数学实践-

Logistic回归分析

二项Logistic回归：问题提出

- 研究二分类变量与其他变量之间的关系
  - 例如：研究吸烟对是否得肺癌的影响，并以年龄和性别作为控制变量，特点：
    - 被解释变量是二值变量
    - 解释变量有分类变量和定距变量
    - 吸烟与肺癌之间并非一种线性关系
- 对二项分类的被解释变量可否直接采用一般多元线性回归分析方法?
  - 结论：不可以

二项Logistic回归

- 当被解释变量为二项(0/1)分类变量时，被变量的取值范围和与自变量的关系问题：
  - 根据回归模型的意义，可知：

$$E(y_i) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \longrightarrow P_{y=1} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

  - 一般回归模型下的被解释变量的取值范围是 -∞~ +∞
  - 这里，被解释变量的取值范围是0~1
  - 一般回归分析建立模型，解释变量与*P*间的关系只能是线性的。

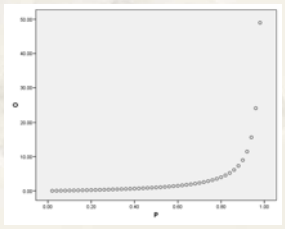
二项Logistic回归

- 解决问题的方向
  - 能否对概率P进行转换处理后，使其取值范围与一般线性回归模型吻合
  - 对概率*P*应采用非线性转化处理
  - 所有的转化都不应改变解释变量和被解释变量之间关系的方向

二项Logistic回归:理论上的处理

- 进行两步转换处理：
  - 第一步，将*P*转换成*Ω*
    - Ω称为优势
    - 对P的转化是非线性的
    - Ω是P的单调增函数
    - 优势的取值范围：0~ +∞

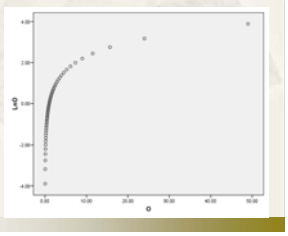
$$\Omega = \frac{P}{1-P}$$



二项Logistic回归:理论上的处理

- 进行两步转换处理：
  - 第二步，*Ω*转换成ln*Ω*
    - ln*Ω*称为Logit *P*
    - Logit *P*与*Ω*仍呈增长（或下降）的一致性关系
    - Logit *P*的取值于 -∞~ +∞

$$\ln(\Omega) = \ln(\frac{P}{1-P})$$



二项Logistic回归:理论上的处理

- **二项Logistic模型：**

$$\ln(\frac{P}{1-P}) = \beta_0 + \sum_{i=1}^{k}\beta_i x_i$$

$$Logit\ P = \beta_0 + \sum_{i=1}^{k}\beta_i x_i$$

二项Logistic回归

- **P与自变量间为非线性关系：**

$$\frac{P}{1-P} = \exp(\beta_0 + \sum_{i=1}^{k}\beta_i x_i)$$

$$P = (1-P)\exp(\beta_0 + \sum_{i=1}^{k}\beta_i x_i)$$

$$P = \exp(\beta_0 + \sum_{i=1}^{k}\beta_i x_i) - P * \exp(\beta_0 + \sum_{i=1}^{k}\beta_i x_i)$$

$$P[1+\exp(\beta_0 + \sum_{i=1}^{k}\beta_i x_i)] = \exp(\beta_0 + \sum_{i=1}^{k}\beta_i x_i)$$

$$P = \frac{\exp(\beta_0 + \sum_{i=1}^{k}\beta_i x_i)}{1+\exp(\beta_0 + \sum_{i=1}^{k}\beta_i x_i)}$$

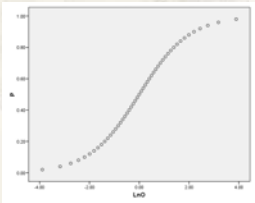
增长曲线

$$P = \frac{1}{1+\exp[-(\beta_0 + \sum_{i=1}^{k}\beta_i x_i)]}$$

二项Logistic回归系数的含义

- **回归系数表示当其他自变量取值保持不变时，某自变量取值增加一个单位引起Logit P平均变化β_i个单位**
  - **在模型的实际应用关心的是自变量变化引起事件发生概率P变化的程度**
  - **当自变量x_i变化时，对概率P的影响程度是非线性的，不易直观理解**

更注重自变量对发生比Ω的影响



二项Logistic回归系数的含义

- **优势：Ω=*P*/(*1-P*)，即某事件发生的概率与不发生的概率之比**
  - **利用优势比可以进行组之间风险的对比分析**
  - **例如，如果吸烟得肺癌的概率是0.25，不吸烟得肺癌得概率是0.10，则两组的优势比为：**

$$OR_{A\ VS.B} = \frac{pr(D_A)}{1-pr(D_A)} / \frac{pr(D_B)}{1-pr(D_B)} = \frac{1}{3} / \frac{1}{9} = 3$$

  - **吸烟的风险近似是不吸烟的三倍，吸烟组得肺癌的风险高于不吸烟组**

二项Logistic回归系数的含义

- **如果被解释变量*y*(肺癌1=得/0=没)，自变量*x*只有一个(x1吸烟1=吸烟/0=不吸烟)，则logistic方程为：**

$$\text{logit}\ [pr(Y=1)] = \beta_0 + \beta_1 X_1$$

  - **吸烟与不吸烟组的方程分别是：**

$$\text{logit}\ [pr(Y=1)] = \ln(odd\ (\text{smokers})) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

$$\text{logit}\ [pr(Y=1)] = \ln(odd\ (\text{nonsmkers})) = \beta_0 + \beta_1 \times 0 = \beta_0$$

  - **两组优势比为：**

$$\text{OR}_{\text{S VS. NS}} = \frac{odds\ (\text{smokers})}{odds\ (\text{nosmokers})} = \frac{e^{(\beta_0+\beta_1)}}{e^{\beta_0}} = e^{\beta_1}$$

  - **可见，当解释变量是1/0二组时，两组间的对比是关于回归方程相应回归系数的对比**

二项Logistic回归系数的含义

- **如果被解释变量*y*(肺癌1=得/0=没)，自变量*x*有三个(x1吸烟/x2年龄/x3性别)，则logistic方程为：**

$$\text{logit}\ [pr(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

  - **X_A=(1,45,1)与X_B=(0,45,1)的方程分别是：**

$$\text{logit}\ [pr(Y=1)] = \ln(odd\ (X_A)) = \beta_0 + \beta_1 \times 1 + \beta_2 \times 45 + \beta_3 \times 1$$

$$\text{logit}\ [pr(Y=1)] = \ln(odd\ (X_B)) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 45 + \beta_3 \times 1$$

  - **两组优势比为：**

$$\text{OR}_{\text{XA VS. XB}} = \frac{odds\ (X_A)}{odds\ (X_B)} = e^{(1-0)\beta_1 + (45-45)\beta_2 + (1-1)\beta_3} = e^{\beta_1}$$

  - **这里的主要目的是研究吸烟对肺癌的影响，年龄和性别是作为控制变量存在的，该比率为调整比率，与不包括控制变量在内的比率不相等。（也可将定距变量作观测变量）**

二项Logistic回归系数的含义

- 自变量对优势Ω的影响

$$\Omega = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)$$

  - 当其他解释变量保持不变而研究观测变量变化一个单位对Ω的影响时，可将新的优势设为Ω*，则有优势比为：

$$\frac{\Omega^*}{\Omega} = \exp(\beta_i)$$

  - 即：当x_i增加一个单位时，将引起优势是原来的$\exp(\beta_i)$倍

二项Logistic回归系数的含义

- 如果被解释变量*y*(肺癌1=得/0=没)，自变量*x*有三个(x1吸烟/x2年龄/x3性别),并考虑吸烟与年龄和对性别的交互作用)，则logistic方程为：

$$\text{logit}\,[\,pr\,(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$$

  - X_A=(1,45,1,1×45,1×1)与X_B=(0,45,1,0×45,0×1)两组的优势比是：

$$\text{OR}_{(\text{XA VS. XB}|\text{age=45,sex=1})} = \frac{odds(X_A)}{odds(X_B)} = e^{(1-0)\beta_1+(45-45)\beta_2+(1-1)\beta_3+(45-0)\beta_4+(1-0)\beta_5}$$

$$= e^{\beta_1+45\beta_4+\beta_5} = e^{\beta_1+\beta_4 x_2+\beta_5 x_3}$$

  - 这里涉及到了多个系数，以及控制变量的不同取值

$$\text{OR}_{(\text{XA VS. XB}|\text{age=35,sex=0})} = e^{\beta_1+35\beta_4}$$

$$\text{OR}_{(\text{XA VS. XB}|\text{age=20,sex=1})} = e^{\beta_1+20\beta_4+\beta_5}$$

二项Logistic回归的参数估计

- 采用极大似然估计法进行参数估计：似然函数值
  - 例如：通过样本数据对购买的比例θ进行估计，其总体服从参数为θ的二项分布。假设θ只有0.2和0.6两个取值，则：

$$pr(Y;\theta) = C_m^y \theta^y (1-\theta)^{m-y} \quad pr(Y;0.2) = C_m^y 0.2^y (1-0.2)^{m-y} \quad pr(Y;0.6) = C_m^y 0.6^y (1-0.6)^{m-y}$$

  - 如果*m*=5，则
  - 如果*y*=4，则θ = 0.6

似然估计函数

$$pr(Y;\hat{\theta}) > pr(Y;\theta^*)$$

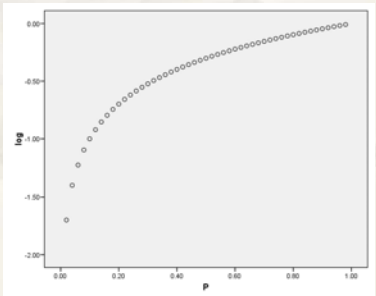
θ	y					
	0	1	2	3	4	5
0.2	0.328	0.409	0.205	0.051	0.007	0.000
0.6	0.010	0.077	0.230	0.346	0.259	0.078

  - 以似然函数值达到最大时的参数值作为总体参数的估计值，似然函数值在0至1间，反映了在所估计参数的总体中抽到特定样本的可能性行，越接近1越好

二项Logistic回归的检验

- 采用极大似然估计法进行参数估计：似然函数值
  - 求似然函数值的对数，得到对数似然函数值

对数似然函数值越大(越接近于0)，意味着模型较好地拟合样本数据的可能性越大，所得模型的拟和优度高；相反，对数似然函数值越小，意味着模型较好地拟合样本数据的可能性越小，所得模型的拟和优度低。



二项Logistic回归的检验

- 回归方程的显著性检验：自变量全体与Logit P的线性关系是否显著，原假设：回归系数同时为0
  - 采用对数似然比测度拟合程度是否提高
  - 设某自变量未引入回归方程前的对数似然函数值为：L_{x_i}
  - 某自变量引入回归方程后的对数似然函数值为：L
  - 对数似然比为：$\frac{L_{x_i}}{L}$
  - 如果对数似然比与1无显著差异，则说明该自变量对Logit *P*的线性解释无显著贡献；如果对数似然比远远大于1，与1有显著差异，则说明解释变量对Logit *P*的线性有显著贡献。

二项Logistic回归的检验

- 回归方程的显著性检验：自变量全体与Logit P的线性关系是否显著
  - 由于对数似然比 $\frac{L_{x_i}}{L}$ 的分布未知，但其函数(似然比卡方) $-\log(\frac{L_{x_i}}{L})^2$
  - 近似服从卡方分布

$$-\log(\frac{L_{x_i}}{L})^2 = -2\log(\frac{L_{x_i}}{L}) = -2\log(L_{x_i}) - (-2\log(L))$$

  - SPSS将自动计算似然比卡方的观测值和对应的概率*p*值

多项Logistic回归：应用案例

- **分析职业、性别在选择品牌（三种）时的倾向性**
  - **利用广义logit模型分析。如果因变量有K个水平，则设定一个对照水平(参照水平)，其他各水平分别与参照水平比较**
  - **例如：因应变量有a、b、c三个水平，以a作为参照，则有：**

$$LogitP_a = \ln[\frac{P_a}{P_a}] = \ln 1 = 0$$

$$LogitP_b = \ln[\frac{P(y=b \mid x)}{P(y=a \mid x)}] = \beta_0 + \sum_{j=1}^{k} \beta_{1j} x_j$$

$$LogitP_c = \ln[\frac{P(y=c \mid x)}{P(y=a \mid x}] = \beta_0 + \sum_{j=1}^{k} \beta_{2j} x_j \qquad P_a + P_b + P_c = 1$$

多项Logistic回归：应用案例

- **分析职业、性别在选择品牌（三种）时的倾向性**

Parameter Estimates

购买品牌[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
A	Intercept	-.656	.296	4.924	1	.026			
	[x1=1.00]	-1.315	.384	11.727	1	.001	.269	.127	.570
	[x1=2.00]	-.232	.333	.486	1	.486	.793	.413	1.522
	[x1=3.00]	0[b]	.	.	0	.	.	.	.
	[x2=1.00]	.747	.282	7.027	1	.008	2.112	1.215	3.670
	[x2=2.00]	0[b]	.	.	0	.	.	.	.
B	Intercept	-.653	.293	4.986	1	.026			
	[x1=1.00]	-.656	.339	3.730	1	.053	.519	.267	1.010
	[x1=2.00]	-.475	.344	1.915	1	.166	.622	.317	1.219
	[x1=3.00]	0[b]	.	.	0	.	.	.	.
	[x2=1.00]	.743	.271	7.533	1	.006	2.101	1.237	3.571
	[x2=2.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: C.

b. This parameter is set to zero because it is redundant.

多项Logistic回归：应用案例

- **分析职业、性别在选择品牌（三种）时的倾向性**

$$\log it \frac{p_a}{p_c} = -0.656 - 1.315x1(1) + 0.747x2(1)$$

•当性别相同时，第一种职业的logit(P_a/P_c)比第三种职业(参照水平)平均减少1.315，第一种职业的(P_a/P_c)是第三种职业的0.269倍。如果以P_c为基准，则第一种职业选择A品牌的倾向不如第三种职业，且统计上显著；

•当职业相同时，男性的logit(P_a/P_c)比女(参照水平)平均多0.747，男性的(P_a/P_c)是女性的2.112倍。如果以P_c为基准，则男性较女性更倾向选择A品牌，且统计上显著，即男性选择A品牌的倾向性与女性的差异显著。

多项Logistic回归：应用案例

- **分析职业、性别在选择品牌（三种）时的倾向性**

$$\log it \frac{p_b}{p_c} = -0.653 - 0.656x1(1) + 0.743x2(1)$$

•当性别相同时，第一种职业的logit(P_b/P_c)比第三种职业(参照水平)平均减少0.653，第一种职业的(P_b/P_c)是第三种职业的0.519倍。如果以P_c为基准，则第一种职业选择B品牌的倾向不如第三种职业，但统计上不显著；

•当职业相同时，男性的logit(P_b/P_c)比女(参照水平)平均多0.743，男性的(P_b/P_c)是女性的2.101倍。如果以P_c为基准，则男性较女性更倾向选择B品牌，且统计上显著，即男性选择B品牌的倾向性与女性的差异显著。