About Final Project…

Econ B2000, Statistics and Introduction to Econometrics

Kevin R Foster, the City College of New York, CUNY

Begin investigating a topic in economics that interests you. This will eventually become the final project that caps the class. Each group will write a final project. The group may be of any size from one person up to four people. The standards for the quality and amount of research are, of course, increased for a larger group.

I want you to use real micro data – not pre-chewed, not data that somebody else has gone through and summarized. For example, the PUMS data are available cross-tabulated and with means by various sub-groups, but that's not what we've been using. Reading somebody else's stats is not doing stats, any more than watching sports makes you athletic.

Step 1: Find a topic.

Step 2: Write a basic overview of some of the recent economic research that has been done on this topic. This initial overview should be just a couple of pages and should demonstrate your basic familiarity with the topic, both the theory and the initial data work that must be done. (The research should be quality academic work, not from newspapers or magazines – so search EconLit not just Google.)

Step 3: refine and repeat.

Often the best way to do a project is to find a cool paper and update it – a paper published in 2016 might have data from 2013 so you can often find a few years of more recent data. It is a useful scientific question to ask if the paper's conclusions still hold with the addition of more data! You can extend the original paper if you want but it's often a good place to start.

Ultimately will submit a project (the written document, with relevant graphs and tables), as well as supplementary material including electronic versions of readings, output files generated by R, and your dataset (if it's not one of the ones I've given you).

Please consult both the CUNY policies on academic integrity and my synopsis about forms of proper citation (below). From my past experience it seems that many students lack this essential background knowledge so please take the time to ensure that you are in compliance. Note that if you work with a group you are jointly liable for any plagiarism.

Don't worry too much about finding the perfect topic: you're not writing a contract, just moving along a particular path for now. If, after a couple of weeks, you change your mind and find another topic that seems better, you can *certainly change! You can also change your study group – you might fragment and/or re-group as we go along. The essential step is to start moving; for too many students 'the best is enemy of the good.'*

Finding a topic:

"Interesting" articles/topics (to me, anyway): [and don't tell me you have trouble finding these articles – use Google to get exact titles; you should be able to use the library to get the journal articles through either EconLit or JStor or Google]

- Household Pulse data on pandemic
- with ATUS – here is some research http://www.bls.gov/tus/research.htm
- look at waves of CPS studies to examine macro effects – how recessions affected different demographic groups?
- National Health Interview Study has all sorts of medical and healthcare data – who has insurance, how often they're sick, doctor visits, pregnancy, weight/height
- BRFSS, Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2010, http://www.cdc.gov/brfss/technical_infodata/surveydata.htm.
- NHANES – National Health And Nutrition Examination Survey, http://wwwn.cdc.gov/nchs/nhanes/bibliography/default.aspx
- Survey of Consumer Finances from Federal Reserve
- IPUMS has historical census data http://www.ipums.org/
- HRS – Health and Retirement Survey from the University of Michigan, has extensive data on people in or near retirement, including their finances, health, and expectations
- These are all US sites; more and more other countries are putting their data online – explore!
- look at labor market outcomes for recent veterans; this really lends itself to a group since each person can take a different part – for example one use ATUS, one use NHIS, BRFSS, CPS, SCF, etc.
- Census has info on small business owners: what makes an entrepreneur?
- IMF/World Bank have good international data
- finance data from WRDS – daily stock price and other data going back years, easily available to look at stock/bond returns even at daily intervals (some have denser); there's even some data on option prices
- lots of sports data (Red Sox and Patriots use cutting-edge statistical and mathematical-optimization techniques); Romer on Bellman Equation in (American) football, Levitt on soccer penalty kicks; look at http://journals.academia.edu/JournalOfSportsEconomics to see the more academic way of approaching these questions, or http://www.sabernomics.com/sabernomics/, www.baseball1.com There are even R packages like nflfastr for NFL and other analogous.
- real estate/housing/mortgage data – very important and relevant to policy
- Dept of Energy has data on fuel prices as well as household choices about energy use

- Fed Reserve has data on household finances – again, very interesting – how does credit card debt, car debt, student load debt and household debt correlate? How vary with socioeconomic status, location, age, education? Veterans have additional mortgage options; are these important?
- Deaton paper on global happiness; www.worldvaluessurvey.org has some data; JEP Winter 2006
- NYC Open Data https://nycopendata.socrata.com/
- on taxis, http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/
- DHS data http://www.dhsprogram.com/What-We-Do/survey-search.cfm?pgtype=main&SrvyTp=country

  But no Kaggle or similar – those datasets are meant for practicing technique not understanding.

Internet Resources

An internet search on a given topic will return a wide variety of hits. The most difficult task is to differentiate the junk (most of it) from the few bits of useful information. Since you are a student, just beginning to learn the field, it is only to be expected that you will have a more difficult time distinguishing the good from the bad. You must be wise, dutiful in checking out sources, and should ask questions.

You might usefully create a web page of your own. This takes about 30 minutes to learn, and removes any mystery. Sometimes students think that publishing online is difficult, so only very high-quality material should be online – FALSE. Anybody imbecile can put any damn thing online – and we do! You've got to be careful. Blogs and wikis have lowered the bar even further. A reader needs to be careful and critical of every source.

There are certain sources that have filtered out much of the worst junk. You can limit your search to only articles published in refereed journals by searching online databases (from the CCNY Library), such as EconLit and JStor. Of course not everything that is published is correct – you must still be diligent in finding recent sources, making your own evaluation of the plausibility of the claims, and arriving at your own judgments.

Both of these links are easily found from the CCNY Library's page, where you can pick them from the list. You need to access them from a CCNY computer, or else you will need a login (which the library can provide you, however this takes time so don't wait for the night before the paper is due!)

EconLit collects citations, most with a detailed abstract, and a large fraction have full text available. "Full text" means that you don't have to find the physical journal; you never touch paper. Just save the .pdf file that it produces.

A hint: one of my favorite journals to recommend to students is the **Journal of Economic Perspectives** (JEP). This gives excellent overviews of particular topics in economics, meant to be accessible to a non-specialist, written by some of the most prominent people in those fields. It is published by the American Economic Association (AEA) and is available through both EconLit and JStor. The **Journal of Economic Literature** (JEL) if also from the AEA and it also has occasional articles that summarize a topic. The library has both journals on the shelves – you can browse through these journals, just skimming to find interesting articles. It's a great way to spend a few hours!

JStor has the full text of articles published in the foremost journals of various disciplines (including Economics, Finance, and Statistics). Generally these articles are at least 3-5 years old, but it gives access to every article in the most important journals for the past several decades.

For news stories, you can search **Lexis/Nexus**. This collects the full text of major newspapers, including the New York Times and the Economist. Again, you need to access them from a CCNY computer, or get a login to work remotely.

There are other outlets, such as prominent and well-regarded thinktanks and policy institutions. In economics, the National Bureau of Economic Research (NBER, at www.nber.org) is highly regarded, as is the Brookings Institution (www.brookings.edu).

Of course all of these sources give "the establishment view" not the ideas and opinions of extremists. This is true by definition: formerly extreme views become mainstream once "the establishment" has published them. I do not want to discourage you from research on the fringes, however many classes at this College will require that you demonstrate a knowledge of the mainstream. (Marx and Keynes and Hayek began their radical writings by first demonstrating their knowledge of what had been written previously, to show where it had holes.)

That said, sometimes if I find an interesting article on EconLit that is not available as full-text there, I can Google it to find a free ungated version.

DATA

There are many sources of data online. Although the principal sources change depending on the particular field, here are some of the basics.

Macro data (maybe not enough for this project but can supplement) at FRED
http://research.stlouisfed.org/fred2/
http://www.federalreserve.gov/– The Federal Reserve
http://www.ny.frb.gov – our own New York Fed
http://www.census.gov/ the US Census
[or http://www.census.gov/main/www/subjects.html is a list of topics on which the Census Bureau has data.]

http://stats.bls.gov/ is the Bureau of Labor Statistics; as well as the CPS at http://www.census.gov/cps/
http://www.oswego.edu/~economic/data.htm is a good portmanteau of links
http://www.worldbank.org/lsms/ World Bank
Overview of BLS data www.bls.gov/bls/inflation.htm
CPI www.bls.gov/cpi/home.htm
CPI Chained http://www.bls.gov/cpi/super_cpi.pdf
PPI www.bls.gov/ppi/home.htm
Unemployment www.bls.gov/cps/home.htm
Wages www.bls.gov/bls/employment.htm
BLS details on coverage www.bls.gov/opub/hom/homch1_a.htm
Health Statistics of Population, National Health Interview Statistics
Census Data Access: http://dataferrett.census.gov/

Write Well

As a general rule, what you read determines how you write. If you don't read much then your writing will be lousy. If you read just texts, tweets, and celebrity gossip web pages, your writing will be trashy. If you run it though some AI, your writing will sound blandly average. If you start to read lots of clear papers that aim to communicate complicated and nuanced ideas, then you've got a good start. Find an economist, whose research is interesting, and read her collected works – you could do worse! If you want to read about economics writing, try Dierdre McCloskey (www.deirdremccloskey.com; her books are also in the library).

Academic Rules for Citations and Avoiding Plagiarism

Read CUNY's policy on academic integrity (on the course syllabus and CCNY web page).
see also http://www.dartmouth.edu/~sources/, http://www.princeton.edu/pr/pub/integrity/index.html

Harvard's guide to 'Writing Economics' is an excellent overall reference. See also NYTimes recent bits on plagiarism.

The essential idea is to differentiate your own contributions, what is new about your analysis or compendium, as distinct from what is taken from other sources.

You must realize what constitutes intellectual achievement: gathering diverse sources and comparing them one to another is such an achievement. But you must be clear about what is gathered, versus what points you are making with your comparison.

Here's a blog post about plagiarism from Reuters,
Jack Shafer, "How to think about plagiarism," Oct 14, 2011,
http://blogs.reuters.com/jackshafer/2011/10/14/how-to-think-about-plagiarism/

> *An editor must have a heart like leather. Not freshly tanned leather—all supple and yielding like a baby's bum—but like an abandoned baseball glove that's been roasting in*

the Sonoran Desert for five or six years. Only those who are hard of heart can properly deal with the plagiarists who violate the journalistic code. ...

"There are no mitigating circumstances for plagiarism," the cold, cold heart of Washington Post Executive Editor Marcus Brauchli stated earlier this year after Post reporter Sari Horwitz got caught stealing copy from the Arizona Republic.

Brauchli got it exactly right. It doesn't matter if you pinched copy because you were tired, you were harried, your spouse or child was sick or dying, you were under deadline pressure, you jumbled up your notes, you took boilerplate or wire copy that nobody should really claim "authorship" over, you have a substance problem, you committed a cut-and-paste error, you were blinded by the "warp speed" of the Internet, you were a victim of the "win the morning" culture, you are young and inexperienced, you had two windows open at the same time and confused them, or any of the excuses tendered by the accused reporters described in Trudy Lieberman's 1995 Columbia Journalism Review article.

These aren't excuses. These are confessions. And they mitigate nothing.

As I've written before, plagiarism doesn't offend me because it exploits the previous hard work of some enterprising writer—even though it does. When you attribute passages to another writer, you're likewise exploiting their work. But at least they receive psychic income from the citation. The quoted writer is enriched by the fact that their work has been acknowledged, that somebody might go back and read their work, and that their reputation is likely to rise because of the credit thrown their way.

Spare the violated writer any pity. He'll be okay. Give your pity to readers, who are the real victims.

The plagiarist defrauds readers by leading them to believe that he has come by the facts of his story first-hand–that he vouches for the accuracy of the facts and interpretations under his byline. But this is not the case. Generally, the plagiarist doesn't know whether the copy he's lifted has gotten the story right because he hasn't really investigated the topic. (If he had, he could write the story himself.) In such cases he must attribute the material he borrows so that at the very least the reader can hold somebody accountable for the facts in a story.

Or to put it another way, a journalist who does original work essentially claims, this is true, according to me. The conscientious journalist who cites the work of others essentially makes the claim that this is true, according to somebody else. The plagiarist makes no such claims in his work. By having no sources of his own and failing to point to the source he stole from, he breaks the "chain of evidence" that allows readers to contest or verify facts. By doing so, he produces worthless copy that wastes the time of his readers. And that's the crime.

RULES:

When directly using someone else's words, these must be in quotation marks with an explanatory reference (either cite, footnote, or endnote)
example of cite:
"A strong, credible body of scientific evidence shows that climate change is occurring" (National Academy of Sciences, 2010).

if long quote (>50 words) then no quotation marks but indent (sometimes different font)
example of long quote with cite:

> Perhaps the most important argument for engaging in alternative monetary policies before lowering the overnight rate all the way to zero is to ensure that the public does not interpret a zero reading for the overnight rate as evidence that the central bank has "run out of ammunition." That is, low rates risk fostering the misimpression that monetary policy is ineffective. As we have stressed, that would indeed be a misimpression, as the central bank has means of providing monetary stimulus other than the conventional measure of lowering the overnight nominal interest rate. However it is also true that policymakers' inexperience with these alternative measures makes the calibration of policy actions more difficult. Moreover, given the important role for expectations in making many of these policies work, the communications challenges would be considerable. Given these difficulties, policymakers are well advised to act preemptively and aggressively to avoid facing the complications raised by the zero lower bound. (Bernanke and Reinhart, 2004)

when using your own words to state someone else's idea or reproducing their image, graph, or data, you don't need quotes but still need a reference.
example:
It is important to address the issues of providing an appropriate decision tree when analyzing game theoretic choices (Aumann, Hart, and Perry 1997).

The cite is of a format that enables the reader to go to your bibliography & find that reference. The usual style is (Author, p. ##) for a book or (Author, Date) for an article or other reference. A footnote or endnote is similar, but placed in a different spot. I think cites work best.

The Bibliography is at the end of the paper, and lists all works used. Include data sources! Also books, articles, web pages, images, graphs, etc. Make sure if your cite is a hyperlink, it gives enough info for a reader of the hard copy to find the reference.

Examples:

Aumann, R., S. Hart and M. Perry, (1997). "The Forgetful Passenger," *Games and Economic Behavior*, 117-20.

Bank of Japan, Flow of Funds Accounts, (2010). http://www.boj.or.jp/en/theme/stat/index.htm accessed July 12, 2010.

Bernanke, B. S., and V. R. Reinhart, (2004). "Conducting Monetary Policy at Very Low Short-Term Interest Rates," *American Economic Review*, May.

Ehrenberg, R. G., and R. S. Smith, (2000). *Modern Labor Economics: Theory and Policy*, seventh edition. New York: Addison-Wesley.

Heckman, J., (1974). "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42(4).

National Academy of Sciences, (2010). "Advancing the Science of Climate Change," *Expert Consensus Report*, National Research Council.

Vames, Steven, (2000). "Income Gains in March Outpaced Spending," *New York Times*, April 28, 2000. http://www.nytimes.com/yr/mo/day/news/financial/28tsc-economy.html

CONSEQUENCES:

Failure to follow these rules is a violation of Academic Integrity. This is a severe violation of the basic principles of the academic community. You may be brought up on academic charges before a Disciplinary Committee of the College, where you are subject to a range of consequences up to expulsion.

Default Guidelines for Final Project

This is a Stats Paper

You're writing a paper for stats class not theory, so concentrate on the stats. There is a certain amount of theory and background information that needs to be stated, in order to show that you understand the basic outlines of the problem being considered. But the large majority of the paper should be looking at the basic statistical results.

Possible Format:

You don't have to follow this format – use your own if you have a better idea. If you don't, though, this will serve as a basis. Also please ensure that your project does not go through and give bullet points in response to each question! You should write a narrative that gracefully includes the answers to these questions. The quality of the writing is a large factor determining the grade you get. There are writing tutors available – use them!

Introduction

A concise description of the project: include the dataset used, the key interesting results (don't reproduce everything), and why those results are interesting. Should be about a page so every word must count!

Literature Review

Describe the papers you've read that also look at this topic. Explain the differences among the results found in different previous studies. You can point out challenges that remain (even if your project doesn't solve them all). Do different authors come to different conclusions? Why might this be? Are their regressions valid (e.g. do they take adequate account of endogeneity issues)? What econometric techniques does each investigation use? Are these appropriate, given what we have learned in class? What are the important assumptions required by this method? Are these reasonable? Are there other techniques (discussed in class) that might be appropriate? (Also, what data sets? Are those data available free online?) These should be academic papers – serious studies not newspaper accounts. You can cite a newspaper to indicate why the result is interesting (e.g. to show that policymakers or the public cares about knowing the real answer, or to give some background on why you're interested in it) but you can't end there.

Means (simple graphs, correlations, differences of means)

First carefully note the dataset you're using, both the original source and any subsequent restrictions (e.g. if you're only looking at children or only those who are working or whatever). Present a table where each important variable in your regression has its mean and standard deviation as well as any other relevant summary statistics (min/max, median, whatever). Verify that the units all make sense.

This is a good place for simple graphs of the sort that we've talked about. Does a two-dimensional scatterplot show your regression results? Why or why not? This is also a good place to discuss functional forms: does the graph show that squared or cubic terms could be useful (or logarithms)? What about subgroups? Medians? (Look over past homework assignments for examples.)

Simple Regressions

Present a few different models in easy-to-read tables. Don't just cut-and-paste the output! That is unacceptable.

Complicated Regressions

Present some more regressions (again, in easy-to-read tables). Show your main conclusion then do some robustness checks (i.e. what if the sample were limited to only males or females or only those of certain ages or whatever is relevant). Go back to the homework assignments from class and do just those sorts of regressions; for example if you have age plus its square and cube, do the results (the coefficients on the variables of interest) change when you put in 5-year age dummies? Show ability to do multiple-variable hypothesis tests. Show off some fancier methodologies.

Conclusion (Explain Results)

Clearly state what you have found and why it is interesting. Do your results confirm what other researchers have found? Or do they contradict earlier research? Why might this be?

Hand in: Paper, dataset, R program [Blackboard for the paper then put link to GitHub onto Slack for all the additional material]

Length? Size doesn't matter! I want excellence; page counts are for high schoolers. Some projects just require more pages than others; some people might just be more verbose while others are laconic. With some projects you'll spend a lot of time getting the data into shape so that will take most time; with others the data is ready to go. Don't worry I know the difference. Since we're economists we care about efficiency: saying the most interesting things in the clearest and most concise way. However I have found that many students still yearn for a page count, despite all of my pleas to ignore it. So, reluctantly, I advise about 15 pages per student (so a three-person group might produce 45 pages). But don't obsess about the page count!

Don't Plagiarize! Remind yourself of the rules for academic honesty (many many previous references are available). The consequences for violations are substantial – up to expulsion.

Rubric of Econometrics Final Projects

1. Does the paper show basic statistics for all of the relevant variables? Does it explain what these mean?
2. Does the paper show simple correlations or conditional means, and demonstrate how to incorporate these basic statistics with the question being considered?
3. What regressions are being done – does the paper demonstrate ability to execute basic linear regressions? (These alone hardly make a C paper) Does the paper explain the meaning of these simple regressions, sufficient to demonstrate that the author understands the concepts?
4. More regressions? Things like interactions, squared or higher-order terms? Again, explain? Variety of specifications, debating considerations like possible endogeneity. These get to B.
5. Advanced statistics and regressions – does the paper demonstrate advanced techniques beyond those carefully covered in class? Is student able to take some advanced concepts, maybe just mentioned in class, learn them by herself?
6. Alt (not required but nice): did students find and assemble their own data, or did they use one of the ones from class?

Example

Using 2010 CPS data, restrict to only fulltime workers with a non-zero wage. Run two sets of regressions to explain earnings: with earnings (annual wage and salary) as the dependent variable; with log of earnings as the dependent.

The first set of basic explanatory variables is hypothesized to be factors such as age, sex, education, race/ethnicity, marital status, veteran status, and if a union member. The means are:

Wage/Salary (annual)	$	49,773.79
Age		41.88
Female		44.5%
White		79.7%
African-American		11.8%
Asian-American		5.8%
Native American/ Indian/ Alaskan/ Inuit/ Hawaiian		2.8%
Hispanic		16.1%
Mexican		9.8%
Puerto Rican		1.4%
Cuban		0.6%
Immigrant		17.5%
1 or more Parents were immigrants		23.8%
Education: no high school		8.6%
Education: High School Diploma		28.9%
Education: Some College (incl no degree or Assoc degree)		27.9%
Education: Some College but no degree		17.5%
Education: Associate in vocational		5.0%
Education: Associate in academic		5.4%
Education: 4-yr degree		22.5%
Education: Advanced Degree		12.1%
Married		62.0%
Divorced or Widowed or Separated		14.8%
Unmarried		23.2%
Union member		2.2%
Veteran (any)		7.4%

The regression estimates are made with three basic specifications: Spec 1 has just the listed variables; Spec 2 included dummies for industry, occupation, and state of residence; Spec 3 has dummy interactions for female*age, African-American*age, female*African-American*age, Hispanic*age, female*Hispanic*age, and female*education.

	Spec 1	Spec 2	Spec 3

	estimated value		estimated value		estimated value	
intercept	-$28,685.56	*	$13,744.52	*	-$10,978.43	*
	1954.106		*3025.180*		*3685.959*	
Age	$2,517.92	*	$2,012.04	*	$3,052.09	*
	93.814		*88.514*		*133.158*	
Age-squared	-$23.60	*	-$18.55	*	-$29.40	*
	1.055		*.994*		*1.504*	
Female	-$17,380.74	*	-$14,587.20	*	$26,912.27	*
	360.019		*393.294*		*4202.955*	
African American	-$6,136.77	*	-$5,315.62	*	$17,924.27	*
	552.138		*545.564*		*7559.610*	
Asian	-$783.89		-$3,140.09	*	-$3,196.33	*
	861.879		*851.007*		*849.324*	
Native American Indian or Alaskan or Hawaiian	-$4,615.72	*	-$3,077.92	*	-$3,030.05	*
	1054.697		*1025.422*		*1022.749*	
Hispanic	-$5,176.56	*	-$4,433.05	*	$32,492.36	*
	596.068		*588.188*		*5715.141*	
Immigrant	-$7,377.88	*	-$4,669.63	*	-$4,080.20	*
	776.395		*731.493*		*733.482*	
1 or more parents were immigrants	$4,513.48	*	$1,231.87		$892.78	
	718.087		*677.532*		*677.771*	
Education: High School Diploma	$7,658.27	*	$3,819.68	*	$4,208.53	*
	701.918		*667.305*		*826.691*	
Education: Some College but no degree	$15,430.94	*	$7,791.73	*	$9,434.14	*
	756.430		*734.022*		*900.898*	
Education: Associate in vocational	$15,719.42	*	$8,376.06	*	$9,873.19	*
	1003.190		*966.454*		*1098.448*	
Education: Associate in academic	$19,907.99	*	$9,660.31	*	$11,310.63	*
	978.304		*948.764*		*1091.644*	
Education: 4-yr degree	$35,565.50	*	$20,756.84	*	$24,651.87	*
	738.325		*761.377*		*949.760*	
Education: Advanced Degree	$63,729.94	*	$40,911.95	*	$46,708.57	*
	815.818		*896.308*		*1109.431*	
Married	$8,100.77	*	$7,074.38	*	$6,912.90	*
	486.083		*459.856*		*459.565*	
Divorced or Widowed or Separated	$1,646.98	*	$1,893.12	*	$1,881.97	*
	633.993		*595.046*		*594.911*	
Union member	-$3,992.75	*	$2,282.96	*	$2,372.64	*
	1169.615		*1108.181*		*1105.552*	

Veteran (any)	-$1,186.63	-$884.41	-$905.22
	687.786	*648.453*	*659.002*
R-squared	0.213	0.315	0.319

Discussion....