

THE UNIVERSITY OF SYDNEY - SCHOOL OF ECONOMICS

ECMT2150 INTERMEDIATE ECONOMETRICS, S1 2022 ASSIGNMENT

Due Date: 22 May 2022 (11:59pm sharp)

Instructions:

- **Anonymous marking:** Do **NOT** put your name anywhere on your assignment or in the file name. Identify yourself only by your student number.
- Answer all questions.
- A total of 100 points are available and marks for each question are indicated throughout.
- The assignment is worth 15% of your final grade for this UoS.
- You will need to use STATA (or another regression software program, e.g. R) to complete this assignment. Do not use Excel.

Submission Instructions:

- **Answers to Parts A-D** are to be submitted via the **Canvas Quiz, "Assignment Quiz..."**
 - I encourage you to work through all of the data analysis following the questions in this document on Stata or another software package before heading to the quiz to answer the questions there. There are no trick questions, so if you have completed each of the following questions, kept a copy of your output and made a note of your answers, there will be no surprises when you are taking the quiz. You should not need to use Stata during the quiz at all. That said, the quiz is not timed, so you could leave and come back to the quiz if you need to.
 - Remember – because it is an untimed quiz, it will not automatically submit at the due date. You must click submit yourself.
 - You will get only one attempt at the quiz.
- You must **upload your Stata output and commands/do file** in the final question of the Canvas quiz.
 - This upload is worth 5 points.
 - This document should be no more than 5 pages long. It should show your commands and your output.
 - Think of this as a way of showing your working.
 - NB: In Stata, if you highlight some of the output in the Results window, then right-click, you can
 - a) copy and then you can paste this into *Word* or some other word processing software. (In *Word*, Font Courier New in size 9 works well), or
 - b) copy as a table or picture. This will capture your commands and output. Then you can paste this table or image into *Word* or some other word processing software.
- You must submit your answer to **Part E through the Assignment dropdown**. Part E must be typed. It will be checked using Turnitin for plagiarism.

Assignment: Multiple Linear Regression Inference, Heteroskedasticity, Endogeneity and IVs

The topic and information on the dataset

This assignment involves the application of a range of econometric methods in analysing the effect of attending a private school on academic achievement. This topic has been the focus of a large research literature by economists, especially in the US, going back decades, and became known as the “school choice” debate. To quote Rouse (1998):

At the cornerstone of many school reform proposals lies the premise that private schools are more efficient than public schools... Proponents of school choice argue that governments should offer tuition vouchers to families who wish to send their children to private, rather than public, schools. If private schools are indeed more effective than public schools, a voucher program may offer a cost-effective way to improve the quality of education.

...

Critics of school choice programs have argued that private schools would not necessarily do a better job educating students who are currently attending public schools. Rather, they argue that the observed superiority of private school students arises from the selection process that leads higher-achieving students to attend private schools.

This quote comes from a published study by Rouse, [Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program](#) (Quarterly Journal of Economics, May 1998, pp.553-602). It is available through the library. Reading the article is not required for this assignment (but if you are interested, I encourage you to take a look).

The data used in our assignment is a subset of the data used in the paper. The data set is named 'Assignment.dta'¹.

Download the data from the [Assignment tab](#) in our Canvas site where you found these instructions.

Note: There are a few different versions of the data and each student will have a link to just one of these. I have edited the data slightly for each version, and by enough that you need to work on your own data. If you work on one of your classmate's data sets, you may answer one or more questions in the quiz incorrectly and lose marks or be referred to the academic integrity office.

¹ Assignment.csv is also available for those using software other than Stata.

The data, background and more info on key variables

The data is a cross-section of data from 1994 on 300 students. There are 300 rows – one for each student - and 9 columns. The columns correspond to the variables:

Variable name	Description
studentid	student identifier
black	= 1 if Black or African America, 0 otherwise
hispanic	= 1 if Hispanic, 0 otherwise
female	= 1 if female, 0 otherwise
mnce	Math NCE score, 1994 – more details are below.
selectyrs	number of years selected to attend a choice school
choicelyrs	number of years attended a choice school
mnce90	the student's Math NCE score in 1990
appyear	year of first application; 1990-1993

What we have called 'choice' schools here are private not-for-profit school. The sample is one of low income students from Milwaukee, Wisconsin. In 1990, Wisconsin ran a school choice lottery in Milwaukee. Available funding was limited – so the lottery was used to randomly select students who would receive a voucher (like a grant) that they could use to pay the fees to attend a choice (i.e. private) school. To be eligible to apply, the student's family had to have a very low income at or below 1.75 times the national poverty line.

Our dataset includes students who applied to the voucher lottery and were accepted, applied to the lottery and were not accepted, and students who did not apply to the lottery.

Our variable *choicelyrs* is the number of years between 1991-1994 that a student attended a choice school.

Our variable *selectyrs* indicates the number of years a student was selected, via the lottery, for a voucher. If the student applied to the lottery in 1990 and received a voucher, then *selectyrs*=4, if she applied in 1991 and received a voucher then *selectyrs*=3, and so on.

The outcome variable of interest is *mnce*. This is a student's percentile score on a maths test administered in 1994. The score takes values from 0-100, and is measured in percentile points.

We also have a *mnce* score from a test the students took 4 years earlier, in 1990, before they had the opportunity to enter a choice school.

Part A: Descriptive Statistics for the Sample [9 marks]

Quiz questions 1-4: [5 marks]

Investigate the distribution of the variables:

mnce, *black*, *hispanic*, *female*, *selectyrs*, *choicelyrs*

For each, find the average, standard deviation, minimum, maximum and median of its sample distribution.

Construct and keep a copy of a histogram for *mnce*.

In the quiz you will be asked to report selected summary statistics either rounded to 2 decimal places or to the nearest whole number. You will also answer a multiple-choice question on the histograms.

Quiz questions 5: [2 marks]

Find out how many students in the sample of 300:

- never received a voucher
- had a voucher for four years
- never attended a choice school
- attended a choice school for four years.

You will report each of these in the quiz.

Quiz question 6: [2 marks]

Pause and think about what you learn from these descriptive statistics. In the quiz you will be asked to briefly describe one useful, unusual or noteworthy thing you discovered from these descriptive statistics.

Part B: Simple & Multiple Regression Model - Estimation and Testing [27 marks]

Quiz question 7: [3 marks]

(1) Estimate the simple regression model in (EQ.1):

$$mnce = \beta_0 + \beta_1 choicelyrs + u \quad (EQ.1)$$

In the quiz you will report selected coefficient estimates, standard errors and the R-squared, rounded to **4 decimal places**.

Quiz questions 8-10: [5 marks]

If private school (i.e. what we are calling choice schools here) do indeed do a better job at educating students (at least as measured by test scores), what sign do we expect for the coefficient on *choicelyrs*? What is the sign of this relationship from your estimates of (EQ.1)? Based on the estimates, is attending a choice school associated with better academic achievement? Interpret the estimated slope coefficient.

Quiz questions 11-13: [3 marks] Is the estimated slope coefficient in (EQ.1) significantly different from zero at the 10% level of significance?

In the quiz, you will not need to set out all the steps of the hypothesis test, but you will need to write down the null and alternative hypotheses for the test, report the p-value, and report whether it is or is not statistically significant.

In your quiz answers, writing H_0 and H_1 , β_1 , $\beta_1\hat{}$, etc is fine – you are not required to use subscript formatting or typeset maths in your quiz answers. But distinguishing between and using $\beta_1\hat{}$ or β_1 is important. To write not equal to 0, you can write it out in words, or write neq or $not=$.

Quiz question 14 [3 marks]:

Do you think the estimated slope coefficient in (EQ.1) is a causal estimate? Briefly explain.

Quiz question 15 [2 marks]:

Now add the available demographic characteristics to the model as control variables. The model is now:

$$mnce = \beta_0 + \beta_1\text{choicerys} + \beta_2\text{female} + \beta_3\text{black} + \beta_4\text{hispanic} + u \quad (\text{EQ.2})$$

In the quiz you will report selected coefficient estimates rounded to **4 decimal places**.

Quiz questions 16-18: [3 marks]

- Find the 90% confidence interval for the coefficient β_1 on *choicerys* in EQ.2.
 - In the quiz, you will report the lower bound of the confidence interval, rounded to **4 decimal places**.
 - You can calculate this yourself – if so, be sure to make any calculations using all of the decimal places given in your Stata regression output.
 - Or, you can use a Stata command – check out the options on the command `regress`. To see all the options for the `regress` command, type `help regress`, in the Stata command window.
- Using your confidence interval, is *choicerys* statistically significant in EQ.2 at the 10% significance level? (Yes/No)
- State how you used the confidence interval you calculated in order to determine whether *choicerys* statistically significant at the 10% significance level?

Quiz question 19: [1 mark]

Based on your estimated results for (EQ.2), is attending a choice school associated with better academic achievement? (Yes/No)

Quiz question 20: [2 marks]

Now add the variable *mnce90* to the model:

$$mnce = \beta_0 + \beta_1\text{choicerys} + \beta_2\text{female} + \beta_3\text{black} + \beta_4\text{hispanic} + \beta_5\text{mnce90} + u \quad (\text{EQ.3})$$

In the quiz you will report selected coefficient estimates to **4 decimal places**.

Quiz question 21: [1 mark]

How does adding the students' earlier test score *mnce90* to the model affect the estimated coefficient on *choyceys*? (MCQ)

Quiz question 22: [1 mark]

Interpret the coefficient on the test score *mnce90*. (MCQ)

Quiz question 23: [2 marks]

We might think of *mnce90* as a proxy variable. Suggest an omitted variable that we could be proxying for with *mnce90*.

Quiz question 24: [1 mark]

What are the assumptions we require a proxy variable such as *mnce90* to satisfy? (MCQ)

Part C: Heteroskedasticity [10 marks]

Quiz questions 25-29: [7 marks]

Apply the *modified White test* for the presence of heteroskedasticity to model (EQ.3), using a 10% significance level. What do you conclude?

- Please use an F-test for your test.
- NB. For full marks, you must conduct all the steps of the test as per the lecture notes or as described in the textbook.

In the quiz you will

- report selected coefficient estimates and the R-squared from your auxiliary regression each to **4 decimal places**,
- report the test statistic, the degrees of freedom and either the critical value or the p-value for the test, and
- provide the conclusion from your test.

Quiz questions 30-31: [3 marks]

Re-estimate the model (EQ.3) with robust standard errors. In the quiz you will report selected standard errors to **4 decimal places**.

You will also answer a MCQ about the differences between the robust standard errors and the regular standard errors you found above in Part B for (EQ.3).

Note: Whatever your findings in Part C, please **do NOT use robust standard errors** for the rest of the analysis.

Part D: Endogeneity and Instrumental Variables [39 marks]

Quiz question 32: [4 marks]

Arguably, *choyceys* is still endogenous in (EQ.3) despite the addition of demographics and the proxy variable *mnce90*. If so, does the multiple regression model in (EQ.3) capture a causal

relationship between attending a choice school and academic achievement? Why or why not? What does this imply about $E(u | \text{choicexprs})$?

Quiz question 33: [3 marks]

What is a plausible reason for, or source of, this endogeneity? Carefully explain.

Quiz question 34: [3 marks]

If the variable *choicexprs* is endogenous in (EQ.3), state the impact on your estimates **and** inference if you estimate model (EQ.3) using OLS.

Quiz question 35: [4 marks]

The variable *selectyrs* provides a potential instrumental variable we could use to cleanly identify the causal effect of attendance at a choice school on student performance. What two key conditions must each instrumental variable satisfy in order for the IV estimator to be consistent? State whether each these conditions can be tested.

Quiz question 36: [4 marks]

Discuss whether, and why or why not, we could expect the IV, *selectyrs*, to satisfy these two conditions that you gave in Question 35. To do this, use intuition or simple economic theory.

Quiz questions 37: [3 marks]

Estimate the first stage (also known as the reduced form) regression if we are going to use *selectyrs* as an IV for *choicexprs* in (EQ.3).

In the quiz you will report selected coefficient estimates to **4 decimal places**.

Quiz questions 38-39: [3 marks]

Using the estimation results from **Question 37**, test the relevance of the IV, *selectyrs* (also known as a test of identification). Use a 1% level of significance.

In the quiz, you will

- report the test statistic and the critical value, **both to 2 decimal places** and
- select the correct formal conclusion for your test (MCQ).

Quiz questions 40: [3 marks]

Re-estimate model (EQ.3) by 2SLS using *selectyrs* as an IV for *choicexprs*.

In the quiz you will report selected coefficient estimates and standard errors to **4 decimal places**.

Quiz question 41: [3 marks]

Interpret the 2SLS-IV estimate for β_1 , the coefficient on *choicexprs*.

Quiz question 42: [3 marks]

Comment on the differences between the 2SLS-IV and OLS estimates and their standard errors for β_1 , the coefficient on *choiceyrs*, in (EQ.3). For reference – these are the estimates from **Question 40** and **Question 20**, respectively.

Quiz question 43: [1 mark]

What can we now conclude? That is, from the 2SLS-IV estimates for (EQ.3) - does it appear that attending a private, i.e. choice, school is associated with better student achievement? (yes/no)

Quiz question 44: [5 marks]

Upload your Stata output/do file in the final question of the Canvas quiz.

- This upload is worth 5 points.
- This document should be no more than 5 pages long. It should show your commands and your output.
- Think of this as a way of showing your working.
- It should be a doc, docx or pdf. No other file types will be accepted.

Part E: Conclusions [15 marks]

Provide a short summary or conclusion for your findings on the research question – the effect of attending a choice (i.e. private) school on academic achievement. Be sure to comment on your conclusions regarding the causal effect of effect of attending a choice school on student academic performance. Explain the reasons for your conclusions.

NB:

- Your answer for Part E should be 1, maybe 2, paragraphs long.
- Answers must be 250 words or less. Include your word count in your document. Answers that exceed the word count may be penalized.
- The best answers are short, to the point, and focused on the key take-aways from our analysis. Imagine you just have 1-2 minutes to tell someone what the research is about and what you found. Do not describe everything you did. Many steps in our analysis are necessary parts of a research project, but we do not need to list all these steps and things we have checked when we are reporting our key results.
- You must type up your answer to this question in your own words and submit it through the assignment dropbox.
- It will be checked using Turnitin for plagiarism.
- It should be a doc, docx or pdf. No other file types will be accepted.