本节可配合第七讲观看

岭回归和lasso回归

在第七讲时，我们介绍了多元线性回归模型，估计回归系数使用的是OLS，并在最后探讨了异方差和多重共线性对于模型的影响。事实上，回归中关于自变量的选择大有门道，变量过多时可能会导致多重共线性问题造成回归系数的不显著，甚至造成OLS估计的失效。

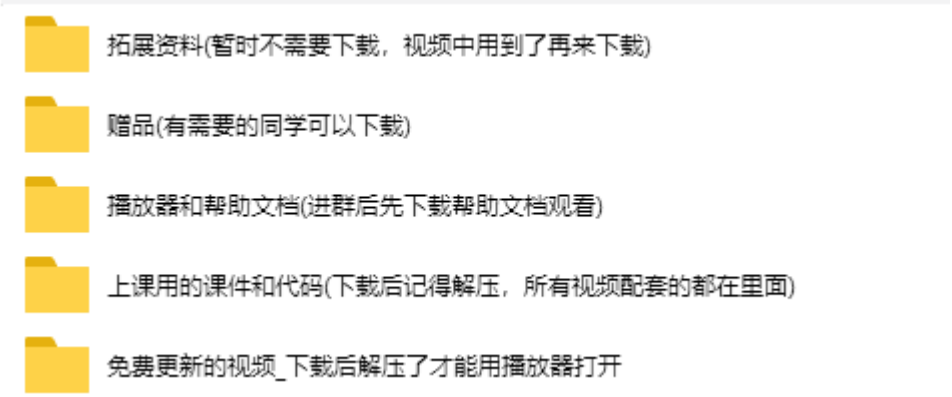
本节介绍到的岭回归和lasso回归在OLS回归模型的损失函数上加上了不同的惩罚项，该惩罚项由回归系数的函数构成，一方面，加入的惩罚项能够识别出模型中不重要的变量，对模型起到简化作用，可以看作逐步回归法的升级版；另一方面，加入的惩罚项能够让模型变得可估计，即使之前的数据不满足列满秩，在稍后的原理推导中我们将更加详细的说明这一点。

注：本讲用到的软件仍为Stata，请没有安装的同学在售后群群文件的拓展资料下载安装。温馨提示，本讲涉及到了一定的数学推导，对模型原理有困难的同学可以直接看应用部分。

温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、我视频中推荐的资料**等。



（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

数学建模学习交流

多元线性回归模型的推导

后面复习就看我的手写讲义

古典回归模型：满足四个假定

假定一：线性假定

假设因变量和自变量之间存在线性关系.

$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$ $(i=1, 2, \cdots, n$，即有n个样本)

注：①当 $x_{i1}=1$ 时，β_1 就是线性方程的截距项

② $\beta_j (j=1, 2, \cdots, k)$ 是未知的回归系数

③ ε_i 是无法观测的且满足一定限制条件的扰动项（有时候用符号 μ_i 表示）

④ $\beta_m = \frac{\partial y_i}{\partial x_{im}}$，所以 β_m 也被称为第m个自变量的偏回归系数.

⑤ 线性假定并不要求初始模型均是以上严格的线性关系，自变量和因变量可通过

文件名：多元线性回归分析的证明和推导.pdf

参考教材：计量经济学林文夫

岭回归的原理

岭回归 (Ridge regression: Hoerl and Kennard, 1970) 的原理和 OLS 估计类似，但是对系数的大小设置了惩罚项。

多元线性回归：$\hat{\beta}=\arg\min_{\hat{\beta}}\sum_{i=1}^{n}(y_i-x_i'\hat{\beta})^2$，其中：$\hat{\beta}=(\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_k)'$

岭回归：$\hat{\beta}=\arg\min_{\hat{\beta}}\left[\sum_{i=1}^{n}(y_i-x_i'\hat{\beta})^2+\lambda\sum_{i=1}^{k}\hat{\beta}_i^2\right]$ (λ为一正常数)

$=\arg\min_{\hat{\beta}}\left[(y-x\hat{\beta})'(y-x\hat{\beta})+\lambda\hat{\beta}'\hat{\beta}\right]$

记 $L=(y-x\hat{\beta})'(y-x\hat{\beta})+\lambda\hat{\beta}'\hat{\beta}$

易知：$\lambda\to 0$ 时，岭回归和多元线性回归完全相同；$\lambda\to+\infty$ 时，$\hat{\beta}=0_{k\times 1}$.

另外：$\frac{\partial L}{\partial\hat{\beta}}=-2x'y+2x'x\hat{\beta}+2\lambda\hat{\beta}=0\Rightarrow(x'x+\lambda I)\hat{\beta}=x'y$

由于 $x'x$ 半正定，则 $x'x$ 特征值均为非负数，加上 λI 后，$x'x+\lambda I$ 特征值均为正数，则 $x'x+\lambda I$ 可逆

所以 $\hat{\beta}=(x'x+\lambda I)^{-1}x'y$ ($\lambda>0$)

Lasso回归的原理

Lasso回归的原理 (Least absolute shrinkage and selection operator)

多元线性回归：$\hat{\beta}=\arg\min_{\hat{\beta}}\sum_{i=1}^{n}(y_i-x_i'\hat{\beta})^2$，其中：$\hat{\beta}=(\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_k)'$

岭回归：$\hat{\beta}=\arg\min_{\hat{\beta}}\left[\sum_{i=1}^{n}(y_i-x_i'\hat{\beta})^2+\lambda\sum_{i=1}^{k}\hat{\beta}_i^2\right]$ (λ为一正常数)

Lasso回归：$\hat{\beta}=\arg\min_{\hat{\beta}}\left[\sum_{i=1}^{n}(y_i-x_i'\hat{\beta})^2+\lambda\sum_{i=1}^{k}|\hat{\beta}_i|\right]$

Lasso回归模型1996年由Robert Tibshirani提出，与岭回归模型相比，其最大的优点是可以将不重要的变量的回归系数压缩至0，而岭回归方法虽然也对原本的系数进行了一定程度的压缩，但是任一系数都不会为0，最终的模型保留了所有的变量。(升级版的逐步回归)

缺点：无显式解，只能使用近似估计算法（坐标轴下降法和最小角回归法）

岭回归和lasso回归的应用

在 Stata 中，我们可以安装 lassopack 命令来实现 Lasso 回归，Lassopack 包含三个与 Lasso 相关的子命令（输入 help lassopack 可以查看详情）： - 子命令 lasso2 可进行 Lasso 估计；**- 子命令 cvlasso 可进行 K 折交叉验证（k-fold cross validation）；** - 子命令 rlasso 可以估计惩罚项由数据决定或者高维情形（变量维度超过样本数）。

注：我们之前在第七讲回归分析中使用的是Stata软件，所以我们这里仍使用Stata软件分析，没有安装Stata软件的同学请在售后群群拓展资料安装。另外，大多数博客或讲义上都是使用Python来做岭回归和lasso回归的，因此有python机器学习基础的同学可以自己查阅相关的调用代码。
另外，Stata中对于岭回归的估计有点bug，因此我们下面只讲Lasso回归的估计，有兴趣的同学可以看这个文章：https://zhuanlan.zhihu.com/p/53905488

安装lassopack命令

（1）电脑联网

（2）输入: findit lassopack 后回车

命令窗口

```
findit lassopack
```

（3）点击蓝色链接进去

```
2 packages found (Stata Journal and STB listed first)
-----------------------------------------------------

lassopack from http://fmwww.bc.edu/RePEc/bocode/l
    'LASSOPACK': module for lasso, square-root lasso, elastic net, ridge,
    adaptive lasso estimation and cross-validation / LASSOPACK is a suite of
    programs for penalized regression / methods suitable for the
    high-dimensional setting where the / number of predictors p may be large
```

（4）在页面中找到这个蓝色的链接点击，进去后就会自动安装
（我的安装速度很慢，大概用了五分钟，可能是下载的服务器在国外的缘故）

```
INSTALLATION FILES                                  (click here to install)
      lassoutils.ado
      lasso2.ado
      lasso2_p.ado
```

如果安装失败的话用手机热点试试

使用lasso回归分析棉花产量例题

先将Excel数据导入到Stata（自变量的量纲相同所以不用标准化）

	年份	单产	种子费	化肥费	农药费	机械费	灌溉费
1	1990	1017	106.05	495.15	305.1	45.9	56.1
2	1991	1036.5	113.55	561.45	343.8	68.55	93.3
3	1992	792	104.55	584.85	414	73.2	104.55
4	1993	861	132.75	658.35	453.75	82.95	107.55
5	1994	901.5	174.3	904.05	625.05	114	152.1
6	1995	922.5	230.4	1248.75	834.45	143.85	176.4
7	1996	916.5	238.2	1361.55	720.75	165.15	194.25
8	1997	976.5	260.1	1337.4	727.65	201.9	291.75
9	1998	1024.5	270.6	1195.8	775.5	220.5	271.35
10	1999	1003.5	286.2	1171.8	610.95	195	284.55
11	2000	1069.5	282.9	1151.55	599.85	190.65	277.35
12	2001	1168.5	317.85	1105.8	553.8	211.05	290.1
13	2002	1228.5	319.65	1213.05	513.75	231.6	324.15
14	2003	1023	368.4	1274.1	567.45	239.85	331.8
15	2004	1144.5	466.2	1527.9	487.35	408	336.15
16	2005	1122	449.85	1703.25	555.15	402.3	358.8
17	2006	1276.5	537	1888.5	637.2	480.75	428.4
18	2007	1233	565.5	2009.85	715.65	562.05	456.9

注：MATLAB中zscore函数可以对数据进行标准化处理。

使用lasso回归分析棉花产量例题

我们使用 K 折交叉验证的方法来选择最佳的调整参数。所谓的 K 折交叉验证，是说将样本数据随机分为 K 个等分。将第 1 个子样本作为“验证集”（validation set）而保留不用，而使用其余 K-1 个子样本作为“训练集”（training set）来估计此模型，再以此预测第 1 个子样本，并计算第1个子样本的“均方预测误差”（Mean Squared Prediction Error）。其次，将第 2 个子样本作为验证集，而使用其余 K-1 个子样本作为训练集来预测第2个子样本，并计算第 2 个子样本的 MSPE。以此类推，将所有子样本的 MSPE 加总，即可得整个样本的 MSPE。最后，选择调整参数，使得整个样本的 MSPE 最小，故具有最佳的预测能力。

```
cvlasso 单产 种子费 化肥费 农药费 机械费 灌溉费, lopt seed(520)
```

其中，选择项“lopt”表示选择使 MSPE 最小的 λ，选择项“seed(520)”表示将随机数种子设为 520（可自行设定），以便结果具有可重复性；默认 K=10（即 10 折交叉验证）。

使用lasso回归分析棉花产量例题

K-fold cross-validation with 10 folds. Elastic net with alpha=1.
Fold **1 2 3 4 5 6 7 8 9 10**

	Lambda	MSPE	st. dev.	
1	**3770.0765**	**20459.787**	**5083.1842**	
2	**3435.1533**	**19908.298**	**4964.2443**	
3	**3129.9837**	**18149.926**	**4346.8097**	
4	**2851.9246**	**16298.956**	**3620.9253**	
5	**2598.5675**	**14780.264**	**3033.957**	
42	**83.135116**	**6467.9448**	**1688.3322**	
43	**75.749621**	**6465.9224**	**1681.9257**	
44	**69.020233**	**6464.6286**	**1676.1033**	*
45	**62.888666**	**6466.2474**	**1671.2623**	
46	**57.30181**	**6528.5216**	**1677.8228**	
47	**52.211275**	**6609.2253**	**1687.2938**	
48	**47.57297**	**6709.7234**	**1705.7325**	

打星号处的 λ=69.02，这是使 MSPE 最小的调整参数。

使用lasso回归分析棉花产量例题

```
Estimate lasso with lambda=69.02 (lopt).
```

Selected	Lasso	Post-est OLS
种子费	0.3205122	0.3065727
农药费	-0.3173749	-0.3437529
灌溉费	0.6905996	0.7388533
Partialled-out*		
_cons	956.8974544	964.0853232

上表右边第 1 列即为 Lasso 所估计的变量系数。其中，除常数项外，只有 3 个变量的系数为非零，而其余变量（未出现在表中）的系数则为 0。考虑到作为收缩估计量的 Lasso 存在偏差（bias），上表右边第 2 列汇报了 “Post Lasso” 估计量的结果，即仅使用 Lasso 进行变量筛选，然后扔掉 Lasso 的回归系数，再对筛选出来的变量进行 OLS 回归。
注意：以上结果可能随着我们之前设置的随机数种子变化，因为lasso回归的估计是近似算法，且剔除的多重共线性变量是相对的。

总结：何时使用lasso回归？

我们首先使用最一般的OLS对数据进行回归，然后计算方差膨胀因子VIF，如果VIF>10则说明存在多重共线性的问题，此时我们需要对变量进行筛选。

在第七讲时我们提到可以使用逐步回归法来筛选自变量，让回归中仅留下显著的自变量来抵消多重共线性的影响，学完本讲后，大家完全可以把lasso回归视为逐步回归法的进阶版，我们可以使用lasso回归来帮我们筛选出不重要的变量，步骤如下：（1）判断自变量的量纲是否一样，如果不一样则首先进行标准化的预处理；（2）对变量使用lasso回归，记录下lasso回归结果表中回归系数不为0的变量，这些变量就是最终我们要留下来的重要变量，其余未出现在表中的变量可视为引起多重共线性的不重要变量。

在得到了重要变量后，我们实际上就完成了变量筛选，此时我们只将这些重要变量视为自变量，然后进行回归，并分析回归结果即可。（注意：此时的变量可以是标准化前的，也可以是标准化后的，因为lasso只起到变量筛选的目的）