RESEARCH METHODS

LECTURE #1 (Mirko Draca section)

FEBRUARY 25TH , 2015

RESEARCH IS THE HARDEST THING TO TEACH!!!

My job is to teach you how to do empirical work.

That is, how to formulate and write an empirically based dissertation or research project.

The problem is that this isn’t a deterministic problem – it’s not an equation you can solve. There’s a lot of ways to go. It takes a good background knowledge, judgement and a bit of creativity.

It’s something that requires experience – and by definition you don’t have experience yet. So it’s HARD.

WHAT WE’LL GO THROUGH TODAY

1. SETTING UP YOUR RESEARCH PROBLEM

- Finding your ‘genre’
- The two most important questions to ask yourself (and your advisor)

2. GETTING IT DONE.

- Starting to do empirical work.
- Framing and presenting empirical results.
- How to design a Table of Results that makes sense.

3. LEARNING FROM EXAMPLES

- We’ll go through some examples of last year’s projects. Why did they get high marks?

WHAT WE'LL GO THROUGH OVER THE FOLLOWING TWO WEEKS

- Basically, more examples. I'll show you what to look for and how to take apart existing papers. How does the engine work?

- If you send me a 2-3 page document with a research question, estimating equation, and table (s) of preliminary results I will go through a selection next lecture and give detailed advice.

- I'll be nice about this and won't embarrass anyone. Send it to me by the end of next Monday night. The idea is to convey detailed advice that will be useful to everyone.

1. SETTING UP YOUR RESEARCH PROBLEM

In the main courses, we basically teach you the tools but not necessarily how to use them.

For example, the properties of estimators; the structure of tests, the importance of assumptions (eg: orthogonality condition).

This is literally the toolbox. There are many, many ways to use these tools to investigate data. So how do you figure out which tools to use for the problem you're interested in?

Different academic literatures focus on particular sets of tools and ways of using them. The way I think of these is as 'genres', like in movies.

Eg: Within Drama, we have: Social Realist, Thriller, Chick Flick etc.
Within Horror, we have: Classic 1930s, Slasher, Italian Giallo, Found Footage.

GENRE IN APPLIED ECONOMICS

There are distinct genres and sub-genres in empirical work. That is, sub-literatures which tend to use a set of tools in certain ways.

eg: Wage equations, event studies, cointegration studies, production function estimation.

A good approach is to work out which genre your work belongs to and figure out the obsessions and tools that dominate the genre.

eg: Production function work is obsessed with getting the parameters on the production function right, and dealing with the endogeniety of input decisions.

Focus on working out how the genre works when choosing how to use the ‘textbook tools’ that are taught in the formal courses.

Yeah, I know you never have any useful instruction for learning how to actually apply tools and interpret results.

The best advice I can give is to READ. In terms of empirical research I recommend the ‘pop science’ economics books that have been coming out (eg: *Freakonomics* by Levitt; *Poor Economics* by Duflo and Banerjee, blah, blah by Acemoglu and Johnson).

OK, let’s get down to thinking up an idea that can be implemented.

Basically, I think there are **two practical questions** that you need to answer for yourself to get going.

That is, if you can answer these questions clearly then you have a viable project or dissertation. This will get you started.

THE TWO CRUCIAL QUESTIONS

1. What are the 3 most important papers to read for my study?

Save the grand, sweeping literature review for the *Times Literary Supplement* or the *New York Review of Books*.

Focus on the key papers that your paper will spin-off from. Try and pick a paper(s) that you can backward engineer and use as a framework for your study.

Yes, read other papers but allocate your attention efficiently. Your literary career can wait for now.

WHAT ARE THE GOOD JOURNALS?

One problem you'll have with identifying three important papers is recognizing whether the paper you are reading and basing your research on is any good.

If the work you are studying is bad it is likely to leave you confused and lead you towards doing meaningless research.

So what are the good journals to look at? Academics usually answer this question by naming the journals that they publish in themselves (it's a joke, but it's also true!). But I think the consensus would be the following journals contain the better applied papers:

Top 5

Quarterly Journal of Economics
American Economic Review
Journal of Political Economy
Econometrica
Review of Economic Studies

GENERAL JOURNALS

American Economic Journal (AEJ): Applied
American Economic Journal (AEJ): Economic Policy
American Economic Journal (AEJ): Macroeconomics
Review of Economics and Statistics
The Economic Journal
Journal of the European Economic Association

FIELD JOURNALS

Journal of Labor Economics
Journal of Human Resources
Journal of Public Economics
Journal of Health Economics
Journal of Law and Economics
Journal of Finance
Journal of Monetary Economics

INDUSTRIAL ORGANISATION

Rand Journal of Economics
International Journal of Industrial Organisation
Journal of Industrial Economics

OTHER JOURNALS
Journal of Development Economics
Journal of Economic Behavior and Organisation
American Journal of Political Science
Management Science
Journal of Economic Psychology

One issue is that because the papers in these journals are good is that they will be too complex for you to do similar analysis – it'll just be too hard to code it up in your statistics package.

Hence, there's a rationale for utilizing papers that are in less good journals but are simpler to follow.

Examples of this would be basic cointegration / VAR studies or cross-country growth regressions which seem to be popular with many of you.

These are legitimate topic areas for project or dissertation work. It's just that they fallen out of fashion with the top journals.

My recommendation: **ask your supervisor / lecturer their brief opinion on a maximum of three papers**. Have them with you so that they can have a quick look at it. It takes minimal effort for them to make a judgement and it can stop you pursuing the wrong approach.

Ok, now let's address the **second crucial question**.

2. What is my estimating equation?

Figure out exactly what regression you want to run: the variables to be included, the key variables of interest , the functional form, and the interpretation of the parameters.

The 2-3 papers you identify in Q1 should help you with this. In 90% of cases someone has run this type of regression before.

Estimating this equation is then your main target: you should organize your work to estimate this equation (s) as soon as possible.

Throughout these 3 weeks I'll show a lot of examples of these estimating equations.

GETTING IT DONE.

1. Starting to do empirical work.

There is a slow, flat learning curve when you learn empirical work.

In particular, you get confused and make mistakes – lots of ‘red ink’ on the screen in STATA.

The best tip is: to watch a professional work, eg: get a job as a Research Assistant. This isn’t easily available and you don’t have the time.

The next best thing: lots of learning materials are online. Youtube is not just funny cat videos.

I recommend the following (search for these on youtube):

William Reed for basic data handling in Stata.

Economicurtis for a few videos on functional form (eg: interpreting log-log models, semi-log).

Econometrics Academy for a massive set of mini-lectures in econometrics and ‘how-to’ demonstrations in Stata.

In the next two weeks, I’ll also be showing you some examples of do-files and spreadsheets from my own work.

2. Framing and Presenting Empirical Results.

What kind of tables to include? A good structure is:

Table 1: Descriptive Statistics

Table 2: Main results.

Table 3: Secondary Results (eg: heterogeneity of coefficients, results for an important sub-sample).

Tables 4-6 : Robustness checks, evidence to defend assumptions behind your analysis (eg: supporting the robustness of the main parameter that you're estimating).

HOW TO PUT TOGETHER A TABLE

This is actually an art. A lot of students make mistakes here

These are a few things I learnt early on from the very shrewd Steve Machin (a Warwick Phd graduate, no less).

A few things to remember when drawing up a table:

- Less is more. Focus on key results rather than printing out pages of output and regression coefficients.

- Tell a story as you move across the columns of a table. The columns should follow some type of logical sequence that usually represents robustness checks or alternative specifications

- Present the coefficient estimates with the standard errors underneath in brackets. Don't report t-statistics. We want to see the coefficients and standard errors separately so we can track how they change across specifications.

- Write a set of notes underneath the table that explain what is going on.

- You want to put together a set of tables that an experienced reader could easily follow without reading the text of the paper in detail.

- Let's look at some examples.

REVOLVING DOOR LOBBYISTS

This paper looks at the revenues brought in by a group of Washington lobbyists who were formerly political staffers in the Congress.

We develop a measure of political connections P_{it} which is based on the number of politicians the staffer-turned-lobbyist worked for who are still serving in the Congress.

Our estimating equation is:

$$R_{it} = \alpha_i + \delta P_{it} + X_{it}\beta + \varepsilon_{it}$$

where R_{it} is revenue; α_i is the lobbyist fixed effect; X_{it} is a vector of controls; and ε_{it} is the error term.

The central idea of the estimation strategy is that once we include the fixed effects then the delta parameter is estimated from the changes in connections P_{it}.

The intuition is that if a lobbyist has no change in connections over the sample then the value of P_{it} in every period. The lobbyist fixed effect α_i (ie: dummy variable for each lobbyist) will then be collinear with P_{it}. That is, there is no separate variation in P_{it} that can't be absorbed into the dummy variable.

The key variable of interest is P_{it} so we set up our table to focus on this variable and test robustness in various ways.

TABLE 2: AVERAGE EFFECTS OF REVOLVING DOOR CONNECTIONS ON LOBBYING REVENUE

	Dependent Variable: (log) revenue per lobbyist			
		Plus Party	Plus Chamber	Plus experience
	(1)	(2)	(3)	(4)
Number of Senators	0.23***	0.23***	0.21***	0.24***
	(0.07)	(0.07)	(0.07)	(0.07)
Number of Representatives	0.09*	0.07	0.08	0.10*
	(0.05)	(0.05)	(0.05)	(0.05)
Individual dummies	Yes	Yes	Yes	Yes
Time	Yes	No	No	No
Time*Party	No	Yes	No	No
Time*Party*Chamber	No	No	No	Yes
Lobbyist Experience	No	No	No	Yes
Individuals	1,113	1,113	1,113	1,113
Observations	10,418	10,418	10,418	10,418

Notes: This table presents the average effects of political connections on ex-staffers lobbying revenue. The dependent variable is the log of the revenue generated from all the clients that an individual lobbyist serves in a time (semester) period. The two main independent variables are the number of senators and representatives that an individual lobbyist worked for previous to entering the lobbying industry *and are serving in Congress in that time period*. All regressions use a sample containing ex-staffers-turned-lobbyists and include both individual lobbyist dummies and time effects (i.e., semester dummies). Column 2 allows for different time effects for lobbyists connected to politicians in different parties (i.e., Democrats versus Republicans). Columns 3 and 4 allow for different time effects for lobbyists connected to politicians in different party/chamber combinations (i.e., Democrats in the Senate, etc.). Column 4 includes lobbyist experience (i.e., number of periods that a lobbyist appears in the sample) in quadratic form. Standard errors are clustered by lobbyist

***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Some things to notice about this table.

1. While there are other variables in the model, I only report the estimates for my key variable of interest P_{it}.

2. As I move across the columns I implement robustness checks whereby I include various controls that could partly explain the observed P_{it} effect. Notice how the structure of the table allows us to clearly track what is happening to the δ coefficient on P_{it}.

3. Finally, look at the set of notes underneath the table. This tells the reader all they need to know to interpret the table, including the goal of the table, the definition of the key dependent and independent variables, and the clustering of the standard errors.

When you read papers in good journals you should now start to notice how good researchers set up their tables in this way.

This is how I recommend that you set up your tables. It makes you think about what you're doing and it makes it easier for the person marking it. They'll be able to work out what you're doing very quickly. Trust me, you will get marks for being clear and making it easy for the marker.

Let's look at another example. We'll look at more and more as the course progresses.

Does Movie Violence Increase Violent Crime?
By Dahl and Della Vigna, Quarterly Journal of Economics (2009)

The idea of this paper is that violent, action-based movies attract a particular type of audience in demographic terms. When these people are in the cinema then they are occupied and are not able to commit crimes.

The estimating equation looks like this:

$$lnV_t = \beta^v A_t^v + \beta^m A_t^m + \beta^n A_t^n + \Gamma X_t + \varepsilon_t$$

where V_t is the number of violent assults, A_t^v is audience numbers if violent movies; A_t^m is audience for mildly violent movies, A_t^n is non-violent movies, X_t is a vector of controls; and ε_t is the error term.

The results are then laid out in two main tables.

Effect of Movie Violence on Same-Day Assaults

Specification:	OLS regressions						IV regressions
Dep. var.:	Log (number of assaults in day t)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audience of strongly violent movies	0.0324	0.0005	−0.0061	−0.0051	−0.0072	−0.0091	−0.0106
(millions of people in day t)	(0.0053)***	(0.0053)	(0.0033)*	(0.0033)	(0.0033)**	(0.0026)***	(0.0031)***
Audience of mildly violent movies	0.0246	0.0017	−0.0084	−0.0042	−0.0056	−0.0079	−0.0102
(millions of people in day t)	(0.0030)***	(0.0029)	(0.0020)***	(0.0026)	(0.0027)**	(0.0022)***	(0.0028)***
Audience of nonviolent movies	0.0082	−0.0164	−0.0062	−0.0023	−0.0029	−0.0035	−0.0050
(millions of people in day t)	(0.0029)***	(0.0030)***	(0.0021)***	(0.0024)	(0.0026)	(0.0024)	(0.0029)*
Control variables							
Year indicators	X	X	X	X	X	X	X
Day-of-week indicators		X	X	X	X	X	X
Month indicators			X	X	X	X	X
Day-of-year indicators				X	X	X	X
Holiday indicators					X	X	X
Weather and TV audience controls						X	X
F-test on additional controls	1,934.02	1,334.31	88.56	13.37	15.05	18.58	
Audience instrumented with predicted audience using next weekend's audience							X
R^2	0.9344	0.9711	0.9846	0.9904	0.9912	0.9931	
N	1,563	1,563	1,563	1,563	1,563	1,563	1,563

Notes. An observation is a Friday, Saturday, or Sunday over the years 1995–2004. Assault data come from the National Incident Based Reporting System (NIBRS), where the sample includes agencies that do not have missing data on any crime (not just assaults) for more than seven consecutive days for that year. The movie audience numbers are obtained from the-numbers.com and are daily box-office revenue divided by the average price per ticket. The ratings of violent movies are from kids-in-mind.com. The audience of strongly violent movies is the audience of all movies with a violence rating 8–10. The audience of mildly violent movies is the audience of all movies with a violence rating 5–7. The specifications in columns (1) through (6) are OLS regressions with the log(number of assaults occurring in day t) as the dependent variable. The specification in column (7) instruments the audience numbers with the predicted audience numbers based on next weekend's audience. Details on the construction of the predicted audience numbers are in the text. Robust standard errors clustered by week are in parentheses.

* Significant at 10%; ** significant at 5%; *** significant at 1%.

A. Benchmark results

Specification:	Instrumental variable regressions			
Dep. var.:	Log (number of assaults in day t in time window)			
	(1)	(2)	(3)	(4)
Audience of strongly violent movies	−0.0050	−0.0030	−0.0130	−0.0192
(millions of people in day t)	(0.0066)	(0.0050)	(0.0049)***	(0.0060)***
Audience of mildly violent movies	−0.0106	−0.0001	−0.0109	−0.0205
(millions of people in day t)	(0.0060)*	(0.0045)	(0.0040)***	(0.0052)***
Audience of nonviolent movies	−0.0033	0.0016	−0.0063	−0.0060
(millions of people in day t)	(0.0060)	(0.0046)	(0.0043)	(0.0054)
Time of day	6 A.M.–12 P.M.	12 P.M.–6 P.M.	6 P.M.–12 A.M.	12 A.M.–6 A.M. next day
Control variables				
Full set of controls	X	X	X	X
Audience instrumented with predicted audience using next week's audience	X	X	X	X
N	1,563	1,563	1,563	1,562

Notice the same structure as I have suggested before.

Focus on key coefficients.

Tell a story across the columns.

Have comprehensive notes at the bottom of the page that explain everything the reader needs to know to interpret the table.

LAST TIP FOR WRITING UP RESULTS

Make sure you think through interpretation of coefficients and magnitude of effect.

Key things here are: a) making sure you understand functional form correctly, b) comparing the magnitude of an 'average' change in the X variable to the standard deviation of outcome Y variable.

I'll talk more about this next week.