

第 2 章 Stata 入门

2.1 为什么使用 Stata

Stata 软件因操作简单且功能强大，为目前在欧美最流行的统计与计量软件，拥有众多用户。

Stata 公司定期升级软件，以适应计量经济学的迅猛发展。

Stata 软件还留有“用户接口”，允许用户自己编写命令与函数，并上传到网上实现共享。一些最新计量方法，可在线查找和下载由用户编写的 Stata 命令程序(user-written Stata commands)。这些“非官方命令”(也称“外部命令”)的使用方法与官方命令完全相同，使得 Stata 的功能如虎添翼。

本教材使用 Stata 13 版本(2013 年 6 月发布)。

对于绝大多数命令与功能，即使用更低的 Stata 版本(如 Stata 11 或 Stata 12)，也几乎没有差别。

2.2 Stata 的窗口

安装 Stata 13 后，在安装的文件夹中将出现如下 Stata 13 图标 (Stata 11 或 Stata 12 的图标大同小异)，参见图 2.1：



图 2.1 Stata 13 的图标

双击此 Stata 图标，即可打开 Stata。

如想在电脑桌面创建开启 Stata 软件的快捷方式，可右键点击 Stata 13 的图标，然后选择“发送到”→“桌面快捷方式”，参见图 2.2。

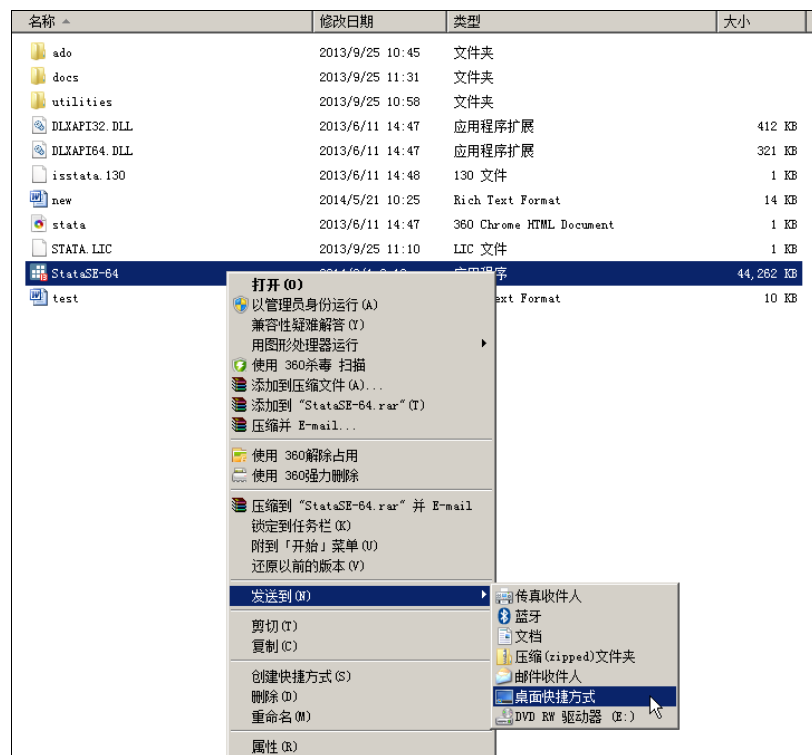


图 2.2 发送 Stata 13 到桌面快捷方式

打开 Stata 后可看到，在最上方有一排“下拉式菜单”(pull-down menu)，参见图 2.3：

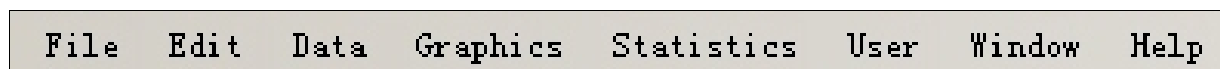


图 2.3 Stata 的下拉式菜单

在 Stata 中运行单个命令主要有两种方式，其一为点击菜单，其二为在“命令窗口”输入命令。

通过菜单执行命令(menu-driven)可能要点击多重菜单，通常还要填写对话框(dialog)，以明确命令参数，不如在命令窗口直接输入命令方便。

在菜单之下，为一系列图标，起着快捷键的作用，参见图 2.4。

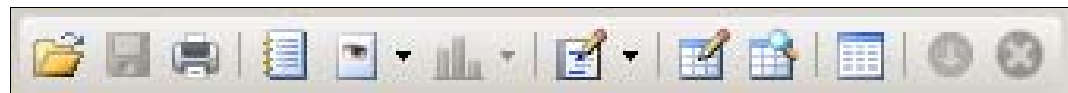
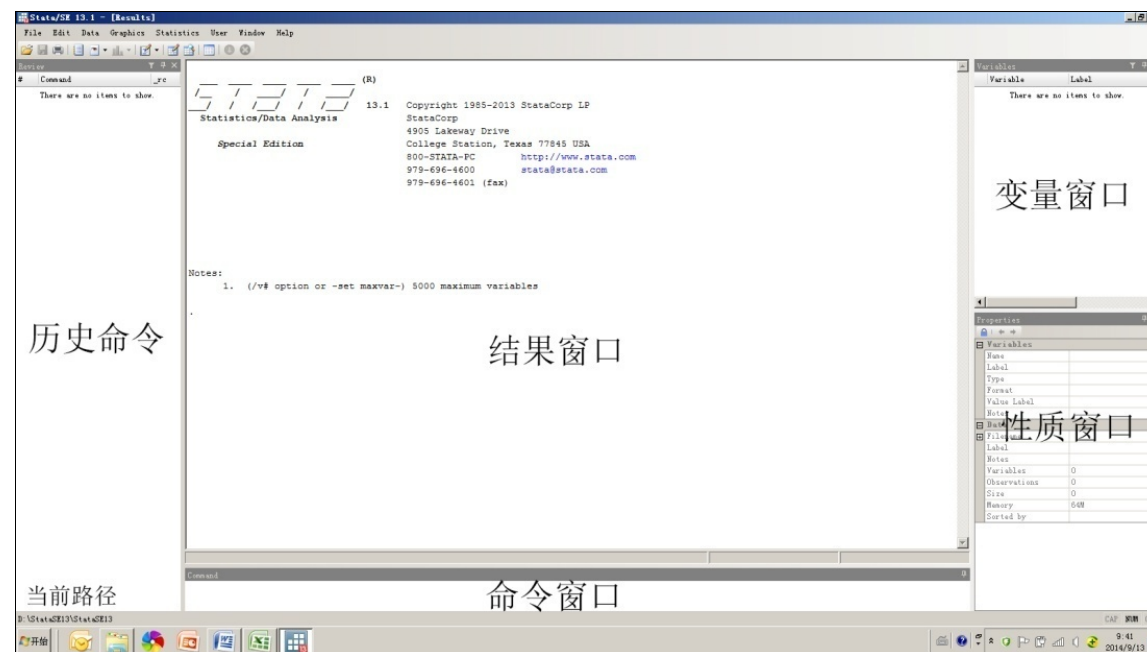


图 2.4 Stata 的快捷键

在快捷键图标之下，有五个窗口，参见图 2.5。



2.3 Stata 操作实例

以数据集 `grilic_small.xls` (Excel 文件)为例, 该文件包含 30 名美国年轻男子的教育投资回报率数据。

1. 导入数据

首先, 打开 Stata 软件, 点击快捷键 Data Editor (Edit)图标(参见图 2.6), 即可打开 Stata 的数据编辑器, 参见图 2.7。

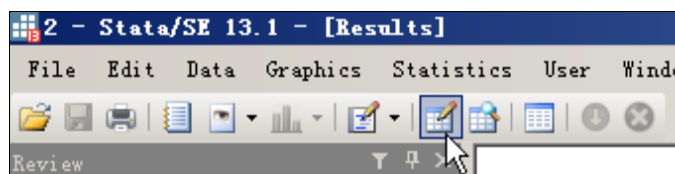


图 2.6 Data Editor (Edit)图标

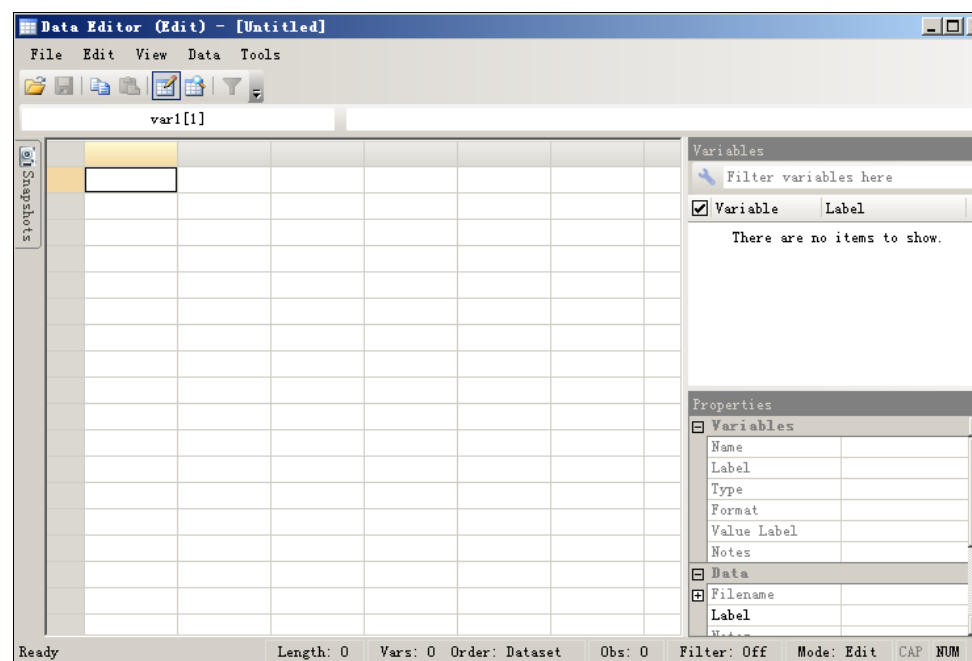


图 2.7 Stata 的数据编辑器

其次，用 Excel 打开文件 “grilic_small.xls”，会看到如下 Excel 格式的数据文件：

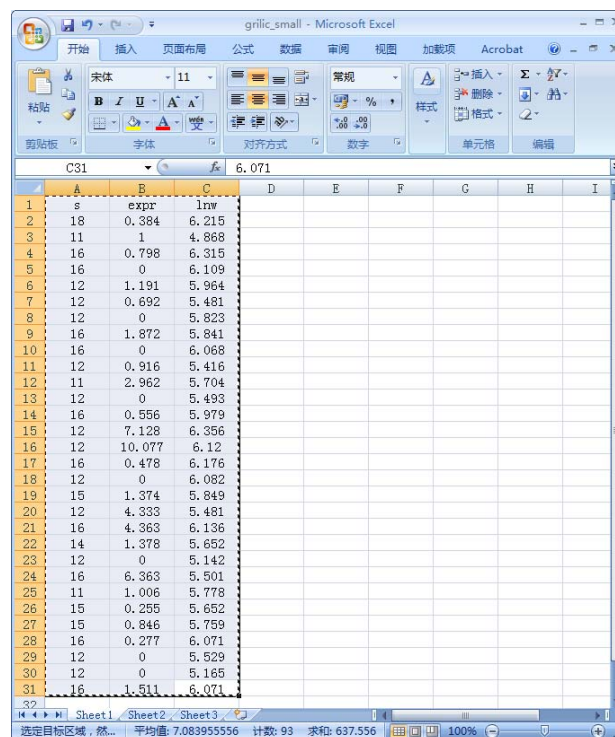


图 2.8 Excel 表中的数据

共有 3 列变量, 分别为 s (schooling, 教育年限), expr (experience, 工龄)与 lnw (lnwage, 工资对数)。

复制此 Excel 表中所有数据(Ctrl + C),粘贴到 Data Editor 中(Ctrl + V)。在 Data Editor 中会出现对话框, 参见图 2.9:

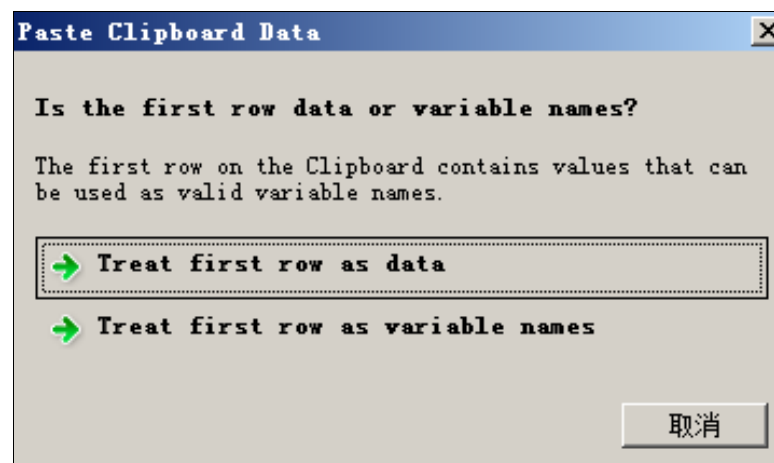


图 2.9 Data Editor 的对话框

此对话框问你“第一行为数据还是变量名”, 点击相应选择即可。

导入数据的另一方法是(特别在数据量很大的情况下), 点击菜单“File” → “Import”, 然后导入各种格式的数据, 参见图 2.10; 但不如直接从 Excel 表中粘贴数据更为方便。

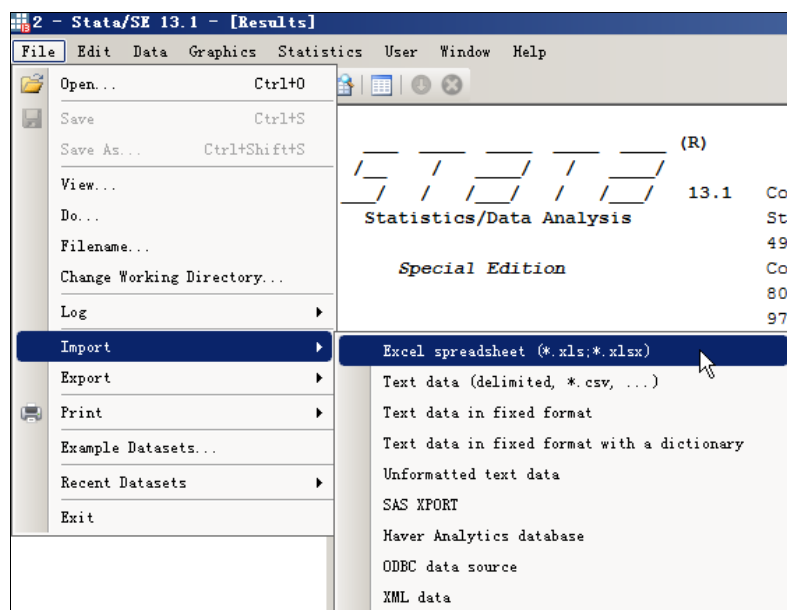


图 2.10 使用 Import 导入数据

关闭 Data Editor (Edit)后, 会看到右上方的变量窗口出现了 3 个变量, 分别为 s, expr 与 lnw。

点击快捷键 Save 图标(参见图 2.11 中鼠标位置, 也可点击菜单“File”→“Save”), 将数据存为 Stata 格式的数据文件(扩展名 dta, 为 data 的缩写), 比如 grilic_small.dta。

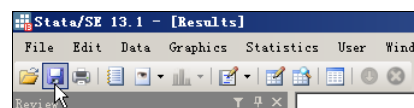


图 2.11 Save 图标

此后可用 Stata 直接打开 grilic_small.dta, 无须再从 Excel 中导入数据。

打开 Stata 数据集的方式有两种。方法之一，点击快捷键 Open 图标(参见图 2.12)，寻找要打开的 dta 文件位置。

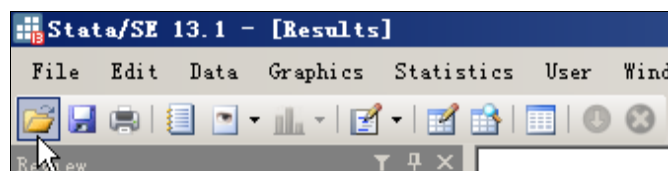


图 2.12 Open 图标

方法之二，在命令窗口输入以下命令(假设文件 `grilic_small.dta` 在 E 盘的根本目录)，然后回车(按 Enter 键)：

```
. use E:\grilic_small.dta,clear
```

逗号 “,” 之后的 “clear” 为 “选择项” (option)，表示可替代内存中的已有数据。

使用命令 `use` 打开 `dta` 数据文件，需输入此文件的路径；一般不如使用快捷键 `Open` 寻找此文件更为方便。

如要关闭一个数据集，以便使用另外一个数据集，可输入命令
`. clear`

内存中数据将被清空，然后可再打开另一数据集。

2. 变量的标签

在变量窗口，变量的“名字”(Name)旁边会显示其“标签”(label)。点击 `Variables Manager` 图标(参见图 2.13)，即可打开变量管理器，然后编辑变量名、标签等。

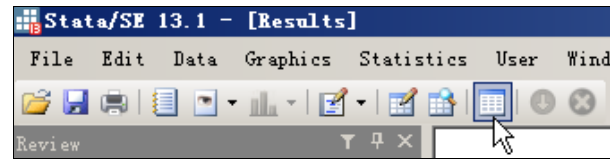


图 2.13 Variables Manager 图标

比如，将变量 s 的标签改为“schooling”，然后点击“Apply” (应用)，参见图 2.14。

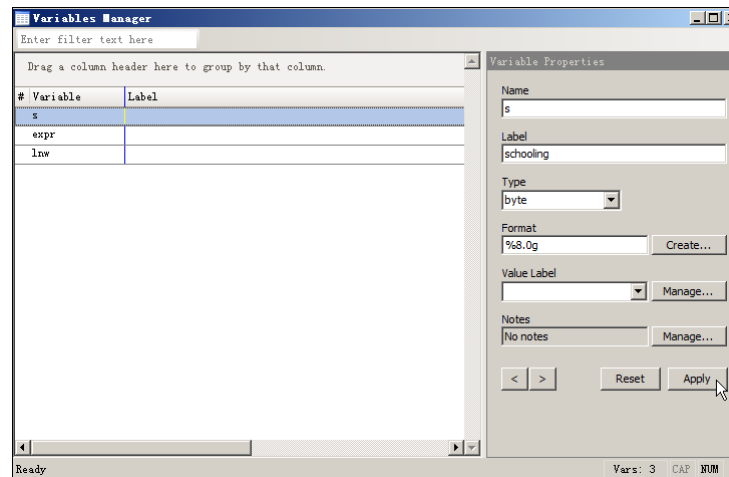


图 2.14 变量管理器的对话框

Stata 严格区分大小写字母(case sensitive)。建议变量名使用小写字母，便于阅读。

3. 审视数据

如想看数据集中的变量名称、标签等，可输入命令

`. describe`

其中，“describe”的下划线表示，可将该命令简写为“d”。

Contains data				
obs:	30			
vars:	3			
size:	270			
variable name	storage type	display format	value label	variable label
s	byte	%8.0g		schooling
expr	float	%8.0g		
lnwage	float	%8.0g		
Sorted by:				
Note: dataset has changed since last saved				

如想看变量 s 与 lnw 的具体数据，可使用命令
`. list s lnw`

	s	lnw
1.	18	6.215
2.	11	4.868
3.	16	6.315
4.	16	6.109
5.	12	5.964
6.	12	5.481
7.	12	5.823
8.	16	5.841
9.	16	6.068
10.	12	5.416
11.	11	5.704
12.	12	5.493
13.	16	5.979
14.	12	6.356
15.	12	6.12
16.	16	6.176
17.	12	6.082
18.	15	5.849
19.	12	5.481
20.	16	6.136
21.	14	5.652
22.	12	5.142
23.	16	5.501
24.	11	5.778
25.	15	5.652

more

在屏幕底端出现带下划线的英文字“more”，用鼠标单击“more”，可翻看下页的结果。

如想连续滚屏显示命令运行结果，可输入命令

```
. set more off
```

如又想恢复分页显示运行结果，可输入命令

```
. set more on
```

如只想对数据集的一部分子集执行命令，比如只看 s 与 lnw 的前 5 个数据，可使用命令

```
. list s lnw in 1/5
```

	s	lnw
1.	18	6.215
2.	11	4.868
3.	16	6.315
4.	16	6.109
5.	12	5.964

如要罗列从第 11-15 个观测值，可输入命令

```
. list s lnw in 11/15
```

	s	lnw
11.	11	5.704
12.	12	5.493
13.	16	5.979
14.	12	6.356
15.	12	6.12

也可通过逻辑关系来定义数据集的子集。比如，要列出所有满足条件 “ $s \geq 16$ ” (教育年限为 16 年及以上)的数据，可使用命令

```
. list s lnw if s>=16
```

	s	lnw
1.	18	6.215
3.	16	6.315
4.	16	6.109
8.	16	5.841
9.	16	6.068
13.	16	5.979
16.	16	6.176
20.	16	6.136
23.	16	5.501
27.	16	6.071
30.	16	6.071

“>=”表示“大于等于”。其他表示关系的逻辑符号为“==”(等于),“>”(大于),“<”(小于),“<=”(小于等于),“≠”(不等于,也可用“!=”表示)。

一个等号“=”表示“赋值”，而两个等号“==”表示“等于”。

查看具体数据的直接方法是，点击 Data Editor (Edit)图标，或右边的 Data Editor (Browse)图标，参见图 2.15。二者的区别在于，Browse 只能看，不能改；而 Edit 还可改数据。

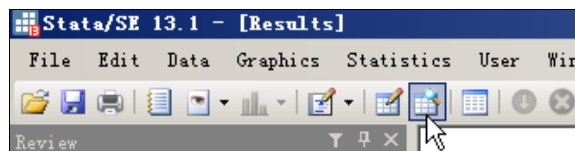


图 2.15 Data Editor (Browse)图标

如要删除满足“ $s \geq 16$ ”条件的观测值，可输入命令

```
. drop if s>=16
```

反之，如只想保留满足“ $s \geq 16$ ”条件的观测值，可使用命令

```
. keep if s>=16
```

删除观测值之后，Stata 不提供类似于 Microsoft Word 的撤销 (undo) 命令。一般建议慎重删除数据，最好先将原始数据备份。

如想将数据按照变量 s 的升序排列，可输入命令

```
. sort s
```

```
. list
```

	s	expr	lnw
1.	11	1	4.868
2.	11	1.006	5.778
3.	11	2.962	5.704
4.	12	0	6.082
5.	12	0	5.529
6.	12	0	5.823
7.	12	7.128	6.356
8.	12	0	5.493
9.	12	0	5.165
10.	12	10.077	6.12
11.	12	.916	5.416
12.	12	4.333	5.481
13.	12	.692	5.481
14.	12	0	5.142
15.	12	1.191	5.964
16.	14	1.378	5.652
17.	15	.255	5.652
18.	15	.846	5.759
19.	15	1.374	5.849
20.	16	0	6.109
21.	16	6.363	5.501
22.	16	1.511	6.071
23.	16	0	6.068
24.	16	.478	6.176
25.	16	.277	6.071
26.	16	4.363	6.136
27.	16	1.872	5.841
28.	16	.798	6.315
29.	16	.556	5.979
30.	18	.384	6.215

命令 `sort` 无法按照变量的降序排列。如想按降序排列，可使用命令 `gsort`:

```
. gsort -s
```

```
. list
```

	s	expr	lnw
1.	18	.384	6.215
2.	16	.556	5.979
3.	16	.798	6.315
4.	16	1.872	5.841
5.	16	4.363	6.136
6.	16	.277	6.071
7.	16	.478	6.176
8.	16	0	6.068
9.	16	1.511	6.071
10.	16	6.363	5.501
11.	16	0	6.109
12.	15	1.374	5.849
13.	15	.846	5.759
14.	15	.255	5.652
15.	14	1.378	5.652
16.	12	1.191	5.964
17.	12	0	5.142
18.	12	.692	5.481
19.	12	4.333	5.481
20.	12	.916	5.416
21.	12	10.077	6.12
22.	12	0	5.165
23.	12	0	5.493
24.	12	7.128	6.356
25.	12	0	5.823
26.	12	0	5.529
27.	12	0	6.082
28.	11	2.962	5.704
29.	11	1.006	5.778
30.	11	1	4.868

4. 画图

看数据的最直观方法是画图。想看变量 *s* 的分布情况，可输入以下命令画直方图(参见图 2.16):

```
. histogram s, width(1) frequency
```

“histogram”表示直方图。

选择项 “width(1)” 表示将组宽设为 1(否则将使用 Stata 根据样本容量计算的默认分组数),

选择项 “frequency” 表示将纵坐标定为频数(默认使用密度)。

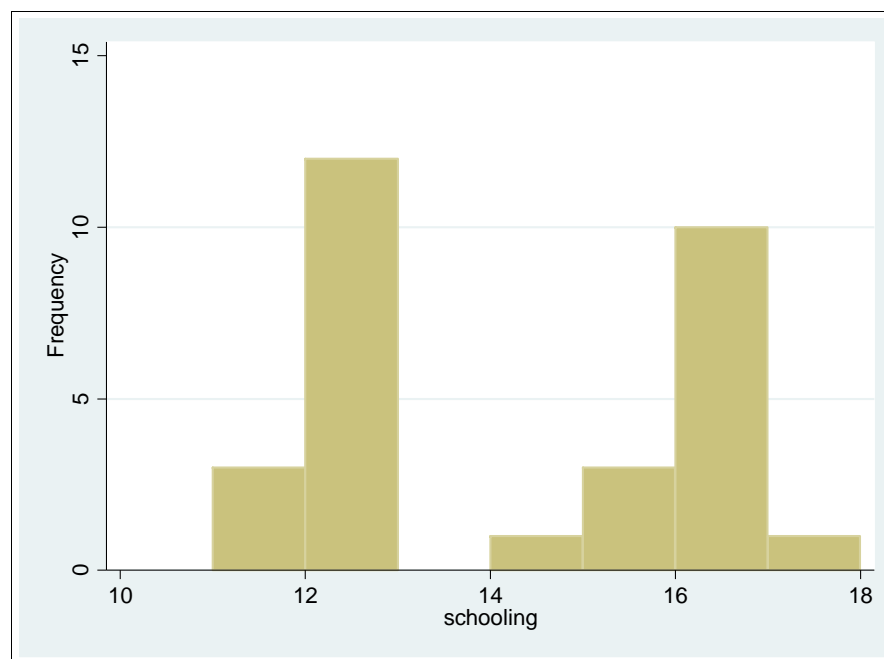


图 2.16 教育年限的直方图

教育年限的分布呈双峰状,受 12 年教育的人数最多(高中毕业),其次为受 16 年教育者(大学毕业)。

如想知道更多有关命令 `histogram` 选项与用法，可输入命令

```
. help histogram
```

对于任何 Stata 命令，只要输入 “`help command_name`” 即可查看该命令的 “帮助文件” (help file)。

如想考察教育年限与工资对数之间的关系，最直观方法是画 `s` 与 `lnw` 之间的散点图，可输入命令(参见图 2.17):

```
. scatter lnw s
```

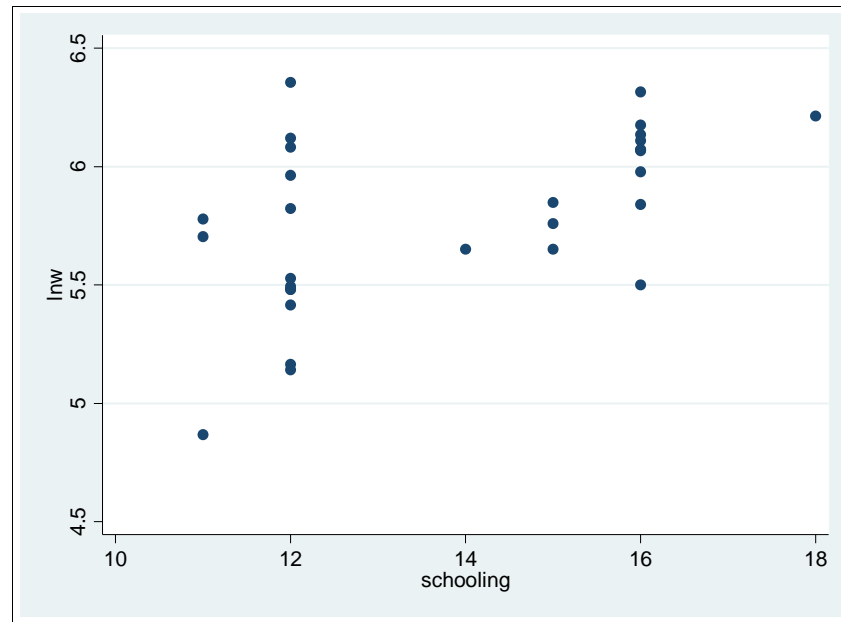


图 2.17 教育年限与工资对数的散点图

工资对数与教育年限似乎存在正相关关系。

如想在散点图上标注出每个点对应于哪个观测值，可先定义变量 n ，表示第 n 个观测值：

```
. gen n=_n
```

“_n”表示第 n 个观测值。然后以变量 n 作为每个点的标签来画散点图，参见图 2.18。

```
. scatter lnw s,mlabel(n)
```

选择项“mlabel(n)”表示，以变量 n 作为标签(mark label)。

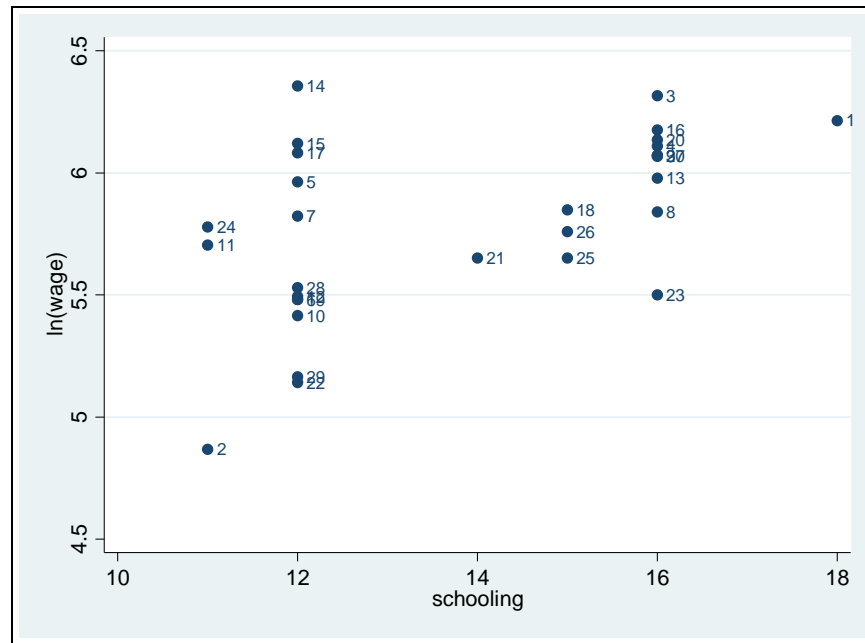


图 2.18 加标签的散点图

Stata 有丰富的作图方法。更多作图方法，参见下拉式菜单“Graphics” (参见图 2.19)。

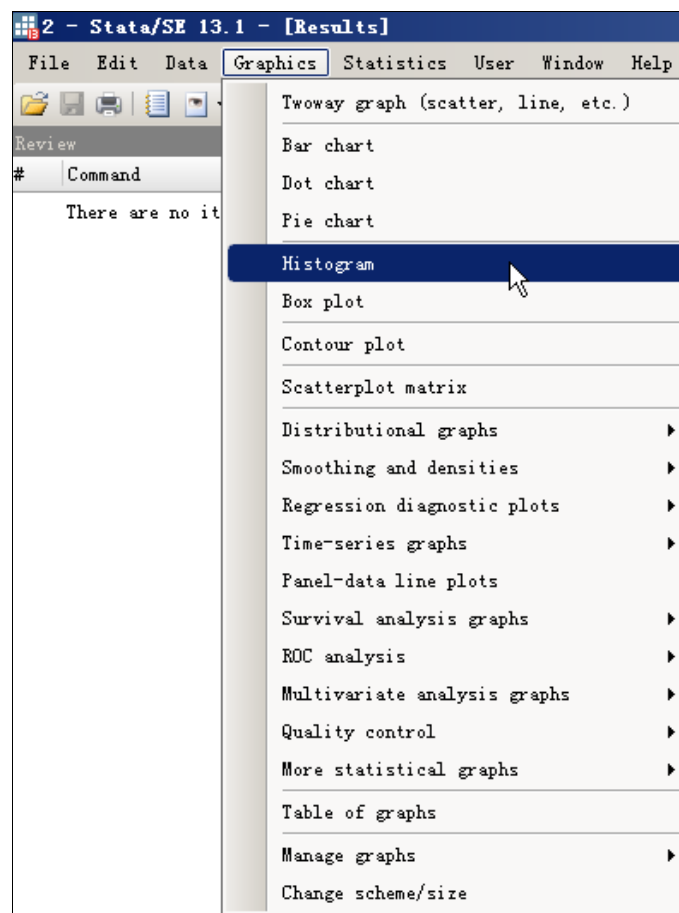


图 2.19 Stata 的作图功能

5. 统计分析

如想看变量 s 的统计特征，可输入命令

```
. summarize s
```

Variable	Obs	Mean	Std. Dev.	Min	Max
s	30	13.8	2.139932	11	18

此结果显示变量 s 的样本容量、平均值、标准差、最小值与最大值。如不指明变量，则显示所有变量的统计指标。

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
s	30	13.8	2.139932	11	18
expr	30	1.658667	2.445213	0	10.077
lnw	30	5.7932	.3679956	4.868	6.356

如要显示变量 s 的经验累积分布函数(empirical cumulative distribution function)，可使用命令

```
. tabulate s
```

schooling	Freq.	Percent	Cum.
11	3	10.00	10.00
12	12	40.00	50.00
14	1	3.33	53.33
15	3	10.00	63.33
16	10	33.33	96.67
18	1	3.33	100.00
Total	30	100.00	

“Freq”表示频数，“Percent”表示百分比，而“Cum.”表示累积百分比。

如要显示工资对数、教育年限、工龄之间的相关系数，可输入命令

```
. pwcorr lnw s expr, sig star(.05)
```

“pwcorr”表示“pairwise correlation”(两两相关)，“sig”表示显示相关系数的显著性水平(即 p 值，列在相关系数的下方)。

“star(.05)”表示给所有显著性水平小于或等于 5% 的相关系数打上星号。

	lnw	s	expr
lnw	1.0000		
s	0.5368* 0.0022	1.0000	
expr	0.2029 0.2823	-0.1132 0.5514	1.0000

lnw 与 s 的相关系数为 0.5368，且在 1%水平上显著(p 值为 0.0022)。

lnw 与 expr 的相关系数也达到 0.2029，但不显著(p 值为 0.2823，可能因为样本容量较小，仅为 30)。

s 与 expr 的相关系数为-0.1132，可能因为上学时间长的年轻人，参加工作时间就不长，但此负相关关系也不显著(p 值为 0.5514)。

6. 生成新变量

在 Stata 中定义新变量，可通过命令 `generate` 来实现。

比如，输入如下命令可定义教育年限的对数。

```
. generate lns=log(s)
```

如要定义 s 的平方项，可使用命令

```
. gen s2=s^2
```

如要生成 s 与 $expr$ 的互动项(interaction term)，可输入命令

```
. gen exprs=s*expr
```

如想根据工资对数 $\ln w$ 计算工资水平 w ，可使用命令

```
. gen w=exp(lnw)
```

在计量经济学中，常使用“虚拟变量” (dummy variable，也称“哑变量”)，即取值只能为 0 或 1 的变量，比如性别。

假设定义“ $s \geq 16$ ”为“受过高等教育”，并使用变量 `college` 来表示：

$$\text{college} \equiv \begin{cases} 1, & \text{如果 } s \geq 16 \\ 0, & \text{其他} \end{cases} \quad (2.1)$$

可使用如下命令

```
. gen colleg=(s>=16)
```

括弧“()”表示对括弧中的表达式“ $s \geq 16$ ”进行逻辑评估：如果此式为真，则取值为 1；如果为假，则取值为 0。

在上面命令中，不慎把 college 打成 colleg 了。可使用如下命令将变量重新命名：

```
. rename colleg college
```

变量 colleg 被重新命名为 college (也可使用变量管理器)。

如想将“受过高等教育”的定义改为“ $s \geq 15$ ”，但仍用 college 作为变量名。

方法之一，去掉现有变量 college，再重新定义一次：

```
. drop college
```

```
. gen college=(s>=15)
```

方法之二，只需一个命令：

```
. replace college=(s>=15)
```

此命令直接将原变量($s \geq 16$)替换为新变量($s \geq 15$)。

对于较长的变量名，输入变量名较麻烦。有如下三个简便方法。

方法一，直接在变量窗口双击需要的变量，该变量名就会出现在命令窗口。

方法二，如有以下变量 s1, s2, s3, s4, s5(比如，对教育年限的 5 种度量方法)，可用 s1-s5 来表示这 5 个变量。

方法三，用 “*” 号来简化变量名的书写。假设有将内存中所有以 “s” 开头的变量都去掉，可输入命令

```
. drop s*
```

这将去掉内存中的 s1, s2, s3, s4, s5 变量(删除之后无法恢复，故应慎重使用)。

7. Stata 的计算器功能

Stata 也可作为计算器使用，命令格式为 “display expression”。

比如，计算 $\ln 2$ ，可输入如下命令

```
. display log(2)
```

```
.69314718
```

如要计算 $\sqrt{2}$ ，则可输入命令

```
. dis 2^0.5
```

```
1.4142136
```

8. 调用命令与终止命令

如果每次都完整地输入整行命令，可能较费时。

较有效率的方法是，调用某个曾经使用过的命令，并在此基础上修改。调用旧命令的方法有两种。

方法一，把光标放在命令窗口，按键盘上的“Pg Up”键调用上一条命令，按“Pg Dn”键调用下一条命令。

方法二，在历史命令窗口单击旧命令，将旧命令调入命令窗口，然后进行编辑；如果用鼠标双击旧命令，将再次执行此旧命令。

有时运行某个命令费时较长(比如, 在数值计算时, 迭代无法收敛)。

如想中途停止该命令的执行, 可点击快捷键 Break 图标(参见图 2.20), 或直接在键盘上同时按 “Ctrl + Break”。

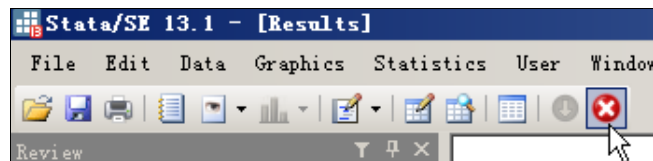


图 2.20 Break 图标

9. Stata 的日志

如希望在每次使用 Stata 时, 储存其运行结果, 可点击菜单“File”→ “Log” → “Begin” 定义 “日志文件” (log file), 参见图 2.21。

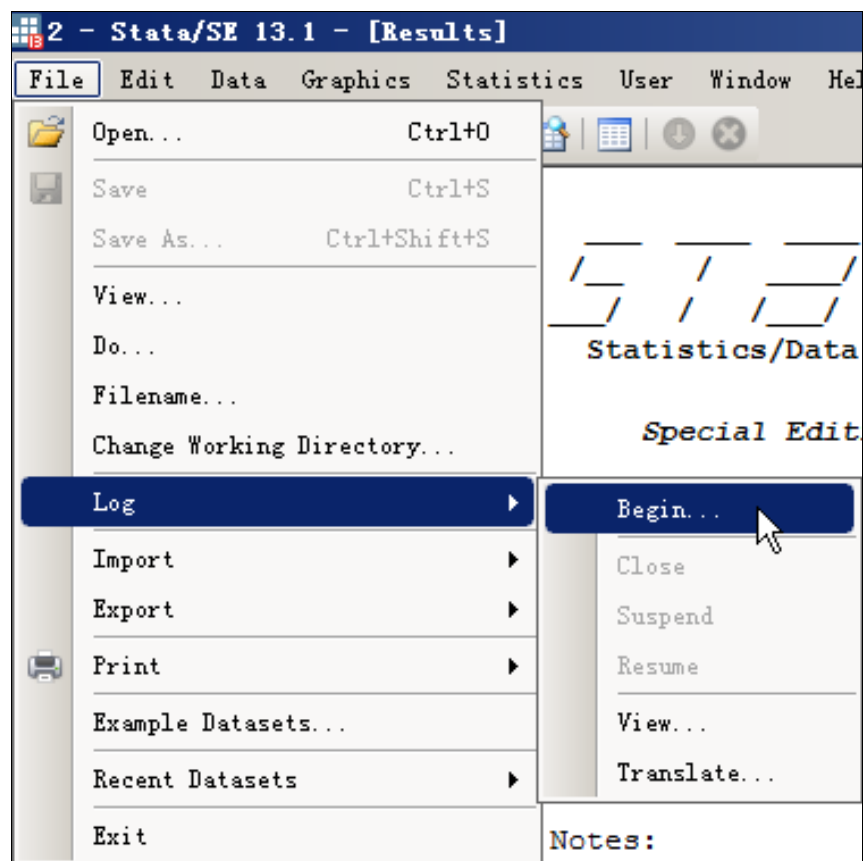


图 2.21 定义日志文件

也可直接点击快捷键 Log 图标，参见图 2.22。

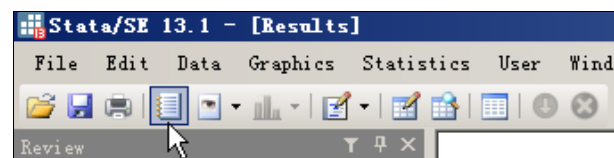


图 2.22 Log 图标

此时会出现如下对话框，参见图 2.23。在对话框中输入日志的文件名，并存储在指定的位置即可：

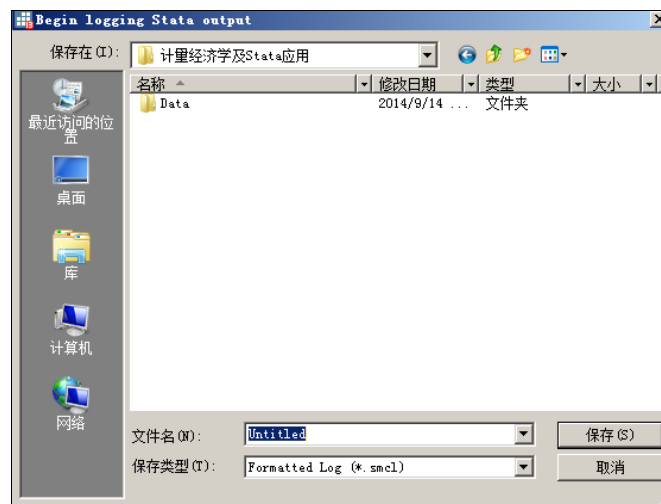


图 2.23 日志文件的对话框

Stata 日志文件的扩展名为 smcl。

也可直接在命令窗口输入如下命令：

```
. log using today
```

```
name: <unnamed>  
log: D:\StataSE13\StataSE13\today.smcl  
log type: smcl  
opened on: 15 Sep 2014, 14:00:15
```

在当前路径就会生成一个名为“today.smcl”的日志文件。

定义日志文件后，在 Stata 中的所有操作及结果，都将记录在日志中，直至退出此日志文件。

如要暂时关闭日志(不再记录输出结果), 可输入命令

```
. log off
```

如要恢复使用日志, 可输入命令

```
. log on
```

如要彻底退出日志, 则可输入命令

```
. log close
```

如要查看日志文件的内容, 可点击菜单 “File” → “Log” → “View”, 然后寻找此日志文件, 参见图 2.24。

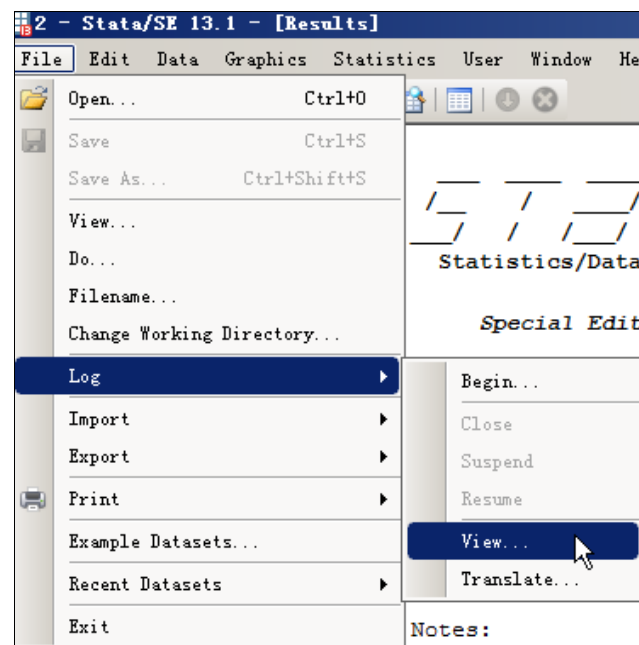


图 2.24 查看日志文件

2.4 Stata 命令库的更新

由于 Stata 版本不同(即使同为 Stata 13),如果发现极少数命令无法运行,可在命令窗口输入,


```
. update all
```

这将更新你的 Stata 命令库(Stata“ado”文件及其他可执行文件)。

Stata 用户还写了大量的外部命令或非官方命令(user-written software)，可直接下载到 Stata 中使用。

最流行的 Stata 非官方命令下载平台为“统计软件成分”(Statistical Software Components, SSC)，由 Boston College 维护，网址为 <http://ideas.repec.org/s/boc/bocode.html>。

从 SSC 下载 Stata 程序的命令为:

```
. ssc install newcommand
```

所有下载与安装过程都将自动完成(包括新命令的帮助文件)。

如非官方命令不是来自 SSC，一般需手工安装，将所有相关文件下载到指定的 Stata 文件夹中即可(通常为 ado\plus\)。

如不清楚应把文件复制到哪个文件夹，可输入以下命令，显示 Stata 的系统路径(system directories):

```
. sysdir
```

会看到类似于以下的结果(取决于 Stata 的安装位置),

```
STATA:  D:\StataSE13\StataSE13\  
BASE:   D:\StataSE13\StataSE13\ado\base\  
SITE:   D:\StataSE13\StataSE13\ado\site\  
PLUS:   c:\ado\plus\  
PERSONAL: c:\ado\personal\  
OLDPLACE: c:\ado\
```

将下载的新命令文件复制到 PLUS 所指示的那个文件夹即可(此处为 “c:\ado\plus\”)。

如想使用某种估计方法，不知道它是否存在，可输入命令

```
. search keyword
```

此命令将搜索 Stata 帮助文件、Stata 常见问题、Stata 案例、*Stata Journal*, *Stata Technical Bulletin* 等。

进一步的搜索可输入以下命令

```
. findit keyword
```

命令 `findit` 的搜索范围比命令 `search` 更广，还包括 Stata 的网络资源。事实上，“`findit`”等价于“`search,all`”。

2.5 进一步学习 Stata 的资源

更多 Stata 知识,将在本书以后章节中逐步介绍。Stata 英文参考书包括 Baum(2006), Cameron and Trivedi(2010), 以及 Stata 出版社(Stata Press)出版的系列书籍。加州大学洛杉矶分校(UCLA)网站(<http://www.ats.ucla.edu/stat/stata/>)有大量 Stata 的资源及实例(搜索“Stata UCLA”即可找到此网站)。

中文参考书包括陈传波《Stata 十八讲》,陈强(2014), 胡咏梅(2010), 兰草(2012), 劳伦斯·汉密尔顿(2008), 李春涛、张璇(2009), 王群勇(2007, 2008), 王天夫、李博柏(2008), 杨菊华(2012), 张鹏伟、李嫣怡(2011)等。

Stata 本身的“帮助”(Help)菜单包含了详细的使用说明，比如，“help histogram”。

更高级的学习，可查看 Stata 手册(Stata manuals)，这些手册对每个 Stata 命令都进行了详尽的说明。