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

2020年12月13日

CONTENTS

CONTENTS

1. 多分类Logit模型需满足“无关选择的独立性”（Independenc of Irrelevant Alternative, IIA）的假定，才能保证模型的正确性。

2. 即：任意两个类别之间的相对发生几率比独立于模型中的其他选择，即第m类结果与第n类结果的发生几率比与其他可能的结果之间彼此独立。
3. 即我们的选择不应受到那些不相关的可选方案的影响。

1. The assumption of independence from irrelevant alternatives, or IIA.
2. IIA holds that the ratio of the choice probabilities of any two alternatives for a particular observation is not influenced systematically by any other alternatives.
3. The red/blue bus paradox.

1. 原有两种交通方案：自驾车（car）和红色巴士（red bus），现在加上“蓝色巴士”（blue bus）的选择之后，将会使选择乘“红色巴士”的概率降低一半。

2. 在引入蓝色巴士后，各种方案的条件概率不再相等，且方案选择彼此不再独立，即违背了IIA假定。

1 Stata中的似不相关（seemingly unrelated estimation）命令“suest”，可以检验IIA假定是否成立。

2 Hausman检验

(1) findit sg155 /*寻找下载地址*/
(2) net install sg155 /*下载安装命令mlogtest*/
(3) mlogit y x1 x2 x3, base(#)
(4) mlogtest, hausman base
/*base表示在检验中包括“去掉参照方案，而以剩余方案中观测值最多的方案为参照方案”*/
(5) 判断是否拒绝IIA的原假设。（p值>0.05，不能拒绝原假设）

违背IIA假定的应对

1 序列Logit模型（Sequential Logit Models）
   (1) 考察受访者是自驾车或坐红色巴士，还是两个都不选择
   (2) 考察那些自驾车及坐红色巴士的两组受访者，他们是否还会分别选择另外的一种交通工具。
2 多分类Probit模型
3 嵌套Logit模型
   (1) 嵌套数据结构是指人们的选择可以成组出现。例如，考察学生对上大学的选择过程。
   (2) 首先要选择的是上三年制大专，还是四年制大学（记为层一水平）。
   (3) 然后再对每一组决定是上公立大学，还是私立大学（记为层二水平）。

1. mlogit y x1 x2 x3, base(#)
   （多分类Logit模型，base(#)用于指定参照组）
2. mlogit y x1 x2 x3, rrr base(#)
   （rrr表示汇报Relative Risk Ratio，即汇报e^{β}，而非β）
3. listcoef
   （listing coefficients，列出回归模型估计的系数）
4. fitstat （拟合优度）
5. lrtest （似然比检验）

STATA软件操作

- 数据清理
- Logistic回归（以论文为例）

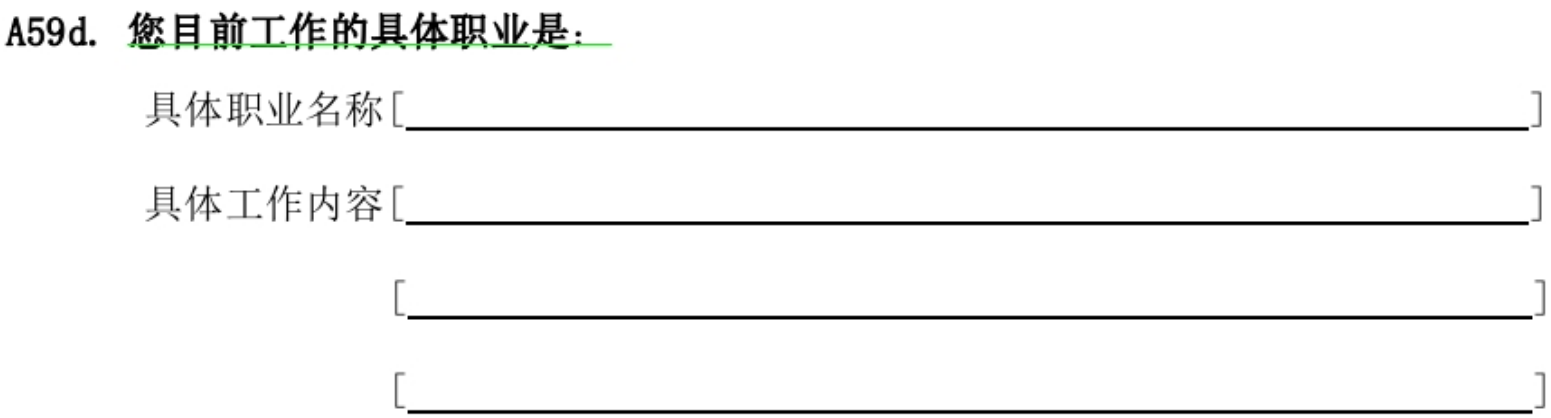

Figure 3.1: CGSS2013

A59j. 您目前工作的单位或公司的单位类型是：

党政机关 ..... 1 → **跳问 A59L**
企业 2
事业单位 3
社会团体、居/村委会 4
无单位/自雇（包括个体户） ..... 5 → **跳问 A61（第 16 页）**
军队 ..... 6 → **跳问 A61（第 16 页）**
其他（请注明：_____） 7

Figure 3.2: CGSS2013

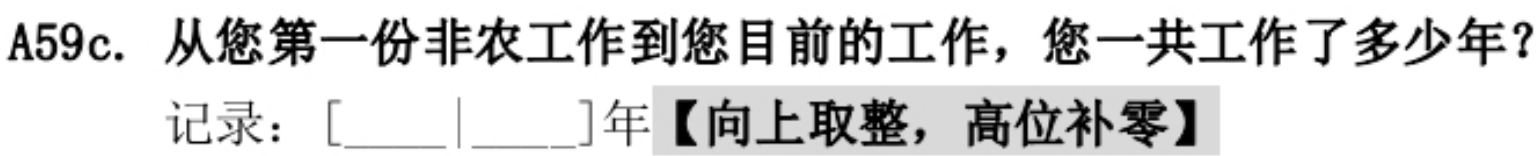

Figure 3.3: CGSS2013

然后以创意专家为参照组，进行多项mlogit回归分析，可以得出“超级创意核心VS创意专家”的情况。再更换参照组进行分析，以非创意阶层为参照组，可以得出“超级创意核心VS非创意阶层”“创意专家VS非创意阶层”的情况。

表4　职业阶层的多项mlogit模型

	模型4 超级创意核心VS创意专家	模型5 创意专家VS非创意阶层	模型6 超级创意核心VS非创意阶层
	偶值比（eb）	偶值比（eb）	偶值比（eb）
男性	0.933（0.154）	1.150（0.108）	1.073（0.164）
年龄	0.953（0.0655）	1.026（0.0386）	0.977（0.0628）
年龄平方/100	1.001（0.000709）	1.000（0.000383）	1.000（0.000667）
非农户口	1.294（0.312）	0.846（0.0972）	1.095（0.249）
教育年限	1.281***（0.0505）	1.196***（0.0227）	1.533***（0.0577）
婚姻状态			
未婚（参照组）			
已婚	1.337（0.342）	0.910（0.141）	1.217（0.284）
离婚	0.465（0.314）	0.915（0.244）	0.426（0.279）

1. 二项定类Logistic回归，是对“非此即彼”的回归。因变量的“变”是受关注的情况是否发生，例如是否当上经理。
2. 如果因变量是定序变量，有几个等级，就可以做定序Logistic回归。
   (1) 又称为累积logit模型（cumulative logit model）和比例发生比模型（proportional odds model）。

1. 一个人的健康状况是定序的例子。健康状况非常好比健康状况良好更好，后者又比健康状况差更好。
2. Likert Scale对态度问题的回答选项。

3. 个人的宗教信仰为非定序的例子（不可以进行排序）。

1 定序Logistic回归的一个关键假设是平行斜率假设（parallel regression assumption）和比例发生比假定（proportional odds assumption）。
   (1) 该假定是指自变量对每一个累积对数发生比（cumulative logits odds）的影响都相同。

2 如果一个变量影响了定序类别中的某一个结果（如饮食对健康状况的影响），就假定这个变量与结果间相关联的系数对所有结果是一样的。

1. 例如，军队的校官分为少校、中校和上校，这是一个定序变量。我们以服役时间为自变量，分析服役时间每增加一年对于官阶的影响。
2. 如果影响一致，就可以通过平行回归检验（test of parallel lines）。

1. 饮食对一个人处于健康状况非常好的可能性的影响程度，与饮食对一个人处于健康状况差的可能性的影响程度是一模一样的。
2. 考查年龄对英语掌握程度的影响时，假定青少年与老年人每增加1岁，他们英语掌握程度向赋值渐高方向（英语程度逐渐变好）的斜率是一致。
3. 考查教育年限对英语掌握程度的影响时，假定在小学阶段和研究生阶段是一致的。

1 定序回归系数意味着：

(1) 当自变量发生一个单位变化时，对于因变量取值从第一层级变为第二层级的影响；
(2) 对于因变量取值从第二层级变为第三层级的影响...
(3) 如果能够通过平行回归假定，意味着上述的影响相同。

1. 第一步把累积概率转变成累积odds；
2. 第二步取累积odds的自然对数。

```
1 ologit y x1 x2 x3
2 findit omodel  /*寻找下载地址*/
3 omodel y x1 x2 x3
   /*零假设：定序回归模型符合平行回归的假定*/
4 brant, detail   /*brant test*/
   /*（若p值<0.05，拒绝原假设，意味着平行回归的假定
   不满足）*/
```

参考文献

1 李连江，2017，《戏说统计：文科生的量化方法》，北京：中国政法大学出版社。
2 李连江，2019，《戏说统计续编：文科生的量化操作指南》，北京：当代世界出版社。