STATA WORKSHOP

ERL Workshop for Sociology Fall 2014

WELCOME TO STATA!

- What are we looking at?
- Opening the data

- Viewing the data
  - Edit vs Browse

- **** Log files ****
  - File → Log → Begin
  - File Name→ whatever makes sense to you
  - **Save as type: Log** *so that you can review the output on a computer without Stata

- Do files
  - For saving a list of commands that you'll use repeatedly

THE GSS: IDENTIFY VARIABLES

- Navigating Codebook
  - Index
  - Codebook
- Can also use lookfor in Stata
  - Say you are interested in the relationship between education and spending money on space exploration
  - *lookfor education* *doesn't give us the respondent's education- what are synonyms?
  - *lookfor school* *highest year of school completed→EDUC; also math and physics courses
- Find variables that are continuous, categorical (nominal), and ordinal
- For your DV, choose variables that can be changed by something
  - age is NOT a good DV because, e.g. the amount of time you spend watching TV isn't going to make you older!

VARIABLES I'LL BE USING

- WRKBABY: "do you think women should work full-time, part-time, or stay at home when they have a child under school age?"
- RHHWORK: "On average, how many hours a week do you personally spend on household work, not including childcare and leisure time activities?"
- SEX: gender (male, female)
- INCOM16: family's income when respondent was 16 (below average, average, above average)
- AGE

UNDERSTANDING YOUR VARIABLES

- Stata uses numbers for categories
  - WRKBABY:
    - 1 = Full-time
    - 2 = Part-time
    - 3 = Stay home
  - SEX
    - 1 = male
    - 2 = female
- But the numbers aren't really meaningful other than to mark the categories
- It can be useful to assign increasing numbers to an ordinal variable

UNIVARIATE STATISTICS

- Useful for understanding your data

- Are there more men or women in the sample?
- Does my sample consist of people with money?
- How old are the people in my sample?
- How much time do they spend on housework?
- How do they feel about women with children working?

UNIVARIATE STATISTICS IN STATA

FREQUENCIES AND DISTRIBUTIONS

- *tabulate SEX* *for percentages of male & female (frequency of each)
  - Gives total respondents

- *tabulate WRKBABY* *for percentages of responses (frequency of each)

- *tabulate INCOM16* *for percentages of income level (what does this mean?)

- *histogram AGE* *for distribution of age

- *histogram RHHWORK* *for distribution of hours spent on housework

LABELING VARIABLES IN STATA

- WRKBABY output
  - 1 is not labeled and can be difficult to interpret

- Stata commands for labeling categories of a variable
  - *label define wrkbabylabel 1 "work full-time" 2 "work part-time" 3 "stay home“*
  - *label values wrkbaby wrkbabylabel*

UNIVARIATE STATISTICS IN STATA

MEAN, MEDIAN, & MODE (MEASURES OF CENTRAL TENDENCY)

- Mode :
  - tabulate WRKBABY, sort
  - tabulate RHHWORK, sort
- Median :
  - summarize INCOM16, detail
  - summarize RHHWORK, detail
- Mean :
  - summarize AGE
  - summarize RHHWORK
- Can all measures of central tendency be calculated for all variables?

UNIVARIATE STATISTICS IN STATA

RANGE, STANDARD DEVIATION (MEASURES OF VARIABILITY)

- Range :
  - Min-max : summarize RHHWORK
  - Deciles or Quartiles : summarize AGE, detail
- Standard Deviation :
  - summarize RHHWORK, detail

TO SUMMARIZE

OR NOT TO SUMMARIZE… THAT IS THE QUESTION

- **summarize** gives you:
  - Median (if you use ‘detail’ argument)
  - Mean
  - Range (deciles & quartiles if you use ‘detail’ argument)
  - Standard deviation
- **tabulate** gives you:
  - Mode (using ‘sort’ argument can help especially with continuous variables)
  - Frequency tables with percentages
- **histogram** give you:
  - A histogram- can help you see the distribution of continuous variables.

BIVARIATE STATISTICS IN STATA

QUESTIONS TO ANALYZE

- Does gender affect people's opinion about women working when they have preschool children?
  - Compare SEX and WRKBABY
  - Categorical or continuous?
- Does age affect hours spent on housework?
  - Compare AGE and RHHWORK
  - Categorical or continuous?
- Does one's income growing up affect hours spent on housework?
  - Compare INCOM16 and RHHWORK
  - Categorical or continuous?

CHI-SQUARE STATISTIC IN STATA

DOES **GENDER** AFFECT PEOPLE'S **OPINION** ABOUT WOMEN WITH PRESCHOOL CHILDREN WORKING?

- Two categorical variables- which is IV & which is DV?
- Need percentages of each response of DV at each level of IV
  - i.e. need percentages for full time, part time, and stay home for each gender
  - The percentages help you 'see' the data

- *tabulate wrkbaby sex, column*
  - OR *tabulate sex wrkbaby, row*

- Chi-square statistics
  - *tabulate wrkbaby sex, chi2*

CORRELATION COEFFICIENT IN STATA

DOES **AGE** AFFECT **HOURS SPENT ON HOUSEWORK**?

- Two continuous variables- IV & DV?

- *scatter rhhwork age*
  - Seeing the data can help you understand the relationship

- *pwcorr rhhwork age, sig*
  - Look at the table where the variables of interest cross
  - First number is r, second is p (if you forget argument 'sig' you will not get the p)
  - 1.000 shows that the variable has a strong correlation with itself (duh!)

ONE-WAY IN STATA

DOES **ONE'S INCOME GROWING UP** AFFECT **HOURS SPENT ON HOUSEWORK**?

- Categorical IV & continuous DV*
  - Oneway ANOVAs should always have a continuous DV

- *oneway rhhwork incom16, means*
  - The means help you 'see' the data

- If you want to see it:
  - *graph bar (mean) rhhwork, over(incom16)*

SUMMARY OF BIVARIATE STATISTICS IN STATA

- Does gender affect people‘s opinion about women with preschool children working?
  - *tabulate DV IV, column*
  - *tabulate DV IV, chi2*

- Does age affect hours spent on housework?
  - *pwcorr DV IV, sig*
  - *scatter DV IV*

- Does one’s income growing up affect hours spent on housework?
  - *oneway DV IV, means*
  - *graph bar (mean) DV, over(IV)*

MULTIVARIATE STAISTICS IN STATA

- Statistical control OR 3rd variable problem (confound or interaction effects)

QUESTIONS TO ANALYZE

- Does gender affect people's opinion about women with preschool children working?
  - Evidence shows that gender roles differ by cultural background; how does race impact gendered opinion?

- Does age affect hours spent on housework?
  - Whether or not have children may also play a role on hours spent on housework

- Does one's income growing up affect hours spent on housework?
  - Since housework is often seen as a gendered role, how does gender impact the relationship between income growing up & hours spent on housework?

CHI-SQUARE STATISTIC IN STATA

Original Question: DOES **GENDER** AFFECT PEOPLE'S **OPINION** ABOUT WOMEN WITH PRESCHOOL CHILDREN WORKING?

Evidence shows that gender roles differ by cultural background; how does RACE impact gendered opinion?

- *tabulate race* *since we haven't seen it before, good to get the univariate stats

- *bysort race: tabulate wrkbaby sex, column chi2*
  - Can combine arguments in Stata (column w chi2)
  - *bysort:* sorts by race

CORRELATION COEFFICIENT IN STATA

Original Question: DOES **AGE** AFFECT **HOURS SPENT ON HOUSEWORK**?

- Whether or not one has children may also play a role on hours spent on housework
- BABIES- number of household members less than 6 years old (0, 1, 2, 3, 4)
  - Continuous or categorical?
  - *summarize babies* AND *tabulate babies* * to get an understanding of the variable

- Recode BABIES into 2 groups; people with children under 6 and people without
  - *recode babies (0 = 0) (1/4 =1), generate(babies2)*
  - generate: creates a new variable which is super duper important!!

  NEVER WRITE OVER YOUR DATA!
  - (1/4) = 1 through 4 & includes all categories between 1 & 4 and replaces them with 1

CORRELATION COEFFICIENT IN STATA

Original Question: DOES **AGE** AFFECT **HOURS SPENT ON HOUSEWORK**?

- Whether or not one has children may also play a role on hours spent on housework
  - You may want to label babies2

- *bysort babies2: pwcorr rhhwork age, sig*
  - 2 tables: babies (1), no babies (0)

- *scatter rhhwork age if babies2 == 0*
  - To see the relationship between rhhwork & age when there are no babies (children under 6) in the household
  - We'll talk more about the if statement in a later slide

ONE-WAY IN STATA

DOES **ONE'S INCOME GROWING UP** AFFECT **HOURS SPENT ON HOUSEWORK**?

- Since housework is often seen as a gendered role, how does SEX impact the relationship between income growing up & hours spent on housework?

- *bysort sex: oneway rhhwork incom16, means*

- If you want to see it:
  - *graph bar (mean) rhhwork, over(incom16) over(sex)*

REGRESSION IN STATA

DOES **AGE** AFFECT **HOURS SPENT ON HOUSEWORK**?

- Income might also play a role in hours spent on housework
- INCOME – 10 categories of income (range: less than $1000 to $25000 or more)
  - *tabulate income*

- *regress rhhwork age income*

IF STATEMENT IN STATA

IF STATEMENT → IF SOMETHING IS TRUE, DO SOMETHING

- *scatter rhhwork age if babies2 == 0*
  - Displayed scatterplot for relevant variables for when the sample had no children aged 6 or younger

- IF syntax
  - > for greater than
  - >= for greater than or equals
  - == for equals
  - & for and (as in this and that)
  - > for greater than
  - >= for greater than or equals
  - != for not equals
  - | for or (as in this or that)
- Examples
  - *summarize educ if race !=1 & sex == 2* **(education for non-white females)

RECODING VARIABLES IN STATA

- Sometimes you may want to look at different categories than provided in the dataset, like with the BABIES variable.

- Another example, maybe you are interested in the impact of at least a bachelor's degree on something. You could recode DEGREE into 2 categories (educational degrees that are lower than a bachelor's & degrees that are greater than or equal to a bachelor's)

- *tabulate degree*
- *recode degree (0/2 = 0) (3/4 =1), generate(atleastba)*
  - You can assign multiple categories (0-2) to one new group (0).
- *tabulate atleastba*

RECODING VARIABLES IN STATA

- Sometimes it is easier to interpret findings when the numbers under the labels increase in a way that makes sense.
- For example HAPPY measures general represented like this in the GSS2012
  - Very happy = 1
  - Pretty happy = 2
  - Not too happy =3
- *pwcorr happy age* * do you get happier as you get older?
  - Even though we are phrasing the question with increases (increase in age, increase in happiness), we'd actually be looking for a negative correlation (the larger the number for happiness, the less happy one is).
- *recode happy (1 = 3) (2 = 2) (3 = 1), generate(happyInc)*
  - Generating a new variable (happyInc) ensures you don't lose your data if you make a mistake (what if I accidently typed 3=3, but didn't generate a new variable?!)

CLOSE YOUR LOG FILE!!!!

When you are done running all of your tests, you need to close your log file.

- *log close*

OR in the menu

- File → Log → Close

Go to your log file and open it. It will open in Notepad and now you can view all of your results from the session!!!

This will be VERY helpful for you when you write up your assignment. ☺

SUMMARY OF MULTIVARIATE STATISTICS & EXTRAS

MULTIVARIATE STATISICS

- Sort by 3rd variable
  - *bysort 3RD VARIABLE:* *Stata test and IV, DV
- Regression
  - *regress DV IV IV*

EXTRAS

- Recoding
  - *recode variable (x = y) (a/c = z), generate newvariable*
    - where a, b, c, x, y, & z represent actual numbers.
- If statement
  - *if babies2 == 0*
- Label variable levels
  - *label define labelvariable 1 “label 1” 2 “label 2”*
  - *label values variablename labelvariable*

THANKS!

Questions about Stata can always be asked at the ERL during our walk-in hours.

- Mon 10-8
- Tues 10-8
- Wed 10-8
- Thurs 10-6
- Fri 10-6
- Sun 1-7
- ERL.barnard.edu/calendar