回归分析:最小二乘统计推断

学业辅导中心

注记. 上一节中为什么要求X列满秩即可? 这是因为$AX = 0$与$A^\top AX = 0$同解. 有$r(X) = r(X^\top X) = p$. 线性方程组同解用高等代数的方法证明即可.

注记. 这个笔记中不区分p和$p+1$, 即有时考虑截距项, 有时不考虑, 这导致了记号混用, 学习时对照应加以区分.

最小二乘估计只是一个方法, 不需要任何假设, 但如果研究最小二乘估计量有什么性质, 则需要一定的假设. 我们想要知道:

1. 最小二乘估计的点估计, 方差估计, 为什么我们选用最小二乘法?
2. 最小二乘估计有什么统计性质，如何做统计推断

研究第一个问题用Gauss-Markov model, 研究第二个问题用Normal linear model.

1 Gauss-Markov模型

定义 (Gauss-Markov model)**.**

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

其中, 设计矩阵X是固定的且有线性无关的列向量, 而且随机误差项ε满足

$$E(\varepsilon) = 0 \quad Var(\varepsilon) = \sigma^2 I_n$$

其中β, σ^2是未知常数.

1. 没有这些随机性假定(stochastic assumptions), 无法讨论期望, 方差, 分布等等;
2. 统计中一般假设X是固定的, 计量经济中假设X随机, 两者理论是等价的, 因为一旦条件在X上就相当于固定X;
3. 不要求ε是normal distribution;
4. Gauss-Markov假定限制了线性性质和同方差性(Homoskedasticity)

$$E(Y) = X\beta \quad cov(Y) = \sigma^2 I_n$$

1.1 OLS估计的均值和方差

- $E(\hat{\beta}) = \beta$
- $\mathrm{cov}(\hat{\beta}) = \sigma^2 (X^{\mathrm{T}} X)^{-1}$.

证明. •

$$E(\hat{\beta}) = E\left\{(X^T X)^{-1} X^T Y\right\} = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X\beta = \beta.$$

•

$$\begin{aligned}\mathrm{cov}(\hat{\beta}) &= \mathrm{cov}\left\{(X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Y\right\} \\ &= (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \mathrm{cov}(Y) X (X^{\mathrm{T}} X)^{-1} \\ &= \sigma^2 (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} X (X^{\mathrm{T}} X)^{-1} \\ &= \sigma^2 (X^{\mathrm{T}} X)^{-1}.\end{aligned}$$

□

推论. *1.* $\mathrm{E}(c^T \hat{\boldsymbol{\beta}}) = c^T \boldsymbol{\beta}$.

2. $\mathrm{cov}(c^{\top} \hat{\boldsymbol{\beta}}) = \sigma^2 c^{\top} (X^{\top} X)^{-1} c$.

注记. 在课上我们使用的记号是:

1. *SST: Sum of Squares Total*
2. *SSR: Sum of Squares Regression*
3. *SSE: Sum of Squares Error*

有时, *SSE*也被称为*RSS(residual sum of squares).* 以后可能出现记号混用.

性质.

$$E\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix} = \begin{pmatrix}X\beta\\ 0\end{pmatrix} \quad \mathrm{cov}\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix} = \sigma^2 \begin{pmatrix}H & 0_{n\times n}\\ 0_{n\times n} & I_n - H\end{pmatrix}$$

注记. 上述性质表明$\hat{Y}$和$\hat{\varepsilon}$是不相关的.

证明. 由于

$$\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix} = \begin{pmatrix}HY\\ (I_n - H)Y\end{pmatrix} = \begin{pmatrix}H\\ I_n - H\end{pmatrix} Y$$

从而

$$\begin{aligned}E\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix} &= \begin{pmatrix}H\\ I_n - H\end{pmatrix} E(Y) \\ &= \begin{pmatrix}H\\ I_n - H\end{pmatrix} X\beta = \begin{pmatrix}HX\beta\\ (I_n - H)X\beta\end{pmatrix} = \begin{pmatrix}X\beta\\ 0\end{pmatrix}\end{aligned}$$

且有 □

$$\begin{aligned}\operatorname{cov}\begin{pmatrix}\hat{Y}\\ \hat{\varepsilon}\end{pmatrix} &= \begin{pmatrix}H\\ I_n-H\end{pmatrix}\operatorname{cov}(Y)\begin{pmatrix}H^{\mathrm{T}} & (I_n-H)^{\mathrm{T}}\end{pmatrix}\\ &= \sigma^2\begin{pmatrix}H\\ I_n-H\end{pmatrix}\begin{pmatrix}H & I_n-H\end{pmatrix}\\ &= \sigma^2\begin{pmatrix}H^2 & H(I_n-H)\\ (I_n-H)H & (I_n-H)^2\end{pmatrix}\\ &= \sigma^2\begin{pmatrix}H & 0\\ 0 & I_n-H\end{pmatrix},\end{aligned}$$

注意: 虽然原始的随机项彼此是不相关的, 但是我们看到$\hat{y}_i, \hat{y}_j$不是独立的, $\hat{\epsilon}_i, \hat{\epsilon}_j$不是独立的, 因为

$$\operatorname{cov}(\hat{y}_i, \hat{y}_j) = \sigma^2 h_{ij}, \quad \operatorname{cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \sigma^2(\delta_{ij} - h_{ij})$$

我们可以看到, 协方差矩阵与σ^2有关, 因此我们要估计它.

1.2 方差估计

σ^2是ε的方差, 一个直观的想法是用$\hat{\varepsilon}$估计方差, 由于$E(\hat{\varepsilon}_i^2) = Var(\hat{\varepsilon}_i) = \sigma^2(1-h_{ii})$. 于是

$$E(RSS) = \sum_{i=1}^{n}\sigma^2(1-h_{ii}) = \sigma^2\{n - tr(H)\} = \sigma^2(n-p)$$

因此方差的无偏估计量是:

$$\hat{\sigma}^2 = \mathrm{RSS}/(n-p) = \sum_{i=1}^{n}\hat{\varepsilon}_i^2/(n-p).$$

注记 (自由度)**.** 上述自由度是$n-p$, 它描述了有多少个随机变量可以独立地变化. 自由度 = 独立随机变量的个数-限制的个数. 在回归中, *RSS(SSE)*为$\sum(y_i - x_i^{\top}\beta)^2$, 有$n$个独立变量$(x_i, y_i)$, 受$p$个参数限制$(\beta)$.

1.3 Gauss-Markov定理

为什么我们选用最小二乘法?

定理 (BLUE)**.** 在β的所有线性(关于Y)无偏估计中, 最小二乘估计量是唯一的具有最小方差的线性无偏估计.

注记. 也就是说, 若$\tilde{\beta}$满足:$\tilde{\beta} = AY, \exists A \in \mathbb{R}^p$与$Y$无关; 且$E(\tilde{\beta}) = \beta, \forall\beta$, 则$cov(\tilde{\beta}) \geq cov(\hat{\beta}_{ols})$.

但是A可以与X有关.

Definition: We call A positive semi-definite, denoted by $A \succeq 0$, if $x^T Ax \geq 0$ for all x; we call A positive definite, denoted by $A \succ 0$, if $x^T Ax > 0$ for all nonzero x.

We call $A \succeq B$ if and only if $A - B \succeq 0$, and we call $A \succ B$ if and only if $A - B \succ 0$.

证明. 由于

$$\begin{aligned}\operatorname{cov}(\tilde{\beta}) &= \operatorname{cov}(\hat{\beta} + \tilde{\beta} - \hat{\beta})\\ &= \operatorname{cov}(\hat{\beta}) + \operatorname{cov}(\tilde{\beta} - \hat{\beta}) + \operatorname{cov}(\hat{\beta}, \tilde{\beta} - \hat{\beta}) + \operatorname{cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}).\end{aligned}$$

于是

$$\begin{aligned}\operatorname{cov}(\hat{\beta}, \tilde{\beta}-\hat{\beta}) = & \operatorname{cov}\left\{\hat{A}Y, (A-\hat{A})Y\right\} \\ = & \hat{A}\operatorname{cov}(Y)(A-\hat{A})^{\mathrm{T}} \\ = & \sigma^2\hat{A}(A-\hat{A})^{\mathrm{T}} \\ = & \sigma^2(\hat{A}A^{\mathrm{T}}-\hat{A}\hat{A}^{\mathrm{T}}) \\ = & \sigma^2\left\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}A^{\mathrm{T}}-(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}\right\} \\ = & \sigma^2\left\{(X^{\mathrm{T}}X)^{-1}I_p-(X^{\mathrm{T}}X)^{-1}\right\} \quad (\text{because } AX=I_p) \\ = & 0.\end{aligned}$$

从而

$$cov(\tilde{\beta}) - cov(\hat{\beta}) = cov(\tilde{\beta}-\hat{\beta}) \geq 0$$

□

注记. 无偏性$\Leftrightarrow AX = I_p$

$$\beta = E(\tilde{\beta}) = E(AY) = AE(Y) = AX\beta$$

1.4 推广: 加权最小二乘

加权最小二乘就是一族无偏的估计量, 但方差未必最小, 因此在应用中, 有效性(efficiency)不及OLS.

$$\tilde{\beta} = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}Y,$$

$$E(\tilde{\beta}) = E\left\{(X^{\mathrm{i}}\Sigma^{-1}X)^{-1}X^{\mathrm{i}}\Sigma^{-1}Y\right\} = (X^{\mathrm{i}}\Sigma^{-1}X)^{-1}X^{\mathrm{i}}\Sigma^{-1}X\beta = \beta.$$

2 Normal Linear Model

在Gauss-Markov model中, 我们不加额外的分布假定, 就得到了关于参数的均值, 方差估计(数字特征). 但是Gauss-Markov模型没有给出$\hat{\beta}$的分布, 因此无法进一步推断, 这就要求我们进一步假设.

定义 (Normal Linear Model)**.**

$$\begin{aligned}& Y \sim \mathrm{N}(X\beta, \sigma^2 I_n) \\ \Longleftrightarrow\ & y_i \overset{\mathrm{IND}}{\sim} \mathrm{N}(x_i^{\mathrm{T}}\beta, \sigma^2), \quad (i=1,\ldots,n)\end{aligned}$$

或

$$Y = X\beta + \varepsilon \Longleftrightarrow y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i, \quad (i=1,\ldots,n),$$

其中

$$\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n) \text{ 或 } \varepsilon_i \overset{\mathrm{IID}}{\sim} \mathrm{N}(0, \sigma^2).$$

2.1 多元正态分布及一些性质

定义 2.2.1 设 $U=(U_1,\cdots,U_q)'$ 为随机向量，$U_1,\cdots,U_q$ 相互独立且同 $N(0,1)$分布；设 μ 为 p 维常数向量，A 为 $p\times q$ 常数矩阵，则称 $X=AU+\mu$ 的分布为 p **元正态分布**，或称 X 为 p **维正态随机向量**，记为 $X\sim N_p(\mu,AA')$.

简单地说，由 q 个相互独立的标准正态随机变量的一些线性组合所构成的随机向量的分布，称其为**多元正态分布**.

在一元统计中，若 $X\sim N(\mu,\sigma^2)$，则 X 的特征函数为

$$\varphi(t)=\mathrm{E}(\mathrm{e}^{\mathrm{i}tX})=\exp\left[\mathrm{i}t\mu-\frac{1}{2}t^2\sigma^2\right].$$

将其推广到多维正态随机向量的情况有如下性质.

性质 1 设 $U=(U_1,\cdots,U_q)'$ 为随机向量，$U_1,\cdots,U_q$ 相互独立且同 $N(0,1)$分布；令 $X=AU+\mu$，则 X 的特征函数为

$$\Phi_X(t)=\exp\left[\mathrm{i}t'\mu-\frac{1}{2}t'AA't\right].$$

定义 2.2.2 若 p 维随机向量 X 的特征函数为

$$\Phi_X(t)=\exp\left[\mathrm{i}t'\mu-\frac{1}{2}t'\Sigma t\right]\quad(\Sigma\geqslant 0),$$

则称 X 服从 p 元正态分布，记为 $X\sim N_p(\mu,\Sigma)$.

性质 2 设 $X\sim N_p(\mu,\Sigma)$，B 为 $s\times p$ 常数矩阵，d 为 s 维常向量，令 $Z=BX+d$，则 $Z\sim N_s(B\mu+d,B\Sigma B')$.

定义 2.2.3 若 p 维随机向量 X 的任意线性组合均服从一元正态分布，则称 X 为 p **维正态随机向量**.

在概率论中大家知道，一元正态随机变量的密度函数是

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(\sigma>0,\ -\infty<x<\infty).$$

这个式子又可改写为

$$f(x)=\frac{1}{(2\pi)^{1/2}|\sigma^2|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'(\sigma^2)^{-1}(x-\mu)\right].$$

作为一元正态随机变量的推广，以下来导出多维正态随机向量的联合密度函数.

性质 5 设 $X\sim N_p(\mu,\Sigma)$，且 $\Sigma>0$(正定)，则 X 的联合密度函数为

$$f(x)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right].$$

推论 设 $X=\begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}\begin{matrix} r \\ p-r \end{matrix} \sim N_p(\mu,\Sigma)$，将 μ,Σ 剖分为

$$\mu=\begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}\begin{matrix} r \\ p-r \end{matrix}, \quad \Sigma=\begin{bmatrix} \overset{r}{\Sigma_{11}} & \overset{p-r}{\Sigma_{12}} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\begin{matrix} r \\ p-r \end{matrix},$$

则 $X^{(1)}\sim N_r(\mu^{(1)},\Sigma_{11})$，$X^{(2)}\sim N_{p-r}(\mu^{(2)},\Sigma_{22})$.

定理 2.3.1 设 p 维随机向量 $X\sim N_p(\mu,\Sigma)$，

$$X=\begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}\sim N_p\left(\begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix},\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),$$

则

$$X^{(1)} \text{ 与 } X^{(2)} \text{ 相互独立} \iff \Sigma_{12}=O$$

(即 $X^{(1)}$ 与 $X^{(2)}$ 互不相关).

定理 2.3.2 设 $X=\begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}\begin{matrix} r \\ p-r \end{matrix}\sim N_p(\mu,\Sigma)$ $(\Sigma>0)$，则当 $X^{(2)}$ 给定时，$X^{(1)}$ 的条件分布为

$$(X^{(1)}|X^{(2)})\sim N_r(\mu_{1\cdot 2},\Sigma_{11\cdot 2}),$$

其中

$$\mu_{1\cdot 2}=\mu^{(1)}+\Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)}),$$
$$\Sigma_{11\cdot 2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

推论 在定理 2.3.2 条件下可得：

(1) $X^{(2)}$ 与 $X^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}X^{(2)}$ 相互独立；

(2) $X^{(1)}$ 与 $X^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}X^{(1)}$ 相互独立；

例 2.2.1 (二元正态分布) 设 $X=\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\sim N_2(\mu,\Sigma)$，记

$$\mu=\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma=\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}\overset{\text{def}}{=\!=}\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}>0$$

(即 $\sigma_1>0,\sigma_2>0,|\rho|<1$).

结论 3 设 $X\sim N_n(0_n,\sigma^2 I_n)$，$A$ 为对称矩阵，且 $\operatorname{rank}(A)=r$，则二次型 $X'AX/\sigma^2\sim\chi^2(r)\iff A^2=A$ (A 为**对称幂等矩阵**).

结论 2 设 $X\sim N_p(\mu,\Sigma),\Sigma>0,A$ 为对称矩阵，$\mathrm{rank}(A)=r$. 则

$$(X-\mu)'A(X-\mu)\sim\chi^2(r)\iff\Sigma A\Sigma A\Sigma=\Sigma A\Sigma.$$

特别的, 上述的A可以取Σ^{-1}.

2.2 参数估计的联合分布

性质 (联合分布)**.**

$$\begin{pmatrix}\hat\beta\\ \hat\varepsilon\end{pmatrix}\sim \mathrm{N}\left\{\begin{pmatrix}\beta\\ 0\end{pmatrix},\right\},\sigma^2\begin{pmatrix}(X^{\mathrm{T}}X)^{-1} & 0\\ 0 & I_n-H\end{pmatrix}$$

且有

$$\hat\sigma^2/\sigma^2\sim\chi^2_{n-p}/(n-p).$$

从而$\hat\beta\perp\!\!\!\perp\hat\varepsilon,\hat\beta\perp\!\!\!\perp\hat\sigma^2$.

证明.

$$\begin{pmatrix}\hat\beta\\ \hat\varepsilon\end{pmatrix}=\begin{pmatrix}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\\ (I_n-H)Y\end{pmatrix}=\begin{pmatrix}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\\ I_n-H\end{pmatrix}Y$$

$$\mathrm{cov}(\hat\beta,\hat\varepsilon)=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{cov}(Y)(I_n-H)^{\mathrm{T}}=\sigma^2(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(I_n-H^{\mathrm{T}})=0$$

□

2.3 统计推断

2.3.1 标量的形式

对于$c^T\hat\beta$推断(特别的, 对单参数推断只需要取特别的c),

$$c^{\mathrm{T}}\hat\beta\sim\mathrm{N}\left\{c^{\mathrm{T}}\beta,\sigma^2c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c\right\}.$$

但是如果σ不知道, 我们就不能采用上面的式子, 可以类似数理统计使用t统计量.

$$T_c\equiv\frac{c^{\mathrm{T}}\hat\beta-c^{\mathrm{T}}\beta}{\sqrt{\hat\sigma^2c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}}\sim t_{n-p}$$

我们经常称$\sqrt{\hat\sigma^2c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}$是参数$c^T\hat\beta$估计的标准差$\hat{se}_c$. 关于置信区间和假设检验, 可以类似数理统计的方法,

$$\mathrm{pr}\left\{c^{\mathrm{T}}\hat\beta-t_{1-\alpha/2,n-p}\hat{\mathrm{se}}_c\le c^{\mathrm{T}}\beta\le c^{\mathrm{T}}\hat\beta+t_{1-\alpha/2,n-p}\hat{\mathrm{se}}_c\right\}=1-\alpha.$$

2.3.2 向量的联合置信区间(confidence region)

引理

证明 当 $b=0$ 或 $d=0$，显然成立. 不妨设 $b\neq 0, d\neq 0$，考虑向量 $b-xd$，这里 x 是一个变数，于是

$$0 \leqslant (b-xd)'(b-xd) = (d'd)x^2 - (2b'd)x + b'b, \tag{7.2}$$

由 x 的二次函数性质知道此时

$$(b'd)^2 - (b'b)(d'd) \leqslant 0.$$

即(7.1)式成立.

若 $b=cd$，立即可知(7.1)式等号成立. 反之若(7.1)式等号成立，则当取 $x=b'd/d'd$ 时(7.2)式为 0，从而存在常数 $c=b'd/d'd$，使得

$$b = cd. \qquad \textbf{(证毕)}$$

引理 7.2（推广的柯西-施瓦茨不等式） 设 b,d 是两个 p 维向量，B 是 p 阶正定矩阵，那么有

$$(b'd)^2 \leqslant (b'Bb)(d'B^{-1}d), \tag{7.3}$$

且等号当且仅当 $b=cB^{-1}d$（或 $d=cBb$）时成立，这里 c 为常数.

证明 由于 B 正定，记 $\tilde{b}=B^{1/2}b$，$\tilde{d}=B^{-1/2}d$，此时 b 与 $\tilde{b}$，d 与 $\tilde{d}$ 同时为零向量或者非零向量，利用引理 7.1 于向量 $\tilde{b}$ 和 $\tilde{d}$，立即可得(7.3)式. **(证毕)**

定理 7.1 设 B 是 p 阶正定矩阵，d 为 p 维向量，对任意 p 维向量 x 下式成立：

$$\max_{x\neq 0} \frac{(x'd)^2}{x'Bx} = d'B^{-1}d, \tag{7.4}$$

且当 $x=cB^{-1}d$ 时达到最大值 $d'B^{-1}d$ ($c\neq 0$ 为常数).

证明 因为 B 是 p 阶正定矩阵，由引理 7.2 知道，对任一 p 维向量 $x\neq 0$，$x'Bx>0$ 成立，以及

$$\frac{(x'd)^2}{x'Bx} \leqslant d'B^{-1}d,$$

且当 $x=cB^{-1}d$ 时等号成立，定理得证. **(证毕)**

对给定的样本 $X_{(t)}(t=1,2,\cdots,n)$和系数向量 a，若全体 $a'\mu$ 值的置信区间是由(3.2.2)式给出的，则不等式

$$t^2 = \frac{n[a'(\bar{X}-\mu)]^2}{a'Sa} \leqslant t_{\alpha/2}^2$$

成立. 若让 a 变化，求所有 $a'\mu$ 的联立置信区间，那么应将(3.2.2)式的右边换上更大的常数才较为合理. 为此来求最大值

$$\max_{a\neq 0} t^2 = \max_{a\neq 0} \frac{n[a'(\bar{X}-\mu)]^2}{a'Sa}.$$

根据附录中定理 7.1 有

$$\max_{a\neq 0}\frac{n[a'(\overline{X}-\mu)]^2}{a'Sa}=n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu)=T^2,\tag{3.2.3}$$

且最大值在 a 与 $S^{-1}(\overline{X}-\mu)$ 成比例时达到.

定理 3.2.2 假设 $X_{(t)}(t=1,2,\cdots,n)$ 为来自 p 元正态总体 $N_p(\mu,\Sigma)$ ($\Sigma>0$ 未知)的随机样本,则对所有的 a,区间

$$[a'\overline{X}-d,a'\overline{X}+d]\quad\left(\text{其中 }d=\sqrt{\frac{(n-1)p}{n(n-p)}F_\alpha a'Sa}\right)$$

包含 $a'\mu$ 的概率为 $1-\alpha$ (其中 F_α 满足(3.2.1)式).

证明 由(3.2.3)式知, $T^2=n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu)\leqslant c^2$ 意味着对一切 a,有

$$\frac{n[a'(\overline{X}-\mu)]^2}{a'Sa}\leqslant c^2,$$

即对一切 a,有

$$a'\overline{X}-c\sqrt{\frac{a'Sa}{n}}\leqslant a'\mu\leqslant a'\overline{X}+c\sqrt{\frac{a'Sa}{n}}.$$

取 $c^2=\frac{(n-1)p}{n-p}F_\alpha$ (F_α 满足(3.2.1)式),故对所有 a,则有

$$P\{T^2\leqslant c^2\}=1-\alpha.\qquad\text{(证毕)}$$

2.4 预测点

若有新的数据(x_{n+1}, y_{n+1}), 我们只观察到x_{n+1}并想基于原有的数据(X, Y)进行预测. 假设数据之间的关系保持不变,

$$y_{n+1}\sim \mathrm{N}(x_{n+1}^{\mathrm{T}}\beta,\sigma^2)$$

且它们有相同的参数(β,σ^2).

以下三种情况的点估计是一样的, 均是$x_{n+1}^T\beta$(取$c=x_{n+1}^T$即可). 不同的是方差估计(置信区间).

2.4.1 y_{n+1}的均值

置信区间, 直接应用数理统计的结论:

$$x_{n+1}^{\mathrm{T}}\hat{\beta}\pm t_{1-\alpha/2,n-p}\hat{\mathrm{se}}_{x_{n+1}}$$

2.4.2 y_{n+1}

由于

$$y_{n+1} \sim \mathrm{N}(x_{n+1}^{\mathrm{T}}\beta, \sigma^2)$$

从而

$$y_{n+1} - x_{n+1}^{\mathrm{T}}\hat{\beta} \sim \mathrm{N}\left\{0, \sigma^2 + \sigma^2 x_{n+1}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_{n+1}\right\}$$

注记. 为加以区分可以采用*PPT*上的记号.

$$\begin{aligned}\frac{y_{n+1} - x_{n+1}^{\mathrm{T}}\hat{\beta}}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 x_{n+1}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_{n+1}}} &= \frac{y_{n+1} - x_{n+1}^{\mathrm{T}}\hat{\beta}}{\sqrt{\sigma^2 + \sigma^2 x_{n+1}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_{n+1}}} \Big/ \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \\ &\sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi_{n-p}^2/(n-p)}} \\ &\sim t_{n-p}\end{aligned}$$

于是它的置信区间是:

$$x_{n+1}^{\mathrm{T}}\hat{\beta} \pm t_{1-\alpha/2,n-p}\hat{\mathrm{pe}}_{x_{n+1}}$$

其中

$$\begin{aligned}\hat{\mathrm{pe}}_{x_{n+1}}^2 &= \hat{\sigma}^2 + \hat{\sigma}^2 x_{n+1}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_{n+1} \\ &= \hat{\sigma}^2\left\{1 + n^{-1}x_{n+1}^{\mathrm{T}}\left(n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}x_{n+1}\right\}\end{aligned}$$

2.4.3 m个y_{n+1}

$$\begin{aligned}\hat{\mathrm{pe}}_{x_{n+1}}^2 &= \frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2 x_{n+1}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_{n+1} \\ &= \hat{\sigma}^2\left\{\frac{1}{m} + n^{-1}x_{n+1}^{\mathrm{T}}\left(n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}x_{n+1}\right\}\end{aligned}$$

3 应用

```
> round(summary(galton.lm)$coef, 3)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       22.636      4.265   5.307        0
midparentHeight    0.637      0.062  10.345        0

> round(head(cbind(ci.new, pi.new)), 3)
     fit    lwr    upr    fit    lwr    upr
1 60.878 59.744 62.012 60.878 54.126 67.630
2 61.197 60.122 62.272 61.197 54.454 67.939
3 61.515 60.499 62.531 61.515 54.782 68.249
4 61.834 60.877 62.791 61.834 55.109 68.559
5 62.153 61.254 63.051 62.153 55.436 68.869
```

```
6 62.471 61.632 63.311 62.471 55.762 69.180


> lalonde.lm <- lm(re78 ~., data = lalonde) # regression on all covariates
> summary(lalonde.lm)

Call:
lm(formula = re78 ~ ., data = lalonde)

Residuals:
   Min     1Q Median     3Q    Max
 -9612  -4355  -1572   3054  53119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.567e+02  3.522e+03   0.073  0.94193
age          5.357e+01  4.581e+01   1.170  0.24284
educ         4.008e+02  2.288e+02   1.751  0.08058 .
black       -2.037e+03  1.174e+03  -1.736  0.08331 .
hisp         4.258e+02  1.565e+03   0.272  0.78562
married     -1.463e+02  8.823e+02  -0.166  0.86835
nodegr      -1.518e+01  1.006e+03  -0.015  0.98797
re74         1.234e-01  8.784e-02   1.405  0.16079
re75         1.974e-02  1.503e-01   0.131  0.89554
u74          1.380e+03  1.188e+03   1.162  0.24590
u75         -1.071e+03  1.025e+03  -1.045  0.29651
treat        1.671e+03  6.411e+02   2.606  0.00948 **
---
Signif. codes:  0 ‘’ *** 0.001 ‘’ ** 0.01 ‘’ * 0.05 ‘’ . 0.1 ‘ ’ 1

Residual standard error: 6517 on 433 degrees of freedom
Multiple R-squared:  0.05822,   Adjusted R-squared:  0.0343
F-statistic: 2.433 on 11 and 433 DF,  p-value: 0.005974


> linearHypothesis(lalonde.lm, c("age=0", "educ=0", "black=0", "hisp=0", "married=0", "nodegr=0",
    "re74=0", "re75=0", "u74=0", "u75=0"))
Linear hypothesis test

Hypothesis:
age = 0
educ = 0
black = 0
hisp = 0
married = 0
nodegr = 0
re74 = 0
re75 = 0
u74 = 0
u75 = 0

Model 1: restricted model
Model 2: re78 ~ age + educ + black + hisp + married + nodegr + re74 +
    re75 + u74 + u75 + treat

```

```
  Res.Df        RSS Df Sum of Sq      F  Pr(>F)
1    443 1.9178e+10
2    433 1.8389e+10 10 788799023 1.8574 0.04929 *
---
Signif. codes:  0 ‘’ *** 0.001 ‘’ ** 0.01 ‘’ * 0.05 ‘’ . 0.1 ‘ ’ 1
```

Dataset used by Dehejia and Wahba (1999) to evaluate propensity score matching.