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

Ngan P. Vu

Western Governors University

Table of Contents

A. Proposal Overview

A.1 Research Question or Organizational Need

This project seeks to answer this question: What is the relative importance of various socioeconomic factors in explaining poverty rates across U.S. counties, and how accurately can these factors be used to predict poverty rates using multiple linear regression?

A.2 Context and Background

Poverty is a complex issue influenced by many factors such as income levels, education, employment rates, healthcare access, public programs, and housing conditions. These variables interact in ways that are not immediately apparent, and their combined effects on poverty rates can vary widely across different regions. Data analysis provides objective insights rather than relying on surface-level perceptions. Through statistical models like multiple linear regression, I can identify and quantify the relationships between these factors and poverty rates. While national statistics provide a high-level view, county-level analysis is essential for identifying localized trends and disparities. Moreover, data analysis enables informed decision-making. By examining this data, policymakers and community leaders can ensure that resources are allocated efficiently and interventions are evidence-based.

A.3 & A.3.A Summary of Published Works and Their Relation to the Project

1. "Poverty in the United States: 2023"

This official report from the U.S. Census Bureau provides comprehensive statistics on poverty in the U.S. in 2023. This report utilizes both the Official Poverty Measure (OPM) and the Supplemental Poverty Measure (SPM). The OPM is based on pre-tax income thresholds that vary by family size and composition, while the SPM accounts for additional factors such as tax payments, work expenses, and non-cash public assistance. In 2023, the OPM reported a national poverty rate of 11.5%, whereas the SPM indicated a higher rate of 12.7% (Shrider, 2024). The results highlight the impact of non-cash benefits and necessary expenses on poverty status.

The dataset I gathered uses the OPM poverty rates across counties. Understanding the difference between OPM and SPM is important, as it highlights the limitations of the OPM in capturing the full range of economic hardship. This awareness helps guide my interpretation of

the regression analysis results and informs the considerations for adding variables that reflect the nuances captured by the SPM.

2. "How Age and Poverty Level Impact Health Insurance Coverage"

This Census Bureau article explores trends in health insurance coverage in the United States in 2023, with a focus on age groups and poverty level. It reports that 8% of the U.S. population lacked health insurance. The study found that children and seniors had higher coverage due to Medicaid and Medicare, respectively. On the other hand, working-age adults and people living below the poverty threshold were more likely to be uninsured (Bunch, 2024).

One of the socioeconomic predictors in this project's dataset is the percentage of people with health insurance. This article highlights the direct connection between poverty and insurance coverage, which supports the idea that health insurance access is a meaningful predictor of poverty. Including this variable in a regression model could help quantify its relationship with poverty rates at the county level.

3. "Multiple Linear Regression"

This tutorial by Michael Brydon provides a practical guide to implementing multiple linear regression using Python. It covers essential steps such as preparing data, handling categorical variables through dummy encoding, and building regression models using the statsmodels library. The tutorial emphasizes the importance of understanding the underlying assumptions of regression analysis and offers useful tips on interpreting model outputs, such as coefficients and statistical significance (Brydon, n.d.).

This tutorial aligns directly with the methodology of this project and serves as a valuable resource for conducting multiple linear regression analysis in Python. It offers practical guidance on data preparation, model implementation, and model diagnostics.

A.4 Deliverables for the Data Analytics Solution

The following deliverables will be produced to address the research question:

1. Final regression model: An improved OLS multiple linear regression model.
2. Model evaluation summary: A detailed summary of the model's performance using metrics such as R-squared, adjusted R-squared, MSE, and statistical test results (F-test, t-tests) to show how well the model explains poverty rates.

3. Model diagnostics report: Visual and statistical diagnostics (e.g., residual vs. fitted plots, residual vs. predictor plots, Q-Q plot, histogram of residuals, multicollinearity check) to verify assumptions and model validity.
4. Visualizations: At least two key visualizations to support interpretation and communication of findings, such as regression plots, a tornado diagram of standardized coefficients and residual plots.
5. Jupyter Notebook: At least one notebook that includes code, analysis steps, visualizations, and interpretations.
6. Written report: A well-documented report that outlines the methodology, analysis steps, interpretation of results, and practical implications of the findings.

A.5 Benefits and Support of Decision-Making Process

The data analytics solution will help identify which socioeconomic factors most strongly influence poverty rates across U.S. counties. The model can guide policymakers, public agencies, and community organizations in prioritizing resources and interventions. For example, if the analysis shows that education level and health insurance coverage are significant predictors, decision-makers may choose to invest more in education and healthcare programs.

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

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.

- 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 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: A 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 Scope of Project

B.2.A Included in 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.

B.2.B Not included in Project Scope

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.

B.3 Standard Methodology

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. 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.4 Timeline and Milestones

Milestone or deliverable	Duration (hours or days)	Projected start date	Anticipated end date
Data Collection	3 days	*4/18/2025*	*4/20/2025*
Data Wrangling	2 days	*4/21/2025*	*4/22/2025*
EDA & Feature Engineering	2 days	*4/23/2025*	*4/25/2025*
Baseline Model Development	2 hours	*4/26/2025*	*4/26/2025*
Model Diagnostics	1 day	*4/27/2025*	*4/27/2025*
Model Refinement	1 day	*4/28/2025*	*4/28/2025*
Final Report	2 days	*4/29/2025*	*4/30/2025*
Panopto Presentation	1 day	*5/1/2025*	*5/1/2025*

B.5 Resources and Costs

Personal computer: $0 (already owned)

Python, Jupyter Notebook: $0 (open-source)

Data source (U.S. Census Bureau datasets): $0 (publicly available)

B.6 Criteria for Success

The project will be considered successful if all important analysis steps are 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 success of the project will be measured based on the completion and quality of the analytical process, not the outcome of statistical significance.

C. Design of Data Analytics Solution

C.1 Hypothesis

To evaluate the role of socioeconomic factors in explaining poverty rates across U.S. counties, both a global and individual hypothesis testing framework will be used.

- **Global hypothesis (F-test):** This test evaluates whether the full set of predictors contributes to explaining poverty rates.
    - **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:

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

    - **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:

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

- **Individual hypotheses (t-test for each coefficient):** These tests assess the significance of each predictor.
    - **Null Hypothesis:** The coefficient for predictor i is zero.

$$H_0: \beta_i = 0$$

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

$$H_0: \beta_i \neq 0$$

C.2 and C.2.A Analytical Method

In exploratory data analysis (EDA) and data wrangling steps, descriptive statistics will be used first to summarize and visualize the dataset. By examining the measures of central tendency and measures of spread, I can make informed decisions on data cleaning. This is followed by calculating a correlation matrix, which is an inferential statistical technique to examine the linear relationships between poverty rate and each explanatory variable. This matrix informs decisions about potential variable transformations.

The main analytical approach is a multiple linear regression method, which allows 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 explanatory variables. Specifically, the project will use ordinary least squares (OLS) model which allows a straightforward interpretation of coefficients, significance testing and model evaluation.

To evaluate the overall predictive power of the model, I will conduct an F-test to determine whether the combination of predictors explains a significant proportion of variance in response variable (Watts, 2022). 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 will be used to assess the statistical significance of each predictor's coefficient. This type of statistical test helps 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 should be used to analyze the predictive power of independent variables.

To evaluate the relative importance of each socioeconomic factor in explaining poverty rates, I also use 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 are obtained by scaling all independent variables using MinMaxScaler or StandardScaler before fitting the model. This will ensure that the predictors are on the same scale.

Finally, to verify model assumptions and ensure the model's validity, I will implement residual analysis. 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.). The analysis will include visualizations such as residual plots, histogram or boxplot of residuals, and Q-Q plot. This diagnostic step helps confirm whether the use of OLS and the resulting hypothesis tests are valid.

C.3 Tools and Environments

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.

C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

1. Multiple linear regression model
    - Model type: Supervised regression model
    - Algorithm: Ordinary least squares (OLS) model using statmodels in Python.
    - Metrics:
        - R-squared: Measures the proportion of variance in the dependent variable explained by the predictors.
        - Adjusted R-squared: A modified version of R-squared that accounts for the number of predictors.
        - Mean Squared Error (MSE): Assesses prediction error on the test dataset.
        - Variance Inflation Factor (VIF): Diagnoses multicollinearity among predictors.
        - Model diagnostic visualizations: Evaluate model assumptions such as linearity, homoscedasticity, and normality.
    - Benchmarks for success:
        - 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 should be close to R and greater than 0.6.
        - MSE as low as possible: A low MSE on the test set reflects good predictive accuracy.
        - VIF for each predictor should be under 5, ideally closer to 1.
        - Diagnostic plots show no major violations: Residual plots should suggest linearity and homoscedasticity, and Q-Q plots should show approximate normality.
    - Justification: Multiple linear regression using the OLS algorithm is appropriate because the goal is to model a continuous dependent variable using several predictors. The chosen

metrics are appropriate because they assess key factors for ensuring valid inference, including model fit, predictor importance, multicollinearity, and residual behavior.

2. F-test for overall model significance
    - Null Hypothesis: All regression coefficients are equal to zero.
    - Planned test: The F-test provided in the OLS regression summary (via statsmodels)
    - Alpha (α): 0.05 or 5% is a standard threshold in statistical hypothesis testing
    - Metric: p-value will be used to assess significance. If the p-value < α, the null hypothesis will be rejected.
    - Justification: F-test helps determine whether the model, as a whole, provides a statistically meaningful explanation of the variation in the dependent variable.
3. t-tests for individual coefficients:
    - Null Hypothesis: Each predictor's coefficient is zero.
    - Planned test: t-tests on each coefficient as part of the OLS regression output.
    - Alpha (α): 0.05 or 5% is a standard threshold in statistical hypothesis testing
    - Metric: p-value for each predictor will be used to assess significance. If the p-value < α, the null hypothesis will be rejected.
    - Justification: t-tests help evaluate the significance of individual coefficients in a regression model. They support the research goal of understanding the relative importance of each socioeconomic variable.

C.5 Practical Significance

To assess the practical significance of the data analytics solution, I will evaluate whether the insights generated are meaningful for making decisions regarding poverty reduction policies. The practical significance lies not just in predicting poverty, but in identifying which levers are likely to produce meaningful improvements in real communities.

The use of standardized coefficients allows for a comparison of the relative impact of each socioeconomic variable on poverty rates. If certain variables show a strong practical effect after controlling for others, policy-makers can use them as high-impact levers for intervention.

Another criterion is whether the model explains a large portion of the variation in poverty rates. A high adjusted R-squared value on the test set indicates that the socioeconomic variables

have statistically predictive power and will be potentially useful for forecasting poverty trends or allocating resources.

C.6 Visual Communication

All visualizations will be created in Python using libraries such as Seaborn, Matplotlib, and Statsmodels, within a Jupyter Notebook environment.

1. Pairplot (Seaborn): To visualize pairwise relationships and potential nonlinear patterns between the poverty rate and each independent variable. This will help highlight non-linear relationships and guide transformation choices.
2. Scatter plot or regression plots (Seaborn): To visualize the individual relationship between the dependent variable (poverty rate) and specific predictors.
3. Histograms of variables (Seaborn): To assess the distribution and skewness of each variable, including the dependent variable (poverty rate) and independent variables.
4. Tornado diagram of standardized coefficients (Matplotlib): To display the relative importance of each predictor variable after standardization. The code for generating a tornado diagram is adapted from Brydon (n.d.).
5. Residuals vs. fitted values plot (Matplotlib): To evaluate model fit and assumption validity, such as linearity and homoscedasticity.
6. Residuals vs. each predictor scatter plot (Matplotlib): To detect potential non-linear relationships or predictors that could benefit from transformations.
7. Histogram of residuals (Seaborn/Matplotlib): To assess whether residuals are normally distributed.
8. Q-Q plot of residuals (Statmodels): To assess normality of residuals by comparing their distribution to a theoretical normal distribution.

D. Description of Dataset

D.1 Source of Data

The dataset used in this project is derived from the U.S. Census Bureau's American Community Survey (ACS) 2023 1-Year Estimates, a nationally recognized source of

socioeconomic data. The dataset was constructed by combining variables from three types of ACS tables (United States Census Bureau, 2024-a):

- Detailed tables: Contain the most granular estimates across a wide range of topics for all geographies.
- Subject tables: Present data organized by specific subject areas, offering both raw values and percentages.
- Data Profiles: Provide high-level snapshots of demographic, social, economic, and housing characteristics.

These tables are publicly available and can be accessed through data.census.gov.

D.2 Appropriateness of Dataset

This dataset is very appropriate for the goals of the project. The ACS provides reliable and up-to-date estimates on a wide range of social and economic indicators, such as education, income, unemployment, housing, and public assistance. These estimates are directly relevant to understanding the predictors of poverty. Additionally, its nationwide county-level coverage allows for insights across diverse U.S. regions.

However, as I mentioned earlier in part A3, one limitation 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.3 Data Collection Methods

The data were collected by making API requests to the U.S. Census Bureau's ACS 2023 1-Year datasets. To construct API request URLs, I will first need to identify variable codes and geographic code (ucgid) for county-level data. The process involved several key steps:

1. Identifying variables:

Relevant variables were selected by browsing tables and variable labels on data.census.gov. Each selected variable label was then mapped to its corresponding variable code using the following variable documents (United States Census Bureau, 2024-a):

- Detailed Tables
- Subject Tables
- Data Profiles

2. Identifying geographic code:

To collect data for all U.S. counties, the API requests include the ucgid predicate which is a geographic identifier used to define the scope of the data request. Based on the guide and resources from the U.S. Census Bureau regarding ucgid predicate (United States Census Bureau, 2024-b), the ucgid that returns data for all counties within the United States is: 0100000US$0500000.

- 0100000US: Fully qualified GEOID of the United States (parent geography)
- $: A separator that links parent and child geographic levels
- 0500000: The summary level code for counties

3. Data assembly:

The data retrieved via API (in JSON format) were transformed into a DataFrame and variables from multiple ACS tables were joined to create a unified dataset.

D.4 Observations on Quality and Completeness of Data

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. While the ACS collects data across the U.S., this dataset includes only 854 counties based on the availability of data for all selected variables. The data were mostly complete, with few missing values, which were handled during preprocessing. Some outliers were identified among counties in Puerto Rico, which will be removed due to the region's distinct socioeconomic characteristics. Remaining outliers should be retained because they are likely to be legitimate, given the careful data processing methods deployed by the U.S. Census Bureau.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

- Data Governance: The ACS dataset is publicly maintained by the U.S. Census Bureau. The data are structured, well-documented, and updated annually.

    - Precaution: I will make sure to document the source of each variable, including the table ID and ACS documentation, to ensure traceability and reproducibility.
- Data privacy and security: All data used are aggregate and non-identifiable**,** representing county-level statistics with no personal identifiers.
    - Precaution: I will avoid inferring information about individuals or small populations.
- Ethical, legal, and regulatory compliance: The data are publicly available under the U.S. government's open data policy.
    - Precaution: I will properly cite the U.S. Census Bureau as the data source in any reports, presentations, or publications.

References

Brydon, M. (n.d.). *Multiple linear regression in Python*. Retrieved from Simon Fraser University: https://www.sfu.ca/~mjbrydon/tutorials/BAinPy/10_multiple_regression.html

Bunch, L. K.-S. (2024, September 10). *How age and poverty level impact health insurance coverage*. Retrieved from U.S. Census Bureau: https://www.census.gov/library/stories/2024/09/health-insurance-coverage.html

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

Shrider, E. A. (2024). *Poverty in the United States: 2023 (Current Population Reports, P60-283).* U.S. Census Bureau. Retrieved from https://www2.census.gov/library/publications/2024/demo/p60-283.pdf

The Pennsylvania State University. (n.d.). *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

United States Census Bureau. (2024, September 12). *American Community Survey 1-Year Data (2005-2023)*. Retrieved from United States Census Bureau: https://www.census.gov/data/developers/data-sets/acs-1year.html

United States Census Bureau. (2024). *Ucgid Predicate: Alternative Option to Specify Geographies*. Retrieved from United States Census Bureau: https://www.census.gov/data/developers/guidance/api-user-guide/ucgid-predicate.html

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