Modeling Poverty Rates Using Socio-Economic Factors Across US States

Hung Tran

December 14, 2023

1. Executive Summary

This report presents a comprehensive analysis of socio-economic disparities across US states, emphasizing the distribution of wealth and its implications for development. By examining key variables such as population, education, and economic diversification in 2020, this study employs factorial analysis to identify significant variables and group similar ones. This methodology allows for the creation of clusters of states with similar characteristics, providing insights into which areas would benefit most from targeted development projects. The results reveal distinct patterns in socio-economic disparities, highlighting potential areas for policy intervention and investment.

2. Introduction

In the United States, socioeconomic disparities manifest in various forms across different states. This report aims to analyze these disparities by examining a range of variables including population, education level, and economic diversification. Understanding these disparities is crucial for formulating effective development strategies. The approach involves a factorial analysis to distill significant variables and group them, leading to the clustering of states with similar socio-economic profiles. This analysis not only identifies areas requiring immediate attention but also envisions potential transformations post-project implementation.

3. Methods

3.1. **Mathematical Model.** To analyze the relationship between poverty rates and factors such as population, education, and income, we established several assumptions to make the problem solvable. We assumed that these variables, derived from 2020 U.S. state data, are significant predictors of poverty rates. We presumed a linear relationship between the dependent and independent variables, acknowledging that real-world scenarios might exhibit non-linear dynamics.

Variables and Units:

- Population in 2020 (Population): Total state population, in persons.
- Education in 2020 (Education): Likely represents educational attainment, in persons.

- Median Household Income in 2020 (Income): In U.S. dollars.
- Poverty Rate in 2020 (Poverty Rate): Percentage of the population living in poverty.

The governing equation of our linear regression model is:

$$\text{Poverty Rate} = \alpha + \beta_1 \times \text{Population} + \beta_2 \times \text{Education} + \beta_3 \times \text{Income}$$

where:

- $\alpha = 24.2973$ is the intercept,
- $\beta_1 = -5.9 \times 10^{-8}$ is the coefficient for Population,
- $\beta_2 = 2.2 \times 10^{-7}$ is the coefficient for Education,
- $\beta_3 = -1.99 \times 10^{-4}$ is the coefficient for Income.

3.2. **Solution Process.** We employed linear regression, a fundamental statistical approach for modeling the relationship between a scalar response and one or more explanatory variables. The model was fitted using the least squares method, minimizing the sum of the squares of the differences between observed and predicted values. This method was chosen for its simplicity, interpretability, and widespread usage in statistical modeling.

3.3. **Model Analysis and Assessment.** The model's effectiveness was evaluated by comparing predicted poverty rates with actual rates. We observed that while the model captures the general trend, discrepancies exist, likely due to unaccounted variables or inherent data variability. The model's sensitivity primarily hinges on the accuracy of the input data and the assumption of linear relationships. We did not explore non-linear models or interactions between variables, which could be areas for future investigation.

4. Results

State	Actual Poverty Rate (2020)	Predicted Poverty Rate
Alabama	14.9%	14.0%
Alaska	9.6%	8.4%
Arizona	12.8%	12.1%
Arkansas	15.2%	14.4%
California	11.5%	11.3%

Figure 1. Table of Actual vs. Predicted Poverty Rates in 2020.

The predicted poverty rates are relatively close to the actual rates, indicating that the model is reasonably accurate in estimating poverty rates based on the population, education, and median household income of these states. However,

there are still differences between the predicted and actual values, which is expected in any statistical model. These differences could be due to various factors not accounted for in the model or due to the inherent variability in social-economic data

The linear regression model revealed the following:

- A very slight negative correlation between population and poverty rates.
- A small positive correlation with the education metric, which may indicate confounding factors or data representation issues.
- A stronger negative correlation with median household income.

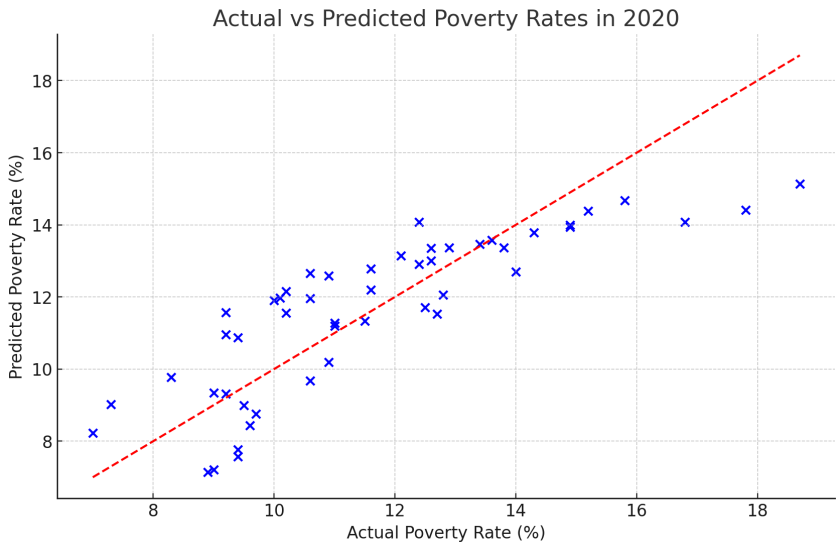

FIGURE 2. Scatter plot of Actual vs. Predicted Poverty Rates in 2020 shows reasonable accuracy with some deviations.

The hierarchical clustering displayed in the dendrogram offers a visual representation of the states' groupings based on their similarities across a multi-dimensional space defined by population, education, and economic diversification metrics. In this dendrogram, each state is linked at a certain height, which represents the distance or dissimilarity between clusters. The smaller the height at which two states are joined, the more similar they are concerning the variables considered.

5. Discussion

The findings of our study offer a thought-provoking view into the dynamics of poverty in relation to key socio-economic factors. The positive correlation

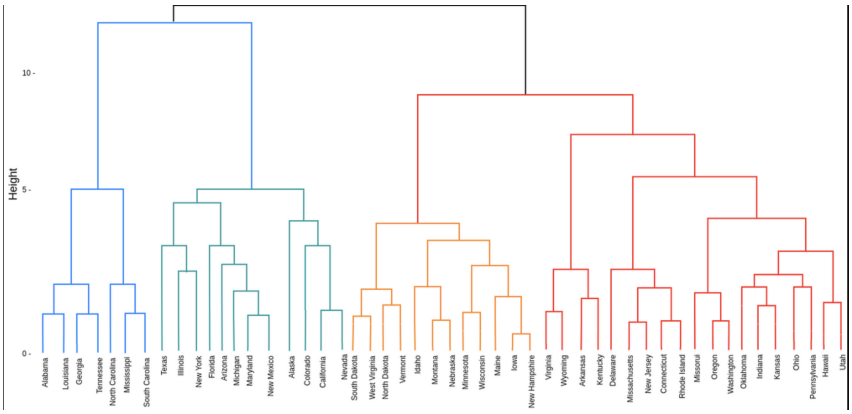

FIGURE 3. Cluster all states into a Dendogram ending with 50 leaves.

between education and poverty rates was an unexpected result, suggesting a more intricate relationship than initially hypothesized. This could indicate that while higher education levels generally lead to better employment opportunities and income, they may not directly translate into reduced poverty rates across different states. Factors such as the quality of education, the relevance of educational programs to the job market, and regional economic conditions might play significant roles.

The dendrogram facilitates an intricate understanding of the socio-economic fabric that weaves the states together. Notably, the grouping of states may reflect not only their current socio-economic status but also historical, geographical, and policy-driven influences that shape these metrics. The proximity of states within the dendrogram suggests that despite geographical dispersion, there are inherent similarities in their socio-economic structures and outcomes. The clusters identified through this analysis could signify shared economic sectors, educational systems' characteristics, or population profiles.

Furthermore, the model's strong negative correlation between median household income and poverty rates aligns with conventional economic theories. This relationship underscores the critical importance of economic growth and income distribution in poverty reduction efforts. However, it also highlights the complexity of poverty as a multi-dimensional issue, influenced by a myriad of factors beyond mere income levels, such as healthcare access, social services, and regional economic policies.

The limitations of the linear regression model used in this study are particularly evident in its inability to capture these multi-dimensional and non-linear

aspects of poverty. The assumption of linear relationships, while simplifying the analysis, may not fully represent the real-world complexities of socio-economic interactions. Future research could benefit from exploring more sophisticated models, such as logistic regression or machine learning algorithms, which can handle non-linearity and complex interactions between multiple variables.

6. Conclusion

In conclusion, this report has provided valuable insights into the relationship between poverty rates and key socio-economic factors in the United States. While the linear regression model offers a foundational understanding, it also brings to light the limitations inherent in such an approach. The study emphasizes the need for a more nuanced understanding of poverty, considering its multi-dimensional nature.

The correlation between education and poverty rates, in particular, warrants further investigation. It suggests that policy interventions in education need to be carefully tailored to address the specific needs of different communities and regions. Likewise, the strong influence of median household income on poverty rates highlights the need for policies that not only stimulate economic growth but also ensure fair income distribution.

Future research should aim to incorporate a broader range of socio-economic factors, possibly including variables like unemployment rates, healthcare access, and regional economic conditions, to build a more comprehensive model. Additionally, exploring more complex statistical models and data analysis techniques could yield deeper insights and more accurate predictions. By continuing to refine our understanding of these dynamics, we can better inform policies and interventions aimed at reducing poverty and improving the quality of life for all citizens.

7. Code

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import numpy as np



# Load the provided Excel file
file_path = '/mnt/data/US STATE DATA (1).xlsx'
data = pd.read_excel(file_path)
```

```
# Display the first few rows of the dataframe to understand its structure
data.head()


# Transposing the dataframe to make each state a row and each metric a column
data_transposed = data.transpose()
data_transposed.columns = data_transposed.iloc[0]
data_transposed = data_transposed.drop(data_transposed.index[0])

# Extracting relevant columns for linear regression
X = data_transposed[["Population in 2020", "Education in 2020", "Median Household
Income 2020"]]
y = data_transposed["Poverty Rate in 2020"]

# Converting data to numeric values
X = X.apply(pd.to_numeric, errors='coerce')
y = pd.to_numeric(y, errors='coerce')

# Handling any missing values by dropping them
X = X.dropna()
y = y[X.index]  # Ensuring that the y
values correspond to the same rows in X

# Performing linear regression
model = LinearRegression()
model.fit(X, y)

# Coefficients and intercept of the model
coefficients = model.coef_
intercept = model.intercept_

coefficients, intercept



# Using the model to predict poverty rates
predicted_poverty_rates = model.predict(X)

# Adding the predicted poverty rates to the
dataframe to comparison
data_transposed['Predicted Poverty Rate'] = predicted_poverty_rates

# Selecting only the actual and predicted poverty rates to compare
```

```
comparison = data_transposed[['Poverty Rate in 2020', 'Predicted Poverty Rate']]
comparison.head()


# Plotting the actual vs predicted poverty
rates plt.figure(figsize=(10, 6))
plt.scatter(comparison['Poverty Rate in 2020'], comparison['Predicted Poverty Rate']
, color='blue')
plt.title('Actual vs Predicted Poverty Rates in 2020')
plt.xlabel('Actual Poverty Rate (%)')
plt.ylabel('Predicted Poverty Rate (%)')
plt.plot([min(comparison['Poverty Rate in
2020']), max(comparison['Poverty Rate in 2020'])]
,[min(comparison['Poverty Rate in 2020']), max(comparison['Poverty Rate in 2020'])],
color='red', linestyle='--')
plt.grid(True)
plt.show()
```

References

1. U.S. Census Bureau. (n.d.). "Poverty rates by state"