《线性回归》
—logisitc回归（模型和估计）

杨　　瑛

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

2019.05.30

主要内容：Logistic回归

1. **Logistic回归**
   - Logistic回归
   - 例子:智利公民投票数据
   - 为什么线性模型不适合0/1数据？
   - 可能的解决方案
   - logit模型中的参数估计
   - 解释：π_i, 优势和对数优势
   - logit模型的参数解释
   - 多重logisitic回归
   - 协变量的类型
   - logisitic回归的检验

Logistic回归

- 当响应变量有两个结果：是/否，0/1，可以使用Logistic回归.
- 预测/描述$\mathbf{E}(Y_i|x_i)$.
- 为什么我们不能使用线性回归?
- 使用logistic回归进行检验

Example (例子:智利公民投票数据)

- 有关的历史:
  - ✓ 1973年: 皮诺切的军事政变
  - ✓ 1988年: 决定政府未来的公民投票:
    Yes-vote = 赞成票：继续维持军政府8年以上
    No-vote = 反对票：改为文官政府。
- 在全民投票前6个月，对随机选出的2700名智利选民进行调查:
  - ✓ 868计划投赞成票
  - ✓ 889计划投反对票
  - ✓ 558是未决定
  - ✓ 187计划弃权
  - ✓ 168没有回答
- 我们只看赞成票/反对票

- 除上面例子外，在医学研究等领域中，我们经常碰到响应变量不是连续而是为分为两个类的变量。例如：
  - ✓ 得病的状态（有病或者没有病）
  - ✓ 生或者死
  - ✓ 低出生体重与否
  - ✓ 健康状况改善与否
- 违约/不违约；诚信/不诚信; ......
- 设$\mathbf{Y}_i$表示上面的分类变量，可以编码：$\mathbf{Y}_i = 1$ 或者0, 对应得病或者不得病，等。$\mathbf{X}_i$记为其它的协变量，$i = 1, \cdots, n$.
- 对于上面的数据建立下面的模型：

$$\mathbf{Y}_i = \mathbf{X}_i^T\theta + \epsilon_i, 1 \leq i \leq n.$$

- 【思考】这个模型合理吗？为什么？

为什么线性模型不适合0/1数据？

- 问题：
  - ✓ 线性模型仅适用于有限的范围。在此范围之外，我们得到的拟合值小于零或大于一。
  - ✓ Y_i只能取值0和1，误差不是正态分布。
  - ✓ 统计误差的方差不是恒定的。

可能的解决方案：

- 使用逻辑回归：
  - ✓ logit$(u) = \log(u/(1-u))$.
  - ✓ 如果$u \in (0,1)$，那么logit$(u) \in (-\infty, \infty)$.
  - ✓ 原则上，可以对y值使用logit变换，但是由于没有定义logit（0）和logit（1），因此必须稍微扰动它们（多少？）。
- 事实上，我们是对$\mathbf{E}[\mathbf{Y}_i|x_i]$做logit转换:

$$
\begin{aligned}
\text{logit}\mathbf{E}[\mathbf{Y}_i|x_i] &= \alpha + \beta x_i, \\
\text{logit}P(\mathbf{Y}_i = 1|x_i) &= \alpha + \beta x_i.
\end{aligned}
$$

对二值响应变量建立模型的基本想法

- 注意这门课程一开始讲的建立回归模型的基本想法：

$$\mathbf{E}[\mathbf{Y}_i|\mathbf{x}_i] = m(\mathbf{x}_i),$$

- 当$\mathbf{Y}_i$是二值(0/1)响应变量时，有$\mathbf{E}[\mathbf{Y}_i|\mathbf{x}_i] = P(\mathbf{Y}_i = 1|\mathbf{x}_i)$. 即，

$$P(\mathbf{Y}_i = 1|\mathbf{x}_i) = m(\mathbf{x}_i).$$

依概率特性，$m(\cdot)$应该是一个取值于$[0,1]$上的函数！同时，

$$\text{Var}[\mathbf{Y}_i] = m(\mathbf{x}_i)(1 - m(\mathbf{x}_i)).$$

- 利用均值–方差的关系有可能建立更为合理的回归模型.

对二值响应变量建立模型的基本想法

- $m(\cdot)$的选择：
  - ✓ $m(u) = \frac{\exp(u)}{1+\exp(u)}$ [logit 模型]
  - ✓ $m(u) = \Phi(u)$, $\Phi(\cdot)$是$N(0,1)$的cdf. [logit模型]

2×2表

为了理解logit模型，这里介绍非常简单的2×2表。

Prob. 暴露 / 得病	Yes	No
Yes	π_{11}	π_{12}
No	π_{21}	pi_{22}

其中$\pi_{ij} = P($暴露$= i$ & 得病$= j)$.
刻画关联最常用的量是relative risk和odds ratio.

Odds ratio

- Relative Risk

$$\mathrm{RR} = \frac{P(\text{得病}|\text{暴露})}{P(\text{得病}|\text{未暴露})} = \frac{\pi_{11}/(\pi_{11}+\pi_{12})}{\pi_{21}/(\pi_{21}+\pi_{22})} = \frac{\pi_{11}(\pi_{21}+\pi_{22})}{\pi_{21}(\pi_{11}+\pi_{12})}$$

- Odds
在暴露的情形下，得病的odds是：

$$\frac{P(\text{得病}|\text{暴露})}{P(\text{未得病}|\text{暴露})} = \frac{\pi_{11}/(\pi_{11}+\pi_{12})}{\pi_{12}/(\pi_{11}+\pi_{12})} = \frac{\pi_{11}}{\pi_{12}}.$$

- Odds Ratio可以表示为：

$$\mathrm{OR} = \frac{\text{得病的odds}|\text{暴露}}{\text{得病的odds}|\text{未暴露}} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{12}} = \frac{\pi_{11}}{\pi_{21}}.$$

得病概率的回归模型

- 如何建立响应**Y**和暴露**X**之间的关系？
- 可以选择合适的函数$g(\cdot)$使得

$$g(E[\mathbf{Y}_i|\mathbf{X}_i=\mathbf{x}_i])=\mathbf{x}_i^T\theta,$$

或者

$$P[\mathbf{Y}_i|\mathbf{X}_i=\mathbf{x}_i])=g^{-1}(\mathbf{x}_i^T\theta)=\pi_i,$$

其中$g(\cdot)$称之为link function.

Y的分布

- 单个观测值的分布为：

$$p(\mathbf{y}_i \quad = \quad \pi_i^{\mathbf{y}_i}(1-\pi_i)^{1-\mathbf{y}_i}=[g^{-1}(\mathbf{x}_i\theta)]^{\mathbf{y}_i}[1-g^{-1}(\mathbf{x}_i\theta)]^{1-\mathbf{y}_i}.$$

两种典型的link function

- $g(\cdot)$的两种典型选择：

$$
\begin{aligned}
g(u) &= \log\frac{u}{1-u}, 0<u<1 \\
g(u) &= \Phi^{-1}(u), 0<u<1,
\end{aligned}
$$

其中$\Phi(\cdot)$是标准正态分布$N(0,1)$的分布函数.

- 它们的逆函数为

$$
\begin{aligned}
g^{-1}(u) &= \frac{\exp(u)}{1+\exp(u)} \\
g^{-1}(u) &= \Phi(u),
\end{aligned}
$$

- 它们分别对应logit和probit模型.

logit模型和probit模型

- logit线性模型

$$P(\mathbf{Y}_i = 1|\mathbf{x}_i) = \frac{\exp(\alpha + \beta\mathbf{x}_i)}{1 + \exp(\alpha + \beta\mathbf{x}_i)}.$$

- probit线性模型

$$P(\mathbf{Y}_i = 1|\mathbf{x}_i) = \Phi(\alpha + \beta\mathbf{x}_i).$$

- 后续问题：
  - ✓ 如何解释参数β和α的含义？【黑板】【试着解释！】
  - ✓ 如何估计参数α和β: MLE
  - ✓ 如何推断（CI和检验）：利用渐近分布或者LRT
  - ✓ Bootstrap

logit模型中的参数估计

- 记logit$P_\theta(\mathbf{Y}_i = 1|x_i) = \mathbf{x}_i^T\theta$, 其中$\theta = (\alpha, \beta)^T$.或者

$$
\begin{aligned}
P_\theta[\mathbf{Y}_i = y_i|\mathbf{x}_i] &= \left(\frac{P_\theta[\mathbf{Y}_i = 1|\mathbf{x}_i]}{P_\theta[\mathbf{Y}_i = 0|\mathbf{x}_i]}\right)^{y_i} P_\theta[\mathbf{Y}_i = 0|\mathbf{x}_i] \\
&= \exp[y_i\mathbf{x}_i^T\theta - \log(1 + \exp(\mathbf{x_i}^T\theta))].
\end{aligned}
$$

- 对数似然是:

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{n} \log P_\theta(\mathbf{Y}_i = y_i|\mathbf{x_i}) \\
&= \sum_{i=1}^{n} [y_i\mathbf{x}_i^T\theta - \log(1 + \exp(\mathbf{x_i}^T\theta))].
\end{aligned}
$$

估计的计算

- 最大值可求解下面的方程:

$$\sum_{i=1}^{n}(y_i - P_{\hat{\theta}}[\mathbf{Y}_i = 1|\mathbf{x_i}])\mathbf{x_i} = \mathbf{0}.$$

- 用迭代法求解.

解释：π_i, 优势和对数优势

- 令$\pi_i = P[\mathbf{Y}_i = 1|\mathbf{x_i}]$表示在$\mathbf{X} = \mathbf{x_i}$的条件下$\mathbf{Y} = 1$的条件概率.
- 注意$\mathbf{E}[\mathbf{Y}|\mathbf{x_i}] = \pi$【黑板】
- $\pi/(1-\pi)$是在$\mathbf{X} = \mathbf{x_i}$的情况下$\mathbf{Y} = 1$的优势(odds)。
- $\log(\pi/(1-\pi))$是对数优势(log odds) [也可以叫做对数赔率]。
- 有关log odds，请参阅表.

logit模型的参数解释

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta \mathbf{x_i}.$$

- Logistic回归是对数优势的的加法模型。这给出了β的一种解释：如果$\mathbf{X}$增加1个单位，则对数优势增加β。
- 逻辑回归是优势的乘法模型：

$$\frac{\pi_i}{1-\pi_i} = \exp(\alpha + \beta \mathbf{X}_i) = \exp(\alpha)[\exp(\beta)]^{\mathbf{X}_i}$$

  这就给出了β的另一种解释：如果$\mathbf{X}$增加1个单位，则胜算将乘以$\exp(\beta)$
- 注意：

$$\pi_i = \frac{1}{1+\exp[-(\alpha+\beta \mathbf{X}_i)]}.$$

logit模型的参数解释(续)

- $$\pi_i = \frac{1}{1 + \exp[-(\alpha + \beta \mathbf{X}_i)]}.$$
- π_i关于$\mathbf{X}_i$求导数【黑板】，得到在$\mathbf{X}_i$处的斜率为$\pi_i(1 - \pi_i)\beta$.
- 因此，拟合图的导数是$\pi_i(1 - \pi_i)\beta$. 这给出了参数β 的第三种解释。如果$\mathbf{X} = \mathbf{x_i}$，若$\mathbf{X}$增加微小的$\epsilon$, 则$\pi_i$ 将增加$\epsilon\pi_i(1 - \pi_i)\beta$。
- 见斜率表。注意，斜率在$\pi = 0.2$和$\pi = 0.8$之间保持稳定。在此范围内，S曲线接近直线。
- 我们不解释α。
- 所有这些是如何对智利数据产生影响的？

多重logisitic回归

$$
\begin{aligned}
\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha + \beta_1 \mathbf{X}_{i1} + \cdots + \beta_k \mathbf{X}_{ik} \\
\frac{\pi_i}{1-\pi_i} &= \exp(\alpha + \beta_1 \mathbf{X}_{i1} + \cdots + \beta_k \mathbf{X}_{ik}) \\
&= \exp(\alpha)\exp(\beta_1 \mathbf{X}_{i1}) \cdots \exp(\beta_k \mathbf{X}_{ik}) \\
&= \exp(\alpha)[\exp(\beta_1)]^{\mathbf{X}_{i1}} \cdots [\exp(\beta_k)]^{\mathbf{X}_{ik}} \\
\pi_i &= \frac{1}{1+\exp(\alpha + \beta_1 \mathbf{X}_{i1} + \cdots + \beta_k \mathbf{X}_{ik})}
\end{aligned}
$$

协变量的类型

- **X**可以与线性回归一样通用：
  - ✓ 定量变量
  - ✓ 定量变量的变换
  - ✓ 定性变量的虚拟回归量
  - ✓ 交互回归

logisitic回归的检验

- Wald检验(类似于t-检验)
- 似然比检验(类似于F-检验)
  - ✓ 全模型m_1
  - ✓ 空模型m_0 (全模型的特殊情形)
  - ✓ 对于这两个模型计算似然：L_1和L_0. $L_1 \geq L_0$. 为什么？