**Multivariate Analysis**

**of Louisiana and Mississippi Parishes.**

**A picture containing text, map

Description automatically generated**

Jorge Lopez Gama

Natalia Mazariegos

Patrick Parham

Oje Williams

**STA 9705**

**Zicklin School of Business**

**Baruch College**

**Spring 2020**

**Table of Contents**

1. **Objective of Analysis…………………………………………………..….. 3**
2. **Data Description ………………………………………………………..… 3**
3. **Multivariate Analysis of Variance ……………………………………..... 3**
4. **Multivariate Regression ………………………………………………….. 6**
5. **Linear Discriminant Analysis ……………………………………………. 9**
6. **Cluster Analysis ………………………………………………………….. 11**
7. **Conclusion ………………………………………………………………... 13**
8. **Appendix …………………………………………………………………...14**
9. **SAS CODE ……………………………………………………………… 20**
10. **Objective of Analysis.**

The purpose of this multivariate analysis is to study the relationship among different attributes that provide information about parishes in two states, Louisiana and Mississippi. It will be analyzed whether there is a significant difference between these two states in terms of their life expectancy and median household income. Furthermore, among various predictors the most important ones will be determined to obtain prediction values of life expectancy and median household income. In addition, it will be analyzed whether there are patterns and relationships between several variables simultaneously.

1. **Data Description.**

The initial data set contains following variables:

* Parish name
* State (LA or MS)
* Life expectancy (in years)
* Population (in numbers)
* Population Under 18 (percentage)
* Population\_African-American (percentage)
* Population\_Female (percentage)
* Population\_Rural (percentage)
* Median Household Income (in dollars)
* Graduation\_Rate (percentage)
* Some\_College\_Rate (percentage)
* Child Poverty Rate (percentage)
* Income Ratio 80pt20pt (ratio)
* Food Insecurity Rate (percentage)
* Uninsured Adults (percentage)
* Average Healthcare Cost (in dollars)
* Physically Inactive Rate (percentage)
* Excessive Drinking Rate (percentage)
* Teen Birth Rate (percentage)

There are total of 146 observations and 17 variables. The group LA contains 64 observations and 82 observations are in group MS. The data has been extracted from CountyHealthRankings.org

The preview of the data set is available in Appendix A.

1. **Multivariate Analysis of Variance.**

**ONE-WAY MODEL**

The one-way multivariate analysis of variance will allow for testing of differences between Louisiana and Mississippi county data for our dependent variables, Life Expectancy and Median Income.

For this data set p = 2, k = 2 and n = 146.

The one non-zero eigenvalue CodeCogsEqn (15).gif = 0.151 (the output is presented in the Appendix B).

The null hypothesis for the test is CodeCogsEqn.gif vs CodeCogsEqn (1).gif: inequality between the states.

**Wilks’ test statistic:**

CodeCogsEqn (2).gif = 1/(1+0.151) = 0.869

The critical value CodeCogsEqn (2).gif 2,1,290 (0.05) > CodeCogsEqn (3).gif2,1,240 (0.05) = 0.869 < 0.975, so the null hypothesis is rejected.

***Note Before Moving on To Other Tests***

When CodeCogsEqn (4).gif = 1 ( k = 2), all four test statistics are equivalent and a two-sample Hotelling’s CodeCogsEqn (5).gif statistic can be used.

Additionally, when s = 1, the test values are calculated as follows:

CodeCogsEqn (6).gif

CodeCogsEqn (7).gif

Therefore:

CodeCogsEqn (8).gif

CodeCogsEqn (10).gif

CodeCogsEqn (11).gif

**Pillai’s test statistic:**

CodeCogsEqn (13).gif = 0.151/(1+0.151) = 0.131

s = 1, m = 0, N = 144

Use Hotelling’s CodeCogsEqn (5).gif

The critical value is F Val 2,144 (0.05) < F Val 2, 120 = 10.810 > 6.106, so the null hypothesis is rejected.

**Hotelling-Lawley Trace test statistic:**

CodeCogsEqn (14).gif= CodeCogsEqn (15).gif = 0.151

CodeCogsEqn (16).gif = 1 p = 2 CodeCogsEqn (17).gif = 290

Use Hotelling’sCodeCogsEqn (5).gif

The critical value is F Val 2,144 (0.05) < F Val 2, 120 = 10.810 > 6.106, so the null hypothesis is rejected.

**Roy’s test statistic:**

= CodeCogsEqn (15).gif = 0.151 and s = 1, m = 0, N = 144



Use Hotelling’s CodeCogsEqn (5).gif

The critical value is F Val 2,144 (0.05) < F Val 2, 120 = 10.810 > 6.106, so the null hypothesis is rejected.

The output for all four tests is presented in the Appendix C.

**Assessment**

All of the four of the tests reject the null. This means that Louisiana and Mississippi counties are different in terms of Life Expectancy and Median Income.

Further, as there is only one non-zero eigenvalue, that value is dominant at 100%. Therefore, the dimensionality is 1. Additionally, there are only two groups, there is no need to assess for contrasts.

**TWO-WAY MODEL**

The populations of the counties might heavily influence the various analyses. Before looking into the population effect more closely, it might be of interest to set an arbitrary threshold for comparison. Therefore, a two-way multivariate analysis of variance test will be conducted for difference between counties with populations over one-hundred thousand citizens and the state effect for the dependent variables, Life Expectancy and Median Income. Additionally, it can also be tested for an interaction effect in this fixed experience for the two factors. This will tell if the effect of one factor is independent of the influence of the other factors.

Below are the null hypotheses for the two main effects (state and ‘population greater than 100k’) and one interaction:

CodeCogsEqn (18).gif



CodeCogsEqn (20).gif

For state (state\_num) effect, CodeCogsEqn (2).gif = .990, CodeCogsEqn (13).gif= 0.010, CodeCogsEqn (14).gif = 0.010 and CodeCogsEqn (12).gif = 0.010.

Per the program output, the p-values (0.494) are below 0.05 for the four tests. The null hypothesis is rejected. It can be concluded that the state effect is significant as seen in the one-way model above.

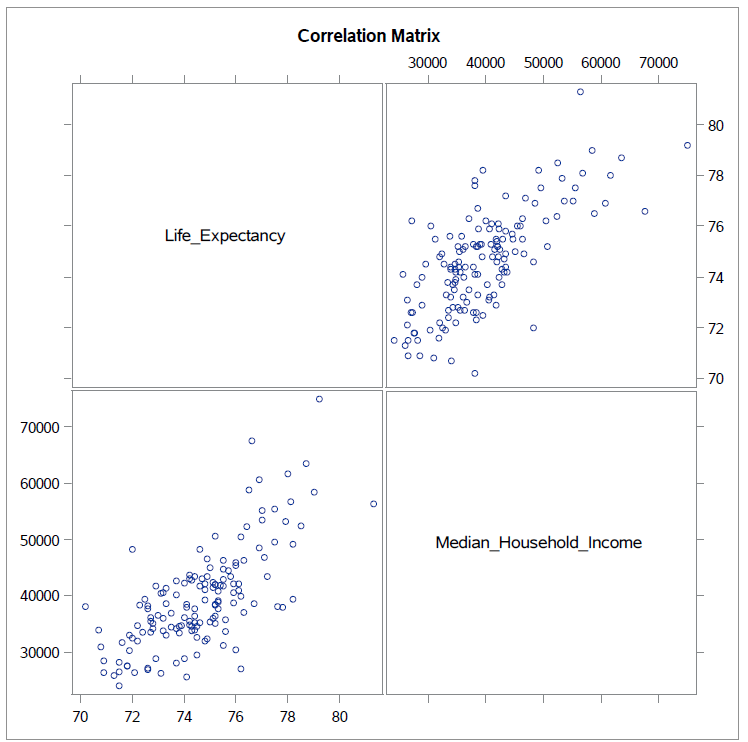
For ‘population greater than 100k’ (pop) effect, Wilks CodeCogsEqn (2).gif = .761, CodeCogsEqn (13).gif= 0.239, CodeCogsEqn (14).gif = 0.315 and CodeCogsEqn (12).gif = 0.315. Per the program output, the p-values are all rather small and below 0.05 for the four tests. The null hypothesis can be rejected. It can be concluded that the ‘population greater than 100k’ effect is significant.

For interaction (state\_num \* pop) effect, Wilks CodeCogsEqn (2).gif = .971, CodeCogsEqn (13).gif= 0.029, CodeCogsEqn (14).gif = 0.030 and CodeCogsEqn (12).gif = 0.030. Per the program output, the p-values are all significantly above 0.05 for the four tests. Therefore, null hypothesis can’t be rejected for interaction effect. The interaction effect is insignificant.

SAS Output for all 3 tests is presented in Appendix D.

1. **Multivariate Regression**

From the data gathered, the response variables chosen were Life Expectancy and Median Household Income. Before multivariate regression was decided upon as a possible analysis method, a correlation plot and table between the two response variables were done in order to check for any linearity. The plot can be seen below. The SAS output is presented in Appendix E.

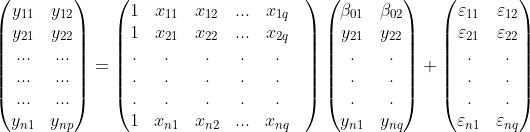


It can be seen from the graph on the left as well as the table on the right that the correlation between the two responses was high. It was more efficient to conduct a multivariate approach rather than a univariate approach which would have been done instead if the correlation between the two was small. Another correlation table was produced also. This time both the predictors and the responses were used. Only the predictors that had an absolute correlation value greater than 0.5 were kept as seen below and then used in conducting the tests.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** |
| Population\_Under\_18 | Population\_African\_American | Child\_Poverty\_Rate | Food\_Insecure\_Rate | Uninsured\_Adults\_Prev | Physically\_Inactive\_Rate | Excessive\_Drinking\_Rate, | Teen\_Birth\_Rate |



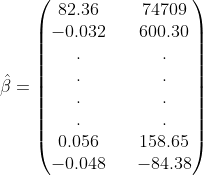
First, the model for the Multivariate Regression has the form:

Which in matrix form can be represented as follows:

**Test of Overall Regression**

The test of overall regression seeks to obtain information on whether there is a linear association between the response variables and the set of predictor variables.

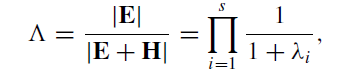
For this data set, n = 146 (number of observations), p = 2 (predictor variables) and q = 8 (number of response variables). Consequently, VH = 8, VE = 137 (n-q-1), s=2, M=2.5, N =67.

 Before the testing of the hypothesis, the estimated coefficients were obtained for the model. Usually, there is a data set with a collection of observations and with no knowledge of the coefficients β0, β1,…,βk. These can be estimated from the data using the least squares principle which minimizes the solution and gives the least value for the sum of squared errors. From the output, a partial list of the least squares estimate was:

**Hypothesis Test**

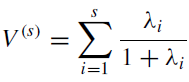
H0: B1 = 0 vs Ha: B1 ≠ 0

Partial SAS output is presented in the Appendix F.

**Wilk’s Test**

Test statistic Λ = 0.0925 (from the table) and the critical value Λ0.5(2,8,137) > Λ0.5(2,8,120) = 0.807.

Since Λ is < Λ.05(2,8,120), *the null hypothesis is rejected.*

**Pillai Test**

Test statistics V(s) = 1.135 (taken from the table) and the critical value V.05(2,3,80)> V.05(2,3,25) = 0.445

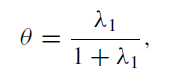
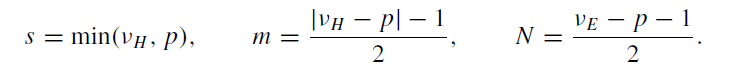
Since V(s) > V.05(2,3,80), *the null hypothesis is rejected*.

**Lawley-Hotelling Test**

Test statistics is (VE/VH) U(s) , where U(s) = 7.35 (taken from the table) and therefore

(VE/VH)U(s) = (137/8)\*7.35 = 125.87 and the critical value is U.05(2,8,137) > U.05(2,8,100) = 3.48

Since U(s) > U.05(2,8,100), *the null hypothesis is rejected.*

**Roy’s Test**

Test statistic θ = (6.99/1+6.99) = 0.875 and the critical value θ.05(2,3,80) = .119.

Since θ > θ.05(2,3,80), *the null hypothesis is rejected.*

Based on the four tests conducted at the α = .05 level, in which all rejected the null hypothesis,it can be concluded that the linear regression between the responses, life expectancy and median household income and the predictors is a satisfactory model.

With the presence of strong linear associations between the responses and the predictors, some tests on subsets of the data can be performed for further analysis.

**Tests of Significance**

 Another area of analysis that can be now attempted is to examine if different subsets of the data will prove to be significant on the responses. When testing subsets of the data the B matrix is split in two and now becomes where Bd contains the last h (the number of variables being subset) rows of B.

****This allows for Xr to contain the last columns in X for Br leading to compare the reduced model to the full model i.e.

In doing these tests of significance, the null and alternate hypothesis take on the form:

H0: Bd = 0 vs Ha: Bd ≠ 0.

First, the significance of each predictor on every other predictor is being tested. The results from the SAS output is presented in Appendix G.

In this case, each of the individual tests has VH = h = 1. This means that each of the four test i.e Wilk’s, Pillai’s, Lawley-Hotelling and Roy’s, are equivalent to each other and they also have exact F transformations. Therefore, by looking at the p-value in each partial test it can be seen that all of the predictors were significant except Excessive Drinking Rate(x7) after being adjusted for the other predictors individually.

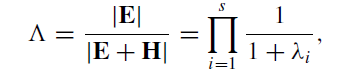
Next, the larger subset such as x1,x2 adjusted for the other predictors will be tested, Keeping in mind that n = 146 p = 2 and VH = 2 (based on subset), VE = 137, s=2, M=0.5, N =67.

**Hypothesis Test**

H0: Bd = 0 vs Ha: Bd ≠ 0 where Bd contains the slopes related to x1 and x2.

The SAS Output for these tests are presented in Appendix H.

**Wilk’s Test**

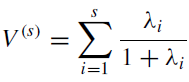


Test statistic Λ = 0.802 (from the table) and the critical value

Λ0.5(2,2,137) > Λ0.5(2,2,120) = 0.924.

Since Λ is < Λ.05(2,2,120), *the null hypothesis is rejected*.

**Pillai Test**

****

Test statistics = .202 (from the table) and the critical value

V.05(2,0.5,67)> V.05(2,1,25) = 0.304.

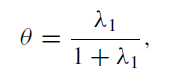
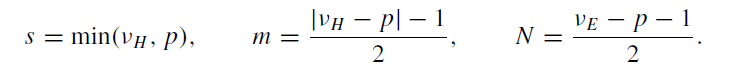
Table A.11 of the Methods of Multivariate Analysis does not have values for Pillai when N is greater than 25. Therefore, the exact p-value is used to conclude the test. Since the Pillai’s p-value as seen in the table above is less than α = .05, *the null is rejected*.

**Lawley-Hotelling Test**

Test statistics is (VE/VH) U(s) , where U(s) = .2403 (taken from the table) and therefore

(VE/VH)U(s) = (137/2)\*.2403 = 16.46 and the critical value is U.05(2,2,137) > U.05(2,2,100) = 4.96

Since U(s) > U.05(2,2,100), *the null hypothesis is rejected*.

**Roy’s Test**

Test statistic θ = (0.213/1+0.213) = 0.176 and the critical value θ.05(2,1,80) = 0.085.

Since θ > θ.05(2,1,80) the null hypothesis is rejected

Based on the four tests conducted at the α = .05 level, in which each test rejected H0 , it can be concluded that the predictors *Population Under 18(x1) and Population African American(x2) are significant predictors when adjusted for the other predictors*.

1. **Linear Discriminant Analysis.**

# Objective

Linear Discriminant Analysis allows to describe the relative separation of the two groups using a linear combination of variables to project the two groups onto the dimension which maximizes the separation between the two groups. This technique also allows to assess the contribution of each variable to the separation of the groups. Linear discriminant analysis does assume that the observations of the two groups are multivariate normal and have equal covariances, and if these conditions are not met then it does a poor job in describing the separation between the two groups.

As it was already observed from the MANOVA tests results done previously, the LA and MS counties have different mean vectors with regards to at least one of the tested variables, Linear Discriminant Analysis is a natural extension to further describe the differences between the two groups.

# 

# Method

In order to maximize the seperation between the LA and MS counties we will transform the observation vectors into a single vector where

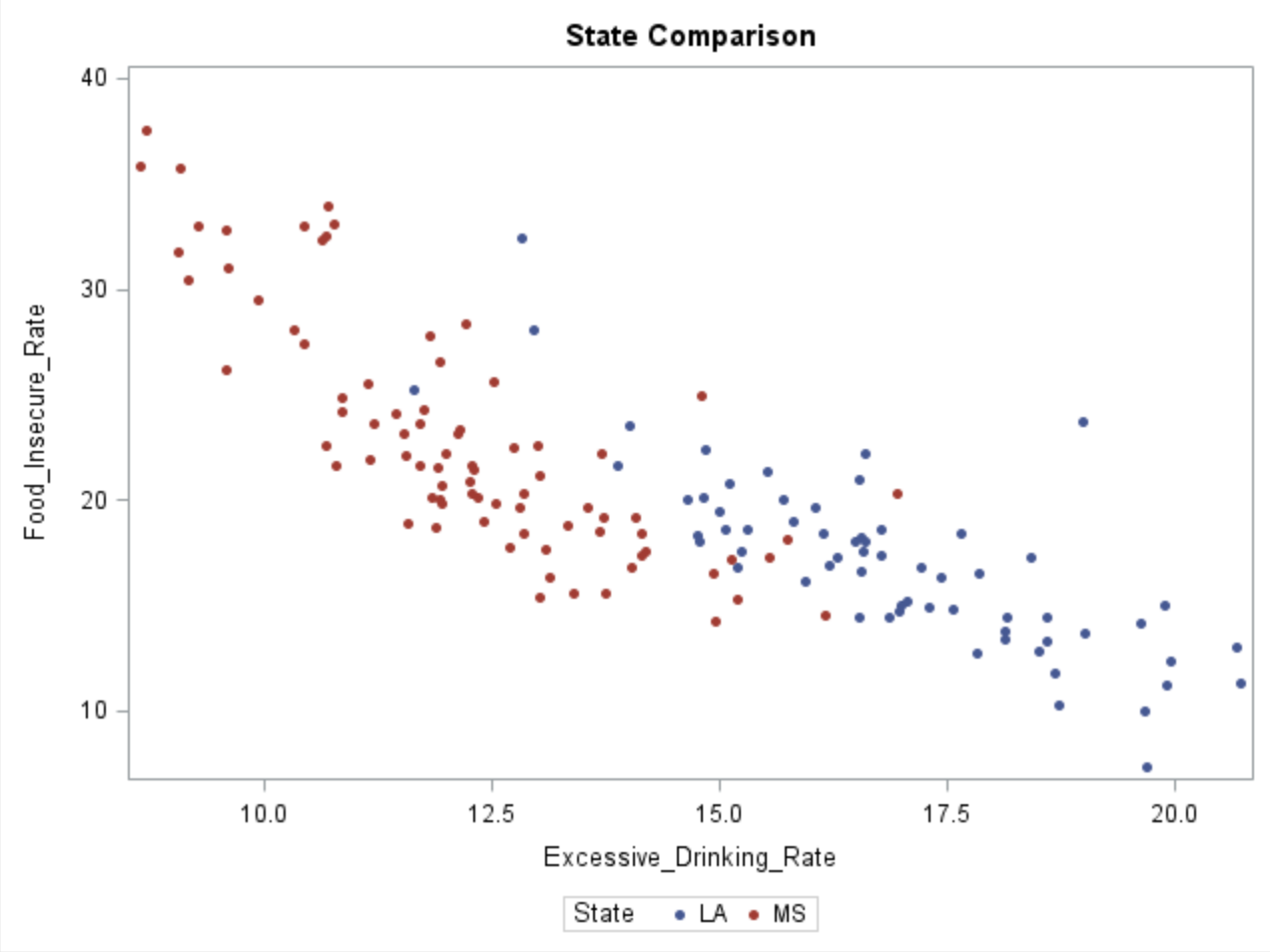
and then find the vector of discriminant function coefficients (*a*) that maximizes the standardized distance which is given by the formula

where is the mean vector of the LA observations and is the mean vector of the MS observations.

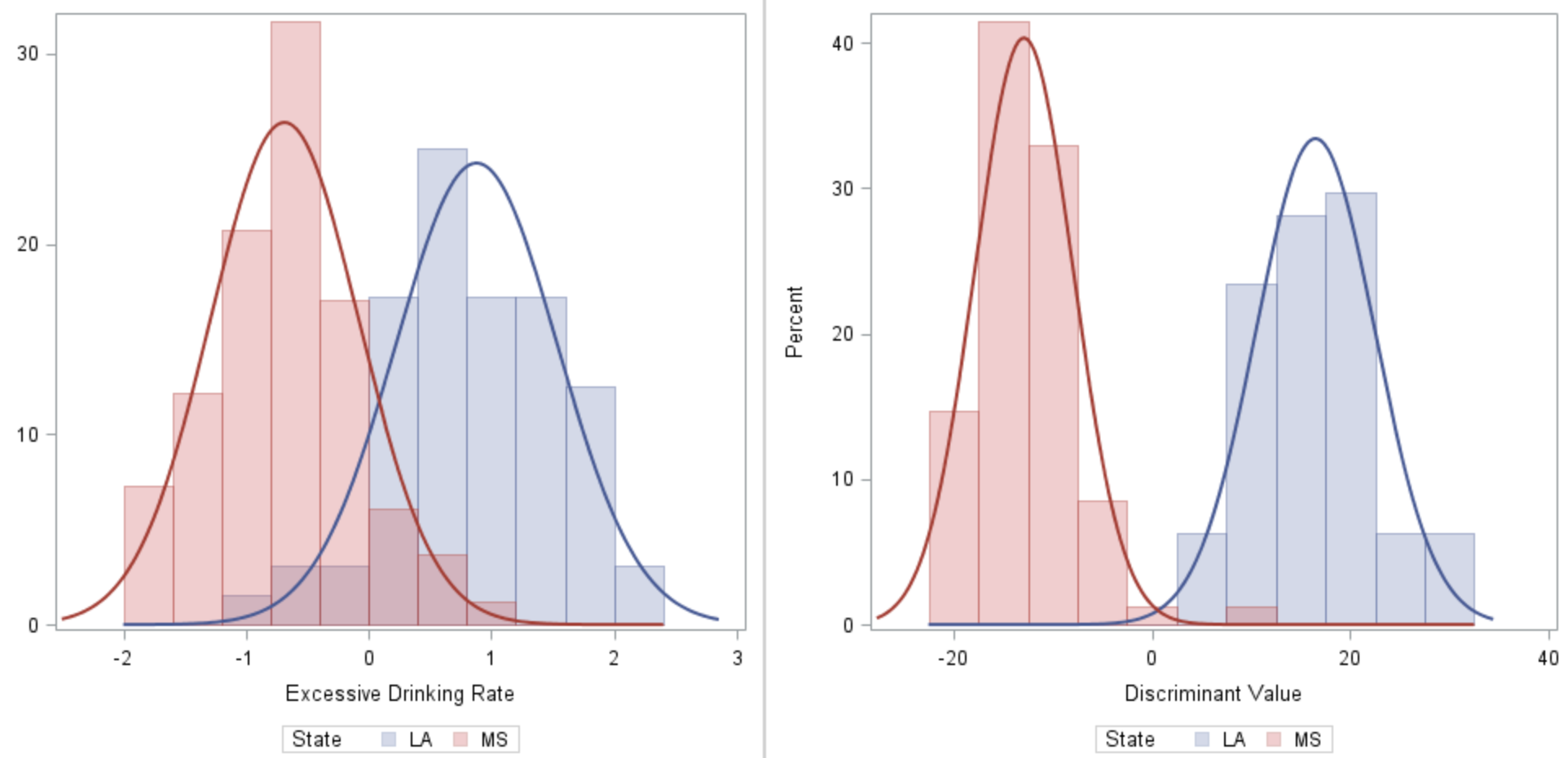
# Assessment

By examining the standardized discriminant function coefficients ranked by absolute value each variable’s individual contribution to the separation between the two groups can be observed.

Excessive Drinking Rate and Food Insecure Rate are the two variables that contribute the most to the separation of the two groups, and by examining a scatterplot of the two variables it can be seen that there is indeed a good amount of separation but there is still some overlap.



When looking at the distribution of the discriminant values after applying the discriminant function and compare it to the single variable that has the most separation, it can be seen that the two groups, except for one single observations, are separated very well after the linear discriminant function.



**Tests of Significance**

From the previous MANOVA test, sufficient evidence was already obtained to believe that the counties of LA and MS have significantly different mean vectors in at least one of the variables used in linear discriminant analysis. Since the test of significant for linear discriminant analysis is equivalent to this MANOVA test, there is no need to carry it out and it can be said that the discriminant function gives a significant separation to the two groups.

1. **Cluster Analysis.**

**OBJECTIVE**

In cluster analysis, that has also been referred to as classification and pattern recognition (specifically, unsupervised learning), patterns in a data set are searched by grouping the observations into clusters. The goal of such analysis is to find an optimal grouping for which the observations within each cluster are similar, but the clusters are dissimilar to each other. In cluster analysis, the number of groups are not known in advance. The similarities are based on some measure of distance in many cases. In others, cluster methods use a preliminary choice for cluster centers or a comparison of within- and between-cluster variability.

**METHOD**

In this part of the analysis, the goal is to group parishes by similarity based on the variables that were determined to be significant in the multivariate regression part. As a reminder, those are as following:

Y1 = median household income

Y2 = life expectancy

X1 = population under 18

X2 = population African-American

X3 = child poverty

X4 = food insecurity rate

X5 = uninsured adults rate

X6 = physically inactive rate

X7 = excessive drinking

X8 = teen birth rate

Also, since some variables are on vastly different scales (i.e Median Household income), the data set will be standardized.

At first approach, since the number of groups are unknown, hierarchical clustering with average link to determine it will be used. Then, the k-means with the numbers of centroids that were determined in step one will be used to analyze if the clustering of parishes is effective.

**ASSESMENT.**

Step 1. Hierarchical Clustering / Average Link /

Partial output of cluster history is presented in Appendix I.

To determine the number of centroids for the next step, the alpha method to calculate the cut off value will be used:

The average of the distances is 0.454051, and the standard deviation is 0.19846, and the k=2.5 is chosen. Thus, the cut off value is 0.950201. Consequently, there will be 4 centroids.

Step 2. The k-means method with 4 centroids was performed. The canonical discriminant analysis of the resulting clusters by examining the plot and the rest of output doesn’t seem to be effective. There are no clear division as it can be observed in the following plot:

A picture containing large, white, table

Description automatically generated

Step 3. At this step, different cluster separations by k-means approach were performed with increasing and decreasing number of seeds, and also testing both approaches, random seeds and first k-seeds.

A close up of a map

Description automatically generated The most noticeable separation of the parishes was observed by using 3 centroids and 3 first observations as seeds.

However, even though 3 groups can be observed, there is no significant distance between the groups themselves as it is desired when perform clustering analysis.

The more detailed separation of the parishes and their assignment to the groups can be seen in the Appendix J.

1. **Conclusion**

Based on different methods of Multivariate Analysis, it has been determined that there is significant difference between Louisiana and Mississippi parishes in terms of life expectancy and median income.

It also has been established that the most difference can be explained by counties where the population is greater than 100K. Furthermore, the most important 8 predictors have been found to determine the life expectancy and median income in the parishes. Moreover, it has been evaluated that Population of African Americans and Under 18 are significant for those levels adjusted for remaining 6. In addition, it was determined that Food Insecure Rate and Excessive Drinking Rate contribute the most for two states separation. Lastly, it was found that the groups constructed from all parishes do not vary vastly among each other which can be explained by similar levels of all factors considering they are all located at the same geographical region.

A picture containing large, room, standing, keyboard

Description automatically generatedAPPENDIX A.

APPENDIX B.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedAPPENDIX C.

APPENDIX D.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

APPENDIX E.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedAPPENDIX F.

APPENDIX G.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedAPPENDIX H.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedAPPENDIX I.

A screen shot of a computer

Description automatically generatedA picture containing white, cellphone

Description automatically generatedA close up of a white background

Description automatically generatedAPPENDIX J.

**SAS CODE.**

proc glm data=lamsdata;

class state\_num;

model lifeexp medinc = state\_num;

manova h=state\_num/printe printh;

run;

proc glm data=lamsdata;

class state\_num pop;

model lifeexp medinc = state\_num pop state\_num\*pop;

manova h=\_all\_/MSTAT=exact short;

run;

TITLE 'Correlation Matrix';

PROC SGSCATTER;

MATRIX LIFE\_EXPECTANCY MEDIAN\_HOUSEHOLD\_INCOME;

RUN;

TITLE 'Correlation Table';

PROC CORR;

VAR LIFE\_EXPECTANCY MEDIAN\_HOUSEHOLD\_INCOME;

RUN;

/\* CORRELATION TABLE OF All VARIABLES \*/

PROC CORR DATA = regression;

VAR Life\_Expectancy Population Population\_Under\_18 Population\_African\_American

Population\_Female Population\_Rural Median\_Household\_Income Graduation\_Rate

Some\_College\_Rate Child\_Poverty\_Rate Income\_Ratio\_80pt20pt Food\_Insecure\_Rate

Uninsured\_Adults\_Prev Average\_Healthcare\_Costs Physically\_Inactive\_Rate

Excessive\_Drinking\_Rate Teen\_Birth\_Rate;

RUN;

/\* MULTIPLE REGRESSION \*/

PROC REG Data=regression;

MODEL Life\_Expectancy Median\_Household\_Income = Population\_Under\_18

Population\_African\_American Child\_Poverty\_Rate Food\_Insecure\_Rate

Uninsured\_Adults\_Prev Physically\_Inactive\_Rate Excessive\_Drinking\_Rate

Teen\_Birth\_Rate;

OVERALL: MTEST /PRINT CANPRINT;

Partial1: MTEST Population\_Under\_18/PRINT MSTAT=EXACT;

Partial2: MTEST Population\_African\_American/PRINT MSTAT=EXACT;

Partial3: MTEST Child\_Poverty\_Rate/PRINT MSTAT=EXACT;

Partial4: MTEST Food\_Insecure\_Rate/PRINT MSTAT=EXACT;

Partial5: MTEST Uninsured\_Adults\_Prev/PRINT MSTAT=EXACT;

Partial6: MTEST Physically\_Inactive\_Rate/PRINT MSTAT=EXACT;

Partial7: MTEST Excessive\_Drinking\_Rate/PRINT MSTAT=EXACT;

Partial8: MTEST Teen\_Birth\_Rate/PRINT MSTAT=EXACT;

Partial12: MTEST Population\_Under\_18,Population\_African\_American/PRINT MSTAT=EXACT;

TITLE ' Multiple Regression ';

RUN;

PROC IMPORT datafile='Pt\_new.csv'  
 out=STATES  
 dbms=csv;  
  
PROC STANDARD DATA=STATES MEAN=0 STD=1 OUT=ZSTATES;  
 VAR \_NUMERIC\_;  
RUN;  
  
PROC IML;  
 USE ZSTATES;  
 READ ALL VAR {Life\_Expectancy Population Population\_Under\_18 Population\_African\_American Population\_Female Population\_Rural Median\_Household\_Income Graduation\_Rate Some\_College\_Rate Child\_Poverty\_Rate Income\_Ratio\_80pt20pt Food\_Insecure\_Rate Uninsured\_Adults\_Prev Average\_Healthcare\_Costs Physically\_Inactive\_Rate Excessive\_Drinking\_Rate Teen\_Birth\_Rate} INTO X;  
 X1 = X[1:64,];  
 X2 = X[65:146,];  
 N1 = NROW(X1);  
 N2 = NROW(X2);  
 X1BAR = 1/N1\*X1`\*J(N1,1);  
 X2BAR = 1/N2\*X2`\*J(N2,1);  
 S1 = 1/(N1-1)\*X1`\*(I(N1)-1/N1\*J(N1))\*X1;  
 S2 = 1/(N2-1)\*X2`\*(I(N2)-1/N2\*J(N2))\*X2;  
 Spl = 1/(N1+N2-2)\*((N1-1)\*S1+(N2-1)\*S2);  
 a = INV(Spl)\*(X1BAR-X2BAR);  
 z1 = a`\*X1`;  
 z1 = z1`;  
 z2 = a`\*X2`;  
 z2 = z2`;  
 PRINT X1BAR,X2BAR, a,z1,z2;  
RUN;  
  
title "State Comparison";  
proc sgplot data=STATES;  
scatter x=Excessive\_Drinking\_Rate y=Food\_Insecure\_Rate / group=State   
markerattrs=(size=6 symbol=CircleFilled);  
run;  
  
PROC IMPORT datafile='disc\_lc.csv'  
 out=TRANSFORMED  
 dbms=csv;  
  
ods graphics / width=500px height=500px;  
ods layout gridded columns=2 advance=table;  
  
proc sgplot data=ZSTATES;  
 title "";  
 histogram Excessive\_Drinking\_Rate / group= State transparency=0.7;  
 Density Excessive\_Drinking\_Rate/ group = State ;  
 xaxis label="Excessive Drinking Rate" ;  
run;  
  
proc sgplot data=TRANSFORMED;  
 title "";  
 histogram Value / group= State transparency=0.7;   
 Density Value / group = State ;  
 xaxis label="Discriminant Value" ;  
  
run;  
  
ods layout end;

**proc** **standard** data=states mean=**0** std=**1** out=states;

var y1 y2 x1 x2 x3 x4 x5 x6 x7 x8;

**RUN**;

**proc** **cluster** data=states method=average outtree=ProTree;

var y1 y2 X1 X2 X3 X4 X5 X6 x7 x8;

id parish;

**run**;

**proc** **tree** data=ProTree nclusters=**4** out=newdata noprint;

id parish;

copy y1 y2 X1 X2 X3 X4 X5 X6 x7 x8;

**run**;

**proc** **sort** data=newdata;

by cluster;

**run**;

**proc** **print** data=newdata;

var parish cluster;

**run**;

**proc** **means** data=newdata noprint;

by cluster;

output out=Seeds mean= y1 y2 X1 X2 X3 X4 X5 X6 x7 x8;

var y1 y2 X1 X2 X3 X4 X5 X6 x7 x8;

**run**;

**proc** **fastclus** data=states maxc=**4** maxiter=**50** seed=Seeds out=Clus\_out;

var y1 y2 X1 X2 X3 X4 X5 X6 x7 x8;

id parish;

**run**;

**proc** **sort** data=Clus\_out;

by cluster distance;

**run**;

**proc** **print** data=Clus\_out;

var parish cluster distance;

**run**;

**proc** **candisc** data=Clus\_out noprint out=ProCan(keep=parish cluster Can1 Can2);

class cluster;

var y1 y2 X1 X2 X3 X4 X5 X6 x7 x8;

**run**;

**proc** **gplot** data=ProCan;

plot Can2\*Can1=cluster / vaxis=axis2 haxis=axis1 nolegend;

axis1 label=("z1" justify=center);

axis2 label=("z2" justify=center a=**90**);

**run**;

**proc** **fastclus** data=states maxc=**3** replace=none maxiter=**10** out=Clus\_out radius=**1**;

var y1 y2 x1 x2 x3 x4 x5 x6 x7 x8;

id parish;

**run**;

**proc** **sort** data=Clus\_out;

by cluster distance;

**run**;

**proc** **print** data=Clus\_out;

var parish cluster distance;

**run**;

**proc** **candisc** data=Clus\_out noprint out=ProCan(keep=parish cluster Can1 Can2);

class cluster;

var y1 y2 x1 x2 x3 x4 x5 x6 x7 x8;

**run**;

goptions reset=all;

goptions csymbol=black;

symbol1 v='1';symbol2 v='2';symbol3 v='3';

**proc** **gplot** data=ProCan;

plot Can2\*Can1=cluster / vaxis=axis2 haxis=axis1 nolegend;

axis1 label=("z1" justify=center);

axis2 label=("z2" justify=center r=**0** a=**90**);

**run**;