

Stata基础知识

1、什么是stata

Stata是经济学研究主流的数据分析软件，它功能强大，程序包丰富，可以说几乎涵盖了应用计量经济学领域所有的功能，另外Stata的help文件非常详细，完全可以自学。

可以说，想要完成规范的现代经济学实证研究，像Stata这样的计量软件是必不可少的工具。

目前Stata的最新版本是15.0版，根据性能差异分为以下几种类型：

Stata/IC For mid-sized datasets.

Stata/SE For large datasets.

Stata/MP 2-core Fast & for the largest datasets.

Stata/MP 4-core Faster.

Stata/MP + cores Even faster.

商用版根据性能不同售价在\$1000~\$6500，学生版顶配为Stata/MP 4-core，售价\$995.

1、什么是stata

为什么选择Stata?

其他软件与Stata相比:

>SPSS 的图像化界面非常友好(同时操作也比较繁琐。当然, 它也可以输入命令或编程开发), 它更侧重于数据的统计描述, 貌似在社会学、心理学领域中比较常用, 比较大的缺点是其输出结果冗杂;

>Eviews 它的特点是专业(时间序列、横截面)但不全面, 图形化界面也很友好, 但面板数据的导入很麻烦, 感觉用的人不多;

>matlab 以处理速度快、语言简洁、自由度高著称, 但语言不够友好。正如其种类繁多的工具包所示, 它更适合工程、金融(如高频交易数据分析)、宏观经济学(Stata15.0已经有了动态随机一般均衡分析工具)、大数据分析等;

>R 与以上收费软件不同, R是开源软件, 因此在企业界普及度较高, R的程序包也十分丰富, 操作难度适中, 绘图十分精美, 处理速度也较快, 且与Office的兼容性非常好, 可以作为Stata的替代软件。

总之, 对于经济专业的学习者(尤其是初学者)而言, Stata和R是最佳选择, 而Stata比R更易上手。

2、Stata的图形化操作界面

Stata/SE 15.1

文件(F) 编辑(E) 数据(D) 图形(G) 统计(S) 用户(U) 窗口(W) 帮助(H)

历史窗口

过滤命令

#	命令	_rc
1	doedit "C:\Users\Sabrin...	

命令窗口

Statistics/Data Analysis

15.1

Copyright 1985-2017 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC <http://www.stata.com>
979-696-4600 stata@stata.com
979-696-4601 (fax)

100-student Stata lab perpetual license:
Serial number: 401506209499
Licensed to: sa
sa

Notes:
1. Unicode is supported; see [help unicode_advice](#).
2. Maximum number of variables is set to 5000; see [help set_maxvar](#).

. doedit "C:\Users\Sabrina\Desktop\计量经济学\计量经济学 stata introduction.do"
>
.

变量窗口

过滤变量

名称	标签
没有项目显示。	

属性窗口

变量

名称	
标签	
类型	
格式	
值标签	
注释	

数据

文件名	
标签	
注释	
变量	0
观测数	0
文件大小	0

C:\Users\Sabrina\Desktop\计量经济学

CAP NUM OVR

3、数据的导入

我们使用计量软件的目的是对“数据”施加“命令”，以得到结果。相应地，在Stata中，最主要的文件类型包括数据(.dta)和命令(.do)，下面我们分别介绍如何在Stata中对二者进行操作。

(1) Stata所直接处理的是扩展名为.dta文件，类似txt文档，占用存储空间小*可以在菜单栏打开

```
clear all
```

```
global root "C:\Users\Sabrina\Desktop\econometrics" /*定义全局宏变量root为文件夹路径。每次开关stata软件，全局宏变量都会被清除，否则全局宏适用于全篇*/
```

```
cd "$root" /*cd设置好工作路径后，以后调用.dta数据只需要 use filename 即可*/
```

```
use nerlove.dta
```

(2)其他兼容的数据类型 csv,txt, xlsx

```
clear
```

```
insheet using Training.csv
```

```
clear
```

```
insheet using Training.txt
```

xlsx文件(stata自身具有一定的数据格式转换功能)

```
clear
```

```
import excel Training.xlsx, sheet() firstrow //第一行为变量名
```

(3)复制粘贴

4、do文件的编辑

4.1 为什么要使用do文件

-图形化界面的局限：

- >命令不易保存、修改，软件关闭，命令即消失；
- >操作繁琐,每次操作都要不断重复点击界面；
- >功能组合有限，自由度低，不能进行软件开发。

-command& review 窗口的局限：

- >命令历史记录保存在Review窗口中，查找苦难；
- >零碎的命令没有条理，无法组织起复杂的操作；
- >与图形化界面类似，command窗口的命令也无法长期保存。

因此我们需要一个记录、编辑命令的编辑器，Stata自带的命令编辑器叫“do文件编辑器”，其功能类似txt文档，所生成的文件扩展名为.do，也就是do文件。

*4.2 do文件的基本编辑规则

do文件中的命令可以直接执行：选中，Ctrl+D

【修改工作路径】

clear all //do文件中的命令默认为蓝色，字符串为红色(双引号中)，变量、语法为黑色

cd "\$root"

use nerlove.dta //这是命令注释-设置好工作路径后直接使用 use filename 即可调取文件

4、do文件的编辑

*行首为星号的命令为绿色，表示不会被执行

如果没有星号，Stata会识别此行的命令、变量名称，如果识别失败则会报错

/*如果不想执行多行的命令

(如注释、说明)，

可以这样*/

或者用 //

注意：

(1)中英文字符的切换，尤其是逗号、引号

(2)stata 是大小写敏感的

(3)等于号==

(4)尽量避免使用系统预留字段作为变量名

5、录屏神器：log文件

log using "\$root\first.log" ///
using在这里表示新建log文件，在其他时候也可以表示打开日志文件
/*log using "\$root\first.log",append表示接着原来的日志记录
log using "\$root\first.log", replace表示覆盖原来的日志文件，重新记录/

```
matrix input a = (1,2\3,4)
matrix list a
matrix input b = (1,2\1,1)
matrix list b
```

```
log off //暂停录制
matrix c=a+b
log on //继续录制
```

```
matrix list c
```

```
log close //结束录制
```


6、基本操作与t检验

6.1 基本操作

```
clear
```

```
import excel "C:\Users\Sabrina\Desktop\econometrics\nerlove.xls", firstrow clear //导入数据
```

```
describe //审视数据
```

```
la data "Nerlove 1963 paper" //给数据集加标签
```

```
d
```

```
list TC Q //列举变量 TC 和 Q 的具体数值
```

```
list TC Q in 1/5 //列举变量 TC 和 Q 的前5个数据
```

```
summarize Q //查看Q的统计特征,也可以只用su
```

```
su Q if Q>=10000 //计算满足条件的子样本的统计指标
```

```
su Q,detail //可以得到更多更详细的统计指标
```

```
su //所有数据的统计特征
```

```
tabulate PL //显示变量PL的经验累积分布函数
```

```
pwcorr //显示所有变量两两相关性
```

```
pwcorr PL PF PK, sig star(.05) //sig 表示显示相关系数的显著性水平
```

6、基本操作与t检验

6.1 基本操作

hist Q ,width(1000) frequency //Q的直方图， 组宽1000， 纵坐标为频数

kdensity Q //连续的经验分布图

scatter TC Q //画TC和Q的散点图

gen n=_n

scatter TC Q, mlabel(n) mlabpos(6) //带标签的散点图

twoway (scatter TC Q)(lfit TC Q) // lfit 指的是 linear fit 将散点图和线性回归图画在一起

graph save scatter1 //保存为scatter1

twoway (scatter TC Q)(qfit TC Q) //散点图加二次回归线

graph save scatter2

graph combine scatter1.gph scatter2.gph

//生成新变量

g lnTC=log(TC)

g lnQ=log(Q)

g lnPL=log(PL)

g lnPF=log(PF)

g lnPK=log(PK)

g Q2=Q^2

g lnPLlnPK=lnPL*lnPK

6、基本操作与t检验

6.1 基本操作

clear all // 清空数据、变量

global root "C:\Users\Sabrina\Desktop\econometrics" // 利用全局宏变量设置根目录

cd "\$root" //设置工作路径

use Training, clear //调取数据文件

tab mostn train //列联表

collapse re74 re75 re78 ,by(train) //按照train分组并保留均值

ssc new

ssc hot

ssc install outreg2

search keyword

6、基本操作与t检验

6.2 t检验

i.单样本t检验

Ho: age 均值为25

use Training, clear

ttest age == 24 //默认置信度为95%，注意赋值符号与等于符号的区别

$t = (\bar{Y} - m) / \text{std dev}(\bar{Y})$

-Ybar act:已知为 【25.37079】

-m:由原假设, $m =$ 【24】

-std dev(Ybar)=总体标准差西格玛/sqrt(n), 西格玛: 【未知】

样本标准差 【S】 作为“西格玛”的估计量: $S^2 = \sum(Y_i - \bar{Y})^2 / (n-1)$

故std dev(Ybar)的估计量 【std error】 = S / \sqrt{n} , 计算为 【0.3428148】

gen ei2=(age-24)^2 //残差平方记为ei2

egen summation=total(ei2) //egen是gen的扩展

gen stdev=sqrt(summation/444)

tab stdev

gen stderr=stdev/sqrt(445)

tab stderr

综上, t的估计值为 $(25.37079 - 24) / \text{stderr} =$ 【3.998631】

6、 基本操作与t检验

ii.双样本均值t检验

use Training, clear

Ho:培训前， 处理组和控制组收入均值无差异

```
ttest re74 , by(train)
```

Ho:培训后， 处理组和控制组收入均值无差异

```
ttest re78 , by(train)
```

7、小建议

(1)文件夹：分类保存不同类型的文件

尽量使用英文做文件夹和文件名(Stata14以后对中文支持变强)

举例：RawData,Dofile,Logfile,

避免污染原始数据

(2)习惯利用do文件进行操作：编辑、修改更容易，操作可保存、复制

(3)定义宏变量：简化命令，方便修改

(4)充分利用搜索引擎Google(baidu很弱)、人大经济论坛、help文档等资源

(5)做好数据备份：Dropbox、onedrive、百度云等网盘，本地介质，甚至邮箱