Final Project Stata Hints

73-374 Econometrics II

11/29/2020

The following hints may help you decipher the code and replicate the results from the papers. The code used to produce the tables and figures is in the .do or .ado files in the replication folders, written in the Stata language. While the goal of the coding component of the project is to figure out what is done in the paper and code it yourself, based on your understanding of the econometrics and applying it in R, the original code may provide a potential source of clarification as to what was done in the paper. You can open the files as text and use them to inform the R code you write. Stata is available on the computers in the computer labs at CMU if you would like to run the code. Information on the variables used can be inferred from this code, from the papers, from the variable names, and the variable labels in the data set, which can be accessed from a .dta file using R command **attr(datasetname, "var.labels")**. Files in Stata's ".dta" format can be imported using libraries **foreign**, **haven**, or, for files created in more recent versions of Stata, **readstata13**.

Stata is a statistical language similar to R (but not free or open source), with slightly different syntax and command names: a glossary of some commonly used commands in Stata and their R equivalents follows.

gen x=(something) (equivalently, **generate** or just **g**) in Stata is the same as **x<-(something)** in R (**egen** creates multiple objects).

reg y x1 x2 is **lm(y~x1+x2)**. Option **,robust** or **vce(something)** after reg corresponds to using a different standard error formula: robust uses heteroskedasticity-robust standard errors, while the **vce()** option allows using some other standard error estimate: these correpond to using the **sandwich** package in R to get alternate stanard error estimates

xtreg y x, fe (or **, re**) is **plm(y~x,model="within")** (or **model="random"**)

tsset var1 var2 sets data as panel data with indices var1 and var2, equivalent to option **index=c("var1" "var2")** in **plm** or **pdata.frame**

rghdfe y x, absorb(z1 z2 z3) is also fixed effects estimation using a different command. It corresponds to using **factor(z1)+factor(z2)+factor(z3)** as covariates (fixed effects) in a least squares dummy variable regression. It can also be implemented using package **lfe** using command **felm y~x | z1+z2+z3**, which allows specifying multiple types of fixed effects. The option **vce(cluster varname)** after this command performs standard error clusering where errors are allowed to be correlated within groups defined by variable "varname" (which may or may not be one of the fixed effects variables). See the help files for **felm** in package **lfe** for details and options. This is implemented in **felm** by augmenting the formula with the cluster variable. In the example above, this is **felm y~x | z1+z2+z3| varname**

areg y x, absorb(z) or **xi: reg y x i.z** are the same as **lm(y~x+factor(z))**

ivregress y (x=z) is **ivreg(y~x|z)** (some Stata code uses a package with additional features, with command **ivreg2** instead, which otherwise has the same syntax)

total gives a **sum**

summ or **sum** (or just **s**) is **summary**

lincom computes point estimates and standard errors for linear combinations of coefficient estimates. **test** just performs (Wald or F) test. Point estimates can be computed manually in R and test statistics and p-values can be computed by performing a Wald or F test using library **coeftest**.

by (index) {command} and **"foreach"** are the same as a **'for'** loop in R, see section 1.8 of your R textbook at URfIE.net

collapse (command) varname, by (varname2) applies group level command "command" (eg mean, max, min, etc) to variable varname within a group specified by varname2 and "collapses" the dataset to contain 1 observation per group specified by varname2. Although this can be done using base R commands, this kind of data manipulation is facilitated by using library **dplyr**: see https://r4ds.had.co.nz/transform.html for guidance.

A $*$ after text, eg $var*$ indicates that all variables starting with the preceding text are included in the variable list, eg $var1$, $var2$, var_US, etc.

[**aw=variable**] option in a regression or similar command is the **weights="variable"** option in **lm/ivreg**

Some of the summary statistics may use weighted means: these can be calculated by using option *weights* in **lm** by running a weighted regression on just a constant.

Many Stata coders use a package called **estout** to make nicely formlated tables (similar to Stargazer or xtable in R): commands **eststo**, **estadd**, **estout**, **esttab** save results for tables, add a number to a table, create tables, and display tables, respectively.

qui before a command runs the command without displaying the output: it has no effect on the actual results

drop, **keep**, and **replace** remove observations, remove all but the selected observations, or change observations to some formula. They can be replicated in R using **subset** or the assignment operator **<-**

Stata includes standard error clustering as an option in regression by **cluster(variablename)** after **xtreg** or **xi areg** or **vce(cluster variablename)** after **reghdfe**. In R, this can be done after estimation; if the regression was run with the command **plm** in library **plm**, you can use the command vcovHC with option cluster="group" so long as the data is set as a panel using pdata.frame with the group index set to be *variablename* (The Stata command for this is **xtset**). Note that for some specifications in the paper, this may not be the same as the index used for differencing or fixed effects estimation. To get around this for first differences, you can create diffferences manually and use option method="pooling" in *plm* to run the pooled regression. For fixed effects regression, you can create dummy variables for the fixed effect groups and include them in the pooled regression. You may also use other packages for clustering, such as the library **multiwaycov** or other commands available in R. Library **lfe**, which provides the alternative fixed effects estimation command **felm** also provides multiple clustering options for fixed effects regressions. Many of these commands differ from the Stata defaults in terms of the normalization of the sample variances (1/n vs 1/n-k or similar): this should not be a major concern, as the difference is asymptotically negligible, but the different **type=** options in **vcovHC** (or options for **exactDOF** in **felm**) may produce more comparable standard errors. Specialized cluster-robust inference packages in Stata may have different syntax. For example, command **clustse** before regress is used for wild cluster bootstrap. This is available as an option in command **felm** in library **lfe**. Similarly, **ritest** performs permutation or randomization based inference.