Final Project: Poverty and County-Level Demographic, Socioeconomic, and Geographic Estimates

Soumyajit De

Department of Communication, University of California, Santa Barbara

PSTAT 220A Advanced Statistical Methods

Instructor: Dr. Alexander Franks

December 8, 2025

Final Project: Poverty and County-Level Demographic, Socioeconomic, and Geographic Estimates

In this project, I analyze county-level patterns of poverty in the United States using a cross-sectional dataset of 3,141 counties drawn from all 50 states. Each observation corresponds to a single county and includes the percentage of residents living below the poverty line, overall population size, and a range of demographic, socioeconomic, and geographic characteristics. Poverty is defined as the percentage of residents whose income falls below the official poverty threshold for a given year. It serves as the primary outcome variable for all analyses. The first research question examines how poverty is related to this broad set of county-level characteristics. Specifically, I ask:

RQ1. How is county-level poverty (percentage of individuals under the poverty level) associated with county demographic, geographic, and socioeconomic characteristics? Which of these variables appear to be most strongly associated with poverty, and which (if any) show little or no predictive value?

In addressing this, I group predictor variables into conceptual blocks (e.g., gender, race/ethnicity, employment sectors, transportation to work, type of work, economic conditions, education, age, and geography) and examine their joint associations with county poverty rates. The second research question narrows the focus to gender composition and state context:

RQ2. To what extent is county-level poverty associated with the gender composition of the population in California and Texas, and does this association differ?

In the following sections, I describe the dataset, key variables, and preprocessing decisions; present exploratory data analyses to summarize the distribution of poverty and its associations with the main predictors; fit multiple linear regression models to address RQ1 and RQ2 using robust standard errors and model diagnostics; apply a multiple-testing correction to account for the number of predictors before summarizing the main patterns of association, and and discuss their implications and limitations.

Methods

Data and Variables

For the analyses, I used the `poverty_data.csv` dataset provided on Canvas. It contains 3,141 cross sectional observations, each corresponding to a single U.S. county. The primary outcome variable is Poverty, which I define as the percentage of county residents whose income falls below the official poverty threshold for a given year. This variable is expressed on a 0–100 scale and serves as the dependent variable for both research questions. All remaining variables are treated as predictors.

Location and geography consists of State, County, and county_fips, which identify each county and allow subsetting for RQ2. Long and Lat provide county-level longitude and latitude that are used to visualize geographic patterns. The variable TotalPop represents the total number of residents in each county. Gender is described by the variables Men and Women, which are later combined into a percentage of women (propWomen). Race and ethnicity are represented by variables Hispanic, White, Black, Native, Asian, and Pacific in per cent. Finally, the AvgAge represents the mean age of residents.

Economic and labor-market conditions are represented by several blocks of variables. Unemployment variable represents the percentage of the labor force that is unemployed. Educational block constitutes variables that are summarized by the percentages of adults with less than a high school diploma (LessThanHighSchool), a high school diploma (HighSchoolDiploma), some college or an associate degree (SomeCollegeOrAssociateDegree), and a bachelor's degree or higher (BachelorDegreeOrHigher). Employment sectors variables include the percentages of employed residents working in Professional, Service, Office, Construction, and Production jobs. The type of employer block includes variables that are the percentages of involvement in private industry (PrivateWork), public jobs (PublicWork), self-employment (SelfEmployed), and unpaid family work (FamilyWork). Transportation to work is described by variables representing the percentages of workers who Drive, Carpool, use Transit, Walk, use other means (OtherTransp), or work from home (WorkAtHome).

MeanCommute variable represents the mean commute time in minutes. Together, these blocks provide a structured description of county-level context for the exploratory analyses and regression models that follow.

Preprocessing and Missing Data

Before conducting exploratory analyses and fitting regression models, I carried out basic preprocessing. All variables defined as percentages were retained on their original 0–100 scale so that a 1-unit change corresponds to a 1–percentage-point change. To summarize gender composition, I constructed a single predictor,

$$\text{propWomen}_i = 100 \times \frac{\text{Women}_i}{\text{TotalPop}_i},$$

where Women_i and TotalPop_i denote the number of women and total population in county i. The percentage of men is then implied as $100 - \text{propWomen}_i$ and is not included separately in the models, avoiding colinearity.

Several other predictor blocks also form compositions whose components sum to approximately 100%: race/ethnicity (Hispanic, White, Black, Native, Asian, Pacific), employment sectors (Professional, Service, Office, Construction, Production), commuting modes (Drive, Carpool, Transit, Walk, OtherTransp, WorkAtHome), type of work or employer (PrivateWork, PublicWork, SelfEmployed, FamilyWork), and educational attainment (LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher). For counties with complete information, I checked the overall composition and they they satisfy the following constraint:

$$\text{Hispanic}_i + \text{White}_i + \text{Black}_i + \text{Native}_i + \text{Asian}_i + \text{Pacific}_i \approx 100.$$

To avoid perfect collinearity in models, I designated one category in each block as a reference and omitted it from the design matrix: White for race/ethnicity, Professional for employment sectors, Drive for transportation modes, PrivateWork for type of employer, and LessThanHighSchool for educational attainment. Coefficients for the included categories are

therefore interpreted as changes in poverty (in percentage points) associated with a 1–percentage-point increase in that category, holding other predictors and the implied reference category constant.

I examined missingness in the data and found that each of the four education variables had eight missing values, corresponding to eight counties with no recorded information on educational attainment. Furthermore, AvgAge had five missing values, and the geographic identifiers Long, Lat, and county–fips each had two missing values. For exploratory plots, I included all available counties, including those with missing data, so that the EDA reflects the full dataset. For regression models, I excluded those 12 counties with missing data and used complete-case analysis with respect to the outcome and the predictors included in each model. Finally, I verified that all percentage variables in the retained modeling dataset fall within the valid range of 0 to 100%; no negative or greater-than-100 values were observed. A brief table listing the eight counties with missing education data and their patterns of missingness is provided in Appendix A.

Exploratory Data Analysis

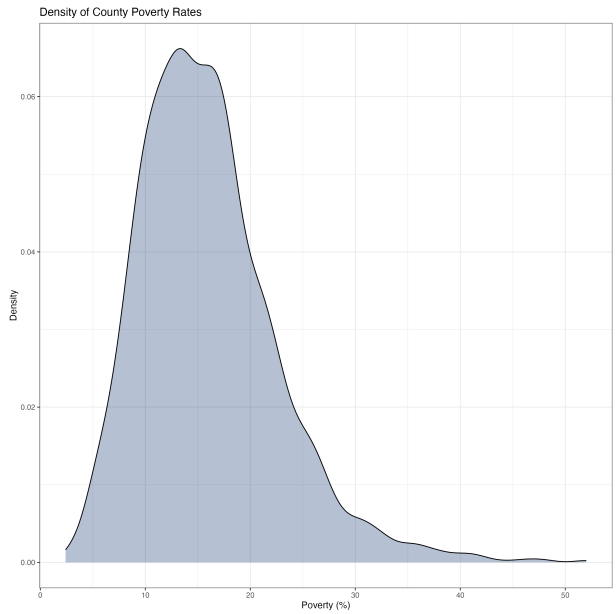

Figure 1

Density plot of county poverty rates (Poverty, in percent).

I first examined the distribution of county poverty rates. The distribution of Poverty is moderately right-skewed: most counties fall between roughly 10% and 25% poverty, with a long right tail including a small number of counties above 40%. This is shown in the density plot in Figure 1. Boxplots of poverty by state (Appendix B, Figure B1) indicate substantial between-state differences in both central tendency and spread, with some states concentrated at relatively low poverty and others centered at much higher levels.

Next, I explored economic and education variables. Counties with higher unemployment rates tend to have substantially higher poverty, and the smoothed curve in Figure 2 is clearly upward-sloping. Educational block showed a similarly strong gradient. Counties with higher percentages of adults holding a bachelor's degree or higher almost always have lower poverty, whereas counties with higher percentages of adults with less than a high school diploma have markedly higher poverty. The intermediate categories (high school diploma and some college or an associate degree) show weaker and more curved relationships as illustrated in Figure 3. Average age and mean commute time have only modest and slightly nonlinear associations with poverty (Appendix B, Figure B2–B3). Overall, these results suggest that unemployment and education composition are core predictors for RQ1, with age and commute time playing more secondary roles.

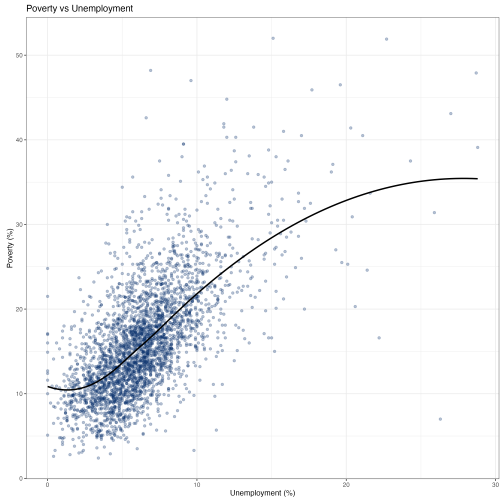

Figure 2

County poverty rate versus unemployment percentage.

I then examined demographic and work-related variables. The derived percentage of women, propWomen, is concentrated near 50% in most counties and has a very weak bivariate association with poverty: the fitted curve is nearly flat, and the point cloud is dense and vertically oriented (Appendix B, Figure B4). When looking into California and Texas, separate fitted lines for each state as indicated in Figure 4, remain shallow and close to horizontal, suggesting that any state-specific differences in the propWomen–poverty association are likely small. In contrast, racial and ethnic composition shows much clearer patterns (Appendix B, Figure B5). Employment sector, work-type, and commuting variables show similar gradients (Appendix B, Figures B6–B7). Together, these exploratory results motivate including unemployment, education, location, race/ethnicity, and selected employment and work-type variables as primary predictors for RQ1, while retaining propWomen to address RQ2 explicitly.

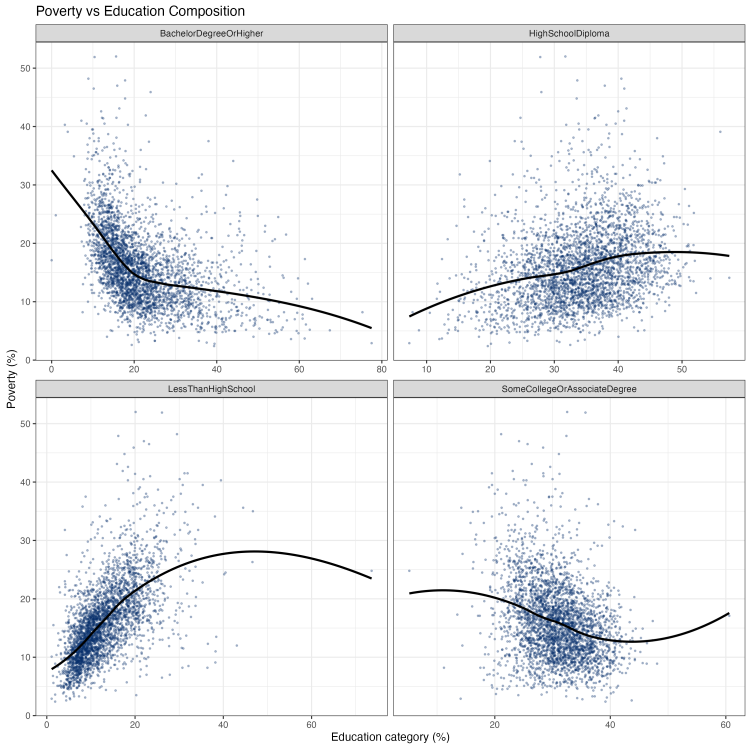

Figure 3

County poverty rate versus percentages in each education category: less than high school, high school diploma, some college or associate degree, and bachelor's degree or higher.

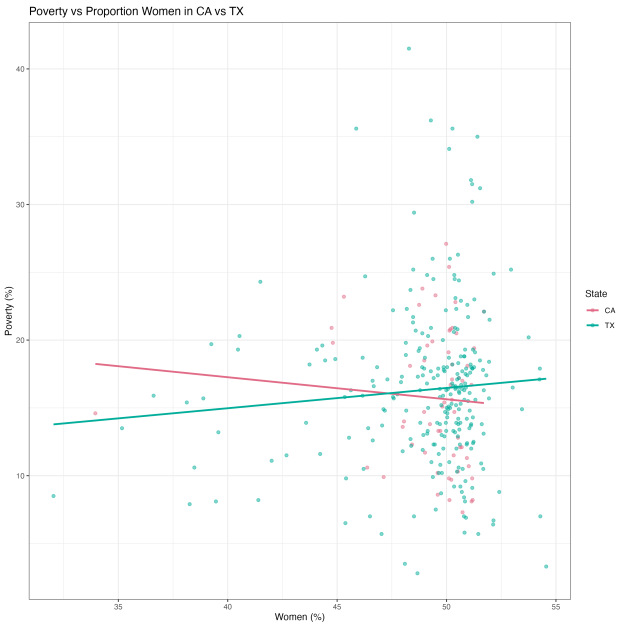

Figure 4

County poverty rate versus percentage women propWomen in California and Texas, with separate fitted lines by state.

Modeling Strategy

To address RQ1, I modeled county poverty as a function of the main demographic, geographic, and socioeconomic blocks described above. Let Poverty_i denote the poverty rate (in percent) for county i. The primary model is a multiple linear regression:

$$\text{Poverty}_i = \beta_0 + \beta_1 \text{propWomen}_i + \beta_2 \text{Unemployment}_i + \mathbf{x}_i^\top + \varepsilon_i,$$

where propWomen_i is the percentage of women, Unemployment_i is the county unemployment rate, and $\mathbf{x}_i$ represents the remaining predictors from the race/ethnicity, education, employment sector, commuting mode, work-type, age, commute-time, and geographic blocks. Within some compositional blocks, one category is omitted as a reference (White, Professional, Drive, PrivateWork, LessThanHighSchool), so coefficients for the included categories can be interpreted as changes in poverty associated with a 1–percentage-point increase in that category, holding other predictors constant. The exclusion also helps to avoid multicolinearity.

For RQ2, I restricted the data to counties in California and Texas and examined whether the association between gender composition and poverty differs by state. State was dummy coded with California as the reference category ($\text{StateTX}_i = 1$ for Texas, 0 for California), so that the StateTX coefficient represents how average poverty in Texas differs from California. I also created a mean-centered version of gender composition, $\text{propWomen}_{c,i} = \text{propWomen}_i - \overline{\text{propWomen}}$, so that $\text{propWomen}_{c,i} = 0$ corresponds to the average percentage of women across CA and TX. This makes the intercept and the StateTX coefficient easier to interpret because they refer to a typical county rather than to the unrealistic case of 0% women.

$$\text{Poverty}_i = \beta_0 + \beta_1 \text{propWomen}_{c,i} + \beta_2 \text{StateTX}_i + \beta_3 (\text{propWomen}_{c,i} \times \text{StateTX}_i) + \mathbf{z}_i^\top + \varepsilon_i,$$

Here, $\mathbf{z}_i$ represents a reduced set of control variables that include unemployment, the education block (HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher), and race/ethnicity (Hispanic, Black, Native, Asian, and Pacific). I included them because they seemed to have a strong association with poverty in EDA and can be plausible confounders. Hence, adjusting for them may help to isolate the association of propWomen_c and its interaction with StateTX.

In the model, β_0 is the mean poverty rate for a California county at the average percentage of women, β_1 is the slope relating gender composition to poverty in California, β_2 is the difference in baseline poverty between Texas and California at the average proportion of women, and β_3 is the difference in slopes between Texas and California. The simple slope for Texas is therefore $\beta_1 + \beta_3$, so in the results I report both the regression coefficients (propWomen_c, StateTX, and their interaction) and derived simple slopes for California (slope $= \beta_1$) and Texas (slope $= \beta_1 + \beta_3$), each with heteroskedasticity-robust confidence intervals.

For both models, I used coefficients from ordinary least squares (OLS) and reported as unstandardized percentage-point units. I found heteroskedasticity and heavy-tailed residuals while testing assumptions and diagnosis, which I will discuss in the next section. Therefore, I based all hypothesis tests and confidence intervals on heteroskedasticity-robust

(HC3; MacKinnon & White, 1985) standard errors. For RQ1, I treated the predictors as a single family and applied the Benjamini–Hochberg false discovery rate procedure (Benjamini & Hochberg, 1995) at $q = .05$ to the robust p-values. For RQ2, I focused on a small number of pre-specified coefficients (the main effect of gender composition and its interaction with state) and therefore report robust p-values without additional multiple-testing correction.

Assumptions and Diagnosis

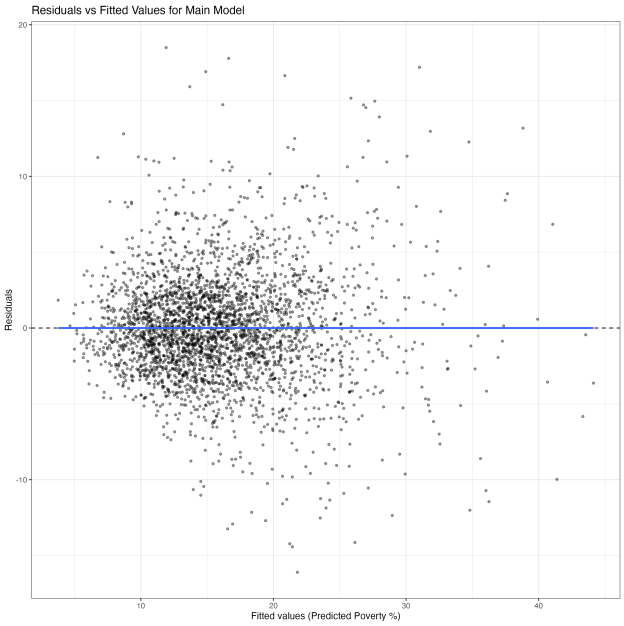

Figure 5

Residuals versus fitted values for the RQ1 model.

For the main RQ1 model, I examined linearity, homoskedasticity, normality, independence, outliers, and multicollinearity using standard regression diagnostics. The residuals versus fitted values plot (Figure 5) shows residuals centered around zero across the fitted range, with no strong curvature, suggesting that the model is fairly linear. However, the vertical spread of residuals increases somewhat for higher fitted poverty values, producing a mild fan shape. This indicated heteroskedasticity, and due to this, I based all hypothesis tests and confidence intervals on heteroskedasticity robust (HC3; MacKinnon & White, 1985) standard errors for both RQ1 and RQ2.

I assessed normality of residuals using a histogram (Figure 6) and a normal Q–Q plot (Appendix C, Figure C1). The histogram is roughly symmetric and centered at zero, but the tails seem heavier than a normal distribution. Given the large sample size and the use of robust standard errors, I did not perform any transformations.

Additional diagnostics are reported in Appendix C. Residual maps and residuals by state (Appendix C, Figures C2–C3) show only mild spatial clustering and no large regions of systematic over or underprediction. Plots of standardized residuals versus leverage and Cook's distance (Appendix C, Figures C4–C5) indicate that only a small number of observations have large residuals or moderate leverage, and no single county has an undue influence on the fit. Variance inflation factors (VIF) (Appendix C, Table C1) are mostly in the low to moderate range, with higher values confined to the education block. This makes sense because if one education category goes up then others should come down. Overall, these diagnostics support using linear models on the original poverty scale combined with robust standard errors and cautious interpretation of tail behavior and spatial dependence.

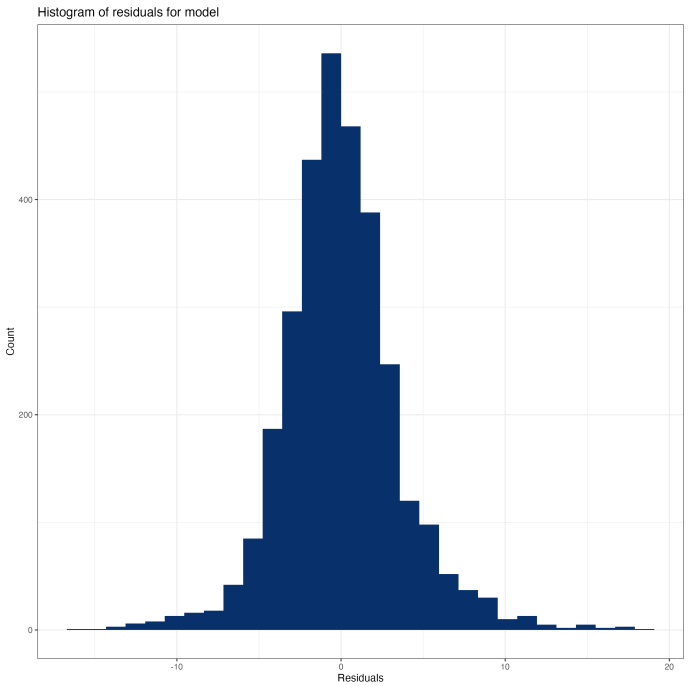

Figure 6

Histogram of residuals for the main RQ1 regression model.

Multiple Testing

Because the main RQ1 model includes a large set of predictors, relying only on raw p-values would increase the chance of false positives. To control false discoveries while retaining reasonable power, I treated the main RQ1 predictors as a single family of hypotheses and applied the Benjamini–Hochberg false discovery rate (FDR) procedure at $q = .05$ (Benjamini & Hochberg, 1995). Let $p_{(1)} \leq \cdots \leq p_{(m)}$ denote the ordered p-values for the m substantive coefficients. The Benjamini–Hochberg procedure finds the largest index k such that

$$p_{(k)} \leq \frac{k}{m} q,$$

and declares all hypotheses with $p_{(j)} \leq p_{(k)}$ as FDR-significant. In the RQ1 results table (Table 1), I report robust p-values for all coefficients and indicate which predictors remain significant after FDR control at $q = .05$. My inference about strong and weak predictors is based on this FDR-adjusted p-values. For RQ2, I examined only a small number of variables, so I did not apply a multiple testing correction and interpreted the robust p-values directly.

Results

RQ1: Associations Between Poverty and County Characteristics

Table 1 reports the multiple regression model for RQ1, with Poverty as the outcome and the full set of demographic, socioeconomic, and geographic predictors. Coefficients are estimated by OLS with HC3 robust standard errors; for inference on the main predictors, I have used Benjamini–Hochberg adjusted p-values.

Unemployment and the educational blocks are the strongest predictors of Poverty. Higher Unemployment is very strongly associated with higher Poverty, whereas longer MeanCommute is associated with lower Poverty. Relative to LessThanHighSchool, higher percentages with HighSchoolDiploma, SomeCollegeOrAssociateDegree, and BachelorDegreeOrHigher are each strongly associated with lower Poverty, with SomeCollegeOrAssociateDegree showing the largest and most negative coefficient in this block.

Table 1

RQ1: County Poverty Regressed on Demographic, Geographic, and Socioeconomic Predictors

RQ1: County Poverty Regressed on Demographic, Geographic, and Socioeconomic Predictors						
Predictor	Estimate	SE (robust)	t	p (raw)	p (BH-adjusted)	95% CI
TotalPop	0.000	0.000	1.000	0.317	0.357	[-0.00, 0.00]
propWomen	0.384	0.041	9.267	< .001	< .001	[0.30, 0.47]
Hispanic	−0.080	0.012	−6.909	< .001	< .001	[-0.10, -0.06]
Black	0.053	0.009	6.211	< .001	< .001	[0.04, 0.07]
Native	0.043	0.020	2.189	0.029	0.046	[0.00, 0.08]
Asian	−0.083	0.032	−2.629	0.009	0.015	[-0.14, -0.02]
Pacific	−0.294	0.152	−1.939	0.053	0.071	[-0.59, 0.00]
Service	0.311	0.034	9.118	< .001	< .001	[0.24, 0.38]
Office	0.031	0.043	0.720	0.472	0.490	[-0.05, 0.12]
Construction	−0.049	0.040	−1.221	0.222	0.261	[-0.13, 0.03]
Production	0.053	0.032	1.666	0.096	0.123	[-0.01, 0.12]
Carpool	−0.012	0.037	−0.325	0.745	0.745	[-0.08, 0.06]
Transit	0.107	0.025	4.330	< .001	< .001	[0.06, 0.16]
Walk	0.041	0.044	0.940	0.347	0.375	[-0.04, 0.13]
OtherTransp	−0.205	0.095	−2.156	0.031	0.047	[-0.39, -0.02]
WorkAtHome	0.097	0.047	2.060	0.039	0.056	[0.00, 0.19]
MeanCommute	−0.180	0.018	−9.867	< .001	< .001	[-0.22, -0.14]
PublicWork	0.130	0.022	5.806	< .001	< .001	[0.09, 0.17]
SelfEmployed	0.117	0.040	2.894	0.004	0.007	[0.04, 0.20]
FamilyWork	0.310	0.250	1.241	0.215	0.261	[-0.18, 0.80]
Unemployment	0.657	0.049	13.399	< .001	< .001	[0.56, 0.75]
HighSchoolDiploma	−0.310	0.037	−8.343	< .001	< .001	[-0.38, -0.24]
SomeCollegeOrAssociateDegree	−0.471	0.032	−14.808	< .001	< .001	[-0.53, -0.41]
BachelorDegreeOrHigher	−0.418	0.035	−12.093	< .001	< .001	[-0.49, -0.35]
AvgAge	−0.285	0.040	−7.089	< .001	< .001	[-0.36, -0.21]
Long	−0.031	0.009	−3.644	< .001	< .001	[-0.05, -0.01]
Lat	−0.142	0.022	−6.529	< .001	< .001	[-0.18, -0.10]

Race and ethnicity also show important patterns. Relative to White, higher percentages of Black and Native residents are associated with higher Poverty, while higher Hispanic and Asian percentages are associated with lower Poverty. Pacific has a relatively large negative point estimate but does not remain significant after FDR adjustment. In the employment-sector block with Professional as the reference, Service is a strong positive predictor of Poverty, whereas Office, Construction, and Production have smaller coefficients are not statistically significant. For type of employer with PrivateWork as reference, PublicWork and SelfEmployed are moderately and significantly positively associated with Poverty, while FamilyWork has the largest positive estimate but is not FDR significant.

The contribution of transportation and remaining demographic/geographic variables remain strong to modest. Relative to Drive, greater use of Transit is positively associated with Poverty, whereas OtherTransp is negatively associated with Poverty; Carpool, Walk, and WorkAtHome have small and insignificant effects. The gender composition variable propWomen is a strong positive predictor, indicating that counties with a higher percentage of women tend to have higher Poverty, even after adjustment. AvgAge is also a strong negative predictor, with older counties having lower Poverty. Finally, the geographic coordinates Long and Lat both have small but significant negative associations with Poverty, while TotalPop has an essentially zero and nonsignificant coefficient, making it the weakest predictor overall.

RQ2: Gender Composition in California vs. Texas

RQ2 tests whether the association between gender composition (propWomen) and Poverty differs between California and Texas. As described before, this model uses only counties in CA and TX, codes StateTX as 0 for California and 1 for Texas, and relies on a mean-centered percentage of women (propWomen_c). The regression includes propWomen_c, StateTX, their interaction propWomen_c:StateTX, and controls for Unemployment, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher, Hispanic, Black, Native, Asian, and Pacific, with HC3 robust

standard errors. Table 2 presents the raw interaction model.

Table 2

County Poverty Regressed on Gender Composition, State, and Controls

County Poverty Regressed on Gender Composition, State, and Controls					
Predictor	Estimate	SE (robust)	t	p	95% CI
(Intercept)	30.716	6.759	4.544	< .001	[17.41, 44.02]
propWomen_c	0.355	0.216	1.645	0.101	[-0.07, 0.78]
StateTX	−1.318	1.131	−1.165	0.245	[-3.54, 0.91]
Unemployment	0.434	0.191	2.275	0.024	[0.06, 0.81]
HighSchoolDiploma	−0.161	0.086	−1.872	0.062	[-0.33, 0.01]
SomeCollegeOrAssociateDegree	−0.262	0.080	−3.251	0.001	[-0.42, -0.10]
BachelorDegreeOrHigher	−0.307	0.072	−4.277	< .001	[-0.45, -0.17]
Hispanic	0.055	0.022	2.453	0.015	[0.01, 0.10]
Black	0.190	0.049	3.894	< .001	[0.09, 0.29]
Native	0.402	0.272	1.481	0.140	[-0.13, 0.94]
Asian	−0.083	0.077	−1.082	0.280	[-0.23, 0.07]
Pacific	0.124	2.117	0.058	0.953	[-4.04, 4.29]
propWomen_c:StateTX	0.090	0.232	0.388	0.698	[-0.37, 0.55]

Here, the coefficient on propWomen_c gives the slope relating gender composition to Poverty in California (the reference state) at the average percentage of women; the StateTX coefficient gives the difference in baseline Poverty between Texas and California at that average; and the propWomen_c:StateTX coefficient gives the difference between the Texas and California slopes ($\text{slope}_{\text{TX}} - \text{slope}_{\text{CA}}$). In Table 2, none of these three coefficients is statistically significant.

To make the state-specific slopes explicit, I computed simple slopes for California and

Texas. Let β_1 denote the coefficient on propWomen_c and β_3 the coefficient on propWomen_c:StateTX. Then:

$$\text{slope}_{\text{CA}} = \beta_1, \qquad \text{slope}_{\text{TX}} = \beta_1 + \beta_3.$$

The variance of the Texas slope is:

$$\text{Var}(\text{slope}_{\text{TX}}) = \text{Var}(\beta_1) + \text{Var}(\beta_3) + 2\,\text{Cov}(\beta_1, \beta_3),$$

so that

$$\text{SE}(\text{slope}_{\text{TX}}) = \sqrt{\text{Var}(\text{slope}_{\text{TX}})}, \quad t = \frac{\text{slope}_{\text{TX}}}{\text{SE}(\text{slope}_{\text{TX}})}, \quad \text{CI}_{95\%} = \text{slope}_{\text{TX}} \pm t_{.975}\,\text{SE}(\text{slope}_{\text{TX}}).$$

Table 3

RQ2: Simple Slopes and Interaction of Gender Composition by State and Controls

RQ2: Simple Slopes and Interaction of Gender Composition by State and Controls					
Predictor	Estimate	SE (robust)	t	p	95% CI
CA	0.355	0.216	1.645	0.101	[-0.07, 0.78]
TX	0.445	0.108	4.107	< .001	[0.23, 0.66]
CA:TX	0.090	0.232	0.388	0.698	[-0.37, 0.55]
Unemployment	0.434	0.191	2.275	0.024	[0.06, 0.81]
HighSchoolDiploma	−0.161	0.086	−1.872	0.062	[-0.33, 0.01]
SomeCollegeOrAssociateDegree	−0.262	0.080	−3.251	0.001	[-0.42, -0.10]
BachelorDegreeOrHigher	−0.307	0.072	−4.277	< .001	[-0.45, -0.17]
Hispanic	0.055	0.022	2.453	0.015	[0.01, 0.10]
Black	0.190	0.049	3.894	< .001	[0.09, 0.29]
Native	0.402	0.272	1.481	0.140	[-0.13, 0.94]
Asian	−0.083	0.077	−1.082	0.280	[-0.23, 0.07]
Pacific	0.124	2.117	0.058	0.953	[-4.04, 4.29]

Table 3 reports these simple slopes together with the same controls as in Table 2. As noted before, the simple slope for propWomen_c in California (CA) is positive but not

statistically significant, indicating little evidence of a gender–poverty association in California. The slope for Texas (TX) is larger and statistically significant, suggesting that in Texas, counties with more women tend to have higher Poverty when unemployment, education, and race/ethnicity are held constant. However, like before the difference between these slopes (the CA:TX interaction) is not significant, indicating no strong evidence that the propWomen–Poverty relationship truly differs between California and Texas.

Discussion and Conclusion

For RQ1, the multiple regression shows that county level poverty is most strongly associated with unemployment, education, race/ethnicity, gender composition, and average age. Counties with higher unemployment and a larger share of adults without a high school diploma tend to have substantially higher poverty, whereas higher levels of educational attainment are linked to notably lower poverty. Racial and ethnic composition also matters: counties with larger Black and Native populations have higher poverty, whereas those with larger Hispanic and Asian populations tend to have lower poverty, net of the other variables. Poverty is additionally higher in counties with more women and lower in counties with older average age, with labor-market (service, public sector, and self-employment) and commuting patterns (greater public transit use) adding smaller but meaningful contributions. Population size and several other employment and transportation categories show little predictive value. Overall, these findings are consistent with the EDA, which already highlighted unemployment, education, and race/ethnicity as the clearest indicators of county poverty.

For RQ2, the California–Texas interaction suggests that gender composition is only weakly related to county poverty. The slope for the percentage of women is small and not significant in California, and somewhat larger and significant in Texas. This indicates that Texas counties with more women tend to have higher poverty after controlling for unemployment, education, and racial/ethnic blocks. However, the difference between the California and Texas slopes is not itself significant, so there is little evidence that the gender–poverty relationship truly differs between the two states. This pattern is consistent

with the EDA, where the fitted lines for California and Texas were closely overlapping (see Figure 4).

Substantively, the results suggest that county poverty is driven primarily by local labor-market conditions, educational attainment, and racial/ethnic composition, with gender composition and many detailed employment and commuting measures playing a more limited role. The analysis is cross-sectional and observational, so the associations should not be interpreted as causal, and residual plots (Appendix B) indicate some remaining spatial structure, suggesting that a full spatial level modeling could provide a better account of geographic dependence. Within these limits, the findings highlight the importance of policies that reduce unemployment and expand educational opportunities as drivers for lowering county-level poverty, and they point to future work that incorporates richer spatial modeling, additional policy covariates, and longitudinal data to study changes in poverty over time.

I used ChatGPT (GPT-5.1 Thinking; OpenAI, 2025) to help me with R, debug code, and polish the overall analysis. I designed the analysis plan by myself, and then asked specific questions to ChatGPT mainly for sanity checks. I asked whether particular modeling choices were appropriate and what problems I might run into, while I implemented the code and verified all results on my own.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325.

OpenAI. (2025, April 29). *GPT-5.1: Advancing reasoning and assistance*. Retrieved December 8, 2025, from https://openai.com/index/gpt-5-1/

Appendix A

Appendix A: Counties Excluded From Regression Models

Counties excluded from regression models due to missing data

State	County	Variables with missing data
Alaska	Hoonah-Angoon Census Area	LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher
Alaska	Kusilvak Census Area	AvgAge
Alaska	Lake and Peninsula Borough	Long, Lat, county_fips
Alaska	Petersburg Borough	LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher, AvgAge
Alaska	Prince of Wales-Hyder Census Area	LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher
Alaska	Skagway Municipality	LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher
Alaska	Wrangell City and Borough	LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher
Illinois	LaSalle	LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher
Louisiana	LaSalle Parish	LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher, AvgAge
New Mexico	Doña Ana	LessThanHighSchool, HighSchoolDiploma, SomeCollegeOrAssociateDegree, BachelorDegreeOrHigher, AvgAge, Long, Lat, county_fips
South Dakota	Oglala Lakota	AvgAge

Figure A1

Counties Excluded From Regression Models Due to Missing Data

Appendix B

Appendix B: Exploratory Data Analysis

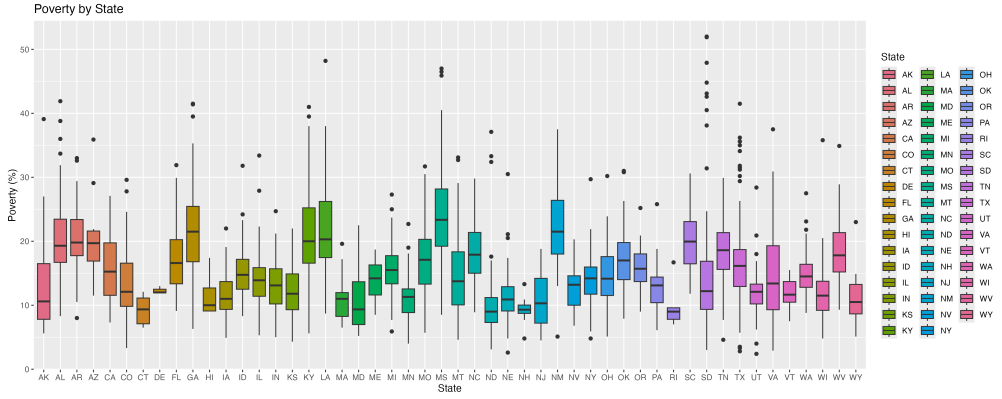

Figure B1

County Poverty Rates by State

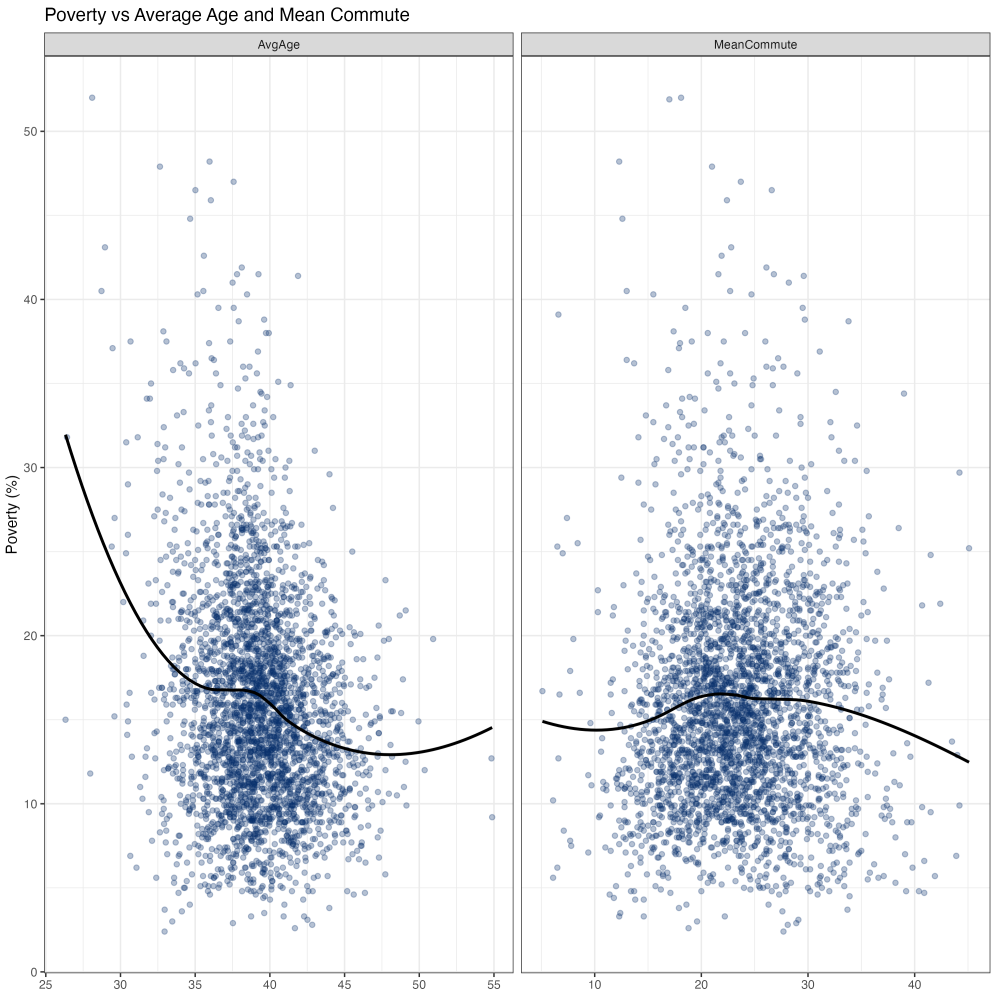

Figure B2

County Poverty Versus Average Age and Mean Commute Time

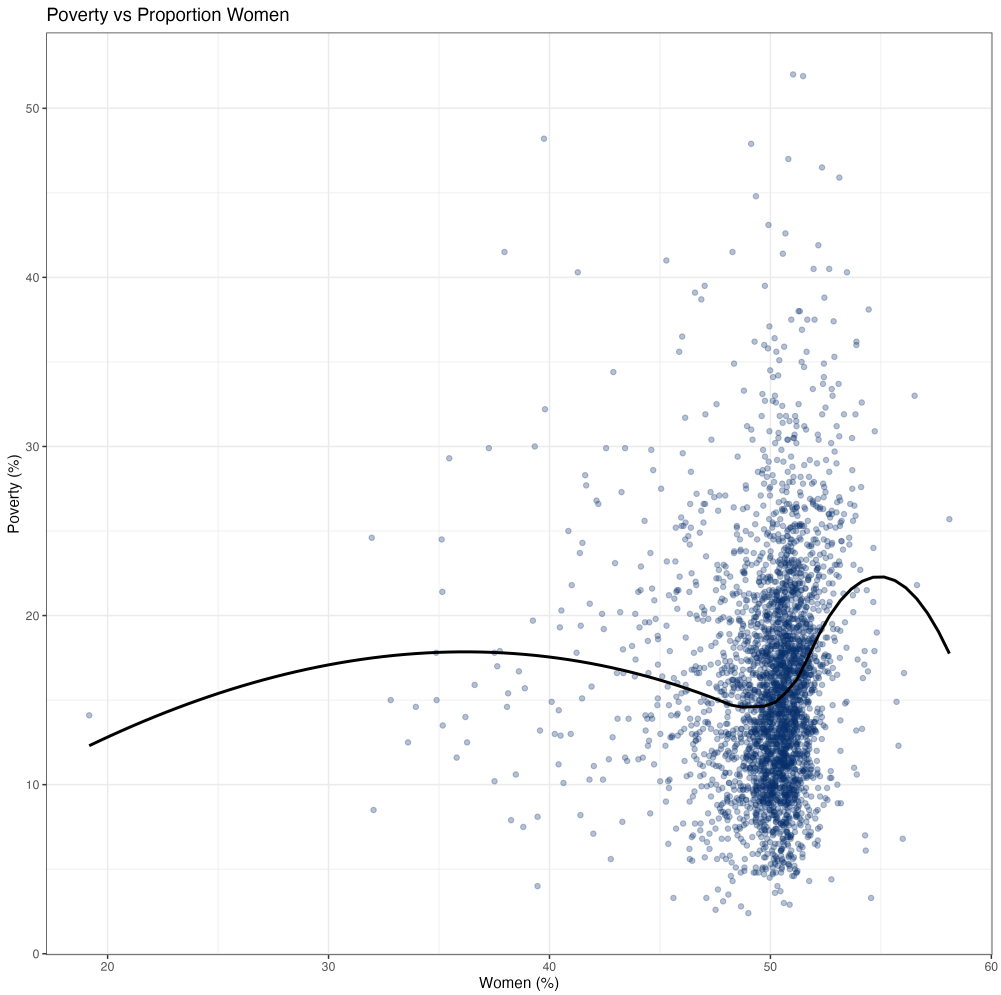

Figure B3

County Poverty Versus Percentage Women

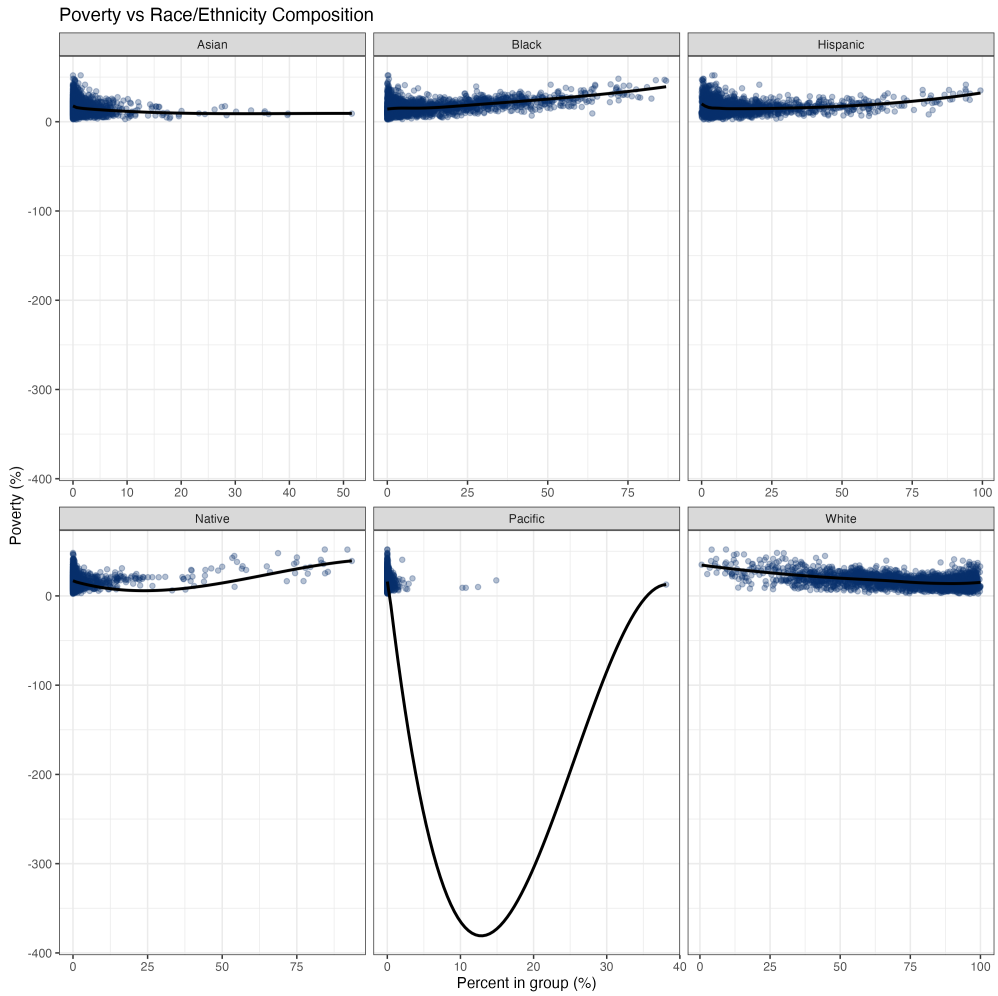

Figure B4

County Poverty Versus Race/Ethnicity Composition

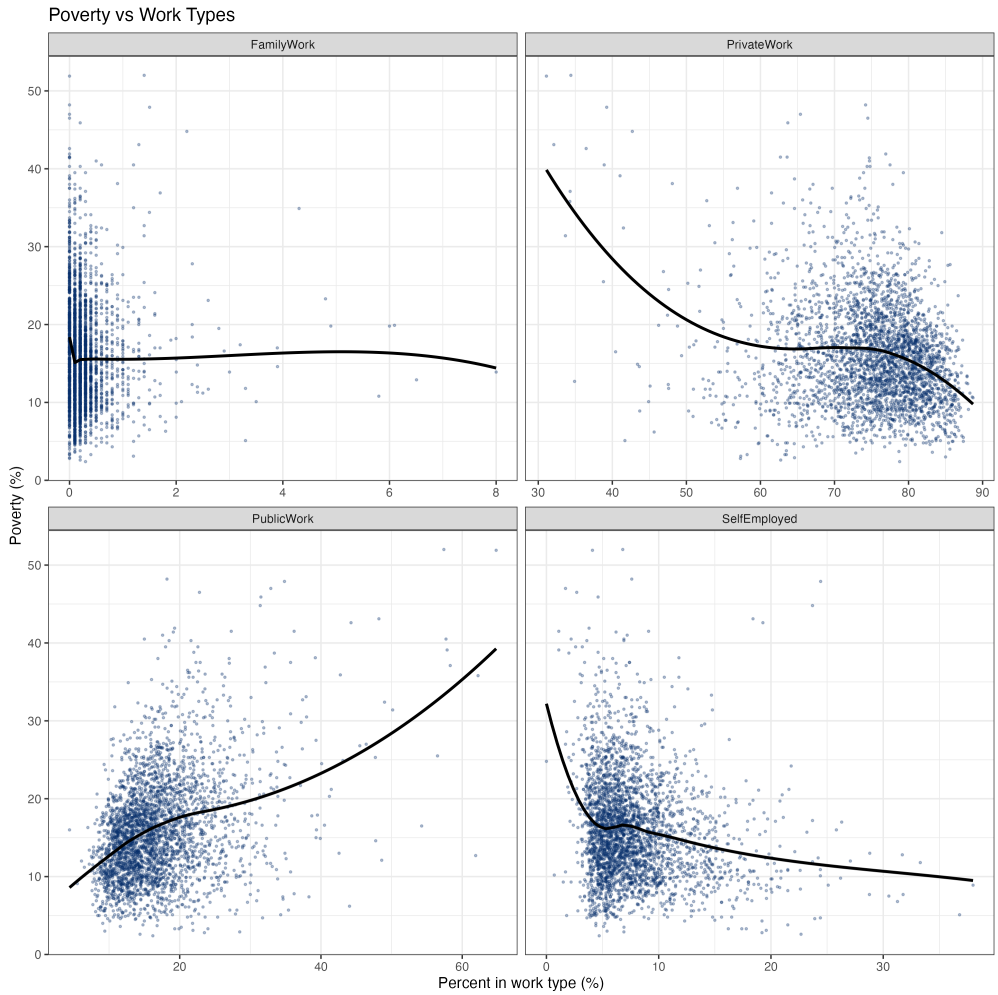

Figure B5

County Poverty Versus Work Type Categories

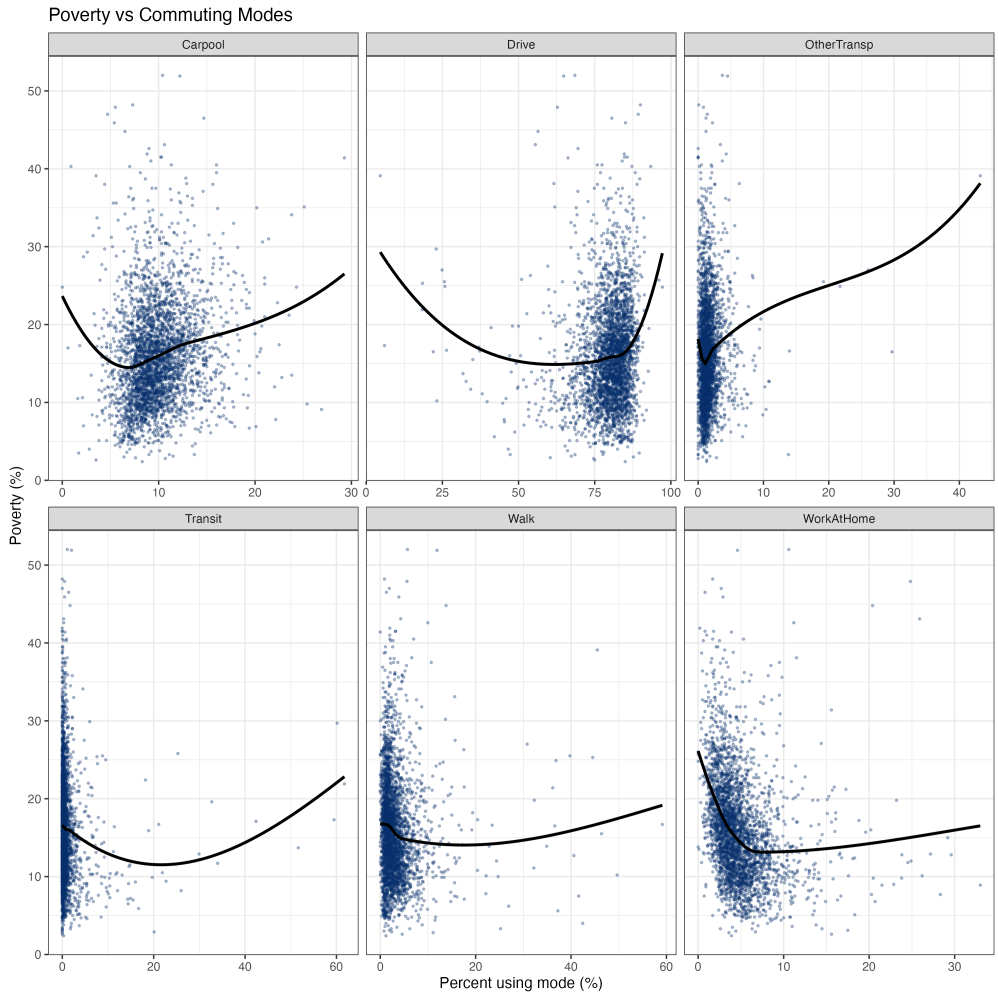

Figure B6

County Poverty Versus Commuting Modes

Appendix C

Appendix C: Assumptions and Diagnostics

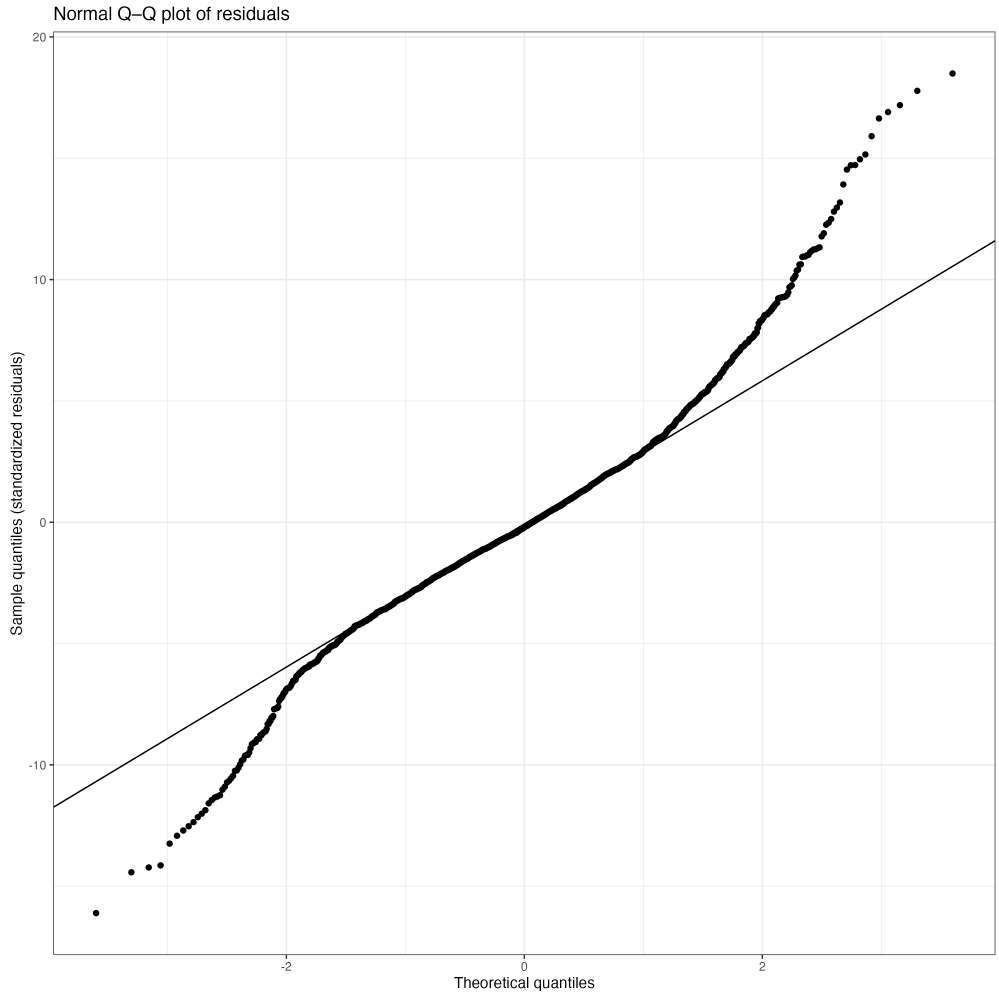

Figure C1

Normal Q–Q Plot of Residuals for the Main Regression Model

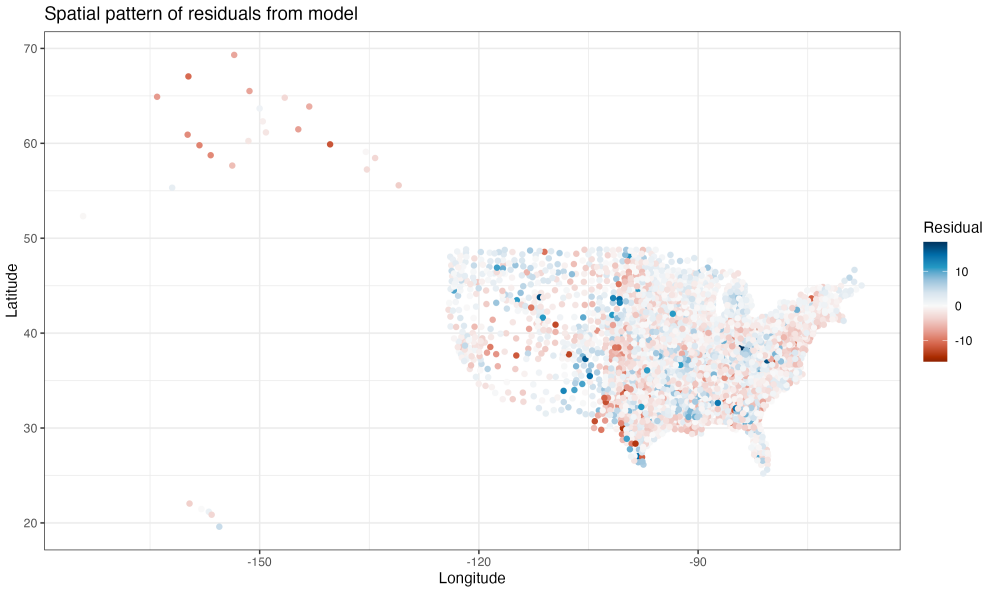

Figure C2

Spatial Pattern of Residuals by Longitude and Latitude

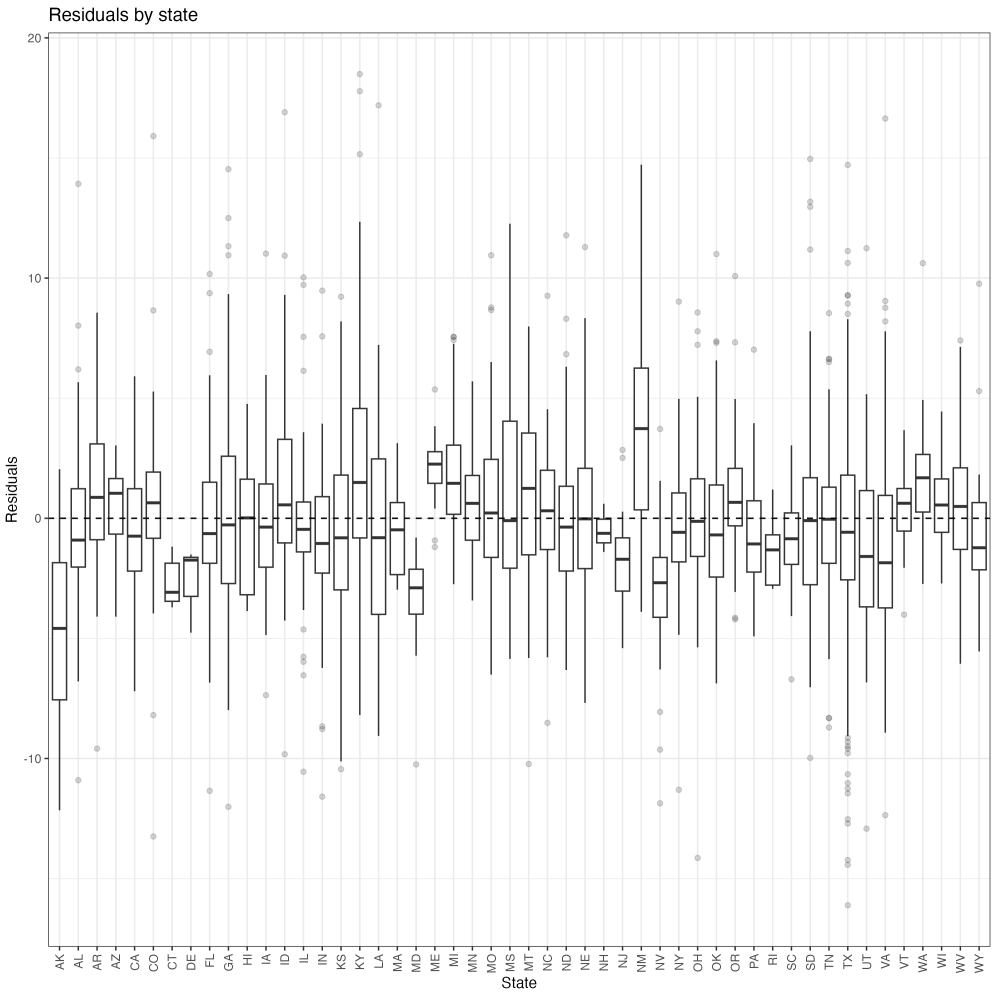

Figure C3

Residuals by State

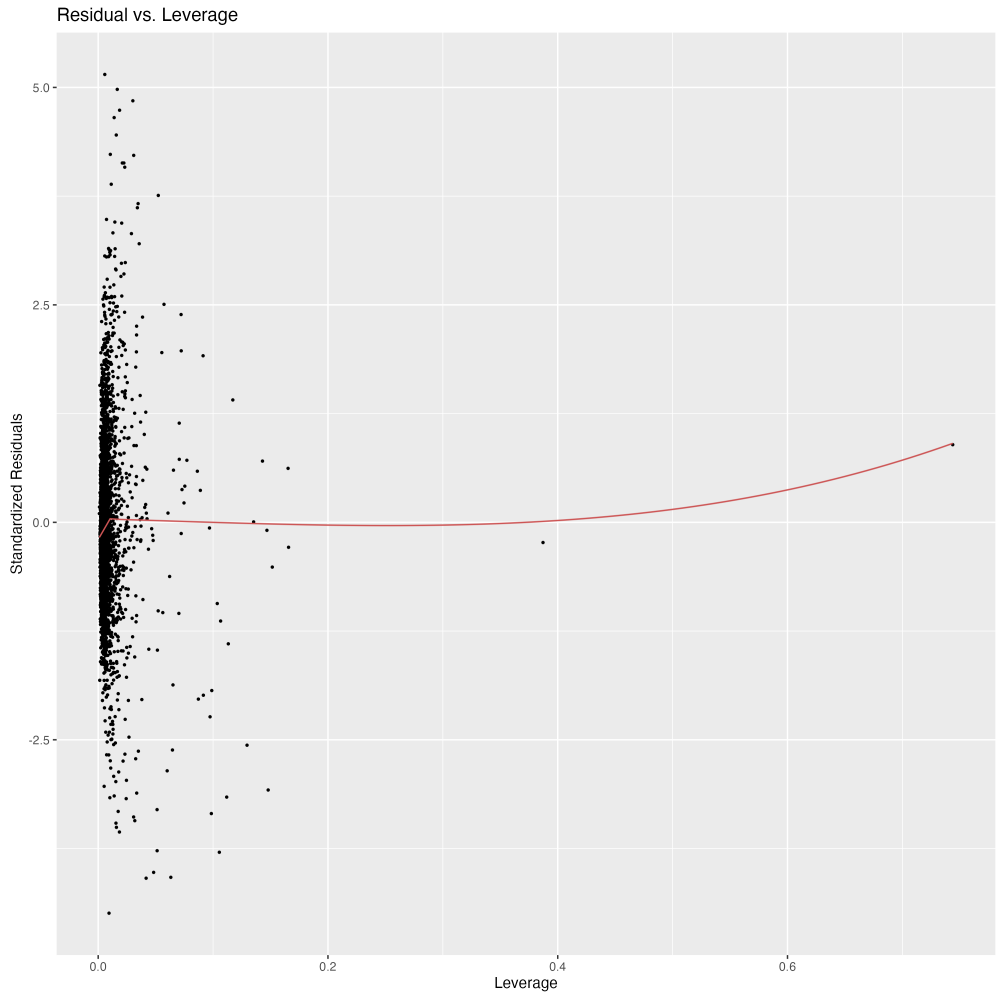

Figure C4

Standardized Residuals Versus Leverage

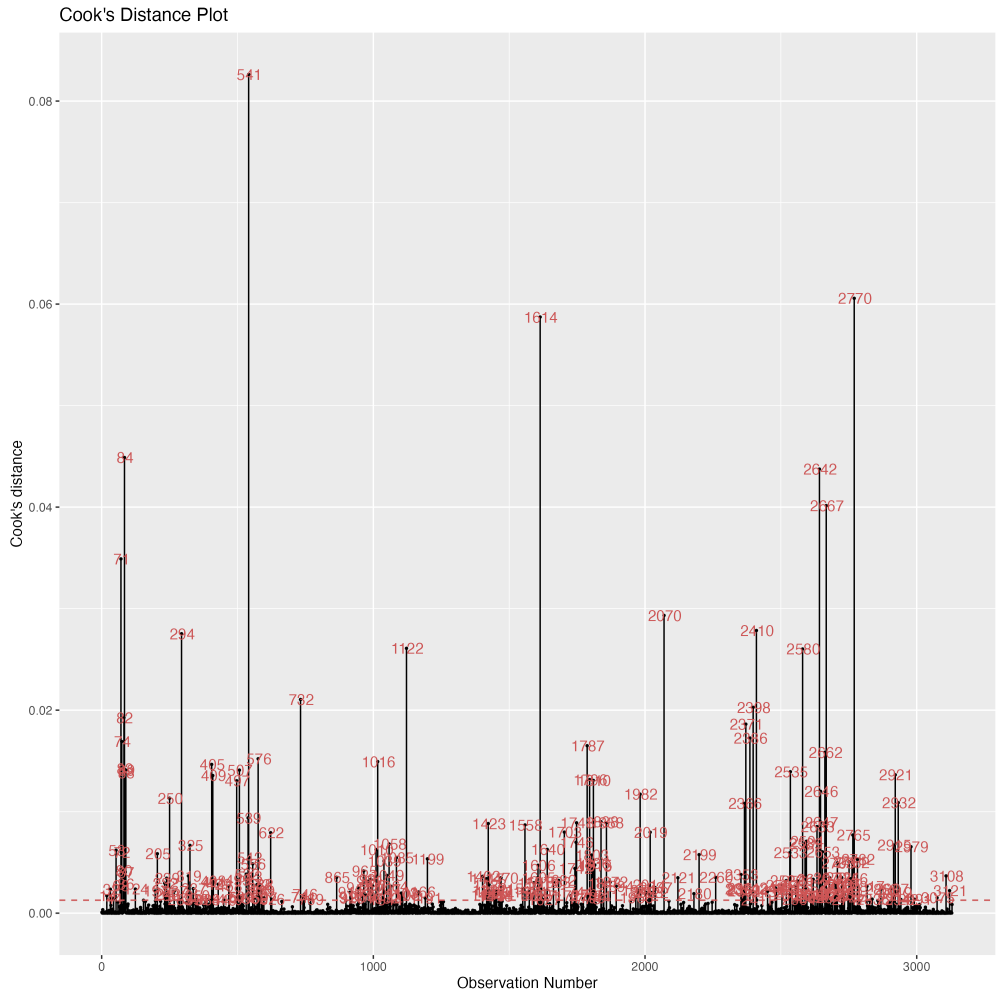

Figure C5

Cook's Distance Values for Influential Counties

Variance Inflation Factors Table for Checking Colinearity		
Predictor	VIF	Note
TotalPop	1.53	
propWomen	1.29	
Hispanic	2.52	
Black	1.93	
Native	1.85	
Asian	2.40	
Pacific	1.58	
Service	1.93	
Office	2.05	
Construction	3.29	
Production	4.48	
Carpool	1.27	
Transit	1.61	
Walk	1.99	
OtherTransp	1.32	
WorkAtHome	2.50	
MeanCommute	1.69	
PublicWork	2.17	
SelfEmployed	2.92	
FamilyWork	1.25	
Unemployment	1.96	
HighSchoolDiploma	6.77	High (≥ 5)
SomeCollegeOrAssociateDegree	2.63	
BachelorDegreeOrHigher	10.04	High (≥ 5)
AvgAge	1.97	
Long	2.44	
Lat	2.39	

Table C1

Variance Inflation Factors for Predictors in the Main Regression Model