统计方法与机器学习

第四章：多重共线性 - 2

倪 葎

DaSE@ECNU
(lni@dase.ecnu.edu.cn)

目录

目录

岭回归

原因

- 线性回归模型的最小二乘估计为

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

  其方差为

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- 当设计矩阵 $\boldsymbol{X}$ 出现多重共线性时，回归系数的最小二乘估计的效果明显变差。

- 原因是

$$|\boldsymbol{X}'\boldsymbol{X}| \approx 0$$

  这导致了 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 计算不稳定。

岭回归

定义

- 为了求解逆矩阵更方便，我们采用

$$\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}, \quad k > 0$$

  代替 $\boldsymbol{X}'\boldsymbol{X}$。

- 优点：
  - k 很小时，$\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I} \approx \boldsymbol{X}'\boldsymbol{X}$
  - $\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}$ 可以避免是奇异矩阵。

岭回归

定义

- 称
$$\hat{\boldsymbol{\beta}}(k) = \left(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$
为回归系数 $\boldsymbol{\beta}$ 的岭回归估计，其中，称 k 为岭参数。

- 注意：
  - 由于 $\boldsymbol{X}$ 已经标准化，所以 $\boldsymbol{X}'\boldsymbol{X}$ 就是自变量样本相关系数矩阵。
  - 因为岭参数 k 不唯一确定，所以岭回归估计
  $$\hat{\boldsymbol{\beta}}(k) = (\hat{\beta}_1(k), \hat{\beta}_2(k), \cdots, \hat{\beta}_p(k))'$$
  是关于回归参数 $\boldsymbol{\beta}$ 的一个估计族。

岭回归

性质 1

- 为什么说岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 是有偏估计吗?

- 我们计算 $\hat{\boldsymbol{\beta}}(k)$ 的期望，即

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}(k)) &= E((\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta \neq \beta
\end{aligned}
$$

定理 3-1

$\hat{\boldsymbol{\beta}}(k)$ 是回归参数 β 的有偏估计。

岭回归

性质 1

- 为什么说岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 是有偏估计吗?
- 我们计算 $\hat{\boldsymbol{\beta}}(k)$ 的期望，即

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}(k)) &= E((\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}
\end{aligned}
$$

定理 3-1

$\hat{\beta}(k)$ 是回归参数 β 的有偏估计。

性质 1

- 为什么说岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 是有偏估计吗?
- 我们计算 $\hat{\boldsymbol{\beta}}(k)$ 的期望，即

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}(k)) &= E((\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}
\end{aligned}
$$

定理 3-1

$\hat{\boldsymbol{\beta}}(k)$ 是回归参数 $\boldsymbol{\beta}$ 的有偏估计。

性质 2

- 岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 与最小二乘估计 $\hat{\boldsymbol{\beta}}$ 有什么关系?
- 根据岭回归估计的定义可知，我们可以得到

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(k) &= \left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\right)\hat{\boldsymbol{\beta}}
\end{aligned}$$

- 如果岭参数 k 是与因变量 $\boldsymbol{y}$ 无关，
  - $\hat{\boldsymbol{\beta}}(k)$ 是 $\hat{\boldsymbol{\beta}}$ 的一种线性变换；
  - 岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 也是 $\boldsymbol{y}$ 的线性函数。

岭回归

性质 2

- 岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 与最小二乘估计 $\hat{\boldsymbol{\beta}}$ 有什么关系?
- 根据岭回归估计的定义可知，我们可以得到

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}(k) &= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}
\end{aligned}
$$

- 如果岭参数 k 是与因变量 $\boldsymbol{y}$ 无关,
  - $\hat{\boldsymbol{\beta}}(k)$ 是 $\hat{\boldsymbol{\beta}}$ 的一种线性变换;
  - 岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 也是 $\boldsymbol{y}$ 的线性函数。

岭回归

性质 3

- 问题：当出现多重共线性时，为什么我们要介绍岭回归估计?
- 由于岭回归估计是有偏的，一般我们以**均方误差**作为标准来比较岭回归估计和最小二乘估计.
- 注意到，最小二乘估计可以看作一种特殊的岭回归估计，即

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X} + 0 \cdot \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = \hat{\beta}(0)$$

性质 3

- 问题：当出现多重共线性时，为什么我们要介绍岭回归估计?

- 由于岭回归估计是有偏的，一般我们以**均方误差**作为标准来比较岭回归估计和最小二乘估计.

- 注意到，最小二乘估计可以看作一种特殊的岭回归估计，即

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + 0 \cdot \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}(0)$$

岭回归

定理 3-2

存在 $k>0$，使得岭估计的均方误差小于最小二乘估计的均方误差，即

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}(k)) \leq \mathrm{MSE}(\hat{\boldsymbol{\beta}}(0))$$

证明： 为了简化符号，我们令 $H(k)=\mathrm{MSE}(\hat{\boldsymbol{\beta}}(k))$，即

$$\begin{aligned}
H(k) &= \mathrm{MSE}(\hat{\boldsymbol{\beta}}(k)) = E\left(\hat{\boldsymbol{\beta}}(k)-\boldsymbol{\beta}\right)'\left(\hat{\boldsymbol{\beta}}(k)-\boldsymbol{\beta}\right) \\
&= E\left(\hat{\boldsymbol{\beta}}(k)-E(\hat{\boldsymbol{\beta}}(k))\right)'\left(\hat{\boldsymbol{\beta}}(k)-E(\hat{\boldsymbol{\beta}}(k))\right) \\
&\quad +(E(\hat{\boldsymbol{\beta}}(k))-\boldsymbol{\beta})'(E(\hat{\boldsymbol{\beta}}(k))-\boldsymbol{\beta}) \\
&=: I_1(k)+I_2(k)
\end{aligned}$$

证明

为了证明存在 $k>0$ 使得

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}(k)) \le \mathrm{MSE}(\hat{\boldsymbol{\beta}}(0)),$$

只需要证明 $H(k)$ 在 $k=0$ 处的导数 $\frac{\partial H(k)}{\partial k}|_{k=0} < 0$ 即可. 根据 $H(k) = I_1(k) + I_2(k)$ 可知,

$$\frac{\partial H(k)}{\partial k} = \frac{\partial I_1(k)}{\partial k} + \frac{\partial I_2(k)}{\partial k}.$$

于是，我们进一步分析这两个导数.

证明：定理 3-2（续）

我们先讨论一些岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 不同的表达形式

$$\begin{aligned}\hat{\boldsymbol{\beta}}(k) &= \left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \overset{\text{def}}{=} \boldsymbol{W}_k\boldsymbol{X}'\boldsymbol{y} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\right)\hat{\boldsymbol{\beta}} \overset{\text{def}}{=} \boldsymbol{W}_k^*\hat{\boldsymbol{\beta}}\end{aligned}$$

于是，$\boldsymbol{W}_k^*$ 与 $\boldsymbol{W}_k$ 之间存在关系，即

$$\begin{aligned}\boldsymbol{W}_k^* &= \boldsymbol{W}_k\left(\boldsymbol{X}'\boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}-k\boldsymbol{I}\right) \\ &= \boldsymbol{I}-k\left(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}\right)^{-1} \\ &= \boldsymbol{I}-k\boldsymbol{W}_k\end{aligned}$$

证明：定理 3-2（续）

假定 $\boldsymbol{X}'\boldsymbol{X}$ 的特征值为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0,$$

而相应正交化后的特征向量记为 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p$，则有

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \quad j = 1, 2, \cdots, p$$

在上式的等式两端同时加上 $k\boldsymbol{I} \cdot \boldsymbol{v}_j$,

$$(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})\,\boldsymbol{v}_j = (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{v}_j + k\boldsymbol{I} \cdot \boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j + k\boldsymbol{I} \cdot \boldsymbol{v}_j = (\lambda_j + k)\boldsymbol{v}_j$$

那么,

$$(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{v}_j = \frac{1}{\lambda_j + k}\boldsymbol{v}_j \Rightarrow \boldsymbol{W}_k \boldsymbol{v}_j = \frac{1}{\lambda_j + k}\boldsymbol{v}_j$$

证明：定理 3-2（续）

另一方面，根据

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \quad j = 1, 2, \cdots, p$$

可知，

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{v}_j = \frac{1}{\lambda_j}\boldsymbol{v}_j.$$

于是，

$$(\boldsymbol{I} + k(\boldsymbol{X}'\boldsymbol{X})^{-1})\boldsymbol{v}_j = \left(1 + \frac{k}{\lambda_j}\right)\boldsymbol{v}_j = \frac{\lambda_j + k}{\lambda_j}\boldsymbol{v}_j.$$

从而

$$\left(\boldsymbol{I} + k(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)^{-1}\boldsymbol{v}_j = \frac{\lambda_j}{\lambda_j + k}\boldsymbol{v}_j \Rightarrow \boldsymbol{W}_k^*\boldsymbol{v}_j = \frac{\lambda_j}{\lambda_j + k}\boldsymbol{v}_j$$

证明：定理 3-2（续）

这里我们总结一下：

- $\boldsymbol{W}_k$ 的特征值分别为 $\frac{1}{\lambda_1+k}, \cdots, \frac{1}{\lambda_p+k}$；
- $\boldsymbol{W}_k^*$ 的特征值分别为 $\frac{\lambda_1}{\lambda_1+k}, \cdots, \frac{\lambda_p}{\lambda_p+k}$；
- $\boldsymbol{W}_k$ 和 $\boldsymbol{W}_k^*$ 的特征向量与 $\boldsymbol{X}'\boldsymbol{X}$ 的特征向量相同，与岭参数 k 无关.

证明：定理 3-2（续）

首先，我们考虑 $I_1(k)$.

$$
\begin{aligned}
I_1(k) &= E\left(\hat{\boldsymbol{\beta}}(k) - E(\hat{\boldsymbol{\beta}}(k))\right)'\left(\hat{\boldsymbol{\beta}}(k) - E(\hat{\boldsymbol{\beta}}(k))\right) \\
&= E(\boldsymbol{W}_k^*\hat{\boldsymbol{\beta}} - \boldsymbol{W}_k^*\boldsymbol{\beta})'(\boldsymbol{W}_k^*\hat{\boldsymbol{\beta}} - \boldsymbol{W}_k^*\boldsymbol{\beta}) \\
&= E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{W}_k^*)'(\boldsymbol{W}_k^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\
&= E(\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{W}_k^*)'(\boldsymbol{W}_k^*)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})
\end{aligned}
$$

最后一个等号成立，是因为

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}
\end{aligned}
$$

岭回归

统计知识（补充）

- 假定 $\boldsymbol{A}$ 是对称矩阵. $\boldsymbol{x}$ 是一个 p 维随机变量，并假定 $\boldsymbol{\mu} = E(\boldsymbol{x})$ 和 $\Sigma = \text{Var}(\boldsymbol{x})$. 那么,

$$E(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = \text{tr}(\boldsymbol{A}\Sigma) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}.$$

证明：定理 3-2（续）

$$
\begin{aligned}
I_1(k) &= E(\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{W}_k^*)'(\boldsymbol{W}_k^*)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}) \\
&= \sigma^2 \mathrm{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{W}_k^*)'(\boldsymbol{W}_k^*)) \\
&= \sigma^2 \mathrm{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{W}_k(\boldsymbol{I} - k\boldsymbol{W}_k)) \\
&= \sigma^2 \left(\mathrm{tr}(\boldsymbol{W}_k) - k\mathrm{tr}(\boldsymbol{W}_k^2)\right) \\
&= \sigma^2 \left(\sum_{j=1}^{p} \frac{1}{\lambda_j + k} - k \sum_{j=1}^{p} \frac{1}{(\lambda_j + k)^2}\right) \\
&= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} \\
\Rightarrow \frac{\partial I_1(k)}{\partial k} &= -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3} < 0 \Rightarrow k \text{ 越大 } I_1(k) \text{ 越小}
\end{aligned}
$$

证明：定理 3-2（续）

接下来，我们考虑 $I_2(k)$.

$$
\begin{aligned}
I_2(k) &= \left(E(\hat{\boldsymbol{\beta}}(k)) - \boldsymbol{\beta}\right)' \left(E(\hat{\boldsymbol{\beta}}(k)) - \boldsymbol{\beta}\right) \\
&= (\boldsymbol{W}_k^* \boldsymbol{\beta} - \boldsymbol{\beta})' (\boldsymbol{W}_k^* \boldsymbol{\beta} - \boldsymbol{\beta}) \\
&= \boldsymbol{\beta}' (\boldsymbol{W}_k^* - \boldsymbol{I})' (\boldsymbol{W}_k^* - \boldsymbol{I}) \boldsymbol{\beta} \\
&= k^2 \boldsymbol{\beta}' \boldsymbol{W}_k^2 \boldsymbol{\beta} \\
&= k^2 \boldsymbol{\beta}' \boldsymbol{V}' \boldsymbol{L} \boldsymbol{V} \boldsymbol{\beta} \qquad \text{令}(\boldsymbol{W}_k^2 = \boldsymbol{V}' \boldsymbol{L} \boldsymbol{V}) \\
&=: k^2 \boldsymbol{\alpha}' \boldsymbol{L} \boldsymbol{\alpha} \qquad \text{令}(\boldsymbol{\alpha} = \boldsymbol{V} \boldsymbol{\beta}) \\
&= k^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j + k)^2}
\end{aligned}
$$

其中，$\boldsymbol{\alpha} = \boldsymbol{V}\boldsymbol{\beta} = (\alpha_1, \alpha_2, \cdots, \alpha_p)'$ 与岭参数 k 无关.

证明：定理 3-2（续）

由于

$$I_2(k) = k^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j + k)^2}$$

因此，

$$\begin{aligned}\frac{\partial I_2(k)}{\partial k} &= 2\sum_{j=1}^{p} \frac{k\alpha_j^2}{(\lambda_j + k)^2} - 2\sum_{j=1}^{p} \frac{k^2\alpha_j^2}{(\lambda_j + k)^3} \\ &= 2k\sum_{j=1}^{p} \frac{\lambda_j\alpha_j^2}{(\lambda_j + k)^3} \geq 0\end{aligned}$$

即当 k 越大时，$I_2(k)$ 越大.

证明：定理 3-2（续）

$$
\begin{aligned}
\frac{\partial H(k)}{\partial k} &= \frac{\partial I_1(k)}{\partial k} + \frac{\partial I_2(k)}{\partial k} \\
&= -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3} + 2k \sum_{j=1}^{p} \frac{\lambda_j \alpha_j^2}{(\lambda_j + k)^3}
\end{aligned}
$$

考虑 $k = 0$ 时，

$$
\left.\frac{\partial H(k)}{\partial k}\right|_{k=0} = \left.\frac{\partial I_1(k)}{\partial k}\right|_{k=0} + \left.\frac{\partial I_2(k)}{\partial k}\right|_{k=0} = -2\sigma^2 \sum_{j=1}^{p} \lambda_j^{-2} < 0
$$

由连续性可知，在以 0 为中心的一个领域内，存在 $k > 0$，使得 $H(k) < H(0)$. 此定理证毕.

说明

- 注意到

$$
\begin{aligned}
\frac{\partial H(k)}{\partial k} &= \frac{\partial I_1(k)}{\partial k} + \frac{\partial I_2(k)}{\partial k} \\
&= -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3} + 2k \sum_{j=1}^{p} \frac{\lambda_j \alpha_j^2}{(\lambda_j + k)^3} \\
&= \sum_{j=1}^{m} \frac{2\lambda_j}{(\lambda_j + k)^3}(k\alpha_j^2 - \sigma^2)
\end{aligned}
$$

- 根据上式，易知使得 $\frac{\partial H(k)}{\partial k} = 0$ 的 k 与 $\sigma^2, \boldsymbol{\beta}$ 有关；
- 但是，σ^2 与 β 均为未知参数，因此无法找到一个对一切 σ^2 及 $\boldsymbol{\beta}$ 都成立的 k 使得 $\boldsymbol{H}(k)$ 达到最小.

另一个角度看岭回归估计

- 由于最小二乘估计 $\hat{\boldsymbol{\beta}}$ 是最小化离差平方和的解，即

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

- 可以证明，岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 是最小化带有 L_2 正则项的离差平方和的解，即

$$\hat{\boldsymbol{\beta}}(k) = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$$

- 等价于最小化

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

另一个角度看岭回归估计

- 考虑带约束的最优化问题：

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

这个约束会对解 $\hat{\boldsymbol{\beta}}(k)$ 带来什么影响?

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \leq s$，那么 $\hat{\boldsymbol{\beta}}$ 就是我们想要的解，也就是说 $\hat{\boldsymbol{\beta}}(k)=\hat{\boldsymbol{\beta}}$；
- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} > s$，那么，我们所得到的解 $\hat{\boldsymbol{\beta}}(k)$ 应该满足

$$\hat{\boldsymbol{\beta}}(k)'\hat{\boldsymbol{\beta}}(k) \leq s < \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$

另一个角度看岭回归估计

- 考虑带约束的最优化问题：

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

  这个约束会对解 $\hat{\boldsymbol{\beta}}(k)$ 带来什么影响?

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \leq s$，那么 $\hat{\boldsymbol{\beta}}$ 就是我们想要的解，也就是说 $\hat{\boldsymbol{\beta}}(k)=\hat{\boldsymbol{\beta}}$；
- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} > s$，那么，我们所得到的解 $\hat{\boldsymbol{\beta}}(k)$ 应该满足

$$\hat{\boldsymbol{\beta}}(k)'\hat{\boldsymbol{\beta}}(k) \leq s < \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$

另一个角度看岭回归估计

- 考虑带约束的最优化问题：

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

  这个约束会对解 $\hat{\boldsymbol{\beta}}(k)$ 带来什么影响?

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \leq s$，那么 $\hat{\boldsymbol{\beta}}$ 就是我们想要的解，也就是说 $\hat{\boldsymbol{\beta}}(k) = \hat{\boldsymbol{\beta}}$；

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} > s$，那么，我们所得到的解 $\hat{\boldsymbol{\beta}}(k)$ 应该满足

$$\hat{\boldsymbol{\beta}}(k)'\hat{\boldsymbol{\beta}}(k) \leq s < \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$

讨论

- 岭回归对应的最优化问题为

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

  其中，$\boldsymbol{\beta}'\boldsymbol{\beta}$ 是 $\boldsymbol{\beta}$ 的 L_2 范数。

- 那么，考虑以下最优化问题：

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq s$$

  其中，$\|\boldsymbol{\beta}\|_1$ 是 $\boldsymbol{\beta}$ 的 L_1 范数，即每一个元素的绝对值之和。这个优化问题的解为 LASSO（Least absolute shrinkage and selection operator）。

- 问题：这两个解有什么差别？

讨论

- 岭回归对应的最优化问题为

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \le s$$

  其中，$\boldsymbol{\beta}'\boldsymbol{\beta}$ 是 $\boldsymbol{\beta}$ 的 L_2 范数。

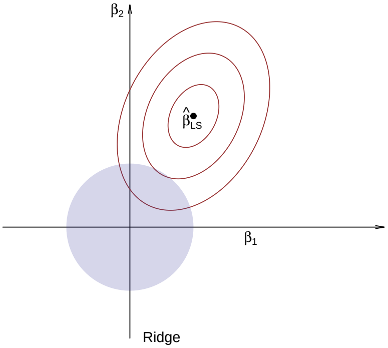
- 那么，考虑以下最优化问题：

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \le s$$

  其中，$\|\boldsymbol{\beta}\|_1$ 是 $\boldsymbol{\beta}$ 的 L_1 范数，即每一个元素的绝对值之和。这个优化问题的解为 LASSO（Least absolute shrinkage and selection operator）。

- 问题：这两个解有什么差别?

讨论：Ridge VS LASSO



岭回归

定理 3-3

对任意 $k>0$，$\|\hat{\boldsymbol{\beta}}\| \neq 0$，总有

$$\|\hat{\boldsymbol{\beta}}(k)\| < \|\hat{\boldsymbol{\beta}}\|$$

证明: 假定 $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ 是 $\boldsymbol{X}'\boldsymbol{X}$ 的特征值,而 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p$ 为其相应的特征向量. 于是，我们有

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V}'\boldsymbol{\Lambda}\boldsymbol{V}$$

其中，$\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_p\}$，$\boldsymbol{V}'$ 是以 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p$ 为列向量的矩阵.

证明：定理 3-3（续）

回归模型可写为

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{X}\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &=: \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \end{aligned}$$

注意到，$\boldsymbol{\alpha} = \boldsymbol{V}\boldsymbol{\beta} \Rightarrow \boldsymbol{\beta} = \boldsymbol{V}'\boldsymbol{\alpha}$.

由此，α 的最小二乘估计为

$$\begin{aligned} \hat{\alpha} &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = (\boldsymbol{V}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}')^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= (\boldsymbol{V}\boldsymbol{V}'\boldsymbol{\Lambda}\boldsymbol{V}\boldsymbol{V}')^{-1}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{\Lambda}^{-1}\boldsymbol{Z}'\boldsymbol{y} \end{aligned}$$

证明：定理 3-3（续）

而 β 的最小二乘估计 $\hat{\beta}$ 与 $\hat{\alpha}$ 存在如下关系

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{V}'\boldsymbol{\Lambda}^{-1}\boldsymbol{V}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{V}'\boldsymbol{\alpha}$$

类似地，关于 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 的岭估计分别为

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(k) &= (\boldsymbol{\Lambda} + k\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ \hat{\boldsymbol{\beta}}(k) &= \boldsymbol{V}'\hat{\boldsymbol{\alpha}}(k) \end{aligned}$$

所以，

$$\|\hat{\boldsymbol{\beta}}(k)\| = \|\hat{\boldsymbol{\alpha}}(k)\| = \|(\boldsymbol{\Lambda} + k\boldsymbol{I})^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}\| < \|\hat{\boldsymbol{\alpha}}\| = \|\hat{\boldsymbol{\beta}}\|$$

由此，定理得证.

说明

- $\hat{\boldsymbol{\beta}}(k)$ 是对 $\hat{\boldsymbol{\beta}}$ 向原点的压缩.
- 这是因为

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\beta}}) &= E\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)\\ &= E(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}})-\boldsymbol{\beta}'\boldsymbol{\beta} = E\|\hat{\boldsymbol{\beta}}\|^2-\|\boldsymbol{\beta}\|^2\end{aligned}$$

- 因此,

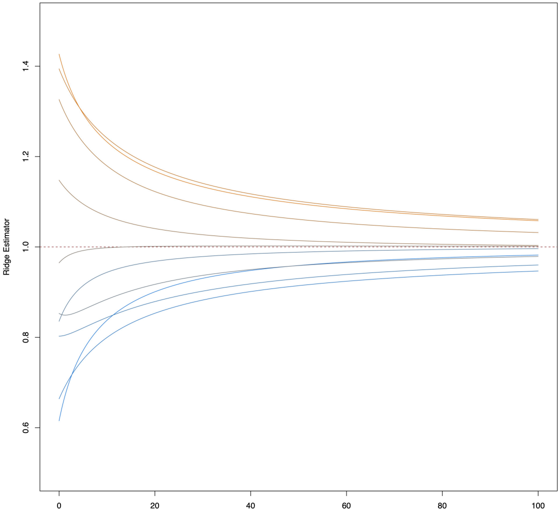
$$E\|\hat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}\|^2+\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{\beta}\|^2+\sigma^2\sum_{j=1}^{p}\lambda_j^{-1}$$

- 当设计矩阵 $\boldsymbol{X}$ 出现多重共线性时，上式中的第二项比较大，因此，对其做压缩是应该的.

岭参数的参数选择（岭迹法）

- 岭估计 $\hat{\boldsymbol{\beta}}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 的分量 $\hat{\beta}_j(k)$ 作为岭参数 k 的函数.
- 当 k 在 $[0, +\infty)$ 变化时，在平面直角坐标系中，我们称 $k - \hat{\beta}_j(k)$ 的图像为**岭迹**.

岭参数的参数选择（岭迹法）



岭回归

岭参数的参数选择（岭迹法）

- 岭迹法的一般原则
  - 各回归系数的岭估计基本稳定；
  - 用最小二乘估计时，符号不合理的回归系数的岭估计的符号变得合理；
  - 回归系数没有不合理的符号；
  - 残差平方和增大不多.
- 优点：容易计算；
- 缺点：具有主观性；

岭参数的参数选择（方差扩大因子法）

- 根据方差扩大因子判定多重共线性，即 $c_{jj} > c_{\text{VIF}}$.
- 岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 的方差为

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}(k)) &= \sigma^2(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1} \\ &\stackrel{\text{def}}{=} \sigma^2\boldsymbol{C}(k)\end{aligned}$$

- 我们可以类似地定义矩阵 $\boldsymbol{C}(k)$ 中对角线的元素 $c_{jj}(k)$ 为岭估计的方差扩大因子.
- $c_{jj}(k)$ 随着 k 的增大而减少.
- 通过选择 k 使得所有方差扩大因子 $c_{jj}(k) \leq c_{\text{VIF}}$，从而确定岭参数 k.

岭参数的参数选择（Hoerl-Kennad 公式）

- 回顾

$$\begin{aligned}\frac{\partial H(k)}{\partial k} &= \frac{\partial I_1(k)}{\partial k} + \frac{\partial I_2(k)}{\partial k} \\ &= \sum_{j=1}^{m} \frac{2\lambda_j}{(\lambda_j + k)^3}(k\alpha_j^2 - \sigma^2)\end{aligned}$$

- 在 1970 年，霍尔和肯纳德提出了

$$k_{\mathrm{HK}} = \frac{\hat{\sigma}^2}{\max_j \hat{\alpha}_j^2}$$

- 易证 $\frac{\partial H(k)}{\partial k}\big|_{k=k_{\mathrm{HK}}} < 0.$

岭参数的参数选择（Mcdorard-Garaneau 公式）

- 回顾

$$E\|\hat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}\|^2 + \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

- 令

$$Q = \|\hat{\boldsymbol{\beta}}\|^2 - \hat{\sigma}^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

- 如果 $Q > 0$，那么认为 $\hat{\boldsymbol{\beta}}$ 中某一分量过大，需要对其进行压缩. 压缩量由 $\sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$ 决定.
- 如果 $Q \leq 0$，那么认为 $\hat{\boldsymbol{\beta}}$ 的各个分量都差不多，此时，对 $\hat{\boldsymbol{\beta}}$ 不进行压缩，选择 $k = 0$.

岭参数的参数选择（Mcdorard-Garaneau 公式）

- Mcdorard 和 Garaneau 建议选择岭参数 k，使得

$$\|\hat{\boldsymbol{\beta}}\|^2 - \|\hat{\boldsymbol{\beta}}(k)\|^2 \approx \hat{\sigma}^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

即选择 k 使得

$$\|\hat{\boldsymbol{\beta}}(k)\|^2 \approx \|\hat{\boldsymbol{\beta}}\|^2 - \hat{\sigma}^2 \sum_{j=1}^{p} \lambda_j^{-1}$$