社会统计学及SPSS软件应用

Instructor:王荣欣

Email: rxwang@qq.com

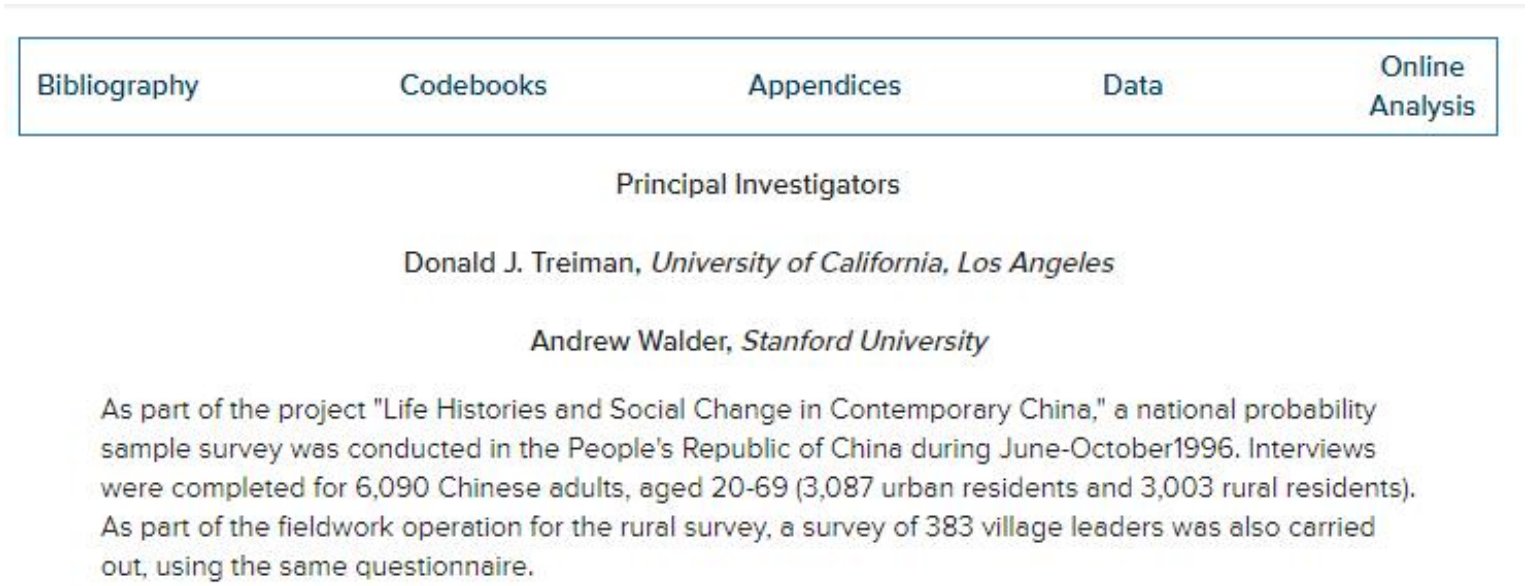
周一3-4节、单周周四3-4节，3A106-2

2020年10月22日

CONTENTS

CONTENTS

Life History and Social Change in Contemporary China, 1996.



Bibliography | Codebooks | Appendices | Data | Online Analysis

Principal Investigators

Donald J. Treiman, *University of California, Los Angeles*

Andrew Walder, *Stanford University*

As part of the project "Life Histories and Social Change in Contemporary China," a national probability sample survey was conducted in the People's Republic of China during June-October1996. Interviews were completed for 6,090 Chinese adults, aged 20-69 (3,087 urban residents and 3,003 rural residents). As part of the fieldwork operation for the rural survey, a survey of 383 village leaders was also carried out, using the same questionnaire.

（教材第385-386页）

https://www.library.ucla.edu/social-science-data-archives/life-histories-social-change-china#Data

因变量：Number of characters correctly identified

1 一万

2 姓名

3 粮食

4 函数

5 雕琢

6 肆虐

7 舛谬

8 耆耄

9 彳亍

Table 6.1. Means, standard deviations, and correlations among variables affecting knowledge of Chinese characters, employed Chinese adults age 20-69, 1996 (N = 4,802).

	V	E	EF	N	U	M	C
V: No. of characters correctly identified (of 10)[a]		.819	.372	.372	.331	.247	.495
E: Years of schooling			.397	.400	.341	.216	.514
EF: Father's years of schooling				.226	.307	.013	.574
N: Nonmanual occupation[b]					.368	.030	.327
U: Urban origin						.003	.399
M: Male							.030
C: Level of cultural capital (range 0-1)[c]							
Mean	3.60	6.47	3.07	.177	.180	.558	.227
Standard deviation	2.22	4.10	3.70	.381	.384	.497	.224

[a] The items, in increasing order of difficulty, are *yiwan* (ten thousand), *xingming* (full name), *liangshi* (grain), *hanshu* (function), *diaozhu* (carve), *sinue* (wreak havoc or wanton massacre), *chuanmiu* (erroneous), *qimao* (octogenarian), *chichu* (walk slowly), and *taotie* (glutton).

```
. reg wordsum educ_hiy feduc_y nm urban male if good==1 [aw=weight], beta
(sum of wgt is   5.0260e+03)
```

Source	SS	df	MS		
				Number of obs	= 4,802
				F(5, 4796)	= 2069.42
Model	16235.9964	5	3247.19928	Prob > F	= 0.0000
Residual	7525.57487	4,796	1.56913571	R-squared	= 0.6833
				Adj R-squared	= 0.6830
Total	23761.5713	4,801	4.94929625	Root MSE	= 1.2527

wordsum	Coef.	Std. Err.	t	P>|t|	Beta
educ_hiy	.4069005	.0053421	76.17	0.000	.7499139
feduc_y	.0296758	.0054404	5.45	0.000	.0494133
nm	.2457067	.0537581	4.57	0.000	.0421185
urban14	.2550815	.0529486	4.82	0.000	.0440488
male	.3695001	.0375141	9.85	0.000	.0824842
_cons	.5794814	.0372324	15.56	0.000	.

```
. reg wordsum educ_hiy feduc_y culcap nm urban male if good==1 [aw=weight], beta
(sum of wgt is   5.0260e+03)
```

Source	SS	df	MS		
				Number of obs	= 4,802
				F(6, 4795)	= 1757.48
Model	16334.0828	6	2722.34714	Prob > F	= 0.0000
Residual	7427.48845	4,795	1.54900698	R-squared	= 0.6874
				Adj R-squared	= 0.6870
Total	23761.5713	4,801	4.94929625	Root MSE	= 1.2446

wordsum	Coef.	Std. Err.	t	P>|t|	Beta
educ_hiy	.393455	.0055701	70.64	0.000	.725134
feduc_y	.0087609	.0060105	1.46	0.145	.0145877
culcap	.8655737	.1087743	7.96	0.000	.0872727
nm	.2113878	.053586	3.94	0.000	.0362356
urban14	.1765357	.0535259	3.30	0.001	.0304851
male	.3847544	.037322	10.31	0.000	.0858894
_cons	.546092	.03723	14.67	0.000	.

Table 6.2. Determinants of the number of Chinese characters correctly identified on a 10-item test, employed Chinese adults age 20-69, 1996 (standard errors in parentheses).[a]

Variable	Model 1	Model 2
Metric regression coefficients		
E: Years of schooling	.407	.393
	(.006)	(.006)
E_F: Father's years of schooling	.030	.009
	(.006)	(.007)
N: Nonmanual occupation	.246	.211
	(.057)	(.057)
U: Urban origin	.255	.177
	(.053)	(.054)
M: Male	.370	.385
	(.045)	(.044)
C: Level of cultural capital (range 0-1)		.866
		(.118)
Intercept	.579	.546
	(.039)	(.039)
R^2	.683	.687
s.e.e.	1.25	1.24

workers from urban origins, we have, for females

$$\begin{aligned}\hat{V} &= a + b(E) + c(\bar{E}_F) + d(N) + e(U) + f(M) + g(\bar{C}) \\ &= .546 + .393(E) + .009(3.07) + .211(1) + .177(1) + .385(0) + .866(.227) \\ &= 1.158 + .393(E)\end{aligned} \tag{6.6}$$

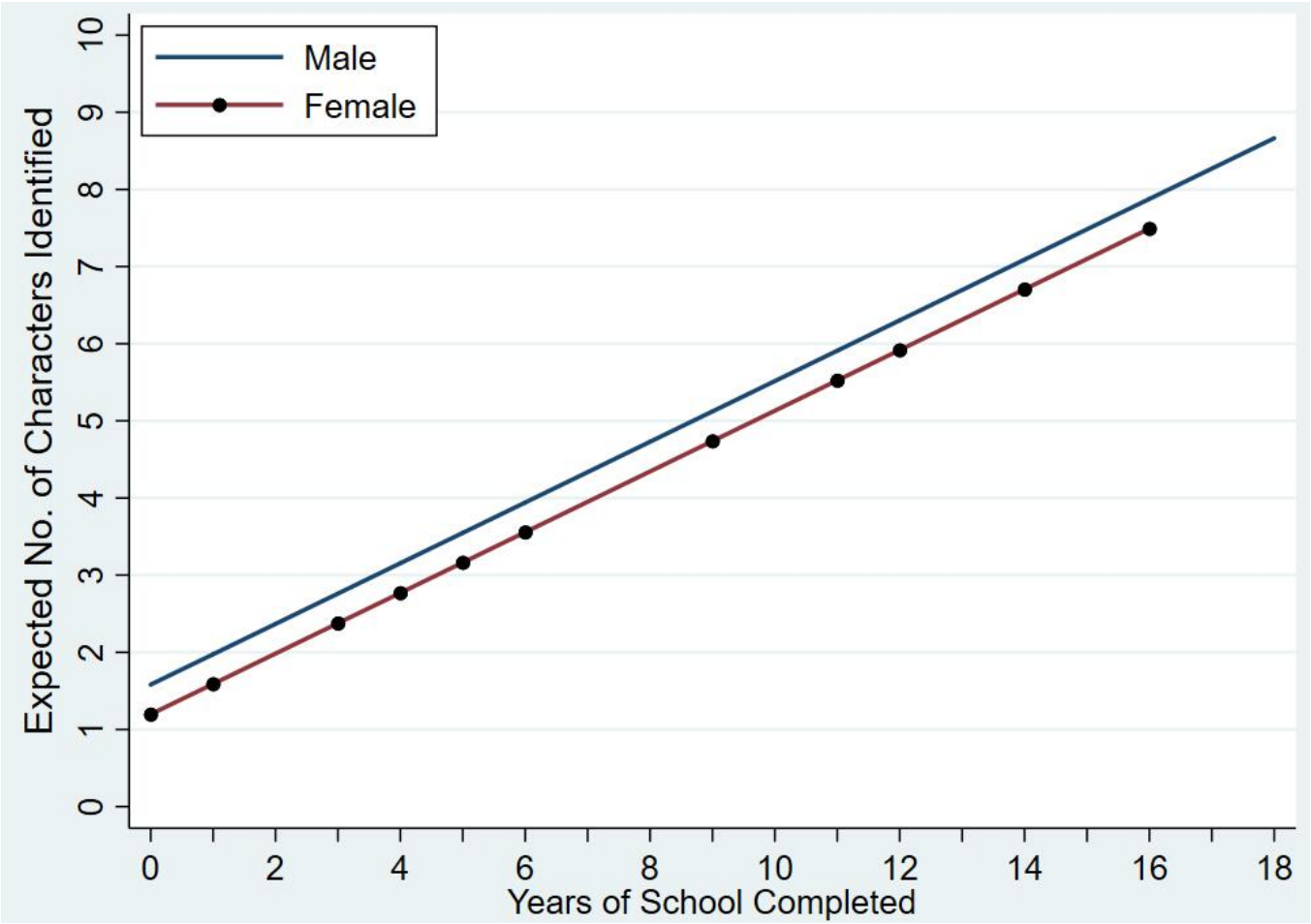
and for males

$$\begin{aligned}\hat{V} &= a + b(E) + c(\bar{E}_F) + d(N) + e(U) + f(M) + g(\bar{C}) \\ &= .546 + .393(E) + .009(3.07) + .211(1) + .177(1) + .385(1) + .866(.227) \\ &= 1.543 + .393(E)\end{aligned} \tag{6.7}$$

$$\begin{aligned}\hat{V} &= a + b(E) + c(\bar{E}_F) + d(\bar{N}) + e(\bar{U}) + f(M) + g(\bar{C}) \\ &= .546 + .393(E) + .009(3.07) + .211(.177) + .177(.180) + .385(0) + .866(.227) \\ &= .839 + .393(E)\end{aligned} \tag{6.8}$$

and for males is

$$\begin{aligned}\hat{V} &= a + b(E) + c(\bar{E}_F) + d(\bar{N}) + e(\bar{U}) + f(M) + g(\bar{C}) \\ &= .546 + .393(E) + .009(3.07) + .211(.177) + .177(.180) + .385(1) + .866(.227) \\ &= 1.224 + .393(E)\end{aligned} \tag{6.9}$$


Male
Female
Expected No. of Characters Identified
0
1
2
3
4
5
6
7
8
9
10
0
2
4
6
8
10
12
14
16
18
Years of School Completed

方程整体显著性检验

- 方程的整体显著性检验，旨在对模型中因变量与自变量之间的线性关系在总体上是否显著成立做出推断。
- 检验模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i$，i=1, 2, …, n中所有参数$\beta_j$都等于0。可提出如下假设：

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \beta_j \text{不全为} 0$$

教材第119-120页

- 根据数理统计学中的知识，在原假设 成立的条件下，统计量

$$F=\frac{ESS/k}{RSS/(n-k-1)}\sim F(k,n-k-1)$$

- 给定显著性水平α，可得到临界值$F_\alpha(k,n-k-1)$ 。如果$F<F_\alpha(k,n-k-1)$，则接受原假设，即该模型的所有回归系数都等于0，该模型是没有意义的；如果$F>F_\alpha(k,n-k-1)$则拒绝原假设，即认为模型是有意义的，但无法确认所有的回归系数是否都显著，这需要做进一步的检验，即需要对单个变量进行显著性检验。

ESS（回归平方和，explained sum of squares）
RSS（残差平方和，residual sum of squares）

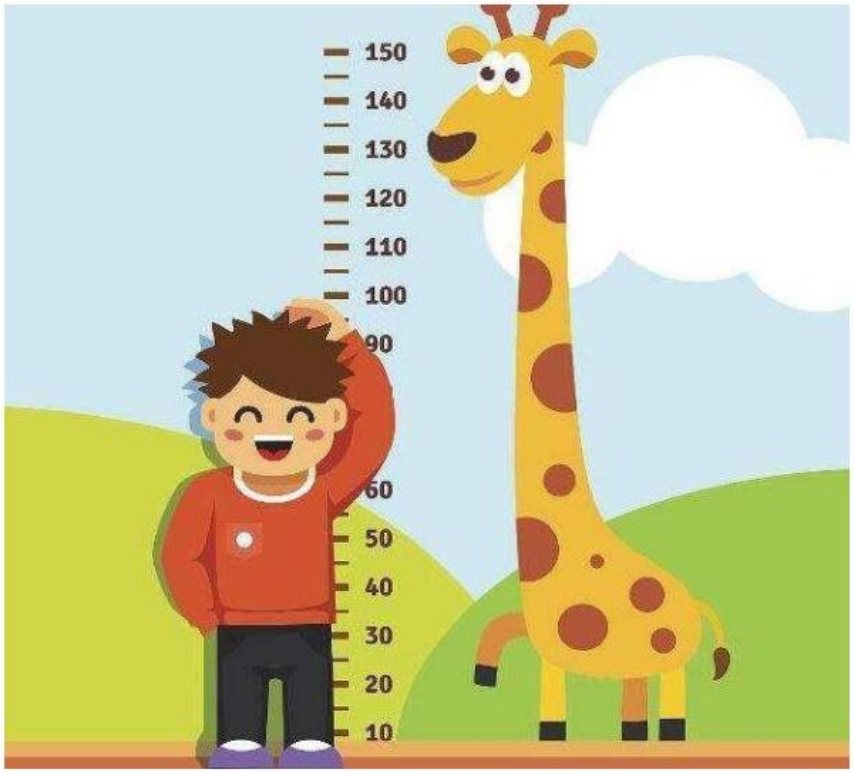
- **F检验的结果**
  - F检验的**原假设**：所有系数（不包括截距）都等于0.
  - 思考：这个原假设如果成立了，说明什么？我们希望这个原假设被拒绝吗？
  - 本案例中，F检验的p值小于0.05（显著性水平），因此可以拒绝原假设，模型整体是显著的（没白忙活）。

变量显著性检验

- 回归分析的目的之一是要判断X是否是Y的一个显著影响因素。这就需要进行变量的显著性检验。我们已经知道回归系数估计量 $\hat{\beta}_1$ 服从正态分布，即 $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2})$ 。又由于真实的σ^2 未知，利用它的无偏估计量$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ 替代时，可构造检验统计量

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2)$$

进行检验。



- 小朋友真实身高：参数 β
- 尺子测量结果：统计量 $\hat{\beta}$
- 尺子的精度： SE
- 尺子的精度估计： $\widehat{\mathrm{SE}}$

因此可以构造统计量:

$$H_0: \beta = \beta_0 \quad \text{v.s.} \quad H_1: \beta \neq \beta_0$$

$$t = \frac{\hat{\beta} - \beta_0}{\widehat{\mathrm{SE}}}$$

单个变量显著性检验

- 单个变量的显著性检验一般利用**t检验**。
- 可提出如下假设：$H_0: \beta_j = 0 (j = 1,2,\dots,k)$

  $H_1: \beta_j \neq 0$

  根据数理统计学中的知识，在原假设 成立的条件下，统计量

$$t = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \sim t(n-k-1)$$

- 给定显著性水平α，可得到临界值$t_{1-\alpha/2}(n-k-1)$ 。若$|t| < t_{1-\alpha/2}(n-k-1)$，则接受原假设，即该模型的$\beta_j = 0$，说明自变量$X_j$对因变量是没有影响的；若$|t| > t_{1-\alpha/2}(n-k-1)$ ，则拒绝原假设，即认为自变量X_j对因变量是有影响的。

教材第103页

加入过多解释变量可提高模型解释力，但会牺牲模型的简洁性（parsimony）。

1. 模型A：Y=beta1*X1+error
2. 模型B：Y=beta1*X1+beta2*X2+error。

这两个模型哪个更好？

模型A的优点是更加简单，表现在它需要的参数个数更少。但是，付出的代价是拟合优度差一些；模型B恰恰相反，拟合优度一定更好，但是更加复杂。

最优的模型一定要在：（1）拟合优度；（2）模型复杂度，二者之间寻求一个平衡。

1. 赤池信息准则（Akaike Information Criterion，简称 AIC）
2. 贝叶斯信息准则（Bayesian Information Criterion，简称 BIC）

命令 estat ic

$$AIC = n \times \ln\left(\frac{SSE}{n}\right) + 2 \times p,$$

$$BIC = n \times \ln\left(\frac{SSE}{n}\right) + \ln(n) \times p.$$

第一项是残差平方和的单调函数。
自变量个数增加，残差平方和降低，
模型的拟合优度变好，第一项减小。

第二项会随着模型复杂程度（自变量个数p）的增加而增大。AIC准则和BIC准则中的系数“2”和“ln(n)”分别叫做“惩罚”。

模型的拟合优度 **VS.** **模型的复杂程度**

1 向前回归：

先从空模型出发，然后从剩下的变量里面，挑选出最好的一个变量（能够让残差平方和最小），从而得到一个自变量大小为1的模型。

接下来，在给定当前模型的前提下，从剩下的变量里面再挑选一个最好的变量，这样得到一个自变量大小为2的模型。

重复这个过程，假设样本量充足，人们就获得了一共p个待选模型。

2 向后回归：从全模型出发，然后依次剔除不靠谱的变量，也能形成p个待选模型。

3 逐步回归

逐步回归的Stata演示

sysuse auto, clear

gen weight2 = weight*weight

1. stepwise, pr(.2): regress mpg weight weight2 displ gear turn headroom foreign price （后退式排除）
2. stepwise, pe(.2): regress mpg weight weight2 displ gear turn headroom foreign price （前进式排除）

- **思考：**直接将所有待选模型的AIC或BIC值计算出来，选择最小的那个即可，为何还要有“向前回归”“向后回归”等实施步骤？
- 计算量：待选模型的个数众多，量级是2^p，不会比较所有待选模型的AIC或BIC值
- 向前回归会产生一条“路径”：M_1，M_2，…，M_p
  - M_0：空模型
  - M_1：只包含一个自变量。此时有p个选择，挑选残差平方和最小的那个。
  - M_2：在M_1的基础上，再加入一个自变量。此时有p-1个选择，同样，挑选残差平方和最小的。
  - 以此类推，直到产生M_p