线性回归 : $X \in \mathbb{R}^{n \times p}$ n个样本，p个特征

low dimension : $\begin{cases} \text{univariate} : p=1 & \text{单变量} \\ \text{multivariate} : 1<p<n & \text{多变量} \end{cases}$ } 低维

high dimension : $p \geq n$ 高维

1. Form of Multivariate Linear Models

① 假设有 n 组样本 : $\{(y_i, x_i) \mid i=1,2,\cdots,n\}$ 其中 标量 $y_i \in \mathbb{R}$ 为 response

向量 $x_i = [x_{i1}, x_{i2}, \cdots, x_{ip}]^T \in \mathbb{R}^p$ 包含 p 个 variate , y_i 与 x_i 满足如下的 Linear Model

向量形式

$$y_i = x_i^T \beta + \varepsilon_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad i=1,2,\cdots,n \tag{1}$$

其中 $\beta = [\beta_1, \beta_2, \cdots, \beta_p]^T \in \mathbb{R}^p$ 是 regression parameters (包含截距 intercept 因为 $x_{i1} = 1$) , ε_i 为第 i 组观测样本的 random error。

称 (1) 为 vector form of the linear model

② 记 design matrix $X = [x_1^T, x_2^T, \cdots, x_n^T]^T \in \mathbb{R}^{n \times p}$

response vector $y = [y_1, y_2, \cdots, y_n]^T \in \mathbb{R}^n$

random error vector $\varepsilon = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n]^T \in \mathbb{R}^n$

regression parameters $\beta = [\beta_1, \beta_2, \cdots, \beta_p]^T \in \mathbb{R}^p$

则有 Matrix form :

矩阵形式

$$y = X \cdot \beta + \varepsilon \tag{2}$$

2. Basic assumptions of multivariate linear models 基本假设

① Assumption 1: fixed designed: designed matrix X is fixed / nonrandom

固定设计：设计矩阵 X 是固定非随机的

② Assumption 2: full rankness of the design matrix: $rank(X) = p < n$

设计矩阵满秩：$rank(X) = p < n$ 即 X 的列向量线性独立

Tips：这一假设是为了保证可识别性（identifiability），例如：若 $x_{i2} = k x_{i3}$

则只能 estimate / learn 到 $(\beta_2 + \beta_3)$，具体的 β_2, β_3 无法得到

Tips：但只对模型的选择、估计、推断是必要的，对预测不必要

③ Assumption 3: Gauss - Markov condition of the random errors

随机误差满足 Gauss - Markov 条件：$E(\varepsilon) = 0$ $cov(\varepsilon) = \sigma^2 I$

$E(\varepsilon) = 0$ ⇒ y 的期望是 X 的线性函数

$cov(\varepsilon) = \sigma^2 I$ ⇒ n 个样本是不相关的（未必独立），且等方差

$$E(y) = X \cdot \beta \qquad cov(y) = \sigma^2 \cdot I \tag{3}$$

3. Interpretation of Regression Parameters 参数的含义

In formula (1),

$$\frac{\partial E(y_i)}{\partial x_{ij}} = \beta_j$$

表示在其他自变量保持不变时，自变量 x_{ij} 每增加 1 单位，因变量 y_i 的平均增幅

4. Estimation of regression parameters 估计

4.1. Ordinary Least Squares Estimation (OLSE) 最小二乘

① 拟合值： fitted values

$$\{\hat{y}_i = x_i^T\hat{\beta} \mid i = 1, 2, \cdots, n\} \subset \mathbb{R}$$

② 残差： residuals

$$\{e_i = y_i - \hat{y}_i \mid i = 1, 2, \cdots, n\} \subset \mathbb{R}$$

③ 估计参数 $\hat{\beta}$

$$\hat{\beta} = \underset{b\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n}(y_i - x_i^T b)^2 = \underset{b\in\mathbb{R}^p}{\arg\min} \|y - Xb\|^2 \tag{4}$$

④ 求导即可： ☆ $\sum x_i e_i = 0$ 而 $x_{i1} = 1 \Rightarrow \sum e_i = 0$

$$\mathbb{R}^p \ni 0 = \frac{\partial Q}{\partial b}\Big|_{b=\hat{\beta}} = \sum_{i=1}^{n} x_i e_i = \sum_{i=1}^{n} x_i(y_i - x_i^T\hat{\beta}) = X^T(y - X\hat{\beta}) \tag{5}$$

$e_i \in \mathbb{R}$

$x_i \in \mathbb{R}^p$

$$\hat{\beta} = (X^TX)^{-1}\cdot X^T\cdot y \tag{6}$$

$X \in \mathbb{R}^{n\times p} \quad y \in \mathbb{R}^n \quad \hat{\beta} \in \mathbb{R}^p$

Proof: $X \in \mathbb{R}^{n\times p}$ 当 $rank(X) = p$ 有 X^TX 可逆

Lemma 对 $\forall$ 矩阵 $A \in \mathbb{R}^{m\times n}$ 有 $A^TAx = 0 \Leftrightarrow Ax = 0$

① $\Leftarrow$ 显然

② $\Rightarrow$ $A^TAx = 0$ 有 $x^TA^TAx = 0$

即 $(Ax)^T(Ax) = 0$ 即 $\|Ax\|_2 = 0$

则 $Ax = 0$ □

回到原命题：

记 $N(X)=\{v \mid Xv=0\}$　　$N(X^TX)=\{v \mid X^TXv=0\}$

由 Lemma，二者相同，即有：$\dim(N(X))=\dim(N(X^TX))$

又∵ $p=\text{rank}(X)+\dim(N(X))$　　$\text{rank}(X)=p$

∴ $\dim(N(X^TX))=\dim(N(X))=0$

故 $\text{rank}(X^TX)=p$　　即 $X^TX\in\mathbb{R}^{p\times p}$

则 X^TX 满秩（可逆）□

Supplementary：$\frac{\partial\|y-Xb\|^2}{\partial b}=\frac{\partial}{\partial b}\{(y-Xb)^T(y-Xb)\}$

$=-X^T\cdot(y-Xb)-(y-Xb)^TX$　　$=-X^Ty+X^TXb-(y^TX)^T+X^TXb$

$=-2X^T(y-Xb)$

4.2 Maximum Likelihood Estimation (MLE) 极大似然：

④ Assumption 4: random error vector follows a multivariate normal

$$\varepsilon\sim N(0,\sigma^2I_n)$$

or

$$y\sim N(X\beta,\sigma^2I_n) \tag{7}$$

① $\{y_i \mid i=1,2,\cdots,n\}$ 相互独立

② 对数似然，函数

$$\log L(\beta,\sigma^2)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}\|y-X\beta\|^2 \tag{8}$$

③ 关于 β 最大化，等价于 (4)，即

在 $\varepsilon \sim N(0, \sigma^2 I)$ 的假设下，β 的 MLE 与 OLSE 相同

$$\hat{\beta} = (X^T X)^{-1} \cdot X^T \cdot y \tag{6}$$

④ 关于 σ^2 最大化

$$\hat{\sigma}^2_{MLE} = n^{-1} \| y - X\hat{\beta} \|^2 = n^{-1} \| e \|^2$$

5. Properties of estimats and residuals 性质

5.1 $\{e_i = y_i - x_i^T \hat{\beta}\}$ 残差的性质

fitted value of y:

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

记 $H = X(X^T X)^{-1} X^T$，则有 $\hat{y} = H \cdot y$

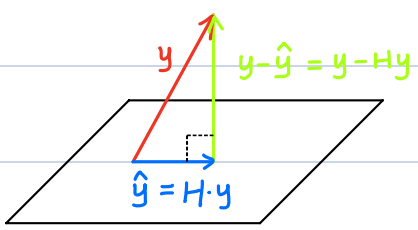
实际上，H 是一个投影矩阵，将 y 投影到设计矩阵 X 的列空间

① H 的性质（投影矩阵）

(1) $H = X(X^T X)^{-1} X^T \in \mathbb{R}^{n \times n}$

(2) $\sum_{i=1}^{n} H_{[i,i]} = p$ ← $tr(AB) = tr(BA)$ 易证

(3) $H^2 = H \quad H^T = H$



(4) $(I-H)^2 = I-H$ ⇒ $I-H$ 特征值为 0 或 1

(5) $(I-H)X = 0$

$\text{rank}(I-H) = \sum \lambda_i = \text{tr}(I-H) = n-p$

$\text{tr}(I-H) = n-p$

② 残差 e_i 的性质

(1) $e_i = y_i - \hat{y}_i$

(2) $E(e_i) = 0$

$\text{Cov}(e) = \text{Cov}(y - \hat{y}) = \text{Cov}((I-H)y)$

$= (I-H)\sigma^2 I \cdot (I-H)^T$

$= \sigma^2(I-H)$

(3) $E(e_i^2) = \text{Var}(e_i) = (1 - H_{[i,i]}) \cdot \sigma^2$

(4) $E(\sum_{i=1}^n e_i^2) = (n - \sum_{i=1}^n H_{[i,i]})\sigma^2 = (n-p)\sigma^2$

表明: $\hat{\sigma}^2 = (n-p)^{-1} \cdot \sum_{i=1}^n e_i^2$ 是 σ^2 的无偏估计

注: 模型建立时, 截距项作为 $x_{i1} = 1$ 的参数, 故与其他样本略有差别

5.2 回归参数的性质

(a). 线性 Linearity: $\hat{\beta}$ 是 y 的线性函数

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

$E(y) = X\beta$

(b) 无偏性 Unbiasedness: $\hat{\beta}$ 是 β 的无偏估计

$$E(\hat{\beta}) = \beta$$

⇒ $\hat{y} = X\hat{\beta}$ 是 $E(y)$ 的无偏估计

$$E(\hat{y}) = E(y)$$

(c) 协方差矩阵 Covariance matrices

$$\mathrm{cov}(\hat{\beta}) = (X^TX)^{-1}X^T \cdot \mathrm{cov}(y) \cdot X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

$$\mathrm{cov}(e) = \sigma^2(I-H)$$ ☆ 见上

(d) Gauss-Markov 定理:

在 Assumption 3 条件下，向量 β 的任意线性组合 $v^T\beta$ 的 最小方差 线性无偏估计 (Best Linear Unbiased Estimator, BLUE) 是 $v^T\hat{\beta}$，其中 $v \in \mathbb{R}^p$

$\hat{\beta}$ 是 β 的 OLSE (见作业)

在 Assumption 4 正态假设也成立时，$v^T\hat{\beta}$ 作为 MLE 是 $v^T\beta$ 的 最小方差 无偏估计 (Best Unbiased Estimator, BUE)，即 $v^T\beta$ 所有 (线性或非线性) 无偏估计中方差最小的

(e). σ^2 的无偏估计 Unbiased estimate for σ^2

$\frac{e^Te}{n-p}$ 是 σ^2 的无偏估计 见前

$$E\left(\frac{e^Te}{n-p}\right) = \sigma^2$$

Lemma: $\mathrm{tr}(H) = p$

(f). 与残差的不相关性:

$$e = y - \hat{y} = y - X(X^TX)^{-1}X^Ty = (I-H)y$$ ☆ ↓ $e = (I-H)y$

$$\mathrm{cov}(\hat{\beta}, e) = 0$$

Note: 当正态假设 Assumption 4 成立时，$\hat{\beta} \perp e$

(g) 分布：当正态假设 Assumption 4 成立时

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

$$\frac{e^Te}{\sigma^2} \sim \chi^2(n-p)$$

$y \sim N(X\beta, \sigma^2 I)$

$e^Te = y^T(I-H)y$

$(X\beta)^T(I-H)(X\beta) = 0$

$(I-H)^2 = I-H$

$X \sim N(\mu, I)$ $\quad rank(A)=r$ $\quad \mu^TA\mu = 0$ $\quad A^2 = A$ $\quad A$ 是实对称

$X^TAX \sim \chi^2(r)$

$rank(I-H) = n-p$

在 Assumption 3 条件下，向量 β 的任意线性组合 $v^T\beta$ 的最小方差线性无偏估计（Best Linear Unbiased Estimator, BLUE）是 $v^T\hat{\beta}$，其中 $v \in \mathbb{R}^p$

$\hat{\beta}$ 是 β 的 OLSE $\quad v^T\hat{\beta}$ 的方差最小

4. 证明课件第五章第 5 节的性质(d)。

设$\tilde{\beta}$为β的无偏估计量，且存在$D \in \mathbb{R}^{p\times n}$以及$B = (X^TX)^{-1}X^T + D$，使得$\tilde{\beta} = By$。由于$\tilde{\beta}$的无偏性，对任意$\beta \in \mathbb{R}^p$，都有

$$\beta = \mathbb{E}(\tilde{\beta}) = \mathbb{E}(By) = (X^TX)^{-1}X^TX\beta + DX\beta = \beta + DX\beta.$$

（$\because \forall\beta$成立）

$\therefore$这表明$DX = 0$。我们进而有

$$\begin{aligned}\mathrm{cov}(\tilde{\beta}) &= B\mathrm{cov}(y)B^T = \sigma^2\{(X^TX)^{-1}X^T + D\}\{X(X^TX)^{-1} + D^T\} \\ &= \sigma^2\{(X^TX)^{-1} + \underline{DX}(X^TX)^{-1} + (X^TX)^{-1}(\underline{DX})^T + DD^T\} \\ &= \sigma^2\{(X^TX)^{-1} + DD^T\}.\end{aligned}$$

故$\mathrm{cov}(\tilde{\beta}) - \mathrm{cov}(\hat{\beta}) = \sigma^2DD^T$，而$\mathrm{var}(v^T\tilde{\beta}) - \mathrm{var}(v^T\hat{\beta}) = \sigma^2v^TDD^Tv = \sigma^2||D^Tv||^2 \geq 0$。

$$E(e^Te) = E\{y^T(I-H)^T(I-H)y\} = E\{y^T(I-H)y\}$$

$$= E(y)^T(I-H)E(y) + tr\{(I-H)cov(y)\}$$

$$= \beta^T X^T (I - H) X\beta + \sigma^2 tr(I - H)$$

idempotent $= \sigma^2 tr(I - H)$

$$= \sigma^2 (tr(I) - tr(H))$$

$$= \sigma^2 (n - p) \quad \square$$

or $E(e^Te) = E(\Sigma e_i^2)$

$$= \Sigma [E(e_i^2) - (E(e_i))^2] \qquad (\because E(e_i) = 0)$$

idempotent $= \Sigma \, var(e_i)$

$$= tr[cov(e, e)]$$

$$= tr[\sigma^2 (I - H)]$$

$$= \sigma^2 (n - p) \quad \square$$