Educating the States:
A Multivariate Statistical Analysis of Education

Sara Blight[1] and Nick Imholte[2]
July 2004

Abstract

Educating the population is important in every state. To measure the quality of education in a state, we examine average Scholastic Aptitude Test scores. We create a model to predict future scores based on variables that affect education. First, we use the multivariate statistical methods of Principal Component Analysis and Factor Analysis to reduce the number of variables. Second, we use both of these methods in conjunction with Discriminant Analysis to create a model that predicts future scores. Finally, we use the results of Discriminant Analysis to conjecture how to improve the quality of education.

Learning is not attained by chance, it must be sought for with ardor and attended to with diligence.

—Abigail Adams [4]

Introduction

Is there a way to predict which states will provide the best education? What factors and influences contribute to students doing well on standardized tests? Using the multivariate statistical methods of Principal Component Analysis (PCA), Factor Analysis (FA) and Discriminant Analysis (DA), we answer these questions and determine whether or not a state is likely to provide a quality education. Further, we reduce the dimensionality of a multifaceted data set in order to discover the underlying patterns and factors. Finally, we use our analysis results to conjecture how to improve the quality of education.

In order to create an accurate model, we gather a data set of many variables which may affect the quality of education. We use several multivariate statistical methods to analyze our data. First, we use the methods of PCA and FA to separately reduce the dimensionality of the data set. Second, we compare the results of both analysis techniques to determine which provides the better results. Third, we apply DA, a classification method for observations, and determine which model produces more accurate results. Finally, three-group DA is used to classify states as high, medium, or low performing.

We obtained our data from the National Education Association, the National Center for Education Statistics, collegeboard.com, and the U.S. Census. We look at variables from all fifty U.S. states and the District of Columbia for the 1998-1999 and 1999-2000 school years. Our fifteen variables are divided into three categories: economic, size and others. These variables are displayed in the table below.

[1] University of Arizona, Tucson, AZ 85721, sblight@email.arizona.edu

Table 1: Categories of Variables

Economic Variables	Size Variables
Average Teacher Salary	Number of School Districts
Per Capita Personal Income	Number of Students who take SAT
Median Income of Family of 4	Enrollment in Institutes of Higher Education
Expenditures per Student	Enrollment in Public High Schools
Revenue per Student	Number of High School Graduates
	Population per Square Mile

Other Variables
Percent Minorities Enrolled
Average Pupil to Teacher Ratio
Percent with High School Diploma
Percent with Bachelor's Degree

We choose these variables with several goals in mind. First, we want to account for as many influences on education as possible. Therefore, we include demographic and economic, as well as educational variables in our analysis. Second, we want to study related variables in order to find patterns and relationships in the data. Finally, we include variables that can be altered by state governments, so that education can be improved based on the results of our analysis.

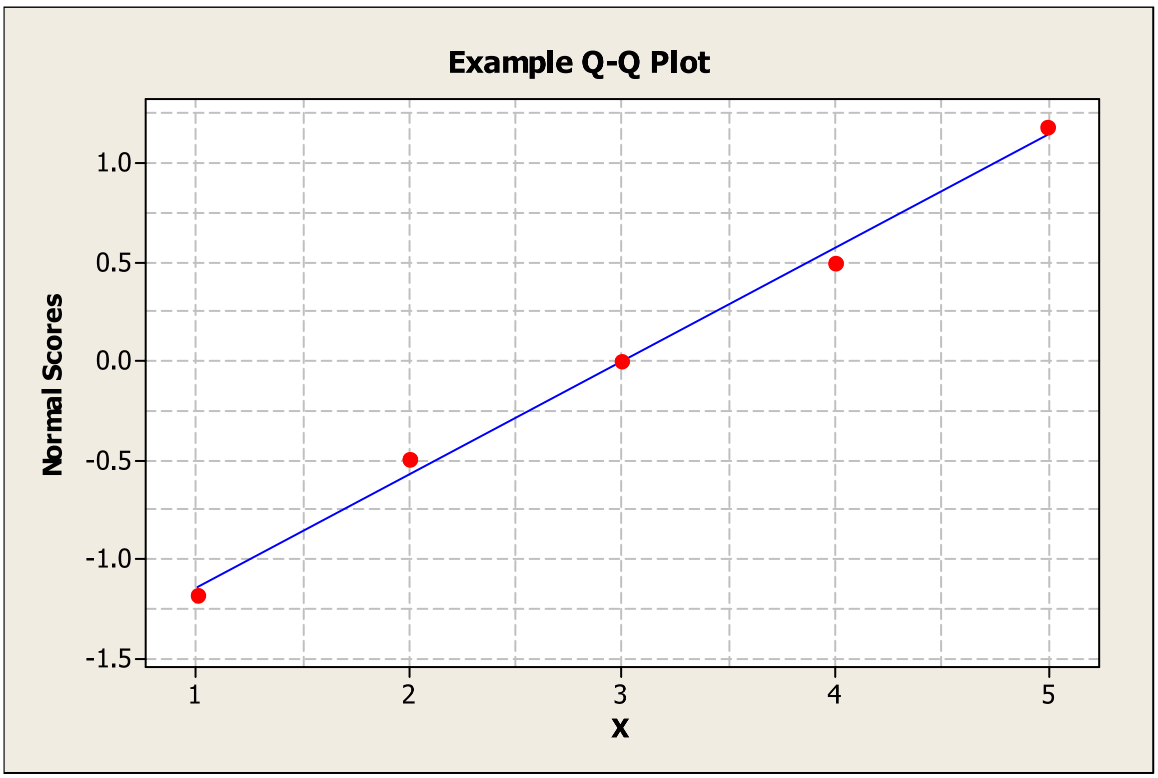
Discussion of Normality

Before we apply any multivariate statistical techniques, we must confirm the normality of our data. A random variable is said to be normally distributed if it has a Probability Density Function (PDF) shaped like a bell curve. One of our main tools for assessing the normality of a variable is the Quantile-Quantile plot, commonly called the Q-Q plot. The Q-Q plot is a way to measure the relationship between a variable and a linear combination of its mean and standard deviation. If the relationship is linear, this gives strong evidence for assuming normality. This conclusion follows from the fact that if X is distributed normally with mean μ and standard deviation σ, denoted $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$. From this we can conclude that $X = \mu + Z\sigma$. While looking at the Q-Q plot and noting a linear relationship is helpful, it is not sufficient evidence for assuming normality. In order to validate our assumption, we measure the linearity of the Q-Q plot by examining the sample correlation coefficient, r_Q, defined by:

$$r_Q = \frac{\text{Covariance}(X_{(j)}, Z_{(j)})}{\sqrt{\text{Variance}(X_{(j)})\text{Variance}(Z_{(j)})}} = \frac{\sum_{j=1}^{n}(X_{(j)} - \overline{X})(Z_{(j)} - \overline{Z})}{\sqrt{\sum_{j=1}^{n}(X_{(j)} - \overline{X})^2}\sqrt{\sum_{j=1}^{n}(Z_{(j)} - \overline{Z})^2}}$$

Note that both the covariance and variance expressions in this formula are sample estimates. In order to understand this expression, first consider a ranked list of the observed values of X. $X_{(1)}$ is the smallest sample value of X, $X_{(2)}$ is the second smallest, and so forth. On the other hand, $Z_{(j)}$ is the *quantile* of the standard normal distribution corresponding to $X_{(j)}$. That is, $P(Z \le Z_{(j)}) = (j - 0.5)/n$, where n is the sample size of X. Further, $\overline{X}$ and $\overline{Z}$ are the sample means of X and Z respectively. Since Z is standard normal, $\overline{Z} = 0$.

To illustrate this process, consider this brief example. Let X be a random variable with sample size 5 and observed values 1, 2, 3, 4 and 5. Using Minitab, we display the corresponding Q-Q plot below.



Notice that the plot looks linear, but we wish to measure exactly how linear it is. Therefore we calculate r_Q. First we rank the observed X values in increasing order. So $X_{(1)} = 1$, $X_{(2)} = 2$ and so on. Second, we see that $\overline{X} = 3$. Next, to determine the $Z_{(j)}$ values we calculate the desired probability levels. These are equal to (j - 0.5)/n, so our probability levels are 0.1, 0.3, 0.5, 0.7 and 0.9. Using a normal distribution table, we calculate the $Z_{(j)}$ values. The inverse normal distribution function gives us the following values: -1.280, -0.524, 0.000, 0.524 and 1.280. Since we have every value that the r_Q

formula calls for, we calculate r_Q to be 0.997. For a sample size of 5, this is very strong evidence for assuming normality.

The process for finding r_Q is long and difficult when done using the above method on a large sample size. Therefore, we find the r_Q value for each variable using a computer software program such as Minitab. If the r_Q-value for a variable is above a pre-determined critical value, then we accept the hypothesis that the variable is normal. With a sample size of 102 and a 0.01 significance level, the critical correlation value for our data is 0.9822. In verifying that each of the fifteen variables is univariate normal we are not guaranteed to have a joint multivariate normal distribution. Nevertheless, this procedure does provide strong enough evidence for our purposes to assume multivariate normality. While stronger tests for determining joint multivariate normality exist, they are beyond the scope of this paper [6].

For our data, we can only accept three variables as normal. These are median income of a family of four, revenue per student, and percentage of population with a high school diploma. The corresponding r_Q-values are 0.990, 0.986, and 0.990 respectively. However, we may perform transformations on the rest of the data to achieve normality. The most common transformations we use are log and square root, which both decrease the variance of the variable and thus increase the likelihood we can accept normality. We transform the remaining twelve variables into new variables that have r_Q-values exceeding the critical value. Note that these variables exist on different scales. For instance, teacher salary is in the tens of thousands range while percent minorities is less than 100. If left unchanged, the salary would dominate the analysis, while percent minorities would be largely ignored. Therefore, in order to ensure that every variable has equal weight in the coming analysis, we scale the data to the range of zero to fifteen. Since scalar multiplication preserves r_Q, this is a legal operation. We display the variables, their transformations and scaling, and their corresponding r_Q-values in the table below.

Table 2: Transformed and Scaled Variables and Normality Tests

Variables	r-values
sqrt(Teacher Salary)/20	0.985
Log(HS Enrollment)	0.991
sqrt(School Districts)/2	0.991
Log(Pupils to Teacher)	0.986
Log(Graduates)	0.991
sqrt(Personal Income)/15	0.986
Med Income/10000	0.990
sqrt(% Minorities)	0.994
Log(Higher Institutes Enrollment)	0.991
(Revenue per Student)/1000	0.988
Log(SAT Takers)	0.990
Log(Population per sq. mile +1)	0.990
sqrt(% with Bachelors Degree)	0.990
(% with HS Diploma)*10	0.990
sqrt(Expenditures / Student)/10	0.984

We display the complete list of the data we use for further analysis in appendix B.

Theory of Principal Component Analysis

A major problem in data analysis is the large number of variables that one encounters. Despite the existence of helpful computer packages, a large data set can still be difficult to manage. To overcome this obstacle and reduce the number of variables we use the technique of Principal Component Analysis (PCA). The goal of PCA is to locate and eliminate the redundancies in the data. Redundancies exist when multiple variables are linearly dependent. The more linear dependence, the more we can reduce the dimensionality of the data set. This process creates new variables called principal components (PCs), which are linear combinations of the original variables. Once we create these new components, we can easily rank them by their variance. Further, we disregard the ones that provide an insignificant fraction of the total variance. By reducing the number of variables, we greatly facilitate additional analysis. In addition, by using PCA we reveal patterns in the data and unobservable relationships between the variables.

Before we apply PCA, we require that the variables are linearly dependent. Clearly if the variables were completely independent, there would be no redundancies in the data, so PCA would be ineffective. If two variables are linearly related, they will have a non-zero correlation. We place the correlations between all of the variables in the population correlation matrix P. Remembering that a variable has a correlation of one with itself, we see that if the variables are completely independent, then P = I, where I is the identity matrix. These observations are critical for the testing of linear independence. In order to use these observations, we employ the Chi-Square test, a likelihood ratio test. We test the null hypothesis, H_0: P = I against H_1: P ≠ I. We use Minitab to calculate a test statistic based on the number of variables and the size of the sample. We then compare this test statistic to a value in the Chi-Square table, allowing for a 1% chance of type one error, which is the chance of rejecting H_0 when H_0 is true. If the test statistic is larger than the table value, then we conclude that the variables are not completely independent. Thus we can use PCA to reduce the dimensionality of the data set.

While PCA is primarily a statistical method, we draw on calculus and linear algebra to create the PCs that have maximum variance. We label these PCs as y_1, y_2,…, y_p, where p is the number of variables. We set up the Lagrangian problem of maximizing var (y_i) = (u^T) R (u) under the constraint that u is a unit vector. Note that u is the vector of coefficients for the i^{th} PC and R is the sample correlation matrix between the variables. This maximization problem leads to the equation $(R - \lambda_i I)\ u = 0$, which is easily recognized from linear algebra as an eigenvalue-eigenvector problem. Since R is a positive definite matrix, the eigenvalues are guaranteed to be positive and real. By solving this eigen-problem for each y_i, we create a new set of vectors, each of which is an eigenvector accounting for a variance equal to its corresponding eigenvalue. Since at this point the number of variables has not been reduced, we must next rank these PCs according to their variance. We focus only on the variables with the greatest eigenvalues, and ignore those whose variance is insignificant to the variance as a whole. The percentage of total variance accounted for by the first k PCs is calculated by $\left(\sum_{i=1}^{k} \lambda_i\right) / \left(\sum_{j=1}^{p} \lambda_j\right)$, where λ_i is the eigenvalue of the i^{th} PC. Generally, we select those vectors that have an eigenvalue greater than one or aim to account for at least 90% of the total variance. These values are arbitrary and are left to the discretion of the researcher; however, we generally wish to use as few principal components as possible, without losing too much information. By using this technique, we reduce the total number of variables while still accounting for a large amount of the total variance [7].

Application of Principal Component Analysis

As discussed above, we check our data set for linear independence. The null hypothesis is
H_0: P = I and we calculate the χ^2 test statistic with the formula

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right)\ln(\det(R))$$

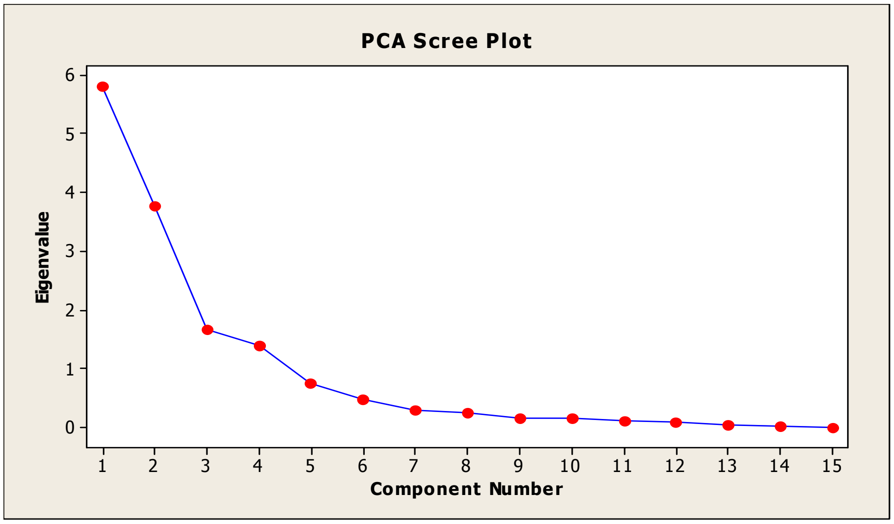
where n is the sample size [7]. Using this formula, the test statistic is found to be 2085.973. We then compare the test statistic with the table value $\chi^2_{105,\ 0.01} \approx 135.807$. Using Minitab, we find the corresponding P-value to be approximately 0. A low P-value is strong evidence to reject H_0. Thus, we overwhelmingly reject H_0, concluding that our variables are not completely independent, and proceed with PCA.

Solving the Lagrangian and eigen-problem by hand is long and tedious. Therefore, we use Minitab to expedite the process, and display the eigenvalues Minitab calculates in the table below.

Table 3: Eigenvalues of Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Eigenvalue	**5.800**	**3.774**	**1.677**	**1.395**	**0.753**	**0.486**	0.304	0.237
Proportion	**0.387**	**0.252**	**0.112**	**0.093**	**0.050**	**0.032**	0.020	0.016
Cumulative	**0.387**	**0.638**	**0.750**	**0.843**	**0.893**	**0.926**	0.946	0.962
	PC9	PC10	PC11	PC12	PC13	PC14	PC15	
Eigenvalue	0.162	0.159	0.101	0.089	0.050	0.011	0.002	
Proportion	0.011	0.011	0.007	0.006	0.003	0.001	0.000	
Cumulative	0.973	0.983	0.990	0.996	0.999	1.000	1.000	

Note that the cumulative variance accounted for by the first six PCs is above 90%, our stated goal, and thus we choose to use only these six. Another tool for deciding how many PCs to keep is the scree plot. The scree plot is a graph of the eigenvalues for each PC versus the component numbers. Typically the graph has two linear portions. At the beginning, it has a steep slope, while near the end it has a shallow slope. The eigenvalues on the shallow slope are insignificant to the total, so we can disregard them. So, by examining the scree plot for a juncture between the two slopes, called the elbow, we can locate and eliminate all of the insignificant PC values. We display the PCA scree plot below.



Looking at this plot, we see that the shallow slope starts at the 7th component, so we call it the elbow, and keep only the six principal components before it. This coincides with our original choice of six PCs. These six PCs correspond to the following vectors of coefficients. The bold numbers indicate a large effect of the given variable on the principal component.

Table 4: Principal Component Coefficients

Variables	PC1	PC2	PC3	PC4	PC5	PC6
sqrt(Teacher Salary)/20	**-0.368**	0.118	0.072	-0.001	0.308	-0.101
Log(HS Enrollment)	-0.235	**-0.407**	-0.135	-0.023	0.021	0.053
sqrt(School Districts)/2	-0.155	-0.231	**-0.488**	0.115	-0.236	**-0.492**
Log(Pupils to Teacher)	0.009	-0.226	0.167	**-0.584**	**0.543**	0.000
Log(Graduates)	-0.230	**-0.401**	-0.180	-0.008	0.002	0.080
sqrt(Personal Income)/15	**-0.348**	0.214	0.042	-0.135	-0.127	0.032
Med Income/10000	**-0.321**	0.231	-0.037	-0.214	-0.009	0.196
sqrt(% Minorities)	-0.120	-0.127	**0.586**	-0.122	-0.117	**-0.655**
Log(Higher Institutes Enrollment)	-0.271	**-0.376**	-0.049	-0.038	-0.100	0.057
(Revenue per Student)/1000	-0.277	0.254	-0.149	0.263	0.365	-0.126
Log(SAT Takers)	**-0.342**	-0.181	0.024	0.029	0.124	0.210
Log(Population per sq. mile +1)	-0.304	0.014	0.325	0.182	-0.269	**0.378**
sqrt(% with Bachelors Degree)	-0.232	0.225	0.027	**-0.392**	**-0.493**	-0.049
(% with HS Diploma)10	-0.008	0.266	-0.446	**-0.474**	0.000	-0.071
sqrt(Expenditures / Student)/10	-0.285	0.281	-0.018	0.287	0.224	-0.239

We examine these coefficients to find patterns in the relationships between the variables and the principal components. The first PC is highly affected by monetary influences and thus we label it the Income Component. We name the second PC the Personal Commitment Component since it is highly related with enrollment and academic success. The third and sixth are impacted by the number of school districts, social diversity and density, so we call them the Diversity Components. The fourth and fifth components are influenced by the education level of the previous generation and pupil to teacher ratio, so we label them the Dedication to Education Components. We have reduced our data set from fifteen to six variables, which is very helpful for further analysis. However, as an alternative to PCA, Factor Analysis provides another method for reducing the dimensionality of a data set.

Theory of Factor Analysis

While PCA creates new components in terms of the original variables, Factor Analysis (FA) expresses the original variables in terms of new factors. This process is based on the theory that there are both unobservable common factors and unique factors in every data set. The common factors affect every variable, while each unique factor is specific to one variable. The main idea is to attempt to factor the population covariance matrix Σ into $LL^T + \Psi$, where Ψ is a diagonal matrix composed of the variances of the unique factors. If Σ can be factored, then each original variable can be expressed as a linear combination of the common factors plus a unique factor. For example, $x_1 = \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \varepsilon_1$, where the F's are common factors, and ε_1 is the unique factor. The matrix L is a $p \times m$ matrix composed of the factor loadings, ℓ_{ij}. Each ℓ_{ij} is the coefficient of the j^{th} factor for the i^{th} original variable.

The goal of factor analysis is to choose m much less than p, subject to the constraint that the chosen m-Factor Model is adequate. To test for adequacy, we start with the case $m = 1$ and apply the χ^2 adequacy test shown below.

$$\chi^2 = \left[n - 1 - \frac{2p+5}{6} - \frac{2}{3}m\right]\ln\left(\frac{\det(\hat{L}\hat{L}^T + \hat{\Psi})}{\det(R)}\right)$$

In this expression, $\hat{L}\hat{L}^T + \hat{\Psi}$ is the maximum likelihood estimator of Σ. As the value of m increases, $\hat{L}\hat{L}^T + \hat{\Psi}$ becomes a better approximation for Σ. Therefore, the $\ln[\det(\hat{L}\hat{L}^T + \hat{\Psi})/\det(R)]$ approaches zero, resulting in a smaller χ^2 value. We test the null hypothesis H_0: $\Sigma = LL^T + \Psi$ against H_1: Σ is any other $p \times p$ positive definite matrix. We reject H_0 if χ^2 is greater than $\chi^2_{\alpha,\nu}$, where α is the chance of type one error, and ν is the degrees of freedom for the χ^2 test, given by $\nu = \frac{1}{2}[(p - m)^2 - p - m]$. If we reject H_0 when $m = 1$, then we proceed to the case of $m = 2$. We repeat this process until H_0 is accepted or we exhaust all possible m values subject to $m < \frac{1}{2}(2p + 1 - \sqrt{8p+1})$. We then use the model with the smallest m value for which H_0 is accepted [6].

Like in PCA, after performing FA we can examine the factor loadings for patterns in the data. High correlations allow for a straightforward classification of the factors. However, sometimes after applying FA and determining which m value is adequate, the factor loadings do not provide enough information to accurately name each factor. Since one of the goals of FA is to clearly label these factors, we may choose to perform a rotation of the factor loadings on an adequate m-model. In theory, a rotation will cause high correlations to increase and low correlations to decrease. If successful, a rotation allows for easier identification of the factors. However, by rotating the factors we risk losing accuracy. Therefore, it is left to the discretion of the researcher to decide when a rotation is desirable. In order to perform a rotation, we must first draw on some knowledge from linear algebra. A matrix T is said to be orthogonal if $TT^T = T^TT = I$. Thus if we have $\Sigma = LL^T + \Psi$, we can write this as $\Sigma = LIL^T + \Psi = LTT^TL^T + \Psi$. Now if we let LT = L^*, then $(L^*)^T = (T^T L^T)$ and $\Sigma = L^*(L^*)^T + \Psi$. Thus, we have rewritten Σ in a different factored form, where L^* is the rotated factor loadings matrix. Since $L^* \neq L$ and yet factors the matrix Σ, we can analyze the factor loadings of L^* for higher correlations to aid in naming [6].

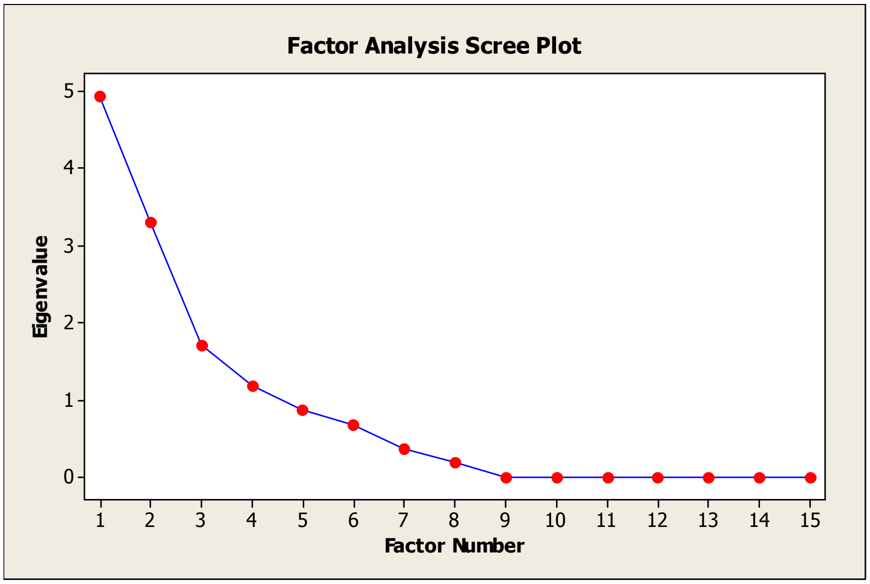
Application of Factor Analysis

As with PCA, we must be sure our variables are linearly dependent to use FA. However, since the variables have not changed and we have already shown their dependence above, we do not need to repeat the test. Using Minitab, we create the factor loadings for the m-Factor Model, starting with m = 1, and continuing until we have shown their adequacy. Since in our case we have p = 15, the adequacy test is restricted to $m < \frac{1}{2}(2p + 1 - \sqrt{8p+1}) = 10$. We set our chance of type one error to $\alpha = 0.05$, and show the adequacy test results for m = 1 through m = 8 below.

Table 5: m-Factor Model, Test for Adequacy

m-Factor Model	Degrees of Freedom	Test Statistic	Critical Value	P-value
1 Factor Model	90	1465.99	113.15	0.000
2 Factor Model	76	611.55	97.35	0.000
3 Factor Model	63	430.55	82.53	0.000
4 Factor Model	51	309.66	68.67	0.000
5 Factor Model	40	147.47	55.76	0.000
6 Factor Model	30	143.53	43.77	0.000
7 Factor Model	21	43.05	32.67	0.003
8 Factor Model	13	17.18	22.36	0.191

The smallest m that is adequate is 8, since the Test Statistic, 17.18, is less than the critical value, $\chi^2_{13,0.05} = 22.36$. Note that the corresponding P-value is indeed greater than 0.05, giving strong evidence for the acceptance of the 8-Factor Model. In addition, we construct a scree plot using Minitab to confirm our choice of 8 factors. We display the FA scree plot below.



We see that there is no elbow before the 9th factor in the scree plot, so we conclude that the 8-Factor Model is best, consistent with the results of the adequacy test. Since the 8-Factor Model is adequate and the scree plot shows similar results, we accept the 8-Factor Model. We first apply factor analysis without rotating the factors. However, the identification of the factors is not perfectly clear, so we try a varimax rotation. Although this rotation makes factor identification easier, the rotated factors are not as useful in Discriminant Analysis as the unrotated factors. Since accuracy is more important for our study than the identification of the factors, we continue our analysis with the 8-factor unrotated model. We display the corresponding unrotated factor loadings, their individual variance, their contribution to the overall variance, and the cumulative variance below.

Table 6: Factor Loadings

Variables	FL1	FL2	FL3	FL4	FL5	FL6	FL7	FL8
sqrt(Teacher Salary)/20	**0.852**	-0.349	-0.022	0.152	0.057	-0.149	0.189	-0.097
Log(HS Enrollment)	0.000	**-0.795**	**-0.607**	0.000	0.000	0.000	0.000	0.000
sqrt(School Districts)/2	0.013	-0.320	**-0.601**	0.167	0.123	**0.421**	0.157	**-0.269**
Log(Pupils to Teacher)	-0.220	-0.460	0.106	**-0.501**	0.143	-0.355	**0.465**	0.055
Log(Graduates)	0.000	**-0.729**	**-0.685**	0.000	0.000	0.000	0.000	0.000
sqrt(Personal Income)/15	**0.910**	-0.192	0.031	-0.072	-0.109	0.141	0.028	-0.001
Med Income/10000	**0.940**	-0.052	-0.122	-0.273	0.011	-0.036	-0.087	-0.019
sqrt(% Minorities)	0.093	-0.509	0.294	-0.010	-0.319	-0.178	0.008	**0.178**
Log(Higher Institutes Enrollment)	0.098	**-0.820**	**-0.522**	0.011	-0.174	0.054	0.069	-0.015
(Revenue per Student)/1000	**0.787**	0.036	-0.092	**0.400**	0.285	-0.016	0.018	0.016
Log(SAT Takers)	0.437	**-0.679**	-0.318	0.120	-0.020	-0.080	-0.030	0.128
Log(Population per sq. mile +1)	0.590	-0.334	-0.047	0.233	**-0.606**	-0.120	0.000	0.005
sqrt(% with Bachelors Degree)	0.683	-0.027	0.087	-0.311	-0.204	**0.373**	0.138	**0.248**
(% with HS Diploma)10	0.329	0.391	-0.057	**-0.473**	0.405	**0.377**	0.210	0.027
sqrt(Expenditures / Student)/10	**0.831**	0.022	0.050	**0.491**	0.162	0.020	0.050	0.056

Variance	4.927	3.297	1.71	1.172	0.863	0.684	0.357	0.199
% Variance	0.328	0.22	0.114	0.075	0.058	0.046	0.024	0.013
Cumulative % Variance	0.328	0.548	0.662	0.737	0.795	0.841	0.865	0.878

Taken together, these 8 factors account for about 88% of the total variance. As with PCA, by examining the factor loadings and their highest correlations, we can discern patterns in the data and give the factors unique names. The first factor is highly correlated with various monetary variables, so we call it the Money Factor. The second and third heavily influence various enrollment variables, so we call them the Enrollment Factors. While the remaining factors contribute a significant amount of variance and indeed are required to make the model adequate, the influences that each exert on the fifteen variables are too widely spread to be given specific names.

Theory of Discriminant Analysis

In everyday practice there are many cases of Discriminant Analysis (DA). Anytime we wish to describe something as good versus bad, or healthy versus sick, we are applying DA. The general idea of DA is to partition a large population into smaller sub-populations, Π_1 and Π_2, which are multivariate normal, each with their own covariance matrix Σ and mean vector $\underline{\mu}$. We then try to create a formula that will determine to which sub-population a new observation belongs. If we assume that the covariance matrices are equal, then we can apply Linear Discriminant Analysis. It follows that if $\Pi_1 \equiv MN(\underline{\mu}_1, \Sigma)$ and $\Pi_2 \equiv MN(\underline{\mu}_2, \Sigma)$ then $2\ln(\lambda) = 2(\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1}\underline{x} - (\underline{\mu}_1 - \underline{\mu}_2)^T\Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2)$, where λ is the ratio of the PDFs of Π_1 and Π_2. In these formulas, $\underline{y}$ indicates that $\underline{y}$ is a p x 1 vector. From here we formulate a linear expression that allows us to classify new observations into the sub-populations with a minimum of error. There are two ways of error occurring, both resulting from misclassifying an observation. We could classify the observation to Π_1 when it really belongs to Π_2 or vice versa. The probability of the first is called α_1 and the probability of the second is called α_2. It can be shown that the optimal linear classification rule is to classify $\underline{x}$ to Π_1 if $\underline{a}^T\underline{x} \geq h$ and to Π_2 otherwise, where $\underline{a}^T = (\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1}$ and h is some yet to be determined constant. It can be shown that $\underline{a}^T\underline{x} \sim N(\underline{a}^T\underline{\mu}_i, \underline{a}^T\Sigma\underline{a})$ under Π_i. We wish to choose h such that $\alpha_1 = \alpha_2$, or $P(\underline{a}^T\underline{x} \leq h | \Pi_1) = P(\underline{a}^T\underline{x} > h | \Pi_2)$. Solving this equation for h we have $h = \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1}(\underline{\mu}_1 + \underline{\mu}_2)$. Now that we have solved for h, we wish to have an exact expression for the Total Probability of Misclassification (TPM). As stated above, TPM = $\alpha = \alpha_1 + \alpha_2 = 2\alpha_1$, which can be shown equals $2\Phi(-\frac{1}{2}\Delta_p)$, where Φ is the cumulative distribution of the standard normal and $\Delta p = \sqrt{(\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)}$. The TPM is very useful in measuring error in linear two-group DA. However, most of the time we cannot assume that the two covariance matrices are equal, so we must use quadratic DA. Still, we usually apply linear DA if the hypothesis is not overwhelmingly rejected [6].

Application of Discriminant Analysis

In our study of education, we classify the states based on their average performance on the SAT into the following three categories: high, medium and low. The SAT consists of two sections testing verbal and math skills, each with a maximum 800 points for a total possible score of 1600. While the appropriateness of the SAT as a measure of academic proficiency is commonly debated, we choose it as our method of classification due to its standardized nature. Most other measures of proficiency, such as individual state tests and the ACT, are either not standardized or are not used nationally, and so are not applicable when classifying states. We could apply a two-group model of high versus low, but there is a large clustering of the scores in the middle region, making it difficult to work with a two-group model. Therefore, we choose to use three-group DA. We classify a state as high when its average SAT score is above 1100 and low when its average score is 1000 or below. Any other state we classify as medium. With this system, we classify 34 states as high, 47 as medium, and 21 as low.

In order to apply linear DA we require that the three covariance matrices corresponding to the three groups are equal. So we test the null hypothesis H_0: $\sum_1 = \sum_2 = \sum_3$ against H_1: Any two covariance matrices are unequal. We compute the pooled unbiased estimate of the common covariance matrix under H_0 given by the expression below.

$$S_p = \frac{1}{N_1 + N_2 + N_3 - 3}\left\{\sum_{i=1}^{3}(N_i - 1)S_i\right\}$$

Here, N_i is the sample size of group i, and S_i is the i^{th} sample covariance matrix. The test statistic for the equality of covariance matrices has a χ^2 distribution and is given by M/c as defined below.

$$M = \left\{\sum_{i=1}^{3}(N_i - 1)\right\}\ln(\det(S_p)) - \left\{\sum_{i=1}^{3}(N_i - 1)\ln(\det(S_i))\right\}$$

$$1/c = 1 - \frac{2p^2 + 3p - 1}{6(p+1)}\left[\left(\sum_{i=1}^{3}\frac{1}{N_i - 1}\right) - \frac{1}{N_1 + N_2 + N_3 - 3}\right]$$

We calculate our test statistic M/c to be 538.4. The degrees of freedom for this test are given by: p(p + 1)(k – 1)/2, where k is the number of groups. Therefore, we have 15(15 + 1)(3 – 1)/2 = 240 degrees of freedom. We then compare our test statistic to the critical value, $\chi^2_{240,0.01} \approx 250$. Since our test statistic is greater than $\chi^2_{240,0.01}$ and our P-value is approximately zero, we must reject H_0 and conclude that all three covariance matrices are not equal. This result gives strong evidence that we should use quadratic DA and not linear. However, it is possible that linear DA could still be effective as a predictive model. With that in mind, we analyze the data using both linear and quadratic DA.

We randomly chose 25 states to use for our test sample and let the remaining 77 states comprise our training sample. Using the six principal components as our linear DA predictors in Minitab for the training sample, we achieve a proportion correct of 0.688. This is an apparent error rate (AER) of 31.2%. Similarly, linear DA applied to the factors results in a proportion correct of 0.727, with an AER of 27.3%. Since each AER is extremely high for the training sample, we opt to not use linear DA. We would like to calculate TPM in addition to AER; however, the TPM formula cannot be generalized for more than two groups. We now continue our analysis with quadratic DA instead of linear. Applying quadratic DA to our training sample with the principal components, we achieve the following results.

Table 7: Quadratic Training (Principal Components)

	True Group High	True Group Medium	True Group Low
Classified into Group			
High	25	2	1
Medium	1	31	0
Low	0	2	15
Total N	26	35	16
N correct	25	31	15
Proportion	0.962	0.886	0.938
	N = 77	N Correct = 71	Proportion Correct = 0.922

The quadratic model has an AER of 7.8%. This is a low error rate, so we accept this model. Next, we wish to see how well our model predicts SAT scores based on observations that are not part of the training sample. For this, we use the principal components of our test sample. The results of the quadratic DA predictions for the test sample are shown below.

Table 8: Principal Components (Quadratic), Prediction Results

Observation	State	Group		Squared Distance		
		Predicted	True	To High	To Medium	To Low
1	Texas, 1998-1999	Low	Low	44.951	22.088	12.305
2	District of Columbia, 1999-2000	Low	Low	1056.733	57.839	21.246
3	South Carolina, 1999-2000	Low	Low	7.137	7.193	2.633
4	Florida, 1999-2000	Low	Low	11.774	13.209	7.573
5	New York, 1998-1999	Low	Low	14.579	22.265	10.999
6	Alaska, 1998-1999	Medium	Medium	333.581	20.421	91.794
7	Kentucky, 1998-1999	Medium	Medium	13.942	7.443	16.576
8	Maine, 1998-1999	Medium	Medium	12.727	5.188	49.294
9	Montana, 1998-1999	Medium	Medium	43.887	12.077	94.662
10	Nevada, 1998-1999	Medium	Medium	54.578	7.042	28.031
11	New Jersey, 1998-1999	Low	Medium	94.544	15.57	8.479
12	Ohio, 1998-1999	High	Medium	4.319	7.649	15.626
13	California, 1999-2000	Low	Medium	42.841	17.861	15.481
14	Kentucky, 1999-2000	Medium	Medium	14.001	6.567	16.295
15	New Jersey, 1999-2000	Low	Medium	97.81	14.801	11.238
16	New Mexico, 1999-2000	Medium	Medium	50.427	12.237	24.537
17	Virginia, 1999-2000	Medium	Medium	55.17	7.742	35.377
18	Iowa, 1998-1999	High	High	0.252	6.103	59.144
19	Michigan, 1998-1999	Low	High	32.21	14.636	7.194
20	Oklahoma, 1998-1999	High	High	2.267	17.506	45.378
21	Kansas, 1999-2000	High	High	-1.803	4.688	50.879
22	Louisiana, 1999-2000	High	High	2.001	7.968	5.899
23	Missouri, 1999-2000	High	High	-0.665	7.172	38.487
24	Tennessee, 1999-2000	Medium	High	12.005	7.042	10.384

25	Utah, 1999-2000	Medium	High	21.276	17.196	101.88

We correctly place 18 of the 25 states in the test sample using the principal component quadratic model. We notice that of the seven misclassified states, six are in a true group adjacent to the predicted group. These six also have similar squared distance values for both the true and predicted groups. Therefore, the misclassification is not extreme. However, Michigan during the 1999-2000 school year is classified as low when its true group is high. We theorize that it shares many characteristics with low performing states, such as high pupil to teacher ratio. This model results in an AER of 28%, which is high compared to the AER of the training sample. This model is a fairly accurate predictor of a state's SAT performance.

Alternatively, since we have already applied FA, we can use the factors in our 8-Factor Model as our predictors in the DA. Once again, we use Minitab to run DA on our training sample and show the results below.

Table 9: Quadratic Training (Factor Analysis)

	True Group High	True Group Medium	True Group Low
Classified Into Group			
High	26	0	0
Medium	0	33	0
Low	0	2	16
Total N	26	35	16
N correct	26	33	16
Proportion	1	0.943	1
N = 77 N Correct = 75 Proportion Correct = 0.974			

With this model we correctly classify all but two of the training sample, accounting for about 97% of the population. Thus, the FA model outperforms the PCA by more than 5%. Next, we run the test sample again, this time using the FA model, with the goal of comparing the results.

Table 10: Factor Analysis, Prediction Results

Observation	State	Group		Squared Distance		
		Predicted	True	To High	To Medium	To Low
1	Texas, 1998-1999	Low	Low	343.101	36.89	1.592
2	District of Columbia, 1999-2000	Low	Low	3075.104	83.074	25.181
3	South Carolina, 1999-2000	Low	Low	24.833	3.596	-8.761
4	Florida, 1999-2000	Low	Low	37.682	7.001	1.374
5	New York, 1998-1999	Low	Low	43.946	18.517	10.588
6	Alaska, 1998-1999	Medium	Medium	692.603	16.54	187.453
7	Kentucky, 1998-1999	Medium	Medium	2.391	-0.032	27.801
8	Maine, 1998-1999	Medium	Medium	4.21	-2.679	337.711
9	Montana, 1998-1999	Medium	Medium	113.345	3.347	472.171

10	Nevada, 1998-1999	Medium	Medium	130.333	0.296	26.799
11	New Jersey, 1998-1999	Medium	Medium	162.345	5.814	15.431
12	Ohio, 1998-1999	High	Medium	-1.934	-0.384	36.577
13	California, 1999-2000	Medium	Medium	148.375	11.882	38.371
14	Kentucky, 1999-2000	Medium	Medium	1.593	0.179	19.463
15	New Jersey, 1999-2000	Medium	Medium	165.362	5.592	39.479
16	New Mexico, 1999-2000	Medium	Medium	138.579	8.061	86.45
17	Virginia, 1999-2000	Medium	Medium	40.057	5.042	206.379
18	Iowa, 1998-1999	High	High	-7.741	0.224	205.808
19	Michigan, 1998-1999	Medium	High	24.757	15.423	38.285
20	Oklahoma, 1998-1999	High	High	2.632	7.601	141.589
21	Kansas, 1999-2000	High	High	-9.838	-4.12	206.29
22	Louisiana, 1999-2000	High	High	-5.997	-0.049	18.984
23	Missouri, 1999-2000	High	High	-7.8	-0.521	165.143
24	Tennessee, 1999-2000	High	High	-0.25	1.791	6.681
25	Utah, 1999-2000	High	High	11.372	24.48	301.78

Using the factor analysis quadratic model, we correctly place 23 of the 25 states in the test sample, resulting in an AER of 8%. In addition, we note that the two misclassified states were close to the borders of the groups. Specifically, Ohio for the 1998-1999 school year has an average SAT score of 1072. Therefore, although Ohio should be classified as medium, its score is only 28 points away from a high classification. Likewise, Michigan for the 1998-1999 school year has an average score of 1122, which is high, but only 22 points away from medium classification. Thus, none of the misclassifications are extreme. This model is an even more accurate predictor of a state's SAT performance than PCA. While the PCA model misclassified seven states, the FA model was only in error two times, and once again, the FA model outperforms the PCA model.

For confirmation of this three-group model, we choose to additionally apply two-group DA. We use the median, 1050.5, as the separation point between high and low SAT scores. We calculate the theoretical TPM to be a shockingly small 5%. However, when we apply linear and quadratic DA to the PCs and Factor Loadings, we find that the AERs are very high in comparison to the TPM. We display the 8 resulting AERs below.

Table 11: AER from Two-Group DA

	Linear	Quadratic
PCA Training	15.6	11.7
FA Training	7.8	5.2
PCA Test	36.0	28.0
FA Test	40.0	24.0

We display the detailed results for the test sample in appendix C. Based on these results, we conclude that the 5% theoretical minimum is almost impossible to attain in actual data analysis. This result confirms our original choice of a three-group model. We theorize that the two-group classification model performs so poorly because of the massive clustering of scores around the median. This made it extremely difficult for any model to predict observations that were near the median. Despite the inability to calculate the TPM for the three-group model, we are still satisfied with our three-group model because of the low observed AERs especially in comparison with the high AERs of the two group model.

Creating an accurate prediction model for average SAT scores is a step in the right direction for improving the educational system. However, we also would like to make some conclusions about what states can do in order to improve their SAT scores. In order to classify states into the groups, high, medium, and low, Minitab creates a weighted linear or quadratic function of the predictors. Unfortunately, Minitab cannot display the equation for our most accurate model, the quadratic FA model. However, Minitab does display the weights of the discriminant function for the linear FA model. Since our linear FA model is more accurate than our principal components model, we examine its weights to draw conclusions. We display the weights of each factor for each classification as well as the constant for each in the table below

Table 12: Linear FA Model Weights

	High	Medium	Low
Constant	-637.527	-640.917	-631.193
F1	-11.103	-10.943	-10.908
F2	-40.106	-39.982	-39.702
F3	8.773	9.11	9.254
F4	55.408	55.141	54.419
F5	-217.852	-216.836	-215.363
F6	-219.531	-220.368	-220.45
F7	45.643	46.751	46.884
F8	-252.649	-252.675	-251.038

As we can see in the table, the weights for Factor 3 decrease as scores increase. We recall that Factor 3, one of the Enrollment Factors, negatively influences high school enrollment, number of public school districts, number of high school graduates and enrollment in higher institutes of learning. Therefore, if high school enrollment, number of districts, number of graduates and higher institute enrollment increase in a state, the average SAT score will most likely increase. From the table, we see that the weights of Factor 4 increase as the scores increase. We also recall that Factor 4 is positively correlated with revenue per student and expenditures per student. In addition, Factor 4 is negatively correlated with average pupil to teacher ratio. Hence, if the revenue and expenditures per student increase, then the scores would probably increase. Also, if there are fewer students per teacher, the average scores would likely increase. Finally, we see that the weights of Factor 7 decrease as scores increase. We remember that Factor 7 is positively correlated with pupil to teacher ratio. This is confirmation of our conclusion that a decrease in the pupil to teacher ratio is likely to increase SAT scores. Due to the nature of the data, it is difficult to draw conclusions from the remaining weights. The positive and negative correlations of each factor combined with the factor signs make stating solid assertions nigh impossible. Nevertheless, we feel that the conclusions shown here are general enough to aid states in education planning and funding.

Conclusion

Although we started with fifteen diverse variables covering over 100 observations, we were able to reduce the number of variables to less than ten using both PCA and FA. While both methods accomplished this task, it is clear that both have advantages and disadvantages. The advantages of PCA are that it reduces the data set to fewer variables than FA and creates components that are easier to name. The disadvantages of PCA are that it does not provide as accurate results when used in DA and there is no definite way to choose the best number of components to use. The advantages of FA are that it produces more accurate results in DA and has a built-in adequacy test for determining the number of factors to use. The disadvantages of FA are that it does not reduce the dimensionality as much as PCA and the factors are more difficult to name. Using these multivariate statistical techniques we are able to create an accurate model to predict SAT scores for the upcoming year. However, there is still room for improvement and much research yet to be done. One way of continuing this work would be to look at additional variables, particularly those that are not directly related to the state education system. Some of these could include voter turnout, crime rate, computers in household, poverty level, cost of living, funding for music education, funding for extracurricular activities such as sports, and political affiliation of state officials. We could also develop our research further by including additional years in order to increase the accuracy of the model. Another way to expand this research is to change the focus of the analysis. That is, instead of examining state data, we could choose a smaller or larger scale, such as districts within a state or countries of the world. This analysis would parallel our research.

We have now successfully created an accurate model for the prediction of SAT scores based on economic, demographic and educational variables. This model can now be applied to estimate future SAT scores. Through analysis of our data and model, we see that if states increase high school enrollment, number of districts, number of graduates and enrollment in higher institutes of learning, then their average SAT scores will most likely increase. In addition, increasing education funding will probably increase scores. Finally, decreasing the average number of students per teacher will most likely increase scores. Intuitively, these conjectures make sense, because with these changes, each student receives more opportunities, choices, and individualized attention. If states apply this knowledge to their education planning decisions, they will be able to improve the quality of education and equip their population with the necessary tools for successful lives.

> *It is the supreme art of the teacher to awaken joy in creative expression and knowledge.*
>
> —Albert Einstein [8]

Acknowledgements

We would like to thank everyone involved in bringing us to Miami University for the SUMSRI program and making this great experience possible. In particular, we wish to thank Dr. Vasant Waikar for his statistical expertise and teaching, Dr. Farmer for his patient help with technical writing, and our Graduate Assistant, James Lawrence, for his support throughout the entire process. Also, thanks to our fellow statistics SUMSRI participants, Ralph Gedeon and Kendall Williams, for their encouragement and assistance. Thanks to all of the other SUMSRI students for making SUMSRI an enjoyable experience. Finally, we would like to thank Miami University for the chance to do mathematical research, and the NSA and NSF for making it all happen.

Bibliography

[1] Census.gov, *Median Income for 4-Person Families, by State*. Retrieved June 2004 from http://www.census.gov/hhes/income/4person.html

[2] Collegeboard.com, *1998 College-Bound Seniors Profile Reports by State*. Retrieved June 2004 from http://www.collegeboard.com/sat/cbsenior/yr1998/states98.html

[3] Collegeboard.com, *1999 College-Bound Seniors Profile Reports by State.* Retrieved June 2004 from http://www.collegeboard.com/sat/cbsenior/yr1999/states99.html

[4] Famous Quotations / Quotes, *Famous Quotes about Liberty*, Retrieved July 2004 from http://quotes.telemanage.ca/

[5] Deirdre A. Gaquin and Katherine A. DeBrandt (Editors), *Education Statistics of the United States (Third Edition),* Bernan (2001)

[6] Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis (Fifth Edition),* Prentice Hall (2002)

[7] James Lattin, J. Douglas Carroll, and Paul E. Green, *Analyzing Multivariate Data*, Thomson: Brooks/Cole (2003)

[8] Mostly Math Enrichment Center, *Education and Math Quotations*, Retrieved July 2004 from http://www.mostly-math.com/links_quotations.shtml

[9] Nces.ed.gov, *Digest of Education Statistics 2000*. Retrieved June 2004 from http://nces.ed.gov/programs/digest/d00/

[10] Nces.ed.gov, *Digest of Education Statistics 2001*. Retrieved June 2004 from http://nces.ed.gov/programs/digest/d01/

[11] *Rankings & Estimates: Rankings of the States 1999 and Estimates of School Statistics 2000*, National Education Association (1999)

[12] *Rankings & Estimates: Rankings of the States 2000 and Estimates of School Statistics 2001*, National Education Association (2001)

[13] *Rankings & Estimates: Rankings of the States 2001 and Estimates of School Statistics 2002*, National Education Association (2002)

[14] Barbara Ryan, Brian Joiner and Jonathon Cryer, *Minitab Handbook: Updated for Release 14 (Fifth Edition)*, Thomson: Brooks/Core (2005)

Appendix A: Using Minitab

In Minitab, there are a few ways to test for normality. The best technique is to compute the normal scores of the variable being tested, and plot them against the variable using the scatterplot option in the Graph menu. Then, the sample correlation coefficient, r_Q, should be produced and compared to the critical value. To calculate r_Q using Minitab, find the correlation between the normal scores and the original variable. This value is r_Q.

Minitab is also helpful in applying PCA. PCA is found in the stat menu under multivariate. The variables, number of principal components, and the type of analysis (correlation or covariance), are selected. Then, Minitab produces the principal component coefficients. Multiplying the $n \times p$ matrix of the original data by the $p \times m$ matrix of coefficients produces the principal components that are later used in discriminant analysis.

Factor analysis is also greatly facilitated through the use of Minitab. FA is found in the stat menu under multivariate. The variables, the number of factors, and the method of extraction (principal components or maximum likelihood) are selected. The type of rotation can also be selected. Also, the columns for storage of L are selected under the storage option of the FA menu. Minitab produces the factor loadings, L, as well as the communalities of the data. The entries, $\Psi_i = 1 -$ (communality of i^{th} variable). With this data, the matrices L^T and Ψ can be calculated and the test of adequacy for the m-Factor Model can be carried out. The factors are created by multiplying the $n \times p$ data matrix by the $p \times m$ matrix L.

Finally, discriminant analysis is carried out using the DA option under the multivariate statistics option in the stat menu. Minitab makes this process straightforward; however, the test sample results must be scored by hand.

Appendix B: Complete Data Tables

Variables	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut
sqrt(Sal)/20, 98-99	9.463	10.822	9.358	8.993	10.654	9.750	11.356
sqrt(Sal)/20, 99-00	9.577	10.870	9.441	9.136	10.939	9.768	11.378
Log(HSEnroll), 98-99	12.234	10.556	12.326	11.798	14.320	12.194	11.887
Log(HSEnroll), 99-00	12.216	10.566	12.342	11.801	14.347	12.214	11.919
sqrt(Districts)/2, 98-99	5.657	3.640	7.483	8.803	15.716	6.633	6.745
sqrt(Districts)/2, 99-00	5.657	3.640	7.483	8.803	15.700	6.633	6.745
Log(P/T), 98-99	2.773	2.868	2.944	2.785	3.109	2.912	2.632
Log(P/T), 99-00	2.721	2.839	2.965	2.667	3.045	2.857	2.632
Log(Grads), 98-99	10.530	8.813	10.397	10.132	12.497	10.529	10.251
Log(Grads), 99-00	10.504	8.831	10.457	10.135	12.623	10.524	10.297
sqrt(Pers Income)/15, 98-99	9.775	10.702	10.144	9.520	11.071	11.318	12.944
sqrt(Pers Income)/15, 99-00	10.046	11.145	10.271	9.705	11.512	11.590	13.082
Med Income/10000, 98-99	5.116	5.973	4.940	4.447	5.521	6.343	7.553
Med Income/10000, 99-00	5.241	7.029	5.304	4.667	6.310	6.286	7.551
sqrt(% Min), 98-99	6.207	6.123	6.710	5.219	7.879	5.425	5.370
sqrt(% Min), 99-00	6.164	5.857	6.148	5.030	6.804	4.615	4.775
Log(HI), 98-99	12.284	10.228	12.619	11.642	14.494	12.458	11.940
Log(HI), 99-00	12.316	10.202	12.695	11.654	14.517	12.475	11.963
Rev/Stu/1000, 98-99	5.272	8.718	5.317	5.184	6.241	6.030	10.133
Rev/Stu/1000, 99-00	5.596	8.915	5.383	5.792	7.662	7.111	11.118
Log(SAT), 98-99	8.235	8.165	9.386	7.402	11.865	9.414	10.201
Log(SAT), 99-00	8.391	8.276	9.553	7.493	11.998	9.541	10.291
Log(pop sq mile +1), 98-99	4.466	0.693	3.738	3.912	5.347	3.664	6.518
Log(pop sq mile +1), 99-00	4.466	0.693	3.761	3.912	5.366	3.689	6.519
sqrt(% Bach), 98-99	4.669	5.050	4.919	4.159	5.206	6.221	5.788
sqrt(% Bach), 99-00	4.517	5.301	4.960	4.290	5.244	5.745	5.657
% HS Dip*10, 98-99	8.110	9.280	8.310	7.890	8.040	9.040	8.370
% HS Dip*10, 99-00	7.750	9.000	8.500	8.170	8.100	9.000	8.820
sqrt(Exp/Stu)/10, 98-99	7.308	9.403	6.695	7.341	7.527	7.269	9.685
sqrt(Exp/Stu)/10, 99-00	7.033	9.517	6.920	7.443	7.958	7.884	9.927
Classification, 98-99	High	Medium	Medium	High	Medium	Medium	Medium
Classification, 99-00	High	Medium	Medium	High	Medium	Medium	Medium

Variables	Delaware	DC	Florida	Georgia	Hawaii	Idaho	Illinois	Indiana	Iowa
sqrt(Sal)/20, 98-99	10.388	10.857	9.476	9.959	10.047	9.228	10.673	10.144	9.344
sqrt(Sal)/20, 99-00	10.540	10.849	9.582	10.127	10.072	9.376	10.780	10.229	9.444
Log(HSEnroll), 98-99	10.414	9.628	13.359	12.826	10.885	11.240	13.236	12.584	11.992
Log(HSEnroll), 99-00	10.391	9.757	13.394	12.845	10.871	11.245	13.245	12.576	11.992
sqrt(Districts)/2, 98-99	2.179	0.500	4.093	6.708	0.500	5.292	15.000	8.544	9.683
sqrt(Districts)/2, 99-00	2.179	2.646	4.093	6.708	0.500	5.315	14.992	8.544	9.683
Log(P/T), 98-99	2.773	2.674	2.890	2.760	2.833	2.907	2.803	2.839	2.708
Log(P/T), 99-00	2.734	2.785	2.907	2.754	2.839	2.890	2.785	2.821	2.701
Log(Grads), 98-99	8.800	7.762	11.501	11.055	9.132	9.665	11.646	10.987	10.396
Log(Grads), 99-00	8.810	7.896	11.509	11.112	9.284	9.661	11.631	10.984	10.483
sqrt(Pers Income)/15, 98-99	11.534	12.880	10.734	10.563	10.793	9.679	11.348	10.393	10.330
sqrt(Pers Income)/15, 99-00	11.475	12.694	10.865	10.857	10.885	9.975	11.600	10.684	10.529
Med Income/10000, 98-99	6.516	6.067	5.258	5.599	6.184	4.917	6.167	5.528	5.323
Med Income/10000, 99-00	6.558	6.228	5.558	5.780	6.640	4.770	6.636	5.852	5.808
sqrt(% Min), 98-99	6.128	9.785	6.688	6.600	8.897	3.588	6.213	3.906	2.934
sqrt(% Min), 99-00	5.630	9.798	5.882	6.269	8.746	2.720	5.496	3.362	2.324
Log(HI), 98-99	10.742	11.190	13.402	12.624	11.030	11.052	13.500	12.610	12.112
Log(HI), 99-00	10.750	11.186	13.437	12.650	11.044	11.077	13.505	12.627	12.138
Rev/Stu/1000, 98-99	8.958	5.573	6.825	6.292	7.028	5.581	6.755	8.162	6.781
Rev/Stu/1000, 99-00	9.798	9.317	7.153	7.997	7.593	6.134	8.344	8.530	7.246
Log(SAT), 98-99	8.576	7.995	10.969	10.784	8.890	7.836	9.733	10.572	7.533
Log(SAT), 99-00	8.659	8.094	11.115	10.862	8.940	7.981	9.785	10.629	7.637
Log(pop sq mile +1), 98-99	5.943	9.050	5.624	4.890	5.231	2.773	5.385	5.112	3.951
Log(pop sq mile +1), 99-00	5.956	9.042	5.638	4.913	5.226	2.773	5.389	5.118	3.951
sqrt(% Bach), 98-99	4.899	6.489	4.648	4.637	5.119	4.561	5.060	4.290	4.658
sqrt(% Bach), 99-00	4.899	6.189	4.775	4.796	5.128	4.472	5.206	4.135	5.060
% HS Dip*10, 98-99	8.450	8.280	8.270	8.070	8.800	8.480	8.540	8.290	8.970
% HS Dip*10, 99-00	8.610	8.320	8.400	8.260	8.740	8.620	8.550	8.460	8.970
sqrt(Exp/Stu)/10, 98-99	8.947	9.109	7.550	7.660	7.802	7.188	8.576	8.233	7.617
sqrt(Exp/Stu)/10, 99-00	9.280	10.053	7.581	9.898	8.090	7.356	8.758	8.484	7.812
Classification, 98-99	Low	Low	Low	Low	Low	Medium	High	Low	High
Classification, 99-00	Low	Low	Low	Low	Medium	Medium	High	Low	High

Variables	Kansas	Kentucky	Louisiana	Maine	Maryland	Massachusetts	Michigan
sqrt(Sal)/20, 98-99	9.670	9.424	9.015	9.342	10.311	10.615	10.978
sqrt(Sal)/20, 99-00	9.805	9.537	9.098	9.429	10.494	10.753	11.034
Log(HSEnroll), 98-99	11.884	12.162	12.256	10.996	12.368	12.460	13.071
Log(HSEnroll), 99-00	11.894	12.153	12.248	11.010	12.386	12.488	13.084
sqrt(Districts)/2, 98-99	8.718	6.633	4.062	7.583	2.450	9.407	13.684
sqrt(Districts)/2, 99-00	8.718	6.633	4.062	7.649	2.450	9.631	13.991
Log(P/T), 98-99	2.688	2.797	2.754	2.639	2.839	2.681	2.923
Log(P/T), 99-00	2.660	2.734	2.809	2.550	2.809	2.526	2.890
Log(Grads), 98-99	10.262	10.546	10.496	9.447	10.773	10.782	11.244
Log(Grads), 99-00	10.270	10.519	10.522	9.413	10.781	10.869	11.300
sqrt(Pers Income)/15, 98-99	10.551	9.787	9.749	10.111	11.551	12.093	10.745
sqrt(Pers Income)/15, 99-00	10.814	10.047	9.954	10.375	11.900	12.380	11.129
Med Income/10000, 98-99	5.534	4.911	4.904	5.106	7.140	6.896	5.902
Med Income/10000, 99-00	5.720	5.219	4.945	5.754	7.481	7.169	6.547
sqrt(% Min), 98-99	4.407	3.405	7.094	1.735	6.705	4.788	5.028
sqrt(% Min), 99-00	3.795	3.286	6.596	1.304	6.348	4.037	4.858
Log(HI), 98-99	12.088	12.104	12.306	10.951	12.488	12.938	13.230
Log(HI), 99-00	12.082	12.110	12.308	10.965	12.502	12.947	13.234
Rev/Stu/1000, 98-99	6.740	6.546	5.983	7.422	7.966	7.941	8.887
Rev/Stu/1000, 99-00	7.286	6.846	6.367	7.690	8.324	9.532	7.150
Log(SAT), 98-99	7.887	8.585	8.268	9.203	10.441	10.760	9.316
Log(SAT), 99-00	7.998	8.614	8.324	9.288	10.533	10.839	9.435
Log(pop sq mile +1), 98-99	3.497	4.605	4.615	3.714	6.265	6.666	5.159
Log(pop sq mile +1), 99-00	3.497	4.615	4.615	3.738	6.273	6.671	5.165
sqrt(% Bach), 98-99	5.148	4.450	4.550	4.785	5.891	5.568	4.615
sqrt(% Bach), 99-00	5.225	4.472	4.796	4.899	5.657	5.718	4.796
% HS Dip*10, 98-99	8.760	7.820	7.830	8.890	8.470	8.510	8.550
% HS Dip*10, 99-00	8.810	7.870	8.080	8.930	8.570	8.500	8.600
sqrt(Exp/Stu)/10, 98-99	7.777	7.681	7.372	8.517	8.402	9.092	8.681
sqrt(Exp/Stu)/10, 99-00	8.016	7.959	7.537	8.565	8.470	9.354	8.379
Classification, 98-99	High	Medium	High	Medium	Medium	Medium	High
Classification, 99-00	High	Medium	High	Medium	Medium	Medium	High

Variables	Minnesota	Mississippi	Missouri	Montana	Nebraska	Nevada	New Hampshire
sqrt(Sal)/20, 98-99	9.932	8.592	9.320	8.854	9.066	9.859	9.670
sqrt(Sal)/20, 99-00	9.975	8.924	9.441	8.961	9.122	9.924	9.713
Log(HSEnroll), 98-99	12.505	11.827	12.479	10.829	11.423	11.312	10.968
Log(HSEnroll), 99-00	12.520	11.816	12.489	10.821	11.421	11.362	11.001
sqrt(Districts)/2, 98-99	9.301	6.164	11.456	10.654	12.207	2.062	6.384
sqrt(Districts)/2, 99-00	9.274	6.164	11.446	10.559	11.948	2.062	6.384
Log(P/T), 98-99	2.760	2.821	2.667	2.754	2.667	2.929	2.728
Log(P/T), 99-00	2.721	2.791	2.660	2.721	2.632	2.929	2.688
Log(Grads), 98-99	10.912	10.086	10.859	9.301	9.917	9.383	9.283
Log(Grads), 99-00	10.925	10.095	10.868	9.293	9.903	9.575	9.285
sqrt(Pers Income)/15, 98-99	11.089	9.189	10.424	9.486	10.496	11.027	11.396
sqrt(Pers Income)/15, 99-00	11.571	9.470	10.711	9.778	10.886	11.330	11.679
Med Income/10000, 98-99	6.714	4.391	5.419	4.474	5.669	5.305	6.101
Med Income/10000, 99-00	6.668	4.792	5.667	5.097	5.569	5.948	6.589
sqrt(% Min), 98-99	3.801	7.230	4.452	3.627	3.893	6.227	1.959
sqrt(% Min), 99-00	2.470	7.490	4.074	2.702	2.933	4.754	1.414
Log(HI), 98-99	12.539	11.794	12.649	10.695	11.618	11.329	11.015
Log(HI), 99-00	12.552	11.799	12.668	10.672	11.616	11.404	11.057
Rev/Stu/1000, 98-99	7.727	4.925	6.597	6.599	5.879	6.472	7.299
Rev/Stu/1000, 99-00	8.378	5.706	7.401	6.765	6.415	6.942	7.819
Log(SAT), 98-99	8.593	7.032	8.505	7.894	7.532	8.351	9.183
Log(SAT), 99-00	8.703	7.218	8.603	7.908	7.651	8.508	9.265
Log(pop sq mile +1), 98-99	4.094	4.094	4.382	1.946	3.136	2.833	4.890
Log(pop sq mile +1), 99-00	4.111	4.094	4.382	1.946	3.136	2.890	4.905
sqrt(% Bach), 98-99	5.657	4.382	4.796	4.899	4.517	4.494	5.215
sqrt(% Bach), 99-00	5.586	4.324	5.119	4.879	4.960	4.393	5.477
% HS Dip*10, 98-99	9.110	7.800	8.500	8.880	8.930	8.640	8.650
% HS Dip*10, 99-00	9.080	8.030	8.660	8.900	9.040	8.280	8.810
sqrt(Exp/Stu)/10, 98-99	8.377	6.811	7.426	7.675	7.514	7.439	8.001
sqrt(Exp/Stu)/10, 99-00	8.659	7.150	7.647	7.830	7.751	7.563	8.233
Classification, 98-99	High	High	High	Medium	High	Medium	Medium
Classification, 99-00	High	High	High	Medium	High	Medium	Medium

Variables	New Jersey	New Mexico	New York	North Carolina	North Dakota	Ohio
sqrt(Sal)/20, 98-99	11.313	9.000	11.117	9.500	8.511	10.071
sqrt(Sal)/20, 99-00	11.421	9.021	11.200	9.927	8.641	10.178
Log(HSEnroll), 98-99	12.715	11.475	13.652	12.719	10.545	13.201
Log(HSEnroll), 99-00	12.723	11.471	13.658	12.740	10.540	13.200
sqrt(Districts)/2, 98-99	12.186	4.717	13.276	5.408	7.566	12.359
sqrt(Districts)/2, 99-00	12.186	4.717	13.276	5.408	7.566	12.359
Log(P/T), 98-99	2.588	2.803	2.646	2.760	2.667	2.803
Log(P/T), 99-00	2.595	2.797	2.660	2.747	2.625	2.760
Log(Grads), 98-99	11.157	9.732	11.833	11.012	9.032	11.653
Log(Grads), 99-00	11.163	9.743	11.838	10.997	9.028	11.653
sqrt(Pers Income)/15, 98-99	12.284	9.430	11.866	10.354	9.822	10.591
sqrt(Pers Income)/15, 99-00	12.413	9.643	12.041	10.607	10.122	10.899
Med Income/10000, 98-99	7.098	4.383	5.714	5.433	5.100	6.017
Med Income/10000, 99-00	7.543	4.495	5.976	5.612	5.100	5.624
sqrt(% Min), 98-99	6.199	7.927	6.664	6.121	3.183	4.304
sqrt(% Min), 99-00	5.559	7.543	5.621	5.621	2.757	4.111
Log(HI), 98-99	12.694	11.599	13.830	12.867	10.583	13.203
Log(HI), 99-00	12.709	11.625	13.836	12.889	10.605	13.215
Rev/Stu/1000, 98-99	10.128	6.324	9.769	6.463	5.960	7.455
Rev/Stu/1000, 99-00	10.326	6.823	10.744	6.572	6.493	8.345
Log(SAT), 98-99	11.055	7.663	11.745	10.597	6.144	10.334
Log(SAT), 99-00	11.128	7.861	11.843	10.671	6.248	10.447
Log(pop sq mile +1), 98-99	6.999	2.708	5.956	5.050	2.303	5.617
Log(pop sq mile +1), 99-00	7.002	2.708	5.956	5.063	2.303	5.620
sqrt(% Bach), 98-99	5.523	4.950	5.187	4.889	4.722	5.050
sqrt(% Bach), 99-00	5.486	4.858	5.357	4.796	4.754	5.000
% HS Dip*10, 98-99	8.740	8.090	8.190	7.980	8.490	8.610
% HS Dip*10, 99-00	8.730	8.000	8.250	8.000	8.550	8.700
sqrt(Exp/Stu)/10, 98-99	9.850	7.533	9.680	7.598	6.780	7.947
sqrt(Exp/Stu)/10, 99-00	10.009	7.769	9.898	7.694	6.715	8.218
Classification, 98-99	Medium	Medium	Low	Low	High	Medium
Classification, 99-00	Medium	Medium	Low	Low	High	Medium

Variables	Oklahoma	Oregon	Pennsylvania	Rhode Island	South Carolina	South Dakota
sqrt(Sal)/20, 98-99	8.825	10.348	11.007	10.683	9.288	8.449
sqrt(Sal)/20, 99-00	8.846	10.114	10.991	10.845	9.498	8.525
Log(HSEnroll), 98-99	12.104	12.002	13.216	10.653	12.138	10.636
Log(HSEnroll), 99-00	12.102	12.023	13.226	10.667	12.118	10.632
sqrt(Districts)/2, 98-99	12.000	7.018	11.180	3.000	4.690	6.577
sqrt(Districts)/2, 99-00	11.662	7.036	11.180	3.000	4.690	6.577
Log(P/T), 98-99	2.741	2.918	2.797	2.565	2.741	2.674
Log(P/T), 99-00	2.715	2.976	2.766	2.653	2.688	2.639
Log(Grads), 98-99	10.514	10.247	11.627	8.978	10.434	9.088
Log(Grads), 99-00	10.529	10.247	11.651	8.968	10.431	9.130
sqrt(Pers Income)/15, 98-99	9.674	10.493	10.932	10.939	9.750	9.933
sqrt(Pers Income)/15, 99-00	10.017	10.789	11.150	11.118	10.089	10.433
Med Income/10000, 98-99	4.744	5.589	5.851	6.234	5.211	4.970
Med Income/10000, 99-00	5.226	5.391	5.955	6.461	5.598	5.225
sqrt(% Min), 98-99	5.742	4.134	4.543	4.861	6.653	3.534
sqrt(% Min), 99-00	4.583	3.194	3.950	3.479	6.738	3.066
Log(HI), 98-99	12.093	12.050	13.298	11.211	12.108	10.635
Log(HI), 99-00	12.095	12.076	13.314	11.223	12.121	10.649
Rev/Stu/1000, 98-99	5.713	7.033	8.439	8.251	6.624	6.196
Rev/Stu/1000, 99-00	5.797	8.068	9.012	8.551	7.110	6.613
Log(SAT), 98-99	8.008	9.697	11.417	8.830	10.042	6.172
Log(SAT), 99-00	8.144	9.789	11.473	8.884	10.108	6.215
Log(pop sq mile +1), 98-99	3.912	3.555	5.595	6.853	4.852	2.398
Log(pop sq mile +1), 99-00	3.912	3.584	5.595	6.855	4.868	2.398
sqrt(% Bach), 98-99	4.868	5.177	4.889	5.177	4.572	5.060
sqrt(% Bach), 99-00	4.690	5.215	4.930	5.138	4.359	5.070
% HS Dip*10, 98-99	8.350	8.620	8.610	8.090	7.860	8.870
% HS Dip*10, 99-00	8.610	8.800	8.570	8.130	8.300	9.180
sqrt(Exp/Stu)/10, 98-99	7.282	8.263	8.458	8.998	7.623	7.265
sqrt(Exp/Stu)/10, 99-00	7.351	8.423	8.732	9.327	7.819	7.391
Classification, 98-99	High	Medium	Low	Medium	Low	High
Classification, 99-00	High	Medium	Low	Medium	Low	High

Variables	Tennessee	Texas	Utah	Vermont	Virginia
sqrt(Sal)/20, 98-99	9.553	9.360	9.076	9.592	9.679
sqrt(Sal)/20, 99-00	9.530	9.691	9.347	9.710	9.763
Log(HSEnroll), 98-99	12.392	13.890	11.936	10.369	12.640
Log(HSEnroll), 99-00	12.436	13.907	11.926	10.382	12.666
sqrt(Districts)/2, 98-99	5.895	16.140	3.162	8.761	5.831
sqrt(Districts)/2, 99-00	5.874	17.197	3.162	8.746	5.745
Log(P/T), 98-99	2.827	2.721	3.096	2.580	2.646
Log(P/T), 99-00	2.715	2.701	3.091	2.510	2.639
Log(Grads), 98-99	10.733	12.215	10.350	8.670	11.042
Log(Grads), 99-00	10.720	12.252	10.385	8.763	11.095
sqrt(Pers Income)/15, 98-99	10.245	10.547	9.683	10.375	11.053
sqrt(Pers Income)/15, 99-00	10.482	10.805	9.963	10.649	11.394
Med Income/10000, 98-99	5.031	5.115	5.495	5.369	6.086
Med Income/10000, 99-00	5.200	5.329	5.725	5.771	6.435
sqrt(% Min), 98-99	5.135	7.474	3.472	1.706	5.926
sqrt(% Min), 99-00	4.848	7.000	2.510	1.265	5.235
Log(HI), 98-99	12.435	13.798	11.927	10.520	12.822
Log(HI), 99-00	12.441	13.806	11.993	10.511	12.843
Rev/Stu/1000, 98-99	4.968	6.552	4.809	7.620	5.899
Rev/Stu/1000, 99-00	5.710	7.251	5.377	9.076	5.862
Log(SAT), 98-99	8.830	11.517	7.299	8.484	10.727
Log(SAT), 99-00	8.980	11.639	7.474	8.542	10.792
Log(pop sq mile +1), 98-99	4.890	4.331	3.296	4.174	5.153
Log(pop sq mile +1), 99-00	4.898	4.357	3.296	4.174	5.165
sqrt(% Bach), 98-99	4.207	4.940	5.282	5.320	5.621
sqrt(% Bach), 99-00	4.690	4.899	5.070	5.385	5.648
% HS Dip*10, 98-99	7.910	7.820	9.100	8.930	8.730
% HS Dip*10, 99-00	8.000	7.920	9.000	9.000	8.660
sqrt(Exp/Stu)/10, 98-99	7.187	7.564	6.331	8.564	7.774
sqrt(Exp/Stu)/10, 99-00	7.340	7.953	6.458	8.904	7.842
Classification, 98-99	High	Low	High	Medium	Medium
Classification, 99-00	High	Low	High	Medium	Medium

Variables	Washington	West Virginia	Wisconsin	Wyoming
sqrt(Sal)/20, 98-99	9.835	9.253	10.082	9.152
sqrt(Sal)/20, 99-00	10.126	9.355	10.143	9.239
Log(HSEnroll), 98-99	12.619	11.426	12.538	10.351
Log(HSEnroll), 99-00	12.641	11.389	12.547	10.324
sqrt(Districts)/2, 98-99	8.602	3.708	10.320	3.464
sqrt(Districts)/2, 99-00	8.602	3.708	10.320	3.464
Log(P/T), 98-99	3.006	2.667	2.741	2.653
Log(P/T), 99-00	2.991	2.625	2.667	2.588
Log(Grads), 98-99	10.926	9.886	10.942	8.754
Log(Grads), 99-00	10.972	9.897	10.975	8.753
sqrt(Pers Income)/15, 98-99	11.169	9.279	10.580	10.160
sqrt(Pers Income)/15, 99-00	11.505	9.596	10.927	10.741
Med Income/10000, 98-99	6.106	4.324	5.789	5.099
Med Income/10000, 99-00	6.262	4.520	6.344	5.562
sqrt(% Min), 98-99	4.884	2.255	4.255	3.379
sqrt(% Min), 99-00	3.937	2.025	3.661	3.050
Log(HI), 98-99	12.608	11.386	12.642	10.299
Log(HI), 99-00	12.634	11.393	12.627	10.275
Rev/Stu/1000, 98-99	7.009	8.047	8.354	8.061
Rev/Stu/1000, 99-00	7.395	7.965	8.870	8.569
Log(SAT), 98-99	10.258	8.257	8.381	6.543
Log(SAT), 99-00	10.316	8.311	8.459	6.627
Log(pop sq mile +1), 98-99	4.454	4.331	4.575	1.792
Log(pop sq mile +1), 99-00	4.477	4.331	4.585	1.792
sqrt(% Bach), 98-99	5.348	4.231	4.858	4.722
sqrt(% Bach), 99-00	5.292	3.912	4.899	4.472
% HS Dip*10, 98-99	9.120	7.510	8.680	9.070
% HS Dip*10, 99-00	9.180	7.710	8.670	9.000
sqrt(Exp/Stu)/10, 98-99	7.776	8.439	8.659	8.252
sqrt(Exp/Stu)/10, 99-00	7.958	8.664	8.818	8.693
Classification, 98-99	Medium	Medium	High	Medium
Classification, 99-00	Medium	Medium	High	Medium

Appendix C: Two-Group DA Test Prediction Results

Observation	State	True Group	Predicted Group			
			Linear PC	Quadratic PC	Linear FA	Quadratic FA
1	California, 1998-1999	L	H	L	H	L
2	Indiana, 1998-1999	L	L	L	L	L
3	Mississippi, 1998-1999	H	H	H	H	H
4	New Hampshire, 1998-1999	L	H	L	H	L
5	New Mexico, 1998-1999	H	H	L	L	L
6	New York, 1998-1999	L	L	L	L	L
7	North Carolina, 1998-1999	L	L	H	L	L
8	North Dakota, 1998-1999	H	H	H	H	H
9	Wisconsin, 1998-1999	H	L	H	H	H
10	Arizona, 1999-2000	L	H	H	H	H
11	California, 1999-2000	L	H	L	H	L
12	Colorado, 1999-2000	H	H	H	L	H
13	Georgia, 1999-2000	L	L	L	L	L
14	Kansas, 1999-2000	H	H	H	H	H
15	Louisiana, 1999-2000	H	L	H	L	L
16	Maryland, 1999-2000	L	L	L	L	L
17	Michigan, 1999-2000	H	H	H	H	H
18	New Mexico, 1999-2000	H	H	L	L	L
19	North Carolina, 1999-2000	L	L	H	L	L
20	Ohio, 1999-2000	H	L	H	H	H
21	Oregon, 1999-2000	H	L	L	L	H
22	Pennsylvania, 1999-2000	L	L	L	L	H
23	Rhode Island, 1999-2000	L	L	L	L	L
24	Utah, 1999-2000	H	H	H	H	H
25	Wyoming, 1999-2000	H	L	L	L	L
	# Correct		16	18	15	19
	Proportion Correct		0.64	0.72	0.6	0.76
	AER		36	28	40	24

Appendix D: Average SAT Scores by State and Year and Overall Ranking over two Years

1998-1999			1999-2000		
State	Average SAT Score	Ranking	State	Average SAT Score	Ranking
Alabama	1116	29	Alabama	1114	31
Alaska	1030	60	Alaska	1034	59
Arizona	1049	53	Arizona	1044	54
Arkansas	1119	26	Arkansas	1117	28
California	1011	72	California	1015	70
Colorado	1076	45	Colorado	1071	48
Connecticut	1019	67	Connecticut	1017	68
Delaware	1000	82	Delaware	998	85
District of Columbia	972	99	District of Columbia	980	97
Florida	997	87	Florida	998	85
Georgia	969	100	Georgia	974	98
Hawaii	995	89	Hawaii	1007	77
Idaho	1082	43	Idaho	1081	44
Illinois	1154	11	Illinois	1154	11
Indiana	994	91	Indiana	999	84
Iowa	1192	3	Iowa	1189	4
Kansas	1154	11	Kansas	1154	11
Kentucky	1094	37	Kentucky	1098	35
Louisiana	1119	26	Louisiana	1120	25
Maine	1010	74	Maine	1004	80
Maryland	1014	71	Maryland	1016	69
Massachusetts	1022	64	Massachusetts	1024	63
Michigan	1122	24	Michigan	1126	22
Minnesota	1184	5	Minnesota	1175	8
Mississippi	1111	33	Mississippi	1111	33
Missouri	1144	16	Missouri	1149	15
Montana	1091	39	Montana	1089	42
Nebraska	1139	17	Nebraska	1131	20
Nevada	1029	61	Nevada	1027	62
New Hampshire	1038	57	New Hampshire	1039	55
New Jersey	1008	76	New Jersey	1011	72
New Mexico	1091	39	New Mexico	1092	38

New York	997	87	New York	1000	82
North Carolina	986	96	North Carolina	988	95
North Dakota	1199	1	**North Dakota**	1197	2
Ohio	1072	46	Ohio	1072	46
Oklahoma	1127	21	Oklahoma	1123	23
Oregon	1050	52	Oregon	1054	49
Pennsylvania	993	92	Pennsylvania	995	89
Rhode Island	1003	81	Rhode Island	1005	79
South Carolina	954	102	**South Carolina**	966	101
South Dakota	1173	10	South Dakota	1175	8
Tennessee	1112	32	Tennessee	1116	29
Texas	993	92	Texas	993	92
Utah	1138	19	Utah	1139	17
Vermont	1020	66	Vermont	1021	65
Virginia	1007	77	Virginia	1009	75
Washington	1051	51	Washington	1054	49
West Virginia	1039	55	West Virginia	1037	58
Wisconsin	1179	7	Wisconsin	1181	6
Wyoming	1097	36	Wyoming	1090	41