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

概率预测

1 estat classification （计算预测准确的百分比）

(1) 若发生概率的预测值$\hat{y} \geq 0.5$，则认为其预测y=1；

(2) 反之，认为其预测y=0。

(3) 再将预测值与实际值（样本数据）进行比较，就能计算出准确预测的百分比。

2 ROC curve（Receiver Operating Characteristic Curve）and AUC (Area Under Curve)

(1) 1-FPR(False Positive Rate): specificity

(2) TPR (True Positive Rate): sensitivity

	实际值	
预测值	true positive $(\hat{y}=1, y=1)$	false positive $(\hat{y}=1, y=0)$
	false negative $(\hat{y}=0, y=1)$	true negative $(\hat{y}=0, y=0)$

sensitivity=$\frac{TP}{TP+FN}$

specificity=$\frac{TN}{FP+TN}$

个体面临的选择为y=1,2,...,J种。

Examples of unordered responses:

1. choice among consumer products
2. occupations
3. academic majors
4. religions
5. modes of transportation
6. political candidates

1 多项定类Logistic回归（multinomial logit regression）。这个回归的分析结果，就是一系列二项定类Logistic回归（binary logit regression）。

(1) 二项定类Logistic回归有一个0（参照类），一个1（关注类）。
(2) 多项定类Logistic回归有一个0（参照类），两个或多个1（关注类）。

1. In the binary case, the comparison is between category 1 and category 2 (or the first vs. the last category).
2. In the multinomial case, the comparison is between category j and J (or any category but the last versus the last).

1. $log\frac{p_1}{p_J} = \sum_{i=1}^{i=k} b_{i1}x_i$
2. $log\frac{p_2}{p_J} = \sum b_{i2}x_i$
3. ...
4. $log\frac{p_{J-1}}{p_J} = \sum b_{i(J-1)}x_i$

当J=2

1. $log\frac{p_1}{p_J} = \sum b_{i1}x_i$
2. 当J=2时，以第一类（y=1）作为参照类，估计一组与y=2相对应的参数。
3. 二项定类Logistic回归是多项定类Logistic回归的特例。
   (1) 二项定类Logistic回归的因变量编码为（0,1）
   (2) 多项定类Logistic回归的因变量编码为（1,2）

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 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）
2. 多分类Probit模型
3. 嵌套Logit模型

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）

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）。
3. 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。
4. 如果影响一致，就可以通过平行回归检验（test of parallel lines）。

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 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。
3 王存同，2017，《进阶回归分析》，北京：高等教育出版社。