广义线性模型GLM

线性回归模型

- 古典的线性回归模型：

$$Y = X\beta + \varepsilon$$

- ε为随机误差项，一般假设为零均值、方差同为为σ^2、且相互独立的正态随机向量。
- u 为因变量均值，一般解释为响应变量线性和的形式：

$$u = X\beta$$

- β 为待估的线性组合参数

- 目标函数：即估计参数的目标，或是偏差最小，或是无偏最小方差等等，在此一般采用的是最小二乘和极大似然。
- 最小二乘：即使得因变量的真值y与拟合值 $X\hat{\beta}$ 间距离最小。
- 矩阵求导基本公式：

$$Y = AX \rightarrow \frac{dY}{dX} = A^T$$

$$Y = XA \rightarrow \frac{dY}{dX} = A$$

$$Y = X^T A \rightarrow \frac{dY}{dX} = A$$

$$\left\| y - X\beta \right\|^2 = (y - X\beta)^T (y - X\beta) = y^T y - y^T X\beta - \beta^T X^T y - \beta^T X^T X\beta$$

- 对β 求导得到：

$$-X^T y - X^T y - X^T X\beta - X^T X\beta = -2(X^T y - X^T X\beta)$$

- 使之为零便可到最小二乘估计：

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- 极大似然
- 对于单个的因变量 $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p + \varepsilon_i$
- 都是同方差的正态随机变量，因此可以容易的得到整体的对数似然函数：

$$l(\beta, \sigma^2) = -n\ln\sigma - \frac{n}{2}\ln 2\pi - \frac{1}{2}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{\sigma})^2$$

- 显然想要最大化似然值，只需最小化 $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- 即可，这也正是前面最小二乘中的距离，故此处极大似然也最小二乘是等价的。

- 估计的性质
  - 无偏性：$E[\hat{\beta}] = E[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T E[y] = \beta$
  - 协方差：

    $$\mathrm{cov}(\hat{\beta}) = (X^T X)^{-1} X^T \,\mathrm{cov}(y)[(X^T X)^{-1} X^T]^T = \sigma^2 (X^T X)^{-1}$$

  - 因此 $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$
- 基本的检验：帽子矩阵H
-

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$$

- H阵的性质：对称而且是幂定的 $H^n = H$
- 考虑拟合的残差：

  $$\hat{\varepsilon} = y - \hat{y} = (I - H)\varepsilon \Rightarrow Var(\hat{\varepsilon}) = \sigma^2(I - H)$$

- 因此H阵的对角元都是0~1的，并且由于幂定阵的一个性质 $rank(H) = trace(H)$，即对角元相加为P。由此可以期望每一个观测值对其预测值的贡献度大约为P/n，对于过大的贡献度则可以认为其存在异常。

- 加权的最小二乘：来源于各个 ε_i 的异方差情形，自然而言，对于方差越小的，我们可以认为其精度较高，在拟合的时候自然考虑增加其比重，反之亦然。
- 假设 $Var(\varepsilon_i) = \sigma^2 / w_i$ ，于是考虑作如下变换

$$y_i^* = \sqrt{w_i}\, y_i, x_{ij}^* = \sqrt{w_i}\, x_{ij,} \varepsilon_i^* = \sqrt{w_i}\, \varepsilon_i$$

- 得到：

$$y_i^* = x_{i1}^* \beta_1 + x_{i2}^* \beta_2 + \ldots + x_{ip}^* \beta_p + \varepsilon_i^*$$

- 再由前面的最小二乘法可得：

$$\hat{\beta}^{*} = (X^{*T}X^{*})^{-1}X^{*T}y^{*} = (X^{T}WX)^{-1}X^{T}Wy$$

- 其中权数矩阵为：

$$W = \begin{pmatrix} w_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{nn} \end{pmatrix}$$

- 和前面一样也有

$$\hat{\beta}^{*} \sim N(\beta, \sigma^{2}(X^{T}WX)^{-1})$$

- 古典线性模型的不足
  - 正态性假设在实践中并不满足
  - 保险实践中，因变量往往为非负，比如赔款次数和数量，这点在正态性假设也无法满足
  - 同方差的假设
  - 解释变量只能通过加法对因变量产生影响

广义线性模型基本理论

- 较之前面的一般线性模型，广义线性模型主要作了两个方面的拓展：
  - 首先是因变量不再只是正态分布，而是夸大为指数分布族中的任一分布
  - 解释变量的线性组合不再直接用于解释因变量的均值u，而是通过一个连接函数g来解释g(u),这里要求连接函数单调可导
- 可以看出，广义线性模型是在古典模型上的一个的推广

- 指数分布族

$$f(y_i;\theta_i,\phi)=\exp\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\}$$

- 其中a,b,c均为已知函数，他们对所有的观测值有相同的形式，且满足以下条件
  - $a(\phi)$ 大于零，连续，通常形式为$\frac{\phi}{w}$ ，其中w为已知先验权重
  - $b(\theta_i)$ 二阶导数存在且大于零
  - $c(y_i,\phi)$ 与参数θ_i 无关

- 指数分布族的均值与方差

  均值：在密度函数两边对θ_i求导，得到

  $$\frac{df}{d\theta_i} = f \times \{\frac{y_i - b'(\theta_i)}{a(\phi)}\}$$然后两边对y_i进行积分有：

  $$0 = \frac{E(y_i) - b'(\theta_i)}{a(\phi)} \Rightarrow E(y_i) = b'(\theta_i)$$

  方差：同法可得

  $$Var(y_i) = b''(\theta_i) \times a(\phi)$$

- 连接函数：单调可导，即可逆的函数。
  - 恒等，g(u)=u，古典线性模型即是正态分布下的恒等连接的广义线性模型
  - 对数，g(u)=ln(u)，因为对数的逆是指数，因此它可将原本的线性和组合转变为乘积的关系
  - Logit，g(u)=ln(u/1-u)，它的特点为可将预测值控制在0~1之间，对于因变量为比率时适合使用

- 参数估计：极大似然和加权最小二乘
- 极大似然：由于已经知道了密度函数，所以可以容易的得到对数似然函数，只是这里的求解会比古典模型复杂，主要为几个分步的求导过程和最后的泰勒逼近于牛顿的迭代算法。
- 加权最小二乘：古典模型中，目标函数为 $\left\| y - X\beta \right\|^2$ 而且y为同方差的随机向量。自然可以想到，广义模型中的目标函数可以为

$$\left\| g(y) - X\beta \right\|^2$$

- 但是需要注意，这里的g(y)已经不再是同方差的，所以需要用到前面的加权最小二乘，首先考虑g(y)的方差，由泰勒展开可得：

$$g(y_i) \approx g(u_i) + g'(u_i)(y_i - u_i)$$

- 将各个分量整合起来写成矩阵的形式有：

$$g(y) \approx g(u) + G(y - u)$$

$$G = \begin{pmatrix} g'(u_1) & & 0 \\ & \ddots & \\ 0 & & g'(u_n) \end{pmatrix}$$

- 于是可以得到g(y)的方差为：

$$Var(g(y)) = GVar(y)G$$

- 采用前面讲到的加权最小二乘即可得到估计为：

$$b \approx \left(X^T W X \right)^{-1} X^T W g(y)$$

- 这里的W为权矩阵：

$$W = \begin{pmatrix} [g'^2(u_1)b''(\theta_1)]^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & [g'^2(u_n)b''(\theta_n)]^{-1} \end{pmatrix}$$

- 构造迭代算法：如前面提到的泰勒逼近，g(y)有如下的形式：

$$g(y) \approx g(u) + G(y-u)$$

- 因此可以构造迭代：

$$b^{m+1} = \left(X^T W X\right)^{-1} X^T W (g(u) + G(y-u))$$

- 将g(u)用上一步的估计值代替即得到最终的迭代：

$$b^{m+1} = \left(X^T W X\right)^{-1} X^T W (Xb^m + G(y-u))$$

- 到此参数的估计基本完成，接下来看看该估计的分布性质，即均值和方差。
- 均值：

$$E[b] = \left(X^T W X\right)^{-1} X^T W E[g(y)]$$

$$\approx \left(X^T W X\right)^{-1} X^T W E[g(u) + G(y-u)]$$

$$= \left(X^T W X\right)^{-1} X^T W g(u)$$

$$= \left(X^T W X\right)^{-1} X^T W X \beta = \beta$$

- 方差：

$$Var[b] = \left(X^T W X\right)^{-1} X^T W Var(g(y))[\left(X^T W X\right)^{-1} X^T W]^T$$
$$= a(\phi)\left(X^T W X\right)^{-1}$$

- 因此在大数据量的情况下，由中心极限可以有如下近似：

$$b \sim N(\beta, a(\phi)\left(X^T W X\right)^{-1})$$

拟合效果检验

- 偏差的检验：将设定模型与饱和模型进行比较。饱和模型是在同种假设、同种连接函数下，拥有最大数量待估参数的模型。好的设定模型应该表现出与饱和模型的差异很小。
- 偏差的定义：

$$D = 2[l(b_{\max}; y) - l(b; y)]$$

- 通过D便可以度量设定模型和饱和模型间的差异，下面考虑D的分布情况。

- 首先将D改写为：

$$D = 2[l(b_{\max}; y) - l(\beta_{\max}; y)] - 2[l(b; y) - l(\beta; y)] + 2[l(\beta_{\max}; y) - l(\beta; y)]$$

- 容易看出，右边的第三为一个与拟合情况有关的正常数，右边的第一和第二项有着相同的结构。
- 考虑$l(b; y) - l(\beta; y)$ 的分布，将$l(\beta; y)$ 在 b 点泰勒展开：

$$l(\beta) = l(b) + (\beta - b)\frac{dl}{d\beta} + \frac{1}{2}(\beta - b)^2 \frac{d^2 l}{d^2 \beta}$$

$$l(\beta) = l(b) - \frac{1}{2}(\beta - b)^2 \frac{1}{a(\phi)} X^T W X$$

$$\Rightarrow l(b) - l(\beta) = \frac{1}{2}(\beta - b)^T \frac{1}{a(\phi)} X^T W X (\beta - b)$$

- 由于$\frac{1}{a(\phi)} X^T W X$为b的方差阵的逆，因此大数据量下有：

$$2[l(b) - l(\beta)] \sim x^2(p)$$

- 同理有： $2[l(b_{\max})-l(\beta_{\max})]\sim x^2(n)$
- 因此D的近似分布为：

$$x^2(n-p)$$

- 模型假设的检验：Anscombe残差和Deviance残差
  - Anscombe残差：通过指数分布族的正态化函数A(.)作用于y将其正态化，残差表示为：

$$A(\bullet)=\int\frac{1}{V^{1/3}(u)}du \qquad \frac{A(y_i)-A(\hat{y}_i)}{A'(\hat{y}_i)\sqrt{V(\hat{y}_i)}}$$

– Deviance残差，y_i的Deviance残差定义为

$$r_{Di} = sign(y_i - u_i)\sqrt{d_i}$$

其中d_i为前面偏差D的第i部分，即：

$$D = \sum d_i$$

上述两个残差，均近似服从正态分布，可将标准化的残差与标准正态分布对比，以作检验。

- 最终选择模型的信息准则
  - AIC准则，即对参数个数进行惩罚：

$$AIC = -2l + 2p$$

  - BIC准则，同时对参数个数和样本量进行惩罚：

$$BIC = -2l + p(\ln n)$$

常用的广义线性模型

- 由先验信息选择分布类型
  常数方差 → 正态分布
  方差等于均值 → 泊松分布
  方差等于均值的平方 → 伽马分布
  方差等于均值的三次方 → 逆高斯分布
- 非寿中常用广义线性模型
  - 索赔次数或索赔频率时，泊松分布和负二项分布，对数连接
  - 索赔强度时，伽马分布和逆高斯分布，对数连接
  - 事故发生率或续保率时，logistic回归

迭代法与GLM的比较

- 迭代法
  - 优点：理论简单易懂，也方便软件计算
  - 缺点：
    - 缺乏完整的统计分析框架
    - 分类变量较多时，计算复杂
    - 费率因子必须是离散的
- 推广的迭代法
  - 无需分布假设，提高了灵活性
  - 同样的易于理解
  - 克服了迭代法中变量多时计算复杂的问题
  - 可以解决混合模型和有限制条件的模型

- GLM
  - 优点
    - 统计模型，有完整的分析框架
    - 同时考虑所有的因素
    - 方便实现，许多软件都能调用
  - 缺点
    - 目前广泛应用的模型局限于指数分布族
    - 无法解决混合模型