社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

December 31, 2020

CONTENTS

1 回归分析（复习）

(1) 嵌套模型

(2) OLS

(3) MLE

归纳与总结

- 统计学包括两部分内容：描述统计和推论统计。
  - 描述是对样本的描述。
  - 推论是对总体的推论。

归纳与总结

测量水平	单个变量	分类变量+连续变量	分类变量+分类变量	多分类变量+连续变量
针对样本	>=<	两个平均数差异的比较	两个分类变量的独立性	多个平均数差异的比较
针对总体	标准误、置信区间	双样本t检验	χ^2 检验	ANOVA：F检验

A Typology of Statistical Test

Statistical Test	Dependent Variable	Independent Variable
One sample t-test	interval or ratio	
Independent samples t-test	interval or ratio	nominal with only two categories
Paired samples t-tests	interval or ratio	nominal with only two categories
ANOVA	interval or ratio	nominal with more than two categories
χ^2	nominal or ordinal	nominal or ordinal

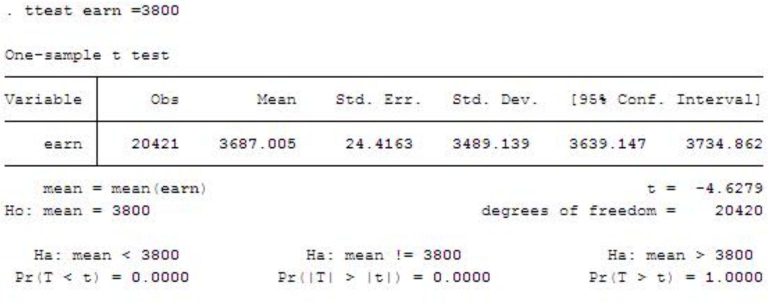
ONE-SAMPLE T TEST

Example 1.1

假设有人提出*1988*年全国城市居民平均收入为*3800*元，而在*1988*年的*CHIP*数据中，我们发现居民的年收入均值为*3687*元，标准差为*3489*元，那么在*0.05*的显著性水平下，这一样本结果与*3800*元的提法一致吗？

- 首先，建立$H_0: \mu = 3800$ 和 $H_1: \mu \neq 3800$
- 根据样本数据，计算t检验统计量
- stata 命令：ttest earn =3800
- 结论：否定H_0，即认为全国城市居民平均收入不是3800元。

ONE-SAMPLE T TEST



```
. ttest earn =3800

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    earn |   20421    3687.005     24.4163    3489.139    3639.147    3734.862
------------------------------------------------------------------------------
    mean = mean(earn)                                             t =  -4.6279
Ho: mean = 3800                                  degrees of freedom =    20420

   Ha: mean < 3800             Ha: mean != 3800               Ha: mean > 3800
 Pr(T < t) = 0.0000       Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Figure 1.1:

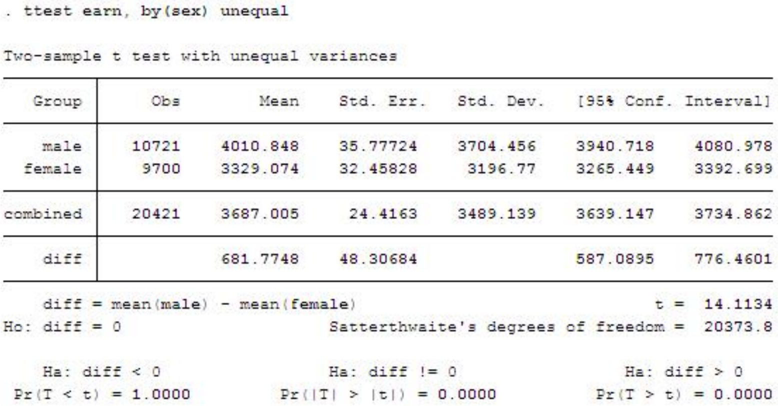
TWO-SAMPLE T TEST

Example 1.2

在*1988*年的*CHIP*数据中，有女性*9700*人，其年收入均值为*3329*元，标准差为*3197*元；有男性*10721*人，其年收入均值为*4011*元，标准差为*3704*元。那么在*0.05* 的显著性水平下，是否存在收入的性别差异?

- 首先，建立$H_0 : \mu_1 = \mu_2$ 和 $H_1 : \mu_1 \neq \mu_2$
- 检验方差是否相等
- 根据样本数据，计算t检验统计量
- stata 命令：ttest earn, by(sex) unequal
- 结论：否定H_0，即认为男性和女性的年收入均值是不相等的。

Two-sample t test



```
. ttest earn, by(sex) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |   10721    4010.848    35.77724    3704.456    3940.718    4080.978
  female |    9700    3329.074    32.45828     3196.77    3265.449    3392.699
---------+--------------------------------------------------------------------
combined |   20421    3687.005     24.4163    3489.139    3639.147    3734.862
---------+--------------------------------------------------------------------
    diff |            681.7748    48.30684                587.0895    776.4601
------------------------------------------------------------------------------
    diff = mean(male) - mean(female)                              t =  14.1134
Ho: diff = 0                     Satterthwaite's degrees of freedom =  20373.8

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Figure 1.2:

1. A variety of social science data come in the form of cross-classified tables of counts, commonly referred to as contingency tables.
2. With the log-linear approach, we model cell counts in a contingency table in terms of associations among the variables and marginal frequencies.

CHI-SQUARE TEST(OBSERVED)

		性别		总数
		男	女	
是否经理	否	290	100	390
	是	80	4	84
总数		370	104	474

CHI-SQUARE TEST(EXPECTED)

		性别		总数
		男	女	
是否经理	否	304.4	85.6	390
	是	65.6	18.4	84
总数		370	104	474

单元格的预期值，等于该单元格所在行的边缘值（marginals）乘以它所在列的边缘值，除以大总数（grand total）

$$\chi^2 = \sum \frac{(observed\ value - expected\ value)^2}{expected\ value}$$

Association between child abuse and substance abuse by parent of abusers (n=49)

Substance Abuse by Parent of Abusers	Child Abuse	
	Yes	No
Yes	18	19
No	2	10

$\chi^2 = 3.84(1), p < 0.05$, read as ”Chi-square with one degree of freedom equals 3.84”.

1 H_0 两个变量无关。

2 H_1 两个变量有关。

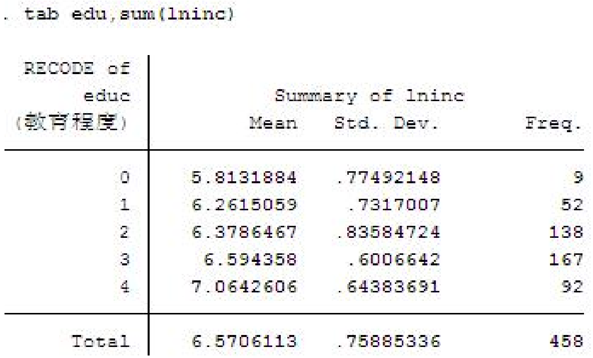
ONE-WAY ANOVA

Example 1.3

在2003年的CGSS数据中，对收入（取对数）进行方差分析，检验不同教育程度的收入水平是否存在差异。

- 首先，建立$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$和$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$
- 根据样本数据，计算F检验统计量
- stata 命令：anova lninc edu
- 结论：否定H_0，即认为不同教育程度的收入水平存在差异。

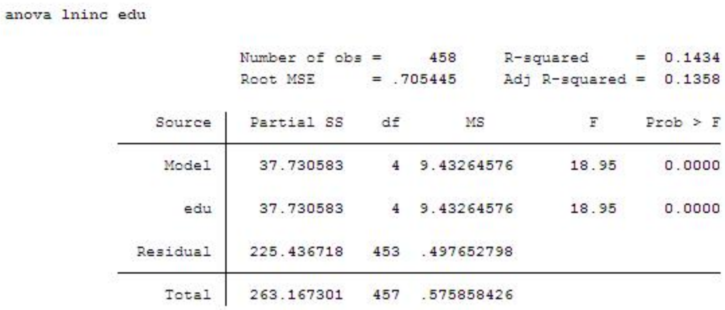
ONE-WAY ANOVA



```
. tab edu,sum(lninc)

 RECODE of  |
      educ  |        Summary of lninc
(教育程度)  |        Mean   Std. Dev.       Freq.
------------+------------------------------------
         0  |   5.8131884   .77492148           9
         1  |   6.2615059    .7317007          52
         2  |   6.3786467   .83584724         138
         3  |    6.594358    .6006642         167
         4  |   7.0642606   .64383691          92
------------+------------------------------------
     Total  |   6.5706113   .75885336         458
```

Figure 1.3:

One-way ANOVA



```
anova lninc edu

                         Number of obs =      458     R-squared     =  0.1434
                         Root MSE      =  .705445     Adj R-squared =  0.1358

               Source |  Partial SS    df       MS           F     Prob > F
           -----------+----------------------------------------------------
                Model |   37.730583     4  9.43264576      18.95     0.0000
                      |
                  edu |   37.730583     4  9.43264576      18.95     0.0000
                      |
             Residual |  225.436718   453  .497652798
           -----------+----------------------------------------------------
                Total |  263.167301   457  .575858426
```

Figure 1.4:

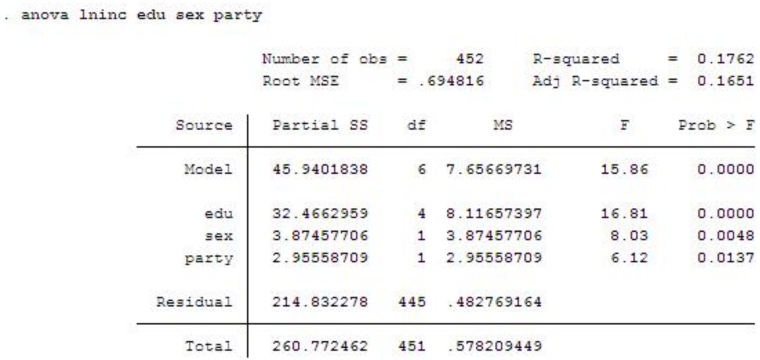
TWO-WAY OR MULTI-WAY ANOVA

Example 1.4

在*2003*年的*CGSS*数据中，对收入（取对数）进行方差分析，检验不同教育程度、不同性别、不同党员身份的收入水平是否存在差异。

- 根据样本数据，计算F检验统计量
- stata 命令：anova lninc edu sex party

Multi-way ANOVA



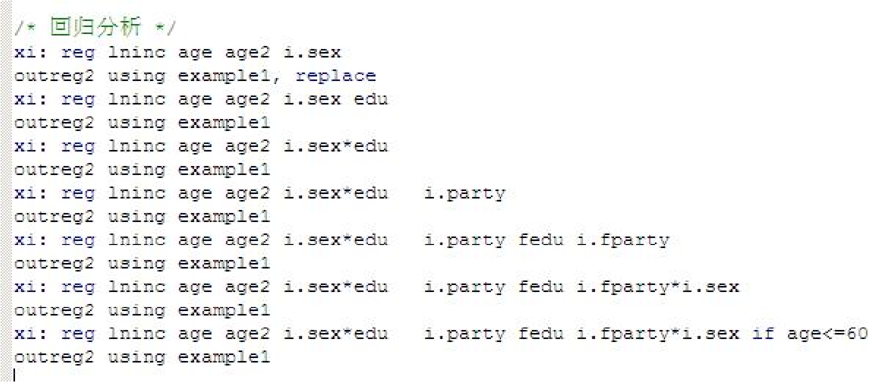
```
. anova lninc edu sex party

                           Number of obs =     452     R-squared     =  0.1762
                           Root MSE      = .694816     Adj R-squared =  0.1651

                  Source | Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  45.9401838     6  7.65669731      15.86     0.0000
                         |
                     edu |  32.4662959     4  8.11657397      16.81     0.0000
                     sex |  3.87457706     1  3.87457706       8.03     0.0048
                   party |  2.95558709     1  2.95558709       6.12     0.0137
                         |
                Residual |  214.832278   445  .482769164
              -----------+----------------------------------------------------
                   Total |  260.772462   451  .578209449
```

Figure 1.5:

COMPARISONS OF CONDITIONAL MEANS

1. Categorical independent variables can be easily handled in a regression model. We can create dummy variables.
2. Analysis fo variance (ANOVA) is statistically equivalent to regression analysis with dummy variable.

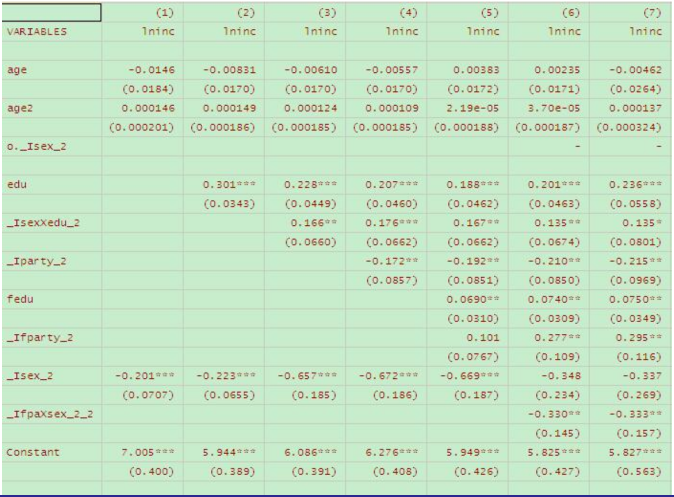
嵌套模型



```
/* 回归分析 */
xi: reg lninc age age2 i.sex
outreg2 using example1, replace
xi: reg lninc age age2 i.sex edu
outreg2 using example1
xi: reg lninc age age2 i.sex*edu
outreg2 using example1
xi: reg lninc age age2 i.sex*edu   i.party
outreg2 using example1
xi: reg lninc age age2 i.sex*edu   i.party fedu i.fparty
outreg2 using example1
xi: reg lninc age age2 i.sex*edu   i.party fedu i.fparty*i.sex
outreg2 using example1
xi: reg lninc age age2 i.sex*edu   i.party fedu i.fparty*i.sex if age<=60
outreg2 using example1
```

Figure 1.6:

嵌套模型



	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	lninc	lninc	lninc	lninc	lninc	lninc	lninc
age	-0.0146	-0.00831	-0.00610	-0.00557	0.00383	0.00235	-0.00462
	(0.0184)	(0.0170)	(0.0170)	(0.0170)	(0.0172)	(0.0171)	(0.0264)
age2	0.000146	0.000149	0.000124	0.000109	2.19e-05	3.70e-05	0.000137
	(0.000201)	(0.000186)	(0.000185)	(0.000185)	(0.000188)	(0.000187)	(0.000324)
o._Isex_2						-	-
edu		0.301***	0.228***	0.207***	0.188***	0.201***	0.236***
		(0.0343)	(0.0449)	(0.0460)	(0.0462)	(0.0463)	(0.0558)
_IsexXedu_2			0.166**	0.176***	0.167**	0.135**	0.135*
			(0.0660)	(0.0662)	(0.0662)	(0.0674)	(0.0801)
_Iparty_2				-0.172**	-0.192**	-0.210**	-0.215**
				(0.0857)	(0.0851)	(0.0850)	(0.0969)
fedu					0.0690**	0.0740**	0.0750**
					(0.0310)	(0.0309)	(0.0349)
_Ifparty_2					0.101	0.277**	0.295**
					(0.0767)	(0.109)	(0.116)
_Isex_2	-0.201***	-0.223***	-0.657***	-0.672***	-0.669***	-0.348	-0.337
	(0.0707)	(0.0655)	(0.185)	(0.186)	(0.187)	(0.234)	(0.269)
_IfpaXsex_2_2						-0.330**	-0.333**
						(0.145)	(0.157)
Constant	7.005***	5.944***	6.086***	6.276***	5.949***	5.825***	5.827***
	(0.400)	(0.389)	(0.391)	(0.408)	(0.426)	(0.427)	(0.563)

1. Although in principle we could design our models to best predict any population parameter(e.g. the median), in practice we commonly use the regression to denote the problem of predicting <u>conditional means</u>.
2. When the regression function is a linear combination of independent variables, we have so-called linear regression, which are widely used for continuous dependent variables.

A Typology of Regression Models

Dependent Variable	Independent	Method of Analysis
Continuous	Continuous	Regression, correlation
Continuous	Categorical	Regression, ANOVA
Binary	Categorical	Logit/probit, loglinear
Binary	Continuous	Logit/probit
Unordered polytomous	Categorical	Multinomial logit, loglinear
Unordered polytomous	Continuous	Multinomial logit
Ordered polytomous	Categorical	Order logit/probit, loglinear
Ordered polytomous	Continuous	Order logit/probit
Cross-classified data	Categorical	Loglinear
Censored duration data	Categorical, continuous	Loglinear, logit/comp. log-log

回归分析的目的是利用变量间的简单函数关系，用自变量对因变量进行“预测”，让“预测值”尽可能地接近因变量的“观测值”。

In regression analysis, the objective is to predict, as closely as possible, an array of observed values of the dependent variable based on a simple function of independent variables.

Obviously, predicted values from regression models are not exactly the same as observed ones. Characteristically, regression partitions an observation into two parts:

回归分析的特点就在于它把观测值分解成两部分：结构部分和扰动部分。

观测项=结构项+随机项

observed=structural + stochastic

1 观测部分代表因变量的实际取值；

2 结构项部分代表因变量和自变量之间的结构关系，表现为“预测值”；

The structural part denotes the relationship between the dependent and independent variable.

3 随机项部分表示观测项中未被结构项解释的剩余部分。

The stochastic part is the random component unexplained by the structural part.

描述性关系：观测项=概括项+残差项

Description: Observed=summary+residual

- The model serves to summarize the basic feature of data. The question is whether the model corresponds to the facts.
- The primary goal of statistical modeling is to summarize massive amounts of data with simple structures and few parameters.

  统计模型的主要目标在于用最简单的结构和尽可能少的参数来概括大量数据所包含的主要信息。

- For the ith observation, we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$$

$$= \sum \beta_k x_{ik} + \epsilon_i$$

$$= x_i' \beta + \epsilon_i$$

- y is decomposed into a linear function of x's with unknown parameters (β) and a residual term(ϵ_i).

1 Since ϵ_i is intrinsically unobservable, simplifying assumptions about the characteristics of ϵ_i are necessary.

2 A key assumption that yields the identification of the unknown parameters is the independence between ϵ and x variables.

3 Other assumptions are often invoked to improve efficiency. For example, it is common to assume ϵ_i to be independent of one another and identically distributed (i.i.d).

   (1) The independence assumption implies that the correlation in ϵ between a pair of observations is zero.

   (2) The identical distribution assumption assures a common variance of σ_ϵ^2 (i.e. homoscedasticity).

ESTIMATION

1. Least squares estimation
   (1) The goal of LS estimation is to find estimates of β's that make the sum of squared errors around the conditional mean as small as possible.
2. Maximum likelihood estimation
   (1) The principal aim of ML estimation is to find parameter values that maximize the sampling likelihood, L, which may be thought of as the formula for the joint probability distribution (or joint density) of the sample.

参数点估计的三个评价标准

1. unbiased: the expected value of an estimator equals the value of the true parameter being estimated
   (1) 无偏性：估计值的平均数恰好等于总体的参数值。
2. efficiency: the estimator with the smallest variance in its sampling distribution
   (1) 有效性：抽样分布方差最小的那个估计就是对总体参数最有效率的估计。
3. consistent：随着样本规模扩大，估计的偏差随之减小，或者说估计值以更大的概率趋近总体参数值。

GAUSS-MARKOV THEOREM

1. Mean zero errors: $E(\epsilon|x) = 0$
2. Homoscedasticity: $var(\epsilon|x) = \sigma^2$
3. Noncorrelated errors: $cov(\epsilon_i, \epsilon_j) = 0, i \neq j$
4. Exogeneity of explanatory variables: $cov(x_i, \epsilon) = 0$

The Gauss-Markov theorem states that OLS estimator is the best (means most efficient) among all linear unbiased estimators (BLUE).

TENSION BETWEEN ACCURACY AND PARSIMONY

1 By ”accuracy” we mean the ability to reproduce the data, measured by goodness-of-fit statistics.

(1) saturated model

2 By ”parsimony” we commonly mean statistical models with few parameters.

(1) $y_i = \beta_0 + \epsilon_i$

1 The objective in model searching is to find models that describe the essential characteristics of the data using as few parameters as possible.

2 Nested models provide a way to motivate model searching.

3 Two models are nested if one model is a special case of the other.

   (1) Model1: $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$

   (2) Model2: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

4 Model1 is said to be nested within Model2, as Model1 is a special case of Model2 with the constraint that $\beta_2 = 0$.

1. The key question is whether the unconstrained model fits significantly better than the constrained model.
2. We need to asses the improvement in fit.
3. For linear regressions, such assessments are conducted using F-tests based on reductions in residual sums of squares, or other proportionate reduction in error (PRE) criteria.
   (1) （削减误差比例）$PRE = \frac{TSS-RSS}{TSS}$
4. For the nonlinear models, most assessments are chi-squared based on different criteria, such as reductions in the log-likelihood-ratio statistics.

MLE

1 If we regard $f(x|\theta)$ as a function of θ for given observed data X, then $L(x|\theta) = f(x|\theta)$ is called a likelihood function.

2 The ML principles states that an admissible θ that maximizes likelihood function probability (discrete case) or density (continuous case), relative to alternative values of θ, provides the θ that is the most ”likely” to have generated the observed data X, given the assumed parametric form.