第三章　回归分析

3.1　简 介

3.1 简 介

- 回归 (regression) 这一概念最早由英国生物统计学家高尔顿 (F.Galton)和皮尔逊 (K.Pearson) 在研究父母和子女的身高遗传特性时提出. 回归分析是处理多个变量间相关关系的一种数学方法, 是机器学习中的监督学习方法.
- 回归模型是一种统计预测模型, 需要预测的变量叫做因变量 (或响应变量 (response variable)), 用来解释因变量变化的变量叫做自变量 (或协变量(covariate)、特征 (feature)). 回归模型通过若干个自变量来预测响应变量.
- 从响应变量个数来看, 回归模型中可以有一个响应变量, 也可以有多个响应变量. 从响应变量取值来看, 回归模型中的响应变量可以是连续变量, 也可以是离散变量.
- 一般的回归模型适用于单个响应变量、取值为连续型且一般服从正态分布的情形. 广义线性模型 (generalized linear models, GLM) 适用于单个响应变量、取值是离散型且分布服从指数型分布族的情形. 多元响应变量协方差广义线性模型 (multivariate covariance generalized linear model, McGLM)是广义线性模型的扩展, 适用于服从非正态分布且不独立的多个响应变量的情形.

3.1 简 介

- 本章主要介绍单响应变量的线性回归模型、单响应变量的广义线性模型和多元响应变量协方差广义线性模型.

3.2　单响应变量的线性回归模型

3.2 单响应变量的线性回归模型

- 本节主要介绍单响应变量的线性回归模型的原理、自变量的选择、多重共线性和岭回归.

3.2.1 线性回归模型的原理

一、线性回归模型的一般形式

- **定义 3.1** p 个自变量 $X_1, X_2, \cdots, X_p$ 对单个响应变量 Y 的线性回归模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{3.2.1}$$

  - 式中, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 是 $p+1$ 个未知参数. β_0 称为回归常数, $\beta_1, \beta_2, \cdots, \beta_p$ 为回归系数.
  - 当 $p = 1$ 时, 式 (3.2.1) 称为一元线性回归模型; 当 $p \geqslant 2$ 时, 式(3.2.1) 称为 p 元线性回归模型.
  - 式中 ε 是随机误差, 常假定 $E(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$.

- 若对单个响应变量的线性回归模型, 观测到样本数据为 $\{(X_{i1}, X_{i2}, \cdots, X_{ip}; Y_i), \mathrm{i} = 1, 2, \cdots, \mathrm{n}\}$, 则基于 n 个样本的线性回归模型可表示为

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \varepsilon_1, \\ Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \varepsilon_2, \\ \cdots\cdots\cdots\cdots \\ Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \varepsilon_n. \end{cases}$$

3.2.1 线性回归模型的原理

一、线性回归模型的一般形式

- 写成矩阵形式为

$$\boldsymbol{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}. \tag{3.2.2}$$

  - 式 (3.2.2) 中 $\boldsymbol{X}$ 为设计矩阵

$$\boldsymbol{Y}=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \mathbf{X}=\begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1\mathrm{p}} \\ 1 & X_{21} & X_{22} & \cdots & X_{2\mathrm{p}} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{\mathrm{n}1} & X_{\mathrm{n}1} & \cdots & X_{\mathrm{np}} \end{pmatrix}, \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{\mathrm{p}} \end{pmatrix}, \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{\mathrm{n}} \end{pmatrix}.$$

3.2.1 线性回归模型的原理

二、线性回归模型的基本假设

- 为了方便地进行模型的参数估计及检验, 对线性回归模型进行如下基本假设:
  - 1. 自变量 $X_1, X_2, \cdots, X_p$ 是确定性变量,而不是随机变量. 且 $\mathrm{rank}(\mathbf{X}) = p + 1 < \mathrm{n}$，即设计矩阵 $\mathbf{X}$ 是列满秩矩阵， 其列之间不线性相关.
  - 2. 随机误差项满足高斯–马尔可夫条件, 即在给定自变量取值的情况下,随机误差项满足零均值、同方差及序列不相关.

$$E(\varepsilon_i) = 0,\ i = 1,\ 2,\ \cdots,\ n,$$

$$\mathrm{Cov}\left(\varepsilon_{\mathrm{i}}, \varepsilon_{\mathrm{j}}\right) = \begin{cases} \sigma^2, & \mathrm{i} = \mathrm{j}, \\ 0, & \mathrm{i} \neq \mathrm{j}, \end{cases} \quad \mathrm{i}, \mathrm{j} = 1, 2, \cdots, \mathrm{n}. \tag{3.2.3}$$

3.2.1 线性回归模型的原理

二、线性回归模型的基本假设

- 为了方便地进行模型的参数估计及检验, 对线性回归模型进行如下基本假设:
  - 3.误差满足正态分布假定, 即 $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \cdots, n$, 且 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 相互独立. 矩阵形式为

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n) \tag{3.2.4}$$

  - 其中 $\boldsymbol{I}_n$ 是 $n \times n$ 单位矩阵.

- 其中条件 1 和 2 是为了最小二乘估计, 条件 3 是为了假设检验过程中研究估计量的分布, 当然条件 3 也可以用于参数的最大似然估计.

3.2.1 线性回归模型的原理

三、参数的估计

- 下面利用最小二乘法求线性回归模型 $\boldsymbol{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 中回归参数 $\boldsymbol{\beta}$ 的最小二乘估计量.

- 最小二乘法 (ordinary least square method, OLS) 就是寻找未知参数$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 的估计值 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots \hat{\beta}_p$ 使得响应变量的离差平方和

$$Q\left(\beta_0, \beta_1, \beta_2, \cdots \beta_p\right)=\sum_{i=1}^{n}\left[Y_i-\left(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}\right)\right]^2$$

达到最小，即满足

$$\begin{aligned}
&Q\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots \hat{\beta}_{\mathrm{p}}\right) \\
&=\sum_{i=1}^{n}\left[Y_i-\left(\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\hat{\beta}_2 X_{i2}+\cdots+\hat{\beta}_p X_{ip}\right)\right]^2 \\
&=\min_{\beta_0, \beta_1, \beta_2, \cdots \beta_p} \sum_{i=1}^{n}\left[Y_i-\left(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}\right)\right]^2 .
\end{aligned}$$

3.2.1 线性回归模型的原理

三、参数的估计

- 根据微积分求极值的原理, 只需求 $Q(\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ 的最小值点. 将 Q 关于 $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 分别求偏导数, 并令其为零, 得到正规方程组

$$
\begin{cases}
\left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum\limits_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip}\right)\right]^2 = 0, \\
\left.\dfrac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum\limits_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip}\right)\right]^2 X_{i1} = 0, \\
\cdots\cdots\cdots\cdots \\
\left.\dfrac{\partial Q}{\partial \beta_p}\right|_{\beta_p=\hat{\beta}_p} = -2\sum\limits_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip}\right)\right]^2 X_{ip} = 0.
\end{cases}
$$

3.2.1 线性回归模型的原理

三、参数的估计

- 整理得矩阵形式的正规方程组

$$\left(\mathbf{X}^{\mathrm{T}}\,\mathbf{X}\right)\hat{\boldsymbol{\beta}}=\mathbf{X}^{\mathrm{T}}\,\boldsymbol{Y}$$

- 由 $\mathbf{X}$ 的列满秩性可知, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 为满秩对称矩阵, 从而回归参数 $\boldsymbol{\beta}$ 的最小二乘估计为

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^{\mathrm{T}}\,\boldsymbol{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\,\boldsymbol{Y} \tag{3.2.5}$$

- 根据唐年胜, 李会琼 [31], 可以证明, 在高斯—马尔可夫假定 (3.2.3) 下最小二乘估计 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的最小方差线性无偏估计. 且在假定 (3.2.4) 下, 我们可以推导出

$$\hat{\boldsymbol{\beta}}\sim N_{p+1}\left(\boldsymbol{\beta},\sigma^{2}\left(\mathbf{X}^{\mathrm{T}}\,\Xi\right)^{-1}\right). \tag{3.2.6}$$

  - 证明可详见唐年胜, 李会琼 [31].

3.2.1 线性回归模型的原理

三、参数的估计

- **定义 3.2** 对于线性回归模型 (3.2.1), 称

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\cdots+\hat{\beta}_pX_p$$

为样本回归方程, 称 $\hat{Y}$ 为响应变量的拟合值, 称 $e=Y-\hat{Y}$ 为回归残差.

- **定义 3.3** 最小二乘法没有给出未知参数σ^2的估计.利用最大似然比原理可得 $\boldsymbol{\beta}$最大似然估计仍为 $\hat{\boldsymbol{\beta}}$,同时给出σ^2的最大似然估计为

$$\frac{\left(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n}=\frac{1}{n}\sum_{i=1}^{n}e_i^2=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \tag{3.2.7}$$

  - 因为上述估计有偏, 通常取

$$\hat{\sigma}^2=\frac{\left(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-p-1}=\frac{1}{n-p-1}\sum_{i=1}^{n}e_i^2=\frac{1}{n-p-1}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \tag{3.2.8}$$

  - 作为 σ^2 的无偏估计, 称 $\hat{\sigma}$ 为回归标准差

3.2.1 线性回归模型的原理

四、显著性检验

1. 回归方程的显著性检验：*F*检验

(1) 平方和分解式

- 下面基于方差分析的思想, 从数据出发来研究响应变量数据变异的原因.
- 数据 $Y_1, Y_2, \cdots, Y_n$ 的波动大小可用总离差平方和$\mathrm{SST}=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$ 来度量, 其中 $\overline{Y}=\sum_{i=1}^{n}Y_i/n$ ， 它反映响应变量 Y 的波动程度或不确定性.
- 称 $\mathrm{SSR}=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2$ 为回归平方和, 它是由回归方程确定的, 是由自变量的波动引起的, 反映了自变量解释响应变量波动的贡献.
- 称 $\mathrm{SSE}=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$ 为残差平方和, 它是不能由自变量解释的波动, 是由自变量之外的因素引起的.
- 易证 $\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE}$, 称此式为平方和分解式.
- 平方和分解式表明, 总离差平方和 SST 中能够由自变量解释的部分为SSR, 不能由自变量解释的部分为SSE. 因此, 回归平方和越大, 回归的效果就越好

3.2.1 线性回归模型的原理

四、显著性检验

1. 回归方程的显著性检验：*F*检验

(2) *F*检验统计量

- 在正态分布假设 (3.2.4) 下, 需要检验自变量 $X_1, X_2, \cdots, X_p$ 从整体上对因变量 Y 是否有显著的影响, 为此提出:
  - 原假设 $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0;$$
  - 备择假设 $H_1 : \beta_1, \beta_2, \cdots, \beta_p$至少有一个不为零.
  - 在原假设 H_0 下, 构造检验统计量 $$F = \frac{SSR / p}{SSE /(n-p-1)} \sim F(p, n-p-1). \tag{3.2.9}$$

3.2.1 线性回归模型的原理

四、显著性检验

1. 回归方程的显著性检验：*F* 检验

(2) *F* 检验统计量

- 给定显著性水平 α $(0 < \alpha < 1)$, 右侧 F 检验的临界值为 $F_{\alpha}(\mathrm{p}, n-p-1)$(上侧 α 分位数). 当检验统计量 F 的观测值落在拒绝域, 即 $F > F_{\alpha}(\mathrm{p}, n-p-1)$ 时, 拒绝原假设, 说明回归方程显著, $\boldsymbol{X}$ 与 Y 有显著的线性关系; 否则,只能接受原假设, 认为回归方程不显著.

- 也可以利用 p 值法进行判断, 其中 $p = P(Z > F)$, $Z \sim F(p, n-p-1)$,当 $p < \alpha$ 时, 拒绝原假设; 否则, 只能接受原假设.

- 通常利用方差分析表 3.1 进行 F 检验.

方差来源	自由度	平方和	均方	F 值	p 值
回归	p	SSR	SSR/p	$\frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)}$	
残差	$n-p-1$	SSE	SSE/$(n-p-1)$		
总和	$n-1$	SST			

◀ 表 3.1
方差分析表

3.2.1 线性回归模型的原理

四、显著性检验

2. 回归系数的显著性检验：t 检验

- 在多元线性回归分析中, 回归方程显著并不意味着每个自变量对 Y 的影响都显著, 所以需要对每个自变量进行显著性检验. 为此, 提出
  - 原假设 $$H_{0j}: \beta_j = 0, j = 1, 2, \cdots, p;$$
  - 备择假设 $$H_{1j}: \beta_j = 0, j = 1, 2, \cdots, p.$$
  - 因 $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots \hat{\beta}_p\right)^{\mathrm{T}} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$,若记

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \left(c_{ij}\right) \triangleq \boldsymbol{C}, \quad i,j = 0,1,2,\cdots,p, \tag{3.2.10}$$

  - 则

$$\hat{\beta}_j \sim N\left(\beta_j, c_{ij}\sigma^2\right), \quad j = 1,2,\cdots,p.$$

3.2.1 线性回归模型的原理

四、显著性检验

2. 回归系数的显著性检验：*t* 检验

- 构造 t 统计量

$$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t(n-p-1), \tag{3.2.11}$$

  - 上式中，$\hat{\sigma}^2 = \frac{(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-p-1}$.

- 给定显著性水平 α $(0<\alpha<1)$, 双侧 t 检验的临界值为 $t_\alpha/2(n-p-1)$.当检验统计量的观测值落在拒绝域, 即 $|t_j|>t_\alpha/2(n-p-1)$ 时, 拒绝原假设,认为 β_j 显著不为 0, X_j 对 Y 影响显著; 否则, 只能接受原假设.

- 还可以利用 p 值法进行判断, 其中 $p=P(|Z|>|t_j|)$, $Z\sim t(n-p-1)$.当 $p<\alpha$ 时, 拒绝原假设; 否则, 只能接受原假设.

3.2.1 线性回归模型的原理

四、显著性检验

3. 拟合优度检验

- 拟合优度检验用于检验回归方程对样本观测值的拟合程度.
- **定义 3.4** 样本决定系数为

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

- 样本决定系数 R^2 是一个回归直线与样本观测值拟合优度的相对指标,反映了响应变量的波动能被自变量解释的比例. 其取值在 0 与 1 之间. R^2 越接近 1, 拟合优度越好.
- 在应用过程中, 残差平方和往往随着解释变量个数的增加而减少. 如果在模型中增加一个解释变量, R^2 往往增大. 而由增加解释变量个数引起的 R^2的增大, 往往与拟合好坏无关, 因此样本决定系数需要调整. 在样本量一定的情况下, 增加解释变量必定使得自由度减少. 故可将残差平方和与总离差平方和分别除以各自的自由度, 以剔除变量个数对拟合优度的影响.

3.2.1 线性回归模型的原理

四、显著性检验

3. 拟合优度检验

- **定义 3.5** 记 $\overline{R}^2$ 为修正的决定系数, 则

$$\overline{R}^2 = 1 - \frac{\mathrm{SSE}/(n-p-1)}{\mathrm{SST}/(n-1)}. \tag{3.2.12}$$

  - $\overline{R}^2$ 越大, 对应的回归方程拟合程度越好.

3.2.1 线性回归模型的原理

五、预测

- 若基于观测数据 $\{(X_{i1}, X_{i2}, \cdots, X_{ip}; Y_i), i = 1, 2, \cdots, n\}$ 建立的回归方程

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

- 通过了回归方程的显著性检验和回归系数的显著性检验, 则对给定的新样本观测点$\boldsymbol{X_0} = (1, X_{01}, X_{02}, \cdots, X_{0p})^{\mathrm{T}}$, 可以用 $\hat{Y}_0 = \boldsymbol{X}_0^{\mathrm{T}} \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 X_{01} + \cdots + \hat{\beta}_p X_{0p}$ 作为 $Y_0 = \beta_0 + \beta_1 X_{01} + \cdots + \beta_p X_{0p} + \varepsilon_0$ 的点预测.

- 因 $\hat{\boldsymbol{\beta}}$ 与 ε_0 独立, $\hat{\boldsymbol{\beta}}$ 与 Y_0 独立, 从而 $\hat{Y}_0$ 与 Y_0 独立. 下面证明可详见唐年胜, 李会琼 [31],

$$\hat{Y}_0 - Y_0 \sim N\left(0, \sigma^2\left(1 + \boldsymbol{X}_0^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\boldsymbol{X}_0\right)\right),$$

- 从而

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + \boldsymbol{X}^{\mathrm{T}}{}_0\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\boldsymbol{X}_0}} \sim t(n - p - 1).$$

3.2.1 线性回归模型的原理

五、预测

3. 拟合优度检验

- 由此得, Y_0 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\hat{Y}_0\text{-}t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{1+\boldsymbol{X}^{\mathrm{T}}{}_0\left(\mathbf{X}^{\mathrm{T}}\,\mathbf{X}\right)^{-1}\boldsymbol{X}_0},\right.$$

$$\left.\hat{Y}+t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{1+\boldsymbol{X}_0^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\,\mathbf{X}\right)^{-1}\boldsymbol{X}_0}.\right)$$

3.2.2 多重共线性

- 一般情况下, 当回归方程的自变量之间存在很强的线性关系, 回归方程的检验高度显著时, 有些回归系数却不能通过显著性检验, 甚至出现有的回归系数所带符号与实际经济意义不符, 此时可认为变量间存在多重共线性.

- 我们已经知道多元线性回归模型 (3.2.2) 的未知参数 $\boldsymbol{\beta}$ 的最小二乘估计为

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$$

- 这就要求 $\mathbf{X}$ 列满秩, 即 $\mathrm{rank}(\mathbf{X}) = p + 1$. 也就是要求 X 的列向量之间线性无关. 然而, 往往 $\mathrm{rank}(\mathbf{X}) < p + 1$, 此时 $|\mathbf{X}^{\mathrm{T}}\mathbf{X}| = 0$, 从而 $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ 不存在, 也就无法得到参数的估计量

- 在实际问题的研究中, 经常见到近似多重共线性的情形. 此时, $\mathrm{rank}(\mathbf{X}) = p + 1$, 但 $|\mathbf{X}^{\mathrm{T}}\mathbf{X}| \approx 0$, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ 的对角线元素很大, 从而 $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ 的对角线元素很大, 使得参数的估计精度很低.

3.2.2 多重共线性

- 多重共线性可以用方差膨胀因子进行判断.

- **定义 3.6**　称式 (3.2.10) 中矩阵 $\boldsymbol{C}=(c_{ij})=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ 中对角线元素 c_{jj}为自变量 X_j 的方差膨胀因子 (variance inflation factor, VIF), 记作 VIF_j .

  - 易证
  $$c_{jj}=\frac{1}{1-R_j^2},$$

  - 其中, R_j^2 是以 X_j 为因变量对其余 $p-1$ 个自变量进行线性回归得到的样本决定系数, 它度量了自变量 X_j 与其余 $p-1$ 个自变量的线性相关程度. 这种相关程度越强, 说明自变量之间的多重共线性越严重. 此时, R_j^2 越接近 1, VIF_j越大. VIF_j 的大小反映了自变量之间存在多重共线性的强弱. 当 $\mathrm{VIF}_j \geqslant 10$时, 说明自变量 X_j 与其余自变量之间存在严重的多重共线性.

3.2.3 岭回归

- 针对自变量之间出现多重共线性时, 普通最小二乘法效果明显变差的问题, 霍尔 (Hoerl) 于 1962 年提出一种改进最小二乘估计的方法[33].

- 当变量 X_j 之间相互独立时, 使用最小二乘法估计的参数 $\hat{\beta}$ 是真实值的无偏估计, 即满足 $E(\hat{\beta})=\beta$, 并且在所有的无偏估计中具有最小的方差. 从求解公式上看, 若要保证该回归系数有解, 必须确保 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 矩阵是满秩的, 即$\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 可逆, 但在实际中, 若是 X_j 间存在高度多重共线性, 或者自变量个数大于等于观测个数, 不论哪种情况, 最终算出来的行列式都等于 0 或者是近似为 0, 此时 $\hat{\beta}$ 的值表现不稳定, 即此时 $\hat{\beta}$ 的方差很大 (虽然在所有无偏估计类中最小, 但其本身仍很大), 导致其均方误差 MSE(MSE($\hat{\beta}$)=Var($\hat{\beta}$) + [bias($\hat{\beta}$)] ^{2}) 很大.

- 当自变量之间存在多重共线性, 即 $|\mathbf{X}^{\mathrm{T}}\mathbf{X}| \approx 0$ 时, 如果给 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 加上一个正的常数矩阵 $\lambda\boldsymbol{I}$ ($\lambda > 0$), 那么 $\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\boldsymbol{I}$ 接近奇异的程度会比 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 接近奇异的程度小得多, 从而使得行列式不再为零, 可以求逆. 即得到岭估计:

$$\hat{\boldsymbol{\beta}}(\lambda)=\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\boldsymbol{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}.$$

3.2.3 岭回归

- 容易看出, 岭估计使用了单位矩阵乘以常数 λ, 这使得对角线元素构成了一条由岭系数组成的 “岭”, 这便是岭回归名称的由来. 从优化角度看, 岭估计量是通过最小化以下目标函数得到:

$$\hat{\boldsymbol{\beta}}(\lambda)=\min_{\boldsymbol{\beta}}\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|^{2}+\lambda\|\boldsymbol{\beta}\|.$$

- **定义 3.7** 称 $\hat{\boldsymbol{\beta}}(\lambda)$ 为参数 $\boldsymbol{\beta}$ 的岭估计, 其中, λ 称为岭参数. 由参数 $\boldsymbol{\beta}$的岭估计所建立的回归方程称为岭回归方程.

- 易证, $\hat{\boldsymbol{\beta}}(\lambda)$ 是回归参数 β 的有偏估计. 且 $\|\hat{\boldsymbol{\beta}}(\lambda)\|<\|\hat{\boldsymbol{\beta}}\|$, 这表明 $\hat{\boldsymbol{\beta}}(\lambda)$可看成由 $\hat{\boldsymbol{\beta}}$ 进行某种向原点的压缩. 在均方误差意义下, $\hat{\boldsymbol{\beta}}(\lambda)$ 优于最小二乘估计 $\hat{\boldsymbol{\beta}}$. 当 $\lambda=0$ 时, 岭估计 $\hat{\boldsymbol{\beta}}(0)$ 即为普通最小二乘估计 $\hat{\boldsymbol{\beta}}$. 证明可详见唐年胜, 李会琼 [31].

- 因为岭参数 λ 不是唯一确定的, 故岭回归估计 $\hat{\boldsymbol{\beta}}(\lambda)$ 是回归参数 $\boldsymbol{\beta}$ 的一个估计族. 当岭参数 λ 在 $(0,\infty)$ 内变化时, $\hat{\beta}_j(\lambda)$ 是 λ 的函数. 在平面坐标系中, 把函数 $\hat{\beta}_j(\lambda)$ 随 $\lambda>0$ 变化所描绘出来的曲线, 称为岭迹.

3.2.3 岭回归

- 选择岭参数 λ 的值一般满足以下原则：
  - (1) 各回归系数的岭估计基本稳定;
  - (2) 用最小二乘估计所得符号不合理的回归系数, 其岭估计的符号变得合理;
  - (3) 回归系数没有不合乎经济意义的绝对值;
  - (4) 残差平方和增加不太多.

- 在实际应用中, 可以利用 $\hat{\beta}(\lambda)$ 的每一分量 $\hat{\beta}_j(\lambda)$ 在同一图形上的岭迹的变化形状来确定适当的 λ. 但是, 岭迹法存在一定的主观性. 此外, 还可以用广义交叉验证 (generalized cross validation, GCV) 来确定岭参数. 通常选取使得 GCV 最小的 λ 值. 具体可参见文献 [34].

3.3　广义线性模型

3.3 广义线性模型

- 在本章的前面几节介绍了多元线性模型以及其假设, 一般的线性模型主要适用于响应变量是连续型随机变量, 并且其一般服从正态分布的情形. 但是我们会发现在实际应用中有些响应变量是离散的, 而且并不服从正态分布.针对这种情况, 内尔德 (Nelder) 和韦德伯恩 (Wedderburn) 将传统线性模型推广到广义线性模型. 广义线性模型将响应变量的分布推广到指数型分布族,从连续型变量拓展到离散型变量, 通过连接函数将服从指数型分布族的响应变量的期望与自变量的线性函数连接起来, 下面开始介绍广义线性模型. 在给出广义线性模型的定义前先介绍指数型分布族与连接函数.

3.3.1 指数型分布族

- 指数型分布族是以指数分布表示的一族分布, 很多常见的分布如正态分布、泊松分布、二项分布等都属于指数型分布族. 下面给出其定义.

- **定义 3.8** 若随机变量 Y 的概率质量 (离散型) 或概率密度 (连续型) 具有如下形式:

$$f(y;\theta,\phi)=\exp\left(\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right), \tag{3.3.1}$$

  - 则称 y 服从指数族分布. 其中 $a(\cdot)$, $b(\cdot)$, $c(\cdot,\cdot)$ 为已知连续函数, θ 和 ϕ 为未知参数. 通常, θ 与 $E(Y)$ 有关, 是我们所感兴趣的参数, 而 ϕ 与 $\mathrm{Var}(Y)$ 有关,称为多余参数 (nuisance parameter).

- 下面以正态分布、泊松分布和二项分布为例说明这两类分布的概率密度或概率质量具有式 (3.3.1) 的形式, 并求出对应的 $a(\cdot)$, $b(\cdot)$, $c(\cdot,\cdot)$ 以及未知参数.

3.3.1 指数型分布族

例 3.1 设 $Y \sim N(\mu, \sigma^2)$, 则 Y 的概率密度可表示为

$$f(y;\theta,\phi)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$=\exp\left(\frac{y\mu-\frac{1}{2}\mu^2}{\sigma^2}-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\ln(2\pi\sigma^2)\right)\right).$$

- 对照式 (3.3.1) 有

$$\theta=\mu,\quad \phi=\sigma^2,\quad a(\phi)=\phi,$$

$$b(\theta)=\frac{1}{2}\mu^2=\frac{1}{2}\theta^2, c(y,\phi)=\frac{1}{2}\left(\frac{y^2}{\phi}+\ln(2\pi\phi)\right).$$

3.3.1 指数型分布族

例 3.2　设 $Y \sim P(\mu)$ (参数为 μ 的泊松分布), 则 Y 的概率质量可表示为

$$P(Y=y)=\frac{\mu^y}{y!}\exp(-\mu)=\exp(y\ln\mu-\mu-\ln(y!)) \quad y=0,1,2,\cdots$$

- 对照式 (3.3.1) 有

$$\theta=\ln\mu,\quad \phi=1, a(\phi)=\phi, b(\theta)=\mu=\exp(\theta), c(y,\phi)=-\ln(y!).$$

3.3.1 指数型分布族

例 3.3 设 $X \sim B(m, \mu)$ (二项分布), 其中 $0 < \mu < 1$, m 为正整数, 且二者均为未知参数. 令 $Y = X/m$, 则 Y 服从指数族分布. 事实上, Y 的所有可能的取值为 $y = 0, 1/m, 2/m, \cdots, 1$, 且

$$\begin{aligned} P(Y = y) &= P(X = my) \\ &= \mathrm{C}_m^{my} \mu^{my} (1-\mu)^{m-my} \\ &= \exp\left(\frac{y \ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu)}{\frac{1}{m}} + \ln\left(\mathrm{C}_m^{my}\right) \right). \end{aligned}$$

- 对照式 (3.3.1) 有

$$\theta = \ln\left(\frac{\mu}{1-\mu}\right), \phi = \frac{1}{m}, \quad a(\phi) = \phi,$$

$$b(\theta) = -\ln(1-\mu) = \ln(1+\exp(\theta)), c(y, \phi) = \ln\left(\mathrm{C}_m^{my}\right).$$

3.3.2 连接函数

- 通过前面的学习, 我们知道“回归”一般用于预测样本的值, 这个值通常是连续的. 在分类问题的应用上效果往往不理想. 为了保留线性回归“简单、效果佳”的特点, 又想让它能够进行分类, 需要对预测值再做一次处理. 这个多出来的处理过程, 就是 GLM 所做的最主要的事. 而处理这个过程的函数,我们把它叫做连接函数.

- 连接函数 $g(\cdot)$ 是将自变量第 i 组观测的线性组合与第 i 个观测值 Y_i 的期望连接起来的函数, 即

$$g\left(E\left(Y_i\right)\right)=g\left(\mu_i\right)=\beta_0+\sum_{j=1}^{p} X_{ij}\beta_j \quad i=1,2,\cdots,n. \tag{3.3.2}$$

3.3.2 连接函数

- 注意, 本小节的期望本质是条件期望. 在线性回归模型中, 由于假定 $Y_i \sim N(\mu_i, \sigma^2)$, 而 $E(Y_i) \in \mathbf{R}$, $\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_{ij}$ 一般也在 **R** 中取值, 因此通常取 $g(E(Y_i)) = E(Y_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_{ij} (i = 1,2,\cdots,n)$. 但是取 $g(E(Y_i)) = E(Y_i)$并不总是合适的, 例如, 当 Y_i 服从泊松分布即 $Y_i \sim P(\mu_i)$ 时, 则 $E(Y_i) = \mu_i > 0$ 而 $\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_{ij}$ 有可能取负值. 因此直接建立模型

$$g(E(Y_i)) = E(Y_i) = \mu_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_{ij}$$

  - 是不合理的.

- 连接函数有很多种, 一般在广义线性模型的建模中我们会使用正则连接函数, 即

$$g(E(Y)) = g(\mu) = \theta.$$

3.3.2 连接函数

- 表 3.2 介绍几种常见的连接函数.

连接函数名称	连接函数英文	连接函数公式
恒等函数	identity	$g(\mu)=\mu$
Logit 函数	logit	$g(\mu)=\ln(\mu/(1-\mu))$
Probit 函数	probit	$g(\mu)=\pi^{-1}(\mu)$
对数函数	log	$g(\mu)=\ln(\mu)$
逆函数	inverse	$g(\mu)=1/\mu$
平方根函数	sqrt	$g(\mu)=\sqrt{\mu}$
逆高斯分布	$1/\text{mu}^2$	$g(\mu)=1/\mu^2$
重对数函数	loglog	$g(\mu)=\ln(-\ln(\mu))$
互补重对数函数	cloglog	$g(\mu)=\ln(-\ln(1-\mu))$
柯西函数	cauchit	$g(\mu)=\tan(\pi(\mu-0.5))$

◀ 表 3.2
常用的连接函数

3.3.3 广义线性模型

- **定义 3.9** 设 $(Y_i; X_{i1}, X_{i2}, \cdots, X_{ip})$ $(i = 1, 2, \cdots, \mathrm{n})$ 为因变量 Y 和自变量 $X_1, X_2, \cdots, X_p$ 的观测值, 若

  - (1) $Y_1, Y_2, \cdots, Y_{\mathrm{n}}$ 相互独立, 且对每个 i, Y_i 服从指数族分布, 即

$$Y_i \sim f(y_i; \theta_i, \phi_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right);$$

  - (2) $g(\mu_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_{ij}$, 其中 $\mu_i = E(Y_i)$ $(i = 1, 2, \cdots, n)$,则称 Y 与 $X_1, X_2, \cdots, X_{\mathrm{p}}$ 服从广义线性模型.

- 表 3.3 是三种常见的广义线性模型. 由于每一个模型都有相应的指数分布, 因此我们可以采用最大似然方法进行参数估计.

分布	函数	模型
正态分布	$E(Y) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$	普通线性模型
二项分布	$E(Y) = \dfrac{\exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})}$	逻辑斯谛模型
泊松分布	$E(Y) = \exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$	对数线性模型

3.3.3 广义线性模型

例 3.4 逻辑斯谛回归是一种广泛用于分类问题的机器学习算法. 为了对数据进行分类, 它使用了一个 Sigmoid 函数将线性模型的输出转换为 0 和1 之间的概率值. 最大似然估计是一种常用的方法来估计逻辑斯谛回归模型的参数. 其基本思想是找到一组参数, 使得在这组参数下, 模型对训练数据的拟合最好. 以下是计算逻辑斯谛回归参数的一般步骤:

1. 定义模型

- 首先需要定义逻辑斯谛回归模型, 一般可以表示为

$$P\left(Y_i=1\middle|\boldsymbol{X}_i\boldsymbol{\beta}\right)=h_{\boldsymbol{\beta}}\left(\boldsymbol{X}_i\right)=\frac{1}{1+e^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i}},$$

- 其中 Y_i 表示第 i 个样本的类别 (0 或 1), $\boldsymbol{\beta}$ 是参数, $h_{\boldsymbol{\beta}}(\boldsymbol{X}_i)$ 是模型预测的 $\boldsymbol{X}_i$ 属于类别 1 的概率.

3.3.3 广义线性模型

2. 定义似然函数

- 假设有 n 个训练样本, 则整个数据集的似然函数可以表示为

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n} h_{\boldsymbol{\beta}}(\boldsymbol{X}_i)^{Y_i}\left(1-h_{\boldsymbol{\beta}}(\boldsymbol{X}_i)\right)^{1-Y_i}.$$

- 该式子可以理解为, 对于每个样本, 我们计算模型预测其类别的概率, 并将其与其真实类别的概率相乘. 由于样本之间是独立同分布的, 因此我们将所有样本的似然函数相乘, 从而得到整个数据集的似然函数.

3. 计算对数似然函数

- 对数似然函数可以方便地计算和优化, 而不会影响最优解. 因此, 可以将似然函数取对数得到

$$l(\boldsymbol{\beta})=\log L(\boldsymbol{\beta})=\sum_{i=1}^{n}\left[Y_i\log\left(h_{\beta}(\boldsymbol{X}_i)\right)+(1-Y_i)\log\left(1-h_{\boldsymbol{\beta}}(\boldsymbol{X}_i)\right)\right].$$

3.3.3 广义线性模型

4. 梯度下降求解最优参数

- 我们的目标是最大化对数似然函数 $l(\boldsymbol{\beta})$. 使用梯度下降法, 可以找到最优的 $\boldsymbol{\beta}$ 值. 具体来说, 首先计算对数似然函数的梯度

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left(h_{\boldsymbol{\beta}}(\boldsymbol{X}_i) - Y_i\right) X_{ij}.$$

- 然后, 通过迭代更新 $\boldsymbol{\beta}$ 的值, 直到达到最大化 $l(\boldsymbol{\beta})$ 的目标.

3.4　多元响应变量协方差广义线性模型

3.4 多元响应变量协方差广义线性模型

- 传统的线性模型 (linear model, LM) 适用于处理单个响应变量的情形.广义线性模型扩展了线性模型, 可用于处理独立非正态数据和单个响应变量情形, 并且假设方差函数是已知的. Anderson [38] 和 Pourahmadi[39] 将GLM方法扩展到处理非独立非正态数据和单个响应变量的情形, 提出协方差广义线性模型 (cGLM). Wagner Hugo Bonat 和 Bent Jørgensen [40] 将 cGLM 扩展到处理非独立非正态数据和多个响应变量的情形, 提出多元响应变量协方差广义线性模型, 它是广义线性模型及协方差广义线性模型的扩展.
- 本节介绍多元响应变量协方差广义线性模型的原理、参数估计和模型选择等.

3.4.1 McGLM 模型的原理

- McGLM 旨在处理具有时间和空间相关结构的多元响应变量, 通过采用传统广义线性模型的方差函数方法来考虑非正态性, 采用连接函数和线性预测器建模均值结构, 采用协方差连接函数和矩阵线性预测器建模协方差结构,为正态和非正态多元数据分析提供通用的统计建模框架.
- McGLM 的每一个响应变量均可由连接函数、方差函数、协方差函数这三种函数及线性预测器和矩阵线性预测器来构成.

- **定义 3.10　McGLM**

  - 设样本量 n, 响应变量个数 R, 响应变量矩阵 $\boldsymbol{Y}_{n\times R}=(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_R)$, 响应变量$\boldsymbol{Y}$ 的期望值矩阵为$\boldsymbol{M}_{n\times R}=(\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_R)$. $n\times n$矩阵 $\boldsymbol{\Sigma}_r$ 表示每个响应变量$\boldsymbol{Y}_r(r=1,2,\cdots,\mathrm{R})$ 的协方差矩阵, 该矩阵描述了每个响应变量内的协方差结构. $R\times R$ 矩阵 $\boldsymbol{\Sigma}_b$ 表示响应变量 $\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_R$ 间的相关系数矩阵.$\boldsymbol{X}_r$ 为设计矩阵的 $n\times k_r$ 子阵. 回归参数向量 $\boldsymbol{\beta}_r$ 为 $k_r\times 1$ 的回归参数子向量. 多元响应变量协方差广义线性模型满足:

3.4.1 McGLM 模型的原理

$$
\begin{aligned}
&E(\boldsymbol{Y})=\boldsymbol{M}=\left\{g_1^{-1}(\boldsymbol{X}_1\boldsymbol{\beta}_1),\cdots,g_R^{-1}(\boldsymbol{X}_R\boldsymbol{\beta}_R)\right\},\\
&\operatorname{Var}(\boldsymbol{Y})=\boldsymbol{C}=\boldsymbol{\Sigma}_R\overset{G}{\otimes}\boldsymbol{\Sigma}_b.
\end{aligned}
\tag{3.4.1}
$$

- 其中$\boldsymbol{\Sigma}_R\overset{G}{\otimes}\boldsymbol{\Sigma}_b=\operatorname{Bdiag}(\widetilde{\boldsymbol{\Sigma}}_1,\cdots,\widetilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b\otimes\boldsymbol{I}_n)\operatorname{Bdiag}(\widetilde{\boldsymbol{\Sigma}}_1^{\mathrm{T}},\cdots,\widetilde{\boldsymbol{\Sigma}}_R^{\mathrm{T}})$是 $\boldsymbol{\Sigma}_R$ 与$\boldsymbol{\Sigma}_b$ 的广义克罗内克积 [41]. 它是 $n_R\times n_R$ 矩阵, 表示所有响应变量的联合协方差矩阵. $\widetilde{\boldsymbol{\Sigma}}_R(r=1,2,\cdots,R)$ 表示协方差矩阵 $\boldsymbol{\Sigma}_r$ 的楚列斯基分解的下三角形矩阵, 是 $n\times n$ 矩阵. 运算符 Bdiag(·) 表示分块对角矩阵. $\boldsymbol{I}_n$ 表示 $n\times n$单位矩阵. ⊗表示克罗内克积. 函数 $g_r(\cdot)$ 为均值结构中的连接函数.

3.4.1 McGLM 模型的原理

1. 响应变量的协方差矩阵 $\boldsymbol{\Sigma}_r$

- 不同取值类型的响应变量 $\boldsymbol{Y}_r$ 可取不同的协方差矩阵:
  - (1) 若 $\boldsymbol{Y}_r$ 取值为连续型、二值型、二项式、比例或指数数据, 其协方差矩阵 $\boldsymbol{\Sigma}_r$ 可定义为

$$\boldsymbol{\Sigma}_r = V\left(\boldsymbol{\mu}_r; \boldsymbol{p}_r\right)^{\frac{1}{2}}\left(\boldsymbol{\Omega}\left(\boldsymbol{\tau}_r\right)\right)V\left(\boldsymbol{\mu}_r; \boldsymbol{p}_r\right)^{\frac{1}{2}}. \tag{3.4.2}$$

  - (2) 若 $\boldsymbol{Y}_r$ 取值为计数数据, 其协方差矩阵 $\boldsymbol{\Sigma}_r$ 可采用如下形式:

$$\boldsymbol{\Sigma}_r = \operatorname{diag}\left(\boldsymbol{\mu}_r\right) + V\left(\boldsymbol{\mu}_r; \boldsymbol{p}_r\right)^{\frac{1}{2}}\left(\boldsymbol{\Omega}\left(\tau_r\right)\right)V\left(\boldsymbol{\mu}_r; \boldsymbol{p}_r\right)^{\frac{1}{2}}. \tag{3.4.3}$$

  - 式 (3.4.2) 与 (3.4.3) 中, $V(\boldsymbol{\mu}_r; \boldsymbol{p}_r) = \operatorname{diag}(\vartheta(\boldsymbol{\mu}_r; \boldsymbol{p}_r))$ 是对角矩阵, 其主对角线元素由方差函数$\vartheta(\cdot\,; \boldsymbol{p}_r)$ 按元素应用于向量 $\boldsymbol{\mu}_r$ 而给出. 方差函数$\vartheta(\cdot\,; \boldsymbol{p}_r)$ 中的参数 $\boldsymbol{p}_r$ 为幂参数 (power parameter).
  - $\boldsymbol{\Omega}(\boldsymbol{\tau}_r)$ 为散度矩阵 (dispersion matrix), 描述了响应变量的协方差矩阵$\boldsymbol{\Sigma}_r$ 中不依赖于均值结构的协方差部分. $\boldsymbol{\tau}_r$ 为散度参数 (dispersion parame-ter).

3.4.1 McGLM 模型的原理

2. 方差函数 $\vartheta(\cdot\,;p_r)$

- 方差函数 $\vartheta(\cdot\,;\boldsymbol{p}_r)$ 在 McGLM 中起着重要作用. 模型选择不同的方差函数, 则对应不同的响应变量边际分布. 下面介绍 McGLM 常用的方差函数.
  - (1) 对于连续响应变量, 方差函数采用幂函数 $\vartheta(\mu_r;\,p_r)=\mu_r p_r$ 来描述Tweedie 分布族. Tweedie 分布族的特例有高斯分布 ($p_r=0$)、伽马分布 ($p_r=2$) 和逆高斯分布 ($p_r=3$).
  - (2) 对于取值为二值型、二项式或比例数据的响应变量, 方差函数常采用扩展二项式函数 $\vartheta(\mu_r;\,\boldsymbol{p}_r)=\mu_r^{pr1}(1-\mu_r)^{pr2}$.
  - (3) 对于计数数据的离散型响应变量, 方差函数采用函数 $\vartheta(\mu_r;\,\boldsymbol{p}_r)=\mu_r+\tau_r\mu_r^{pr}$ (其中 τ_r 是散度参数) 来描述 Poisson-Tweedie 分布族. 由于散度参数仅出现在第二项中, 因此协方差矩阵 $\boldsymbol{\Sigma}_r$ 采用式 (3.4.3) 的形式. Poisson-Tweedie分布族的特例有埃尔米特 (Hermite) 分布 ($p_r=0$)、奈曼 (Neyman) 分布($\mathrm{p}_r=1$)、负二项分布 ($p_r=2$) 和泊松逆高斯 (Poisson-inverse Gaussian) 分布 ($p_r=3$).

3.4.1 McGLM 模型的原理

3. 幂参数 p_r

- 在 McGLM 二阶矩假设下, 模型估计幂参数 p_r 的信息来自均值和方差的关系, 因此估计幂参数 p_r 时需要均值向量有足够的变化, 即要求线性预测器中存在显著的协变量. 对于 McGLM 中所有的方差函数, 幂参数 p_r 是区分重要分布的指标. 模型中的算法允许估计幂参数 p_r, 该参数可用于自动选择分布.

4. 协方差连接函数 $h(\cdot)$ 与矩阵线性预测器

- Anderson[38] 和 Pourahmadi[39], Bonat 和 Jørgensen[40] 提出使用矩阵线性预测器结合协方差连接函数来对散度矩阵 $\boldsymbol{\Omega}(\boldsymbol{\tau}_r)$ 进行建模, 即根据已知矩阵的线性组合建立一个模型.

$$h\left(\boldsymbol{\Omega}\left(\boldsymbol{\tau}_r\right)\right)=\tau_{r0}\boldsymbol{Z}_{r0}+\cdots+\tau_{rD}\boldsymbol{Z}_{rD}. \tag{3.4.4}$$

- 这种结构与均值结构的线性预测器相似, 称为矩阵线性预测器 (matrixlinear predictor). 将矩阵线性预测器 (3.4.4) 代入式 (3.4.1) 中, 即得到多元响应变量协方差广义线性模型.

3.4.1 McGLM 模型的原理

- 式 (3.4.4) 中, $h(\cdot)$ 为协方差连接函数 (covariance link function); $\boldsymbol{Z}_{rd}(d=0,\cdots,D)$ 是反映响应变量 $\boldsymbol{Y}_r$ 内协方差结构的已知矩阵, $\boldsymbol{\tau}_r=(\tau_{r0},\cdots,\tau_{rD})^{\mathrm{T}}$为 $D+1$ 维散度参数向量.
- McGLM 中协方差连接函数常采用恒等函数 (identity function)、逆函数(inverse function) 和指数矩阵函数 (exponential-matrix function).
- 实际上, 很难定义散度参数向量 $\boldsymbol{\tau}_r$ 的参数空间. Bonat 和 Jørgensen 利用倒数似然算法, 使用一个调整常数来控制算法的步长, 并避免参数向量 $\boldsymbol{\tau}_r$的不合理值.
- 矩阵预测器中的协方差结构常采用协方差矩阵的线性模型, 用于处理非高斯数据、纵向自相关数据 (如复合对称、移动平均和一阶自回归等)、空间数据及区域数据等. 关于如何指定矩阵 $\boldsymbol{Z}_{rd}$, 详见文献 [40].

5. 连接函数 $g(\cdot)$

- McGLM 中, 连接函数 $g(\cdot)$ 将响应变量的期望与协变量联系起来, 反映了模型的均值结构. 常采用的连接函数也如表 3.2.

3.4.2 参数估计

- 设参数向量 $\boldsymbol{\theta}=(\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})^{\mathrm{T}}$, 其中回归参数向量 $\boldsymbol{\beta}_{K\times 1}=(\boldsymbol{\beta}_1{}^{\mathrm{T}},\cdots,\boldsymbol{\beta}_R{}^{\mathrm{T}})^{\mathrm{T}}$,散度参数向量 $\boldsymbol{\lambda}_{Q\times 1}=(\rho_1,\cdots,\rho_{R(R-1)/2},p_1,\cdots,p_R,\boldsymbol{\tau}_1{}^{\mathrm{T}},\cdots,\boldsymbol{\tau}_R{}^{\mathrm{T}})^{\mathrm{T}}$.
- McGLM 采用基于二阶矩假设的估计函数法进行拟合. 模型基于拟得分(quasi-score) 函数和皮尔逊估计(Pearson estimating) 函数的有效牛顿评分算法 (Newton scoring algorithm), 针对回归参数和散度参数, 利用修正的追赶算法 (modified chaseralgorithm) 以及改进的倒数似然算法 (reciprocal likelihoodalgorithm), 通过调节常数 α 来调节步长, 从而进行参数估计及统计推断, 进而拟合 McGLM.
- 具体来讲, 我们让 $y=\left(\boldsymbol{Y}_1^T,\cdots,\boldsymbol{Y}_R^T\right)^T$ 和 $\mathcal{M}=\left(\boldsymbol{\mu}_1^T,\cdots,\ \boldsymbol{\mu}_R^T\right)^T$ 表示$nR\times 1$ 列y向量, 其中的元素分别由响应变量矩阵 $\boldsymbol{Y}_{n\times R}$ 和期望值矩阵 $\boldsymbol{M}_{n\times R}$按列排序得到.
- 为了进行参数估计, 我们采取拟得分函数[42]

$$\psi_{\beta}(\boldsymbol{\beta},\boldsymbol{\lambda})=\boldsymbol{D}^{\mathrm{T}}\boldsymbol{C}^{-1}(\mathcal{Y}-\mathcal{M}), \tag{3.4.5}$$

3.4.2 参数估计

- 其中 $\boldsymbol{D}=\nabla_{\beta}\mathcal{M}$ 是一个 $nR\times K$ 矩阵, 并且 ∇_{β} 表示对相应参数求梯度. 此外, 可以得到 ψ_{β} 的 $K\times K$ 敏感性矩阵和特异性矩阵

$$\boldsymbol{S}_{\boldsymbol{\beta}}=E\left(\nabla_{\boldsymbol{\beta}}\boldsymbol{\psi}_{\boldsymbol{\beta}}\right)=-\boldsymbol{D}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{D}\text{和}V_{\boldsymbol{\beta}}=\operatorname{Var}\left(\boldsymbol{\psi}_{\boldsymbol{\beta}}\right)=\boldsymbol{D}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{D}. \tag{3.4.6}$$

- 散度参数所采用的皮尔逊估计函数的各个部分定义如下:

$$\psi_{\lambda_i}(\boldsymbol{\beta},\boldsymbol{\lambda})=\operatorname{tr}\left(\boldsymbol{W}_{\lambda_i}\left(\boldsymbol{r}^{\mathrm{T}}\boldsymbol{r}-\boldsymbol{C}\right)\right),\qquad i=1,2,\cdots Q, \tag{3.4.7}$$

- 其中 $\boldsymbol{W}_{\lambda i}=-\partial\boldsymbol{C}^{-1}/\partial\lambda_i$ 并且 $\boldsymbol{r}=\mathcal{Y}-\mathcal{M}$.

- ψ_{λ} 的 $Q\times Q$ 敏感性矩阵的第 i 行第 j 列元素定义如下:

$$\boldsymbol{S}_{\lambda_{ij}}=E\left(\frac{\partial}{\partial\lambda_i}\psi_{\lambda_i}\right)=-\operatorname{tr}\left(\boldsymbol{W}_{\lambda_i}\boldsymbol{C}\boldsymbol{W}_{\lambda_j}\boldsymbol{C}\right). \tag{3.4.8}$$

3.4.2 参数估计

- ψ_λ 的 $Q \times Q$ 特异性矩阵的第 i行第 j 列元素定义如下：

$$V_{\lambda_i} = \mathrm{Cov}\left(\psi_{\lambda_i}, \psi_{\lambda_j}\right) = 2\mathrm{tr}\left(\boldsymbol{W}_{\lambda_i}\boldsymbol{C}\boldsymbol{W}_{\lambda_i}\boldsymbol{C}\right) + \sum_{l=1}^{nR} k_l^{(4)}\left(\boldsymbol{W}_{\lambda_i}\right)_u\left(\boldsymbol{W}_{\lambda_i}\right)_u, \tag{3.4.9}$$

- $k_l^{(4)}$ 是 y_l 的第四阶累积量 (the fourth cumulant). 2004 年, Jørgensen 和Knudsen[43] 提出修正的追赶算法用来解 $\boldsymbol{\psi}_\beta = \mathbf{0}$ 和 $\boldsymbol{\psi}_\lambda = \mathbf{0}$. 具体更新公式如下：

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \boldsymbol{S}_\beta^{-1}\psi_\beta\left(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}\right),$$
$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)} - \alpha\boldsymbol{S}_\lambda^{-1}\psi_\lambda\left(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}\right).$$

- 2016 年, Bonat 和 Jørgensen[40] 提出倒数似然算法 (reciprocal likelihoodalgorithm), 通过调节常数 α 来调节步长, 从而对参数 λ 估计. 具体更新公式如下：

$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)} - \left[\alpha\boldsymbol{\psi}_\lambda\left(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}\right)^{\mathrm{T}}\boldsymbol{\psi}_\lambda\left(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}\right)\boldsymbol{V}_\lambda^{-1}\boldsymbol{S}_\lambda + \boldsymbol{S}_\lambda\right]^{-1}\psi_\lambda\left(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}\right).$$

3.4.2 参数估计

- McGLM 中回归参数估计量对协方差结构的形式依赖性相对较小, 而回归参数估计量的标准误差则直接依赖于协方差结构的选择. McGLM 还可以进行回归参数的稳健和偏差校正标准误差、残差分析等. 详见文献 [40].

3.5　回归分析实践

实践代码