第11章　逻辑回归

第11章 逻辑回归分析——目录

应用背景

- 许多社会科学问题中的因变量往往是分类变量。比如，政治学中经常研究的是否选举某候选人，候选人的类型等；

- 又如，经济学研究中所涉及的是否销售或购买某种商品、是否签订 一份合同，保险是否违约，违约有哪些类型等等。

这些分类变量中有一类特殊的变量，遵循二值取值原则，要么“是”或“发生”，要么“否”或“未发生”。统计上我们将这样的变量称作二分类变量（Binary variable）。

11.1 逻辑回归模型

- 多元回归模型分析二分类变量的局限性
  - 被解释变量的取值区间受限制
  - 自变量的边际分析不符合实际

- 分析二分类变量的方法

利用概率转化模型调整二分类变量使其线性化，也即，使其随着自变量的变化，这一概率的值总是在0到1之间变化。

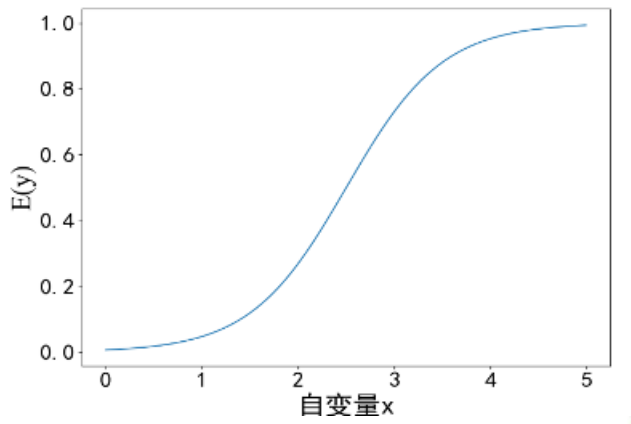
11.1 逻辑回归模型——概率转换方法

- Logistic 函数

$$f(x)=\frac{1}{1+\exp[-(a+bx)]}$$

例：假定模型仅包括一个自变量 x，并且模型的参数。$\beta_0=-5$，$\beta_1=2$

也即：



$$P=E(y)=\frac{1}{1+\exp[-(-5+2x)]}$$

- Logistic 回归方程

$$P=E(y)=\frac{1}{1+\exp[-(\beta_0+\beta_1x_1+\beta_2x_2\,\mathrm{L}\;\;\beta_kx_k)]}$$

11.2 估计逻辑回归模型——最大似然估计

➢ 似然函数

$$L(Y;\theta)=\prod_{i=1}^{n}\Pr(Y_i;\theta_i)=\prod_{i=1}^{n}[\theta_i^{Y_i}(1-\theta_i)^{1-Y_i}]$$

其中，

$$\theta_i=1\Big/\{1+\exp[-(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\mathrm{L}\ +\beta_kx_{ik})]\},i=1,2,\mathrm{L}\ \ ,n$$

追求似然函数最大值的过程就是追求对数似然函数值最大值的过程。**对数似然函数值越大**，意味着**模型较好地拟合**样本数据的可能性也越大，所得模型的拟合优度越高。

11.3 显著性检验

➢ 整体模型的检验和评价

- 似然比 L_0 / L_f

L_0 = 解释变量均未引入回归方程前的似然函数值
L_f = 解释变量引入回归方程后的似然函数值

如果似然比与1无显著差异，则说明当前模型中的解释变量全体对LogitP的线性解释无显著贡献

- 似然比 χ^2

$$-\ln\left(\frac{L_0}{L_f}\right)^2 = -2\ln\left(\frac{L_0}{L_f}\right) = -2\ln(L_0) - \left(-2\ln(L_f)\right)$$

11.3 显著性检验

➢ 整体模型的检验和评价

• 判错矩阵

		预测值		
		0	1	正确率
观测值	0	f_{11}	f_{12}	$\frac{f_{11}}{f_{11}+f_{12}}$
	1	f_{21}	f_{22}	$\frac{f_{22}}{f_{21}+f_{22}}$
		总体正确率		$\frac{f_{11}+f_{22}}{f_{11}+f_{12}+f_{21}+f_{22}}$

判错矩阵是一种极为直观的评价模型优劣的方法，它通过矩阵表格的形式展现预测值与实际观测值的吻合程度。

11.3 显著性检验

➢ 整体模型的检验和评价

- **Cox & Snell R^2统计量**

$$\text{Cox \& Snell } R^2 = 1 - \left(\frac{L_0}{L}\right)^{\frac{2}{n}}$$

Cox & Snell R2统计量与一般线性回归分析中的R2有相似之处，也是方程对被解释变量变差解释程度的反映。缺点是取值范围不易确定，因此使用时不方便。

- **Nagelkerke R2统计量**

$$\text{Nagelkerke R}^2 = \frac{\text{Cox \& Snell R}^2}{1 - (L_0)^{\frac{2}{n}}}$$

Cox & Snell R2统计量取值在0～1之间，越接近1说明方程的拟合优度越高。

11.3 显著性检验

➢ 回归系数的显著性检验

对于模型中某个自变量参数估计值的统计检验，我们可以采用Wald统计量。其原假设H_0：$\beta_i=0$。$S_{\beta i}$是回归系数的标准误。**Wald检验统计量**服从自由度为1的卡方χ^2分布。

- **Wald 统计量** $Wald_i=\left(\beta_i/S_{\beta_i}\right)^2$

- **多重共线性检验**

应当注意，如果解释变量存在**多重共线性**会对Wald检验统计量产生影响。由于用于logistic回归建模的很多软件包，如 Excel, SPSS, 和 R 并不提供共线性的问题检验，所以如果用户想检验共线性问题，可以就给定的自变量做一个线性回归模型，并输出共线性诊断指标，就可以了解自变量的相关情况。