Lab5: Tables and OLS

Introduction to Econometrics,Fall 2020

Yi Wang

Nanjing University

21/10/2020

Section 1

Matrix

Subsection 1

Definition of Matrix

Matrix

- 基本定义方式
  - Stata中的数据可以视为矩阵。
  - 规则:【,】分列,【\】分行。

```
. matrix drop _all
. matrix a = (1,2,3 \ 4,5,6)
. mat list a
a[2,3]
    c1  c2  c3
r1   1   2   3
r2   4   5   6
```

Subsection 2

Management of Matrix

Matrix

- 矩阵的管理
  - 1.矩阵的名称
  - 可以内存中的变量同名。
  - 不可以和单值重名(一旦重名，会自动覆盖)。
  - 注意：
    数学运算中，同时为**变量名称**和**矩阵名称**的名称，Stata会将其解释为变量名称。

Matrix

- 矩阵的管理
  - 1.矩阵的名称

```
. mat b = (2,3)


. mat dir
               b[1,2]
               a[2,3]

. matrix rename a c

. mat dir
               c[2,3]
               b[1,2]
```

Matrix

- 矩阵的管理
  - 2.矩阵的列式

```
. mat list b
b[1,2]
    c1  c2
r1   2   3
. mat list b, format(%3.1f)
b[1,2]
     c1   c2
r1  2.0  3.0


. mat d = J(3,3,2)
. mat list d, nohalf nonames title("一个3*3的对称矩阵")
symmetric d[3,3]:  一个3*3的对称矩阵
  2  2  2
  2  2  2
  2  2  2
```

- 矩阵的管理
  - 2.矩阵的列式
  - 更加灵活的设定方式*-matlist-*（了解即可）

```
. matrix x = (1, 2 \ 3, 4 \ 5, 6)
. matlist x, border(rows) rowtitle(rows) left(4)

    ---------------------------------
           rows |        c1        c2
    ------------+--------------------
             r1 |         1         2
             r2 |         3         4
             r3 |         5         6
    ---------------------------------

. matlist 2*x, border(all) format(%6.1f) names(rows) twidth(8) title(x)
x
+----------------------------+
|       r1 |    2.0      4.0 |
|       r2 |    6.0      8.0 |
|       r3 |   10.0     12.0 |
+----------------------------+
```

Matrix

- 矩阵的管理
  - 2.矩阵的列式
  - 更加灵活的设定方式-*matlist-*（了解即可）

```
. matrix Htest = ( 12.30, 2, .00044642 \                ///
>                  2.17, 1, .35332874  \                ///
>                  8.81, 3, .04022625  \                ///
>                  20.05, 6, .00106763 )


. matrix rownames Htest = trunk length weight overall     //定义行名
. matrix colnames Htest = chi2 df p                       //定义列名
```

Matrix

- 矩阵的管理
  - 2.矩阵的列式
  - 更加灵活的设定方式-*matlist-*（了解即可）

```
. matlist Htest, title("检验结果") rowtitle("变量名称")        ///
> cspec(o4& %12s | %8.0g & %5.0f & %8.4f o2&) rspec(&-&&--)
检验结果
       变量名称 |     chi2      df          p
    ------------+---------------------------
          trunk |     12.3       2     0.0004
         length |     2.17       1     0.3533
         weight |     8.81       3     0.0402
    ------------+---------------------------
        overall |    20.05       6     0.0011
    ------------+---------------------------

 *cspec()特指列的格式和列的分隔符:
 *Sep（分隔符）:
 * |  指定绘制一条垂直线。
 * &  指定不画线。
 * o# 指定分隔符之前和之后的空格数，默认为一个空格，但第一列之前和最后一列之后不包含
空格。
```

Matrix

- 矩阵的管理
  - 2.矩阵的列式
  - 更加灵活的设定方式-*matlist-*（了解即可）

```
. matlist Htest, title("检验结果(New)") rowtitle("变量名称") ///
> cspec( o4&o2 %10s | b t %8.0g & %4.0f & i c %7.4f o2& )    ///
> rspec( & - & & - & )
检验结果(New)
        变量名称 |      chi2     df          p
      ---------+--------------------------------
          trunk |      12.3      2     0.0004
         length |      2.17      1     0.3533
         weight |      8.81      3     0.0402
      ---------+--------------------------------
        overall |     20.05      6     0.0011


 * b   加粗(bold)
 * t   绿色(text color)
 * i   斜体(italic)
 * c   白色(command color)
```

Matrix

- 矩阵的管理
  - 3.矩阵的行数和列数

```
. matrix x = (1, 2 \ 3, 4 \ 5, 6)
. display colsof(d)
3
. display rowsof(c)
2
```

  - 4.矩阵的行名和列名

```
. mat rownames x = "一" "二" "三"
. mat colnames x = x1 x2
. mat list x
x[3,2]
     x1  x2
 一   1   2
 二   3   4
 三   5   6
```

- 矩阵的管理
  - 5.矩阵的查找和删除

```
. mat dir
            x[3,2]
        Htest[4,3]
            d[3,3]
            c[2,3]
            b[1,2]


. mat drop b c d x
. mat drop _all
```

Subsection 3

Operation of Matrix

Matrix

- 矩阵的操作
  - 1.选取1个元素

```
. matrix e = (1,2,3,4,5 \ 2,3,4,5,6 \ 3,4,5,6,7 \ 4,5,6,7,8 \ 5,6,7,8,9)
. mat list e, nohalf
symmetric e[5,5]
    c1  c2  c3  c4  c5
r1   1   2   3   4   5
r2   2   3   4   5   6
r3   3   4   5   6   7
r4   4   5   6   7   8
r5   5   6   7   8   9


. mat e1 = e[2,3]
. mat list e1
symmetric e1[1,1]
    c1
r1   4
```

Matrix

- 矩阵的操作
  - 2.选取子矩阵

```
. mat e2 = e[....,3...]

. mat list e2

e2[5,3]
    c3  c4  c5
r1   3   4   5
r2   4   5   6
r3   5   6   7
r4   6   7   8
r5   7   8   9
```

Matrix

- 矩阵的操作
  - 2.选取子矩阵

```
. mat e3 = e[4...,3...]
. mat list e3
e3[2,3]
    c3  c4  c5
r4   6   7   8
r5   7   8   9
```

Matrix

- 矩阵的操作
  - 2.选取子矩阵

```
. mat e4 = e[1..3,3...]
. mat list e4
symmetric e4[3,3]
    c3  c4  c5
r1   3
r2   4   5
r3   5   6   7
```

Matrix

- 矩阵的操作
  - 3.矩阵元素的修改

```
. mat e[4,5] = (1)
. mat e[2,2] = (-29, 78)
. mat list e
e[5,5]
     c1   c2   c3   c4   c5
r1    1    2    3    4    5
r2    2  -29   78    5    6
r3    3    4    5    6    7
r4    4    5    6    7    1
r5    5    6    7    8    9
```

Matrix

- 矩阵的操作
  - 4.分块矩阵的操作

```
. mat a1 = (1, 2, 3 \ 42, 50, 63)
. mat a2 = (-3,-5,-7 \ -9 , -11, -13)
. mat list a1
a1[2,3]
    c1  c2  c3
r1   1   2   3
r2  42  50  63
. mat list a2
a2[2,3]
     c1   c2   c3
r1   -3   -5   -7
r2   -9  -11  -13
```

Matrix

- 矩阵的操作
  - 4.分块矩阵的操作

```
. mat aa = [a1, a2]   //横向合并
. mat list aa
aa[2,6]
     c1   c2   c3   c1   c2   c3
r1    1    2    3   -3   -5   -7
r2   42   50   63   -9  -11  -13
. mat aaa = [a1 \ a2] //纵向追加
. mat list aaa
aaa[4,3]
     c1   c2   c3
r1    1    2    3
r2   42   50   63
r1   -3   -5   -7
r2   -9  -11  -13
```

Subsection 4

Common Matrixes

Matrix

- 常用矩阵的定义
  - 1.单位矩阵

```
. mat I = I(4)
. mat list I
symmetric I[4,4]
    c1  c2  c3  c4
r1   1
r2   0   1
r3   0   0   1
r4   0   0   0   1
```

Matrix

- 常用矩阵的定义
  - 2.常数矩阵

```
. mat f = J(2,6,-1)

. mat list f

f[2,6]
    c1  c2  c3  c4  c5  c6
r1  -1  -1  -1  -1  -1  -1
r2  -1  -1  -1  -1  -1  -1
```

Matrix

- 常用矩阵的定义
  - 3.对角矩阵

```
. mat g = (1, 2, 3, 4, 5)

. mat list g

g[1,5]
    c1  c2  c3  c4  c5
r1   1   2   3   4   5

. mat dg = diag(g) // 取出对角元素

. mat list dg

symmetric dg[5,5]
    c1  c2  c3  c4  c5
c1   1
c2   0   2
c3   0   0   3
c4   0   0   0   4
c5   0   0   0   0   5
```

Subsection 5

Conversion Between Variables and Matrices

Matrix

- 变量与矩阵的转换
  - 1.变量—>矩阵*-mkmat-*

```
. sysuse auto,clear
(1978 Automobile Data)
. mkmat price, mat(Y)
. gen cons= 1
. mkmat cons weight length foreign, mat(X)


. mat list Y
Y[74,1]
      price
 r1    4099
 r2    4749
 r3    3799
 r4    4816
 r5    7827
 r6    5788
 r7    4453
 r8    5189
 r9   10372
r10    4082
r11   11385
r12   14500
```

Matrix

- 变量与矩阵的转换
  - 1.变量—>矩阵*-mkmat-*

```
. mat list X
X[74,4]
           cons   weight   length  foreign
 r1           1     2930      186        0
 r2           1     3350      173        0
 r3           1     2640      168        0
 r4           1     3250      196        0
 r5           1     4080      222        0
 r6           1     3670      218        0
 r7           1     2230      170        0
 r8           1     3280      200        0
 r9           1     3880      207        0
r10           1     3400      200        0
r11           1     4330      221        0
r12           1     3900      204        0
r13           1     4290      204        0
r14           1     2110      163        0
r15           1     3690      212        0
r16           1     3180      193        0
r17           1     3220      200        0
r18           1     2750      179        0
r19           1     3430      197        0
r20           1     2120      163        0
```

Matrix

- 变量与矩阵的转换
  - 1.变量—>矩阵*-mkmat-*
  - 实例：OLS系数估计

```
. mat b = inv(X´*X)*X´*Y  //inv() 逆矩阵函数
. mat list b
b[4,1]
             price
   cons  4838.0206
 weight  5.7747117
 length  -91.37083
foreign  3573.0919
```

- 变量与矩阵的转换
  - 1.变量—>矩阵-*mkmat*-
  - 实例：OLS系数估计

```
. reg price weight length foreign
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(3, 70)        =     28.39
       Model |   348565467         3   116188489   Prob > F        =    0.0000
    Residual |   286499930        70  4092856.14   R-squared       =    0.5489
-------------+----------------------------------   Adj R-squared   =    0.5295
       Total |   635065396        73  8699525.97   Root MSE        =    2023.1

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   5.774712   .9594168     6.02   0.000     3.861215    7.688208
      length |  -91.37083   32.82833    -2.78   0.007    -156.8449   -25.89679
     foreign |   3573.092    639.328     5.59   0.000     2297.992    4848.191
       _cons |   4838.021    3742.01     1.29   0.200    -2625.183    12301.22
------------------------------------------------------------------------------
```

- 变量与矩阵的转换
  - 2.矩阵—>变量*-svmat-*

```
. svmat b, names(coef)
. list coef in 1/5

     +-----------+
     |     coef1 |
     |-----------|
  1. |  4838.021 |
  2. |  5.774712 |
  3. | -91.37083 |
  4. |  3573.092 |
  5. |         . |
     +-----------+


. svmat X, names(var)  //自行定义统一的变量名
. drop weight length foreign cons
. svmat X, names(col)  //用矩阵的列名作为变量的名称
```

Subsection 6

Simple Calculation of Matrices

Matrix

- 矩阵的简单计算

```
help matrix operators
```

  - 1.加(+)

```
. matrix A = (1,2\3,4)
. matrix B = (5,7\9,2)
. matrix C = A+B
. matrix list C
C[2,2]
    c1  c2
r1   6   9
r2  12   6
```

Matrix

- 矩阵的简单计算
  - 2.减(-)

```
. matrix B = A-B

. matrix list B

B[2,2]
    c1  c2
r1  -4  -5
r2  -6   2
```

Matrix

- 矩阵的简单计算
  - 3.乘(*)

```
. matrix X = (1,1\2,5\8,0\4,5)

. matrix C = 3*X*A´*B

. matrix list C

C[4,2]
      c1    c2
r1  -162    -3
r2  -612   -24
r3  -528    24
r4  -744   -18
```

Matrix

- 矩阵的简单计算
  - 4.矩阵与单值的运算

```
. scalar y = 5
. mat D = J(4,4,1)


. mat D_y = D/y
. mat list D_y
symmetric D_y[4,4]
    c1  c2  c3  c4
r1  .2
r2  .2  .2
r3  .2  .2  .2
r4  .2  .2  .2  .2
```

Matrix

- 矩阵的简单计算
  - 5.矩阵的转置

```
. matrix F = (-1, 2 \ 3, 4 )
. matrix H = ( 4, 1 \ 2, 5 )


. mat P = (F*H)´
. mat Q = H´*F´
. mat list P
P[2,2]
    r1  r2
c1   0  20
c2   9  23
. mat list Q
Q[2,2]
    r1  r2
c1   0  20
c2   9  23
```

Matrix

- 矩阵的简单计算
  - 6.矩阵的行列式

```
. mat I = (-1, 2 \ 3, 4)

. mat list I

I[2,2]
    c1  c2
r1  -1   2
r2   3   4

. scalar detI = det(I)

. dis detI
-10
```

Matrix

- 矩阵的简单计算
  - 7.矩阵的逆

```
. mat invI = inv(I)

. mat list invI

invI[2,2]
     r1   r2
c1  -.4   .2
c2   .3   .1
```

  - 8.矩阵的迹(trace)

```
. mat trI = trace(I)

. mat list trI

symmetric trI[1,1]
    c1
r1   3
```

Matrix

- 矩阵的简单计算
  - 9.矩阵的秩(rank)

```
. mata
------------------------------------------------- mata (type end to exit) ------
: A = (1,2,3 \ 3,2,1)´
: A
       1   2
    +---------+
  1 |  1   3  |
  2 |  2   2  |
  3 |  3   1  |
    +---------+

: rank(A)
  2

: end
--------------------------------------------------------------------------------
```

  - ...
  - ...

Section 2

Return Values

Return Values

- Stata命令分为三种类型:
  - r-class 与模型估计无关的命令,e.g.sum
  - e-class 与模型估计有关的命令,e.g.reg
  - s-class 其它命令,e.g.list
  - c-class 存储系统参数

- 内存中结果的显示方法:
  - r-class:

```
return list
```

  - e-class:

```
ereturn list
```

  - s-class:

```
sreturn list
```

  - c-class:

```
creturn list
```

Return Values

- 留存值分为四种类型：
  - scalars: r(mean), r(N), e(r2), e(F)
  - matrices: e(b), e(V)
  - macros: e(cmd), e(depvar)
  - functions: e(sample)

Return Values

- r-class

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

. sum price

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906

. return list

scalars:
                  r(N) =  74
              r(sum_w) =  74
               r(mean) =  6165.256756756757
                r(Var) =  8699525.974268788
                 r(sd) =  2949.495884768919
                r(min) =  3291
                r(max) =  15906
                r(sum) =  456229

. dis "汽车的平均价格是: " in g r(mean) //或者写成`r(mean)'
汽车的平均价格是: 6165.2568
```

Return Values

- e-class

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

. regress price weight length foreign

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(3, 70)        =     28.39
       Model |   348565467         3   116188489   Prob > F        =    0.0000
    Residual |   286499930        70  4092856.14   R-squared       =    0.5489
-------------+----------------------------------   Adj R-squared   =    0.5295
       Total |   635065396        73  8699525.97   Root MSE        =    2023.1

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   5.774712   .9594168     6.02   0.000     3.861215    7.688208
      length |  -91.37083   32.82833    -2.78   0.007    -156.8449   -25.89679
     foreign |   3573.092    639.328     5.59   0.000     2297.992    4848.191
       _cons |   4838.021    3742.01     1.29   0.200    -2625.183    12301.22
------------------------------------------------------------------------------

. ereturn list

scalars:
                  e(N) =  74
               e(df_m) =  3
               e(df_r) =  70
                  e(F) =  28.38811943068357
                 e(r2) =  .5488654690386703
               e(rmse) =  2023.080852875841
```

Return Values

- e-class

```
. dis "系数的方差-协方差矩阵为: "
系数的方差-协方差矩阵为:
.     mat list e(V), format(%6.2f)
symmetric e(V)[4,4]
           weight    length   foreign      _cons
 weight      0.92
 length    -28.94   1077.70
foreign    123.05    753.80   4.1e+05
  _cons   2623.60  -1.2e+05  -6.3e+05    1.4e+07
```

Return Values

- c-class
  - 提供了大量提供系统参数的返回值

```
. dis `c(pi)´
3.1415927

. dis "`c(sysdir_plus)´"
D:\Stata16\ado\plus/

. dis "`c(current_date)´"
17 Oct 2020
```

Section 3

T-Test and Table

T-Test and Table

- Review the Theory

Hypothesis Test of of $\bar{Y}$

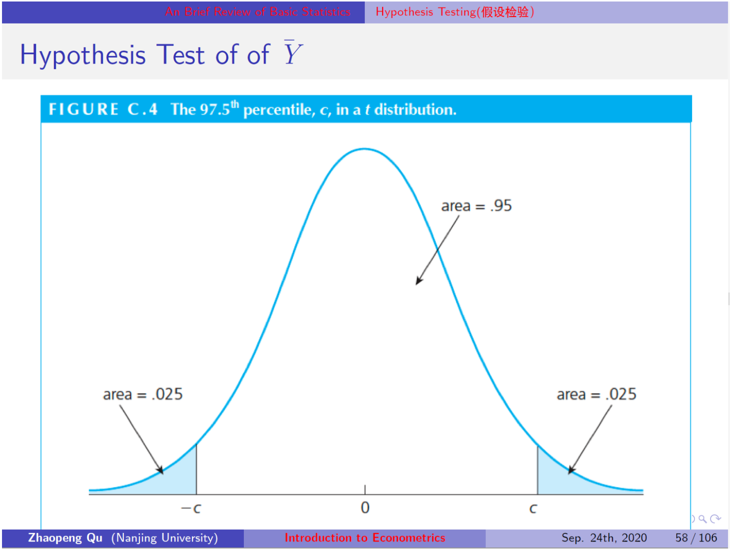
- Specify H_0 and H_1

$$H_0 : E[Y] = \mu_{Y,0} \; H_1 : E[Y] \neq \mu_{Y,0}$$

- Choose the significance level α and define a decision rule (critical region or critical value)
  - eg. if we choose $\alpha = 0.05$, then the critical value is 1.96, then the region is $(-\infty, -1.96]$ and $[1.96, +\infty)$

T-Test and Table

- Review the Theory



T-Test and Table

- Review the Theory

Hypothesis Test of of $\bar{Y}$

- Given the data compute the test statistic
  - Step1: Compute the sample average $\bar{Y}$
  - Step2: Compute the **standard error** of $\bar{Y}$

$$SE(\bar{Y}) = \frac{s_Y}{\sqrt{n}}$$

  - Step3: Compute the **t-statistic**

$$t^{act} = \frac{\bar{Y} - \mu_{Y,0}}{SE\left(\bar{Y}\right)}$$

  - Step4: Reject the null hypothesis if
    - $\mid t^{act} \mid > critical\ value$
    - or if $p - value < significance\ level$

- Review the Theory

Hypothesis Tests for the Difference Between Two Means

- To illustrate a test for the difference between two means, let μ_w be the mean hourly earning in the population of women recently graduated from college and let μ_m be the population mean for recently graduated men.
- Then the **null hypothesis** and **the two-sided alternative hypothesis** are

$$\begin{aligned} H_0 : \mu_m &= \mu_w \\ H_1 : \mu_m &\neq \mu_w \end{aligned}$$

- Consider the null hypothesis that mean earnings for these two populations differ by a certain amount, say d_0. The null hypothesis that men and women in these populations have the same mean earnings corresponds to H_0 : $H_0 : d_0 = \mu_m - \mu_w = 0$

T-Test and Table

- Review the Theory

The Difference Between Two Means

- Suppose we have samples of n_m men and n_w women drawn at random from their populations. Let the sample average annual earnings be $\overline{Y}_m$ for men and $\overline{Y}_w$ for women. Then an estimator of $\mu_m - \mu_w$ is $\overline{Y}_m - \overline{Y}_w$.
- Let us discuss the distribution of $\overline{Y}_m - \overline{Y}_w$.

$$\sim N(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w})$$

- if σ_m^2 and σ_w^2 are known, then the this approximate normal distribution can be used to compute p-values for the test of the null hypothesis. In practice, however, these population variances are typically unknown so they must be estimated.
- Thus the *standard error* of $\overline{Y}_m - \overline{Y}_w$ is

$$SE(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

- Review the Theory

The Difference Between Two Means

- The t-statistic for testing the null hypothesis is constructed analogously to the t-statistic for testing a hypothesis about a single population mean, thus *t-statistic* for comparing two means is

$$t_{act} = \frac{\overline{Y}_m - \overline{Y}_w - d_0}{SE(\overline{Y}_m - \overline{Y}_w)}$$

- If both $n_m$and n_m are large, then this t-statistic has a standard normal distribution when the null hypothesis is true,thus $\overline{Y}_m - \overline{Y}_w = 0$.

T-Test and Table

- Review the Theory

Confidence Intervals for the Difference Between Two Means

- the 95% two-sided confidence interval for d consists of those values of d within ± 1.96 standard errors of $\overline{Y}_m - \overline{Y}_w$, thus $d = \mu_m - \mu_w$ is

$$(\overline{Y}_m - \overline{Y}_w) \pm 1.96 SE(\overline{Y}_m - \overline{Y}_w)$$

- 单样本t检验

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

. ttest price == 6000 if foreign == 0 ,level(90)

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
   price |      52    6072.423    429.4911    3097.104    5352.903    6791.943
------------------------------------------------------------------------------
    mean = mean(price)                                            t =   0.1686
Ho: mean = 6000                                  degrees of freedom =       51

   Ha: mean < 6000             Ha: mean != 6000               Ha: mean > 6000
 Pr(T < t) = 0.5666         Pr(|T| > |t|) = 0.8668          Pr(T > t) = 0.4334

* level默认95%的水平
* 结果p值大于0.1，不能拒绝H0
```

- 独立样本t检验
  - 一个变量利用另一个变量来分组比较

```
. sdtest price, by(foreign)
Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      52    6072.423    429.4911    3097.104    5210.184    6934.662
 Foreign |      22    6384.682    558.9942    2621.915     5222.19    7547.174
---------+--------------------------------------------------------------------
combined |      74    6165.257    342.8719    2949.496    5481.914      6848.6
------------------------------------------------------------------------------
    ratio = sd(Domestic) / sd(Foreign)                            f =   1.3953
Ho: ratio = 1                                    degrees of freedom =   51, 21

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.7963         2*Pr(F > f) = 0.4073          Pr(F > f) = 0.2037


* 方差齐性检验(F检验)
* 对两个独立样本进行比较的时候，首先要判断两总体方差是否相同，即方差齐性。
* 若两总体方差相等equal variances(方差齐)，则直接用t检验;
* 若不等unequal variances(方差不齐)， 选择unequal variances(方差不齐)的均值T检验去
做,加unequal选项。
```

T-Test and Table

- 独立样本t检验
  - 一个变量利用另一个变量来分组比较

```
. ttest price, by(foreign)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      52    6072.423    429.4911    3097.104    5210.184    6934.662
 Foreign |      22    6384.682    558.9942    2621.915     5222.19    7547.174
---------+--------------------------------------------------------------------
combined |      74    6165.257    342.8719    2949.496    5481.914      6848.6
---------+--------------------------------------------------------------------
    diff |           -312.2587    754.4488               -1816.225    1191.708
------------------------------------------------------------------------------
    diff = mean(Domestic) - mean(Foreign)                         t =  -0.4139
Ho: diff = 0                                     degrees of freedom =       72

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.3401         Pr(|T| > |t|) = 0.6802          Pr(T > t) = 0.6599
```

T-Test and Table

- 独立样本t检验
  - 在两个变量间进行比较

```
. webuse fuel,clear
```

```
. sdtest mpg1 == mpg2
Variance ratio test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    mpg1 |      12          21    .7881701    2.730301    19.26525    22.73475
    mpg2 |      12       22.75    .9384465    3.250874    20.68449    24.81551
---------+--------------------------------------------------------------------
combined |      24      21.875    .6264476    3.068954    20.57909    23.17091
------------------------------------------------------------------------------
    ratio = sd(mpg1) / sd(mpg2)                                   f =   0.7054
Ho: ratio = 1                                    degrees of freedom =   11, 11

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.2862         2*Pr(F < f) = 0.5725           Pr(F > f) = 0.7138
```

T-Test and Table

- 独立样本t检验
  - 在两个变量间进行比较

```
. ttest mpg1 == mpg2, unpaired
Two-sample t test with equal variances
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    mpg1 |      12          21    .7881701    2.730301    19.26525    22.73475
    mpg2 |      12       22.75    .9384465    3.250874    20.68449    24.81551
---------+--------------------------------------------------------------------
combined |      24      21.875    .6264476    3.068954    20.57909    23.17091
---------+--------------------------------------------------------------------
    diff |               -1.75    1.225518               -4.291568    .7915684
------------------------------------------------------------------------------
    diff = mean(mpg1) - mean(mpg2)                                t =  -1.4280
Ho: diff = 0                                     degrees of freedom =       22

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0837         Pr(|T| > |t|) = 0.1673          Pr(T > t) = 0.9163


* unpaired 表示对两个不同变量检验，不是配对检验
```

- 配对样本t检验(单样本t检验的扩展)
  - 检验对象是配对样本观测值之差

```
. ttest mpg1==mpg2

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    mpg1 |      12          21    .7881701    2.730301    19.26525    22.73475
    mpg2 |      12       22.75    .9384465    3.250874    20.68449    24.81551
---------+--------------------------------------------------------------------
    diff |      12       -1.75    .7797144     2.70101    -3.46614   -.0338602
------------------------------------------------------------------------------
     mean(diff) = mean(mpg1 - mpg2)                               t =  -2.2444
 Ho: mean(diff) = 0                              degrees of freedom =       11

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0232         Pr(|T| > |t|) = 0.0463          Pr(T > t) = 0.9768


* 没有unpaired选项
* 结果p值小于0.05，拒绝H0
```

T-Test and Table

- -ttest-的局限
  - 每次只能对一个变量进行检验，无法批量对多个变量检验。
  - 汇报结果过于详细，有时我们只需要一个相对精简的结果，如两组各自均值，均值差异，T-Statistic或者P-Value。
  - 当待检验变量增加，ttest 命令费时费力。

- 多变量均值比较表格输出-ttable2-

```
ssc install ttable2

. sysuse auto,clear
(1978 Automobile Data)

. ttable2 price wei len mpg, by(foreign) f(%6.2f)
---------------------------------------------------------------------------
Variables   G1(Domestic)      Mean1     G2(Foreign)     Mean2      MeanDiff
---------------------------------------------------------------------------
price          52           6072.42        22         6384.68    -312.26
weight         52           3317.12        22         2315.91    1001.21***
length         52           196.13         22         168.55      27.59***
mpg            52            19.83         22          24.77      -4.95***
---------------------------------------------------------------------------
```

- 多变量均值比较表格输出-ttable2-

```
. tab rep78
     Repair
Record 1978 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          2        2.90        2.90
          2 |          8       11.59       14.49
          3 |         30       43.48       57.97
          4 |         18       26.09       84.06
          5 |         11       15.94      100.00
------------+-----------------------------------
      Total |         69      100.00
. ttable2 price wei len mpg if rep78==3|rep78==4, by(rep78)
--------------------------------------------------------------------------
Variables    G1(3)          Mean1     G2(4)          Mean2       MeanDiff
--------------------------------------------------------------------------
price          30        6429.233        18       6071.500     357.733
weight         30        3299.000        18       2870.000     429.000*
length         30         194.000        18        184.833       9.167
mpg            30          19.433        18         21.667      -2.233*
--------------------------------------------------------------------------
* 当组类别大于两类时，可以通过指定样本范围进行比较
```

T-Test and Table

- 结果导出-logout-

```
ssc install logout

logout, save(ttable) excel replace : ttable2 price ///
        wei len mpg, by(foreign) f(%6.2f)
logout, save(ttable) word replace : ttable2 price  ///
        wei len mpg, by(foreign) f(%6.2f)
logout, save(ttable) tex replace : ttable2 price   ///
        wei len mpg, by(foreign) f(%6.2f)
```

- 结果导出-t2docx-

```
ssc install t2docx

t2docx price weight length mpg  ///
    using ttable1.docx,replace     ///
    by(foreign)                 ///
    title("表1: t检验")
```

- 结果导出-esttab-

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


. local var price wei len mpg
. qui estpost ttest `var´, by(foreign)
. esttab ., cell("mu_1(fmt(2)) mu_2(fmt(2)) b(star fmt(2)) t(fmt(2))")   ///
>           starlevels(* 0.10 ** 0.05 *** 0.01) replace noobs compress   ///
>           title(esttab_Table: T_test)
esttab_Table: T_test by group
-----------------------------------------------------
                    (1)

                  mu_1       mu_2          b            t
-----------------------------------------------------
price          6072.42    6384.68    -312.26        -0.41
weight         3317.12    2315.91    1001.21***      6.25
length          196.13     168.55      27.59***      5.89
mpg              19.83      24.77      -4.95***     -3.63
-----------------------------------------------------
```

T-Test and Table

- 结果导出-esttab-

```
sysuse auto,clear

local var price wei len mpg
qui estpost ttest `var´, by(foreign)
esttab using ttable2.rtf, cell("mu_1(fmt(2)) mu_2(fmt(2)) b(star fmt(2)) t(fmt(2))")  ///
        starlevels(* 0.10 ** 0.05 *** 0.01) replace noobs compress                    ///
        title(esttab_Table: T_test)
```

Section 4

Descriptive Statistics Table

Descriptive Statistics Table

- 描述性统计表格导出
  - -logout-

```
logout, save(Desc1) word replace:           ///
tabstat price wei len mpg rep78,            ///
   stats(mean sd min p50 max) c(s) f(%6.2f)
```

Descriptive Statistics Table

- 描述性统计表格导出
  - -sum2docx-

```
sum2docx price wei len mpg rep78 using Desc2.docx,replace            ///
    stats(N mean(%9.2f) sd(%9.3f) min(%9.2f) median(%9.2f) max(%9.2f)) ///
    title(sum2docx_Table: Descriptive statistics)


*仅sum2docx支持中文,其余命令不支持
*能设置每个统计量的小数点位数
```

Descriptive Statistics Table

- 描述性统计表格导出
  - -outreg2-

```
outreg2 using Desc3, sum(detail) replace word                    ///
    keep(price wei len mpg rep78) eqkeep(N mean sd min p50 max) ///
    fmt(f) sortvar(wage age grade)                               ///
    title(outreg2_Table: Descriptive statistics)


*若变量里有字符串变量,outreg2命令的处理最智能化:
*会在窗口说明什么变量是字符型，并在报告列表中自动剔除该变量
*支持变量排序
```

Descriptive Statistics Table

- 描述性统计表格导出
  - -esttab-

```
estpost summarize price wei len mpg rep78, detail
esttab using Desc4.rtf,                                                   ///
    cells("count mean(fmt(2)) sd(fmt(2)) min(fmt(2)) p50(fmt(2)) max(fmt(2))") ///
    noobs compress replace title(esttab_Table: Descriptive statistics)


*能设置每个统计量的小数点位数
```

Section 5

Correlation Matrix Table

Correlation Matrix Table

- 相关系数矩阵导出

```
*<方法一> -logout-
logout, save(Corr1) word replace: pwcorr price wei len mpg rep78, star(.05)


*<方法二> -esttab-
estpost correlate  price wei len mpg rep78, matrix
esttab using Corr2.rtf,                                                  ///
    unstack not noobs compress nogaps replace star(* 0.1 ** 0.05 *** 0.01) ///
    b(%8.3f) p(%8.3f) title(esttab_Table: correlation coefficient matrix)


*<方法三> -corr2docx-
corr2docx price wei len mpg rep78 using Corr3.docx,                      ///
        replace spearman(ignore) pearson(pw) star                        ///
        title(corr2docx_Table: correlation coefficient matrix)
```

Section 6

OLS Regression-Estimation

Subsection 1

Data Analysis Flow

OLS Regression-Estimation

- Data Analysis Flow
  - Open the data, find the variables, and see the base case.
  - Data Cleaning.
  - Summary Statistics: Figures and Tables.
  - Model Estimation and Hypothesis Testing.
  - Report results, explain and analyze.

Subsection 2

Review the Theory

OLS Regression-Estimation

- Review the Theory

Terminology for Simple Regression Model

- The linear regression model with one regressor is denoted by

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Where
  - Y_i is the **dependent variable**(Test Score)
  - X_i is the **independent variable** or regressor(Class Size or Student-Teacher Ratio)
  - $\beta_0 + \beta_1 X_i$ is the **population regression line** or the **population regression function**

- Review the Theory

Review for the previous lectures

The OLS Estimator

- The estimators of the slope and intercept that *minimize the sum of the squares* of $\hat{u}_i$,thus

$$arg\,min_{b_0,b_1} \sum_{i=1}^{n} \hat{u}_i^2 = min_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

are called the **ordinary least squares (OLS) estimators** of β_0 and β_1.

OLS estimator of β_1:

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})}$$

OLS Regression-Estimation

- Review the Theory

Least Squares Assumptions

1. Assumption 1: Conditional Mean is Zero
2. Assumption 2: Random Sample
3. Assumption 3: Large outliers are unlikely

- If the 3 least squares assumptions hold the OLS estimators will be
  - **unbiased**
  - **consistent**
  - **normal sampling distribution**

OLS Regression-Estimation

- Review the Theory

Multiple regression model with k regressors

- The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n \quad (4.1)$$

where

- Y_i is the **dependent variable**
- $X_1, X_2, ...X_k$ are the **independent variables(includes one is our of interest and some control variables)**
- $\beta_i, j = 1...k$ are slope coefficients on X_i corresponding.
- β_0 is the estimate *intercept*, the value of Y when all $X_j = 0, j = 1...k$
- u_i is the regression *error term*, still all other factors affect outcomes.

OLS Regression-Estimation

- Review the Theory

Multiple Regression: Assumption

- Assumption 1: The conditional distribution of u_i given $X_{1i}, ..., X_{ki}$ has mean zero,thus

$$E[u_i|X_{1i}, ..., X_{ki}] = 0$$

- Assumption 2: $(Y_i, X_{1i}, ..., X_{ki})$ are i.i.d.
- Assumption 3: Large outliers are unlikely.
- Assumption 4: No perfect multicollinearity.

OLS Regression-Estimation

- Review the Theory
  - The OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1 ... \widehat{\beta}_k$ are *unbiased*.
  - The OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1 ... \widehat{\beta}_k$ are *consistent*.
  - The OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1 ... \widehat{\beta}_k$ are *normally distributed* in large samples.

  - Multiple OLS estimator

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{j,i} Y_i}{\sum_{i=1}^{n} \tilde{X}_{j,i}^2} \; for \; j = 1, 2, .., k$$

Subsection 3

OLS in stata

OLS Regression-Estimation

- 普通最小二乘法(OLS)

```
. *help  reg
. *regress depvar [indepvars] [if] [in] [weight] [, options]  //因变量, 自变量


. sysuse auto, clear
(1978 Automobile Data)

. reg price weight mpg turn foreign
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(4, 69)        =     19.23
       Model |   334771309         4  83692827.3   Prob > F        =    0.0000
    Residual |   300294087        69  4352088.22   R-squared       =    0.5271
-------------+----------------------------------   Adj R-squared   =    0.4997
       Total |   635065396        73  8699525.97   Root MSE        =    2086.2

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   4.284532   .7404967     5.79   0.000     2.807282    5.761783
         mpg |  -.4660076   73.51407    -0.01   0.995    -147.1226    146.1905
        turn |  -229.2059   114.2423    -2.01   0.049    -457.1131   -1.298676
     foreign |   3221.415   706.4847     4.56   0.000     1812.017    4630.813
       _cons |   1368.197   4887.597     0.28   0.780    -8382.292    11118.69
------------------------------------------------------------------------------
```

OLS Regression-Estimation

- 普通最小二乘法(OLS)

```
. regress weight length, noconstant  //不包括截距项（constant）
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 73)        =   3450.13
       Model |   703869302         1   703869302   Prob > F        =    0.0000
    Residual |  14892897.8        73  204012.299   R-squared       =    0.9793
-------------+----------------------------------   Adj R-squared   =    0.9790
       Total |   718762200        74   9713002.7   Root MSE        =    451.68

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   16.29829   .2774752    58.74   0.000     15.74528     16.8513
------------------------------------------------------------------------------
```

OLS Regression-Estimation

- 普通最小二乘法(OLS)

```
. reg price weight mpg turn foreign, robust  //稳健标准误（robust）
Linear regression                                    Number of obs     =         74
                                                     F(4, 69)          =      12.46
                                                     Prob > F          =     0.0000
                                                     R-squared         =     0.5271
                                                     Root MSE          =     2086.2

------------------------------------------------------------------------------
             |               Robust
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   4.284532   .9164881     4.67   0.000     2.456188    6.112876
         mpg |  -.4660076   84.34373    -0.01   0.996    -168.7271    167.7951
        turn |  -229.2059   136.4962    -1.68   0.098    -501.5084    43.09658
     foreign |   3221.415   690.7001     4.66   0.000     1843.506    4599.324
       _cons |   1368.197   6008.419     0.23   0.821    -10618.27    13354.66
------------------------------------------------------------------------------
```

OLS Regression-Estimation

- 回归结果

```
. regress price mpg weight foreign
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(3, 70)        =     23.29
       Model |   317252881         3   105750960   Prob > F        =    0.0000
    Residual |   317812515        70  4540178.78   R-squared       =    0.4996
-------------+----------------------------------   Adj R-squared   =    0.4781
       Total |   635065396        73  8699525.97   Root MSE        =    2130.8

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |    21.8536   74.22114     0.29   0.769    -126.1758     169.883
      weight |   3.464706    .630749     5.49   0.000     2.206717    4.722695
     foreign |    3673.06   683.9783     5.37   0.000     2308.909    5037.212
       _cons |  -5853.696   3376.987    -1.73   0.087    -12588.88    881.4934
------------------------------------------------------------------------------
```

OLS Regression-Estimation

- 回归结果

```
predict yhat, xb        //price的拟合值
predict e, residual     //残差
vce                     //获取变量的方差—协方差矩阵

. test mpg = 20               //单变量检验
 ( 1)  mpg = 20
       F(  1,    70) =    0.00
            Prob > F =    0.9801
. test mpg weight foreign   //联合检验
 ( 1)  mpg = 0
 ( 2)  weight = 0
 ( 3)  foreign = 0
       F(  3,    70) =   23.29
            Prob > F =    0.0000
```

OLS Regression-Estimation

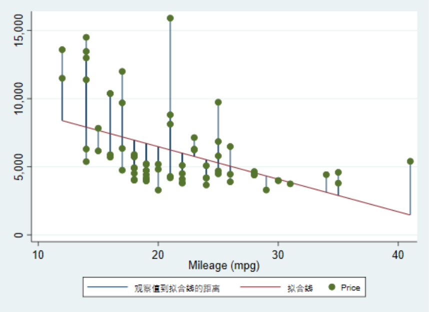
- 回归结果

```
. qui reg price mpg

. predict yhat_p, xb
(option xb assumed; fitted values)

. twoway (rspike price yhat_p mpg )     ///
>        (lfit price mpg)               ///
>        (scatter price mpg ),          ///
>        legend(label(1 "观察值到拟合线的距离") label(2 "拟合线") row(1) size(small))

. graph export olsf.png, width(500) replace
(note: file olsf.png not found)
(file olsf.png written in PNG format)
```



Subsection 4

OLS Result Table

OLS Regression-Estimation

- 回归结果输出-esttab-
  - word文档

```
sysuse nlsw88, clear

reg wage age married occupation
est store m1
reg wage age married collgrad occupation
est store m2
xi: reg wage age married collgrad occupation i.race
est store m3

esttab m1 m2 m3 using ols.rtf, scalar(r2 r2_a N F) compress  ///
    star(* 0.1 ** 0.05 *** 0.01)                              ///
    b(%6.3f) t(%6.3f) r2(%9.3f) ar2                           ///
    mtitles("OLS-1" "OLS-2" "OLS-3")                          ///
    title(esttab_Table: regression result)
```

OLS Regression-Estimation

- 回归结果输出-esttab-
  - Tex文档

```
    esttab m1 m2 m3 using ols.tex, replace          ///
        star( * 0.10 ** 0.05 *** 0.01 ) compress    ///
        b(%6.3f) t(%6.3f) r2(%9.3f) ar2             ///
        mtitles("OLS-1" "OLS-2" "OLS-3")            ///
        title(esttab_Table: regression result)      ///
        booktabs page width(\hsize)

/*
esttab 的 LaTeX 输出的专有选项:
1. booktabs: 用 booktabs 宏包输出表格(三线表格)。
2. page[(packages)]: 创建完成的 LaTeX 文档以及添加括号里的宏包
3. 如果写了 booktabs 选项, 则 page[(packages)] 将自动添加\usepackagebooktabs。
4. alignment(cccccc): 定义从第二列开始的列对齐方式(默认居中)。
5. width(\hsize): 可以使得表格宽度为延伸至页面宽度
6. fragment: 不输出表头表尾, 只输出表格本身内容, 其不能与 page[(packages)] 选项共存。
*/
```

OLS Regression-Estimation

表 1: esttab_Table: regression result

	(1) OLS-1	(2) OLS-2	(3) OLS-3
age	-0.064	-0.059	-0.067*
	(-1.637)	(-1.579)	(-1.796)
married	-0.469*	-0.472**	-0.629**
	(-1.873)	(-1.983)	(-2.578)
occupation	-0.284***	-0.384***	-0.370***
	(-8.055)	(-11.251)	(-10.756)
collgrad		4.220***	4.133***
		(15.444)	(15.051)
_Irace_2			-0.784***
			(-2.897)
_Irace_3			-0.224
			(-0.210)
_cons	11.910***	11.168***	11.753***
	(7.654)	(7.545)	(7.878)
N	2237	2237	2237
R^2	0.031	0.125	0.128
adj. R^2	0.030	0.123	0.126

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$