回归模型的补救措施

1 误差项方差不相等的补救措施–加权最小二乘法

我们在第3章和第6章解释了Y的变换如何有助于减少或消除误差项的不等方差. Y的转换的一个困难是, 它们可能会创建不适当的回归关系. 当找到了适当的回归关系, 但误差项的方差不相等时, 转换的替代方法是加权最小二乘, 这是一种基于多元回归模型推广的过程. 我们现在用σ_i^2表示误差项ϵ_i的方差, 以认识到不同的误差项可能有不同的方差. 多元回归模型可表示为:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{p-1} + \epsilon_i$$

其中

- $\beta_0, \cdots, \beta_{p-1}$为参数
- $X_{i1}, X_{i1}, \cdots, X_{i,p-1}$为已知的常数
- ϵ_i彼此独立并分别服从$N(0, \sigma_i^2)$(i.n.d.)
- 有$i = 1, \cdots, n$个观测值

因此可以写误差向量ϵ的方差-协方差矩阵:

$$\underset{n\times n}{\sigma^2(\epsilon)} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} = \sigma^2 \Lambda$$

根据方差是否已知将考虑的情况分为一下两种:

1.1 误差方差已知

当误差方差σ_i^2已知时, 可以用极大似然法得到回归模型中回归系数的估计量. 对其中误差方差相等情况下的似然函数进行修正, 将σ^2项替换为各自的方差σ_i^2

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2}\left(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1}\right)^2\right] \text{ 令: } w_i = \frac{1}{\sigma_i^2}$$

$$\Rightarrow \quad L(\boldsymbol{\beta}) = \left[\prod_{i=1}^{n} \sqrt{\frac{w_i}{2\pi}}\right] \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1}\right)^2\right]$$

由于σ已知, 为取得最大化的$L(\boldsymbol{\beta})$, 我们需要最小化$Q_w := \sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}\right)^2$. 注意到那些有较小方差$\sigma_i^2$的那些观测值在加权最小二乘判别式中具有较大的权重$w_i$.

写成一个矩阵的形式, 令

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} = W^\top$$

因此方差协方差矩阵可以表示为$\sigma^2(Y) = \sigma^2(\epsilon) = W^{-1}$. 此时正规方程可以写为

$$(X^\top W X) b_w = X^\top W Y$$

注记. *在等方差,* $W = \sigma^{-2} I$*的情况下, 消掉*σ^{-2}*, 就是一般最小二乘的正规方程.*

注意这里面W包含了原本的方差σ, 因此对应参数点估计量, 期望, 点估计方差的估计分别可以写出:

- $b_w = (X^\top W X)^{-1} X^{-\top} W Y := AY$, 其中$A := (X^\top W X)^{-1} X^\top W$
- $E(b_w) = AE(Y) = AX\beta (X^\top W X)^{-1} X^\top W X \beta = \beta$
- $\sigma^2\{b_w\} = A\sigma^2\{Y\}A^\top = \underbrace{(X^\top W X)^{-1} X^\top W}_{A} W^{-1} \underbrace{W X (X^\top W X)^{-1}}_{A^\top} = (X^\top W X)^{-1}$

也可以写成代数形式, 记$x_1^\top, \cdots, x_n^\top$分别为设计矩阵$X$的$n$个行向量, 此时$\beta$的广义最小二乘估计(generalized least squares estimates, GLSE, 有时也称为Gauss-Markov估计)为

$$\beta^* = \left[\sum_{i=1}^{n} \frac{x_i x_i^\top}{\sigma_i^2}\right]^{-1} \left[\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} x_i\right]$$

从向量代数形式的表达式可以看出两个和式分别是$x_i x_i^\top, y_i x_i^\top$的加权和, 而所用的权重都是$\frac{1}{\sigma_i^2}$, 因此$\beta^*$往往被称为加权最小二乘估计.

1.2 误差方差未知

1.2.1 当发现方差与自变量之间的回归关系时

比如使用了BP检验, 有 $\ln \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \epsilon_i^*$ 则使用plug-in方法, 带入$\hat{\sigma}_i^2$.
由于对于误差项$\epsilon_i \sim N(0, \sigma_i^2)$, 有$E|\epsilon_i| = \sqrt{\frac{2}{\pi}}\sigma_i$. 因此可以使用$|\epsilon_i|$估计标准差$\hat{s}_i$.
我们给出一个计算方法加以说明,

1. 在原始数据集上直接应用一般最小二乘法
2. 将绝对残差或平方残差关于X进行回归
3. 使用X拟合的值作为标准差的估计插入W矩阵中.

4. 应用加权最小二乘法得到新的回归直线

5. 重复步骤2, 之道某两次得到的直线较为接近

1.2.2 当同方差部分有重复数据(近邻数据)时

我们给出了一个例子说明情况,

例 (加权最小二乘法的插入估计)**.** 若观测向量y是有n个观测值组成的样本, 这n个观测值分成k组, 每一组的分类个体都有某种分类性质, 第$i(i = 1, \cdots, k)$组的第$j, j = 1, \cdots, n_i$个个体的观测值记为y_{ij}. 得到模型为

$$y_{ij} = x_{ij}^\top \beta + \epsilon_{ij}, i = 1, \cdots, k; j = 1, \cdots, n_i$$

我们假设

$$var(\epsilon_{ij}) = \sigma_i^2, i = 1, \cdots, k; j = 1, \cdots, n_i$$

$$cov(\epsilon_{ij}, \epsilon_{i'j'}) = 0, i \neq i', j \neq j'$$

若记$\epsilon = (\epsilon_{11}, \cdots, \epsilon_{1n_1}, \cdots, \epsilon_{k1}, \cdots, \epsilon_{kn_k})^\top$, 则$\epsilon$的协方差矩阵有形式

$$\underset{n\times n}{\sigma^2(\epsilon)} = \begin{pmatrix} \sigma_1^2 I_{n_1} & & \cdots & 0 \\ & \sigma_2^2 I_{n_2} & & \\ \vdots & & & \vdots \\ 0 & & \cdots & \sigma_k^2 I_{n_k} \end{pmatrix}$$

对每个不同的分块,

$$y_i = (y_{i1}, y_{i2}, \cdots, y_{i,n_i})^\top$$
$$X_i = (x_{i1}, x_{i2}, \cdots, x_{i,n_i})^\top$$
$$\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{i,n_i})^\top$$

此时模型具有如下的形式

$$y_i = X_i\beta + \epsilon_i, \quad E(\epsilon_i) = 0, cov(\epsilon_i) = \sigma_u^2 I_{n_i}, i = 1, \cdots, k$$

由于广义最小二乘估计的表达式为

$$\beta^* = \left[\sum_{i=1}^n \frac{x_i x_i^\top}{\sigma_i^2}\right]^{-1} \left[\sum_{i=1}^n \frac{y_i}{\sigma_i^2} x_i\right]$$

其中β的参数个数为p, 当在实际中σ_i^2未知的时候, 可以对于固定的i, 利用

$$\hat{\sigma}_i^2 = \frac{\left\|y_i - X_i\hat{\beta}_i\right\|^2}{n_i - p}$$

其中$\hat{\beta}_i = (X_i^\top X_i)^{-1} X_i^\top y_i$, 再将这样的估计式带入$\beta^*$的式子中便得到对$\beta$的一个估计,

$$\widetilde{\beta} = \left[\sum_{i=1}^n \frac{x_i x_i^\top}{\hat{\sigma}_i^2}\right]^{-1} \left[\sum_{i=1}^n \frac{y_i}{\hat{\sigma}_i^2} x_i\right]$$

上述过程可以进一步被写成迭代估计的形式, 即首先获得β的估计式$\widetilde{\beta}^{(1)}$, 再代回$\hat{\sigma}_i^2$的估计式中计算出对σ_i^2新的估计, 从而计算出新的估计$\widetilde{\beta}^{(2)}$, 重复上述过程直到相邻两次迭代求得的$\widetilde{\beta}$相差不多时停止. 一般情况下, 这个迭代过程是收敛的.

注记. 当误差方差$\sigma_i^2 = kx_i^2$时, 其中$k > 0$是某未知常数, 可以考虑使用$w_i = \dfrac{1}{x_{i1}^2}$

在现实生活中遇到的情况往往是方差未知的情况, 由于

$$\sigma_i^2 = E(\epsilon_i^2) - (E\epsilon_i)^2 = E(\epsilon_i^2)$$

因此残差平方e_i^2是σ_i^2的估计量, 进一步的, 残差的绝对值$|e_i|$是标准差σ_i的估计量.

2 多重共线性的补救措施–岭回归方法

2.1 共线性的理论机理

设待估参数θ为$p\times 1$的未知参数向量, $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2, \cdots, \widetilde{\theta}_p)^\top$是$\theta$的一个估计. 定义$\widetilde{\theta}$的均方误差为

$$MSE(\widetilde{\theta}) = E\|\widetilde{\theta} - \theta\|^2 = E(\widetilde{\theta} - \theta)^\top(\widetilde{\theta} - \theta)$$

MSE本质上度量了一个参数估计$\widetilde{\theta}$与真参数θ之间距离平方的数学期望, 因此一个好的估计应当有较小的均方误差.

关于一般意义下MSE的偏倚-方差分解可以写成:

$$\boxed{MSE(\widetilde{\theta}) = tr\left(cov(\widetilde{\theta})\right) + \|E\widetilde{\theta} - \theta\|^2}$$

请在整理笔记的时候补完这个证明, 使用加一项减一项, 证明的技巧在于$tr(AB) = tr(BA)$, 特别的当AB是一个数时. 此外$tr(\cdot)$与E可以交换.

进一步, 其可以写成

$$\boxed{MSE(\widetilde{\theta}) = \underbrace{\sum_{i=1}^{p} var(\widetilde{\theta}_i)}_{\text{方差}} + \underbrace{\sum_{i=1}^{p}\left(E\widetilde{\theta}_i - \theta_i\right)^2}_{\text{偏倚}}}$$

一个较好的估计应当具有较小的偏倚和方差.

考虑中心化线性回归模型

$$\boldsymbol{y} = \alpha\mathbf{1} + \boldsymbol{X}\beta_r + e, E(\boldsymbol{e}) = \mathbf{0}, \operatorname{cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}.$$

这里 $\operatorname{rk}(\boldsymbol{X}_c) = p$, 此时

$$\hat{\alpha} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\hat{\boldsymbol{\beta}}_c = \left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)^{-1}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{y}.$$

因为 $\Delta_2 = 0$, 故

$$\mathrm{MSE}\left(\hat{\boldsymbol{\beta}}_c\right) = \Delta_1 = \sigma^2 \operatorname{tr}\left(\boldsymbol{X}_c^{\top}\boldsymbol{X}_c\right)^{-1}.$$

因为 $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$ 是对称正定阵, 于是存在 $p \times p$ 正交阵 $\boldsymbol{\Phi}$ 将 $\boldsymbol{X}_c^{\top}\boldsymbol{X}_c$ 对角化, 即

$$\boldsymbol{X}_c^{\top}\boldsymbol{X}_c = \boldsymbol{\Phi}\begin{bmatrix} \lambda_1 & 0 & \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}\boldsymbol{\Phi}^{\top},$$

这里 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$, 为 $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$ 的特征值. 记 $\boldsymbol{\Phi} = (\varphi_1, \varphi_2, \cdots, \varphi_p)$, 则 $\varphi_1, \varphi_2, \cdots, \varphi_p$ 分别为对应于 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 的标准正交化特征向量. 从式 (5-4) 得

$$\left(\boldsymbol{X}_c^{\top}\boldsymbol{X}_c\right)^{-1} = \boldsymbol{\Phi}^{\top}\begin{bmatrix} \frac{1}{\lambda_1} & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \frac{1}{\lambda_p} \end{bmatrix}\boldsymbol{\Phi},$$

由此可知 $\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_p^{-1}$ 为 $\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)^{-1}$ 的特征值. 再利用 $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$ 以及 $\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi} = \boldsymbol{I}$, 得到

$$\operatorname{tr}\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)^{-1} = \sum_{i=1}^{p}\frac{1}{\lambda_i},$$

我们证明了

$$\mathrm{MSE}\left(\hat{\boldsymbol{\beta}}_c\right) = \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i},$$

由上式可以看出, 如果 $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$ 至少有一个特征根非常小, 即非常接近于零, 那么 $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$ 就会很大.从均方误差的标准来看, 这时的最小二乘估计 $\boldsymbol{\beta}$ 就不是一个好的估计. 另一方面

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

于是

$$\boldsymbol{E}\left\|\hat{\boldsymbol{\beta}}_c\right\|^2 = \|\boldsymbol{\beta}_c\|^2 + \mathrm{MSE}\left(\hat{\boldsymbol{\beta}}_c\right) = \left\|\hat{\boldsymbol{\beta}}_c\right\|^2 + \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i},$$

这就是说, 当 $X_c^{\mathrm{T}}X_c$ 至少有一个特征值很小时, 最小二乘估计 $\hat{\boldsymbol{\beta}}_c$ 不再是一个好的估计. 若记 $\boldsymbol{X}_c = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$, 其中 $\boldsymbol{x}_i$ 为设计阵 $\boldsymbol{X}_c$ 的第 i 列. 设 λ 为 $\boldsymbol{X}_c^{\top}\boldsymbol{X}_c$ 中的一个特征值, $\boldsymbol{\varphi} = (c_1, c_2, \cdots, c_p)^{\mathrm{T}}$ 为其特征向量, 其长度为 1 , 即 $\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\varphi} = 1$. 若 $\lambda \approx 0$,则

$$\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\boldsymbol{\varphi} = \lambda\boldsymbol{\varphi} \approx \boldsymbol{0},$$

用 $\boldsymbol{\varphi}^{\mathrm{T}}$ 左乘上式,得

$$\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{X}_c^{\top}\boldsymbol{X}_c\boldsymbol{\varphi} = \lambda\boldsymbol{\varphi}^{\top}\boldsymbol{\varphi} = \lambda \approx 0,$$

因为

$$\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\boldsymbol{\varphi} = \|\boldsymbol{X}_c\boldsymbol{\varphi}\|^2 = \lambda \approx 0,$$

于是有

$$X_c \boldsymbol{\varphi} \approx \mathbf{0},$$

上式即为

$$c_1 x_1 + c_2 x_2 + \cdots + c_p x_p \approx 0$$

即向量$x_1, x_2, \cdots, x_p$ 之间存在着近似线性关系. 这种关系就是前面所说的自变量之间的共线性关系或多重共线性关系 (multicollinearity).

这里需要强调 3 点:

- 共线性问题不是一个模型设定错误, 而是所获得的数据本身自带的缺陷, 不能通过考察残差的办法去探测共线性问题;
- 由于没有模型的误设问题, 我们不能在模型误设的基础上继续进行分析, 只有在模型的设定比较正确时, 才可以进一步讨论模型的共线性所带来的问题;
- 相关系数存在两两变量之间的比较关系, 而多重共线性能体现出多个预测变量之间的关系, 用相关系数往往并不能发现这种线性关系. 弄清楚数据的共线性这种缺陷以及这种缺陷对数据分析造成的严重后果是十分重要的. 当数据出现共线性问题时, 我们必须十分谨慎地对待回归分析中得到的所有结论.

2.2 岭估计

解决共线性情况下的参数估计的一个重要方法是岭估计(ridge estimation)方法. 岭估计方法的主要设计思路是适当允许参数估计是有偏的估计, 通过将偏差(的平方)$\|E\widetilde{\theta} - \theta\|^2$控制在可以接受的范围内, 然后尽量降低估计的方差项$tr\left(cov(\widetilde{\theta})\right)$. 具体的, 回归系数$\beta$的岭估计为

$$\hat{\boldsymbol{\beta}}(k) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + k\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y},$$

这里 $k > 0$ 是可选参数, 称为岭参数或偏参数. 当 k 取不同的值时. 我们得到不同岭估计.

我们给出如下的定理,

定理 (Ridge)**.** *存在$k > 0$, 使得在均方误差意义下, 岭估计优于最小二乘估计.*

岭估计等价于最小化

$$\begin{aligned}
Q &= \sum_{i=1}^{n}\left[Y_i^* - \left(\beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^*\right)\right]^2 + c\left[\sum_{j=1}^{p-1}{\beta_j^*}^2\right] \\
&= (Y^* - X^*\beta^*)^\top (Y^* - X^*\beta^*) + c{\beta^*}^\top \beta^* \\
&= {Y^*}^\top Y^* - 2{\beta^*}^\top {X^*}^\top X^* + {\beta^*}^\top {X^*}^\top X^* \beta^* + c{\beta^*}^\top \beta^*
\end{aligned}$$

因此

$$\begin{aligned}
\frac{\partial Q}{\partial \beta^*} &= -2{X^*}^\top Y^* + 2{X^*}^\top X^* \beta^* \\
&=
\end{aligned}$$

2.3 岭迹图

记$\hat{\beta}_i(k)$为$\hat{\boldsymbol{\beta}}(k)$ 的第 i 个分量, 它是 k 的一元函数. 当 k 在 $[0,+\infty)$ 上变化时, $\boldsymbol{\beta}(k)$ 的图形称为岭迹(ridge trace). 选择 k 的岭迹法是将 $\hat{\beta}_1(k), \hat{\beta}_2(k), \cdots, \hat{\beta}_p(k)$ 的岭迹画在同一个图上, 根据岭迹的变化趋势选择 k 值, 使得各个回归系数的岭估计大体上稳定, 并且各个回归系数岭估计值的符号比较合理. 我们知道, 最小二乘估计是使残差平方和达到最小的估计. k 愈大, 岭估计与最小二乘估计偏离愈大. 因此, 它对应的残差平方和也随着 k 的增加而增加. 当我们用岭迹法选择 k 值时, 还应考虑使得残差平方和不要上升太多. 在实际处理上, 上述几点原则有时可能会有些互.相不一致, 顾此失彼的情况也经常出现, 这就要根据不同情况灵活处理.

```r
dat = read.table(file = './Data_5e/Data_5e/CH07TA01.txt'); dat = scale(dat)
X1 = dat[,1]; X2 = dat[,2]; X3 = dat[,3]; Y = dat[,4]
library(MASS)
X = cbind(dat[,1:3])
lamlist = seq(from=0, to=1, by=0.01)
fit = lm.ridge(Y~X1+X2+X3-1, lambda=lamlist)
plot(fit)
```

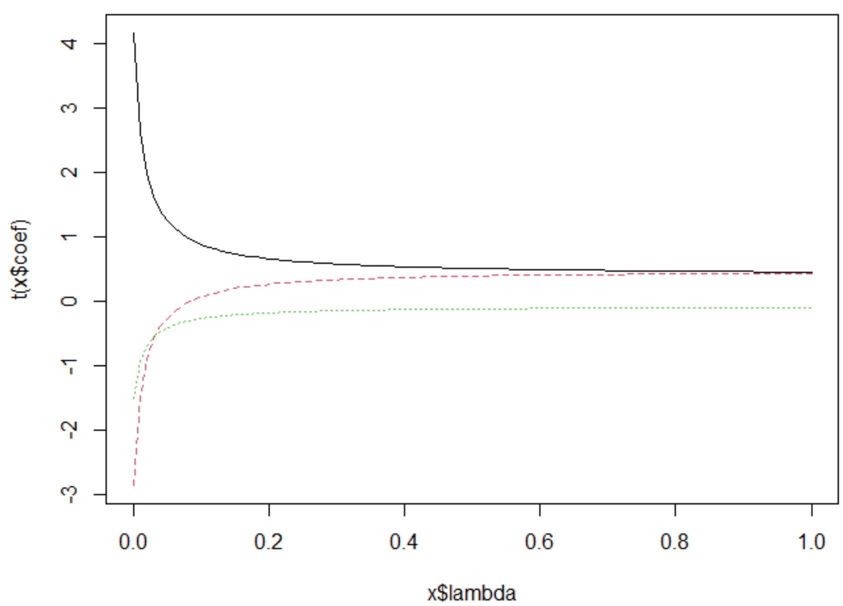

图 1: 用PPT中的代码实现岭迹图

3 强影响点的补救措施–稳健回归

最小一乘

$$\min_{\alpha} \sum_{i=1}^{N} \left| x_i^T \alpha - y_i \right|$$

最小中位数

$$\min_{\alpha} med\left(x_i^T\alpha - y_i\right)^2$$

迭代再加权最小二乘(IRLS)用于解决特定的最优化问题，这个最优化问题的目标函数如下所示：

$$\arg\min_{\beta}\sum_{i=1}^{n}|y_i - f_i(\beta)|^p$$

这个目标函数可以通过迭代的方法求解。在每次迭代中，解决一个带权最小二乘问题，形式如下：

$$\beta^{t+1} = \arg\min_{\beta}\sum_{i=1}^{n} w_i(\beta^{(t)}))|y_i - f_i(\beta)|^2 = (X^T W^{(t)} X)^{-1} X^T W^{(t)} y$$

在这个公式中，$W^{(t)}$是权重对角矩阵，它的所有元素都初始化为1。每次迭代中，通过下面的公式更新。

$$W_i^{(t)} = |y_i - X_i\beta^{(t)}|^{p-2}$$

3.1 Huber加权函数

$$w = \begin{cases} 1 & |u| \le 1.345 \\ \dfrac{1.345}{|u|} & |u| > 1.345 \end{cases}$$

4 非参数回归–Lowess方法和回归树

4.1 Lowess(局部加权回归, q近邻)

思路: 局部的响应曲面可以近似为平面.

$$E(Y|X_1, X_2) = m(X_1, X_2)$$

- 距离度量$d_i = \left[(X_{i1} - X_{h1})^2 + (X_{i2} - X_{h2})^2\right]^{\frac{1}{2}}$
- 选取超参数q为比例, 原始数据中有比例q的数据与数据点(X_{h1}, X_{h2}), 更大的q将导致更为光滑的拟合, 建议q取值在0.4到0.6之间.
- 加权函数可以取

$$w_i = \begin{cases} \left[1 - \left(\dfrac{d_i}{d_q}\right)^3\right]^3, & d_i < d_q \\ 0, & d_i \ge d_q \end{cases}$$

4.2 回归树

最后将数据集划分的最小分块称为: 终端结点(terminal node)或树叶(leaf). 沿树将预测变量空间分开的点称为内部结点(internal node). 树内部各个结点的连接部分称为分支(branch). 回归树都优势是解释起来简便, 能很好的用图形表示.

通过特征空间分层预测

建立回归树都过程大致分为两步,

1. 将预测变量空间(即$X_1, \cdots, X_p$的可能取值的集合)分割成J个互不重叠的区域$R_1, \cdots, R_J$.
2. 对落入区域R_j的每个观测值都作同样的预测, 预测值等于R_j上训练集的响应值都算数平均.

树的剪枝

5 非标准情况下的精度估计–Bootstrap方法(自助法)

比如想要估计回归系数b_1的方差 ,

1. 设置Bootstrap样本$B = 500$
2. 对每个Bootstrap样本$k = 1, \cdots, B$, 有放回的(replacement)从原始数据中抽取n个观测值, 每个抽样提供了一个新的估计$b_1(k)$.
3. 用新参数估计$\{b_1(1), \cdots, b_1(B)\}$中的标准差估计参数的标准差.