Quantitative Data Analysis II

SOC 781

Course overview

Today we will…

- Cover syllabus and expectations
- Start reviewing data management

COURSE GRADING

Item	Points	Percent of grade	Cut-points for final grades
Pre-assessment	10	3%	A = 90% to 100%
Assignments	175 (35 each)	50%	B = 80% to 89%
Presentation	45	13%	C = 70% to 79%
Final paper	120	34%	D = 60% to 69%
Total	350	100%	E < 60%

Statistical software

- Course taught in Stata
  - Comfort level with Stata?
- Stata access issues?
  - Any other computing issues?
    - e.g., inside and outside of class
- MUST be done in Stata, but
  - I’m open to helping you learn R

Textbooks and other readings

- *Regression Models for Categorical Dependent Variables Using Stata (3rd ed.)* J. S. Long & J. Freese
- *Generalized Linear Models: An Applied Approach*. J. P. Hoffmann
- Complete assigned reading BEFORE class
- Other resources: UCLA

Assignment 1

General overview

A. Select a topic that you want to study through secondary data analysis

B. Identify and download survey data that can be used to examine your topic

C. Begin cleaning your data (for this assignment, include at least five variables)

D. Compute and table descriptive statistics

E. Describe descriptive statistics table in text

- Assignments build on one another up to final project

- You have two weeks to identify, download, and become familiar with data
- No data exists on exact interest, then recommend scouring the GSS for something relatable
- Consult with me about data
  - More later…

Assignment 2

General overview

A. Build and describe an OLS model

B. Describe the regression diagnostics you performed and whether this model is ideal

C. Table OLS results

D. Outline the results in a professional journal-style format

Assignment 3

General overview

A. State two testable hypotheses: one to assess moderation and another to assess mediation

B. Conduct moderation and mediation tests and table results

C. Describe the results

Assignment 4

General overview

A. Examine a dichotomous DV using a logit and a probit model

B. Compare the logit and probit results

C. Select the most appropriate results, and describe them

Assignment 5

General overview

A. Examine an ordinal DV using an OLS model and an ordered logit model

B. Compare the OLS and ordered logit results

Presentation

This presentation exercise serves two purposes: First, conference presentations typically involve papers that are in progress, so this will help get you used to presenting research that is not yet complete. Moreover, being prepared to present your final paper will help keep you on track and prevent you from starting your paper at the last minute. Second, the process of writing research papers is repetitive and involves multiple revisions that are responsive to feedback, such as the feedback you may receive after presenting at a conference. There will be time for Q&A after each presentation, as well as discussion and consultation after all the presentations are complete. This will provide you the opportunity to learn how to receive and use feedback, as well as how to provide constructive feedback. Your presentation should be 15-minutes (not counting Q & A) and follow a conference style. Guidelines are provided below:

- Introduction & Background
- Data & Methods
- Results
- Discussion & Limitations

Final paper

This project was designed for you to demonstrate your knowledge of GLM techniques and to hone your academic writing skills, and for me to assess your achievement of the course learning goals and objectives. There is no required length (minimum or maximum), and no required format but try to adhere to a format commonly used in your subfield. Feel free to copy and paste some from your assignments (this is not self-plagiarism in this case), but your final paper should be self-contained and cohesive. I don't need your .do file but save it in case we (or you) need to reference it later (it should be readily available if I request it). This paper should be written as if you were writing a journal article.

- Introduction & Background
- Data & Methods
- Results
- Discussion & Limitations

Data

- Use data that you're interested in
  - existing dataset with at least 5 variables of interest
  - outcome can be manipulated into various levels of measurement, and/or
  - more than one outcome measure (binary, ordinal, nominal, and count)

- Unit of analysis = individual (at least a few hundred)
  - we will not deal with repeated observations

- Final project: analyses must be GLM technique

- Consult with me as you choose your data

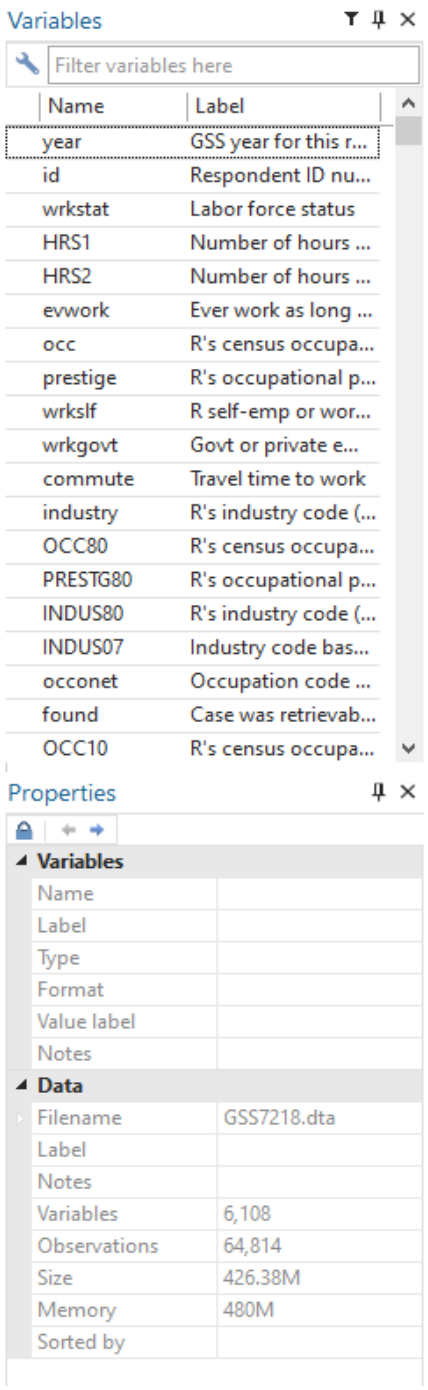
Data management

- Open Stata 781_8.do

- Read in data and let me know when ready or if any issues
  - don’t forget to change the file extension

Example data

- The GSS is a nationally representative cross-sectional survey conducted biannually since 1972 with approximately 2,000 respondents each wave
- We will be working with the cumulative waves 1972-2018
- How many variables?
- How many observations?
  - What is an observation?



DV and IV: for many examples

- Happiness: Taken all together, how would you say things are these days—would you say that you are (1) very happy, (2) pretty happy, or (3) not too happy

- This may be difficult to examine when focused on various GLM techniques (e.g., binary, categorical, count). You will come across similar issues in your assignments. Be prepared to manipulate your DV, examine different outcomes and consider if your DV could be an IV.

- Age: Measured in years from 18 to 89-plus

Data cleaning and management

- Save original dataset and documentation
  - somewhere that it will NOT be corrupted

- Clearly document .do files
  - initial .do file to clean data and ID analytic sample
    - document missing
  - save analytic dataset
    - often faster than working w/ original dataset
  - create new .do file for analysis
    - less worries about corrupting analytic sample

Rename and recode

- Create new variables rather than recode original
  - meaningful name and coding scheme
    - ordinal: corresponding direction (e.g., negative-positive)
    - binary [0,1]: group 1 (e.g., 0=male, 1=female then var=female)

- Make sure to deal with non-response
  - don’t know; no answer; not sure; refused; etc.
  - and common numeric denotations (e.g., 99, 999, 9999)
    - also consider not applicable/inapplicable

Rename and recode: Happiness

- One of the first things we want to do is see what our variables look like
  - try a crosstab
    - note difference between “tab *var*” and “tab *var*, m”

```
. tab happy
```

General happiness	Freq.	Percent	Cum.
VERY HAPPY	18,823	31.34	31.34
PRETTY HAPPY	33,563	55.89	87.23
NOT TOO HAPPY	7,668	12.77	100.00
Total	60,054	100.00	

```
. tab happy, m
```

General happiness	Freq.	Percent	Cum.
VERY HAPPY	18,823	29.04	29.04
PRETTY HAPPY	33,563	51.78	80.83
NOT TOO HAPPY	7,668	11.83	92.66
DK	39	0.06	92.72
IAP	4,383	6.76	99.48
NA	338	0.52	100.00
Total	64,814	100.00	

- Inapplicable is NOT considered missing

Rename and recode: Happiness

- Let's put this ordinal variable in an intuitive direction

```
/*recode in positive direction*/
gen hap=.
replace hap=1 if happy==3 /*not so happy*/
replace hap=2 if happy==2 /*pretty happy*/
replace hap=3 if happy==1 /*very   happy*/
/*check recode: crosstab w/ original*/
*tab hap happy,m
```

```
. tab hap happy
```

hap	General happiness			Total
	VERY HAPP	PRETTY HA	NOT TOO H	
1	0	0	7,668	7,668
2	0	33,563	0	33,563
3	18,823	0	0	18,823
Total	18,823	33,563	7,668	60,054

- Check recoding with original variable
  - this is one reason to create new and not recode original variable

Rename and recode: Happiness (0,1)

- We can also dichotomize this ordinal measure

```
/*dichotomize happiness (very/pretty vs. not so happy)*/
gen hap_dic=.
replace hap_dic=0 if hap==1
replace hap_dic=1 if hap==2 | hap==3
/*check recode*/
*tab hap_dic hap,m
```

```
. tab hap_dic hap
```

hap_dic	hap 1	2	3	Total
0	7,668	0	0	7,668
1	0	33,563	18,823	52,386
Total	7,668	33,563	18,823	60,054

- Must have good justification to manipulate any variable
  - in our case, to meet assignment requirements it's okay

Rename and recode: Labels

- We can add variable and value labels

```
/*variable labels*/
label variable hap "Recode original happiness variable in positive direction"
/*value labels*/
label define happy 1 nottoohappy 2 prettyhappy 3 veryhappy
label values hap happy
tab hap happy
```

```
. tab hap happy
```

Recode original happiness variable in positive direction	General happiness			
	VERY HAPP	PRETTY HA	NOT TOO H	Total
nottoohappy	0	0	7,668	7,668
prettyhappy	0	33,563	0	33,563
veryhappy	18,823	0	0	18,823
Total	18,823	33,563	7,668	60,054

- If intuitive and good notes, then no need. I like to see the values rather than the labels without having to use an extra command.

Rename and recode: Age

- Age already has a meaningful name and intuitive coding
  - so, let's keep it and not manipulate the coding other than missing data
- But we may want to make some polynomial transformations
  - Why?

```
replace age=. if age>89
/*polynomial transformations: capture non-linear patterns*/
/*quadratic (one inflection point)*/
gen age2=age*age
/*cubic (two inflection points)*/
gen age3=age*age*age
/*quartic (three inflection points)*/
gen age4=age*age*age*age
```

- Make sure to recode any missing values to "." first. Sometimes Stata can treat other codes as a value, especially when computing a new variable

Recode the rest of the variables

- Often original measurement is best, but sometimes it’s not
  - for either statistical or interpretive reasons

- Make sure you have detailed notes so when you look back later you know what you’re working with

Missing data

- There is no consensus surrounding missing data
- Generally, NOT suggested to impute DV or IV
  - we will NOT deal with imputation
- Missing on other covariates may cause sample reduction
- General rule of thumb 20% threshold
- Descriptive statistics useful for identifying patterns of missingness
  - when reporting, make sure same N (analytic sample)

Missing data

- Identify analytic sample
  - all respondents w/o missing values on any variable in ALL models
  - keep track in .do file

- Easiest, and arguably best, solution: listwise deletion
  - either drop those with any missing values or create an indicator

Missing data: indicator example

```
**********************************Missing Data*************************************
/*Generate variable to store information on missing data*/
egen nmiss=rmiss(hap age female nonwhite educ married)
*tab nmiss
/*What's missing? Tab through each variable to figure out*/
*keep if nmiss==0 /*drops only those w/ any missing information on listed vars*/
*sum hap female nonwhite if nmiss==0 /*comand to include only analytic sample*/
/*if you don't limit to analytic sample when computing descriptive statistics
then you will obtain different Ns for each variable w/ missing data*/
```

```
. tab nmiss
```

nmiss	Freq.	Percent	Cum.
0	59,725	92.15	92.15
1	5,022	7.75	99.90
2	58	0.09	99.99
3	9	0.01	100.00
Total	64,814	100.00	

Descriptive statistics

- Central tendency and dispersion
  - mean, median, mode
    - reflects “center” value, or “typical” case in the distribution
  - range, variance, standard deviation
    - reflects the spread around the center

- Typically report mean and SD for those that are or treated as continuous
  - percentages in each category for others

Descriptive statistics

- The summary command is very useful

```
. sum hap age female if nmiss==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hap	59,725	2.185835	.6373977	1	3
age	59,725	46.04745	17.5749	18	89
female	59,725	.5577899	.4966533	0	1

- Provide clean and informative tables in assignments
  - NOT raw output

Table 1a. Summary of the analytic sample: GSS 1972-2018 (n = 59,725)

	Mean	SD	Range
Happy	2.19	0.64	1-3
Age	46.05	17.57	18-89+
Female	0.56		0-1

Descriptive statistics

- Consider whether the statistics you're reporting are informative

```
. sum i.hap age female if nmiss==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hap					
2	59,725	.5591963	.4964876	0	1
3	59,725	.3133194	.4638469	0	1
age	59,725	46.04745	17.5749	18	89
female	59,725	.5577899	.4966533	0	1

Table 1b. Summary of the analytic sample: GSS 1972-2018 (n = 59,725)

	Mean	SD	Range
Happy			
Very	0.31		0-1
Pretty	0.56		0-1
Not too	0.13		0-1
Age	46.05	17.57	18-89+
Female	0.56		0-1

Descriptive statistics

- You may want information for specific subgroups (e.g., sex)
  - crosstabs are useful

```
. tab female hap if nmiss==0
```

female	hap 1	2	3	Total
0	3,339	14,970	8,102	26,411
1	4,275	18,428	10,611	33,314
Total	7,614	33,398	18,713	59,725

```
tab female hap if nmiss==0, ro co cell /* percentage options*/
```

female	hap 1	2	3	Total
0	3,339	14,970	8,102	26,411
	12.64	56.68	30.68	100.00
	43.85	44.82	43.30	44.22
	5.59	25.06	13.57	44.22
1	4,275	18,428	10,611	33,314
	12.83	55.32	31.85	100.00
	56.15	55.18	56.70	55.78
	7.16	30.85	17.77	55.78
Total	7,614	33,398	18,713	59,725
	12.75	55.92	31.33	100.00
	100.00	100.00	100.00	100.00
	12.75	55.92	31.33	100.00

- row, column, and cell percentages are informative
  - Who's happier, females or males?

Descriptive statistics

Group differences	Nominal	Ordinal	Interval	Ratio
t-test (grouping var has 2 categories) compares means	No	No	Yes	Yes
ANOVA (grouping var has 3+ categories) compares variances	No	No	Yes	Yes
Crosstab and chi-square Compare probability of sets of responses	Yes	Yes	In theory	In theory

- These are common techniques, but many others based on measurement of each variable: resource

Descriptive statistics

- Types of correlation

- Pearson's: continuous-continuous
- Polychoric or Spearman: ordinal-ordinal
- Kendall's: continuous-ordinal
- Point-biserial: continuous-nominal
- Rank-biserial: ordinal-nominal
- Chi-square: nominal-nominal

- Some may require special packages in Stata

Descriptive statistics

- Pearson's most common: assumes
  - Both measures are continuous
  - Linear relationship
  - No major outliers
  - Approximately normally distributed

Coefficient Value	Strength of Association
$0.1 < \lvert r \rvert < .3$	small correlation
$0.3 < \lvert r \rvert < .5$	medium/moderate correlation
$\lvert r \rvert > .5$	large/strong correlation

Descriptive statistics

- Often see correlation matrix. Sometimes with appropriate tests, but often just with Pearson's. Not huge fan - but starting point

```
. cor hap age female nonwhite educ married if nmiss==0
(obs=59,725)
```

	hap	age	female	nonwhite	educ	married
hap	1.0000					
age	0.0265	1.0000				
female	0.0077	0.0347	1.0000			
nonwhite	-0.1054	-0.0964	0.0344	1.0000		
educ	0.0935	-0.1809	-0.0355	-0.0963	1.0000	
married	0.2366	0.0348	-0.0739	-0.1508	0.0294	1.0000

```
. pwcorr hap age female nonwhite educ married if nmiss==0, sig /*need "pwcorr"
> for ",sig" option*/
```

	hap	age	female	nonwhite	educ	married
hap	1.0000					
age	0.0265 0.0000	1.0000				
female	0.0077 0.0607	0.0347 0.0000	1.0000			
nonwhite	-0.1054 0.0000	-0.0964 0.0000	0.0344 0.0000	1.0000		
educ	0.0935 0.0000	-0.1809 0.0000	-0.0355 0.0000	-0.0963 0.0000	1.0000	
married	0.2366 0.0000	0.0348 0.0000	-0.0739 0.0000	-0.1508 0.0000	0.0294 0.0000	1.0000

- Need pwcorr command for significance levels

Next class we will…

- start working on Assignment 1,
  - must have your own data to do this
  - check in with me if any questions or concerns
- read Long & Freese CH 2 before class

- Download R and RStudio before next class
  - if haven't yet