Dean&Francis
ISSN 2959-6130

Modeling Income: A Regression-Based Econometric Analysis of How Age, Gender, and Education Influence Income

Qianyi Xie

Suzhou Science & Technology Town Foreign Language High School, Suzhou, China
xieqianyi1223@outlook.com

Abstract:

Income inequality is a persistent issue in economics and public policy, as it influences opportunities for individuals and affects broader patterns of social mobility and economic fairness. Understanding the factors that drive differences in earnings is therefore essential for both researchers and policymakers. This study uses data from the Panel Study of Income Dynamics (PSID) to build a multiple nonlinear regression model that explores how age, gender, and education affect income. To capture the possible nonlinear relationship between age and income, a quadratic term is added for age, and a log transformation is applied to income to reduce skewness. The regression results show that income generally follows an inverted U-shaped trend with age. While gender shows a borderline significant effect, education does not appear to be statistically significant in this model. After applying regression without education, the fit of the model did not improve noticeably. Although the overall explanatory power of the model is limited, it still reflects important ideas from human capital theory. Therefore, the study highlights several limitations, including the oversimplification of the model and the exclusion of relevant variables such as race and occupation. Taking these factors into account and drawing on the insights from previous research, the report concludes with policy implications and recommendations for future studies.

Keywords: Income inequality; Human capital; Education; Gender; PSID.

1. Introduction

Understanding why people earn different incomes has long been a central concern in economics. Income inequality not only shapes individual lives, but also affects broader issues such as social mobility, access to education, and perceptions of fairness in society. Economists therefore devote much attention

ISSN 2959-6130

to the forces that drive earnings.

One influential perspective is human capital theory, which argues that education, work experience, and skills raise productivity and are rewarded with higher wages [1]. Guided by this idea, many empirical studies investigate how age, gender, and education contribute to earnings differences [2-4]. These variables are frequently used in models that attempt to explain or predict income patterns.

Yet the links between these factors and income are not always linear or simple. In practice, earnings often rise with age but eventually plateau or decline. Several studies, for example, document an inverted U-shaped income profile across the life course [5,6]. Likewise, while education is generally associated with higher wages, the size of this advantage can depend on gender, occupation, and broader social context. Such findings suggest that straightforward linear models may not capture the full complexity of wage determination [7].

With the availability of richer and more recent datasets, researchers can revisit these questions using stronger evidence. This study draws on the 2023 wave of the Panel Study of Income Dynamics (PSID), a nationally representative U.S. survey that follows individuals and families over time [8]. Using these data, the analysis focuses on how age, gender, and education relate to income. To address nonlinear patterns, income is log-transformed and a squared term for age is added. Two main questions guide the analysis: whether a simple nonlinear model can capture earnings over the life cycle, and how much each factor, especially education, contributes to income inequality. By testing the model both with and without education, the study also evaluates how essential schooling is in explaining wage gaps. In linking economic theory with fresh evidence, the research aims to clarify current income dynamics and contribute to ongoing debates on education and inequality.

2. Literature Review

A large body of research examines the links between individual characteristics and income, with age frequently taken as a starting point. For example, Rawal used a linear regression model to examine the link between age and income in Godawari Municipality, Nepal [7]. His results showed a positive relationship, meaning that income tends to increase as people get older [7]. However, the model did not fit the data very well, suggesting that this relationship might not be fully explained by a simple straight-line (linear) pattern [7].

To better understand this, Ozhamaratli, Kitov, and Barucca proposed a model that looks at how age and income are distributed together in a population. Their findings showed that income typically rises in early adulthood, peaks around middle age, and then declines in later years. This creates a curve that looks like an upside-down "U". Such a pattern better reflects real-world life stages, including gaining experience and eventually retiring. This supports the idea that the connection between age and income is not linear, but curved or nonlinear [9].

Since age alone cannot explain all the differences in income, researchers have also focused on other factors—especially education. Card found strong evidence that more years of schooling lead to higher earnings. He also showed that education doesn't just relate to income—it actually causes changes in earnings, which is important for understanding how policies on schooling might affect economic outcomes [10].

Xie added to this by studying how factors like education, gender, race, marital status, age, and occupation together influence wages. His findings confirmed that income depends on a mix of characteristics, not just one [11]. Zhou and Ramezani used machine learning to go even further. Their model tested which personal traits matter most for financial success. They found that education, age, job type, and gender consistently stood out as the most important [12].

All of this research points to the same conclusion: it is not enough to look at just one factor in a simple way. However, several gaps remain in the existing literature. First, many studies still rely on simple linear models, which fail to capture non-linear patterns such as the inverted U-shaped relationship between age and income. Second, some analyzes omit key variables—particularly education and gender—making it difficult to isolate their independent effects. Third, much of the available evidence is based on older datasets, which may not accurately reflect current labor market dynamics.

These gaps point to the importance of using more up-to-date and flexible methods. This study therefore relies on the 2023 wave of the PSID, transforms income into logarithmic form to handle skewness, and adds both a quadratic term for age and controls for education and gender. With this design, the analysis aims to give a more accurate and current picture of the factors shaping income.

3. Methodology

3.1 . Descriptive Statistics

This study looks at the effects of age, gender, and education on income in the United States, using data from the 2023 wave of the Panel Study of Income Dynamics (PSID), a long-running national survey widely applied in social science research [8].

Since income distributions are typically skewed by a few very high earners, the dependent variable is expressed in logarithmic form. Taking the natural log reduces the weight of outliers and makes percentage differences in in-

come easier to interpret.

The main explanatory variables are age, gender, and years of schooling. To test whether earnings rise and then decline with age, both age and age squared are included. Gender is coded as a binary indicator (male = 1, female = 0), while education is measured by the number of completed school years.

Table 1: Descriptive Statistics of Key Variables

Variable	Mean	Std.Dev.	Min	Max	Obs.
log(income)	9.57	1.11	1.95	12.03	1698
Age	27.67	10.90	18.00	65.00	1698
Age^2	884.32	812.52	324.00	4225.00	1698
Education	11.89	4.00	0.00	17.00	1698
Gender (Male=1)	0.52	0.50	0.00	1.00	1698

Table 1 presents the descriptive statistics for the variables used in the regression analysis. The sample includes 1,698 individuals aged 18 to 65. The mean log income is 9.57 with a standard deviation of 1.11, suggesting moderate variability in earnings. The average age is 27.67 years, and individuals have completed an average of 11.89 years of schooling. The gender distribution is fairly balanced, with the mean of 0.52. Given the wide age range and prior evidence of a nonlinear relationship between age and income, the model includes an age-squared term to capture potential curvature in the income-age profile.

3.2 . Model Specification

This study applies a standard Ordinary Least Squares (OLS) regression model to examine the relationship between income and three main variables: age, education, and gender. The dependent variable is the logarithm of income, which helps correct for skewed income distribution and allows for percentage-based interpretation of the results.

The model is specified as follows:

$$\log(Income_i) = \beta 0 + \beta 1 \cdot Age_i + \beta 2 \cdot Age_i^2 + \beta 3 \cdot Education_i + \beta 4 \cdot Gender_i + \varepsilon i \quad (1)$$

The inclusion of both age and age squared allows the model to capture a possible non-linear (inverted U-shaped) relationship between age and income. Education and gender are added to assess their independent effects on earnings. Other variables, such as region, occupation, or marital status, are not included, to clearly focus on the three predictors of interest.

3.3 . Robustness Check: Excluding Education

To test the reliability of the main findings, a reduced model was estimated by removing the education variable from the regression. This step helps assess whether the effect of education is independent or overlapping with the effects of age and gender.

The reduced model removes the education term from equation (1), and is specified as:

$$\log(Income_i) = \beta 0 + \beta 1 \cdot Age_i + \beta 2 \cdot Age_i^2 + \beta 3 \cdot Gender_i + \varepsilon i \quad (2)$$

By comparing this version with the full model that includes education, it becomes possible to see how much education contributes to explaining differences in income. Changes in the size or statistical significance of the age and gender coefficients also indicate whether the original estimates were partly influenced by shared variation with education. This approach adds confidence to the results by showing how stable the model is when key variables are added or removed.

4. Results and Discussion

4.1 . Model with Education

Table 2 reports the results of the full OLS regression model, which includes age, age squared, education, and gender as independent variables. The coefficient of age is positive (0.1583) and statistically significant at the 1% level, indicating that income tends to increase as people age. However, the negative coefficient for age squared (−0.0017, also highly significant) suggests that this trend eventually reverses, forming an inverted U-shape relationship between age and income. This is consistent with the idea that income grows in early and middle adulthood but tends to plateau or decline later in life.

Education shows a small positive coefficient (0.0143), but the p-value (0.157) is above the conventional thresholds ($p < 0.05$ or $p < 0.01$), indicating that the effect is not statistically significant. Gender has a positive coefficient (0.0966), implying that men earn more on average than women, though the effect is only marginally significant ($p = 0.057$). The model explains approximately 14.4% of the variation in log income ($R^2 = 0.144$).

Table 2: OLS Regression Results for log(Income).

ISSN 2959-6130

Variable	Coef.	Std.Err.	t	p-value	0.025	0.975
Intercept	6.4884	0.216	29.990	0.000	6.064	6.913
Age	0.1583	0.016	9.823	0.000	0.127	0.190
Age^2	-0.0017	0.000	-7.315	0.000	-0.002	-0.001
Education	0.0143	0.010	1.415	0.157	-0.006	0.034
Gender (Male=1)	0.0966	0.051	1.904	0.057	-0.003	0.196

Notes: Dependent variable is log(Income). R^2 = 0.144, Adjusted R^2 = 0.142, N = 1698. Gender coded as Female = 0, Male = 1.

The residual plot for this model (Figure 1) shows a fairly even spread of residuals around the zero line, with no major pattern or curvature, suggesting that the model fits reasonably well.

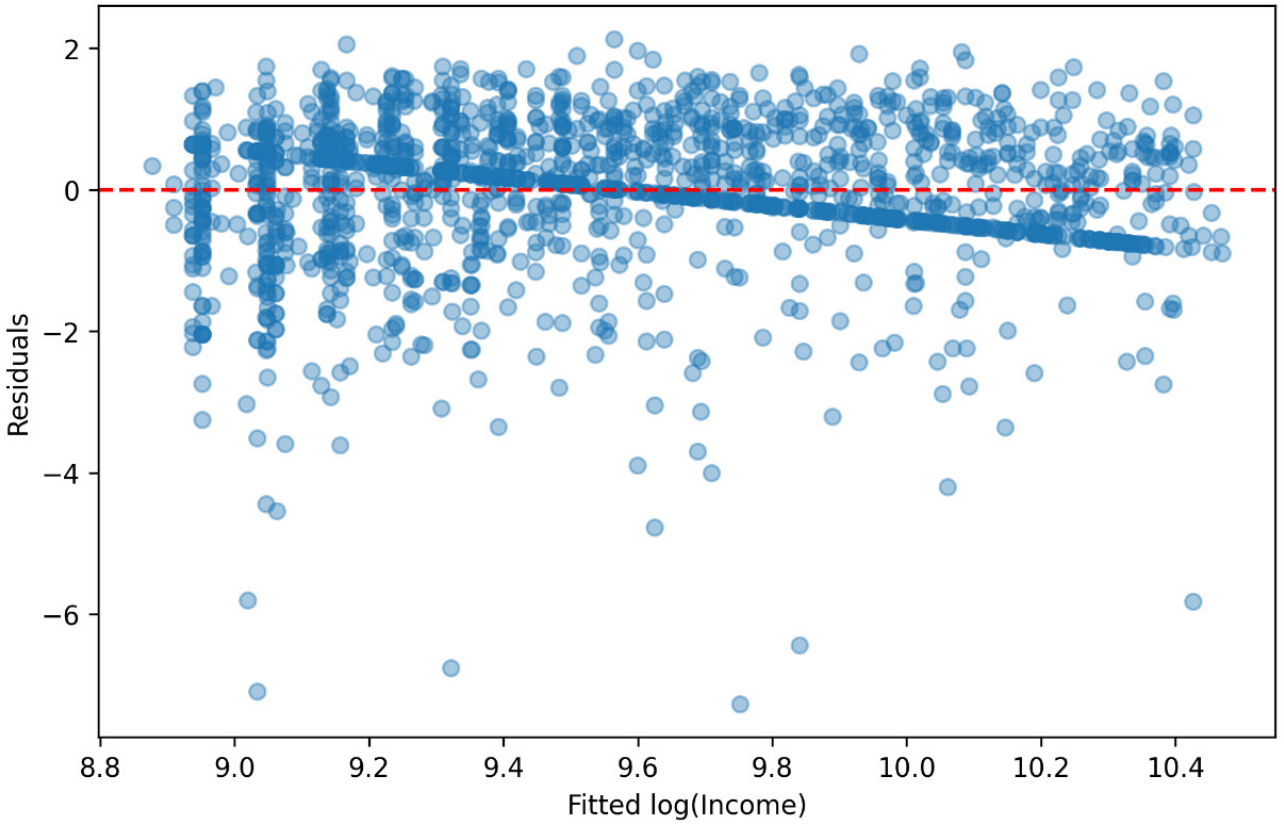

Fig. 1 Residuals vs. Fitted Values.

4.2 . Model without Education

To assess how education influences the results, a second model is estimated without the education variable. The results are presented in Table 3. After removing education, the coefficients for age (0.1710) and age squared (−0.0019) become slightly larger in absolute value but remain statistically significant. This suggests that some of the effects of age on income in the full model may have overlapped with education.

The gender coefficient slightly decreases to 0.0878 and becomes less statistically significant (p = 0.081), implying that part of the gender effect in the full model may have been linked to differences in education levels. The model fit drops only slightly (R^2 = 0.143), indicating that removing education has only a small effect on the model's explanatory power.

Table 3: OLS Regression Results for log (Income) without Education.

Variable	Coef.	Std.Err.	t	p-value	0.025	0.975

Intercept	6.5016	0.216	30.071	0.000	6.078	6.926
Age	0.1710	0.013	12.753	0.000	0.145	0.197
Age^2	-0.0019	0.000	-10.748	0.000	-0.002	-0.002
Gender (Male=1)	0.0878	0.050	1.743	0.081	-0.011	0.187

Notes: Dependent variable is log (Income). R^2 = 0.143, Adjusted R^2 = 0.141, N = 1698. Gender coded as Female = 0, Male = 1.

The residual plot for the reduced model (Figure 2) shows a similar pattern to the full model, though the spread of residuals appears slightly wider, with more clustering and some outliers. This supports the idea that including education helps improve the model's fitness, even though its coefficient was not statistically significant.

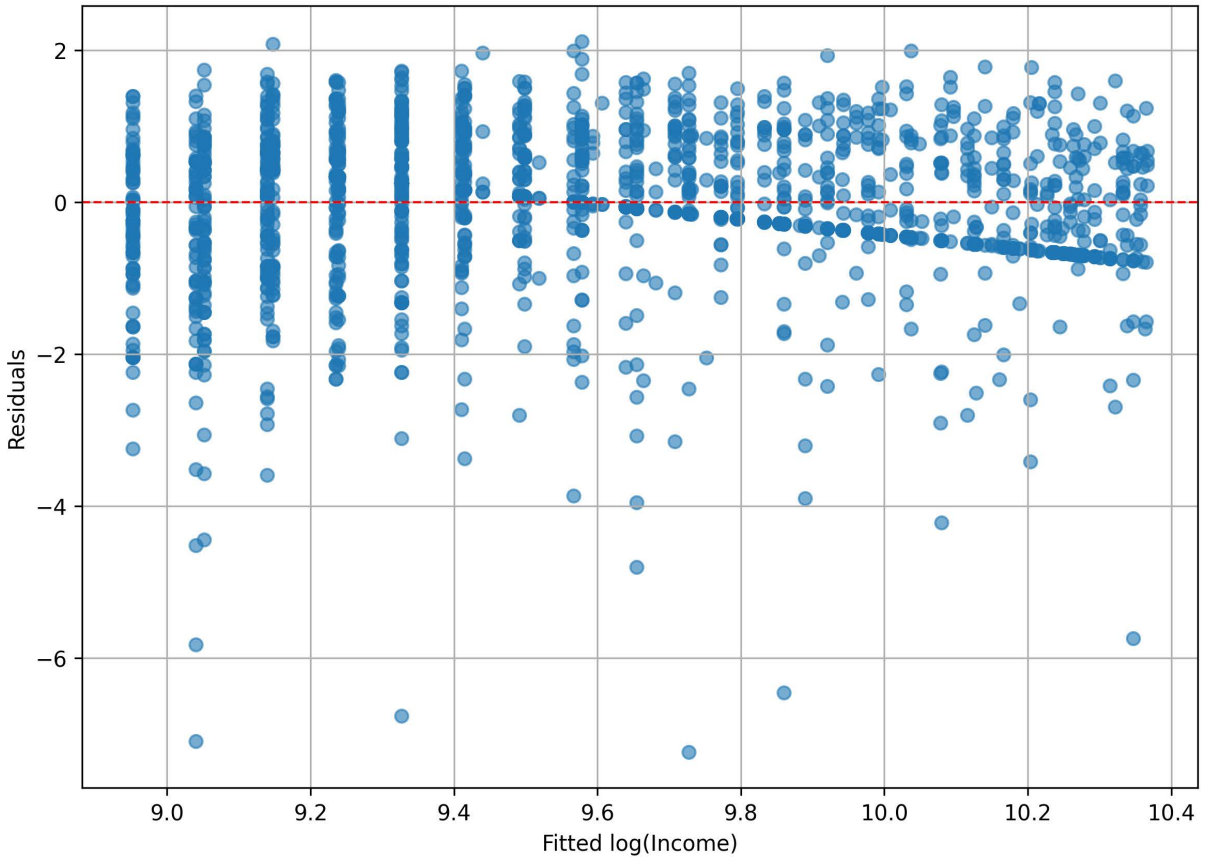

Fig. 2 Residuals vs. Fitted Values (without Education).

5. Discussion

The regression results show that age, education, and gender all influence income. The positive coefficient for age and the negative one for age squared confirm the inverted U-shaped age–earnings profile: income grows in early and middle adulthood, peaks, then falls before retirement. This matches the findings of Murphy and Welch and is similar to results in other countries [6, 13]. Education also has a clear positive effect, supporting human capital theory [10]. Gender differences remain even after controlling other factors, which lead to continuing inequality in the labour market [11].

Residual plots suggest that the model works reasonably well, with no strong signs of non-linearity or unequal variance. When education is removed from the model, the residuals become slightly larger, showing that education helps explain income and improves model stability.

6. Conclusion

This study looked at how age, education, and gender affect income using recent PSID data. The results show that income rises in early and middle adulthood, peaks, and then declines in later years. Education has a clear positive

ISSN 2959-6130

effect, and gender differences remain even after other factors are considered. These patterns match human capital theory and past research, and they show why non-linear terms can improve income models. While the study leaves out some factors, it underlines the role of education and the need to close gender pays gaps.

The model leaves out some important variables, such as occupation, industry, work experience, and region, which could affect income. Past research warns that cross-sectional data can overstate the age–income curve because of differences between generations. Long-term studies also suggest that retirement decisions and who stays in the labour market may shape the income drop seen in older age. Survey-based income data can contain reporting errors, which may bias results.

From a methods perspective, adding an age-squared term helps show non-linearity, but more complex trends may still be missed. OLS cannot fully solve problems such as omitted variable bias or the possible two-way relationship between education and income. Multicollinearity between variables could also make the coefficients less stable.

Future research could use longitudinal data to separate generation effects from real age-related changes, as suggested by Thornton, Rodgers, and Brookshire. Adding interaction terms (such as age × education or gender × occupation) may reveal patterns hidden in simpler models. Non-parametric methods or machine learning could capture more complex links, as seen in recent inequality studies. Work by Castelló-Climent and Doménec shows that expanding education can affect income distribution in non-linear ways, so policy needs to balance fairness and efficiency.

The clear positive link between education and income supports investment in high-quality, accessible education. But as Goldin and Katz note, education policy must also keep up with technological change to avoid growing skill gaps. Closing gender pay gaps remain important, which means enforcing pay equity laws and opening higher-paying jobs to underrepresented groups. Such steps could help reduce inequality and support more inclusive growth.

References

[1] Borjas, George J., and Jan C. Van Ours. *Labor Economics*. McGraw-Hill/Irwin, 2010.

[2] Checchi, Daniele. "Education, Inequality and Income Inequality." 2001.

[3] Lee, Jong-Wha, and Hyeok Yong Lee. "Human Capital and Income Inequality." *Journal of the Asia Pacific Economy*, vol. 23, no. 4, 2018, pp. 554–583.

[4] Aqil, Muhammad, and Diah Wahyuniati. "The Effect of Human Capital Inequality on Income Inequality: Evidence from Indonesia: An Application of Generalized Method of Moment Estimation." *Proceedings of The International Conference on Data Science and Official Statistics*, vol. 1, 2021, pp. 358–372.

[5] Lee, Ronald, et al. "Charting the Economic Life Cycle." 2006.

[6] Myck, Michal. "Wages and Ageing: Is There Evidence for the 'Inverse-U' Profile?" *Oxford Bulletin of Economics and Statistics*, vol. 72, no. 3, 2010, pp. 282–306.

[7] Rawal, Shalik Ram. "A Linear Regression Study of the Effects of Age on Income in the Godawari Municipality, Lalitpur." *Pakistan Social Sciences Review*, vol. 6, no. 4, 2022, pp. 52–61.

[8] *Panel Study of Income Dynamics: Public Use Data*. Panel Study of Income Dynamics, 2023. https://psidonline.isr.umich.edu/. Accessed 26 July 2025.

[9] Ozhamaratli, Filiz, et al. "A Generative Model for Age and Income Distribution." *EPJ Data Science*, vol. 11, no. 1, 2022, pp. 1–26.

[10] Card, David. "The Causal Effect of Education on Earnings." *Handbook of Labor Economics*, vol. 3, 1999, pp. 1801–1863.

[11] Xie, Rui. "The Influence of Education Level, Gender, Race, Marital Status, Age, and Occupation on the Wage of the General Population." *2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)*, Atlantis Press, 2022, pp. 926–932.

[12] Zhou, Michael, and Ramin Ramezani. *A Deep Dive into the Factors Influencing Financial Success: A Machine Learning Approach*. arXiv:2405.08233, 2024.

[13] Murphy, Kevin M., and Finis Welch. "Empirical Age-Earnings Profiles." *Journal of Labor Economics*, vol. 8, no. 2, 1990, pp. 202–229.