《系统工程导论》主成分分析作业

1. 编程实现 PCA 算法，具体要求如下：

(1) 实现函数（以 MATLAB 函数为例）

```
function [pcs, cprs_data, cprs_c] = pca_compress(data, rerr)
```

其中输入输出变量含义如下

变量名	含义
data	输入的原始数据矩阵，每一行对应一个数据点
rerr	相对误差界限，即相对误差应当小于这个值，用于确定主成分个数
pcs	各个主成分，每一列为一个主成分
cprs_data	压缩后的数据，每一行对应一个数据点
cprs_c	压缩时的一些常数，包括数据每一维的均值和方差等。利用以上三个变量应当可以恢复出原始的数据

(2) 实现函数（以 MATLAB 函数为例）

```
function recon_data = pca_reconstruct(pcs, cprs_data, cprs_c)
```

其中输入输出变量含义如下

变量名	含义
pcs	各个主成分，每一列为一个主成分
cprs_data	压缩后的数据，每一行对应一个数据点
cprs_c	压缩时的一些常数，包括数据每一维的均值和方差等。利用以上三个变量应当可以恢复出原始的数据
recon_data	恢复出来的数据，每一行对应一个数据点

2. 利用上面编写的函数，以及线性回归章节作业中编写的函数，对附件的数据进行建模。附件的数据为美国 1992 年总统竞选各个 county 的投票情况，数据说明如下

Name	Labels	Storage
county		character
state		character
pop.density	1992 pop per 1990 miles^2	double
pop	1990 population	double
pop.change	% population change 1980-1992	double
age6574	% age 65-74 1990	double
age75	% age >= 75 1990	double
crime	serious crimes per 100000 1991	double
college	% with bachelor's degree or higher of those age>=25	double
income	median family income 1989 dollars	double
farm	farm population % of total 1990	double
democrat	% votes cast for democratic president	double
republican	% votes cast for republican president	double
Perot	% votes cast for Ross Perot	double
white	% white 1990	double

black	% black 1990	double
turnout	1992 votes for president / 1990 pop x 100	double

数据来源：http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/

请将从 pop.density 到 black 一共 14 个变量作为 x，将 turnout 作为 y，试建立 y 关于 x 的线性回归模型，给出 y 的表达式和置信区间（为书写方便，可以在有明确说明的情况下只给出 y 表达式中的系数和置信区间的半长度）。

提示：可以利用所学的知识检查自变量之间是否有线性相关关系，利用 pca 对自变量进行压缩后即可认为消除了自变量之间的线性相关。

作业要求：

1. 独立完成不得抄袭
2. 提交电子版。如题意不明请随时联系助教
3. 所有代码请独立于作业报告存放，不要贴在作业报告中；
   报告中请附加实验结果与分析。
4. 第一题不用写在报告中，第二题需要详细说明解题过程和思路，还可以包括解题中发现的问题以及你的解决方法
5. 编程语言可用MATLAB和python（建议使用MATLAB）。如果使用其他语言，在输入输出上可做变通，但必须包含上面提到的所有内容。