PROJECT TITLE: IMPACT OF EDUCATIONAL ATTAINMENT ON INCOME LEVELS – A MULTIVARIATE LOGISTIC REGRESSION IN STATA

1. Background and Problem Statement

A policy research institute aimed to understand how different levels of education affect individuals' probability of falling into higher income brackets in urban India. The institute had access to a nationally representative household survey dataset. While income is a continuous variable, for practical and policy communication purposes, it was categorized into binary classes: low and high income. Traditional correlation analysis failed to isolate the unique contribution of educational attainment after controlling for other variables such as age, gender, occupation, and region. Therefore, a multivariate logistic regression model was deemed suitable to explore the adjusted effects of education levels on income status.

2. Objectives

- To assess the relationship between education level and the probability of being in the high-income category
- To control for potential confounding variables such as gender, age, region, and occupation type
- To calculate odds ratios for each education level
- To visualize predicted probabilities across education categories
- To provide policy-relevant interpretation and suggest targeted education-based interventions

3. Methodology

3.1 Data Overview

- Source: National Sample Survey (NSSO) Unit-Level Data (2021)
- Sample Size: 15,200 individuals from urban households
- Dependent Variable:
    - Binary income classification: High income = 1 (above ₹7.5 lakh/year), Low income = 0 (below ₹7.5 lakh/year)

- Independent Variables:
    - Education Level (no schooling, primary, secondary, higher secondary, graduate, postgraduate)
    - Gender (male, female, other)
    - Age (continuous)
    - Employment Type (government, private, self-employed, unemployed)
    - Region (north, south, east, west, northeast, central)

3.2 Analysis in Stata

- Logistic regression using the logit command
- Post-estimation margins for predicted probabilities
- Test for multicollinearity using collin command
- ROC curve to assess classification accuracy
- Odds ratios with 95% confidence intervals
- Clustered standard errors at the household level

3.3 Model Specification

logit high_income i.education_level i.gender c.age i.occupation_type i.region, vce(cluster household_id)

4. Results

- **Odds Ratios**
    - Individuals with graduate-level education were 4.2 times more likely to be in the high-income group than those with only secondary education.
    - Postgraduates had the highest odds ratio: 6.1, statistically significant at $p < 0.01$.
    - No schooling showed a negative association with high income (OR = 0.42).
- **Marginal Effects**
    - Probability of high income increased from 12.4% (no schooling) to 65.3% (postgraduate)
    - The largest jump in predicted probability was between secondary and graduate levels

- **Model Accuracy**
    - ROC AUC = 0.71, indicating good model performance
    - Pseudo R-squared = 0.32
- **Significant Confounders**
    - Gender: Males had higher odds of being in high-income group (OR = 1.48)
    - Occupation: Government employees had higher odds compared to private sector

5. Interpretation and Insights

- Educational attainment is a statistically significant predictor of income class, even after controlling for key demographics and employment type
- The transition from secondary to tertiary education is crucial for upward income mobility
- Gender and employment type still influence income status, indicating systemic inequality
- Region-specific variations suggest that educational returns vary geographically

6. Deliverables

- .do file with the full analysis pipeline
- .dta and .csv files of cleaned and transformed dataset
- Visualizations of marginal effects and ROC curve
- Comprehensive Word report with full statistical tables and policy discussion

7. Stakeholder Relevance

Academic:

- Suitable for inclusion in public policy, economics, and development studies coursework
- Demonstrates application of multivariate logistic regression using real data

Corporate/Government:

- Provides insight for HR planning, workforce development, and education-related CSR policies
- Useful for designing targeted upskilling programs and government income-equality policies