Geneva Academics in Management and Economics

THE DO'S AND DON'TS OF A DO-FILE

tips on how to do empirical work with Stata

Emmanuel Milet (emmanuel.milet@unige.ch)

15th Nov. 2016

The do's and don'ts of a do-file

Why bother with a learning how to write a do-file?

- **THE** most important tool of your empirical analysis.
- You may share it with co-authors
- You will return to it several month after submitting your paper (and you'll possibly do this multiple times too)
- You may be asked by an editor to make your code and data publicly available

The do's and don'ts of a do-file

Some important steps:

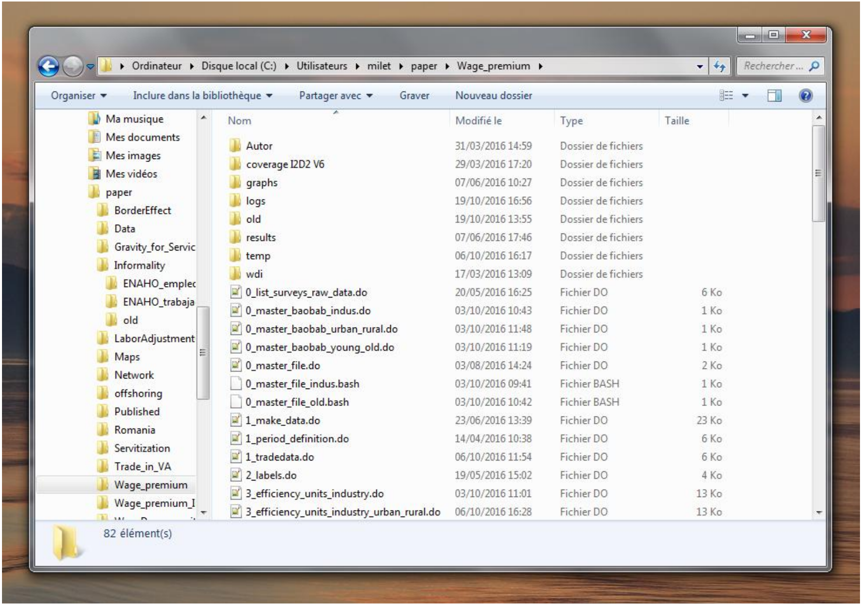
1. Folder
2. Master do-file
3. Logs
4. Names and label variables
5. Regression output
6. Various commands

The do's and don'ts of a do-file: FOLDER

You are going to generate TONS of output while doing your empirical analysis. To make things *easier* and *clearer*, I recommend to create a folder for each type of output:

- logs
- graphs
- regression results
- tables and other output for the stylized facts
- manuscript
- an "old" folder where you can store do-files and other output that you do not need anymore

Your data and do-file can be stored in the root folder.

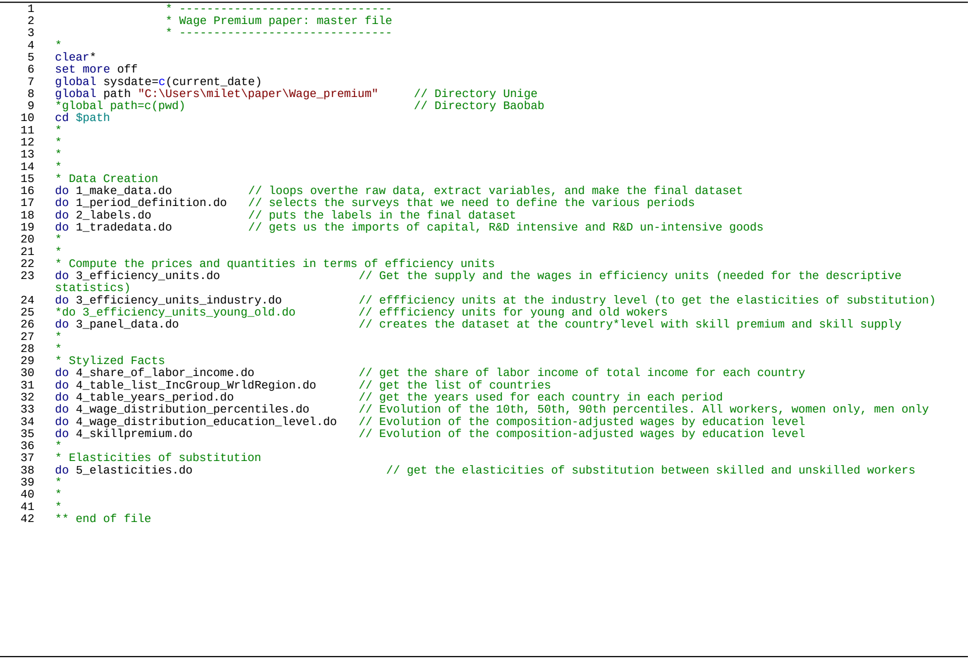


Nom	Modifié le	Type	Taille
Autor	31/03/2016 14:59	Dossier de fichiers	
coverage I2D2 V6	29/03/2016 17:20	Dossier de fichiers	
graphs	07/06/2016 10:27	Dossier de fichiers	
logs	19/10/2016 16:56	Dossier de fichiers	
old	19/10/2016 13:55	Dossier de fichiers	
results	07/06/2016 17:46	Dossier de fichiers	
temp	06/10/2016 16:17	Dossier de fichiers	
wdi	17/03/2016 13:09	Dossier de fichiers	
0_list_surveys_raw_data.do	20/05/2016 16:25	Fichier DO	6 Ko
0_master_baobab_indus.do	03/10/2016 10:43	Fichier DO	1 Ko
0_master_baobab_urban_rural.do	03/10/2016 11:48	Fichier DO	1 Ko
0_master_baobab_young_old.do	03/10/2016 11:19	Fichier DO	1 Ko
0_master_file.do	03/08/2016 14:24	Fichier DO	2 Ko
0_master_file_indus.bash	03/10/2016 09:41	Fichier BASH	1 Ko
0_master_file_old.bash	03/10/2016 10:42	Fichier BASH	1 Ko
1_make_data.do	23/06/2016 13:39	Fichier DO	23 Ko
1_period_definition.do	14/04/2016 10:38	Fichier DO	6 Ko
1_tradedata.do	06/10/2016 11:54	Fichier DO	6 Ko
2_labels.do	19/05/2016 15:02	Fichier DO	4 Ko
3_efficiency_units_industry.do	03/10/2016 11:01	Fichier DO	13 Ko
3_efficiency_units_industry_urban_rural.do	06/10/2016 16:28	Fichier DO	13 Ko

The do's and don'ts of a do-file: MASTER DO FILE

A master do-file is a do-file which launches your other do-files in a specific order.

- It allows you to organize your do-files, and add comments to them.
- It makes sure that your final results (in your paper) are produced correctly.
- It allows you to declare a directory path, as well as other options that you may use across all your do-files.



```
                * -------------------------------
                * Wage Premium paper: master file
                * -------------------------------
*
clear*
set more off
global sysdate=c(current_date)
global path "C:\Users\milet\paper\Wage_premium"     // Directory Unige
*global path=c(pwd)                                 // Directory Baobab
cd $path
*
*
*
*
* Data Creation
do 1_make_data.do           // loops overthe raw data, extract variables, and make the final dataset
do 1_period_definition.do   // selects the surveys that we need to define the various periods
do 2_labels.do              // puts the labels in the final dataset
do 1_tradedata.do           // gets us the imports of capital, R&D intensive and R&D un-intensive goods
*
*
* Compute the prices and quantities in terms of efficiency units
do 3_efficiency_units.do                    // Get the supply and the wages in efficiency units (needed for the descriptive statistics)
do 3_efficiency_units_industry.do           // effficiency units at the industry level (to get the elasticities of substitution)
*do 3_efficiency_units_young_old.do         // effficiency units for young and old wokers
do 3_panel_data.do                          // creates the dataset at the country*level with skill premium and skill supply
*
*
* Stylized Facts
do 4_share_of_labor_income.do               // get the share of labor income of total income for each country
do 4_table_list_IncGroup_WrldRegion.do      // get the list of countries
do 4_table_years_period.do                  // get the years used for each country in each period
do 4_wage_distribution_percentiles.do       // Evolution of the 10th, 50th, 90th percentiles. All workers, women only, men only
do 4_wage_distribution_education_level.do   // Evolution of the composition-adjusted wages by education level
do 4_skillpremium.do                        // Evolution of the composition-adjusted wages by education level
*
* Elasticities of substitution
do 5_elasticities.do                            // get the elasticities of substitution between skilled and unskilled workers
*
*
*
** end of file
```

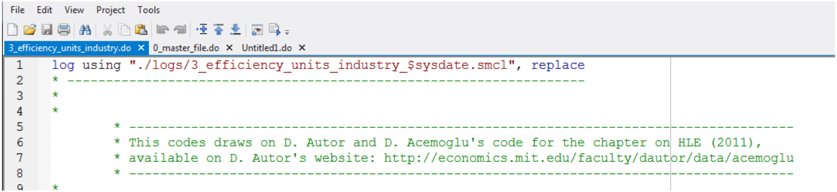
The do's and don'ts of a do-file: LOGS

Log files are used to keep track of EVERYTHING your do-file does, and more:

- It saves what Stata erases in the result window.
- You should ALWAYS start your do-file by opening a log file, and end by closing it.
- Your log files should be stored in a specific log folder.
- Give the SAME name to your do-files and your logs.
- You can also add the date of the day in the log's name, so that you do not erase them each time your run your do-file.

The do's and don'ts of a do-file: LOGS

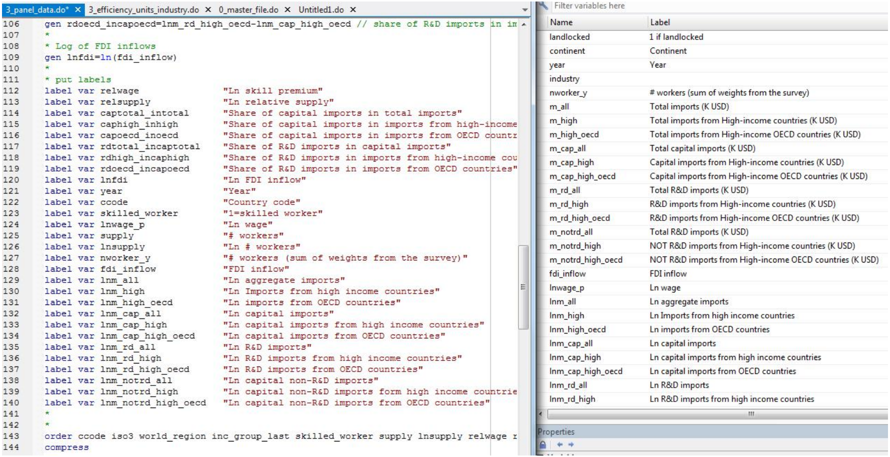
Example:



```
log using "./logs/3_efficiency_units_industry_$sysdate.smcl", replace
* -------------------------------------------------------------------
*
*
        * -------------------------------------------------------------------------------------
        * This codes draws on D. Autor and D. Acemoglu's code for the chapter on HLE (2011),
        * available on D. Autor's website: http://economics.mit.edu/faculty/dautor/data/acemoglu
        * -------------------------------------------------------------------------------------
*
```

The do's and don'ts of a do-file: NAMES & LABELS

1. Give explicit names to your variables.
2. Adopt a naming rule and stick to it! For instance:
   - log variables can be named: ln_*varname*
   - dummy variables created from variables: *varname*_dum
3. Attach labels to the variables.
4. Ideally, you should have a do-file with all the labels.
5. Labels will show up in any output that you produce.



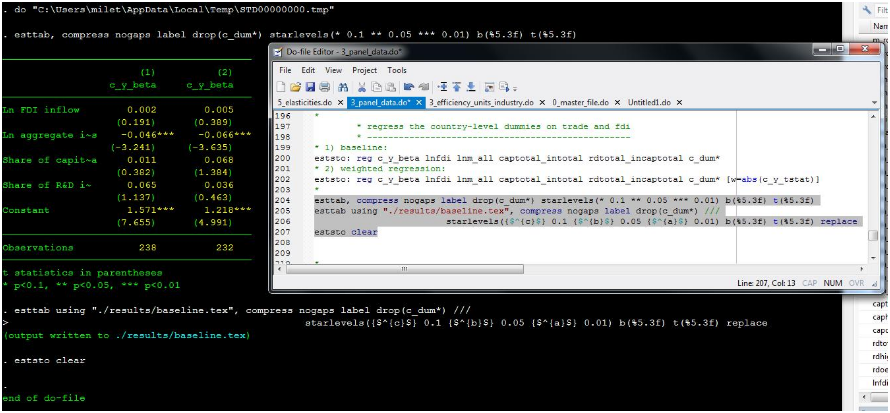
```
gen rdoecd_incapoecd=lnm_rd_high_oecd-lnm_cap_high_oecd // share of R&D imports in im
*
* Log of FDI inflows
gen lnfdi=ln(fdi_inflow)
*
* put labels
label var relwage               "Ln skill premium"
label var relsupply             "Ln relative supply"
label var captotal_intotal      "Share of capital imports in total imports"
label var caphigh_inhigh        "Share of capital imports in imports from high-income
label var capoecd_inoecd        "Share of capital imports in imports from OECD countr
label var rdtotal_incaptotal    "Share of R&D imports in capital imports"
label var rdhigh_incaphigh      "Share of R&D imports in imports from high-income cou
label var rdoecd_incapoecd      "Share of R&D imports in imports from OECD countries"
label var lnfdi                 "Ln FDI inflow"
label var year                  "Year"
label var ccode                 "Country code"
label var skilled_worker        "1=skilled worker"
label var lnwage_p              "Ln wage"
label var supply                "# workers"
label var lnsupply              "Ln # workers"
label var nworker_y             "# workers (sum of weights from the survey)"
label var fdi_inflow            "FDI inflow"
label var lnm_all               "Ln aggregate imports"
label var lnm_high              "Ln Imports from high income countries"
label var lnm_high_oecd         "Ln imports from OECD countries"
label var lnm_cap_all           "Ln capital imports"
label var lnm_cap_high          "Ln capital imports from high income countries"
label var lnm_cap_high_oecd     "Ln capital imports from OECD countries"
label var lnm_rd_all            "Ln R&D imports"
label var lnm_rd_high           "Ln R&D imports from high income countries"
label var lnm_rd_high_oecd      "Ln R&D imports from OECD countries"
label var lnm_notrd_all         "Ln capital non-R&D imports"
label var lnm_notrd_high        "Ln capital non-R&D imports form high income countrie
label var lnm_notrd_high_oecd   "Ln capital non-R&D imports from OECD countries"
*
*
order ccode iso3 world_region inc_group_last skilled_worker supply lnsupply relwage r
compress
```

Name	Label
landlocked	1 if landlocked
continent	Continent
year	Year
industry	
nworker_y	# workers (sum of weights from the survey)
m_all	Total imports (K USD)
m_high	Total imports from High-income countries (K USD)
m_high_oecd	Total imports from High-income OECD countries (K USD)
m_cap_all	Total capital imports (K USD)
m_cap_high	Capital imports from High-income countries (K USD)
m_cap_high_oecd	Capital imports from High-income OECD countries (K USD)
m_rd_all	Total R&D imports (K USD)
m_rd_high	R&D imports from High-income countries (K USD)
m_rd_high_oecd	R&D imports from High-income OECD countries (K USD)
m_notrd_all	Total R&D imports (K USD)
m_notrd_high	NOT R&D imports from High-income countries (K USD)
m_notrd_high_oecd	NOT R&D imports from High-income OECD countries (K USD)
fdi_inflow	FDI inflow
lnwage_p	Ln wage
lnm_all	Ln aggregate imports
lnm_high	Ln Imports from high income countries
lnm_high_oecd	Ln imports from OECD countries
lnm_cap_all	Ln capital imports
lnm_cap_high	Ln capital imports from high income countries
lnm_cap_high_oecd	Ln capital imports from OECD countries
lnm_rd_all	Ln R&D imports
lnm_rd_high	Ln R&D imports from high income countries

The do's and don'ts of a do-file: REGRESSION OUTPUT

Regression results are probably the most important output from your do-files.
It is important that they look nice!

- The "combo" esttab/estout commands is a good way to produce and export regression results.
- There is also outreg2 (personally not a fan)



```
. do "C:\Users\milet\AppData\Local\Temp\STD00000000.tmp"

. esttab, compress nogaps label drop(c_dum*) starlevels(* 0.1 ** 0.05 *** 0.01) b(%5.3f) t(%5.3f)

--------------------------------------------
                         (1)          (2)
                    c_y_beta     c_y_beta
--------------------------------------------
Ln FDI inflow          0.002        0.005
                     (0.191)      (0.389)
Ln aggregate i~s   -0.046***    -0.066***
                    (-3.241)     (-3.635)
Share of capit~a       0.011        0.068
                     (0.382)      (1.384)
Share of R&D i~        0.065        0.036
                     (1.137)      (0.463)
Constant            1.571***     1.218***
                     (7.655)      (4.991)
--------------------------------------------
Observations             238          232
--------------------------------------------
t statistics in parentheses
* p<0.1, ** p<0.05, *** p<0.01

. esttab using "./results/baseline.tex", compress nogaps label drop(c_dum*) ///
>                                                starlevels({$^{c}$} 0.1 {$^{b}$} 0.05 {$^{a}$} 0.01) b(%5.3f) t(%5.3f) replace
(output written to ./results/baseline.tex)

. eststo clear

.
end of do-file
```

```
*
        * regress the country-level dummies on trade and fdi
        * ---------------------------------------------------
* 1) baseline:
eststo: reg c_y_beta lnfdi lnm_all captotal_intotal rdtotal_incaptotal c_dum*
* 2) weighted regression:
eststo: reg c_y_beta lnfdi lnm_all captotal_intotal rdtotal_incaptotal c_dum* [w=abs(c_y_tstat)]
*
esttab, compress nogaps label drop(c_dum*) starlevels(* 0.1 ** 0.05 *** 0.01) b(%5.3f) t(%5.3f)
esttab using "./results/baseline.tex", compress nogaps label drop(c_dum*) ///
                          starlevels({$^{c}$} 0.1 {$^{b}$} 0.05 {$^{a}$} 0.01) b(%5.3f) t(%5.3f) replace
eststo clear


*
```

The do's and don'ts of a do-file: REGRESSION OUTPUT

```
{
\def\sym#1{\ifmmode^{#1}\else\(^{#1}\)\fi}
\begin{tabular}{l*{2}{c}}
\hline\hline
                &\multicolumn{1}{c}{(1)}&\multicolumn{1}{c}{(2)}\\
                &\multicolumn{1}{c}{Dem. shifters}&\multicolumn{1}{c}{Dem. shifters
\hline
Ln FDI inflow   &     0.002         &     0.005         \\
                &   (0.191)         &   (0.389)         \\
Ln aggregate imports&    0.046{$^{a}$}&    0.066{$^{a}$}\\
                &  ( 3.241)         &  ( 3.635)         \\
Share of capital imports in total imports&    0.011         &     0.068
\\
                &   (0.382)         &   (1.384)         \\
Share of R\&D imports in capital imports&    0.065         &     0.036
\\
                &   (1.137)         &   (0.463)         \\
Constant        &     1.571{$^{a}$}&     1.218{$^{a}$}\\
                &   (7.655)         &   (4.991)         \\
\hline
Observations    &       238         &       232         \\
\hline\hline
\multicolumn{3}{l}{\footnotesize \textit{t} statistics in parentheses}\\
\multicolumn{3}{l}{\footnotesize {$^{c}$} p<0.1, {$^{b}$} p<0.05, {$^{a}$} p<0.01}
\end{tabular}
}
```

The do's and don'ts of a do-file: REGRESSION OUTPUT

	(1) Dem. shifters	(2) Dem. shifters
Ln FDI inflow	0.002	0.005
	(0.191)	(0.389)
Ln aggregate imports	-0.046^{a}	-0.066^{a}
	(-3.241)	(-3.635)
Share of capital imports in total imports	0.011	0.068
	(0.382)	(1.384)
Share of R&D imports in capital imports	0.065	0.036
	(1.137)	(0.463)
Constant	1.571^{a}	1.218^{a}
	(7.655)	(4.991)
Observations	238	232

t statistics in parentheses

^{c} p<0.1, ^{b} p<0.05, ^{a} p<0.01

The do's and don'ts of a do-file: USEFUL COMMANDS

The **COLLAPSE** command:

- It reduces the dimensionality of your dataset and calculates many statistics (count, median, mean, standard deviation, percentiles, first obs. . . .) base on certain dimensions (useful when using panel data, or data with more than 2 dimensions: firm×product×destination×year for instance)
- BEWARE: Weight normalization impacts only the sum, count, sd, semean, and sebinomial statistics
- A weighted average cannot be obtained directly from collapse, despite the fact that you can ask for the mean of a variable, and specify weights.

The do's and don'ts of a do-file: USEFUL COMMANDS

The **TAG** command:

- It creates a dummy variable taking the value 1 for each occurrence of a variable.
- It is useful when you have nested dimensions (say individual/household/county/region), and want to quickly get statistics at a specific level.

The do's and don'ts of a do-file: USEFUL COMMANDS

The **TAG** command:

- It creates a dummy variable taking the value 1 for each occurrence of a variable.
- It is useful when you have nested dimensions (say individual/household/county/region), and want to quickly get statistics at a specific level.

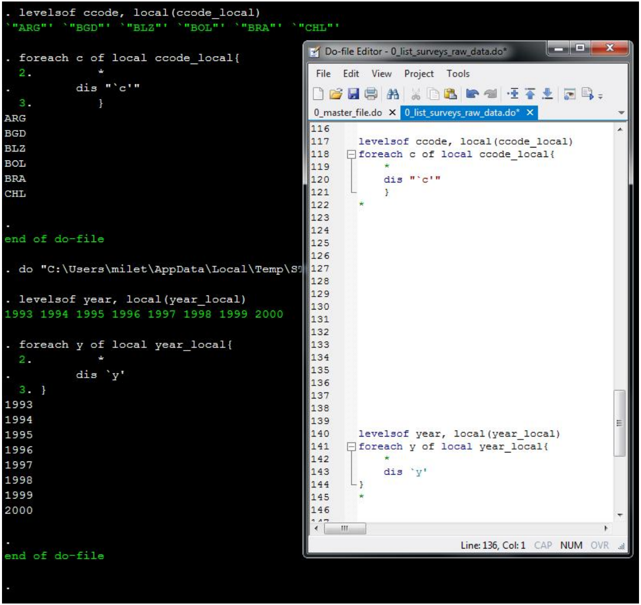
The **GROUP** command:

- Creates a variable taking integer values from 1 to $\mathbb{N}$ for each occurrence of a given variable or list of variables (i.e. an occurrence is therefore a combination of various variables).
- This is handy to create fixed-effect variables.
- Turns a string variable into a numeric variable (the command **encode** does this too, but only for 1 variable).

The do's and don'ts of a do-file: USEFUL COMMANDS

The **LEVELSOF** command:

- Especially useful for loops
- IT lists all occurrences of a variable, and stores the list into a local variable.
- You can then loop over the elements of this list.
- The nice feature is that Stata does not create any variable, the elements of the list (i.e. the occurrences) are *local* elements.



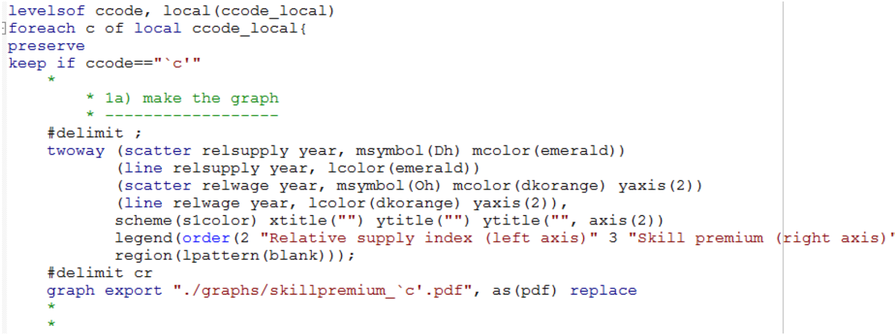
```
. levelsof ccode, local(ccode_local)
`"ARG"' `"BGD"' `"BLZ"' `"BOL"' `"BRA"' `"CHL"'

. foreach c of local ccode_local{
  2.         *
.         dis "`c'"
  3.         }
ARG
BGD
BLZ
BOL
BRA
CHL


.
end of do-file

. do "C:\Users\milet\AppData\Local\Temp\ST

. levelsof year, local(year_local)
1993 1994 1995 1996 1997 1998 1999 2000

. foreach y of local year_local{
  2.         *
.         dis `y'
  3. }
1993
1994
1995
1996
1997
1998
1999
2000

.
end of do-file

.
```

Do-file Editor - 0_list_surveys_raw_data.do*
File
Edit
View
Project
Tools
0_master_file.do
0_list_surveys_raw_data.do*

```

levelsof ccode, local(ccode_local)
foreach c of local ccode_local{
	*
	dis "`c'"
	}
*

















levelsof year, local(year_local)
foreach y of local year_local{
	*
	dis `y'
}
*

```

The **#delimit** command:

- This is a line breaker.
- Essentially the same as using three forward slash bars: ///
- It is more convenient though (personal opinion).



```
levelsof ccode, local(ccode_local)
foreach c of local ccode_local{
preserve
keep if ccode=="`c'"
    *
        * 1a) make the graph
        * ------------------
    #delimit ;
    twoway (scatter relsupply year, msymbol(Dh) mcolor(emerald))
           (line relsupply year, lcolor(emerald))
           (scatter relwage year, msymbol(Oh) mcolor(dkorange) yaxis(2))
           (line relwage year, lcolor(dkorange) yaxis(2)),
           scheme(s1color) xtitle("") ytitle("") ytitle("", axis(2))
           legend(order(2 "Relative supply index (left axis)" 3 "Skill premium (right axis)"
           region(lpattern(blank)));
    #delimit cr
    graph export "./graphs/skillpremium_`c'.pdf", as(pdf) replace
    *
    *
```

- That's it for today.
- You can find a written version fo all this on my webpage: `http://emmanuelmilet.weebly.com/`
- Thank you for your attention