Lab1: Introduction to Stata

Introduction to Econometrics,Spring 2024

Zishu Wang

Nanjing University

March 1,2024

Section 1

What is Stata?

What is Stata?

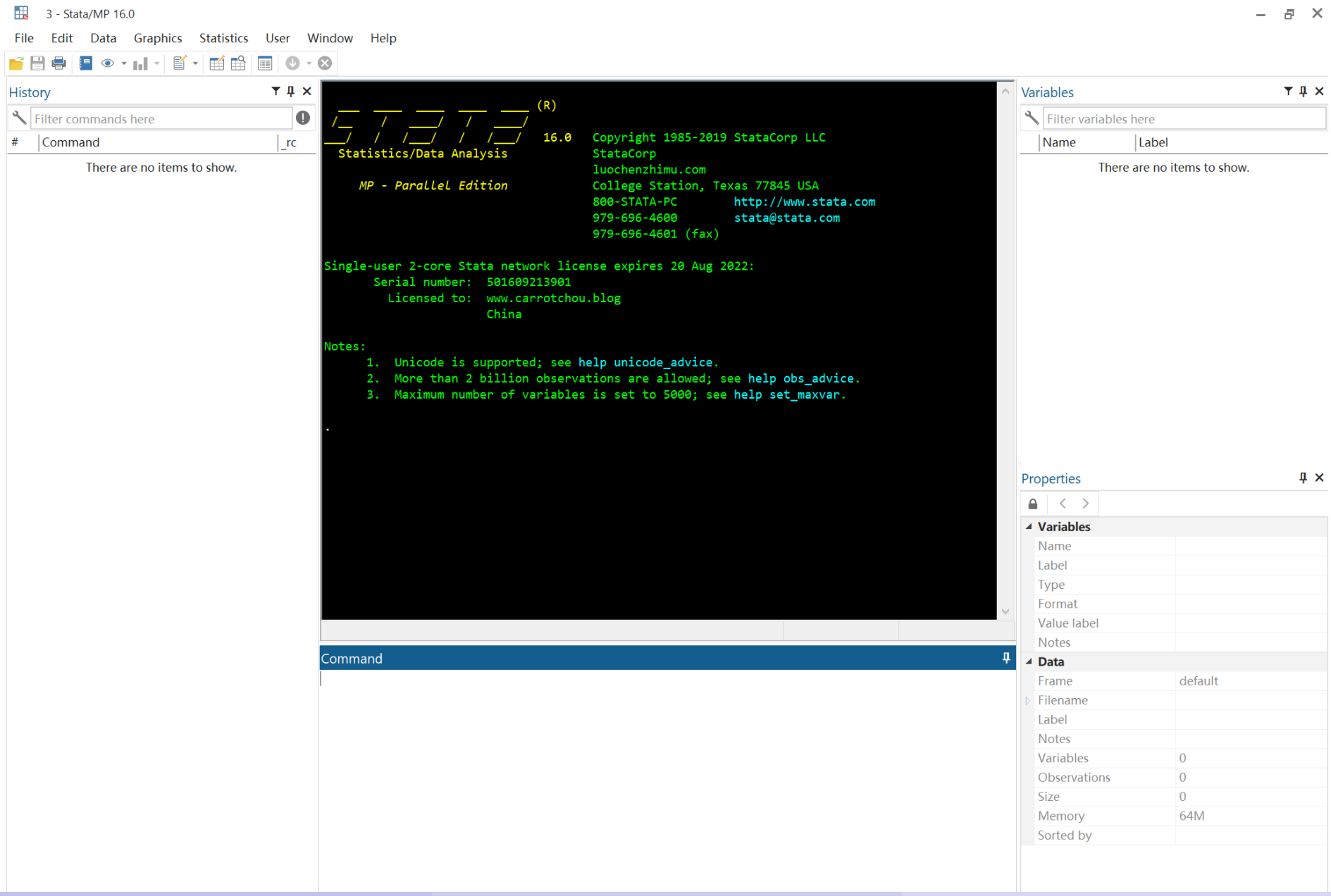
What

- statistics + data = Stata
- Stata 最初由美国计算机资源中心（Computer Resource Center）研制，现在为 Stata 公司的产品，其最新版本为 Stata 17。

Why

- Stata 是经济学研究主流的数据分析软件，功能强大快捷，程序包丰富，更新快，平台宽，几乎涵盖了应用计量经济学领域所有的功能。
- Stata 的 **help** 文件非常详细，可以自学应用。
- Stata 使现代经济学实证研究更加规范且具有可重复性。

Interface of Stata



Interface of Stata

- 五个窗口，两组菜单条
  - Command–命令窗口
  - Results–结果窗口
  - History–历史窗口
  - Variables–变量窗口
  - Properties–属性窗口
- 两种执行命令的方式
  - 菜单
  - 命令

Interface of Stata

- An easy example

```
sysuse auto,clear
des
sum
twoway (scatter mpg weight)
reg price wei len mpg
```

Learning Materials of Stata

- help 文档（详尽, 精确查找, 最佳选择）
- 搜索引擎（Google,Bing,Baidu...）
- 线上线下交流学习
- 网络学习资源汇总
  - Stata website: http://www.Stata.com/
  - Stata journal: https://www.stata-journal.com/
  - Stata FAQs: https://www.stata.com/support/faqs/
  - Stata bookstore: https://www.stata.com/bookstore/books-on-stata/
  - Code and resources: https://geocenter.github.io/StataTraining/
  - 人大经济论坛【Stata 专版】:https://bbs.pinggu.org/forum-67-1.html

Section 2

Basic Settings of Stata

Basic Settings of Stata

- 当前工作路径

```
help cd                                    //更改工作文件路径
help dir                                   //显示当前路径下的文件列表

cd "D:\Teaching2024"                       //修改文件路径
dir                                        //显示当前路径下的文件信息
cd _data
cdout                                     //打开当前工作路径对应文件夹

use "GTAs_2008.dta", clear                 //读入当前工作路径下的文件
use "D:\Teaching2024\_data\GTAs_2008.dta", clear

pwd                                        //查看当前工作路径
```

Basic Settings of Stata

- stata 系统文件设定
  - stata 命令的执行过程
  - 每次启动 Stata 都需要执行的命令如何设定

```
sysdir                                  //显示当前系统文件的放置位置

STATA:  D:\Stata17\                      // stata安装根目录
BASE:  D:\Stata17\ado\base\             //【官方命令】存储地址
SITE:  D:\Stata17\ado\site\             //【自编命令】存储地址
PLUS:  D:\Stata17\ado\plus\             //【外部命令】存储地址
PERSONAL:  D:\Stata17\ado\personal\     //【自有文件夹】没有可自建

adopath
adopath + "D:2024_plus"
```

Basic Settings of Stata

- profile 文档
  - 工作原理
  - profile.do 参考

```
基本参数设定:
    set type double           //后续产生的变量都将是双精度型
    set memory 50m            //为stata分配50M内存
    set matsize 2000          //矩阵维度
    set scrollbufsize 50000   //设定屏幕的最大显示行数
    set more off, perma       //不分屏显示

log文件设定:
    log using    "D:\stata17\personal\stata.log", text replace
    cmdlog using "D:\stata17\ado\personal\command.log", append

文件目录设定:
    sysdir set PLUS     "D:\stata17\ado\plus"  //外部命令的存放地址
    sysdir set OLDPLACE "D:\ado"
    sysdir set PERSONAL "D:\stata17\ado\personal" //个人文件夹

ado文档查找路径(可以添加其他路径):
    adopath + "D:\stata17\ado\personal"
    adopath + "D:\stata17\ado\personal\_Myado"

指定默认工作路径:
    cd "D:\stata17\ado\personal"
```

Basic Settings of Stata

- **程序运行过程中的目录管理（参考）**

```
global root "D:\毕业论文\实证部分"    //设置自己的根目录

global do   $root/Dofiles           //保存do文档
global log  $root/Logfiles          //保存log文档
global work $root/WorkData          //保存临时数据
global raw  $root/RawData           //保存原始数据
global save $root/SaveData          //保存修改后的最终数据
global fig  $root/Figures           //保存图片
global tab  $root/Tables            //保存表格

cd ${work}     //如在对数据进行处理时,可以直接指定该路径
```

Basic Settings of Stata

- Stata 的帮助文件和外部命令
  - 简介

```
help help
help guide
help tabstat
```

  - 命令检索与下载

```
help findit
findit dynamic panel

help ssc          //-ssc: Statistical Software Components
ssc install winsor, replace  //replace可以对旧版本进行更新

help ado                //查询已安装的外部命令
ado                     //呈现本机上安装的所有外部命令
ado, find(winsor)       //仅呈现包含特定关键词的外部命令
```

Basic Settings of Stata

- Stata 命令的语法格式
  - 样本范围的限制 if，in，by

```
help summ

sysuse nlsw88, clear

sum wage                          // 多数命令和选项都可以简写
sum wage if race == 1             // if 限定样本
sum wage hours if race == 1
sum wage if race != 3
sum wage if hours >= 40
sum wage in 1/10                  // in 很少用
sum wage in -5/-1                 //倒数第1至第5个
sum wage, detail                  // 选项,灵活性

bysort race: sum wage hours age  //对race进行排序再分组描述
bysort marr collgrad: sum wage hours
```

Basic Settings of Stata

- Stata 命令的语法格式
  - 一般语法格式

```
help language                          //完整的语法命令格式参考

* summarize [varlist] [if] [in] [weight] [, options]
注意：逗号后为 options，整条命令只能有一个裸露在外的逗号。
```

  - 变量的引用：通配符: *, ?, -

```
help varlist

sysuse nlsw88, clear
sum age race married never_married grade
sum age-grade     // 按顺序出现的变量，列出头尾两个变量即可
sum s*            // "*"表示任何长度的字母或数字
sum *arr*         // "*"可以用在任何位置
sum ?a?e          // "?"只能表示一个长度的字母或数字，很少使用
```

Section 3

Variables and Basic Statistics

Variables and Basic Statistics

- **变量命名**
  - **新变量命名**

```
help gen
help egen
help varname      //新变量命名规则
help newvarlist   //多个新变量
```

  - **变量重命名**

```
help rename
help rename group  //批量处理

sysuse auto, clear
rename make mk
rename (price rep78) (Price REP78)
rename mpg foreign trunk, upper    //大写
```

Variables and Basic Statistics

- 变量名称基本规则
  - 由英文字母、数字或 _ 组成，至多不超过 32 个;
  - 首字母不能为数字;
  - 英文字母大小写具有不同含义;
  - 尽量不要使用 _ 作为第一个字母, 因为许多 stata 的内部变量都是以“_” 开头, 如 _n, _N,_cons,_b 等等。

```
help _variables
```

Variables and Basic Statistics

- 查看数据结构

```
. sysuse auto, clear
(1978 Automobile Data)
. describe
Contains data from D:\Stata16\ado\base/a/auto.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          13 Apr 2018 17:45
                                              (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
make            str18   %-18s                 Make and Model
price           int     %8.0gc                Price
mpg             int     %8.0g                 Mileage (mpg)
rep78           int     %8.0g                 Repair Record 1978
headroom        float   %6.1f                 Headroom (in.)
trunk           int     %8.0g                 Trunk space (cu. ft.)
weight          int     %8.0gc                Weight (lbs.)
length          int     %8.0g                 Length (in.)
turn            int     %8.0g                 Turn Circle (ft.)
displacement    int     %8.0g                 Displacement (cu. in.)
gear_ratio      float   %6.2f                 Gear Ratio
foreign         byte    %8.0g      origin     Car type
-------------------------------------------------------------------------------
Sorted by: foreign
```

Variables and Basic Statistics

- 变量的存储类型
  - 字符型数据：字母 + 特殊符号。
    表示姓名、住址（文字信息）；性别（定性）；身份证号（数字）等。
    一般用 str# 来表示字符。
    每个汉字占两个字符。
    str18 表示 make 变量最多容纳的字符个数是 18。
  - 数值型数据：便于进行数字的算数运算。
    整数的存储类型
    byte 字节型 (-100, +100)
    int 一般整数型 (-32000,+32000)
    long 长整数型 $(-2.14^{*}10\hat{}10, +2.14^{*}10\hat{}10)$，即，正负 21 亿
    小数的存储类型
    float 浮点型 8 位有效数字
    double 双精度 16 位有效数字
  - 缺失数据：”.” 被认为大于任何数。

Variables and Basic Statistics

- 定义变量的显示格式
  - 字符型变量%#s(提示符 + 字符数 + 显示格式)

    %-18s 靠左列印;
    %18s 靠右列印;
    %~18s 居中列印。
  - 数值变量%w.d+3 种基本显示格式 (c 要求 stata 给出",")

    e.g.12345
    g 一般格式: %9.0g(12345) %9.2gc(12,345)
    f 固定格式: %9.4f(12345.0000) %9.0fc(12,345)
    e 科学计数法格式: %9.2e(1.23e+04)

    %6.2f 总共占 6 个空格，小数位占两个空格。

Variables and Basic Statistics

- 定义变量的显示格式

```
list price gear in 1/5
format price %6.1f
format gear %6.4f
```

Variables and Basic Statistics

- 数据和变量的标签
  - 样本标签

```
    . sysuse auto, clear
    (1978 Automobile Data)

    . label data "这是一份汽车价格资料"

    . des

Contains data from D:\Stata16\ado\base/a/auto.dta
  obs:            74                          这是一份汽车价格资料
 vars:            12                          13 Apr 2018 17:45
                                              (_dta has notes)
------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
make            str18   %-18s                 Make and Model
price           int     %8.0gc                Price
mpg             int     %8.0g                 Mileage (mpg)
rep78           int     %8.0g                 Repair Record 1978
headroom        float   %6.1f                 Headroom (in.)
trunk           int     %8.0g                 Trunk space (cu. ft.)
weight          int     %8.0gc                Weight (lbs.)
length          int     %8.0g                 Length (in.)
turn            int     %8.0g                 Turn Circle (ft.)
displacement    int     %8.0g                 Displacement (cu. in.)
gear_ratio      float   %6.2f                 Gear Ratio
foreign         byte    %8.0g       origin    Car type
------------------------------------------------------------------------------
Sorted by: foreign
```

Variables and Basic Statistics

- **数据和变量的标签**
  - **变量标签**

```
. label var price "汽车价格"
. label var foreign "汽车产地(1 国外; 2 国内)"
. des
Contains data from D:\Stata16\ado\base/a/auto.dta
  obs:            74                          这是一份汽车价格资料
 vars:            12                          13 Apr 2018 17:45
                                              (_dta has notes)
------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
make            str18   %-18s                 Make and Model
price           int     %8.0gc                汽车价格
mpg             int     %8.0g                 Mileage (mpg)
rep78           int     %8.0g                 Repair Record 1978
headroom        float   %6.1f                 Headroom (in.)
trunk           int     %8.0g                 Trunk space (cu. ft.)
weight          int     %8.0gc                Weight (lbs.)
length          int     %8.0g                 Length (in.)
turn            int     %8.0g                 Turn Circle (ft.)
displacement    int     %8.0g                 Displacement (cu. in.)
gear_ratio      float   %6.2f                 Gear Ratio
foreign         byte    %8.0g       origin    汽车产地(1 国外; 2 国内)
------------------------------------------------------------------------------
Sorted by: foreign
```

Variables and Basic Statistics

- **数据和变量的标签**
  - **值标签（数字—文字对应表）**

```
browse rep78

Step1: label define          //定义标签内容
label define rep78 1 "很好" 2 "较好" 3 "中等" 4 "较差" 5 "很差"

Step2: label value          //将变量与标签内容关联起来
label value rep78 rep78

browse rep78

label list rep78      //列出值标签的名称和内容
label drop rep78      //删除标签

browse rep78
```

Variables and Basic Statistics

- **查看和查找变量**

```
help des          //变量概况 (样本数; 变量数; 变量名称…)
help ds           //列示变量名称
help des2         //外部命令,提供链接便于查看数据特征

sysuse nlsw88, clear
des
ds
ds, alpha         //按照字母先后顺序显示

ssc install des2
des2              //更加方便获取变量信息

help lookfor      //查找变量
```

Variables and Basic Statistics

- 基本统计量
  - -summarize-命令

```
. sysuse auto, clear
(1978 Automobile Data)
. summarize mpg weight if foreign
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mpg	22	24.77273	6.611187	14	41
weight	22	2315.909	433.0035	1760	3420

Variables and Basic Statistics

- 基本统计量
  - -codebook-命令

```
. codebook price
-------------------------------------------------------------------------------
price
-------------------------------------------------------------------------------

                  type:  numeric (int)
                 range:  [3291,15906]                 units:  1
         unique values:  74                       missing .:  0/74
                  mean:   6165.26
             std. dev:    2949.5
           percentiles:        10%       25%       50%       75%       90%
                              3895      4195    5006.5      6342     11385
```

Variables and Basic Statistics

- 基本统计量
  - -codebook-命令

```
. codebook rep78          //变量中的非重复值小于9，视为类别变量
```

```
rep78                                                                 Repair Rec
```

```
                  type:  numeric (int)
                 range:  [1,5]                        units:  1
        unique values:  5                    missing .:  5/74
           tabulation:  Freq.  Value
                          2  1
                          8  2
                         30  3
                         18  4
                         11  5
                          5  .
```

Variables and Basic Statistics

- **基本统计量**
  - **列表统计** table，tabulate

```
webuse nhanes2l, clear
des

table (highbp) ()        //为行变量创建表，默认显示频率和总频率
                         //第二个括号可省略
table () (highbp)        //为列变量创建表,第一个括号必需

table (sex) (highbp)      //为行变量和列变量创建交叉表

table (sex) (highbp), nototals
table (sex) (highbp), totals(highbp)  //totals中包含指定变量名称

table (sex) (highbp),                ///
    statistic(frequency)             ///
    statistic(percent)               ///
    statistic(mean age)              ///
    statistic(sd age)                ///
    nototals
```

Variables and Basic Statistics

- 基本统计量
  - 列表统计 table，tabulate

```
table (sex) (highbp),                  ///
    stat(frequency)                   ///
    stat(percent)                     ///
    stat(mean age)                    ///
    stat(sd age)                      ///
    nototals                          ///
    nformat(%9.0fc frequency)         ///
    nformat(%6.2f mean sd)
```

Variables and Basic Statistics

- 基本统计量
  - 列表统计-table-, -tabulate-

```
. sysuse auto,clear
(1978 Automobile Data)
. tabulate foreign
```

Car type	Freq.	Percent	Cum.
Domestic	52	70.27	70.27
Foreign	22	29.73	100.00
Total	74	100.00	

```
. table foreign
```

Car type	Freq.
Domestic	52
Foreign	22

Variables and Basic Statistics

- 基本统计量
  - 列表统计-table-, -tabulate-

```
. tabulate foreign rep78, summarize(mpg)
```

Means, Standard Deviations and Frequencies of Mileage (mpg)

	Repair Record 1978					
Car type	1	2	3	4	5	Total
Domestic	21	19.125	19	18.444444	32	19.541667
	4.2426407	3.7583241	4.0856221	4.5856055	2.8284271	4.7533116
	2	8	27	9	2	48
Foreign	.	.	23.333333	24.888889	26.333333	25.285714
	.	.	2.5166115	2.7131368	9.367497	6.3098562
	0	0	3	9	9	21
Total	21	19.125	19.433333	21.666667	27.363636	21.289855
	4.2426407	3.7583241	4.1413252	4.9348699	8.7323849	5.8664085
	2	8	30	18	11	69

Variables and Basic Statistics

- **基本统计量**
  - **列表统计** table，tabulate

```
help tabulate

sysuse auto, clear

table foreign
tabulate foreign            //会报告变量所占比重和累积比重
tabulate foreign, nol       //不显示值标签
tabulate rep78,sort         //按频数降序排序
tab rep78,g(new)            //对变量进行频率分析并生成多个虚拟变量
```

Variables and Basic Statistics

- 基本统计量
  - 统计表格-tabstat-

```
. sysuse auto,clear
(1978 Automobile Data)
. tabstat price weight length
```

stats	price	weight	length
mean	6165.257	3019.459	187.9324

```
. tabstat price weight length, stats(mean med min max) col(s) format(%6.2f)
```

variable	mean	p50	min	max
price	6165.26	5006.50	3291.00	15906.00
weight	3019.46	3190.00	1760.00	4840.00
length	187.93	192.50	142.00	233.00

```
. tabstat price weight length, s(mean p25 med p75 min max) c(s) f(%6.2f)
```

variable	mean	p25	p50	p75	min	max
price	6165.26	4195.00	5006.50	6342.00	3291.00	15906.00
weight	3019.46	2240.00	3190.00	3600.00	1760.00	4840.00
length	187.93	170.00	192.50	204.00	142.00	233.00

Variables and Basic Statistics

- 基本统计量
  - 统计表格-tabstat-

```
. tabstat price weight length, s(mean sd p25 med p75 min max) c(s) f(%6.2f) by(foreign)
Summary for variables: price weight length
     by categories of: foreign (Car type)
```

foreign	mean	sd	p25	p50	p75	min	max
Domestic	6072.42	3097.10	4184.00	4782.50	6234.00	3291.00	15906.00
	3317.12	695.36	2790.00	3360.00	3730.00	1800.00	4840.00
	196.13	20.05	179.50	200.00	209.50	147.00	233.00
Foreign	6384.68	2621.92	4499.00	5759.00	7140.00	3748.00	12990.00
	2315.91	433.00	2020.00	2180.00	2650.00	1760.00	3420.00
	168.55	13.68	156.00	170.00	175.00	142.00	193.00
Total	6165.26	2949.50	4195.00	5006.50	6342.00	3291.00	15906.00
	3019.46	777.19	2240.00	3190.00	3600.00	1760.00	4840.00
	187.93	22.27	170.00	192.50	204.00	142.00	233.00

Section 4

Do Files

Do Files

- 图形化界面的局限：
  - 命令不易保存、修改，软件关闭，命令即消失；
  - 操作繁琐，每次操作都要不断重复点击界面；
  - 功能组合有限，自由度低，不能进行软件开发。
- command&review 窗口的局限：
  - 命令历史记录保存在 Review 窗口中，查找困难；
  - 零碎命令无条理，无法组织复杂的操作；
  - 与图形化界面类似，command 窗口命令也无法长期保存。
- 所以我们需要一个记录、编辑命令的编辑器，Stata 自带的编辑器即 do 文件编辑器，功能类似 txt 文档，所生成的文件扩展名为【.do】，也就是 do 文件。
- do 文件实际上是 Stata 命令的集合，方便我们一次性执行多条命令，且使我们的分析工作具有可重复性。

Do Files

- **打开和新建 do 文档**
  - **方法一：一些快捷键**

Ctrl-key(Windows)	Ctrl-key(Mac)	Definition
Ctrl+D	Command+Shift+D	执行(Do)选中的命令(*)
Ctrl+R	Command+Shift+R	运行程序(Run)(*)
Ctrl+F	Command+F	在do-editor中搜索特定的关键词
Ctrl+O	Command+O	打开do文档
Ctrl+N	Command+N	新建do文档
Ctrl+S	Command+S	保存do文档(*)

(*) 表示仅适用于do-editor

Do Files

- **打开和新建 do 文档**
  - **方法二**

```
doedit           //打开do-editor
doedit auto.do //打开一个已存在的do文档，可指定完整路径
```

  - **方法三**：Rusults **窗口按钮**
  - **设置界面属性**

Do Files

- 执行 do 文档
  - 部分执行：选中需要执行的命令点击点击 doedit 窗口中的图标或使用快捷键
    `Ctrl+D` (Windows) ,`Command+shift+D`(Mac);
  - 整体执行：

```
do auto.do
```

Do Files

- **注释语句**

```
help comments       //永远不要相信自己的记忆力...

clear all
sysuse auto

*三种注释示例:
*第一种注释方式
sum price weight    /*查看price与weight变量部分统计量*/
gen x = 5           //生成取值为5的变量x
```

Do Files

- **断行（“///” ,“/* */”,#delimit 命令）**

```
*-第一种断行方式： /// 物理断行，逻辑一行

sysuse auto, clear  //调用数据
sum price weight length gear turn
tabstat price weight length gear turn ,                ///
       stats(mean sd p5 p25 med p75 p95 min max)   ///
       format(%6.2f) c(s)

*-第二种断行方式： /* */

sysuse auto, clear
sum price weight length gear turn
tabstat price weight length gear turn ,          /*
*/ stats(mean sd p5 p25 med p75 p95 min max)   /*
*/ format(%6.2f) c(s)
```

Do Files

- **三种断行方式**：“///” ,“/* */”,#delimit **命令**

```
*-第三种断行方式：  #delimit 命令

#delimit ;          //delimit声明，表示出现";"才结束

sysuse auto, clear; des; sum;
twoway (scatter price wei)
       (lfit price wei),
       title("散点图和线性拟合图");

#delimit cr         //退回到回车断行模式
```

Do Files

- 注意事项
  - 各段代码采用一个或多个空行加以分隔，每一行的语句不要过长。
  - 在一段代码内适当位置妙用空格加以分隔提升可读性和美观性
  - 注意中英文字符的切换，尤其是逗号，引号
  - 等于号 ==
  - 尽量避免使用系统预留字段作为变量名，如"_"
  - Stata 对大小写敏感

Section 5

Log File

5. 录屏神器:log 文件

```
log using "$root\lab1_0916.log" //新建Log文件

log using "$root\lab1_0916.log",append //接着原来记录

log using "$root\lab1_0916.log",replace   //覆盖掉原来

matrix input a = (1,2\3,4)
matrix list a
matrix input b = (1,2\1,1)
matrix list b
```

5. 录屏神器:log 文件

```
log off // 暂停录制

matrix c = a+b

log on // 继续录制

matrix list c

log close //结束录制

shellout "$root\lab1_0916.log"
```

Reference

- 连玉君 Stata 初级教程讲义
- Stata 统计分析与应用 (第 3 版). 电子工业出版社
- https://data.princeton.edu/stata/markdown