CDA 统计课程

==

4.logistic 回归

4.1 logistic 回归与卡方

4.2 最大似然估计

4.3 logistic 回归解析

***模型与线性回归对比**

***两种 R2 对比：线性回归和 logistic 回归**

***回归系数与 or 值、rr 值**

***or 值与 gamma 值**

4.4 评分与预测

***老样本预测与市场细分**

***数据的分箱化**

==

4.1 logistic 回归与卡方

卡方分析原理

a、表示观测值与理论值间的偏离程度。

H0：观察频数与期望频数没有差别。

Pearson 卡方计算公式：

$$\sum_{i=1}^{k}\frac{(A_i-np_i)^2}{np_i} \quad ,i=1,2,3\cdots k$$

注：A_i为 i 水平的观察频数；n为总频数；ρ_ε为 i 水平的期望频数；k为单元格数；

单元格期望数的计算：单元格期望数的计算：

	Y1	Y2	合计
X1	频数 1	频数 2	r_i
X2	频数 3	频数 4	
合计	c_j		W

$$E_{ij}=np_i=\frac{r_i c_j}{W}$$

注：E_{ij}表示单元格期望值，或理论值；r_i用于表示第 i 行的子汇总；c_j用于表示第 j 列的子汇总；w用于表示总体汇总。期望次数是虚无假设成立时的数值。

b、统计量近似服从 k-1 自由度的卡方分布。

c、样本要求：

单元格最小期望频数需大于 1，

单元格期望频数小于 5 的不能超过 20%。

如果不满足，可以接受确切概率法的结果。

--------------------------------相关系数指标-------------------------------

统计量：

Kappa：内部一致性系数，取值在[0 1]，一般认为大于 0.75 表示一致性较好；在[0.4 0.75]间一致性一般；小于 0.4 较差。

风险：OR（比数比）和 RR（相对危险度），用于度量行列间的关联强调。

OR=（反应阳性组实验因素阳性人数/阴性人数）/（反应阴性组实验因素阳性人数/阴性人数）*--大于 1 表示试验因素更容易导致结果为阳性--

RR=（实验组反应阳性人数/实验组总人数）/（对照组反应阳性人数/对照组总人数）

Mcnemar：用于配对卡方检验。

===

4.2 最大似然估计

在满足独立同分布，并且随机变量分布参数已知的情况下，最大似然估计是最佳无偏估计。是所有无偏估计中最有效的估计量。

第一、确定模型随机误差项的分布。一般为 logistic 分布。

第二、构建拟然函数，观测数据出现的概率可以表述成未知模型参数的函数。因为独立同分布的假定，所以样本为 n 的联合分布可以表示成边际分布的连乘，以 logit 模型为例，其拟然函数为：

$$L = \prod p_i^{y_i}(1-p_i)^{(1-y_i)}$$

$$= \prod[(\mathrm{e}^{\sum_{k=0}^{j}\beta_k x_{ik}})/(1+\mathrm{e}^{\sum_{k=0}^{j}\beta_k x_{ik}})]^{y_i}[1/(1+\mathrm{e}^{\sum_{k=0}^{j}\beta_k x_{ik}})]^{(1-y_i)}$$

第三、取对数，即对数拟然函数。由于 L 的计算量复杂，lnL 可以将指数转化成加法形式，大大减少计算量，并且 lnL 与 L 为单调递增函数，所以 lnL 取最大时，同样使得 L 取最大。

第四、求偏导。令其为 0，解方程组，求得对应一组回归参数β_k的最优解。

==

4.3 logistic 回归解析

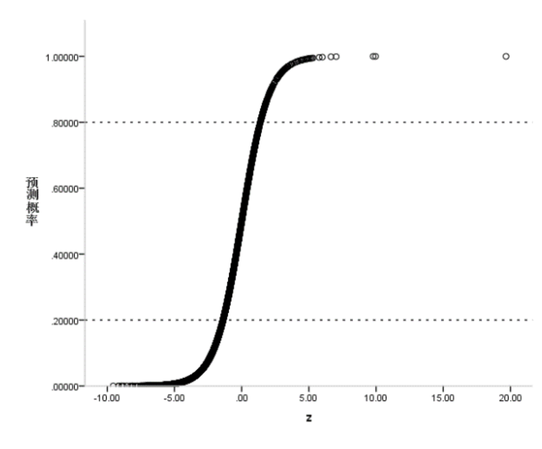
(1) Logistic 的回归方程式：

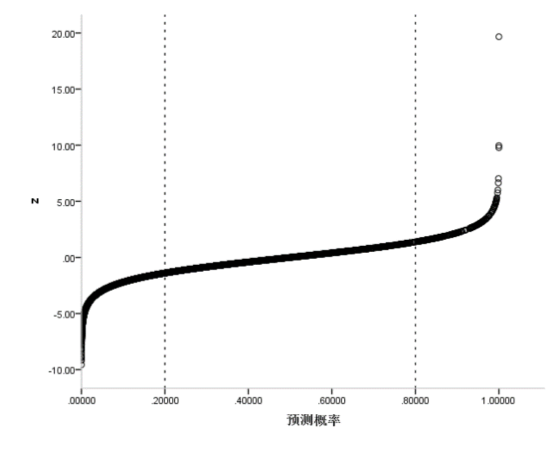
$$\text{Logit}(p_i)=\log（p_i/1\text{-}p_i）=\sum_{k=0}^{j}\beta_k x_{ik} \qquad (1\text{-}1)$$

$$\Rightarrow p_i=(e^{\sum_{k=0}^{j}\beta_k x_{ik}})/(1+e^{\sum_{k=0}^{j}\beta_k x_{ik}})$$

此外，由于二分因变量并没有尺度信息，误差方差具有人为设定的任意性，所以构造线性回归中的 R 方，着实不易，因此更多的是利用拟合优度产生的卡方指标判断模型优劣。

Logit 的概率函数曲线：0.2-0.8 概率间是线性关系，此外是非线性关系。





(2)由假设公式（1-1）中只包含一个自变量。

$\log（p_i/(1\text{-}p_i)）=\beta_0+\beta_1 x$

$\log（p_i'/(1-p_i')）=\beta_0+\beta_1(x+1)$

等式两侧同时取对数的反函数：

$\Rightarrow（p_i/(1\text{-}p_i)）=e^{\beta_0}e^{\beta_1 x}$

$（p_i'/(1-p_i')）=e^{\beta_0}e^{\beta_1 x}e^{\beta_1}$

两等式相除：

or=$（p_i'/(1-p_i')）/（p_i/(1\text{-}p_i)）=e^{\beta_0}e^{\beta_1 x}e^{\beta_1}/e^{\beta_0}e^{\beta_1 x}=e^{\beta_1}$

e^{β_1}表示为，x 每增加一个单位风险增加的倍数，如果β_1为 0.44，即e^{β_1}为 1.53，解释为其他变量处于控制的状态下，x 每增加一个单位风险相比原来增加 1.53 倍。

或者：

其他变量处于控制的状态下，x 每增加一个单位风险相比原来增加 53%（（1.53-1）/1*100%=53%）。

关于“风险”的含义。但概率取值偏大，结果的差异变大。

如若 1:
rr=p_i'/p_i=0.2/0.18=1.111111111
or=（$p_i'/(1-p_i')$）/（$p_i/(1-p_i)$）=1.13888888
如若 2:
rr=p_i'/p_i=0.1/0.08=1.25
or=（$p_i'/(1-p_i')$）/（$p_i/(1-p_i)$）=1.277777777
如若 3:
rr=p_i'/p_i=0.01/0.03=0.3333333333
or=（$p_i'/(1-p_i')$）/（$p_i/(1-p_i)$）=0.3265993266

```
%let x=0.01;
%let y=0.03;
data _null_;
    rr=&x./&y.;
    or=(&x./(1-&x.))/(&y./(1-&y.));
    put rr= or=;
run;
```

===

4.4 评分与预测

数据离散化

优点:

（1）通俗性
（2）影响速度
（3）避免过拟合
（4）消除异常值
（5）最优离散（有监督）
（6）保密

评分卡

（1）寻找最大拐点处。最大拐点处往往具有对称性。 （2）异常值诊断。为什么会出现异常，该异常和这组群体间的关系。 （3）拐点与业务意义