EC228 Research Project: A few tips

Note: The assignment varies from semester to semester, and may or may not include *Improvement.*

Picking a paper: There is no shortage of outstanding candidate papers… academic papers published in an academic journal. The key is finding one that a) interests you, and b) is replicable (the most important question wrt replication is: Can you get your hands on the raw data?)

Where to look for papers:

1. Try a *Google search* by general topic area… for instance *economics of abcxyz* or *econometrics abcxyz*
2. And continuing in this vein, look at recent issues of Journals… e.g. *Journal of Applied Economics, Journal of Health Economics, Journal of Sports Economics*, etc. etc. etc.
3. *Google Scholar* provides a wealth of candidates: https://scholar.google.com/ just enter some search terms, or perhaps the citation for a specific paper of interest
4. If you've identified an interesting candidate, use the *Cited by* feature in Google Scholar to see which papers cited the paper that you identified. This allows you to move forward in time.
5. And you can use bibliographies to move backwards in time.
6. If you've identified a dataset that researchers have employed, and you have determined that you can access the data, try a Google search of the dataset to see who has done what with the data.
7. And don't forget those amazing papers that I've posted. Some of those are excellent candidates! But beware: Some are not! In fact, with some there is zero chance of replication in finite time.

Replication

I've encouraged everyone to submit a PowerPoint presentation (preferably in hardcopy form)… but any (hardcopy) format is fine. Please keep everything in your submission brief and concise. While some submissions have had 20-30 PowerPoint slides, I'd say that the average has about 10-15 slides. And as you'll see below, it's hard to have fewer than nine slides.

Remember: The goal of replication is independent replication… by which I mean independently replicating from primary sources the construction of the data used in analysis, and then replicating the published result using your own programs and analyses. For more details: See Important I and II below.

At a minimum, your submission should include:

1. Please put your Team # in your file name… so I know who did what.

2. Team number, team member names and a full citation for the paper of interest
3. A brief discussion of what question/s issue/s topic/s are/was/were addressed in the paper and what was concluded
4. A brief discussion of the methodology employed in the paper and the regression model(s) of interest
5. A brief discussion of the raw data they worked with (including geographies, years and levels of aggregation), and specific sources for the data.
6. Identify the specific sources for the data that you worked with… including citations/links to specific sources.

   Important I: In some cases, teams work with data provided by the authors, often because there is no alternative.. As discussed elsewhere, you will not receive full credit for doing so. And if you do so, you must clearly state such in your presentation. Please do not pretend that you constructed the replication dataset when you just downloaded a dataset that the authors or others may have posted. To not give credit is to plagiarize.

   If you do use a dataset provided by the authors, you must in your presentation explain why that was your only alternative, and why you were unable to reconstruct the dataset on your own.
7. ***Summary Stats***: A side-by-side comparison of the summary stats they provided in their paper, compared to your summary stats.

   If you are having difficulty finding the data, you might try Google's dataset search tool (which they rolled out in the fall of 2018): https://toolbox.google.com/datasetsearch ... or the Wayback Machine internet archive: https://archive.org/web/
8. ***Brag, Brag, Brag***: Some bragging about what matches exactly or closely, and speculating about why some stats did not match up.
9. ***Regression Results***: Another side by side comparison, this time of their and your regression results.

   Important II: In some cases, teams work with programs (e.g. do files) provided by the authors. That is not independent replication… it is plagiarism, unless you give full credit to the authors… in which case you will receive little/no credit for independent replication. You ought to be able to at least write your own do files.
10. ***… and Brag Some More:*** More bragging some more about what matches exactly or closely, and speculating about why some of the results did not match up quite as nicely as you might have hoped for

Improvement

To start, it might be useful to think about where the bodies might be buried:

- Your going-in assumption should be that the analysis in the published paper is rife with errors and skullduggery… and you will almost always be wrong! (Remember: someone with an agenda ran hundreds of regressions before they settled on the ones they are presenting.)
- Is this a favorite coefficient model? If it is, then you have a head start in the investigation.
- Low R-squared usually means there's lots of room for improvement… but not always!
- Are there any suspicious RHS coeffs? (variables that have no right to be in the model; highly statistically significant variables with the wrong sign; peculiar RHS variable constructions (e.g. x_1=sin(z^2.379); x_2 lagged 12 days; x_3 is an avg over the last 7 wks; x_4 on full moons)?
- What obvious RHS vars didn't make it into the final model? Hmmm, wonder why they didn't do that? It's worth a look-see.
- Look at the discussion: Do they discuss heteroskedasticity or multicollinearity? What did they do about these? Do they run weighted regressions… and do the weights make sense?
- And if it's a time series model, did they worry about things like serial correlation?
- Look at all reported regression results, not just the one of interest. How do things move around with the different variations? Which obvious regression results were not reported?
- Look for the footnotes saying things like "we dropped the ABC observations because of course… ." or "it's obvious that the right way to handle this is to …" … Of course not! … obvious not!... and beware of any claims of "sophisticated analysis"

Further, there are some obvious candidates for improvement:

- You know about omitted variable bias… find those omitted variables and see how they impact the model…. all the more important if you are working with a favorite coefficient model.
- Grab more data: other countries, states, years, etc.
- Grab similar RHS variables from other sources: e.g. try the World Bank's macro data instead of the IMF's; grab Freedom House's corruption index instead of transparency.org's
- Run VIF to see how much multicollinearity you have? And if it's there, do something about it (drop vars to assess impacts; get more data) ... or maybe not!
- If you are working with a favorite coefficient model and the multicollinearity involves your favorite coefficient, then that's a big deal! … so deal with it. But if you've got lots of multicollinearity in your model, none of which relates to your favorite coefficient, then no big deal! … and don't worry about it. Because after all, you really only care about that favorite coefficient
- If there are obvious F tests (or Chow tests) to run, do those as well and see what you conclude. Relatedly: Run the regression on different subsets of data to see how robust the model is.
- If data is over time, make sure that dollar denominated vars are in real $s
- The heteroskedasticity correction is easy (, robust) so why not do it?
- Functional forms: RHS vars: try x^2 , x^3, ln(x) etc. ; interactions… for LHS, try ln(y) (but be careful as it's a challenge to compare models with different dependent variables)
- And dummies, dummies, … dummies: try out lots of intercept and slope dummies to see what the model missed… especially useful if it's a favorite coefficient model.
- Remember that dummies just capture what the rest of the model missed; so figure out what was missed and build a better model!

EC228 Research Project: ***A few tips***

The Improvement (PowerPoint) Presentation

Your presentation should tell a story… things you observed… things you tried… what worked… what didn't work… and what you would do had you had more time and money?

Keep your eye on the ball! While it's great to hear you brag about how much better your analysis is… there was a key finding/result/conclusion in the original paper, yes?. So how does your analysis impact that finding/result/conclusion? So, for example: If you are working with a favorite coefficient model, don't just brag about how you improved adj R-sq… talk about how your improved analysis impacts the estimated favorite coefficient, and the important key finding/result/conclusion in the original paper.

So include lots of esttab's showing the progression of results… and ultimately land on your preferred model, and give it the discussion it deserves!

The number of PowerPoint slides here is mostly driven by the length of the journey… but I'd say that in the past, teams usually have 6-8 slides at a minimum… and I've seen once or twice more than 20 slides.

As always, keep everything simple, brief and concise.

Most Common Mistakes/Regrets

1. Read the paper, will ya? No seriously, read the paper! It's amazing how much detail authors provide. Sometimes, though, it can take a while to pin things down. So read the paper … and pay close attention to footnotes, tables, figures etc. etc. etc.
2. ***Keep your eye on the ball!*** (see above) There was a key finding/result/conclusion in the original paper, yes?. So how does your analysis impact that finding/result/conclusion?
3. It is almost always the case that papers/presentations could be significantly improved with just a little more work. So take that little bit of extra time at the end to turn an OK submission into a terrific work product. You did all that work… ***Don't fumble the ball at the goal line!***