# Intermediate Stata Workshop

Hsueh-Sheng Wu

CFDR Workshop Series

Spring 2009

BGSU

Center for Family and Demographic Research

# Outline of the Workshop

- A brief review of topics covered in the Introduction to Stata workshop (Fall 2008)
- Commands for data management
- Commands for variable construction
- Commands for statistical analyses
- Helpful tips

# Topics Covered in the Last Stata Workshop

- Four window screens
- Opening data
- Basic commands
  - *describe*
  - *summarize*
  - *tabulate gender*
  - *codebook gender*
  - *generate income2 = income +100*
  - *sort gender*

# Topics Covered in the Last Stata Workshop

- Operators
  - Examples
    - \>     Greater than
    - \<     Less than
    - = =   Is or is equal to (equality symbol)
    - \> =   Greater than or equal to
    - \< =   Less than or equal to
    - &     And
    - |     Or
    - ~     Not
  - Symbols only, no mnemonic equivalents
  - Symbols need to be in correct order
    - >= NOT =>

# Topics Covered in the Last Stata Workshop

- "do" file – Stata command file (.do)
  - Write and save commands into a text file
- "log" file – Stata output file (.log)
  - log using c:\temp\stata.log,replace
  - log close
- Command structure
  - No punctuation at end of command
  - Case sensitive

# Topics Covered in the Last Stata Workshop

- Sources for help
  - http://www.stata.com
  - http://www.ats.ucla.edu/stat/stata/
  - http://www.princeton.edu/~erp/stata/main.html
  - http://library.syr.edu/information/mgi/nds/software/stata/intro/index.html
  - Acock, Alan C. 2006. *A Gentle Introduction to Stata.* Stata Press.

# Commands for Data Management

- *list* – list values of variables
  - *list* name
  - *list* name id
- *if* – a qualifier
  - *list* name *if* educ == 1
  - *list* name *if* educ ~=1
  - *list* name *if* educ == 1 & income >= 40000
  - *list* name *if* capital ==.
  - *list* name *if* county ==""

# Commands for Data Management

- *in* – a qualifier
  - *list* name *in* f/10 (list the 1$^{st}$ case through the 10$^{th}$ case)
  - *list* name *in* 10/l (list the 10$^{th}$ case through the last case)
  - *list* name *in* 5/10 (list the 5$^{th}$ case through the 10$^{th}$ case)

# Commands for Data Management

- *recode* – change values of numerical variables
  - *recode* educ (1 = 2)
  - *recode* educ (1 = .)
  - *recode* educ (1 = 2) *if* gender == 1
  - *recode* educ (1 = 2) *if* name == "Amy"
  - *recode* educ (1 = 4) (3 = 5), *gen*(educ2)

# Commands for Data Management

- *replace* – change the contents of existing string or numerical variables
    - *replace* educ = 2
    - *replace* educ = 2 *if* id == 10
    - *replace* educ = 2 *if* name == "David"
    - *replace* educ =.
    - *replace* county = "wood"
    - *replace* county = "wood" *if* id == 10
    - *replace* county= "wood" *if* name == "David"
    - *replace* county = ""

# Commands for Data Management

- *keep* and *drop* –
  - *keep* id gender educ income
  - *drop* income

  - *keep* if gender == 1
  - *drop* if gender  == 2

- dataset 1:  id year month day state county
- dataset 2:  id name gender educ employed income capital assets cars

# Commands for Data Management

- *merge* – merge datasets
  - *use* dataset2
  - *sort* id
  - *save* dataset2, *replace*
  - *use* dataset 1, *clear*
  - *sort* id
  - *merge* id *using* dataset2
  - *tab* _merge

# Commands for Variable Construction

- *generate* – creates new variable
    - *generate* logincome = log(income)
    - *generate* wealth = income + capital + assets
    - *generate* newid = state + county
    - *generate* newid2 =state + " " +county

# Commands for Variable Construction

- *egen* – extensions to generate variables
  - *egen* mean_inc = *mean*(income)
  - *egen* med_inc = *median*(income)
  - *egen* std_inc = *std*(income)
  - *egen* total_inc = *rowtotal*(income capital)
  - *egen* miss_inc = *rowmiss*(income capital)
  - egen nmiss_inc = *rownonmiss*(income capital)

# Commands for Statistical Analyses

- Descriptive statistics
  - *summarize name income*
  - *tab1 name income*
  - *tab2 name income*
- Continuous dependent variables
  - *ttest* income*, by (*gender)
  - *pwcorr* income capital assets
  - *anova* income gender educ
  - *regress* income gender educ

# Commands for Statistical Analyses

- Categorical dependent variables
  - logistic regression
    - *employed* educ gender
  - ordered logistic regression
    - ologit educ income
  - Multinomial logistic regression
    - *mlogit* newcar income

  - Predicted values – obtain predicted values after a model is estimated
  - *regress* income gender educ
  - *predict* phat
  - *predict* resid, *residuals*
  - *sum* phat resid

# Helpful Tips

- help and search
    - *help* tabulate
    - *search* tabulate
    - *search* rc 198
- *set memory* 100m
- *set maxvar* 6000
- *set more on* and *set more off*
- *exit* – exit from a program or do-file

# Helpful Tips

- *xi* – interaction expansion
  - *xi*: *regress* income  educ
  - *xi*: *regress* income i.educ
  - *xi*: *regress* income i.educ gender