Notes 20: Research

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

This class wraps up everything that we learned so far. It will give you a good idea of how far we've come. It provides a guide of how to go about pursuing a research project. Finally, it is the best guide to answer the essay question in the Final.

Notice: This guide is prepared for someone with the knowledge of econometrics taught in this course. It is possible to do a much better job if you learn more advanced techniques both in econometrics and in economic theory.

1 Deciding on a topic

Often the research topic is given to you. For example, your boss may want you to answer a very specific question. However, many times you get to decide what you want to research, or even if you are given strict guidelines, there is still enough freedom that you may have doubts about what to choose. This section is about how to choose a topic when you have a lot of freedom.

1. If you have no restrictions, _____

   Yes, this is obvious, but students usually disregard it. Students decide to research a Finance topic because it may help them get a job, or they choose topics others consider hard, or a topic that is popular right now. The topic of research is nothing compared to the actual research work. By now you know that you will have to read data sets descriptions, codebooks, and papers. You will also have to run many regressions, experiment with several models, attempt many changes. More than anything, you will have to think about the topic. You will have to think a great deal. If you don't care about the topic, if you are not genuinely curious, you will get bored and do a worse job of it. No matter what anyone tells you about which topic is better, nothing trumps the quality of the work. Sure, for the same quality of work some topics are better than others, but don't think you will put the same amount of effort for any topic, it just doesn't work like that. So, make an effort to pick a topic you like.

2. If you like a few choices the same, if you really have no idea about your preferences, or if you are entirely convinced of what you want to do next, then _____

_____

Taste comes first, second comes the game. However, if it is all the same, by all means choose a topic that will give you the best chances to get what you want. However, don't underestimate the difficulty of trying to do things to please someone else. It is usually a better idea to stick to the things you truly like, and let the opportunities arise to match your taste, instead of changing yourself to fit. Again, nothing is more important than the quality of the work. Do whatever with passion, put in the hours and the deep thinking to make it good, and the opportunities will arise.

3. _____

After you choose a topic, what would you like to know? Is there something you are curious about? Do you have an opinion about something you would like to check? You are beginning to think about the research question. However, don't get too obsessed. Most of the time you will have to revise your question later. For example, you may not get good data, or somebody else may have done exactly what you intended. Choose a question very freely. It does not need to be brilliant, it just needs to be something to guide you in the beginning steps. Students think that the question is very important, however the most important thing is to begin working. The sooner you begin working, the sooner you start to learn about the topic. When you learn more, you will be able to have much better ideas. Even if you change the question, you won't have wasted time, you will have learned about the topic.

4. _____

Now you have to write your question in a causal format. What causes what? Put it in the specific units: what is the effect of one extra cigarette per day during pregnancy on the baby's birth weight? You can also state it in categories: what is the causal difference of a college degree in comparison to having a high school degree? Again, it does not need to be the best question, it just needs to be precise. You will see the advantages of precision in the next section.

5. _____

What I mean here is what directly causes the dependent variable. For example, what causes birth weight? Genetics, mother's health, nourishment, substances consumed, and substances to which the mother is exposed. Income, education, marital status,

etc. are not causal variables. Money alone cannot make your baby fatter. In an ideal world you would like to control for these things. It doesn't matter that many of those are unobservable, think about them anyway. It helps to think of the ideal experiment. What is the essence of *ceteris paribus* here? You can't control for things that are caused by smoking. Otherwise, how could you vary smoking and keep them constant? Think deeply about this. Even if you don't love your current question, this kind of thinking will teach you about the topic, and perhaps give you new ideas of questions. If you have a new idea, and like it better, change now and repeat the previous step as well as this one.

6. _____

Write down the following information: (1) the causal variable, including the units, (2) the dependent variable, including the units, (3) the causal variables which you think are likely to exist in a data set, and (4) proxies for the variables which you don't expect to find in a data set. When you write down the proxies, write down the causal variable, and then list the proxies you think you will find. As you write them down, notice which variables you think it will be harder to proxy for.

2 Searching for a data set

1. _____

With the list in hand, use the data bases we taught you, as well as all your goggle searching skills to look for a data set. Try many search strategies. You cannot be lazy about this step. You must find the absolute best data set you can access for your case. Read descriptions, compare very carefully. Remember that the sample size is very important, consider it as an important component in your search. Sometimes there is no absolute best, because one data set may have advantages and disadvantages in comparison to another. If that is the case, take notes so that you don't forget what you found. Be exhaustive in your search. The point is to come up with the very best choice, but it is also to learn. What kind of information can be usually found in data sets on the topic? Is there a data set that asks unusual questions? You need to know these things for many reasons: (1) you need to justify the choice of your data set in your paper, (2) it will help you if you want to change your question, or in your next project, (3) it is the type of knowledge that makes you a specialist, it is part of knowing the field, and becoming an authority in it.

2. _____

Before you do anything else, take ten minutes to think about what you expect to find. Which problems do you anticipate in the data? Download the chosen data set and begin exploring it. Check the important variables. Are there many missing values? Do they seem to be missing-at-random? Is any variable censored? Are the important variables in the right units? If not, can they be transformed? For example, your data set could have only information on whether the woman smoked or not, but no information on the quantity. Is the data set organized, or will it be a nightmare to use?

3. _____

Same as in the last item, think about what you expect to find in the codebook. How do you think the survey was collected? Which problems do you expect to find? Think especially about the following issues: sample selection, measurement error in the main variables and controls, and near multicollinearity (rare categories, especially for the main variables and controls). You must do this step very carefully. Think not only of the problems, but also of how they could have solved them. Now go to the codebook and check what they really did. Read the description of how the data was collected for clues, read the most important questions, how they were phrased, when they appeared in the questionnaire. If you found problems in the previous item, try to look for clues about what is going on. Was the question badly phrased? Write down what you think, and what you find. Keep track, because even if you cannot solve these issues, you will have to discuss them all in your final paper.

4. _____

Now that you understand this data set, go back to your search. Get the runner ups. Read their description with more attention. Are they better than your first pick? Do they solve some of the problems? You may need to download them and try either or both of the last two steps on them.

Notice: If your data set is longitudinal, it is very worthwhile to use its structure, as you will be able to control for a lot of unobservable things. Search for information on panel data models, fixed effects, differences-in-differences. Get help if you can. This is actually a good thing, and I think you are ready to tackle this.

5. _____

You know which is the best data set now, and you also understand the reality of the information available. Now you must refine your question. Given your searches, you

may have seen that there was information you did not anticipate. You may decide to change the question completely. If that is the case, then you must go back to step 4 in the previous section (don't worry, everything will go much faster now). If you still like your question, you may need to modify it. Do you need to restrict the population? Do you need to change the causal variable (for example effect of smoking versus not smoking as opposed to the effect of an extra cigarette)? Do you need to study a slightly different dependent variable? If you have to make any modification, check steps 4 to 6 of the previous section again to make sure you don't have to expand your data set search.

3 Regressions and tests

1. _____

   Write down the main model you would like to run. Also, write down some of the alternative models you would like to explore. This is the time to think about linearity. You can also think of several specifications, some with more controls, and others with less.

2. _____

   Make a table of summary statistics. Is there something strange there? Do the means make sense? Make sure that the data matches the values of the population you want to study. Are the races proportionally represented? Perhaps you missed that the data set is stratified. In this case, make sure to get the weights, and run the regressions using them. If the proportions are off and there is no stratified sampling, it can be a sign of selection. Think carefully, and try to figure out what is happening. If there is selection, can you solve it? Should you refine your question yet again? Should you give up on this data set and use a different one?

3. _____

   Run the regressions you planned. Make sure that you assume that there is heteroskedasticity, no point doing anything else.

4. _____

   Look at the results you are finding. Anything interesting showing up? Are you surprised by any of the coefficients? If yes, then what do you think is causing this?

Can you think of something you can do to check if your thoughts are right? Perhaps at this point you may decide to run some additional regressions to investigate further.

5. _____

This step can be done at the same time as the last one. Your table should have all your models from the simplest (usually without any controls) to the most complex (usually the one with the most controls) on each column. Write down the estimated coefficients and the standard errors underneath. Think about what you found. What changes because of the controls? What does that teach you? Also, are your standard errors too large? Perhaps you may have a problem of near multicollinearity, or perhaps you are using too many controls for the amount of data you have. Can you combine some controls into a few categories? Run the new regression, see if something changes, and incorporate it into the table in the right order of complexity.

6. _____

Calculate which coefficients are significant, and run the tests that are relevant to your particular question. Are you surprised by the results? What are the confidence intervals?

7. _____

From what you learned in the previous steps, you may have more ideas of different specifications you can try. Think about the model, should you have introduced some non-linearity? Make plots, for example scatter plots and graphs of averages

4 Interpretations

1. _____

Understand the meaning of your results. If your model is correct, what do they imply for policy? What would you recommend that the government do? What should companies do? What should people do? Write the results in terms that can be understood.

2. _____

Carefully study how the results change with the different specifications. It is a good idea to imagine what should happen first, then look at the results. Do the results

change in the direction you anticipated? Do they make economic sense? Can you give a convincing argument for why this is happening?

3. _____

   What are you not accounting for? Which endogeneity do you still fear? Which variables do you suppose aren't properly proxied? Is there selection of any form? What about measurement error? Consider all these issues, and study the bias you expect them to generate. What would this imply for the interpretation of your results? Would this completely invalidate your point? As you learn more about this, make notes of possible ideas you have for solutions.

4. _____

   Give the mater of endogeneity some further thought. Search for solutions. What does the literature say on the matter? Speak with other people, somebody may give you an idea of what to look for. Every attempt you can do to solve the problems of your model is very worthwhile. Consider learning about instrumental variables. This is another topic that is just a bit more advanced than this course, but you are ready to learn this.

5. _____

   Now that you have your final results, can you explain them? Are they as the theory predicts? Are they similar to the literature, or not? Can you rule out some models? Do the results open more questions? Can you answer them? If not, can you at least pose new hypotheses for future research?

5 Final Remarks

There is a sensation of power that comes from knowing how to deal with data, how to write models and estimate them. Never forget that I spent as much time in this course speaking about what could be wrong as I spent teaching the techniques. In fact, we've been discussing confounders since the second class! There is a reason for that. You can easily run regressions and give results, but that doesn't mean that the interpretations are correct. You are responsible for the claims you make, and before you enthusiastically make any claim, happy to be able to handle this new toy, consider the cost of a wrong interpretation to you, to your employer, and to society.

That said, don't be afraid to use what you learned. As long as you understand and report the possible problems with your results, you are ready to use these tools. Moreover,

as you study any question – thinking deep, working with the data, reading papers – you will begin to know the subject more than most people around you. This is the point when you can benefit from help from people that know more techniques than you do. You may decide to go even further and tackle some of the problems you identified head on.