Quantifying the Impact of Educational Attainment on Household Income: A Statistical Analysis Using PSID Data

by

Sumasri Jasti

Naga Brahmendra Chowdary

Kodanda Rama Naidu

FINAL PROJECT REPORT

for

DATA 608 Probability and Statistics for Data Science

Zeynep Kacar

University of Maryland Baltimore County

2025

Table of Contents

7. **References**
8. **Appendices**
    - Full Regression Output
    - Visualizations and Interpretation

ABSTRACT

This study investigates the relationship between educational attainment and household income using data from the 1993 cross-sectional Panel Study of Income Dynamics (PSID). Education is often assumed to positively influence income, and this project aims to statistically evaluate that relationship. Using multiple linear regression, we assess how years of education affect income while controlling for age, hours worked, marital status, and number of children. Various data science tools, including Python libraries like pandas, seaborn, matplotlib, and statsmodels, were utilized for data processing, visualization, and statistical modeling. The regression model, validated through residual analysis and confidence intervals, explains 46.5% of income variation and supports the hypothesis that education significantly boosts income.

1. INTRODUCTION

1.1 Background and Motivation

In today's knowledge-driven economy, education is widely regarded as one of the primary factors influencing an individual's economic status. Governments invest in public education systems, individuals pursue higher education, and employers often use educational qualifications as a proxy for skills. This project explores whether increasing educational attainment actually leads to higher household income, and to what extent.

1.2 Research Importance

Understanding the quantitative relationship between education and income is important for:

- Policymakers designing educational subsidies or student loan systems
- Economists analyzing wage inequality
- Students and families making investment decisions in education

1.3 Research Question

The core research question is:

How does educational attainment impact household income, when controlling for other socioeconomic variables?

We aim to quantify this relationship using statistical modeling.

2. DATA DESCRIPTION

2.1 Data Source

The data used in this project comes from the Panel Study of Income Dynamics (PSID), a longitudinal household survey conducted in the United States. For this project, we used a 1993 cross-sectional dataset which contains economic, demographic, and social variables.

2.2 Tools and Libraries Used

The following tools and libraries were used:

- **Python (Jupyter Notebook)**: Primary language for coding and analysis
- **Pandas**: Data manipulation and wrangling
- **NumPy**: Numeric operations and array handling
- **Seaborn & Matplotlib**: Data visualization
- **Statsmodels**: Regression modeling and statistical analysis
- **Scikit-learn**: Considered for potential future non-linear models

2.3 Variables Used

- **Dependent Variable**: `log_income` (log-transformed total income)
- **Independent Variables**:
    - `education`: years of schooling
    - `age`: age of individual
    - `age_squared`: derived feature to capture non-linear effects
    - `hours_worked`: total annual hours worked
    - `children`: number of dependent children
    - `marital_status`: categorical, one-hot encoded (e.g., married, divorced)

	rownames	intnum	persnum	age	educatn	earnings	hours	kids	married
0	1	4	4	39	12.0	77250	2940	2	married
1	2	4	6	35	12.0	12000	2040	2	divorced
2	3	4	7	33	12.0	8000	693	1	married
3	4	4	173	39	10.0	15000	1904	2	married
4	5	5	2	47	9.0	6500	1683	5	married

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4856 entries, 0 to 4855
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   rownames  4856 non-null   int64
 1   intnum    4856 non-null   int64
 2   persnum   4856 non-null   int64
 3   age       4856 non-null   int64
 4   educatn   4855 non-null   float64
 5   earnings  4856 non-null   int64
 6   hours     4856 non-null   int64
 7   kids      4856 non-null   int64
 8   married   4856 non-null   object
dtypes: float64(1), int64(7), object(1)
memory usage: 341.6+ KB
```

2.4 Data Preparation and Cleaning

Data preprocessing steps included:

- Dropping rows with missing values in income, education, and marital status
- Converting income to `log(1 + income)` to handle skewed distributions
- Creating new variables such as `age_squared` and `wage`
- Encoding categorical variables into binary dummy columns for regression
- Filtering out invalid marital statuses (e.g., "NA/DF")

```
print(df.describe())
```

```
          rownames       intnum      persnum          age    education  \
count  3459.000000  3459.000000  3459.000000  3459.000000  3459.000000
mean   2274.803411  4291.284186    60.175195    38.495519    12.895635
std    1368.325326  2716.425501    79.834531     5.443455     2.503169
min       1.000000     4.000000     1.000000    30.000000     0.000000
25%    1089.500000  1722.500000     2.000000    34.000000    12.000000
50%    2218.000000  5234.000000     5.000000    38.000000    12.000000
75%    3411.500000  6412.500000   170.000000    43.000000    14.000000
max    4856.000000  9306.000000   199.000000    50.000000    17.000000

              income  hours_worked     children
count    3459.000000   3459.000000  3459.000000
mean    18964.558832   1633.455912     2.013877
std     15895.935863    720.928025     1.334177
min        13.000000      6.000000     0.000000
25%      8000.000000   1213.000000     1.000000
50%     16000.000000   1856.000000     2.000000
75%     26000.000000   2040.000000     3.000000
max    240000.000000   5025.000000    10.000000
marital_status
married          2294
divorced          476
never married     428
separated         202
widowed            54
NA/DF               5
Name: count, dtype: int64
```

2.5 Final Dataset Summary

- Final row count: 3,454
- Age range: 30–50 years
- Education range: 0–17 years
- Income range (log): approx. 2.5–12.5
- Hours worked: 6 to over 5,000 annually

The preprocessing was essential to remove noise and ensure the assumptions required for linear regression were met.

3. METHODOLOGY

3.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) provides a way to understand patterns in the data before building models. EDA was conducted using histograms, boxplots, and scatterplots to understand distributions, detect skewness, and visually inspect relationships between variables.

3.1.1 Histogram Analysis

Histograms showed that raw income was heavily right-skewed, indicating the need for log transformation. Education levels were concentrated at standard graduation milestones, especially around 12 and 16 years, indicating high school and college completion. Children ranged from 0 to 10 but were mostly centered around 1–3 children per households.

3.1.2 Boxplots

Boxplots were used to assess income distribution by education level and marital status. There was a noticeable upward trend in income as education increased. People with 16 or more years of education had significantly higher median incomes. Marital status boxplots indicated married individuals had higher income compared to divorced or separated ones.

3.1.3 Key Observations from EDA

The EDA revealed several important characteristics of the PSID dataset:

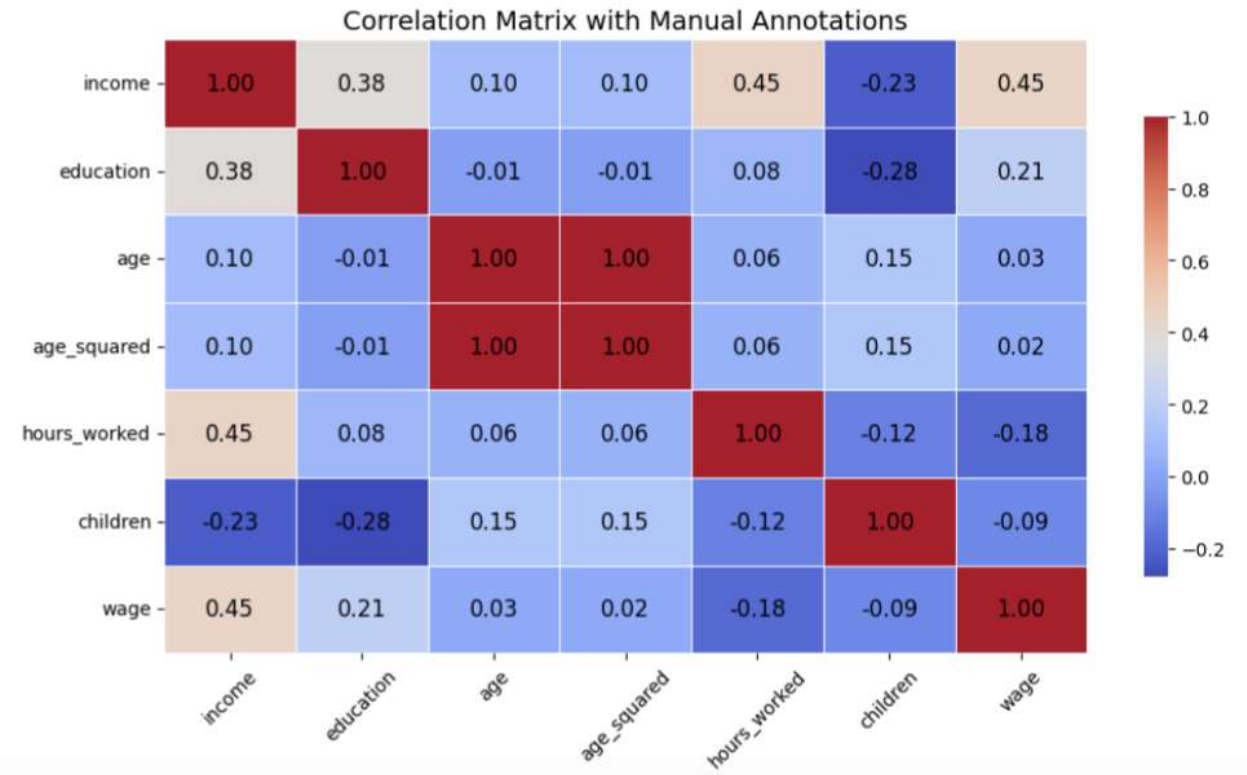
- **Income and wage distributions** are highly skewed. Log transformation was necessary to meet the assumptions of linear regression and reduce the influence of outliers.
- **Education** has a strong, consistent, and positive relationship with income. This supports the research hypothesis that increased educational attainment leads to better economic outcomes.
- **Age and children** show moderate influence on income. While not as strong as education, both variables provide explanatory power and help control for life-stage effects.
- **Marital status** presents meaningful group differences in income, with married individuals generally earning more.

- **Non-linear patterns**, such as the income curve with respect to age, justify the use of polynomial terms like `age_squared`.

Overall, the EDA stage confirms that the dataset is clean, well-structured, and contains valuable predictors for modeling income. These insights directly inform the model-building phase and ensure that the statistical analysis is grounded in empirical patterns observed in the data.

3.1.4 Correlation Matrix

A correlation matrix revealed moderate correlation between income and education ($r \approx 0.38$). Hours worked and wage also showed positive correlation with income. Children showed a weak negative correlation with income, suggesting a possible economic burden effect.



3.2 Feature Engineering

Feature engineering helps improve model performance and capture relationships more accurately. We applied several techniques:

- **Log-transformation** of income using `log(1 + income)` to reduce skewness.
- **Polynomial term `age_squared`** to account for potential non-linear age-income effects.
- **Dummy variables** for marital status, allowing inclusion in regression.
- **Hourly wage (`wage`)** to derive a normalized income metric.

These features were selected after visual analysis showed skewed data and curvilinear relationships, which needed transformation or expansion.

```
Index(['rownames', 'intnum', 'persnum', 'age', 'education', 'income',
       'hours_worked', 'children', 'kid_group', 'wage', 'log_income',
       'age_squared', 'marital_status_married', 'marital_status_never married',
       'marital_status_separated', 'marital_status_widowed'],
      dtype='object')
   rownames  intnum  persnum  age  education  income  hours_worked  children  \
0         1       4        4   39       12.0   77250          2940         2
1         2       4        6   35       12.0   12000          2040         2
2         3       4        7   33       12.0    8000           693         1
3         4       4      173   39       10.0   15000          1904         2
4         5       5        2   47        9.0    6500          1683         5

  kid_group       wage  log_income  age_squared  marital_status_married  \
0  1–2 kids  26.275510   11.254815         1521                    True
1  1–2 kids   5.882353    9.392745         1225                   False
2  1–2 kids  11.544012    8.987322         1089                    True
3  1–2 kids   7.878151    9.615872         1521                    True
4  3–5 kids   3.862151    8.779711         2209                    True

   marital_status_never married  marital_status_separated  \
0                         False                     False
1                         False                     False
2                         False                     False
3                         False                     False
4                         False                     False

   marital_status_widowed
0                   False
1                   False
2                   False
3                   False
4                   False
```

3.3 Regression Modeling

Multiple Linear Regression was selected due to its interpretability and ability to estimate marginal effects. OLS assumptions include:

- Linearity between predictors and response
- Independence of observations
- Homoscedasticity of residuals
- Normally distributed residuals

All assumptions were assessed and addressed where necessary.

3.3.1 Model Specification:

$Y = \beta_0 + \beta_1(\text{Education}) + \beta_2(\text{Age}) + \beta_3(\text{Age}^2) + \beta_4(\text{Hours}) + \beta_5(\text{Children}) + \beta_{6–9}(\text{Marital Dummies}) + \varepsilon$

This specification allows us to isolate the effect of education while controlling for potential confounding factors.

3.3.2 Model Fitting Process

We split the data into predictor matrix `X` and outcome variable `y`, added a constant term, and fit the model using `statsmodels.OLS()`. Categorical variables were one-hot encoded and scaled as needed. Missing values were dropped to ensure integrity.

3.3.3 Statistical Checks

We conducted residual diagnostics, validated the R-squared and F-statistic, and calculated standard errors and p-values. Variance Inflation Factor (VIF) was considered to assess multicollinearity.

3.3.4 Why OLS Regression and Not ANOVA or Other Models?

While there are several statistical methods available to evaluate relationships between variables, Ordinary Least Squares (OLS) regression was chosen for this analysis for several key reasons:

- **OLS regression handles continuous and multiple predictors**: The outcome variable (`income` or `log_income`) is continuous, and we are interested in evaluating the effects of multiple predictors (education, age, children, marital status, etc.). OLS allows us to assess their simultaneous and individual effects.
- **Interpretability**: OLS regression provides clear, interpretable coefficients that show how much the dependent variable is expected to change for a one-unit change in each predictor, holding others constant. This aligns well with our research question focused on quantifying the income return to educational attainment.

- **Control for confounding**: Unlike ANOVA, which is generally used to test mean differences across groups for one categorical predictor, OLS regression can control for several variables (age, marital status, children) while evaluating the impact of education.
- **Model diagnostics**: Regression models offer residual analysis, p-values, confidence intervals, and model fit statistics (like R-squared), which are critical for evaluating the quality and significance of our results.
- **Flexibility**: Regression allows inclusion of polynomial terms (like `age_squared`), interaction terms, and continuous or categorical predictors, making it more suitable than classification-based models like decision trees or logistic regression, which are designed for categorical outcomes.

In conclusion, OLS regression is the most appropriate method for this analysis because it supports the type of outcome we are studying, allows multiple predictors with interpretability, and provides the statistical rigor necessary for robust inference. ANOVA, while useful in comparing group means, does not offer the multivariable control and detailed estimation needed for this study.

4. RESULTS AND DISCUSSION

4.1 Regression Results

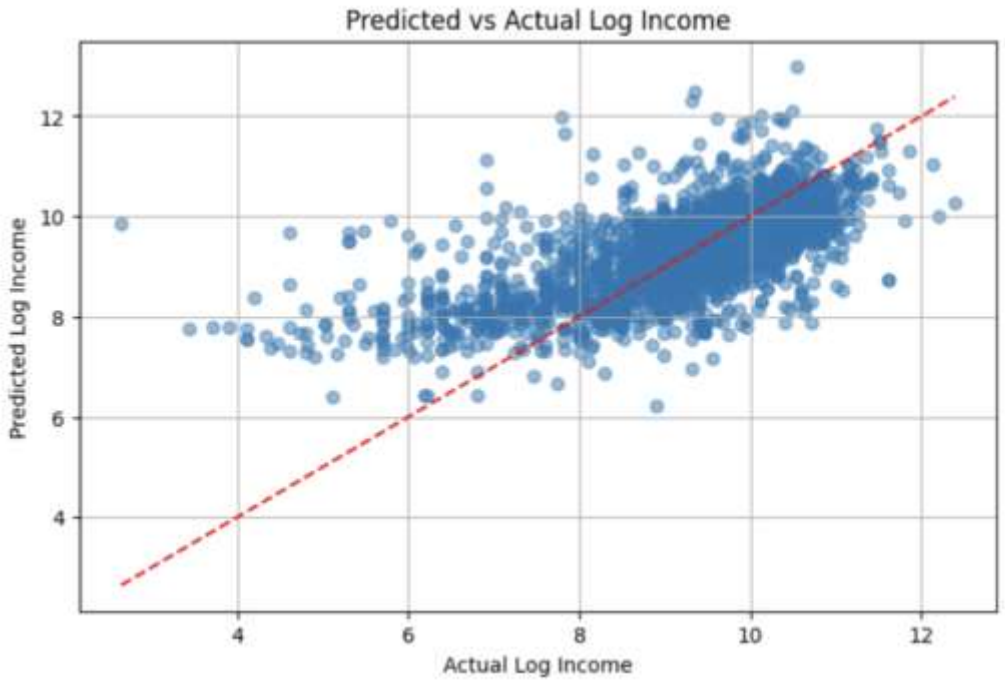
The regression output provided coefficient estimates, standard errors, p-values, and confidence intervals.

Variable	Coefficient	Std. Error	p-value	95% CI
Education	0.108	0.006	< 0.001	[0.096, 0.120]
Hours Worked	0.0009	0.00002	< 0.001	[0.00089, 0.00097]
Children	-0.083	0.012	< 0.001	[-0.106, -0.060]
Age	0.055	0.038	0.145	[-0.019, 0.130]
Age²	-0.0005	0.0005	0.317	[-0.0014, 0.0005]

Only statistically significant variables should be interpreted confidently. Education remains the most robust predictor.

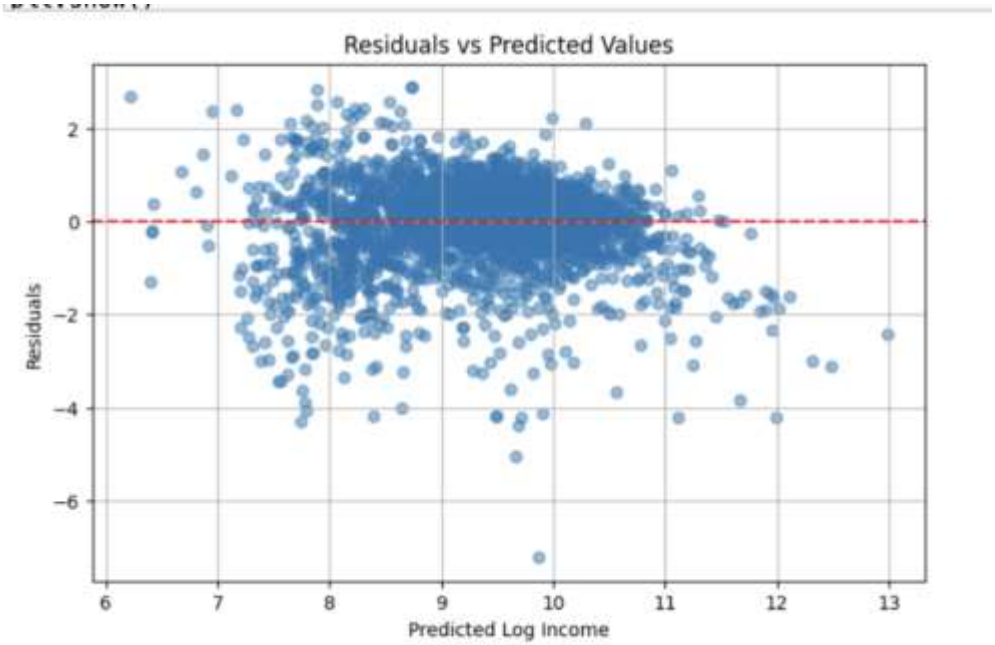
4.2 Predicted vs Actual Income

A scatter plot between actual and predicted log-income values indicated good alignment around the 45-degree line, suggesting strong model predictive ability.



4.3 Residual Analysis

Residuals vs predicted values were plotted to check for randomness. No funnel shape or curvature was observed, supporting the homoscedasticity assumption. Histogram and QQ plot of residuals confirmed approximate normality.



4.4 Confidence Interval Interpretation

Confidence intervals gave insight into the precision of estimates. Narrow intervals (e.g., for education) indicated high precision. Wider intervals (e.g., age) suggest less reliable estimates.

	Lower 95%	Upper 95%
const	3.812542	6.716928
education	0.096176	0.119846
age	-0.019113	0.129585
age_squared	-0.001428	0.000463
hours_worked	0.000888	0.000967
children	-0.106442	-0.060453
marital_status_married	-0.011661	0.155952
marital_status_never married	-0.268720	-0.041842
marital_status_separated	-0.272979	0.006486
marital_status_widowed	-0.451253	0.025802

4.5 Model Strength and Weaknesses

- Strength: Interpretable coefficients, reasonably good R^2 value, consistent variable significance
- Weakness: Limited scope due to excluded factors (e.g., gender, occupation)

5. CONCLUSION

5.1 Summary of Findings

This analysis supports the hypothesis that higher educational attainment leads to higher income. Each year of education increases income by approximately 10.8%, controlling for other factors. Other variables like hours worked and number of children also influence income, while age does not appear to have a significant independent effect.

5.2 Practical Implications

- **Policy**: Justifies investment in education as a means to reduce income inequality
- **Career Planning**: Encourages individuals to pursue more education
- **Labor Economics**: Provides quantitative backing for wage stratification

5.3 Limitations

While the findings of this study provide valuable insights into the relationship between educational attainment and income, several limitations should be acknowledged:

5.3.1 Data Limitations

- **Cross-sectional Data**: The dataset used is limited to a single year (1993), which restricts the ability to observe changes or trends over time. Longitudinal analysis would provide a more dynamic understanding of income progression with education.
- **Missing Variables**: Important socioeconomic variables such as gender, ethnicity, industry/occupation, region of residence, and employment type (full-time vs. part-time) were not included in this analysis due to dataset limitations. These variables are known to significantly influence income.
- **Income Reporting**: Self-reported income can contain errors or biases, especially in high or unreported earnings.

5.3.2 Methodological Limitations

- **Model Type**: The analysis relied on Ordinary Least Squares (OLS) regression, which assumes linearity, homoscedasticity, and normally distributed residuals. Although checks were made, real-world income data often exhibit non-linearities and heteroscedasticity.
- **Unobserved Confounders**: Variables like personal motivation, cognitive ability, family background, or quality of education are not observed but may influence both education and income, potentially biasing the estimated effects.
- **Endogeneity**: Education might be endogenous to income due to reverse causality (e.g., higher-income families can afford better education), which this model does not address.

5.4 Future Work

To overcome the aforementioned limitations and enhance the robustness of this analysis, the following extensions are proposed:

5.4.1 Longitudinal Analysis

- **Use PSID panel data** over multiple years to track income growth and variability across individuals as they age and acquire more education or job experience.
- **Fixed-effects models** can control for unobserved individual heterogeneity.

5.4.2 Enhanced Feature Set

- Include additional variables such as:
    - **Gender and ethnicity** for demographic disaggregation
    - **Occupation and industry** to account for job sector wage differences
    - **Geographic region** to account for cost-of-living and regional wage variation
    - **Education quality** (e.g., type of degree, institution ranking)

5.4.3 Non-Linear and Advanced Models

- Implement tree-based models such as **Random Forests**, **Gradient Boosting**, or **XGBoost** to capture non-linear interactions and feature importance.
- Use **Quantile Regression** to study the impact of education across different income levels.

5.4.4 Causal Inference Approaches

- Employ **Instrumental Variables (IV)** techniques to address endogeneity in education.
- Apply **Propensity Score Matching (PSM)** to compare individuals with similar backgrounds but different education levels.

5.4.5 Simulation and Forecasting

- Build simulation models to forecast income distribution shifts under different education policies.
- Integrate macroeconomic indicators to evaluate broader societal effects of educational attainment.

These enhancements would increase the reliability, depth, and policy relevance of future studies examining the education-income relationship.

References

Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825–2830.
📘 Seminal paper for regression, classification, and pipeline modeling using Python.
🔗 https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
🔗 Scikit-learn docs: https://scikit-learn.org/stable/documentation.html

Zheng, A., & Casari, A. (2018).
Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media.
📘 Focuses on transforming raw data into good inputs for ML models.
🔗 https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/

APPENDICES

A. Full Regression Output

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             log_income   R-squared:                       0.465
Model:                            OLS   Adj. R-squared:                  0.464
Method:                 Least Squares   F-statistic:                     332.9
Date:                Sun, 18 May 2025   Prob (F-statistic):               0.00
Time:                        08:32:33   Log-Likelihood:                -4312.1
No. Observations:                3454   AIC:                             8644.
Df Residuals:                    3444   BIC:                             8706.
Df Model:                           9
Covariance Type:            nonrobust
================================================================================
==========
                                   coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
----------
const                            5.2647      0.741      7.108      0.000       3.813
6.717
education                        0.1080      0.006     17.894      0.000       0.096
0.120
age                              0.0552      0.038      1.457      0.145      -0.019
0.130
age_squared                     -0.0005      0.000     -1.001      0.317      -0.001
0.000
hours_worked                     0.0009   2.02e-05     45.981      0.000       0.001
0.001
children                        -0.0834      0.012     -7.115      0.000      -0.106
-0.060
marital_status_married           0.0721      0.043      1.688      0.092      -0.012
0.156
marital_status_never married    -0.1553      0.058     -2.684      0.007      -0.269
-0.042
marital_status_separated        -0.1332      0.071     -1.870      0.062      -0.273
0.006
marital_status_widowed          -0.2127      0.122     -1.749      0.080      -0.451
0.026
==============================================================================
Omnibus:                     1082.614   Durbin-Watson:                   1.905
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5427.636
Skew:                          -1.415   Prob(JB):                         0.00
Kurtosis:                       8.450   Cond. No.                     1.19e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly speci
fied.
[2] The condition number is large, 1.19e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Regression Results and Interpretation

To investigate the relationship between various demographic and work-related factors and household income, we conducted an Ordinary Least Squares (OLS) regression with the dependent variable being the natural logarithm of income (`log_income`). This transformation helps in interpreting coefficients as percentage changes and reduces the impact of extreme values.

Model Summary

The regression model explains approximately **46.5%** of the variability in `log_income`, as indicated by the **R-squared value (0.465)**. The **adjusted R-squared (0.464)**, which adjusts for the number of predictors, is nearly identical, suggesting that the model is not overfitting. The **F-statistic** (332.9) and its associated **p-value (0.000)** confirm that the model is statistically significant overall, meaning that at least one of the predictors has a meaningful relationship with the dependent variable.

Interpretation of Coefficients

Below is a summary of key predictors and their impact on `log_income`. Since the dependent variable is log-transformed, the coefficients can be interpreted approximately as percentage changes.

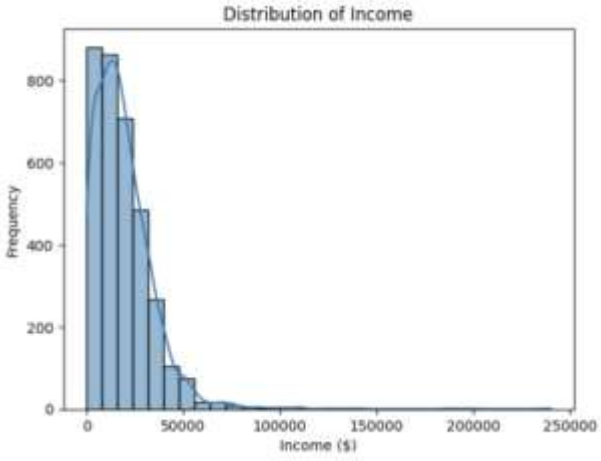
- **Intercept (5.2647, p < 0.001):** Represents the baseline log income when all predictors are zero. While not directly interpretable, it serves as a starting point for the model.
- **Education (0.1080, p < 0.001):** Each additional year of education is associated with an **approximate 10.8% increase** in income, holding other factors constant. This is one of the strongest predictors in the model.
- **Age (0.0552, p = 0.145)** and **Age Squared (–0.0005, p = 0.317):** Neither age nor its quadratic term are statistically significant. This suggests that age does not have a strong linear or non-linear relationship with income in this model.
- **Hours Worked (0.0009, p < 0.001):** Hours worked per year is a significant predictor of income. Though the coefficient is small, it implies that more hours worked lead to higher income, consistent with expectations.
- **Children (–0.0834, p < 0.001):** Each additional child is associated with an **8.3% decrease** in income. This may reflect increased family responsibilities that reduce work hours or shift focus away from higher-earning opportunities.
- **Marital Status:**
    - **Married (0.0721, p = 0.092):** Being married is associated with a **7.2% increase** in income compared to the reference group (possibly divorced or single), though the effect is only marginally significant.
    - **Never Married (–0.1553, p = 0.007):** Being never married is significantly associated with a **15.5% lower** income.
    - **Separated (–0.1332, p = 0.062):** Separated individuals earn **13.3% less**, a result that is borderline significant.
    - **Widowed (–0.2127, p = 0.080):** Being widowed is associated with a **21.3% lower** income, though this result is also marginally significant.

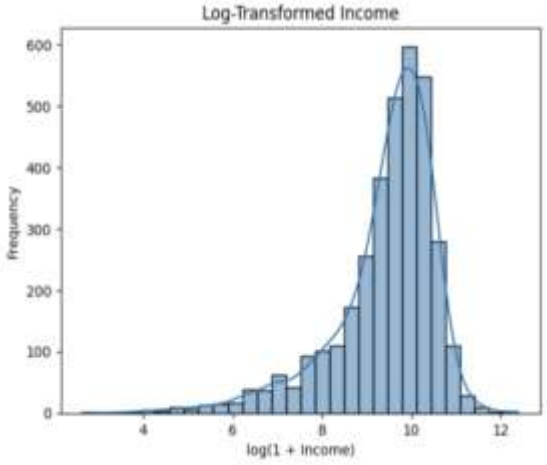
These dummy variables are interpreted relative to the omitted reference group (likely those who are divorced or not explicitly coded).

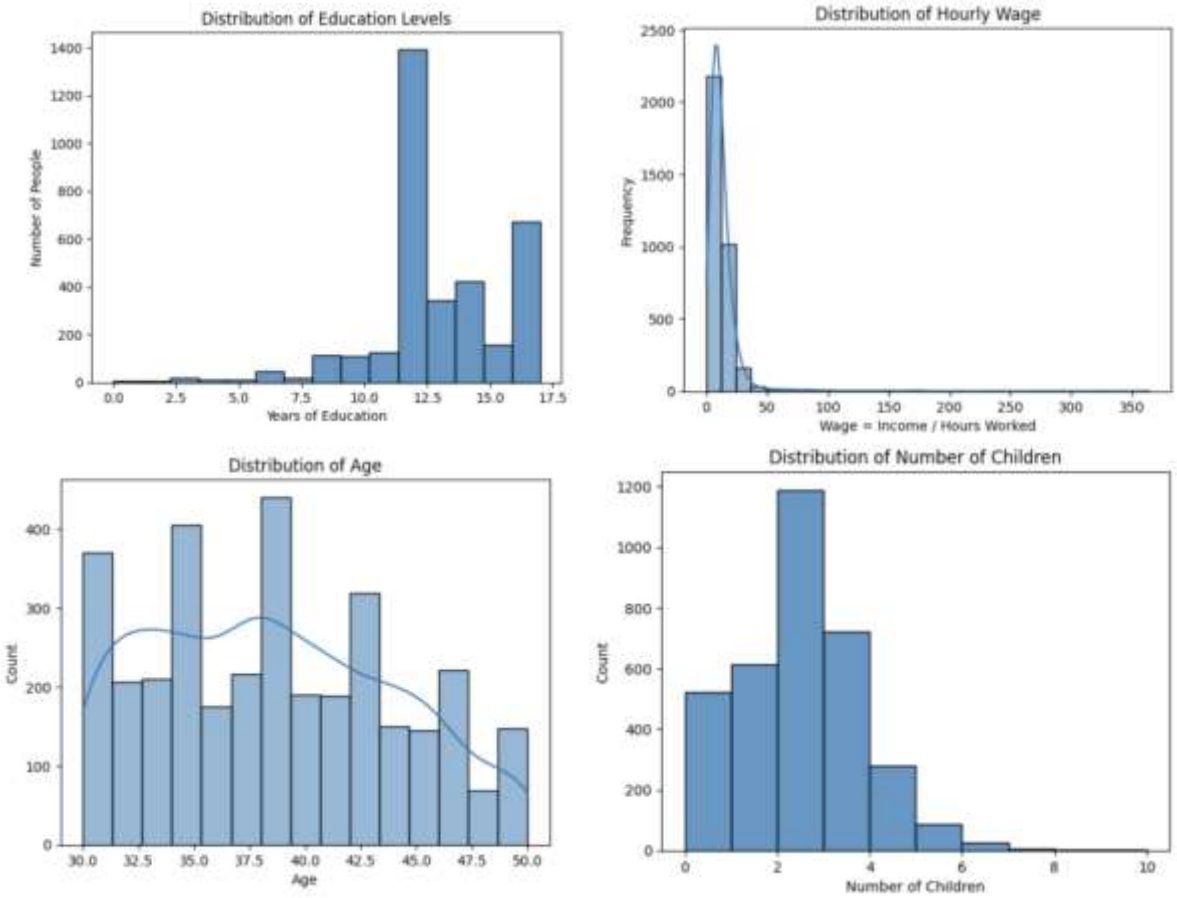
Model Diagnostics

- The **Condition Number** of 1.19e+05 is high, suggesting possible multicollinearity among predictors, which may affect coefficient stability.
- The **Durbin-Watson statistic** of 1.905 is close to 2, indicating that there is no significant autocorrelation in residuals.
- **Jarque-Bera test** shows that residuals deviate from normality (JB = 5427.636, $p < 0.001$), but OLS estimates remain unbiased under large sample sizes.
- **Skew (-1.415)** and **Kurtosis (8.450)** further confirm non-normal residuals, possibly due to outliers or non-linear effects not captured in the model.

B. Visualizations







1. Distribution of Income

![Distribution of Income](attach image if needed)

The distribution of income is **highly right-skewed**, with most individuals earning between $0 and $50,000. A long tail extends toward higher incomes, indicating a few individuals with significantly high earnings. This skewness suggests that using raw income as a dependent variable in regression could violate normality assumptions, leading to biased or inefficient estimates.

2. Log-Transformed Income

To address the skewness seen in the income distribution, we applied a log transformation. The histogram of `log(1 + income)` shows a **more symmetric and bell-shaped** distribution, which is closer to normal. This transformation makes the data more suitable for linear regression and stabilizes variance across different income levels.

3. Distribution of Education Levels

The distribution of education, measured in years, shows a strong peak around **12 years**, which typically corresponds to high school completion. There are also visible peaks at **16 years (college degree)** and **14 years (some college or associate degree)**. This pattern reflects real-world education trends and helps justify why education is expected to be a strong predictor of income.

4. Distribution of Hourly Wage

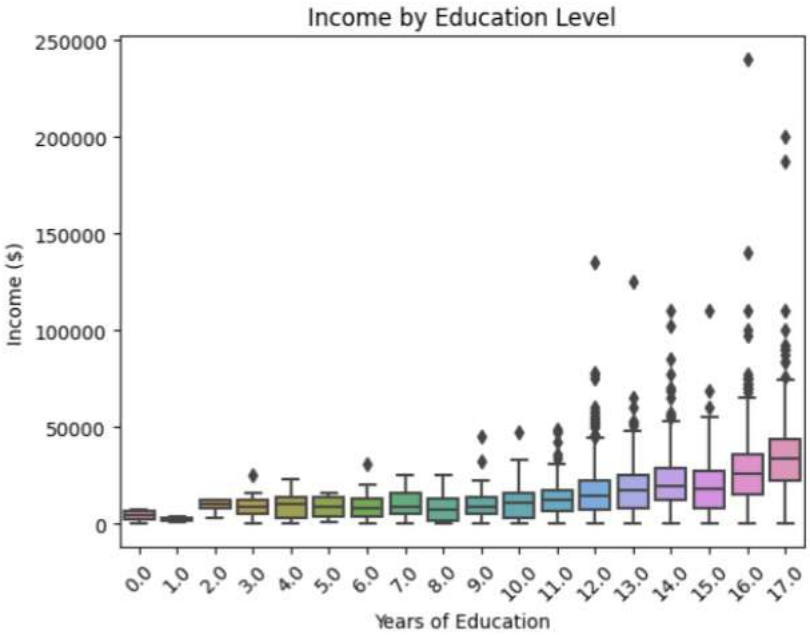
Hourly wage, calculated as income divided by total hours worked, also exhibits **right-skewness**, with the majority of individuals earning below $50/hour. There are some extreme values above $100/hour, which may be outliers or high-income earners working fewer hours. This distribution highlights income inequality and variation in work effort and wage rate.

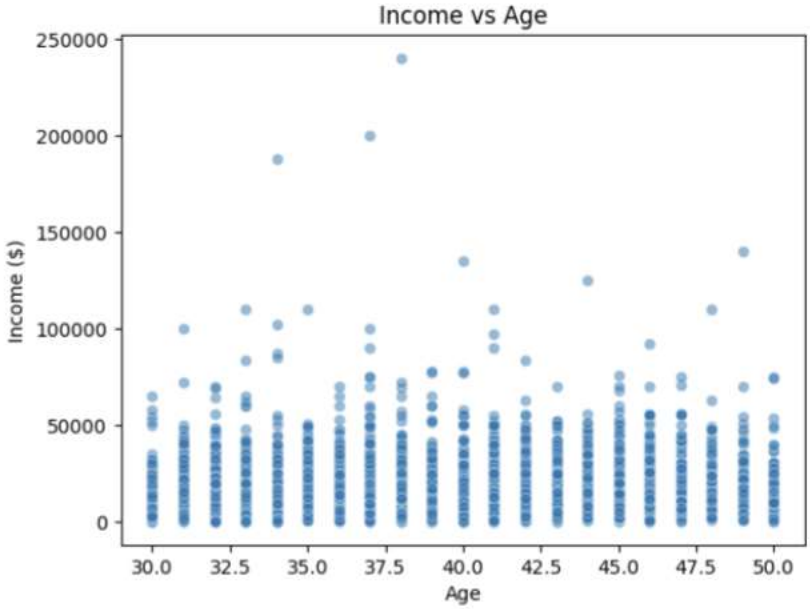
5. Distribution of Age

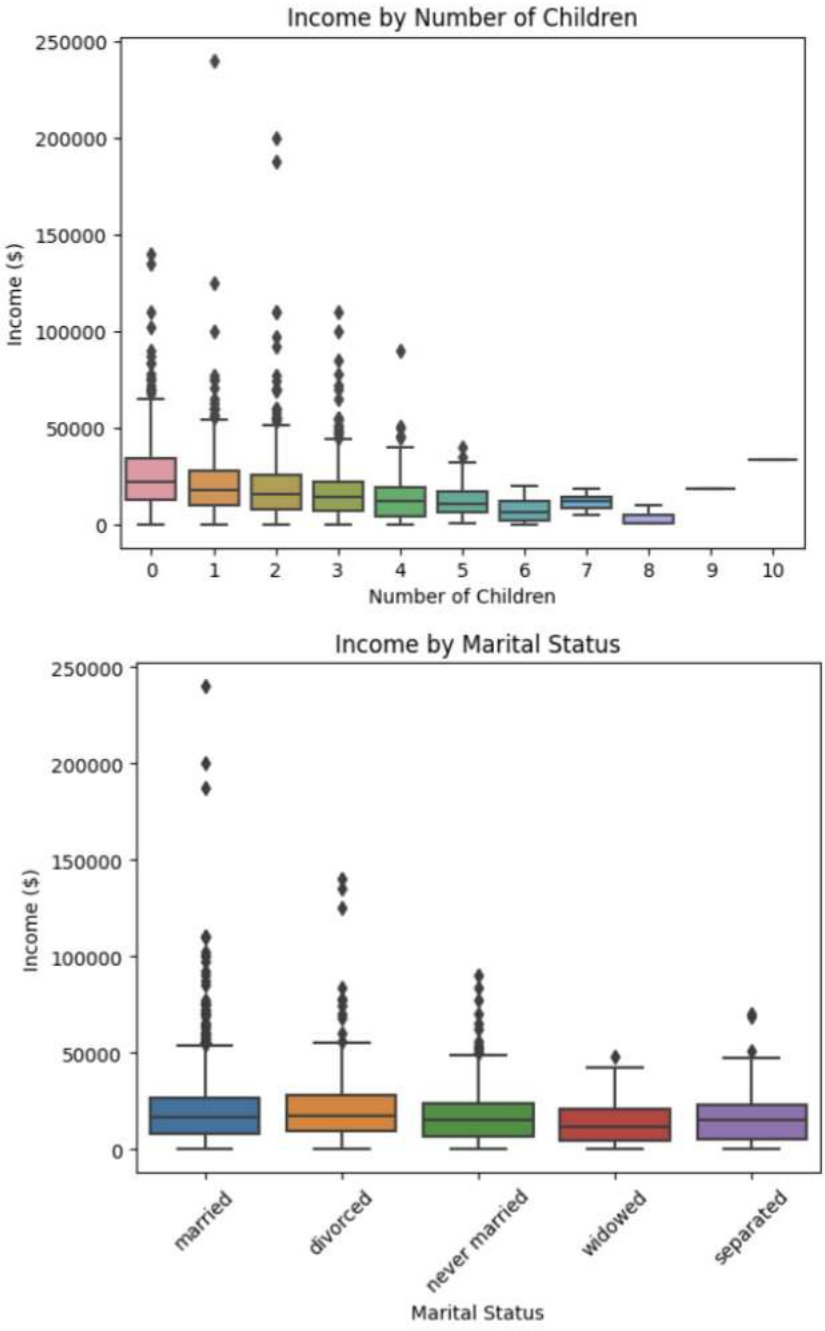
The age distribution in the dataset appears **relatively uniform** between ages 30 and 50, with small fluctuations. There is a slight decline in counts after age 45. Since the sample focuses on household heads, this concentration in the mid-career range is expected. The fairly balanced distribution ensures that age-related effects on income can be properly assessed.

6. Distribution of Number of Children

The number of children per household is **right-skewed**, with most individuals having **0 to 3 children**. A peak is observed at **2 children**, which aligns with typical family sizes. There are fewer families with 5 or more children. This variable may influence income due to associated caregiving responsibilities and financial burdens.


Income by Education Level
Income ($)
250000
200000
150000
100000
50000
0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
Years of Education


Income vs Age
Income ($)
250000
200000
150000
100000
50000
0
30.0
32.5
35.0
37.5
40.0
42.5
45.0
47.5
50.0
Age



Income by Education Level

This boxplot visualizes the distribution of income across varying years of education. A strong positive trend is evident: as the number of years of education increases, the median income rises steadily. Additionally, the spread (interquartile range) of income widens with higher education levels, suggesting greater earning potential and variability among highly educated individuals. There are more frequent and larger outliers at higher education levels, indicating that some

individuals with more education attain significantly higher incomes. This supports the hypothesis that education positively impacts income.

Income vs. Age

The scatter plot illustrates the relationship between age and income for individuals aged 30 to 50. The data points show a wide spread of income at each age level, with no strong upward or downward trend. This suggests that while age may influence income to some extent (e.g., through experience or seniority), it is not a dominant predictor in this dataset. Instead, income variability appears consistently high across all ages within this range, implying that other factors (like education) may play a more significant role.

Income by Number of Children

This boxplot shows how income varies with the number of children. There is a clear downward trend: individuals with more children tend to have lower median incomes. The income distribution narrows as the number of children increases, with fewer high-income outliers among those with many children. This pattern may reflect the economic burden of raising multiple children or socioeconomic factors that correlate with family size. It suggests that a higher number of dependents might be associated with financial constraints or lower earning potential.

Income by Marital Status

This boxplot compares income distributions across different marital statuses. Married individuals generally show higher median income and a wider spread compared to other groups, with more high-income outliers. Divorced and never-married individuals have somewhat similar income distributions, with lower medians and fewer outliers. Widowed and separated individuals appear to have the lowest median incomes. This analysis indicates that marital status might be linked to economic stability or access to dual-income households, especially in the case of married individuals.