Assignment 2: Panel Data Analysis

FEM 11087 – Applied Microeconometrics

Fall 2025

In this assignment, you will apply various estimation techniques used in panel data analysis, discuss their underlying assumptions, and evaluate the appropriateness of each estimator.

The assignment consists of three parts and carries a maximum score of 10 points.

Part 1 – Empirical Analysis (7.5 points)

In the first part, you will follow a set of guiding questions and implement different panel data estimators to examine the relationship between childbirth and income. You will use longitudinal real-world data from the National Longitudinal Survey of Youth (NLSY) in the United States.

You are required to submit the following on Canvas:

- A **PDF** Report named 'groupnumber.pdf' under 'Assignment 2. Report'. This report should include:
  - All commands used
  - Relevant STATA output
  - Answer to the questions
- A **.do** file named 'groupnumber.do' under 'Assignment 2. Do-file', containing all commands used in your analysis.

We will review the correct functioning of your .do file. Only **one submission per group** is required. All group members are jointly responsible for the contents of the submitted material.

Deadline: September 30th at 19:00

Part 2 – Literature discussion (2 points)

In this part you will:

- Individually read a published academic article that applies a panel data estimator
- Submit a report (maximum one page) under 'Assignment 2. Discussion' in Canvas responding to a set of discussion questions shared during the tutorial session

The discussion groups will be assigned during the tutorial session. **Attendance to the tutorial session is mandatory** to receive a grade for this part. Only **one submission per discussion group** is required. All group members are jointly responsible for the contents of the submitted material.

Deadline: October 3rd at 17:00

Part 3 – Peer assessment (0.5 points)

After completing the group work, you will individually reflect on the contributions of your fellow group members in a constructive and respectful manner.

Your grade for this component will be based on the evaluations you receive from your group members.

Note:

- Lecturers reserve the right to disregard peer assessments in cases of unusual or inappropriate grading patterns.
- **Failure to submit peer assessments for all group members will result in a zero** for this part of the assignment.

Submit your peer assessments individually under *'Assignment 2. Peer assessment'* on Canvas.

Deadline: October 3rd at 17:00

Part 1. Empirical Application [7.5 points]

In this assignment, you will evaluate the effect of years of education on earned income. The dataset **Assignment2.dta** contains data from 1997 to 2017 for a representative sample of the U.S. population who were between 12 and 16 years old in 1996. The **National Longitudinal Survey of Youth 1997 (NLSY97)** is a longitudinal survey that provides detailed information on the life cycle of young individuals, including variables related to employment, income, educational attainment, health, and participation in welfare programs, among others.

Table 1: List of variables in dataset childbirth.dta

pid	Personal identifier variable: Allows to group observations for the same person
wave	Wave in which the information was collected (2-18 waves)
age	Age in completed years at the time of interview
male	Indicator of gender: male (1) or female (0)
edyears	Number of years of schooling
child_birth	Indicator whether a child was born in after the past wave (1) or not (0)
income	Wage or salary income in US dollars of the individual in the past year
Mstatus	Categorical variable of the marital status of the respondent: (0) Never married, (1) Married, (2) Separated or divorced and (3) Widowed
Ethnicity	Categorical variable of the self-reported ethnicity of the respondent: (1) Black (2) Hispanic, (3) Mixed-race non-Hispanic and (4) Non-black/Non-Hispanic

The **do-file Stata Application.do** from the panel data models module contains a list of Stata commands that may be helpful for completing this assignment. To open a do-file, **do not double-click** the file directly. Instead, follow these steps:

1. In Stata, go to the **Window** menu and select **Do-file Editor**.
2. In the Do-file Editor, go to **File > Open**, and then select the .do file.

You should always open a **log file** at the start of your Stata session—and close it at the end—to save all your output. The log file automatically records all commands used and results obtained, **except for graphs**. You can open the log file in Word and edit it by adding your answers to the assignment and removing any code or output that is not relevant.

As explained above, a complete assignment submission must include:

- The Stata output,
- The commands used to generate this output,
- Your answers to the questions provided.

Remember, the main objective of the assignment is to demonstrate your ability to **interpret** the results you obtain.

You are expected to answer all questions in your own words. All assignments will be reviewed for plagiarism. According to ESE's regulation, "(plagiarism is) also understood to mean to copy from one's own or someone else's (group)work an extract larger than a couple of words literally or translated for the purpose of a paper, thesis or any other form of text being part of the teaching without indicating this by means of quotation marks or another univocal typographic means, even if bibliographically traceable and correct acknowledgements are included" (https://my.eur.nl/en/ese/information-desk/regulations/rules-and-regulations)

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.

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)

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.

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?

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?

Part 2. Academic paper discussions [2 points]

Each student will be randomly assigned one academic article from the list below, each of which employs a panel data estimation technique. You are expected to prepare **individually** for the tutorials, where you will explain your assigned paper to a group of students who are unfamiliar with it.

When reading your paper, consider the following questions:

1. What is the research question addressed in the paper?
2. What data is used?
3. What model(s) are estimated? What assumptions underlie the chosen method, and are these assumptions satisfied?
4. What are the main findings of the paper?
5. How do you evaluate the methodology? Do you think the estimation technique used was appropriate?

During the tutorial sessions, you will be randomly assigned to a new discussion group. In these groups, you will discuss the different papers and collaboratively draft a one-page report responding to a set of guiding questions.

List of articles

1. Chay, K. Y., & Greenstone, M. (2003). The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. *The quarterly journal of economics*, *118*(3), 1121-1167.
2. McInerney, Melissa, and Jennifer M. Mellor. "Recessions and seniors' health, health behaviors, and healthcare use: Analysis of the Medicare Current Beneficiary Survey." *Journal of health economics* 31.5 (2012): 744-751.
3. Ruhm, Christopher J. "Are recessions good for your health?." *The Quarterly journal of economics* 115.2 (2000): 617-650.
4. Ásgeirsdóttir, Tinna Laufey, et al. "Lifecycle effects of a recession on health behaviors: Boom, bust, and recovery in Iceland." *Economics & Human Biology* 20 (2016): 90-107.

Part 3. Individual peer assessment [0.5 points]

After completing the group assignment, you will reflect constructively on the contributions of each of your fellow group members. This process is designed to enhance your teamwork skills and promote more effective collaboration in the future. Please be fair and honest in your assessments, and provide written, constructive feedback to help everyone learn and improve their ability to work in teams.

You provide your peer assessment using this rubric:

	Capstone 10	**Milestone high 9**	**Milestone low 8**	**Benchmark 6**	**Does not meet the benchmark 0**
Contributes to team meetings	Helps the team move forward by articulating the merits of alternative ideas or proposals.	Offers alternative solutions or courses of action that build on the ideas of others.	Offers new suggestions to advance the work of the group.	Shares ideas but does not advance the work of the group.	
Facilitate the contributions of team members	Engages team members in ways that facilitate their contributions to meetings by both constructively building upon or synthesizing the contributions of others as well as noticing when someone is not participating and inviting them to engage.	Engages team members in ways that facilitate their contribution to meetings by constructively building upon or synthesizing the contributions of others.	Engages team members in ways that facilitate their contribution to meetings by restating the views of other team members and/or asking questions for clarification.	Engages team members ty taking turns and listening to others without interrupting.	
Individual contributions outside of team meetings	Completes all assigned tasks by deadline; work accomplished is thorough, comprehensive, and advances the project. Proactively helps other team members complete their assigned tasks to a similar level of excellence.	Completes all assigned tasks by deadline; work accomplished is thorough, comprehensive and advances the project.	Completes all assigned tasks by deadline; work accomplished advances the project.	Completes all assigned tasks by deadline.	
Fosters constructive team climate	Supports a constructive team climate by doing all of the following:	Supports a constructive team climate by doing any three of the following: Treats team members respectfully by being polite	Supports a constructive team climate by doing any two of the following: Treats team members respectfully by being polite	Supports a constructive team climate by doing any one of the following: Treats team members respectfully by being polite	

	Treats team members respectfully by being polite and constructive in communication. Uses positive vocal or written tone, facial expressions, and/or body language to convey a positive attitude about the team and its work. Motivate teammates by expressing confidence about the importance of the task and the team's ability to accomplish it. Provides assistance and/or encouragement to team members.	and constructive in communication. Uses positive vocal or written tone, facial expressions, and/or body language to convey a positive attitude about the team and its work. Motivate teammates by expressing confidence about the importance of the task and the team's ability to accomplish it. Provides assistance and/or encouragement to team members.	and constructive in communication. Uses positive vocal or written tone, facial expressions, and/or body language to convey a positive attitude about the team and its work. Motivate teammates by expressing confidence about the importance of the task and the team's ability to accomplish it. Provides assistance and/or encouragement to team members.	and constructive in communication. Uses positive vocal or written tone, facial expressions, and/or body language to convey a positive attitude about the team and its work. Motivate teammates by expressing confidence about the importance of the task and the team's ability to accomplish it. Provides assistance and/or encouragement to team members.	
Responds to conflict	Addresses destructive conflict directly and constructively, helping to manage/resolve it in a way that strengthens overall team cohesiveness and future effectiveness.	Identifies and acknowledges conflict and stays engaged with it.	Redirecting focus toward common ground, toward task at hand (away from conflict).	Passively accepts alternate viewpoints/ideas/opinions.	