《线性回归》
—广义线性模型(GLM)

杨　　瑛

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

2019.06.11

主要内容：广义线性模型（GLM）

1. **广义线性模型(GLM)**
   - GLM：结构
   - 系统部分
   - 定义GLM
   - 利用变换建立近似的GLM
   - GLM: 估计
   - GLM: 推断
   - GLM: 诊断
   - 特殊类型数据的统计分析

- 在线性模型中通常假定响应变量的方差为常数。
- 而在广义线性模型（GLM）中假定响应值来自于更一般的分布族。我们将讨论非常有用的离散度模型(dispersion model). 广义线性模型是回归模型，其由随机部分和系统部分组成。在线性模型$y_i = x_i^T\beta + \epsilon_i$中，$E[y_i|x_i](= x_i^T\beta)$是系统部分, Var$(\epsilon_i) = \sigma^2$是随机部分。通常随机部分和系统部分都是未知的，需要利用数据估计出来.
- 广义线性模型中的随机部分和系统部分取什么样的形式，取决于要回答的问题：
  - ✓ 什么概率分布是合适的?答案决定了模型的随机部分。概率分布的选择可以由响应数据的类型来确定或由方差如何随均值变化的知识来确定。
  - ✓ 解释变量与响应变量的均值μ是如何相关联的? 其答案可以说明模型的系统部分是如何构成。GLM假设一个函数连接线性预测量$\eta = \beta_0 + \sum_{j=1}^p \beta_j x_j$与均值$\mu$之间的关系。例如，$\log(\mu) = \eta$. 即，GLM是关于参数线性的回归模型。

随机部分：指数离散度模型(Exponential Dispersion Models (EDM))

- EDM族的分布假定具有形式（y可以是连续的或者离散的）

$$\mathcal{P}(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta-\kappa(\theta)}{\phi}\right\}, \tag{1}$$

其中

- ✓ θ是典型参数(canonical parameter);
- ✓ $\kappa(\theta)$是已知的函数，称为半不变量函数(cumulant function);
- ✓ $\phi > 0$称为离散度参数(dispersion parameter);
- ✓ $a(y,\phi)$是正则化函数，保证(1)是概率密度函数;

根据具体的分布，可以确定支撑集合和参数空间.

例子

Example (1)

均值为μ, 方差为σ^2的正态密度函数是：

$$\begin{aligned}\mathcal{P}\left(y;\mu,\sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{y\mu-\left(\mu^2/2\right)}{\sigma^2}-\frac{y^2}{2\sigma^2}\right\}. \qquad (2)\end{aligned}$$

很容易看出，$\theta=\mu$是典型参数，$\kappa(\theta)=\mu^2/2=\theta^2/2$是半不变量，$\phi=\sigma^2$是离散度参数，
$a(y,\phi)=\left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-y^2/\left(2\sigma^2\right)\right\}$是正则化函数。正态分布是EDM.

Example (2)

Poisson概率密度函数是：

$$\mathcal{P}(y;\mu) = \frac{\exp(-\mu)\mu^y}{y!}, \mu > 0, y = 0, 1, 2, \cdots$$

可以表示为(1)的形式：

$$\mathcal{P}(y;\mu) = \exp\{y \log \mu - \mu - \log(y!)\}.$$

很容易看出，$\theta = \log \mu$是典型参数，$\kappa(\theta) = \mu$是半不变量，$\phi = 1$是离散度参数，
$a(y, \phi) = 1/y!$是正则化函数。Poisson分布是EDM.

Example (3)

二项概率密度函数是：

$$\begin{aligned}\mathcal{P}(y;\mu,m) &= \binom{m}{my}\mu^y(1-\mu)^{m(1-y)} \\ &= \binom{m}{my}\exp\left[m\left\{y\log\frac{\mu}{1-\mu}+\log(1-\mu)\right\}\right],\end{aligned} \tag{3}$$

其中$y=0,1/m,2/m,\cdots,1$, $0<\mu<1$ 比较(1), 得到：$\theta=\log\{\mu/(1-\mu)\}$是典型参数，$\kappa(\theta)=-\log(1-\mu)$是半不变量，$\phi=1/m$是离散度参数，$a(y,\phi)=\binom{m}{my}$是正则化函数。当$m$已知时，二项分布是EDM.

Example (4)

Weibull分布的概率密度函数是：

$$\mathcal{P}(y;\alpha,\gamma)=\frac{\alpha}{\gamma}\left(\frac{y}{\gamma}\right)^{\alpha-1}\exp\left\{-\left(\frac{y}{\gamma}\right)^{\alpha}\right\}, y>0, \alpha>0, \gamma>0.$$

可以表示为(1)的形式：

$$\mathcal{P}(y;\alpha,\gamma)=\exp\left\{-(y/\gamma)^{\alpha}+\log(\alpha/\gamma)+(\alpha-1)\log y/\gamma\right\}.$$

在指数函数内，残差$y\theta$的形式得不出来，除非$\alpha=1$. 因此，一般来说，Weibull分布不是EDM。当$\alpha=1$时，概率密度函数是：

$$\mathcal{P}(y;\gamma)=\exp(-y/\gamma)/\gamma=\exp\{-(y/\gamma)-\log\gamma\},$$

这是指数分布。很容易看出，$\theta=-1/\gamma$是典型参数，$\kappa(\theta)=\log\gamma$是半不变量，$\phi=1$是离散度参数.

EDM的性质

- EDM有很多重要的性质:
  - ✓ 矩母函数（MGF）有很简单的形式：

$$M(t) = \mathrm{E}\left[e^{ty}\right] = \begin{cases} \int_S \mathcal{P}(y) e^{ty} dy, & \text{当}y\text{ 连续} \\ \sum_{y \in S} \mathcal{P}(y) e^{ty}, & \text{当}y\text{ 离散}, \end{cases}$$

其中$t \in \{M(t) < \infty\}$.

  - ✓ 半不变母函数(cumulant generating function, CGF)定义为：

$$K(t) = \log M(t) = \log \mathrm{E}\left[e^{ty}\right].$$

  - ✓ 第r阶半不变量是：$\kappa_r = \left.\frac{\mathrm{d}^r K(t)}{\mathrm{d}t^r}\right|_{t=0}$.
  - ✓ 利用CGF，可以得到均值和方差是：

$$\mathrm{E}[y] = \kappa_1 = \left.\frac{\mathrm{d}K(t)}{\mathrm{d}t}\right|_{t=0}, \quad \mathrm{var}[y] = \kappa_2 = \left.\frac{\mathrm{d}^2 K(t)}{\mathrm{d}t^2}\right|_{t=0}.$$

系统部分：link 函数

- GLM假定系统部分是线性预测量

$$\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j,$$

  它通过link函数$g()$与均值联系起来，$g(\mu) = \eta$. 这一系统部分说明GLM是关于参数线性的回归模型.
- link函数$g(\cdot)$是关于μ的单调和可微函数。可微性主要是为了估计目的。

offsets

在一些应用中，线性预测变量中有一个量是不需要估计的，这个量称之为offsets. 在线性预测变量中，$\beta_j x_{ji}$看作offset，则β_j是已知的. 在R的函数lm()或者glm()中都有offset的选项，注意使用方法.

定义GLM

- GLM由系统部分和随机部分组成：
  - ✓ 随机部分：观测值y_i独立地来自于一个确定的EDM，使得$y_i \sim \text{EDM}(\mu_i, \phi/w_i), i = 1, \cdots, n$. 其中$w_i$是已知非负的先验权。通常，所有的权都等于1.
  - ✓ 系统部分：线性预测变量是$\eta_i = o_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}$，其中$o_i$是offsets，通常都等于0, $g(\mu) = \eta$是已知的单调、可微的link函数.
- GLM定义为：

$$\begin{cases} y_i \sim \text{EDM}\left(\mu_i, \phi/w_i\right) \\ g\left(\mu_i\right) = o_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}. \end{cases} \tag{4}$$

GLM的核心结构是从EDM类中根据具体情况**选定分布和link函数**。

定义GLM (续)

- 给定刻画随机部分的EDM，以及建立均值μ与解释变量之间联系的link 函数，就可以制定GLM：

$$\text{GLM(EDM; link函数)}.$$

Example

y_i表示是否患病，x_i表示是否吸烟，$i = 1, \cdots, n$. 则

$$\begin{cases} y_i \sim \text{Binomial}\,(1, \mu_i) & \text{(随机部分)} \\ \log \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 x_i & \text{(系统部分)} \end{cases}.$$

这是前面已经讲过的logistic 模型. 在R中可以利用GLM 来实现，但是要说明'family('binomial',link='logit')'.

方差稳定化方法

- 利用在《统计推断》中学习过方差稳定化方法，当已知方差结构V(μ) 之后【这里的V(μ)表示响应变量均值和方差之间的关系】，使变换响应$h(y)$的方差为常数, $h(\cdot)$是变换。
- 下面列出一些常见的变换。

方差稳定化变换列表

稳定化变换 (Box-Cox变换中的λ)	近似的GLM 方差函数	link函数
$y* = \sin^{-1}(\sqrt{y})$	$V(y) = \mu(1-\mu)$ (二项GLM)	$g(\mu) = \sin^{-1}(\mu)$
$y* = \sqrt{y}(\lambda = 1/2)$	$V(\mu) = \mu$ (Poisson GLM)	$g(\mu) = \sqrt{\mu}$
$y* = \log(y)(\lambda = 0)$	$V(\mu) = \mu^2$ (gamma GLM)	$g(\mu) = \log(\mu)$
$y* = 1/\sqrt{y}(\lambda = -\frac{1}{2})$	$V(\mu) = \mu^3$ (逆高斯GLM)	$g(\mu) = 1/\sqrt{\mu}$
$y* = 1/y(\lambda = 1)$	$V(\mu) = \mu^4$ (Tweedie GLM)	$g(\mu) = 1/\mu$

上面主要讨论的是如何建立GLM的一般原则。

GLM中参数估计概要：

- 这里非常简要的概述GLM中未知参数(包括回归参数和离散度参数)的估计问题。因为GLM中假定了响应值的概率分布来自于EDM，因此，MLE可以用来求出参数估计。主要内容包括：
  - ✓ 推导出得分方程和GLM情形下的Fisher信息矩阵.
  - ✓ 计算回归参数的算法.
  - ✓ 拟合模型之后，定义residual deviance【作用？】.
  - ✓ 计算回归参数的se【作用？】.
  - ✓ 离散度参数的估计【类似于线性模型中σ^2的估计】.
  - ✓ 如何评价估计量的性质【相合性，渐近正态性，$\cdots$】.
  - ✓ R中的实现：函数glm(). 格式：例如，GLM(Poisson, log), 需要制定发布和link函数. 当然glm()中还有更多的选项，请help(glm)和example(glm)查看glm使用的使用方法和示例.
- 更多细节请参考：
  Peter K. Dunn and Gordon K. Smyth (2018). Generalized Linear Models With Example in R. New York: Springer. (Chapter 6).

GLM中参数推断概要：

- 这里非常简要的概述GLM中未知参数(包括回归参数和离散度参数)的推断问题。因为GLM中假定了响应值的概率分布来自于EDM，因此，基于似然理论的推断方法：Wald检验，score检验，LRT可以完全用于GLM的参数推断。主要内容包括：
  - ✓ 当离散度参数ϕ已知时，利用大样本渐近结果推断.
  - ✓ 拟合优度检验，确定线性预测变量是否充分的描述了数据中的系统趋势.
  - ✓ 考虑离散度参数ϕ未知时, 参数的推断问题，Wald检验，score 检验，LRT.
  - ✓ 非嵌套模型的比较.
  - ✓ GLM中的变量选择.
- 更多细节请参考：
  Peter K. Dunn and Gordon K. Smyth (2018). Generalized Linear Models With Example in R. New York: Springer. (Chapter 7).

GLM：诊断

- 像线性模型要对模型的各种假设做诊断一样，glm的建模过程中也有各种假设，需要对模型进行诊断，在模型的假定违反和不成立时，提供可能的解决方案。线性模型诊断的基本工具是各种残差。在glm中也要定义各种类型的残差以实现对模型的诊断。
  - ✓ 首先要搞清楚glm的假设.
  - ✓ 定义残差的三种基本类型(Pearson、偏差和分位数).
  - ✓ 检查模型假设的各种诊断工具.
  - ✓ 检查异常值和有影响的观察结果的技术.
  - ✓ 诊断共线性的方法.
- 更多细节请参考：
  Peter K. Dunn and Gordon K. Smyth (2018). Generalized Linear Models With Example in R. New York: Springer. (Chapter 8).

- 下面几种重要类型的数据也是回归分析的内容：
  - ✓ 比例数据（proportions）
  - ✓ 计数数据(counts)
  - ✓ 正的连续数据(positive continuous data)
- 更多细节请参考：
  Peter K. Dunn and Gordon K. Smyth (2018). Generalized Linear Models With Example in R. New York: Springer. (Chapters 9-12).