

Stata 基础训练营

第一讲 软件基础与数据处理

江俊澄

jc_jiang@stu.zuel.edu.cn

中南财经政法大学财政税务学院
学生学术研究中心 副主编

2024.11.17



课前问卷

STATA: STATistics Aided by Technology, Applications, and Software.



训练营的设计

经济学的学术研究在如今，理论模型离不开实证的检验，而随着大数据时代的来临，处理经济数据是经济学学生的必修技能。

在以往的课程学习中，更多地注重于计量方法的介绍和讲解，但运用计量方法处理数据的实操环节较为贫乏。

根据学院师生们的反馈和建议，似乎有必要组织一些研讨培训，对基本的数据处理和 stata 操作进行入门性的介绍。于此，Stata 基础训练营应运而生。

本训练营的定位是 Stata 的入门小白，因此不会涉及过为复杂的操作技术，从最基本的安装和认识 stata 开始，逐步讲解如何对数据进行清洗、整理、可视化、回归和结果输出；最后以 top 期刊上公开的 code 为例，向大家演示复现论文的实操过程。

学员构成

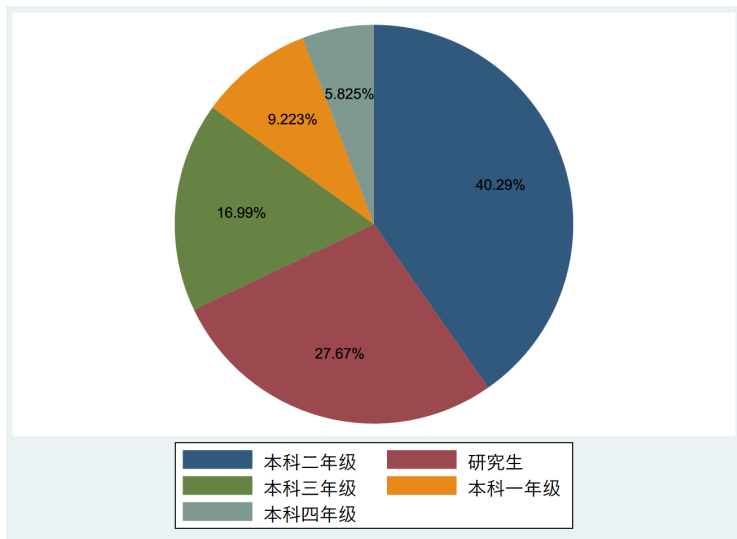


图: 学员构成

Stata 软件版本介绍

Stata 每一期常见的有三个版本：MP SE BE

- ▶ Stata/MP：最快的 Stata 版本（适用于四核、双核和多核/多处理器计算机），可以分析最大的数据集
- ▶ Stata/SE：标准版；对于更大的数据集
- ▶ Stata/BE：基础版；用于中型数据集

简单来说，三个版本功能比较 $MP > SE > BE$ ，而此次我们分享的版本为 MP 版本，为 Stata 最高配版。

目前，Stata 已经更新到第 18 版。安装步骤参见训练营预备篇安装教程。

Stata 第一版发行于 1985 年 1 月。

Stata 的窗口和菜单栏

- ▶ 命令窗口
- ▶ 结果窗口
- ▶ 历史窗口
- ▶ 变量窗口
- ▶ 属性窗口



Stata 中的文件格式

- ▶ Stata 中可以识别的数据文件名以.dta 作为后缀；你可以在搜集整理数据是以.xml 或.csv 先行储存，之后导入到 stata 中进行保存，导入的方式会即将介绍。

Stata 中的文件格式

- ▶ Stata 中可以识别的数据文件名以.dta 作为后缀；你可以在搜集整理数据是以.xml 或.csv 先行储存，之后导入到 stata 中进行保存，导入的方式会即将介绍。
- ▶ 与其他编程语言类似，你可以在将所有的命令写进 do file 里，从而执行该文件而不上选择人机交互地逐次输入命令，do 文件的存储后缀为.do

Stata 中的文件格式

- ▶ Stata 中可以识别的数据文件名以.dta 作为后缀；你可以在搜集整理数据是以.xml 或.csv 先行储存，之后导入到 stata 中进行保存，导入的方式会即将介绍。
- ▶ 与其他编程语言类似，你可以在将所有的命令写进 do file 里，从而执行该文件而不上选择人机交互地逐次输入命令，do 文件的存储后缀为.do
- ▶ 此外，当你运行 do file 时可以选择记录你的运行过程，stata 有专门的日志命令可以帮你记录，它的后缀是.smcl(只能用 stata 打开)，不过你可以通过 translate 将其转换为.log 或者.pdf 格式

数据整理形式——Tidy Data

一般我们将“在列中包含变量，在行中包含观测值”的数据成为整洁数据 (tidy data);

有三个相互关联的规则使数据集整洁 (Wickham, H. 2014):

1. 每个变量都是一列; 每一列都是一个变量
2. 每个观测值都是一行; 每一行都是一个观测值
3. 每个值都是一个单元格; 每个单元格都是一个值

country	year	cases	population
Afghanistan	1999	181	199871
Afghanistan	2000	266	2045360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21996	12852583

variables

country	year	cases	population
Afghanistan	1999	181	199871
Afghanistan	2000	266	2045360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21996	12852583

observations

country	year	cases	population
Afghanistan	1999	181	199871
Afghanistan	2000	266	2045360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21996	12852583

values

其他组织方式

table1

```
#> # A tibble: 6 × 4
#>   country    year cases population
#>   <chr>      <dbl> <dbl>      <dbl>
#> 1 Afghanistan 1999    745  19987071
#> 2 Afghanistan 2000   2666  20595360
#> 3 Brazil      1999  37737  172006362
#> 4 Brazil      2000  80488  174504898
#> 5 China       1999 212258 1272915272
#> 6 China       2000 213766 1280428583
```

table3

```
#> # A tibble: 6 × 3
#>   country    year rate
#>   <chr>      <dbl> <chr>
#> 1 Afghanistan 1999 745/19987071
#> 2 Afghanistan 2000 2666/20595360
#> 3 Brazil      1999 37737/172006362
#> 4 Brazil      2000 80488/174504898
#> 5 China       1999 212258/1272915272
#> 6 China       2000 213766/1280428583
```

table2

```
#> # A tibble: 12 × 4
#>   country    year type      count
#>   <chr>      <dbl> <chr>      <dbl>
#> 1 Afghanistan 1999 cases         745
#> 2 Afghanistan 1999 population 19987071
#> 3 Afghanistan 2000 cases         2666
#> 4 Afghanistan 2000 population 20595360
#> 5 Brazil      1999 cases         37737
#> 6 Brazil      1999 population 172006362
```

#> # i 6 more rows

Spread across two tibbles

table4a # cases

```
#> # A tibble: 3 × 3
#>   country    `1999` `2000`
#>   <chr>      <dbl> <dbl>
#> 1 Afghanistan    745    2666
#> 2 Brazil        37737  80488
#> 3 China          212258 213766
```

table4b # population

```
#> # A tibble: 3 × 3
#>   country    `1999` `2000`
#>   <chr>      <dbl> <dbl>
#> 1 Afghanistan 19987071  20595360
#> 2 Brazil      172006362  174504898
#> 3 China      1272915272  1280428583
```

常见问题

五种常见的混乱的数据集问题（补救措施参见原文）：

- ▶ 列标题是值，而不是变量名。
- ▶ 多个变量存储在一列中。
- ▶ 变量存储在行和列中。
- ▶ 同一表中存储了多种类型的观测单位。
- ▶ 一个观测单位存储在多个表中

经济数据的结构

通常来说, 可以将经济数据分为三/四类:

- ▶ **横截面数据 cross-section data**
 - ▶ 由在给定时间点采集的个人、家庭、公司、城市、国家或各种其他单位的样本组成。
 - ▶ 忽略微小的时差, 在一年内不同月份对不同的人进行了调查。
- ▶ **时间序列数据 time series data**
 - ▶ 由对一个或多个变量随时间变化的观察组成。
 - ▶ 很少假设经济数据在时间上是独立的。
 - ▶ 需要关注数据搜集时的数据频率。(每天/周/月/季/年?)
- ▶ **(混合横截面数据 pooled cross section)**
 - ▶ 同时具有横截面和时间序列特征。
如: 2022 年调查了一组人, 2024 年调查了新的一组人。
- ▶ **面板数据 panel data**
 - ▶ 由数据集中每个横截面单位的一个时间序列组成。

判断结构类型

Data Editor (Browse)												
海报[7]											http://img5.mtime.cn/mtime/2023/07/26/144520.68448668_1280X720X2.jpg	
	电影名	电影id	年代	海报	时光评分	评论人数	导演	编剧	类型	国家地区	Variable	
4	丛林大冒险	274737	2019	http://img5.mtime.cn	0.0	1人评分	姚瑶·梅纳德	Martin BehnkeMarina...	动画/冒险/喜剧	比利时	Name	
5	阿尔米塔什博物馆：艺术的力...	274325	2019	http://img5.mtime.cn	0.0	0人评分			纪录片	意大利	☒ 抓取	
6	京门烽火	268322	2020	http://movie.mtime.cn	0.0	0人评分	刘子铭		2D	中国	☒ 电影	
7	野猫入保	270412	2021	http://img5.mtime.cn	0.0	7人评分	陈翠梅	陈翠梅	剧情	中国香港	☒ 年代	
8	拉莫尔：年轻的画圣	274660	2021	http://img5.mtime.cn	0.0	0人评分			纪录片	意大利	☒ 海报	
9	涌公河巨兽	274736	2022	http://img5.mtime.cn	0.0	0人评分	李·透坎	李·透坎	惊悚/科幻/恐怖	澳大利亚	☒ 时光	
10	热烈	272254	2022	http://img5.mtime.cn	7.7	201人评分	大鹏	大鹏苏彪	剧情/喜剧	中国	☒ 评论	
11	女王的肖像	274659	2022	http://img5.mtime.cn	0.0	0人评分			纪录片	意大利	☒ 导演	
12	消失的她	273294	2022	http://img5.mtime.cn	6.8	327人评分	崔睿	陈思诚张翥杨阳一德		中国	☒ 编剧	
13	火山挚恋	272736	2022	http://img5.mtime.cn	8.4	73人评分	Sara Dosa	Shane BorisErin Cas...		中国	☒ 类型	
14	爱犬奇缘	270236	2022	http://img5.mtime.cn	0.0	0人评分	陈国辉			中国	☒ 国家	
15	欢乐丛林之外星人大冒险	274180	2023	http://img5.mtime.cn	0.0	1人评分	邢海鹏			中国	☒ 发行	
16	最遗憾是错过你	274094	2023	http://img5.mtime.cn	0.0	0人评分			爱情/剧情	中国台湾	☒ 片长	
17	学爸	270250	2023	http://img5.mtime.cn	0.0	34人评分	苏亮	苏亮	剧情/喜剧	中国	☒ 上映	
18	忠爱无疆2	262838	2023	http://img5.mtime.cn	0.0	0人评分	谈宜之	谈宜之	剧情	中国	☒ 又名	
19	潜艇总动员：环游地球80天	274622	2023	http://img5.mtime.cn	0.0	1人评分				中国	☒ mtime	
20	北京2022	273015	2023	http://img5.mtime.cn	0.0	4人评分	陆川		纪录片/剧情/运动	中国	Property	
21	雪域少年	272275	2023	http://img5.mtime.cn	0.0	0人评分	路奇	王运生路武梅		中国	☒ Name	
22	巨齿鲨2：深渊	270273	2023	http://img5.mtime.cn	6.5	209人评分	本·维特利	乔·霍贝尔埃里克·霍...	动作/科幻/冒险	中国	☒ Label	
23	长安三万里	272782	2023	http://img5.mtime.cn	8.3	423人评分	谢君伟	红泥小火炉	动画/历史	中国	☒ Type	
24	力量密码	274483	2023	http://img5.mtime.cn	0.0	2人评分			剧情/历史/传记	中国	☒ Form	
25	看不见的他们	274797	2023	http://img5.mtime.cn	0.0	0人评分	李飞		家庭/惊悚/恐怖	中国	☒ Value	
26	川流不“熄”	270506	2023	http://img5.mtime.cn	0.0	7人评分	冯钶予		剧情	中国	☒ Notes	
27	暗杀风暴	255768	2023	http://img5.mtime.cn	0.0	37人评分	邱礼涛	沈锡然	悬疑/犯罪	中国	☒ Data	
28	GT赛车：极速狂飙	273470	2023	http://img5.mtime.cn	0.0	6人评分	尼尔·布洛姆坎普	杰森·霍尔扎克·贝林...	动作/剧情/冒险	美国	☒ Frame	
29	我经过风暴	270490	2023	http://img5.mtime.cn	0.0	34人评分	秦海燕	秦海燕	家庭/剧情	中国	☒ File	
30	百变马丁：怪咖大作战	274470	2023	http://img5.mtime.cn	0.0	3人评分	张天晓		动画/冒险	中国	☒ Label	
31	超能一家人	262067	2023	http://img5.mtime.cn	6.7	143人评分	张阳	张阳张皓然李楠	喜剧/家庭/喜剧	中国		
Vars: 17 Order: Dataset Obs: 53											Length: 65	

图：混合截面数据

最常用的类型--面板数据

平衡面板数据是指在一个特定的时间段内，每个观测对象都有完整的观测值。即对于**每一个体**，**每一时间点上的数据都是齐全的**。

Data Editor (Browse) — wc14.dta

var620[11]

	pid	year	age	male	drink_lastyr
1	010104102001	2011	57	0	0.None
2	010104102001	2013	59	0	0.None
3	010104102001	2015	61	0	0.None
4	010104102001	2018	64	0	0.None
5	010104102002	2011	60	1	1.Yes
6	010104102002	2013	62	1	1.Yes
7	010104102002	2015	64	1	1.Yes
8	010104102002	2018	67	1	1.Yes
9	010104104001	2011	61	1	0.None
10	010104104001	2013	63	1	0.None
11	010104104001	2015	64	1	0.None
12	010104104001	2018	67	1	0.None
13	010104104002	2011	53	0	.m:missing
14	010104104002	2013	55	0	0.None
15	010104104002	2015	57	0	0.None
16	010104104002	2018	60	0	1.Yes
17	010104105001	2011	66	0	0.None
18	010104105001	2013	69	0	0.None
19	010104105001	2015	70	0	0.None
20	010104105001	2018	73	0	0.None
21	010104105002	2011	68	1	1.Yes
22	010104105002	2013	70	1	1.Yes
23	010104105002	2015	72	1	1.Yes
24	010104105002	2018	75	1	0.None
25	010104107001	2011	79	1	0.None
26	010104107001	2013	81	1	0.None
27	010104107001	2015	83	1	0.None
28	010104107001	2018	86	1	0.None

Vars: 5 of 617 Order: Dataset Obs: 36,080

图：平衡面板数据

Data Editor (Browse) — wc14.dta

var626[9]

	pid	year	age	male	drink_lastyr
1	010104101001	2015	50	0	0.None
2	010104101001	2018	53	0	0.None
3	010104101002	2013	50	1	1.Yes
4	010104101002	2015	52	1	1.Yes
5	010104101002	2018	55	1	0.None
6	010104102001	2011	57	0	0.None
7	010104102001	2013	59	0	0.None
8	010104102001	2015	61	0	0.None
9	010104102001	2018	64	0	0.None
10	010104102002	2011	60	1	1.Yes
11	010104102002	2013	62	1	1.Yes
12	010104102002	2015	64	1	1.Yes
13	010104102002	2018	67	1	1.Yes
14	010104103001	2013	50	1	1.Yes
15	010104103001	2015	52	1	1.Yes
16	010104103001	2018	55	1	1.Yes
17	010104103002	2015	50	0	0.None
18	010104103002	2018	53	0	0.None
19	010104104001	2011	61	1	0.None
20	010104104001	2013	63	1	0.None
21	010104104001	2015	64	1	0.None
22	010104104001	2018	67	1	0.None
23	010104104002	2011	53	0	.m:missing
24	010104104002	2013	55	0	0.None
25	010104104002	2015	57	0	0.None
26	010104104002	2018	60	0	1.Yes
27	010104105001	2011	66	0	0.None
28	010104105001	2013	69	0	0.None

Vars: 5 of 617 Order: Dataset Obs: 63,135

图：非平衡面板数据

非平衡的转换

在计量分析中，平衡面板数据相比非平衡面板更易于处理。但并不是说一定要用平衡面板才能跑回归。

如果你想仅保留非平衡面板数据在给定年限中均存在的个体，从而将其转为平衡面板，具体用到的命令为：xtbalance

外部命令下载安装为：ssc install xtbalance

语法格式为：xtbalance, range(numlist) [miss(varlist)]

注意事项为：使用前需要先进行面板数据设定声明。

转化的具体操作可以参见连享会，这里不作展开讲解：

Stata 数据处理：xtbalance-非平衡面板之转换

<https://blog.csdn.net/arlionn/article/details/118815391>

数据类型

数据的三种类型（在 stata 中以颜色区分）

- ▶ 红色：文本（可以按照字母顺序进行排序） `type:string(str)`
字符串
 - ▶ 一种：纯数字 (stkcd 股票代码)
 - ▶ 另一种：纯文字 (省份名称、行业名称)
- ▶ 黑色：数字（可以做回归分析）
- ▶ 蓝色：数字-文字对应表（虚拟变量）实质上也是数字

数据整理常见步骤 Stata command

- ▶ 查看工作路径 (pwd); 改变工作路径 (cd)
- ▶ 导入 Excel 文件 (import excel), CSV 文件 (import delimited), dta 文件 (use)
- ▶ 添加变量或变量数值标签 (label)
- ▶ 生成新变量 (gen) 或统计量变量 (egen)
- ▶ 将观测值按照变量数值大小排序 (sort; gsort)
- ▶ 改变变量前后顺序 (order)
- ▶ 将数据集进行长宽转换 (reshape)
- ▶ 合并数据集 (merge)
- ▶ 删除 (drop) 或保留 (keep) 观测值或变量
- ▶ 导出 Excel, CSV 文件 (export) 或 dta 文件 (save)

STATA 环境命令举例

```
1 use xxx.dta, clear
2 keep a b c
3 gen d
4 gen dummy=1 if year>=2022 // 生成虚拟变量
5
6 .....
7
8 outreg2 using statistic, word replace sum(log) /*
9 */ sortvar(income price interest) /*
10 */ keep(price income interest)
11 * 基本回归
```

基本命令 list

```
1 use "file.dta", clear // 读取 .dta 数据
2 sort air_time dest // 对列进行排序
3 order month day // 对行进行排序
4 rename (old) (new) // 重命名
5 keep in 1/200
6 keep if day>5 & day<10
7 keep if in range(day,5,10)
8 keep month day carrier // 保留子集行或列
9 gen dist_sq=distance^2 // 创建新变量
10 replace a=dep_delay+arr_delay // 修改现有变量
11 drop varlist [if] [in] // 删除变量
12 .....
```

本讲你将学到:

```
1 use xxx.dta, clear
2 log
3 global local
4 rename
5 label
6 encode
7 destring
8 keep
9 gen
10 replace
11 sort
12 .....
```

DO!