*Brad Tweardy*

*April 29, 2016*

*MSDS 6372 – 401*

*Project 3 – Cluster Analysis*

**Introduction**

The data being analyzed contains data from a study by Richard Hernstein and Charles Murray. Their study was originally meant to prove that general intelligence is a better indicator of life success rather than one’s education and/or socioeconomic background. In this analysis, I will perform a cluster analysis in order to find a good description of the data with regards to a survey taken by the individual subjects who rated their self-esteem in 2006. These same subjects had data taken about them in the late 1970’s and early 1980’s as part of this dataset.

**Descriptive Statistics**

The data comes from a study in which psychologist Richard Hernstein and political scientist Charles Murray had disputed that intelligence was a better predictor to future success than was a person’s education and family’s socioeconomic status. They published their findings in the book *The Bell Curve: Intelligence and Class Structure*.

The data includes results from 2,584 Americans selected by the National Longitudinal Study of Youth in 1979. It included test scores of individuals from the Armed Services Vocational Battery of Tests in 1981, their socioeconomic status (family class, education, and family income) in 1979, personal demographics, and variables related to life success from 2006. The measures of life success was income in 2005 and ten self-esteem measures from a survey taken by each subject in 2006. For the purpose of this study, we will look at the self-esteem factors of success as they relate to education. The self-esteem measures are listed below:

1: I am a person of worth.

2: I have a number of good qualities.

3: I am inclined to feel like a failure.

4: I do things as well as others.

5: I do not have much to be proud of.

6: I take a positive attitude towards myself and others.

7: I am satisfied with myself.

8: I wish I could have more respect for myself.

9: I feel useless at times.

10: I think I am no good at all.

Each individual scored the ten measures of self-esteem listed above in the following way:

1: Strongly Agree

2: Agree

3: Disagree

4: Strongly Disagree

Figure 1 shows the descriptive statistics of the data on the left and the eigenvalues for each self-esteem measure on the right.

*fig. 1*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Descriptive Statistics** | | | | | **Eigenvalues of the Covariance Matrix** | | | |
| **Variable** | **Mean** | **Std. Dev.** | **Skewness** | **Kurtosis** | **Bimodality** | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **Esteem1** | 1.4261 | 0.519 | 0.6473 | -0.3962 | 0.5442 | 1.8601487 | 1.351108 | 0.4864 | 0.4864 |
| **Esteem2** | 1.4234 | 0.5119 | 0.5784 | -0.6866 | 0.576 | 0.5090406 | 0.179015 | 0.1331 | 0.6195 |
| **Esteem3** | 3.5101 | 0.5695 | -0.8449 | 0.8376 | 0.4462 | 0.3300259 | 0.043413 | 0.0863 | 0.7058 |
| **Esteem4** | 1.5704 | 0.5555 | 0.3845 | -0.1852 | 0.4073 | 0.2866133 | 0.062875 | 0.0749 | 0.7807 |
| **Esteem5** | 3.4524 | 0.6597 | -1.2657 | 2.2118 | 0.4989 | 0.2237384 | 0.065092 | 0.0585 | 0.8392 |
| **Esteem6** | 1.62 | 0.5617 | 0.3122 | -0.0668 | 0.3737 | 0.1586467 | 0.014228 | 0.0415 | 0.8807 |
| **Esteem7** | 1.7496 | 0.5991 | 0.3026 | 0.1729 | 0.3436 | 0.1444186 | 0.004383 | 0.0378 | 0.9184 |
| **Esteem8** | 3.1289 | 0.7631 | -0.6296 | 0.0917 | 0.4511 | 0.1400355 | 0.023248 | 0.0366 | 0.9551 |
| **Esteem9** | 3.1618 | 0.7284 | -0.5715 | 0.0486 | 0.4346 | 0.1167878 | 0.061703 | 0.0305 | 0.9856 |
| **Esteem10** | 3.3936 | 0.6616 | -0.9894 | 1.2701 | 0.4631 | 0.0550844 |  | 0.0144 | 1 |

**Analysis**

To begin the analysis, I looked at several different clustering methods including complete, centroid, average, and Ward’s. Ultimately, I decided to use Ward’s hierarchical clustering strategy. This is an agglomerative clustering method that begins with each feature in its own cluster and then the closest features by distance merge until we get the optimal number of clusters to describe the data. Each time an agglomeration occurs, the distances are recalculated. After looking at the complete, centroid, average, and Ward’s strategies it was clear that Ward’s method presented the cleanest description of the data.

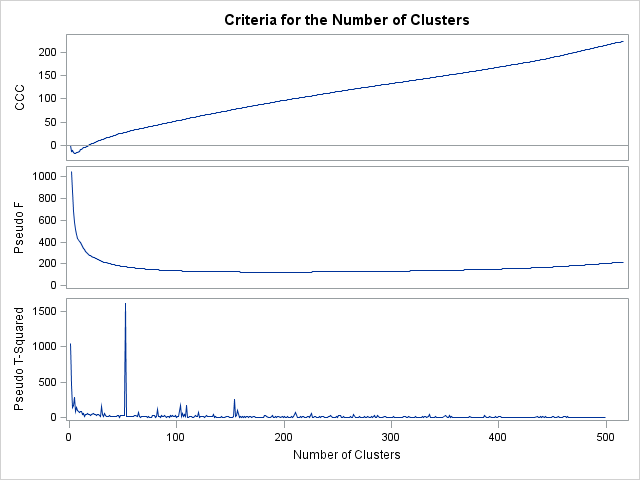
A root-mean-square standard deviation was computed to be 0.618. This was a relatively low value and tells us that there is homogeneity among the clusters, which is what we want. To follow this statistic, I used Ward’s method to analyze the dataset for the most optimal number of clusters to explain 70% of the variance. This came out with 26 clusters (*q = 26*). A Tree Diagram displaying the clusters can be seen in Appendix1. After determining the number of clusters, it was important to merge the original dataset with the new dataset in order to perform an analysis of variance on each self-esteem measure. We do so by comparing the means for each esteem measure across clusters. While doing so, we apply the Bonferroni method in order to control the experiment-wise error rate. It is important to note that our hypothesis statement states:

*There are equal means across clusters at a 95% confidence level with a p-value of less than 0.005.*

Since we have 10 different measures of self-esteem, we can set our *p* = 10 and α =0.05. Using the Bonferroni method, we use the formula α/*p* to determine the rejection level of the p-value. In this case, we set the rejection level at a p-value < 0.005.

In Figure 2 below, I display a graph of the *Criteria for the Number of Clusters*. This compares the Cubic Clustering Criterion, Pseudo F, and Pseudo T-Squared statistics. We can see that the elbow for each of the three statistics falls around 26 clusters. This confirms the number of clusters chosen. For the final computed model, the corresponding values for each cluster is displayed in Appendix 3.

*fig. 2*



After performing our analysis of variance, I displayed the F-Value, p-Value, and for each self-esteem measure in Figure 3.

*fig. 3*



**Conclusion**

To conclude the analysis, it may be best to do so by looking at the statistics in Figure 4. Here I have displayed the mean averages of the Cluster Analysis Tree in order to interpret the level of self-esteem within each cluster. Highlighted in yellow is the cluster (or group of clusters if a tie) with the highest mean average. The highlighted cells indicate a stronger presence of that self-esteem measure within each cluster.

*fig. 4*



A few observations may be worth noting:

*Cluster 2*: Esteem 8, 9, 10 have a tie which may indicate a strong relationship among these measures. With an understanding of what each of these measure, we can say that that would relate to a high level of self-confidence is present within the subjects categorized to this cluster.

*Cluster 11*: Esteem 9 and 10 have the strongest presence in this cluster. This again may display a high level of self-confidence and self-worth within this cluster.

*Cluster 26*: Esteem 3 and 5 have the strongest presence in this cluster. We can deduct from these results that a high level of confidence in one’s abilities and successes is strong in this cluster.

Only one of the five measures that represent a positive level of self-esteem (Esteem 7) is most prevalent in any of the clusters, and even that only happened one time (cluster 20).

**Appendix 1**

***SAS Code***

Title "Cluster Analysis - Career Success - Hiearchial Method";

**proc** **import** out = demo2

datafile = 'C:\Users\Brad\_2\Desktop\Data for Cluster Analysis Happiness Indicator.csv'

dbms = csv replace;

getnames=yes;

datarow = **2**;

**run**;

**data** demo2;

set demo2(keep= Esteem1 Esteem2 Esteem3 Esteem4 Esteem5 Esteem6 Esteem7 Esteem8 Esteem9 Esteem10);

cid = \_n\_;

**run**;

**proc** **sort**;

by cid;

**run**;

**proc** **print** data=demo2;

**run**;

ods graphics on;

**proc** **cluster** method=ward outtree=cluster1;

var Esteem1 Esteem2 Esteem3 Esteem4 Esteem5 Esteem6 Esteem7 Esteem8 Esteem9 Esteem10;

id cid;

**run**;

ods graphics off;

goptions vsize=**9**in hsize=**6.4**in htext=**.9**pct htitle=**3**pct;

axis1 order=(**0** to **1** by **0.2**);

title'"Effect of the Level of Education on Happiness"';

title'Cluster Analysis Tree Diagram';

**proc** **tree** horizontal level=**0.7** out=cluster2;

height \_rsq\_;

id cid;

**run**;

**proc** **sort**;

by cid;

**run**;

**proc** **print**;

**run**;

**data** combine;

merge demo2 cluster2;

by cid;

**run**;

**proc** **glm**;

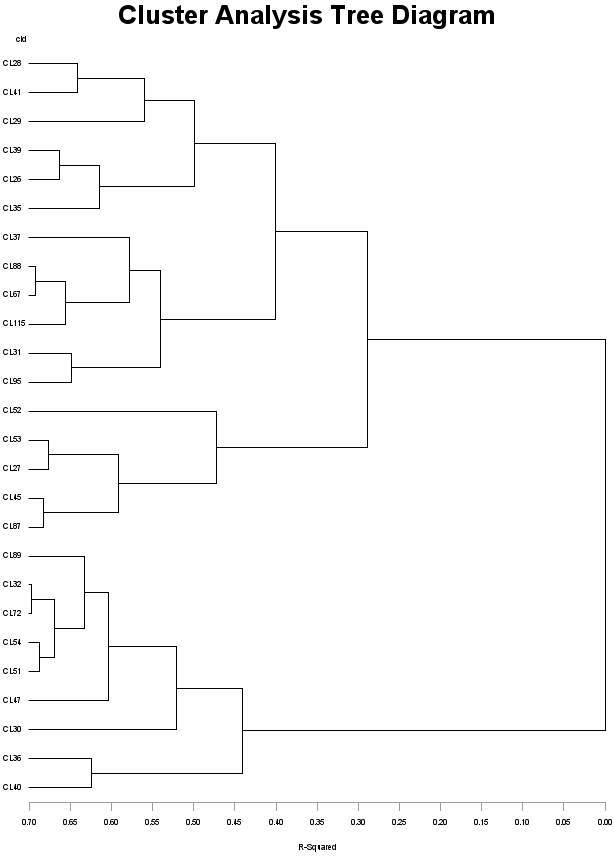
class cluster;

model Esteem1 Esteem2 Esteem3 Esteem4 Esteem5 Esteem6 Esteem7 Esteem8 Esteem9 Esteem10 = cluster;

means cluster;

**run**;

**Appendix 2**

***Cluster Analysis Tree Diagram***

**Appendix 3**

***Cluster History***

