Lab 2 - Describing the Relationship Between SES and Factors Associated with Poverty and Rates of Drug Overdose Related Deaths Across U.S Counties

Grace Suh, Randell Smith, Tarun Yeddula

12/9/25

Introduction

Across the United States, drug overdose mortality tells two very different stories. In some counties, overdose deaths occur at rates so high that they shape housing systems, health departments, and even local economies. In others, the issue remains comparatively distant. These disparities are not random; they map onto differences in local economic conditions, service capacity, and resource access at the county level. Understanding this variation requires looking beyond individual behavior and toward the broader material and economic conditions that shape daily life. County-level socioeconomic status (SES) can be observed in several straightforward indicators, including unemployment levels, insurance gaps, and other limits on economic security. These measures are not meant to imply hidden mechanisms but simply describe the material conditions counties operate within. While prior research has noted socioeconomic gradients in a range of health outcomes, national patterns specific to overdose mortality remain less descriptively clear. Our aim is to describe how overdose mortality varies across counties with different economic profiles using the 2020 Social Determinants of Health (SDOH) dataset.

This descriptive focus guides the rest of the analysis. Rather than attempting to explain why overdoses are concentrated where they are, we use the SDOH dataset to examine how consistently overdose mortality aligns with observable county-level economic conditions. The goal is not to resolve mechanisms or isolate SES as a singular driving force, but to provide a clearer national portrait of how overdose loss distributes itself across communities with differing access to treatment capacity, economic stability, and institutional support. By clarifying where overdose mortality is most heavily concentrated across the socioeconomic landscape, we can better understand the scope of variation that public health systems must navigate, even without making causal claims.

Data

We used the 2020 Social Determinants of Health (SDOH) County File which was compiled by the Agency for Healthcare Research and Quality (AHRQ). This dataset provides a standardized set of socioeconomic and health indicators and kpis for all U.S. counties. Using the AHRQ public repository, we imported the "Data" sheet and restricted our analysis to variables that we felt were relevant to socioeconomic status, which includes household income across 6 bands, unemployment, uninsured rates, and mental health provider availability. We converted percentage fields to numeric, removed counties missing values for either overdose mortality or poverty, and constructed a clean analytical dataset of approximately 2,200 counties. As a baseline, we produced histograms of drug overdose death rates (Figure 1) and Each Income-Share variable (Figure 2). These distributions show substantial right-skew in overdose mortality and approximately normal or mild skew distributions across income bands, supporting the use of Linear Regression Models.

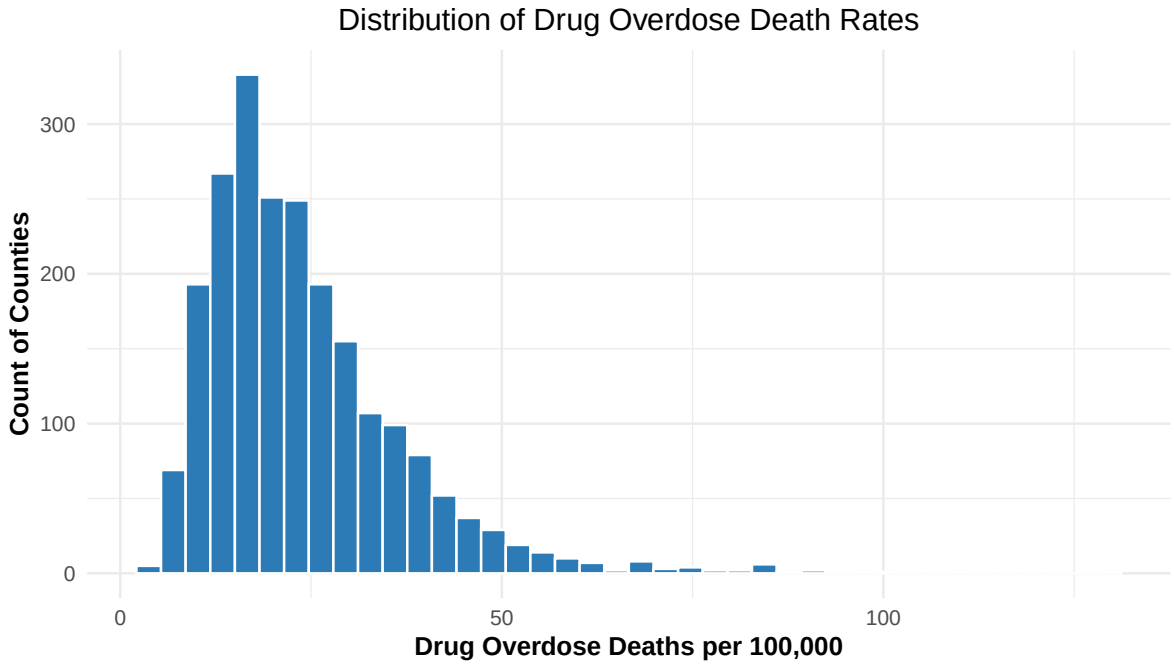

Figure 1: Histogram illustrating the distribution of drug overdose death rates per 100,000 residents across U.S. counties. The distribution is right-skewed, with most counties experiencing relatively low overdose mortality rates and a long tail representing a smaller number of counties with substantially higher rates.

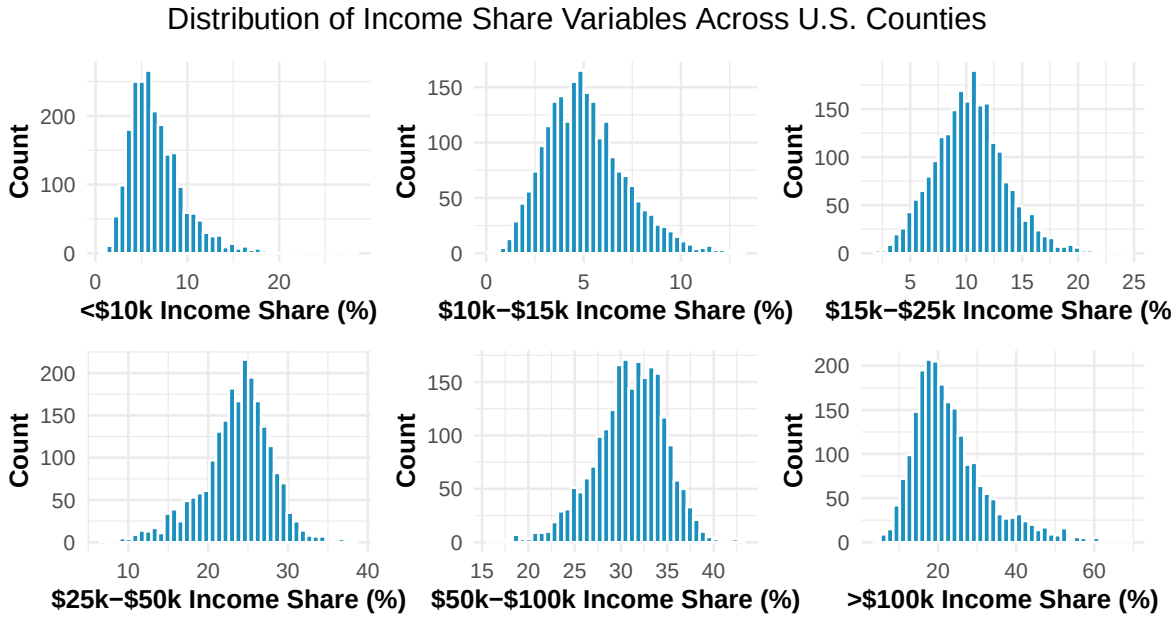

Figure 2

Figure 2: Histograms showing the distribution of income share variables across U.S. counties. Lower-income brackets (<$10k, $10k–$15k, and $15k–$25k) exhibit right-skewed distributions with most counties having relatively small shares of households in these ranges. Middle-income brackets ($25k–$50k and $50k–$100k) are more symmetric and concentrated around their respective modal values, while the highest-income bracket (>$100k) again demonstrates right skewness

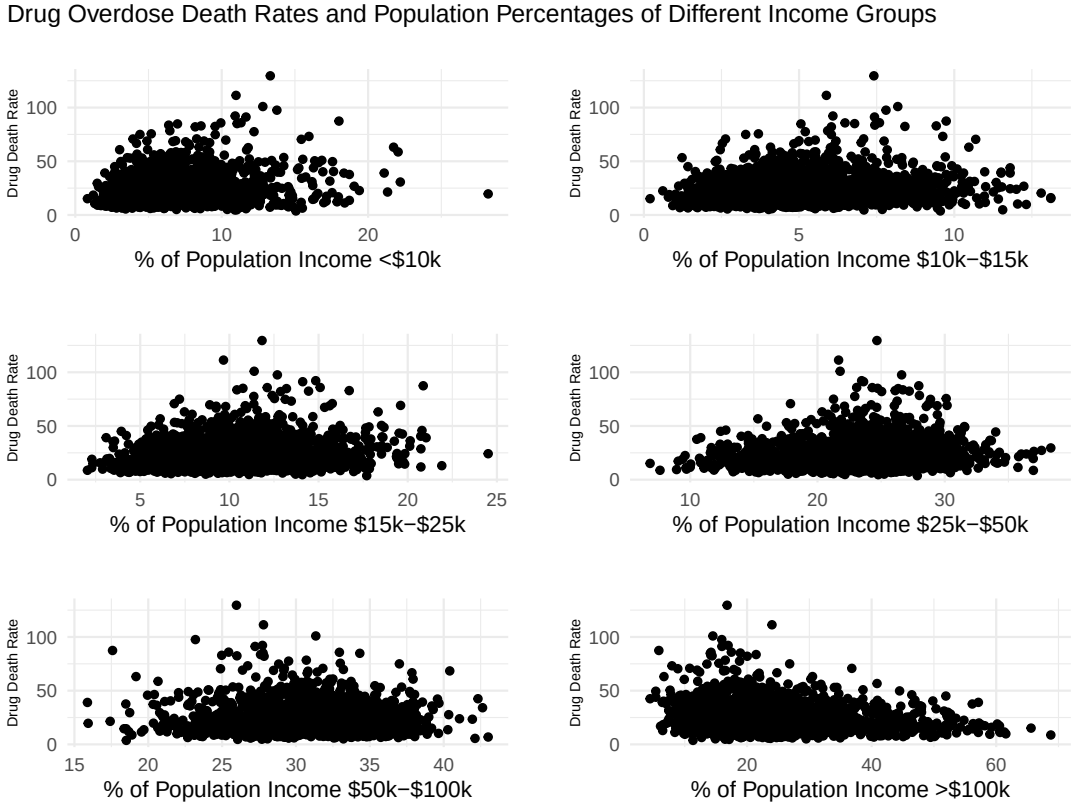

Figure 3: Scatterplots showing the relationship between drug overdose death rates and county-level income shares across six household income brackets. Overdose rates appear to be positively correlated in counties with larger proportions of lower- and middle-income households, while no clear pattern is observed for the highest-income group.

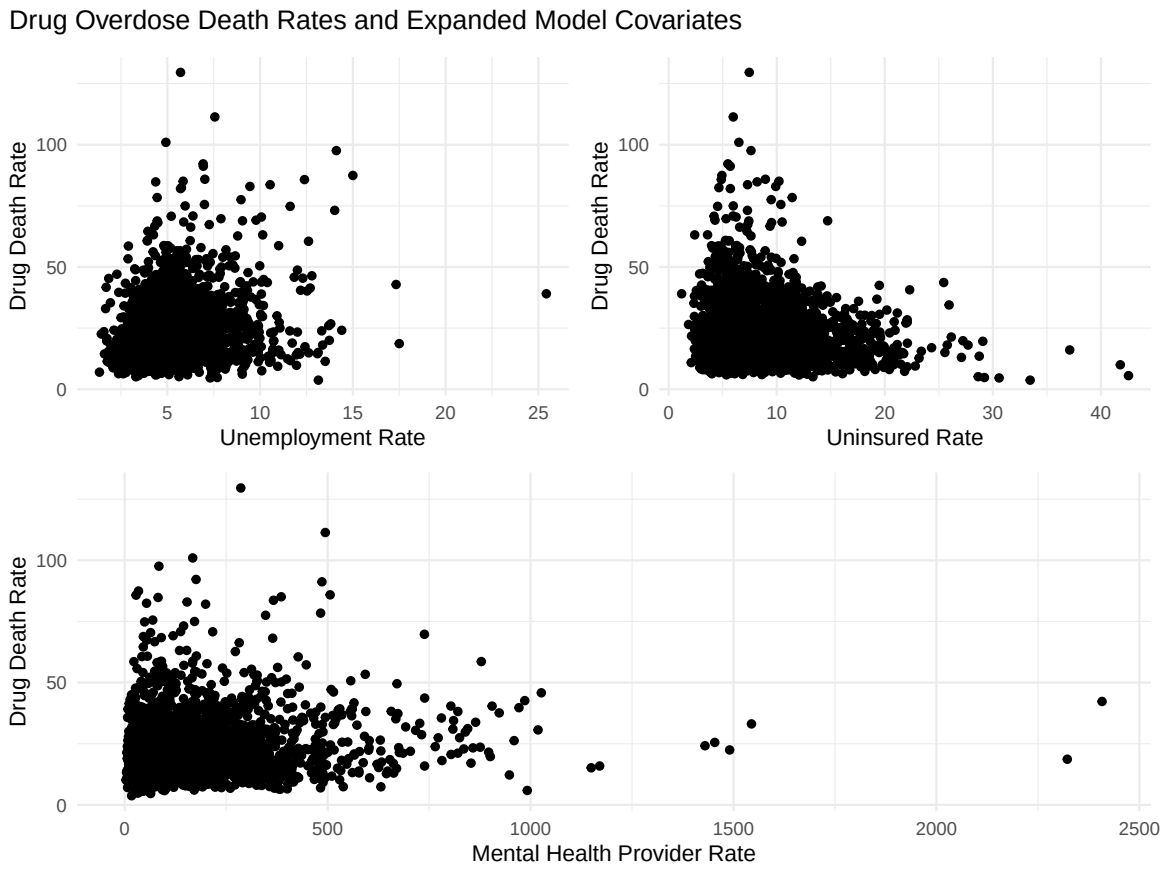

Figure 4: The relationship between drug overdose death rates and additional socioeconomic and health-related covariates. Higher overdose rates are generally observed in counties with higher unemployment and uninsured rates, while no clear pattern emerges with mental health provider availability.

Model Specification & Assumptions

The primary regression model of this study tests the strength of the relationship between a county's composition of shares of various income bands and drug overdose death rates. Since the dataset records percentages of income bands ranging from less than $10,000 to over $100,000, these are ratio variables that will yield perfect collinearity if all bands are included in the model. To further verify the optimal form of the model, simple EDA was conducted examining the distribution of percentage of household income across the six bands against drug overdose death rates to determine if variable transformations would be required. Figure 3 indicates that linear terms for all variables would be sufficient. The final primary model was as followed:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{15k,i} + \hat{\beta}_2 X_{25k,i} + \hat{\beta}_3 X_{50k,i} + \hat{\beta}_4 X_{100k,i} + \hat{\beta}_5 X_{>100k,i} + \epsilon_i$$

Such that the indicator variables (with $X_{<15k,i}$ as baseline) included are:

- $X_{15k,i}$ = Percentage of population with household income between \$10,000 and \$14,999
- $X_{25k,i}$ = Percentage of population with household income between \$15,000 and \$24,999
- $X_{50k,i}$ = Percentage of population with household income between \$25,000 and \$49,999
- $X_{100k,i}$ = Percentage of population with household income between \$50,000 and \$99,999
- $X_{>100k,i}$ = Percentage of population with household income greater than \$100,00

This model performed satisfactorily in terms of it's ability to explain the variance of drug overdose death rates (R^2=0.033) since income is only a narrow facet of risk factors associated with drug use. The model suggests that counties with higher proportions of households with incomes above \$10,000 experience lower drug overdose death rates. Our second model sought to include more variables associated with SES and that could further describe drug overdose death rate variance. The expanded model was as followed:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{15k,i} + \hat{\beta}_2 X_{25k,i} + \hat{\beta}_3 X_{50k,i} + \hat{\beta}_4 X_{100k,i} + \hat{\beta}_5 X_{>100k,i} + \hat{\beta}_6 X_{Unemploy,i} + \hat{\beta}_7 X_{Unins,i} + \hat{\beta}_8 X_{Prov,i} + \epsilon_i$$

- $X_{Unemploy,i}$ = Percentage of labor force that is unemployed
- $X_{Unins,i}$ = Percentage of population with no health insurance coverage
- $X_{Prov,i}$ = Mental health care providers per 100,000 population

The these variables were not transformed for similar reasons mentioned for the primary model as the distribution of all three against the dependent variable suggested linear terms were sufficient (Figure 4). In comparison to the base, the expanded model performed better as it yielded a more favorable R^2=0.133.

In order for both models to be valid, data must be IID. Independence is not upheld due to strong geographic clustering as the data set includes all counties within the U.S which counties within the same state or region will often share similar environments (i.e. political, social, and economic) that will bias our variables. The lack of independence will reduce our ability to extract information from the county-level due to geographic clustering. In terms of identical distribution, this is upheld since each county comes from the same population distribution of being within the U.S. While IID is not upheld due to the violation of independence, our model can still provide utility when paired with robust standard errors to enable us to make

inferences. In terms of unique BLP, this assumption is upheld as our regression results indicate the lack of perfect collinearity and because all of our variables have finite means and variance. Our variables are population percentages which must be bounded between 0-100 and cannot grow infinitely in variance with unbounded support.

Results

In the base model, the coefficients for all income bands except for 15k-25k were found to be statistically significant (Table 1). However, once the model accounted for unemployment, insurance levels, and mental health provider access rates, the coefficient for 10k-15k is no longer significant while the three additional predictors and income bands ranging from 25k to +100k remained significant. The base model experienced a form of omitted variables bias where the low income band appears significant since it is likely highly correlated with poorer outcomes such as higher unemployment and lower insurance coverage which will more directly impact drug overdose rates. Once the expanded model accounts for these omitted variables and decouples it from the 10-15k variable, this income band loses explanatory power.

The regression results of the expanded model indicate that a 10% increase in county shares of 25k-50k, 50k-100k, and >100k leads to 3.6, 3.6, and 5 fewer deaths per 100,000 respectively holding all other variables constant (Table 1). For every percent increase in unemployment, it is associated with increase in deaths by 1 while, every percent increase in uninsured rates lead to 0.7 decrease in deaths. This negative association may reflect structural differences in healthcare access where counties with higher uninsured populations may have fewer medical providers and lower exposure to opioids. Lastly, although the coefficient for mental health provider density is statistically significant, the effect size is negligible. A 100-provider increase per 100,000 residents is associated with only a 0.5-point change in overdose deaths, a magnitude unlikely to be meaningful in practical terms.

Conclusion

Across U.S. counties, overdose mortality aligns with meaningful, if modest, socioeconomic patterns. In both the base and expanded models, counties with larger shares of higher-income households consistently reported lower overdose death rates, while those facing greater economic constraint showed elevated mortality levels. These results do not indicate a singular pathway or fully account for the drivers of overdose risk, but they do clarify how uneven overdose loss is distributed across the national socioeconomic landscape. The expanded model, which incorporated unemployment, uninsured rates, and mental health provider capacity, strengthened explanatory power relative to the income-only specification, suggesting that county-level institutional and economic supports matter alongside income composition itself. As noted in the model discussion, overall explanatory strength remained limited, which is expected given that overdose vulnerability extends beyond economic structure alone. Geographic clustering

Table 1: Regression Models of Drug Overdose Death Rates per 100,000 Population

	Dependent variable:	
	Drug Overdose Death Rate	
	Base Model	Expanded Model
	(1)	(2)
Household Income Percentage: 10k–15k	−0.644**	−0.563**
	(0.297)	(0.285)
Household Income Percentage: 15k–25k	−0.362	−0.265
	(0.224)	(0.215)
Household Income Percentage: 25k–50k	−0.676***	−0.393**
	(0.168)	(0.164)
Household Income Percentage: 50k–100k	−0.533***	−0.325**
	(0.135)	(0.137)
Household Income Percentage: >100k	−0.647***	−0.532***
	(0.141)	(0.139)
Unemployment Rate (%)		1.176***
		(0.153)
Uninsured Rate (%)		−0.795***
		(0.061)
Mental Health Providers per 100,000		0.004***
		(0.001)
Constant	78.633***	61.341***
	(13.200)	(13.328)
Observations	2,202	2,186
R^2	0.033	0.130
Adjusted R^2	0.031	0.127
Residual Std. Error	12.853 (df = 2196)	12.168 (df = 2177)
F Statistic	15.095*** (df = 5; 2196)	40.747*** (df = 8; 2177)
Note:		*p<0.1; **p<0.05; ***p<0.01

also indicates shared regional environments, underscoring that counties do not experience these conditions in isolation.

Taken together, the models offer a clearer descriptive portrait rather than a definitive explanation: overdose burden concentrates more heavily in counties with fewer economic and institutional resources, and more lightly where those supports are stronger. Even without establishing causality, documenting these disparities allows public health work to begin with a grounded sense of where the burden is highest and what structural contexts accompany it. In this sense, the analysis contributes a map that traces how overdose mortality distributes itself across widely different local capacities, and clarifies the uneven terrain that national response efforts must navigate.

Appendix

Data Source

Social Determinants of Health Database by Agency for Healthcare Research and Quality - 2020 County Level Dataset

Source Code

Lab 2 Markdown

Residuals-vs-Fitted-Values Plot

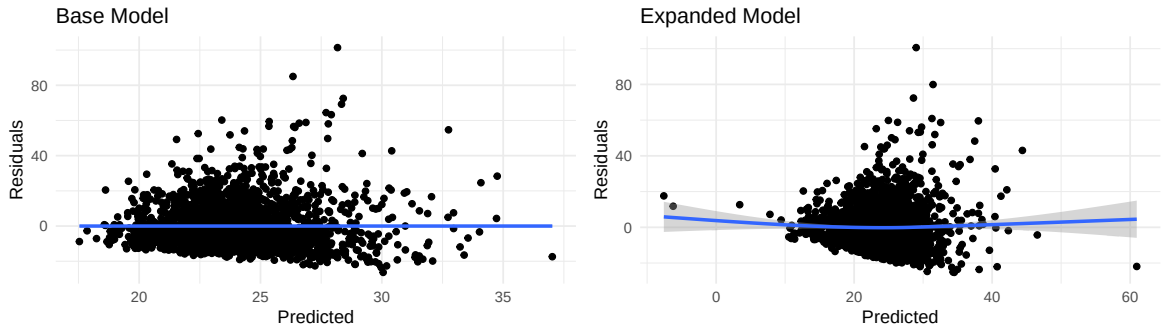

Figure 5: Residual plots for the base and expanded models. The expanded model shows a slight curvature in residuals across predicted values but is largely captured by the model.