系统工程第 5 次作业

张博睿　自 75　2017011537

程序详见

$$./code/..$$

1.

（1）使用 PCA+线性回归结果

算法流程

输入：数据$data$和标签Y。
（1）对数据进行归一化并保存数据的均值和方差 $$\tilde{X} = \frac{X - \bar{X}}{std(X)}$$ （2）计算归一化数据的协方差矩阵 $$\Sigma = \frac{1}{n-1} X^T X$$ （3）对协方差矩阵进行特征值分解，并根据$rerr$选择占比较大的特征值和特征向量（主成分）pcs。 （4）计算降维后的数据 $$\hat{X} = \tilde{X} \times pcs$$ （5）通过 OLS 计算$\hat{X}$和$\tilde{Y}$的回归系数并进行 F 检验。 （6）根据保存的均值和方差恢复X关于Y的回归系数。
输出：最终的回归系数和相关中间过程结果。

最终结果如下

项目	结果
显著性水平α	0.05
显著性检验F_0	1.69
计算F	208.45
显著性检验	存在线性
病态特征值阈值	0.95
降维维度	10
置信区间	$y \in [\hat{y} - 10.724, \hat{y} + 10.724]$

回归系数为

项目	$pop.density$	pop	$pop.change$	$age6574$
系数	-0.000378817	$-2.16E-06$	-0.001496108	0.607704585

项目	$age75$	$crime$	$college$	$income$
系数	0.680461857	−0.000420616	0.329745055	0.000252818
项目	$farm$	$democrat$	$republican$	$Perot$
系数	0.161117621	−0.000167626	−0.096145826	0.155669417
项目	$white$	$black$	$intercept$	
系数	0.05514121	−0.030452773	19.589064	

运行的原始结果如下

```
Dimension reduction:  10
Linearity is considered under the condition of significance 0.05
======== OLS result ===============
omega:            [ 0.20013795  0.20374793 -0.24534054  0.0444032  -0.05773427  0.18024992
  0.27955154 -0.12137926  0.19162341 -0.00420912]
F:                208.4529095257804
F0:               1.6949614976166913
interval:         10.723590157482047


======== recover result ===========
re_omega:         [-3.78816653e-04 -2.15781284e-06 -1.49610831e-03  6.07704585e-01
  6.80461857e-01 -4.20616476e-04  3.29745055e-01  2.52818439e-04
  1.61117621e-01 -1.67626174e-04 -9.61458256e-02  1.55669417e-01
  5.51412104e-02 -3.04527729e-02]
intercept:        19.589064434611366
```

（2）直接使用病态回归建模

最终结果如下

项目	结果
显著性水平α	0.05
显著性检验F_0	1.83
计算F	292.21
显著性检验	存在线性
病态特征值阈值	0.95
降维维度	10
置信区间	$y \in [\hat{y} - 10.717, \hat{y} + 10.717]$

回归系数为

项目	$pop.density$	pop	$pop.change$	$age6574$
系数	−0.000378817	$-2.16E-06$	−0.001496108	0.607704585
项目	$age75$	$crime$	$college$	$income$
系数	0.680461857	−0.000420616	0.329745055	0.000252818
项目	$farm$	$democrat$	$republican$	$Perot$
系数	0.161117621	−0.000167626	−0.096145826	0.155669417
项目	$white$	$black$	$intercept$	
系数	0.05514121	−0.030452773	19.589064	

运行的原始结果如下

```
Decomposed dimension: 10
Linearity is considered under the condition of significance 0.05
omega         [ 0.20013795  0.20374793 -0.24534054  0.0444032  -0.05773427  0.18024992
  0.27955154 -0.12137926  0.19162341 -0.00420912]
re_omega      [-7.14372789e-02 -7.51337839e-02 -3.99718610e-03  1.72879389e-01
  2.11106945e-01 -1.27942808e-01  2.84949246e-01  2.33785569e-01
  1.55614234e-01 -2.37453228e-04 -1.08535378e-01  1.40831051e-01
  1.12213217e-01 -5.75733293e-02]
final_omega   [-3.78816653e-04 -2.15781284e-06 -1.49610831e-03  6.07704585e-01
  6.80461857e-01 -4.20616476e-04  3.29745055e-01  2.52818439e-04
  1.61117621e-01 -1.67626174e-04 -9.61458256e-02  1.55669417e-01
  5.51412104e-02 -3.04527729e-02]
intercept     19.589064434611366
S_sigma       0.7189139358874499
Z             1.959963984540054
error         10.716676172135639
```

（3）分析

可以看到，使用 **PCA+线性回归**和**直接使用病态回归**的结果基本一致，下面对比分析两者的不同点和相同点。

不同点：

	PCA+线性回归	病态回归
优化目标	$\max \sum_{t=1}^{N}\sum_{k=1}^{m}\left(y_k(t)\right)^2$ $s.t.\begin{cases} y_k(t)=\sum_{i=1}^{n} l_i(k)\tilde{x}_i(t) \\ \sum_{i=1}^{n}\left(l_i(k)\right)^2=1 \\ \sum_{i=1}^{n} l_i(k)l_i(j)=0 \end{cases}$	$\sum_{i=1}^{N}\left(x(t)-Lv(t)\right)^T\left(x(t)-Lv(t)\right)$
特征分解	计算协方差矩阵（除以 n-1）	没有除以 n-1

相同点：

在处理实际问题是，两者都是通过计算协方差矩阵的特征分解，并通过选择较大特征值的方式来达到优化目标。

分析：

通过实验发现，两种方法得到的结果基本是一致的。回到两种方法的目标函数可以发现，尽管两种方法的优化目标存在一定的差异，但是在实际运算的时候可以用相似的方法（特征值分解）来优化目标函数。

由于主成分分析的时候强调了协方差矩阵，计算时除以了系数 n-1，而病态回归问题没有。但是这在选择特征值和特征向量（主成分）时并没有太多的影响（可能存在数值计算的差别）。最终导致两种方法的效果基本是一致的。