Risky Behavior: A Multivariate Statistical Analysis of the United States Based on Health Risk Factors

Christina McIntosh, Spelmen College, Atlanta, GA, 30314
Alicia Smith, Winston-Salem State University, Winston-Salem, NC, 27110
Ashley Swandby, Longwood University, Farmville, VA, 23909

0 Abstract

We study a number of variables associated with health risk factors in the United States. We use the 2006 Centers for Disease Control's Behavioral Risk Factor Surveillance System survey data to analyze each state based on these variables. We use Principal Component Analysis, Factor Analysis, and Discriminant Analysis in order to analyze the multivariate data. Furthermore, we provide a ranking of relative health for some of the states based on the analysis.

1 Introduction

Throughout the United States, there is an increasing awareness of individual health. Health care has become a major political issue, and many people attempt to obtain an optimal weight and live a healthy lifestyle. In this study, we use multivariate statistical analysis techniques in MINITAB to assess, state-by-state, the health of the United States and to categorize each state as healthy or unhealthy. We use Principal Component Analysis, Factor Analysis, and Discriminant Analysis to develop and analyze the data. Principal

Component Analysis allows us to reduce the dimensionality of the data set, thus simplifying the process of classification of the states' health. Using Factor Analysis, we are able to detect factors that are not easily measurable and then group the correlated variables. Finally, Discriminant Analysis allows us to classify each state properly into different groups based upon its collective measurements.

In this analysis of health factors by state, we consider variables that have been considered as high risk to one's health. Some of these variables are obesity, exercise, health care access, health insurance, alcohol consumption, tobacco use, poverty levels, mean and median income, and education attainment. Our data is provided by the Behavioral Risk Factor Surveillance System (BRFSS), a survey system monitored by the Centers for Disease Control (CDC) and administered in all fifty states. This data provides a representative sample of each state by the methods enlisted by the survey to include a diverse population in the results, and is obtained through a telephone survey with an attempt to minimize bias.

2 Multivariate Statistical Techniques

2.1 Assessing Normality

The multivariate statistical techniques we use to analyze the data require the assumption that the data set comes from a multivariate normal distribution. In order to make this assumption, we must first verify that the data set is in fact from a multivariate normal distribution. We achieve this by testing normality of each variable. Toward this end, we create a quantile-quantile (Q-Q) plot of each variable in our data and check for the linearity of the Q-Q plot. The Q-Q plot is a plot of the normal quantiles versus the ordered x-variable values. The normal quantiles are computed from the z-scores of the corresponding percentiles of the ordered x-variables. We give an example of a Q-Q plot for the variable *Former Smoker* in Figure 1. If the data comes from a normal distribution, the equation for the regression line is $x_{(i)} = \mu + \sigma z_i$, where $x_{(i)}$ is the observed sample data point and z_i is the corresponding standard normal quantile. Thus, if the plot of our data is linear, we accept the assumption of normality in the variable. However, this test of normality is a visual inspection, so we desire a statistical test of linearity as well.

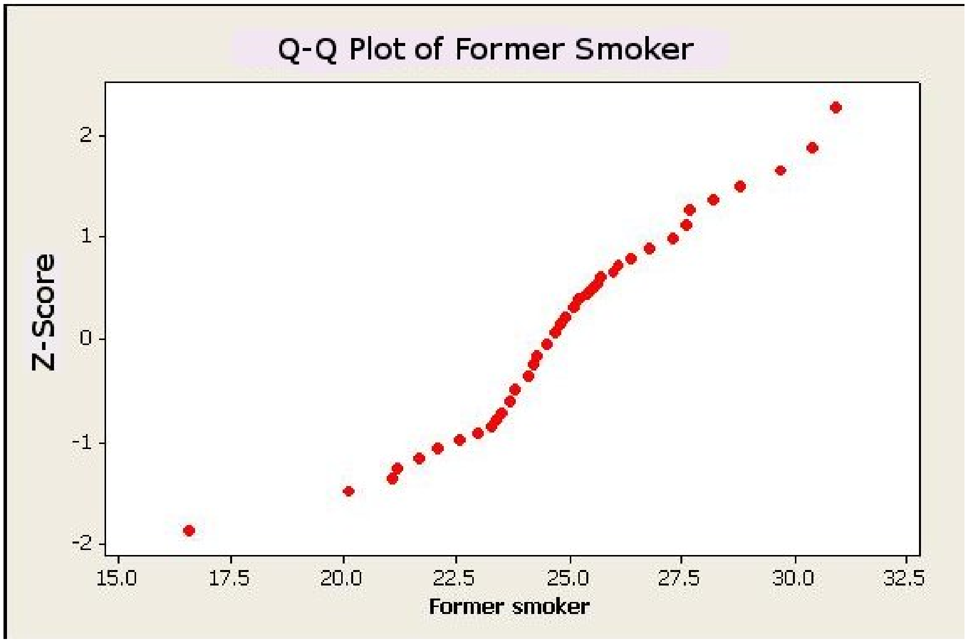

Figure 1: Q-Q Plot of *Former Smoker*

We use the correlation coefficient test for normality to provide us with the evidence of normality in our data. To find the sample correlation coefficient r_q, we compute:

$$0 < r_q = \frac{\hat{\text{Cov}}(x_{(i)}, z_i)}{\sqrt{\hat{\text{Var}}(x_{(i)})\hat{\text{Var}}(z_i)}} = \frac{\sum_{i=1}^{n}(x_{(i)} - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(z_z - \bar{z})^2}} < 1.$$

When we test the the null hypothesis (H_0) that the population is from a normal distribution, we are testing the hypothesis that the population correlation coefficient ρ equals 1. The alternate hypothesis (H_A) is that the population is not from a normal distribution, or that $0 < \rho < 1$. The test is carried out as:

$$H_0 : \rho = 1$$

$$H_A : \rho < 1.$$

Thus, a strong sample correlation coefficient r_q is a value close to 1, representing the strength of the linear relationship between $x_{(i)}$ and z_i. To determine the critical value c for the hypothesis test, we consider the sample size and significance level of the test α. If $r_q > c$, we accept the null hypothesis that the population is normal. For the variables that do not follow a normal distribution, we apply transformations to the data set in order to achieve a normal distribution in each variable of our data.

2.2 Assessment of Normality for BRFSS Variables

For our data set we have a sample size of $n = 51$ representing the 50 states and the District of Columbia. We have 17 variables, a significance level of $\alpha = 0.05$ and a corresponding critical value $c = .9768$ for the correlation coefficient of each variable. Out of the seventeen variables, twelve tests (*heavy drinkers, binge drinkers, High School or G.E.D., $ 15,000-24,999 income, $ 50,000+ income, physical activities in last month, health care coverage, dental visit in last year, teeth extraction in lifetime, neither overweight or obese, obese,* and *smokes everyday* do not reject H_0, the hypothesis that the data comes from a normal distribution. The tests of the five remaining variables (*former smoker, never smoked, less than $15,000 yearly income, college +, and less than high school education*) reject H_0.

To normalize these five variables, we use transformations. We transform the variables in several ways, and calculate the resulting correlation coefficients. In each case we choose the transformation that provides the highest r_q value. By this method, we apply logarithmic transformations to *less than high school education* and *less than $15,000 income,* and apply a square-root transformation to the *never smoked* variable. After experimenting with several transformations of the*former smoker* and *college* + variables, we do not see an improvement in the test of normality. Since $r_q = 0.970$ for *former smoker* and $r_q = 0.967$ for *college* + without transformation, we allow these variables in our analysis by lowering the α value. The two variables *never smoked* and *less than high school education,* after transformation, yield r_q values lower than the critical value of 0.9768, but we keep these variables because we consider them to be important to our data set. The r_q values of 0.970

Variable	r_q	**Transformation**	r_q **after transformation**
Heavy Drinkers	0.989		
Binge Drinkers	0.987		
Less than High School	0.975	log	0.931
High School/ G.E.D	0.993		
College +	0.967	none	
Less than $15000 income	0.768	log	0.994
$15000-24999	0.998		
$50000+	0.986		
Physical Activities	0.990		
Health Care Coverage	0.986		
Dental Visit in last year	0.996		
Teeth Extraction	0.986		
Neither Overweight or Obese	0.981		
Obese	0.997		
Smokes Everyday	0.982		
Former Smoker	0.975	none	
Never Smoked	0.890	square root	0.970

Table 1: Correlation Coefficient Values and Transformations

and 0.931 can be accepted with a lower α value. The summary of the sample correlation coefficients, transformations used, and r_q values after transformation is given in Table 1.

2.3 Principal Component Analysis

When we desire to obtain data reduction and to interpret the relative importance of the variables in our data, we consider Principal Component Analysis (PCA). This method involves linear combinations of the variables in order to explain the variance-covariance structure of those variables. PCA depends solely on either the covariance matrix Σ or the correlation matrix R of the p variables $x_1, x_2, \ldots, x_p$. We define $\underline{x}$ as the vector $(x_1, x_2, \ldots, x_p)$ distributed as a multivariate normal with a corresponding covariance matrix Σ and eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$.

Let $[a_{lj}]$ be any $p \times p$ matrix and let $\mathbf{a}_i$ and $\mathbf{a}_k$ be the i^{th} and k^{th} columns, respectively, of the matrix. If we consider the linear combinations:

$$y_1 = a_1^T \underline{x} = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$$

$$y_2 = a_2^T \underline{x} = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p$$

$$\vdots$$

$$y_p = a_p^T \underline{x} = a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p,$$

we obtain:

$$\text{Var}(y_i) = \mathbf{a_i^T}\Sigma\mathbf{a_i} \text{ for } i = 1, 2, \ldots, p \text{ and}$$

$$\text{Cov}(y_i, y_k) = \mathbf{a_i^T}\Sigma\mathbf{a_k} \text{ for } i, k = 1, 2, \ldots, p.$$

We choose coefficients a_{lj} in order to explain the maximum variance of our data set using the linear equations and resulting matrices. The first principal component y_1 accounts for the most variance, the second principal component y_2 accounts for the most in the remaining variance, and so on, with the last principal component y_p accounting for the least amount of variance.

In order to have a unique solution of $y_1, \ldots, y_p$, we must maximize the variance with respect to $\mathbf{a_i}$ with the condition that $\mathbf{a_i^T a_i} = 1$ and Cov$(\mathbf{a_i^T x}, \mathbf{a_k^T x}) = 0$ for $i \neq k$. To obtain a non-null solution, a_i for $i = 1, 2, \ldots, p$ we must satisfy the condition of $|\Sigma - \lambda I| = 0$, λ being the characteristic root of Σ.

We can also express our principal components as $y_i = \mathbf{e}_\mathbf{i}^\mathbf{T}\mathbf{x}, i = 1, 2, \ldots, p$ where each $\mathrm{e_i}$ is the normalized eigenvector associated with the i^{th} largest eigenvalue, λ_i.

Because the eigenvalues $\lambda_1, \ldots, \lambda_p$ correspond to the variances of the components, we can compute the proportion of total variance explained by each y_i. It is given by:

$$\frac{explained}{total} = \frac{\lambda_i}{\sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

2.4 Principal Component Analysis Application

With our 17 normally-distributed variables, we apply PCA to the data using the sample covariance matrix S. We consider a scree plot (Figure 2) to identify the appropriate number of components to extract. We notice the plot begins to decrease slowly after component 6, so we consider the first six principal components for describing our data.

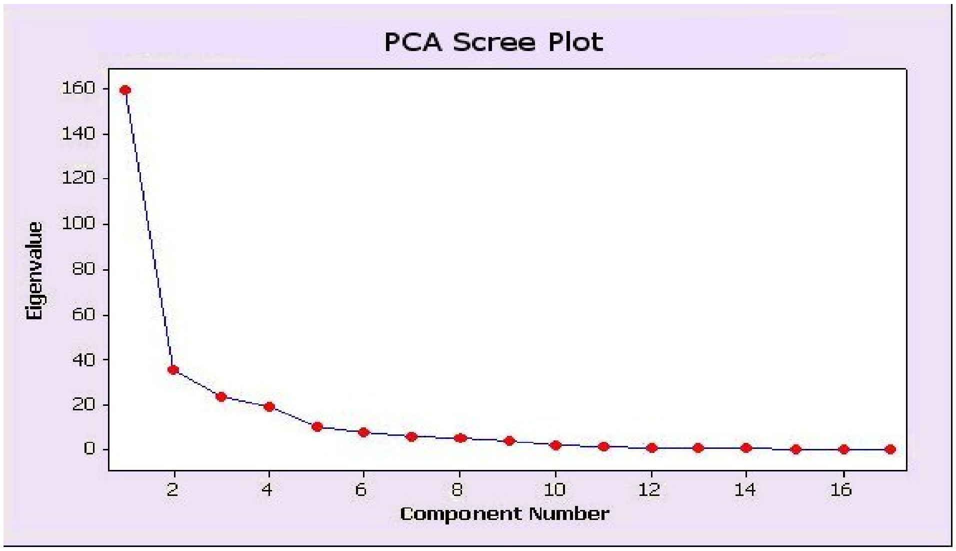

Figure 2: PCA Scree Plot

We are able to determine the amount of variance described by our model by observing the eigenvalues of the covariance matrix and the proportion of total variance that each eigenvalue contributes. In Table 2, we give the first 6 eigenvalues in bold, noting the proportions and cumulative proportions of each value. The information in Table 2 shows

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	159.22	35.61	23.85	19.45	9.91	7.40
Proportion	0.583	0.130	0.087	0.071	0.036	0.027
Cumulative	0.583	0.713	0.800	0.872	0.908	0.935

Table 2: PCA Eigenvalues and Proportions

	Label
PC1	Positive Lifestyle
PC2	Education
PC3	Health Care
PC4	Income
PC5	Poor Lifestyle
PC6	Alcohol and Cigarettes

Table 3: PCA Labels

the relative importance of each of the 6 principal components and indicates that together they provide an explanation of 93.5% of the total variance.

Because the PCA provides us with a linear combination of the variables, we can label each component based on the most significant variables—those variables with coefficients of large magnitude. We give these labels in Table 3.

2.5 Factor Analysis

The general purpose of factor analysis (FA) is to give a description of the covariance relationship among many variables. This covariance relationship is expressed in terms of underlying and unobservable random quantities (factors). FA is used to reduce the dimensionality of a data set as well. Let $\underline{x}$ be an observable random vector with p components, mean vector $\underline{\mu}$, and a covariance matrix $\mathbf{\Sigma}$. We want to make an m-factor model in which $\underline{x}$ is linearly dependant on random variables (common factors) $F_1, F_2, \ldots, F_m$, and has p additional specific factors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ corresponding to the errors.

Thus we create the m-factor model:

$$x_1 = \mu_1 + \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \varepsilon_1$$

$$x_2 = \mu_2 + \ell_{21}F_1 + \ell_{22}F_2 + \ldots + \ell_{2m}F_m + \varepsilon_2$$

$$\vdots$$

$$x_p = \mu_p + \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \varepsilon_p$$

The m-factor model in matrix notation is:

$$\underline{x} - \underline{\mu} = \mathbf{L}\,\underline{F} + \underline{\varepsilon}$$

The coefficient ℓ_{ij} is the *loading* of the i^{th} variable on the j^{th} common factor. This model is based on three assumptions on the random vectors $\underline{F}$ and $\underline{\varepsilon}$ which have a multivariate normal distribution with

1. $E(\underline{F}) = 0$;

2. $\text{Cov}(\underline{F}) = E(\underline{F}\,\underline{F}^T) = \mathbf{I}$ and $\text{Cov}(\underline{\varepsilon}) = E(\underline{\varepsilon}\,\underline{\varepsilon}^T) = \Psi = \text{diag}(\Psi_1, \Psi_2, \ldots, \Psi_p)$;

3. $\text{Cov}(\underline{\varepsilon}, \underline{F}) = E(\underline{\varepsilon}\,\underline{F}^T) = 0$, implying that $\underline{\varepsilon}$ and $\underline{F}$ are independent.

Based on these assumptions, we can make an m-factor model (with $m < p$) and show $\mathbf{L}\,\mathbf{L}^T + \mathbf{\Psi} = \mathbf{\Sigma}$, where $\mathbf{\Sigma}$ is the unknown covariance matrix. To obtain $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$, we must use the likelihood function:

$$L(\mathbf{L}, \mathbf{\Psi}) = \frac{1}{(2\pi)^{p/2}|\mathbf{L}\,\mathbf{L}^T + \mathbf{\Psi}|^{1/2}} e^{-(1/2)(\underline{x}_i - \underline{\mu})^T(\mathbf{L}\,\mathbf{L} + \Psi)^{-1}(\underline{x}_i - \underline{\mu})}$$

to find the maximum likelihood estimates of the two parameters $\mathbf{L}$ and $\mathbf{\Psi}$.

The null hypothesis H_0 is $\mathbf{\Sigma} = \mathbf{L}\,\mathbf{L}^T + \mathbf{\Psi}$. The alternate hypothesis H_A is that $\mathbf{\Sigma}$ must be equal to some positive definite matrix. The test is the likelihood ratio test. The null hypothesis is rejected if the test statistic χ^2 is greater than $\chi^2_{\nu,\alpha}$, where α is the level of significance and ν is the degrees of freedom.

The χ^2 test statistic is:

$$\chi^2 = [n - 1 - (2p + 4m + 5)/6] \ln \frac{|\hat{\mathbf{L}}\,\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}}|}{|\mathbf{S}|}$$

and ν is given by:

$$\nu = \frac{1}{2}[(p - m)^2 - p - m],$$

where n is the large sample size, p is the number of variables, m is the number of common factors, and $\mathbf{S}$ is the sample covariance matrix. We begin by testing the null hypothesis H_0 with value $m - 1$. If H_0 is rejected when $\chi^2 > \chi^2_{\nu,\alpha}$, we increase m until H_0 is not rejected. If H_0 is not rejected, we claim that our m-factor model is adequate.

2.6 Factor Analysis Application

We now give a description of the health data as a factor model. After we compute the correlation matrix R and the factor loadings of the variables, the scree plot (Figure 2) provides us with an estimate of how many factors to include in our model.

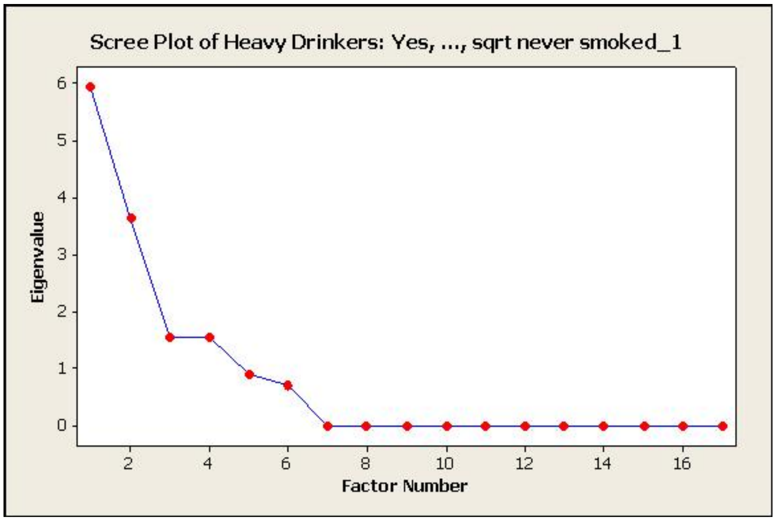

Figure 3: Scree Plot for $m = 6$ factor model

m	ν	**Test Statistic**	$\chi^2_{\nu,\mathbf{0.05}}$	**Comment**
5	61	267.7427	80.2321	reject
6	49	713.4095	66.3381	reject
7	38	149.8616	53.3835	reject
8	28	446.6309	41.3371	reject
9	19	103.2540	30.1435	reject
10	11	112.4175	19.6751	reject

Table 4: FA Tests of Adequacy Statistics

m	Variance Explained
5	0.761
6	0.837
7	0.881
8	0.888
9	0.908
10	0.913

Table 5: FA Variances for $m = 5, 6, \ldots, 10$ models

We note here, however, that according to Johnson & Wichern, with a large number of variables, our factor analysis is likely to fail the tests of adequacy. We observe this in our tests. Listed in Table 4 are the statistics of the $m = 5$ to $m = 10$ tests of adequacy.

Because we reject the null hypothesis for all of the models tested, we support the use of the the 6-factor model based on the statements of Johnson and Wichern, our scree plot, and the total variance described by each model. We give these variance percentages in Table 5.

With our 6-factor model, we consider the factor loadings and label each factor according to the weights placed on each variable. For example, we label the first factor as

Factor	Label
F1	Lifestyle
F2	Alcohol and Cigarettes
F3	Health care
F4	Physical Appearance
F5	Education
F6	Positive Lifestyle

Table 6: FA Labels

Lifestyle by observing that the largest positive coefficients primarily correspond to those variables that should be present in a healthy lifestyle, while strong negative coefficients correspond to those variables associated with an unhealthy lifestyle. We provide these labels in Table 6.

2.7 Discriminant Analysis

Discriminant Analysis is a multivariate statistical technique that is used to classify an observation into a previously defined group. Let us label two populations Π_1 and Π_2, and let $\underline{x}$ represent a set of measured variables. Our goal is to classify $\underline{x}$ into either Π_1 or Π_2. We will accomplish this goal by using Fisher's linear discriminant function, which is a linear combination of the components of $\underline{x}$. Fisher's linear discriminant function maximizes the distance between the two populations Π_1 and Π_2.

Let Ω be the p-dimensional space that contains all the values of $\underline{x}$. We want a rule that partitions Ω into two parts, R_1 and R_2, where $R_1 \cup R_2 = \Omega$ and $R_1 \cap R_2 = \emptyset$. Our classification procedure classifies $\underline{x}$ into Π_1 if $\underline{x}$ is in R_1 and $\underline{x}$ into Π_2 if $\underline{x}$ is in R_2. We would like to find an optimal partition that minimizes the *total probability of misclassification* (TPM). The TPM can be written as $\alpha = \alpha_1 + \alpha_2$, where α_1 is the probability of classifying an object as Π_2 when it is from Π_1, and α_2 is the probability of classifying an object as Π_1 when it is from Π_2. That is,

$$\alpha_1 = P(2|1) = \int_{R_2} f_1(\underline{x})dx$$

and

$$\alpha_2 = P(1|2) = \int_{R_1} f_2(\underline{x})dx,$$

where $f_i(\underline{x}) =$ p.d.f. of $\underline{x}$ under $\Pi_i, i = 1, 2$.

In order to minimize the TPM subject to $\alpha_1 = \alpha_2$, we classify $\underline{x}$ into Π_1 if $\underline{x}$ is more likely to fall under Π_1 than Π_2. We can assume that our populations are multivariate normal with mean vectors $\underline{\mu}_1$ and $\underline{\mu}_2$, respectively, after we have tested our data for normality. The two populations have p variables each and the covariance matrices are Σ_1 and Σ_2, respectively. In practice, $\underline{\mu}_1$ and $\underline{\mu}_2$ are unknown, so we use training samples of sizes n_1 and n_2 from Π_1 and Π_2. Let $\underline{\bar{x}}_1$ be an estimate of $\underline{\mu}_1$ and $\underline{\bar{x}}_2$ be an estimate of $\underline{\mu}_2$. Furthermore, let S_1 an estimate of Σ_1 and S_2 and estimate of Σ_2.

The calculation of the TPM involves calculating the Mahalanobis distance, Δp between two populations, where $\Delta p^2 = (\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$, assuming $\Sigma_1 = \Sigma_2 = \Sigma$. Let D_i^2 be the square of the Mahalanobis distance of $\underline{x}$ from Π_i. We classify $\underline{x}$ into Π_1 if D_1^2 is less than D_2^2; otherwise, we classify $\underline{x}$ into Π_2. This rule is known as the minimum distance classification rule.

Now, we need to determine whether or not to use linear ($\Sigma_1 = \Sigma_2$) or quadratic ($\Sigma_1 \neq \Sigma_2$) discriminant analysis. For this reasone, we test to see if the covariance matrices Σ_1 and Σ_2 are equal. Thus, the null hypothesis is $H_0 : \Sigma_1 = \Sigma_2 = \Sigma$(unknown) and the alternate hypothesis $H_A : \Sigma_1 \neq \Sigma_2$. If we accept H_0, we are able to use the linear discriminant analysis.

To see if we can use the linear discriminant analysis, we carry out the test of H_0 using the χ^2 test. In this test, we use the unbiased estimate of the common covariance matrix Σ given by S_{pooled} (the pooled sample covariance matrix.) The S_{pooled} matrix is defined by:

$$S_{pooled} = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} (n_i - 1) S_i,$$

The test statistic is given by M/c, where

$$M = \sum_{i=1}^{2}(n_i - 1)S_i \ln |S_{pooled}| - \sum_{i=1}^{2}(n_i - 1)\ln |S_i|$$

and

$$1/c = 1 - \frac{2p^2 + 3p - 1}{6(p+1)}[(\sum_{i=1}^{2} \frac{1}{n_i - 1}) - \frac{1}{n_1 + n_2 - 2}].$$

Under H_0, the statistic M/c has a χ^2 distribution with $\nu = \frac{p(p+1)}{2}$ degrees of freedom. We accept the null hypothesis if $M/c < \chi^2_{\nu,\alpha}$, where α is the significance level and p is the number of variables in $\underline{x}$.

After we determine the classification rule, we must find any misclassification in our training samples using the apparent error rate (APER). The APER measures the accuracy of the model and we use it to give us the percentage of the observations that are misclassified. Our objective is to aim for a small APER and minimize the percentage of misclassifications.

2.8 Discriminant Analysis Application

The first step in running the discriminant analysis is to divide the population in half. Thus, we divide the states into two populations, unhealthy (Π_1) and healthy (Π_2). We do this by using the median of the percentage of obese people who live in the state since obesity is an important variable. In practice, this criterion is given elsewhere, but for illustration purposes, we classify the training sample based on the median obesity value of all of the states. The median percentage of obese people in the states is 25.1%. Thus, the states with an obesity percentage greater than the median are classified into Π_1, the unhealthy group. On the other hand, the states with an obesity percentage less than the median are classified into Π_2, the healthy group. In our training sample, we set $n_1 = 18$ and $n_2 = 18$. Our testing sample is the remaining group of 15 states.

We test $H_0 : \Sigma_1 = \Sigma_2$ against $H_A : \Sigma_1 \neq \Sigma_2$ to determine if the linear discrimination is appropriate. Our test statistic is $M/c = 293.0905$ with a p-value of 7.2626×10^-11. Our critical value χ^2 with 153 degrees of freedom and $\alpha = 0.05$ is 182.865. Since $M/c =$

$293.0905 > 182.865 = \chi^2$, we reject H_0. After rejecting H_0, we decided to decrease the number of variables from seventeen to eleven. With the reduced number of variables, we retested H_0 and obtained $M/c = 101.7070$ and a p-value of 0.003141. This was still greater than our critical value with 66 degrees of freedom, 85.9648. Therefore, we concluded that we cannot use the linear discriminant analysis and must use the quadratic discriminant analysis. We give the results of the quadratic discriminant analysis for the training sample in Table 7. Thirty-six of thirty-six observations are classified correctly giving an APER of 0%. In Table 8, we show the classification of the states in our training sample, from which we created the rule for classification of the remaining states. This rule is applied to the remaining states, given in Table 9 along with the relative rankings of these test states (with a rank 1 being the healthiest) based on the squared distance to the group, either Π_1 or Π_2. We comment that these rankings are only available for the states in the test sample because the rule for discrimination is based on the training sample. We are treating the two samples as independent, and we do not reclassify the training sample to avoid redundancy in the results. If we were to rank the training sample as well, we would be providing unnecessary results, as the training sample developed the rule for classification and should not be included in the test sample as well. We also note that in this analysis, we chose our training sample based on the median of one variable, but this could have been chosen based on some other criterion.

	True Group	
Predicted Group	1	2
1	18	0
2	0	18
Total N	18	18
N Correct	18	18
Proportion Correct	1.000	1.000

Table 7: **Classification Summary of Quadratic Discriminant Analysis**

Healthy		**Unhealthy**	
Arizona	New Hampshire	Alabama	Nebraska
Colorado	New Jersey	Alaska	North Dakota
Connecticut	New Mexico	Arkansas	Ohio
District of Columbia	New York	Delaware	Oklahoma
Idaho	Oregon	Indiana	South Carolina
Maine	Pennsylvania	Kansas	Tennessee
Maryland	Rhode Island	Michigan	Texas
Massachusetts	Utah	Mississippi	West Virginia
Minnesota	Washington	Missouri	Wisconsin

Table 8: **Training Sample of Healthy and Unhealthy States**

State	**Predicted Group**	**Squared Distance**	**Health Rank**
Hawaii	2	8.595	1
Wyoming	2	11.134	2
Vermont	2	17.163	3
Nevada	2	19.408	4
Florida	2	22.745	5
California	2	24.965	6
Montana	1	30.567	7
Georgia	1	28.692	8
Kentucky	1	20.156	9
South Dakota	1	13.835	10
North Carolina	1	11.621	11
Iowa	1	10.964	12
Louisiana	1	10.789	13
Virginia	1	10.052	14
Illinois	1	9.436	15

Table 9: **Classification of Test Sample and Rankings**

3 Conclusion

Through the use of Factor Analysis, Principal Component Analysis, and Discriminant Analysis, we have assessed the overall health of each state, provided a classification of the United States into healthy and unhealthy groups, and created a ranking for a subset of the United States in order to focus on the progress of health improvements and on the most important variables in determining a state's health. We were surprised by the segmentation of the United States into the two groups. The Southeast and Midwest were almost entirely classified as unhealthy, while the West and Northeast were healthy. Our analysis did not account for geography, but the classification closely follows the geographic regions of the United States. We observe that this Midwest and Southeast are largely rural and agriculturally-based, perhaps encouraging lower income levels and less education. Since these variables largely contributed to the analysis, we can explain this partitioning. We would like to consider more variables in classification, such as the population with diseases, hospital visits, and health education in schools. Given more data, we could present a stronger classification and yield further conclusions of the relative health of the states.

4 Acknowledgements

We would like to express our deepest gratitude to Dr. Vasant Waikar for his instruction and guidance throughout our research. We would also like to thank Dr. Tom Farmer for preparing us in mathematical writing and for his numerous comments that helped us complete this paper. To Kevin Tolliver, our Graduate Assistant, we extend our thanks for all of the assistance and support as we faced difficulties in our journey. We greatly appreciate all of the faculty and staff in the SUMSRI program, without whom we would not have been able to have this experience. Thank you to Ms. Bonita Porter for your hard work and wonderful organizational skills that allowed us to have a memorable experience. We also thank Miami University, the National Security Agency, and the National

Science Foundation for the funding and grants that allowed us to participate in this program.

References

[1] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall, Upper Saddle River, New Jersey, 6th edition, 2007.

[2] Centers for Disease Control and Prevention (CDC), *Behavioral Risk Factor Surveillance System Survey Data*, Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2007.