Lab1: Introduction to Stata

Introduction to Econometrics,Fall 2021

Mingxin Zhao

NJU

9/8/2021

What is Stata?

- Stata是经济学研究主流的数据分析软件，它功能强大，程序包丰富，几乎涵盖了应用计量经济学领域所有的函数和模型。
- Stata的**help**文件非常详细，完全可以自学，上手容易。
- Stata使现代经济学实证研究更加规范且具有可重复性。
- 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**。

Stata的使用界面

- 四个窗口，两组菜单条
  - 1.Command
  - 2.Results
  - 3.History
  - 4.Variables
- 两种执行命令的方式
  - 1.菜单
  - 2.命令

Stata的使用界面

- An easy example

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

Stata的一些基本设定

- 初次使用时Preference的设定
  - 1.设定方法

Edit–>preference–>General Preference 按喜好设定

注：可进一步设定图形偏好、do-editor的风格等

  - 2.保存设定

Edit–>Preference–>Save...–>New... 任意输入一个名称，如kobold

```
window manage prefs save kobold
```

  - 3.调入已有的界面偏好设定

Edit–>Preference–>Load...–>选择你喜欢的设定

```
window manage prefs load kobold
```

- 4.删除设定

Edit–>Preference–>Delete...–>选择你要删除的设定

Stata基本设定

- 文件目录

```
. sysdir                    //显示当前系统目录的设定
   STATA:  F:\Stata16\
    BASE:  F:\Stata16\ado\base\
    SITE:  F:\Stata16\ado\site\
    PLUS:  F:/Stata16/ado/plus\
PERSONAL:  F:/Stata16/ado/personal\
OLDPLACE:  F:/Stata16/ado/oldplace\
. pwd                       //当前工作路径
F:\0001\Teaching\Labs\Drafts\Lab1_0831
. sysdir set PLUS "F:/Stata16/ado/plus"  // 外部命令的存放地址
. sysdir set PERSONAL "F:/Stata16/ado/personal" // 个人文件夹
. sysdir set OLDPLACE "F:/Stata16/ado/oldplace" // 旧址
```

数据导入

- 实证分析的第一步：数据处理！收集数据、存储、修改、分析、输出结果。Stata所直接处理的是dta文件，类似txt文档，占用存储空间小，可以直接在菜单栏打开。
- 导入数据的三种方式：
  - 1.手动输入
  - 2.从txt或Excel文档中粘贴
  - 3.使用Stata命令

数据导入

- 手动输入 (极少使用)

```
clear
input x y z
1 2 3
4 5 6
end
save mydata, replace // 保存数据
use mydata, clear // 调入数据
```

数据导入

- 从txt或Excel文档中粘贴

```
shellout d1.txt //copy-paste(【!open】 in Mac)
shellout d1.xls //copy-paste

edit    // 打开数据编辑器
```

数据导入

- 使用Stata命令
- dta文件的导入

```
clear all
use "F:\Teaching\Stata\lab1\auto1.dta"          //修 改成自己的工作路径

use "F:\Teaching\Stata\lab1\auto1.dta", clear   //修 改成自己的工作路径

cd "F:\Teaching\Stata\lab1\"
use auto1.dta,clear

global root "F:\Teaching\Stata\lab1\"           //global定义全局宏
cd "$root"
use auto1.dta,clear
```

数据导出

- -export-导出

```
sysuse auto, clear
export excel auto2.xlsx
export excel make mpg weight using auto, replace
```

- *记得删除电脑上保存输出的数据，免得占内存

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 F:\Stata16\ado\base/a/auto.dta
  obs:             74                          1978 Automobile Data
 vars:             12                          13 Apr 2018 17:45
                                               (_dta has notes)
```

variable name	storage type	display format	value 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

- 更改变量的存储类型

```
sysuse auto, clear
list gear_ratio in 1/5
d gear_ratio
recast int gear_ratio, force
d gear_ratio
list gear_ratio in 1/5

compress // 自动精简资料的存储格式
```

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：靠左列印于屏幕上；若%18s，则靠右列印；若%~18s, 则居中列印。
  - 数值变量–%w.d+3种基本显示格式(c要求stata给出“，”)
  - 一般格式g：%w.dg %w.dgc
  - 固定格式f: %w.df %w.dfc
  - 科学计数法格式e：%w.de
  - %6.2f 总共占6个空格，小数位占两个空格。
  - :%9.0g(12345);%9.2gc(12,345);%9.4f(12345.0000);%9.0fc(12,345);%9.2e(1

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
list price gear in 1/5
```

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 F:\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
```

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 F:\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)
```

Variables and Basic Statistics

- 数据和变量的标签
  - 值标签（数字和文字相对应）

```
browse
label define repair 1 "好" 2 "较　　好" 3 "中" 4 "较差" 5 "差" //定义一个标签名repair
label values rep78 repair   ////将变量值和标签联系起来
browse
```

Variables and Basic Statistics

- 数据和变量的标签
  - 管理标签

```
label list              //列出值标签的名称和内容
label drop repair       //删除repair
label list
```

```
. labelbook          // 推荐使用
```

```
value label origin
```

```
      values                                  labels
       range:  [0,1]                  string length:  [7,8]
           N:  2                  unique at full length:  yes
        gaps:  no                   unique at length 12:  yes
missing .*:  0                             null string:  no
                                 leading/trailing blanks:  no
                                      numeric -> numeric:  no

definition
           0  Domestic
           1  Foreign
 variables:  foreign
```

Variables and Basic Statistics

- 数据和变量的标签
  - 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 — 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 Record 1978
```

```
                  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-

```
. 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-

```
. table foreign rep78, c(mean price) f(%9.2f) center row col
```

Car type	Repair Record 1978					
	1	2	3	4	5	Total
Domestic	4564.50	5967.63	6607.07	5881.56	4204.50	6179.25
Foreign			4828.67	6261.44	6292.67	6070.14
Total	4564.50	5967.63	6429.23	6071.50	5913.00	6146.04

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

Variables and Basic Statistics

- 基本图形分析
  - 直方图: 样本的总体分布情况

```
. sysuse auto,clear
(1978 Automobile Data)
. histogram mpg
(bin=8, start=12, width=3.625)
. graph export h1.png, width(500) replace
(note: file h1.png not found)
(file h1.png written in PNG format)
```

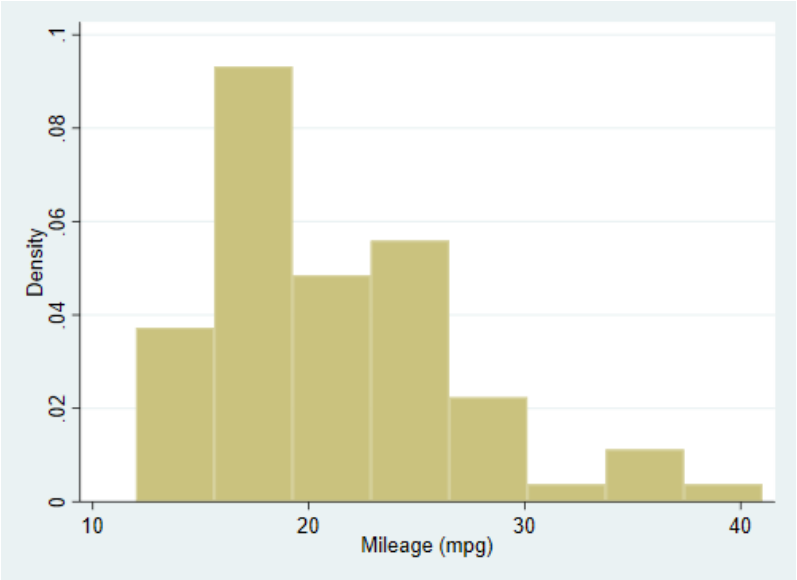
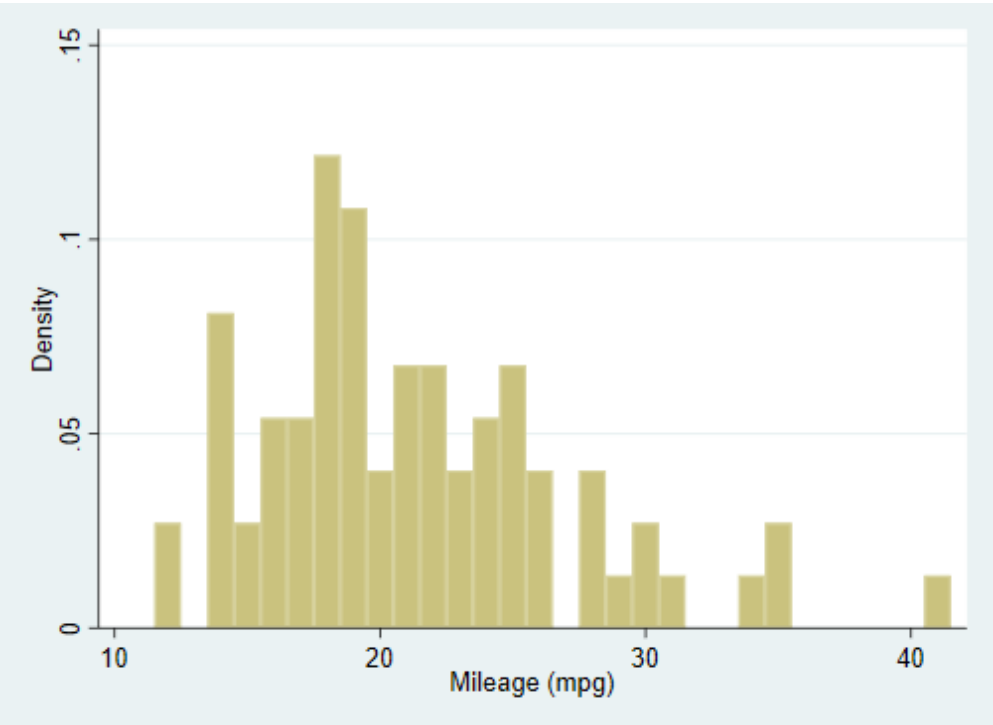


图 1: price&weight

Variables and Basic Statistics

- 基本图形分析
  - 直方图: 样本的总体分布情况

```
. histogram mpg, discrete    //discrete makes a histogram with a bin for each of the 21
>  unique values.
(start=12, width=1)
. graph export h2.png, width(500) replace
(note: file h2.png not found)
(file h2.png written in PNG format)
```



Variables and Basic Statistics

- 基本图形分析
  - 直方图: 样本的总体分布情况

```
. histogram mpg, discrete by(foreign)
. graph export h3.png, width(500) replace
(note: file h3.png not found)
(file h3.png written in PNG format)
```

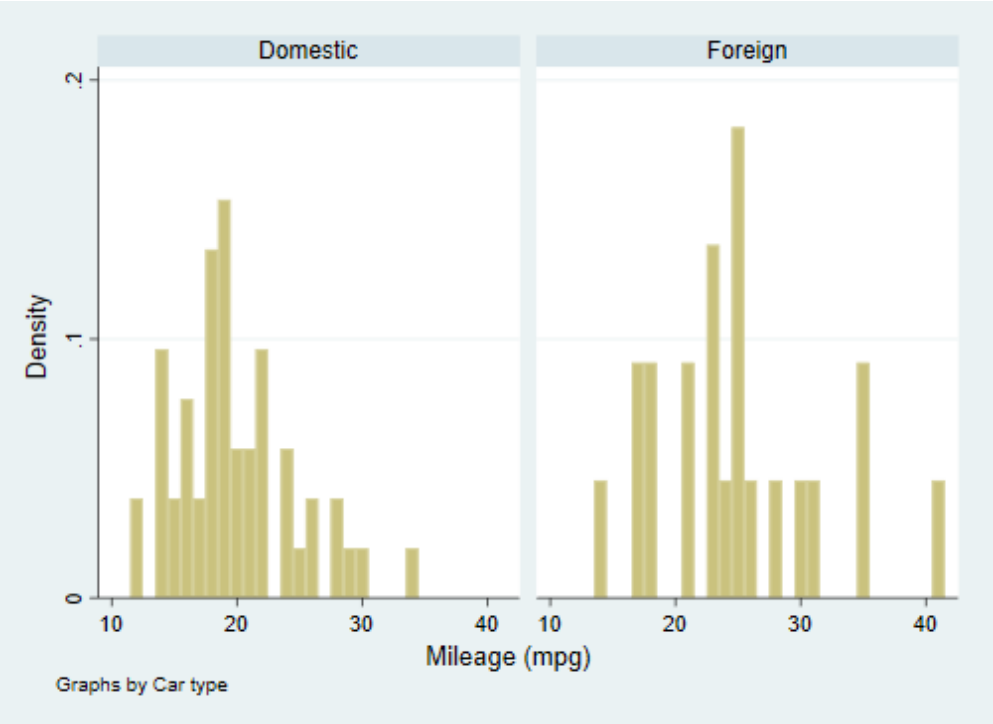
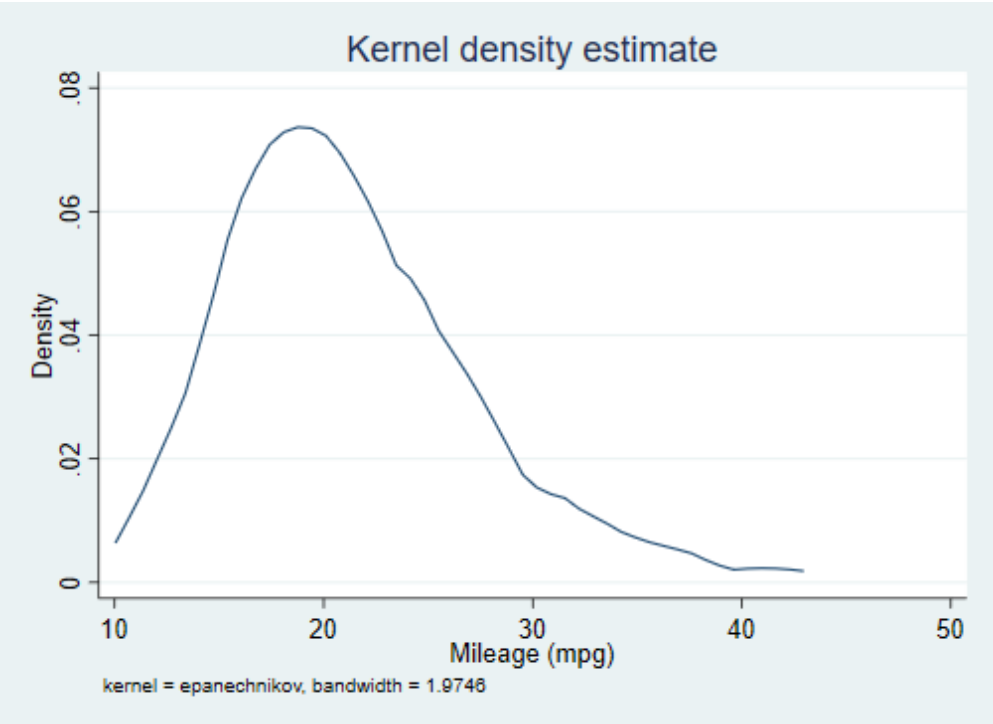


图 3: price&weight

Variables and Basic Statistics

- 基本图形分析
  - 密度函数图

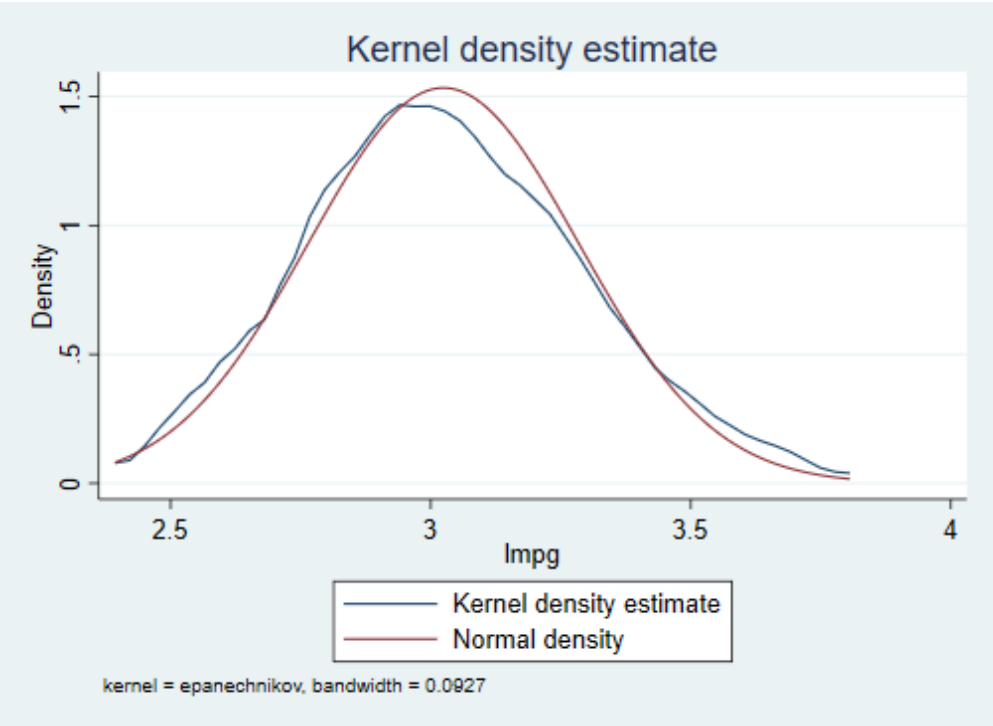
```
. sysuse auto,clear
(1978 Automobile Data)
. kdensity mpg
. graph export k1.png, width(500) replace
(note: file k1.png not found)
(file k1.png written in PNG format)
```



Variables and Basic Statistics

- 基本图形分析
  - 密度函数图

```
. gen lmpg=ln(mpg)
. kdensity lmpg,normal
. graph export k2.png, width(500) replace
(note: file k2.png not found)
(file k2.png written in PNG format)
```



Variables and Basic Statistics

- 散点图

```
. sysuse auto, clear
(1978 Automobile Data)
. scatter mpg turn
. graph export s1.png, width(500) replace
(note: file s1.png not found)
(file s1.png written in PNG format)
```

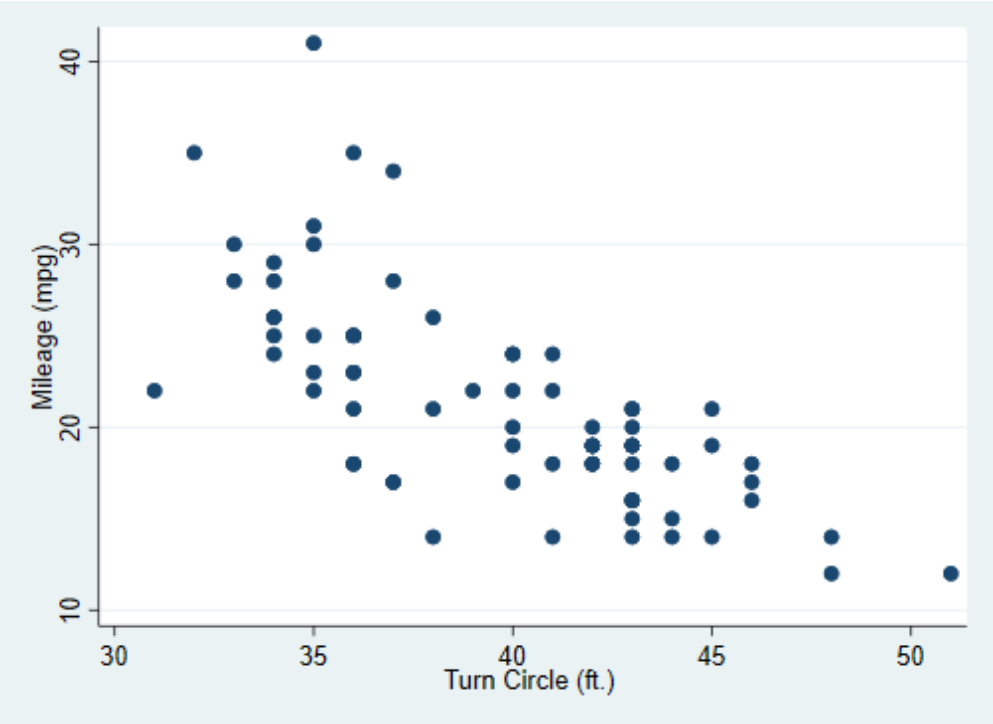


图 6: price&weight

Variables and Basic Statistics

- 相关系数矩阵

```
. sysuse auto, clear
(1978 Automobile Data)
. graph matrix price wei len mpg
. graph export m1.png, width(500) replace
(note: file m1.png not found)
(file m1.png written in PNG format)
```

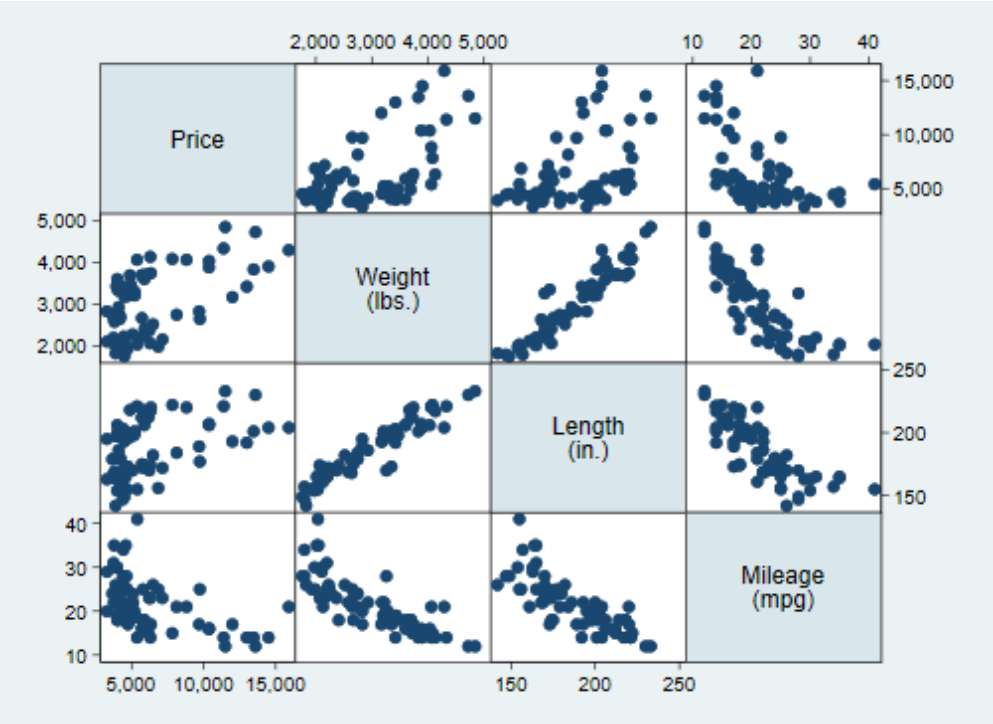


图 7: price&weight

Do Files

- 图形化界面的局限：图形命令不易保存和修改；操作繁琐；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:快捷键（常用）

Ctrl-key(Windows)	Ctrl-key(Mac)	Definitio
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-e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
- 执行do文档
  - 部分执行快捷键：选中需要执行的命令，ctrl+D（Windows）,Command+shift+D(Mac);
  - 整体执行：

```
do auto.do
```

Do Files

- 注释语句

```
help comments
clear all
sysuse auto

*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

- 三种断行方式：“///”、“/* */”、 #delimit 命令

```
*-第一种断行方式:  ///
sysuse auto, clear  //调用数据
sum price weight length gear turn
tabstat price weight length gear turn ,                  /// 物
理断行, 逻辑一行
        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)

*-第三种断行方式:  #delimit 命令      ///表示出现;才结束
```

Do Files

- 注意事项
  - Stata对大小写敏感
  - 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"_"
  - 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，不用拖动下方引导条即可阅读

5.录屏神器:log文件

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

```
log using "$root\first.log",replace 表示覆盖原来的日志文件重新记录
```

```
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
```