Kilger

MKT6971 2 credit course

Practicum I second semester

Exercise #4

Name: Collin Real (yhi267)

This forth exercise is to give you practice at clustering market segments using k means clustering.

You can reuse the code from exercise 3. You will need to add some extra code in the PROC FACTOR procedure to create a temporary SAS file that creates a temporary SAS system file that contains all of your original variables plus the two PCA factor variables that were created in PROC FACTOR.

1. Run your k means cluster analysis using PROC CLUSTER using the temporary file just created above. Use all of your single driver variables plus the two factor variables you created in the PROC FACTOR procedure. Run it each time for k = 3 to k= 8 (that is 6 runs). Be sure to include these runs in your submission.
2. Build a table that contains the results for the k=3 to k=8 runs with values for the following statistics:

|  |  |  |  |
| --- | --- | --- | --- |
| **clusters** | **r\_square** | **ccc** | **psuedo\_f** |
| 3 | 0.25612 | 29.128 | 4922.45 |
| 4 | 0.32182 | 33.858 | 4520.92 |
| 5 | 0.36786 | 34.811 | 4113.62 |
| 6 | 0.40759 | 30.854 | 3803.78 |
| 7 | 0.44005 | 15.327 | 3432.19 |
| 8 | 0.4636 | 19.697 | 3265.71 |

1. Use whatever plotting software you like to plot the following against the number of clusters
2. Examine the CCC plot and tell me what the rule is to suggest how many clusters to keep. How many does it suggest for this k means analysis. Note that it may not suggest any specific number of clusters.

For the CCC plot, we should pick the first local maximum. The local maximum for my CCC plot is k=5 clusters.

1. Examine the Pseudo F plot and tell me what the rule is to suggest how many clusters to keep. How many does it suggest for this k means analysis. Note that it may not suggest any specific number of clusters.

For the Pseudo F plot, we should also pick the first local maximum. The highest value begins at the first cluster (k=3) and slightly declines as a cluster is added; therefore, it does not suggest any specific number of clusters. (Note: if I run k=2 to k=8 (i.e. 7 instead of 6 runs), the pseudo F statistic for k=2 is 3841.14. Since this value is below the Pseudo F value of k=3 (4922.45), k=3 would be the first local maximum.)

1. Examine the means of the driver variables for the solution that is suggested by the CCC plot. Do they look like this might be a good solution? Why or why not? Cut and paste the driver means table for the best solution.

A table with numbers and text

Description automatically generated

Only four cluster combination means (onlineshopper k=1 & k=2, clothes\_last\_long\_time k=3 & k=4, buy\_clothes\_dont\_need k=2 & k=3, and dress\_to\_please\_myself k=1 & k=2) do not satisfy the > 0.1 difference and are a lot closer in value compared to the rest of the variables. However, it looks like most of the variables discriminate and spaced well.

1. Examine the means of the driver variables for the solution that is suggested by the Pseudo F plot. Do they look like this might be a good solution? Why or why not? Cut and paste the driver means for the best solution.

A table with numbers and symbols

Description automatically generated

Three cluster combinations (onlineshopper k=1 & k=2, disc\_clothes\_good k=1 & k=2, and dress\_to\_please\_myself k=1 & k=2) do not satisfy the > 0.1 difference and are a lot closer in value compared to the rest of the variables. However, it looks like most of the variables discriminate and spaced well.

1. If neither the CCC or the Pseudo F plot suggest a best solution, examine the driver means and tell me which one you think would work best. Cut and paste the driver means table for that solution.

N/A