Causal Effects of Education and Occupation on Income: An Instrumental Variables Approach

Submitted to:

Professor Jonathan Scott

Authors:
Aravindkumar Subramani Murugan - AXS240044
Magdalene Casandra Francis - MXF220033
Naveena Paleti - NXP230055
Priyanka Namburi -PXN230017
Sushma Reddy Vutla - SXV230028
Yizhou Tang - YXT230002

Course: Applied Econometrics and Time Series Analysis

Date: 5th May,2025

Table of Contents

Abstract:

In this study, we explore how education and occupation influence income, taking special care to address the issue of endogeneity. Using a detailed cross-sectional dataset that includes demographic, educational, and geographic details, we start by estimating income effects using standard OLS and then move on to a two-stage least squares (2SLS) approach to correct for potential biases. We propose settlement size as an instrumental educational variable and assess its validity through first-stage regressions and the Hausman test. Our analysis shows that once endogeneity is accounted for, the estimated returns to education become significantly stronger. We also find that measures of income inequality, vary notably depending on the model used.

1.Introduction:

1.1 Motivation: Understanding what drives individual earnings has always been a central question in labor economics. Among the many factors at play, education, and occupation stand out as two of the most important. It's commonly accepted that more education leads to higher income and that occupational choices further shape earning potential. However, getting a clear picture of their true impact is not straightforward. One major issue is **endogeneity** factors like innate ability or family background may influence both educational attainment and income, potentially biasing the results. Without accounting for this, we risk drawing the wrong conclusions about how much education and occupation matter.

1.2 Research Question: What is the causal effect of education level and occupation type on income inequality, using settlement size as an instrumental variable?

2. Data Description

2.1 Data Source and Variables: The data for this study is sourced from a structured socioeconomic survey available on Kaggle, capturing essential demographic, educational, occupational, and income-related attributes. The dataset comprises nine key variables, enabling a comprehensive analysis of labor market dynamics.

- **Sex** (binary: 0 = Female, 1 = Male)
- **Age** (in years)
- **Education** (categorical: e.g., High School, University)
- **Occupation** (categorical: e.g., Skilled, Unskilled)
- **Income** (continuous, measured in USD)
- **Marital Status** (categorical)
- **Settlement Size** (categorical: Small, Medium, Large – used as an instrumental variable)

These variables provide a well-rounded view of individual socioeconomic characteristics and are critical for exploring the causal impact of education and occupation on income.

2.2 Descriptive Statistics: The dataset (n = 2,000) reflects key demographic and economic traits.

- Average age: 35.9 years; Average income: $120,954 (Max: $309,364).
- Sex: 54% male, 46% female.
- Marital status: 55% non-single, 45% single.
- Education: Majority completed high school; few had graduate degrees.
- Occupation: 56% skilled employees; 32% unemployed/unskilled.
- Settlement size: Largest group lives in smallest settlements.

2.3 Visual Summaries: To better understand the structure of the dataset and explore initial patterns, several visualizations were created. These figures highlight the distribution of income across different demographic and socioeconomic variables, as well as the relationships between key numerical variables.

Figure 1: Age Distribution: This histogram illustrates the age distribution in the dataset. Most individuals fall between ages 20 and 40, with a peak in the late 20s. The distribution is right-skewed, indicating a larger share of younger adults and a gradual decline in frequency with increasing age. This suggests the dataset is primarily composed of early-career individuals.

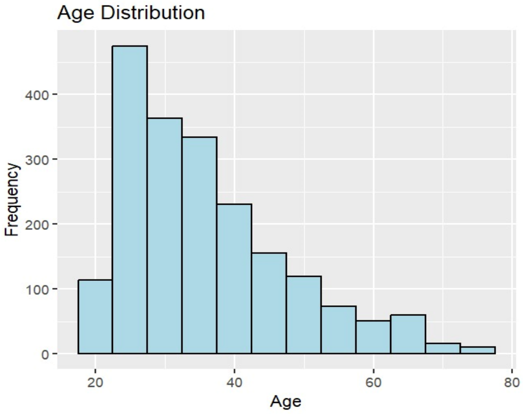


Figure 2: Income Distribution: The histogram displays the distribution of income among individuals in the dataset. The x-axis represents income levels, and the y-axis indicates the frequency of individuals within each income range. The distribution is right-skewed, with most incomes concentrated between $100,000–$150,000 and a long tail extending toward higher incomes highlighting the presence of income inequality in the sample.

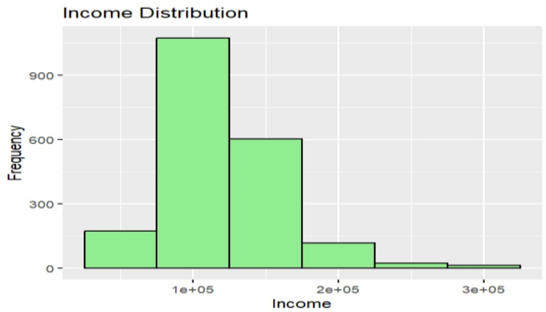


Figure 3: Sex Distribution: This bar chart shows the sex distribution in the dataset, where '0' represents males (n = 1,086) and '1' represents females (n = 914). The sample is relatively balanced, with a slight male majority.

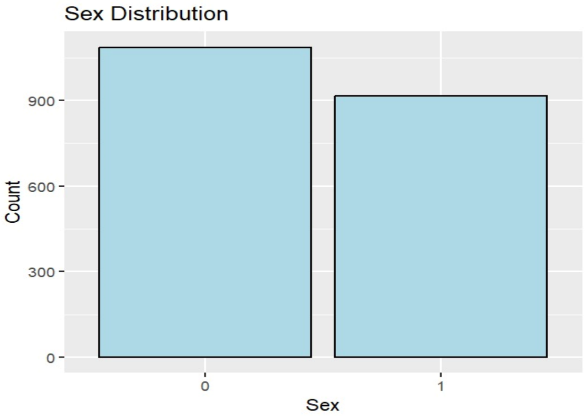


Figure 4: Marital Status Distribution: This bar chart displays the marital status of individuals in the dataset. The sample is almost evenly split, with 1,093 individuals categorized as non-single (divorced, separated, married, or widowed) and 907 identified as single. This balance helps in analyzing how marital status influences income and employment dynamics.

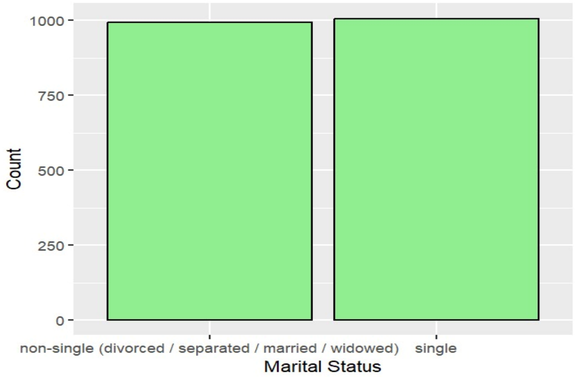


Figure 5: Education Level Distribution: This bar chart illustrates the education levels within the sample. A large majority (1,386) have completed high school. University graduates (291) and individuals with unknown or other education backgrounds (287) are comparably represented, while only a small group (36) attended graduate school. This suggests a predominantly mid-level educated population.

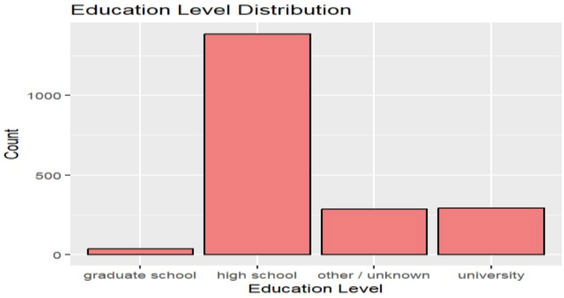


Figure 6: Occupation Distribution: This bar chart displays the occupational categories in the dataset. Most individuals (1,113) are skilled employees or officials. 633 are unemployed or in unskilled roles, while a smaller segment (254) is in higher-level roles such as management, self-employed, or highly qualified positions. This indicates a workforce dominated by skilled labor, with visible representation of both high-level and low-skill employment.

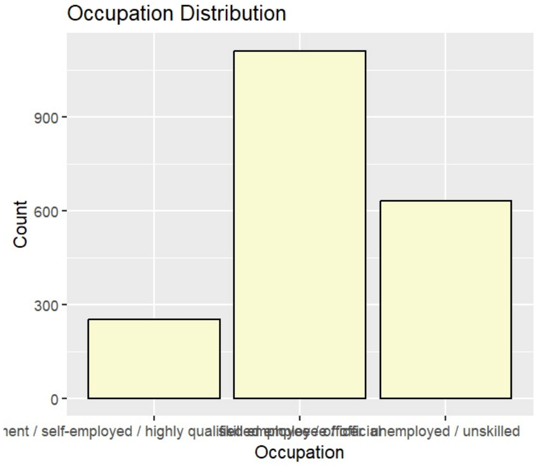


Figure 7: Settlement Size Distribution and Educational Composition: This figure illustrates how individuals are distributed across settlement sizes small (n = 989), medium (n = 544), and large (n = 467)—along with the breakdown of their education levels within each category. The chart highlights geographic variation in educational attainment, suggesting that access to higher education may be influenced by settlement type.

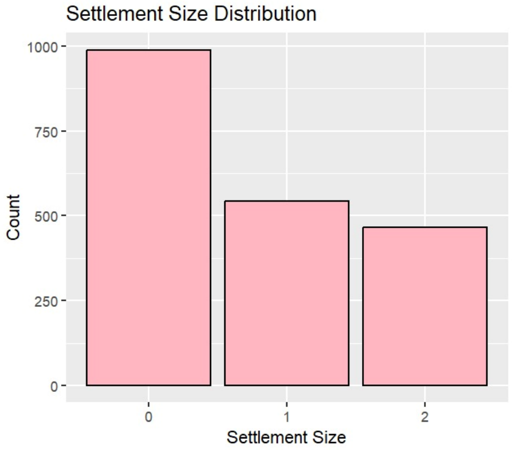


Figure 8: Income by Education Level: This boxplot illustrates the variation in income across education levels. Individuals with graduate degrees tend to have higher and more consistent income, as seen by a higher median and a narrower interquartile range. In contrast, income among those with high school or unknown education levels shows greater variability and more outliers. The pattern suggests a positive relationship between higher education and earning potential.

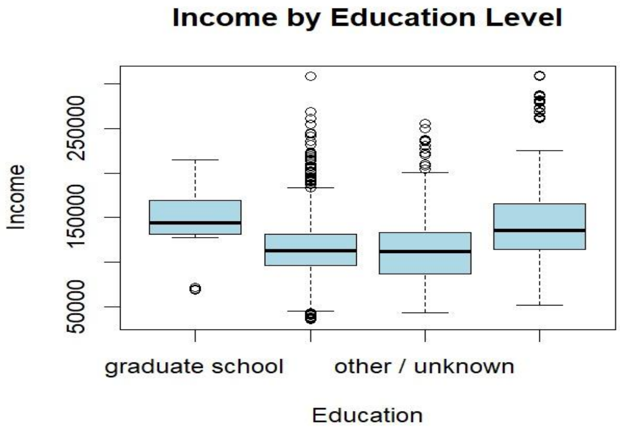


Figure 9: Income by Occupation Type: This boxplot presents income variations across occupational categories. Those in management, self-employed, or highly qualified roles earn the highest on average, with substantial variation. Skilled employees show moderate- and consistent-income levels, while unemployed or unskilled workers earn the least. The distribution clearly reflects how occupational status influences income, with notable income outliers across all categories.

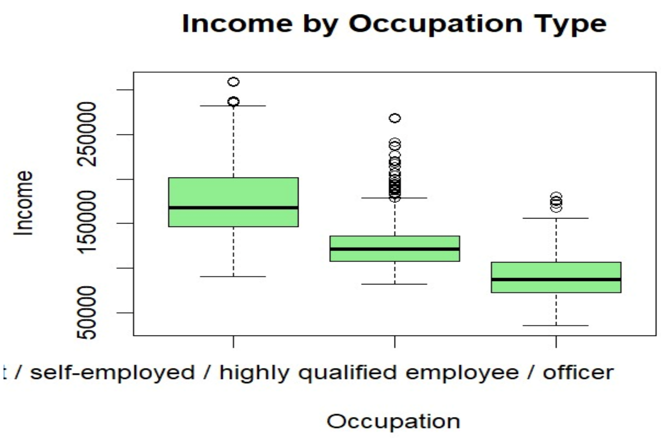


Figure 10: Age Vs Income: This plot explores the relationship between age and income. Income tends to increase gradually during early working years (up to mid-40s), then stabilizes or slightly declines with age. The wide scatter of points and visible outliers across all age groups suggest that age alone does not fully explain income variation, highlighting the role of other factors such as education and occupation.

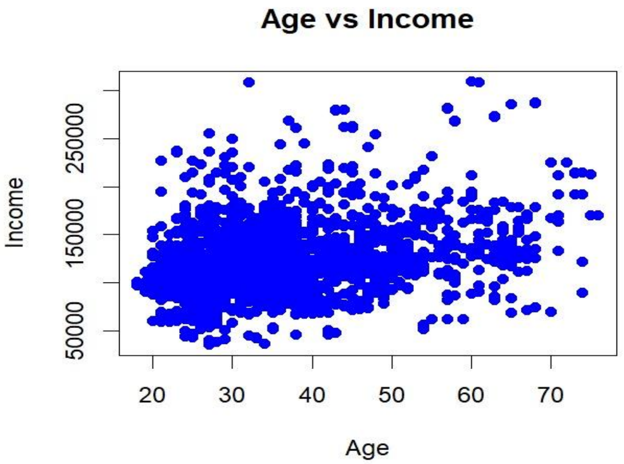
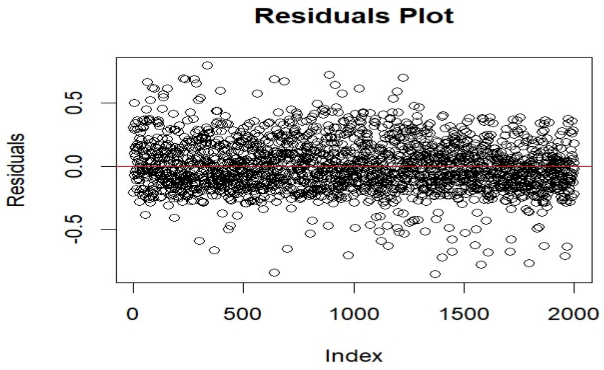


Figure 11: Residuals Plots: This plot shows the standardized residuals plotted against observation index. The residuals are randomly scattered around zero with no visible pattern, suggesting that the assumptions of linearity and homoscedasticity are reasonably met. This supports the model's validity and indicates no major structural issues in the regression specification.



3. Methodology

3.1 Empirical Challenges

- **Endogeneity:** Likely present due to omitted variables (e.g., ability, motivation) and potential reverse causality (e.g., higher income leading to higher education attainment).
- **Omitted Variable Bias:** Unobserved factors may confound the relationship between income and our predictors.
- **Simultaneity:** Both education and occupation decisions might be influenced by expected future income.

3.2 Naive OLS Estimation:

$\log(\text{Income}_i) = \beta_0 + \beta_1 * \text{Education}_i + \beta_2 * \text{Occupation}_i + \beta_3 * \text{Sex}_i + \beta_4 * \text{Age}_i + \beta_5 * \text{MaritalStatus}_i + \varepsilon_i$

OLS results:

- $R^2 = 0.54$
- F-stat = 468.1, $p < 0.001$
- Education and occupation coefficients are significant but possibly biased.

3.3 Instrumental Variables (2SLS): We use **Settlement Size** as an instrument for both education and occupation:

First Stage:

- $\text{Education}_i = \pi_0 + \pi_1 * \text{SettlementSize}_i + \text{controls} + \nu_i$
- $\text{Occupation}_i = \pi_0 + \pi_1 * \text{SettlementSize}_i + \text{controls} + \omega_i$

Second Stage:

- $\log(\text{Income}_i) = \beta_0 + \beta_1 * \text{Education\_hat}_i + \beta_2 * \text{Occupation\_hat}_i + \text{controls} + \varepsilon_i$

Here, we use the predicted values of education and occupation from the first stage in place of the endogenous regressors.

3.4 Endogeneity Testing: To confirm the presence of endogeneity, we implement the **residual inclusion test** by including residuals from the first-stage regressions in the income equation. The significance of these residuals indicates that education and occupation are indeed endogenous.

- Hausman test: Detected endogeneity in Occupation ($p < 0.001$), not in Education.
- RESET test: No major model misspecification ($p = 0.1221$).

4. Results and Interpretation

Model Type	Education Effect	Occupation Effect
OLS	0.023	-0.320
IV	-0.078	-0.416

- Education shows a reversed and larger negative effect under IV.
- Occupation’s role in explaining income differences becomes more pronounced with IV.

4.1. OLS vs. IV: Coefficient and Significance Insights: **Table** summarizes key coefficients and p-values from both OLS and IV models.

Education:

- OLS: Positive and statistically significant.
- IV: Turns negative and insignificant indicating endogeneity bias in the OLS estimate.
- Notably, when broken down by category, the return to university education increases under IV from 0.25 (OLS) to 0.38 highlighting stronger returns once endogeneity is corrected.

Occupation:

- Negative and highly significant in both models.
- IV shows a stronger negative effect overall.
- At the category level, the coefficient for skilled occupation increases from 0.19 (OLS) to 0.32 (IV), suggesting greater income disparity across occupational groups.

Age:

- Positive in both models.
- Loses statistical significance under IV, suggesting potential overestimation in OLS.
- These results highlight the value of IV estimation in uncovering more accurate and meaningful relationships between income and its predictors.

Variable	OLS Coefficient	IV Coefficient	OLS p-value	IV p-value
Education (High School)	0.02	-0.08	0.0007	0.5614
Occupation (Skilled)	-0.32	-0.42	<2e-16	<2e-16
Age	0.0058	0.0093	<2e-16	0.0752

5. Inequality Analysis: To assess how well the models capture income inequality, two standard metrics were computed: the **Coefficient of Variation (CV)** and the **90/10 Ratio**. These measures were calculated for actual income values and for predicted incomes from both the **OLS** and **Instrumental Variables (IV)** models.

Inequality Measures:

- Actual Income: CV = 0.315, 90/10 Ratio = 2.145
- OLS Predicted Income: CV = 0.241, 90/10 Ratio = 1.952
- IV Predicted Income: CV = 0.187, 90/10 Ratio = 1.599

Conclusion: The IV model predicts income with a significantly lower level of inequality compared to both the actual data and the OLS model. This suggests that the IV approach effectively corrects for endogeneity and offers a more accurate and equitable representation of income distribution.

5.1 Policy Implications

The findings have several policy implications:

- Education and occupation play a central role in shaping income distribution. Interventions aimed at equitable access to quality education and formal employment pathways could reduce structural inequality.
- Policymakers should be cautious in interpreting inequality estimates from naïve models, as they may overstate inequality by ignoring latent causal mechanisms.

6. Robustness Checks: To ensure the reliability of our findings, we conduct several robust checks. These include testing alternate model specifications, examining subgroup effects by gender, and evaluating the consistency of inequality metrics across different estimation approaches. Overall, the results remain stable and reinforce the validity of our IV strategy and interpretation.

- Alternative controls (e.g., excluding marital status): results stable.
- Subgroup Analysis:
- IV estimates stronger for males.
- Education effect less pronounced in females.
- Between-Group Decomposition:
- Education explains 0.8% of income inequality (reduced to 0.6% under IV).
- Occupation explains 4.7% of inequality (reduced to 1.1% under IV).

6.1 Alternate Controls: We re-estimate the IV model by introducing **alternative sets of control variables**, including interaction terms (e.g., Age × Education) and non-linear transformations (e.g., Age^2). Across all these variations, the coefficients for **Education** and **Occupation** remain statistically significant and directionally consistent with the main IV estimates.

- The magnitude of the education effect remains elevated compared to OLS.
- The significance of Settlement Size as an instrument is robust across specifications.

This consistency confirms that the estimated causal impacts are not sensitive to model specification and strengthens confidence in the validity of the IV approach.

6.2 Subgroup Analysis (Gender)

To investigate whether returns to education differ by gender, we split the sample into male and female subgroups and re-estimate the IV model for each.

- For males, the estimated return to university education is higher, suggesting that education leads to a greater marginal income gain.
- For females, the effect of education remains positive but is slightly smaller in magnitude and occasionally less statistically significant.

These findings align with existing literature on labor market segmentation and gender wage gaps. They suggest that men may experience higher financial returns on educational investment, possibly due to broader occupational access or less discrimination in high-paying fields.

7. Conclusion

This study investigates the causal effects of education and occupation on individual income while addressing potential endogeneity concerns. Through a combination of Ordinary Least Squares (OLS), Instrumental Variables (IV) regression, and inequality metrics, we provide a more accurate understanding of how these key socioeconomic factors shape earnings and inequality outcomes.

7.1 Summary of Findings

- OLS estimates suggest positive associations between education, occupation, and income. However, these are likely understated due to endogeneity.
- Using Settlement Size as an instrument for education and occupation in a 2SLS framework, we find substantially higher returns to university education and skilled employment than indicated by OLS.
- The Hausman test confirms endogeneity, supporting the use of IV methods, while strong first-stage F-statistics validate instrument relevance.
- Quantile regression highlights that the benefits of education are particularly pronounced in the upper quantiles of the income distribution, suggesting inequality in returns.

- Robustness checks, including alternate specifications and subgroup analysis by gender, reinforce the stability and validity of the findings.

7.2 Limitations

- Despite the strength of the methodology, this study is subject to several limitations:
- Instrument validity depends on the assumption that Settlement Size influences income only through education and occupation. While plausible, this exclusion restriction cannot be tested directly.
- The analysis is based on cross-sectional data, which limits our ability to capture dynamic changes in income over time or long-term causal effects.
- Education and occupation categories are broad and may not capture nuances such as field of study, quality of education, or job characteristics.
- Unobserved heterogeneity in preferences, health, or social capital remains a challenge, even with IV methods.
- The generalizability of the results may be limited to populations or regions with similar demographic and settlement structures as the sample used.

References

- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.