Introduction To STATA Part II

Adrian Rohit Dass

January 19th, 2024

Recap: STATA Part I Session

- Why use STATA?
- Reading/Cleaning data
- Regression Analysis
- Post-estimation Diagnostic Checks
- Other Topics in STATA
- Applied Example
- STATA Resources

STATA for Health Economics

- A survey conducted to IHPME health economics students in late 2021 suggested the following research interests
  - Working with data
    - Common tasks: reading in data, creating new variables, data subsets, etc.
  - Applied econometrics
    - Common tasks: descriptive analysis, regression analysis, etc.
  - Economic Evaluation
    - Common tasks: model building (Markov, Microsim, etc.), sensitivity analysis, etc.

Outline

- Working with matrices in STATA using mata
- Programming in STATA
  - Working with matrices, data, applied econometrics etc.
  - Applied examples throughout
- Additional STATA resources

mata

mata (continued)

What is mata?

“From STATA manual: Mata is a matrix programming language that can be used by those who want to perform matrix calculations interactively and by those who want to add new features to Stata.”

Source: https://www.stata.com/manuals/m.pdf

To start and stop a mata session:

```
. mata //start mata session
/*Insert STATA mata matrix commands here*/
end // end mata session
```

mata Example Functions

- Create general nxk matrix (named A) with same value across rows and columns
  A = J(n,k,val)
- Create matrix of any dimensions manually (named B)
  B = (0.95,0.05\0,1)
- Extract i*th* row
  A[i,]
- Extract jth column
  A[,j]
- Matrix multiplication
  *
- Element wise multiplication
  :*

For more commands, please see https://www.stata.com/manuals/m.pdf

Applied Example

Markov model with the following transitional probability matrix

H	S	D
0.9	0.08	0.02
0	0.8	0.2
0	0	1

Everybody in the model starts in H

Can we use mata and matrix algebra to solve for the second period health states?

Code for applied example

```
. mata

mata clear

A = J(2,3,0) // 2X3 matrix of 0s (Two time periods, 3 health states)
P = 0.9, 0.08, 0.02\0,0.8, 0.2\0,0,1 //3x3 transitional probability matrix

A[1,] = 1, 0, 0 //Initial health states

A[2,] = A[1,] * P //Health states in period 2

A //Display entire matrix

end
```

Looping in STATA

forvalues command

forvalues *lname = range* {

Stata commands referring to ‘lname’

}

Source: https://www.stata.com/manuals13/pforvalues.pdf

Example:

```
forvalues i = 1/10 {
        display `i'
}
```

Applied Example

- The code for the Markov model would be difficult to use for multiple time periods
- Would need multiple lines of mata code:
  - A[3,] = A[2,] * P
  - A[4,] = A[3,] * P
  - …and so on
- Can we use forvalues function to loop instead?

Code for applied example

```
clear

. mata

mata clear

A = J(10,3,0) // 10X3 matrix of 0s (Ten time periods, 3 health states)
P = 0.9, 0.08, 0.02\0,0.8, 0.2\0,0,1 //3x3 transitional probability matrix

A[1,] = 1, 0, 0 //Initial health states

A //Display entire matrix

end

forvalues i = 2/10 {
                        mata: A[`i',] = A[`i' - 1,] * P //Single line version of mata
}

mata: round(A, 0.01)
```

Macros

- In STATA, a macro is a string of characters that stands for another string of characters (Camerson & Trivedi, 2021)
- Leads to code that is shorter, easier to read, and can be adapted to similar problems
- Macros can be global or local
  - Global: accessible across STATA do-files or throughout a STATA session
    - Ex: Variable list that is required across entire analysis
  - Local: Can be accessed only within given do-file or in the interactive session
    - Particularly useful for programming

Applied Example

- Analysis of Health Expenditure Data in Jones et al. (2013) *Chapter Three*
- The data covers the medical expenditures of US citizens aged 65 years and older who qualify for health care under Medicare.
  - Outcome of interest is total annual health care expenditures (measured in US dollars).
  - Other key variables are age, gender, household income, supplementary insurance status (insurance beyond Medicare), physical and activity limitations and the total number of chronic conditions.
  - Can we use macros to help analyze?
- Data can be downloaded from here (mus03data.dta): https://www.stata-press.com/data/musr.html

Code for applied example

```
log using "mylogfile.smcl", replace //start log file
clear //remove variables from STATA
 use "mus03data.dta" //Load Data
global xvars age female income suppins phylim actlim totchr
global xvarssub female income suppins phylim actlim totchr
drop if posexp==0 //Remove individuals with $0 in health expenditures (following example)
*Regression*
reg totexp $xvarssub //Regression without age
eststo reg1 //Store results
reg totexp $xvars //Regression with age (following example)
eststo reg2 //Store results
esttab reg1 reg2 using "Results/myresults.csv", cells(b(fmt(3)star) se(par)) stats (N r2) replace //export results
*Robust regression*
reg totexp $xvarssub, robust //Regression without age, HC robust
eststo robust1 //Store results
 reg totexp $xvars, robust //Regression with age (following example), HC robust
eststo robust2 //Store results
esttab robust1 robust2 using "Results/myresultsrobust.csv", cells(b(fmt(3)star) se(par)) stats (N r2) replace //export results
```

Program in STATA

- Creating a program in STATA allows for the creation of a custom-made command
- This command can be used to call for the running of lines of STATA code to produce an output
- The program can take inputs, but is not necessary to do so
- For more information, please see here https://www.stata.com/manuals/u18.pdf

Program Example

- Instrumental Variable (IV) estimation is typically estimated using two-stage least squares (2SLS), which uses a linear function in both stages
- In many health applications, the second stage is non-linear
- The control function approach, or two-stage residual inclusion, has been suggested as an alternative to 2SLS in non-linear models (see for example Papke and Wooldridge (2008); Basu et al. (2018))
- Standard errors need to be adjusted for the first step estimation, which can be done by jointly bootstrapping both steps (Cameron and Trivedi, 2022)
- Can we use a STATA program to assist?

Applied Example

- Analysis of health expenditure data for individuals 65+ in the U.S. (Medicare) (Cameron and Trivedi (2022) *Chapter Seven)*
- Outcome is log of total out-of-pocket expenditures on prescribed medications
- Endogenous variable is indicator for whether individual holds employer or union-sponsored health insurance
- Instrument is ratio of individual's social security income to income from all sources
- Other model variables include number of chronic conditions, age, female, whether black or Hispanic, and log of annual household income (in thousands of dollars).
- Data can be downloaded from here: https://www.stata-press.com/data/mus2.html

Code for applied example

```
clear
use "mus207mepspresdrugs"
global xvar totchr age female blhisp linc
capture program drop ivboot
program ivboot, rclass
regress hi_empunion ssiratio $xvar
predict v1hat, resid
regress ldrugexp hi_empunion $xvar v1hat, vce(robust)
return scalar blavgrexp = _b[hi_empunion]
return scalar btotchr = _b[totchr]
return scalar bage = _b[age]
return scalar bfemale = _b[female]
return scalar bblhisp = _b[blhisp]
return scalar blinc = _b[linc]
return scalar bcons = _b[_cons]
drop v1hat
end

bootstrap r(blavgrexp) r(btotchr) r(bage) r(bfemale) r(bblhisp) r(blinc) r(bcons), seed(123) reps(999): ivboot
```

Summary of applied example

- Bootstrapped standard errors are similar to those produced by 2SLS
- Control function example can be extended to case where second stage is non-linear
  - Example: Papke and Wooldridge (2008) investigate the impact of school funding on student math test pass rates
  - The authors use a control function approach with a fractional response model for the second stage
- Method can also be extended to discrete endogenous variables, but careful attention needs to be placed on the form of the residual. Some useful references:
  - Binary: 2SLS VS 2SRI: APPROPRIATE METHODS FOR RARE OUTCOMES AND/OR RARE EXPOSURES - PMC (nih.gov)
  - Multinomial: Testing Exogeneity of Multinomial Regressors in Count Data Models: Does Two-stage Residual Inclusion Work? (degruyter.com)

Looping in STATA (Continued)

foreach command

foreach lname {in | of *listtype*} *list* { *commands referring to 'lname'* {

}

- Allows for looping over items in a list
- Example: frequency tables for a list of variables

```
foreach var of varlist $xvars {
tab `var'
}
```

Can this looping structure be helpful in applied econometrics?

Panel Data Econometrics

- In STATA, one area of specialization is said to be panel data (https://sites.google.com/a/nyu.edu/statistical-software-guide/summary)

$$y_{it} = \beta_0 + \beta_1 X_{it} + \mu_i + \epsilon_{it}$$

Estimation is typically performed by pooled estimation, random effects, and fixed effects

  - NB: fixed and random effects are estimators, as referred to in econometrics (https://www.jstatsoft.org/article/view/v027i02)
- By using fixed effects estimator, we can control for unobserved time invariant heterogeneity (μ_i).

Panel Data Econometrics (Continued)

Panel data econometrics

- Linear model

xtreg *depvar indepvars if in weight*, fe *FE options*

- Non-linear models
  - Conditional likelihood considered to be the ”gold standard” (Allison, 2014). See Allison (2009) for details
  - This can be implemented in STATA for logit and poisson

xtlogit *depvar indepvars if in weight*, fe *FE options*

xtpoisson *depvar indepvars if in weight*, fe *FE options*

Panel Data Econometrics (Continued)

- Some potential issues with traditional fixed effects approaches
  - Cannot obtain estimates of time-invariant variables
  - Hausman test may fail to compute (https://www.statalist.org/forums/forum/general-stata-discussion/general/1406912-hausman-test-using-suest-for-xtlogit)
- One approach is to use Allison's (2009) hybrid method, which involves including time averages of all time-varying variables, as well as deviations from these time averages. This involves the creation of many new variables.
- Allison (2009) creates these variables line-by-line. Can we use the looping structure to help?
- How do the results compare to fixed effects estimator, as well as Mundlak's (1978) correction (commonly used in economics, see Cameron and Trivedi (2022))

Applied Example

- Analysis of wage in Cameron and Trivedi (2022) *Chapter Eight*
- Time varying explanatory variables include weeks worked and experience (including quadratic)
- Time invariant explanatory variable is education
- Data can be downloaded from here: https://www.stata-press.com/data/mus2.html

Code for Applied Example

```
clear
use "mus2/mus208psid"
global xvars exp exp2 wks
foreach var of varlist $xvars {
                by id: egen mean`var' = mean(`var')
                gen d`var' = `var' - mean`var'
}

*Fixed Effects*
xtreg lwage $xvars ed, fe vce(cluster id)

*Hybrid (Allison)*
xtreg lwage dexp dexp2 dwks meanexp meanexp2 meanwks ed, re vce(cluster id)
test (dexp = meanexp) (dexp2 = meanexp2) (dwks = meanwks) *Alternative to Hausman - see Allison (2009)*

*Mundlak Correction*
xtreg lwage $xvars meanexp meanexp2 meanwks ed, re vce(cluster id)
test meanexp meanexp2 meanwks *Alternative to Hausman - see Wooldridge (2010)*
```

Summary of applied example

- In fixed effects specification, education was omitted from regression
- In hybrid and Mundlak specifications, education could be included. We also obtained fixed effects estimates on the time-varying variables
- Key difference between hybrid and Mundlak specifications is the interpretation of the group mean variables. Allison discusses it here: https://statisticalhorizons.com/problems-with-the-hybrid-method/
- Mundlak correction more common in economics textbooks (i.e., Cameron and Trivedi (2022)) but hybrid method has been used to estimate sibling fixed effects in economics literature (see Lebenbaum (2022))

Conclusions

- mata facilitates matrix programming in STATA, thereby allowing us to expand the capabilities of existing methods
  - Examples in this presentation are for Markov modelling, but programming new statistical routines is possible as well
- Programming statements (loops, macros, etc.) facilitate data analysis and provide efficiency gains
- Preprogrammed commands allow for the estimation of complex models using STATA syntax
  - Can easily create group mean variables, perform joint tests of statistical significance, etc. as part of the analysis

Additional Resources

Applied Econometrics

- Jones, A.M., Rice, N., d'Uva, T.B., Balia, S. 2013. Applied Health Economics - Second Edition, Routledge Advanced Texts in Economics and Finance. Taylor & Francis
- Cameron, A.C., Trivedi, P.K. 2022. Microeconometrics Using Stata – Volume 1: Cross-sectional and panel regression methods, Stata Press books.
- Allison, P.D. 2009. Fixed Effects Regression Models, Quantitative Applications in the Social Sciences. SAGE Publications.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press

Mathematical Economics

- Hoy, M, Livernois, J, Mckenna, C, Rees, R, Stengos, T. 2011. Mathematics for Economics – Third Edition. MIT Press Books

Medical Decision Making (R code)

- Alarid-Escudero, F., Krijkamp, E. M., Enns, E. A., Yang, A., Hunink, M. G., Pechlivanoglou, P., & Jalal, H. (2021). A Tutorial on time-dependent cohort state-transition models in R using a cost-effectiveness analysis example. *arXiv preprint arXiv:2108.13552*.

Thanks for Listening

Good luck with STATA!