Introduction to Computational Statistics INSH 5301

Homework 11

03/30/2020

For this homework you'll need to use some real world data to answer some research questions using multiple regression. The data, along with the data description, can be downloaded from the course material section in Blackboard. You can also downlod this dataset using the AER package.
AER package: Just
Run 'library(AER)' first, and then 'data("CollegeDistance")'

For this homework use the bfi dataset (25 personality items thought to boil down to a few core personality types) from the psych package. You can load the data using, for instance, data(bfi) after loading the psych package; you may need to clean it a bit first with na.omit() to remove the observations with na items, or else impute those missing items. It might also help to use scale() on your dataset before analysis. scale() takes all your variables (columns) and rescales them to have a mean of 0 and a sd of 1, so that you can more easily compare all your factors or clusters to see which are larger or smaller.
For the factor analysis, you may use any of the methods covered in the lesson – they should all produce similar results, though princomp and prcomp might be simplest. You don't have to interpret everything, say, fa() outputs, which produce a lot of output – is easier to use str() to examine the output of your function and find the quantities you want.
Also, because some of the methods when deciding the number of factors and number of clusters are not objective, don't worry about not getting the right number. But provide an explanation to why you choose a the particular number, your reasoning should be based on the described methodologies in the learning module.

After running a factor analysis or PCA, be sure to discuss and interpret the results:

1. Examine the factor eigenvalues in the dataset. Plot these in a scree plot and use the "elbow" test to guess how many factors one should retain.

2. How many factors are needed to explain 50 percent of the total variance in the dataset?

3. Examine the loadings of the factors on the variables (sometimes called the "rotation" in the function output) – ie, the projection of the factors on the variables – focusing on just the first one or two factors. Sort the variables by their loadings, and try to interpret what the first one or two factors "mean." This may require looking more carefully into the dataset to understand exactly what each of the variables were measuring. You can find more about the data in the psych package using ?psych.

Next perform a cluster analysis with the same dataset.

4. First use k-means and examine the centers using two and three centers. How are they similar to and different from the factor loadings of the first couple factors?

5. Next use hierarchical clustering. Print the dendrogram, and use that to guide your choice of the number of clusters. Use cutree to generate a list of which clusters each observation belongs to. Aggregate the data by cluster and then examine those centers (the aggregate means) as you did in (4). Can you interpret all of them meaningfully using the methods from (4) to look at the centers?

6. From the factor and cluster analysis, what can you say more generally about what you have learned about your data?