《线性回归》
—logistic回归（估计和诊断）

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.06.04

主要内容：Logistic回归

1. **Bernoulli情形的MLE回顾**
2. **假设检验**
   - 似然比检验
   - Wald检验
   - Score检验
   - Bernoulli情形的结果与logistic 回归的关系
   - logit模型的系数估计
   - R中简单logistic回归系数的检验
   - 回归系数和odds ratios的置信区间
   - 简单logistic回归系数的LRT
   - deviance分析表
3. **logistic回归的诊断**
   - 残差
   - 影响分析
   - 模型选择
   - 预测

在处理较为复杂问题的MLE时，一定要搞清楚最简单情形。

Bernoulli情形的MLE

- 假定$Y_1, \cdots, Y_n$ *iid* Binomial$(1, p)$, $0 \le p \le 1$. $Y_i = 1$ 表示第i个事件发生，否则$Y_i = 0$表示不发生。
- 似然函数为：

$$L = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} = p^{\sum_{i=1}^{n} Y_i}(1-p)^{n-\sum_{i=1}^{n} Y_i}. \tag{1}$$

对数似然函数为：

$$\ell = \log(L) = \sum_{i=1}^{n} Y_i \log(p) + \left(n - \sum_{i=1}^{n} Y_i\right) \log(1-p). \tag{2}$$

Bernoulli情形的MLE(续):

- ℓ关于p的似然函数是:

$$U(p) = \frac{\partial l}{\partial p} = \sum_{i=1}^{n} Y_i/p - \left(n - \sum_{i=1}^{n} Y_i\right)/(1-p) \tag{3}$$

定义为**得分函数**(score function).

- 为计算p的MLE,令得分函数$U(p)=0$,关于p求解。

$$\hat{p} = \sum_{i=1}^{n} Y_i/n.$$

信息矩阵

- 另外一个重要的函数可以从似然推导出来，就是关于未知参数的Fisher Information. 信息函数是$\ell = \log L$的曲率负的值. 即

$$
\begin{aligned}
I(p) &= E\left[-\frac{\partial^2 l}{\partial p^2}\right] \\
&= E\left[\sum_{i=1}^{n} Y_i/p^2 + \left(n - \sum_{i=1}^{n} Y_i\right)/(1-p)^2\right] \\
&= \frac{n}{p(1-p)}
\end{aligned}
$$

注意， 信息阵通常依赖于未知的参数，需要估计出来。

信息阵的估计

将p的MLE带入$I(p)$, 得到$I(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}$.

参数p的推断:

我们利用在MLE处的信息函数逆来估计$\hat{p}$的方差:

$$\widehat{\text{var}}(\hat{p}) = I(\hat{p})^{-1} = \frac{\hat{p}(1-\hat{p})}{n}$$

对于充分大的n, $\hat{p}$近似地服从均值为p, 方差为$p(1-p)/n$的正态分布. 因此，我们可以构造p的置信水平为$100(1-\alpha)\%$的置信区间:

$$\hat{p} \pm Z_{1-\alpha/2}[\hat{p}(1-\hat{p})/n]^{1/2}.$$

假设检验

- 似然比检验(LRT)
- Wald 检验
- Score 检验

似然比检验

- LRT统计量为：

$$LR = -2\log\left(\frac{L \text{ at } H_0}{L \text{ at } MLE(s)}\right) = -2l\,(H_0) + 2l(MLE).$$

对于充分大的n, 近似地有：

$$LR \sim \chi^2,$$

其中χ^2的自由度是待估计参数的个数.

♠ 对于上面讨论的二值的情形，考虑假设：$H_0: p = p_0$ vs $H_A: p \neq p_0$, 则

$$\begin{aligned}\ell\left(H_0\right) &= \sum_{i=1}^{n} Y_i \log\left(p_0\right) + \left(n - \sum_{i=1}^{n} Y_i\right) \log\left(1 - p_0\right) \\ \ell(MLE) &= \sum_{i=1}^{n} Y_i \log(\hat{p}) + \left(n - \sum_{i=1}^{n} Y_i\right) \log(1 - \hat{p}).\end{aligned}$$

则LRT统计量为

$$LR = -2\left[\sum_{i=1}^{n} Y_i \log\left(p_0/\hat{p}\right) + \left(n - \sum_{i=1}^{n} Y_i\right) \log\left\{\left(1 - p_0\right)\left(1 - \hat{p}\right)\right\}\right],$$

其中$LR \sim \chi_1^2$.

Wald检验

- Wald检验统计量是：

$$W = \frac{(\hat{p} - p_0)^2}{\hat{p}(1-\hat{p})/n},$$

其中，对于充分大的n, $W \sim \chi^2$, 自由度是1. 在R中，报告的是$\sqrt{W} \sim N(0,1)$的结果.

Score检验

- 对于前面说的二值的情形，score检验统计量是：

$$S = U\left(p_0\right)^2 / I\left(p_0\right),$$

其中$S \sim \chi_1^2$.

Bernoulli情形的结果与logistic 回归的关系

- 到目前为止，我们学习了Bernoulli情形下如何估计p和检验关于p的假设。这些结果与logistic 回归是什么关系？
  - ✓ β的MLE；
  - ✓ 关于β的检验假设
  - ✓ 构造β的置信区间

- 对于logistic模型：

$$E\left[Y_i|\mathbf{X}_i=\mathbf{x}_i\right]=\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}.$$

n个观测值的似然为：

$$L=\prod_{i=1}^{n}\left(\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right)^{\sum_{i=1}^{n}Y_i}\left(1-\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right)^{n-\sum_{i=1}^{n}Y_i}$$

对数似然为

$$\ell=\sum_{i=1}^{n}\left[Y_i\log\left(\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right)+\left(1-Y_i\right)\log\left(1-\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right)\right]$$

MLE的计算

- logisitic回归模型的β的$p+1$个得分函数不能解析的求出。通常使用数值的方法求解。例如, Newton-Raphson算法【黑板】
- ℓ关于β的二阶导数的$(p+1)\times(p+1)$矩阵是信息阵。而信息阵的逆是$\hat{\beta}$的协方差矩阵.

R中简单logistic回归系数的检验

- 为检验logistic回归系数，利用Wald检验

$$\frac{\hat{\beta}_j - \beta_{j0}}{\hat{\text{se}}(\hat{\beta})} \sim N(0,1),$$

  其中$\hat{\text{se}}(\hat{\beta})$通过信息阵估计的逆得到。
- R中的实现与简单回归模型记为相似。但是要使用函数'glm'. 例如，
  $L = \text{glm}(y \sim x1 + x2$, data=mydat, binomial(link = "logit"))
  summary(L)可以查看结果.

- 在logistic模型

$$\text{logit}\left(P(Y_i = 1)\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

中，β_j的$(1-\alpha)\times 100\%$的CI为

$$\hat{\beta}_j \pm Z_{1-\alpha/2}\hat{se}\left(\hat{\beta}_j\right), j = 1, \cdots, p.$$

- x_j改变1个单位对应的odds ratio的$(1-\alpha)\times 100\%$的CI为

$$\left[\exp\left(\hat{\beta}_j - Z_{1-\alpha/2}\hat{\text{se}}\left(\hat{\beta}_j\right)\right), \exp\left(\hat{\beta}_j + Z_{1-\alpha/2}\hat{\text{se}}\left(\hat{\beta}_j\right)\right)\right].$$

简单logistic回归系数的LRT

- 假设
$$H_0 : \beta_2 = 0 \text{ vs. } H_A : \beta_2 \neq 0$$
的LRT统计量为
$$LR = -2\left(l\left(\hat{\beta}|H_0\right) - l\left(\hat{\beta}|H_A\right)\right)$$
- 检验程序与Wald检验非常的相似.
- LRT还可以检验关于多个变量的假设。例如，检验假设
$$H_0 : \beta_1 = \beta_2 = 0 \text{ vs. } H_A : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$
【想一下，检验统计量是什么。渐近分布的自由度是多少？（自由度为2，为什么？）】

deviance分析表

- 类似于线性模型中anova, 当得到上面的L (logistic回归的结果)后，可以运行：

  $$\text{anova(L)}$$

  得到所谓的deviance分析表
- 【请借助于线性模型中anova去解释这里的结果】

logistic回归的诊断

主要内容

- 模型拟合评估
- 残差
- 影响分析
- 模型选择
- 预测

主要内容

定义记号：

$$E\left[Y_i\right]=\pi_i,$$

$$\text{logistic}\left(\pi_i\right)=\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}=x_i'\beta$$

$x_i=\left(1,x_{i1},\ldots x_{ip}\right)'$, $\beta=\left(\beta_0,\ldots,\beta_p\right)'$. 因此

$$P\left[Y_i=1\right]=E\left[Y_i\right]=\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}.$$

模型拟合评估

- 拟合模型的deviance：

$$\text{DEV} = -2\sum_{i=1}^{n}\left[Y_i \log\left(\hat{\pi}_i\right) + (1 - Y_i)\log\left(1 - \hat{\pi}_i\right)\right]$$

  可以用来评估模型，其中$\hat{\pi}_i$是π_i的拟合值.
- 理想的模型是DEV趋于0. 注意到DEV ≥ 0.
- DEV越接近与0，模型拟合的越好；DEV越大，模型拟合的越差。

Hosmer-Lemeshow (goodness of fit test)拟合优度检验

- 检验的假设是：

$$\begin{aligned} H_0 : E[Y] &= \frac{\exp(X'\beta)}{1+\exp(X'\beta)} \\ H_a : E[Y] &\neq \frac{\exp(X'\beta)}{1+\exp(X'\beta)} \end{aligned} \tag{4}$$

- 检验统计量的计算步骤如下：
  - ✓ 拟合值排序
  - ✓ 拟合值分为c组（c通常介于6和10之间）
  - ✓ 计算每组中的观则数和期望数
  - ✓ 计算χ^2检验(回顾正态性检验中χ^2统计量)
    $X^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk}-E_{jk})^2}{E_{jk}} \sim \chi^2_{c-2}$
  - ✓ 将提供R代码【课下可以练习！】

残差

- 残差在甄别潜在的异常值（不能用模型很好的拟合的观测值）和模型错误设定时可能有用。
- logistic回归分析中有两种常用的残差：
  - ✓ Deviance residuals
  - ✓ partial residuals (略)

Deviance residual

- deviance residual在确定单个点不能很好的拟合模型时有用。
- deviance residual定义为：$\mathrm{dev}_i = \pm\left\{-2\left[Y_i \log\left(\hat{\pi}_i\right) + \left(1 - Y_i\right)\log\left(1 - \hat{\pi}_i\right)\right]\right\}^{1/2}$ 正负号的确定：若$Y_i \geq \hat{\pi}_i$，去正号，否则取负号.
- R: residuals()
- 可以作图查看残差图（纵坐标为残差，合作表可以为标号或者拟合值$\hat{\pi}_i$）.【可尝试作图！】

影响分析

- 利用R命令dffits()和dfbetas()可以甄别有影响的观测值【这两个函数在线性模型中也有用！】

模型选择

- 像线性模型中一样，可以利用AIC，逐步回归等方法进行模型选择或者比较.
- R: stepAIC(glm(y∼x1+x2,family=binomial,data=mydata))【找一个数据集，尝试这个方法！】
- 要搞清楚我们喜欢的模型是什么？

预测

- logistic回归的主要选取常常在于预测(或者解释)。假定我们估计了个体的概率，我们如何将这些概率转化为预测结果？
- 对预测而言，最基本的规则是：
  - ✓ 使用0.5作为分界点。如果对新的协变量预测得到的$\hat{\pi} > 0.5$，则预测结果为$y = 1$. 否则为$y = 0$.

定量评估预测能力

- logistic模型的预测可以用ROC（receiver operating characteristic）曲线来衡量。这个曲线的横坐标是1-specificity, 纵坐标是sensitivity.
- ROC曲线下的面积可以给出模型的预测能力。如果面试是0.5，此时ROC曲线时斜率为1的直线，模型随机的预测结果。如果面积接近于1，则模型的预测能你很强。

名词解释：特异性（specificity）和灵敏度（sensitivity）

- 特异性（specificity）和灵敏度（sensitivity）。logistic模型可以用于分类. 考虑二分类的情况，类别为1和0，我们将1和0分别作为正类（positive）和负类（negative），则实际分类的结果有4种. 表格如下：

横：类别；纵：分类	1	0
1	T	F
1	F	T

  - ✓ 敏感度（sensitivity）—TPR：true positive rate，描述识别出的所有正例占所有正例的比例（实际有病而按该筛检试验的标准被正确地判为有病的百分比。它反映筛检试验发现病人的能力）。
  - ✓ 特异度（specificity）—TNR: true negative rate，描述识别出的负例占所有负例的比例（实际无病按该诊断标准被正确地判为无病的百分比，它反映筛检试验确定非病人的能力。）

名词解释：特异性（specificity）和灵敏度（sensitivity）

✓ 在医学中是这样定义的：
灵敏度=真阳性人数/（真阳性人数+假阴性人数）*100%。正确判断病人的率。
特异度=真阴性人数/（真阴性人数+假阳性人数））*100%。正确判断非病人的率。

第二次测验题目：logistic模型参数估计的数值解