STA380.18
Homework on Cluster Analysis

Directions: *This homework should be submitted on Canvas in the Quiz Module. Each question has a numerical answer. First, solve all of the questions, then log on to the quiz in Canvas and enter your solutions there. You will have to enter all of your solutions on Canvas without interruption. After you have submitted your solutions on Canvas, you will get instant feedback. Your homework solutions are to be entirely your own effort. You may not communicate with anyone about the homework, except for the TA and/or the instructor. You may use the Canvas postings, in-class discussion, any of the recommended textbooks, and computer software, if necessary, but no other resources.*

This homework references HWcluster.xlsx on Canvas, which contains two sheets of data. Data1 is for questions 1-20. Data2 is for questions 21-35. [Hint: For some or all of the questions, you may find it helpful to perform computations in Excel.]

Context for questions 1-20:

For a given clustering of the data, the total sum of squares (TSS) is the sum of squared distances between all of the cases and the centroid of all of the cases. In symbols,

$$\text{TSS} = \sum_{i=1}^{m}\sum_{k=1}^{p}\sum_{l=1}^{n_i}(X_{ikl} - \bar{X}_{.k.})^2 ,$$

where X_{ikl} is the value of case l on variable k in cluster i, there are p variables, m clusters, and n_i is the number of cases in cluster i, and $\bar{X}_{.k.}$ is the mean of variable k over all cases in the dataset.

The between-cluster sum of squares (BSS) is the sum of weighted squared distances between each cluster's centroid and the centroid of all cases, where the weight given to cluster i is n_i . In symbols,

$$\text{BSS} = \sum_{i=1}^{m}\sum_{k=1}^{p} n_i(\bar{X}_{ik.} - \bar{X}_{.k.})^2$$

where $\bar{X}_{ik.}$ is the mean of the n_i values of variable k in cluster i.

The within-cluster sum of squares for a given cluster i is the sum of squared distances from each of the cases in cluster i to the centroid of cluster i. In symbols, it is $\sum_{k=1}^{p}\sum_{l=1}^{n_i}(X_{ikl} - \bar{X}_{ik.})^2$ for cluster i. The within-cluster sum of squares for the dataset (WSS) is the sum of within-cluster sums of squares over all m clusters. In symbols,

$$\text{WSS} = \sum_{i=1}^{m}\sum_{k=1}^{p}\sum_{l=1}^{n_i}(X_{ikl} - \bar{X}_{ik.})^2$$

Note: For questions 1-20, do not standardize the data in HWcluster.xlsx.

For the 8 numbered cases in the Data1 sheet of *HWcluster.xlsx*, calculate:

1. The within-cluster sum of squares for Cluster 1
2. The within-cluster sum of squares for Cluster 2
3. The within-cluster sum of squares for Cluster 3
4. WSS
5. BSS
6. TSS
7. The explanatory power ("R-square") of the 3-cluster solution.

If each of the 8 numbered cases in the Data1 sheet of *HWcluster.xlsx* were in its own singleton cluster, what would be the values of:

8. TSS
9. BSS
10. WSS
11. The explanatory power ("R-square") of the 8-cluster solution.

If all 8 numbered cases in the Data1 sheet of *HWcluster.xlsx* were in one super-cluster, what would be the values of:

12. TSS
13. BSS
14. WSS
15. The explanatory power ("R-square") of the 1-cluster solution.

Given the clustering of the 8 numbered cases into the three clusters as shown in the Data1 sheet of *HWcluster.xlxs*, suppose that a 9th case becomes available. Its coordinates are X = 5, Y = 2. To which of the three clusters would the 9th case be assigned by:
[Use squared Euclidean distance as the measure of (dis)similarity for these questions.]

16. Single linkage
17. Complete linkage
18. Average linkage
19. Centroid method
20. Ward's method

Context for questions 21-35:

Please refer now to the sheet Data2 in *HWcluster.xlsx*. Data2 contains data on a random sample of the customers of a Mid-west based chain of clothing stores. The variables are described in the file. The objective is to try to identify market segments of customers to which the chain appeals.

Run the following SAS program, after modifying the first line to suit your set-up:

```
PROC IMPORT DATAFILE="/home/tomsager/CLUSTER/HWcluster.xlxs"
                OUT=customers
                DBMS=XLSX
                REPLACE;
                RANGE='Data2$A14:G288';
RUN;
* PRINT=20 limits the printed history to the final 20 generations;
PROC CLUSTER DATA=customers METHOD=ward STANDARD CCC PSEUDO PRINT=20
OUTTREE=customers_tree;
  VAR EducLevel--SCIndex;
RUN;
PROC TREE DATA=customers_tree OUT=customer_NCl4 NCLUSTERS=4;
  COPY EducLevel--SCIndex;
RUN;
PROC MEANS DATA=customer_NCl4 n mean std;
  VAR EducLevel--SCIndex;
  CLASS cluster;
RUN;
```

Regarding the cluster history, when five clusters became four clusters:

21. The higher number of the two clusters that joined then is Cluster # ___
22. The higher-numbered cluster had how many customers?
23. The lower number of the two clusters that joined then is Cluster # ___
24. The lower-numbered cluster had how many customers?
25. What was the value of "R-Square" after the joining?
26. What was the value of "Semipartial R-square" after the joining?

How many clusters are there? Search for all solutions suggested by each of the pseudo F, pseudo T2, and CCC clustering criteria over the final 15 generations of the cluster history. Include possible one-sided peaks at the boundaries if Gen 1 > Gen 2 and/or Gen 15 > Gen 14. (Also include possible one-sided peaks next to missing values.) After finding all possible suggestions, answer the following:

27. What is the largest number of clusters suggested by the pseudo F criterion? *[Exclude 15 as a possible answer.]*
28. What is the smallest number of clusters suggested by the pseudo F criterion? *[Exclude 1 as a possible answer.]*
29. What is the largest number of clusters suggested by the pseudo T2 criterion? *[Exclude 15 as a possible answer.]*
30. What is the smallest number of clusters suggested by the pseudo T2 criterion? *[Exclude 1 as a possible answer.]*
31. What is the largest number of clusters suggested by the CCC? *[Exclude 15 as a possible answer.]*
32. What is the smallest number of clusters suggested by the CCC? *[Exclude 1 as a possible answer.]*

Which one of the four clusters is a customer most likely to belong to, if that customer is described as:

33. Masters degree, medium salary, medium concern about social status
34. High school diploma, low salary, unconcerned about social status
35. Masters degree, high salary, low savings