data formats/conversion and manipulation

[version: Friday 21st September, 2018 08:31]

warning, this may be a heavy lift and may take time, again, start right away and ask questions

1

1. as always, load data from online–either from third party url or put it yourself on your website (eg google drive–directions below)
2. use your dataset–again if you do not have a dataset, email me (and possibly classmates): "Hi ! I would like to study ???, where can i find data?"
3. don't forget about preamble and comments
4. don't forget to cd to working directory and avoid using paths to files
5. write code that would read these data into Stata, and then save it in at least 2 different formats
6. have a good look at your data by doing basic descriptive statistics; it is very important to get familiar with your data

2

1. use at least one time each of the following : **`recode`** , **`replace`** , **`bys: egen`** ; for **`bys: egen`** –calculate some (eg median, sd, iqr, etc) group statistics that are interesting and make sense–briefly, say a sentence, interpret results in a comment

---

how to put data online?

using google drive:

- go to `drive.google.com`
- first upload the file, then right-click on it and select "Share...", go to "Advanced" at the bottom-right of the window and then click "Change..." and check "On - Public on the Web" and hit "Save" button and "Done" button
- then right-click the file again and "Get shareable link", and paste link into dofile; it should look like `https://drive.google.com/open?id=0B5Y56f52-YHrMEpQX2` and then copy the FILE_ID from it, ie everything that follows "id="
- and then paste that FILE_ID into `https://docs.google.com/uc?id=FILE_ID&export=download`
- so it would become `https://docs.google.com/uc?id=0B5Y56f52-YHrMEpQX2ZwVDV0QVE&export=download`
- in this example it's a .dta file so to load it, you'd say
  `use "https://docs.google.com/uc?id=0B5Y56f52-YHrMEpQX2ZwVDV0QVE&export=download"`

other ideas: may try RU website `https://oit-nb.rutgers.edu/service/publishing-world-wide-web`
and many of theirs:
`http://www.cloudwards.net/top-10-secure-dropbox-alternatives/`
`http://www.lifehack.org/articles/technology/running-out-room-dropbox-here-are-11-dropbox-alternatives-that-offer-way-more-free-cloud-sto`
`html`
`http://beebom.com/2015/03/best-dropbox-alternatives-for-cloud-storage`

---

tips/general comments on ps from past year

1. remember:

○ have preamble

○ cd, mkdir etc

- typically only one **`cd`** at the beginning
- and then no paths
- can check if runs at the library or apps.rutgers.edu

○ that it runs on your pc does not mean it will on mine!

○ again, the only thing i need to change (once!) is path

- it needs to run without any problems!
- I'll be giving very low grades if code breaks!

- keep it simple especially when learning new things!

○ much easier to figure things out

- say keep 5 vars and 50 obs:
- **`sample, 50 count`**
- **`keep Country GDPlat GDPqtr GDP11`**

- it's easier to figure things out with a small and handy data
- so not only simplicity in code but also in data is good
- ○ later, we'll complicate, but always try to simplify

- if you have questions on my comments on your ps
- do ask for clarification!
- i tend to be overly parsimonious...

- yes, you cannot overdo with comments
- ○ but super detailed comments are not necessary
- ○ the point is to put only the comments that are useful to you!
- no need to put comments about everything you do (unless this really helps you)

- always cite data!
- at a minimum say where exactly it come from, ie the url
- if ambiguous say which year, wave, version etc...

---

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
- 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, email us, stop by our offices, etc
- 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! 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 the Sakai's dropbox, or 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
- 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