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3日

CONTENTS

1 对数和指数的代数基础

2 线性概率模型的局限

3 logistic转换

(1) 概率

(2) odds

(3) log odds

CONTENTS

1. $b^n = X$（以b为底，以n为指数）
2. n就是X的对数，$b^{logX} = X$
3. 常用对数以10为底，写成log X。

X>0	log X
(0,1)	负数
1	0
10	1
100	2
1000	3
10000	4
100000	5

1 自然对数就是以e为底的对数，e ≈ 2.718，写成ln X。

2 以2.718为底的对数称为自然对数，因为这个对数的变化能描述自然现象的增长或衰退的速度。

$$e^{ln(X)} = X$$

$$ln(X \times Y) = ln(X) + ln(Y)$$

$$e^{X+Y} = e^{X} \times e^{Y}$$

$$ln(X^{P}) = P \times ln(X)$$

回归分析

1. 普通线性回归：因变量Y必须是连续型数据
2. 0-1回归： 因变量Y是0-1型数据（binary data）
3. 定序回归：因变量Y是定序数据（ordered data）
4. 计数回归：因变量Y是计数的数据（非负的整数）（count data）
5. 生存数据回归：因变量Y是生存数据（教材第302页例三）

客户流失：Y = 流失与否

征信：Y = 是否逾期

购买决策：Y = 是否购买

客户流失：X =

在线时长
活跃程度
朋友个数
…

征信：X =

消费记录
工作背景
教育程度
…

购买决策：X =

性别
购买记录
产品特征
…

线性概率模型

1. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i = \beta_0 + x_i' \beta + \epsilon_i$（i=1,...,n）
2. 其中，自变量和参数分别为：

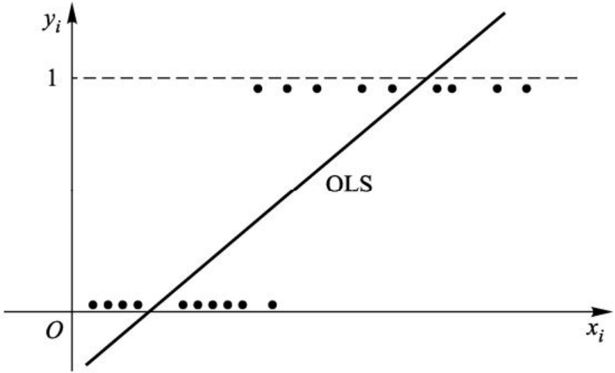
$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \ldots \\ x_{ik} \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \ldots & \beta_k \end{pmatrix}$$

线性概率模型的局限

1. 概率在0至1之间，线性回归方程不能做到这点。
2. 线性概率假定，概率随着自变量变化而发生线性变化。
   - 对于年薪1万元、10万元、100万元的三个人，当政府给购车者提供5000元补贴时，他们各自购买汽车的概率会以同样的幅度增加吗？

$$Y = \beta_0 + \beta_1 X_1 + +\beta_2 X_2 + ... + \beta_k X_k + \epsilon$$

1. 回归系数（coefficient）β_i是直线的斜率，它表示一个单位的x值变化所引起y值变化的大小。
   (1) 回归系数表示边际效应（marginal effect）。
2. ϵ是随机误差项（randomized error term）。
   (1) $E(\epsilon) = 0$
   (2) $\sigma_\epsilon^2 = \sigma^2$
   (3) $Cov(\epsilon_i, \epsilon_i') = 0$


y_i
1
OLS
O
x_i

继续使用线性回归？

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y = 0 或 1，线性回归仍然可以求解最小二乘估计
- **问题**：估计出来的Y不是 0 或 1！
- 解决办法：选取一个阈值c，
  - 如果Y的预测值大于c，将Y预测成1
  - 如果Y的预测值小于c，将Y预测成0
- 虽然不正确，但是这种方法在实践中经常被使用！
- **思考**：你觉得可以怎么办？

在线性回归的基础上充分考虑其特定的概率分布、连结函数等，并对传统的线性回归进行改造或扩展，从而形成广义线性模型（Generalized Linear Models）。

广义线性模型包括三个要素:

1 随机要素 (random component)
 (1) 因变量Y的特定概率分布形态

2 系统要素 (systematic component)
 (1) $\beta_0 + \beta_1 X_1 + +\beta_2 X_2 + ... + \beta_k X_k$

3 连接函数 (link function)
 (1) OLS回归: $g(\mu) = \mu$
 (2) Logit回归: $g(\mu) = log[\mu/(1-\mu)]$
 (3) 泊松回归: $g(\mu) = log(\mu)$

1. Modern developments of the logit and probit model were developed in the filed of bioassay or dose-response methodology.
2. Binomial response models can be motivated by considering an experiment in which different amounts of a drug or other chemical compound are applied to batches of experimental subjects.
3. The purpose of the experiment is to determine the lethal dosage levels (or response rates) or levels at which we would expect a certain proportion of the population to respond (by dying) to a given dosage level.

线性回归不适用，逻辑回归了解一下？

- 逻辑回归对 Y=1 的概率进行建模
- P(Y=1) 是0至1之间的数
- 这个函数叫做Logistic函数，这个回归模型又叫做**逻辑回归** ➡ $$P(Y=1)=\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$$
- 如果用其他形式建模，如：标准正态分布的累积分布函数，就会得到**Probit回归** ➡ $$P(Y=1)=\Phi(\beta_0+\beta_1 X)$$

1. $odds = \frac{p}{1-p}$
   (1) odds可以理解为事件发生的可能性与未发生可能性的比值。
   (2) odds来源于流行病学和人口学中用于评价死亡和患病等事件的相对风险，后来被社会科学借用。
2. $p = \frac{odds}{1+odds}$
3. odds ratio

1 转换因变量，使得估计的Y值在接近上限和下限时，X对Y的影响变小。
   (1) 首先取p与1-p之比，称为odds。
   (2) 其次，取它的自然对数，即log odds。

2 由此形成Logit模型，其基本方程为：
   (1) $log(\frac{p}{1-p}) = a + bX$
   (2) 其中，$log(\frac{p}{1-p})$称为链接函数（link function）。

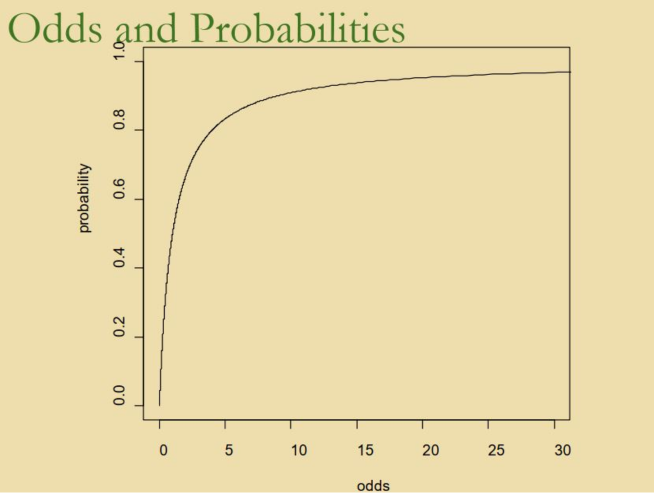
3 假定事件发生的概率与自变量的关系服从logistic函数分布。

什么是ODDS？

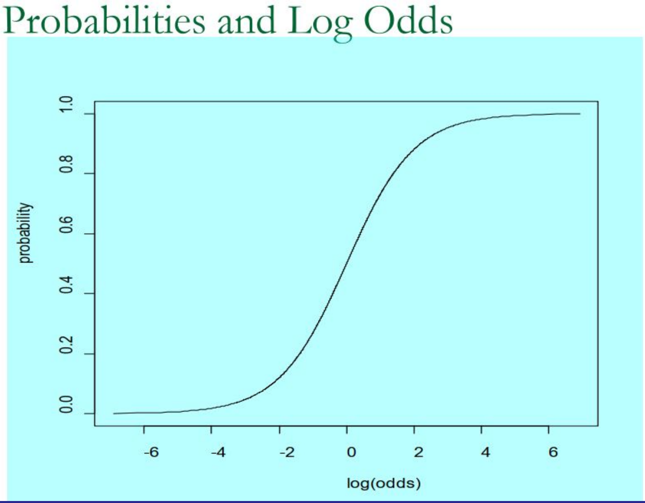
1. 从probability开始，概率变化[0,1]
2. 比如，成功的概率p=0.8 ,那么失败的概率q = 1-0.8 = 0.2;
3. 所以，成功的胜率（几率）为：
$$odds(success) = \frac{p}{1-p} = \frac{0.8}{0.2} = 4;$$
4. 失败的概率为： $odds(failure) = \frac{q}{1-q} = \frac{0.2}{0.8} = 0.25;$
5. 结论：很显然，成功的几率和失败的几率互为倒数。

概率 有上下限	odds 有下限	log odds 没有上下限
0.0001	$\frac{1}{9999}$	-9.21
0.5	1	0
0.8	4	0.6
0.9	9	0.95
0.9982	554	2.74
0.9983	586	2.76
0.9998	4999	3.7
0.9999	9999	+9.21

概率$\longrightarrow$odds$\longrightarrow$log odds
概率$\longrightarrow$odds and odds ratio$\longrightarrow$log odds


Odds and Probabilities
probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
odds

SIGMOID CURVE



Example 1

共有10个male、10个female考大学。

被录取的结果为：
7个male、3个female。

因变量（admit）	自变量（sex）
录取，admit=1	男性=1
不录取，admit=0	女性=0

1. $logit(p) = log(odds) = log(\frac{p}{1-p})$
（所谓logit，即logistic probability unit。这里的log是以e为底，相当于ln）

2. $logit(p) = a + bX$

X是自变量（sex），b为系数;
p 为admit（录取结果）= 1 的概率。

3. 变换为：

$$log(\frac{p}{1-p}) = a + bX$$
$$log(odds) = a + bX$$

这就意味着逻辑斯蒂回归的系数b in term of log(odds)。

1. 第一步把概率转变成odds（译成：优势比、发生比、几率比）；
   (1) 不足：概率变化与odds变化的不对称。概率变化一点点，odds会发生巨变。
2. 第二步取odds的自然对数。
   (1) 概率变化与odds自然对数的变化不仅对称，还以0为中间点。
      - $p = 0.5$, log odds$= 0$
      - $p < 0.5$, log odds < 0
      - $p > 0.5$, log odds > 0

```
. logit admit sex

Iteration 0:   log likelihood = -13.862944
Iteration 1:   log likelihood = -12.222013
Iteration 2:   log likelihood = -12.217286
Iteration 3:   log likelihood = -12.217286

Logistic regression                               Number of obs   =         20
                                                  LR chi2(1)      =       3.29
                                                  Prob > chi2     =     0.0696
Log likelihood = -12.217286                       Pseudo R2       =     0.1187
```

admit	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
sex	1.694596	.9759001	1.74	0.082	-.2181333	3.607325
_cons	-.8472978	.6900656	-1.23	0.220	-2.199801	.5052058

$$log(\frac{p}{1-p}) = a + bX$$
$$log(odds) = a + bX$$

1 这就意味着逻辑斯蒂回归的系数b in term of log(odds);
2 系数b为1.69意味着:
   (1) X(即性别)变化一个单位,导致log(odds)变化1.69个单位。
   (注:很少人用对数想问题)
   (2) X(即性别)变化一个单位,odds变化$e^{1.69} = 5.44$个单位。
   (即:男性成功录取的odds是女性成功录取odds的5.44倍)

举例：什么是ODDS RATIO?

1 从录取这个角度而言，

$$odds(male) = \frac{p}{1-p} = \frac{0.7}{0.3} = 2.333$$
$$odds(female) = \frac{q}{1-q} = \frac{0.3}{0.7} = 0.428$$

2 所以录取的odds ratio为：

$$OR = \frac{odds(male)}{odds(female)} = \frac{\frac{p}{1-p}}{\frac{q}{1-q}} = \frac{2.333}{0.428} = 5.44;$$

3 结论：对一个male而言，录取的成功率是female的5.44倍；男生考入大学的odds是女生考入大学odds的5.44倍。

$$OR = \frac{odds(male)}{odds(female)} = \frac{\frac{p}{1-p}}{\frac{q}{1-q}} = \frac{2.333}{0.429} = 5.44;$$

对一个male而言，录取的成功率（几率）是female的5.44倍。

odds ratio = OR= e^b

$$OR(sex) = e^{1.69} = 5.44$$

OR 代表胜率比（几率比），计算公式为：OR = $e^{coefficients}$

	大学录取	
	是	否
男性=1	7	3
女性=0	3	7

1 OR is the cross-product ratio (compare sex = 1 group to sex = 0 group)

$$odds\ ratio = \frac{\frac{7}{3}}{\frac{3}{7}} = \frac{7 \times 7}{3 \times 3} = \frac{49}{9} = 5.44$$

2 odds of y = 1 are 5.44 times higher when sex =1 than when sex = 0.

3 对一个male而言，录取的成功率（几率）是female的5.44倍。

计算ODDS RATIO的三种方式

1. OR = $\frac{OR_1}{OR_2}$
2. OR = e^b

   b是逻辑斯蒂回归的系数in term of log(odds)。
3. OR is a cross-product ratio.

```
. logit admit sex, or

Iteration 0:   log likelihood = -13.862944
Iteration 1:   log likelihood = -12.222013
Iteration 2:   log likelihood = -12.217286
Iteration 3:   log likelihood = -12.217286

Logistic regression                               Number of obs   =         20
                                                  LR chi2(1)      =       3.29
                                                  Prob > chi2     =     0.0696
Log likelihood = -12.217286                       Pseudo R2       =     0.1187
```

admit	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	5.444444	5.313233	1.74	0.082	.8040182	36.86729
_cons	.4285714	.2957424	-1.23	0.220	.1108252	1.657327

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

$$log\frac{p}{1-p} = (a + bX)$$

$$e^{(log\frac{p}{1-p})} = e^{(a+bX)}$$

$$\frac{p}{1-p} = e^{(a+bX)}$$

$$p = \frac{e^{(a+bX)}}{1+e^{(a+bX)}}$$

odds ratio = OR= e^b

$$log(\frac{p}{1-p}) = a + bX$$

系数b为1.69意味着，

1 X（即性别）变化一个单位，导致log(odds)变化1.69个单位。
2 X（即性别）变化一个单位，odds变化 $e^{1.69} = 5.44$个单位。
   (1) 把log odds还原为odds。
   (2) 一个是X未变时的odds，一个是X变化一个单位的odds。
   (3) 然后看两个odds的比，也就是odds ratio。
3 更直观的是把log odds的变化还原为概率的变化。

按ODDS来解释LOGISTIC回归系数

$$odds = \frac{p}{1-p} = e^{(a+b_1 SEX+b_2 INCOME)} = e^a \times e^{b_1 SEX} \times e^{b_2 INCOME}$$

1. $b_1 > 0$，意味着男性录取的odds大于女性。
2. $b_1 = 0$，意味着男性与女性具有相同的录取odds。
3. $b_1 < 0$，意味着男性录取的odds小于女性。

按ODDS RATIO来解释LOGISTIC回归系数

1. $e^{b_1} > 1$，意味着男性录取的odds大于女性。
2. $e^{b_1} = 1$，意味着男性与女性具有相同的录取odds。
3. $e^{b_1} < 1$，意味着男性录取的odds小于女性。

小结：LOGISTIC回归系数的解释

1 回归系数的正负

(1) 如果系数估计为正，相应的自变量的增加（控制其他因素不变）会导致 Y=1 的可能性的增加；
(2) 如果系数估计为负，相应的自变量的增加（控制其他因素不变）会导致 Y=0 的可能性的增加。

2 log odds的变化是一条曲线。

(1) 自变量对因变量影响的大小，取决于影响发生在曲线的位置。
(2) （控制其他因素不变）自变量增加1个单位，log-odds增加b。
(3) （控制其他因素不变）自变量增加1个单位，odds ratio增加e^b。

1. logit x1 x2 x3 （logit模型）
2. logit x1 x2 x3, or （logit模型，显示odds ratio）
3. logistic x1 x2 x3 （显示odds ratio）
4. predict prob （计算发生概率的预测值）
5. list prob y if x1==0 & x2==0 （列出在给定条件下y=1的概率预测值）
6. estat classification （计算预测准确的百分比）
7. margins, dydx(*) （计算所有自变量的平均边际效应）
8. margins, dydx(*) at (x1=0) （计算所有自变量在$x_1 = 0$的平均边际效应）