of this assumption.

4. **Normality of residuals** with mean zero: departures from the diagonal line the Q–Q plot (Figure 3) and skew in the residual histogram (Figure 4) indicate deviations from normality for the residuals.

Section 4 will show that while the simple model shows a relationship between the variables, our diagnostic checks suggest potential violations of error independence, variance constancy, and normality. Overall, the diagnostics provide evidence about the level of estimated coefficient reliability and inference validity from the model (Kutner et al. 2005), and Section 5 will highlight how assumption violations impact our model inductions.

Some of our possible pitfalls and **limitations** include spatial clustering of tracts (that would violate independence), measurement error from ACS survey margins, and mismatch in denominators (education measured for adults, and poverty measured for all residents).

In future work, we 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 potential predictors of poverty by exploring more robust multiple linear regression models (Wickham 2016).

Together, these methods helped us create data cleaning, regression modeling, and visualization that can be reproducible.

4 Results

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

$\widehat{pov} = 14.255 + 0.979 \times (edu)$ with $R^2 = 0.616$ and number of tracts = 7906.

With a sample size of7906 tracts, the intercept $\beta_0 \approx 14.255$. Therefore, 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 $\beta_1 \approx 0.979$, indicating a predicted average 0.98% increase in tract poverty rate for one-percent increase in adults without a high-school diploma in that tract. The model fit $R^2 = 0.62$, so about 61.59% of the variation in the poverty rate we are measuring is explained by low-education attainment percentage within a census-tract. The estimated effect of education appears to be statistically significant ($p < 0.001$), which provides evidence of a positive association between low education attainment and higher poverty rates.

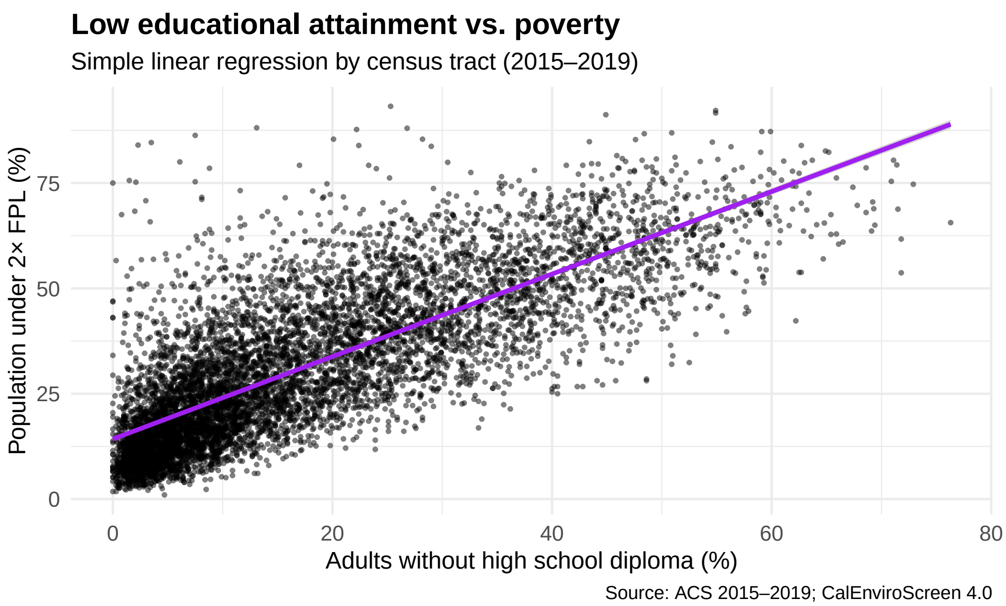

Figure 1: California census tracts (2015–2019): each +1 pp in adults without HS diploma is associated with 0.98 pp higher poverty (OLS). R^2 = 0.62.

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) vs. H_a: $\beta_1 \neq 0$. The t-test for the education coefficient yields a large test statistic ($t \approx 113$) with a p-value <2e-16 ($p < 0.001$). The p-value shows that the observed data is not more extreme than common rejection region thresholds like $< \alpha = 0.05$, we reject the null hypothesis and conclude that low educational attainment is associated with higher tract-level poverty rates.

From our model, we find that the 95% confidence interval (CI) for the education slope is approximately [0.96166, 0.99574]. This means that if we were to repeat this study many times using different random samples, about 95% of those intervals would contain the true value of the slope. In other words, we are 95% confident that for each percentage point increase in adults without a high school diploma, the poverty rate of the same census tract will increase between 0.96166 and 0.99574 percentage points.

We assess the reliability of our inferences for our primary simple regression with our diagnostic checks in Figure 1, Figure 2, Figure 4, and Figure 3. The residuals versus fitted plot shows curvature and clustering which indicates some unequal variances. The residuals histogram appears roughly bell-shaped, but its skew and the Q-Q plots's tail deviations indicate a deviation from error normality.

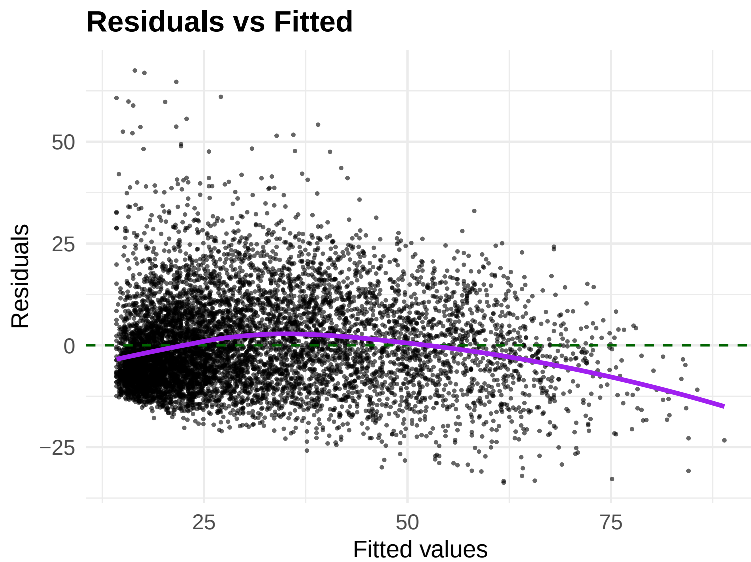

Figure 2: Residuals vs. fitted values for the simple OLS (poverty ~ education). While the scatter appears roughly centered around zero, curvature in the residuals suggests heteroskedasticity, and the clustering may reflect geographic dependence between census tracts.

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 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: The regression results (Section 4) show positive relationship between educational attainment and poverty. This model predicts that census tracts with higher percentages of adults without high school diplomas tend to experience higher rates of poverty. While unemployment is also positively correlated with poverty, its effect shows to be weaker than that

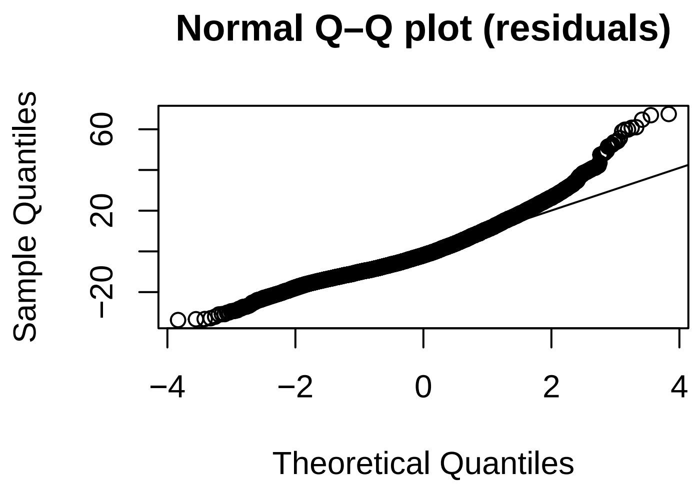

Figure 3: Normal Q–Q plot of residuals. Points near the line indicate approximate normality; curvature indicates deviations from normality.

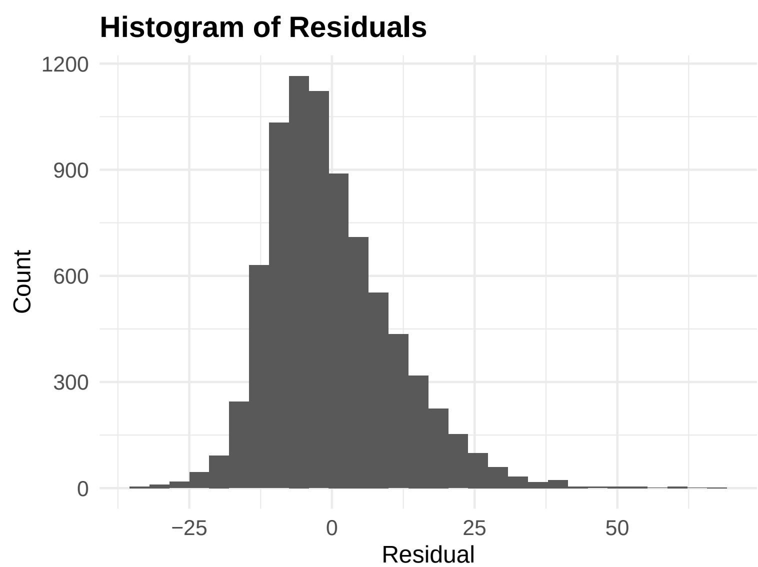

Figure 4: Histogram of residuals. A roughly bell-shaped, symmetric distribution supports the normal-errors assumption, skew indicates deviations from normality.

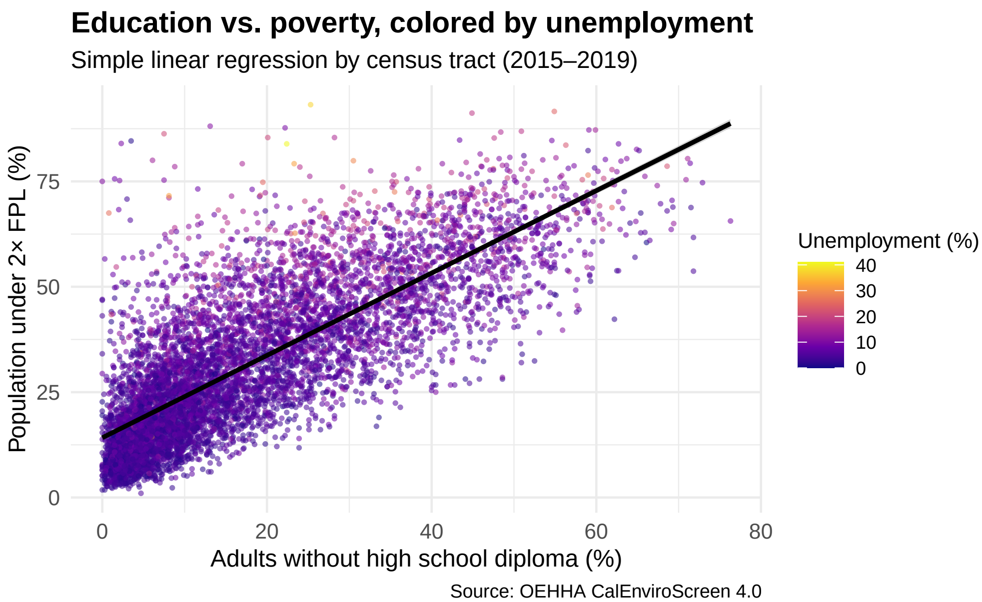

Figure 5: California census tracts (2015–2019). Poverty is higher where educational attainment is lower: each +1 % point in adults without a HS diploma is associated with 0.84 pp higher poverty (OLS), R^2 = 0.69.

of education. These findings share the importance of education as a key factor in community socioeconomic well-being.

Model assumptions: Our diagnostic plots show us several violations of the classical linear regression assumptions. Linearity appears reasonable from the Figure 1, but independence of errors is unlikely given clustering in Figure 2, and the residuals show variability and non-normality in Figure 2 and Figure 3. If the residuals were approximately normal, the black dots would fall close to the straight diagonal line of the Q-Q plot. With thousands of tracts, coefficient estimates are stable, but standard errors may be understated under OLS assumptions. However, the plot bends upwards both in the lower and upper tails. Additionally, the histogram of residuals (Figure 4) appears approximately bell-shaped and centered near zero, which supports the assumption that the errors have mean zero. However, the distribution is not perfectly symmetric: the right tail is longer than the left, and there are some extreme positive residuals. This indicates mild right skew and the presence of high-poverty tracts where the model underpredicts. Overall, the violation of these assumptions do bring us to conclude our the inference from our confidence interval with caution. However, while the residuals do not follow a perfect normal distribution, the large sample size (n = 7906) reduces concerns about inference validity due to the Central Limit Theorem. Nevertheless, the skewness suggests that robust standard errors or a variance-stabilizing transformation (e.g., square-root of the response) might provide more reliable inference in future analyses. Overall though, the non-normality and changes in variance, our 95% confidence interval should only be drawn cautiously, so we might want to consider using more robust standard errors.

Comparing education and unemployment: When we added unemployment to the model, it did not meaningfully change the estimated effect of education. The unemployment has a weaker positive association with poverty, but the education coefficient remains consequential. This result challenges a possible assumption that unemployment is a primary driver of poverty. We see from Figure 5 that a high percentage of individuals are employed but remain below twice the federal poverty threshold OEHHA (2021) set. This could indicate communities where individuals are working but still fall below the poverty threshold in our dataset. Comparatively, educational attainment shows a stronger relationship with poverty, which highlights its relevance as a relevant factor for socioeconomic well-being.

Generally speaking, the simple linear regression analysis for this research question can only draw questionable inferences, so the linear regression is likely not a full picture of the relationship between education and poverty. With that being said, we did see that the positive correlation in the linear regression model is likely statistically significant due to the p-value and R^2 value. Therefore, there is a positive relationship between low education attainment percentage and poverty rate percentage.

Implications: These findings suggest that education is a central driver of poverty across California communities. Policy interventions that expand access to high school completion programs or adult education initiatives may be effective tools for reducing poverty. Additionally, because CalEnviroScreen is used to guide environmental and equity-focused resource allocation, incorporating education as a socioeconomic vulnerability factor can help target resources more effectively, ensuring that communities facing both environmental and socioeconomic burdens receive appropriate support.

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 heteroskedasticity. 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 unadjusted 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 means that our poverty measure may overstate poverty in some rural areas and understate it in high-cost metropolitan regions, which could feasibly introduce additional variation not explained by education or unemployment in our models. Another limitation is that the education measure applies only to adults aged 25 and older, but the poverty measure covers the entire population. This mismatch means that our predictor and outcome are not measured on exactly the same group. For example, tracts with many children in poverty but relatively well-educated adults could weaken the observed association. On the other side, 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.

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

Office of Environmental Health Hazard Assessment (OEHHA). 2021. "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.

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.