Socioeconomic Factors Impacting Poverty in U.S. Counties: A Regression Approach

Ngan P. Vu

Western Governors University

Table of Contents

Project Overview

A. Project Highlights

A.1. Research Question

Poverty is a complex issue influenced by many factors such as income levels, education, employment rates, healthcare access, public programs, and housing cost. These factors interact in ways that are not immediately obviosu, and their combined effects on poverty rates can vary widely across different regions. This project aims to answer: *What is the relative importance of various socioeconomic factors in explaining poverty rates across U.S. counties, and how well can we predict poverty rates based on these factors using multiple linear regression?* By quantifying these relationships through data analysis, the findings can help regional policymakers and community leaders allocate resources efficiently to reduce poverty.

A.2. Project Scope

This project focuses on analyzing how socioeconomic factors are associated with poverty rates across U.S. counties using multiple linear regression. It includes data collection, exploratory analysis, feature engineering, model fitting, model evaluation and diagnostics, model refinement, and statistical tests. The goal is to assess how well these factors, collectively and individually, explain variation in poverty rates.

The project will not apply other modeling techniques such as other machine learning algorithms, generalized linear models, or regularized regression. It will also not attempt to establish causal relationships as the main focus is associations and predictive insights.

A.3. Solution Overview

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to guide each step of the analysis. The process begins with defining the research question, followed by understanding and preparing the data, building and refining a statistical model, evaluating its performance, and finally presenting the findings to inform decision-making.

The main analytical approach is multiple linear regression using the ordinary least squares (OLS) method. With this model, I can quantify the relationships between multiple socioeconomic variables and poverty rates. To evaluate the model, both an F-test (for overall significance) and t-tests (for individual predictors) are used. Standardized coefficients are also calculated to assess the relative importance of each predictor.

To ensure the reliability of the model and its statistical tests, diagnostic checks are conducted using residual analysis, including plots to assess linearity, normality, and homoscedasticity.

I used Jupyter Notebook as the primary development platform. Jupyter Notebook is an easy-to-use and visually intuitive environment for conducting analysis and code presentation. The whole analytical process will be done using Python libraries such as pandas for data manipulation, Matplotlib and Seaborn for visualization, NumPy for numerical operations, statsmodels for regression and statistical testing, and scikit-learn for scaling and validation. All code used in this project is included in the submitted notebooks (Appendix A).

Project Execution

B. Project Plan Execution and Milestones

B.1. Project Plan

The execution of my project closely followed the plan outlined in Part B of the project proposal. The following objectives were completed as intended, and there was no major change to the project's scope, methods, or steps.

- Objective 1: Data preparation and exploratory data analysis (EDA)
  - Deliverable 1.1: Data retrieval from the U.S. Census Bureau
  - Deliverable 1.2: A cleaned dataset with selected variables and handled missing or inconsistent entries
  - Deliverable 1.3: A summary of EDA including descriptive statistics, correlation analysis and visualizations
  - Deliverable 1.4: Transformed variables to address nonlinearity and skewness, and documented justification for each transformation

- Objective 2: Build and evaluate a baseline multiple linear regression model
  - Deliverable 2.1: An OLS model fitted on transformed and scaled predictors with train/test split
  - Deliverable 2.2: A summary of baseline model performance
  - Deliverable 2.3: Preliminary interpretation of coefficients and statistical significance

- Objective 3: Perform model diagnostics to validate model assumptions
  - Deliverable 3.1: Residual plots (residuals vs. fitted values, residuals vs. predictors) to assess homoscedasticity and linearity
  - Deliverable 3.2: Histogram, boxplot and Q-Q plot of residuals to assess normality
  - Deliverable 3.3: A summary of baseline model validity and limitations

- Objective 4: Refine the regression model to improve reliability and account for identified issues.

  - Deliverable 4.1: A new multiple linear regression model that incorporates improvements such as variable transformation, interaction terms, and robust standard errors (HC3).
  - Deliverable 4.2: A summary of updated model performance metrics
  - Deliverable 4.3: Statistical output including hypothesis test results
  - Deliverable 4.4: A summary of the relative importance of predictors, including a tornado diagram of standardized coefficients
  - Deliverable 4.5: Revised model diagnostics with visualizations to confirm improvements in model assumptions

- Objective 5: Present results and communicate findings
  - Deliverable 5.1: A written project report summarizing methodology, findings, and conclusions
  - Deliverable 5.2: At least one well-documented Jupyter Notebook with code, outputs, and narrative explanations
  - Deliverable 5.3: A video presentation explaining the project process, results, and insights

B.2. Project Planning Methodology

I followed the CRISP-DM framework as planned and each phase of CRISP-DM maps directly to steps in this project:

1. Business understanding: Define the research question and objectives around explaining poverty rates using socioeconomic factors.
2. Data understanding: Explore the data with summary statistics, histograms, pairplots, and correlation analysis.
3. Data preparation: Clean the dataset, transform variables, and scale features.
4. Modeling: Build and refine an OLS multiple linear regression model using statsmodels.
5. Evaluation: Assess model performance and validate assumptions with statistical metrics and diagnostic plots.
6. Deployment: Present findings in a clear, visual format to support data-driven decision-making.

B.3. Project Timeline And Milestones

The only difference from the original plan was the timeline. I was able to start and complete each phase of the project ahead of schedule.

Milestones/ Deliverables	Projected start date	Projected end date	Actual start date	Actual end date	Actual duration
Data Collection	4/18/2025	4/20/2025	4/7/2025	4/9/2025	3 days
Data Wrangling	4/21/2025	4/22/2025	4/10/2025	4/11/2025	2 days
EDA & Feature Engineering	4/23/2025	4/24/2025	4/13/2025	4/14/2025	2 days
Baseline Model Development	4/25/2025	4/25/2025	4/15/2025	4/15/2025	1 hour
Model Evaluation	4/25/2025	4/24/2025	4/15/2025	4/15/2025	1 hour
Model Diagnostics	4/26/2025	4/26/2025	4/16/2025	4/16/2025	1 day
Model Refinement	4/27/2025	4/27/2025	4/17/2025	4/18/2025	2 days
New Model Validation & Final Results	4/28/2025	4/28/2025	4/19/2025	4/19/2025	1 day
Official Report	4/29/2025	4/30/2025	4/20/2025	4/21/2025	2 days
Panopto Presentation	5/1/2025	5/1/2025	4/22/2025	4/22/2025	1 day

Methodology

C. Data Collection Process

The data collection process for this project was largely consistent with the plan outlined in the proposal. I used the U.S. Census Bureau's American Community Survey (ACS) 2023 1-Year Estimates, a publicly available and trusted source for socioeconomic data.

There were no major changes in the source of data or the overall approach, but I did make one important improvement during execution. In the proposal, I initially planned to manually identify variable codes by searching ACS tables and variable documentation. However, during the actual project, I developed a more efficient and reliable method. I wrote Python code in a separate notebook (Appendix A1) that automatically looks up variable codes in downloaded JSON format metadata files (Appendix B). This script allows the user to input a table ID and parts of a variable label, then returns the matching variable codes and full labels. It handles case sensitivity and spacing issues, which reduces the risk of mistakes compared to manual search. A portion of this code was adapted from a helpful thread on Stack Overflow, which I cited directly in my comments.

There were no major obstacles during data collection, but the process did require a deeper understanding of how ACS data is structured. I spent time reading through various Census Bureau guides to fully understand table formats, variable naming conventions, and geographic codes (like the ucgid used to request county-level data). This learning curve helped construct accurate API calls.

No unplanned data governance issues occurred. The dataset contains only public, aggregate data with no personal identifiers. As noted in the proposal, I took care to document sources, cite the U.S. Census Bureau properly, and avoid making any inferences about individuals.

C.1. Advantages and Limitations of Data Set

The dataset is of high quality as it is sourced from ACS, a reputable federal statistical program. The ACS is conducted annually using thorough sampling and estimation techniques. The ACS provides reliable and up-to-date estimates on a wide range of social and economic

indicators which are directly relevant to understanding the predictors of poverty. Additionally, its nationwide county-level coverage allows for insights across diverse U.S. regions.

One limitation of the dataset is that the poverty rate used in this dataset is based on the Official Poverty Measure (OPM) rather than the Supplemental Poverty Measure (SPM). The OPM does not account for the cost of living or public assistance programs such as housing subsidies or SNAP. This limitation may affect the interpretation of socioeconomic influences, and therefore, should be kept in mind when drawing conclusions from the analysis.

D. Data Extraction and Preparation

Data extraction began with API requests to the U.S. Census Bureau's ACS 2023 1-Year Estimates, using variable codes identified during the collection process. The data from three types of ACS tables (Detailed Tables, Subject Tables, and Data Profiles) were saved as .json files, parsed with Python, and loaded into Pandas DataFrames. I then merged these DataFrames into a single dataset at the county level.

Data preparation included renaming coded headers to descriptive variable names, converting data types, and handling missing or special values by imputing values based on the 2022 ACS datasets. Redundant or unnecessary columns were dropped, and related variables were combined to improve usability of the dataset. The cleaned dataset was saved as a .csv file for modeling. These processes were done in a separate data wrangling notebook (Appendix A2).

E. Data Analysis Process

E.1. Data Analysis Methods

In exploratory data analysis (EDA), descriptive statistics were used first to summarize and visualize the dataset. By examining the measures of central tendency and measures of spread, I could make informed decisions on data cleaning.

To detect multicollinearity, I first examined the correlation matrix to identify highly correlated predictor pairs. I also calculated the Variance Inflation Factor (VIF), which measures how much the variance of a regression coefficient is inflated due to multicollinearity (Singh, 2024). A VIF value exceeding 5 is commonly considered a sign of problematic collinearity. High

VIF values suggest that a predictor may be linearly dependent on other predictors, leading to unstable coefficient estimates and reduced interpretability (Frost, 2023; Singh, 2024).

To explore relationships between the response and explanatory variables, I used pairplots and scatter plots. Non-linear patterns observed in these plots guided the choice of variable transformations to better satisfy linear model assumptions. To address nonlinearity and skewness in the data, I applied appropriate transformations based on the distribution and each variable's relationship with the response. When a variable was highly right-skewed and showed a nonlinear, curved pattern in scatter plots, a natural logarithm transformation was used. This approach compresses large values while expanding smaller ones, helping to linearize relationships and stabilize variance (Atomise Biostats, 2017). In cases where variables included zero values or small positive counts, I used the numpy's log1p transformation, which calculates the natural log of (1 + x). This is especially useful when the data includes zeros, since the regular logarithm is undefined at zero (Fatima, 2024). For variables with moderate skewness, I applied a square root transformation, which can reduce skewness and improve linearity without a overly compressing the data (Atomise Biostats, 2017).

The main analytical approach was a multiple linear regression method, which allowed me to quantify relationships between socioeconomic factors and poverty rates while considering the influence of multiple variables simultaneously. Linear regression is suitable for predicting a continuous response variable based on several predictors. Specifically, the project used ordinary least squares (OLS) model which allowed a straightforward interpretation of coefficients, significance testing and model evaluation.

To evaluate the model's predictive performance, I used metrics such as R-squared on both the training and test sets, adjusted R-squared, and mean squared error (MSE) on the test set. These metrics provided insight into how well the model fit the training data and how accurately it predicted poverty rates on new, unseen data.

To assess the overall significance of the model, I conducted an F-test to determine whether the combination of predictors explains a significant proportion of variance in response variable (Watts, 2022). The hypotheses for this test were:

- Null hypothesis: None of the socioeconomic factors have a statistically significant effect on county-level poverty rates. In other words, all regression coefficients are equal to zero.

- Alternative hypothesis: At least one of the socioeconomic factors has a statistically significant effect on county-level poverty rates. That is, at least one regression coefficient is not equal to zero.

This directly supports the research question regarding how well we can predict poverty rates using the selected predictors.

Additionally, individual t-tests on each independent variable were used to assess the statistical significance of each predictor's coefficient. The hypotheses for each test were:

- Null hypothesis: The coefficient for the predictor is zero (no effect).
- Alternative hypothesis: The coefficient is not zero (the predictor has a significant effect).

This type of statistical test helped me determine which factors contribute significantly to the model when accounting for the effects of all other variables. This aligns with the research question's objective of assessing the relative importance of each socioeconomic factor. It is important to note that, as discussed by a reputable contributor on Cross Validated (Stack Exchange), a model may show a significant F-test even when individual t-tests are not significant, due to multicollinearity or redundant predictors (whuber, 2011). This is why both F-test and t-test were used to analyze the predictive power of independent variables. The alpha value used for both types of tests is 0.05 or 5%, which is a standard threshold in statistical hypothesis testing.

To evaluate the relative importance of each socioeconomic factor in explaining poverty rates, I also used the standardized coefficients from the multiple linear regression model. As Frost (2021) points out, while p-values help identify statistically significant predictors, standardized coefficients are more appropriate for indicating the practical importance of a variable in the model, as they allow for comparison of predictors on a common scale. These coefficients were obtained by scaling all independent variables using MinMaxScaler before fitting the model. This will ensure that the predictors are on the same scale.

For the baseline model and the improved model, I implemented residual analysis to verify model assumptions and ensure the model's validity. OLS regression relies on several assumptions: linearity of relationships between predictors and the dependent variable, homoscedasticity (constant variance of residuals), independence of residuals, and normally distributed errors (The Pennsylvania State University, n.d.-a). The analysis included

visualizations such as residual plots, histogram and boxplot of residuals, and Q-Q plot. This diagnostic step helped confirm whether the use of OLS and the resulting hypothesis tests were valid.

Finally, to address issues identified in model diagnostics, I applied several refinement methods. The response variable was transformed to reduce heteroscedasticity and improve linearity (Frost, 2022). An interaction term was added to improve model accuracy when the effect of one variable depends on another (Frost, 2024; Berman, n.d.). Robust standard errors (HC3) were used to produce more reliable p-values in the presence of heteroscedasticity (Lucich, 2024; Fiveable, 2024).

E.2. Advantages and Limitations of Tools and Techniques

I used Jupyter Notebook as the primary development platform. Jupyter Notebook is an easy-to-use and visually intuitive environment for conducting analysis and code presentation. The whole analytical process was done using Python libraries such as pandas for data manipulation, Matplotlib and seaborn for visualization, NumPy for numerical operations, statsmodels for regression and statistical testing, and scikit-learn for scaling and validation. The tools used were great for statistical analysis and data visualization, but each had trade-offs. For example, visualization libraries like Matplotlib and seaborn are flexible but needed extra customization for advanced visuals, and scikit-learn does not provide detailed statistical summaries. Therefore, I needed to use a combination of libraries throughout the analysis process.

Multiple linear regression was used to examine the relationship between poverty rates and socioeconomic factors. This technique is interpretable and widely used for estimating the impact of several predictors at once. However, linear regression assumes linearity, normality, and constant variance, which may not always hold. This required me to conduct residual analysis to make sure that the model and results are valid.

E.3. Application of Analytical Methods

The data analysis process was done in the main analysis notebook (Appendix A3).

1. Exploratory Data Analysis (EDA)

   The analysis began with an exploration of the dataset's structure:

- Sample size checks: Verified the number of counties, states, and counties per state.
- Outlier evaluation: Extreme values in poverty rate were found to be concentrated in Puerto Rico whose socioeconomic characteristics differ significantly from U.S. states. Puerto Rico was excluded to improve model generalizability. Other outliers were retained as they likely represent legitimate values from a trusted source (U.S. Census Bureau).
- Multicollinearity check: A correlation matrix and Variance Inflation Factor (VIF) were used to examine relationships between predictors. One variable (median_gross_rent) was dropped due to redundancy (a high VIF).
- Linearity assessment: Pairplots and scatter plots were used to assess relationships between each predictor and the response variable. Nonlinear patterns led to targeted transformations:
  - median_household_income: log transformation
  - public_transit: log(x+1) transformation
  - median_house_value: square root transformation

2. Feature Scaling

All predictors were scaled using MinMaxScaler to normalize their ranges and enable meaningful coefficient comparisons in the regression model.

3. Baseline Model Construction

A baseline multiple linear regression model was built using the Ordinary Least Squares (OLS) method:

- A constant term was added to the feature matrix.
- The data were split into training and test sets.
- The model was fitted using the training data, and predictions were generated on the test set.

4. Model Evaluation

Key performance metrics were calculated (see Appendix C for model's summary):

- R-squared (both training and test sets) to measure the proportion of variance in the dependent variable explained by the predictors. R-squared should be greater than 0.6,

which means that at least 60% of the variance in poverty rates is explained by the socioeconomic variables.

- Adjusted R-squared: A modified version of R-squared that accounts for the number of predictors.
- Mean Squared Error (MSE) on the test set to evaluate prediction error.
- R-squared and adjusted R-squared to assess variance explained.
- Coefficients and p-values from the model summary provided insights into the direction, magnitude, and significance of relationships.

5. Model Diagnostics

Residual analysis was conducted to verify OLS assumptions (see Appendix D for visualizations):

- Residuals vs fitted values plot: To check for linearity and homoscedasticity. Ideally, the residuals should be randomly scattered around zero with no clear pattern. Patterns or funnel shapes may indicate issues with linearity or constant variance assumptions (The Pennsylvania State University, n.d.-b).
- Residuals vs individual predictors plots: To check for the linearity assumption between each predictor and the response variable. "The (vertical) spread of the residuals remains approximately constant as we scan the plot from left to right. …violation of either of these for at least one residual plot may suggest the need for transformations of one or more predictors and/or the response variable" (The Pennsylvania State University, n.d.-b).
- Normality of residuals: To verify the normality assumption of the residuals.
    - Histogram: Should resemble a bell-shaped curve.
    - Boxplot: Should be symmetric with few or no outliers.
    - Q-Q Plot: Points should fall approximately along the 45-degree reference line.

6. Model Refinement

Several refinement techniques were first tested incrementally in a separate notebook (see Appendix A4) before selecting the final improvements included in the main analysis.

- To address potential issues and improve performance:

- The response variable was transformed using a square root transformation.
- An interaction term between log_income and sqrt_median_house_value was added.
- A new OLS model was fitted with robust standard errors (HC3) to account for heteroscedasticity.

- The refined model was validated again to make sure that the model's performance is improved and assumptions are met.

7. Hypothesis Testing and Interpretation

- An F-test was conducted to confirmed the model's overall significance ($\alpha = 0.05$).
  - Null hypothesis: All regression coefficients are equal to zero

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

  - Alternative hypothesis: At least one regression coefficient is not equal to zero

$$H_1: At\ least\ one\ \beta_i \neq 0$$

- t-tests identified which predictors were statistically significant ($\alpha = 0.05$).
  - Null hypothesis: The coefficient for predictor i is zero

$$H_0: \beta_i = 0$$

  - Alternative hypothesis: The coefficient for predictor i is not zero.

$$H_0: \beta_i \neq 0$$

- Standardized coefficients were used to assess the relative importance of predictors. These were visualized using a tornado diagram.

Project Results

F. Data Analysis Results

F.1. Statistical Significance

1. Model Performance

- Model type: Supervised regression using multiple linear regression (OLS with HC3 robust standard errors)
- Test set R-squared: 0.831, indicating the model explains 83.1% of the variance in poverty rates on unseen data.
- Mean Squared Error (MSE): 0.072, a substantial improvement over the baseline model (MSE = 4.32).
- Training R-squared: 0.796, indicating the model explains 79.6% of the variance in the training data (see Appendix E).
- Model diagnostics results (see Appendix F2 – F4):
    - Residual vs fitted plot: Residuals are randomly scattered around zero with consistent spread across fitted values. This means there is no major heteroscedasticity. A few mild outliers are present.
    - Residuals vs predictors: No strong patterns or funnel shapes. This supports the assumption of constant variance and linearity across predictors.
    - Histogram & boxplot of residuals: Distribution is roughly normal, bell-shaped, and symmetrical. The median is near zero, with a few mild upper-end outliers.
    - Q-Q plot: Residuals closely follow the normal line with a slight deviation in the upper tail, which is acceptable.

  Overall, the final model met the linear regression assumptions and its results (coefficients, p-values) can be trusted for inference.

2. Overall Model Significance (F-test)

- Null hypothesis: All regression coefficients are equal to zero

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

- Alternative hypothesis: At least one regression coefficient is not equal to zero.

$$H_1: At\ least\ one\ \beta_i \neq 0$$

- Alpha level (α): 0.05
- Test result: F-statistic = 312.2, p < 0.001 (see Appendix E1)

Since the p-value is well below 0.05, we reject the null hypothesis. This indicates the model as a whole is statistically significant.

3. Individual Predictor Significance (t-tests)

- Null hypothesis: The coefficient for predictor i is zero:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

- Alternative hypothesis: The coefficient for predictor i is not zero:

$$H_1: At\ least\ one\ \beta_i \neq 0$$

- Alpha level (α): 0.05
- Predictors with p-values below 0.05 include (see Appendix E1):
  - log_median_income
  - sqrt_median_house_value
  - log_public_transit, health_insurance
  - The interaction term income × house value.

These results mean I have enough evidence to conclude these variables are significantly associated with poverty rate when other predictors are controlled.

The percentage of bachelor holders and public assistance were not statistically significant in the refined model, which mean these factors have limited unique contribution when controlling for other factors.

4. Relative Importance of Predictors

The model's summary (Appendix E) and tornado diagram of standardized coefficients (Appendix G) showed that:

- The interaction between income and house value had the largest impact.
- The most impactful predictors of poverty rates are "median house value" and "median household income" as they both show strong negative relationships. Counties with higher home values and incomes tend to have lower poverty rates.
- Other variables had smaller coefficients but still measurable effects.

- Higher unemployment associates with higher poverty
- Greater health insurance coverage comes with lower poverty
- More public transit use relates to higher poverty (possibly reflecting urban conditions)

F.2. Practical Significance

The results are practically meaningful. Income and home value had strong negative relationships with poverty, suggesting that higher-income areas and those with more expensive housing tend to have lower poverty rates. Counties with more unemployment and less health insurance coverage tend to have higher poverty.

For example, a local government or policymaker could use this model to identify counties most at risk based on these factors and effectively target resources or policies, like workforce programs or health coverage plans.

Even though public assistance rates were not statistically significant, this may reflect the way poverty is defined (excluding non-cash benefits), not the lack of practical effect. These insights are still valuable for understanding how poverty is measured and how different variables interact.

F.3. Overall Success

The project was successful as all important analysis steps were conducted properly and completed. The analytical process includes data preparation, OLS model implementation, statistical testing, model diagnostics, model refinement and delivery of visualizations and documentation.

The refined model passed significance tests, met model assumptions, and explained over 79% of the variation in training data and over 83% on the test set. It also highlighted which factors have the strongest relationships with poverty. This project successfully provided answer to the research question and clear insights for decision making.

G. Conclusion

G.1. Summary of Conclusions

The goal of this project is to understand how well socioeconomic factors, collectively and individually, explain variation in poverty rates across U.S. counties using multiple linear regression. The key findings that support the project goal are as follow:

- The final model worked well and explained 79.6% of the variance in training data and 83.1% in the test set.
- Based on the F-test result, the final model was statistically significant overall ($p < 0.001$).
- Significant predictors included median income, median house value, health insurance coverage, public transit use, and an income × house value interaction.
- Counties with higher incomes and home values tend to have significantly lower poverty rates. On the other hand, higher unemployment, greater health insurance coverage and more use of public transit (a potentially tied to urban poverty) were associated with higher poverty.
- Public assistance rates and the percentage of bachelor's degree holders were not statistically significant in the final model, which mean these factors have limited unique contribution when controlling for other factors. It is important to note that unlike SPM, the OPM does not include non-cash public assistance (like SNAP, housing subsidies, TANF) in its calculation. So counties with high assistance rates may not show reduced poverty under OPM even though aid might be helping in reality. For instance, the U.S. Census Bureau reports that based on SPM data, "Social Security kept 27.6 million people out of poverty and had the largest anti-poverty impact" (Creamer & King, 2024).

G.2. Effective Storytelling

To communicate findings effectively, I used a combination of graphical and statistical visualizations throughout the analysis. All visualizations were created using Python's matplotlib, seaborn, statmodels. In the EDA phase, scatter plots, pairplots, and histograms helped reveal skewed distributions and nonlinear relationships that guided data transformation decisions. During modeling and diagnostics, residual plots, histograms, boxplots, and Q-Q plots visually

validated model assumptions. After fitting the final model, a tornado diagram of standardized coefficients was used to visually display the relative importance of predictors on a common scale.

G.3. Recommended Courses of Action

- Invest in income growth and housing affordability: These were the strongest predictors of lower poverty. Policies targeting wage growth, job access, and affordable housing could significantly reduce poverty rates.
- Expand access to health insurance: Health insurance coverage was significantly linked to lower poverty. Improving coverage could reduce financial hardship and support poverty reduction efforts.

H. Panopto Presentation

Panopto video link:
https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8e4a4413-384d-4804-a3c4-b2c8003e2161

The video is accompanied by a presentation slide deck submitted as a separate PDF file (see Appendix H).

References

Atomise Biostats. (2017, June 28). *Transforming skewed data: How to choose the right transformation for your distribution.* Retrieved from Anatomise Biostats: https://anatomisebiostats.com/biostatistics-blog/transforming-skewed-data/

Berman, H. (n.d.). *Interaction effects in regression*. Retrieved from Stat Trek: https://stattrek.com/multiple-regression/interaction

Creamer, J., & King, M. D. (2024, November 14). *How Do Policies and Expenses Affect Supplemental Poverty Rates?* Retrieved from Census.gov: https://www.census.gov/library/stories/2024/11/supplemental-poverty-measure-visualization.html

Fatima, N. (2024, June 28). *Understanding np.log and np.log1p in NumPy*. Retrieved from Medium: https://medium.com/@noorfatimaafzalbutt/understanding-np-log-and-np-log1p-in-numpy-99cefa89cd30

Fiveable. (2024, August 20). *7.5 Robust standard errors – Intro to Econometrics*. Retrieved from Fiveable: https://www.citationmachine.net/apa/cite-a-website/new

Frost, J. (2021, November 27). *Identifying Important Independent Variables in Regression Models.* Retrieved from Statistics By Jim: https://statisticsbyjim.com/regression/identifying-important-independent-variables/

Frost, J. (2022, July 22). *Heteroscedasticity in Regression Analysis*. Retrieved from Statistics By Jim: https://statisticsbyjim.com/regression/heteroscedasticity-regression/

Frost, J. (2023, January 29). *Multicollinearity in regression analysis: Problems, detection, and solutions*. Retrieved from Statistics By Jim: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Frost, J. (2024, September 21). *Understanding interaction effects in statistics*. Retrieved from Statistics By Jim: https://statisticsbyjim.com/regression/interaction-effects/

Lucich, M. (2024, July 29). *Addressing heteroskedasticity in time-series modeling with robust standard errors (in python)*. Retrieved from Medium:

https://medium.com/@matt.lucich/addressing-heteroskedasticity-in-time-series-modeling-with-robust-standard-errors-in-python-937317490370

Singh, V. (2024, November 18). *Variance Inflation Factor (VIF): Addressing Multicollinearity in Regression Analysis*. Retrieved from DataCamp: https://www.datacamp.com/tutorial/variance-inflation-factor

The Pennsylvania State University. (n.d.-a). *7.3 – MLR model assumptions. STAT 501: Regression methods.* (The Pennsylvania State University) Retrieved from Penn State. Eberly College of Science: https://online.stat.psu.edu/stat501/lesson/7/7.3

The Pennsylvania State University. (n.d.-b). *7.4 - assessing the model assumptions: Stat 501*. Retrieved from PennState: Statistics Online Courses: https://online.stat.psu.edu/stat501/lesson/7/7.4

Watts, V. (2022, September 1). *13.5 Testing the significance of the overall model. Introduction to statistics.* Retrieved from eCampusOntario: https://ecampusontario.pressbooks.pub/introstats/chapter/13-5-testing-the-significance-of-the-overall-model/

whuber. (2011, August). *Why is it possible to get significant F statistic (p<.001) but non-significant regressor t-tests?* Retrieved from Stack Exchange - Cross Validation: https://stats.stackexchange.com/q/14528

Appendix A – Code and Notebooks

A.1. retrieving_variable_codes.ipynb

This notebook retrieves ACS variable codes from variables metadata to assist API request construction.

A.2. data_extraction_&_data_wrangling.ipynb

This notebook covers the process of acquiring and cleaning data.

A.3. main_analysis.ipynb

Main analysis notebook includes EDA, data preparation, model building, model validation, final model refinement, hypothesis testing and visualizations.

A.4. model_refinement_experiments.ipynb

This notebook explores various model refinement strategies such as transformation testing, interaction term, and robust standard errors.

Appendix B – Variable Metadata

Each variable label was mapped to its corresponding variable code using these metadata resources:

JSON:

- Detailed Tables Variables (JSON):
  https://api.census.gov/data/2023/acs/acs1/variables.json
  Saved as “b_variables.json” in project folder
- Subject Tables Variables (JSON):
  https://api.census.gov/data/2023/acs/acs1/subject/variables.json
  Saved as “s_variables.json” in project folder
- Data Profiles Variables (JSON):
  https://api.census.gov/data/2023/acs/acs1/profile/variables.json
  Saved as “dp_variables.json” in project folder

HTML:

- Detailed Tables Variables (HTML):
  https://api.census.gov/data/2023/acs/acs1/variables.html
- Subject Tables Variables (HTML):
  https://api.census.gov/data/2023/acs/acs1/subject/variables.html
- Data Profiles Variables (HTML):
  https://api.census.gov/data/2023/acs/acs1/profile/variables.html

Appendix C – Baseline Model Summary

Baseline Model's OLS Regression Results

OLS Regression Results

Dep. Variable:	poverty_rate	R-squared:	0.763
Model:	OLS	Adj. R-squared:	0.761
Method:	Least Squares	F-statistic:	307.1
Date:	Fri, 18 Apr 2025	Prob (F-statistic):	1.04e-203
Time:	15:33:43	Log-Likelihood:	-1513.3
No. Observations:	674	AIC:	3043.
Df Residuals:	666	BIC:	3079.
Df Model:	7		
Covariance Type:	nonrobust		

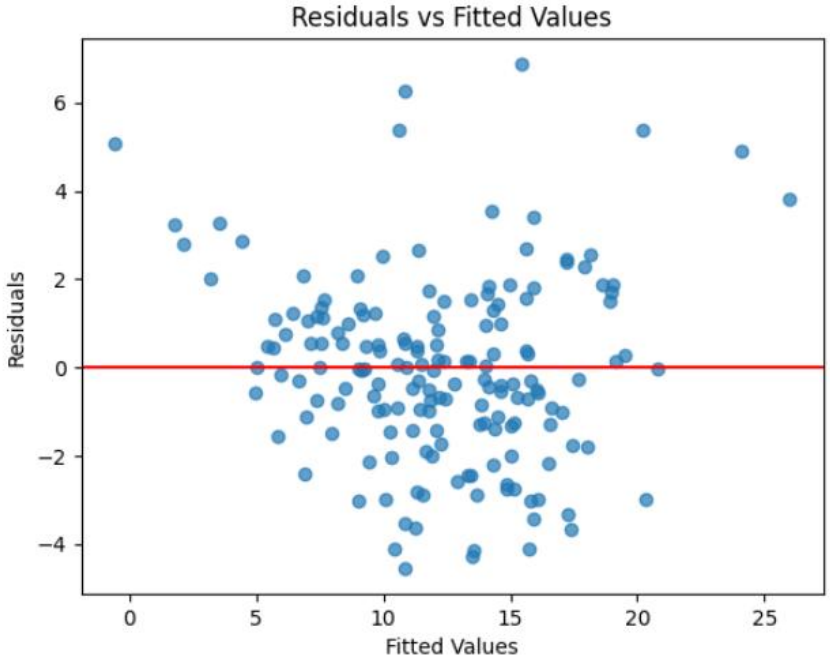
	coef	std err	t	P>|t|	[0.025	0.975]
const	20.8661	0.641	32.554	0.000	19.608	22.125
health_insurance	-2.7029	0.710	-3.807	0.000	-4.097	-1.309
unemployment_rate	6.7305	0.769	8.753	0.000	5.221	8.240
bachelor_holders	1.2994	0.854	1.521	0.129	-0.378	2.977
public_assistance	0.1061	0.838	0.127	0.899	-1.540	1.752
log_median_income	-26.6265	1.060	-25.125	0.000	-28.707	-24.546
log_public_transit	5.6042	0.698	8.027	0.000	4.233	6.975
sqrt_median_house_value	4.7093	1.205	3.908	0.000	2.343	7.075

Omnibus:	53.971	Durbin-Watson:	1.804
Prob(Omnibus):	0.000	Jarque-Bera (JB):	121.840
Skew:	0.456	Prob(JB):	3.49e-27
Kurtosis:	4.873	Cond. No.	25.0

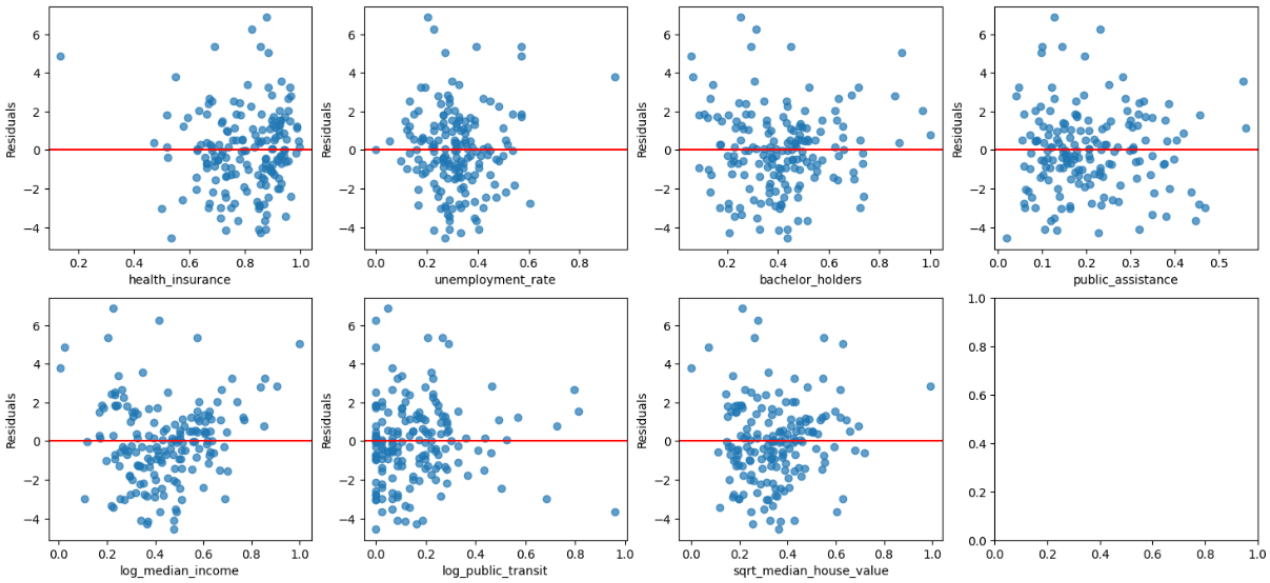
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Appendix D – Baseline Model Diagnostics

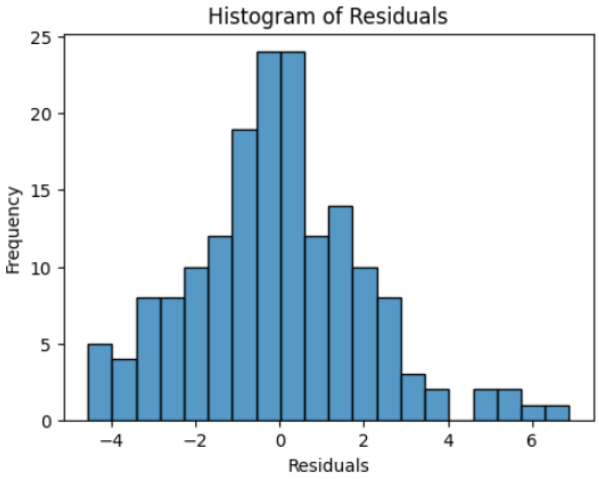
C.1. Residuals vs. Fitted Values

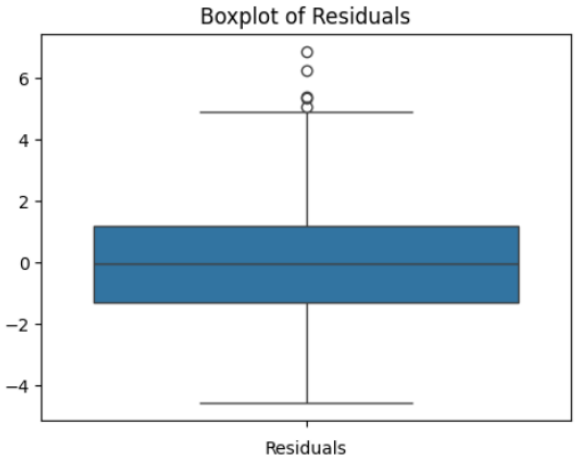


C.2. Residuals vs. Predictors

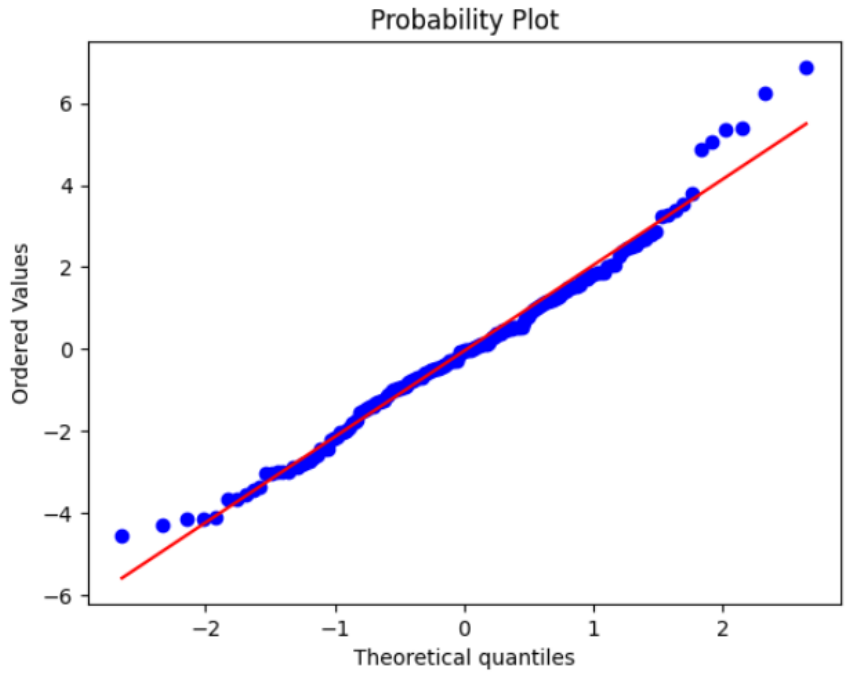


C.3. Histogram and Boxplot of Residuals





C.4. Q-Q Plot



Appendix E – Refined Model Summary

Refined Model's OLS Regression Results

OLS Regression Results

Dep. Variable:	sqrt_poverty_rate	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.793
Method:	Least Squares	F-statistic:	312.2
Date:	Sun, 20 Apr 2025	Prob (F-statistic):	2.10e-219
Time:	18:34:51	Log-Likelihood:	-139.94
No. Observations:	674	AIC:	297.9
Df Residuals:	665	BIC:	338.5
Df Model:	8		
Covariance Type:	HC3		

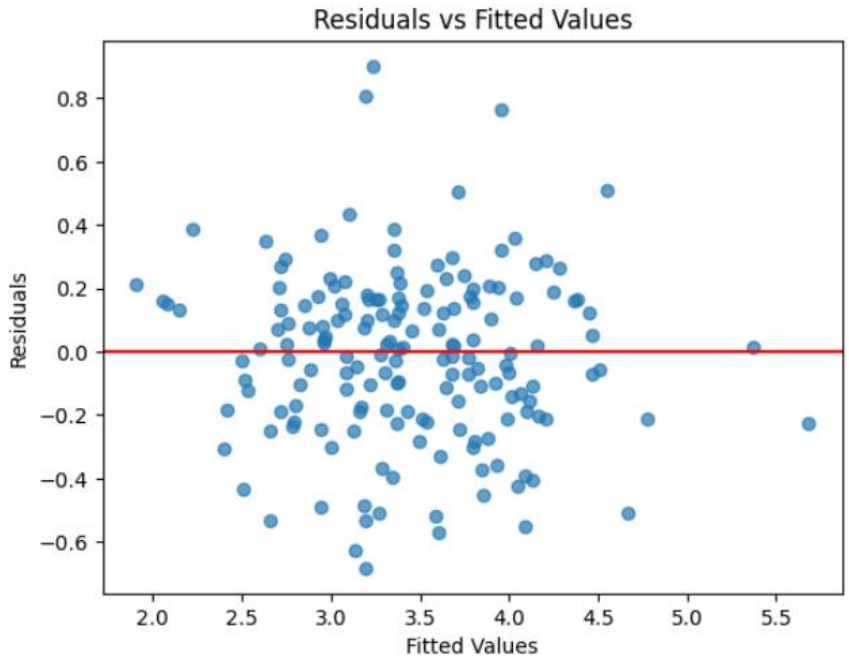
	coef	std err	z	P>|z|	[0.025	0.975]
const	5.0828	0.125	40.570	0.000	4.837	5.328
health_insurance	-0.3674	0.107	-3.445	0.001	-0.576	-0.158
unemployment_rate	0.8111	0.109	7.411	0.000	0.597	1.026
bachelor_holders	0.1332	0.116	1.149	0.251	-0.094	0.360
public_assistance	0.0735	0.120	0.613	0.540	-0.162	0.309
log_median_income	-5.2078	0.252	-20.651	0.000	-5.702	-4.713
log_public_transit	0.7545	0.095	7.940	0.000	0.568	0.941
sqrt_median_house_value	-18.3796	3.244	-5.666	0.000	-24.738	-12.021
income_x_house_value	20.4523	3.399	6.018	0.000	13.791	27.113

Omnibus:	16.933	Durbin-Watson:	1.854
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34.163
Skew:	-0.042	Prob(JB):	3.82e-08
Kurtosis:	4.100	Cond. No.	677.

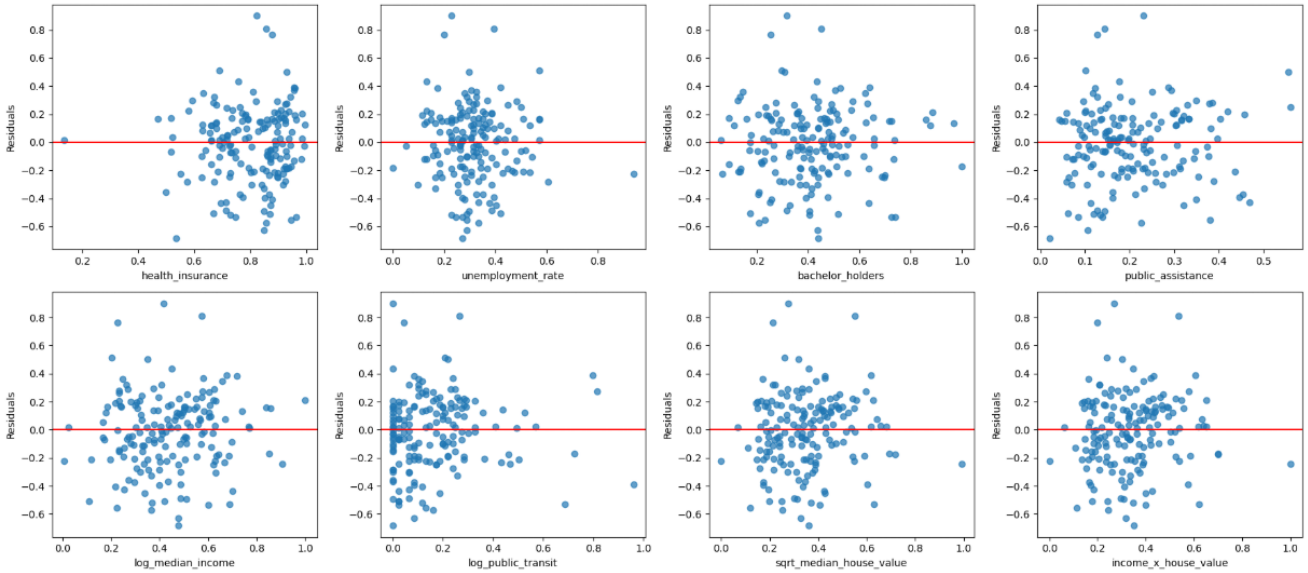
Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Appendix F – Refined Model Diagnostics

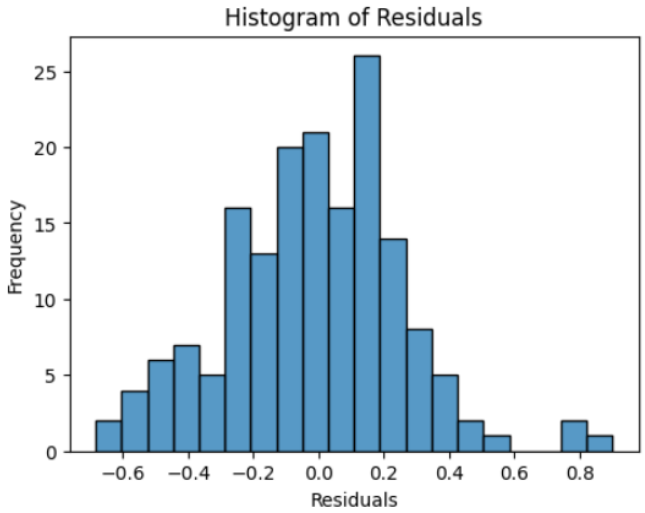
E.1. Residuals vs. Fitted Values

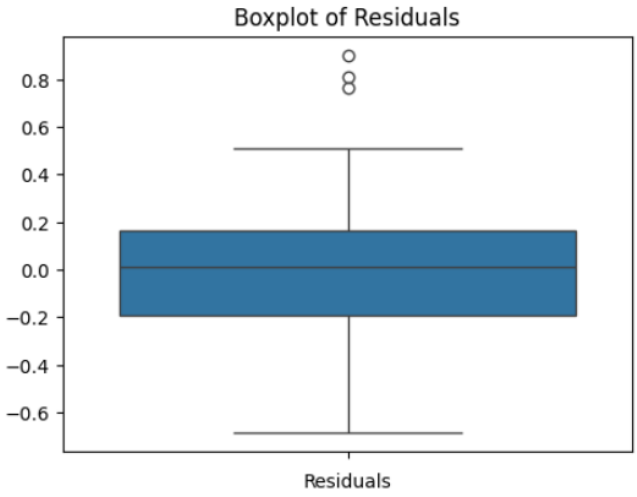


E.2. Residuals vs. Predictors

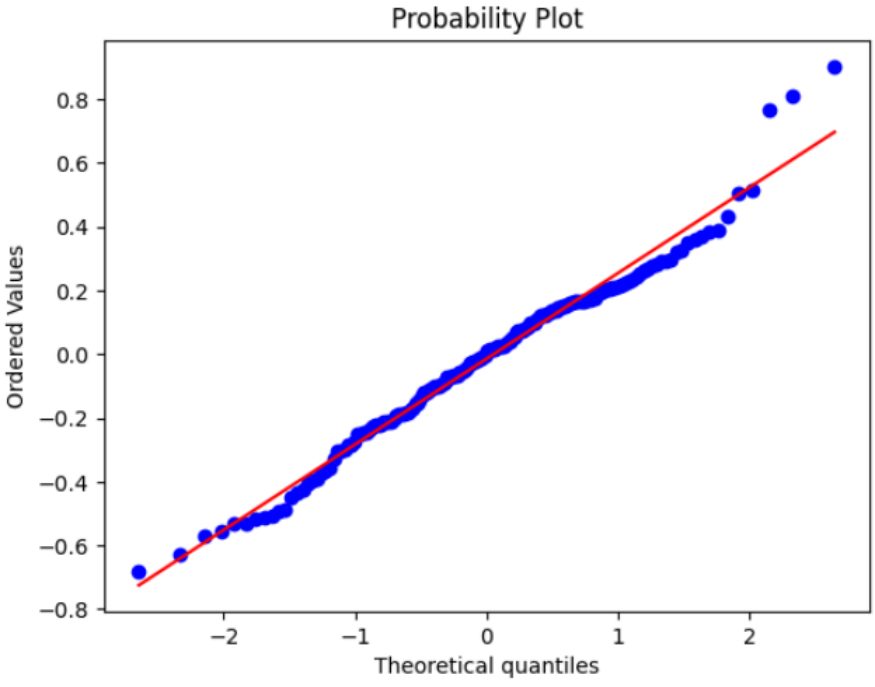


E.3. Histogram and Boxplot of Residuals



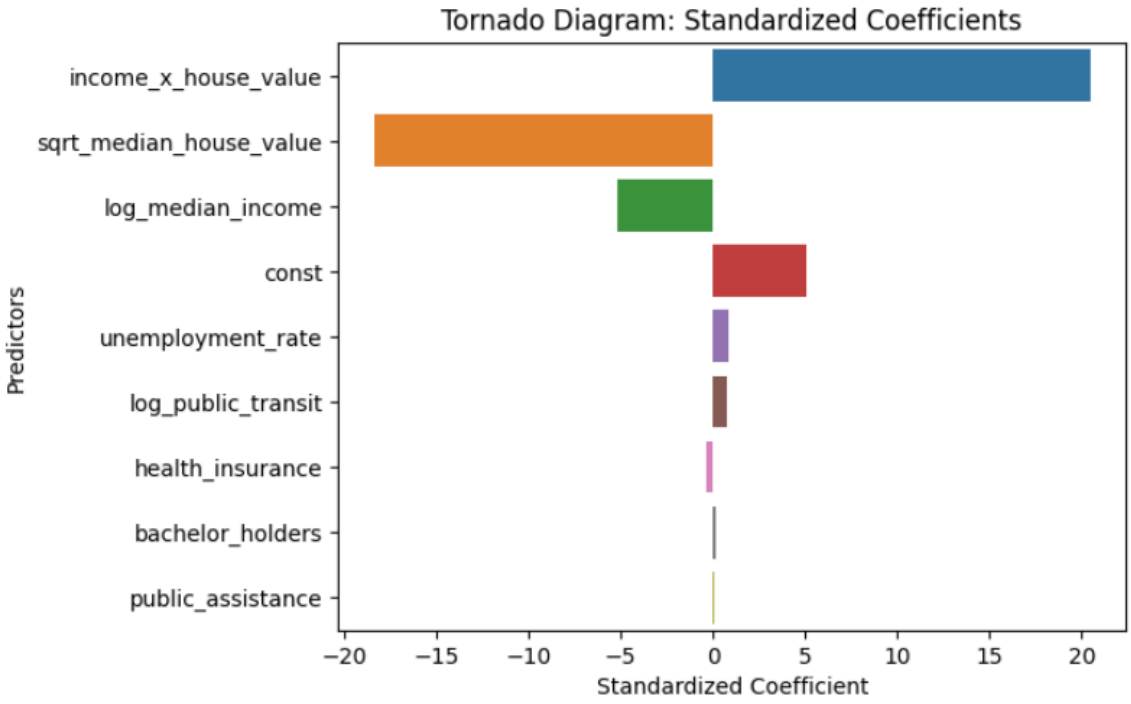


E.4. Q-Q Plot



Appendix G – Tornado Diagram of Standardized Coefficients

Tornado diagram of standardized coefficients displays the relative importance of each predictor variable after standardization.



Appendix H – Slide Deck

The slide deck accompanying the Panopto presentation is submitted as a separate file titled: “Panopto Presentation.pdf”