FEM 11087 - Applied Microeconometrics

Assignment 2: Panel Data Analysis

Empirical Application

Group 33

Alejandra Betancourt 778686
Yinli Hu 744678
Teun Mulder 565586
Andrés Piñón 775387

30 September 2025

Question 1 [0.7 points]

A central question in labor economics is: ***How much more do individuals earn with higher levels of education?*** *Economists often estimate the* returns to education*—that is, the increase in earnings associated with completing high school, college, or additional years of schooling.*

Using the panel data provided, begin by constructing a ***bar chart*** *showing* ***mean income by education group****. Group individuals based on their* ***highest level of educational attainment*** *(e.g., less than high school, high school graduate, some college, college degree or more), and plot the* ***average income*** *for each category.*

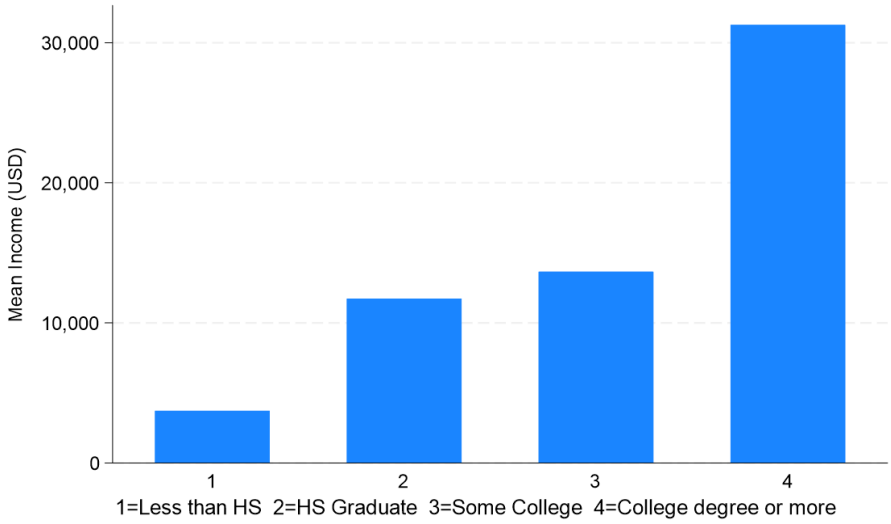

Figure 1.1: Mean income by education level

Recent debates around student debt and the value of higher education often assume that education "pays off" equally for everyone. ***Does your analysis support that assumption?*** *To explore this, create* ***separate plots by gender*** *to highlight any differences in the relationship between education and earnings. Discuss your findings.*

Note: *For this question, create and use a categorical education variable based on each individual's highest level of education completed across the panel. Construct four categories:*

- *Less than high school (11 or fewer years)*
- *High school graduate (exactly 12 years)*
- *Some college (13 to 15 years)*
- *College degree or more (16 or more years)*

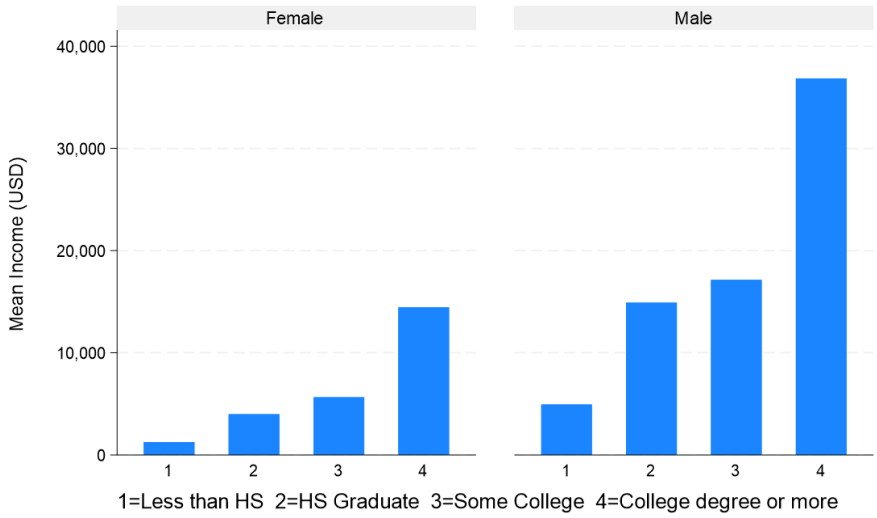

Figure 1.2: Mean income by education level, by gender

Question 2 [1 point]

Now, we turn to formally estimating the effect of years of education on income. Use ***pooled OLS*** *to examine the impact of years of education (edyears) on* ***log(income)****, controlling for age, gender (male), marital status categories, ethnicity categories, and childbirth.*

Table 2.1: Pooled OLS

	Coefficient	Robust std. err.	P>|t|
Treatment variable			
Education years	0.1458	(0.0031)	0.000
Control variables			
Age	0.1263	(0.0017)	0.000
Male	0.6000	(0.0115)	0.000
Childbirth	-0.0795	(0.0198)	0.000
Black	-0.4286	(0.0139)	0.000
Hispanic	-0.0825	(0.0147)	0.000
Mixed Race (Non-Hispanic)	-0.3744	(0.0627)	0.000
Married	0.6189	(0.0211)	0.000
Separated or divorced	0.0767	(0.0472)	0.104
Widowed	0.0647	(0.2690)	0.810
Constant			
Constant	3.0156	(0.0318)	0.000
Observations	55874		
Prob > F	0.0000		
R-squared	0.4089		
Adj. R-squared	0.4088		

Dependent variable: log(income)

a) What is the estimated return to an additional year of education? Interpret the coefficient on years of education in terms of its ***sign***, ***magnitude***, *and* ***statistical significance***.

b) Differences in returns to schooling by gender are sometimes interpreted as potential evidence of ***labor market discrimination***. *Test whether the effect of years of education using the categorical variable created in Question 1 on log(income) is the* ***same for men and women***. *Based on your results, do you find any evidence consistent with discrimination?*

Table 2.2: Pooled OLS with interaction effects

	Coefficient	Robust std. err.	P>|t|
Treatment variable			
HS graduate	0.3802	(0.0219)	0.000
Some college	0.6488	(0.0248)	0.000
College degree or more	1.0239	(0.0409)	0.000
HS graduate × Male	0.3836	(0.0278)	0.000
Some college × Male	0.1664	(0.0307)	0.000
College degree or more × Male	0.2206	(0.0454)	0.000
Control variables			
Age	0.1225	(0.0017)	0.000
Male	0.4472	(0.0147)	0.000
Childbirth	-0.0759	(0.0196)	0.000
Black	-0.4143	(0.0138)	0.000
Hispanic	-0.0732	(0.0146)	0.000
Mixed Race (Non-Hispanic)	-0.3752	(0.0619)	0.000
Married	0.6013	(0.0209)	0.000
Separated or divorced	0.0631	(0.0468)	0.178
Widowed	0.0932	(0.2497)	0.709
Constant			
Constant	4.4713	(0.0307)	0.000
Observations	55874		
Prob > F	0.0000		
R-squared	0.4195		
Adj. R-squared	0.4193		

Dependent variable: log(income)

Table 2.3: Joint test

F-statistic	66.60
Prob > F	0.0000

(1) HS graduate × Male = 0
(2) Some college × Male = 0
(3) College degree or more × Male = 0

c) Under what conditions is the pooled OLS estimate of the effect of years of education ***unbiased and efficient****? Do you believe these conditions are likely to hold in this context?*

Question 3 [0.5 points]

So far, the panel structure of the data has been largely unexploited. Random effects (RE) estimation can improve the efficiency of the estimates compared to pooled OLS.

a) Estimate the effect of years of education (edyears) on ***log(income)*** *using the* ***random effects*** *(RE) model, controlling for age, gender (male), marital status categories, ethnicity categories, and childbirth. Interpret the estimated coefficient for years of education in terms of its* ***sign***, ***magnitude***, *and* ***statistical significance****. Then, compare the RE estimate and standard error of the education coefficient with those obtained from the* ***pooled OLS model****.*

b) Under which conditions and why can the random effects estimator be ***more efficient*** *than pooled OLS?*

Question 4 [1.55 points]

Alternatively, the panel structure of the data can be used to perform ***fixed effects (FE)*** *estimation.*

a) Based on theoretical considerations, would you ***prefer*** *fixed effects or random effects estimation? Justify your answer.*

b) Use a ***fixed effects estimator*** *to examine the impact of years of education (edyears) on* ***log(income)****, controlling for age, gender (male), marital status categories, ethnicity categories, and childbirth. Interpret the coefficient on years of education in terms of its* ***sign***, ***magnitude***, *and* ***statistical significance****. Compare your results with those from the* ***pooled OLS*** *and* ***random effects*** *models.*

c) Perform the ***Hausman*** *test. What do the results indicate? Based on the test outcome,* ***which estimator*** *(RE or FE) is more appropriate in this context?*

Question 5 [0.9 points]

*Next, estimate a **Correlated Random Effects (CRE)** model to examine the effect of years of education (edyears) on **log(income)**.*

*a) What is one advantage of the **CRE estimator** compared to the **random effects (RE)** estimator?*

*b) What is one advantage of the **CRE estimator** compared to the **fixed effects (FE)** estimator?*

*c) Compare the estimated coefficient for years of education from the **CRE model** with those from the **RE** and **FE** models. Are the coefficients similar or different? Explain why this is the case.*

*d) Based on your CRE estimates, does the assumption of **exogeneity** appear to hold? Which estimator would you consider most appropriate in this context?*

Question 6 [0.9 points]

Recent research provides compelling evidence that after the birth of a first child, women's earnings decline sharply and remain persistently lower, while men's earnings remain largely unaffected.

*a) Estimate the effect of childbirth on **log(income)** using the **most appropriate model**. Control for age, gender (male), marital status categories, ethnicity categories, and years of education (edyears). Interpret the estimated coefficient for childbirth in terms of its **sign**, **magnitude**, and **statistical significance**.*

*b) Test whether the effect of childbirth on log(income) **differs** between males and females. What conclusions can you draw from your results?*

Question 7 [1.2 points]

Without conducting any empirical analysis:

*a) Compare the key assumptions underlying **pooled OLS**, **fixed effects (FE)**, and **random effects (RE)** estimators. Discuss theoretically in which scenarios you would prefer to use each method.*

b) Within the practical context of this assignment (effect of education on earnings), provide an example situation for each estimator in the form of a ***Directed Acyclic Graph (DAG)****. For each case (Pooled OLS, FE, and RE), explain why the assumptions required for the respective method hold in that example, and why that method would be preferred.*

Question 8 [0.75 points]

Finally, revisit your data and evaluate whether ***attrition*** *is present in your sample. Based on your preferred model, discuss the likelihood of* ***attrition bias****. What conclusions can you draw regarding its presence, and how might it affect the validity of your results?*