

```

1
2
3      *=====
4      *              Stata基础训练营讲义
5      *=====
6      *              =====
7      *              Ch1  软件基础与数据处理
8      *              =====
9
10
11     *              主讲人：江俊澄
12     *              中南财经政法大学财政税务学院
13     *              学生学术科研中心
14     *              2024年11月17日
15
16
17
18

```

```

19 *-----
20 *->  1 软件须知
21 *-----
22

```

```

23 *1.1  stata软件简介
24

```

```

25 *-----
26 *stata界面
27

```

```

28 **标题栏、菜单栏、状态栏、工具栏
29

```

```

30 **另外界面和spss、evIEWS、sas以及office软件一致，都包括file、edit
31

```

```

32 **在最上方有一排菜单，即"File、Edit、data、Graphics、Statistics、User、Window、Help"
33

```

```

34 **左上"Review"（历史窗口），可以查询，
35

```

```

36 **正上方"Results"（结果窗口）
37

```

```

38 **正下方"Command"（命令窗 ）
39

```

```

40 **右上"Variables"（变量窗口）
41

```

```

42 **右下"Properties"（属性窗口）
43
44

```

```

45 *1.2 do文档
46

```

```

47 * 打开do文档或者do-file
48

```

```

49 *点击
50

```

```

51 *快捷键
52

```

```

53 *   ctrl+9  新建dofile
54

```

```

55 *   ctrl+0  打开文件夹
56

```

```

57 *-拓展知识点：ctrl+8，打开数据编辑界面
58

```

```

59 *-do文档编写习惯
60
61

```

```

62 *1.3-stata 功能一览
63

```

```

64 view browse "http://www.stata.com/features/
65

```

```

66 //帮助文件的完整目录help contentshelp contents utilities basic //基本命令
67

```

```

68 help contents_data //数据处理
69

```

```

70 help contents_stat //统计回归和分析
71

```

```

72 help graph //绘图
73

```

```

74 help program //编程
75

```

```

76 help contents matrix //矩阵
77

```

```

78 help estimationcommands //估计命令概览
79

```

```

80 help guide //一些最基本的命令和概念
81

```

```

82 help functions //函数
83

```

```

84 help language //stata的语法格式
85

```

```

86 help query //系统参数及设定
87

```

```

69 help winfonts //操作界面和字体设定
70 help contents interface //更全面的设定
71 ssc hot,n(50) //最热门的外部命令
72
73
74 *1.4-注释语句
75
76 help comments
77 sysuse auto, clear
78 *-示例:
79
80 sum price weight /* 第一种注释方式
81 gen x = 5 /* 第二种注释方式 */
82 // 第三种注释方式
83
84 /* begin the line with *;
85 begin the comment with //; or
86 place the comment between /* and */ delimiters.
87 */
88
89
90 *1.5-断行, 三种方式: "///" 、 "/* */" 、 #delimit 命令
91
92 *-第一种断行方式: ///
93 sysuse auto, clear
94 sum price weight length gear turn
95 tabstat price weight length gear turn , ///
96 stat(mean sd p5 p25 med p75 p95 min max) ///
97 format(%6.2f) c(s)
98
99 *-第二种断行方式: /* */
100 sysuse auto, clear
101 sum price weight length gear turn
102 tabstat price weight length gear turn , /*
103 */ stat(mean sd p5 p25 med p75 p95 min max) /*
104 */ format(%6.2f) c(s)
105
106 *-第三种断行方式: #delimit 命令 ///表示出现;才结束
107 #delimit ;
108 tabstat price weight length gear turn ,
109 stat(mean sd p5 p25 med p75 p95 min max)
110 format(%6.2f) c(s) ;
111 #delimit cr
112
113 *1.6 其他语法须知
114
115 *.1.6.1 注意
116 * Stata 区分大小写! Stata is case-sensitive!
117 * Stata 还区分中英文标点 “ ” " " , , ` '
118 * 错误的大小写,中英拼写会使你的屏幕“满江红”
119
120 * 1.6.2. 常见符号:
121 * =是赋值 (e.g. gen x= 1)
122 * ==是恒等 (e.g. gen x=1 if y== 2)
123 * |是或者 (e.g. gen x=1 if y==2|y== 3)
124 * &是且 (e.g. gen x=1 if y==2 & z==3)
125 * 不等于 ! 或者 ~
126 * 加减乘除幂 + - * \ ^
127 * 大小比较 < > == <=
128 * 通配符 wildcard *
129
130
131
132
133 *-----
134 *-> 2 数据清理准备工作
135 *-----
136

```

```

137 *2.1 设置工作路径
138 *(1) cd "" 把工作的文件路径复制进来
139 cd "E:\Desktop\Stata ch1"
140 pwd // 查看当前的工作目录
141 *(2) 设置root路径 调用global (global等会儿讲)
142 global root=" "
143 /*
144 global dofiles= "Sroot\dofiles"、
145 global logfiles="Sroot\logfiles"
146 global raw data= "Sroot\raw data"
147 global working data= "$root\working data"
148 global temp data= "$root\temp data"
149 global tables= "Sroot\tables"
150 global figures= "$root\figures"
151 */
152
153
154
155
156 *2.2 log file
157 cap log use// 出错时停止 stop when error
158
159 //log file
160 log using "draft.smcl" // 使用日志记录整个运行过程
161 log off //暂停日志
162
163 log on // 继续
164
165 *2.3 导入和储存数据
166 *可以在窗口打开，也可以使用命令调入
167 * 如果不是dta格式，可以导入后储存
168 import excel "地市级人口数03-19.xls", sheet("Sheet1") firstrow
169
170 drop in 1/2 // drop line 1 to 2
171
172 save sample_nov17.dta, replace
173
174 *save与save as不同；save将覆盖原始数据并保存新数据；
175 *覆盖和另存为
176
177
178 *2.4 加标签
179 label data "This is the first datasent we created in Nov 17, 2024"
180
181
182 *2.5 导出日志
183 *假设你已经做完了很多操作，把回归结果和需要的都操作完成了。这里我们插入display为例
184 *stata可以当成计算器， 在display后面输入算数，可以帮你计算
185 dispaly 2+2
186 display 589/8932
187 *但当带引号后，你输入啥它输出啥
188 display "589/8932"
189 display "今天是第一次课"
190
191 log off //暂停日志
192 display "这行会运行，但不会出现在日志里"
193 log on // 继续
194 display "这行会运行，也会出现"
195 log close // 关闭
196
197 translate "draft.smcl" "draft.log",replace //
198 translate "draft.smcl" "draft.pdf"
199
200 clear all
201
202 *-----
203 cd "E:\Desktop\Stata ch1"
204 use sample_nov17.dta

```

```

205
206 log using "nov 17.smcl" // 使用日志记录整个运行过程
207
208 label data "This is the first datasent we created in Nov 17, 2024"
209
210
211 *2.6 global 与 local
212
213 global /*
214 作为永久性盒子，全局暂元 (global macros) 一旦被定义，那么就可以在全局被调用，
215 包括：
216 直接在命令窗口中输入命令
217 运行整个 do 文件或 do 文件中的某段代码
218 编写的程序 (program) */
219
220
221 local /*
222 作为一次性盒子，局部暂元 (local macros) 在 do文件中只能执行一次，
223 执行过后会释放暂元内容，下次使用时需要重新定义，
224 并且在其他的 do 文档或者同一 do 文档其他代码块中无法识别*/
225
226 sysuse auto, clear
227 * 将表示汽车特征的一组变量放入全局暂元 size 中
228 global size " weight length trunk displacement "
229 regress price $size
230
231 clear
232 *-----
233
234
235 *-----
236 *-> 3 整理清洗数据的语法
237 *-----
238
239
240 *3.1 浏览数据 describe edit
241 cd "E:\Desktop\Stata ch1 "
242 use sample_nov17.dta
243
244 rename 人口集聚能力 population
245
246 edit
247
248 log using "nov 17.smcl" // 使用日志记录整个运行过程
249
250 label data "This is the first datasent we created in Nov 17, 2024"
251
252 describe
253 codebook
254
255
256
257
258 *****
259 encode 城市序号,gen(city)
260 encode 城市,gen(cityID) // it looks like string, but essentially it is stored as "numbers"
261
262 sum city
263 brow city
264 label variable city "city id"
265
266 destring 年份,gen(year)
267 tab year
268
269 /*
270 if string values are all numbers (e.g., 年份, 工资), use destring.
271 If string values are not numbers (e.g., 城市, 性别), use encode.
272

```

```

273 */
274 destring 人均GDP元, gen(gdp_per)
275
276
277 destring 公共图书馆图书总藏量千册件中国城市数据库, gen(books)
278
279
280 destring 行政区域土地面积平方公里中国城市数据库, gen(km_sq)
281
282 destring 专利授权数, gen(patent)
283
284 destring population, replace //converst data type only, do not generate new variable
285
286 encode 所属省份, gen(pr)
287
288 *****
289
290
291
292
293 *3.2 给变量重命名 /加标签
294 rename pr province
295 label var province "看起来是文字, 实际上是数字的省份变量"
296
297 rename cityID cityname
298 label var cityname "city name"
299
300 *3.3 描述性分析
301 summarize tabstat
302
303 /*
304 1. sum命令
305 (1) sum命令功能sum是summarize的缩写。
306 sum命令的功能为计算并显示各种单变量摘要统计信息。
307 如果未指定变量名, stata默认计算数据集中所有变量的摘要统计信息
308 默认计算的统计量包括: 平均值、标准差、最小值和最大值。
309 若需要计算更多其他的统计量, 则需要使用option中的detail。
310
311 (2) 命令语法sum 【变量名】 [if] [in] [weight] [, options]
312 if: 样本条件。in: 样本范围。weight: 样本权重。
313 options: detail: 显示其他统计信息, 包括偏度, 峰度, 四个最小值和四个最大值以及各种百分位数。
314 meanonly: 隐藏结果; 只计算平均值。
315 format: 使用变量的显示格式。
316 separator(#): 在每个#变量之后绘制分隔线, 指定将分隔线插入输出的频率。
317 默认值为separator(5), 这意味着每五个变量后画一条线。
318 separator(10)会在每10个变量后画一条线。
319 separator(0)抑制分离线。display_option: 控制间距, 线宽, 基本和空单元格。
320
321 2. tabstat命令 (1) tabstat命令功能tabstat可以计算变量的相关基本统计量。
322 (2) 命令语法tabstat 【变量名】 [if] [in] [weight] [, options]
323 if: 样本条件。in: 样本范围。weight: 样本权重。options: by 【变量名】: 按照该变量分组。
324 stat(statname): 输出统计量。
325 c(s): 结果表格以列形式展示。nototal: 不报告总计部分。format: 显示格式。
326 */
327
328
329 browse
330 brow 城市序号
331
332 brow 城* // 查看含有xx的变量
333
334 sum //summarize
335 sum 城市序号
336 ta 城市序号
337
338
339 *常用的指标有平均值(mean)、中位数(median)、四分位数(p25), 方差(variance)、标准差(sd)等等。
340 *数据的集中趋势一般采用平均值、中位数表示。数据的离散程度一般采用方差、标准差表示

```

```

341
342 *3.4 生成 保留变量 改变顺序
343 generate // gen 生成新的变量
344 keep //保留一部分变量
345 order //改变数据位置顺序
346
347 *3.5 数据转换  encode 红转蓝（限于第二种红色）
348 encode 城市序号,gen (city)
349 encode 城市,gen (cityID)
350
351 sum city
352 brow city
353 label variable city "city id"
354
355 destring 年份, gen(year)
356
357 tab year
358 /*
359 If string values are all numbers (e.g., 年份, 工资), use destring.
360 If string values are not numbers (e.g., 城市, 性别), use encode.
361 string -> red text
362 encode -> blue the numbers have meaning
363 destring -> black number
364 */
365
366 tab 城市
367
368 destring 人均GDP元, gen(gdp_per)
369
370 //找到最大gdp的城市
371 brow 城市 所属省份 年份 if gdp_per==467749
372
373
374 destring 公共图书馆图书总藏量千册件中国城市数据库, gen(books)
375 sum books
376
377 destring 行政区域土地面积平方公里中国城市数据库, gen(km_sq)
378 destring 专利授权数, gen(patent)
379 destring population, replace //仅转换数据类型，不生成新变量
380
381 encode 所属省份, gen(pr)
382
383 *3.6 利用sum的结果
384 sum books
385 gen books_mean=r(mean) //r(mean)代表前一变量的均值
386
387 gen books_min=r(min)
388 brow books_mean
389
390 *3.7 对变量进行运算操作
391 gen lgdp_per=ln(gdp_per) //取对数
392 gen gdp_per2=gdp_per^2 // 平方
393 gen pop_intense=population/km_sq //算比率
394
395 *3.8 if condition语句
396 gen gdp_per2013=gdp_per if year==2013
397 gen gdp_perother=gdp_per if year!=2013
398
399
400
401 *-----
402 *-> 4 其他数据处理基本操作
403 *-----
404
405 *4.1 生成虚拟变量 （0 1 dummy）
406
407 tab year
408

```

```

409 gen temp=1 if year >=2010 & year<2017
410
411 *replace temp=0 if temp==. // be careful
412
413 replace temp=0 if year<2010 | year>=2017 //safer
414
415 *快速为每个xx创建虚拟变量
416
417 brow 所属省份 pr
418 gen pro=pr
419
420 tab 所属省份, gen(p) //create dummy variable for each province
421
422 describe p1-p27
423
424 *4.2 给数据按组别xxxx
425 brow pr year population
426 *sort data
427
428 sort pr
429 sort pr year
430 sort pr year population
431
432 bys pr: tab cityID
433 bys year: tab 所属省份
434 bys year: sum population
435 *bys和discribe不能连用
436 * bysort xxx: sum xxx
437 * bysort xxx: tab xxx 按谁分类进行分组展示
438
439
440
441 gen c=city // link blue srting to number
442 brow city cityID 所属省份 c
443
444 brow cityID 所属省份 if c==5
445
446 *每个省有多少个城市?
447 bys pr year: gen city_ny=_N
448
449 gen unknow=_N //total obs
450 count if pro==1&year==2004 //count
451
452 *会展现计数过程(from 1 to n)
453 bys pr year: gen city_littlen=_n
454
455
456
457
458 *4.3 loop循环 (选讲)
459 foreach x of varlist 专利授权数 政府科学支出万元 从事科技活动的人数{
460     destring `x',gen(`x'1)
461 }
462
463
464 forvalues i = 1(1)5 {
465     display `i'
466 }
467
468 forvalues i = 2010(2)2017{
469     replace previousy=`i'-1 if year==`1'
470 }
471
472 drop previousy
473
474 gen previousy=.
475 forvalues i =2010(2)2017{
476     replace previousy=`i'-1 if year==`1'

```



```

477 }
478
479 *4.4 合并联合数据库（选讲）
480 /*
481 append & merge
482 append: 不同种变量直接合并
483 merge: 变量是相同的；需要找到ID
484 */
485
486 clear all
487
488 use "sample_nov17.dta"
489
490 keep 年份 所属省份 城市 政府科学支出万元 从事科技活动的人数
    公共图书馆图书总藏量千册件中国城市数据库 行政区域土地面积平方公里中国城市数据库
491
492 save "keep.dta"
493
494 clear all
495 cap log close
496 use "sample_sep10.dta"
497
498 drop 政府科学支出万元 从事科技活动的人数 公共图书馆图书总藏量千册件中国城市数据库
    行政区域土地面积平方公里中国城市数据库
499
500 save "drop.dta"
501
502 ** 将上述两个数据集通过识别信息合并 merge the above two datasets with the identification
    information
503 clear all
504 cap log close
505
506 use "drop.dta"
507 merge 1:1 年份 所属省份 城市 using "keep.dta"
508
509
510 destring 人均GDP元, gen(gdp_per)
511
512 hist gdp_per
513
514
515
516
517 *****
518 *附注参考归纳
519
520 *1、快捷键
521 Ctrl+D 执行Do文档中命令
522 *2、打开auto美国1978年汽车行业横截面数据
523 sysuse "auto.dta", clear
524 *3、数据导入和导出
525 cd E:\stata\results
526 sysuse use auto.dta
527 export excel auto, firstrow(variables)
528 import excel auto.xls, firstrow cleardescribe
529 *4、编辑查看数据
530 edit
531 desc
532 browse
533 *5、数据类型转换
534 destring 字符转数值
535 destring [varlist],[generate (newvarlist)|replace] [options]
536 tostring 数值转字符
537 tostring [varlist],[generate (newvarlist)|replace] [options]
538 *6、数据标签
539 label data " " //表示将该组stata数据起标签为
540 label var y " " //表示将变量y起标签为
541 *为观测值加标签Define value label

```



```
542 label define lblname # "label" [# "label" ...] [, add modify replace nofix]
543 *7、升序降序
544 sort varlist [in] [, stable]
545 *案例
546 sysuse auto
547 keep make mpg weight
548 sort mpg
549 help gsort
550 gsort [+|-] varname [[+|-] varname ...] [, generate(newvar) mfirst]
551
552 sysuse auto
553 #in ascending order of price升序
554 gsort price
555 gsort +price
556 *降序
557 *in descending order of price降序
558 gsort -price
559 *8、数据合并
560 *数据合并merge
561 *1:1合并
562 merge 1:1 varlist using filename [, options]
563 *m:1合并
564 Many-to-one merge on specified key variables
565 merge m:1 varlist using filename [, options]
566 *1:m合并
567 One-to-many merge on specified key variables
568 merge 1:m varlist using filename [, options]
569 *m:m合并
570 Many-to-many merge on specified key variables
571 merge m:m varlist using filename [, options]
572
573 *9、数据追加
574 append using filename [filename ...] [, options]
575
576
577
578
579
580
581
582
583
```