Summer 2023
EC204 Empirical Economics 2

Empirical Project Guideline

This is a guideline for your empirical project. The goal is to propose your own research question and apply the techniques learned during the lecture. You may work individually or in a group of two students. Please note that the main objective of this project is for you to learn, not just for me to grade. Therefore, do not hesitate to reach out to the instructor or TF if you have any questions. The empirical project will receive a good grade if the following requirements on formatting and analysis are appropriately addressed.

Only one person in each team needs to make a submission. Late submissions will not be accepted. The project will involve the following two steps:

1. Choose the team, research topic, and potential data sources: Submit by 11:59 pm on Sunday, August 6th. Use the following link for submission.
https://docs.google.com/spreadsheets/d/1WGjwwW4iEJiFF25Hi0mKUmmNUOuu515ZpVoifD8CRt8/edit?usp=sharing
   - If you are using the textbook dataset, I want to avoid two groups working on the same dataset and research topic. Therefore, I created the spreadsheet so you can see which topics your classmates have chosen. The research topic is on a first-come-first-served basis. You can use the same dataset, but your research question should be different. If your classmate has already chosen the same dataset and topic, please consider selecting a different one.

2. Submission of the paper: Submit by 11:59 pm on Sunday, August 13th. Make sure to submit **two** items using Blackboard:
   • The final draft of the paper as a **pdf** file
   • The data set in Stata format

Format

• Final draft
- The length of the paper should be up to 14 pages including all tables and graphs, plus (not counting toward your 14 pages) a one-page "Reference Cited" section and your do-file in the "Appendix" section. It should be double-spaced, with a 12-point font, and 1" margins. Make sure to number the pages.
- Please note that reaching the maximum page limit is not a requirement for grading. The important thing is to complete all the required tasks in your paper.
- You must include the do-file used for the analysis in the appendix. Simply copy and paste your do-file.

• Data set
- Organize your dataset as a Stata file with the extension ".dta." Ensure that the variables are clearly labeled, and only include the variables relevant to your analysis. The dataset will be part of the grade. I will evaluate your skills in data cleaning, which you learned during Stata sessions, such as renaming variables, keeping/dropping variables, and labelling variables. Please indicate the source of each variable in the data description section of the paper.

Finding the data set

While you should find your own data set for this course, I want to offer the option of using the data sets provided by our textbook, considering the intensity of the summer course. However, since finding your own data set requires a significant amount of effort, I will grant a 5-point bonus to the total score of your research project in such cases. This means that you can achieve a total score of 100 using the textbook data set, but by using your own data set, you could compensate for any potential deductions in your scores.

1. Using the textbook data set
   I uploaded the data files and description on Blackboard folder "Research project." After reading the description, you can choose the data set to use and your research topic.

2. Using your own data sets
   Good sources of data are

- IPUMS (https://ipums.org/);
- World Bank Indicators (https://data.worldbank.org/indicator?tab=all),
- American Economic Association (https://www.aeaweb.org/resources/data),
- OECD (https://data.oecd.org/).

Requirement for your data set

- Make sure your sample size is at least 100 observations so you can use the asymptotic sampling distribution for hypothesis testing.
- Make sure your data set is a cross-sectional data. For example, if your original data set is a panel data, you should focus on one period and make it into a cross-sectional data set.

Outline for the paper

1. Title page
2. Introduction
3. Literature review
4. Data description
5. Econometric analysis
6. Summary and potential extensions
7. References
8. Appendix

1. Title page

The title page counts as page 1 within your total page limit. The title page must include your name(s), course number/title, semester, the title of your paper, and a short abstract of your work (maximum 100 words). An abstract is a brief introduction of the purpose and main results of your research. I encourage you to write the abstract as the final component of the project, after you have completed the paper.

2. Introduction

This section consists of 2-3 paragraphs and provides background information on the economic issue under examination. Here, you will describe your economic question and the econometric hypothesis to be tested in your paper. Identify your dependent and independent variables, and

specify the expected relationship you aim to test (positive or negative). Additionally, provide rationale for your hypotheses by explaining why you expect the variables to be related as stated. Finally, summarize the results of your analysis and outline the structure of the remaining sections of your paper.

3. Literature review

This is a (VERY) brief summary of previous research findings relevant to your topic, consisting of no more than 2-3 paragraphs. It is not necessary to cover every single aspect of your topic in the literature review. However, your literature review section should include a minimum of two references to academic journal articles.

To find academic articles, you can utilize resources such as Google Scholar, the Journal of Economic Literature (JEL), and the Econlit database. These sources can provide access to various types of publications, including working papers, theses or dissertations, and published economic papers. Additionally, you may find examples of working papers and academic articles online through the NBER (http://www.nber.org/papers) and RePEc (repec.org) websites.

4. Data description

Describe the sources and structure of your data. Indicate who or what your observations represent (e.g., country, workers, states), including the total number of observations. Please note that your data set must consist of at least 100 observations. Present this information in paragraph form.

Next, create a table to provide descriptions of each variable, including a brief explanation and the units of measurement. This table will be referred to as "**Table 1**: Variable Descriptions." Additionally, prepare a well-organized summary statistics table that includes the number of observations, mean, standard deviation, minimum, and maximum. This table will be named "**Table 2**: Summary Statistics." Discuss the type (continuous, binary, categorical) of your key variables, Y and X_1, and comment briefly about the interesting features of their summary statistics.

5. Econometric analysis

The instruction to follow will depend on whether your main variable of interest X_1 is a quantitative or qualitative (Ch 7) variable.

<If X_1 is a quantitative variable>

5.1. Model selection: Exploring the functional form of main relationship

First, include a scatterplot graph of your dependent variable Y and your main variable of interest X_1, including a line of best fit laying over the scatterplot. This will be your **Figure 1**.

If the scatter plot shows nonlinear relationship, you can try including the quadratic term or log-transformations of variables. Draw a scatter plot with variable transformation as well as the line of best fit. This will be your **Figure 2**.

The model that explains the relationship between Y and X_1 the best will be your main specification. It is important to include your rationale for choosing your final model (economic interpretation, fit of the regression model, …).

5.2. Empirical analysis

Now, we will include controls variables to your main specification found in Section 5.1 to mitigate the omitted variable bias. First, present your main regression model as an equation. For example, if your chosen specification in Section 5.1 is

$$wage_i = \beta_0 + \beta_1 exper_i + \beta_2 exper_i^2 + u_i,$$

your main regression model will be something like

$$wage_i = \beta_0 + \beta_1 exper_i + \beta_2 exper_i^2 + \beta_3 age_i + \beta_4 educ_i + u_i$$

when I include the control variables *age* and *educ*. Please do not indicate the numerical values of your parameter estimates, just the population regression model here.

- Explain why you included/excluded the particular explanatory variables.
- Provide an economic justification for the expected signs of the coefficients (both X_1 and control variables).
- Is there any remaining concern for omitted variable bias?

Then, present your regression results in a table format. Please use the heteroskedasticity-robust standard error unless you have a strong reason to validate homoskedasticity assumption. This will be your **Table 3**.

The table will contain <u>four</u> model specifications, and therefore the table will have five columns. The model specifications are as follows.

1. Simple regression on y on x_1
2. Regression exploring nonlinear relationship in Section 5.1.
3. Adding control variables to model 2
4. Adding interaction term involving x_1 to model 3

Now start the discussion of your regression results.

- First, interpret the effect of x_1 focusing on the third model.
    - Discuss the statistical and economic significance.
        - If insignificant, comment on whether multicollinearity could be a culprit. (You can try using "variance inflation factor (VIF)" to detect multicollinearity.)
    - How does the estimate change in different regression models?
    - Are the signs of coefficients same as your expectation? If not, suggest an explanation of why this is the case (ex. small data set, data set has some peculiar features related to the time of collection, omitted variable bias).
    - If you used the regression model with a quadratic term as your main specification in Section 5.1., calculate the turnaround point.
    - Discuss the effect with the interaction term. Why did you include that interaction term (Why did you expect the interested effect would depend on that another variable)? Is the interaction term statistically significant?
- Now, turn your attention to other regressors.
    - Are your expectations about the signs of other coefficients met? Comment (briefly) on their sign, magnitude, and statistical significance.

Finally, include the result of <u>one F-test</u> you conducted. The F-test could be about the hypothesis of parameters in your main regression model. For example, if you have race dummy variables as

your control variables, you might want to test whether the effects of races on the dependent variable are equal. Clearly state the hypothesis you are testing, and the result given using Stata.

<If X_1 is a qualitative variable>

5.1. Summary statistic by category

First, choose between Y and log (Y) as your dependent variable. Discuss the rationale for this decision such as economic interpretation or mitigating the outlier problem.

Next, explore the frequency of X_1. Comment on frequencies of categories (You can include a frequency table if there are many categories). Include a bar graph of mean of dependent variable according to category of X_1. This will be your **Figure 1**. Comment on the relationship discovered by the bar graph.

5.2. Empirical analysis

Now, we will include controls variables to your main specification found in Section 5.1 to mitigate the omitted variable bias. First, present your main regression model as an equation. For example, if your chosen specification in Section 5.1 is

$$\log(wage_i) = \beta_0 + \beta_1 female_i + u_i,$$

your main regression model will be something like

$$\log(wage_i) = \beta_0 + \beta_1 female_i + \beta_2 married_i + \beta_3 educ_i + u_i.$$

Please do not indicate the numerical values of your parameter estimates, just the population regression model here.

- Explain why you included/excluded the particular explanatory variables.
- Provide an economic justification for the expected signs of the slope coefficients.
- Is there any remaining concern for omitted variable bias?

Then, present your regression results in a table format. Please use the heteroskedasticity-robust standard error unless you have a strong reason to validate homoskedasticity assumption. This will be your **Table 3**.

The table will contain four model specifications, and therefore the table will have five columns. The model specifications are as follows.

1. Simple regression on y (or $\log(y)$) on x_1(or dummy variables of x_1 categories)
2. Adding control variables to model 1
3. Adding interaction term of x_1 with another dummy variable to model 2
4. Adding interaction term of x_1 with another continuous variable to model 2

Now start the discussion of your regression results.

- First, focus on the effect of x_1.
    - Discuss the statistical and economic significance focusing on model 1 and 2.
        - If insignificant, comment on whether multicollinearity could be a culprit. (You can try using "variance inflation factor (VIF)" to detect multicollinearity.)
        - How does the estimate change in different regression models?

        - Are the signs of coefficients same as your expectation? If not, suggest an explanation of why this is the case (ex. small data set, data set has some peculiar features related to the time of collection, omitted variable bias).

    - Discuss the effect with the interaction term.
        - Why did you include that interaction term (Why did you expect the interested effect would depend on that another variable)?
        - Is the interaction term statistically significant?
        - Interpret the effect of x_1 taking interaction term into account.
        - For the interaction term using continuous variable, it is helpful to plug-in a meaningful value for interpretation.
- Now, turn your attention to other regressors.
    - Are your expectations about the signs of other coefficients met? Comment (briefly) on their sign, magnitude, and statistical significance.

Finally, include the result of <u>two F-tests</u> you conducted. The F-test could be about the hypothesis of parameters in your main regression model. For example, if you have race dummy variables as your control variables, you might want to test whether the effects of races on the dependent variable are equal. Clearly state the hypothesis you are testing, and the result given using Stata.

6. Summary and potential extensions

First, summarize your findings. Remember you just detailed them in the previous section, but here you simply provide a (brief) overview of what you found. Discuss policy implications and the potential contribution of the research.

Next, talk about what could have made your research better. For example, if data limitations may be causing omitted variable bias, discuss the direction of the bias that you expect, and whether your coefficient may be over- or underestimating the effect of X on Y. Also discuss things that may be able to be handled with future research, or different samples to be studied, or anything you can think of that could improve our knowledge beyond what your research taught us.

(3-point Bonus) Discuss whether your research could be improved by using instrument variable regression or panel data we learned in Chapter 14 and 15.

7. References

List the literature you cite in your study in alphabetical order. Articles should be cited according to the following format:

Author, A.A., Author, B.B., and Author, C.C. (Year of publication). Title of article. *Title of periodical*, volume number (issue number), pages.

Example:

Adalja, A., Hanson, J., Towe, C., and Tselepidakis, E. (2015). An examination of consumer willingness to pay for local products. *Agricultural and Resource Economics Review,* 43(3), 253-274

8. Appendix

Include your do-file here (as an insert in the Word document) simply by copying and pasting your commands into Word.

9. Example of tables

[Example of Table 1]

Table 19.1 Variable Descriptions

salary	annual salary (including bonuses) in 1990 (in thousands)
sales	firm sales in 1990 (in millions)
roe	average return on equity, 1988–1990 (in percent)
pcsal	percentage change in salary, 1988–1990
pcroe	percentage change in roe, 1988–1990
indust	= 1 if an industrial company, 0 otherwise
finance	= 1 if a financial company, 0 otherwise
consprod	= 1 if a consumer products company, 0 otherwise
util	= 1 if a utility company, 0 otherwise
ceoten	number of years as CEO of the company

[Example of Table 2]

	count	mean	sd	min	max
price	179	76628.04	30626.44	26000	300000
dist	179	21195.53	8849.76	5200	40000
rooms	179	6.58	0.96	4	10
area	179	1999.64	635.01	750	5078
baths	179	2.31	0.74	1	4
nbh1	179	0.32	0.47	0	1
nbh2	179	0.08	0.28	0	1
nbh3	179	0.17	0.38	0	1
nbh4	179	0.02	0.15	0	1
nbh5	179	0.20	0.40	0	1
nbh6	179	0.11	0.31	0	1
nbh7	179	0.09	0.29	0	1

[Example of Table 3]

	price	lprice	lprice	lprice
dist	0.845***			
	(0.207)			
ldist_miles		0.317***	0.140***	0.166
		(0.0375)	(0.0445)	(0.105)
rooms			0.0274	0.0273
			(0.0220)	(0.0221)
area			0.000170***	0.000169***
			(0.0000515)	(0.0000508)
baths			0.169***	0.188**
			(0.0334)	(0.0852)
nbh2			0.0865	0.0840
			(0.0660)	(0.0688)
nbh3			-0.00126	0.0000791
			(0.0532)	(0.0524)
nbh4			-0.201**	-0.201**
			(0.0895)	(0.0906)
nbh5			-0.0922	-0.0915
			(0.0653)	(0.0657)
nbh6			-0.0269	-0.0231
			(0.0571)	(0.0562)
nbh7			0.0499	0.0519
			(0.0493)	(0.0477)
distbaths				-0.0142
				(0.0526)
_cons	58715.5***	10.77***	10.10***	10.07***
	(5699.3)	(0.0577)	(0.136)	(0.189)
N	179	179	179	179
R^2	0.060	0.184	0.621	0.621