Stata 入门教程：从0到1

第一次上机课

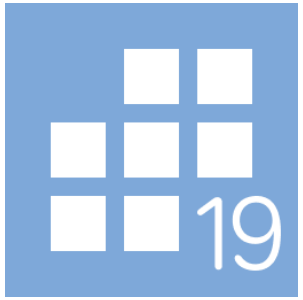
赵震宇（2120253538）
南开大学 国际经济研究所
zzynankai@outlook.com
xishanyu2.github.io
2025 年 10 月 16 日

Stata 简介与界面认识

Stata是一款功能强大的统计软件，广泛应用于经济学、社会学等领域（1.0于1985年1月发布）。

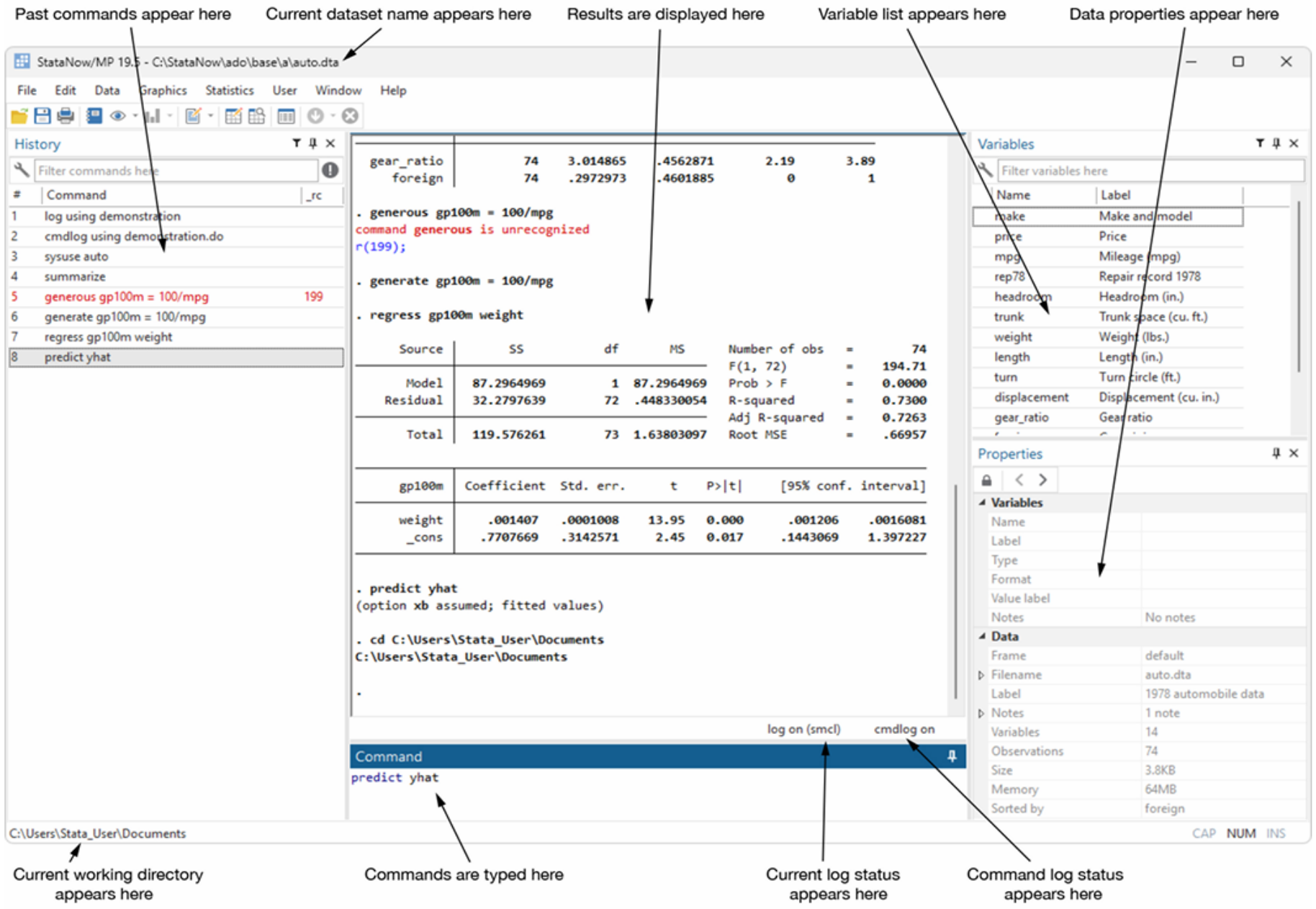
它以**命令行**为主，同时也提供图形化界面，非常适合进行数据管理、统计分析和可视化。

Version	Release date	
19.5	April 2025	StataNow
19	April 2025	Stata 19
18.5	April 2024	StataNow
18.0	April 2023	Stata 18
17.0	April 2021	Stata 17



启动Stata后，你会看到五个窗口：

- **命令窗口**：输入命令，按Enter键后执行，相应的结果会呈现在结果窗口。
- **结果窗口**：运行的命令以及运行的结果和错误都会在此展示。
  - 窗口中不同颜色的文本有不同的含义：
  - 黑色表示命令与结果、蓝色代表链接、红色表示错误。
- **历史窗口**：依次列出所有执行过的命令。
  - 单击后命令即被自动复制到命令窗口中；
  - 双击相应的命令则会重复执行。
- **变量窗口**：列出当前数据中的所有变量名称与标签。
- **属性窗口**：在点击变量窗口中某个变量后，可以显示该变量的具体性质，如变量的类型、格式、备注信息，以及数据集的信息。


Past commands appear here
Current dataset name appears here
Results are displayed here
Variable list appears here
Data properties appear here
StataNow/MP 19.5 - C:\StataNow\ado\base\a\auto.dta
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
Command _rc
1 log using demonstration
2 cmdlog using demonstration.do
3 sysuse auto
4 summarize
5 generous gp100m = 100/mpg 199
6 generate gp100m = 100/mpg
7 regress gp100m weight
8 predict yhat
gear_ratio 74 3.014865 .4562871 2.19 3.89
foreign 74 .2972973 .4601885 0 1
. generous gp100m = 100/mpg
command generous is unrecognized
r(199);
. generate gp100m = 100/mpg
. regress gp100m weight
Source SS df MS Number of obs = 74
F(1, 72) = 194.71
Model 87.2964969 1 87.2964969 Prob > F = 0.0000
Residual 32.2797639 72 .448330054 R-squared = 0.7300
Adj R-squared = 0.7263
Total 119.576261 73 1.63803097 Root MSE = .66957
gp100m Coefficient Std. err. t P>|t| [95% conf. interval]
weight .001407 .0001008 13.95 0.000 .001206 .0016081
_cons .7707669 .3142571 2.45 0.017 .1443069 1.397227
. predict yhat
(option xb assumed; fitted values)
. cd C:\Users\Stata_User\Documents
C:\Users\Stata_User\Documents
.
Variables
Filter variables here
Name Label
make Make and model
price Price
mpg Mileage (mpg)
rep78 Repair record 1978
headroom Headroom (in.)
trunk Trunk space (cu. ft.)
weight Weight (lbs.)
length Length (in.)
turn Turn circle (ft.)
displacement Displacement (cu. in.)
gear_ratio Gear ratio
Properties
Variables
Name
Label
Type
Format
Value label
Notes No notes
Data
Frame default
Filename auto.dta
Label 1978 automobile data
Notes 1 note
Variables 14
Observations 74
Size 3.8KB
Memory 64MB
Sorted by foreign
log on (smcl)
cmdlog on
Command
predict yhat
C:\Users\Stata_User\Documents
CAP NUM INS
Current working directory appears here
Commands are typed here
Current log status appears here
Command log status appears here

Stata命令的语法格式

```
[by varlist:] command [varlist] [=exp] [if exp] [in range] [weight] [using filename] [, options]
```

只有`command`（命令）是必不可少的，`[ ]`表示可以省略或者根据需要使用的选项。

`varlist`：变量列表，可以设置一个或者多个变量，多个变量之间用空格分隔开。

`by varlist`：按照变量值分类操作，表示对分类变量的子集分别执行相应的操作。

```
help by //Stata的内置帮助文档非常强大，在任何命令前加上help即可查看

/*
Syntax

        by varlist: stata_cmd

        bysort varlist: stata_cmd

    The above diagrams show by and bysort as they are typically used.  The full syntax of the commands is

        by varlist1 [(varlist2)] [, sort rc0]:  stata_cmd

        bysort varlist1 [(varlist2)] [, rc0]:  stata_cmd
*/

* by without the sort option requires that the data be sorted by varlist.
```

`=exp`：用来生成新变量或替换原变量的值，主要包括`generate`和`replace`两个命令。

`if exp`：条件表达式，用于对样本集进行筛选，只对符合条件的样本子集执行相应的操作。

`in range`：同样用于对样本集进行筛选，不依赖变量，而是直接作用于使用范围内的样本观测值。

`weight`：对样本观测值进行加权，通常用于加权最小二乘回归分析。

`options`：具体命令具体分析，注意一行命令只有一个逗号！

如果命令太长，可以使用`///`换行。

第一步：管理你的工作目录和日志

在开始分析前，养成良好的工作习惯！

- **设置工作目录**：告诉Stata你的数据文件和输出文件放在哪里。
  - **图形化操作**：`File` -> `Change Working Directory`
  - **命令**：在命令窗口中输入（将路径替换为你自己的文件夹路径）

```
cd "C:\Users\YourName\Desktop\StataProject"
```

- **开启日志**：记录你所有的操作和结果，便于复现和检查。

```
log using "my_first_analysis.log", replace
```

- 这会将之后所有操作记录到`my_first_analysis.log`文件中。
- `replace`选项表示如果文件已存在，则覆盖它。

- 结束工作时，**关闭日志**：

```
log close
```

第二步：数据管理

1. 导入数据

Stata的默认数据格式是`.dta`，但也可以导入Excel、CSV等格式的数据。

- **使用Stata格式数据：**

```
use "my_data.dta", clear
```

```
sysuse auto, clear
```

- **导入CSV文件：**

```
import delimited using "my_data.csv", clear
```

- **导入Excel文件：**

```
import excel using "my_data.xlsx", sheet("Sheet1") firstrow clear //firstrow表示将第一行作为变量名
```

2. 认识数据

导入数据后，先用一些基本命令了解数据的概况。

- **打开数据编辑器**（像Excel一样查看）：

```
browse
```

- **查看数据概览**：

```
describe //显示变量名、类型、格式等信息
```

- **查看数据内容**：

```
list //列出所有数据，如果数据很大，会刷屏
```

```
list in 1/10 //只列出前10行
```

- **查看变量编码**（对于分类变量很重要）：

```
codebook //显示变量的详细信息，包括取值和标签
```

3. 基本的变量操作

- 生成新变量:

```
generate bmi = weight / (height^2) //generate可简写为gen
generate age_sq = age^2
generate old = (age >= 60) if !missing(age) //(age >= 60)会生成一个取值为1(真)或0(假)的虚拟变量
*- if !missing(age)确保只在age不为缺失值时进行计算
```

- 修改变量:

```
replace income = income * 1.05 if year == 2025 //将2025年的income增加5%
```

- **重命名变量：**

```
rename income wage
```

- **删除变量：**

```
drop wage
```

- **给变量和取值添加标签**（非常好的习惯！）：

```
label variable income "家庭年收入（万元）"
label define gender_label 1 "男" 2 "女"
label values gender gender_label
```

第三步：基础统计分析

1. 生成摘要统计

```
summarize //对所有连续变量计算观测数、均值、标准差、最小值和最大值
summarize age income //只对age和income两个变量进行计算
```

2. 更详细的描述性统计

```
tabstat age income, stats(mean sd p50 min max n) col(stat)
```

- **单变量频数表：**

```
tabulate gender
```

- **交叉表：**

```
tabulate gender married //查看性别和婚姻状况的交叉分布
tabulate gender married, row col //显示行和列百分比
```

3. 绘图（略）

第四步：保存与退出

- 保存处理后的数据：

```
save "my_data_cleaned.dta", replace
```

- 退出Stata：

```
exit, clear //clear表示退出时不保存当前内存中的数据
```

下一步

当你掌握了以上基础后，可以继续学习：

- **Do-file**：编写脚本文件来执行一系列命令，保证分析的可重复性。
- **循环与函数**：使用`foreach`和`forvalues`等命令来处理重复性与复杂性任务。
- **结果导出**：使用`esttab`等命令将描述性统计结果导出为Word或Excel格式的表格。

.....

祝你Stata学习之旅顺利！

zzynankai@outlook.com

Bilibili：西山yu

xishanyu2.github.io