第5章 数据整理

公管教研室 马冉

- 主要内容：
  - 基本操作
  - 数据集相关操作
  - 变量相关操作
  - 表达式与函数
  - 数据集合并

一、Stata基本操作

- 1.使用界面
  - 结果窗口
  - 命令窗口
  - 命令回顾窗口（单击复制，双击重复执行）
  - 变量名窗口
  - 数据编辑/浏览窗口

- 2.基本命令
  - help(精确查找)
  - search（模糊查找）
  - ssc install (下载非官方命令)
  - clear(清空内存)
  - set mem 100m(适用于低版本，设定内存空间)
  - set more off（自动执行命令）
  - cd （查看/更改当前工作目录）
  - unicode analyze 文件名—unicode encoding set gb18030—unicode translate 文件名 （用于对乱码转码，必须先清理内存并将文件放置到默认浏览器中）

二、变量的取值类型

- 字符型变量储存格式：str# #代表最多可容纳字符数
- 数值型变量：字节（byte）、整数（int）、长整数（long）、浮点（float）(默认储存格式)、双浮点（double）（计算结果的储存格式）
- 缺失值：“.”

三、创建新的数据集

- edit/browse
- 变量重命名：rename old_varname new_varname
- 储存数据：save filename (,replace)
- 对数据集的简单描述：describe (varlist) (, memory_option)
- 显示变量数值：list （varlist） （if） (in) (,option)
- 详尽显示变量内容：codebook （varlist） （if） (in) (,option)

例1：数据集auto.dta

①改变变量名make到var1

②使用describe和codebook查看数据概貌

③保存数据到硬盘的工作目录

四、导入已有数据集

- edit
- Excel格式：复制-粘贴
- Stata格式：use

五、表达式

- 关系符号：==等于　　!=不等于

  >大于　　<小于

  >=大于等于　　<=小于等于

  注：=让左边的值和右边的值相等

- 逻辑符号：!或　　&且　　|非

六、常用函数

- log(x) 取对数
- abs(x) 绝对值
- round(x) 四舍五入
- sqrt(x)开根
- autocode（x,n,xmax,xmin）将x的值域等分为n份，并求出各x值所在区间的上限
- group(x) 建立一个分类变量，将排序后的数据分为尽量等规模的x个子样本

七、定义数据子集

- in用于指定观测值

  command in # #/#

- if用于在命令后选择满足条件的数据子集

  command if exp

- by前置语句，用来对某些变量具有相同赋值的样本子集重复执行命令

  by varlist: stata_cmd

  bysort varlist: stata_cmd

  e.g. bysort female： stata_cmd

八、变量相关操作

- 1.建立新的变量——generate

generate (type) newvar=exp (if) (in)

注：新变量可能存在缺失值，这既可能受表达式的影响，也可能是设定了条件

例2：wage.dta

①生成educ的平方，并命名为educ2

②对educ大于9的观测值生成educ与exper的交叉项条件，并命名为logexper。

③生成exper的对数值，并命名为logexpereduc

④生成female与educ的交叉项，并命名为fem_educ

⑤生成educ的根号项并保留整数位

- 2.更改已有变量——replace

  replace oldvar =exp（if）(in)

  例3：wage.dta

  生成一个变量educat,将受教育年限变为定类变量,具体划分如下：小于3年为0；4-6为1；7-9年为2；10-12年为3；13-15年为4；16-18年为5

- egen—generate的扩展

egen (type) newvar =fcn(exp) (if) (in) (,option)

注：by选项仅对某些egen的函数适用

函数清单详见help

例4：wage.dta

①生成新变量educde,为educ与平均受教育年数之差—mean

②生成educed2，为educ与分性别的受教育年数中位数之差—median

③生成stdeduc,为标准化后的受教育年数—std

④生成higheduc，educ大于12则包含该变量的数值，否则为缺失值—anyvalue

⑤生成sexmar，为female和married的各种组合赋值类别编号—group

九、数值和字符串的转换

- 字符型—数值型

encode varname (if) (in) ,gen(newvar) (label(name))

- 数值型—字符型

decode varname (if) (in) ,gen(newvar) (maxlength(#))

从字符串表达中得到数值：real（）

例5 hbp2.dta

①用sex生成新的数值型变量gender，其标签为1“female”、0“male”

②将gender转换为字符型变量gender2，对比gender2与sex

十、生成分类变量和虚拟变量

- 1.生成虚拟变量

（1）使用generate+replace生成虚拟变量（参考例3）

例6：wage.dta

生成college，读过大学为1，否则为0

（2）使用gen newvar=(varname>#)生成虚拟变量

续例6：令教育年数在1和6之间的为小学，令教育年数在7和9之间的为初中，令教育年数在10和12之间的为高中，令教育年数在13和16之间的为大学本科，令教育年数在17和18之间的为研究生

- 2.生成分类变量

（1）使用generate+replace生成分类变量

（2）recode命令—对已有变量重新赋值

recode varlist (rule) (rule)…（，gen(newvar)）

赋值规则 #=# 3=1

##=# 2.=9

#/#=# 1/5=4 1到5等于4

nomissing=#

missing=#

例7：recxmpl.dta

①改变x的赋值，将1变为2，生成新变量nx

②改变x1的赋值，将1变为2、2改为1，生成nx1

③改变x2的赋值，将1和2变为1、3改为2、4到7为3，生成nx2

（3）使用autocode()、recode()、group()3个函数生成分类变量

autocode（x,n,xmin,xmax）将x的值域等分为n份，并求出各x值所在区间的上限，分组原则为取值区间宽度相等

group(x)则为案例数目的均匀分组，且必须先进行排序

recode（x,x1,x2,x3……）x为被分类变量，x1等为各组别取值上限

例8：wage.dta

①将exper分为5组，间隔为10年，生成exper1

②将exper按案例数量等分为5组，生成exper2

③将exper分为1-5,6-15,16-25,26-40，41-50等5组

十一、数据的整理

- 1.数据的横向合并

样本数量不变，变量数目增加

merge 1:1/1:m/m:1/m:m varlist using filename [, options]

注：两个数据都要有序号，且同一个样本序号一致

gen id=_n

options: merge: 记录变量1— master，2—using，3—matched

update: 内存内的缺失数值会被内存外的相应数值覆盖

1—不吻合，只内存内存在的个案　2—不吻合，只内存外存在的个案

3—一致　4—主数据缺失被更新　5—不一致，主数据被替换

replace：内存内的非缺失数值会被内存外的相应数值覆盖

nokeep：去掉内存外数据的非吻合个案

例9-1：autotech.dta数据：

make—生产厂商　　mpg—行驶里程

weight—车身重量　　length—车身长度

autocost数据：

make—生产厂商　　price—汽车价格

rep78—年度修理次数

要求：合并两个数据

例9-2：original为主数据，updates为调用数据，合并中出现不一致的地方依据调用数据进行修改

- 2.数据的纵向合并

变量数目不变，样本数量增加

前提：（1）变量名一致；（2）变量类型一致；（3）个案序号不重复；（4）为每个数据生成一个新变量来辨别合并后该数据的样本。

append using filename （，options）

options: keep(varlist) ：指定保留内存外多出来的变量

nalabel：不复制内存外数据的数值标签

nonotes：不复制内存外数据的数值说明

例10-1：运用append命令合并odd.dta和even.dta

例10-2：将auto.dta拆分为两个数据集，一个为国产汽车（foreign=0），另一个为国外产汽车（foreign=1）,然后将两者合并，并仅保存make、price、mpg、rep78、foreign等5个变量

- 3.数据的交叉合并

把一个数据的个案和另一个数据的个案交叉搭配成新的数据

(1) 组内交叉

两个数据样本有分组，每一个组里的样本做一对一交叉

joinby (varlist) using filename (,options)

options：样本吻合时：update replace

样本非吻合时：unmatched(none)不保留非吻合样本

unmatched(both)两个数据的非吻合样本都保存

unmatched(master)保存内存内非吻合样本

unmatched(using)保存内存外非吻合样本

merge(newvar)自动生成合并吻合度的记录变量

Nolabel不复制内存在数据的数值标签

例11：parent.dta父母数据，child.dta子女数据，得到亲子关系

（2）一一交叉

一个数据的每一个个案和另外一个数据的每一个个案交叉，合并后的样本数为两个样本数的积

最典型的应用：婚姻配对

cross using filename

前提：两个数据没有相同的文件名，否则会沿用内存内数据

例12：sex.dta 性别数据，agecat.dta年龄数据，用cross进行交叉合并。

- 4.数据抽取

sample # (if) (in) (,count by(groupvars))

随机从内存里的数据中抽取样本

指定count时#为抽取样本容量，不指定时为抽取百分比

by(groupvars)用来进行组内抽样，比如某样本随机抽取可能会出现某一教育水平的样本过多而其他教育水平过少的情况，这是可以用by（educ）表明在各个教育水平上做平均的抽取

例13：wage.dta

①简单抽取10个观测值，罗列wage、female、married这3个变量；

②在女性数据中抽取10个样本并保留所有男性样本；

③在男性和女性样本中各抽取10个样本；

④对female和married的各种组合各抽取5个样本。