# 统计学（Stata实现）

## 第三次上机课

赵震宇（2120253538）
南开大学 国际经济研究所
zzynankai@outlook.com
xishanyu2.github.io
2026 年 1 月 8 日

CONTENT

# 目录

# 1. 列联分析

# 1.1 列联表分析

列联表是分析两个分类变量（名义变量或者定序变量）之间关系的基本统计方法。设两个变量A和B分别有r和c个类型，则它们可以构成一个的列联表。

在对两变量进行列联表分析时，首先要检验它们的独立性，这就涉及到独立性检验统计量，常用的统计量有皮尔逊卡方统计量（Pearson Chi-Square）和似然比统计量（Likelihood Ratio）。

$$\chi_p^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2[(r-1)(c-1)]$$

# 1.2 列联表中的相关测量统计量

对于名义变量，通常采用基于卡方统计量的各种统计量来进行关联性度量，包括$\varphi$相关系数、列联相关系数、Cramer’s V相关系数。

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

$$c = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$V = \sqrt{\frac{\chi^2}{n \times \min[(r-1),(c-1)]}}$$

对于有序变量，关联性度量的统计量通常是基于观测中的一致对和不一致对的个数。可以定义以下统计量来描述有序变量之间的有序关联性的大小：gamma统计量和Kendall统计量。

## 1.3 使用 table 命令生成列联表

table命令可以用于生成一维到多维的列联表，表中不仅可以包含常见的频数，还可以包含任意其他变量的描述性统计量。

```
* option contents() not allowed since Stata 17 in command "table"
* option format() not allowed since Stata 17 in command "table"
```

# 例题1

数据集auto.dta是Stata系统自带的关于1978年汽车市场的一个调查数据，该数据集共有74个观测值，12个变量。变量的具体情况是：make是字符型变量，表明生产厂商；其他均为数值型变量，price是汽车的价格（万元），mpg是行驶里程（英里），rep78是在该年度修理次数，foreign是虚拟变量（赋值1代表国外产，0代表国产），headroom、trunk、length、turn、displacement、gear_ratio依次代表汽车内部空间大小、后备箱空间大小、身长、转弯半径、排气量和变速比。

（1）使用`table`命令创建关于rep78的频数表，并在表中显示mpg的观测值个数、平均数、标准差和中位数。

```
table rep78, statistic(n mpg) statistic(mean mpg) statistic(sd mpg) statistic(median mpg)
tabstat mpg, by(rep78) statistics(count mean sd median)
```

（2）使用`table`命令创建关于foreign和rep78的二维表格，并在表中显示mpg的平均数。

```
table foreign rep78, statistic(mean mpg)
```

# 例题2

使用一个关于肺炎的调查数据集`byssin1.dta`。数据集中每个观测值代表了一类人，变量prob是该类人罹患肺炎的概率，smokes是代表是否吸烟的虚拟变量，race是关于是否是白人的虚拟变量，workplace按照工作场所的烟尘等级划分为三类（1为最少，3为最多），sex是性别（1代表男性）。最关键的变量是pop，表示具有该观测值的个体的数量，比如，pop=3，则意味着样本中有3个人具有该观测值所揭示的性质。在这个数据集中，pop的取值从0到507不等，其中0说明没有任何个体具有该观测值所揭示的性质，507则说明有507个个体具有同样的性质。

（1）创建一个关于workplace、smokes和race的三维列联表。

```
table workplace smokes race [fweight=pop], statistic(mean prob) nformat(%9.3f)
```

（2）创建一个关于workplace、smokes、race和sex的四维列联表。

```
table workplace smokes race sex [fweight=pop], statistic(mean prob) nformat(%9.3f)
table (sex workplace) (race smokes) [fweight=pop], statistic(mean prob) nformat(%9.3f)

* 不同变量组合、不同分组先后顺序、不同行和列?
```

## 1.4 使用 tabulate 命令进行列联表分析

tabulate的一维命令主要用于生成含有频数的一维表格。

tabulate一维命令还有一个配套的方便工具——tab1，在其后可以添加多个变量，然后Stata会对每个变量创建一维表格，也就是相当于多次执行tabulate命令。

# 例题3

使用`auto.dta`数据创建rep78和foreign的一维频数表。

```
sysuse auto,clear
tabulate rep78,sort
tabulate rep78
tabulate foreign,nolabel
tabulate foreign

tab1 rep78 foreign
tab1 rep78 foreign, sort
tab1 rep78 foreign, nolabel
tab1 rep78 foreign, sort nolabel
```

二维`tabulate`命令在生成二维表格的同时，可以计算多种独立性检验统计量和相关测量统计量，包括常用的Pearson's chi-squared、likelihood-ratio chi-squared、Cram 's V、Goodman and Kruskal's gamma、Kendall's tau-b

同`tabulate`的一维表格一样，二维`tabulate`命令也有`tab2`，可用于快速生成所罗列的变量所有可能的二维表格。

# 例题4

利用关于肺炎的调查数据byssin1.dta来检验得肺炎概率与是否抽烟、工作场所空气质量这两个变量之间的独立性（当然，凭借常识，它们应该是有关系的），即分别对probcat和smokes以及probcat和workplace做列联分析。

```
sort prob
gen probcat=group(5)
tabulate probcat smokes [fw=pop], chi2
tabulate probcat smokes [fw=pop], all
tabulate probcat workplace [fw=pop], chi2
tabulate probcat workplace [fw=pop], all
```

# 例题5

对于已经打印的表格，手头又没有具体的数据，可以使用`tabi`做即时的独立性检验。

下表汇总了1912年4月15日泰坦尼克号沉没时，乘客们和船员的命运，请检验性别和存活之间在统计上是否关联。

| 项目 | 男人 | 女人 | 男孩 | 女孩 | 总计 |
|---|---|---|---|---|---|
| 幸存 | 332 | 318 | 29 | 27 | 706 |
| 死亡 | 1360 | 104 | 35 | 18 | 1517 |
| 总计 | 1692 | 422 | 64 | 45 | 2223 |

```
tabi 332 318 29 27\ 1360 104 35 18, all
```

# 2. 方差分析

# 2.1 方差分析

方差分析是基于样本方差对总体均值进行统计推断的方法，它是通过实验观察某一种或多种因素的变化对实验结果是否带来显著影响，进而鉴别各种因素的效应，从而选取一种最优方案。

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1 : \mu_1, \mu_2, \cdots, \mu_k \text{不全相等}$$

$$SST = SSA + SSE$$

$$SST = SSR + SSC + SSE$$

$$F = \frac{MSA}{MSE} = \frac{SSA/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k)$$

## 2.2 单因素方差分析

单因素方差分析用于比较多组样本的均值是否相同，并假定：每组的数据服从正态分布，具有相同的方差，且相互独立。

实现单因素方差分析的两个命令为`oneway`和`longway`。

# 例题1

大学生信息表.dta来自于Ward和Ault（1990）对在校大学生的抽样调查，其中year代表所处的大学年级，gender代表性别，drink用一个33级别表来衡量学生喝酒频度和程度，gpa代表学分绩点，belong表示是否是大学生联谊会的会员，employed为是否就业。检验大学生饮酒行为平均数是否会因为是否就业而有所变化。

```
oneway drink employed
/*
                        Analysis of variance
    Source              SS         df      MS             F      Prob > F
------------------------------------------------------------------------
Between groups        273.74774      1    273.74774       6.19     0.0136
 Within groups       10574.7336    239   44.2457472
------------------------------------------------------------------------
    Total            10848.4813    240   45.2020055

Bartlett's equal-variances test: chi2(1) =   0.0683    Prob>chi2 = 0.794
*/

oneway drink employed,tabulate scheffe
//其中tabulate可以产生平均数和标准差表
//Scheffe选项生成了一个表来显示在每一对平均数之间的差异
```

## 2.3 双因素和多因素方差分析

多因素方差分析用来研究两个及两个以上控制变量是否对观测变量产生显著影响。研究多个因素对观测变量的影响，称为多因素方差分析；若研究的是两个变量，则称为双因素方差分析。
多因素方差分析不仅能够分析多个因素对观测变量的独立影响，更能够分析多个控制因素的交互作用能否对观测变量的分布产生显著影响，进而最终找到利于观测变量的最优组合。

在Stata中用命令`anova`实现。

# 例题2

员工信息表.dta介绍了某单位的员工信息，其中minority代表是否属于少数民族（0是非少数民族，1表示为少数民族），educ代表教育年限，salary为年薪，beginsalary为起始年薪，gender为性别。考察工资是否因性别、少数民族与否的不同而存在差异。

```
gen gender_minority = gender * minority
anova salary gender minority gender_minority
/*
                          Number of obs =       474    R-squared     =  0.2578
                          Root MSE      =   14758.2    Adj R-squared =  0.2530

                  Source | Partial SS         df         MS        F    Prob>F
              -----------+----------------------------------------------------
                   Model |  3.555e+10          3   1.185e+10     54.40  0.0000
                         |
                  gender |  2.914e+10          1   2.914e+10    133.77  0.0000
                minority |  7.197e+09          1   7.197e+09     33.04  0.0000
           gender_mi~y |  1.432e+09          1   1.432e+09      6.58  0.0106
                         |
                Residual |  1.024e+11        470   2.178e+08
              -----------+----------------------------------------------------
                   Total |  1.379e+11        473   2.916e+08
*/
```

## 2.4 协方差分析

不论是单因素方差分析还是多因素方差分析，控制因素都是可控的，其各个水平可以通过人为的努力得到控制和确定。但在许多实际问题中，有些控制因素很难人为控制，但它们的不同水平确实对观测变量产生了较为显著的影响。

协方差分析将那些人为很难控制的控制因素作为协变量，并在排除协变量对观测变量影响的条件下，分析控制变量（可控）对观测变量的作用，从而更加准确地对控制因素进行评价。

协方差分析拓展了多因素方差分析，使之可以包含分类变量和连续变量的情况。当出现连续变量时，定义此变量，方差分析便可进行。anova具有处理连续变量和分类变量的能力。

# 例题3

继续使用“员工信息表.dta”中的数据，检验薪水水平的差异是否还由起始薪水的差异所引起，其中起始薪水为连续变量。

在分析薪水差异的原因之后，对变量进行回归分析，得到回归方程。

```
* 用c.和i.来区分连续变量和分类变量
anova salary gender minority gender_minority c.beginsalary
/*
                      Number of obs =       474    R-squared     =  0.7803
                      Root MSE      =   8037.97    Adj R-squared =  0.7784

                 Source | Partial SS         df         MS        F    Prob>F
            ------------+----------------------------------------------------
                  Model |  1.076e+11          4   2.690e+10    416.41  0.0000
                        |
                 gender |  5.576e+08          1   5.576e+08      8.63  0.0035
               minority |  3.461e+08          1   3.461e+08      5.36  0.0211
          gender_mi~y |   52910806          1    52910806      0.82  0.3660
          beginsalary |  7.207e+10          1   7.207e+10   1115.42  0.0000
                        |
             Residual |  3.030e+10        469    64608890
            ------------+----------------------------------------------------
                Total |  1.379e+11        473   2.916e+08
*/
```

# 3. 时间序列分析

1. 时间数据的处理
2. 时间数据的声明
3. 时间序列的成分
4. 时间序列的修匀
5. 时间序列的预测

# 3.1 Date variable

```
use date.dta
* For ‘date1’ type: 1-Jan-95
gen datevar1 = date(date1,"DMY", 2020)
format datevar1 %td
* For ‘date2’ type: 1/1/1995
gen datevar2 = date(date2,"MDY")
format datevar2 %td
* For ‘date3’ type: 19950101
tostring date3, gen(date3a)
gen datevar3 = date(date3a,"YMD")
format datevar3 %td
```

```
* Year, month, day
tostring date4, gen(date4a)
gen len = length(date4a)
gen year = substr(date4a,1,4)
* When len=6, month is in 5th position and day in 6th
gen month = substr(date4a,5,1) if len == 6
gen day = substr(date4a,6,1) if len == 6
* When len=7 is hard to distinguish month/day, we skip
* When len=8, month is in 5th/6th positon and day in 7th/8th
replace month = substr(date4a,5,2) if len == 8
replace day = substr(date4a,7,2) if len == 8
destring month day year, replace
* Creating datevar
gen datevar4 = mdy(month,day,year)
format datevar4 %td
* Filling in the missing dates
replace datevar = datevar[_n-1] + 1 if datevar == .
```

To extract days of the week (Monday, Tuesday, etc.):

```
gen dayofweek = dow(datevar1)
```

This will create the variable ‘dayofweek’ where 0 is ‘Sunday’, 1 is ‘Monday’, etc.

**Self-study**

From daily to weekly/quarterly and getting monthly/yearly?

## 3.2 Setting as time series

Once you have the date variable in a ‘date format’ you need to declare your data as time series in order to use the time series operators:

```
tsset datevar
```

# 3.3 The four components of a time series

- trend
- cycle
- seasonal
- random noise

## Trend

For a real-world example of trend, let's look at U.S. GDP. The data are quarterly measurements of nominal GDP in billions of dollars, seasonally adjusted and converted to an annual rate. There are important nontrend components to GDP, but the trend is the dominant visual feature.

```
use quarterly.dta, clear
describe
keep if !missing(gdp)
tsset
tsline gdp
```

The upward trend of GDP over time is unmistakable. The downward "hook" near the end of the series is evidence of the recession that began in 2008.

## Cycle

For an example of cycle, let's look at the civilian unemployment rate.

```
use monthly.dta, clear
keep if !missing(unrate)
tsset
tsline unrate
```

The sharp increase in the unemployment rate at the right-hand side of the graph is another indicator of the post-2007 recession.

The lack of an obvious trend in this graph suggests (but does not prove) that the rate of unemployment accounted for by the sum of **frictional and structural unemployment** has been relatively stable during the post-World War II era. Instead the shape of the time line is dominated by the **cyclical** component of the unemployment rate.

## Seasonal

Another type of cycle is seasonality, the tendency of some series to increase or decrease in predictable ways at the same time of the day, the same day of the week, the same month of the year, etc. The average daily high temperature follows a seasonal pattern, as does the average temperature in a 24-hour period. This latter example of a seasonal component happens to have a smoothly oscillating character. Temperatures tend to rise during the morning, peak in the afternoon, and decline in the evening and overnight. Other seasonal variations are irregular, not resembling a smooth cycle at all. For instance, bank deposits rise and fall sharply around tax payment dates.

# 3.4 Filtering time-series data

Smoothers provide tools for filtering out the noise component in a series, making it easier to isolate the systematic components. Simple smoothers are motivated by the assumption that the signal in a series evolves smoothly over time, while the noise component is erratic. Additional time-series smoothers add incrementally more assumptions about the structure of the signal in the series, ultimately separating out the trend, cycle, and seasonal components.

Stata offers a wide variety of time-series smoothers through its family of `tssmooth` commands.

```
tssmooth smoother [type] newvar = exp [if] [in] [, ...]

   Smoother category             smoother
   ------------------------------------------------------
   Moving average                ma

   Recursive
     exponential                 exponential
     double exponential          dexponential
     nonseasonal Holt-Winters    hwinters
     seasonal Holt-Winters       shwinters

   Nonlinear filter              nl
   ------------------------------------------------------
```

# Nonlinear filter

We begin with the `tssmooth nl` command, which provides median smoothers and a special case of a weighted-average smoother. For instance, a span-5 median smoother calculates the t-th value of the fit as:

$$y_t^* = median\ of\ (y_{t-2}, y_{t-1}, y_t, y_{t+1}, y_{t+2})$$

It allows to calculate median smoothers with spans from one to nine observations. It also offers a span-3 weighted mean, called the **Hanning smoother**, that is defined as:

$$y_t^* = (y_{t-1} + 2y_t + y_{t+1})/4$$

```
webuse sales2, clear
sort t
tsset t
tsline sales, name(g1, replace)
tssmooth nl nl1=sales, smoother(5)
tsline nl1, name(g2, replace)
tssmooth nl nl2=sales, smoother(3RSSH)
tsline nl2, name(g3, replace)
tssmooth nl nl3=sales, smoother(3RSSH, twice)
tsline nl3, name(g4, replace)
graph combine g1 g2 g3 g4
```

# Moving-average filter

$$y_t^* = (y_{t-5} + y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1})/5$$

```
webuse sales1, clear
tsset
tsline sales, name(g1, replace)
tssmooth ma sm1=sales, window(2 1 3)
tsline sm1, name(g2, replace)
tssmooth ma sm2=sales, weights(1/2 <3> 2/1)
tsline sm2, name(g3, replace)
graph combine g1 g2 g3
```

# Single-exponential smoothing

$$y_t^* = \alpha y_t + (1-\alpha) y_{t-1}^*$$

```
webuse sales1, clear
tsset
tsline sales, name(g1, replace)
tssmooth exponential double sm2a=sales
tsline sm2a, name(g2, replace)
tssmooth exponential double sm2b=sales, p(.4)
tsline sm2b, name(g3, replace)
tssmooth exponential double sm2c=sales, p(.4) forecast(3)
tsline sm2c, name(g4, replace)
graph combine g1 g2 g3 g4
```

# Double-exponential smoothing

$$y_t^{**} = \alpha y_t^{*} + (1-\alpha) y_{t-1}^{**}$$

```
webuse sales2, clear
tsset
tsline sales, name(g1, replace)
tssmooth dexponential double sm2a=sales
tsline sm2a, name(g2, replace)
tssmooth dexponential double sm2b=sales, p(.7)
tsline sm2b, name(g3, replace)
tssmooth dexponential double sm2c=sales, p(.7) s0(1031 1031)
tsline sm2c, name(g4, replace)
tssmooth dexponential double sm2d=sales, p(.7) s0(1031 1031) forecast(4)
tsline sm2d, name(g5, replace)
graph combine g1 g2 g3 g4 g5
```

# Holt-Winters smoothing

## nonseasonal

## seasonal

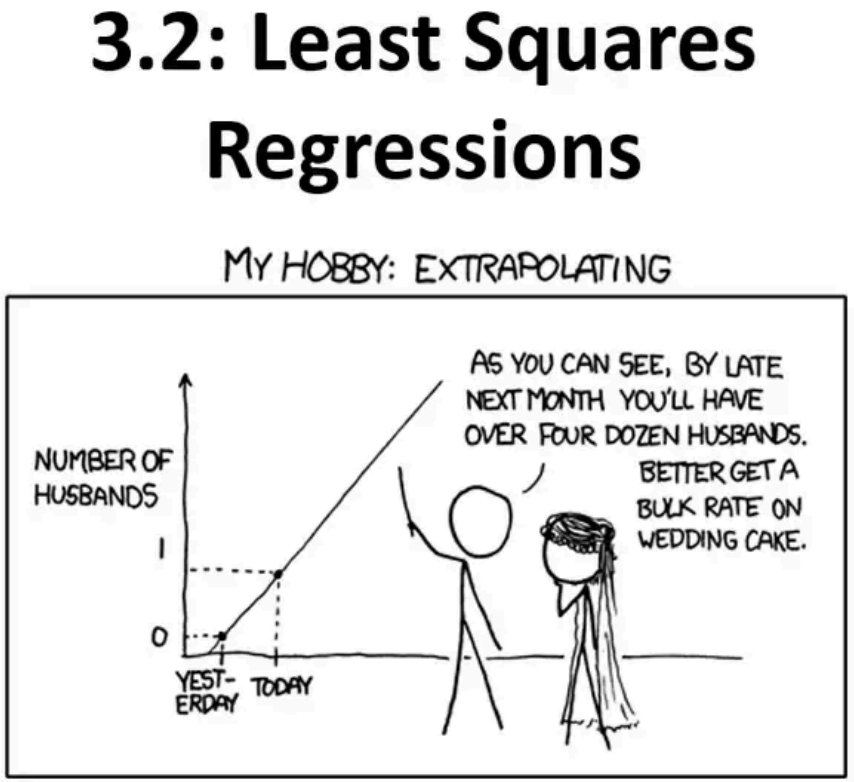
# 3.5 Forecasting

Once someone said: Forecasting is the art of saying what will happen in the future and then explaining why it didn't.

The reality is that forecasting is a really tough task, and you can do really bad, justlike in this cartoon:

But we can do definitely better using quantitative methods and common sense!

In the previous part, we introduced a wide variety of filters that help in highlighting patterns in time series. Some of these filters also can be used for forecasting.



# 4. 课后习题

# 习题1

文件CEOSAL2.dta包含了177位首席执行官的数据，可以用来考察企业业绩对CEO薪水的影响。请利用该数据完成下列操作：

(1)将salary分成三组（分组的方式可以自由设定，变量名设定为salarygroup），然后利用salarygroup和graduate两变量做二维列联表，并作卡方检验。

(2)利用salarygroup和graduate两变量生成二维表格，要求表中数值为lsalary的均值。

(3)将mktval分成三组（分组方式同上），利用salarygroup、mktvalgroup和graduate三个变量做三维列联表。

(4)请利用tabstat命令生成一个报告salary和age的均值、方差的表格，并按照grad变量分类。

# 习题2

1. 利用usaauto.dta数据进行单因素方差分析，分析内容为美国汽车的价格price是否受进口还是国产的影响，即以price为因变量，以foreign为分类变量进行单因素分析，并且进行结果的解读。

2. 利用womenwork.dta数据进行多因素方差分析，分析内容为妇女受教育水平education是否受结婚married和是否有子女children以及二者交互项的影响，即以education为因变量，以married、children和married*children为自变量进行多因素分析，并且进行结果的解读。

3. 利用usaauto.dta数据进行协方差分析，分析内容为美国汽车的价格price是否受进口车foreign、重量weight和长度length的影响，即以price为因变量，foreign为分类变量，weight和length为连续变量进行协方差分析，并且进行结果的解读。

# 作业要求

- 只需要提交代码（扫码填写问卷）
- **DDL**：1月11日24点前

# 欢迎交流~


zzynankai@outlook.com

Bilibili: 西山yu

xishanyu2.github.io