4.12 HW 12

作业 12 链接

✍ **练习 4.1** 假设 $(y_i, \mathbf{x}_i), i = 1, \ldots, n$ 满足线性模型

$$y_i = \beta_0 + \mathbf{x}_i^\top \mathbf{b} + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2), \quad i = 1, \ldots, n.$$

若对某个 $1 \le k \le n, y_k = \bar{y} = \sum_{i=1}^n y_i/n, \mathbf{x}_k = \bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$, 证明删除 $(y_k, \mathbf{x}_k)$ 不影响最小二乘法。

证明 线性模型可写为 $y = X\beta + \varepsilon$, $\hat{\beta}$ 为 β 的 LS 估计

设 $\widetilde{\beta}$ 为删除 $(\mathbf{x}_k, y_k)$ 后 β 的 LS 估计, 根据 ppt week13 P29 命题 2

$$\widetilde{\beta} = \hat{\beta} - \frac{\left(X^\top X\right)^{-1} \mathbf{x}_k e_k}{1 - h_{ii}}$$

$$\begin{aligned} \text{其中} e_k &= y_k - \mathbf{x}_k^\top \beta \\ &= \bar{y} - \bar{\mathbf{x}}^\top \beta \\ &= \frac{1}{n}\left(\mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}\right) \\ &= \frac{1}{n} \mathbf{1}^\top \left(\mathbf{I} - \mathbf{P}_\mathbf{X}\right) \mathbf{y} \end{aligned}$$

由于 $\mathbf{1} \in \mathbf{C}(\mathbf{X})$, 则 $\mathbf{1}^\top \left(\mathbf{I} - \mathbf{P}_\mathbf{X}\right) = \mathbf{0}$, 所以 $\tilde{\beta} = \hat{\beta}$

✍ **练习 4.2** 假设 $\mathbf{y}_1, \ldots, \mathbf{y}_n$ iid $\sim N\left(\theta, \sigma^2 I_p\right)$, 其中 θ 为 $p \times 1$ 未知参数向量, 假设 σ^2 已知。令 $\tilde{\theta} = \lambda \bar{\mathbf{y}}$, 其中 $0 \le \lambda \le 1$ 是常数。

(a) 求 $\tilde{\theta}$ 的均方误差 $m(\lambda) = E\|\tilde{\theta} - \theta\|^2$.

(b) 求 $\lambda_{opt} = \arg\min_\lambda m(\lambda)$ (即求使得 $m(\lambda)$ 达到最小的 λ).

(c) λ_{opt} 中含有未知 $\|\theta\|^2$, 以其无偏估计代入 (注意 $\|\bar{\mathbf{y}}\|^2$ 不是 $\|\theta\|^2$ 的无偏估计), 得到的 $\tilde{\theta} = \hat{\lambda}_{opt} \bar{\mathbf{y}}$ 是否等于或接近 James-Stein 估计?

解

(a) 因为

(i) $\mathrm{Var}(\tilde{\theta}) = \lambda^2 \mathrm{Var}(\bar{y}) = \lambda^2 \sigma^2 I_p/n$,

(ii) bias $= E\tilde{\theta} - \theta = \lambda\theta - \theta$

所以

$$m(\lambda) = E\|\tilde{\theta} - \theta\|^2 = \mathrm{tr}(\mathrm{Var}(\tilde{\theta})) + \|bias\|^2 = \lambda^2 \sigma^2 p/n + (\lambda - 1)^2 \|\theta\|^2$$

(b)

$$\lambda_{\text{opt}} = \underset{\lambda}{\text{argmin}} m(\lambda) = \frac{\theta^\top \theta}{\frac{p\sigma^2}{n} + \theta^\top \theta}$$

(c)

$$\begin{aligned} E\|\bar{y}\|^2 &= E\bar{y}^\top \bar{y} \\ &= \frac{np\sigma^2 + n^2 \theta^\top \theta}{n^2} \\ &= \frac{p\sigma^2}{n} + \theta^\top \theta \end{aligned}$$

$\|\bar{y}\|^2 - \frac{p\sigma^2}{n}$ 为 $\theta^\top \theta$ 的无偏估计

$$\begin{aligned} \tilde{\theta} = \hat{\lambda}_{\text{opt}} \bar{y} &= \frac{\|\bar{y}\|^2 - \frac{p\sigma^2}{n}}{\|\bar{y}\|^2} \bar{y} \\ &= \left(1 - \frac{p\sigma^2}{n\|\bar{y}\|^2}\right) \bar{y} \end{aligned}$$

$$\text{James - Stein 估计}\hat{\theta}_{J-S}=\left(1-\frac{(p-2)\sigma^2}{n\|\bar{y}\|^2}\right)\bar{y}$$

✍ **练习 4.3** 假设模型

$$\mathbf{y}_{n\times 1}=X_{n\times p}\beta_{p\times 1}+\epsilon_{n\times 1},\epsilon\sim\left(0,\sigma^2 I_n\right),$$

其中 X 的第一列为向量 $\mathbf{1}$. 回归系数的最小二乘估计记为 $\hat{\beta}=\left(X^\top X\right)^{-1}X^\top\mathbf{y}$ 。假设有“新”数据 $\mathbf{x}_0,y_0$ 满足模型 $y_0=\mathbf{x}_0^\top\beta+\epsilon_0,\epsilon_0\sim\left(0,\sigma^2\right)$, 其中 $\mathbf{x}_0$ 已知, 需要预测 y_0 。预测统计量取为 $\hat{y}_0=\mathbf{x}_0^\top\hat{\boldsymbol{\beta}}$ 证明如下结论

(a) $\hat{y}_0$ 的预测误差为 $pe\left(\hat{y}_0\right)=E\left(\hat{y}_0-y_0\right)^2=\sigma^2\left[1+\mathbf{x}_0^\top\left(X^\top X\right)^{-1}\mathbf{x}_0\right]$.

(b) 当 $\mathbf{x}_0=\mathbf{x}_i(X$ 的第 i 行 $),pe\left(\hat{y}_0\right)=\left(1+h_{ii}\right)\sigma^2\overset{\text{记作}}{=}e\left(\mathbf{x}_i\right)$, 其中 h_{ii} 为 $H=X\left(X^\top X\right)^{-1}X^\top$ 的 (i,i) 元;

(c) 当 $\mathbf{x}_0=\bar{\mathbf{x}},\hat{y}_0=\bar{y}=\mathbf{1}^\top\mathbf{y}/n$, 且 $pe\left(\hat{y}_0\right)=(1+1/n)\sigma^2\overset{\text{记作}}{=}e(\bar{\mathbf{x}})$, 说明 $e(\bar{\mathbf{x}})\leq e\left(\mathbf{x}_i\right),i=1,\ldots,n$.

(d) 当 $\mathbf{x}_0=X^\top\mathbf{a}$ (这里 $\mathbf{a}\in R^n$) $,pe\left(\hat{y}_0\right)\leq\sigma^2\left(1+\|\mathbf{a}\|^2\right)$.

证明

(a)

$$\begin{aligned}\mathrm{pe}\left(\hat{y}_0\right)&=E\left(\hat{y}_0-y_0\right)^2\\&=E\left(\mathbf{x}_0^\top\hat{\beta}-\mathbf{x}_0^\top\beta-\varepsilon_0\right)^2\\&=E\left(\mathbf{x}_0^\top\hat{\beta}-\mathbf{x}_0^\top\beta\right)^2+E\varepsilon_0^2\\&=\mathrm{Var}\left(\mathbf{x}_0^\top\hat{\beta}\right)+\sigma^2\\&=\mathbf{x}_0^\top\mathrm{Var}(\hat{\beta})\mathbf{x}_0+\sigma^2\\&=\mathbf{x}_0^\top\mathrm{Var}\left(\left(X^\top X\right)^{-1}X^\top y\right)\mathbf{x}_0+\sigma^2\\&=\left(1+\mathbf{x}_0^\top\left(X^\top X\right)^{-1}\mathbf{x}_0\right)\sigma^2\end{aligned}$$

(b) 当 $\mathbf{x}_0=\mathbf{x}_i$ 时,

$$\begin{aligned}\mathbf{x}_i\left(X^\top X\right)^{-1}\mathbf{x}_i&=h_{ii}\\\Rightarrow pe\left(\hat{y}_0\right)&=\left(1+h_{ii}\right)\sigma^2\end{aligned}$$

(c) 当 $\mathbf{x}_0=\bar{\mathbf{x}}$ 时

$$\begin{aligned}\bar{\mathbf{x}}^\top\left(X^\top X\right)^{-1}\bar{\mathbf{x}}&=\frac{1}{n^2}\mathbf{1}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{1}\\&=\frac{1}{n^2}\mathbf{1}^\top\mathbf{P}_\mathbf{X}\mathbf{1}\\&=\frac{1}{n^2}\mathbf{1}^\top\mathbf{1}\\&=\frac{1}{n}\\\mathrm{pe}\left(\hat{y}_0\right)&=\left(1+\frac{1}{n}\right)\sigma^2\end{aligned}$$

根据 ppt weck 13 命题 1 可知$h_{ii}\geq\frac{1}{n}$

所以 $e(\bar{\mathbf{x}})\leqslant e\left(\mathbf{x}_i\right)$

(d) 当 $\mathbf{x}_0=X^\top a$ 时

$$\begin{aligned}pe\left(\hat{y}_0\right)&=\left(a^\top P_X a+1\right)\sigma^2\\&=\left(\|P_X a\|^2+1\right)\sigma^2\\&\leqslant\left(\|a\|^2+1\right)\sigma^2\end{aligned}$$

✍ **练习 4.4** 条件同上一题, 对任何给定的 $\mathbf{x}_0\in R^p$, 我们需要预测对应的 y_0, 但 y_0 的预测取为 $\tilde{y}_0=\mathbf{x}_0^\top\tilde{\boldsymbol{\beta}}$, 其中 $\tilde{\beta}=\hat{\beta}\lambda,0\leq\lambda\leq 1$ 。

(a) 证明 $\tilde{y}_0$ 的预测误差为

$$\mathrm{pe}\left(\tilde{y}_0\right)=E\left(\tilde{y}_0-y_0\right)^2=(1-\lambda)^2\left(\mathbf{x}_0^{\top}\beta\right)^2+\lambda^2\sigma^2\mathbf{x}_0^{\top}\left(X^{\top}X\right)^{-1}\mathbf{x}_0+\sigma^2.$$

(b) 证明如果 $\|X\beta\|^2\leq\frac{1+\lambda}{1-\lambda}\sigma^2$, 则 $pe\left(\tilde{y}_0\right)\leq pe\left(\hat{y}_0\right)$, 其中 $\hat{y}_0=\mathbf{x}_0^{\top}\hat{\boldsymbol{\beta}}$ 为基于 LS 估计的预测。

证明

(a)

$$\begin{aligned}
\tilde{y}_0&=\mathbf{x}_0^{\top}\tilde{\beta}=\lambda\mathbf{x}_0^{\top}\hat{\beta}\\
p\left(\tilde{y}_0\right)&=E\left(\tilde{y}_0-y_0\right)^2\\
&=E\left(\lambda\mathbf{x}_0^{\top}\hat{\beta}-\mathbf{x}_0^{\top}\beta-\varepsilon_0\right)^2\\
&=E\left(\lambda\mathbf{x}_0^{\top}\hat{\beta}-\lambda\mathbf{x}_0^{\top}\beta+(\lambda-1)\mathbf{x}_0^{\top}\beta-\varepsilon_0\right)^2\\
&=\mathrm{Var}\left(\lambda\mathbf{x}_0^{\top}\hat{\beta}\right)+(1-\lambda)^2\left(\mathbf{x}_0^{\top}\beta\right)^2+E\varepsilon_0^2\\
&=\lambda^2\mathbf{x}_0^{\top}\mathrm{Var}(\hat{\beta})\mathbf{x}_0+(1-\lambda)^2\left(\mathbf{x}_0^{\top}\beta\right)^2+\sigma^2\\
&=\lambda^2\mathbf{x}_0^{\top}\left(X^{\top}X\right)^{-1}\mathbf{x}_0\sigma^2+(1-\lambda)^2\left(\mathbf{x}_0^{\top}\beta\right)^2+\sigma^2
\end{aligned}$$

(b)

$$\begin{aligned}
\mathrm{pe}\left(\tilde{y}_0\right)&=(1-\lambda)^2\left(\mathbf{x}_0^{\top}\beta\right)^2+\lambda^2\mathbf{x}_0^{\top}\left(X^{\top}X\right)^{-1}\mathbf{x}_0\sigma^2+\sigma^2\\
pe\left(\hat{y}_0\right)&=\mathbf{x}_0^{\top}\left(X^{\top}X\right)^{-1}\mathbf{x}_0\sigma^2+\sigma^2\\
\mathrm{pe}\left(\hat{y}_0\right)-\mathrm{pe}\left(\tilde{y}_0\right)&=\left(1-\lambda^2\right)\mathbf{x}_0^{\top}\left(X^{\top}X\right)^{-1}\mathbf{x}_0\sigma^2-(1-\lambda)^2\left(\mathbf{x}_0^{\top}\beta\right)^2.
\end{aligned}$$

由于$\beta^{\top}X^{\top}X\beta\leqslant\dfrac{1+\lambda}{1-\lambda}\sigma^2$和 ppt week15 p5 引理 1

$$\begin{aligned}
\beta\beta^{\top}&\leqslant\frac{1+\lambda}{1-\lambda}\sigma^2\left(X^{\top}X\right)^{-1}\\
\mathbf{x}_0^{\top}\beta\beta^{\top}\mathbf{x}_0&\leqslant\frac{1+\lambda}{1-\lambda}\mathbf{x}_0^{\top}\left(X^{\top}X\right)^{-1}\mathbf{x}_0\sigma^2\\
&\text{所以}pe\left(\hat{y}_0\right)\geqslant pe\left(\tilde{y}_0\right)
\end{aligned}$$

✍ **练习 4.5** 假设模型 $\mathbf{y}_{n\times1}=X_{n\times p}\beta_{p\times1}+\epsilon_{n\times1},\epsilon\sim\left(0,\sigma^2I_n\right)$ 中 $X,\mathbf{y}$ 都已经中心化 (因此模型中没有截距项), 对设计阵 X 进行奇异值分解:

$$X_{n\times p}=U_{n\times p}D_{p\times p}V_{p\times p}^{\top},$$

其中 $U^{\top}U=I_p,V^{\top}V=I_p,D=\mathrm{diag}\left(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_p}\right)$, 这里 $\lambda_1\geq\ldots\geq\lambda_p>0$ 是 $X^{\top}X$ 的特征根。所以 $X\beta=UDV^{\top}\beta=U\gamma$, 其中 $\gamma=DV^{\top}\beta$ 。由此, 我们改写原模型为

$$\mathbf{y}=U\gamma+\epsilon,\epsilon\sim\left(0,\sigma^2I_n\right),$$

此模型称为主成分回归模型。记 U 的各列为 $U=(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_p)$ 。

(a) 证明主成分回归模型中回归系数 γ 的最小二乘估计为 $\hat{\gamma}=U^{\top}\mathbf{y}$, 由此求出原模型中的 β 的 LS 估计。

(b) 主成分回归模型中, 响应变量的基于最小二乘估计 $\hat{\gamma}$ 的拟合值向量为 $\hat{\mathbf{y}}=U\hat{\gamma}=\sum_{j=1}^{p}\mathbf{u}_j\left(\mathbf{u}_j^{\top}\mathbf{y}\right)$, 求其均方误差 $m(\hat{\mathbf{y}})=E\|\hat{\mathbf{y}}-U\gamma\|^2$.

(c) 设下标集合 $A_q=\{i_1,\ldots,i_q\}\subset\{1,2,\ldots,p\},1\leq q\leq p-1$, 令

$$\tilde{\mathbf{y}}^{(A_q)}=\sum_{j\in A_q}\mathbf{u}_j\left(\mathbf{u}_j^{\top}\mathbf{y}\right).$$

求其均方误差 $m\left(\tilde{\mathbf{y}}^{(A_q)}\right)=E\left\|\tilde{\mathbf{y}}^{(A_q)}-U\gamma\right\|^2$ 。假设 $\|\gamma\|^2\leq\sigma^2$, 证明

$$m\left(\tilde{\mathbf{y}}^{(A_q)}\right)\leq m(\hat{\mathbf{y}}).$$

(d) 写出 C_q 准则的具体表达, 给出最优子集搜索算法。

解

(a)

$$\begin{aligned}
\hat{\gamma} &= \left(U^\top U\right)^{-1} U^\top y \\
&= U^\top y \\
&= DV^\top \beta \\
X\beta &= U\hat{\gamma} \\
&= UU^\top y \\
X\left(X^\top X\right)^{-1} X^\top &= UDV^\top \left(VDU^\top UDV^\top\right)^{-1} VDU^\top \\
&= UU^\top \\
\Rightarrow X\beta &= X\left(X^\top X\right)^{-1} X^\top y \\
\hat{\beta} &= \left(X^\top X\right)^{-1} X^\top y
\end{aligned}$$

(b)

$$\begin{aligned}
m(\hat{y}) &= E\|\hat{y} - U\gamma\|^2 \\
&= E\left\|UU^\top y - U\gamma\right\|^2 \\
&= E\left\|UU^\top (y - U\gamma)\right\|^2 \\
&= E\left\|UU^\top \varepsilon\right\|^2 \\
&= E\left(\varepsilon^\top UU^\top \varepsilon\right) \\
&= \operatorname{tr}\left(UU^\top\right)\sigma^2 \\
&= \operatorname{tr}(U^\top U)\sigma^2 \\
&= p\sigma^2
\end{aligned}$$

(c)

$$
\begin{aligned}
\tilde{y}^{(A_q)} &= \sum_{j\in A_q} u_j \left(u_j^\top y\right) \\
m\left(\tilde{y}^{(A_q)}\right) &= E\left\|\sum_{j\in A_q} u_j\left(u_j^\top y\right) - \sum_{i=1}^{p} u_j\gamma_i\right\|^2 \\
&= E\left\|\sum_{j\in A_q} u_j\hat{\gamma_j} - \sum_{i=1}^{p} u_i\gamma_i\right\|^2 \\
&= E\left\|\sum_{j\in A_q} u_j\left(\hat{\gamma_j}-\gamma_j\right) - \sum_{i\notin A_q} u_i\gamma_i\right\|^2 \\
&= E\left\|\sum_{j\in A_q} u_j\left(\hat{\gamma_j}-\gamma_j\right)\right\|^2 + E\left\|\sum_{i\notin A_q} u_i\gamma_i\right\|^2 \\
&= E\left(\sum_{j\in A_q} u_j\left(\hat{\gamma_j}-\gamma_j\right)\right)^\top\left(\sum_{j\in Aq} u_j\left(\hat{\gamma_j}-\gamma_j\right)\right) + \sum_{i\notin A_q}\gamma_i^2 \\
&= \sum_{j\in A_q} E\left\|\hat{\gamma_j}-\gamma_j\right\|^2 + \sum_{i\notin A_q}\gamma_i^2 \\
&= \sum_{j\in A_q} E\varepsilon^\top u_j u_j^\top\varepsilon + \sum_{i\notin A_q}\gamma_i^2 \\
&\leqslant q\sigma^2 + \|\gamma\|^2 \\
&\leqslant (q+1)\sigma^2 \leqslant p\sigma^2 = m(\hat{y})
\end{aligned}
$$

(d)

$$
\begin{aligned}
C_q &= \frac{RSS_q}{\hat{\sigma}^2} - n + 2q \\
\text{其中}\hat{\sigma}^2 &= \frac{RSS_p}{x-p} \\
RSS_q &= \left\|y-\tilde{y}^{(A_q)}\right\|^2 \\
&= \|y\|^2 - \sum_{j\in A_q}\left(u_j^\top y\right)^2 \\
RSS_p &= \|y-\hat{y}\|^2 \\
&= \|y\|^2 - \sum_{j=1}^{p}\left(u_j^\top y\right)^2
\end{aligned}
$$

$$
C_q = \frac{\|y\|^2 - \sum_{j\in A_q}\left(u_j^\top y\right)^2}{\|y\|^2 - \sum_{j=1}^{p}\left(u_j^\top y\right)^2}(n-p) - n + 2q
$$

1. 计算 $u_j^\top y, j=1,\ldots,p$, 并排序 $u_{i_1}^\top y \leqslant \cdots \leqslant u_{i_p}^\top y$

2. $q^* = \underset{q}{\operatorname{argmin}} = \frac{\|y\|^2 - \sum_{j\in A_q}\left(u_j^\top y\right)^2}{\hat{\sigma}^2}(n-p) - n + 2q$

3. $A_{q^*} = \{i_1,\ldots,i_{q^*}\}$