《线性回归》
—稳健回归

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.05.28

主要内容：稳健回归

1. 稳健回归
   - 最小二乘回归
   - LS回归的问题
   - 稳健回归
   - L_1 回归
   - Huber 回归
   - L_1/Huber估计
   - Mallows/Schweppe回归
   - 崩溃点(breakdown point)
   - LMS回归
   - MM-估计
   - 结束语

稳健回归

稳健回归

这一讲的材料来源于：

- Seber and Lee (2003). Section 3.13 [p. 77-96]
- Draper and Smith (1998). Chapter 25 [567-584]
- J. J. Faraway (2002). Practical Regression and Anova using R.

最小二乘回归

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 \\ &= \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \arg\min_{\theta} \sum_{i=1}^{n} \widehat{\epsilon}_i^2\end{aligned}$$

为什么要使用最小二乘回归？既有历史的原因，又有其本身的优良性质

- 历史的原因（自1800年以来一直在使用）
- 最小二乘估计量$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$具有显式解，并且易于计算。
- 如果$y = \mathbf{X}\theta + \epsilon$, $\epsilon \sim N_n(\mathbf{0}, \sigma^2 I)$：
  - ✓ 最小二乘估计=MLE
  - ✓ 在所有的无偏估计中最小二乘估计的方差最小（Gauss-Markov）

LS回归的问题

- 当统计误差不服从正态分布时，与线性模型有关的置信区间和检验的水平差不多是正确的，但检验的功效可能很低（功效= $P(\text{reject } H_0|H_a\text{为真})$.
- LSE它对异常值很敏感, 因为平方之后的大残差占有很大的权重。

稳健回归

- Robust回归可以（部分）解决这些问题。我们将研究以下方法：
  - ✓ L_1回归（=最小绝对偏差（LAD）回归。）
  - ✓ Huber回归
  - ✓ Mallows回归
  - ✓ Sweweppe回归
  - ✓ 最小平方中位数（LMS）回归

L_1回归

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} |y_i - x_i^T \theta|$$

- L_1 回归比LS更古老：最早可以追溯到Boscovich (1760)，Laplace (1789)
- L_1之所以没有变得很流行，因为没有显式解（但是，对现代计算机不再是问题了;可以用内点法有效地解决计算问题）
- 在最简单的位置模型$y_i = \alpha + \epsilon_i$中，$L_1$回归的解释数据的中位数.
- 对y方向的异常值更加稳健，但对x方向的异常值仍非常敏感.
- 当误差是正态分布时，L_1回归的效率较低;对于相同的精度，需要大约多50%的观测值才能达到【为什么？试用统计推断的知识解释之】

Huber 回归

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \rho_c(y_i - x_i^T \theta),$$

其中

$$\rho_c(u) = \begin{cases} u^2/2, & \text{如果}|u| \le c, \\ c(|u| - c/2), & \text{如果}|u| > c. \end{cases}$$

- 它是L_1和L_2回归之间的一种折中方案：
  - ✓ $c = \infty \longrightarrow L_2$回归（最小二乘）.
  - ✓ $c = 0 \Longrightarrow L_1$回归(使用$\rho_c(u) = |u|$).
- 想法：对小残差用二次方式惩罚，对大残差用线性方式惩罚
- 计算：解方程$\sum_{i=1}^{n} \psi_c(y_i - x_i^T \theta) x_i = 0$ 其中$\psi_c(u) = \rho_c'(u) = \mathrm{sign}(u) \min(|u|, c)$.
- 对于残差而言，应该选择合适的变点c.可用迭代加权最小二乘法计算.

L_1/Huber估计

- 估计量的精确分布不能准确的确定$\Longrightarrow$ 使用渐近结果或bootstrap做推断
- y方向上的异常值影响有限，但x方向上的异常值则不然. 解决方案：Mallows / Schweppe 估计

Mallows/Schweppe 回归

$$\sum_{i=1}^{n} \eta\left(x_i, \frac{y_i - x_i^T\theta}{\widehat{\sigma}}\right) x_i = 0$$

- Mallows:

$$\eta(x, r) = \min\left(1, \frac{a}{\|\mathbf{A}x\|}\right)\psi_c(r)$$

- Schweppe:

$$\eta(x, r) = \frac{1}{\|\mathbf{A}x\|}\psi_c(\|\mathbf{A}x\| r),$$

  $\|\mathbf{A}x\|^2 = c \cdot x^T(\mathbf{X}^T\mathbf{X})^{-1}x$, 然后稳健化，其中$c$是一个正的常数

- $\psi_c = \rho'(x)$由Huber回归得到

崩溃点(breakdown point)

估计的崩溃点(breakdown point)是指：为使估计值产生任意大的结果，估计量可以处理的不正确观察（即任意大的观察值）的比例.【需要查阅更精确的定义】

- 平均的崩溃点：0
- 中位数的崩溃点：1/2
- 最小二乘回归的崩溃点：0
- L_1和Huber回归的的崩溃点：0（在x方向上）
- Mallows/Schweppe的崩溃点：$\leq 1/p$

LMS回归

$$\hat{\theta} = \arg\min_{\theta} \mathrm{median}((y_i - x_i^T\theta)^2).$$

- 【示意图】.
- Hampel (1975)，Rousseeuw (1984)
- Breakdown point约为0.5
- 由于存在许多局部最小值而难以计算
- 当统计误差服从正常分布时，估计的效率较低(收敛速度为$n^{-1/3}$). 这时可以用α截断平均来代替中位数.即, 将剔除残差最大的αn个之后做平均.【最小截断二乘估计】

MM-估计

- 首先找到σ的高度稳健的M估计（第一个M）
- 然后保持$\hat{\sigma}$固定, 然后找到θ的M估计. 例如使用Newton法（第二个M）

结束语(一些想法，见Faraway Ch 13.)

- Robust估计可以对付长尾错误分布，但不能对付模型选择和方差结构的问题. 后面这些问题可能比非正态误差更为严重.
- θ的推断更加困难.可以使用bootstrap或者类似bootstrap的方法(随机扰动法).
- 除了最小二乘法之外，还可以使用稳健方法.如果这两个估计值相差很大，我们有理由表示担心.

作业

针对简单线性模型实施稳健的L_1估计，试设计随机模拟方案给出简单线性模型参数的bootstrap置信区间（置信水平取为0.95）.
[选作：] 你认为这样做的理论基础是什么？