Assignment 1: Regression Analysis with Cross-Sectional Data, Endogeneity and Instrumental Variable Estimation

FEM 11087 - Applied Microeconometrics

Fall 2025

In this assignment you will apply the OLS estimator of the linear regression model, discuss the assumptions necessary for unbiasedness of this estimator, and apply the instrumental variable (IV) estimator.

This assignment consists of three parts and has a maximum score of 10 points.

Part 1 – Empirical Analysis (7.5 points)

In the first part, you will apply OLS and IV estimators to study the relationship between BMI and income using a modified dataset from the Add Health study.

You are required to submit the following on Canvas:

- A **PDF** Report named 'groupnumber.pdf' under 'Assignment 1. Report'. This report should include:
  - All commands used
  - Relevant STATA output
  - Answer to the questions
- A **.do** file named 'groupnumber.do' under 'Assignment 1. 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 16th at 19:00

Part 2 – Presentation (2 points)

In this part you will:

- Prepare a 7-minute presentation on an academic article that uses the IV estimator. You will present this during the tutorial sessions on week 3.
- Submit a .pdf file 'groupnumber_presentation.pdf' with the slides of your presentation (under "Assignment 1. Slides")

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: September 16th at 19:00

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 1. Peer assessment' on Canvas.

Deadline: September 19th at 17:00

Part 1. Empirical application [7.5 points]

The focus of this case-study is the relationship between an individual's income and their body mass index (BMI). The dataset **income_bmi.dta** contains information on income and other individual characteristics, including information needed to compute individual BMI. The **Add Health study** covers a representative sample of US citizens who were adolescents in 1994-1995. We use data collected from 2015-2017 and includes information on participants and their parents (biological, adoptive, or stepparent). More information on the datasets is available here: https://dataverse.unc.edu/dataverse/addhealth.

For this assignment, we will use a modified dataset that contains information on participants' parents. Therefore, our unit of analysis is the parents. We select parents who report receiving an income.

Table 1: List of variables in dataset income_bmi.dta and the corresponding questions

Variable name	**Question asked to the respondent**
income	How much income in total did you receive in the past 12 months before taxes and deductions from working at any jobs you have or may have had or from self-employment? Please give us your best estimate that includes all wages, salaries, income from self-employment, professional practice and trade, bonuses, commissions and overtime pay. Income is in USD.
height	How tall are you? Answer: in cm.
weight	What is your current weight? Answer: in kg.
black	What is your race (Black or African-American)? Answers: 0 – not black, 1 – black.
drinks	In the past 7 days, how many regular (non-diet) sweetened drinks did you have? Include regular soda, juice drinks, sweetened tea or coffee, energy drinks, flavored water, or other sweetened drinks.

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 [0.6 points]

First generate the variable BMI, where BMI equals weight in kg divided by height in meters squared (BMI = weight/(height)^2). Construct a categorical variable for BMI that considers the commonly used categories: i) underweight, BMI below 18.5; ii) normal weight, BMI larger or equal to 18.5 and lower than 25; iii) overweight, BMI larger or equal to 25 and lower than 30; iv) obese, BMI of 30 or higher.

a) Compute and report the **prevalence of overweight and obesity** by ethnic group (black vs non-black). What differences do you observe?
b) Make an appropriate graph to **compare income distributions** across ethnic groups and discuss what you see.
c) If there are **missing values** for BMI, discuss how they may impact the validity of your regression analysis.
d) If there are **unreliable or implausible** values in any of the variables of interest, describe how you would identify them and drop them before proceeding with the rest of the analysis.

Question 2 [0.9 points]

Estimate a multivariate regression model explaining income as a function of BMI and whether the individual is black. Thus, estimate:

$$\text{income} = \beta_0 + \beta_1 \text{BMI} + \beta_2 \text{black} + u \qquad (1)$$

a) Interpret the estimated coefficients of all the explanatory variables (**sign, magnitude, and significance**).
b) What is the estimated 95% Confidence Interval of β_1? What can you conclude based on the information in this confidence interval about the **effect of BMI on income**?
c) What is the **relative magnitude** of the effect of being black on income? Interpret the relative magnitude.

Question 3 [0.9 points]

In the previous question, we have assumed that the association between income and BMI is linear.

a) Do you think this assumption is likely to hold? Explain.
b) Add **BMI^2** to the regression of question 2 and estimate it. What is the estimated effect of BMI on income? In your answer, interpret the effect at two different points of the BMI distribution.
c) How does adding BMI^2 capture non-linearities in the relationship between BMI and income?
d) What is your **preferred specification** (2 or 3)? Explain.

Question 4 [0.7 points]

Use the **log of income** as the dependent variable. Start by creating this variable.

a) Provide one reason why a logarithmic transformation of income may be useful in (linear) regression analysis.
b) Estimate a regression model using OLS explaining the **log of income** as a function of **BMI as categorical variable** and whether the individual is black.
c) Interpret the estimated coefficients of all the explanatory variables (**sign, magnitude, and significance**).

d) Extend the model to estimate if the relationship between **BMI (in categories)** and income is different **across ethnic groups**. What do you conclude?

Question 5 [0.5 points]

Are any of the models in Question 2, Question 3 and Question 4 **correctly specified**? Explain with your own words. (Note: For this question, focus on functional form).

Question 6 [0.9 points]

a) Reflect on the interpretation of the BMI coefficient. Does it capture the **causal effect** of body fat on income?
b) Draw a **DAG** illustrating how an **omitted variable** could bias the estimated effect of BMI on income and indicate one such variable.
c) Based on your DAG, would the resulting bias in the BMI coefficient be **upward or downward**? Explain.
d) Identify a variable (not necessarily observed in your dataset) that could act as a **collider** in this context. Use a DAG to illustrate and explain why conditioning on it would lead to bias.

Question 7 [0.6 points]

Do you think that **reverse causality** influences the OLS estimate of BMI? Explain.

Estimate a multivariate regression model explaining **BMI as a function of income** and whether the individual is black.

Is the coefficient of income statistically significant? Does this provide evidence that the variable *BMI* suffers/does not suffer from reverse causality in the previous models? Discuss.

Question 8 [1 point]

We now consider using a variable that indicates how many times in the past seven days an individual has had sweetened drinks as an instrument for *BMI*.

Estimate the model of Question 2 by 2SLS, using ***drinks*** as an instrumental variable for *BMI*.

a) Write down the estimated first stage of the model.
b) **Interpret** the estimated coefficient for *drinks* in the first stage regression. Is the estimated coefficient of *BMI* and its significance obtained with the IV estimator different from the OLS estimator in question 2? Does this suggest that the variable is **endogenous** or not? Discuss [Note: For this question, do not perform any additional analysis]
c) Perform a **formal test** of the null hypothesis that *BMI* is **exogenous**. What do you conclude? Explain.

Question 9 [0.7 points]

a) Draw a **DAG** that illustrates the assumptions required for ***drinks*** to be a good instrument for *BMI*.
b) Explain these assumptions in your own words.
c) In your opinion, do these assumptions hold? Discuss without any further analysis. You may use evidence from previous questions.

Question 10 [0.3 points]

What is your preferred model? Why?

Question 11 [0.4 points]

Researchers want to estimate the effect of **peers' risky health behaviors** (alcohol consumption) on the **academic performance of adolescents** by exploiting the quasi-exogenous assignment of high-school students across classes. They want to use the **peers' fathers' drinking behavior** as an instrumental variable for peers' risky health behaviors. They estimate a linear regression model with peers' risky health behaviors as the dependent variable (measured as the average number of times peers consumed alcohol in the past month) and the peers' fathers' drinking as an explanatory variable. The coefficient of this explanatory variable is **positive and statistically significant**.

Based on this information, what is your **assessment** of this identification strategy to estimate the effect of peers' risky health behaviors on the academic performance of adolescents?

Part 2. Presentation [2 points]

For this part of the assignment, your group will prepare a presentation of one academic article that uses a 2SLS instrumental variable estimator. Each group is assigned one article from the list below (see Excel file Groups and papers in Canvas). Your presentation should be no longer than 8 minutes (Title slide+ 3/4 content slides) and include the following information:

1) Research question and data used in the paper
2) DAG illustrating the assumptions. Do the assumptions – in your opinion – hold?
3) Critical assessment of other aspects of the paper. This critical assessment is an important part of your training as not all published papers are equally good. Focus on the following discussion points.
    a. One strength of the paper (and why)
    b. Are the findings economically relevant?
    c. External validity of the results.
    d. What would you improve in the study (and why)?

Please note that you should:

- Make sure the font and format used for the slides are clearly readable.
- Be concise in your presentation. Any content presented after the allocated time will not be assessed.
- Justify your answers.
- Include the names of all group members and your group number in the presentation title page.

Note: For papers using more than one identification strategy, please focus on the analysis using Instrumental Variables.

<u>List of papers:</u>

1) Daron Acemoglu, Simon Johnson, James A. Robinson (2001) The Colonial Origins of Comparative Development: An Empirical Investigation. The American Economic Review, 91(5): 1369-1401.

2) Ana Maria Costa-Ramon, Ana Rodriguez-Gonzalez, Miquel Serra-Burriel, Carlos Campillo-Artero (2018) It's about time: Caesarean sections and neonatal health. Journal of Health Economics, 59: 46-59.

3) Devon Gorry (2016) Heterogenous effects of sports participation on education and labor market outcomes. Education Economics, 24(6): 622-638.

4) J. Möller and M. Zierer (2018) Autobahns and jobs: A regional study using historical instrumental variables. Journal of Urban Economics, 103: 18-33.

5) Jangho Yoon and Stephanie Bernell (2013) The effect of self-employment on health, access to care, and health behavior. *Health*, **5**, 2116-2127.

6) Miguel Edward, Satyanath Shanker and Sergenti Ernest (2004) Economic shocks and civil conflict: an instrumental variables approach. Journal of Political Economy, 112(4): 725-753.

7) Thomas Daniel Robert, Harish SP, Kennedy Ryan and Urpelainen (2020) The effects of rural electrification in India: an instrumental variable approach at the household level. Journal of Development Economics, 146: 102520.

Part 3. Individual peer assessment [0.5 points]

After completing the group assignment, you reflect in a constructive way upon the contributions of each of your fellow working group members. This is expected to improve your groupwork skills and to lead to even more productive cooperations in the future. Be fair and honest to your peers, and provide written constructive feedback so you can all learn and improve your skills to work in teams.

You provide your peer assessment using this rubric:

	Capstone 0.5	Milestone high 0.45	Milestone low 0.4	Benchmark 0.3	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	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.	

	advances the project. Proactively helps other team members complete their assigned tasks to a similar level of excellence.				
Fosters constructive team climate	Supports a constructive team climate by doing all of the following: - 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	Supports a constructive team climate by doing any three of the following: - 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	Supports a constructive team climate by doing any two of the following: - 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	Supports a constructive team climate by doing any one of the following: - 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.	

	team’s ability to accomplish it. - Provides assistance and/or encouragement to team members.	encouragement to team members.	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.	