岭回归

马学俊 (主讲) 晁越 (助教)

苏州大学

数学科学学院

https://xuejunma.github.io/

岭回归估计的定义

- 普通最小二乘估计带来的问题
  - 当自变量间存在复共线性时，回归系数估计的方差就很大，估计值就很不稳定。在具体取值上与真实值有较大的偏差，有时甚至会出现与实际意义不服的正负号。
- 岭回归的定义
  - 岭回归 (Ridge Regression, 简记为 RR) 提出的想法是很自然的。当自变量间存在复共线性时，$\boldsymbol{X}'\boldsymbol{X} \approx 0$，我们设想给 $\boldsymbol{X}'\boldsymbol{X}$ 加上一个正常数矩阵 $k\boldsymbol{I}, (k > 0)$，那么 $\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}$ 接近奇异的程度就会比 $\boldsymbol{X}'\boldsymbol{X}$ 接近奇异的程度小得多。
  - 考虑到变量的量纲问题，我们先对数据做标准化，为了记号方便，标准化后的设计阵仍然用 $\boldsymbol{X}$ 表示，定义为：

$$\hat{\beta}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{7.2}$$

    为 β 的岭回归估计，其中 k 称为岭参数。
  - 由于假设 $\boldsymbol{X}$ 已经标准化，所以 $\boldsymbol{X}'\boldsymbol{X}$ 就是自变量样本相关阵，(7.2) 式计算的实际是标准化岭回归估计。
  - (7.2) 式中因变量观测向量 $\boldsymbol{y}$ 可以经过标准化也可以未经标准化。显然，$\hat{\beta}(k)$ 岭回归做为 β 的估计应比最小二乘估计稳定

岭回归估计的性质

性质 $\hat{\beta}(k)$ 是回归参数 β 的有偏估计。

证明：

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

显然只有当 $k=0$ 时，$E(\hat{\beta}(0))=\beta$；当 $k\neq 0$，$\hat{\beta}(k)$ 是回归参数 β 的有偏估计。要特别强调的是 $\hat{\beta}(k)$ 不再是 β 的无偏估计，有偏性是岭回归估计的一个重要特性。

岭回归估计的性质

- 性质 在认为岭参数 k 是 $\boldsymbol{y}$ 的无关的常数时，$\hat{\beta}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是最小二乘估计 $\hat{\beta}$ 的一个线性变换，也是 $\boldsymbol{y}$ 的线性函数。

证明：

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

所以，岭估计 $\hat{\beta}(k)$ 是最小二乘估计 $\hat{\beta}$ 的一个线性变换，根据定义式 $\hat{\beta}(k)=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 知 $\hat{\beta}(k)$ 也是 $\boldsymbol{y}$ 的线性函数。

- 这里需要注意的是，在实际应用中，由于岭参数 k 总是要通过数据来确定，因而 k 也依赖于 $\boldsymbol{y}$，
- 本质上来说，$\hat{\beta}(k)$ 并非 $\hat{\beta}$ 的线性变换，也不是 $\boldsymbol{y}$ 的线性函数。

岭回归估计的性质

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

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

这里 $||\cdot||$ 是向量的模，等于向量各分量的平方和。这个性质表明 $\hat{\beta}(k)$ 可看成由 $\hat{\beta}$ 进行某种向原点的压缩。从 $\hat{\beta}(k)$ 的表达式可以看到，当 $k\to\infty$ 时，$\hat{\beta}(k)\to\mathbf{0}$，即 $\hat{\beta}(k)$ 化为零向量。

- 以 MSE 表示估计向量的均方误差，则存在 $k>0$，使得

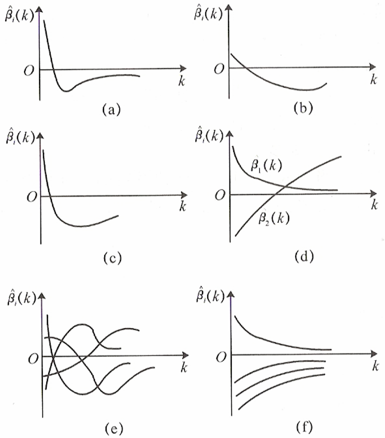
$$MSE[\hat{\beta}(k)] < MSE[\hat{\beta}]$$

即

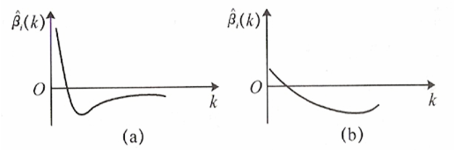
$$\sum_{j=1}^{p} E[\hat{\beta}_j(k)-\beta_j]^2 < \sum_{j=1}^{p} D(\hat{\beta}_j)$$

岭迹分析

可以根据岭迹线的形状变化来确定适当的 k 值和进行自变量的选择。

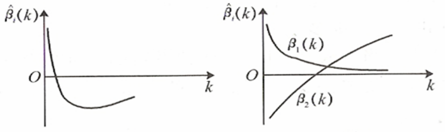


岭迹分析



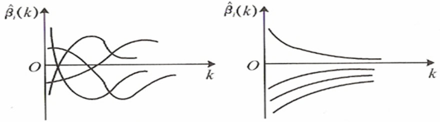
1. 在图 (a) 中，$\hat{\beta}_j(0) = \hat{\beta}_j > 0$，且比较大。从古典分析的观点看，应将 x_j 看作对 y 有重要影响的因素。但 $\hat{\beta}_j(k)$ 的图形显示出相当的不稳定性，当 k 从零开始增加时，$\hat{\beta}_j(k)$ 显著地下降，而且迅速趋于 0，因而失去预测能力。从岭回归的观点看，x_j 对 y 不起重要作用，甚至可以剔除这个变量。
2. 在图 (b) 中的情况与图 (a) 中相反，$\hat{\beta}_j = \hat{\beta}_j(0) > 0$，但很接近 0. 从古典回归分析的观点看，$x_j$ 对 y 的作用不大。但随着 k 略增加，$\hat{\beta}_j(k)$ 骤然变为负值，从岭回归的观点看，x_j 对 y 有显著影响。

岭迹分析



3. 在图 (c) 中，$\hat{\beta}_j = \hat{\beta}_j(0) > 0$，说明 x_j 比较显著，但当 k 增加时，$\hat{\beta}_j(k)$ 迅速下降，且稳定为负值。从古典回归分析的观点看，x_j 对 y 有正影响的显著因素。从岭回归的观点看，x_j 对 y 有负影响的因素。
4. 在图 (d) 中，$\hat{\beta}_1(k)$ 和 $\hat{\beta}_2(k) > 0$ 都很不稳定，但其和却大体上稳定。这种情况往往发生在自变量 x_1 和 x_2 的相关性很强的场合，即在 x_1 和 x_2 之间存在多重共线性。因此，从自变量选择的观点看，两者只要保留一个就够了。这可以用来解释某些回归系数估计的符号不合理的情形，从实际观点看，β_1 和 β_2 不应有相反的符号。岭回归分析的结果对这一点提供了一种解释。

岭迹分析



(e) (f)

5. 从全局看，岭迹分析可以用来估计在某一具体实例中最小二乘估计是否适用。把所有回归系数的岭迹都描绘在一张图上，如果这些岭迹线的不稳定性很强，整个系统呈现比较“乱”的局面，往往就使人怀疑最小二乘估计是否很好的反映了真实情况，如图 (e) 所示。如果情况如图 (f) 那样，则我们对最小二乘估计可以有很大的信心。当情况介于 (e) 和 (f) 之间时，我们必须适当选择 k 值。

岭参数 k 的选择

- 岭迹法

岭迹法选择 k 值的一般原则是：

1. 各回归系数的岭估计基本稳定；
2. 用最小二乘估计时符号不合理的回归系数，其岭估计的符号变得合理；
3. 回归系数没有不合乎经济意义的绝对值；
4. 残差平方和增大不太多。

岭迹法

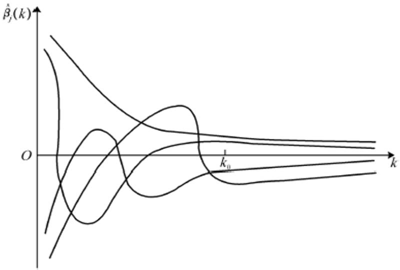


图 7－3

从图中可以看出，当 k 取 k_0 时，各回归系数的估计值基本上都能相对稳定。

岭参数 k 的选择

- 方差扩大因子法

方差扩大因子 c_{jj} 可以度量多重共线性的严重程度。计算岭估计 $\hat{\beta}(k)$ 的协方差矩阵

$$
\begin{aligned}
D(\hat{\beta}(k)) &= cov(\hat{\beta}(k), \hat{\beta}(k)) \\
&= cov((\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}, (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}) \\
&= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}' cov(\mathbf{y}, \mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \\
&= \sigma^2\mathbf{c}(k)
\end{aligned}
$$

矩阵 $\mathbf{c}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}$，其对角元素 $c_{jj}(k)$ 为岭估计的方差扩大因子。不难看出，$c_{jj}(k)$ 随着 k 的增大而减小。

选择 k 的经验做法：选择 k 使所有的方差扩大因子 $c_{jj}(k) \leq 10$。

岭参数 k 的选择

- 由残差平方和来确定 k 值

岭估计在减小均方误差的同时增大了残差平方和，我们希望岭回归的残差平方和 $SSE(k)$ 的增加幅度控制在一定的限度以内，可以给定一个大于 1 的 c 值，要求：

$$SSE(k) < cSSE$$

寻找使上式成立的最大的 k 值。

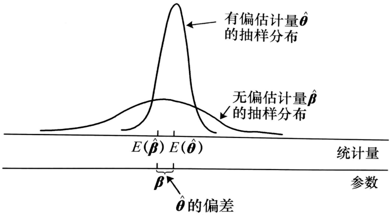
用岭回归选择变量

岭回归选择变量的原则:

1. 在岭回归中设计矩阵 $\boldsymbol{X}$ 已经中心化和标准化了，这样可以直接比较标准化岭回归系数的大小。可以剔除掉标准化岭回归系数比较稳定且绝对值很小的自变量。
2. 随着 k 的增加，回归系数不稳定，震动趋于零的自变量也可以剔除。
3. 剔除标准化岭回归系数很不稳定的自变量. 如果依照上述去掉变量的原则，有若干个回归系数不稳定，究竟去掉几个，去掉哪几个，这并无一般原则可循，这需根据去掉某个变量后重新进行岭回归分析的效果来确定。

本章小节与评注

岭回归与普通最小二乘的本质区别是：普通最小二乘估计是线性无偏估计中最好的，岭回归估计则是有偏估计中一个较好的估计。



如果一个估计量只有很小的偏差，但它的精度大大高于无偏估计量，人们更愿意选择这个估计量，因为它接近真实参数值的可能性更大。由上图可以看到，估计量 $\hat{\beta}$ 是无偏的，但不精确，而估计量 $\hat{\theta}$ 精确度高却有小的偏差。$\hat{\theta}$ 落在真值 β 附近的概率远远大于无偏估计量 $\hat{\beta}$

作业

- p.181 7.6
- 扩展阅读：CH10,11