DSC324/424
Assignment #2 (DUE SUNDAY, August 2nd, 2020 by Midnight)

Deliverables: Turn in your answers in a single PDF file. Use KnitR or Copy any R output relevant to your answer into your Word document and explain your answer thoroughly and include a copy of the full analysis in your report along with your conclusions. Also, provide your R code files.

Problem 1 (10 points) Answer each of the following questions:

a) What are regularized regressions? What are the differences between ridge and lasso regressions?

b) What are some causes of overfitting? How do we diagnose and treat overfitting in regression models?

c) What is multicollinearity? How do we diagnose and treat multicollinearity in regression models?

Problem 2 (10 Points): Have 1 Group member post the answers to the below questions **to the final project forum under the discussion section of D2L**:

- Project Team: Group Members
- Data:
    - Subject Area or Field of Interest
    - Source of Data (provide link to data)
    - Specific dataset(s)
    - description of its scope (# metric variables, #categorical variables, #samples, multiple related tables?)
    - Technology group plans to use for Project (i.e. Python, R, SPSS, Tableau, etc.)
        - How do you plan to use the technology?

    - In addition, as you are forming your groups, remember the following requirements for datasets and groups:

        a. Your group should have 4-5 people in it.
        b. Please to make sure to have 1 liaison person for the group, who can submit assignments and ask me questions on behalf of the group.
        c. Your dataset should be a real and rich dataset with at least 15 to 20 variables metric (continuous). It should have at least
        (10 * #var, but better yet 20 * #var) samples (we will see that some techniques like PCA require this for significance/stability). You will need a large sample size if you have a large n umber of variables. See me if you have any doubts about your dataset.

Problem 3 (Paper review) (10 Points) An academic paper from a conference or Journal will be posted to the Homework 2 content section of D2L. Review the paper and evaluate their usage of Factor Analysis. In particular address the following: **(See article on Understanding and measuring coping with cyberbullying in adolescents: exploratory factor analysis of the brief coping orientation to problems experienced inventory)**

- How are they applying Factoring Analysis?
- What kind of factor rotation do they use?
- How many factors do they concentrate on in their analysis? How did they arrive at these number of factors?
- Explain the breakdown of the factors and the significance of their names.
- How do they evaluate the stability of the components (i.e. factorability)?
- Do they use these factors in later analysis, such as regression? If so, what do they discover?
- What overall conclusions does Factor Analysis allow them to draw?

Problem 4 (Principal Component Analysis - 20 points): The data given in the file 'Big5.csv' are 5-point Likert items taken from the Big Five Personality Test web-based personality assessment. Techniques, such as Principal Component Analysis (PCA), can be used to determine different types of personalities. There are 19,719 subjects in the file and 50 variable items as follows:

E1 I am the life of the party.

E2 I don't talk a lot.

E3 I feel comfortable around people.

E4 I keep in the background.

E5 I start conversations.

E6 I have little to say.

E7 I talk to a lot of different people at parties.

E8 I don't like to draw attention to myself.

E9 I don't mind being the center of attention.

E10 I am quiet around strangers.

N1 I get stressed out easily.

N2 I am relaxed most of the time.

N3 I worry about things.

N4 I seldom feel blue.

N5 I am easily disturbed.

N6 I get upset easily.

N7	I change my mood a lot.
N8	I have frequent mood swings.
N9	I get irritated easily.
N10	I often feel blue.
A1	I feel little concern for others.
A2	I am interested in people.
A3	I insult people.
A4	I sympathize with others' feelings.
A5	I am not interested in other people's problems.
A6	I have a soft heart.
A7	I am not really interested in others.
A8	I take time out for others.
A9	I feel others' emotions.
A10	I make people feel at ease.
C1	I am always prepared.
C2	I leave my belongings around.
C3	I pay attention to details.
C4	I make a mess of things.
C5	I get chores done right away.
C6	I often forget to put things back in their proper place.
C7	I like order.
C8	I shirk my duties.
C9	I follow a schedule.
C10	I am exacting in my work.
O1	I have a rich vocabulary.
O2	I have difficulty understanding abstract ideas.
O3	I have a vivid imagination.
O4	I am not interested in abstract ideas.
O5	I have excellent ideas.
O6	I do not have a good imagination.

O7 I am quick to understand things.

O8 I use difficult words.

O9 I spend time reflecting on things.

O10 I am full of ideas.

A) How many components are need to explain 100% of total variation for this data? How many components are determined from the scree plot? What number of components would you use in the model?

B) For the number of components in part A, give the formula for each component and a brief interpretation after rotating the components. What names might you give for each of the components?

C) What subjects have the highest and lowest values for each principal component (only include the number of components specified in part A. For each of those subjects, give the principal component scores (again only for the number of components specified in part A).

D) Finally, run a common factor analysis on the same data. What difference, if any, do you find? Does the factor analysis change your ability to interpret the results practically?