**MKT 6323 Database Marketing**

**Assignment 2**

**Chih-Kai Wang**

**Kuo-Chi Lin**

**Te-Hsuan Lung**

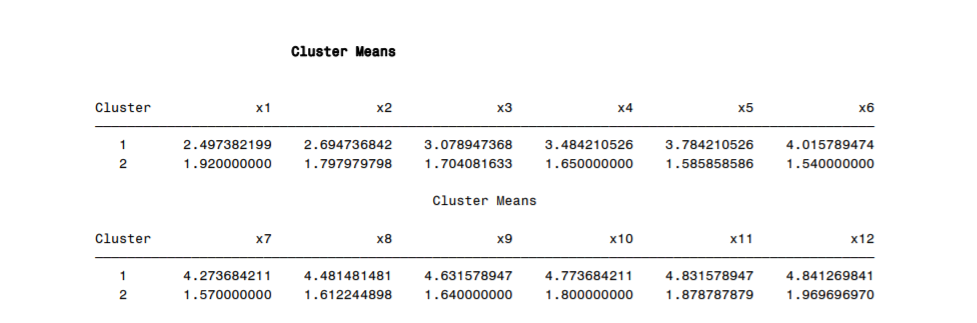
**Wei-Min Huang**

**Yu-Min Wang**

**Oct. 26, 2017**

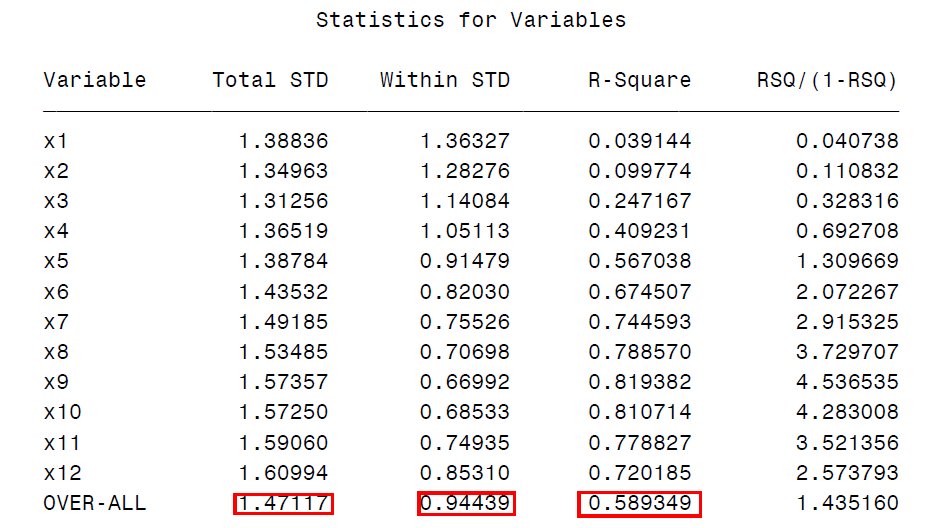
**(i) Based on your results in (a) how would you determine which cluster represents**

**consumers that are users (non-users) of public transportation?**



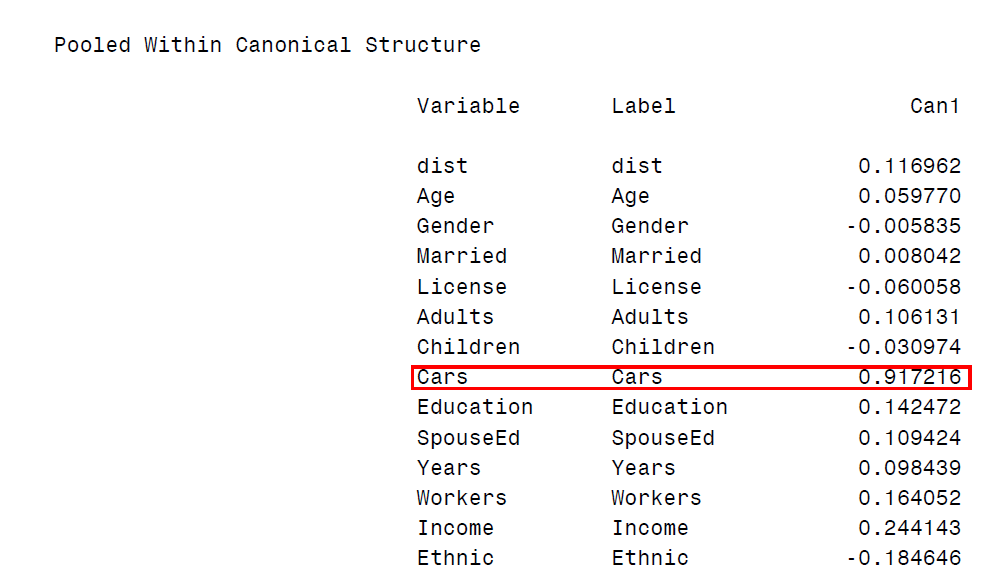
The questionnaire indicates that the selection 1 means using mass transit very frequently while 5 means no use. By looking at the cluster means, cluster 1 shows that the average of each variable is much larger than the average of cluster 2, which can be concluded that cluster 2 represents consumer who uses public transportation frequently, while cluster 1 represents non-user of public transportation.

**(ii) What criteria would you use to assess the goodness of your segmentation in (a)?**



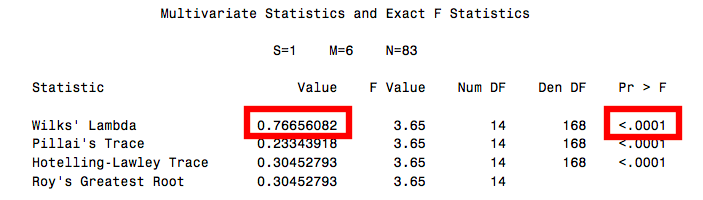
We use 2 criteria to assess the goodness of the segmentation. First, we look at the total R Square, the more RSQ is closer to 1, the better the segmentation is. Here the value of over-all R- Square is 0.589349 while the R-Square of some variables is larger than 0.7. Second, we look at the ratio of within-STD to total STD. The closer the ratio of within-STD to total STD is to zero the better the segmentation is. Here the ratio is 0.94439/1.47117 = 0.6419.

**(iii) Based on your results in (b) which demographic variables help discriminate potential users of public transportation from non-users. Why?**



Based on the results of (b), it can be observed that the variable “Cars” helps discriminate potential users of public transportation from non-users. We examine the standardized canonical coefficient in Pooled Within Canonical Structure table. The closer the value of Can1 is to 1, the better discrimination. Since the variable will discriminate the two clusters effectively. Variable “Cars” has the significantly highest standardized canonical coefficient, which is 0.917216. This means that “Cars” is the most important discriminant variable.

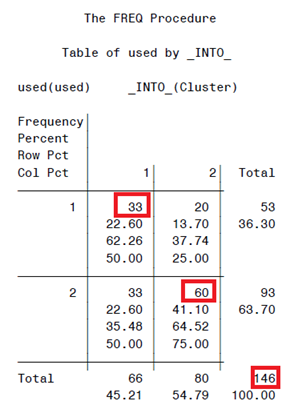
**(iv) How good is your discriminant analysis classification – what criteria would you use to ascertain this?**



The Wilk`s Lambda is used to assess the discriminant analysis classification. Due to lower P-value of Wilk`s Lambda(<0.001), the value of Wilk`s Lambda is significant. Hence, we can conclude that this classification is not good because of the high value of Wilk`s Lambda(0.766).

**(v) Comment on the cross-tabulation in (c). What does that tell you about the**

**effectiveness of your classification procedure.**



It indicates that the proportion of correct clustering is (33+60)/146 = 63.7% by adding actual consumers not using mass transit who are classified as cluster 1, 33 samples, and actual consumers using mass transit who are classified as cluster 2, 60 samples, then dividing by total samples ,146 samples.