Multivariate Classification Methods:
The Prevalence of Sexually Transmitted Diseases

Summer Undergraduate Mathematical Sciences Research Institute (SUMSRI)

Lindsay Kellam, Queens College
kellaml@queens.edu
Brandon McKenzie, Centre College
bdmcke00@centre.edu
Candace Porter, Albany State University
cporter@asurams.edu
Michael Sotelo, California Polytechnic University of Pomona
mjsotelo@csupomona.edu

In recent years, public attention regarding the prevalence of, effects of, and concerns for some sexually transmitted diseases have diminished. For many, the STD of greatest importance is HIV, the precursor to AIDS. However, although not in the visual forefront, the pervasiveness of certain sexually transmitted diseases, such as chlamydia, gonorrhea, and syphilis, continue to plague a large portion of America's adolescent population. The *National Overview of Sexually Transmitted Diseases, 1997*, a paper published by the Center for Disease Control and Prevention, highlights the "multifaceted, national dimensions of the morbidity, mortality, and costs that result from sexually transmitted diseases (STDs) in the United States."[i] National control programs have been established for chlamydia, gonorrhea, and syphilis. Each year, thousands of federal and state dollars are allocated for education programs, medical treatments, and preventative measures.

We used the STD situation to illustrate how multivariate classification methods can be used. First, we used principal component analysis to simplify the interpretation and summary of those variables which aid in predicting STD rates. Principal component analysis allowed us to depict our set of data using a number of descriptive factors that was less than the number of variables. We began with measurements of ten racial, ethnic, socioeconomic, and educational variables for each case and were able to combine them into four components that provide a clearer picture of the factors that predict the rate of STDs. Second, using discriminant analysis, we created a model that consisted of two groups: a group with a high rate of STDs and another with a low rate of STDs. Members (cases) in each group share similar racial, ethnic, socioeconomic, and educational variables. Using this discriminant model, we can predict an unknown observation's group classification.

When deciding which variables would be best for STD prediction, many factors were considered. One variable we knew would be of particular importance was a state's population. However, the population as a raw score was not a beneficial variable for our particular model. In discussing and predicting the spread of diseases, an important factor we took into consideration was population density. The most important result of using population density instead of population is that the model can then be applied to areas of varying size. In order to obtain the population density, we divided the population[ii] by the total land area[iii] in square miles. Due to the difficulty in locating full sets of data, we were forced to use land area data from 1996 and population data from 1997, but the change in land area from 1996 to 1997 was considered to be negligible.

Aside from population density, we decided that racial, ethnic, educational, and socioeconomic factors would serve as accurate predictors of STD cases. Therefore, the remaining nine variables were female population[iv], population in poverty[v], African American population[vi], Hispanic population[vii], birth rate[viii], per capita personal income[ix], Medicaid recipients[x], unemployment rate[xi], and percentage of high school graduates[xii]. (Female population, African American population, Hispanic population, Medicaid recipients, population in poverty, and high school graduates were recorded as proportions of the whole population). Due to the limited access of recent statewide data, all the measurements of the above nine variables were from 1997.

Of the ten variables collected on each of the fifty United States, only one state, Hawaii, contained a missing value. Due to the nature of the multivariate tests we conducted, the state of Hawaii was dismissed and all tests were performed on forty-nine cases. For a complete listing of the data[xiii] refer to the endnote section.

PRINCIPAL COMPONENT ANALYSIS

I. Theoretical Background

Principal component analysis (PCA) is a dimensionality reduction method, a technique used to reexamine data, and a process conducted to simplify data summarization. In PCA, the original variables are transformed into linear combinations called principal components. The goal of this multivariate technique is to find a few principal components that explain a large proportion of the total sample variance. The transformed variables (principal components) are defined to be orthogonal and uncorrelated.

Principal components can be thought of as dimensions that maximally separate the individual vectors. The extraction of principal components amounts to a variance maximizing rotation of the original variable space. Each principal component accounts for as much of the variability as possible, with the first component accounting for the most variability and each succeeding component accounting for as much of the remaining variability as possible.

The principal component method is derived as follows; we seek a linear combination $z = \vec{a}'\vec{x}$, where $\vec{a}$ is an arbitrary coefficient vector and $\vec{x}$ is the vector of original variables, and whose sample variance $S_z^2 = \vec{a}'S\vec{a}$ is a maximum relative to the length of $\vec{a}$. We want to maximize $\lambda = \frac{\vec{a}'S\vec{a}}{\vec{a}'\vec{a}}$ with respect to $\vec{a}$. Thus, we set $\frac{\partial\lambda}{\partial\vec{a}} = \frac{\vec{a}'\vec{a}(2S\vec{a}) - \vec{a}'S\vec{a}(2\vec{a})}{(\vec{a}'\vec{a})^2} = 0$. Simplifying, we obtain $S\vec{a} - \frac{\vec{a}'S\vec{a}}{\vec{a}'\vec{a}}\vec{a} = S\vec{a} - \lambda\vec{a} =$ $(S - \lambda I)\vec{a} = 0$. Since λ is an eigenvalue of S, the largest eigenvalue will maximize λ and $\vec{a}$ will be the corresponding eigenvector to λ_1, the largest eigenvalue. The argument is similar for the second principal component, which will be the eigenvector corresponding to the second largest eigenvalue of S, i.e. λ_2, etc.

We will now show some properties of principal components.

1. For $i = 1,2,..., p$, the eigenvector a_i is scaled so that $\lambda_i = \frac{\vec{a}_i' S \vec{a}_i}{\vec{a}_i' \vec{a}_i} = \vec{a}_i' S \vec{a}_i = S_{zi}^2$.

2. The total sample variance of the components is equal to the total sample variance of the variables:

$\sum_{j=1}^{p} Var(x_j) = \sum_{j=1}^{p} S_j^2 = \sum_{j=1}^{p} \lambda_j = tr(S)$, where $tr(S)$ is the trace (sum of the main diagonal) of the covariance matrix.

3. In principal component analysis, we can report the percent of variance explained by a ratio of eigenvalues. The proportion of total sample variance accounted for by the first k principal components is $\sum_{j=1}^{k} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j$.

II. Computational Background

Principal components can be obtained from either computing the covariance matrix or the correlation matrix. To achieve optimal results when using the covariance matrix, all variables (predictors) must be recorded in equivalent units. The widely accepted and understood reasoning behind this concept evolves from the idea that those variables with large variance relative to the others' variance will unjustly dominate the principal components. Therefore, variables that have relatively little significance as predictors, but have large variances will skew the results.

In general, if scaling the original variables is inappropriate or not convenient, it is conventional to take the principal components of the correlation matrix, implicitly rescaling all the variables to have unit sample variance. After evaluating our variables, we decided that calculating the correlation matrix would produce optimal results. The following table consists of the results that we obtained using Minitab.

TABLE 1.1
Eigenanalysis of the Correlation Matrix

Component Number	1	2	3	4	5	6
Eigenvalue	3.0613	1.9820	1.3405	1.1882	0.7359	0.5932
Proportion	0.306	0.198	0.134	0.119	0.074	0.059
Cumulative Proportion	0.306	0.504	0.638	0.757	0.831	0.890

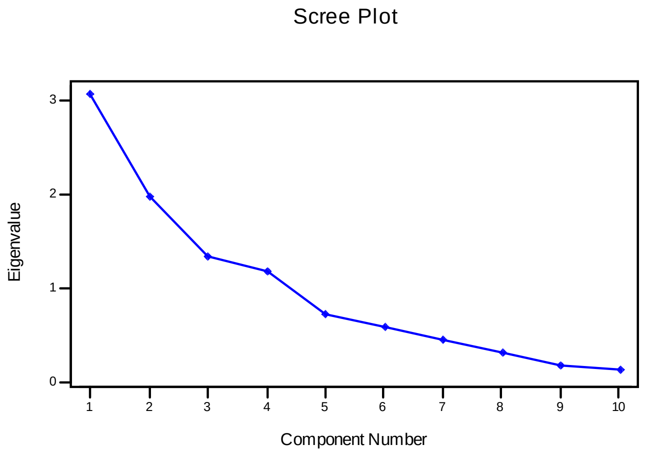
Component Number	7	8	9	10
Eigenvalue	0.4606	0.3204	0.1751	0.1428
Proportion	0.046	0.032	0.018	0.014
Cumulative Proportion	0.936	0.968	0.986	1.000

As shown in Table 1.1, we find in the second row the variance on the new factors that were successively extracted from the Minitab statistical software. In the third row, these values are expressed as a proportion of the total variance. We can see from Table 1.1, component 1 accounts for 30.6% of the variance. As expected, this individual component explains the largest portion of the variance. Component 2 accounts for 19.8% and the cumulative sum of the first four components is 75.7%.

Now that we have a measure of how much variance each successive factor extracts, we must decide how many components to preserve. Although determining the number of components to retain is often an arbitrary and relatively unformulaic process, there are some established guidelines that yield the best results.

The two most widely accepted and used methods of analysis for evaluating the principal components are the Kaiser[xiv] criterion and the Scree[xv] test. The Kaiser criterion relies on the use of the correlation matrix for extracting the principal components. When the correlation matrix is used, the variance of each variable is equal to one and therefore, the total variance is equal to the number of original variables. Based on Kaiser's method, only components with eigenvalues greater than one are retained. Consequently, unless a component extracts at least as much variance as the equivalent of one variable, it is eliminated. The basis for evaluation of the Scree plot relies on visual analysis. As illustrated in Figure 1.1, the Scree test plots the number of components versus their corresponding eigenvalues. In order to determine the number of components to retain, Cattell, who first proposed this graphical method, suggested that the values to the right of the place where the smooth decrease occurs and a reasonably horizontal line begins should be omitted.

FIGURE 1.1



III. Analysis

According to both criteria, Kaiser and Scree, we can say with relative certainty that four principal components should be retained. The eigenanalysis (Table 1.1) shows that that the first four eigenvalues are of size greater than one and accordingly should be used as principal components. The last few components have the smallest variance and are relatively close to zero. These latter components define a linear relationship among the variables that is nearly constant over the sample. Although the Scree plot (Figure 1.1) is graphically ambiguous as to the number of components we should retain, it seems appropriate to preserve four components. We consider the fifth component to be the graph's *elbow*: the point at which each successive component is relatively small and accounts for an almost negligible portion of the variance.

In summary, roughly 75.7% of the population variance is explained by four components. This percentage is significantly large for data that has not been numerically manipulated or computer generated.

We are now concerned with naming each component. In order to determine the property or name of each component, it is necessary to evaluate the variable's coefficients. The weight of a variable's coefficient predicts the importance of that variable to the model. The coefficients for the four principal components appear in Table 1.2.

TABLE 1.2

Variable	**PC1**	**PC2**	**PC3**	**PC4**
Population Density (persons/sq. mile)	0.144	-0.592	0.096	0.203
Proportion of Females	-0.085	-0.372	-0.012	-0.452
Proportion of Population in Poverty	-0.484	0.103	-0.126	0.122
Proportion of African Americans	-0.254	-0.337	0.200	-0.447
Proportion of Hispanics	-0.164	-0.199	-0.633	0.253
Births per 1000	-0.188	0.026	-0.645	-0.445
Per Capita Personal Income	0.317	-0.519	-0.105	0.173
Proportion of Medicaid Recipients	-0.465	0.007	0.213	0.170
Unemployment Rate	-0.330	-0.176	-0.095	0.459
High School Graduates (%)	0.456	0.215	-0.228	0.094

Naming the component is a very subjective procedure, but it seems most likely that the first principal component is the economic and education factor. Those variables that carry a significantly large portion of the weight are population in poverty, Medicaid recipients, per capita personal income, unemployment rate, and high school graduates. All these predictors lead us to conclude that the first principal component, which accounts for 30.6% of the variation, is the economic and education factor.

The second component appears to explain the demography. The largest coefficients correspond to the variables African Americans, females, and Hispanics. Therefore, we would name the second component the demographic factor. The third component is a slightly more difficult to name. The variables with the largest coefficients are proportion of Hispanics and birth rate. Since however, we have already used the Hispanic variable to describe the second component, this third component could be classified as the reproductive factor. The reproductive factor seems like a very logical component used to describe the rate of STDs in a population. Lastly, due to the ambiguity of the coefficients, we were unable to specifically name the fourth component.

In conclusion, principal component analysis allows for the opportunity to depict a set of data using a number of descriptive factors, which is less than the number of original variables. In all, PCA provides statisticians with a much more convenient method for describing and summarizing sets of data.

DISCRIMINANT ANALYSIS

I. Theoretical Background

In general, discriminant analysis is used to investigate differences between groups, determine the most parsimonious way to distinguish between groups, discard variables which are little related to group distinctions, classify cases into groups, and test theory by observing whether cases are classified as predicted. In contrast to principal component analysis, essentially a one-sample procedure, discriminant analysis, which maximally separates groups of observations, deals with two or more groups. When performing discriminant analysis, an attempt is made to delineate based upon maximizing between group variance while minimizing within group variance.

This multivariate technique is used to build a model for optimal group prediction. If the discriminant analysis function is effective for a set of data, the classification table of correct and incorrect estimates will yield a high percentage correct.

The discriminant function is a function of the random vector $\vec{x}$. In the simplest case, we have the linear discriminant function $z = \vec{a}'\vec{x}$, where $\vec{a}$ is an arbitrary coefficient vector and $\vec{x}$ is a vector of the original variables, and two populations π_1 and π_2. Assuming the availability of two random samples of p variate observation vectors $\vec{x}_1$ and $\vec{x}_2$ from populations with mean vectors $\vec{\mu}_1$ and $\vec{\mu}_2$ respectively, and common covariance matrix $\sum$, the goal of the discriminant technique is to seek a linear combination $z = \vec{a}'\vec{x}$ of the p variables that maximizes the distance between the two means $\bar{z}_1$ and $\bar{z}_2$.

We show $\bar{z}_1 - \bar{z}_2 = \vec{a}'\bar{\vec{x}}_1 - \vec{a}'\bar{\vec{x}}_2 = a'\left(\bar{\vec{x}}_1 - \bar{\vec{x}}_2\right)$. Without restrictions on a, $\bar{z}_1 - \bar{z}_2$ has no maximum. So we then standardize $\bar{z}_1 - \bar{z}_2$ by dividing by the standard deviation $S_z = \sqrt{\left(\vec{a}'S_{p1}\vec{a}\right)}$ where S_{p1} is the pooled covariance matrix from the two samples. Hence we seek $\vec{a}$ that maximizes $\left(\bar{z}_1 - \bar{z}_2\right)/S_z$.

Note: For two groups, the coefficient vector $\bar{a}$, which maximizes the above expression subject to all the conditions, is given by $\vec{a} = S_{p1}{}^{-1}\left(\bar{\vec{x}}_1 - \bar{\vec{x}}_2\right)$ or any multiple of $\vec{a}$. $\vec{a}$ is not unique, but its “direction” is and in order for $S_{p1}{}^{-1}$ to exist, we must have $n_1 + n_2 - 2 > p$ for n_1, n_2 equal the number of observations from π_1 , π_2 respectively.

We are assuming $\vec{x} \sim MN\left(\vec{\mu}, \sum\ \right)$ where $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_p \end{bmatrix}$ and $E\left(\vec{x}\right) = \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_p \end{bmatrix}$.

Two groups π_1 and π_2 are such that a vector $\underset{px1}{\vec{x}}$ of measurements $x_1,\ x_2, \ldots, x_p$ is to be classified into π_1 or π_2 using a linear discriminant function, $f(\vec{x}) = \vec{a}'\vec{x}$ where $\vec{a}'$ is a $1xp$ coefficient vector consisting of scalars.
We assume $\pi_1 \sim MN(\vec{\mu}_1, \sum)$ and $\pi_2 \sim MN(\vec{\mu}_2, \sum)$. To obtain the discriminant function, we choose $\vec{a}'$ subject to the constraint of minimizing $\alpha_1 + \alpha_2$, where $\alpha_1 + \alpha_2$ equals the total probability of misclassification and $\alpha_1 = \alpha_2$.

We obtain $\vec{a}' = 2(\vec{\mu}_1 - \vec{\mu}_2)'\sum^{-1}$. Since $\vec{a}'\vec{x} \sim N(\vec{a}'\vec{\mu}, \vec{a}'\sum\vec{a})$, our method becomes: classify $\vec{x}$ into π_1 if $\vec{a}'\vec{x} > h$ and classify into π_2 if $\vec{a}'\vec{x} \leq h$, where $h = \frac{1}{2}(\vec{\mu}_1 - \vec{\mu}_2)'\sum^{-1}(\vec{\mu}_1 + \vec{\mu}_2)$.
Note: We define $\Delta_p{}^2 = (\vec{\mu}_1 - \vec{\mu}_2)'\sum^{-1}(\vec{\mu}_1 - \vec{\mu}_2)$ to be the Mahalanobis distance between the two populations π_1 and π_2.
When $\vec{\mu}_1$, $\vec{\mu}_2$, and $\sum$ are unknown, we use a training sample $\vec{x}_1^{(1)}.........\vec{x}_{n1}^{(1)}$, $\vec{x}_1^{(2)}.........\vec{x}_{n2}^{(2)}$ which are from π_1 and π_2, respectively. Furthermore, $S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(\vec{x}_i^{(1)} - \bar{\vec{x}}_1)(\vec{x}_i^{(1)} - \bar{\vec{x}}_1)'$ (similarly for S_2), $S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{(n_1+n_2-2)}$ and it can be shown that $E(S_p) = \sum$.

In the case of k populations, we generalize the notion of Mahalanobis distance and obtain a classification rule that minimizes the total probability of misclassification.

The rule is to first obtain training samples $\begin{bmatrix} x_1^{(1)}...x_{n1}^{(1)} \approx MN(\vec{\mu}_1, \sum) \\ . \\ . \\ . \\ x_1^{(k)}...x_k^{(1)} \approx MN(\vec{\mu}_k, \sum) \end{bmatrix}$ and estimate both $S_1, S_2, ..., S_k$ and $\bar{\vec{x}}_1, \bar{\vec{x}}_2, ..., \bar{\vec{x}}_k$. Then for each unclassified vector $\vec{x}$, we check its Mahalanobis distance from each population $1, ..., k$ and classify $\vec{x}$ into π_j if $\Delta_j^2 = \min\left\{\Delta_1^2, \Delta_2^2, ..., \Delta_k^2\right\}$.

II. Computational Background

The same variables that were used for principal component analysis were used for discriminant analysis. For this multivariate technique, the variables are known as model predictors.

In order to create a discriminant model, we obtained a training sample of thirty states for which we knew the chlamydia, gonorrhea, and syphilis rates. The way in which we created group classifications was by assigning values, 1-low, 2-medium, and 3-high, for each of the three diseases and then obtained the sum of these values for each observation. By assigning these new values for each STD rate within each state, we were able to create a model that would accurately account for a state with an incredibly high rate or incredibly low rate for a particular disease. Next, we selected a cutoff value for group division. If an observation had a combined sum of less than six, it was classified as belonging to group 0, the low rate group. Similarly, if an observation had a combined sum of six or greater, it was classified as belonging to group 1, the high rate group. Six was selected as the cutoff value since it was the median of the range of possible raw scores (3-9) and would therefore provide both groups with equivalent ranges.

From the forty-nine states that did not contain missing values, we performed, using Minitab, a random selection of thirty cases. These cases were then used to build the discriminant function. The subsequent table, Table 1.3, is a summary of classification of the model.

TABLE 1.3

	…True Group…	
Classified into Group	0 1	
0	22 0	
1	0 8	
Total N	22 8	
N Correct	22 8	
Proportion	1.000 1.000	
N=30 N Correct=30 Proportion Correct =1.000		
Squared Distance Between Groups		
	0	1
0	0	29.1901
1	29.1901	0

According to this report, the Apparent Error Rate (AER) of misclassification is 1-1.000=0.000. This model will misclassify unknown cases 0% of the time, i.e. the rate of assigning a case to group 1 when it should actually be classified into group 0 is 0.000 and similarly when a case is assigned to group 0 and should instead be indexed as group 1. Clearly it is important to know the AER of a function, but it is equally important to know whether or not the discriminant function discriminates well. This is indicated by D^2 or the squared distance between groups. If D^2 is large then the function discriminates well, yet if D^2 is small then the function does not discriminate well. In the above case, the squared distance between groups is 29.1901, which is significantly large, indicating that this function discriminates well.

III. Functional Application

The essence of creating a functional model is to apply it to unknown observations to determine their correct classification. Therefore, we used this model to classify the remaining nineteen states as either belonging to group 1 or group 0. Table 1.4 shows how each observation was classified and the numerical basis for group prediction.

TABLE 1.4

Observation	Corresponding State	Predicted Group	From Group	Squared Distance	Probability
1	Virginia	1	0	18.662	.019
			1	10.761	.981
2	Vermont	0	0	19.673	1.000
			1	54.781	0.000
3	Utah	0	0	117.116	1.000
			1	187.021	0.000
4	Texas	0	0	15.171	1.000
			1	31.441	0.000
5	South Dakota	0	0	5273.244	1.000
			1	5439.274	0.000
6	Rhode Island	0	0	146.986	1.000
			1	206.579	0.000
7	Pennsylvania	0	0	12.061	1.000
			1	27.932	0.000
8	Oregon	0	0	6.478	1.000
			1	30.715	0.000
9	Oklahoma	0	0	4.468	1.000
			1	34.062	0.000
10	North Dakota	0	0	34.457	1.000
			1	90.005	0.000
11	New Jersey	0	0	127.655	1.000
			1	166.251	0.000
12	Missouri	0	0	9.468	.999
			1	23.613	.001
13	Massachusetts	0	0	73.885	1.000
			1	108.493	0.000
14	Maine	0	0	3835.888	1.000
			1	3980.083	0.000
15	Louisiana	1	0	47.631	0.000
			1	7.331	1.000
16	Illinois	1	0	13.134	.361

			1	11.988	.639
17	Florida	0	0	9.491	.970
			1	16.428	.030
18	Delaware	1	0	14.880	.243
			1	12.609	.757
19	Arkansas	1	0	18.774	.051
			1	12.941	.949

The values located in the column entitled “Squared Distance” are statistically known as the Mahalanobis distances. The Mahalanobis distance is the distance between an experimental case and the centroid[xvi] for each group, 0 or 1. Each case will have one Mahalanobis distance for each group, and it will be classified as belonging to the group for which its Mahalanobis distance is smallest. Moreover, the smaller the Mahalanobis distance, the closer the case is to the group centroid and more likely it is to be classified as belonging to that group.

Likewise, the probabilities indicate the likelihood that the specific case being tested belongs to one of the two groups. For example, observation 18 has a probability of .243 of being classified as belonging to group 0 and a probability of .757 of being classified as belonging to group 1. For all nineteen observations, both the Mahalanobis distances and the probabilities yielded the same conclusion.

IV. Analysis

In order to determine the efficiency of our model, we compared the classifications of each state as published in the report by the Center for Disease Control and Prevention with the groupings determined by our discriminant model. Three cases, observation 1 (Virginia), observation 16 (Illinois), and observation 19 (Arkansas) were classified differently. Our model classified Virginia, Illinois, and Arkansas as belonging to group 1, while according to the *National Overview of Sexually Transmitted Diseases, 1997,* all three cases are members of group 0.

In conclusion, we were successfully able to build a discriminant function for STD rate classification. By showing that this model provides accurate and reliable results for classifying states, we can in the future examine counties and cities within a particular state.

Acknowledgements

We would like to thank Dr. Vasant Waikar for overseeing the progress of this paper and Natasha Trofimenko for assisting with Minitab work. We would also like to thank Dr. Dennis Davenport, director of SUMSRI 2000, for establishing this program and giving us the opportunity to participate. Lastly, we would to extend our thanks to our graduate assistants, Angela Grant and Duke Hutchings, for their help and advice.

Bibliography

[i] *National Overview of Sexually Transmitted Diseases, 1997*, STD Surveillance 1997.

[ii] The 1997 population estimates were posted on the Census web page (http://www.census.gov) as *Estimates of the Population of States, Annual Time Series, July 1, 1990 to July 1, 1997*.

[iii] The Total Area of States in Square Miles in 1996 came from State Rankings 1997.

[iv] The 1997 female population data came from CQ's State Fact Finder 1999 (Table A-13).

[v] The 1997 Population in Poverty data came from CQ's State Fact Finder 1999 (Table A-11).

[vi] The 1997 African American Population data came from CQ's State Fact Finder 1999 (Table A-9).

[vii] The 1997 Hispanic Population data came from CQ's State Fact Finder 1999 (Table A-10).

[viii] The 1997 Birth Rate data came from CQ's State Fact Finder 1999 (Table A-14).

[ix] The 1997 Per Capita Personal Income data came from CQ's State Fact Finder 1999 (Table B-3).

[x] The 1997 Medicaid Recipients data came from CQ's State Fact Finder 1999 (Table I-12).

[xi] The 1997 Unemployment Rate data came from http://epinet.org/datazone/urates_bystdiv.html (Source: Economic Policy Institute analysis of BLS data).

[xii] The 1997 Percentage of Population Over 25 with a High School Diploma data came from CQ's State Fact Finder 1999 (Table H-6).

[xiii]

State	Population 1997 (in 1000)	Population Density (People/Sq. Mile)	Proportion of Females 1997	Proportion in Poverty 1997
New Jersey	8053	1085.46	0.514715	0.091519
Rhode Island	987	944.5	0.518744	0.121581
Massachusetts	6118	780.56	0.516999	0.119647
Connecticut	3270	674.92	0.513761	0.086239
Maryland	5094	521.13	0.513153	0.082843
New York	18137	384.06	0.517947	0.16425
Delaware	732	374.42	0.512295	0.098361
Ohio	11186	273.14	0.515645	0.110048
Florida	14654	271.69	0.513921	0.140303
Pennsylvania	12020	268.18	0.518386	0.111231
Illinois	11896	213.98	0.511516	0.113399
California	32268	206.88	0.499132	0.169177
Michigan	9774	172.05	0.512482	0.102926
Virginia	6734	170.06	0.510247	0.127413

Indiana	5864	163.48	0.512449	0.087824
North Carolina	7425	152.41	0.51367	0.112997
New Hampshire	1173	130.78	0.507246	0.092924
Tennessee	5368	130.23	0.51658	0.147355
Georgia	7486	129.25	0.512423	0.148143
South Carolina	3760	124.87	0.516755	0.132979
Louisiana	4352	99.89	0.517923	0.158778
Kentucky	3908	98.36	0.513818	0.159417
Wisconsin	5170	95.19	0.507737	0.081625
Alabama	4319	85.1	0.519102	0.153971
Washington	5610	84.26	0.501604	0.094296
Missouri	5402	78.41	0.514809	0.116068
West Virginia	1816	75.39	0.51707	0.157489
Texas	19439	74.22	0.505582	0.169607
Vermont	589	63.68	0.50764	0.091681
Minnesota	4686	58.86	0.506402	0.097525
Mississippi	2731	58.21	0.51959	0.166606
Iowa	2852	51.04	0.512272	0.09467
Arkansas	2523	48.45	0.515656	0.204122
Oklahoma	3317	48.3	0.510702	0.137474
Maine	1242	40.24	0.293881	0.099839
Arizona	4555	40.08	0.504281	0.174973
Colorado	3893	37.53	0.503725	0.082199
Oregon	3243	33.78	0.505396	0.117792
Kansas	2595	31.71	0.507514	0.096339
Utah	2059	25.06	0.502186	0.089849
Nebraska	1657	21.55	0.509958	0.098371
Nevada	1677	15.27	0.490161	0.113298
Idaho	1210	14.62	0.5	0.15124
New Mexico	1730	14.25	0.506936	0.223699
South Dakota	738	9.72	0.263279	0.158537
North Dakota	641	9.29	0.50078	0.135725
Montana	879	6.04	0.501706	0.158134
Wyoming	480	4.94	0.495833	0.1375
Alaska	609	1.07	0.474548	0.091954

State	Proportion in Poverty 1997	Proportion of African Americans 1997	Proportion of Hispanics 1997	Births per 1000
New Jersey	0.091519	0.145287	958.9	14
Rhode Island	0.121581	0.048126	61.5	12.5
Massachusetts	0.119647	0.0627	358.5	13.5
Connecticut	0.086239	0.091682	259.2	13.1
Maryland	0.082843	0.274205	179.4	13.8
New York	0.16425	0.176893	2570.4	14.5
Delaware	0.098361	0.19153	24.1	14
Ohio	0.110048	0.114214	172.7	13.6
Florida	0.140303	0.153726	2105.7	13.1
Pennsylvania	0.111231	0.096847	302.3	12
Illinois	0.113399	0.152581	1183	15.2
California	0.169177	0.074269	9941	16.3
Michigan	0.102926	0.142398	253.8	13.7
Virginia	0.127413	0.199569	238.9	13.7
Indiana	0.087824	0.082469	136.6	14.2
North Carolina	0.112997	0.432094	149.4	14.4
New Hampshire	0.092924	0.007161	16.9	12.3
Tennessee	0.147355	0.164773	56.6	13.9
Georgia	0.148143	0.28401	207.1	15.8
South Carolina	0.132979	0.300638	46.3	13.8
Louisiana	0.158778	0.320887	113.2	15.2
Kentucky	0.159417	0.072467	30.1	13.6
Wisconsin	0.081625	0.055338	127.7	12.9
Alabama	0.153971	0.259389	39.3	14.1
Washington	0.094296	0.034938	339.6	14.1
Missouri	0.116068	0.112292	82.2	13.8
West Virginia	0.157489	0.031718	10.1	11.4
Texas	0.169607	0.122136	5722.5	17.2
Vermont	0.091681	0.005263	5.2	11.3
Minnesota	0.097525	0.028404	80.7	13.8
Mississippi	0.166606	0.36375	21.7	15.7
Iowa	0.09467	0.019495	53.1	12.9
Arkansas	0.204122	0.16084	45.1	14.6
Oklahoma	0.137474	0.07748	122.1	14.5
Maine	0.099839	0.004911	8.5	11

Arizona	0.174973	0.035324	998.6	16.6
Colorado	0.082199	0.043257	556.1	14.5
Oregon	0.117792	0.017946	189.8	13.5
Kansas	0.096339	0.058882	132.6	14.4
Utah	0.089849	0.008499	133.4	21.3
Nebraska	0.098371	0.039952	67.9	14.1
Nevada	0.113298	0.074717	253.3	16.1
Idaho	0.15124	0.005455	86	15.4
New Mexico	0.223699	0.025723	692.6	15.5
South Dakota	0.158537	0.006775	8	13.8
North Dakota	0.135725	0.00624	6.8	13
Montana	0.158134	0.003641	15.1	12.3
Wyoming	0.1375	0.008333	28.4	13.4
Alaska	0.091954	0.03908	23.3	15.9

State	**Per Capita Personal Income 1997**	**Proportion of Medicaid Recipients 1997**	**Unemployment Rate**
New Jersey	32233	0.066807	5.1
Rhode Island	26689	0.118541	5.3
Massachusetts	31207	0.118176	4
Connecticut	35954	0.061774	5.1
Maryland	28671	0.078916	5.1
New York	30299	0.173788	6.4
Delaware	28443	0.114754	4
Ohio	24203	0.124799	4.6
Florida	24795	0.10898	4.8
Pennsylvania	25678	0.085275	5.2
Illinois	27929	0.117687	4.7
California	26218	0.150459	6.3
Michigan	24998	0.11592	4.2
Virginia	26172	0.088358	4
Indiana	23183	0.087824	3.5
North Carolina	23174	0.149899	3.6
New Hampshire	27806	0.080989	3.1
Tennessee	22752	0.263785	5.4
Georgia	23893	0.161368	4.5
South Carolina	20651	0.138298	4.5

Louisiana	20473	0.171415	6.1
Kentucky	20599	0.169908	5.4
Wisconsin	24199	0.075822	3.7
Alabama	20699	0.126418	5.1
Washington	26412	0.112299	4.8
Missouri	23723	0.099963	4.2
West Virginia	18734	0.197687	6.9
Texas	23647	0.130614	5.4
Vermont	23018	0.185059	4
Minnesota	26295	0.079172	3.3
Mississippi	18087	0.184548	5.7
Iowa	23177	0.103086	3.3
Arkansas	19602	0.146651	5.3
Oklahoma	20214	0.095267	4.1
Maine	21928	0.134461	5.4
Arizona	21994	0.118771	4.6
Colorado	27015	0.064475	3.3
Oregon	23984	0.163737	5.8
Kansas	24014	0.089788	3.8
Utah	20246	0.070423	3.1
Nebraska	23656	0.122511	2.6
Nevada	26553	0.063208	4.1
Idaho	20394	0.095041	5.3
New Mexico	19250	0.184971	6.2
South Dakota	20651	0.101626	3.1
North Dakota	20213	0.095164	2.5
Montana	19704	0.109215	5.4
Wyoming	22612	0.102083	5.1
Alaska	24945	0.119869	7.9

State	**Percentage of High School Graduates 1997**	**Chlamydia Rate per 100,000**	**Gonorrhea Rate per 100,000**
New Jersey	84.8	129.5	95
Rhode Island	77.5	208.9	42.6
Massachusetts	85.9	131	36.5
Connecticut	84	185.2	92.4
Maryland	84.7	271.4	228.1

New York	80	385.7	123.1
Delaware	84.4	360.5	175.6
Ohio	86.2	204.3	133.9
Florida	81.4	186	132.5
Pennsylvania	82.4	164.5	82.7
Illinois	84.4	194.4	155.5
California	80.7	215.3	56.3
Michigan	86	223	164
Virginia	81.3	174	130.8
Indiana	81.9	164.4	105.4
North Carolina	78.4	233.6	230.6
New Hampshire	85.1	70.2	8.3
Tennessee	76.1	235	207.2
Georgia	78.8	216.4	251.2
South Carolina	77.3	338.2	310.6
Louisiana	75.7	265.4	247.8
Kentucky	75.4	163	103.7
Wisconsin	87.1	185.2	83.6
Alabama	77.6	203.7	281.6
Washington	88.8	173	35.6
Missouri	80.1	229.7	148.2
West Virginia	77.3	170.2	52.4
Texas	78.5	264.9	139.1
Vermont	84.4	73.7	9
Minnesota	87.9	142.4	51.9
Mississippi	77.5	290.8	306.6
Iowa	86.7	172.1	46
Arkansas	76.9	99.7	174.6
Oklahoma	85.2	224.7	144.1
Maine	85.8	85.7	5.3
Arizona	82.6	243.5	85.9
Colorado	87.6	188.2	60.7
Oregon	84.7	164.5	24.1
Kansas	88.1	155.6	67.6
Utah	89.5	88.7	13.9
Nebraska	86	167.5	73.2
Nevada	85.4	121.8	34.2
Idaho	85.7	143.7	13.3

New Mexico	78	234.7	50
South Dakota	85.6	198	23.6
North Dakota	82.6	140.2	10.6
Montana	88.6	130.3	7.5
Wyoming	91.3	131.9	11.2
Alaska	92.1	266.1	64.6

State	**Syphilis Rate per 100,000**
New Jersey	1.9
Rhode Island	0.2
Massachusetts	1.3
Connecticut	1.9
Maryland	17.6
New York	0.8
Delaware	3
Ohio	2
Florida	2.1
Pennsylvania	1
Illinois	3.7
California	1.2
Michigan	1.6
Virginia	3.5
Indiana	2.6
North Carolina	9.8
New Hampshire	0
Tennessee	14
Georgia	7
South Carolina	10.2
Louisiana	8.4
Kentucky	3.5
Wisconsin	1.7
Alabama	9.6
Washington	0.3
Missouri	2.1
West Virginia	0.1
Texas	3.5

Vermont	0
Minnesota	0.3
Mississippi	14.4
Iowa	0.2
Arkansas	6.9
Oklahoma	3.5
Maine	0.2
Arizona	3
Colorado	0.4
Oregon	0.3
Kansas	1.1
Utah	0.2
Nebraska	0.3
Nevada	0.6
Idaho	0.1
New Mexico	0.5
South Dakota	0.1
North Dakota	0
Montana	0
Wyoming	0
Alaska	0.2

[xiv] Kaiser (1960)

[xv] The scree test was introduced by Cattell (1966). Scree is the geological term referring to a slope of loose rock debris at the base of a steep incline or cliff.

[xvi] The centroid is the mean value for the discriminant scores for a given category of the dependent variable.