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7日

Contents

CONTENTS

1 $odds = \frac{p}{1-p}$
  (1) $\frac{p}{1-p}$称为几率（odds）或相对风险（relative risk）。
  (2) 若y=1表示生，y=0表示死，odds为2，则意味着存活的概率是死亡概率的两倍，故存活概率为2/3，死亡概率为1/3。
2 取它的自然对数，即log odds。
  (1) 由此形成Logit模型，其基本方程为：
  (2) $log(\frac{p}{1-p}) = \alpha + \beta_1 x_1 + \beta_2 x_2 ... + \beta_p x_p$

https://www.bettingodds.com/football/live-scores

解释LOGISTIC回归系数

1. 按odds来解释logistic回归系数
   (1) X变化一个单位，导致log(odds)变化b个单位。
2. 按odds ratio来解释logistic回归系数
   (1) X变化一个单位，odds变化 e^b 个单位。
3. 按概率来解释logistic回归系数
4. 按边际效应解释

$$\frac{\dfrac{p^*}{1-p^*}}{\dfrac{p}{1-p}} = \frac{\exp[\beta_1 + \beta_2 x_2 + \cdots + \beta_j (x_j + 1) + \cdots + \beta_K x_K]}{\exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_K x_K)} = \exp(\beta_j)$$

$\exp(\hat{\beta}_j)$表示变量x_j增加一单位引起几率比的变化倍数。

Stata 也称$\exp(\hat{\beta}_j)$为几率比(odds ratio)。

$$logit(\hat{p}) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 * sex$$

$$OR = \frac{odds(male)}{odds(female)} = \frac{\frac{p}{1-p}}{\frac{q}{1-q}} = \frac{e^{(\beta_0+\beta_1)}}{e^{\beta_0}} = \frac{e^{\beta_0}e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

1 e^{β_1}表示关注组的odds与参照组odds之间的倍数关系。
   (1) $e^{\beta_1} > 1$，意味着男性录取的odds大于女性。
   (2) $e^{\beta_1} = 1$，意味着男性与女性具有相同的录取odds。
   (3) $e^{\beta_1} < 1$，意味着男性录取的odds小于女性。

2 若自变量为连续变量，e^{β_1}表示该自变量变化一个单位，带来的odds的倍数变化。

边际效应

1. 边际效应是指物品的后一单位比前一单位的效用。
2. 若后一单位的效用比前一单位的效用小，是边际效用递减。
3. 对于非线性模型而言，边际效应本身不是常数，它随着自变量x的变化而变化。
4. x对Pr(y=1)的边际效应，会随着x的变化而变化。
5. 当Pr(y=1)=0.5时，logit模型和probit模型取得最大的边际效应。

边际效应

1 平均边际效应（average marginal effect）

(1) margins, dydx(*) （计算所有自变量的平均边际效应）

2 样本均值处的边际效应（marginal effect at mean）

(1) margins, dydx(*) atmeans （计算所有自变量在样本均值处的边际效应）

3 某个代表值处的边际效应（marginal effect at a representative value）

(1) margins, dydx(*) at (x1=0) （计算所有自变量在$x_1 = 0$的边际效应）

概率预测

1. logit x1 x2 x3 x4 x5 （logit模型）
2. predict prob （计算发生概率的预测值）
3. list prob y if x1==0 & x2==0 （列出在给定条件下y=1的概率预测值）
4. predict pmale if sex==1（男性的概率）
5. predict pfmale if sex==0（女性的概率）
6. sum pmale pfemale

Example 1

以数据集 titanic.dta 为例。该数据集包括泰坦尼克号乘客的存活数据。

变量	变量值	变量值
survive	存活=1	死亡=0
child	儿童=1	成年=0
female	女性=1	男性=0
class1	头等舱=1	其他=0
class4	船员=1	其他=0

1. sum [fweight=freq] （描述统计）
2. sum survive if child==1 [fweight=freq] /*小孩的存活率*/
3. sum survive if female==1 [fweight=freq] /*女士的存活率*/
4. sum survive if class1==1 [fweight=freq] /*头等舱旅客的存活率*/
5. logit survive class1 class2 class3 child female [fweight=freq] /*logit模型，以重复次数（freq）作为权重*/
6. logit survive class1 class2 class3 child female [fweight=freq], or /*logit模型，显示odds ratio*/

1. margins, dydx(*) （计算所有自变量的平均边际效应）
2. margins, dydx(*) atmeans （计算在均值处的边际效应）
3. estat classification /* 计算预测准确的百分比*/
4. predict prob
   /*计算发生概率的预测值，预测每位乘客的存活概率*/
5. list prob survive freq if class1==1 & child==0 & female==1
   /*头等舱、成年、女士的存活率*/
6. list prob survive freq if class3==1 & child==0 & female==0
   /*三等舱、成年、男性乘客的存活概率*/

1. 对广义线性模型的估计，一般采用MLE。
2. MLE（Maximum Likelihood Estimation）最早由Ronald Fisher在1912年至1922年间开始使用。
3. 通过最大似然函数，就可以得出“最大似然估计”。

1. 该估计方法使用概率模型，其目标是寻找能够以较高的概率产生观察数据的系统发生数。
2. 当从模型总体随机抽取n组样本观测值后，最合理的参数估计量应该使得从模型中抽取该n组样本观测值的概率最大，而不像OLS旨在得到使模型更好地拟合样本数据的参数估计量。

最大似然估计

给定样本取值后，该样本最有可能来自参数θ为何值的总体。

1. 先确定一个函数来说明未知函数的似然函数（likelihood function）
2. 再找出此未知参数的观测值，使此似然函数达到最大值。

最大似然估计

1. The likelihood function is defined to be the probability density function treated as a function θ.
2. Estimate the unknown parameter θ by the value which makes the observed data most likely.
3. The value of θ at which the likelihood function achieves its maximum is called the maximum likelihood estimate, or MLE.

1 似然度（likelihood）是过去发生的概率，指的是在特定分布下出现的概率，简单来说，就是某件事在在限定的大背景下发生的概率。

2 likelihood（似然性）：现实世界在过去发生的概率。

整体模型的评价

1. -2 log likelihood
   (1) 这个指标值越小，说明模型的拟合程度越好。
2. 方程的拟合优度：pseudo R^2 by McFadden(1974)

$$pseudo\ R^2 = \frac{lnL_0 - lnL_1}{lnL_0}$$

3. 正确预测的百分比（percent correctly predicted）

$$pseudo\ R^2 = \frac{lnL_0 - lnL_1}{lnL_0} = \frac{lnL_1 - lnL_0}{lnL_{max} - lnL_0}$$

1. lnL_0是以常数项为唯一解释变量的对数似然函数的最大值。
2. lnL_1是原模型的对数似然函数的最大值。
3. 由于y为离散的两点分布，似然函数的最大可能值为1，故对数似然值的最大可能值为0，记为lnL_{max}。
4. $pseudo\ R^2$ 可以视作对数似然函数的实际增加值（$lnL_1 - lnL_0$）占最大可能增加值（$lnL_{max} - lnL_0$）的比重。

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

模型检验

1 嵌套模型（likelihood ratio test）：看两模型是否有显著统计差异。若$p < 0.05$，则选择大模型。

  (1) logit x1 x2 x3 x4 x5
  (2) estimates store m1 /*将第一次回归结果储存到m1*/
  (3) logit x1 x2 x3 x4
  (4) est store m2 /*将第二次回归结果储存到m2*/
  (5) lrtest m1 m2

2 非嵌套模型：比较AIC和BIC

  (1) estat ic

统计检验

1. Wald检验：对模型中自变量参数估计值的统计检验
   (1) 模型全局检验（global test）。如果Waldχ^2检验所对应的P值小于0.05，说明模型的设定有一定的意义，其结果可以推断到总体，且模型中至少有一个自变量呈现显著统计学差异。
2. Likelihood-ratio 检验
   (1) $\chi^2 = -2(logL_0 - logL_a)$
   (2) ln0.0001=-0.921，ln0.5=-0.69，ln0.9999=-0.0001，乘以-2以后都变成了正数。-2 loglikelihood的分布与卡方值的分布相似。
   (3) 卡方值分布表说明，任意一个卡方值在某个自由度下发生的概率。

1. 检验零假设的指标：-2 loglikehood （负二倍）
2. 用-2 loglikehood作为指标，参考卡方值的分布，可以算出一个概率，即当初始模型为真时，现实世界发生的概率。
3. 如果这概率极小，就可以放弃关于初始模型的零假设。
4. 不断修改模型，逐渐减小-2 loglikelihood，直到不能“显著”减少，找到最合适的模型，达到最大似然。

最大似然估计的逻辑

1. 初始模型。
   (1) 零假设：不管自变量怎么变，因变量发生的概率都不受影响。
2. 如果初始模型与现实世界有显著差异，就放弃初始模型。
   (1) 判断两者是否有显著差距，借助于类似卡方检验的检验。

最大似然估计涉及两种零假设。

1. 变量之间关系的零假设。
2. 模型与数据之间差距的零假设。
   (1) 预测与观察差距越大，-2 loglikelihood越大，模型越不准确。
   (2) 对应于一个概率。若概率很小，我们放弃零假设（零假设是指初始估计与数据完美契合），提出非零的假设。
   (3) 继续改进模型。

PREDICTING PREVALENCE OF ARMED THREATS

Example 2

以数据集 gunx.dta 为例。

```
. svy:logit gun male educ year black if good==1
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =        15                Number of obs     =     19,260
Number of PSUs     =     1,404                Population size   = 19,272.123
                                              Design df         =      1,389
                                              F(   4,    1386)  =     246.21
                                              Prob > F          =     0.0000

------------------------------------------------------------------------------
             |             Linearized
         gun |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.423456   .0465642    30.57   0.000     1.332112      1.5148
        educ |  -.0185324   .0065632    -2.82   0.005    -.0314073   -.0056575
        year |   .0101219   .0038004     2.66   0.008     .0026667     .017577
       black |   .4463046   .0687437     6.49   0.000      .311452    .5811573
       _cons |  -2.894143   .3178486    -9.11   0.000    -3.517658   -2.270628
------------------------------------------------------------------------------
```

Figure 5.1: 教材表13-3

```
. svy:logit gun male educ year black if good==1, or
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =        15                  Number of obs      =     19,260
Number of PSUs     =     1,404                  Population size    = 19,272.123
                                                Design df          =      1,389
                                                F(    4,    1386)  =     246.21
                                                Prob > F           =     0.0000
```

gun	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	4.151443	.1933085	30.57	0.000	3.789039	4.54851
educ	.9816383	.0064427	-2.82	0.005	.9690808	.9943585
year	1.010173	.003839	2.66	0.008	1.00267	1.017732
black	1.562527	.1074139	6.49	0.000	1.365406	1.788107
_cons	.0553464	.0175918	-9.11	0.000	.0296688	.1032473

Note: _cons estimates baseline odds.

Figure 5.2: 教材表13-3

1. In model 2, the expected odds of males being threatened by a gun are 4.15 (=$e^{1.4235}$) times greater than the odds of females being threatened, holding constant race, education, and the year of the survey.
2. The odds of having been threatened increase by 1.0102 each year, net of sex, race, and education.
3. The expected net odds of having being threatened in 1994 are about 25 percent larger than in 1973 (=$e^{0.0102(1994-1973)}$ 1.2363).

1. Net of other factors, the expected odds of males ever having being threatened are four times greater than for females.
2. The odds of having been threatened decline slightly with increasing education.
3. The odds of Blacks having been threatened in any given year are more than 1.5 times as great as for non-Blacks of the same sex with the same amount of education (=$e^{0.4463} = 1.56$).

```
. svy:logit gun male educ year black blkmale if good==1
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =        15              Number of obs     =     19,260
Number of PSUs     =     1,404              Population size   = 19,272.123
                                            Design df         =      1,389
                                            F(   5,    1385)  =     194.10
                                            Prob > F          =     0.0000
```

gun	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	1.454341	.0507311	28.67	0.000	1.354823	1.553859
educ	-.0191231	.006547	-2.92	0.004	-.0319661	-.0062801
year	.0100748	.0037996	2.65	0.008	.0026212	.0175284
black	.5689946	.1007161	5.65	0.000	.3714225	.7665667
blkmale	-.2125267	.125905	-1.69	0.092	-.4595112	.0344577
_cons	-2.903652	.3178482	-9.14	0.000	-3.527167	-2.280138

Figure 5.3: 教材表13-3

```
. svy:logit gun male educ year black blkmale if good==1, or
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =        15              Number of obs      =     19,260
Number of PSUs     =     1,404              Population size    = 19,272.123
                                            Design df          =      1,389
                                            F(   5,    1385)   =     194.10
                                            Prob > F           =     0.0000
```

gun	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	4.281661	.2172133	28.67	0.000	3.876076	4.729686
educ	.9810586	.006423	-2.92	0.004	.9685394	.9937396
year	1.010126	.0038381	2.65	0.008	1.002625	1.017683
black	1.76649	.177914	5.65	0.000	1.449796	2.152364
blkmale	.8085387	.1017991	-1.69	0.092	.6315923	1.035058
_cons	.0548226	.0174253	-9.14	0.000	.0293881	.1022701

Note: _cons estimates baseline odds.

Figure 5.4: 教材表13-3

1. In model 4, the odds of non-Black males having being threatened are 4.3 times as large as the odds of non-Black females having been threatened.
2. The odds of Black males having been threatened are about 3.5 times as large as the odds of females having been threatened.

参考文献

1 陈强，2014，《高级计量经济学及Stata应用（第二版）》，北京：高等教育出版社。
2 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。
3 王存同，2017，《进阶回归分析》，北京：高等教育出版社。