Carrying Out an Empirical Project

ECONOMETRICS (ECON 360)

BEN VAN KAMMEN, PHD

Introduction

This lecture steps back from estimation and inference and reminds you to consider why you're doing econometrics in the first place.

- To test a theory about how the economy (or a part of it) works?
- To evaluate whether a policy or program accomplished what someone thought it would?
- To predict out of sample values of a variable (forecasting)?
- To complete a work assignment or term paper?
- To win at fantasy sports?

Estimating the regression is just one of the most technical parts of completing an empirical project, and this lecture attempts to put the methods from ECON 360 in context:

- "Why do we need to learn this?"

Outline

Posing a Question.

Literature Review.

Data Collection.

Econometric Analysis.

Writing an Empirical Paper.

It's science

A good question is . . .

Specific.

Overly Broad	Getting Better	Appropriately Specific
Labor Supply	Why is the number of workers claiming disability insurance rising?	How much have expansions in the generosity of claims (since 2008) increased D.I. claims?
Human Capital	Does grade inflation distort students' college major choices?	Does grade inflation decrease the probability that a student chooses a "STEM" major?
Unemployment	Do the unemployed search harder as the expiration of their UI benefits approaches?	Does the job finding rate increase at the individual level as proximity to UI benefits expiration gets closer?

Capable of being answered with data.
◦ Notice how the 3rd column is phrased in terms of <u>variables</u> that can, in principle, be observed.

Interesting. There are obviously shades of grey here. Ranging from:
◦ "Is the Affordable Care Act (major federal health law) 'good' for labor market efficiency?"

To:
◦ "Are quarterbacks 'undervalued' in my fantasy football league?"

Interesting-ness

If the question is interesting to you, it's worth trying to answer with statistics.
  ◦ Or is it vice versa?

If the results are interesting to enough *other* people, you might be able to publish the results somewhere or get rewarded materially for (a job? promotion? grant?) them.

Somewhere in between is the threshold for a good term paper/thesis.
  ◦ You're probably not aiming as high as publication, although there are plenty of journals that are not difficult to publish in.
  ◦ But you want the results to be <u>non-obvious</u>, non-trivial, and to attempt to answer something that your professor/supervisor has not read about 1000 times.
    ◦ Test this by talking to him/her or pitching the idea to your friends or classmates. You should be able to explain to them in a 1 minute "pitch" why you chose this topic without inducing yawns or glazed over eyeballs.

Interesting-ness (cheat codes)

If you are having a hard time identifying an interesting topic, these suggestions can help you come up with an acceptable, if less than ideal, idea:

◦ Estimate a familiar model with the addition of a new regressor,

◦ "Update" an old finding by estimating the model with more recent data,

◦ Estimate a model with new data from a different country/state to see if the result still holds,

◦ Replicate an existing study,

  ◦ This may or may not satisfy your supervisor's requirements, but there is too little of replication/verification in the economics field.

  ◦ Sometimes you will even discover errors in the original by doing this!

◦ Use different estimators on an existing model.

I refer to these as "cheat codes" because they are reliable ways to get a workable paper, without being as rewarding to you as pursuing a topic that you are passionately interested in.

Literature review

One way to explain why your topic is interesting, at least to academics, is to put it in context with other preceding papers.

◦ Lists what has already been done on this subject, so

◦ Your contribution(s) can be distinguished from the existing knowledge.

◦ Previous papers ("the literature") establish the convention(s) for studying a topic, e.g.,

  ◦ How to specify the dependent variable,

  ◦ What level of data has been (should be) used,

and this can provide justification for the choices you make (model, specification, data). "The log-log specification is standard in this literature, so it is the one I use in the present paper," you might say.

Literature review (what to do)

The idea is to read (and take notes on) all related academic (maybe some popular, as the situation warrants) sources. The ones found in scholarly journals.

- ◦ Found on search engines such as Google Scholar and EconLit (or the University library, generally).
- ◦ Usually the articles you want are in journals that have the word *economic(s)* in the title, but there are notable exceptions, e.g., *Journal of Human Resources*, *Industrial Relations*, and *Demography*.

This is also a good way to tell if your topic is too broad. If you find yourself making a list of 100 papers to read, you should narrow it down.

Summarize "the literature" in a few paragraphs of text, emphasizing the most similar (to yours) sources, so you can compare and contrast your results to their authors'.

Data collection

What kind do you need to answer the question of interest?

- ◦ Think about the "gold standard" or "experimental ideal" and how you could mimic it using observational data.
- ◦ Which data sets have others used?
- ◦ Do you need to use longitudinal ("panel") data or would a cross section suffice? The latter might require more control variables be observed . . .
- ◦ What does your professor/supervisor recommend?

Consider what level: individuals/firms? Aggregates by industry? By occupation? By state?

The National Bureau of Economic Research (NBER) has a nice representative list of data sets, including links: http://www.nber.org/data/.

- ◦ Even confining attention to commonly used sources, it would be a long list to try naming them all.

Loading and storing data

This is one major difference between in-class exercises and "the real world." Data probably will not arrive in sanitized formatted (for your stats package) files, ready to use.

Hopefully they arrive, at least, in electronic form with some structure, so they can be *import* –ed or copied into the spreadsheet Data Editor in Stata.

◦ Structure means text with lines differentiating observations and either: a specified character (e.g., comma or tab) separating variables, or variables that have a fixed width so the observation can be "chopped up" into its parts.

If they arrive in printed form, you can (for reasonable numbers of cells) enter them manually in a spreadsheet; this is labor-intensive but at least you have a unique data set that few others have likely used.

The only norm, when it comes to loading data, is that your data set will be an exception. But there's always a way, and Stata's help files do an excellent job of explaining your options.

Loading and storing data (hacks)

Panel Data should be stored:

◦ In the long form, as opposed to the wide form. See Stata help file on the *reshape* command.

◦ Using the *ts* and *xt* settings in Stata (which is <u>highly</u> desirable) requires a unique (numeric) identifying variable for the individuals in your cross sections ("i") and a variable differentiating the time periods in which you observe person "i".

◦ You can always use *sort* {varlist} to arrange observations in such a way that the settings will work, provided you have valid "i" and "t" variables.

"Match" variables from one data set to a 2nd one observing the same units, e.g., if you want the unemployment rate in the state where person "i" lives, use the *joinby* command.

◦ This thing is a miracle cure for combining data.

Missing values should be represented by a period and not a numerical "code".

Cleaning data

Probably the #1 thing you spend (lots of) time on in the real world that you don't have to when doing in-class exercises.

Obtain any published "codebooks" or documentation files that accompany your data set (from the agency that collected the data.

◦ It's not their responsibility to guarantee accuracy and cleanliness of each observation.

◦ Reasonable to expect them to explain what they did to encode non-responses, censored, top-coded, imputed values.

Use your own senses: study the summary stats, look for outliers, see how earlier authors in the literature handled their data problems

◦ Does the distribution pass the "smell test"?

Daniel Hamermesh's "The Craft of Labormetrics" (pp. 363-366) is an excellent explanation of the risks resulting from failure to do this.

Cleaning data (hacks)

Watch out for missing observations coded numerically.

Learn how each variable was arrived at, e.g.,
- Hourly/weekly/monthly/annual?
- Units?
- Real/nominal?
- Seasonally adjusted?

Remember what you're doing:
- Why does taking the *log* return missing values?
- Why is there no difference for the first time period in a series?
- Difference between the *d2* operator and *s2*?

$$\text{first one is: } (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}); \text{second is: } (x_t - x_{t-2}).$$

Econometric analysis

This is what all the other lectures are about, but when an exercise isn't telling you which method to employ, you have to make a choice. Kennedy* lists a few criteria to consider when choosing:

- Computational cost (this will rarely be binding for novice empiricists),
- "Fit": are the residuals small? Is R^2 high?
- Unbiasedness: are the assumptions necessary to imply this realistic? Also there is a trade-off between unbiasedness and . . .
- Efficiency, least variance, least wide distribution of $\hat{\beta}$; minimizing the mean squared error can help.
- Asymptotic properties.

Even if it's not the *preferred* method, OLS is the place to start.

You can "lean on" the literature to guide you. What have others done?

* Kennedy, Peter. 2003. A Guide to Econometrics, 5th Ed. MIT Press, Cambridge, MA.

Econometric analysis (robustness)

Look out for influential observations; is your estimate sensitive to excluding a single observation?

- *lvr2plot*
- Cook's Distance for each observation: *predict {name}, cooksd*.
- Robust regression (*rreg*) to see how sensitive estimates are to re-weighting.

Is it sensitive to adding regressors?

- Wooldridge: "If a variable is statistically significant in only a small faction of the models estimated, it is quite likely that the variable has no effect in the population."

Writing an empirical paper

As a resource for helping compose a paper, the following are highly recommended.
◦ Miller, Jane E. 2004. The Chicago Guide to Writing About Numbers. University of Chicago Press: Chicago, IL.
◦ McCloskey, Deirdre. 2000. Economical Writing. Waveland Press: Long Grove, IL.

Also my other class notes that discuss presenting estimates in tables.

Reading published scholarly articles also reveals organization and style tips. Organization goes:
◦ Introduction
◦ Literature Review
◦ Conceptual Framework
◦ Methods and Data
◦ Results
◦ Conclusion

Writing an introduction

An introduction identifies the topic, explains its place among the overall literature, and <u>motivates the reader to continue through the rest of the paper</u>.

- ◦ "Why should I care?" is a question you want to answer as early in the paper as possible.
- ◦ This means giving the reader a glimpse of how public policy or our economic welfare hinges on the findings you will present.
- ◦ Get to the point and quickly; the introduction should be no longer than 1 page.

Miller likens this part of the paper to the "opening argument" in a legal trial.

- ◦ You back up your generalizations later on with specific evidence.

One can assume that the reader is sufficiently interested that he is reading an article about "The Economics of [fill in the blank]", but you have to explain why the facet you are writing about is so important.

Writing a literature review

Unlike Wooldridge I say present this as a 2nd section.

The point is:
◦ List what is already know about your topic.
◦ Reveal what is not already known, i.e., something you will contribute with the present paper.
◦ Give the reader context in which to judge your results to see if they're plausible and consistent.
  ◦ If not, why?

Writing about the conceptual framework

Identify the relevant <u>theoretical predictions about the topic</u>, e.g.:

◦ "The compensating wage differential for a job requiring work-related travel will be positive if workers consider travel a disamenity. However if some workers enjoy traveling—and they are the only ones who accept the traveling jobs—the compensating differential could be negative, as well."

Explain any jargon that is necessary so someone not acquainted with the literature can understand.

You don't have to include mathematical derivations from theoretical models as the authors of academic journal articles often do, but you should be able to find the key assumptions of their models and the testable implications.

◦ Common sense reasoning is usually enough to make your point.

◦ Emphasize things that could "confound" your results and should be controlled for.

Writing about the data and methods

Get to the point and write out your regression <u>model</u>. E.g., a <u>fixed effects</u> model would look like this:

$$y_{it} = \beta_0 + \delta_0 t + \beta_1 \boldsymbol{X}_{it} + \gamma D_{it} + a_i + u_{it}.$$

In the next paragraph, explain how y is defined, which control variables are included in the vector X, and how "gamma" is the effect of interest, e.g., associated with a policy's ($D = 1$) effect.

◦ You may also opt for writing it in "equation" form (à la Wooldridge's text) with the variable abbreviations. But this makes the expression longer, and often it's better to just explain in words what the notation means.

Writing about the data and methods (continued)

Explain how you will estimate the parameters in the model.

Discuss the assumptions necessary to estimate them consistency and without bias.

Discuss any modifications to OLS that you need to deal with unobserved variables, e.g.,
◦ Instrumental variables: assumptions of validity and strength.
◦ FE or FD or regression discontinuity and how they mitigate biases.

What do you need to do about heteroskedasticity?

Writing about the data and methods (concluded)

Describe the <u>data</u>.

- ◦ Cite the source and how you accessed the sample you're using.
- ◦ Discuss units of measure and any measurement or missing data problems. Justify any modifications you made to the sample.
- ◦ Include a table of summary statistics. It may be necessary to show detail on your variable(s) of interest.
    - ◦ "When did each state in your sample adopt the [policy in question]?"
- ◦ Explain the longitudinal structure, i.e., number in the cross section and how many cross sections?
    - ◦ "The sample consists of 2 cross sections (2001 and 2005) of 72 counties each, for a total sample size of 144."

Writing about the results

Organize this section into paragraphs, one per result.

Begin each paragraph with a generalization of the result, e.g., "earnings are positively related to veteran status, with veterans earning a premium of roughly 4% compared to similar non-veterans."

◦ Then you can list and discuss a representative example and any exceptions: "most of my estimates are within $\pm$ 1 percentage point of 4% and are statistically significant at a minimum of 95% confidence; the only exception is the one that restricts the sample to peace-time veterans. I suspect this may be . . . ."

The quantitative estimates will appear on self-contained tables, so no need to repeat all of the "gory details" in the text, but direct the reader's attention to the "most important" figures.

Writing about the results (concluded)

Discuss practical significance and magnitude.

◦ Compare to other papers in the literature if possible.

Discuss whether coefficients have the expected signs.

◦ Is there an omitted variable or other explanation?

Can different methods' results be traced to the assumptions underlying them?

◦ "The rate of return to education should get smaller when you control for (positive) ability bias, e.g., with an instrumental variable."

◦ You can discuss the strengths and "limitations" of the methods you used honestly.

  ◦ What would you need to overcome the limitations? Different data?

Writing the conclusion

This doesn't have to say anything particularly new, unless you've discovered some interesting directions to extend your analysis in future projects.

◦ In fact all of your important findings should already have been listed and discussed in the preceding section!

Summarize the main findings for emphasis

◦ And what practical conclusions you draw from them.

◦ "You know? I learned something today . . ."

List and discuss any "grains of salt" the reader should take along with your results.

General instructions (from my ECON 385 project instructions)

As a minimum for a good grade, **the caliber of written communication will befit a college graduate**.

- A paper that is incomprehensible (because of poor sentence structure, grammar, using words out of context, or subject-verb disagreement, et al.) will earn you no points.
- I will not (nor will any reader) waste time trying to decipher poorly written paragraphs.
- I have to read over 100 papers from the class, and **I reserve the right to award a zero to any paper that is too hard to read for grammatical or mechanical reasons**.

If you are concerned about your writing ability,
- visit the writing center;
- get a friend, sibling, or co-worker to read your paper and proofread it;
- run spellcheck (!) and search your paper for incorrect homonyms (spellcheck won't find these).

Do whatever it takes to avoid handing in a poorly written paper.

General instructions (continued)

Cite your sources in the text, (Author year) and include a *works cited* page.

Use active voice.
◦ Avoid the following phrases: "I think", "I believe", "I feel". You're writing the thing; you wouldn't be writing it if you didn't think it.

Double space your submission.
◦ It should go without saying, but type your assignment and hand in a printed paper copy. Hand-written submissions are not acceptable.

Generally do all the other good things you learned in English composition classes. Remember it's your job to communicate your thoughts to the reader—not the reader's job to divine what you are trying to say.

Checklist

Title that is descriptive of the topic.

Double-spaced.

Page numbers.

Equations, figures, and tables all numbered (consecutively).

Citations for all sources and a works cited page.

Tables "self contained" so you could understand them even without reading the supporting text.
- Usually in a draft of a (non-class) paper, you put all of the graphs and tables at the end of the document.
- Little placeholders like [Table 1 here] go in the text to show where to put them in the final version.

Sections have headings, probably numbered, i.e.: *1. Introduction*.

Conclusion

The application of econometrics is usually part of a much more complex project.

Producing a convincing empirical finding is the ultimate test of your skills as an economist, drawing on your knowledge of :

- ◦ (micro/macro) economic theory,

- ◦ econometric theory,

- ◦ how to use software,

- ◦ how to communicate effectively using words <u>and</u> numbers, and

- ◦ most of all, how to solve the problems you encounter along the way.

There are too many potential problems to cover them exhaustively, but hopefully this class has identified many of the most common solutions and shown you ways to think about solving the remaining ones.