first draft of a final project

[version: Tuesday 29th October, 2019 16:38]

due in 2 weeks (fall 2019: due in 3 weeks: we're behind)

1. produce a draft of your final project; reuse code from earlier, polish it–quality matters! make sure you have a paragraph or so saying what you are doing here and why, e.g. your research idea, initial findings, etc; something like an extended abstract; code has to be at least 400 lines of **tight code** (excluding comments and blank lines) that makes sense, no code padding, drop things that are not necessary; **this is important: i will penalize code that is just written to add lines!**; indeed, better fail the 400 lines minimum than generate boilerplate
2. have plenty of graphs! likewise use lavishly outreg2 and/or estout

for instance you may have something like that in your preamble

```
/* For Problem Set 4 I began to work on my final project by organizing and
simpliying my exisiting work.

I also began to see a clearer picture about the impact of environment (food
access, poverty) on health in NJ emerge through descriptive statistics.


yes, descripive stats is great for that!
-------------------------------------------------------------------------

great to talk about this upfront:

Research questions include the following:
1- Does inaccces to healthy food impact behavior and mental health?
2- Does increased green space decrease poverty and mental health ?
3- Who is most impacted by pollution (by race, gender, income)?
4- Do counties with higher pollution experience worse health outcomes (physical
and mental)?

-------------------------------------------------------------------------

and great to cite data, either here or when you load it!

My completed dataset includes data from the following sources:

1- NJ County Health Rankings Data
(http://www.countyhealthrankings.org/rankings/data/nj)

2- New Jersey Behavioral Risk Factor Survey, Center for Health Statistics, New
Jersey Department of Health Statistics, New Jersey State Health Assessment Data
(NJSHAD) (http://nj.gov/health/shad)

3- U.S. Census Bureau, 2016 American Community Survey 1-Year Estimates

4- Center for Disease Control and Prevention. Environmental Public Health
Tracking Network. Acute Toxic Substance Releases (www.cdc.gov/ephtracking) note:
I would love to have data from 2015 since most of my other datasets are from
this year but this was the most recent I could find. This would be good to
research further.

5- EPA Outdoor Air Quality Report
(https://www.epa.gov/outdoor-air-quality-data/air-quality-statistics-report)
```

6- Food Access and Research Center (FARC) and it is County SNAP (food stamp)
usage from 2011-2015 and simply shows the use of the Supplemental Nutrition
Assistance Program.

7- U.S. Census Bureau population counts by County for 2010-2016
(https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk)

---

general directions (always the same):

- i will show your code in class and possibly post some of your code or link to it–again, as per our core values–opensource, transparency, sharing; but if you'd like to keep your code private, that's fine–just let me know, and i will keep your code secret (no penalty, except that you may get little less feedback–usually if we discuss your code in the class, you will benefit from it!)
- you must submit all the code that was executed from the very beginning starting with the very raw data as per replication principle; unless data is too big to fit online, then just start with a comment, eg "to fit data online i had to take a random sample of 10perc"
- all ps are mostly cumulative–you can, and should, include much of previous code you've written for this class; can also use code you've written outside of this class (other classes, projects, etc)–but you have to clearly mark the code that has not been written for this class–otherwise, scholastic dishonesty!
- use your own dataset; again if you do not have a dataset, ask for help finding it
- because you are only submitting code, it must load data from Internet–just put your data onto your own website, wordpress, google drive, etc; (when you put data into any public space, try not to violate data copyrights... I haven't heard of anyone having problems with that, but be careful–for instance you may subset dataset to few vars and smaller sample using **sample**); and it is also easier to experiment on small datasets
- keep it simple! at the beginning of your dofile drop unnecessary vars; and even retain only certain, say most important, observations; keep it manageable; it is much easier to learn using simple data; can always complicate later!; much better to do it right using simple data than do it wrong using complex data!
- have nice structure in your file: sections, subsections, etc; may also have multiple files
- great idea to submit ps as early as possible–we will probably give you some comments; if not, email us and ask for comments!
- it is great to copy code from others; again, one of the rules for this class is 'be lazy': don't reinvent the wheel, whatever you are coding, it has already been done, google things often; but of course you cannot submit 100% code by someone's else.
- if you do something extra/fancy that is relevant and closely related to the assignment questions, it will be extra credit
- use coding rules that we've learned so far
- submit (only) the code into git repo; ps are due by the beginning of the next class unless indicated otherwise, eg "due in 2 weeks"; late ps are not accepted; NOTE: push to github early and send email to listerv with the link to your submission and ask for comments and ask any questions–the surest way to get the ps right!
- we are on the way to developing the final project with these ps: as we progress, your ps should start resembling a coherent and logical project where you use data management techniques to build new a dataset that can be used to answer interesting questions– say in few sentences (as a comment) why are you doing what you are doing–that is, answer the "so what question": "ok, you're gonna run all that code, and so what?" what's the goal of all that, why are you doing this? you need a compelling justification for what you are doing; typically: to develop a new dataset (that has not existed before) that can be used to answer some exciting questions: say what are those questions you want to answer; be brief, say couple sentences, and definitely not more than say 100 lines, typically 10-50 lines is enough; related: even at the beginning, already in ps1, say why you use data you are using, is it best, does it serve the purpose; also, feel free to ask me questions in comments
- be prepared do present your code in class (if time), just briefly, key points, couple minutes