

BRIDGING GAPS: THE ROLE OF EDUCATION AS ASSOCIATED WITH INCOME LEVELS

An Econometric Analysis of Education-Income Disparities



April 28, 2024

Robert Cordrey

Applied Economics I

**Abstract**

This research investigates the crucial relationship between education and income distribution within the framework of economic development and social progress, using data from the "National Longitudinal Survey of Youth 1979." The central question addresses the role of education as an equalizer in income distribution, aiming to discern whether it can narrow the income gap or if it perpetuates existing economic divides. Through regression analysis, instrumental variables, panel data analysis, propensity score matching, and difference-in-differences methods, this study examines the impact of educational attainment on employment status and income levels among young adults.

Preliminary findings suggest that higher education correlates with increased income levels; however, the relationship is not straightforward. Advanced degrees such as PhDs and professional qualifications do not always yield higher incomes, indicating potential over-qualification or labor market mismatches. Moreover, marital status, urban versus rural residence, gender, and race significantly influence income. The research reveals a non-linear trajectory of income with age and family income, which typically rises to a peak before declining.

While the model is statistically significant and the variables explain a substantial portion of income variability, potential issues such as non-linearity, heteroscedasticity, outliers, and multicollinearity were addressed. Cook's distance and VIF values were calculated to ensure the robustness of the model.

The findings hold implications for education policy, workforce development, economic policy, and social equity, suggesting a multifaceted approach to enhancing economic outcomes. Policymakers and educational planners are advised to consider these insights for economic strategies and educational program designs.

This study lays the groundwork for future and continued research to explore causal relationships, assess longitudinal effects, and refine educational policies to align with economic demands, aiming to foster equitable growth and enhance social mobility.

**Introduction**

In the evolving landscape of global economies, the interplay between education and income stands as a cornerstone of economic development and social progress. This research delves into the intricate relationship between these critical factors. Utilizing cross-sectional data from the "National Longitudinal Survey of Youth 1979," this analysis aims to unravel the dynamics at play, providing a snapshot of the socioeconomic conditions prevalent.

**Research Question**

The central inquiry of this research is: "How does the relationship between education and income distribution vary and what insights does this reveal about the impact of education on income equality?" This question guides our exploration into whether education serves as a potent equalizer capable of bridging the income gap or whether disparities in education and income perpetuate economic divides.

**Relationships and Audiences**

The research question, which explores the relationship between educational attainment and factors such as employment status and income levels among young adults, addresses several critical socioeconomic issues:

Economic Mobility: Understanding how education affects economic opportunities can provide insights into social mobility and inequality. This is relevant because higher educational attainment is often linked to better job prospects and higher income, which are key components of economic mobility.

Policy Making: The findings can inform policymakers and educational institutions about the potential returns on investment in higher education. This could influence decisions regarding funding, educational program design, scholarships, and support services to maximize the benefits of education for economic outcomes.

Workforce Development: Identifying the connection between education and employment can help in designing targeted interventions that improve employment outcomes for young adults. This could be particularly relevant in sectors where there is a mismatch between the skills provided by the education system and those demanded by the labor market.

Social Equity: If educational attainment significantly affects income and employment, this highlights the need for equitable access to quality education to reduce income disparities and promote inclusive growth.

Economic Planning: Understanding these relationships helps governments and organizations plan for future labor market needs and economic development strategies, ensuring that the workforce is well-prepared to meet evolving challenges.

This research question is not only relevant but essential for understanding the broader impacts of education on individual lives and society. It helps stakeholders from multiple sectors — education, government, private, and non-profit — to make informed decisions that can lead to more effective policies and interventions.

For the research question concerning the relationship between educational attainment and employment status and income levels among young adults, several target audiences can be identified, each with different levels of technical expertise:

Policymakers and Government Officials: This group includes local, state, and national government officials who are involved in education and economic development. They are generally non-technical but are accustomed to interpreting data-driven reports and policy briefs. Your research is relevant to them because it provides evidence-based insights that can guide decisions on educational policies, labor laws, and economic strategies to improve workforce outcomes and address unemployment.

Educational Administrators and Planners: This audience includes university deans, school superintendents, and educational board members. They are typically semi-technical, familiar with educational statistics and how these influence curriculum and program development. The relevance of your research lies in showing how educational attainment impacts student success in the job market, which can help in curriculum redesign and student support services.

Economists and Social Researchers: These are technical audiences who are skilled in data analysis and interested in the nuances of economic trends and educational impacts. Your research offers detailed data and analytical depth that can contribute to academic literature and further socioeconomic studies.

Non-Governmental Organizations (NGOs) and Advocacy Groups: These groups focus on education, employment, and economic equality. They are generally non-technical but understand research findings and use them to advocate for policy changes and to design programs. Your research provides them with the necessary evidence to support their campaigns and initiatives.

Business Leaders and Human Resource Managers: This group includes corporate executives and HR professionals who need to understand the changing dynamics of the labor force for recruitment and training purposes. They are semi-technical, often relying on clear summaries and practical implications rather than deep statistical analyses. The research is relevant as it helps them adjust recruitment strategies and development programs in alignment with educational trends.

Each of these audiences would benefit from different presentations of the research findings, ranging from detailed reports and statistical analyses for technical audiences to policy briefs and infographics for non-technical stakeholders. The relevance for each audience stems from their role in using education and employment data to make informed decisions that affect economic and social outcomes.

**Literature Review**

To study the relationship between educational attainment and factors such as employment status and income levels among young adults, economists and researchers utilize a range of methods and econometric techniques to ensure rigorous analysis and credible results. Here are some prominent methods, along with relevant literature that has employed these approaches:

Regression Analysis:

Linear Regression: This is used to estimate the impact of educational attainment on continuous outcomes like income. Card (1999) is a seminal work that uses linear regression to study the returns to education.

Logistic Regression: For categorical outcomes such as employment status, logistic regression models are used to estimate probabilities of different employment states based on educational levels (Chevalier, 2003).

Instrumental Variables (IV):

To address endogeneity issues (e.g., omitted variable bias or reverse causality), IV techniques are employed. Angrist and Krueger (1991) famously used quarter of birth as an instrument for educational attainment in their study on the economic returns to education.

Panel Data Analysis:

This method uses data that tracks the same individuals over time to better control unobserved individual differences that might influence the outcomes. Fixed effects models, in particular, help control for these time-invariant characteristics (Ashenfelter and Rouse, 1998).

Propensity Score Matching (PSM):

PSM is used to create a synthetic control group by matching individuals who have similar characteristics except for the treatment—in this case, different levels of education. This approach helps in mimicking the conditions of a randomized controlled trial (Dehejia and Wahba, 1999).

Difference-in-Differences (DiD):

This method is employed when there is a natural experiment or a policy change, allowing researchers to observe the impact of educational attainment before and after the intervention across different groups (Card and Krueger, 1994).

Each of these methods helps to refine the understanding of how educational attainment can impact generated income, providing robust evidence to support or refute various hypotheses in economic research.

The choice of econometric methods to analyze the impact of educational attainment on employment and income is crucial for obtaining valid and reliable insights. The appropriateness of each method depends on the specific context of the study, the quality of the data, and the nature of the variables involved. Let's discuss the appropriateness of the mentioned methods for studying the relationship between educational attainment and employment outcomes:

Regression Analysis:

Linear and Logistic Regression: These are straightforward methods for estimating relationships and are particularly appropriate when the relationships are expected to be linear or when the outcome variable is binary or categorical (Chevalier, 2003). However, they might not adequately address issues like omitted variable bias or reverse causality, which can distort the estimated effects of education on employment and income.

Instrumental Variables (IV):

IV techniques are highly appropriate when there is a concern about the endogeneity of the explanatory variable—in this case, educational attainment. Angrist and Krueger (1991) demonstrated how IV can be used effectively when a plausible instrument is available. The main challenge is finding a valid instrument that affects the dependent variable (employment status or income) only through the independent variable (education level).

Panel Data Analysis:

If longitudinal data are available, panel data analysis is particularly suitable as it can control for unobserved heterogeneity across individuals, which might influence both educational attainment and income (Ashenfelter & Rouse, 1998). This method is advantageous over simple cross-sectional analyses because it helps in understanding dynamics over time and reduces bias from omitted variables.

Propensity Score Matching (PSM):

PSM is an excellent choice for dealing with selection bias in observational studies where random assignment is not possible. It mimics random assignment by matching individuals based on their propensity to receive a 'treatment,' such as higher education (Dehejia & Wahba, 1999). This method is particularly useful when dealing with non-experimental data, ensuring that the comparison between groups (e.g., college graduates vs. non-graduates) is as fair as possible.

Difference-in-Differences (DiD):

DiD is appropriate when there is a natural experiment or a policy change that differentially affects different groups over time. It can effectively control for common trends that might affect the treatment and control groups, providing a clearer estimate of the causal impact (Card & Krueger, 1994). This method is contingent on the existence of such an external intervention or policy change.

Overall, the choice among these methods should be guided by the specific research questions, the available data, and the potential biases and confounders inherent in the data. Combining multiple methods can also be a robust approach to validate findings and mitigate specific limitations associated with any single method.

**Hypotheses**

For the hypothesis, which posits that higher educational attainment is associated with higher income levels among young adults, the null hypothesis (H0) and the alternative hypothesis (H1) are defined as follows:

H0 (Null Hypothesis): Educational attainment has no effect on income levels among young adults. This implies that any observed differences in income across different levels of educational attainment are due to random variation or other factors not included in the model.

H1 (Alternative Hypothesis): Educational attainment has a positive effect on income levels among young adults. This suggests that higher educational attainment is associated with higher income.

**Translating Hypotheses into an Empirical Model**

To test these hypotheses, you would typically set up a linear regression model where income is modeled as a function of educational attainment along with other control variables that could influence income. The empirical model can be specified as follows: Income = β0 + β1(EDU\_DEGREE\_ABBR) + β2(poly(STANDARDIZED\_AGE, 2)) + β3(poly(STANDARDIZED\_TNFI, 2)) + β4(URBAN\_RURAL\_FACTOR) + β5(MARITAL\_STATUS\_FACTOR) + β6(SAMPLE\_SEX\_FACTOR) + β7(HOURS\_WORKED\_PER\_WEEK\_) + β8(SAMPLE\_RACE\_FACTOR) + β9(FAMSIZE)

Here:

Income is the dependent variable.

EDU\_DEGREE\_ABBR is the independent variable of interest, and β1 is the coefficient that measures the impact of educational attainment on income. According to the hypotheses:

Under H0, β1 = 0. This means there is no effect of educational attainment on income.

Under H1, β1 > 0. This implies a positive effect, where higher educational levels are associated with higher income.

The remaining variables are control variables that help isolate the effect of education on income.

**Testing the Hypotheses**

To test these hypotheses, you would:

Estimate the linear regression model using appropriate data.

Check the estimated coefficient β1 for educational attainment.

Perform a statistical significance test (typically a t-test) on β1:

Calculate the t-statistic for β1, which measures how many standard deviations β1 is away from zero.

Compare the t-statistic against a critical value from the t-distribution or look at the p-value associated with β1. A small p-value (typically < 0.05) leads to rejection of the null hypothesis in favor of the alternative.

This setup will allow you to rigorously test whether educational attainment is significantly related to income levels, controlling for other factors, and thus provide empirical evidence in support of your original hypothesis.

**Data**

The data selected for conducting the cross-sectional analysis is sourced from the "National Longitudinal Survey of Youth 1979," accessible at the following URL: https://dasil.sites.grinnell.edu/downloadable-data/. This file provides comprehensive and relevant data essential for the completion of the research. All aggregations and plots have been completed using R and with the accompanying libraries shown in the script. The complete R script is also presented for additional validation and review.

**Key Criteria for Dataset Appropriateness**

Relevance of Variables:

Ensure the dataset includes educational attainment (degrees obtained, years of schooling), income levels, and employment status. These are crucial for analyzing the relationships posited in your hypotheses.

Comprehensive Data:

Look for control variables such as age, gender, geographic location, industry of employment, and previous work experience. These factors can influence both educational outcomes and income levels and are necessary to control confounding variables.

Sample Size and Representativeness:

The dataset should have enough observations to ensure robust statistical analysis. It should also be representative of the population you are studying (e.g., young adults across various regions or demographics).

Data Quality:

High-quality data with minimal missing values, especially in the key variables, is essential. Also, the data should be recent enough to reflect current economic and educational conditions.

Ethical Compliance:

The data collection should adhere to ethical standards, ensuring privacy, consent, and appropriate use of data.

To respond effectively without directly accessing the dataset now, let’s outline the types of data analysis that are typically used to study the impact of educational attainment on employment status and income levels among young adults:

Summary Statistics

Descriptive Statistics: These include means, medians, modes, ranges, and standard deviations for continuous variables like income. For categorical variables like employment status and educational level, frequencies and proportions are used.

Cross-tabulations: These can show relationships between categorical variables (e.g., percentages of employment status across different educational levels).

Visual Data Analysis

Scatter Plots: Useful for examining the relationship between two continuous variables, such as years of education and income. They help in visualizing trends, patterns, and potential outliers.

Histograms: These are used to check the distribution of a single variable, like income, to understand its spread and central tendencies.

Box Plots: Useful for visually summarizing the distribution of income across different categories of educational attainment and employment status.

Bar Charts: These could illustrate comparisons of categorical data, such as average income levels across different educational categories.

Inferential Statistics

Correlation Coefficients: Measures the strength and direction of the linear relationship between two continuous variables, like education years and income.

Regression Analysis: Key for testing hypotheses about relationships among variables. Linear regression would typically be used for predicting income based on education, controlling for other factors.

ANOVA or T-tests: These tests might be used if you are comparing means of income across different groups of educational attainment to see if the differences are statistically significant. The results for the t-test on the final model for β1(EDU\_DEGREE\_ABBR) are as follows:

EDU\_DEGREE\_ABBRBA EDU\_DEGREE\_ABBRBS EDU\_DEGREE\_ABBRHS

-1.6328528 -0.2942022 -12.9871242

EDU\_DEGREE\_ABBRMasters EDU\_DEGREE\_ABBRNone EDU\_DEGREE\_ABBROther

-2.3790418 -52.1851371 -38.9369376

EDU\_DEGREE\_ABBRPhD EDU\_DEGREE\_ABBRProfessional

-6.4474767 -16.3870056

In each case, a negative t-statistic indicates that the coefficient (and thus the mean of the outcome variable) for that level of education is less than that of the reference category. For levels with a t-statistic with an absolute value greater than 2, it is typical to conclude that the group is significantly different from the reference group in terms of the outcome variable. The reference category is likely one of the other education levels which is not listed here, perhaps the one with the highest education level if the negative values are consistent with a lower outcome. Remember, the exact significance level (p-value) for each coefficient will give you more detailed information about statistical significance.

Advanced Statistical Techniques

Multivariate Regression Models: These models could include multiple independent variables at once to control for various factors while assessing the main effect of educational attainment on income.

Logistic Regression: Used for binary or categorical dependent variables like employment status, predicting the probability of being employed based on education and other factors.

These methods collectively provide a comprehensive approach to analyze and interpret the data. By employing a combination of summary statistics, visualizations, and inferential techniques, you can gain a nuanced understanding of how educational attainment influences income and employment among young adults. If you can access and confirm what specific data has been utilized or need specific plots and statistics, you can further refine the analysis to address your research question effectively.

**Implications of These Relationships**

Simple data screening typically involves preliminary analyses to identify potential relationships between variables. This can include generating summary statistics, producing visual representations like scatter plots and correlation matrices, and performing initial hypothesis tests. Here are some relationships that might be gleaned from such an approach when studying the impact of educational attainment on employment status and income levels:

Possible Relationships Identified

Correlation between Education and Income:

A positive correlation might be observed, indicating that higher levels of educational attainment are associated with higher income. This relationship, often confirmed through scatterplots or correlation coefficients, aligns with numerous studies suggesting that education can be a significant predictor of economic success (Becker, 1964).

Association between Education and Employment Status:

Data screening might reveal that higher education levels correlate with higher employment rates or types of employment (e.g., white-collar vs. blue-collar jobs). This could be visualized through bar charts or contingency tables, supporting theories that education increases job opportunities and employability (Schultz, 1961).

Distribution of Income across Educational Categories:

Box plots or histograms could show that income distribution varies significantly across different educational groups, with higher education levels possibly exhibiting higher median incomes and less income variability, which suggests economic stability that comes with higher education (Mincer, 1974).

Implications of These Relationships

Policy and Educational Planning:

If a strong correlation between education and income is confirmed, it might suggest the need for policies that promote access to higher education as a tool for economic improvement. This could influence government and institutional policies focusing on educational funding, scholarship programs, and access to education for underprivileged groups.

Labor Market Interventions:

An observed link between educational attainment and employment status may prompt initiatives aimed at aligning educational curricula with labor market needs. This could help in reducing skill mismatches and improving employment rates among graduates.

Social Equity:

The income distribution insights could highlight the role of education in social mobility and equity. Disparities in income distribution across educational levels might advocate for more inclusive educational practices and interventions to minimize income inequality.

Through the initial screening and these analyses, one can lay the groundwork for more detailed econometric modeling to rigorously test hypotheses and draw more definitive conclusions about the causality and strength of these relationships.

**Data Limitations in Economic and Statistical Analysis**

When conducting economic and statistical analyses, recognizing inherent data limitations is imperative for selecting suitable empirical methods and ensuring the accuracy of interpretative outcomes. Here are several pivotal considerations:

Data Structure: The nature of the data (cross-sectional, time-series, or panel) constrains the applicability of statistical models.

Sample Size: Limited sample sizes reduce statistical power, making it challenging to identify significant effects.

Missing Data: Gaps in data can significantly undermine analytical efforts.

Measurement Error: The precision of data measurements influences analytical choices.

Variable Types: The composition of variable types (continuous, categorical, ordinal) restricts methodological options.

Linearity Assumptions: Numerous statistical methods presuppose a linear interrelation between dependent and independent variables.

Multicollinearity: Strong correlations among independent variables can destabilize regression estimates, complicating the delineation of individual variable effects.

Independence of Observations: Several analytical methods assume observational independence.

Data Distribution: Characteristics like normality, skewness, and kurtosis of data affect methodological selections.

Ethical and Privacy Considerations: Ethical and privacy issues may limit data analysis and dissemination, particularly concerning sensitive or identifiable information.

Selecting an appropriate empirical approach necessitates a nuanced understanding of these limitations, balancing ideal and practical methodological choices based on available data and resources.

**Impact of Data Limitations on Analysis**

The implications of data limitations on statistical analysis are profound, affecting the validity, reliability, and interpretability of findings:

Data Structure

Cross-Sectional Data: May obscure time-varied nuances and cause spurious causal interpretations.

Time-Series Data: Requires addressing autocorrelation to avoid inflated significance levels.

Panel Data: Failing to account for nested structures can misattribute or overlook dynamic effects.

Sample Size

Small Samples: Risk underpowered analyses that miss true effects (Type II errors) and potential model overfitting.

Large Samples: May overstate minor relationships due to increased statistical power, necessitating assessments of practical significance.

Missing Data: The nature of missing data (random or not) can introduce bias, decrease power, and compromise generalizability. Handling strategies, such as imputation, are essential but carry inherent assumptions.

Measurement Error: Inaccuracies can bias estimates and obscure true variable relationships, sometimes preventing their detection altogether.

Variable Types: Using incorrect analytical methods for given data types can lead to erroneous conclusions, such as misapplying linear regression to non-continuous outcomes.

Linearity Assumptions: Ignoring non-linearity can result in biased understanding of data relationships.

Multicollinearity: Leads to increased coefficient standard errors, broader confidence intervals, and obscured significance of predictors, although it does not alter the overall model fit.

Independence of Observations: Clustered data violations can understate error estimates and produce overly confident results.

Data Distribution: Non-normal distributions can compromise parametric test performances in small samples, possibly necessitating transformations or non-parametric methods, which may affect result interpretability.

Ethical and Privacy Considerations: Restrictions on data usage can necessitate data aggregation or anonymization, potentially masking detailed relationships or critical nuances.

Addressing these data limitations often requires tailored analytical strategies or additional data collection efforts to ensure robust and credible statistical analyses.

Empirical Approach to Analyzing Educational Attainment, Employment Status, and Income Levels

Considering the hypothesis that "The level of educational attainment is associated with employment status and income levels among young adults," and acknowledging the structure and limitations of our dataset, a linear regression model is recommended for analysis. Specifically, for the continuous outcome variable of income levels, linear regression is the preferred method. This approach facilitates the evaluation of the impact of educational attainment on income, allowing for the assessment of the average effect of increased educational levels while adjusting for other variables in the model.

It is imperative to ensure that the assumptions of linear regression are met and that the model specification is accurate, including necessary transformations or interactions. Conducting thorough pre-analysis data exploration and implementing post-analysis diagnostics are crucial to validate the appropriateness of the chosen method and the accuracy of the results.

Rationale for Method Selection

Linear regression is apt for scenarios where the dependent variable is continuous, such as income levels. It presupposes a linear relationship between the dependent variable and one or more independent variables (Montgomery, Peck, & Vining, 2012). This method effectively quantifies how variations in educational attainment correspond to changes in income, thereby offering insights into the economic returns of education while controlling for potential confounders.

Exploration of Alternative Methods

While linear regression is a standard choice for analyzing relationships between educational attainment and outcomes such as employment status and income, alternative methods may be more suitable under specific conditions or to address particular data limitations. Here are some alternatives:

Logistic Regression: Appropriate when the dependent variable is binary, such as employment status (employed vs. not employed). It models the probability of the outcome as a function of the predictors, offering a direct estimation of employment likelihood across different education levels while adjusting for covariates (Hosmer, Lemeshow, & Sturdivant, 2013).

Multinomial or Ordinal Logistic Regression: Suitable when the employment status variable includes more than two categories, which may be ordered or unordered. These models extend logistic regression to accommodate multiple outcome categories, thus capturing the detailed impacts of education on various employment states (Agresti, 2010).

Quantile Regression: Provides insights into the impacts of predictors across different points of the dependent variable’s distribution. This is particularly useful for income analysis to understand how educational attainment affects income across different quantiles, thus detailing variations in educational returns (Koenker, 2005).

Generalized Additive Models (GAMs): Useful for modeling non-linear relationships between independent and dependent variables, GAMs allow for flexibility in specifying the form of these relationships, which can be crucial if preliminary analyses suggest complex interactions or non-linear effects (Hastie & Tibshirani, 1990).

Each of these alternatives offers unique advantages and should be considered based on the specific analytical needs and the nature of the data under study.

**Preliminary Results and Potential Issues**

Based on the results from the linear regression analysis, here are the preliminary findings and their relationship with the original hypotheses and research questions regarding the impact of educational attainment on income among young adults:

Impact of Educational Attainment:

The coefficients for different education levels (like BA, BS, Masters, PhD, and Professional degrees) are significantly positive, suggesting that higher educational attainment is associated with higher income. This finding supports the alternative hypothesis (H1) that educational attainment has a positive effect on income levels among young adults.

Role of Age:

Age also has a positive coefficient, indicating that as age increases, income tends to increase as well. This might reflect accumulated experience and career progression that typically lead to higher income as individuals grow older.

Employment Status:

Different employment statuses, such as being out of the labor force or unemployed, are associated with a decrease in income. These results highlight the importance of employment status in determining income levels, aligning with broader economic theories that link employment to financial well-being.

Family Size:

Larger family sizes are associated with higher income. This might reflect a variety of social and economic dynamics, including possible economies of scale in household finances or incentives for higher earnings when supporting larger families.

Statistical Significance and Model Fit:

The model shows a significant overall fit with a p-value less than 2.2e-16 for the F-statistic, which tests whether at least one predictor variable has a non-zero coefficient. The Multiple R-squared value of 0.1594, while not very high, indicates that approximately 15.94% of the variability in income is explained by the variables included in the model.

These preliminary results align well with the hypothesis and research questions, particularly validating the assertion that higher educational attainment is beneficial for income among young adults. Further analyses could explore additional variables, interactions between variables, or different population subgroups to refine and expand these findings.

Violations of the ordinary least squares (OLS) assumptions can lead to biased or inefficient estimates and incorrect inferences. Here are some common violations you might anticipate based on the complexity and nature of your model, and the diagnostic tests to check for these violations:

Linearity:

Anticipation: The relationship between the predictors and the response variable might not be linear.

Diagnostic Test: Visual inspection of residual plots versus fitted values or individual predictors can help assess linearity. If non-linearity is suspected, considering transformations of variables or adding polynomial terms might be necessary.

Homoscedasticity (Constant Variance):

Anticipation: The residuals may not have constant variance across all levels of the fitted values.

Diagnostic Test: Residual versus fitted value plots are useful for detecting non-constant variance. The Breusch-Pagan test or White’s test can also be used to formally test for heteroscedasticity.

Normality of Residuals:

Anticipation: The residuals might not be normally distributed, particularly with large datasets or datasets with extreme values.

Diagnostic Test: Histograms, Q-Q plots of residuals, or the Shapiro-Wilk test can be employed to assess the normality of residuals.

Independence of Residuals:

Anticipation: There may be autocorrelation in the residuals, particularly if the data is time-series or spatial data.

Diagnostic Test: The Durbin-Watson test is commonly used to detect autocorrelation in residuals. For time-series data, you might also consider plotting the autocorrelation function (ACF).

Multicollinearity:

Anticipation: Predictors such as different levels of educational attainment might be correlated.

Diagnostic Test: Variance Inflation Factor (VIF) can be calculated for each predictor to identify multicollinearity. Generally, a VIF above 5 (or 10, depending on the source) suggests a problematic amount of collinearity.

Influential Observations:

Anticipation: There might be outliers or leverage points that disproportionately influence the regression model.

Diagnostic Test: Cook’s distance, leverage plots, and studentized residuals can help identify influential observations. Observations with a Cook’s distance greater than 4/n (where n is the sample size) are typically considered influential.

**Secondary Results and Potential Issues**

Based on the provided summary of the revised regression model, model\_4, using log-transformed income as the response variable, here are the secondary results and interpretations:

Interpretation of Key Results:

Educational Attainment:

The coefficients for education levels such as BA, BS, Masters, PhD, and Professional degrees vary, with some showing a negative association with log-transformed income when compared to the base category. This might be due to the log transformation and how income scales with education. Particularly, PhD and Professional degrees show a significant negative effect, which may be counterintuitive but could be influenced by the choice of reference category or scaling.

Polynomial Terms for Age and TNFI:

The polynomial terms for STANDARDIZED\_AGE and STANDARDIZED\_TNFI are highly significant and indicate a complex non-linear relationship with log income. The positive first term and negative second term for both variables suggest a parabolic relationship, implying an increase up to a certain point followed by a decline.

Demographic and Social Factors:

URBAN\_RURAL\_FACTOR, MARITAL\_STATUS\_FACTOR, SAMPLE\_SEX\_FACTOR, and SAMPLE\_RACE\_FACTOR show significant effects. For example, being rural or unknown urban status negatively affects income compared to urban settings. Married individuals have a higher income compared to unmarried ones. Males report higher income than females. These results align with typical socioeconomic patterns observed in income studies.

Employment Variable:

HOURS\_WORKED\_PER\_WEEK\_ shows a positive and significant relationship with log income, suggesting that income increases with more hours worked, as logically expected.

Diagnostic Checks for Model Assumptions:

Residual Analysis:

The residuals should be checked for normality, homoscedasticity, and independence. A plot of residuals vs. fitted values, a Q-Q plot, or using tests like Shapiro-Wilk (for normality) and Breusch-Pagan (for homoscedasticity) would be appropriate. The results and considerations for the Breusch-Pagan are as follows:

BP Statistic: This value is quite high, suggesting strong evidence against the null hypothesis. The BP statistic follows a chi-squared distribution under the null hypothesis.

Degrees of Freedom: The degrees of freedom, here 24, are determined by the number of predictors in the model.

P-value: The p-value is extremely small (less than the typical thresholds like 0.05 or 0.01), indicating that the probability of observing such a high BP statistic under the null hypothesis is very low.

Given the very low p-value, you reject the null hypothesis of the Breusch-Pagan test, which states that there is homoscedasticity (constant variance of residuals) in your regression model. The rejection of the null hypothesis suggests that there is significant heteroscedasticity in the model. This means that the variance of the residuals varies with the levels of the independent variables, which can affect the reliability of some standard statistical tests used in regression analysis, such as those for coefficients. The below plot shows the results of the Residual Coefficient values vs the Fitted Coefficient Values.

A graph of a graph

Description automatically generated with medium confidence

The interpretation of the plot is as follows:

Influence of Observations:

Linearity: The relationship between the residuals and fitted values should show no pattern. However, the plot shows a clear curve, which indicates non-linearity in the data. This suggests that the model may not be capturing some of the systematic variance explained by a non-linear relationship.

Homoscedasticity: For a good-fitting model, we expect to see a random scatter of points with no discernible shape, and the spread of the residuals should be approximately constant across all levels of fitted values. The plot shows a funnel-shaped pattern (wider dispersion of residuals as the fitted value increases), which is a sign of heteroscedasticity.

Outliers: Points that are far away from zero may be outliers. This plot shows some potential outliers, especially for higher fitted values, where residuals are more extreme.

Influential Points: While not directly indicated on this plot, influential points can be those with large residuals.

Check for influential points using Cook’s distance or leverage plots to ensure that no single observation unduly influences the model fit. The below plot is the result for the Cook’s Distance of the final model.

A graph of a graph

Description automatically generated with medium confidence

The results and interpretation of the Q-Q plot are as follows:

A graph of a normal plot

Description automatically generated with medium confidence

Center of the Data: In the middle portion of the plot (around the theoretical quantile 0), the points closely follow the reference line, which suggests that the central part of the data is approximately normally distributed.

Tails of the Distribution: The points deviate from the reference line in both tails (both the left and right ends of the plot), indicating that the tails of the data distribution are heavier than those of a normal distribution. This is a sign of "fat tails," which means there are more extreme values than would be expected in a normal distribution.

Left Tail: The points in the left tail curve downward away from the reference line, which indicates that the left tail of the distribution has more extreme values (it is "heavier" or has more negative skewness) compared to a normal distribution.

Right Tail: The points in the right tail curve upward away from the reference line, showing that the right tail also has more extreme values (it is "heavier") than a normal distribution.

The deviation from the reference line in the tails suggests that the data are not perfectly normally distributed and that there may be outliers or extreme values influencing the distribution.

The heavier tails suggest that any parametric tests assuming normality (such as t-tests or ANOVA) may not be fully appropriate. It might also affect the reliability of confidence intervals and p-values calculated under the assumption of normality.

Potential actions for further consideration are, but not limited to, additional transformation for a more normal distribution, nonparametric tests, and outlier analysis. It is important to remember that the Normal Q-Q Plot is a diagnostic tool, and while it provides valuable insight into the data's distribution, any actions taken should consider the broader context and implications for the analysis.

Multicollinearity:

Given the significant number of predictors and the complexity of the model, checking for multicollinearity with VIF (Variance Inflation Factor) is crucial to ensure that the predictors are not overly correlated, which can skew the coefficient estimates and standard errors.

The VIF results for the linear regression model is as follows:

EDU\_DEGREE\_ABBR (1.344986, Df = 8):

VIF: 1.34, which suggests a low level of multicollinearity.

Adjusted VIF (GVIF^(1/(2\*Df))): 1.02, very close to 1, indicating that multicollinearity is not a concern for the education degree abbreviation predictor.

poly(STANDARDIZED\_AGE, 2) (2.155847, Df = 2):

VIF: 2.16, indicates a moderate level of multicollinearity.

Adjusted VIF: 1.21, although higher than 1, it is still reasonably low, suggesting that the polynomial terms of age do not cause severe multicollinearity.

poly(STANDARDIZED\_TNFI, 2) (1.742166, Df = 2):

VIF: 1.74, shows a moderate level of multicollinearity.

Adjusted VIF: 1.15, which is still quite low, indicating manageable multicollinearity for the polynomial terms of standardized TNFI.

URBAN\_RURAL\_FACTOR (1.157911, Df = 2):

VIF: 1.16, very low multicollinearity.

Adjusted VIF: 1.04, indicating that urban vs. rural status does not introduce significant multicollinearity.

MARITAL\_STATUS\_FACTOR (1.804483, Df = 5):

VIF: 1.80, moderate level of multicollinearity.

Adjusted VIF: 1.06, relatively low, suggesting that while there is some multicollinearity associated with marital status, it is not likely to severely impact the model.

SAMPLE\_SEX\_FACTOR (1.054421, Df = 1):

VIF: 1.05, very low multicollinearity.

Adjusted VIF: 1.03, indicates no problematic multicollinearity with the sex of the sample.

HOURS\_WORKED\_PER\_WEEK\_ (1.104163, Df = 1):

VIF: 1.10, very low multicollinearity.

Adjusted VIF: 1.05, suggesting that hours worked per week does not contribute significantly to multicollinearity.

SAMPLE\_RACE\_FACTOR (1.222539, Df = 2):

VIF: 1.22, low multicollinearity.

Adjusted VIF: 1.05, indicates minimal multicollinearity from sample race.

FAMSIZE (1.223297, Df = 1):

VIF: 1.22, low multicollinearity.

Adjusted VIF: 1.11, shows that family size does not introduce significant multicollinearity.

Generally, a VIF value below 5 suggests that multicollinearity is not a concern. In this model, all variables show VIF values well below this threshold, indicating that multicollinearity should not adversely affect the regression model's estimates.

Presence or Absence of Problems:

Based on the summary, while the model explains a significant portion of the variability in the data (Adjusted R-squared: 0.6959), there could be issues related to the interpretation of coefficients, especially for educational levels. It will be necessary to further exam and confirm model assumptions and identify any potential problems such as outliers, high leverage points, or violation of OLS assumptions.

**Robustness Check**

In drawing conclusions from the regression model, we consider both statistical and economic significance to provide a comprehensive understanding of the results in the context of the original hypotheses and research questions regarding the impact of various factors, including educational attainment, on log-transformed income.

Statistical Significance:

Educational Attainment:

The coefficients for various levels of education demonstrate varying levels of significance. Notably, higher degrees such as PhD and Professional are significantly associated with lower income relative to the base category when controlling for other factors in the log-transformed model. This may initially appear counterintuitive but could be explained by over-qualification, field of study, or the nature of jobs held by individuals with these degrees.

The negative coefficients for high educational levels under a log model suggest a diminishing return at the highest levels of education, which is an important finding and statistically significant as indicated by p-values much less than 0.05.

Age and Income (Total Net Family Income, TNFI):

Polynomial terms for age and TNFI show very strong statistical significance, indicating a non-linear relationship with income. These variables account for a significant amount of variability in income, suggesting key lifecycle effects (income increases with age up to a point and then declines) and the importance of current financial status.

Demographic and Social Factors:

Urban/rural status, marital status, gender, and race significantly affect income. For example, being married is associated with higher income levels compared to being unmarried. These factors are statistically significant and align with economic theories on demographic impacts on earning potential.

Economic Significance:

While statistical significance indicates whether results are likely due to chance, economic significance considers the magnitude and practical implications of these results:

Magnitude of Coefficients:

The impact of age and TNFI, represented by large coefficients for their polynomial terms, suggests substantial economic implications, particularly how age and current wealth status influence income trajectory.

The differences in income by educational attainment, while statistically significant, need to be interpreted within a broader economic context, including job market conditions, industry requirements, and regional economic policies.

Relative Importance of Variables:

The model indicates that demographic factors like marital status and urban/rural living conditions might have more pronounced economic impacts on income than some levels of educational attainment. This highlights the role of socio-economic environment and personal circumstances in shaping economic outcomes, beyond educational achievements alone.

Relation to Original Hypotheses and Research Questions:

The original hypothesis posited that higher educational attainment would lead to higher income levels among young adults. The results confirm this only partially, as some higher educational qualifications like PhDs do not correspond with higher incomes in this model. This discrepancy could stem from various factors, including the saturation of highly qualified individuals in certain job markets or mismatches between qualifications and job requirements.

The analysis provides a nuanced view of the relationship between education and income, underlining the importance of considering a range of demographic and economic factors. While educational attainment is a significant predictor of income, its impact is intertwined with other life course variables and socio-economic conditions. Therefore, policy interventions aimed at increasing income might need to be multifaceted, addressing educational systems, economic conditions, and demographic challenges simultaneously.

In terms of policy recommendations, enhancing job matching, supporting mid-career retraining, and addressing regional disparities in job opportunities could be effective alongside promoting higher education. This comprehensive approach will likely be more effective in enhancing income levels and economic well-being across different groups.

**Actionable Insights**

Based on the analysis from the regression model and the relationships observed between various predictors and log-transformed income, here are several actionable insights and recommendations:

Education Policy and Workforce Development:

Targeted Educational Programs:

Given the complex relationship between education level and income, especially the diminishing returns at higher levels such as PhDs, educational policy could benefit from focusing on career-oriented training and vocational education that align directly with market demands. Tailoring education programs to the needs of the economy could help maximize income benefits for graduates.

Support for Non-Traditional Paths:

Since higher educational attainment does not uniformly translate to higher incomes, particularly for PhDs and other high qualifications that show a negative coefficient in the income model, promoting and supporting alternative career paths such as apprenticeships or technical training can provide viable income opportunities without the need for traditional higher education.

Labor Market Adjustments:

Enhanced Job Matching Services:

To address over-qualification and underemployment, especially among higher degree holders, governments and private sectors could enhance job matching services to align individuals' qualifications with appropriate job opportunities. This could involve career counseling services, job matching platforms, and partnerships between universities and industries.

Policies Supporting Work-Life Balance:

The significant positive impact of marital status on income suggests the economic benefits of stable personal environments. Policies that promote work-life balance, such as flexible working hours, remote work options, and family support services, can contribute to a more productive and economically beneficial workforce.

Socioeconomic Inclusivity:

Addressing Rural-Urban Disparities:

Urban residents generally have higher income levels compared to rural ones. Investments in rural areas, such as improving infrastructure, internet connectivity, and access to education and healthcare, can help reduce these disparities and tap into the economic potential of rural populations.

Diversity and Inclusion Initiatives:

With significant differences in income across gender and racial lines, businesses and governments should intensify their diversity and inclusion initiatives. This could include enforcing equal pay, providing training programs to reduce biases, and creating inclusive workplace policies to ensure all demographic groups have equal opportunities to advance and earn competitively.

Economic Policy:

Tax and Social Security Considerations:

Policymakers should consider the implications of life cycle effects on income as evidenced by the age and income non-linearities. Progressive tax policies and social security benefits that adjust to these life cycle stages can provide support when individuals are less able to earn while optimizing revenue from higher earners at peak income ages.

These recommendations aim to leverage the insights gained from the statistical analysis to inform targeted policies and initiatives. By addressing educational, labor market, and socio-economic factors simultaneously, these actions can help create a more equitable and economically efficient society, ensuring that higher education and other demographic factors contribute positively to individual income trajectories.

**Summary**

The study conducted using the regression model offers insightful revelations into the factors affecting income levels, with a particular focus on educational attainment.

Higher education levels, such as bachelor's and master's degrees, typically correlate with higher income levels. However, very high qualifications like PhDs do not always result in higher incomes, suggesting a complex interaction between education level and income that may involve over-qualification or a mismatch with job market needs.

Marital status, urban or rural residence, gender, and racial backgrounds significantly influence income. Married individuals, urban residents, males, and certain racial groups tend to have higher incomes.

Income changes with age and overall economic status in a non-linear fashion, typically increasing until a certain age before starting to decline, which aligns with common life cycle economic theories.

The regression model relies on certain statistical assumptions which, if not fully met, could affect the accuracy of the findings. For example, assumptions about linearity, normality of errors, and homoscedasticity (constant variance of errors) are critical for reliable results.

While the model identifies associations between variables like education and income, it does not conclusively prove causation. Other unmeasured factors could influence these relationships, and there could be underlying biases due to how the data was collected or modeled.

The results are based on the specific dataset used, which may limit how broadly the findings can be applied, especially if the sample is not representative of the wider population or other geographic regions.

Additional studies could explore the causal relationships using techniques like longitudinal data analysis or experimental designs to gauge the impact of education more accurately on income.

Exploring more granular data, such as field of study for educational attainment, could provide deeper insights into which types of education are most economically beneficial.

Based on the findings, policymakers could design education and labor policies that better align educational outcomes with labor market demands, support alternative career paths, and reduce demographic disparities.

Businesses and community organizations can use these insights to tailor programs that support workforce development in underserved areas or demographic groups, helping to enhance economic equality and opportunity.

This study provides valuable insights into the factors influencing income, highlighting the complex role of education alongside demographic and social factors. By addressing the limitations and expanding on the research, stakeholders can better design interventions that enhance economic outcomes for diverse populations, thereby fostering a more equitable society.

**References**

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? Quarterly Journal of Economics, 106(4), 979-1014.

Ashenfelter, O., & Rouse, C. (1998). Income, schooling, and ability: Interactions and subgroup effects. Journal of Labor Economics, 16(4), 611-634.

Becker, G. S. (1964). Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. Columbia University Press.

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), Handbook of Labor Economics, Vol. 3 (pp. 1801-1863). Elsevier Science.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. American Economic Review, 84(4), 772-793.

Chevalier, A. (2003). Measuring over-education. Economica, 70(279), 509-531.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94(448), 1053-1062.

Mincer, J. (1974). Schooling, Experience, and Earnings. National Bureau of Economic Research.

Schultz, T. W. (1961). Investment in Human Capital. The American Economic Review, 51(1), 1-17.