**A Research Policy Report Based on the American Community Survey Dataset,**

**Investigating the Association between Personal Income and Education**

Chelsea Chen

University of Pennsylvania, School of Social Policy & Practice

MSSP 8970: Applied Linear Modeling

Instructor: Momin Malik

Policy Research Report 4

December 17, 2024

**Introduction**

Personal income is a key determinant of living standards and a crucial indicator of a country's economic development. In the United States, there are significant differences between individual incomes. Persistent economic inequalities are influenced by various demographic factors such as education and citizenship status (Chetty et al., 2020; Blau & Kahn, 2017). Education has long been recognized as a key driver of income, with higher educational attainment generally leading to higher earnings (Goldin, 2014). Citizenship disparity further exacerbates income inequalities, as naturalized citizens often have different economic opportunities than non-citizens (Chetty et al., 2020).

Using data from the 2023 American Community Survey (ACS) from IPUMS USA, this study examines the factors influencing personal income by focusing on educational attainment as the primary independent variable and controlling citizenship status, age, and working hours through linear regression models. The analysis also explores how education interacts with gender, working hours, and family structure. Additionally, we conduct hierarchical regression models clustered at the family level to capture the within-family and between-family variations in personal income.

**Methods**

**Data**

We use data from IPUMS USA, specifically the 2023 American Community Survey (ACS), representing 3,405,809 respondents. We selected a subsample of 2,800 observations from this dataset that was more representative of our analysis.

**Variables**

We consider the following variables:

* **INCTOT**: Total pre-tax personal income or losses from all sources reported for the previous year.
* **INCTOTR**: Total pre-tax personal income, excluding negative values. Only positive incomes are investigated as the dependent variable in the regression model.
* **CITIZEN**: Citizenship status of respondents, ranging from 0 to 9.
* **CITIZENR**: Citizenship is recoded into two categories: 0 represents non-citizen; 1 represents Citizen.
* **AGE**: Respondent’s age in 2023.
* **Agec**: Centered age variable to make the intercept meaningful.
* **EDUC**: Highest educational attainment or education years completed by the respondents. 0 represents no schooling; 1 represents the completion of Grades 1 to 4 in primary school. 2 represents the completion of primary school and middle school, covering Grades 5 to 8; from 3 to 10, each number represents a specific school year from high school to college; 11 represents education beyond college, such as graduate or professional studies.
* **UHRSWORK**: The usual number of hours worked per week is
* **SEX**: Respondent's gender: Male or Female.
* **SEXR**: Recoded gender. 0 represents Female; 1 represents Male.
* **RACE**: Respondent’s race, ranging from 1 to 9.
* **RACER**: Recoded race. 0 represents non-white; 1 represents white.
* **NCHILD**: Number of own children living in the household, including stepchildren and adopted children.

Limiting the dataset to only non-missing observations across all these variables, we have 2,800 observations. Our primary explanatory variable is educational attainment; however, we will control whether the person has citizenship, centered age, and usual hours worked per week.

**Analysis**

We will conduct diagnostic tests for the linear regression model to identify potential violations of key assumptions, including linearity, homoscedasticity, the normality of residuals, and whether omitting any valuable variables. Additionally, we will check for potential outliers and apply appropriate corrections. Interaction terms between educational attainment and other variables will be explored to determine if these interactions improve model performance. We will compare the fit of these revised models to the original model. Furthermore, we will perform hierarchical regression to analyze personal income about educational attainment and other variables, accounting for clustering at the family level.

**Results**

**Descriptive Statistics**

Descriptive statistics are given in Table (1).

**Table 1.** Descriptive statistics for a limited sample

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Scale** | **Unit** | **Min** | **Max** | **Mean** | **Median** | **S.D.** |
| INCTOTR | Continuous | Dollars | 1 | 9999999 | 1127518 | 41600 | 3086271.9886 |
| CITIZENR | Dummy | No/Yes | 0 | 1 | 0.4839 | 1 | 0.4998 |
| EDUC | Continuous | Years | 0 | 11 | 6.238 | 6 | 3.4074 |
| UHRSWORK | Continuous | Hours | 0 | 99 | 31.32 | 40 | 19.0803 |
| RACER | Dummy | No/Yes | 0 | 1 | 0.1643 | 0 | 0.3706 |
| SEXR | Dummy | No/Yes | 0 | 1 | 0.5786 | 1 | 0.4939 |
| NCHILD | Continuous | Scale | 0 | 9 | 0.9164 | 0 | 1.2615 |
| agec | Continuous | Years | -29.86 | 65.138 | 8.869 | 8.138 | 17.3341 |

There are *n* = 2,800 observations with no missing values across all variables.

For the continuous variables, INCTOTR (total personal income) ranges widely from $1 to $9,999,999, with a mean of $1,127,518 and a median of $41,600, indicating a highly skewed distribution as evidenced by the large standard deviation of $3,086,271.99. EDUC (educational attainment) spans from 0 to 11 years, with a mean of 6.24 years and a median of 6 years, suggesting that most individuals have completed high school education. NCHILD (number of children) ranges from 0 to 9, with a mean of 0.92 children and a standard deviation of 1.26, indicating a significant disparity with most respondents having zero or one child. Agec (centered age) varies significantly from -29.86 to 65.14 years, with a mean of 8.87 years and a standard deviation of 17.33 years, highlighting a broad age distribution.

For dummy variables, CITIZENR (citizenship status) shows that 48.39% of respondents are citizens (mean = 0.4839). RACER (race) indicates that only 16.43% of respondents are categorized as white (mean = 0.1643). SEXR (gender) shows that 57.86% of respondents are male (mean = 0.5786). These dummy variables provide binary indicators, with standard deviations close to 0.5, reflecting a relatively even distribution between categories for citizenship and gender, while race shows a more skewed distribution.

**Regression Diagnostics**

Before analyzing the relationship between variables, we evaluate the quality of the model,

and examine whether gender (as measured in the variable ), race (as measured in the variable ), and the number of children (as measured in the variable ), are relevant variables.

***Linearity Diagnostics***

The scatterplot matrix reveals varying degrees of linearity between INCTOTR (total personal income) and the predictor variables. A strong linear relationship is observed between INCTOTR and EDUC (education level), with the regression line fully covered by the confidence band, indicating that higher educational attainment is associated with higher incomes. In contrast, the relationships between INCTOTR and agec (centered age) and UHRSWORK (usual hours worked) appear non-linear.

**Figure 1.** Scatterplot of Continuous Variables Before Correction

***A collage of graphs

Description automatically generated***

To satisfy the linearity assumption, we applied a log transformation to INCTOTR. We excluded observations where the log-transformed income exceeded 15 or was less than 6. Consequently, our analysis focuses on log-transformed income values ranging between 6 and 15. The final sample for the following investigation consists of 2,483 observations.

After the log transformation, all three continuous variables exhibit linear relationships with log-transformed income.

**Figure 2.** Scatterplot of Continuous Variables After Correction

*A collage of graphs

Description automatically generated*

***Normality of Errors***

The histogram displays the distribution of residuals (df\_resid) for the linear regression model. The residuals are approximately normally distributed, centered around zero, and exhibit a bell-shaped curve.

**Figure 3.** Histogram of Residuals

*A graph of a person with numbers and a line

Description automatically generated with medium confidence*

***Homoscedasticity***

The scatter plot displays residuals (df\_resid1) against the regression model’s fitted values (df\_fitted1). The residuals are dispersed around the horizontal line at zero, indicating that the model's predictions are relatively unbiased. There is a slight funnel shape, with residuals showing more variability as the fitted values increase. This pattern is close to homoscedasticity, so no further corrective action was taken.

**Figure 4.** Scatterplot of Residuals vs Fitted Values

A diagram of a graph

Description automatically generated with medium confidence

***Omitted Relevant Variables***

Numerous studies have explored the relationship between race, sex, and the number of children owned on personal income, highlighting systemic disparities rooted in socio-economic and demographic factors. Studies have shown that non-white individuals, mainly Black, Hispanic, and Indigenous populations, experience lower average earnings compared to their white counterparts, even after controlling for education and experience (Chetty et al., 2020; Killewald & Bearak, 2014). Discriminatory hiring practices, occupational segregation, and disparities in access to quality education contribute to these income gaps (Pager & Shepherd, 2008). Sex also influences income, with women typically earning less than men, a disparity known as the gender pay gap. Research indicates that even when controlling for job type, education, and experience, women earn approximately 82 cents for every dollar earned by men (Blau & Kahn, 2017). The number of children in a household can also impact personal income. The “motherhood penalty” refers to the reduction in earnings women experience after having children, partly due to time taken off work, reduced working hours, or employers’ biases against mothers (Budig & Hodges, 2010). Conversely, fathers may experience a “fatherhood premium,” where they receive higher wages compared to non-fathers, driven by perceptions of increased responsibility and commitment (Killewald, 2013).

We examine correlations between variables. The correlation matrix is shown in Table (2).

**Table 2.** Correlation Matrix Including New-added Variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | CITIZENR | agec | EDUC | UHRSWORK | RACER | SEXR | NCHILD |
| CITIZENR | 1 |  |  |  |  |  |  |
| agec | 0.2315 | 1 |  |  |  |  |  |
| EDUC | 0.2078 | 0.2424 | 1 |  |  |  |  |
| UHRSWORK | -0.0024 | 0.1385 | 0.3776 | 1 |  |  |  |
| RACER | 0.1011 | 0.0623 | 0.0799 | -0.0044 | 1 |  |  |
| SEXR | -0.0897 | -0.0037 | -0.0517 | 0.1354 | -0.0100 | 1 |  |
| NCHILD | -0.0016 | 0.1776 | 0.0448 | 0.2388 | -0.0295 | 0.0134 | 1 |

There are no strong correlations between variables (> 0.6), so we do not have to worry about potential collinearity.

Next, we tested the correlations of race (RACER), gender (SEXR), and the number of children (NCHILD) with the log-transformed total income (INCTOTlog1). The correlation with race was not statistically significant. Sex (correlation = 0.16) and the number of children (correlation = 0.07) were significant. Therefore, sex and the number of children were included in the subsequent models.

***Outlier Diagnostics***

The influence was used to identify potential outliers using Cook’s distance method. The initial cut-off point was set at 4/2483 = 0.0016, which identified approximately 7% of the observations as outliers - an unacceptably high proportion. The quantile distribution of Cook's distance values above the initial threshold was examined to refine this. We saw a significant jump at 95%; therefore, the cut-off threshold was finalized at 0.0064. Consequently, nine observations were identified as outliers and subsequently removed from the dataset.

**Table 3.** Influence-based Outliers for Model 4

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **INCTOTlog1** | **CITIZENR** | **agec** | **EDUC** | **UHRSWORK** | **SEXR** | **NCHILD** | **Cook’s *D*** |
| 12.7799 | 1 | 9 | 6 | 0 | 0 | 1 | 0.0086 |
| 11.6439 | 1 | 37 | 0 | 0 | 1 | 0 | 0.0064 |
| 10.3089 | 1 | 24 | 6 | 99 | 1 | 0 | 0.0106 |
| 7.6962 | 0 | 18 | 11 | 20 | 0 | 5 | 0.0088 |
| 7.1701 | 1 | 23 | 11 | 50 | 0 | 1 | 0.0082 |
| 6.8023 | 0 | 4 | 6 | 0 | 0 | 5 | 0.0069 |
| 6.3969 | 0 | 47 | 6 | 0 | 0 | 1 | 0.0080 |
| 6.3969 | 0 | 20 | 11 | 0 | 0 | 1 | 0.0077 |
| 6.2146 | 1 | -10 | 6 | 40 | 1 | 1 | 0.0082 |

**Interaction**

Interaction terms between education level and other key variables - the number of children, usual working hours per week, and gender – are analyzed to capture how the effect of education on income may vary based on different personal and social contexts. Prior research has shown that educational attainment positively influences earning potential, but childcare responsibilities may moderate this effect (Budig & England, 2001). Including the interaction between education level and the number of children can help clarify whether the income benefit of education diminishes as the number of children increases. Higher educational attainment typically leads to jobs that offer higher wages, which may encourage individuals to work longer hours (Goldin, 2014). This interaction term can identify whether the marginal effect of education on income is amplified for individuals who work more hours. Gender disparities in income persist despite similar educational achievements, often due to occupational segregation, discrimination, or differing career trajectories (Blau & Kahn, 2017). Including this interaction can reveal whether higher education yields different income benefits for men and women, shedding light on gender-based inequalities in the labor market.

**Table 4.** Correlation Matrix of Variables in Interactions and the Interaction Terms without Centering

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CITIZENR | agec | EDUC | UHRSWORK | SEXR | NCHILD | EDUC:  NCHILD | EDUC:  UHRSWORK | EDUC:  SEXR |
| CITIZENR | 1 |  |  |  |  |  |  |  |  |
| agec | 0.2133 | 1 |  |  |  |  |  |  |  |
| EDUC | 0.1807 | -0.1250 | 1 |  |  |  |  |  |  |
| UHRSWORK | -0.0854 | -0.3294 | 0.1127 | 1 |  |  |  |  |  |
| SEXR | -0.1054 | -0.0284 | -0.0821 | 0.1581 | 1 |  |  |  |  |
| NCHILD | -0.0345 | 0.0292 | -0.1103 | 0.1171 | 0.0103 | 1 |  |  |  |
| EDUC:  NCHILD | 0.0403 | 0.0235 | 0.2555 | 0.1254 | -0.0148 | **0.8310** | 1 |  |  |
| EDUC:  UHRSWORK | 0.0704 | -0.2633 | **0.7123** | **0.7128** | 0.0301 | -0.0075 | 0.2485 | 1 |  |
| EDUC:SEXR | -0.0038 | -0.0628 | 0.3913 | 0.1524 | **0.8087** | -0.0375 | 0.1146 | 0.3553 | 1 |

**Table 5.** Correlation Matrix of Variables in Interactions and the Interaction Terms with Centering

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CITIZENR | agec | EDUC | UHRSWORK | SEXR | NCHILD | EDUC:  NCHILD | EDUC:  UHRSWORK | EDUC:  SEXR |
| CITIZENR | 1 |  |  |  |  |  |  |  |  |
| agec | 0.2133 | 1 |  |  |  |  |  |  |  |
| EDUC | 0.1807 | -0.1250 | 1 |  |  |  |  |  |  |
| UHRSWORK | -0.0854 | -0.3294 | 0.1127 | 1 |  |  |  |  |  |
| SEXR | -0.1054 | -0.0284 | -0.0821 | 0.1581 | 1 |  |  |  |  |
| NCHILD | -0.0345 | 0.0292 | -0.1103 | 0.1171 | 0.0103 | 1 |  |  |  |
| EDUCc:  NCHILD | 0.1283 | -0.0068 | 0.6337 | 0.0276 | -0.0432 | -0.1870 | 1 |  |  |
| EDUCc:  UHRSWORK | 0.0416 | 0.1257 | -0.0036 | -0.1000 | -0.0794 | -0.0529 | 0.0963 | 1 |  |
| EDUCc:SEXR | 0.1429 | -0.0663 | **0.7810** | 0.0347 | -0.0438 | -0.0783 | 0.5103 | 0.1327 | 1 |

Table 5 still shows one large correlation - between EDUC and the EDUC and SEXR interaction term - after centering that could potentially lead to collinearity issues, but this is not as large (nothing as high as 0.8) as the correlations we got without centering, which includes two correlations above 0.8.

**Regression Results**

Model 1 presents a simple regression between the independent variable, EDUC (education level), and the dependent variable, INCTOTR (total personal income). In Model 2, controls for CITIZENR (citizenship status), agec (centered age), and UHRSWORK (usual hours worked per week) are included. These two models (Model 1 and Model 2) analyze the untransformed version of the income variable (INCTOTR). Due to violations of linear regression assumptions observed in the first two models, we applied a log transformation to the dependent variable, resulting in INCTOTlog1. Model 3 builds on this transformation. In Model 4, SEXR (gender) and NCHILD (number of children) are added to further refine the model. Outliers identified through diagnostic methods are excluded in Models 5 and 6 to enhance model validity. In Model 6, EDUC is centered, and interaction terms—EDUCc:NCHILD, EDUCc:UHRSWORK, and EDUCc:SEXR—are introduced to explore how the effect of education on income is moderated by the number of children, working hours, and gender.

Since the first two models do not satisfy the assumptions of linear regression, we will focus on the last four models.

We interpret *p*-values at the 0.05 level, where any *p*-values above the critical value of 0.05 indicate insufficient evidence to reject a null hypothesis that there is no relationship between a given independent variable and the response, and *p*-values below the critical value of 0.05 indicate sufficient evidence to reject a null hypothesis. Furthermore, we interpret adjusted *R*2 as a measure of how much variance in the response is explained by the given combination of the independent variables.

**Table 6.** Regression Models

========================================================================================================================================================================================

Dependent variable:

--------------------------------------------------------------------------------------------------------------------------------------------------------------------

INCTOTR INCTOTlog1 INCTOTlog1

(1) (2) (3) (4) (5) (6)

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Constant 4,054,788.000\*\*\* (104,095.700) 5,234,077.000\*\*\* (82,597.000) 8.346\*\*\* (0.066) 8.192\*\*\* (0.070) 8.174\*\*\* (0.069) 8.794\*\*\* (0.057)

EDUC -469,274.900\*\*\* (14,645.820) -231,801.100\*\*\* (11,209.780) 0.085\*\*\* (0.006) 0.090\*\*\* (0.006)

CITIZENR 324,909.100\*\*\* (71,202.740) 0.087\*\* (0.038) 0.108\*\*\* (0.038)

agec -90,073.370\*\*\* (2,064.613) 0.014\*\*\* (0.001) 0.013\*\*\* (0.001)

UHRSWORK -64,461.470\*\*\* (1,936.814) 0.036\*\*\* (0.001) 0.034\*\*\* (0.001)

SEXR 0.243\*\*\* (0.037)

NCHILD 0.032\*\* (0.014)

EDUC 0.093\*\*\* (0.006)

EDUCc 0.116\*\*\* (0.011)

CITIZENR 0.098\*\*\* (0.037) 0.097\*\*\* (0.037)

agec 0.014\*\*\* (0.001) 0.014\*\*\* (0.001)

UHRSWORK 0.034\*\*\* (0.001) 0.034\*\*\* (0.001)

SEXR 0.235\*\*\* (0.037) 0.247\*\*\* (0.037)

NCHILD 0.036\*\*\* (0.014) 0.034\*\* (0.014)

EDUCc\_NCHILD -0.010\*\* (0.004)

EDUCc\_UHRSWORK 0.001\*\*\* (0.0004)

EDUCc\_SEXR -0.020 (0.012)

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Observations 2,800 2,800 2,483 2,483 2,474 2,474

R2 0.268 0.661 0.329 0.341 0.353 0.356

Adjusted R2 0.268 0.660 0.328 0.340 0.351 0.354

Residual Std. Error 2,640,214.000 (df = 2798) 1,798,611.000 (df = 2795) 0.908 (df = 2478) 0.900 (df = 2476) 0.880 (df = 2467) 0.879 (df = 2464)

F Statistic 1,026.662\*\*\* (df = 1; 2798) 1,361.581\*\*\* (df = 4; 2795) 303.565\*\*\* (df = 4; 2478) 213.772\*\*\* (df = 6; 2476) 224.030\*\*\* (df = 6; 2467) 151.519\*\*\* (df = 9; 2464)

========================================================================================================================================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

**Association of individual variables**

In Model 1 and Model 2, we can see that adding controls for CITIZENR (citizenship status), agec (centered age), and UHRSWORK (usual hours worked per week) reduced the effect magnitude of education level, indicating that citizenship status, age, and working hours are reasonable proxies for more significant opportunities and competence. EDUC (education level) positively correlates with income, as evidenced by its consistently positive coefficient from Model 3 to Model 6. Citizenship status also shows a positive association, suggesting that citizens tend to have higher incomes than non-citizens. Age and working hours display positive associations, indicating that greater age and longer working hours correspond to higher income. NCHILD (number of children) also demonstrates positive coefficients, implying that being male and having more children are linked to higher income. All the variables, the interactions between education level and the number of children, and the interaction between education level and working hours show significant correlations with p-values smaller than 0.05. Those interaction terms reveal that the effects of education can vary depending on these factors. However, the interaction between education level and gender is insignificant, suggesting that education's impact on income is similar for both males and females.

**Variance explained**

The education level alone explained 26.8% of the variance in personal income. Adding controls for citizenship status, age, and usual working hours significantly improved the model fit. However, the adjusted R² value of 66% appeared unusually high. After applying a log transformation to personal income, the adjusted R² decreased to a more reasonable level of 32.8%. Including the number of children and gender as additional predictors did not substantially improve the model fit. Removing outliers further increased the adjusted R² to 0.351. Finally, incorporating interaction terms between education level and other predictors slightly improved the model fit, yielding an adjusted R² of 0.354, which means the last model explains 35.4% of the variance in personal income.

Adding controls for citizenship status, age, and working hours significantly improved the model, but it is less clear if the addition of the number of children and gender leads to a significant improvement in the model. We investigated this by a series of partial *F*-tests, shown in the ANOVA table in Table (6).

**Table 7**. ANOVA Comparisons of the Models Before Outliers Dropping

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Res *df* | RSS | *df* | Sum of Sq | *F* | Pr(>*F*) |
| (Intercept) | 2482 | 3042.1 |  |  |  |  |
| Model 3 | 2478 | 2041.6 | 4 | 1000.43 | 309.022 | <0.001 |
| Model 4 | 2476 | 2004.0 | 2 | 37.67 | 23.272 | <0.001 |

As we can see, although the addition of the number of children and gender does not add a substantial amount to the Adjusted *R*-squared, a partial *F*-test shows that the addition does significantly improve the model fit at the 0.05 level, *F*(2, 2476) = 37.67, p <0.001. Thus, we prefer Model 4 to Model 5.

**Final Model**

We used the AIC/n (Akaike Information Criterion per observation) to further compare Model 4, Model 5, and Model 6, with values of 2.63, 2.59, and 2.58, respectively. The lower AIC/n for Model 6 indicates a better balance between model fit and complexity, which aligns with the findings from the adjusted R² values. Consequently, Model 6 is selected as our final model.

**Contribution of Education Level**

In Model 6, the coefficient of 0.116 for EDUCc (centered educational attainment) suggests that for each additional year of education, the log of income increases by 0.116 points, holding all other variables constant. The range of EDUC is 11 -0 =11, meaning that a person with no formal education is expected to have a log of income 11 \*0.116 =1.276 points lower than someone with five or more years of college education, holding all other variables constant. However, this effect is moderated by the number of children (NCHILD) and gender (SEXR) due to the interaction terms. The interaction between education and the number of children shows a coefficient of **-**0.01, meaning that the positive effect of education on income decreases slightly with each additional child. The interaction between education and usual hours worked per week has a coefficient of 0.001, indicating that the longer weekly working hours can further augment the advantage of a higher education level.

**Hierarchical Regression**

The data were clustered at the family level, and a hierarchical linear mixed-effects model was employed to analyze the relationship between educational attainment and personal income. In this model, EDUC was treated as a Level-1 predictor with random effects, while other variables were included as fixed effects. To assess the random effect of educational attainment, we first ran a baseline model with only an intercept in random effects. We compared it to a model including EDUC using an ANOVA test, shown in Table (8).

**Table 8.** ANOVA Comparisons of the Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | df | AIC | BIC | logLik | L.Ratio | p-value |
| (Intercept) | 11 | 6375.623 | 6439.572 | -3176.811 |  |  |
| Model Random | 13 | 6349.924 | 6425.501 | -3161.962 | 29.699 | <0.001 |

The random-effects model shows improved model fit, as indicated by a lower Akaike Information Criterion (AIC = 6349.924) and Bayesian Information Criterion (BIC = 6425.501). The log-likelihood increases from -3176.811 in the intercept-only model to -3161.962 in the random-effects model. The likelihood ratio test (L. Ratio = 29.699, *p* < .001) confirms that the addition of random effects for educational attainment significantly improves the model fit. These results justify the inclusion of random effects for EDUC, suggesting that variation in personal income attributable to education differs across families.

We will only focus on the random effects of the hierarchical regression model, shown in Table (9).

**Table 9.** Random Effects of Educational Attainment on Family Level

|  |  |
| --- | --- |
| Random Effects | SD |
| Intercept (Family Level) | 0.4303 |
| EDUCc (Family Level) | 0.0835 |
| Residual | 0.7670 |

The between-family variation was 0.4303, and the within-family variation was 0.8945. The intraclass correlation coefficient (ICC) was calculated as 0.4303 / (0.4303 + 0.7670) = 35.94%. This indicates that 35.94% of the variation in the log-transformed personal income is attributable to differences between families, while the remaining variation is within families. Therefore, more variation in personal income exists within families than between families.

**Limitation**

This study has several limitations. First, the dataset includes respondents across all age groups, meaning children, students, and retirees are not excluded. This could introduce bias in the analysis of personal income since these groups typically do not participate in the labor market the same way as working-age adults. Second, missing data were handled by direct deletion, which may have excluded respondents with missing values for key variables without investigating the underlying reasons for missingness. This could lead to potential selection bias and affect the generalizability of our findings.

Additionally, while the study incorporates interaction terms between education and other variables (such as gender, working hours, and number of children), these interactions may not fully capture the complexity of socio-economic relationships. The model also assumes a linear relationship between predictors and income, which may oversimplify more nuanced, non-linear relationships. Lastly, despite accounting for family-level clustering in the hierarchical model, other contextual factors, such as regional differences or occupational categories, were not included, limiting the scope of our analysis.

**Conclusion**

This study provides a detailed examination of the factors influencing personal income from 403.43 dollars to around 3.3 million dollars (log of personal income from 6 to 15) in the United States, focusing on education, citizenship status, gender, working hours, and family structure. The analysis highlights that education significantly contributes to income, but this relationship is influenced by additional factors such as citizenship and gender. Hierarchical regression models revealed that 38.27% of the variation in personal income is attributable to differences between families, while most variation occurs within families.

The inclusion of interaction terms showed that family responsibilities and work patterns might moderate the positive effects of education on income, although some interactions, such as education and gender, were not significant. Despite the identified limitations, the findings underscore the importance of addressing socioeconomic disparities through targeted policies that consider educational attainment, citizenship status, and family dynamics. Future research should explore non-linear relationships and include a broader range of contextual variables to provide a more comprehensive understanding of income inequality in the United States.

**Reference**

* Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2020). Race and Economic

Opportunity in the United States: An Intergenerational Perspective. *Quarterly Journal of Economics, 135*(2), 711–783. <https://doi.org/10.1093/qje/qjz042>

* Killewald, A., & Bearak, J. (2014). Is the Motherhood Penalty Larger for Low-Wage

Women? A Comment on Quantile Regression. *American Sociological Review, 79*(2), 350–357. <https://doi.org/10.1177/0003122414524577>

* Pager, D., & Shepherd, H. (2008). The Sociology of Discrimination: Racial

Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annual Review of Sociology, 34*, 181–209. <https://doi.org/10.1146/annurev.soc.33.040406.131740>

* Blau, F. D., & Kahn, L. M. (2017). The Gender Wage Gap: Extent, Trends, and

Explanations. *Journal of Economic Literature, 55*(3), 789–865. https://doi.org/10.1257/jel.20160995

* Budig, M. J., & Hodges, M. J. (2010). Differences in Disadvantage: Variation in the

Motherhood Penalty across White Women’s Earnings Distribution. *American Sociological Review, 75*(5), 705–728. https://doi.org/10.1177/0003122410381593

* Killewald, A. (2013). A Reconsideration of the Fatherhood Premium: Marriage,

Coresidence, Biology, and Fathers’ Wages. *American Sociological Review, 78*(1), 96–116. <https://doi.org/10.1177/0003122412472341>

* Budig, M. J., & England, P. (2001). The wage penalty for motherhood. *American*

*Sociological Review*, 66(2), 204–225. https://doi.org/10.2307/2657415

* Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic*

*Review*, 104(4), 1091–1119. https://doi.org/10.1257/aer.104.4.1091

**Appendix**

|  |
| --- |
| install.packages('ipumsr')  library(ipumsr)  ddi <- read\_ipums\_ddi("/Users/mac/Desktop/usa\_00002.xml")  data <- read\_ipums\_micro(ddi)  df <- as.data.frame(data)  # set the representative larger than 500  df1 <- na.omit(data[df$PERWT > 500,])  table(df1$CITIZEN)  table(df1$EDUC)  # data cleaning  df1$CITIZENR <- ifelse(df1$CITIZEN == 1, 1,  ifelse(df1$CITIZEN == 2, 1,  ifelse(df1$CITIZEN == 0, NA,0)))  df1$RACER <- ifelse(df1$RACE == 1,1,0)  df1$SEXR <- ifelse(df1$SEX == 1,1,0)  df1$INCTOTR <- ifelse(df1$INCTOT <= 0, NA, df1$INCTOT)  df1$agec <- df1$AGE - mean(df1$AGE)  dfR <- na.omit(df1[,c('INCTOTR','CITIZENR','EDUC','UHRSWORK',  'RACER','SEXR','NCHILD','SERIAL','agec')])  # regression model  lm\_education <- lm(dfR$INCTOTR ~ dfR$EDUC)  summary(lm\_education) # control model  lm\_full <- lm(dfR$INCTOTR ~ dfR$EDUC + dfR$CITIZENR + dfR$agec + dfR$UHRSWORK)  summary(lm) # model with iv  # diagnosis  ##### Linearity #####  library(car)  scatterplotMatrix(dfR[,c('INCTOTR','agec','EDUC','UHRSWORK')],  cex = .5,  pch = 16,  col = rgb(0,0,0,1/32),  diagonal=list(method ="histogram",  breaks = 20),  cex.labels = 0.5,  regLine=list(method = lm,  lty = 1,  lwd = 1,  col = 1),  smooth = list(method = "loessLine",  lty.smooth = 2,  lwd.smooth = 1,  col.smooth = 2,  lty.spread = 3,  lwd.spread = 1,  col.spread = 2))  # correction of INICTOT  dfR$INCTOTlog <- ifelse(dfR$INCTOTR <= 1, 0, ifelse(dfR$INCTOTR > 1, log(dfR$INCTOTR),NA))  dfR$INCTOTlog1 <- ifelse(dfR$INCTOTlog > 15, NA,  ifelse(dfR$INCTOTlog < 6, NA, dfR$INCTOTlog))  # only investigate log of income between 6 and 15  hist(dfR$INCTOTlog1)  scatterplotMatrix(dfR[,c('INCTOTlog1','agec','EDUC','UHRSWORK')],  cex = .5,  pch = 16,  col = rgb(0,0,0,1/32),  diagonal=list(method ="histogram",  breaks = 20),  cex.labels = 0.5,  regLine=list(method = lm,  lty = 1,  lwd = 1,  col = 1),  smooth = list(method = "loessLine",  lty.smooth = 2,  lwd.smooth = 1,  col.smooth = 2,  lty.spread = 3,  lwd.spread = 1,  col.spread = 2))  lm1 <- lm(dfR$INCTOTlog1 ~ dfR$EDUC + dfR$CITIZENR + dfR$agec + dfR$UHRSWORK)  summary(lm1) # new model after correction  # family size is no longer significant  ##### Normality of residuals #####  df\_resid <- resid(lm1)  df\_fitted <- fitted(lm1)  hist(df\_resid,breaks = 50)  ##### Homoscedasticity #####  df\_resid1 <- resid(lm1)  df\_fitted1 <- fitted(lm1)  scatterplot(df\_fitted1,df\_resid1,  jitter = list("x" = 1,  "y" = 0))  ##### Omitted Relevant Variables #####  ## Correlation analysis  # multicollinearity  cor\_matrix <- cor(dfR[,c('CITIZENR','agec','EDUC','UHRSWORK','RACER','SEXR','NCHILD')])  cor\_matrix[upper.tri(cor\_matrix)] <- NA  round(cor\_matrix, 4)  cor.test(dfR$SEXR,dfR$INCTOTlog1,use = "complete.obs") #significantly positive  cor.test(dfR$RACER,dfR$INCTOTlog1,use = "complete.obs") # not significant  cor.test(dfR$NCHILD,dfR$INCTOTlog1,use = "complete.obs") #significantly positive  lm2 <- lm(dfR$INCTOTlog1 ~ dfR$EDUC + dfR$CITIZENR + dfR$agec + dfR$UHRSWORK +  dfR$SEXR + dfR$NCHILD) # new model with an omitted variable - gender  summary(lm2)  summary(dfR)  round(apply(dfR, 2, sd), 4)  # outlier  # influence will be used  dfR1 <- na.omit(dfR[,c('INCTOTlog1','CITIZENR','agec','EDUC','UHRSWORK',  'SEXR','NCHILD','SERIAL')])  dfR1$coosd <- cooks.distance(lm2)  large\_cooksd <- subset(dfR1,abs(dfR1$coosd) > 4/2483)  177/2483 # 7% too large  plot(lm2,which = 4)  round(quantile(dfR1$coosd[dfR1$coosd > 4/2483],probs = seq(0,1,0.05)),4)  large\_cooksd1 <- subset(dfR1,abs(dfR1$coosd) > 0.0064)  lm3 <- lm(INCTOTlog1 ~ EDUC + CITIZENR + agec + UHRSWORK+ SEXR + NCHILD, data = dfR1[dfR1$coosd <= 0.0064,])  summary(lm3)  # interaction  dfR1$EDUC\_CITIZENR <- dfR1$EDUC \* dfR1$CITIZENR  dfR1$EDUC\_agec <- dfR1$EDUC \* dfR1$agec  dfR1$EDUC\_UHRSWORK <- dfR1$EDUC \* dfR1$UHRSWORK  dfR1$EDUC\_SEXR <- dfR1$EDUC \* dfR1$SEXR  dfR1$EDUC\_NCHILD <- dfR1$EDUC \* dfR1$NCHILD  cor\_matrix <- cor(dfR[,c('CITIZENR','agec','EDUC','UHRSWORK','SEXR','NCHILD','EDUC\_NCHILD','EDUC\_UHRSWORK','EDUC\_SEXR')])  cor\_matrix[upper.tri(cor\_matrix)] <- NA  round(cor\_matrix, 4)  dfR1$EDUCc <- dfR1$EDUC-mean(dfR1$EDUC)  dfR1$EDUCc\_NCHILD <- dfR1$EDUCc \* dfR1$NCHILD  dfR1$EDUCc\_SEXR <- dfR1$EDUCc \* dfR1$SEXR  dfR1$EDUCc\_UHRSWORK <- dfR1$EDUCc \* dfR1$UHRSWORK  cor\_matrix <- cor(dfR1[,c('CITIZENR','agec','EDUCc','UHRSWORK','SEXR','NCHILD','EDUCc\_NCHILD','EDUCc\_UHRSWORK','EDUCc\_SEXR')])  cor\_matrix[upper.tri(cor\_matrix)] <- NA  round(cor\_matrix, 4)  lm4 <- lm(INCTOTlog1 ~ EDUCc + CITIZENR + agec + UHRSWORK + SEXR + NCHILD + EDUCc\_NCHILD +  EDUCc\_UHRSWORK + EDUCc\_SEXR, data = dfR1[dfR1$coosd < 0.0064,])  summary(lm4)  # regression models summary  lm\_intercept2 <- lm(dfR$INCTOTlog1 ~ 1)  lm\_intercept3 <- lm(INCTOTlog1 ~ 1, data = dfR1[dfR1$coosd < 0.0064,])  library(stargazer)  stargazer(lm\_education, lm\_full, lm1, lm2, lm3, lm4,  type = "text",  header = F,  intercept.bottom = F,  no.space = T,  single.row = T)  anova(lm\_intercept2,lm1,lm2)  AIC(lm2)/(2483)  AIC(lm3)/(2483-9)  AIC(lm4)/(2483-9)  # Hierarchical modeling  library(nlme)  dfR1$family <- dfR1$SERIAL  table(table(dfR1$family))  ##Fully Unconditional Model##  model1 <- lme(INCTOTlog1 ~ 1,  random = ~ 1| family,  data = dfR1[dfR1$coosd < 0.0064,],  method = 'ML')  summary(model1)  # 0.5274321/(0.5274321+0.971962) = 35.18%  ## multiple model ##  # educ: level 1 and random effects  # CHILDR: level 2 and fixed effects  # Interaction is based on the assumption that  # the difference in EDUC impact may be explained by the level-2 IV  model2 <- lme(INCTOTlog1 ~ EDUCc + CITIZENR + agec + UHRSWORK + SEXR + NCHILD + EDUCc\_NCHILD +  EDUCc\_UHRSWORK,  random = ~ EDUC| family,  data = dfR1[dfR1$coosd < 0.0064,],  method = 'ML')  summary(model2)  0.4303 / (0.4303 + 0.7670)  ## to test if EDUC has random effects  model3 <- lme(INCTOTlog1 ~ EDUCc + CITIZENR + agec + UHRSWORK + SEXR + NCHILD + EDUCc\_NCHILD +  EDUCc\_UHRSWORK,  random = ~ 1| family,  data = dfR1[dfR1$coosd < 0.0064,],  method = 'ML')  summary(model3)  anova(model3,model2) # yes |