第八讲 线性回归模型

回归函数

y: 响应变量

$\mathbf{x}$: 自变量, 1维或多维

以$\mathbf{x}$的某个函数$f(\mathbf{x})$逼近y，误差为$\mathrm{E}(y-f(\mathbf{x}))^2$
最优逼近是$f(\mathbf{x})=\mathrm{E}(y|\mathbf{x})$,这是因为
$$\mathrm{E}(y-f(\mathbf{x}))^2 \ge \mathrm{E}(y-\mathrm{E}(y|\mathbf{x}))^2$$

$\mathrm{E}(y|\mathbf{x})$称为回归函数. 令$\varepsilon=y-\mathrm{E}(y|\mathbf{x})$，则容易验证$\varepsilon$与$\mathbf{x}$不相关。所以响应变量$y$可以表示为：

$$y=\mathrm{E}(y|\mathbf{x})+\varepsilon。$$

对回归函数假设结构，即得到各种回归模型。

一般线性回归模型

模型表示1：
y: 响应变量, $\mathbf{x}$: 自变量(向量)。线性回归模型假设:
(i) 线性回归函数: $\mathrm{E}(y|\mathbf{x})=a+\mathrm{b}'\mathbf{x}$,
(ii) 方差常数/齐性：$\mathrm{var}(y|\mathbf{x})=\sigma^2$.
其中a,b,σ^2是未知参数。

- 自变量$\mathbf{x}$一维时,称为简单线性回归模型 (simple linear regression model)
- 自变量$\mathbf{x}$多维时,称为多重线性回归模型 (multiple linear regression model)

令$\varepsilon=y-\mathrm{E}(y|\mathbf{x})=y-(a+\mathrm{b}'\mathbf{x})$,
若要求ε与$\mathbf{x}$独立， 则有"等价的"/稍强的模型表示:

模型表示2:
$$y=a+b'\mathbf{x}+\varepsilon,$$
其中
(1) $E(\varepsilon)=0$
(2) $\mathrm{var}(\varepsilon)=\sigma^2$ (方差齐性, Homoscedasticity)
(3) ε 与 $\mathbf{x}$ 独立 (外生性, Exogeneity)

条件(1),(2)加上样本独立性，称为Gauss-Markov假设

注：
(1) 误差0均值(均值常数,不依赖于x)：
保证可识别性(Identifiability), 截距参数a可识别

(2)方差齐性(方差常数,不依赖于x)：
保证参数估计的最优性($Gauss-Markov$定理)

(3)外生性：保证参数估计的无偏性 /因果

为什么线性？

1.如果x, y服从联合正态分布：$\begin{pmatrix} y \\ \mathrm{x} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_y \\ \mu_{\mathrm{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_{yy} & \Sigma_{y\mathrm{x}} \\ \Sigma_{\mathrm{x}y} & \Sigma_{\mathrm{xx}} \end{pmatrix}\right)$,则

$$\mathrm{y} \mid \mathrm{x} \sim \mathrm{N}(\mu_y + \Sigma_{y\mathrm{x}}\Sigma_{\mathrm{xx}}^{-1}(\mathrm{x} - \mu_{\mathrm{x}}), \Sigma_{yy} - \Sigma_{y\mathrm{x}}\Sigma_{\mathrm{xx}}^{-1}\Sigma_{\mathrm{x}y}),$$

(a) $\mathrm{E}(y \mid \mathrm{x}) = \mu_y + \Sigma_{y\mathrm{x}}\Sigma_{\mathrm{xx}}^{-1}(\mathrm{x} - \mu_{\mathrm{x}}) \hat{=} a + b'\mathrm{x}$,
(b) $\mathrm{var}(y \mid \mathrm{x}) = \Sigma_{yy} - \Sigma_{y\mathrm{x}}\Sigma_{\mathrm{xx}}^{-1}\Sigma_{\mathrm{x}y} \hat{=} \sigma^2$.

或等价地(令 $\varepsilon = y - E(y \mid x) = y - (a + b'\mathrm{x})$)
$$y = a + b'\mathrm{x} + \varepsilon,$$
其中$\varepsilon \sim N(0, \sigma^2)$,且与x独立.

2. 如果$x(=0,1)$是因子变量,代表随机化分组,
而每一组内y服从正态分布：
$$y|_{\mathrm{x}=1} \sim N(\mu_1, \sigma^2), \quad y|_{\mathrm{x}=0} \sim N(\mu_0, \sigma^2)$$

则可以表示为：
$$y|_x \sim N(\mu_0 + (\mu_1 - \mu_0)x, \sigma^2) \hat{=} N(a + bx, \sigma^2), x = 0,1$$
或$y = a + bx + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$,且与$x$独立.

模型的数据形式

数据：$(y_i, \mathbf{x}_i)$，$i = 1,2...,n$独立，其中
向量$\mathbf{x}_i$为自变量的第i个观察值(第一分量为1,对应于截距项);
y_i为响应变量,$i = 1,2...,n$.

假设$(y_i, \mathbf{x}_i)$，$i = 1,2...,n$,独立，满足线性回归模型：
$$y = \mathbf{x}'\beta + \varepsilon, \quad \varepsilon \sim (0, \sigma^2), \quad \varepsilon\text{与}\mathbf{x}\text{独立}$$
其中β为$p \times 1$回归系数列向量(第一个分量是截距项)。即
$$y_i = \mathbf{x}_i'\beta + \varepsilon_i, i = 1,2...,n$$
其中$\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ iid $\sim (0, \sigma^2)$，且ε_i与$\mathbf{x}_i$独立。

记

$$\mathbf{Y}=\begin{pmatrix}y_1\\y_2\\\cdots\\y_n\end{pmatrix},\quad X=\begin{pmatrix}\mathbf{x}_1'\\\mathbf{x}_2'\\\cdots\\\mathbf{x}_n'\end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix}\beta_1\\\beta_2\\\cdots\\\beta_p\end{pmatrix},\quad \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\cdots\\\varepsilon_n\end{pmatrix},$$

其中 X 称为设计阵(第i行为$\mathbf{x}_i'$)。

模型等价地以矩阵-向量形式表示为：

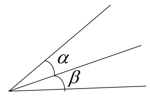
$$\mathbf{Y}_{n\times1}=X_{n\times p}\boldsymbol{\beta}_{p\times1}+\boldsymbol{\varepsilon}_{n\times1},\quad \boldsymbol{\varepsilon}\sim(\mathbf{0},\sigma^2 I_n),\ \ \boldsymbol{\varepsilon}\text{与}X\text{独立。}$$

或以前两阶矩表示为：

(a) $\mathrm{E}(\mathbf{Y}|X)=X\boldsymbol{\beta}$,

(b) $\mathrm{var}(\mathbf{Y}|X)=\sigma^2 I_n$

例1. 一个航海员使用六分仪测量下图中的角度 α，β，他共测量了3次，第一、二次分别测量 α，β，得到测量值 y_1,y_2 第三,四次测量角度 α+β 得到测量值y_3, y_4, 假设四次测量独立, (可加)误差均值为0， 方差为σ², 写出线性模型



$y_1=\alpha+\varepsilon_1$

$y_2=\beta+\varepsilon_2$

$y_3=\alpha+\beta+\varepsilon_3$

$y_4=\alpha+\beta+\varepsilon_4$

ε_i独立$\sim(0,\sigma^2)$

$$\mathbf{Y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon},E\boldsymbol{\varepsilon}=0,\mathrm{var}(\boldsymbol{\varepsilon})=\sigma^2 I_4$$

$$\mathbf{Y}=\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix},\quad X=\begin{pmatrix}1&0\\0&1\\1&1\\1&1\end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix}\alpha\\\beta\end{pmatrix},\quad \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\end{pmatrix}.$$

例2(单因素方差分析模型). n个个体被随机地分配于K组分别接受某种处理(*treatment*),并测量某指标。假设第k组有n_k个个体，指标测量为$y_{k1},..,y_{kn_k}$ iid $\sim \mathrm{N}(\mu_k,\sigma^2), k=1,....K$

记 $\varepsilon_{kj}=y_{kj}-\mu_k$, $\mathrm{j}=1,...,\mathrm{n_k}$ iid $\sim\mathrm{N}(0\ ,\sigma^2), k=1,....K$,即

$$y_{kj}=\mu_k+\varepsilon_{kj},$$

模型可表示为：$\mathbf{Y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon},E\boldsymbol{\varepsilon}=0,\mathrm{var}(\boldsymbol{\varepsilon})=\sigma^2 I_n$, 其中

$$Y=\begin{pmatrix}y_{11}\\\vdots\\y_{1n_1}\\\vdots\\y_{K1}\\\vdots\\y_{Kn_K}\end{pmatrix},\mathrm{X}=\begin{pmatrix}1&0&\cdots&0\\\vdots&\vdots&\cdots&\\1&0&\cdots&0\\\vdots&\vdots&\cdots&\vdots\\0&0&\cdots&1\\\vdots&\vdots&\cdots&\vdots\\0&0&\cdots&1\end{pmatrix},\beta=\begin{pmatrix}\mu_1\\\mu_2\\\vdots\\\mu_K\end{pmatrix},\varepsilon=\begin{pmatrix}\varepsilon_{11}\\\vdots\\\varepsilon_{1n_1}\\\vdots\\\varepsilon_{K1}\\\vdots\\\varepsilon_{Kn_K}\end{pmatrix}$$

为了检验K个均值相同,通常重新参数化，比如令

$$\beta_1=\mu_1,\ \beta_2=\mu_2-\mu_1,\ \ldots,\ \beta_{\mathrm{K}}=\mu_{\mathrm{K}}-\mu_1,$$

对于上述新参数$\tilde{\boldsymbol{\beta}}=(\beta_1,...,\beta_K)'$,写出模型

$\mathbf{Y}=X\tilde{\boldsymbol{\beta}}+\boldsymbol{\varepsilon},E\boldsymbol{\varepsilon}=0,\mathrm{var}(\boldsymbol{\varepsilon})=\sigma^2 I_{\mathrm{n}}$

中的设计阵X