Factors Contributing to Income Levels Across the United States

Caroline Drury, Connor Prather, and Wendell Rogers

Georgia Institute of Technology

ECON 2250

Professor Afi Ramadhani

Fall 2025

I. Introduction and data:

Income inequality in the United States has drastically widened in recent years, with substantial growth occurring disproportionately among high-income earners. According to *A Guide to U.S. Income Inequality* by American Compass, high-income earners have experienced steeper income gains per year compared to the middle and lower income earners. Given these clear differences in income growth, we want to identify the factors that help explain this behavior. Specifically, the purpose of our project is to discover what economic, educational, and social factors contribute to these differences. Because income inequality is persistent across the United States and influenced by several factors such as cost of living, state tax policies, educational disparities, crime rate, and demographics, our goal is to find which of these various factors contribute most strongly to average income. Therefore, our research question is:

"What economic, educational, and social factors best predict income levels across the United States?"

For our data sets, we gathered county-level data from the United States Census Bureau that includes median income, education level, race, median age, income per capita, and county GDP. Median income was our outcome/dependent variable, while the other variables were used as predictors to assess their association with income levels. Education level was broken down into the proportion of residents with a high school diploma, bachelor's degree, and advanced degrees, and race was categorized by Asian, black, Hispanic, and white populations. After performing our analysis, income per capita had the strongest (and positive) relationship with median income. This was expected as income per capita directly contributes to the median income in a county. In contrast, we found that education did not have as strong a correlation with median income as expected, which suggests that its influence may be indirect or vary based on the context of each county.

II. Methodology:

The primary goal of our multi-hypothesis is to test the joint null hypothesis that all predictor variables, those being per capita income, education count, median age, and the four race/ethnicity percentages, have no linear relationship with the outcome variable: median household income. The specific statistical test used is the F-test for overall model significance. The Null Hypothesis (H_o) for the overall model is that all coefficients are simultaneously equal to zero ($\beta_1 = \beta_2 = \ldots = \beta_8 = 0$), implying the model has no power in predicting the outcome variable accurately. The Alternative Hypothesis (H_A) is that at least one coefficient is non-zero ($\beta_x \neq 0$), suggesting the model has at least some accuracy in predicting the outcome variable. In addition, the significance of each individual predictor is assessed using a t-test, where the individual null hypothesis is that a specific coefficient is zero ($\beta_x = 0$), and the alternative hypothesis is that the

coefficient is non-zero ($\beta_x \neq 0$), indicating a significant relationship for that single variable after accounting for all others. In the context of our model, it holds true that when all other variables are held constant, there is still an impact on the output variable, on average.

The multiple linear regression model relies on the underlying assumptions of the Classical Linear Regression Model (CLRM) for the F-tests and t-tests to be valid. The four core assumptions are: linearity, which assumes the relationship between the predictors and the outcome is linear; independence, which assumes the residuals are uncorrelated with one another, homoscedasticity (or constant variance), which assumes the variance of the residuals is constant across all levels of the predicted outcome, and finally, normality of residuals, which assumes the residuals are normally distributed. While minor violations of the normality assumption are often tolerated given the large sample size of 3,220 county observations, systematic violations of linearity or homoscedasticity would require model adjustments, such as variable transformation or the use of weighted least squares.

III. Results:

For the results, we wanted to see how the data helped answer our research question, “What economic, educational, and social factors best predict income levels across the United States?” After performing the test, the variable that stuck out the most was our per capita income. This showed that there was a strong correlation between one's per capita income and a household's income. Intuitively, this makes sense, but it was nice to see it backed up with data. This also helps tell us that households that make a lot often tend to have everyone in the family making a lot, a good case for the power of a wealthy socio-economic background.

Another really interesting result we had regarded the correlation between a county’s wealth and income. As the wealth of the overall county increased, there was a strong increase in personal income as well. This can help expose how segregated urban areas are by wealth. The data makes a convincing argument for the concentration of wealth contained in cities and how the wealth of the county directly leads to a wealthy household.

Interestingly, we had some results that were unexpected. We found that race was a strong negative predictor, similar to education. Education being negative was very counterintuitive, as common sense has long held that going to school for longer improves how much you earn. I doubt our model disproves decades of evidence about the benefits of education. These results can instead be interpreted as evidence of how high the personal income and county income are for predicting wealth in a household. The strength of these two likely explains most of the variance, so the effect education has is minimized, and the sign is able to flip. Ultimately, our model was able to show the strong effects of a person's environment and personal income, and household income, marking a successful research project.

IV. Discussion + Conclusion:

Overall, our results suggest that income per capita and county GDP play the most influential role in predicting median household income. Counties with higher amounts of accumulated wealth produce higher household earnings, which reflects that opportunity and socioeconomic advantage are associated with geography (county). The significance of the F-test confirms that our model explains variation in income, and the individual t-tests identify economic indicators as the strongest predictors.

In contrast, education and race behaved differently than we expected. When economic factors were included, these variables had weak or negative correlations, which may indicate that their effects happen indirectly through county-level wealth rather than having an independent influence. This suggests that environmental context is central to explaining income disparities. Overall, our findings highlight the dominant role of broader economic conditions in shaping income levels across U.S. counties.

Much like many other experiments, there are necessary limitations of our analysis that must be addressed. The most prevalent, multicollinearity, has a strong impact on the results found between per capita income, total population, and the education count. As a result of this overlap, coefficient estimates tend to be unstable and can lead to certain results that are hard to interpret their true impact. There is also significant omitted variable bias, with variables such as cost of living and housing costs being left out of the experiment. Oftentimes, these variables are correlated with large, educated, and diverse populations and are crucial determinants in median household incomes. Lastly, there is likely an assumption violation with there not being constant variance in the residuals, as over 3,000 counties were represented in this experiment. Even if the coefficients themself are unbiased, this assumption violation likely can lead to biased t-statistics.

To counter these limitations, each one has a necessary implementation for if we were to rerun this experiment. First, to address multicollinearity, removing or combining highly correlated variables could further stabilize the model and improve the interpretability of the remaining coefficients. Next, to address omitted variable bias, we can simply include data such as cost of living and housing costs to mitigate the negative bias present in the current experiments. Finally, to address the assumption violation, robust standard errors could be implemented. Essentially, this would make the t-statistics and p-values more reliable without altering coefficient estimates.

This project ended up being a lot of fun and a solid model. However, after finishing the project, we realized how much more we could do to expand our research. One big thing we talked about was how we could easily add some more variables. We were interested in looking at geography and population density as additional variables and seeing the effect they could have on our model.

Another thing we were interested in reviewing was the multicollinearity of the model. After getting some strange results, it seemed possible that some of our variables may have poorly influenced our overall model. By testing this, we could see whether the larger variables were having negative effects on the total model. This would allow us to refine our model further and improve its overall accuracy.

Lastly, it'd be helpful to run some other tests to continue improving the model. One potential test is Cook's model, which looks at outliers affecting the data. Since income tends to follow a power law, there are frequently outliers far along the data set. By using Cook's, this could help reduce any data we have using the mean. Another interesting thing to do could be some residual plot tests that would allow us to determine linearity.

References:

"A Guide to U.S. Economic Inequality." *American Compass*, 8 Nov. 2023, americancompass.org/economic-inequality-guide/?gad_source=1&gad_campaignid=11199510140&gbraid=0AAAAACR4DslLJW23DHpkikMGkF-6RIsu8&gclid=Cj0KCQiA_8TJBhDNARIsAPX5qxRHC2yRT3pd_TAmoSsn01BOHqY-hCkAADAnoV5AJqLx23va3WYyHigaAqEHEALw_wcB.

Wijaya, Cornellius. "Understanding Residual Plots." Statology, 21 Apr. 2025, www.statology.org/understanding-residual-plots/.