Stata 软件入门教程

李昂然
浙江大学社会学系
Email: angranli19@gmail.com

版本：2020/02/05

1. 导论

本教程将快速介绍 Stata 软件（版本 16）的一些基本操作技巧和知识。对于详细的 Stata 介绍和入门，小伙伴们可以参考 Stata 官方的英文手册以及教程所提供的学习资料。跟其他大多数统计软件一样，Stata 可以同时通过下拉菜单以及命令语句来操作。初学者可以通过菜单选项来逐步熟悉 Stata，但是命令语句的使用是 Stata 用户的最佳选择。因此，本教程将着重介绍命令语句的使用。

对于中文用户来讲，在打开 Stata 之后，可以通过下拉菜单选项中的用户界面语言选择将中文设置为默认语言。同时，也可以在命令窗口中输入 `set locale ui zh_CN` 来设置中文显示。在选择完语言后，记得重新启动 Stata。需要提醒大家，虽然 Stata 用户界面可以显示中文，但是统计分析的结果仍然将以英文显示。

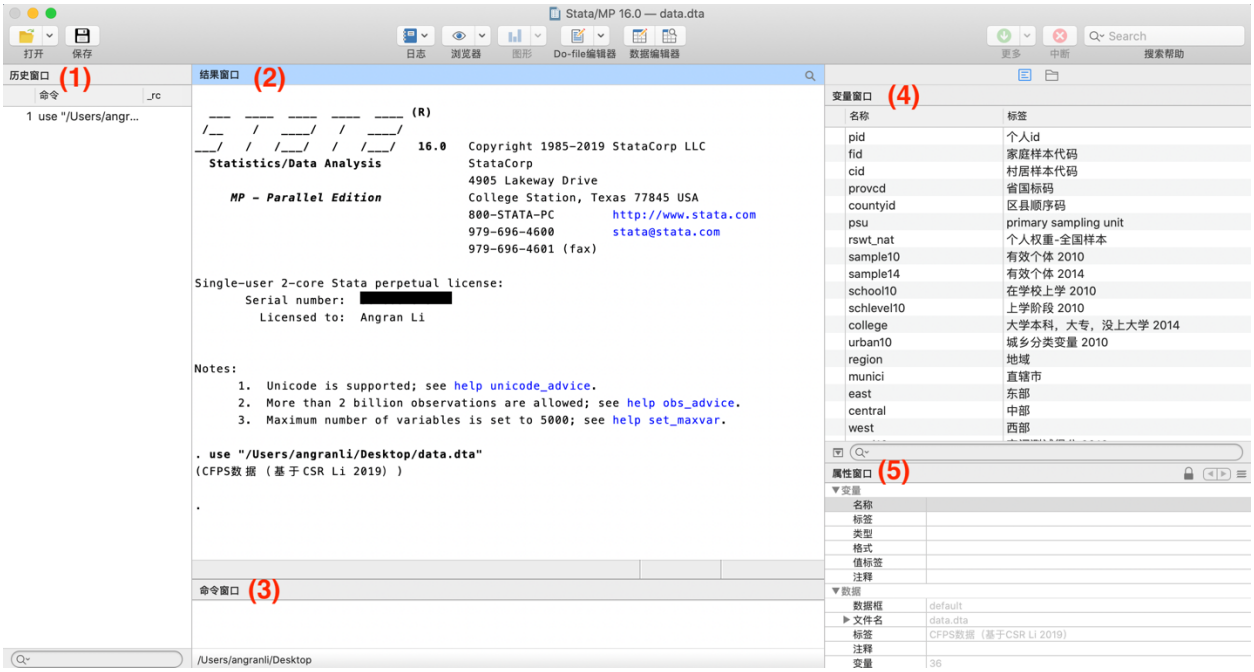
本教程中使用的案列数据源自**中国家庭追踪调查（China Family Panel Studies）**。具体数据出自本人于 2019 年发表于 *Chinese Sociological Review* 上“Unfulfilled Promise of Educational Meritocracy? Academic Ability and China’s Urban-Rural Gap in Access to Higher Education”一文中使用的数据。关于数据的具体问题，请联系本人。

同时，本教程提供相应的 do file 和数据文件给同学们下载，同学们可以根据 do file 复制本教程的全部内容。下载地址为我个人网站：https://angranli.me/teaching/

温馨提示：关于 Stata 操作的大多数疑问，都可以在官方手册上找到答案。同时，在 Stata 中输入 `help [command]` 便可以查看关于命令使用的详细信息。

2. Stata 用户界面

Stata 的用户操作界面主要包括五个部分：（1）历史窗口、（2）结果窗口、（3）命令窗口、（4）变量窗口、（5）属性窗口。



以上为 Mac 版本 Stata 的用户界面。

*（1）历史窗口*显示已经执行过的命令语句，可以帮助用户回顾已经使用过的命令。如果命令输入错误，窗口中会显示红色的提示。

当遇到错误提示，仔细检查一下 code。大多数时候，对命令语句的熟悉是在不断犯错中加深的，所以，遇到错误提示，请淡定。

*（2）结果窗口*显示数据分析的结果。如果想要清除当前显示的结果，可以点击结果窗口，会出现菜单选项，然后进行选择。

*（3）命令窗口*的使用很直接，无需介绍太多。记住一点，当从别处复制语句并粘贴到命令窗口时，如果语句有多行，Stata 会逐行读取，如果语句中断，会出现错误提示。

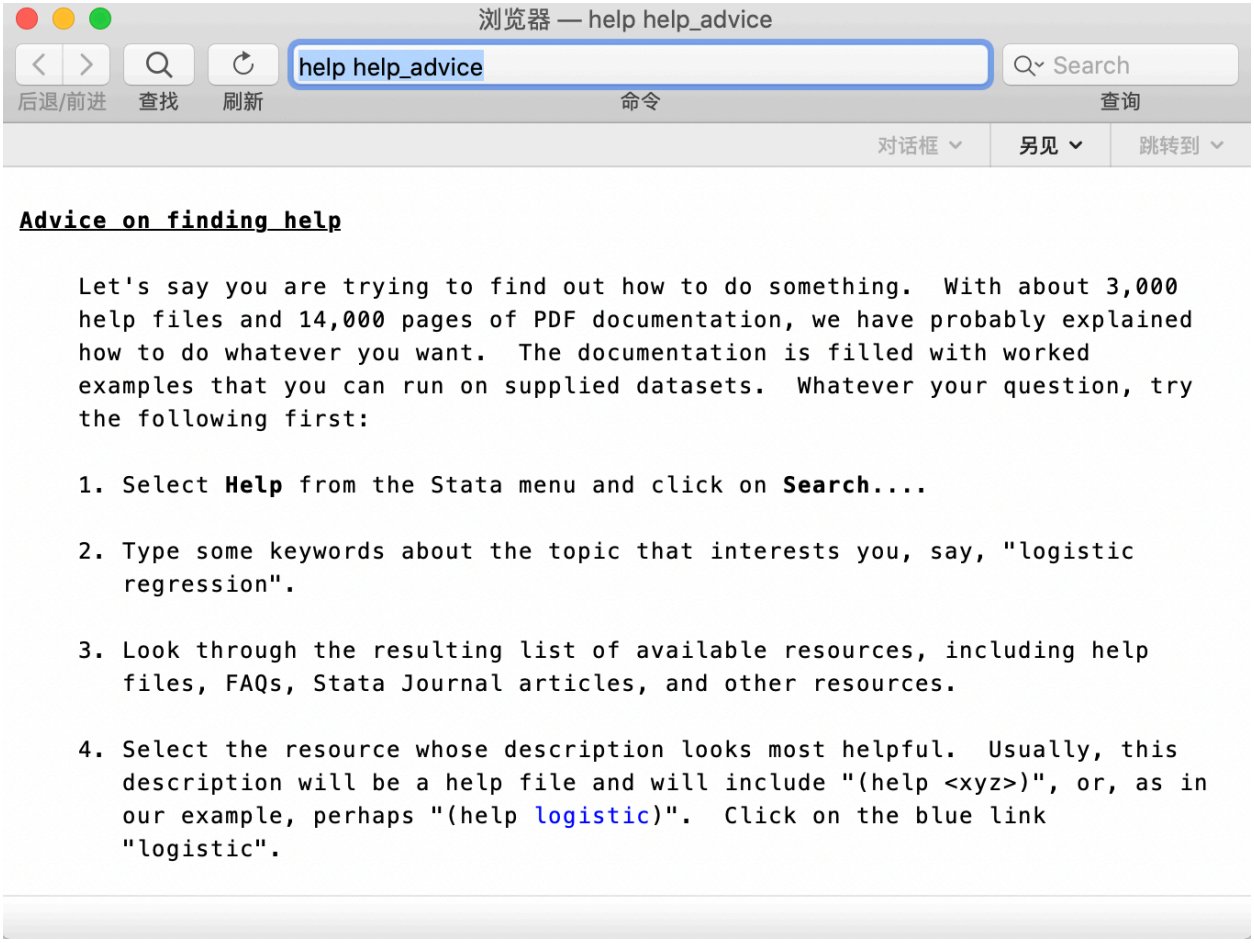
*（4）变量窗口*显示数据中变量的基本信息。当点击变量左边的按钮时，变量的名称会自动出现在命令窗口。

*（5）属性窗口*显示更多关于数据和变量的详细信息。可以在此处手动对变量的信息进行修改。

3. Stata 浏览器以及 Help 功能

Stata 浏览器的作用类似搜索引擎，提供对于 Stata 的全方位帮助。主要有两个作用：1）通过游览器，可以查询 Stata 用户手册中的各种信息。2）游览器可以连接网络，在网络中下载并安装其他骨灰级用户自己编写的命令语句（community-contributed commands）。

开启浏览器可以点击工具栏中的浏览器按钮，也可以在命令窗口输入 `view`。



作为初学者，我认为 Stata 中最重要的命令是 `help`。遇到任何关于命令的疑问时，只需输入 `help` 加上命令语句便可。例如，如果我们不知道命令 `merge` 怎么使用，只需要输入 `help merge`，在浏览器中便会出现关于 `merge` 的所有详细信息。

4. 数据使用、存储以及编辑

工作目录

在每次使用 Stata 之前，应该确认设置好工作目录（Working Directory）。工作目录可以通过点击**文件->更改工作目录**来修改，也可以通过在命令窗口输入 `cd` 修改。例如

```
cd "c:\angranli\Stata16"
```

记住，命令 `cd` 之后是文件所在的文件夹位置，Windows 版本和 Mac 版本的输入格式是有差别，很容易犯错。

工作目录显示于命令窗口的下方：



也可以在命令窗口中输入 `pwd` 来检查当下的工作目录是哪个文件夹。

打开和导入数据

Stata 的数据是`.dta` 格式。直接双击数据文件就可以打开数据。在打开数据的同时，系统会默认数据所在文件夹为工作目录。也可以在命令窗口输入：

```
use "data.dta", clear
```

`clear` 的意思是告诉系统先清除之前打开的数据，然后再打开新的数据。

大多数情况下，同学们会处理不同格式的数据。这个时候，我们需要在 Stata 中导入不同格式的数据，如`.csv` `.sav` `.xlsx` 等等。点击**文件->导入**，便可进行选择导入不同格式数据。或者，可以在命令窗口输入 `import` 来导入数据。例如，`import spss "data.sav"`导入 SPPS 数据。

存储和导出数据

存储数据很直接，直接点击保存。也可以输入指令：

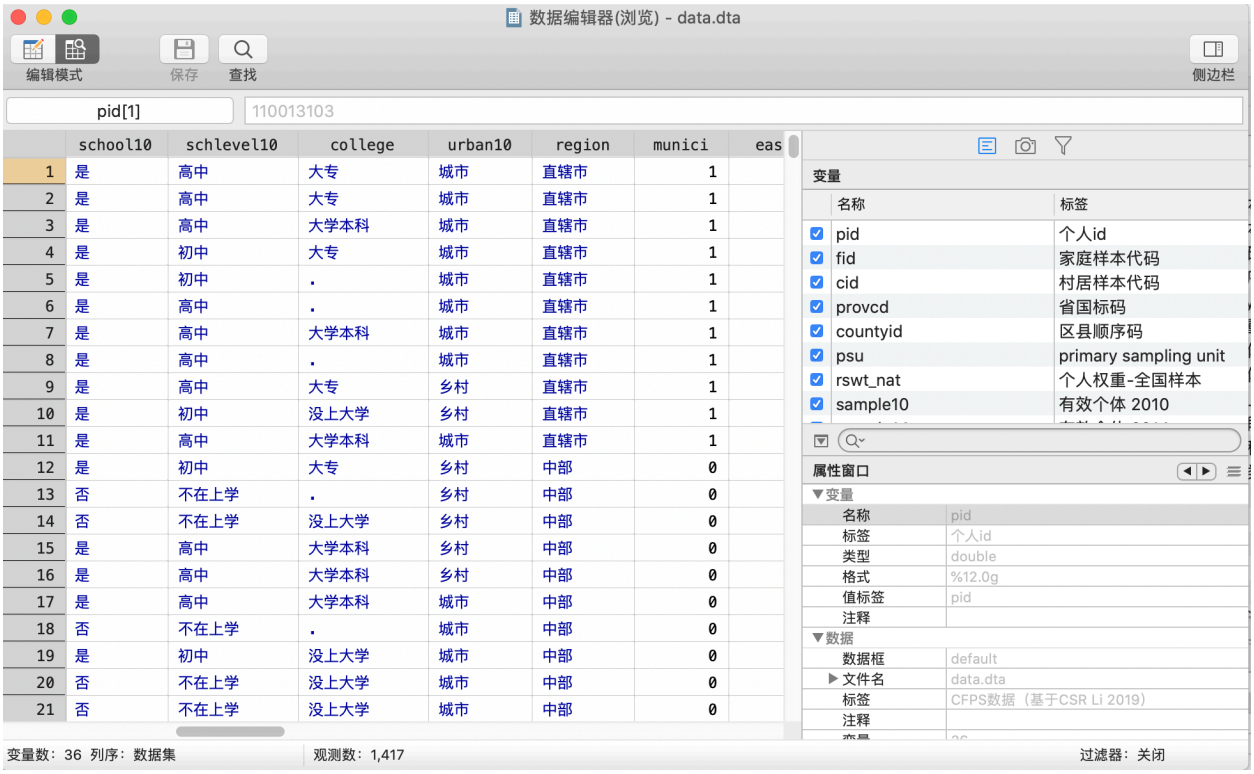
```
save "data.dta", replace
```

replace 的意思是告诉系统存储时替代之前的数据，如果没有 replace 并且文件夹中已经存在相同名称的数据，会出现错误提示。

导出数据很简单，点击**文件->导出**，选择不同的导出格式。或者使用 export 指令，方法和 import 一样。

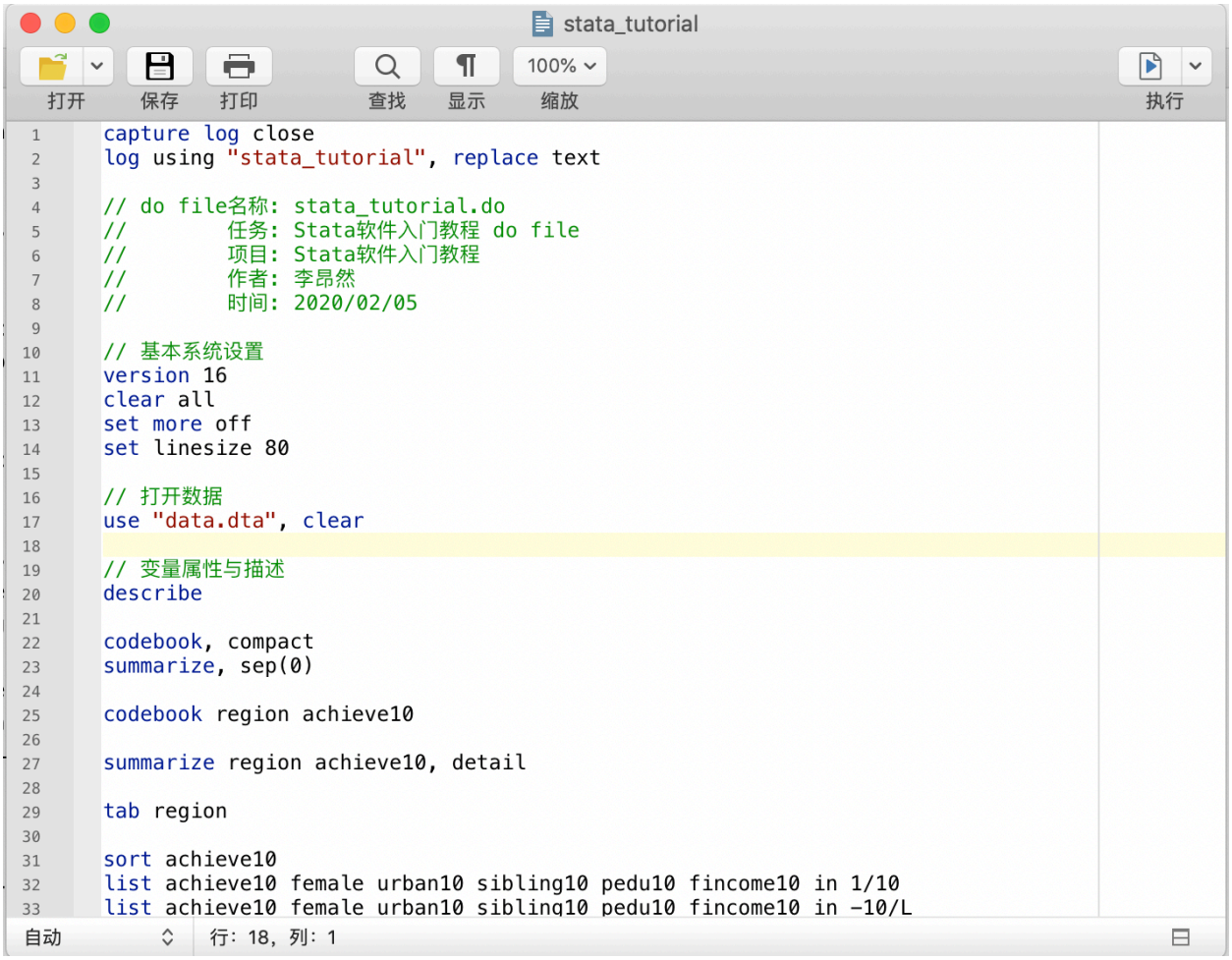
数据编辑

有时候，同学们需要手动输入数据。打开数据编辑器便可进行修改。记住，数据编辑器有两种模式：浏览和编辑。浏览模式下，只能看原始数据，不能改。修改需要打开编辑模式。浏览模式可以用 browse 来打开，编辑模式用 edit 打开。



5. Do File 和 Log File

Do file 是保存 Stata 命令语句的文件。大家可以把通过命令窗口输入的语句全部保存在 do file 中。这样做的目的是为了更好的重复工作，避免每次都去手动输入以及规避可能出现输入错误，在研究中做到统计分析结果科学的复制（replication）。打开 do file 可以点击 do file 编辑器也可以直接输入 `doedit`。



Log file 是用来)存储命令反馈结果的，简单讲，就是出现在结果窗口的内容都可以保存在 log file 中。

我在这里给大家分享一个简单的 do file 模板，可以自动创建 log file 并保存分析结果。可以根据个人需要修改。

```
1) capture log close
2) log using "stata_template", replace text

3) // do file 名称: stata_template.do
4) //         任务: 创建 do file 模板
5) //         项目: Stata 软件入门教程
6) //         作者: 李昂然
7) //         时间: 2020/02/05

8) // 基本系统设置
9) version 16
10) clear all
11) set more off
12) set linesize 80

13) // 打开数据
14) use "data.dta", clear

15) // 保存数据
16) save "data.dta", replace

17) log close
18) exit
```

第 1 行指令是告诉 Stata 关闭目前打开的 log file，然后开始记录新的结果。

第 2 行指令是创建一个新的 log file 并且其文件名和 do file 相同。同时，Stata 会自动保存新的结果去替代旧的文件。

第 3 行到第 7 行给出文件的一些基本信息，便于快速的了解这个 do file 的工作任务是什么。注意：在//之后的内容，Stata 是不会读取的，可以当做文本保存在 log file 里，相当于用来做笔记。也可以用*，效果是一样的。如果要记录很长的笔记，并且要用好多行，可以用/*作为开始，*/作为结束。例如：

```
/* 关于 Stata 软件使用
更多的内容，请查看
https://www.stata.com */
```

另外，对于过长的命令语句，需要换行时，应该使用///。

第 8 行到第 12 行是一些基本的系统设置，用来更好的跑文件。`version 16` 是指定版本 16。`clear all` 是清除目前系统保存的数据及其相关信息。`set more off` 是告诉 Stata 在跑结果的时候，不要分段报告结果，直接全部跑完。`set linesize 80` 是设置 log file 报告结果的文本宽度格式，让结果能看的清楚。如果有的结果在 log file 里面出现了排版拥挤的现象，可以调高宽度。

17 和 18 两行是结束当前的 log 记录，并退出。

6. 数据管理与清理

变量属性与描述

当打开 Stata，你首先需要知道数据中有哪些变量。具体讲，变量的名称、标签以及值标签要熟悉。其次，可以进一步去了解这些变量的统计值和分布。

这里要介绍几个常用的命令，包括 `describe`，`codebook`，`summarize`，`tabulate`，`list`。

`describe` 会给出变量具体的信息，包括变量名称，标签，储存类型，格式，以及值标签的名称。同时，也提供数据的基本信息。其输出结果如下：

```
. describe
```

```
Contains data from data.dta
  obs:         1,417                          CFPS 数据（基于 CSR Li 2019）
 vars:            36                          5 Feb 2020 15:16
-------------------------------------------------------------------------------
-
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
-
pid             double  %12.0g     pid        个人 id
fid             double  %12.0g     fid        家庭样本代码
cid             double  %12.0g     cid        村居样本代码
provcd          double  %24.0g     provcd     省国标码
countyid        double  %12.0g     countyid   区县顺序码
psu             double  %12.0g     psu        primary sampling unit
rswt_nat        double  %12.0g     rswt_nat   个人权重-全国样本
sample10        float   %9.0g                 有效个体 2010
sample14        float   %9.0g                 有效个体 2014
school10        float   %9.0g      school10   在学校上学 2010
schlevel10      float   %12.0g     schlevel10
                                               上学阶段 2010
college         float   %12.0g     college    大学本科，大专，没上大学 2014
urban10         float   %9.0g      urban10    城乡分类变量 2010
region          float   %9.0g      region     地域
munici          byte    %8.0g                 直辖市
east            byte    %8.0g                 东部
central         byte    %8.0g                 中部
west            byte    %8.0g                 西部
```

```
word10          float   %9.0g                 字词测试得分 2010
math10          float   %9.0g                 数学测试得分 2010
achieve10       float   %9.0g                 个人学习能力 2010
female          float   %9.0g      female     女
age10           float   %9.0g                 年龄 2010
ethnicity       float   %9.0g      ethnicity
                                              少数民族
hukou10         float   %9.0g      hukou10    农业户口 2010
sibling10       float   %9.0g                 兄弟姐妹数 2010
pedu10          float   %9.0g                 父母教育程度（年）2010
fincome10       float   %9.0g                 家庭总收入 2010
lnfincome10     float   %9.0g                 家庭总收入 (ln) 2010
eduexpense10    float   %9.0g                 教育总支出 2010
lneduexpense10  float   %9.0g                 教育总支出（ln）2010
yexpect10       float   %9.0g                 个人教育期望（年）2010
expect10        float   %15.0g     expect10   个人教育期望（分类）2010
tutoring10      float   %9.0g      tutoring10
                                              参加课外辅导 2010
schtype10       float   %9.0g      schtype10
                                              重点/私立学校 2010
class10         float   %9.0g      class10    就读重点班 2010
------------------------------------------------------------------------------
-
Sorted by:
```

如果想要输出变量的一些描述性统计，例如平均值、标准差、最小值、最大值，可以通过 `codebook, compact` 和 `summarize` 命令实现。`sep(0)`规定 `summarize` 的显示结果不需要用区块隔开。结果如下：

```
. codebook, compact
```

```
Variable       Obs Unique       Mean        Min       Max  Label
------------------------------------------------------------------------------
pid           1417   1417   4.00e+08   1.10e+08  6.22e+08  个人 id
fid           1417   1328   400074.1     110013    621872  家庭样本代码
cid           1417    504   16675.02      10010     23200  村居样本代码
provcd        1417     25   39.94001         11        62  省国标码
countyid      1417    159    83.247          1       162  区县顺序码
psu           1417    171   89.90049          1       176  primary sampling
unit
rswt_nat      1417   1159   39602.72    691.162  110910.4  个人权重-全国样本...
sample10      1417      1          1          1         1  有效个体 2010
sample14      1417      2   .7198306          0         1  有效个体 2014
school10      1417      2   .7148906          0         1  在学校上学 2010
schlevel10    1415      6    2.90742          1         6  上学阶段 2010
```

```
college       1029      3   1.563654          1          3  大学本科，大专，没...
urban10       1417      2   .3973183          0          1  城乡分类变量 2010...
region        1417      4   2.922371          1          4  地域
munici        1417      2   .0649259          0          1  直辖市
east          1417      2   .2724065          0          1  东部
central       1417      2   .3380381          0          1  中部
west          1417      2   .3246295          0          1  西部
word10        1417     35   25.92096          0         34  字词测试得分 2010...
math10        1417     25     15.494          0         24  数学测试得分 2010...
achieve10     1417    350  -1.57e-09  -3.195165   1.325824  个人学习能力 2010...
female        1417      2   .4932957          0          1  女
age10         1417      3   16.95131         16         18  年龄 2010
ethnicity     1417      2    .089626          0          1  少数民族
hukou10       1417      2   .7854622          0          1  农业户口 2010
sibling10     1417      9   1.148192  -.0124451          8  兄弟姐妹数 2010
pedu10        1417     22   6.387157          0       19.5  父母教育程度（年）...
fincome10     1417   1003   32299.67  -49131.13     517400  家庭总收入 2010
lnfincome10   1417   1003    10.0029   5.298317   13.15657  家庭总收入 (ln) 20...
eduexpense10  1417    868   2870.077  -7850.915      31400  教育总支出 2010
lneduexpe~10  1417    868   7.034553          0   12.68215  教育总支出（ln）20...
yexpect10     1417     90   14.02635          0         21  个人教育期望（年）...
expect10      1417      5   2.994354          1          5  个人教育期望（分类...
tutoring10    1417      2   .0917431          0          1  参加课外辅导 2010...
schtype10     1417      2   .1813691          0          1  重点/私立学校 2010...
class10       1417      2   .2180663          0          1  就读重点班 2010
-------------------------------------------------------------------------------

. summarize, sep(0)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         pid |      1,417    4.00e+08    1.54e+08   1.10e+08   6.22e+08
         fid |      1,417    400074.1    153810.9     110013     621872
         cid |      1,417    16675.02    5132.573      10010      23200
      provcd |      1,417    39.94001    15.37745         11         62
    countyid |      1,417      83.247    47.93385          1        162
         psu |      1,417    89.90049    54.20222          1        176
    rswt_nat |      1,417    39602.72     29708.2    691.162   110910.4
    sample10 |      1,417           1           0          1          1
    sample14 |      1,417    .7198306    .4492404          0          1
    school10 |      1,417    .7148906    .4516259          0          1
  schlevel10 |      1,415     2.90742    1.303046          1          6
     college |      1,029    1.563654    .8067569          1          3
     urban10 |      1,417    .3973183    .4895157          0          1
      region |      1,417    2.922371    .9226669          1          4
      munici |      1,417    .0649259    .2464821          0          1
        east |      1,417    .2724065    .4453551          0          1
     central |      1,417    .3380381    .4732086          0          1
```

```
        west |      1,417    .3246295    .4684016          0          1
      word10 |      1,417    25.92096    7.181325          0         34
      math10 |      1,417      15.494    5.571708          0         24
   achieve10 |      1,417   -1.57e-09    .9140147  -3.195165   1.325824
      female |      1,417    .4932957    .5001316          0          1
       age10 |      1,417    16.95131    .8172056         16         18
   ethnicity |      1,417     .089626     .285746          0          1
     hukou10 |      1,417    .7854622    .4106462          0          1
   sibling10 |      1,417    1.148192    1.032285  -.0124451          8
      pedu10 |      1,417    6.387157    3.771436          0       19.5
   fincome10 |      1,417    32299.67    36394.94  -49131.13     517400
 lnfincome10 |      1,417     10.0029    .9018059   5.298317   13.15657
eduexpense10 |      1,417    2870.077    3515.814  -7850.915      31400
lneduexpe~10 |      1,417    7.034553    1.951467          0   12.68215
   yexpect10 |      1,417    14.02635     3.12407          0         21
    expect10 |      1,417    2.994354    1.236782          1          5
  tutoring10 |      1,417    .0917431    .2887649          0          1
   schtype10 |      1,417    .1813691    .3854597          0          1
     class10 |      1,417    .2180663    .4130785          0          1
```

单纯使用 `codebook` 不添加 `compact` 选项，我们会得到详细的关于变量的描述信息。如下：

```
. codebook region achieve10
```

```
-------------------------------------------------------------------------------
region                                                                    地域
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  region

                 range:  [1,4]                        units:  1
         unique values:  4                        missing .:  0/1,417

            tabulation:  Freq.   Numeric  Label
                           92         1  直辖市
                          386         2  东部
                          479         3  中部
                          460         4  西部

-------------------------------------------------------------------------------
achieve10                                                  个人学习能力 2010
-------------------------------------------------------------------------------

                  type:  numeric (float)
```

```
                range:  [-3.1951654,1.325824]        units:  1.000e-11
        unique values:  350                      missing .:  0/1,417

                 mean:  -1.6e-09
            std. dev:   .914015

          percentiles:        10%       25%       50%       75%       90%
                           -1.18377   -.37767   .160764   .578514   1.04732
```

输入 `summarize, detail` 会得到更加详细的描述性信息。如下：

```
. summarize region achieve10, detail
```

```
                            地域
-------------------------------------------------------------
      Percentiles      Smallest
 1%           1              1
 5%           1              1
10%           2              1       Obs               1,417
25%           2              1       Sum of Wgt.       1,417

50%           3                      Mean           2.922371
                        Largest       Std. Dev.      .9226669
75%           4              4
90%           4              4       Variance       .8513141
95%           4              4       Skewness      -.3423151
99%           4              4       Kurtosis       2.102561

                    个人学习能力 2010
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -3.195165      -3.195165
 5%    -1.910969      -3.195165
10%    -1.183774      -3.195165       Obs               1,417
25%    -.3776704      -3.195165       Sum of Wgt.       1,417

50%     .1607642                      Mean          -1.57e-09
                        Largest       Std. Dev.      .9140147
75%     .5785144       1.325824
90%     1.047324       1.325824       Variance       .8354228
95%      1.16646       1.325824       Skewness      -1.298654
99%     1.325824       1.325824       Kurtosis       5.091919
```

对于分类变量，我们可以用 `tabulate` 来检查频数的分布情况。方便起见，可以使用缩写 `tab`。其结果如下

```
. tab region
```

```
       地域 |      Freq.     Percent        Cum.
------------+-----------------------------------
     直辖市 |         92        6.49        6.49
       东部 |        386       27.24       33.73
       中部 |        479       33.80       67.54
       西部 |        460       32.46      100.00
------------+-----------------------------------
      Total |      1,417      100.00
```

list 可以用来查看某些具体的观测单位，从而更好的了解数据。比如，我们想要知道学习能力最好的和最差的学生的家庭背景大体是什么样的，这个时候我们可以用 list 来查看。首先，我们用 sort 命令把数据按照学习能力测试分数进行排序，Stata 默认的是由低到高低升序排列。

以下语句给我列出了数据中学习能力测试分数最低的 10 个观测个体的家庭背景变量。

```
. sort achieve10

. list achieve10 female urban10 sibling10 pedu10 fincome10 in 1/10
```

```
     +-------------------------------------------------------------+
     | achieve10   female   urban10   sibli~10   pedu10   finco~10 |
     |-------------------------------------------------------------|
  1. | -3.195165       否       乡村          3        0       7763 |
  2. | -3.195165       否       乡村          1        9     113410 |
  3. | -3.195165       是       乡村          3        0   28953.55 |
  4. | -3.195165       是       乡村          1        0     2017.5 |
  5. | -3.195165       否       乡村          6        3    20508.6 |
     |-------------------------------------------------------------|
  6. | -3.195165       是       乡村          1        0       4200 |
  7. | -3.195165       是       乡村          3        0      11305 |
  8. | -3.195165       是       乡村          6        0    43114.3 |
  9. | -3.195165       否       乡村          3        3   37887.32 |
 10. | -3.195165       否       乡村          0        6       4000 |
     +-------------------------------------------------------------+
```

以下语句给我列出了数据中学习能力测试分数最高的 10 个观测个体的家庭背景变量。

```
. sort achieve10

. list achieve10 female urban10 sibling10 pedu10 fincome10 in -10/L
```

```
      +-------------------------------------------------------------+
      | achie~10   female   urban10   sibli~10   pedu10   fincome10 |
      |-------------------------------------------------------------|
1408. | 1.325824       否       城市          1        9       58500 |
1409. | 1.325824       是       乡村          0      7.5       61094 |
1410. | 1.325824       是       乡村          2       12   -13334.41 |
1411. | 1.325824       否       乡村          2      4.5    104772.7 |
1412. | 1.325824       是       乡村          1        9      121000 |
      |-------------------------------------------------------------|
1413. | 1.325824       是       城市          0        3    16061.49 |
1414. | 1.325824       否       乡村          1      7.5       30000 |
1415. | 1.325824       否       乡村          1     10.5        9874 |
1416. | 1.325824       否       城市          0      7.5       33240 |
1417. | 1.325824       是       乡村          3        6       25187 |
      +-------------------------------------------------------------+
```

变量创建与修改

generate 命令是用来创建新的变量，可以使用简写 gen。以下的例子展示了如何创建变量。

（1）可以通过简单的运算来创建变量。

```
. gen totalscore = word10+math10

. gen meanscore = (word10+math10)/2

. sum totalscore meanscore

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  totalscore |      1,417    41.41496    11.67473          0         58
   meanscore |      1,417    20.70748    5.837367          0         29
```

（2）根据已有连续变量，创建新的分类变量。recode 命令是用来重新赋值。

```
. gen groupscore = totalscore

. recode groupscore (0/20=1) (21/40=2) (41/58=3)
(groupscore: 1416 changes made)

. tab groupscore
```

```
 groupscore |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         91        6.42        6.42
          2 |        436       30.77       37.19
          3 |        890       62.81      100.00
------------+-----------------------------------
      Total |      1,417      100.00
```

（3）根据已有分类变量，可以合并相关类别。

```
. gen dcity = region

. recode dcity (1=1) (2/4=0)
(dcity: 1325 changes made)

. tab dcity

      dcity |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,325       93.51       93.51
          1 |         92        6.49      100.00
------------+-----------------------------------
      Total |      1,417      100.00
```

也可以使用 `replace` 对新的分类变量赋值。

```
. replace dcity = 1  if region==1
(0 real changes made)

. replace dcity = 0 if region==2 | region==3 | region==4
(0 real changes made)



. tab dcity

      dcity |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,325       93.51       93.51
          1 |         92        6.49      100.00
------------+-----------------------------------
      Total |      1,417      100.00
```

egen 命令可以通过明确不同的函数运算来创建变量。比如，我们可以对连续变量进行标准化。

```
. egen z_totalscore = std(totalscore)

. sum z_totalscore

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
z_totalscore |      1,417    3.93e-09           1    -3.5474   1.420592
```

变量名称与标签

label var 用来给变量建立标签。

```
. label var totalscore "字词与算数分数总和"

. label var meanscore "平均分数"

. label var groupscore "分数分组"

. label var dcity "直辖市"

. label var z_totalscore "标准化分数"

. codebook totalscore meanscore groupscore dcity z_totalscore, compact

Variable       Obs Unique      Mean      Min       Max  Label
------------------------------------------------------------------------------
totalscore    1417     58  41.41496        0        58  字词与算数分数总和...
meanscore     1417     58  20.70748        0        29  平均分数
groupscore    1417      3  2.563867        1         3  分数分组
dcity         1417      2  .0649259        0         1  直辖市
z_totalscore  1417     58  3.93e-09  -3.5474  1.420592  标准化分数
------------------------------------------------------------------------------
```

label define 是用来给分类变量建立值标签。label value 为数值赋予值标签。

```
. label define group 1 "低分数组" 2 "中等分数组" 3 "高分数组"

. label value groupscore group

. tab groupscore

     分数分 |
```

```
          组 |      Freq.     Percent        Cum.
------------+-----------------------------------
    低分数组 |         91        6.42        6.42
  中等分数组 |        436       30.77       37.19
    高分数组 |        890       62.81      100.00
------------+-----------------------------------
       Total |      1,417      100.00
```

变量删除与保留

变量的删除和保留可以使用 `drop` 和 `keep`。

```
. drop totalscore meanscore groupscore dcity z_totalscore
```

如果想要根据某条件，删除或保留部分观测值，需要 `if` 来明确条件。

```
. keep if college~=.
(388 observations deleted)
```

`~=`为不等于符号。这里为删除变量 `college` 缺失的数据。

数据合并与转换

这里只是简单的介绍数据合并转换的命令，不作具体例子解释，想要深入了解如何使用这些命令的同学可查看官方手册。

`merge` 为横向合并，即通过有相同观测 ID 的情况下合并数据。

`append` 为纵向合并，即在有相同变量的情况下合并数据。

`reshape` 为数据形式的转换指令，可以在宽数据（wide format）和长数据（long format）之间互相转换。

7. 数据可视化

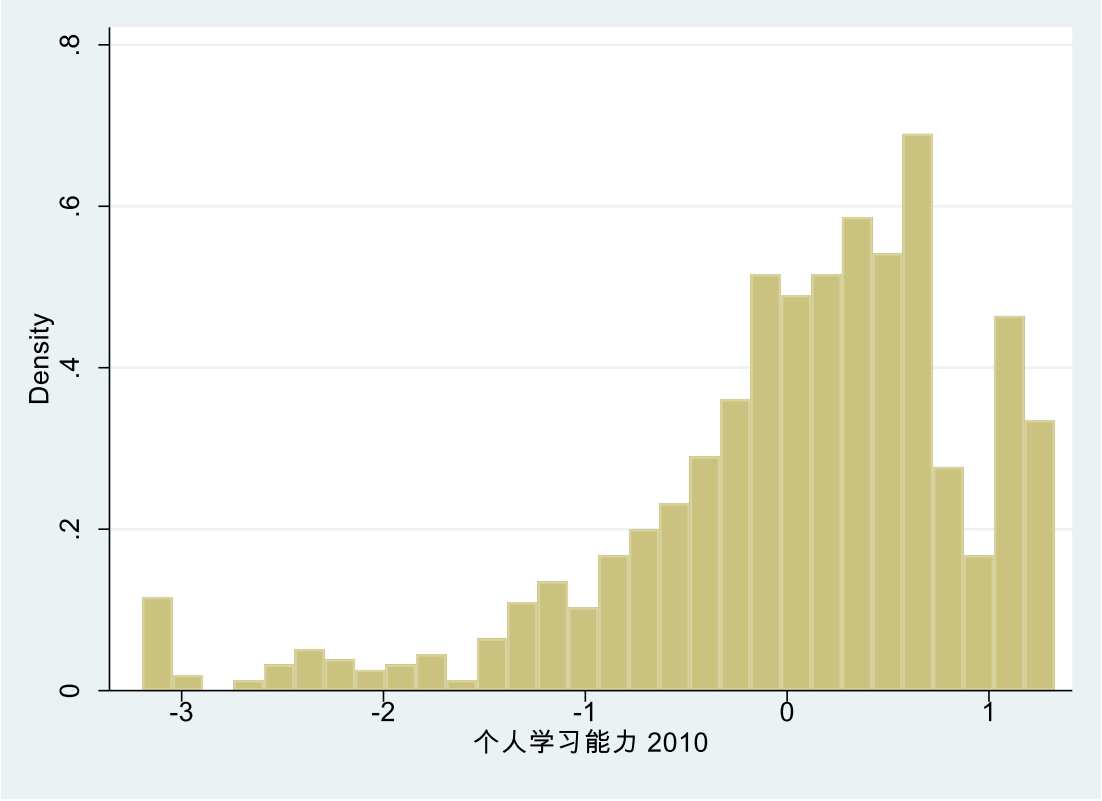
这里介绍简单常用的 Stata 的绘图命令。

单变量可视化

（1）连续变量

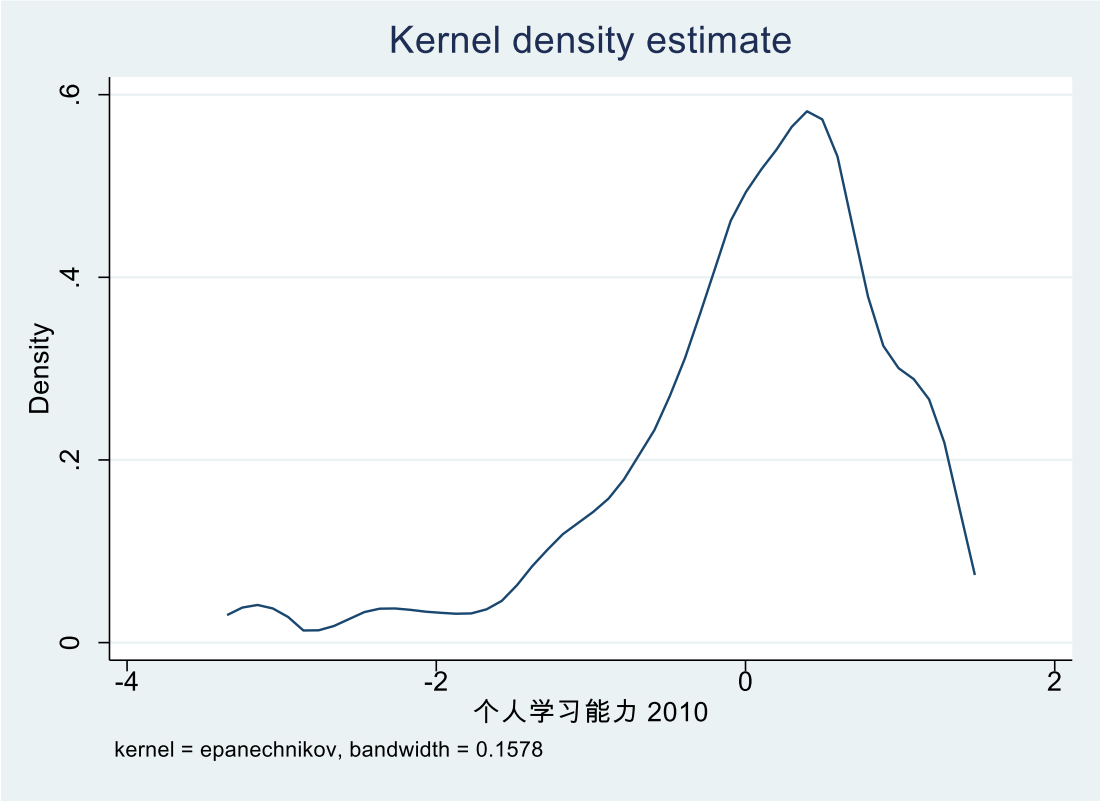
`histogram` 用来生成分布直方图。

```
. histogram achieve10
(bin=30, start=-3.1951654, width=.15069965)
```



kdensity 用来生成核密度分布估计图。

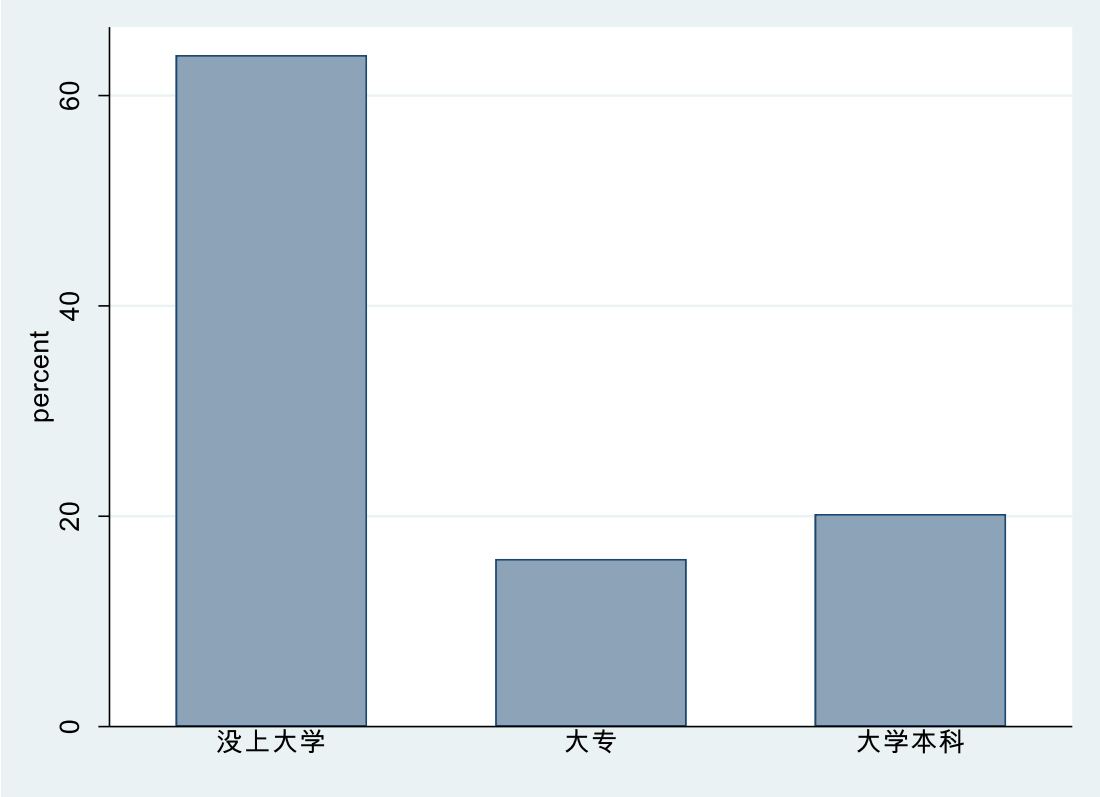
```
. kdensity achieve10
```



（2）离散变量

`graph bar` 可以生成条形图。

```
. graph bar (percent), over(college) intensity(*0.5)
```

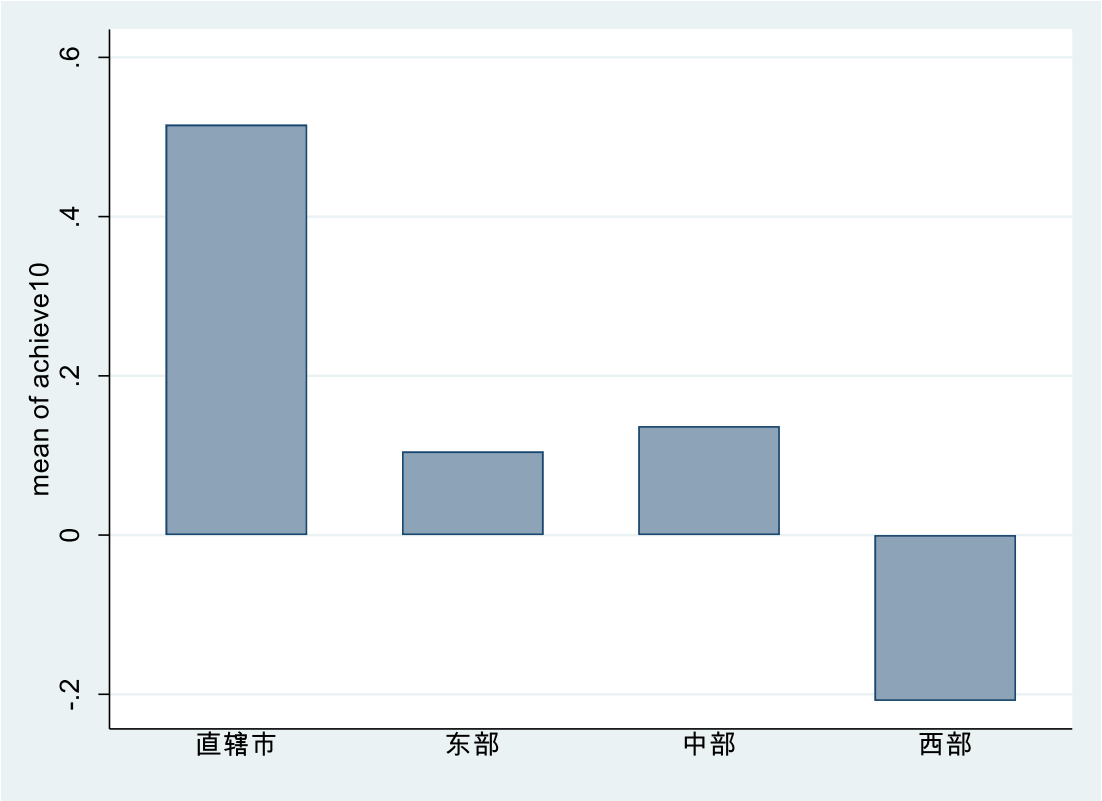


变量间关系可视化

（1）离散和连续变量

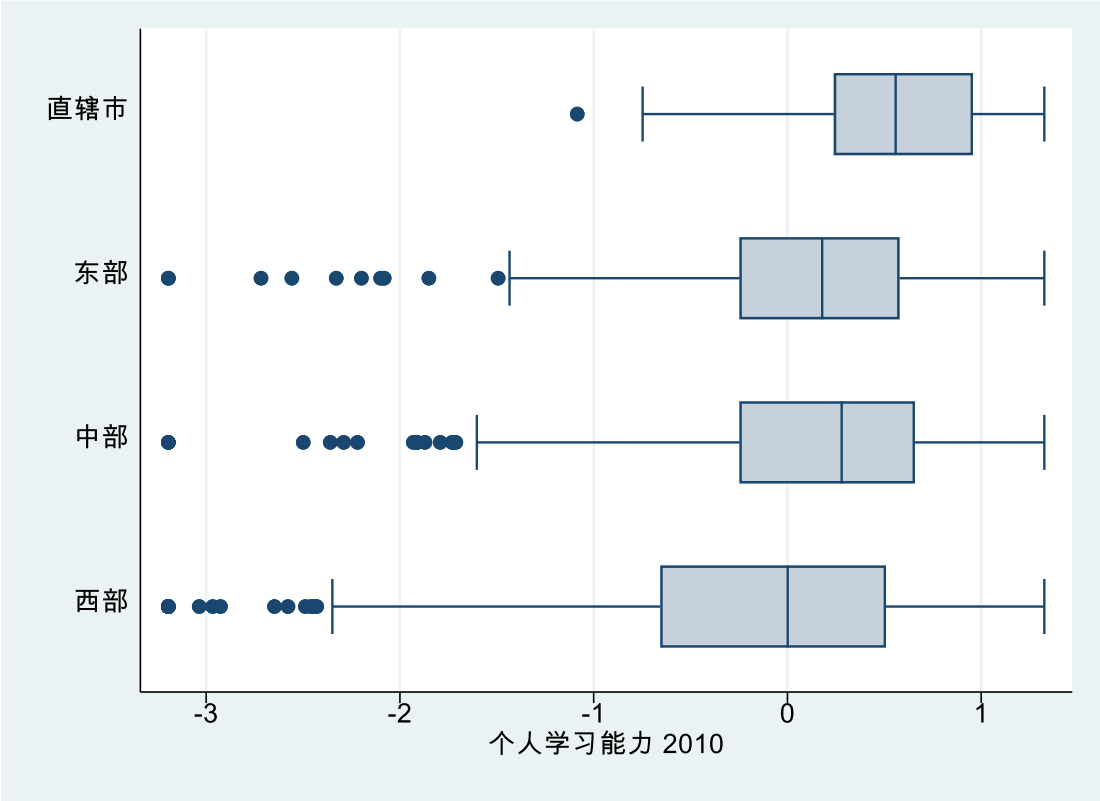
如果我们想要查看学习能力测试分数的地域差异，我们可以用以下的指令。

```
. graph bar (mean) achieve10, over(region) intensity(*0.5)
```



`graph hbox` 可以生成箱线图。

```
. graph hbox achieve10, over(region) intensity(*0.5)
```



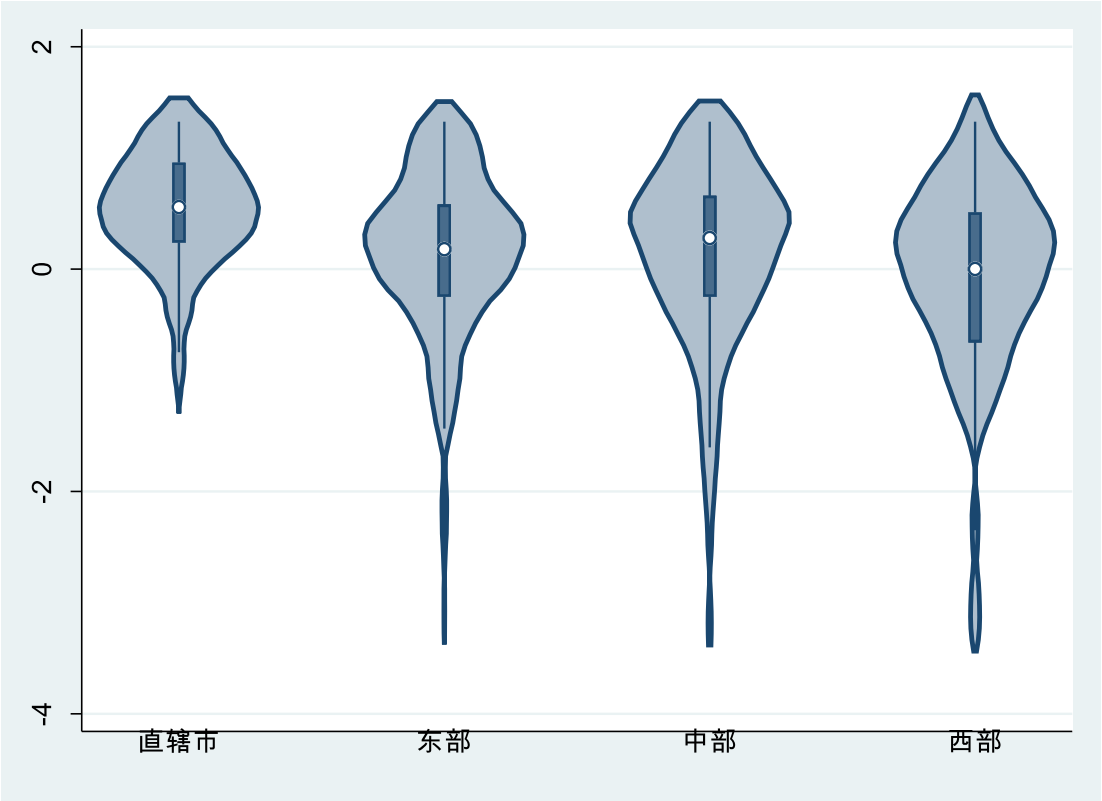
另外，我们可以安装 community-contributed command `vioplot` 并生成小提琴分布图，因为图形的形状酷似小提琴而得名。

首先我们安装以下这个指令：

```
. ssc install vioplot, replace
checking vioplot consistency and verifying not already installed...
all files already exist and are up to date.
```

小提琴分布图主要是箱线图和核密度分布图的结合，可以更直观的看到变量在不同组的分布情况。

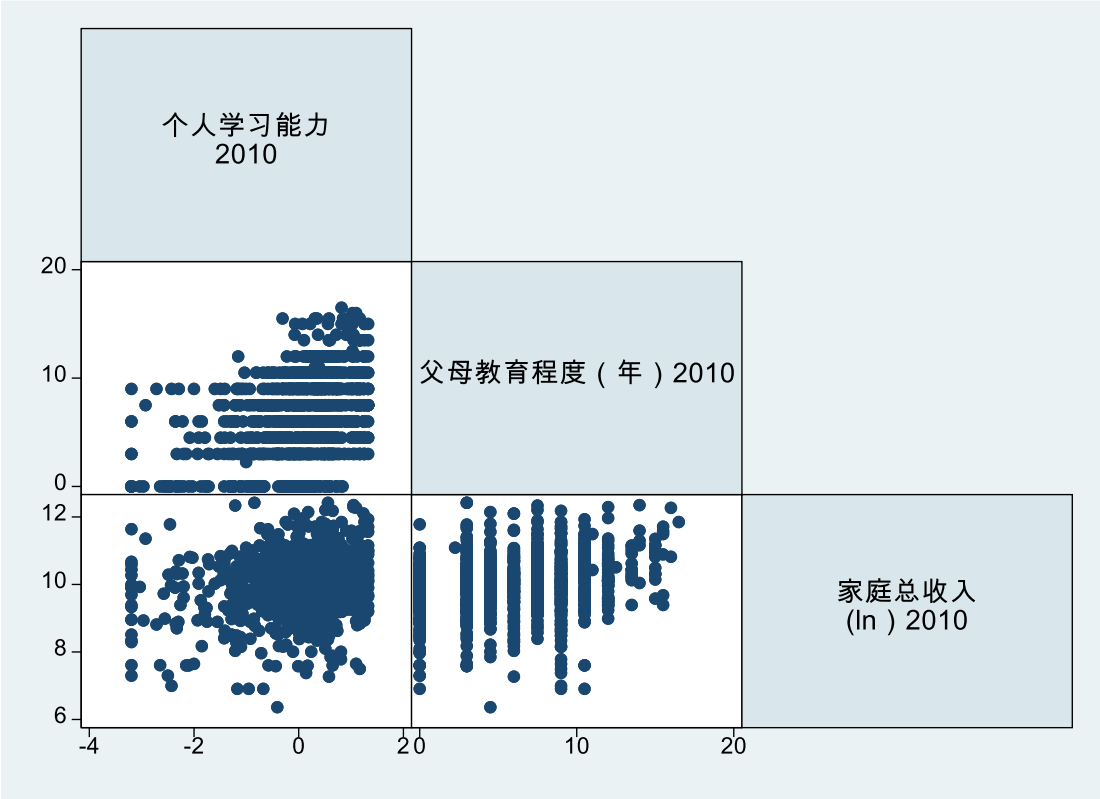
```
. vioplot achieve10, over(region)
```



（2）多个连续变量间关系

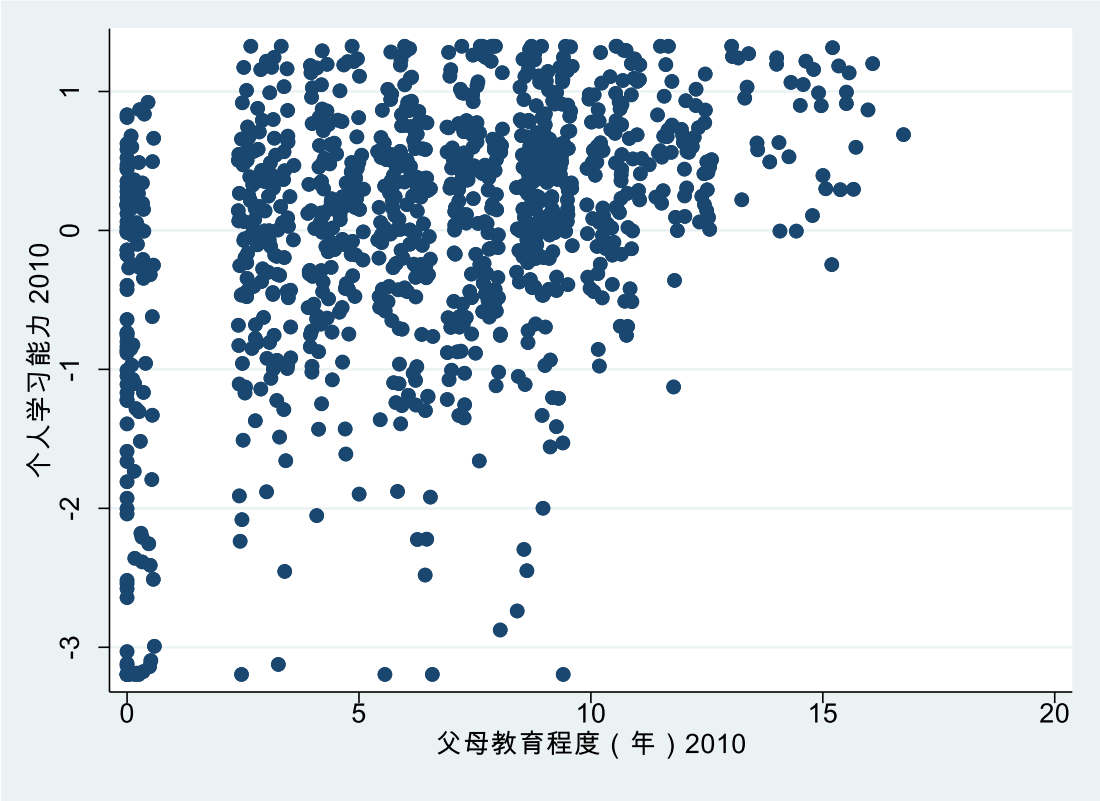
连续变量间的关系可以用散点图来描绘，有多个连续变量时，散点图矩阵 `graph matrix` 可以给我们展示变量间的关系。

```
. graph matrix achieve10 pedu10 lnfincome10, half
```



当然，我们可以直接看二维散点图 `twoway scatter`。

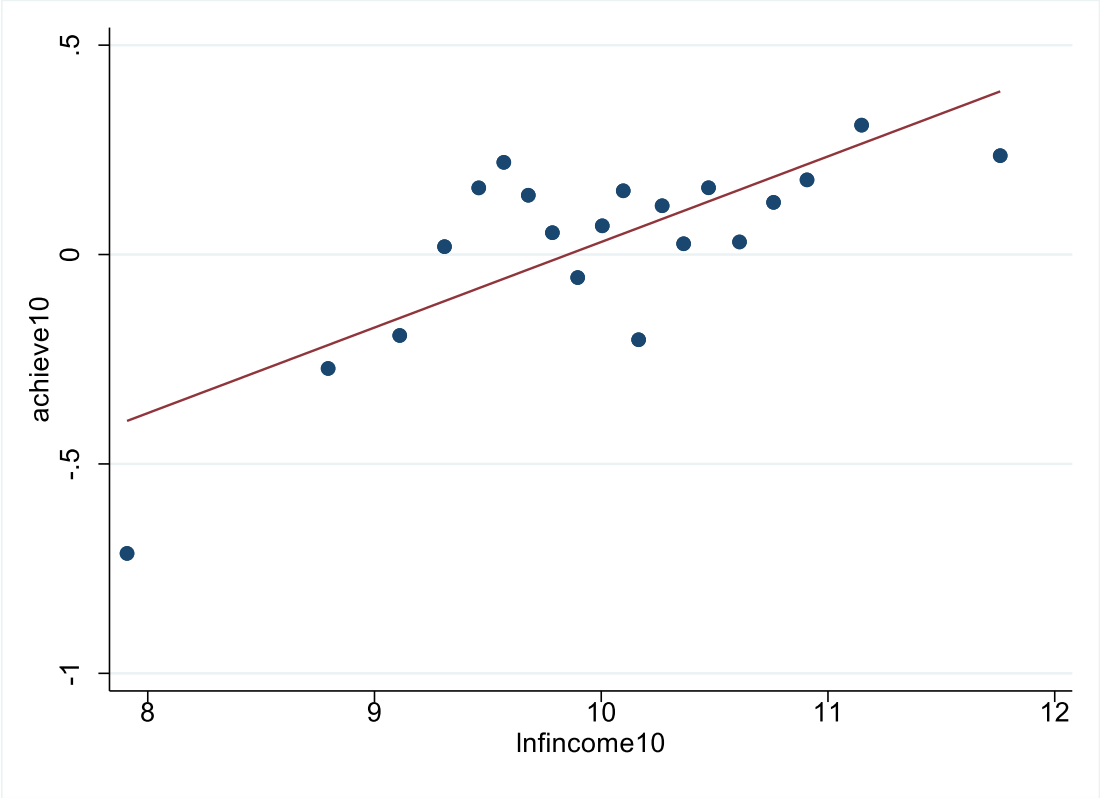
```
. twoway scatter achieve10 pedu10, jitter(7)
```



大家很容易就看到，当观测点很多时，散点图就显得过于拥挤了，很难看出来有什么直接的联系。我们可以用安装 community-contributed command `binscatter` 并生简化版的散点图，以及描绘变量间的拟合趋势。

```
. ssc install binscatter, replace
checking binscatter consistency and verifying not already installed...
all files already exist and are up to date.

. binscatter achieve10 lnfincome10
```



8. 学习资料

- 书籍

Acock, Alan. C. 2008. *A Gentle Introduction to Stata*. Stata Press.

Daniels, Lisa and Nicholas MinotAn. 2020. *Introduction to Statistics and Data Analysis Using Stata*. Sage Publications.

Kohler, Ulrich and Frauke Kreuter. 2012. *Data Analysis Using Stata, Third Edition*. Stata Press.

Long, J. Scott. 2009. *The Workflow of Data Analysis Using Stata*. Stata Press.

Mitchell, Michael N. 2012. *Interpreting and Visualizing Regression Models Using Stata*. Stata Press.

- 网站

（1）Stata 官方网站：https://www.stata.com/features/documentation/

（1）加州大学洛杉矶分校 UCLA IDRE：https://stats.idre.ucla.edu/stata/

（2）普利斯顿大学 Princeton Stata： http://data.princeton.edu/stata/