Stata 入门介绍

转载，原作者不详。

(1) Stata 要在使用中熟练的，大家应该多加练习。

(2) Stata 的很多细节，这里不会涉及，只是选取相对重要的部分加以解释，大家在使用 Stata 过程中留心积累。作为入门性质的介绍，本文只选取和中级计量经济学作业相关的内容和一些处理数据所使用的基本命令。对于更高深的内容，请大家参看 STATA manual.”

界面

当我们把 Stata 装好以后，首先需要了解的是它的界面。打开 Stata 后我们便可以看到它常用的四个窗口：Stata Results; Review; Variables; Stata Command。我们所有的运行结果都会在 Stata Results 界面中显示；而命令的输入则在 Stata Command 窗口；Review 窗口记录我们使用过的命令；最后 Variables 窗口显示存在于当前数据库中的所有变量的名称。可以直接点击 Review 窗口来重新输入已使用过的命令，我们所需变量可以通过点击 Varaibles 窗口来得到，这些都可以简便我们的操作。

Stata 命令

Stata 软件功能强大，体现在它提供了丰富的命令，可以实现许多功能。每一个 Stata 命令都相应的命令格式。我们在这里介绍常用的一些命令的功能和相应的格式，大家在使用 Stata 的过程中会不断积累命令的相关知识。

需要对命令的帮助时可以用 help 命令查询。例如了解命令： “reg” ，就可以在 Stata Command 窗口输入 “help reg” ，也可以在 Help 选项下 content 中查找我们需要的相关命令。用 help 查询，则窗口会显示关于该命令的详尽说明。更直接的办法是看 Examples 中的范例是如何使用该命令，阅读一些相关的说明并加以模仿。

重要习惯

我们使用 Stata 进行回归分析时，需要养成一些好的习惯。在进行一些数据量很大，过程复杂的分析时尤其重要。

（1）使用日志（log）。它可以帮助我们记录 stata 的运行结果。

格式：log using c:\stata8\logfiles\10.21.5_30.log

（注意：我们需要先建好文件夹 c:\stata8\logfiles ）

关闭 log 的命令为“log close”。

格式: log close

那么“10.21.5_30.log”文件就记录了从“log using”命令 到“log close”命令之间 stata 运行的所有结果。

（2）Do-file。在 command 窗口输入命令的方式很受限制，我们使用工具栏中“Do-file-editor”（第 8 个）在 Do-file 中编程。直接的好处便是我们可以很方便的执行以前写过的命令，并记录我们需要的命令，方便下一次的使用和分析。在复杂的分析中，采用 Command 窗口输入的方式会是非常的困难，我们必须用 do-file 去编程。

在 do-file 文件中，用*表示注释内容，Stata 在运行 do-file 时会跳过这些注释语句。加入注释语句能增强 do-file 的可读性。我们应该养成习惯为每一个 do-file 文件写详细的注释内容。比如要说明文件名称，回归分析的目的，时间和存放位置。如果过程中生成并保存了数据文件，应写出相应数据文件的名称等。如果中途对 do-file 文件进行过修改，最好将修改过文件保存为另一个文件，以便于将来对比分析原文件和修改后的文件。

格式：

```
*Wage_analysis.do
*The program is written for the analysis of wage determination.
*Data management: reshape the data to panel.
*This rsult will be saved in the data file: wage1.dta
* written: 10/21/05
```

在调试 do-file 文件时，可以选择部分命令让 Stata 只运行选中部分。

我们可以保存当前使用的 do-file 文件。Review 窗口中的命令也可以保存为 do-file。方法是右键点击 Review 窗口，选择 Save Review Contents。

（3）存储数据。在分析一个大的数据库时，中途对数据有改动和删减，有必要在分析过程中将数据进行保存，可以用 File 选项中“save as”，同时要为中途保存的数据文件写一个详尽的说明文件，此外还可以在 do-file 文件中或 command 窗口中使用命令 “save” 来实现。

格式：save c:\stata\datasets\2.dta

打开数据文件

我们用 Stata 做回归的第一步便是打开一个数据库。我们可以用工具栏“Open”（第 1 个），打开相应数据文件。也可以使用命令 “use” 。

格式: use c:\data\datasets\1.dta

Stata 有自己的数据格式，我们课上一般会给大家 Stata 格式的数据库。有时候，我们手头的数据格式不符合 Stata 的格式，就需要用相关软件进行转换，比如 transfer，对这个问题感兴趣的同学可以课后和我们联系。如果我们的数据是 Excel 格式，那么可以直接把里面的数据拷贝粘贴到 Stata 中：只需要点开数据工具栏“Data Editor”（第 9 个），就可以进行粘贴。

打开数据后我们可以用工具栏“Data Browse”（第 10 个）浏览数据。浏览数据可以帮助我们了解具体每一个数据。要了解数据具有的特征，我们必须借助 Stata 命令。

了解数据特征

“describe” 命令可以告诉我们每一个变量的含义。

格式：describe

具体了解每一个变量的特征，我们可以用 tabstat 命令。例如我们可以计算 wage 的均值，方差，中位数，范围，具体可以用 help tabstata 查询。

格式:
tabstat wage, stats(mean)
tabstat wage, stats (sd median range) (注意不要逗号)

如果我们想要了解不同教育水平的工资的均值，可以用如下命令：

格式: tabstat wage, by (educ) stats(mean)

此外可以使用 “Sum”，它是命令 “summarize” 的简写。Summarize（Sum）将汇报数据的均值和方差等信息。

格式：
summarize wage
sum educ exper

需要了解如“中位数”(median)，我们可以进一步使用后缀 detail。此时会详细报告百分比所对应的样本值。

格式：sum wage educ, detail

此外 Stata 还提供了别的命令帮助我们了解数据，如 “codebook” 命令，它与带 detail 后缀的 “sum” 命令相似。 “table”，它将报告数据取值和相应的频率。 “tabulate”(或简写为 ta)是一个很有用的命令。与 table 相比，ta 将进一步报告数据分布的百分比。

格式:
codebook wage educ
table wage
ta educ

利用“by”命令，我们可以了解数据更细致的特征。例如我们想知道受不同教育的人群中工资的分布。

格式:
sort educ（这一步不可缺，一定需要先排序）
by educ: table wage
by educ: tabulate wage

画图

很多时候，画图能够直观地看到数据分布和它们之间关系。比如我们可以 “histogram” 命令画出数据分布的柱状图(histogram)。

格式：
histogram wage

“scatter”命令可以画出两个变量之间的分布关系。例如我们想直观的看到教育水平变化时工资的变化，可以用 “scatter” 命令或者 “graph twoway scatter” 命令。

格式:

```
scatter wage educ
graph twoway scatter wage educ
```

“graph twoway”命令可以带别的后缀，例如 “graph twoway line” 则画的是线状图。

格式:

```
graph twoway line wage educ
```

“graph”命令还有很多别的功能。例如使用“graph matrix”可以了解更多的变量之间的关系。“graph bar (mean) y, over(x)”就可以了解 y 的平均值关于 x 分布的柱状图。

格式:

```
graph matrix wage educ
graph matrix wage educ exper
graph bar (mean) wage, over (educ)
```

右键点击 graph 窗口可以将图片进行保存和复制。

变量

在分析的过程中，有些变量并没有在数据中提供，需要我们用原始数据或者回归的结果构造。常用的命令是 “gen” 和 “egen” 。

格式

```
gen educsqr=educ^2
```

egen 命令相对复杂一些，它能生成一些“gen”命令无法生成的变量。例如可以生成 wagesum 为每个人的工资和，以及生成 wagemedian 为工资的中位数(median)，wagemax 为工资的最大值。

格式:

```
egen wagesum=sum(wage)
egen wagemedian=median(wage)
```

egen wagemax=max(wage)

更复杂的如想产生一个变量“wagemax”为相同教育水平里的最高工资。

格式: egen wagemaxeduc=max (wage),by (educ)

如果我们需要替换某一变量，我们可以用的命令是“replace”。

格式:
replace wagemax=wage
replace wagemax=1

有时候我们在生成变量时可以加上一定条件，例如如果一个样本工资超过 3，我们就定义它的变量 wagehigh 的取值为 1，否则为 0。

格式:
gen wagehigh=1 if wage>=10
replace wagehigh=0 if wagehigh ==.（注意是两个等号）

我们也需要去掉过程中的暂用的变量，以方便我们浏览数据和重新定义变量。我们可以用 drop 命令。

格式: drop educsqr wagesum wagemedian wagemax wagemaxeduc wagehigh

我们可以用“keep”或“drop”命令来删除一些样本,在删除之前，我们需要了解删除带来的影响，则可以用“count”命令来了解样本取值的情况。

格式:
count if wage<100
count if wage<10

我们可以用“sort”和“list”命令来了解数据分布的细节。例如我们想知道工资值从小到大排列在第 50 到 70 的样本的工资值。

格式:
sort wage

list wage in 50/70

如果我们想保留工资小于 100 的样本，可以有两种命令。

格式:
keep if wage<100
drop if wage>=100

有时我们关心变量之间的相关性，可以使用“correlate”命令，它将报告变量之间的相关系数。

格式：
correlate wage educ exper tenure

回归

现在我们以进入最重要的环节：回归分析。
进行 OLS 回归的命令为“reg”。

格式: reg wage educ

Stata Results 窗口将报告这一回归的相关结果:

. reg wage educ

```
Source  |       SS       df       MS              Number of obs =     526
--------+------------------------------           F(  1,   524) =  103.36
Model   |  1179.73204     1  1179.73204           Prob > F      =  0.0000
Residual|  5980.68225   524  11.4135158           R-squared     =  0.1648
--------+------------------------------           Adj R-squared =  0.1632
Total   |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593   .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

表格中最后两行报告回归的斜率和截距的系数，相应的标准差、t 值和 P 值，同时给出 95%的置信区间。在表格左上方，报告了回归的总变异、解释变异和残差变异。表格右上方报告回归的 R 方和调整后的 R 方。其中 F 是自变量所有的系数都为 0（即自变量完全没有解释力）这样一个零假设对应的 F 分布值。

回归会产生很多我们感兴趣的值,例如回归的拟合值以及回归的残差。Stata 提供了 predict 命令帮助我们存储这些变量。例如我们把拟合值定义为 wagehat，残差定义为 wageresid。

格式:
predict wagehat
predict wageresid, re

我们常常需要检验某一个零假设，例如在我们作了如下回归

格式: reg wage educ exper tenure nonwhite female

之后,我们想要知道 nonwhite 的系数是否显著,我们可以直接看回归结果报告,也可以用 test 命令。

格式: test nonwhite
test 命令报告的结果为 F 值。而回归结果报告的为 t 值。它们之间是平方关系，而 p 值是一样的。对于更复杂的零假设，比如 nonwhite 和 female 是否同时为 0。exper 的系数和 tenure 的系数是否相等，则只能借助“test”命令。

格式:
test nonwhite female
test exper=tenure

报告回归结果

一般需要报告回归系数和相应的残差，同时报告系数的显著性。此外根据需要往往还要报告回归的拟合优度和使用的样本个数。对于回归系数的符号和大小变化，要给出相应的分析和解释。许多时候还会把检验的结果附在表格中。

下面是一个报告回归结果的表格（摘自经济学论文）。其中括号里报告的是系数的方差， All Women

和 Married Women 表示两个总体，(1)(2)(3)对应不同的模型设定。

计算器

Stata 可以充当计算器用，使用 “display” 命令:

格式: display sqrt(5)*sin(0.5)

关于 Stata 的数学函数的命令格式，可以查询 help function。