

Lab2: Data Management

Introduction to Econometrics, Spring 2023

Jiayi Cheng

Nanjing University

10/03/2023

Section 1

Preparation

Preparation

- Change your own path.

```
global root "D:\Teaching\Stata\lab2\  
cd ${root}
```

- Download data and slides to the current directory.

Section 2

Command

- 1.stata 命令的通用格式

- ▶ `command varlist [if] [in] [, options]`

[if] [in] 用于限制样本范围

[options] “可选项”，增加了命令的弹性

“[]” 为可选项，可以不填

```
help sum //帮助文件的解读
```

```
sysuse nlsw88, clear
```

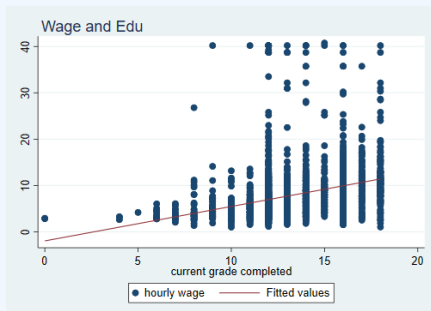
```
sum wage hours ttl_exp if race==2, detail
```

Command

1.stata 命令的通用格式

- ▶ 整条命令“裸露”的逗号只能有一个，“,”前为命令主体，“,”后为选项
- ▶ 选项中可能有子选项，但子选项前的逗号并未“裸露”

```
. twoway (scatter wage grade)           ///  
(lfit wage grade),                   ///  
  title("Wage and Edu", place(left))  
  
. graph export Wage_Edu.png, width(500) replace
```



1.Command

- 2. 变量列举 [varlist]

```
. sum age race married never_married grade
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|--------------|-------|----------|-----------|-----|-----|
| age | 2,246 | 39.15316 | 3.060002 | 34 | 46 |
| race | 2,246 | 1.282725 | .4754413 | 1 | 3 |
| married | 2,246 | .6420303 | .4795099 | 0 | 1 |
| never_marr_d | 2,246 | .1041852 | .3055687 | 0 | 1 |
| grade | 2,244 | 13.09893 | 2.521246 | 0 | 18 |

```
. sum age-grade
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|--------------|-------|----------|-----------|-----|-----|
| age | 2,246 | 39.15316 | 3.060002 | 34 | 46 |
| race | 2,246 | 1.282725 | .4754413 | 1 | 3 |
| married | 2,246 | .6420303 | .4795099 | 0 | 1 |
| never_marr_d | 2,246 | .1041852 | .3055687 | 0 | 1 |
| grade | 2,244 | 13.09893 | 2.521246 | 0 | 18 |

1.Command

2. 变量列举 [varlist]

▶ 通配符匹配规则 [*] [?]

```
. sum s* // "*"可以表示【任何长度】的字母或数字
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|----------|-----------|-----|-----|
| south | 2,246 | .4194123 | .4935728 | 0 | 1 |
| smsa | 2,246 | .7039181 | .4566292 | 0 | 1 |

```
. sum ?a?e // "?"只能替代【一个长度】的字母或数字
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|----------|-----------|----------|----------|
| race | 2,246 | 1.282725 | .4754413 | 1 | 3 |
| wage | 2,246 | 7.766949 | 5.755523 | 1.004952 | 40.74659 |

1.Command

● 3. 样本范围的限制 **[if]** **[in]** **[by]**

```
. sum in 10/20           // 第10至第20个观察值之间的观察值
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|--------------|-----|----------|-----------|----------|----------|
| idcode | 11 | 21.09091 | 6.155559 | 14 | 36 |
| age | 11 | 40.09091 | 1.445998 | 37 | 42 |
| race | 11 | 1 | 0 | 1 | 1 |
| married | 11 | .8181818 | .4045199 | 0 | 1 |
| never_marr_d | 11 | .0909091 | .3015113 | 0 | 1 |
| grade | 11 | 14.63636 | 1.026911 | 12 | 16 |
| collgrad | 11 | .3636364 | .504525 | 0 | 1 |
| south | 11 | 0 | 0 | 0 | 0 |
| smsa | 11 | .9090909 | .3015113 | 0 | 1 |
| c_city | 11 | .0909091 | .3015113 | 0 | 1 |
| industry | 11 | 10.27273 | 1.678744 | 6 | 11 |
| occupation | 11 | 1.636364 | 1.433369 | 1 | 5 |
| union | 10 | .3 | .4830459 | 0 | 1 |
| wage | 11 | 9.731538 | 3.78571 | 4.180602 | 16.79548 |
| hours | 11 | 33.72727 | 13.72655 | 4 | 50 |
| t1l_exp | 11 | 14.35198 | 3.471585 | 7.384615 | 17.75 |
| tenure | 11 | 5.757576 | 3.630782 | .9166667 | 13.83333 |

1.Command

● 3. 样本范围的限制 **[if]** **[in]** **[by]**

```
. sum wage in -5/-1 // 倒数第1至第5个
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|----------|----------|
| wage | 5 | 6.654148 | 4.638684 | 2.447664 | 14.32367 |

```
. sum wage hours if race == 1 // 等于
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|----------|-----------|----------|----------|
| wage | 1,637 | 8.082999 | 5.955069 | 1.004952 | 40.19808 |
| hours | 1,635 | 36.90398 | 11.28842 | 1 | 80 |

```
. sum wage if race _= 3 // 不等于(the same as !=)
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|----------|-----------|----------|----------|
| wage | 2,220 | 7.757769 | 5.762044 | 1.004952 | 40.74659 |

```
. sum wage if hours >= 40 // 大于等于
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|---------|-----------|----------|----------|
| wage | 1,487 | 8.41767 | 6.281296 | 1.392914 | 40.19808 |

1.Command

- 3. 样本范围的限制 **[if]** **[in]** **[by]**

```
. bysort race:sum wage //先对race进行排序，再分组进行描述
```

```
-> race = white
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|----------|-----------|----------|----------|
| wage | 1,637 | 8.082999 | 5.955069 | 1.004952 | 40.19808 |

```
-> race = black
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|----------|----------|
| wage | 583 | 6.844558 | 5.076187 | 1.151368 | 40.74659 |

```
-> race = other
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|---------|----------|
| wage | 26 | 8.550781 | 5.20943 | 1.80602 | 25.80515 |

1.Command

3. 样本范围的限制 **[if]** **[in]** **[by]**

```
. sum wage if (race==2) & (married==1) // 且
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|----------|----------|
| wage | 274 | 6.968853 | 5.309318 | 1.344605 | 40.74659 |

```
. sum wage if (race==3)|(married==0) // 或
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|----------|----------|
| wage | 822 | 8.092208 | 6.313993 | 1.151368 | 40.19808 |

4. 命令作用的增减 **[, options]**

- ▶ 多数命令和选项都支持缩写, “_” 的部分表示可缩写的程度。

Section 3

Data Management

- 1. 数学表达式

- ▶ 关系运算符 `[==] [>] [<] [>=] [<=] [!=] [~=]`
- ▶ 逻辑运算符 `[& (and)] [| (or)]`
- ▶ 算术运算符 `[+ - * / ^]`

```
. display 5^2  
25
```

```
. dis 1-3*2+4/5-9^3  
-733.2
```

```
. dis 2*_pi  
6.2831853
```

```
. display cos(_pi)  
-1
```

- 2. 建立新变量-gen-

```
sysuse nlsw88.dta, clear

generate ttl_exp2 = ttl_exp^2 //生成工作经验二次项

gen ttl_exp3 = ttl_exp^2 if race==2
//仅生成黑人妇女的工作经验二次项

gen wwage = wage*hours //生成周工资
```

- 2. 建立新变量-gen-
 - ▶ 常用数学函数转换

```
gen ln_wage = ln(wage)           // 取对数
gen sqrt_hours = sqrt(hours)     // 开根号
gen int_wage = int(wage)         // 取整
gen floor_wage = floor(wage)     // 等价于取整
gen ceil_wage = ceil(wage)       // 取整数上限
list *wage in 1/5
```


- 3. 更改旧变量和观察值

- ▶ 变量重命名-`rename`-

```
rename grade edu_level //rename 【旧变量名】 【新变量名】
```

- 3. 更改旧变量和观察值

- ▶ 修改观察值-replace-

```
replace hours = 40 if (hours > 40)
//假如法定工作周时间不超过40小时

sysuse auto, clear
list make in 50/59
replace make="宝马 320i" if (make=="BMW 320i")
//文字变量观察值的修改要加【""】

list make in 50/59
```

● 4.An Example

```
. sysuse nlsw88.dta, clear
(NLSW, 1988 extract)

. list married if never_married==1

. *生成新变量married2, 并赋值
. gen married2 = 0 if never_married==1 //未婚
(2,012 missing values generated)
. replace married2 = 1 if married==1 //在婚
(1,442 real changes made)
. replace married2 = 2 if (married==0) & (never_married==0)
//离婚
(570 real changes made)

. label var married2 "婚姻状况2" //生成变量标签

. label define marr 0 "未婚" 1 "在婚" 2 "离婚" //生成值标签
. label value married2 marr
```

- 4.An Example

```
. tab married2
```

```
//列表统计
```

| 婚姻状 况2 | Freq. | Percent | Cum. |
|-----------|-------|---------|--------|
| 未婚 | 234 | 10.42 | 10.42 |
| 在婚 | 1,442 | 64.20 | 74.62 |
| 离婚 | 570 | 25.38 | 100.00 |
| Total | 2,246 | 100.00 | |

Data Management

- 5. 变量与样本的保存、删除
 - ▶ 变量与样本的保存-keep-

```
sysuse nlsw88.dta, clear
```

```
keep wage race ttl_exp //只保留这三个变量数据  
keep in 1/5
```

- ▶ 变量与样本的删除-drop-

```
drop if race==2  
drop wage-ttl_exp  
drop _all //删除内存中的所有变量  
  
drop wage //由于已经不存在任何变量，所以报错  
capture drop wage //加capture，不会报错
```

- 5. 变量与样本的保存、删除

- ▶ Expansions—【capture 的作用】

不显示结果（类似于 `qui`）；

将错误代码返回给 `__rc`（系统标量），若该命令未出错，则 `__rc` 值为 0，程序会跳出 `capture` 段，执行后续命令；

如果既想显示结果，又想不让程序因错误而终止并返回错误值，则使用 `capture noisely +cmd`。

同样适用于其它命令。

- 6. 变量的移动、克隆

- ▶ 变量的移动-order-

```
sysuse nlsw88.dta, clear  
  
order wage race ttl_exp  
order _all,alpha           //按字母排序
```

- ▶ 变量的克隆-clonevar-

```
clonevar race2=race  
//把已有变量的标签，数字-文字对应表等都复制过去  
gen race3=race
```

- 7. 样本的排序-sort-

```
sort wage // 默认为升序排列
list wage in 1/10
dis "max = " wage[_N]

gen nag_wage = -wage
sort nag_wage // 降序排列

gsort -wage // 降序排列
list wage in 1/10
```


Section 4

More About Creating Variables

More About Creating Variables

- 1. `_n` 和 `_N`

- ▶ `_n` 是一个变量: 1,2,3,...
- ▶ `_N` 是一个单值: 样本数

```
sysuse nlsw88.dta, clear
list _n                // 错误
gen id_n = _n

dis _N                 //单值
gen id_N = _N
```

More About Creating Variables

- 1. `_n` 和 `_N`

- ▶ 应用

```
sort wage
gen wage_diff = wage[_N] - wage[1] //range

bysort race: gen gid = _n           //分组
```

More About Creating Variables

- 2. 生成虚拟变量

- ▶ 常规操作

```
sysuse nlsw88.dta, clear

gen dum_race2=0
replace dum_race2=1 if race==2
gen dum_race3 = 0
replace dum_race3=1 if race==3
```

More About Creating Variables

- 2. 生成虚拟变量

- ▶ -tab-命令

```
sysuse nlsw88.dta, clear  
tab race, gen(dum_r)
```

More About Creating Variables

- 2. 生成虚拟变量

- ▶ 利用条件生成 0-1 虚拟变量的函数 (可自行学习)

`cond()`

`inlist()`

`inrange()`

e.g. `inrange(x, a,b)`

1 if $a \leq x \leq b$;

0 otherwise.

More About Creating Variables

- 3. 将连续变量转为类别变量
 - ▶ 等分样本 group ()

```
. sysuse nlsw88.dta, clear
(NLSW, 1988 extract)
. sort wage //必须先排序
. gen g_wage = group(5) //等分为5组
. tab g_wage
```

| g_wage | Freq. | Percent | Cum. |
|--------|-------|---------|--------|
| 1 | 450 | 20.04 | 20.04 |
| 2 | 449 | 19.99 | 40.03 |
| 3 | 449 | 19.99 | 60.02 |
| 4 | 449 | 19.99 | 80.01 |
| 5 | 449 | 19.99 | 100.00 |
| Total | 2,246 | 100.00 | |

More About Creating Variables

- 3. 将连续变量转为类别变量

- ▶ 等分样本 group ()

```
. tabstat wage, stat(N mean med min max) by(g_wage) f(%4.2f)  
Summary for variables: wage  
by categories of: g_wage
```

| g_wage | N | mean | p50 | min | max |
|--------|---------|-------|-------|-------|-------|
| 1 | 450.00 | 3.12 | 3.22 | 1.00 | 4.03 |
| 2 | 449.00 | 4.68 | 4.69 | 4.03 | 5.43 |
| 3 | 449.00 | 6.32 | 6.27 | 5.43 | 7.31 |
| 4 | 449.00 | 8.73 | 8.67 | 7.32 | 10.27 |
| 5 | 449.00 | 16.00 | 12.78 | 10.32 | 40.75 |
| Total | 2246.00 | 7.77 | 6.27 | 1.00 | 40.75 |

More About Creating Variables

- 3. 将连续变量转为类别变量
 - ▶ 指定区间-recode-

```
sum age
```

*左开右闭

```
recode age (min/39 = 1) (39/42 = 2) (42/max = 3), ///  
gen(g_age)
```

*自行查看结果

```
list age g_age in 1/50, sepby(g_age)
```

More about creating Variables

- 4. 交叉类别变量的生成-xgroup-

```
. tab race
```

| race | Freq. | Percent | Cum. |
|-------|-------|---------|--------|
| white | 1,637 | 72.89 | 72.89 |
| black | 583 | 25.96 | 98.84 |
| other | 26 | 1.16 | 100.00 |
| Total | 2,246 | 100.00 | |


```
. tab married
```

| married | Freq. | Percent | Cum. |
|---------|-------|---------|--------|
| single | 804 | 35.80 | 35.80 |
| married | 1,442 | 64.20 | 100.00 |
| Total | 2,246 | 100.00 | |

More about creating Variables

4. 交叉类别变量的生成-xgroup-

```
. ssc install xgroup
checking xgroup consistency and verifying not already installed...
all files already exist and are up to date.
. xgroup race married, gen(race_marr2) label lname(race_marr_lab)
. //生成一个新的类别变量，取值为1-6，是race和married的组合
. labelbook race_marr_lab
```

```
value label race_marr_lab
```

| values | labels |
|---------------|-----------------------------|
| range: [1,6] | string length: [12,13] |
| N: 6 | 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

```
1 white single
2 white married
3 black single
4 black married
5 other single
6 other married
```

```
variables: race_marr2
```

More About Creating Variables

- 5.-egen-命令

- ▶ 与-gen-的区别

sum() 函数

```
. sysuse nlsw88.dta, clear  
(NLSW, 1988 extract)  
. sort wage  
. gen sum_wage1 = sum(wage)           // 累加  
. egen sum_wage2 = sum(wage)         // 总体加总
```

More About Creating Variables

- 5.-egen-命令

- ▶ 与-gen-的区别

sum() 函数

```
. list wage sum_wage1 sum_wage2 in 1/10
```

| | wage | sum_wa_1 | sum_wa_2 |
|-----|----------|----------|----------|
| 1. | 1.004952 | 1.004952 | 17444.57 |
| 2. | 1.032247 | 2.037199 | 17444.57 |
| 3. | 1.151368 | 3.188567 | 17444.57 |
| 4. | 1.344605 | 4.533172 | 17444.57 |
| 5. | 1.392914 | 5.926086 | 17444.57 |
| 6. | 1.501798 | 7.427885 | 17444.57 |
| 7. | 1.545893 | 8.973778 | 17444.57 |
| 8. | 1.561996 | 10.53577 | 17444.57 |
| 9. | 1.571983 | 12.10776 | 17444.57 |
| 10. | 1.59261 | 13.70037 | 17444.57 |

More About Creating Variables

- 5.-egen-命令

- ▶ 与-gen-的区别

对缺漏值的处理

```
. clear
. input v1 v2
      v1 v2
      1  5
      2  .
      .  3
      2  4
      4  .
      .  6
      end
. gen mean = (v1+v2)/2
(4 missing values generated)
. egen mean_egen = rmean(v1 v2)
```

More About Creating Variables

- 5.-egen-命令

- ▶ 与-gen-的区别

对缺漏值的处理

```
. list
```

| | v1 | v2 | mean | mean_e_n |
|----|----|----|------|----------|
| 1. | 1 | 5 | 3 | 3 |
| 2. | 2 | . | . | 2 |
| 3. | . | 3 | . | 3 |
| 4. | 2 | 4 | 3 | 3 |
| 5. | 4 | . | . | 4 |
| 6. | . | 6 | . | 6 |

More About Creating Variables

- 5.-egen-命令

- ▶ 丰富的特有函数功能

```
help egen          //extended generate

sysuse nlsw88.dta, clear
egen x1 = seq(), from(-1)    //等差数列
egen r2 = fill(2 4)         //间隔2的递增数列
egen avg_w_r = mean(wage), by(race) //组内均值
egen med_w = median(wage), by(race) //组内中位数
egen std=sd(wage)
egen max=max(wage)
egen min=min(wage)
```


More About Creating Variables

- 5.-egen-命令

- ▶ 丰富的特有函数功能

```
egen sum=sum(wage)
//得到wage的列总和
egen per=pc(wage),prop
//wage中每个观测值的值占列总和的比例
egen per_1=pc(wage)
//wage中每个观测值的值占列总和的百分数
egen pct=pctile(wage),p(25)
//生成wage第25百分位上的值

... ..
```

Section 5

Grouping Statistics

Grouping Statistics

- 1. 单维分组-bysort-

```
. sysuse nlsw88.dta,clear  
(NLSW, 1988 extract)  
. bysort race: sum wage
```

-> race = white

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|----------|-----------|----------|----------|
| wage | 1,637 | 8.082999 | 5.955069 | 1.004952 | 40.19808 |

-> race = black

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|----------|----------|
| wage | 583 | 6.844558 | 5.076187 | 1.151368 | 40.74659 |

-> race = other

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|---------|----------|
| wage | 26 | 8.550781 | 5.20943 | 1.80602 | 25.80515 |

Grouping Statistics

- 1. 单维分组-tabstat-

```
. tabstat wage hours ttl_exp, by(race)   ///  
> stat(n mean sd med min max)         ///  
> format(%6.2f) columns(statistics)
```

Summary for variables: wage hours ttl_exp
by categories of: race (race)

| race | N | mean | sd | p50 | min | max |
|-------|---------|-------|-------|-------|-------|-------|
| white | 1637.00 | 8.08 | 5.96 | 6.55 | 1.00 | 40.20 |
| | 1635.00 | 36.90 | 11.29 | 40.00 | 1.00 | 80.00 |
| | 1637.00 | 12.47 | 4.62 | 12.91 | 0.12 | 28.88 |
| black | 583.00 | 6.84 | 5.08 | 5.43 | 1.15 | 40.75 |
| | 581.00 | 38.12 | 7.79 | 40.00 | 3.00 | 70.00 |
| | 583.00 | 12.72 | 4.54 | 13.60 | 0.40 | 26.54 |
| other | 26.00 | 8.55 | 5.21 | 7.56 | 1.81 | 25.81 |
| | 26.00 | 36.81 | 11.80 | 40.00 | 10.00 | 60.00 |
| | 26.00 | 12.60 | 5.61 | 13.70 | 2.48 | 21.22 |
| Total | 2246.00 | 7.77 | 5.76 | 6.27 | 1.00 | 40.75 |
| | 2242.00 | 37.22 | 10.51 | 40.00 | 1.00 | 80.00 |
| | 2246.00 | 12.53 | 4.61 | 13.13 | 0.12 | 28.88 |

- 1. 单维分组-tabulate-

```
. tabulate race, sum(wage)
```

| race | Summary of hourly wage | | Freq. |
|-------|------------------------|-----------|-------|
| | Mean | Std. Dev. | |
| white | 8.0829994 | 5.9550691 | 1,637 |
| black | 6.8445578 | 5.0761866 | 583 |
| other | 8.5507813 | 5.2094301 | 26 |
| Total | 7.766949 | 5.7555229 | 2,246 |

- 2. 二维、三维分组

```
bysort race married: sum wage
```

```
bysort race married: tabstat wage,by(union)    ///  
                    s(n mean sd p50 min max)
```

```
bysort race married: tab union, sum(wage)
```

Grouping Statistics

3. 多维分组

```
. table race married union, by(collgrad) c(mean wage freq) format(%4.2f)
```

| college graduate and race | union worker and married | | | |
|------------------------------|--------------------------|----------------|----------------|----------------|
| | nonunion | | union | |
| | single | married | single | married |
| not college grad | | | | |
| white | 7.10 215.00 | 6.62 577.00 | 8.23 63.00 | 7.56 131.00 |
| black | 5.34 143.00 | 5.32 155.00 | 7.95 66.00 | 7.90 48.00 |
| other | 7.25 3.00 | 6.74 8.00 | 5.29 2.00 | 8.49 3.00 |
| college grad | | | | |
| white | 10.74 80.00 | 9.85 179.00 | 11.70 34.00 | 9.89 74.00 |
| black | 10.10 32.00 | 8.81 20.00 | 10.78 16.00 | 10.68 21.00 |
| other | 15.15 2.00 | 15.16 3.00 | | 7.92 3.00 |

- 4. 转换数据为分组统计量-collapse-

```
help collapse
```

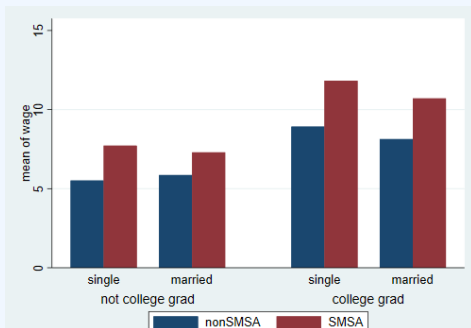
```
sysuse nlsw88.dta,clear  
collapse (mean) wage hours    ///  
(count) n_w=wage n_h=hours,  ///  
by(industry)
```


Grouping Statistics

5. Some Statistical Graphs

▶ 柱状图

```
. sysuse nlsw88.dta, clear  
(NLSW, 1988 extract)  
. graph bar (mean) wage, over(smsa) over(married) over(collgrad)  
  
. graph export bar1.png,width(500) replace  
(file bar1.png written in PNG format)
```



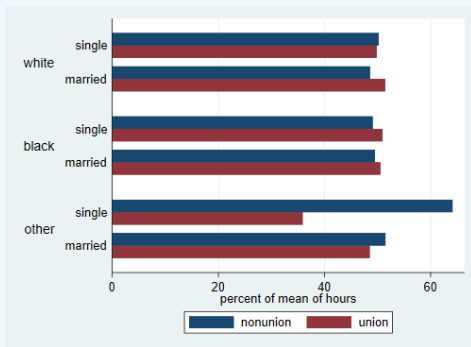
Grouping Statistics

5. Some Statistical Graphs

▶ 柱状图

```
. graph hbar (mean) hours, over(union) over(married) ///  
>          over(race) percent asyvars
```

```
. graph export bar2.png,width(500) replace  
(file bar2.png written in PNG format)
```



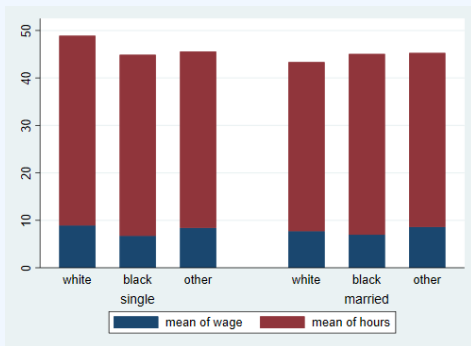
Grouping Statistics

5. Some Statistical Graphs

▶ 柱状图

```
. graph bar wage hours, over(race) over(married) stack
```

```
. graph export bar3.png,width(500) replace  
(file bar3.png written in PNG format)
```



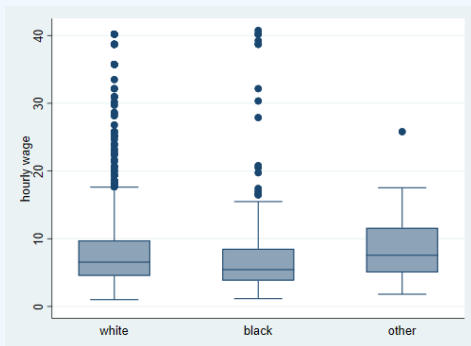
Grouping Statistics

5. Some Statistical Graphs

▶ 箱形图

```
. graph box wage, over(race)
```

```
. graph export box1.png,width(500) replace  
(file box1.png written in PNG format)
```



Section 6

Duplicate Observations

Duplicate Observations

- 1. 检查重复的样本

```
. sysuse nlsw88.dta, clear  
(NLSW, 1988 extract)
```

- isid-命令

```
cap isid race age  
isid idcode
```

Duplicate Observations

- 1. 检查重复的样本

- ▶ -duplicates list-命令

```
. duplicates list race married in 1/20  
Duplicates in terms of race married
```

| group: | obs: | race | married |
|--------|------|-------|---------|
| 1 | 7 | white | single |
| 1 | 14 | white | single |
| 1 | 20 | white | single |
| 2 | 4 | white | married |
| 2 | 5 | white | married |
| 2 | 6 | white | married |
| 2 | 8 | white | married |
| 2 | 9 | white | married |
| 2 | 10 | white | married |
| 2 | 11 | white | married |

Duplicate Observations

- 1. 检查重复的样本

- ▶ -duplicates list-命令

| | | | |
|---|----|-------|---------|
| 2 | 12 | white | married |
| 2 | 13 | white | married |
| 2 | 15 | white | married |
| 2 | 16 | white | married |
| 2 | 17 | white | married |
| 2 | 18 | white | married |
| 2 | 19 | white | married |
| 3 | 1 | black | single |
| 3 | 2 | black | single |
| 3 | 3 | black | single |

Duplicate Observations

- 1. 检查重复的样本

▶ -duplicates report-命令

```
. duplicates report race married occupation  
Duplicates in terms of race married occupation
```

| copies | observations | surplus |
|--------|--------------|---------|
| 1 | 6 | 0 |
| 2 | 12 | 6 |
| 3 | 12 | 8 |
| 4 | 12 | 9 |
| 5 | 15 | 12 |
| 6 | 18 | 15 |
| 7 | 7 | 6 |
| 8 | 8 | 7 |
| 9 | 9 | 8 |
| 10 | 10 | 9 |
| 13 | 13 | 12 |
| 14 | 14 | 13 |
| 15 | 30 | 28 |
| 16 | 16 | 15 |
| 20 | 20 | 19 |
| 25 | 25 | 24 |
| 26 | 26 | 25 |

Duplicate Observations

- 1. 检查重复的样本

- ▶ -duplicates report-命令

| | | |
|-----|-----|-----|
| 27 | 54 | 52 |
| 34 | 34 | 33 |
| 41 | 41 | 40 |
| 44 | 44 | 43 |
| 51 | 51 | 50 |
| 59 | 59 | 58 |
| 60 | 60 | 59 |
| 63 | 63 | 62 |
| 67 | 67 | 66 |
| 75 | 75 | 74 |
| 84 | 168 | 166 |
| 86 | 86 | 85 |
| 91 | 91 | 90 |
| 118 | 118 | 117 |
| 120 | 120 | 119 |
| 139 | 139 | 138 |
| 140 | 140 | 139 |
| 174 | 174 | 173 |
| 409 | 409 | 408 |

Duplicate Observations

- 2. 标记重复的样本

- ▶ -duplicates tag-命令

```
. duplicates tag race married occupation, gen(rm_dtag) //重复值的个数
Duplicates in terms of race married occupation
. list rm* in 1/20
```

| | rm_dtag |
|-----|---------|
| 1. | 66 |
| 2. | 5 |
| 3. | 85 |
| 4. | 119 |
| 5. | 83 |
| 6. | 408 |
| 7. | 90 |
| 8. | 139 |
| 9. | 408 |
| 10. | 173 |
| 11. | 173 |
| 12. | 173 |

Duplicate Observations

- 2. 标记重复的样本

- ▶ -duplicates drop-命令

```
. duplicates drop race married occupation, force  
Duplicates in terms of race married occupation  
(2,188 observations deleted)
```

Section 7

Missing Values

- 1. 缺漏值的简介

```
help missing //缺漏值的简介
```

- ▶ “.” 大于任何自然数
- ▶ 有些命令, 如 `sum,regress,generate` 等, 会自动忽略缺漏值
- ▶ 有些命令, 如 `count,keep` 等会将 “.” 视为一个无穷大的数值

- 1. 缺漏值的简介

```
sysuse auto,clear
sort rep78
list rep78
sum rep78 if rep78>4 //obs=11
count if rep78>4 //obs=16
keep if rep78>4
list rep78
```

Missing Values

2. 查找缺漏值

```
. sysuse nlsw88.dta,clear  
(NLSW, 1988 extract)  
. sum
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|--------------|-------|----------|-----------|----------|----------|
| idcode | 2,246 | 2612.654 | 1480.864 | 1 | 5159 |
| age | 2,246 | 39.15316 | 3.060002 | 34 | 46 |
| race | 2,246 | 1.282725 | .4754413 | 1 | 3 |
| married | 2,246 | .6420303 | .4795099 | 0 | 1 |
| never_marr_d | 2,246 | .1041852 | .3055687 | 0 | 1 |
| grade | 2,244 | 13.09893 | 2.521246 | 0 | 18 |
| collgrad | 2,246 | .2368655 | .4252538 | 0 | 1 |
| south | 2,246 | .4194123 | .4935728 | 0 | 1 |
| smsa | 2,246 | .7039181 | .4566292 | 0 | 1 |
| c_city | 2,246 | .2916296 | .4546139 | 0 | 1 |
| industry | 2,232 | 8.189516 | 3.010875 | 1 | 12 |
| occupation | 2,237 | 4.642825 | 3.408897 | 1 | 13 |
| union | 1,878 | .2454739 | .4304825 | 0 | 1 |
| wage | 2,246 | 7.766949 | 5.755523 | 1.004952 | 40.74659 |
| hours | 2,242 | 37.21811 | 10.50914 | 1 | 80 |
| ttl_exp | 2,246 | 12.53498 | 4.610208 | .1153846 | 28.88461 |
| tenure | 2,231 | 5.97785 | 5.510331 | 0 | 25.91667 |

Missing Values

- 2. 查看缺漏值

```
. misstable summarize //查看所有变量
```

| Variable | Obs=. | Obs>. | Obs<. | Unique values | Min | Max |
|------------|-------|-------|-------|---------------|-----|----------|
| grade | 2 | | 2,244 | 16 | 0 | 18 |
| industry | 14 | | 2,232 | 12 | 1 | 12 |
| occupation | 9 | | 2,237 | 13 | 1 | 13 |
| union | 368 | | 1,878 | 2 | 0 | 1 |
| hours | 4 | | 2,242 | 62 | 1 | 80 |
| tenure | 15 | | 2,231 | 259 | 0 | 25.91667 |

Missing Values

2. 查看缺漏值

```
. misstable sum age-union //查看指定变量
```

| Variable | Obs=. | Obs>. | Obs<. | Obs<. | | |
|------------|-------|-------|-------|---------------|-----|-----|
| | | | | Unique values | Min | Max |
| grade | 2 | | 2,244 | 16 | 0 | 18 |
| industry | 14 | | 2,232 | 12 | 1 | 12 |
| occupation | 9 | | 2,237 | 13 | 1 | 13 |
| union | 368 | | 1,878 | 2 | 0 | 1 |

Missing Values

3. 删除缺漏值

```
. drop if missing(grade,indus,occup,union,hours,tenure)  
(398 observations deleted)
```

```
. sum
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|--------------|-------|----------|-----------|----------|----------|
| idcode | 1,848 | 2614.384 | 1486.31 | 1 | 5159 |
| age | 1,848 | 39.21429 | 3.041416 | 34 | 46 |
| race | 1,848 | 1.291667 | .4823869 | 1 | 3 |
| married | 1,848 | .6515152 | .4766194 | 0 | 1 |
| never_marr_d | 1,848 | .1087662 | .31143 | 0 | 1 |
| grade | 1,848 | 13.17208 | 2.550548 | 0 | 18 |
| collgrad | 1,848 | .2478355 | .4318727 | 0 | 1 |
| south | 1,848 | .4242424 | .4943612 | 0 | 1 |
| smsa | 1,848 | .7083333 | .4546527 | 0 | 1 |
| c_city | 1,848 | .2938312 | .4556388 | 0 | 1 |
| industry | 1,848 | 8.255952 | 3.042377 | 1 | 12 |
| occupation | 1,848 | 4.62013 | 3.479021 | 1 | 13 |
| union | 1,848 | .2467532 | .4312386 | 0 | 1 |
| wage | 1,848 | 7.60597 | 4.173447 | 1.344605 | 39.23074 |
| hours | 1,848 | 37.61905 | 9.957783 | 1 | 80 |
| t1l_exp | 1,848 | 12.86178 | 4.576879 | .4038461 | 28.88461 |
| tenure | 1,848 | 6.582882 | 5.631957 | 0 | 25.91667 |

Section 8

Outliers

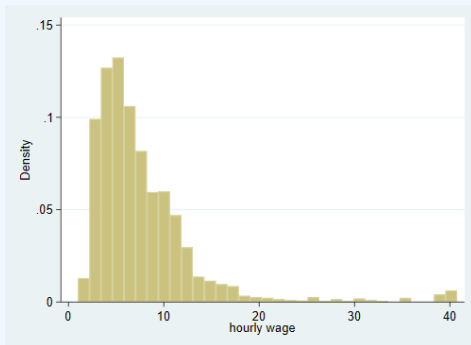
- 1. 离群值的影响

- ▶ 离群值 (outliers) 是指在一份数据中, 与其他观察值具有明显不同特征的那些观察值。
- ▶ 通常对回归结果的影响却很大。
- ▶ 下图小时工资在 20 以上的观察值比重较小, 是所谓的高薪个体。

Outliers

- 1. 离群值的影响

```
. sysuse nlsw88.dta, clear  
(NLSW, 1988 extract)  
. histogram wage, ylabel(,angle(0))  
(bin=33, start=1.0049518, width=1.2042921)  
. graph export hi_wage.png,width(500) replace  
(file hi_wage.png written in PNG format)
```



Outliers

2. 查找离群值-

▶ -adjacent-命令

```
. ssc install adjacent //安装外部命令
checking adjacent consistency and verifying not already installed...
all files already exist and are up to date.
```

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

| price | lower adjacent | upper adjacent |
|-------|----------------|----------------|
| . | 3291 | 8814 |

*注:

*四分位间距(interquartile range): $iqr = p75 - p25$

*上界(upper adjacent) = $p75 + 1.5 * iqr$

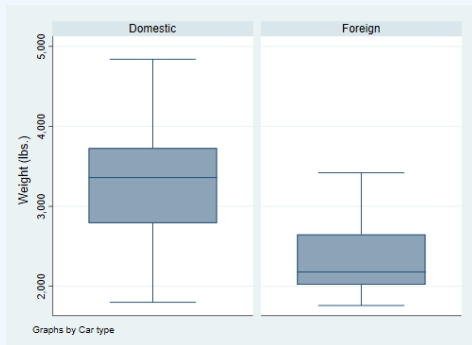
*下界(lower adjacent) = $p25 - 1.5 * iqr$

Outliers

- 2. 查找离群值-

- ▶ 箱型图

```
. graph box weight, by(foreign)  
. graph export bo_wei.png,width(500) replace  
(file bo_wei.png written in PNG format)
```



3. 删除离群值

```
. sysuse auto, clear  
(1978 Automobile Data)  
. adjacent price, by(foreign)
```

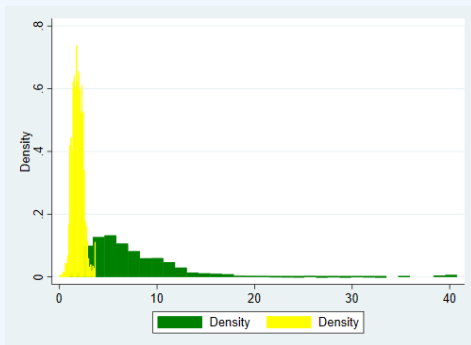
| foreign | lower adjacent | upper adjacent |
|----------|----------------|----------------|
| Domestic | 3291 | 8814 |
| Foreign | 3748 | 9735 |

```
. drop if (price>8814&foreign==0) | (price>9735&foreign==1)  
(10 observations deleted)
```

Outliers

4. 取对数

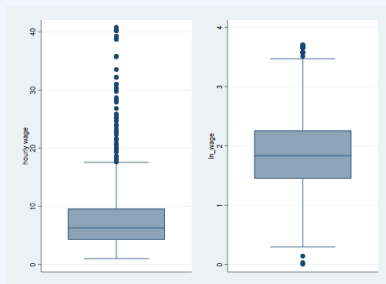
```
. sysuse nlsw88, clear  
(NLSW, 1988 extract)  
. gen ln_wage = ln(wage)  
  
. twoway (histogram wage,color(green))(histogram ln_wage,color(yellow))  
. graph export hi_alwage.png,width(500) replace  
(file hi_alwage.png written in PNG format)
```



Outliers

4. 取对数

```
. graph box wage  
. graph save bo_w.gph, replace  
(file bo_w.gph saved)  
  
. graph box ln_wage  
. graph save bo_lw.gph, replace  
(file bo_lw.gph saved)  
  
. cap graph combine bo_w.gph bo_lw.gph, saving(box_wage)  
. graph export box_wage.png,width(500) replace  
(file box_wage.png written in PNG format)
```



● 5. 缩尾

- ▶ 将超出变量特定百分位范围的数值替换为其特定百分位数值的方法。

```
. sysuse nlsw88.dta, clear
(NLSW, 1988 extract)

. ssc install winsor
checking winsor consistency and verifying not already installed...
all files already exist and are up to date.

. winsor wage, gen(wage_w) p(0.025)

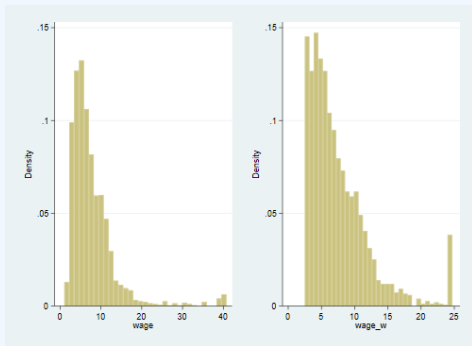
. histogram wage, ylabel(,angle(0)) xtitle("wage") name(fig1, replace)
(bin=33, start=1.0049518, width=1.2042921)

. histogram wage_w, ylabel(,angle(0)) xtitle("wage_w") name(fig2, replace)
(bin=33, start=2.5083611, width=.67131721)
```

Outliers

5. 缩尾

```
. cap graph combine fig1 fig2, saving(1_2)  
. graph export 1_2.png,width(500) replace  
(file 1_2.png written in PNG format)
```

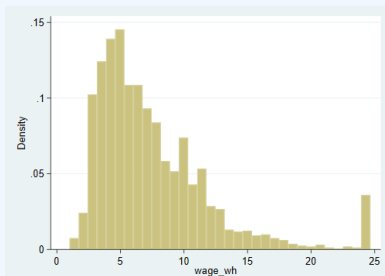


Outliers

5. 缩尾

- ▶ 左侧似乎并不存在离群值，右侧缩尾。

```
. sysuse nlsw88.dta, clear
(NLSW, 1988 extract)
. winsor wage, gen(wage_wh) p(0.025) highlyonly
//highlyonly或lowonly选项来进行单侧缩尾处理
. histogram wage_wh, ylabel(, angle(0)) xtitle("wage_wh") saving(fig3, replace)
(bin=33, start=1.0049518, width=.71687507)
(file fig3.gph saved)
. graph export 3.png,width(500) replace
(file 3.png written in PNG format)
```



6. 截尾

- ▶ 将超出变量特定百分位范围的数值予以删除。

```
. sysuse nlsw88.dta, clear
(NLSW, 1988 extract)

. ssc install winsor2
checking winsor2 consistency and verifying not already installed...
all files already exist and are up to date.

. winsor2 wage, cut(2.5 97.5) trim    //trim指定进行截尾处理（否则默认进行缩尾处理）

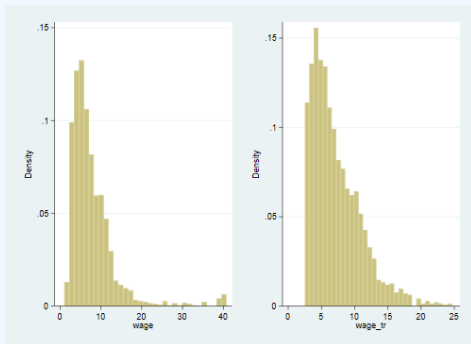
. histogram wage, ylabel(, angle(0)) xtitle("wage") name(fig4, replace)
(bin=33, start=1.0049518, width=1.2042921)

. histogram wage_tr, ylabel(, angle(0)) xtitle("wage_tr") name(fig5, replace)
(bin=33, start=2.5201283, width=.67096063)
```

Outliers

- 6. 截尾

```
. cap graph combine fig4 fig5, saving(4_5)  
. graph export 4_5.png,width(500) replace  
(file 4_5.png written in PNG format)
```



Outliers

- 6. 截尾

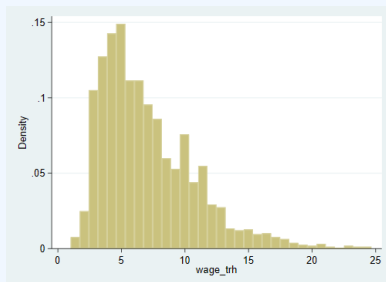
- ▶ 右侧截尾

```
. sysuse nlsw88.dta, clear
(NLSW, 1988 extract)

. winsor2 wage, cut(0 97.5) trim suffix(_trh)

. histogram wage_trh, ylabel(, angle(0)) xtitle("wage_trh") saving(fig6.png, replace)
(bin=33, start=1.0049518, width=.71687507)
(file fig6.png saved)

. graph export 6.png,width(500) replace
(file 6.png written in PNG format)
```



Section 9

Database Append and Merge

Database Append and Merge

1. 数据纵向合并-append-

```
. clear  
. use even,clear    //调用even使用数据  
                    (6th through 8th even numbers)  
. list              //看数据
```

| | number | even |
|----|--------|------|
| 1. | 6 | 12 |
| 2. | 7 | 14 |
| 3. | 8 | 16 |

```
. use odd,clear    //主数据  
                    (First five odd numbers)  
. list
```

| | number | odd |
|----|--------|-----|
| 1. | 1 | 1 |
| 2. | 2 | 3 |
| 3. | 3 | 5 |
| 4. | 4 | 7 |
| 5. | 5 | 9 |

Database Append and Merge

1. 数据纵向合并-append-

```
. append using even, gen(append_id) //将even加在odd之后  
. list
```

| | number | odd | append_d | even |
|----|--------|-----|----------|------|
| 1. | 1 | 1 | 0 | . |
| 2. | 2 | 3 | 0 | . |
| 3. | 3 | 5 | 0 | . |
| 4. | 4 | 7 | 0 | . |
| 5. | 5 | 9 | 0 | . |
| 6. | 6 | . | 1 | 12 |
| 7. | 7 | . | 1 | 14 |
| 8. | 8 | . | 1 | 16 |

```
. save all,replace  
(note: file all.dta not found)  
file all.dta saved
```

//对修改后数据进行保存

- 1. 数据纵向合并-append-
 - ▶ 两个数据库中的变量名称要相同
 - ▶ 两个数据库中的同名变量要具有相同的存储类型

Database Append and Merge

2. 数据横向合并-merge-

▶ 1:1 合并 (合并关键变量取值完全相同的数据)

```
. use autotech, clear //autotech主数据
(1978 Automobile Data)
. desc
Contains data from autotech.dta
  obs:          74          1978 Automobile Data
  vars:          4          10 Jul 2010 00:06
```

| variable name | storage type | display format | value label | variable label |
|---------------|-----------------|-------------------|----------------|----------------|
| make | str18 | %18s | | Make and Model |
| mpg | int | %8.0g | | Mileage (mpg) |
| weight | int | %8.0g | | Weight (lbs.) |
| length | int | %8.0g | | Length (in.) |

Sorted by: make

Database Append and Merge

- 2. 数据横向合并-merge-

- ▶ 1:1 合并 (合并关键变量取值完全相同的数据)

```
. use autocost, clear //autocost使用数据
(1978 Automobile Data)
. desc
Contains data from autocost.dta
  obs:          74          1978 Automobile Data
  vars:          3          10 Jul 2010 00:07
```

| variable name | storage type | display format | value label | variable label |
|---------------|--------------|----------------|-------------|--------------------|
| make | str18 | %18s | | Make and Model |
| price | int | %8.0g | | Price |
| rep78 | int | %8.0g | | Repair Record 1978 |

```
Sorted by: make
```

Database Append and Merge

2. 数据横向合并-merge-

▶ 1:1 合并 (合并关键变量取值完全相同的数据)

```
. use autotech, clear
(1978 Automobile Data)
. merge 1:1 make using autcost    //make是牵引变量
```

| Result | # of obs. |
|-------------|----------------|
| not matched | 0 |
| matched | 74 (_merge==3) |

```
. tabulate _merge
```

| _merge | Freq. | Percent | Cum. |
|-------------|-------|---------|--------|
| matched (3) | 74 | 100.00 | 100.00 |
| Total | 74 | 100.00 | |

Database Append and Merge

2. 数据横向合并-merge-

- ▶ m:1(或 1:m) 合并 (合并关键变量取值重复的数据)

```
. use cgss13m1, clear //多的那个(调查对象家庭成员的基本情况)
. desc
Contains data from cgss13m1.dta
  obs:          9
  vars:         4                13 Jan 2016 10:33
```

| variable name | storage type | display format | value label | variable label |
|---------------|--------------|----------------|-------------|----------------|
| id | float | %9.0g | | 问卷编号 |
| numid | byte | %9.0g | | 家庭成员序号 |
| a12 | byte | %21.0g | lab1 | 家庭成员的性别 |
| a14 | int | %14.0g | a1402lab | 家庭成员的年龄 |

Sorted by: id

Database Append and Merge

2. 数据横向合并-merge-

- ▶ m:1(或 1:m) 合并 (合并关键变量取值重复的数据)

```
. use cgss13m2, clear    //(调查对象家庭成员的全家收入)
```

```
. desc
```

```
Contains data from cgss13m2.dta
```

```
obs:
```

```
5
```

```
vars:
```

```
2
```

```
13 Jan 2016 10:33
```

| variable name | storage type | display format | value label | variable label |
|---------------|--------------|----------------|-------------|--------------------|
| id | float | %9.0g | | 问卷编号 |
| a62 | long | %42.0g | lab32 | 您家2011年全年家庭总收入是多少? |

```
Sorted by: id
```

Database Append and Merge

2. 数据横向合并-merge-

▶ m:1(或 1:m) 合并 (合并关键变量取值重复的数据)

```
. use cgss13m1, clear
. merge m:1 id using cgss13m2
```

| Result | # of obs. |
|-------------|---------------|
| not matched | 0 |
| matched | 9 (_merge==3) |

```
. describe
```

Contains data from cgss13m1.dta

| | | |
|-------|---|-------------------|
| obs: | 9 | |
| vars: | 6 | 13 Jan 2016 10:33 |

| variable name | storage type | display format | value label | variable label |
|---------------|--------------|----------------|-------------|--------------------|
| id | float | %9.0g | | 问卷编号 |
| numid | byte | %9.0g | | 家庭成员序号 |
| a12 | byte | %21.0g | lab1 | 家庭成员的性别 |
| a14 | int | %14.0g | a1402lab | 家庭成员的年龄 |
| a62 | long | %42.0g | lab32 | 您家2011年全年家庭总收入是多少? |
| _merge | byte | %23.0g | _merge | |

Database Append and Merge

2. 数据横向合并-merge-

- ▶ m:1(或 1:m) 合并 (合并关键变量取值重复的数据)

```
. list, sepby(id) //合并两个数据
```

| | id | numid | a12 | a14 | a62 | _merge |
|----|-------|-------|-----|-----|--------|-------------|
| 1. | 3179 | 1 | 男 | 51 | 30000 | matched (3) |
| 2. | 3179 | 2 | 男 | 26 | 30000 | matched (3) |
| 3. | 3179 | 3 | 男 | 24 | 30000 | matched (3) |
| 4. | 3932 | 1 | 男 | 40 | 8000 | matched (3) |
| 5. | 3932 | 2 | 女 | 15 | 8000 | matched (3) |
| 6. | 3932 | 3 | 男 | 1 | 8000 | matched (3) |
| 7. | 5592 | 1 | 男 | 48 | 9000 | matched (3) |
| 8. | 6242 | 1 | 男 | 70 | 30000 | matched (3) |
| 9. | 10902 | 1 | 男 | 59 | 130000 | matched (3) |

Database Append and Merge

● 3. 配对合并-joinby-

- ▶ 适用于 m: m 的交叉匹配合并
- ▶ 比如把父母的数据和孩子的数据进行配对合并

```
. use child, clear
(Data on Children)
. desc
Contains data from child.dta
  obs:          5          Data on Children
  vars:         4          30 Apr 2017 20:07
```

| variable name | storage type | display format | value label | variable label |
|---------------|--------------|----------------|-------------|------------------|
| family_id | int | %8.0g | | Family ID number |
| child_id | byte | %8.0g | | Child ID number |
| x1 | byte | %8.0g | | |
| x2 | int | %8.0g | | |

```
Sorted by: family_id
```

Database Append and Merge

- 3. 配对合并-joinby-

```
. list
```

| | family_d | child_id | x1 | x2 |
|----|----------|----------|----|-----|
| 1. | 1025 | 3 | 11 | 320 |
| 2. | 1025 | 1 | 12 | 300 |
| 3. | 1025 | 4 | 10 | 275 |
| 4. | 1026 | 2 | 13 | 280 |
| 5. | 1027 | 5 | 15 | 210 |

Database Append and Merge

- 3. 配对合并-joinby-

```
. use parent, clear
(Data on Parents)
. desc
Contains data from parent.dta
  obs:          6          Data on Parents
  vars:         4          30 Apr 2017 20:07
```

| variable name | storage type | display format | value label | variable label |
|---------------|--------------|----------------|-------------|------------------|
| family_id | int | %8.0g | | Family ID number |
| parent_id | float | %9.0g | | Parent ID number |
| x1 | float | %9.0g | | |
| x3 | float | %9.0g | | |

Sorted by:

Database Append and Merge

- 3. 配对合并-joinby-

```
. list, sep(0)
```

| | family_d | parent_d | x1 | x3 |
|----|----------|----------|----|-----|
| 1. | 1030 | 10 | 39 | 600 |
| 2. | 1025 | 11 | 20 | 643 |
| 3. | 1025 | 12 | 27 | 721 |
| 4. | 1026 | 13 | 30 | 760 |
| 5. | 1026 | 14 | 26 | 668 |
| 6. | 1030 | 15 | 32 | 684 |

Database Append and Merge

• 3. 配对合并-joinby-

```
. sort family_id  
. joinby family_id using child  
. list, sepby(fam)
```

| | family_d | parent_d | x1 | x3 | child_id | x2 |
|----|----------|----------|----|-----|----------|-----|
| 1. | 1025 | 12 | 27 | 721 | 1 | 300 |
| 2. | 1025 | 12 | 27 | 721 | 4 | 275 |
| 3. | 1025 | 12 | 27 | 721 | 3 | 320 |
| 4. | 1025 | 11 | 20 | 643 | 4 | 275 |
| 5. | 1025 | 11 | 20 | 643 | 1 | 300 |
| 6. | 1025 | 11 | 20 | 643 | 3 | 320 |
| 7. | 1026 | 14 | 26 | 668 | 2 | 280 |
| 8. | 1026 | 13 | 30 | 760 | 2 | 280 |