WEEK 6 HOMEWORK – SAMPLE SOLUTIONS

IMPORTANT NOTE

These homework solutions show multiple approaches and some optional extensions for most of the questions in the assignment. You don't need to submit all this in your assignments; they're included here just to help you learn more – because remember, the main goal of the homework assignments, and of the entire course, is to help you learn as much as you can, and develop your analytics skills as much as possible!

Question 1

Using the same crime data set as in Homework 5 Question 2, apply Principal Component Analysis and then create a regression model using the first 4 principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your solution to Homework 5 Question 2. You can use the R function prcomp for PCA. (Note that to first scale the data, you can include scale. = TRUE to scale as part of the PCA function.)

Here's one possible solution. Please note that a good solution doesn't have to try all of the possibilities in the code; they're shown to help you learn, but they're not necessary.

The file HW6-Q1-fall.R shows how to calculate the principal components, run a regression, and find the resulting coefficients in terms of the original variables.

The model using the first four principal components is the following:

Crime ~ 1667 – 16.9M + 21.3So + 12.8Ed + 21.4Po1 + 23.1Po2 – 347LF – 8.3MF + 1.0Pop + 1.5NW – 1510U1 + 1.7U2 + 0.040Wealth – 6.9Ineq + 144.9Prob – 0.9Time

Its R^2 is just 0.309, much lower than the 0.803 found by the regression model found in the previous homework. But, remember that cross-validation showed a lot of overfitting in the previous homework. The R code shows the results using the first k principal components, for k=1..15:

Model	R-squared on training data	Cross-validation R-squared
Top 1 principal component	0.17	0.07
Top 2 principal components	0.26	0.09
Top 3 principal components	0.27	0.07
Top 4 principal components	0.31	0.11
Top 5 principal components	0.65	0.49

Top 6 principal components	0.66	0.46
Top 7 principal components	0.69	0.46
Top 8 principal components	0.69	0.37
Top 9 principal components	0.69	0.33
Top 10 principal components	0.70	0.30
Top 11 principal components	0.70	0.19
Top 12 principal components	0.77	0.39
Top 13 principal components	0.77	0.39
Top 14 principal components	0.79	0.47
Top 15 principal components	0.80	0.41
All original variables [from HW3]	0.80	0.41

From the table above, it seems clear that there's still over-fitting (not surprising with so few data points for the number of factors, as we saw in HW5). And it seems clear that adding the 5th principal component is important.