第五章：模型及诊断

5.1 含常数项的线性模型

前面研究的都是一般形式下的线性模型。在实际中，往往线性模型含有常数项(截距项)。设此时线性模型为

$$y=\beta_0+\beta_1x_1+\cdots+\beta_{p-1}x_{p-1}+e,$$

这里β_0为常数项，$x_1,x_2\cdots x_{p-1}$为协变量，e为随机误差。若记$x_0=1$，$x=(x_0,x_1,\cdots x_{p-1})'$，$\beta=(\beta_0,\beta_1,\cdots\beta_{p-1})'$，则即为通常线性模型情形$y=x'\beta+e$。

设有n次观测，写成矩阵形式，令

$$Y=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix},\quad X=\begin{pmatrix}1 & x_{11} & \cdots & x_{1,p-1}\\ 1 & x_{21} & \cdots & x_{2,p-1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{n,p-1}\end{pmatrix},$$

$$\beta=\begin{pmatrix}\beta_0\\ \beta_I\end{pmatrix}=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_{p-1}\end{pmatrix},\quad e=\begin{pmatrix}e_1\\ \vdots\\ e_n\end{pmatrix},\ \text{则}$$

$$Y=X\beta+e$$

假设$Ee=0$，$Cov(e)=\sigma^2I_n$，$rank(X_{n\times p})=p$。最小二乘估计$\hat\beta=(X'X)^{-1}X'Y$，令$\hat\sigma^2=\dfrac{\|Y-X\hat\beta\|^2}{n-p}$。

在上述假定下，将第三章中的结论搬过来有

1.$E\hat\beta=\beta$，$Var(\hat\beta)=\sigma^2(X'X)^{-1}$，$E\hat\sigma^2=\sigma^2$；

2. (Gauss-Markov 定理)对$\forall c'\beta$，$c'\hat\beta$是其唯一的 BLUE；

若进一步假定误差为正态分布，则

3.对$\forall c'\beta$，$c'\hat\beta$是其唯一的 MVUE；

4.$\hat\beta\sim N_p\left(\beta,\sigma^2(X'X)^{-1}\right)$，$\dfrac{(n-p)\hat\sigma^2}{\sigma^2}\sim\chi^2_{n-p}$，$\hat\beta$与$\hat\sigma^2$独立。

对此模型，主要感兴趣的是回归系数β_I的估计，常数项β_0单独考虑。令$E_{n\times1}=(1,\cdots,1)'$，$X_{n\times p}=\left(E_n\vdots\tilde X_{n\times(p-1)}\right)$，则模型为$Y=\beta_0E_n+\tilde X\beta_I+e$。

在实际应用中，对数据中心化是常用的手段。所谓中心化就是把自变量的度量起点移至到n次试验中所取值的中心点处。记

$\bar{x}_j=\dfrac{\sum_{i=1}^{n}x_{ij}}{n},1\le j\le p-1$，$\bar{x}=(\bar{x}_1,\cdots,\bar{x}_{p-1})'$，

$\bar{y}=\dfrac{\sum_{i=1}^{n}y_i}{n}$，则中心化后模型分量形式为：

$$y_i=\alpha+\beta_1(x_{i1}-\bar{x}_1)+\cdots+\beta_{p-1}(x_{i,p-1}-\bar{x}_{p-1})+e_i$$

其中$\alpha=\beta_0+\bar{x}'\beta_I$，写成矩阵形式为

$Y=\alpha E_n+\tilde{X}_c\beta_I+e$，$Ee=0$，$Cov(e)=\sigma^2I_n$，

其中$\tilde{X}_c=\left(I_n-\dfrac{E_nE_n'}{n}\right)\tilde{X}$。$\tilde{X}_c$称为中心化了的设计矩阵，易见$\tilde{X}_c'E_n=0$。此时线性回归模型称为中心化的线性回归模型。正规方程：

$$\begin{pmatrix}n & 0\\ 0 & \tilde{X}_c'\tilde{X}_c\end{pmatrix}\begin{pmatrix}\alpha\\ \beta_I\end{pmatrix}=\begin{pmatrix}n\bar{y}\\ \tilde{X}_c'Y\end{pmatrix}$$

解得

$$\hat{\alpha}=\bar{y},\quad \hat{\beta}_I=(\tilde{X}_c'\tilde{X}_c)^{-1}\tilde{X}_c'Y,$$

$$Cov\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}_I\end{pmatrix}=\sigma^2\begin{pmatrix}1/n & 0\\ 0 & (\tilde{X}_c'\tilde{X}_c)^{-1}\end{pmatrix}。$$

中心化的线性模型，常数项由样本均值估计，回归系数β_I的估计等价于线性回归模型$Y=\tilde{X}_c\beta_I+e$的参数估计。若误差正态分布，则中心化后的模型估计$\hat{\alpha}$与$\hat{\beta}_I$独立。

定理 5.1.1：中心化后给出的回归系数估计与没有中心化时给出的估计是一致的。

除了中心化，对协变量经常作另一种处理。

令 $s_j^2=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$，$1\le j\le p-1$，

$Z=(z_{ij})_{n\times(p-1)}$，其中$z_{ij}=\dfrac{x_{ij}-\bar{x}_j}{s_j}$，则$\sum_{i=1}^{n}z_{ij}^2=1$。

z_{ij}是将x_{ij}中心化后再标准化，易见$E_n'Z=0$。

令$R=(r_{ij})_{(p-1)\times(p-1)}=Z'Z$，则

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{s_i s_j}$$

若把协变量看成随机的，则 r_{ij}正好是协变量x_i与x_j的样本相关系数。中心化后标准化的好处在于：

1.R 可以分析协变量之间的相关关系；

2.消去了单位和取值范围的差异(R无量纲)。

用Z作为设计矩阵，此时分量形式为：$1\le i\le n$，

$y_i=\alpha^{(0)}+\frac{x_{i1}-\bar{x}_1}{s_1}\beta_1^{(0)}+\cdots\frac{x_{i,p-1}-\bar{x}_{p-1}}{s_{p-1}}\beta_{p-1}^{(0)}+e_i$。

这里$\alpha^{(0)}=\alpha$，$\beta_i^{(0)}=s_i\beta_i$，$1\le i\le p-1$。记$\beta_I^{(0)}=\left(\beta_1^{(0)},\cdots,\beta_{p-1}^{(0)},\right)'$，写成矩阵形式：

$$Y=\alpha^{(0)}E_n+Z\beta_I^{(0)}+e,$$

最小二乘估计

$$\hat{\alpha}^{(0)}=\bar{y},\quad \hat{\beta}_i^{(0)}=s_i\hat{\beta}_i,\quad 1\le i\le p-1。$$

5.2 哑(或虚拟)变量(dummy variable)处理

在实际应用中，常常会遇到一些协变量为属性变量，设其属性有k个状态，固然可以用数字$1,2\cdots k$来标识，但不可用来计算，因为它们无数量意义。解决办法是引进哑变量(dummy variable)，$x_{(1)},x_{(2)},\cdots x_{(q)}$，其中$q=k-1$，

$$x_{(i)}=\begin{cases}1,\text{若处在}i\text{状态}\\0,\qquad\text{其它}\end{cases},\quad i=1,2,\cdots q.$$

故$x_{(1)}=x_{(2)}=\cdots x_{(q)}=0$表示处在$k$状态。

若对响应变量y只考察此一因素，模型为

$Ey=\beta_0+\beta_1x_{(1)}+\cdots+\beta_qx_{(q)}$，则易见

$E(y|\text{状态}j)=\beta_0+\beta_j,\ j=1,2,\cdots q.$

$E(y|\text{状态}k)=\beta_0$，故此取哑变量的方法是以状态k为标准，β_j衡量状态j超出状态k的值。

另一种取法是$x_{(j)}=\begin{cases}1,\text{若处在状态}j\\-1,\text{若处在状态}k\\0,\qquad\text{其它}\end{cases}$

此时$x_{(1)}=x_{(2)}=\cdots x_{(q)}=-1$表示状态$k$。因而

$$E(y|\text{状态}j)=\beta_0+\beta_j,\ j=1,2,\cdots q.$$

$$E(y|\text{状态}k)=\beta_0-(\beta_1+\beta_2\cdots+\beta_q)$$

于是$\dfrac{\sum_{j=1}^{k}E(y|\text{状态}j)}{k}=\beta_0$，故$\beta_0$为平均效应，而$\beta_1,\cdots,\beta_q$衡量状态超出平均的效应。

5.3 显著性检验

对回归系数作出估计后就可以得到经验回归方程。所建立的经验回归方程是否真正地刻画了响应变量与协变量之间的实际依赖关系呢？

对线性回归模型：$1\le i\le n$，

$y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_{p-1}x_{i,p-1}+e_i$，$e_i\sim N(0,\sigma^2)$。

首先考虑响应变量 y是否线性地依赖协变量$x_1,\cdots,x_{p-1}$这个整体，即检验假设

$$H_0:\ \beta_1=\cdots=\beta_{p-1}=0。$$

上检验称为回归方程的显著性检验。若假设H_0被接受，意味着相对误差 e而言，所有协变量对响应变量 Y的影响是不重要的。

将模型中心化，写成矩阵形式：

$$Y=\alpha E_n+\tilde{X}_c\beta_I+e=(E_n\vdots\tilde{X}_c)\begin{pmatrix}\alpha\\ \beta_I\end{pmatrix}+e,\ e\sim N(0,\sigma^2I_n)。$$

要检验的假设为

$$H_0:\ H\begin{pmatrix}\alpha\\ \beta_I\end{pmatrix}=0，\text{其中}H_{(p-1)\times p}=(0\vdots I_{p-1})。$$

该假设可以由F检验来给出拒绝域。具体地

$$F=\frac{\hat{\beta}_I'\tilde{X}_c'Y/(p-1)}{(Y'Y-n\bar{y}^2-\hat{\beta}_I'\tilde{X}_c'Y)/(n-p)},$$

在假设H_0下，$F\sim F_{p-1,n-p}$，故给定水平$\alpha\in(0,1)$，当$F>F_{p-1,n-p}(\alpha)$时拒绝假设H_0，否则接受H_0。

当回归方程的显著性检验结果是拒绝原假设时，仅说明至少有一个$\beta_j\neq0$，并不排除响应变量 y不依赖其中某些协变量。

于是在整体的回归方程显著性检验被拒绝后还需对每个自变量逐一地作显著性检验，即对固定的某个 i，作如下假设检验H_i：$\beta_i = 0$。

对线性模型$Y = X\beta + e$，$e \sim N(0, \sigma^2 I_n)$，估计$\hat{\beta} \sim N_p(\beta, \sigma^2 (X'X)^{-1})$，$\hat{\sigma}^2 = \dfrac{\|Y - X\hat{\beta}\|^2}{n-p}$，令$C = (c_{ij})_{p\times p} = (X'X)^{-1}$，则$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$。

在假设H_i下，$t_i = \dfrac{\hat{\beta}_i}{\sqrt{c_{ii}}\hat{\sigma}} \sim t_{n-p}$，故给定水平$\alpha$，当$|t_i| > t_{n-p}(\alpha/2)$时拒绝$H_i$，否则接受$H_i$。

若经过检验，接受原假设H_i：$\beta_i = 0$，认为协变量 x_i对响应变量 y无显著影响，因而可以将其从回归方程中剔除，此时 y对剩余协变量重新作回归，回归系数的估计也随之变化，然后再检验剩余回归系数是否为零，再剔除经检验对y无显著影响的协变量，这样的过程一直下去。

5.4 回归协变量的选择

通常在作回归分析(以后若不特别指明，假设模型都含有常数项，即为 5.1 节中的形式)时，根据问题本身的专业理论及有关经验，常常把各种与响应变量有关或可能有关的协变量引入到回归模型。其结果是把一些对响应变量影响很小，甚至无影响的协变量都选入回归模型中，不但计算量大，而且估计和预测的精度也会下降。

此外，在一些情况下，某些协变量观测数据的获得代价昂贵，若这些协变量对响应变量影响很小或根本没有影响，若不加选择的引进回归模型，势必造成观测数据收集和模型应用的费用不必要的增大。

因此，对模型协变量的精心选择是十分有必要的。

设响应变量 y以及一系列的协变量$x_1, \cdots, x_s$以及这些量的 n次观测值，要识别哪些协变量 x_j对响应变量 y是重要的。

定义 5.4.1：令$R^2=\frac{RSS}{TSS}=\frac{\hat{\beta}_I'\tilde{X}_c'Y}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$，称 R^2 为决定系数(coefficient of determination)。

注 1：在线性回归分析中，β_I是关注的焦点。一般线性模型总和是$Y'Y$，在回归分析中常数项的回归平方和为$n\bar{y}^2$，因此这里总和实际上是去掉常数项的回归平方和，即$TSS=Y'Y-n\bar{y}^2$。

注 2：由等式 $TSS=RSS+ESS$，$(ESS=\left\|Y-X\hat{\beta}\right\|^2)$因此$0\le R^2\le 1$。$R^2$反映了回归和在总和所占的比例，$R^2$越大，表示回归协变量解释的越好。

注 3：将协变量看成随机的，则 y与$(x_1,\cdots,x_{p-1})'$ 的复相关系数 (multiple correlation coefficient)定义为

$$\rho=\sqrt{\frac{Cov(y,x)Var(x)^{-1}Cov(x,y)}{Var(y)}}$$

$\frac{1}{n}\tilde{X}_c'(Y-\bar{y}E_n)$，$\frac{1}{n}\tilde{X}_c'\tilde{X}_c$，$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$ 分别作为$Cov(x,y),Var(x),Var(y)$相应的样本估计，这样得到复相关系数 ρ的估计

$$\hat{\rho}=\sqrt{\frac{\hat{\beta}_I'\tilde{X}_c'Y}{\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

恰好为R，所以 R称为(样本)复相关系数。

注 4：F统计量与R^2的关系：$F=\frac{R^2}{1-R^2}\cdot\frac{n-p}{p-1}$。

若回归协变量个数固定时，则应选择 R^2大的那个回归。但当协变量个数不一样时，用R^2来选择协变量就失效了，因为全部变量都作为协变量，R^2的值将达到最大。

Adjusted R^2 criterion (调整 R^2准则)：回归协变量集的选择应使得$Adj\ R^2$ 达到最大，其中$Adj\ R^2=1-\frac{n-1}{n-p}(1-R^2)$，$p$为协变量的个数(含常数项)。

同样的道理，若回归协变量个数固定时，则应选择误差平方和ESS小的那个回归。但当协变量个数不一样时，用ESS来选择协变量就失效了，因为全部变量都作为协变量，此时ESS的值将达到最小，故必须对协变量的个数加一个“惩罚因子”。令

$$RMS_p = \frac{ESS}{n-p} = \frac{\|Y - X\beta\|^2}{n-p} = \hat{\sigma}^2$$

RMS_p criterion(平均残差平方和准则，residual mean squares criterion)：回归协变量集的选择应使 RMS_p达到最小。

令$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$为 y的样本方差，则

$Adj\ R^2 = 1 - \frac{RMS_p}{s_y^2}$，故$Adj\ R^2$准则与$RMS_p$准则本质上是一样的。

上面的准则是从回归拟合角度来看，Mallows'C_p准则则是从预测角度出发。设使用所有协变量的完全模型(含常数项)为：

$$Y = X_{n\times s}\beta + e,$$

给定一个含有 p个参数的子集模型(含常数项)，得到经验方程$\hat{Y} = X_p\hat{\beta}_p$，用 $\hat{Y}$去预测EY，其均方误差为：

$$MSE_p = E(\hat{Y} - EY)'(\hat{Y} - EY)$$
$$= (X_p\beta_p - EY)'(X_p\beta_p - EY) + p\sigma^2$$

去掉刻度的影响，即无量纲化，令

$$J_p = \frac{MSE_p}{\sigma^2} = \frac{(X_p\beta_p - EY)'(X_p\beta_p - EY)}{\sigma^2} + p,$$

从预测角度来看，应该选择J_p最小的那个回归子集。但J_p不是统计量，因此要找到J_p的一个合理估计。由于对于子集模型，

$$E(ESS) = E(Y - \hat{Y})'(Y - \hat{Y})$$
$$= (EY - X_p\beta_p)'(EY - X_p\beta_p) + (n-p)\sigma^2$$

从而有

$$J_p = \frac{E(ESS)}{\sigma^2} - n + 2p,$$

因此定义 $C_p = \frac{ESS}{\hat{\sigma}^2} - n + 2p$，其中$\hat{\sigma}^2$为全模型的方差估计，即$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}_s\|^2}{n-s}$，统计量$C_p$为$J_p$的一个合理估计。

C_p criterion：回归协变量集的选择应使 C_p 达到最小。

Akaike Information Criterion(AIC 准则)

设$y_1, \cdots, y_n$为一组样本，服从某个含p个参数的模型，参数用向量$\theta_{p\times 1}$表示，似然函数为$l_p(Y|\theta)$，设参数 θ的极大似然估计为$\hat{\theta}$，令

$$AIC_p = -2\log l_p(Y|\hat{\theta}) + 2p,$$

AIC 准则：(回归)模型应选择使统计量AIC_p达到最小的一组参数。

对正态线性模型，具体地有似然函数

$$l(\beta_p, \sigma^2|Y) = (2\pi\sigma^2)^{-n/2}\exp-\frac{\|Y - X_p\beta_p\|^2}{2\sigma^2},$$

极大似然估计 $\beta_p^* = (X_p'X_p)^{-1}X_p'Y$，$\sigma^{*2} = \frac{ESS}{n}$，其中$ESS = \|Y - X_p\beta_p^*\|^2$，略去与$p$无关的常数得到

$$AIC_p = n\log ESS + 2p,$$

回归协变量集的选择应使得AIC_p达到最小的那个。

以上无论哪一种回归协变量的选择准则都需要对不同协变量的子集进行比较，计算量相当大。

5.5 回归诊断与 Box-Cox 变换

到目前为止得到的估计、检验及其他分析都是在认为模型以及关于模型的假设都是正确的情况下得到的。在许多实际问题中，有时这些关于模型的假设是令人怀疑的，需要作一些诊断。

回归诊断关心的是两个相互有关的问题。首先是模型在多大程度上与观测数据相一致；其次令人感兴趣的问题是，每个案例在估计及综合分析的影响。在某些数据集中，若一个案例被删除，估计或分析的统计量可能有重要改变，这样一个案例称为有影响性的(强影响点)，需要检测出这样的案例并分析原因。

设线性回归模型(含截距项)

$$Y = X\beta + e,\ Ee = 0,\ Cov(e) = \sigma^2 I_n$$

$rank(X_{n\times p}) = p$，X 的第一列为 $E_n = (1,\cdots,1)'$。

或者写成中心化形式 $Y = \alpha E_n + \tilde{X}_c\beta_I + e = \left(E_n \vdots \tilde{X}_c\right)\begin{pmatrix}\alpha\\ \beta_I\end{pmatrix} + e$（见 5.1 节）

β 的估计为 $\hat{\beta} = (X'X)^{-1}X'Y$，对应观测值 Y 的拟合值为 $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$，其中 $H = X(X'X)^{-1}X'$ 称为“帽子矩阵”(hat matrix)。

$H = (h_{ij})_{p\times p}$ 的性质：

1. H 对称幂等，$rank(H) = tr(H) = p$；
2. $HX = X$，$HE_n = E_n$。

令 $\hat{e} = Y - X\hat{\beta} = (I_n - H)Y$，称为残差向量(residual vector)，有性质：

1. $E\hat{e} = 0, Cov(\hat{e}) = \sigma^2(I_n - H)$；
2. $Cov(\hat{e}, \hat{Y}) = 0$；
3. $E_n'\hat{e} = 0$。

记 $\hat{e}=(\hat{e}_1,\cdots,\hat{e}_n)'$，其中 $\hat{e}_i=y_i-x_i'\hat{\beta}$，则 $Var(\hat{e}_i)=\sigma^2(1-h_{ii})$。若 $h_{ii}\approx 1$，则 $Var(\hat{e}_i)=Var(y_i-\hat{y}_i)\approx 0$，由于 $E(\hat{e}_i)=E(y_i-\hat{y}_i)=0$，因此 $y_i-\hat{y}_i\approx 0$。

表明点 (x_i',y_i)，其实际值 y_i 与理论值 $\hat{y}_i$ 拟合的特别好，或者说，这样的点有把拟合的回归平面拖向自己的倾向。这样的点常成为**高杠杆点**(high leverage point)。

进一步，由于

$$h_{ii}=(1\quad x_i')\left(X'X\right)^{-1}\begin{pmatrix}1\\ x_i\end{pmatrix}$$
$$=\frac{1}{n}+(x_i-\bar{x})'\left(\tilde{X}_c'\tilde{X}_c\right)^{-1}(x_i-\bar{x})$$

这里，$\bar{x}=\frac{1}{n}E_n'X$ 为数据中心。若点 x_i 距数据中心越远，则 h_{ii} 值越大，越有将回归平面拉向自己的倾向，如下图。

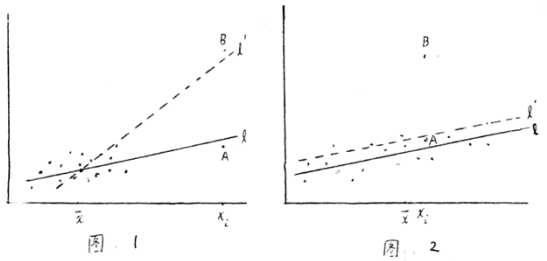
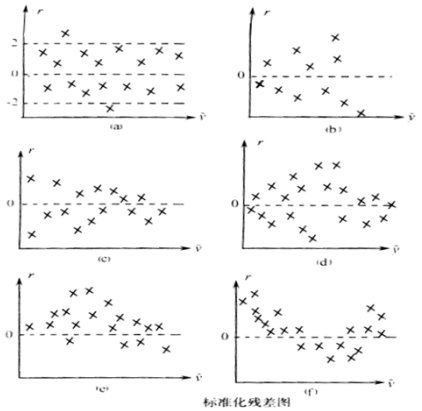


图 1，x_i 远离中心 $\bar{x}$，其 $h_{ii}\approx 1$ 为高杠杆点。若 (x_i,y_i) 在 A 处(属于正常)，则尚不能引起回归直线的基本走向。若在 B 处，则把回归直线由 l 拖向 l'(拉向自己)。好像有一个力将 l 的右端以 $(\bar{x},\bar{y})$ 为定点的杠杆抬上去。这正式高杠杆点这一名称的由来。

而图 2，x_i 离中心 $\bar{x}$ 近，非高杠杆点。此时即便 (x_i,y_i) 处在离群位置 B 处，其作用把回归直线由 l 平行的带上去一点点到 l'。意味着不会较大影响回归系数的估计，仅仅影响了截距项的估计。

若误差正态分布，则$\hat{e}_i \sim N(0,\sigma^2(1-h_{ii}))$，标准化$\frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$，由于 σ^2未知，用 $\hat{\sigma}^2$代替，令$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$，称为“学生化内残差”(因为$\hat{\sigma}^2$用到第 i个案例在内的全部数据)，由于r_i与$\hat{y}_i$轻微相关，故分布比较复杂，但可以近似的认为 r_i相互独立服从$N(0,1)$分布。将点$(\hat{y}_i, r_i)$，$i=1,\cdots,n$描在平面上就可以得到残差图。



线性模型剔除第 i组数据后，剩余$n-1$组数据线性回归模型记为

$Y_{(i)} = X_{(i)}\beta + e_{(i)}$，$Ee_{(i)} = 0$，$Cov(e_{(i)}) = \sigma^2 I_{n-1}$。

β 的 LS 估计记为$\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)}$，为刻画第 i组数据对回归系数估计影响定义

$$D_i = \frac{(\hat{\beta}-\hat{\beta}_{(i)})' X'X (\hat{\beta}-\hat{\beta}_{(i)})}{p\hat{\sigma}^2}$$

称为 Cook 距离。

关于D_i

1. D_i的等值线是椭圆，与置信椭圆具有相同形状；

2.通过线性变换改变X的列，D_i的值不变；

3. 令 $\hat{Y} = X\hat{\beta}$ ， $\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$ ， 则 $D_i = \frac{(\hat{Y}-\hat{Y}_{(i)})'(\hat{Y}-\hat{Y}_{(i)})}{p\hat{\sigma}^2}$刻画了$\hat{Y}$与 $\hat{Y}_{(i)}$的距离。

D_i 大的案例对 $\hat{\beta}$ 及对拟合值 $\hat{Y}$ 都有实质性的影响，删除它们可能会导致结论的重大改变，这样的点称为强影响点。

定理 5.5.1：$D_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}} r_i^2$。

用 Cook 统计量给出判定强影响点的临界值是困难的，在实际中要视具体情况而定。

若一个案例不遵从某个模型，但其余数据遵从，则该案例称为异常值(outliers)。产生异常值的原因一般有二：

1. 用线性模型来近似实际，可能在一定范围是比较好的，当超出该范围时，会产生异常而不适合线性模型；
2. 由于变量测量误差或者数据不正确。

假设第 i 个案例可能是异常值，则

1. 从数据中删除第 i 个案例，余下的 $n-1$ 个案例拟合线性模型；
2. 使用删除后的数据集估计 β 和 σ^2，记为 $\hat{\beta}_{(i)}$ 和 $\hat{\sigma}^2_{(i)}$，以表示第 i 个案例没有用于估计(注 $\hat{\sigma}^2_{(i)}$ 的自由度为 $n-p-1$)；
3. 对于被删除的案例，计算其拟合值 $\hat{y}_{(i)} = x_i'\hat{\beta}_{(i)}$，由于第 i 个案例没有用于估计，y_i 与 $\hat{y}_{(i)}$ 相互独立，

方差 $Var(y_i - \hat{y}_{(i)}) = \sigma^2\left[1 + x_i'\left(X_{(i)}'X_{(i)}\right)^{-1}x_i\right]$，其估计为 $\hat{\sigma}^2_{(i)}\left[1 + x_i'\left(X_{(i)}'X_{(i)}\right)^{-1}x_i\right]$；

4.若 y_i 不是异常值，则 $E(y_i - \hat{y}_{(i)}) = 0$，在误差正态时，检验假设 $E(y_i - \hat{y}_{(i)}) = 0$ 的 t 检验统计量为

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x_i'\left(X_{(i)}'X_{(i)}\right)^{-1}x_i}},$$

给定水平 α，当 $|t_i| > t_{n-p-1}(\alpha/2)$ 时拒绝原假设，即认为 i 数据异常。

称 t_i为“学生化外残差”(因为σ^2的估计没有用到第 i个案例)。

定理 5.5.2：$t_i = \dfrac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$。

通常研究者对异常值无先验选择，若检验具有最大 t_i值的案例为异常值，则事实上进行了n次检验，对每个案例都进行了 t检验。

Box-Cox 变换

对观测得到的试验数据集(x_i', y_i)，$i = 1, \cdots, n$.若经回归诊断后发现不满足 GM 条件，就需要对数据采取“治疗”措施。数据变换是处理有问题数据的一种好的方法。Box 和 Cox(1964)对选择变换的问题给出了一个系统化的处理方法。实践证明，Box-Cox 变换对许多实际数据都是行之有效的，一般可明显地改善数据的正态性、方差齐性。

Box-Cox 变换是对回归响应变量作如下变换：

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}。$$

λ是待定的变换参数。Box 和 Cox 建议检验模型 $Y^{(\lambda)} = X\beta + e$，$Ee = 0$，$Var(e) = \sigma^2 I_n$，可用极大似然估计来确定$\lambda$。具体地，假设$Y^{(\lambda)} \sim N(X\beta, \sigma^2 I_n)$，似然函数为：

$$L(\beta, \sigma^2 | Y^{(\lambda)}) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(Y^{(\lambda)} - X\beta)'(Y^{(\lambda)} - X\beta)\right\}。$$

由于$\left|\dfrac{\partial Y^{(\lambda)}}{\partial Y}\right| = \prod_{i=1}^{n} y_i^{\lambda-1}$，故

$$L(\beta, \sigma^2, \lambda | Y) \propto \sigma^{-n} \prod_{i=1}^{n} y_i^{\lambda-1} \exp\left\{-\frac{1}{2\sigma^2}(Y^{(\lambda)} - X\beta)'(Y^{(\lambda)} - X\beta)\right\}。$$

先固定λ，β, σ^2的极大似然估计分别为

$$\hat{\beta}(\lambda) = (X'X)^{-1}X'Y^{(\lambda)},$$

$$\hat{\sigma}^2(\lambda) = \frac{Y^{(\lambda)'}(I_n - P_X)Y^{(\lambda)}}{n} = \frac{ESS(\lambda, Y^{(\lambda)})}{n}。$$

对应似然函数最大值为

$$L(\lambda)=\max_{\beta,\sigma^2} L(\beta,\sigma^2,\lambda|Y)$$

$$=\left(\frac{2\pi e}{n}\right)^{-\frac{n}{2}}\left[ESS(\lambda,Y^{(\lambda)})\right]^{-\frac{n}{2}}\prod_{i=1}^{n}y_i^{\lambda-1}。$$

对上式求最大值，最大值点 $\hat{\lambda}$作为 λ的极大似然估计。取对数似然，略去无关常数得

$$\ln L(\lambda)=-\frac{n}{2}\ln ESS(\lambda,Y^{(\lambda)})+\ln\prod_{i=1}^{n}y_i^{\lambda-1}。$$

令

$$z_i^{(\lambda)}=\begin{cases} y_i^{(\lambda)}\left(\prod_{i=1}^{n}y_i\right)^{\frac{1-\lambda}{n}}, & \lambda\neq 0 \\ (\ln y_i)\left(\prod_{i=1}^{n}y_i\right)^{\frac{1}{n}}, & \lambda=0 \end{cases},$$

$$Z^{(\lambda)}=(z_1^{(\lambda)},\cdots,z_n^{(\lambda)})',$$

则$\ln L(\lambda)=-\frac{n}{2}\ln ESS(\lambda,Z^{(\lambda)})$，故

$$\hat{\lambda}=Arg\min_{\lambda} ESS(\lambda,Z^{(\lambda)}),$$

其中$ESS(\lambda,Z^{(\lambda)})=Z^{(\lambda)'}(I_n-P_X)Z^{(\lambda)}$。最后得到$\beta,\sigma^2$的估计$\hat{\beta}(\hat{\lambda}),\hat{\sigma}^2(\hat{\lambda})$。

5.6 共线性、岭估计与主成分分析

矩阵X的列向量线性相关等价于方阵$X'X$为奇异矩阵，也等价于$X'X$有 0 特征值。一般称X的列向量近似线性相关，若$X'X$至少有一个很小(接近 0)的特征值。

考虑中心化标准化后的线性回归模型

$$y_i=\alpha+\beta_1\frac{x_{i1}-\bar{x}_1}{s_1}+\cdots+\beta_{p-1}\frac{x_{i,p-1}-\bar{x}_{p-1}}{s_{p-1}}+e_i,$$

$i=1,\cdots,n$，这里 $\bar{x}_j=\sum_{i=1}^{n}x_{ij}/n$，$s_j^2=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$，$1\le j\le p-1$.写成矩阵形式：$Y=\alpha\cdot E_n+X\beta+e$，$Ee=0$，$Var(e)=\sigma^2 I_n$，$rank(X_{n\times(p-1)})=p-1$。

常数项与回归系数的估计为 $\hat{\alpha}=\bar{y}$，$\hat{\beta}=(X'X)^{-1}X'Y$。若设计矩阵$X$的列向量近似线性相关，则称回归模型协变量之间是(近似共线性的(collinearity)，或者称设计矩阵X是共线性的(一般说来，设计矩阵精确共线性是偶然的)。

度量共线性的程度一般有两种。一种度量是 $X'X$ 有特征值$<\varepsilon$，此时称矩阵 X 为ε ill-defined。另外一个是从条件数(condition number)，矩阵 X 的条件数定义为

$cond(X)=\sqrt{\frac{\lambda_{\max}(X'X)}{\lambda_{\min}(X'X)}}$，条件数越大表明矩阵越病态，近似共线性程度越高。

设计矩阵共线性程度对回归系数的估计有着重要的影响。一般衡量估计$\hat{\beta}$的好坏用均方误差(mean squared error)$MSE(\hat{\beta})=E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$。由于

$$\begin{aligned}MSE(\hat{\beta})&=tr\left[E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right]\\&=tr\left[Var(\hat{\beta})\right]\\&=\sigma^2 tr(X'X)^{-1}\end{aligned},$$

设$\lambda_1\ge\cdots\ge\lambda_{p-1}>0$为$X'X$的特征值，则

$$MSE(\hat{\beta})=\sigma^2\sum_{i=1}^{p-1}\frac{1}{\lambda_i}。$$

若$X'X$有一个特征值非常小，则$MSE(\hat{\beta})$就会相当大，从均方误差来看，$\hat{\beta}$的 LS 估计就不是好的估计。

共线性产生的一般可能与数据收集有关，或者可能与回归协变量的选择有关。Hoerl 和 Kennard(1970)建议用

$$\hat{\beta}(k)=(X'X+kI_{p-1})^{-1}X'Y$$

作为β的估计(选择适当的常数$k>0$)，称为岭估计(ridge estimate)。

注：$\hat{\beta}(k)$是有偏的估计。

直观上看，当 X 呈病态时，$X'X$ 的特征值至少有一个非常接近与 0，故$X'X+kI_{p-1}$的特征值接近于 0 的程度将得到改善。

定理 5.6.1：$\exists k>0$使得

$$MSE(\hat{\beta}(k))<MSE(\hat{\beta})。$$

为证明上面的基本定理需要用到线性回归模型的典则形式(canonical form)。设$\lambda_1,\cdots,\lambda_{p-1}$为$X'X$的特征值，$\phi_1,\cdots,\phi_{p-1}$为对应的标准正交化特征向量，即$X'X=\Phi\Lambda\Phi'$，这里$\Lambda=diag(\lambda_1,\cdots,\lambda_{p-1})$，$\Phi=(\phi_1,\phi_2,\cdots\phi_{p-1})$。线性模型可以写成

$Y=\alpha\cdot E_n+Z\gamma+e$，$Ee=0$，$Var(e)=\sigma^2 I_n$，这里$Z=X\Phi$，$\gamma=\Phi'\beta$，此形式称为线性模型的典则形式。

此时参数估计为$\hat{\alpha}=\bar{y}$，$\hat{\gamma}=\Lambda^{-1}Z'Y$，$Var(\hat{\gamma})=\sigma^2\Lambda^{-1}$。典则形式回归系数$\gamma$的岭估计为

$$\begin{aligned}\hat{\gamma}(k)&=(Z'Z+kI_{p-1})^{-1}Z'Y\\&=(\Lambda+kI_{p-1})^{-1}Z'Y\end{aligned},$$

$$\hat{\beta}(k)=(X'X+kI_{p-1})^{-1}X'Y=\Phi\hat{\gamma}(k),$$

因此

$$MSE(\hat{\gamma})=MSE(\hat{\beta}),$$
$$MSE(\hat{\gamma}(k))=MSE(\hat{\beta}(k))。$$

定理 5.6.1 只需对典则形式证明即可。由于

$$MSE(\hat{\gamma}(k))=\sum_{i=1}^{p-1}\frac{k^2\gamma_i^2}{(\lambda_i+k)^2}+\sigma^2\sum_{i=1}^{p-1}\frac{\lambda_i}{(\lambda_i+k)^2},$$

$$\frac{\partial MSE(\hat{\gamma}(k))}{\partial k}=2\sum_{i=1}^{p-1}\frac{\lambda_i(k\gamma_i^2-\sigma^2)}{(\lambda_i+k)^3},$$

由于$\frac{\partial MSE(\hat{\gamma}(k))}{\partial k}\Big|_{k=0}<0$，$\exists k^*$使得$k\in[0,k^*)$，$\frac{\partial MSE(\hat{\gamma}(k))}{\partial k}<0$，$MSE(\hat{\gamma}(k))$为单调减函数，故$MSE(\hat{\gamma}(k))<MSE(\hat{\gamma}(0))=MSE(\hat{\gamma})$。

在实际应用中，岭参数 k的选择很重要。由于$MSE(\hat{\beta}(k))$依赖于未知参数β,σ^2，故 k不能从$\frac{\partial MSE(\hat{\gamma}(k))}{\partial k}=0$得到。Hoerl 和 Kennard 建议选择 k的估计为$\hat{k}=\frac{\hat{\sigma}^2}{\max\limits_{1\le i\le p-1}\hat{\gamma}_i^2}$。这个方法基于如下考虑，若对$i=1,\cdots,p-1$，$k\gamma_i^2-\sigma^2<0$，则一定有$\frac{\partial MSE(\hat{\beta}(k))}{\partial k}<0$。

当设计阵X存在多重共线性时（往往协变量维数过高），即有一些$X'X$的特征值很小时，一个解决多重共线性常用的方法是主成分分析(principal component analysis，PCA)。对线性模型中心化得到$Y=\alpha\cdot E_n+X\beta+e$，同上设$\lambda_1\ge\cdots\ge\lambda_{p-1}>0$为$X'X$的特征值，$\phi_1,\cdots,\phi_{p-1}$为对应的标准正交化特征向量，利用线性模型典则形式$Y=\alpha\cdot E_n+Z\gamma+e$，所谓主成分分析是指：选择$\lambda_1,\cdots,\lambda_q$使得

$\sum_{i=1}^{q}\lambda_i/\sum_{i=1}^{p-1}\lambda_i\ge c$，通常$c=80\%$，85%或者更大，由研究者自己选择；再选取$\lambda_i$所对应的$\phi_i$，$z_i=X\phi_i$，用新模型$Y=\alpha\cdot E_n+\tilde{Z}\tilde{\gamma}+e$

其中$\tilde{Z}_{n\times q}=X(\phi_1,\cdots,\phi_q)$，$\tilde{\gamma}=(\gamma_1,\cdots,\gamma_q)'$，得到LS 估计$\hat{\alpha}=\bar{y}$，$\hat{\tilde{\gamma}}=(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'Y$。其中$z_i$称为第$i$个主成分。当设计矩阵存在多重共线性时，适当选择保留主成分的个数可使主成分估计比原模型 LS 估计有较小的均方误差。主成分估计也是有偏估计。

主成分也等价于下面优化问题的解。

设X已中心化，第一个主成分$z_1=X\phi_1$，其中

$$\phi_1=Arg\max_{\|\phi\|=1}\phi'X'X\phi,$$

第i个主成分$z_i=X\phi_i$，其中

$$\phi_i=Arg\max_{\substack{\|\phi\|=1\\ \phi_l'X'X\phi=0,\ l=1,\cdots i-1.}}\phi'X'X\phi,$$

选择合适的$q\le p-1$个主成分，最后得到回归$\hat{y}^{PCR}=\bar{y}+\sum_{i=1}^{q}\hat{\gamma}_i z_i$。

5.7 偏最小二乘与逐步回归

当协变量维数较高时，设计矩阵往往容易存在共线性，不像上一节的主成分分析(PCA)只用了协变量的数据，并没有考虑响应变量，**偏最小二乘**(partial least squares，PLS)回归利用响应变量Y来构造偏最小二乘方向。

PLS 方法等价于下面优化问题的解。

设X，Y都已中心化，第一个 PLS 方向为$z_1 = X\phi_1$，其中

$$\phi_1 = Arg \max_{\|\phi\|=1} \phi' X'YY'X\phi,$$

第i个 PLS 方向为$z_i = X\phi_i$，其中

$$\phi_i = Arg \max_{\|\phi\|=1 \atop \phi' X'X\phi=0,\ l=1,\cdots i-1.} \phi' X'YY'X\phi,$$

选择合适的$q \le p-1$个 PLS 方向，最后得到回归$\hat{y}^{PLS} = \bar{y} + \sum_{i=1}^{q} \hat{\gamma}_i z_i$。

逐步回归

当协变量维数较高时，若通过所有可能协变量子集来确定最好的回归，例如 5.4 节中的方法，则在计算时间上可能不可行。**逐步回归**(stepwise regression)是解决此问题的一个常用方法。逐步回归一般可以采取向前选择(forward selection, FS)、向后削去(backward elimination, BE)或者交叉逐步(hybrid stepwise)进行。

向前选择 (forward selection, FS)

此方法从只有截距项开始，依次添加协变量。设当前已有k个协变量(含截距项)，回归系数估计为$\hat{\beta}$，残差平方和为$ESS(\hat{\beta})$，设新加一个协变量后回归系数估计为$\tilde{\beta}$，残差平方和为$ESS(\tilde{\beta})$，这样计算F值

$$F=\frac{ESS(\hat{\beta})-ESS(\tilde{\beta})}{ESS(\tilde{\beta})/(n-k-1)},$$

若有$F>F_{1,n-k-1}(\alpha)$，则具有最大F值所对应的协变量将添加进去；否则停止。

向后削去 (backward elimination)

此方法从所有协变量开始，依次剔除协变量。设当前已有$k+1$个协变量(含截距项)，回归系数估计为$\tilde{\beta}$，残差平方和为$ESS(\tilde{\beta})$，设剔除一个协变量后回归系数估计为$\hat{\beta}$，残差平方和为$ESS(\hat{\beta})$，这样计算F值

$$F=\frac{ESS(\hat{\beta})-ESS(\tilde{\beta})}{ESS(\tilde{\beta})/(n-k-1)},$$

若有$F<F_{1,n-k-1}(\alpha)$，则具有最小F值所对应的协变量将被剔除；否则停止。

交错逐步(hybrid stepwise)回归就是轮流使用添加和剔除的方式来选择变量。

5.8 Linear Regression with Regularization

设具有s个协变量的线性回归模型：

$$y_i=\beta_1 x_{i1}+\cdots+\beta_s x_{i,s}+e_i,\ i=1,2\cdots n$$

考虑一般形式地 penalized least-squares 方法，对给定非负罚函数(penalty function)$P(\beta)$，可以写成下面形式

$$\phi(\beta)=(Y-X\beta)'(Y-X\beta)+\lambda P(\beta)$$

其中$\beta=(\beta_1,\cdots\beta_s)'$为回归系数，$\lambda>0$称为 regularization parameter 或者 tuning parameter。$\hat{\beta}=Arg\min_{\beta}\phi(\beta)$。

如果$\lambda \to 0$，这样就得到通常最小二乘估计

$$\hat{\beta}_{LS} = Arg\min_{\beta}(Y - X\beta)'(Y - X\beta)。$$

一般常见罚函数取$P(\beta) = \sum_{j=1}^{s} p(\beta_j)$，其中$p(x) = p(|x|)$为偶函数。例如一般取$P_q(\beta) = \sum_{j=1}^{s} p_q(\beta_j) = \sum_{j=1}^{s} |\beta_j|^q$，相当于$p(x) = |x|^q$。

特别$q = 2$时，$p_2(x) = x^2$，将得到 ridge estimates

$$\hat{\beta}(\lambda) = (X'X + \lambda I_s)^{-1} X'Y。$$

求解

$$\min_{\beta} \phi(\beta) = \min_{\beta}(Y - X\beta)'(Y - X\beta) + \lambda P(\beta) \Leftrightarrow$$

对某个$t > 0$，在$P(\beta) \le t$约束条件下来求解$\min(Y - X\beta)'(Y - X\beta)$。

若当罚函数取$P_q(\beta)$时，定义相应的$\phi_q(\beta)$。取$p(x) = |x|^q$，称为L_q型罚函数。可见当$q > 1$时，$p_q(x) = |x|^q$，此时$\phi_q(\beta)$为光滑的凸函数；当$q = 1$时，$p_1(x) = |x|$，$\phi_q(\beta)$为凸函数，此时$\min_{\beta} \phi_q(\beta)$称为 Lasso(Tibshirani, 1996)。

当$q \ge 1$，可以用传统优化方法来求解$\min_{\beta} \phi_q(\beta)$。而当$0 < q < 1$时，$p_q(x) = |x|^q$，此时$\phi_q(\beta)$不是凸函数，求解较复杂。当$q = 0$时，此时$p_0(x) = I(x \ne 0)$，$P_q(\beta) = \sum_{j=1}^{s} I(\beta_j \ne 0)$，$\min_{\beta} \phi_q(\beta)$等价于$AIC$或 BIC 类型的变量选择准则。

注意：若模型含有常数项，通常是不被惩罚的。

注意到

$$(Y - X\beta)'(Y - X\beta) =$$

$$(Y - X\hat{\beta}_{LS})'(Y - X\hat{\beta}_{LS}) + (\beta - \hat{\beta}_{LS})' X'X (\beta - \hat{\beta}_{LS})$$

因此，

在$P_q(\beta) \le t$条件下，$\min(Y - X\beta)'(Y - X\beta) \Leftrightarrow$

$$\min_{\beta}(\beta - \hat{\beta}_{LS})' X'X (\beta - \hat{\beta}_{LS})$$

$$s.t. \quad \sum_{i=1}^{s} |\beta_i|^q \le t.$$

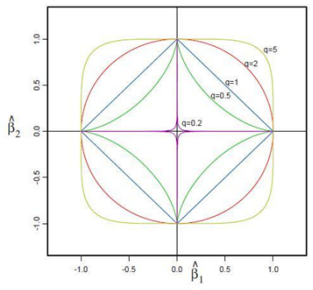
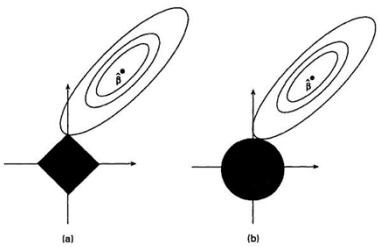


FIGURE *Two-dimensional contours of the symmetric penalty function $p_q(\beta) = |\beta_1|^q + |\beta_2|^q = 1$ for $q = 0.2, 0.5, 1, 2, 5$. The case $q = 1$ (blue diamond) yields the lasso and $q = 2$ (red circle) yields ridge regression.*

当$0 \le q \le 1$时，区域$\sum_{j=1}^{s}|\beta_j|^q \le t$有“角”(corners)，(与坐标轴的交点)，最小值通常都会在此处达到。这样所得的估计$\hat{\beta}$，其某些分量就是 0，从而表明所对应的协变量不在回归协变量中，同时达到变量选择和回归系数的估计。这种自动将某些回归系数估计为 0，从而降低模型复杂程度的估计方法，称为估计的稀疏性(sparsity)。下图解释了 Lasso($q=1$)具有稀疏性。



Estimation picture for (a) the lasso and (b) ridge regression

Fan & Li(2001)指出，一个好的罚函数选择应该使得估计具有以下三个性质：1)近似无偏性(approximately unbiased)（when the true unknown parameter is large to avoid unnecessary modeling bias）；2)稀疏性(sparsity)；3)连续性(continuity)(to avoid instability in prediction)。同时给出

1）满足近似无偏性的充分条件是罚函数满足$\lim_{|x|\to\infty} p'(|x|) = 0$；

2）满足稀疏性的一个充分条件是$\min_x |x| + p'(|x|) > 0$；

3）满足连续性的一个充分必要条件是$\min_x |x| + p'(|x|)$在$x = 0$处达到。

由 2)和 3)可知，同时满足稀疏性和连续性必须是罚函数$p(x)$在 0 点奇异，即在 0 处导数不存在。由此可见对于L_q型的罚函数$p_q(x)=|x|^q$，只有当$0\le q\le 1$，才有稀疏性；$q>1$不具有稀疏性。$q\ge 1$具有连续性，而$0\le q<1$不具有连续性。因此，$q=1$，即 Lasso 具有连续性，又具有稀疏性，而且还是凸函数。

Lasso 的缺点：

Lasso shrinks the estimates for the nonzero coeficients too heavily.

假设真实的模型所对应的回归系数为β_0，其中可以分成两个部分$(\beta_{10}',\beta_{20}')'$，其中$\beta_{20}=0$，即真实模型只包含$\beta_{10}$所对应的协变量。Fan & Li(2001)提出，一个好的变量选择程序得到的估计$\hat{\beta}=(\hat{\beta}_1',\hat{\beta}_2')'$在通常正则条件下应该具有如下两个性质，也称为 Oracle Properties：

1) (model selection consistency):

$$\lim_{n\to\infty} P(\hat{\beta}_2=0)=1;$$

2) (asymptotic normality):

$$\sqrt{n}(\hat{\beta}_1-\beta_{10})\xrightarrow{d.f.}N(0,\Sigma),$$

其中Σ是真实模型下的渐近方差。

在通常条件下，Lasso 不具有 Oracle 性质，只是在一个特殊条件下才有，见 Zhao and Yu (2006)。而对于$q<1$，其解具有 Oracle 性质，但不具有连续性，求解是非凸优化，较复杂。

Zou(2006)提出 Adaptive Lasso，对系数β给予不同的惩罚，对最小二乘估计$\hat{\beta}_{LS}$其绝对值大的，惩罚应该较小，即最小化

$$\min_{\beta}(Y-X\beta)'(Y-X\beta)+\lambda\sum_{j=1}^{s}w_j|\beta_j|$$

其中$w_j=1/\left|(\hat{\beta}_{LS})_j\right|^k, if\ (\hat{\beta}_{LS})_j\neq 0$，否则$w_j=0$。