岭回归和lasso

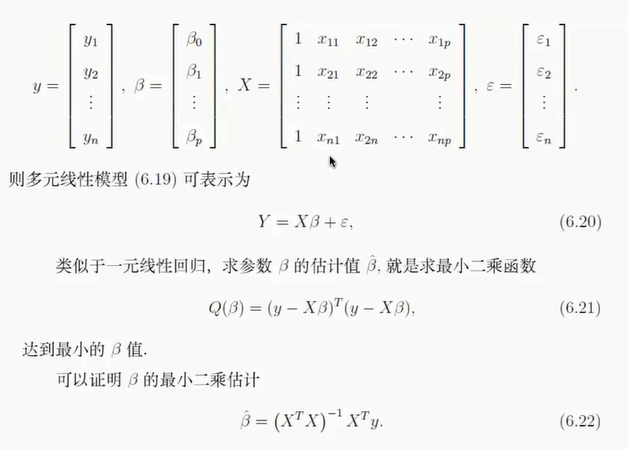
回归里面的两个重大难题：多重共线性和变量选择。

lasso和岭回归的作用：排除多重共线性和进行变量选择

多重共线性也可以归结为变量选择问题。

回归问题的数学背景：

========多元线性回归的最小二乘法(无偏估计)==========



X可以看成是p个变量的n个样本。

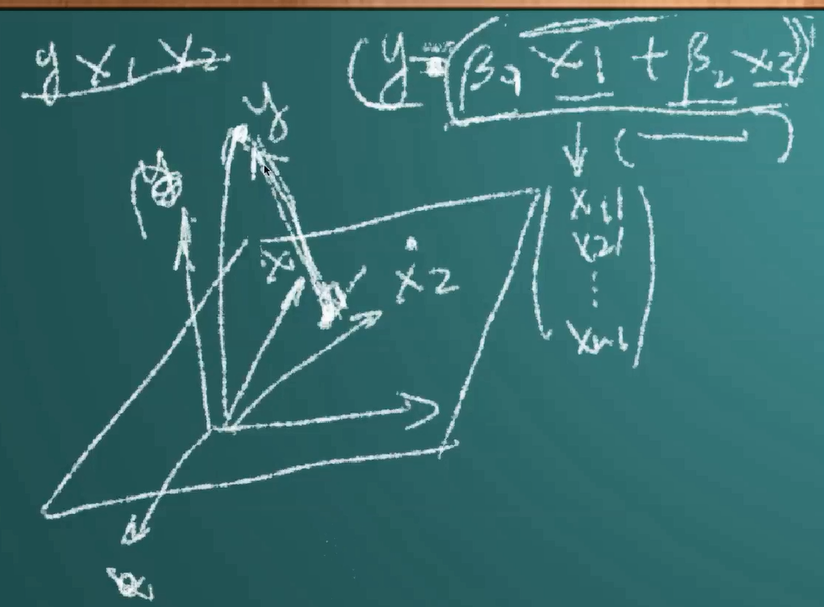
假设样本存在多元线性关系，可使误差项ε达到最低。

最小二乘函数展开就是残差的平方和！多元线性回归目的就是求解最小二乘函数最小时的β值。这里，Y和X都是已知的，都是样本。

我们需要选一组合适的β使得残差平方和达到最小。

可以使用配方法求解，或者用微积分方法求偏导数，两者结果是一样的。

===集合意义==========



假如存在公式y= β1x1+β2x2 因为x1和x2都是变量那么存在一个经过向量x1和x2的平面，向量y的顶点到该平面的垂直距离的平方和即我们所要求得解。

向量y到该平面的投影的坐标即我们所要找出的β1和β2，向量y到该平面的垂直距离就是残差向量，它的长度就是残差值。向量y的顶点投影到该平面上一点称为拟合点

如果x有p个平面，那么形成的是p维的超平面。

Lasso里边有最小角回归。

最后求出的ββ是如下结果：

X的转置乘以X再求逆再乘X的转置再乘y。

这个解叫做最小二乘解，括号部分我们称为广义逆（推广这个矩阵求逆），

如果X是一个矩阵的话，（X的转置乘以X）的逆展开后是X的转置的逆乘以X的逆，然后在和括号后边的X的转置相乘，可以消去X 的转置，结果就变成X的逆乘以y。当然前提是X可求逆。

如果不是 一个方阵的话，X的转置乘以X也会得到一个方阵，这个可以求逆，然后再乘以X的转置，最后还是得到一个方阵。

可以把广义逆看成是矩阵逆的一种推广，原先矩阵的逆只能面对方阵n\*n来求，现在可以面向非方阵来求，并不是任意X都存在广义逆，哪些情况不存在广义逆呢？

1、x本身存在线性相关关系，x的秩并不是满秩的，秩是衡量线性无关程度的一项指标。

2、P\*N矩阵和N\*P矩阵相乘会得到一个P\*P矩阵，如果它的秩不是P，而是<P，那么这个矩阵是不可逆的，此时也不能求解.

X的转置乘以X的秩和原来X的秩是一样的

3、变量比样本多。P>N，则x最大是N，说明也是不可逆的，无法求解。

广义逆的奇异性：



无偏估计：假如Y和X真的存在这样一种关系(Y=Xβ+ε)，可以利用抽样每次抽出来估算一个β^（读作贝塔one），做很多次抽样，可能计算出很多β^，然后取其平均值，该平均值会无限逼近Xβ+ε的值。

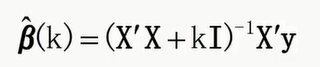
岭回归和lasso都是有偏的。

=========岭回归(Ridge Regression，RR)========================

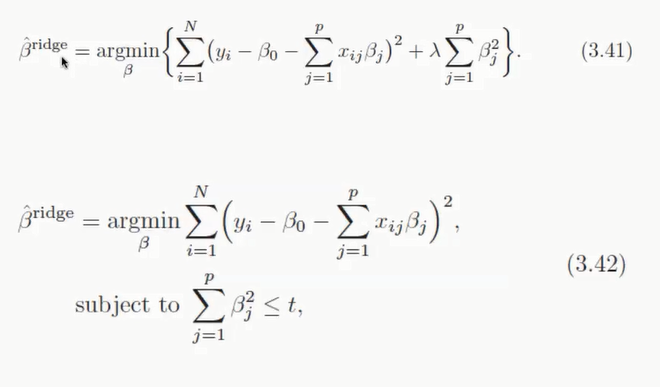
为什么叫岭回归？做岭回归关键是要画一张岭迹图。该图很像山的形状。

1962年由Heer首先提出，1970年后他与肯纳德合作进一步发展。

先对数据做标准化(中心化)，为了记号方便，标准化后的学习集仍然用X表示。

我们称：为β的岭回归估计，其中k称为岭参数。X’就是X的转置。I是扰动。关于岭参数k的岭回归估计。

========等价模型：惩罚函数======



3.41的lambda就是岭参数k，为惩罚函数。和6.21一样。

Lambda和t是一一对应的。

=========岭回归的几何意义=============

残差平方和RSS，β1，β2坐标。

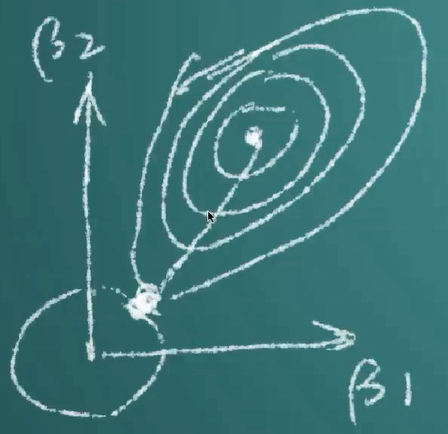
RSS和β1β2的关系就是一个抛物面（内部椭圆）。抛物面的最低点就是最小二乘解。没有约束情况下：



有约束的情况下(β1的平方+β2的平方<=t)：

这样会产生一个圆柱，圆柱和抛物面会产生一个焦点就是我们要求的最小二乘解。

或者将该抛物面横切直到切点恰好与圆柱相交，该焦点就是最小二乘解，如下图：

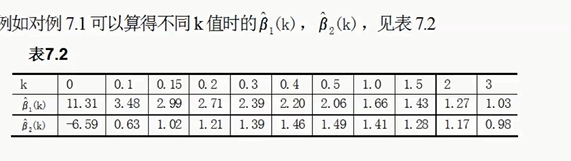


估计族

当岭参数为0，得到最小二乘解

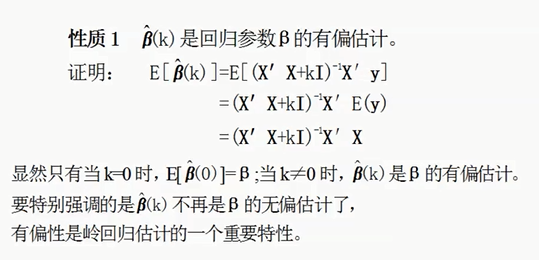
当岭参数取向更大时，岭回归系数估计趋向于0

因为岭参数k不是唯一确定的，所以我们得到的岭回归估计β^(k)实际是回归参数β的一个估计族



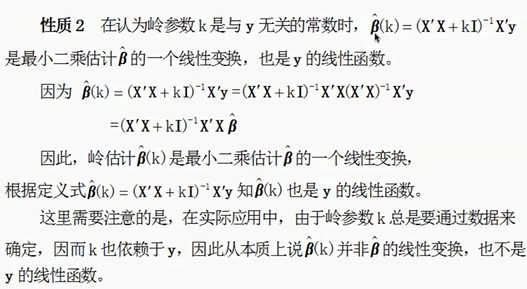
=========岭回归估计的性质=============

性质1 β^(k)是回归参数β的有偏估计。



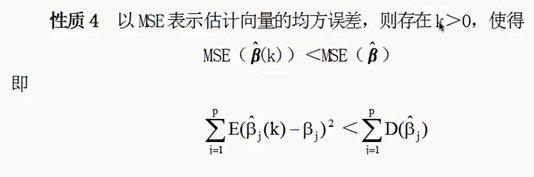
岭回归一般会使得残差平方和变大，但是岭回归会使系数检验变好。

性质2



性质4： 以MSE表示估计向量的均方误差，则存在k>0,使得MSE(β^(k))<MSE(β^)

K比最小二乘法能更加逼近理想回归系数β

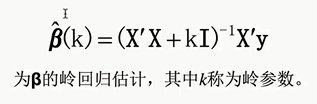


岭迹图：

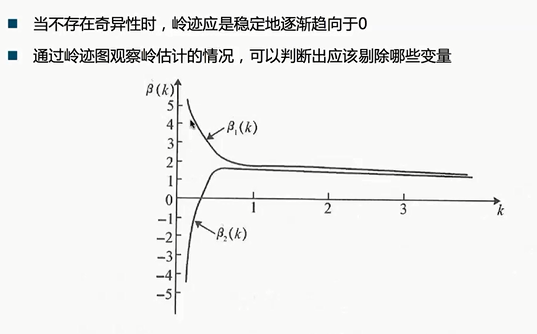
当不存在奇异性时，岭迹应是稳定地逐渐趋向于0。

通过岭迹图观察岭估计的情况，可以判断出应该是剔除哪些变量。

K变化时，左边β这个岭估计会跟着变化如下：

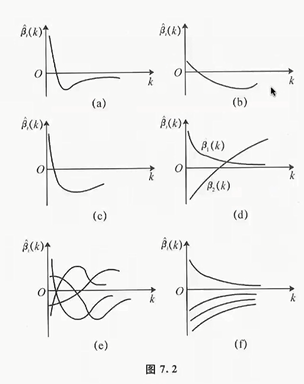


画成图形表示，横坐标k，纵坐标β(k)是向量，有很多的分量



岭迹图能够帮我决定k，当图形呈现喇叭形状时，存在多重共线性，当越来越平缓成一条线越来越接近于0，k就在喇叭口附近的值。岭回归还是主要用来解决多重共线性的。国内有一两本书有教筛选变量的，国外就没有。所以，岭回归做筛选变量还是不靠谱的。

岭迹分析：



=====岭参数的一般选择原则============

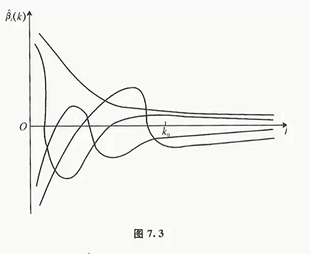
选择k(或lambda)值，使到：

(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

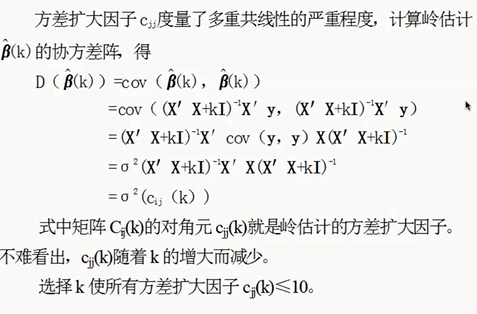
(3)回归系数没有不合乎实际意义的绝对值

(4)残差平方和增大不太多

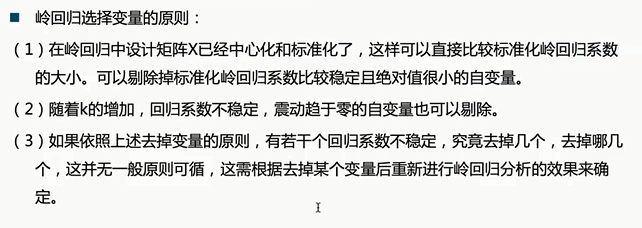


Y轴正负号一定要稳定。

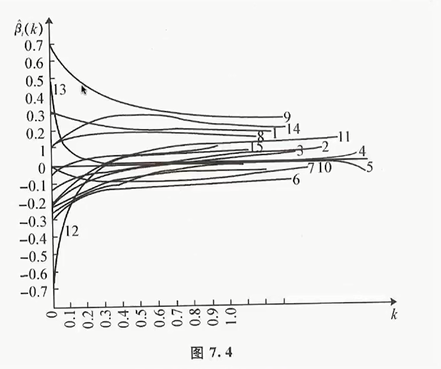
===========方差扩大因子法(解决k的方法)==================

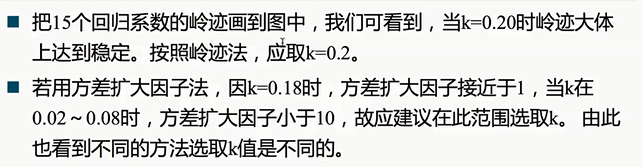


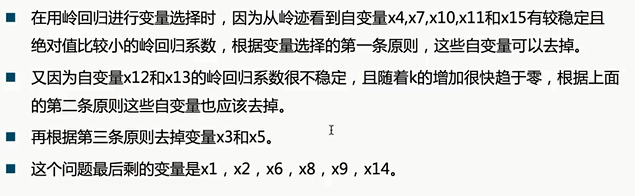
========用岭回归选择变量(并不合适，看看就可以不必认真)======



例子：



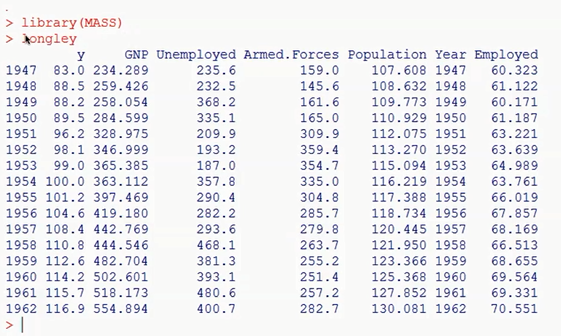




个人认为这样筛选主观性比较强，感觉不大科学。

=================用R语言做岭回归====================

Longley数据集来自J.W.Longley(1967)发表在JASA上的一篇论文，具有强共线性的宏观经济数据，包含GNP defalter(GNP平减指数)、GNP(国民生产总值)、Unemployed(失业率)、ArmedForces(武装力量)、Population(人口)、year(年份)，Employed(就业率)。数据集存在严重的多重共线性问题。





Employed与其他所有变量.的线性关系。

Residuals(残差)：

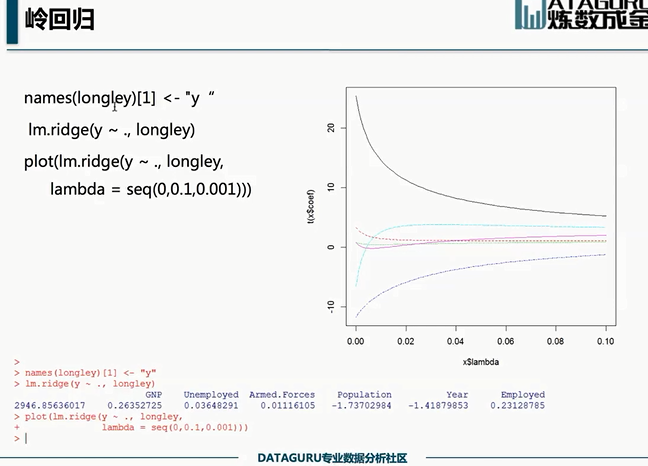
给出最大最小值以及上四分位数3Q和下四分位数1Q

Coefficients：给出估计值、标准误差、t值、p值，p值后边\*\*为显著性检验，\*越多越显著。

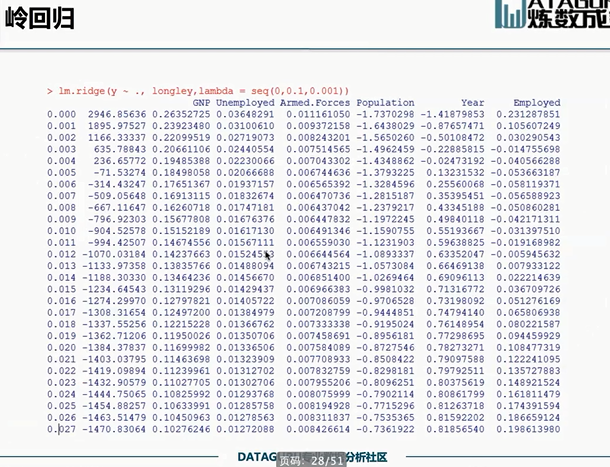
Residual standard error:残差的标准误差，自由度为9。

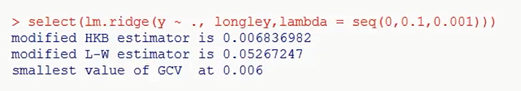
Multiple R-squared：R^2越接近1越优

F-statistic和p-value : f统计量和P值

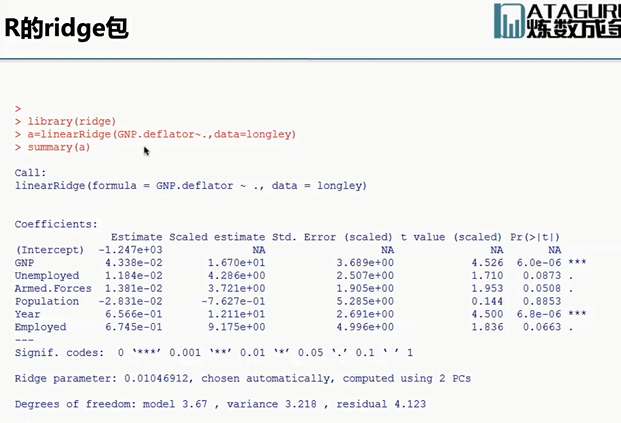


Lambda=seq(0,0.1,0.001) 从0开始到0.1结束，间隔0.001





估计出的岭参数k有三种，而且有10倍之差，可见其用于筛选变量不靠谱。



Lasso是对岭回归改进的一个方法。

==============================================

相关代码：

> library(MASS)

> longley

GNP.deflator GNP Unemployed Armed.Forces Population Year Employed

1947 83.0 234.289 235.6 159.0 107.608 1947 60.323

1948 88.5 259.426 232.5 145.6 108.632 1948 61.122

1949 88.2 258.054 368.2 161.6 109.773 1949 60.171

1950 89.5 284.599 335.1 165.0 110.929 1950 61.187

1951 96.2 328.975 209.9 309.9 112.075 1951 63.221

1952 98.1 346.999 193.2 359.4 113.270 1952 63.639

1953 99.0 365.385 187.0 354.7 115.094 1953 64.989

1954 100.0 363.112 357.8 335.0 116.219 1954 63.761

1955 101.2 397.469 290.4 304.8 117.388 1955 66.019

1956 104.6 419.180 282.2 285.7 118.734 1956 67.857

1957 108.4 442.769 293.6 279.8 120.445 1957 68.169

1958 110.8 444.546 468.1 263.7 121.950 1958 66.513

1959 112.6 482.704 381.3 255.2 123.366 1959 68.655

1960 114.2 502.601 393.1 251.4 125.368 1960 69.564

1961 115.7 518.173 480.6 257.2 127.852 1961 69.331

1962 116.9 554.894 400.7 282.7 130.081 1962 70.551

> summary(fm1 <- lm(Employed ~ ., data = longley))

Call:

lm(formula = Employed ~ ., data = longley)

Residuals:

Min 1Q Median 3Q Max

-0.41011 -0.15767 -0.02816 0.10155 0.45539

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.482e+03 8.904e+02 -3.911 0.003560 \*\*

GNP.deflator 1.506e-02 8.492e-02 0.177 0.863141

GNP -3.582e-02 3.349e-02 -1.070 0.312681

Unemployed -2.020e-02 4.884e-03 -4.136 0.002535 \*\*

Armed.Forces -1.033e-02 2.143e-03 -4.822 0.000944 \*\*\*

Population -5.110e-02 2.261e-01 -0.226 0.826212

Year 1.829e+00 4.555e-01 4.016 0.003037 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3049 on 9 degrees of freedom

Multiple R-squared: 0.9955, Adjusted R-squared: 0.9925

F-statistic: 330.3 on 6 and 9 DF, p-value: 4.984e-10

> names(longley)[1]<-"y"

> lm.ridge(y~.,longley)

GNP Unemployed Armed.Forces Population

2946.85636017 0.26352725 0.03648291 0.01116105 -1.73702984

Year Employed

-1.41879853 0.23128785

> plot(lm.ridge(y~.,longley, lambda = seq(0,0.1,0.001)))

> select(lm.ridge(y~.,longley, lambda = seq(0,0.1,0.001)))

modified HKB estimator is 0.006836982

modified L-W estimator is 0.05267247

smallest value of GCV at 0.006