Stata 基础教程：从数据处理到实证分析

本科高年级与硕士低年级适用

2025 年 12 月 21 日

目录

1 引言 **3**

2 Stata 基础操作 **3**

2.1 界面介绍 3

2.2 基本命令 3

2.2.1 数据管理命令 3

2.2.2 数据导入导出 4

3 数据处理 **4**

3.1 变量操作 4

3.1.1 创建和修改变量 4

3.1.2 变量标签 4

3.2 数据清理 5

3.2.1 缺失值处理 5

3.2.2 异常值处理 5

3.3 数据合并 6

4 Do 文件编写规范 **6**

4.1 Do 文件基本结构 6

4.2 编写规范 7

4.3 常用编程技巧 7

5 基本回归方法 **8**

5.1 普通最小二乘法（OLS） 8

5.1.1 基本命令 8

5.1.2 结果解释 8

5.1.3 完整示例 9

5.2 工具变量法（IV） 9
5.2.1 基本命令 9
5.2.2 内生性检验 10
5.2.3 过度识别检验 10
5.2.4 完整示例 10
5.3 双重差分法（DID） 11
5.3.1 基本命令 11
5.3.2 平行趋势检验 11
5.3.3 完整示例 11
5.4 断点回归设计（RDD） 12
5.4.1 基本命令 12
5.4.2 带宽选择 12
5.4.3 稳健性检验 13
5.4.4 完整示例 13

6 结果输出与可视化 **14**
6.1 结果输出 14
6.2 图形绘制 14

7 常见问题与解决方案 **14**
7.1 内存不足 14
7.2 大样本处理 15
7.3 面板数据处理 15

8 总结 **16**

9 参考文献与扩展阅读 **16**

10 附录：常用命令速查表 **17**

1 引言

Stata 是一款广泛应用于经济学、社会学、医学等领域的统计软件，特别适合进行实证研究和数据分析。本教程面向本科高年级和硕士低年级学生，系统介绍 Stata 的基本使用方法，包括数据处理、do 文件编写以及常用的计量经济学方法（OLS、IV、DID、RDD）的实现。

2 Stata 基础操作

2.1 界面介绍

Stata 的界面主要包含以下几个部分：

- **命令窗口（Command）**：输入 Stata 命令
- **结果窗口（Results）**：显示命令执行结果
- **变量窗口（Variables）**：显示当前数据集中的变量
- **属性窗口（Properties）**：显示变量的属性信息
- **历史窗口（Review）**：记录已执行的命令

2.2 基本命令

2.2.1 数据管理命令

```
* 查看当前工作目录
pwd

* 改变工作目录
cd "C:\Users\YourName\Documents\Stata"

* 查看当前数据
describe
summarize

* 查看前几行数据
list in 1/10

* 查看特定变量
list var1 var2 in 1/10
```

2.2.2 数据导入导出

```
* 导入Excel文件
import excel "data.xlsx", sheet("Sheet1") firstrow clear

* 导入CSV文件
import delimited "data.csv", clear

* 导入Stata格式数据
use "data.dta", clear

* 保存数据
save "newdata.dta", replace

* 导出为CSV
export delimited using "output.csv", replace
```

3 数据处理

3.1 变量操作

3.1.1 创建和修改变量

```
* 生成新变量
generate newvar = var1 + var2
gen log_income = log(income)

* 条件生成变量
gen high_income = 1 if income > 50000
replace high_income = 0 if income <= 50000

* 使用egen命令
egen mean_income = mean(income)
egen sd_income = sd(income)
egen group_id = group(city year)
```

3.1.2 变量标签

```
* 给变量添加标签
```

```
label variable income "个人年收入（元）"
label variable age "年龄"

* 给变量值添加标签
label define gender 1 "男" 2 "女"
label values gender gender
```

3.2 数据清理

3.2.1 缺失值处理

```
* 查看缺失值
misstable summarize
misstable patterns

* 删除包含缺失值的观测
drop if missing(var1)

* 删除所有变量都缺失的观测
egen miss = rowmiss(_all)
drop if miss == 0

* 用特定值替换缺失值
replace var1 = 0 if missing(var1)
```

3.2.2 异常值处理

```
* 识别异常值（使用IQR方法）
summarize income, detail
gen outlier = 1 if income > r(p99) | income < r(p1)
tab outlier

* 删除异常值
drop if outlier == 1

* Winsorize处理（缩尾处理）
winsor2 income, cuts(1 99) replace
```

3.3 数据合并

```
* 横向合并（merge）
use "data1.dta", clear
merge 1:1 id using "data2.dta"
drop _merge

* 纵向合并（append）
use "data1.dta", clear
append using "data2.dta"
```

4 Do 文件编写规范

Do 文件是 Stata 中保存命令序列的脚本文件，良好的编写习惯对于提高工作效率和代码可读性至关重要。

4.1 Do 文件基本结构

```
* ============================================
* 项目名称：XXX研究
* 作者：XXX
* 日期：2024年
* 说明：本文件用于XXX分析
* ============================================

* 清理环境
clear all
set more off

* 设置工作目录
cd "C:\Users\YourName\Documents\Stata"

* 设置输出路径
global output ".\output"
global data ".\data"

* 第一部分：数据导入
* ============================================
use "$data\raw_data.dta", clear
```

```
22
23 * 第二部分：数据清理
24 * ============================================
25 * 处理缺失值
26 drop if missing(income)
27
28 * 处理异常值
29 winsor2 income, cuts(1 99) replace
30
31 * 第三部分：描述性统计
32 * ============================================
33 summarize income age education
34
35 * 第四部分：回归分析
36 * ============================================
37 regress income age education
38
39 * 保存结果
40 save "$output\cleaned_data.dta", replace
```

4.2 编写规范

1. **注释充分**：使用星号（*）添加注释，说明每部分代码的功能
2. **分段清晰**：使用分隔线和标题将代码分为不同部分
3. **变量命名**：使用有意义的变量名，避免使用缩写
4. **全局宏**：使用 global 定义常用路径，便于修改
5. **错误处理**：使用 capture 命令捕获可能的错误
6. **结果保存**：及时保存中间结果，避免重复计算

4.3 常用编程技巧

```
1 * 使用循环处理多个变量
2 foreach var of varlist income age education {
3     summarize `var'
4     gen log_`var' = log(`var')
5 }
```

```
6
7 * 使用条件执行
8 if `condition' {
9     regress y x1 x2
10 }
11 else {
12     regress y x1 x2 x3
13 }
14
15 * 使用临时文件
16 tempfile temp1 temp2
17 save `temp1'
18 use "data2.dta", clear
19 save `temp2'
```

5 基本回归方法

5.1 普通最小二乘法（OLS）

普通最小二乘法是最基础的回归方法，用于估计线性回归模型。

5.1.1 基本命令

```
1 * 简单回归
2 regress y x
3
4 * 多元回归
5 regress y x1 x2 x3
6
7 * 添加选项
8 regress y x1 x2 x3, robust           // 稳健标准误
9 regress y x1 x2 x3, cluster(id)       // 聚类标准误
10 regress y x1 x2 x3, vce(hc3)          // HC3标准误
```

5.1.2 结果解释

回归结果主要关注：

- **系数（Coef.）**：解释变量的边际效应

- **标准误（Std. Err.）**：系数估计的不确定性
- **t 值**：用于检验系数显著性
- **P 值**：显著性水平
- **R-squared**：模型拟合优度

5.1.3 完整示例

```
* 示例：研究教育对收入的影响
use "wage_data.dta", clear

* 描述性统计
summarize wage education experience age

* OLS回归
regress wage education experience age, robust

* 保存结果
estimates store ols1

* 添加控制变量
regress wage education experience age married urban, robust
estimates store ols2

* 结果对比
estimates table ols1 ols2, star(0.1 0.05 0.01) b(%9.3f) se(%9.3f)
```

5.2 工具变量法（IV）

当解释变量与误差项相关时（内生性问题），需要使用工具变量法。

5.2.1 基本命令

```
* 两阶段最小二乘法（2SLS）
ivregress 2sls y (x = z) controls

* 有限信息最大似然法（LIML）
ivregress liml y (x = z) controls

```

```
* 广义矩估计（GMM）
ivregress gmm y (x = z) controls
```

5.2.2 内生性检验

```
* 第一阶段回归
regress x z controls
test z  // 检验工具变量的相关性（F统计量应大于10）

* 内生性检验（Hausman检验）
ivregress 2sls y (x = z) controls
estimates store iv
regress y x controls
estimates store ols
hausman iv ols
```

5.2.3 过度识别检验

```
* 当工具变量数量多于内生变量时
ivregress 2sls y (x = z1 z2) controls
estat overid  // Sargan检验或Hansen J检验
```

5.2.4 完整示例

```
* 示例：研究教育对收入的影响（使用出生季度作为工具变量）
use "wage_data.dta", clear

* 第一阶段：工具变量对内生变量的影响
regress education quarter controls, robust
test quarter  // 检查F统计量

* 第二阶段：IV回归
ivregress 2sls wage (education = quarter) experience age, robust

* 内生性检验
ivregress 2sls wage (education = quarter) experience age, robust
estimates store iv_model
regress wage education experience age, robust
estimates store ols_model
```

```
hausman iv_model ols_model
```

5.3 双重差分法（DID）

双重差分法用于评估政策或处理的效果，通过比较处理组和对照组在处理前后的差异。

5.3.1 基本命令

```
* 手动构建DID模型
gen did = treat * post
regress y treat post did, robust

* 使用reghdfe命令（推荐，可控制固定效应）
reghdfe y did, absorb(id year) cluster(id)

* 使用diff命令
diff y, t(treat) p(post)
```

5.3.2 平行趋势检验

```
* 生成事件时间变量
gen time_to_treat = year - treatment_year
replace time_to_treat = -5 if time_to_treat < -5
replace time_to_treat = 5 if time_to_treat > 5

* 事件研究法
forvalues i = -5/5 {
    gen pre`i' = (time_to_treat == `i' & treat == 1)
}
drop pre0  // 以政策实施前一年为基准

regress y pre* treat post, robust
```

5.3.3 完整示例

```
* 示例：评估最低工资政策对就业的影响
use "employment_data.dta", clear

```

```
* 生成DID变量
gen post = (year >= 2015)  // 假设2015年实施政策
gen treat = (province == "北京")  // 处理组
gen did = treat * post

* DID回归
regress employment treat post did controls, robust cluster(province)

* 使用固定效应
reghdfe employment did, absorb(province year) cluster(province)

* 平行趋势检验
gen time_to_treat = year - 2015
forvalues i = -3/3 {
    gen pre`i' = (time_to_treat == `i' & treat == 1)
}
drop pre0
regress employment pre* treat post controls, robust cluster(province)
```

5.4 断点回归设计（RDD）

断点回归利用处理变量在某个临界值处的跳跃来识别因果效应。

5.4.1 基本命令

```
* 使用rdrobust命令（推荐）
rdrobust y x, c(0)  // c(0)表示断点在0处

* 手动实现RDD
gen x_centered = x - cutoff
gen treat = (x >= cutoff)
gen x_treat = x_centered * treat

regress y treat x_centered x_treat, robust
```

5.4.2 带宽选择

```
* 使用rdrobust自动选择最优带宽
rdrobust y x, c(0) p(1)  // p(1)表示使用线性模型
```

```
* 手动指定带宽
rdrobust y x, c(0) h(10)  // h(10)表示带宽为10
```

5.4.3 稳健性检验

```
* 不同带宽
rdrobust y x, c(0) h(5)
rdrobust y x, c(0) h(10)
rdrobust y x, c(0) h(15)

* 不同多项式阶数
rdrobust y x, c(0) p(1)  // 线性
rdrobust y x, c(0) p(2)  // 二次
rdrobust y x, c(0) p(3)  // 三次

* 协变量平衡性检验
rdrobust covariate x, c(0)  // 检验协变量在断点处是否跳跃
```

5.4.4 完整示例

```
* 示例: 研究高考分数线对大学录取的影响
use "college_data.dta", clear

* 生成断点变量（假设分数线为500分）
gen score_centered = score - 500
gen above_cutoff = (score >= 500)

* 基本RDD回归
regress admission above_cutoff score_centered
   c.score_centered#c.above_cutoff, robust

* 使用rdrobust命令
rdrobust admission score, c(500) p(1)

* 不同带宽的稳健性检验
rdrobust admission score, c(500) h(10)
rdrobust admission score, c(500) h(20)
rdrobust admission score, c(500) h(30)
```

```
* 协变量平衡性检验
rdrobust gender score, c(500)  // 检验性别在断点处是否跳跃
rdrobust age score, c(500)     // 检验年龄在断点处是否跳跃
```

6 结果输出与可视化

6.1 结果输出

```
* 使用estout输出回归结果
ssc install estout
eststo clear
eststo: regress y x1 x2, robust
eststo: regress y x1 x2 x3, robust
esttab using "results.rtf", replace star(* 0.1 ** 0.05 *** 0.01) ///
    b(%9.3f) se(%9.3f) r2 ar2
```

6.2 图形绘制

```
* 散点图
scatter y x

* 添加回归线
scatter y x || lfit y x

* 箱线图
graph box income, over(education)

* 直方图
histogram income, normal

* 保存图形
graph export "figure.png", replace width(800) height(600)
```

7 常见问题与解决方案

7.1 内存不足

```
* 增加内存
set max_memory 2g

* 使用preserve和restore
preserve
* 进行某些操作
restore
```

7.2 大样本处理

```
* 使用sample命令抽取样本
sample 10  // 抽取10%的样本

* 使用collapse命令汇总数据
collapse (mean) mean_income=income (sd) sd_income=income, by(city
   year)
```

7.3 面板数据处理

```
* 声明面板数据
xtset id year

* 固定效应回归
xtreg y x, fe robust

* 随机效应回归
xtreg y x, re robust

* 豪斯曼检验
xtreg y x, fe
estimates store fe
xtreg y x, re
estimates store re
hausman fe re
```

8 总结

本教程系统介绍了 Stata 的基本使用方法，包括：

1. **基础操作**：界面介绍、基本命令、数据导入导出
2. **数据处理**：变量操作、数据清理、数据合并
3. **Do 文件编写**：编写规范、编程技巧
4. **回归方法**：OLS、IV、DID、RDD 的基本命令和应用
5. **结果输出**：结果表格和图形绘制

掌握这些基本技能后，学生应该能够独立完成大部分实证研究工作。建议通过实际项目不断练习，加深对 Stata 的理解和运用。

9 参考文献与扩展阅读

- Baum, C. F. (2006). *An Introduction to Modern Econometrics Using Stata*. Stata Press.
- Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics Using Stata*. Stata Press.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.
- Stata 官方文档：`https://www.stata.com/manuals/`

10 附录：常用命令速查表

表 1: Stata 常用命令速查表

功能	命令
数据管理	use, save, import, export
变量操作	generate, replace, egen, label
数据清理	drop, keep, missing, winsor2
数据合并	merge, append, joinby
描述统计	summarize, tabulate, correlate
回归分析	regress, ivregress, reghdfe
面板数据	xtset, xtreg, xtabond
结果输出	estimates, estout, esttab
图形绘制	scatter, line, histogram, graph box