Problem Set 5

BUAN 6356

Due: Tuesday, 2017-12-05-11:59pm

Deliverable:

an R source-code file named ps5.r

Question 1

Data

The WAGE1.csv data is the data from your first problem set.

Analysis

- Read the data WAGE1.csv into a new variable: context1
- Consider using $k-$means to segment the WAGE1.csv data. Plot the within-group sum of squares for $k = 1, 2, \ldots, 10$.
- Set the seed to be 2 as we did in class and use $k-$means with 10 initial starting positions to estimate $k = 3$ means (which may or may not be the correct number) using context1. Store the $k-$means result in: model1 (Hint: be careful about your seed, or you might get a different cluster order than the grading system does.)
- Find the estimated means from model1.
- Using model1, segment the data into three groups and run the following linear model for clusters 1, 2, and 3. Store the results in model2, model3, and model4 respectively.

$$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{tenure}_i + e_i$$

Interpretations

a. Using the elbow test on the within- sum of squares plot, find the optimal number o clusters for this data set.

b. Looking at the means from model1, describe the different clusters. [Hint: Look at the education, experience, and tenure variables in particular.]

c. Discuss the differences between models 2, 3, and 4.

Question 2

Data

The data in ffportfolios.csv contains market returns for 32 stock market portfolios from July 1963 to September 2017.

Analysis

- Read the data ffportfolios.csv into a new variable: context2
- Run the level KPSS test on all 32 time series and verify that every series is level stationary without the need for any differencing.
- Run principal components analysis on the 32 portfolios and store the result in: model5.
- Generate the scree plot for model5.
- Store the first principal component inside of context2 as: factor
- Standardize the factor variable to have variance equal to one.
- Find the year values where the standardized factor is less than -2.58

Interpretations

a. Based on the scree plot, how many principal componenets should we use for this data?

b. Looking at the years where the standardized factor is less than the first percentile (-2.58), how would you characterize this principal component?