Education By Nation: A Multivariate Statistical Analysis

Ashley Brooks
Winston Salem State Univ.
Winston Salem, NC 27110
Abroo811@wssu.edu

Amber Shoecraft
Johnson C. Smith Univ.
Charlotte, NC 28216
jcsumath@aol.com

Anthony Franklin
Coastal Carolina Univ.
Myrtle Beach, SC 29588
amfrankl@coastal.edu

Abstract

We analyze education systems of 64 countries using multivariate statistical techniques such as principal component analysis, factor analysis, and discriminant analysis. Our goal is to classify countries into two populations, one where the educational system of the country is exceptional and the other where the educational system is fair. Reducing the dimensionality of the data set simplifies this process.

Education is our passport to the future, for tomorrow belongs to the people who prepare for it today.
- Malcolm X

Introduction

Do you wonder if the education you or your children are receiving in your state is highly rated? What about how it compares to the school systems around the world? The United States consists of nearly 298 million citizens out of 6.5 billion people in the world. Although we are among the most populated nations, how does the United States rank in terms of education?

In this paper we analyze the education systems of a collection of nations by using multivariate statistical techniques. We use principal component analysis, factor analysis, and discriminant analysis to develop and analyze our data. Principal component analysis helps to reduce the dimensionality of our data set. Factor analysis helps detect unobservable factors, or those that cannot easily be measured. Discriminant analysis allows us to place countries in their proper categories. To predict the quality of education systems we measure factors such as: enrollment, student-teacher ratio, civic knowledge and time spent during lessons. There remain other variables to measure, and it is by analysis that we decide if some of the data is redundant.

To obtain our data set, we have searched in many national census statistical databases. One website is the National Center for Education Statistics, which is funded and run by the United States Department of Education. This link contains information about international education systems. Since we cannot get data from every nation, we select the countries in the data set based on school life expectancy, the expected number of years of schooling to be completed by the student. To further reduce the number of countries, we have excluded them based on lack of information. We focus on primary and secondary institutions around the world.

Normality Assessment

The statistical techniques used in this project, principal component analysis and factor analysis, require a few assumptions on the data set before their application. When the data is multivariate normal the interpretations from these techniques have a strong foundation. Thus, to assess multivariate normality we test the univariate normality for each variable. If the variable is not normal, we use transformations to force normality. If transformations do not succeed then rejection of the specific variable may be desirable. A variable is considered to be normal if the probability density function is a bell shaped curve.

To test normality, we use a Quantile-Quantile plot (Q-Q Plot). The Q-Q plot is a scatter plot that displays the ordered data of the x quantiles versus the normal quantiles. The normal quantiles are the values from the observations if the variables are normally distributed. So we look for a linear relationship among the plotted values. A variable that is normally distributed with mean μ and standard deviation σ has corresponding standard normal scores expressed as $Z = (x- \mu)/ \sigma$. When we solve for x we obtain $x = Z\sigma + \mu$, which is a linear equation.

In most cases the human eye is not the best predictor of strong linear relationships. So to verify the linearity of a Q-Q plot we use the correlation coefficient test for normality. We analyze the sample correlation coefficient r_Q between x_i and z_i. We can express r_Q as

$$r_Q = \frac{Cov(x_i, z_i)}{\sqrt{Var(x_i)Var(z_i)}}$$

Where x_i is a sample quantile and z_i is the corresponding normal quantile. The coefficient r_Q has values between zero and one. We aim to have the value of r_Q be as close to one as possible indicating a strong linear relationship. In this test, the null hypothesis is $\rho = 1$ and the alternative hypothesis is $\rho < 1$. The given sample size and desired level of significance determines the critical value c for the test. Table 4.2 from [2] gives a list of the critical points. We accept the null hypothesis if $r_Q \geq c$, thus implying normality.

For our data set we have a sample size n=64, with 21 variables, and we use a significance level $\alpha = 0.01$. The corresponding critical value is c=0.9720. From our list of twenty one variables, thirteen fail to reject H_o: $\rho = 1$ and are considered to be normally distributed. These normal variables include school life expectancy, duration of school cycle, school starting age for primary schools, secondary school starting age, average 8th grade mathematics and science scores, percentage of parents with at least a secondary education, number of teaching hours per year for primary schools, number of teaching hours per year for secondary, average class size, average teacher salary for primary schools, number of days of instruction for primary schools, number of days of instruction

for secondary schools, and secondary graduation rate. Table 1 displays the correlation coefficients for the normal variables.

Table 1: Correlation Coefficients for Normality Test

Variables	Correlation
School Life Expectancy	0.992
Average Gross Enrollment for Primary and Secondary	0.957
Primary School Teacher/Student Ratio	0.919
Duration of School Cycle	0.998
School Starting Age Primary	0.998
Secondary School Teacher/Student Ratio	0.914
Secondary School Starting Age	1
Total Percentage of Repeaters in School	0.81
Average 8th grade Math/Science Scores	0.98
Percentage of Parents with at least Secondary Education	0.974
Number of Teaching Hours per year (Secondary)	0.991
Number of Teaching Hours per year (Primary)	0.987
Education Expenditure as % of GDP	0.936
Average Class Size	0.989
Average Teacher Salary (Primary)	0.987
Average Teacher Salary (Secondary)	0.971
Number of Days Instruction (Primary)	0.977
Number of Days Instruction (Secondary)	0.977
Secondary Graduation Rate	0.993
Households with calculators, computers, dictionaries, 25 + books and study desks	0.936

The remaining eight variables rejected $H_{o:}\ \rho = 1$. Transformations are needed to be performed in order to obtain normality for these variables. Three of the remaining variables test normal after a log transformation. These include secondary school student/teacher ratio, total percentage of repeaters in school, and education expenditures as a percent of GDP. Three variables test normal after a square root transformation. These include average gross enrollment for primary and secondary schools, primary school teacher/student ratio, and average teacher salary for secondary schools. Table 2 displays the correlation coefficients of the variables before and after the transformation.

Table 2: Correlation Coefficients for Transformed Variables

Variable	Transformation	Correlation	New Correlation
Average Gross Enrollment for Primary and Secondary	square root	0.957	0.967
Primary School Teacher/Student Ratio	square root	0.919	0.96
Secondary School Teacher/Student Ratio	log 10	0.914	0.97
Total Percentage of Repeaters in School	log 10 (x+1)	0.81	0.977
Education Expenditures as % of GDP	log 10	0.936	0.978
Average Teacher Salary Secondary	square root	0.971	0.986
Households with calculators, computers etc.	none	0.936	N/A

Figure 1 displays the Q-Q plot of the variable average teacher salary before the transformation. Figure 2 is the Q-Q plot of the transformed variable. Although the overall reading literacy rate and the number of households with calculators, computers, dictionaries, twenty five plus books, and study desks does not test normal but it is important to our analysis so we do not discard it.

Figure 1: Before

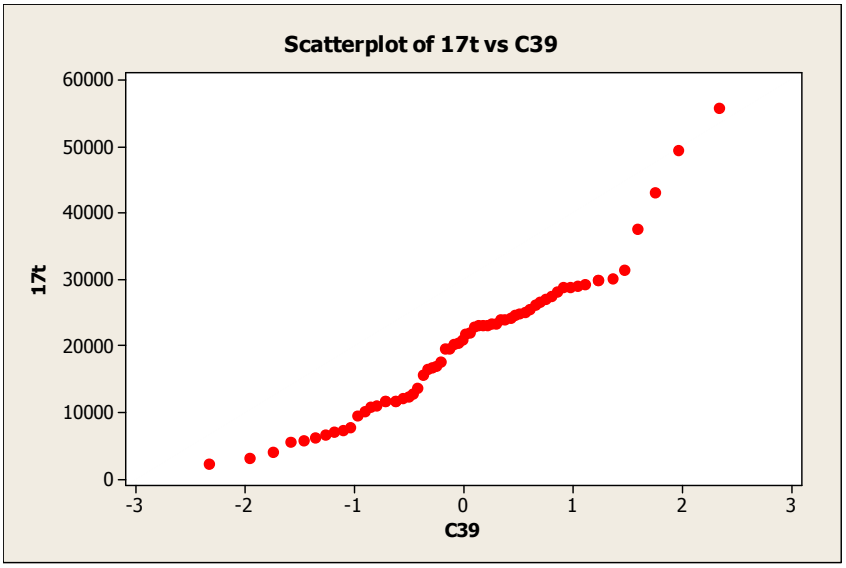
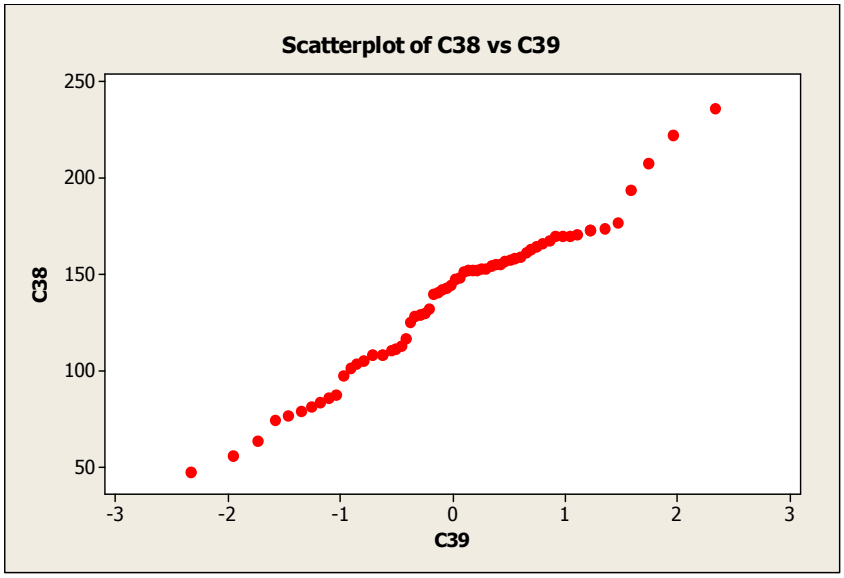


Figure 2: After



Principal Component Theory

In principal component analysis there are two objectives: data reduction and interpretation of the components. The method uses linear combinations of the variables to explain the variance- covariance structure of the variables. Using the principal components is a way of taking intermediate steps to further investigation.

Let $\underline{\mathbf{X}}$ be a multivariate normal random vector with mean $\underline{\mu}$ and covariance matrix $\mathbf{\Sigma}$. We want linear combinations of the p random variables that represent the rotated coordinate system in the directions that maximize the variability of the data set. Let

$$Y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p = \underline{\mathbf{a}}^T{}_1\underline{\mathbf{X}}$$
$$Y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p = \underline{\mathbf{a}}^T{}_2\underline{\mathbf{X}}$$
$$\vdots$$
$$Y_p = a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p = \underline{\mathbf{a}}^T{}_p\underline{\mathbf{X}}$$

We want to choose linear coefficients $\underline{\mathbf{a}}_i$ such that it accounts for the maximum variance possible. Thus we define principal components as the uncorrelated linear combinations $Y_1, Y_2, \ldots Y_p$ whose variances are as large as possible.

We define the first principal component Y_1 to be the linear combination with the maximum explained variance. We call the second principal component Y_2 as the linear combination of the random variables that explains the maximum amount of variance remaining. This implies that the p^{th} principal component explains the least amount of the variance.

In practice we want to maximize the variance with respect to a_1 so that we have a unique solution. For a nonnull solution to exist we must satisfy the equation $|\Sigma - \lambda_1 I| = 0$. Since $\underline{\mathbf{X}}$ is a random vector with a covariance matrix $\sum$ it has characteristic roots $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Notice $(\Sigma - \lambda_1 I)\underline{\mathbf{a}}_1 = 0$ implies $\underline{\mathbf{a}}_1$ is a characteristic vector of $\mathbf{\Sigma}$ corresponding to the characteristic root λ_1. Thus it is convenient to normalize the coefficients vectors. The contribution of the first principal component Y_1 to the total variation is given by

$$\frac{Explained}{TotalVariance} = \frac{\lambda_1}{\sum_1^p \sigma_{ii} = \sum_1^p \lambda_i} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

It can be shown that the j^{th} principal component can be expressed as $Y_j = \underline{\mathbf{a}}^T{}_j\underline{\mathbf{x}}$, where $\underline{a}_j$ is given by the characteristic vector corresponding to λ_j, the j^{th} largest characteristic root of $\mathbf{\Sigma}$.

Principal Component Analysis Application

The objective is to reduce the amount of variables yet still explaining most of the variance. The Minitab software has the capability of running the principal component analysis. From the Minitab software, we are able to explain 65% of the variance using

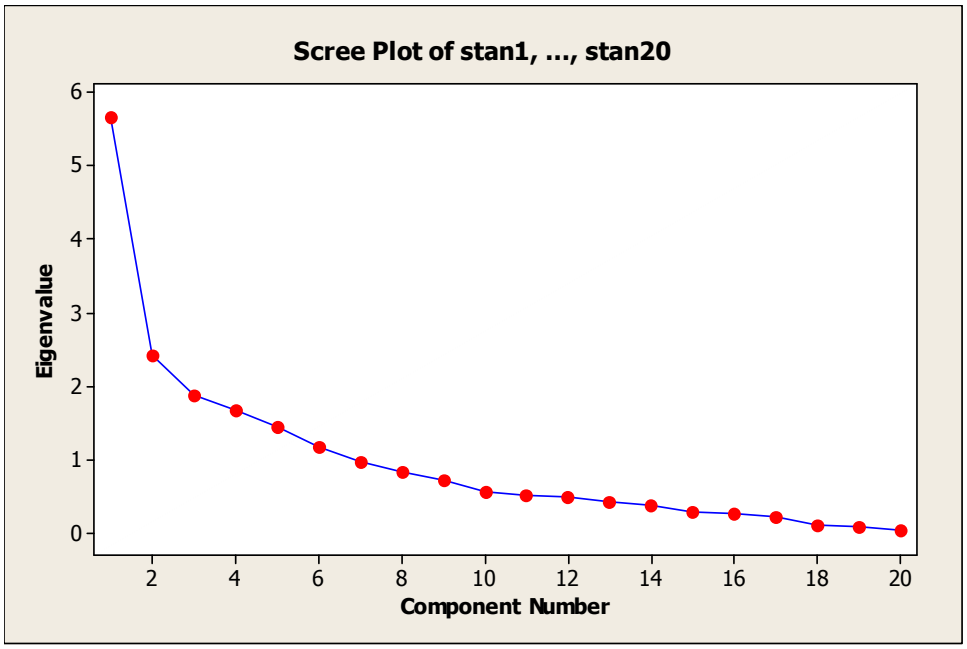
only 5 principal components. This reduces the dimensionality of the data set by 75%. Table 3 shows the coefficients of the principal components.

Table 3: Coefficients for Principal Components

Variable	PC 1	PC 2	PC 3	PC 4	PC 5
Sch Life Exp	0.347767	0.040706	0.291958	0.057661	-0.103028
Avg. Enroll (Prm./ Sec.)	0.266987	0.021487	0.347139	0.178275	0.032899
Prm. School Tch/Std Ratio	-0.253601	0.269996	-0.069009	-0.284135	0.341077
Duration of Sch Cycle	0.170707	0.154802	0.07601	-0.36037	-0.484636
School Starting Age Prm.	-0.036519	-0.367834	-0.070294	0.096943	0.072172
Sec. Sch Tch/Std Ratio	-0.249474	0.228222	-0.105736	-0.175063	0.320655
Sec. Sch Starting Age	0.016758	0.16157	0.386814	-0.213175	-0.171166
Total % of Rep in Sch	-0.257647	0.034617	0.137388	0.066337	-0.185317
Avg. 8th gr Math/Sci	0.320306	0.005756	0.028226	-0.130357	0.169914
% of Prt at least Sec Edu.	0.280654	0.036959	-0.026414	0.027371	0.075141
Avg. Adult Literacy	0.300416	-0.137012	0.084799	0.196198	0.077709
# of Tch Hr per yr (Sec.)	-0.031669	0.418187	-0.158712	0.317834	-0.228917
# of Tch Hr per yr (Prm.)	0.02273	0.331358	-0.313531	0.389576	-0.240937
Edu. Exp as % of GDP	0.107	0.266856	-0.087061	0.335856	0.072294
Avg. Class Size	-0.090705	-0.430754	-0.189749	0.268107	-0.015514
Avg Tch Salary (Prm.)	0.256764	-0.025733	-0.466645	-0.256996	-0.126242
Avg Tch Salary (Sec.)	0.242008	-0.060717	-0.438395	-0.279793	-0.162393
# of Days Inst. (Prm./Sec.)	0.132028	0.344781	0.020219	0.006224	0.29783
Sec. Graduation Rate	0.328777	0.036131	-0.038277	-0.08541	0.283436
Households Items	0.203121	0.013781	-0.107285	0.137968	0.312417

Also, to choose how many principal components is sufficient we glance at the scree plot. We look for the elbow of the graph. The elbow is the most distinct turning point from the vertical to horizontal of the given plots. Figure 3 shows the scree plot for the principal component analysis. The scree plot shows there is a range of components that can be considered the elbow. We chose component five.

Figure 3



The 5 principal components have been selected. Now they must be labeled. Principal component 1 is labeled as 'expected student performance' based on the high coefficient values of graduation rate and school life expectancy. Principal component 2 is labeled as 'teacher importance' because of the high coefficient value of teaching salary. Principal component 3 is labeled as 'parent's participation' based on the high coefficient value of average gross enrollment for primary and secondary schools. Principal component 4 is labeled as 'overall importance of education' based on the high coefficient value of education expenditure as a percent of GDP. Principal component 5 is labeled as 'student/teacher relationship' based on the high coefficient value of student/teacher ratio. Overall the first 5 principal components account for nearly 65% of the variance in the data. Since the data set consists of measurements of social sciences, 65% is sufficient in explaining the variance. Appendix A gives a list of the labels corresponding to the principal components.

Factor Analysis Theory

The ultimate goal of factor analysis is to explain the covariance relationships among the variables in terms of some unobservable and non measurable random factors. Factor Analysis is a means of describing groups of highly correlated variables by a single underlying construct, or factor that is responsible for the observed correlations. Then, once the groups of correlated variables are identified we must interpret and label each factor.

Consider an observable multivariable normally distributed random vector $\underline{\mathbf{X}}$ with mean $\underline{\boldsymbol{\mu}}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\underline{\mathbf{X}}$ consist of *p* random variables. We want to create a factor model that expresses X_i as a linear combination of common factors F_1, F_2, …,F_m , les and p additional terms ε_1, ε_2, …,εp called errors or specific factors. Let the factor model be:

$$\begin{aligned} X_1 &= \mu_1 + a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + \varepsilon_1 \\ X_2 &= \mu_2 + a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + \varepsilon_2 \\ &\vdots \qquad\qquad \vdots \\ X_p &= \mu_p + a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pm}F_m + \varepsilon_p \end{aligned}$$

The coefficient a_{ij} is called the loading of the j^{th} factor on the i^{th} variable, thus the matrix $\underline{\mathbf{L}}$ is the pxm matrix of factor loadings. The measure of the total variance of x_i explained by the m common factors can be expressed as $\sum_{j=1}^{m} a_{ij}^2$ called communality. In matrix rotation our factor model becomes:

$$\underline{\mathbf{X}} = \underline{\boldsymbol{\mu}} + \underline{\mathbf{LF}} + \underline{\boldsymbol{\varepsilon}},$$

where $\underline{\mathbf{F}}$ is the mx1 vector of unobservable common factors and $\underline{\boldsymbol{\varepsilon}}$ is the px1 vector of specific factors.

In order to guarantee certain covariance relationships among the variables, it is important that the random vectors $\underline{\mathbf{F}}$ and $\underline{\boldsymbol{\varepsilon}}$ satisfies the following assumptions.

1. $E(\underline{\mathbf{F}}) = E(\underline{\boldsymbol{\varepsilon}}) = \underline{\mathbf{0}}$.
2. $Cov(\underline{\mathbf{F}}) = \underline{\mathbf{I}}$ and $Cov(\underline{\boldsymbol{\varepsilon}}) = \underline{\boldsymbol{\Psi}}$ (a diagonal matrix).
3. $Cov(\underline{\boldsymbol{\varepsilon}}, \underline{\mathbf{F}}) = \underline{\mathbf{0}}$ ($\underline{\boldsymbol{\varepsilon}}$ and $\underline{\mathbf{F}}$ are independent).

Based on these assumptions and the initial conditions of given vectors we can describe our original factor model as an orthogonal factor model with m common factors. Using these assumptions, it can be shown that $\boldsymbol{\Sigma} = \mathbf{L}\,\mathbf{L}^T + \boldsymbol{\Psi}$.

Theoretically, $\boldsymbol{\Sigma}$ is unknown. Thus we must estimate $\mathbf{L}$ and Ψ from a sample. We are looking for estimates $\hat{L}$ and $\hat{\Psi}$, so we use maximum likelihood estimates. To obtain these, we maximize the likelihood function

$$L(\mathbf{L}, \boldsymbol{\Psi}) = \prod_{i=1}^{n} \frac{e^{\frac{-1}{2}\left(\underline{x_i}-\underline{\mu}\right)^T \left(LL^T+\Psi\right)^{-1}\left(\underline{x}-\underline{\mu}\right)}}{(2\pi)^{\frac{p}{2}}\left|LL^T+\Psi\right|^{\frac{1}{2}}}.$$

Notice there are two parameters $\mathbf{L}$ and Ψ in the equation. Thus, in order to maximize the function we must use partial derivatives. Computation shows that the first partial derivative does not give explicit solutions. Thus we must use statistical software such as SAS or Minitab to approximate the solutions.

We want to test the adequacy of the m-factor model that we create. Let's keep in mind our goal is to reduce the dimensionality of the data set, so we start with m=1 and test for adequacy. The null hypothesis is $\mathbf{\Sigma} = \mathbf{L}\,\mathbf{L}^{T} + \mathbf{\Psi}$ and the alternative hypothesis is $\mathbf{\Sigma}$ must be equal to some other positive definite matrix. This is a likelihood ratio test. We reject the null hypothesis if the test statistic χ^2 is greater than $\chi^2_{\alpha,v}$, where α is the level of significance and *v* is is the number of degrees of freedom.

The test statistic is given by,

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

and

$$V = \frac{1}{2}\left[(p-m)^2 - p - m\right].$$

Where n is the large sample size, p is the number of variables, m is the number of common factors, and S represents the sample covariance matrix. Also $\hat{L}$ and $\hat{\Psi}$ maximum likelihood estimates of *L* and Ψ**.** If we do not reject the null hypothesis m =1 then we use one common factor. If we do reject the null then we run the test with m=2 and repeat the process until we find an adequate model value of m for which we accept H_o.

Since the degrees of freedom must be a positive value, m must satisfy $m \prec \frac{1}{2}\left(2p+1-\sqrt{8p+1}\right)$. This places an upper bound on the number of common factors. When the appropriate factor model is obtained, the m factors must be interpreted and labeled. Now the data set is reduced from p variables to m factors.

Factor Analysis Application

The aim of factor analysis is to reduce the dimensionality of the data set. The factor model created must still account for most of the variance of the data set. We use Minitab software to run factor analysis and create a 6 - factor model.

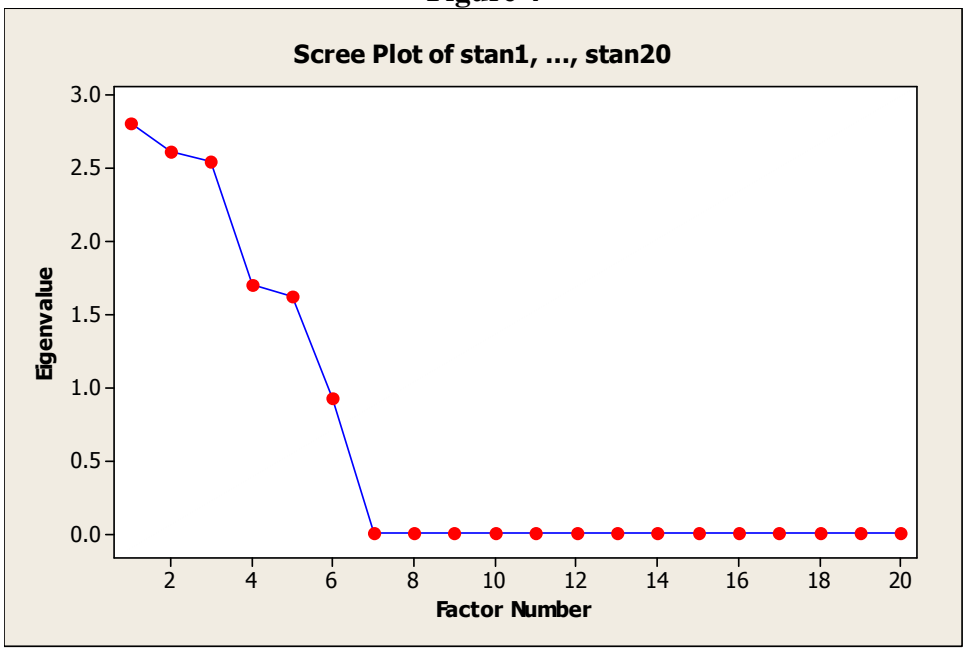
Table 4: Adequacy Test of *m*-factor model

m Factor Model	Degrees of Freedom	Test Statistic	Critical Value	P-value
1	170	423.1	201.423	0
2	151	282	180.676	0
3	136	218.82	164.216	0.0001
4	116	171.56	142.138	0.0007
5	100	128.85	124.342	0.037

6	85	80.91	107.52	0.605

We start at m=1 and we test the adequacy of the corresponding model. Table 4 lists the m factor models and their corresponding degrees of freedom, test statistic, critical values at $\alpha = 0.05$, and p-value. When using the chi-square test of adequacy we want the critical value to be greater than the test statistic. This implies we accept the null hypothesis and the corresponding m factor model is adequate. In table four we see that the 6 factor model gives a p-value greater than 0.05. Thus we accept the null hypothesis $\Sigma = LL^T + \Psi$.

Figure 4



The scree plot for factor analysis shows that the sixth factor is a candidate for the elbow of the graph. The number of factors has been selected and now they must be labeled. Table 5 gives the list of variables and their corresponding factor loadings for the 6 factor model. Based on the magnitudes, signs, and similarities of the values, we use table five to label the factors.

Table 5: Factor Loadings for 6-Factor Model

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Sch Life Exp	0.519	0.213	-0.634	-0.409	0.016	-0.177
Avg. Enroll (Prm./ Sec.)	0.416	0.065	-0.698	-0.153	0.04	-0.319
Prm. School Tch/Std Ratio	-0.991	-0.131	0	0	0	0
Duration of Sch Cycle	0.207	0.32	0.178	-0.889	-0.015	0.014

School Starting Age Prm.	0.103	0.065	0.046	0.422	-0.205	-0.166
Sec. Sch Tch/Std Ratio	-0.746	-0.156	0.105	0.083	0.094	-0.046
Sec. Sch Starting Age	-0.058	-0.063	-0.096	-0.387	-0.051	-0.28
Total % of Rep in Sch	-0.249	-0.295	0.295	0.118	0.086	-0.414
Avg. 8th gr Math/Sci	0.203	0.408	-0.564	-0.181	-0.093	0.16
% of Prt at least Sec Edu.	0.3	0.344	-0.457	-0.048	0.102	0.029
Avg. Adult Literacy	0.516	0.209	-0.459	-0.001	-0.076	0.219
# of Tch Hr per yr (Sec.)	-0.097	-0.025	0.026	-0.049	0.797	-0.085
# of Tch Hr per yr (Prm.)	0.068	0.065	0.105	-0.011	0.79	0.144
Edu. Exp as % of GDP	-0.008	0.062	-0.277	-0.027	0.377	0.154
Avg. Class Size	0.235	-0.024	0.296	0.494	-0.19	-0.063
Avg Tch Salary (Prm.)	0.132	0.944	-0.017	-0.015	0.053	0.137
Avg Tch Salary (Sec.)	0.1	0.995	0	0	0	0
# of Days Inst. (Prm./Sec.)	-0.11	0.05	-0.427	-0.187	0.276	0.226
Sec. Graduation Rate	0.214	0.371	-0.563	-0.187	-0.105	0.507
Households Items	0.277	0.209	-0.324	0.129	0.064	0.242

Appendix E gives a list of the new labels for each factor. Factor 1 is labeled as 'student teacher relationship' based on the very high negative loading on primary teacher/student ratio. Factor 2 is labeled as 'teacher importance' based on the high value of the loading on teacher salary and the extreme low value on the GDP. Factor 3 has the highest negative loadings between graduation rate and average 8th grade mathematics and science scores; therefore we label this factor as 'school performance'. Factor 4 is labeled as the 'overall seriousness put on education' based on the high negative loading of school life expectancy. Factor 5 is labeled as 'education time' based on the high loading on teaching hours per year for both primary and secondary schools. Factor 6 is labeled as the 'overall expected performance of students' based on the high loading on graduation rate and average adult literacy rate.

Overall the 6 factor model explains approximately 61% of the variance in the data. This is sufficient in most social science models.

Discriminant Analysis Theory:

Discriminant analysis is a multivariate technique used to separate distinct sets of objects and/or allocate new objects to previously defined groups. It is often used on a one time basis to look into observed differences when connecting relationships are not well understood.

Let's label the classification populations as Π_1 and Π_2. Consider $\underline{\mathbf{X}}$ as a set of measured variables. We want to classify each observed value of $\underline{\mathbf{X}}$ into either Π_1 or Π_2. In order to classify our measurements we produce a linear combination of component $\underline{\mathbf{X}}$ called Fishers' linear discriminant function that should separate Π_1 and Π_2 as much as possible.

Let $\mathbf{\Omega}$ be the p-dimensional space of all $\underline{\mathbf{X}}$. We want a rule that partitions $\mathbf{\Omega}$, into two parts, R_1 and R_2. Due to the properties of a partition, $R_1 \cup R_2 = \mathbf{\Omega}$ and $R_1 \cap R_2 = \emptyset$. Our classification procedure classifies $\underline{\mathbf{X}}$ into Π_1 if $\underline{\mathbf{X}}$ is in R_1, and classifies $\underline{\mathbf{X}}$ into Π_2 if $\underline{\mathbf{X}}$ is in R_2. Let $f_1(\underline{\mathbf{x}})$ and $f_2(\underline{\mathbf{x}})$ be the probability density function of $\underline{\mathbf{X}}$ under Π_1 and Π_2 respectively. Since there are infinitely many such partitions we want an optimal partition that minimizes the total probability of misclassification (TPM). The TPM can be written as $\alpha = \alpha_1 + \alpha_2$, where α is the conditional probability of classifying an object as Π_2 when it is from Π_1, α_2 is the conditional probability of classifying an object as Π_1 when it is from Π_2. Note that

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

and

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

We aim to minimize the TPM subject to $\alpha_1 = \alpha_2$. Thus, we classify $\underline{\mathbf{X}}$ into Π_1 if $\underline{\mathbf{X}}$ is more likely to fall under Π_1 than Π_2. We classify $\underline{\mathbf{X}}$ into Π_2 if $\underline{\mathbf{X}}$ is more likely to fall under Π_2 than Π_1. We can express this optimal partition as $R_1 = \{\underline{\mathbf{X}} | f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) > C\}$ and $R_2 = \{\underline{\mathbf{X}} | f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) < C\}$, where C is chosen such that $\alpha_1 = \alpha_2$.

We now set restrictions on our populations where Π_1 and Π_2 are multivariate normal with p variables, with mean vectors $\underline{\mu}_1$ and $\underline{\mu}_2$ respectively, and covariance matrices $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$. In practice $\underline{\mu}_1$, $\underline{\mu}_2$, and $\mathbf{\Sigma}$ are unknown so we use training samples of sizes n_1 and n_2 from Π_1 and Π_2. Let $\underline{\bar{x}}_1$ be an estimate of $\underline{\mu}_1$ and S_1 an estimate of $\mathbf{\Sigma}_1$ and S_2 an estimate of $\mathbf{\Sigma_2}$.

To determine whether linear or quadratic discriminant analysis is used, we test whether the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are equal. If the covariance matrices are equal then we apply the linear discriminant analysis. Thus the null hypothesis is H_o: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ and the alternate is H_a: $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. For this test we use the unbiased, pooled estimate Sp of the common covariance matrix. Sp can be expressed by

$$\text{Sp} = \frac{1}{n_1 + n_2 - 2}\left\{\sum_1^2 (n_i - 1)S_i\right\}.$$

This is a chi-square likelihood ratio test. The test statistic is given by $\frac{M}{C}$ where

$$\text{M} = \left\{\sum_1^2 (n_i - 1)S_i\right\}\ln\left|S_p\right| - \left\{\sum_1^2 (n_i - 1)\ln\left|S_i\right|\right\}$$

and

$$\frac{1}{c} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)}\left[\left(\sum_1^2 \frac{1}{n_i - 1}\right) - \frac{1}{n_1 + n_2 - 2}\right].$$

Under H_o, $\frac{M}{C}$ has a x^2 distribution with $p(p+1)/2$ df. We accept the null hypothesis if $\frac{M}{C} < \chi^2_{p(p+1)/2,\alpha}$, where α is the significance level and p is the number of variables in the data set. The quantity $p(p+1)/2$ represents the degrees of freedom and is denoted by *v*.

Now we must define the classification rules. For the linear discriminant function we classify as follows:

$$\text{Classify } \underline{x} \text{ into } \Pi_1 \text{ if } \hat{\underline{a}}^T\underline{x} > \hat{h},$$

and

$$\text{Classify } \underline{x} \text{ into } \Pi_2 \text{ if } \hat{\underline{a}}^T\underline{x} \le \hat{h},$$

$$\text{where } \hat{\underline{a}} = S^{-1}{}_p(\underline{\square}_1 - \underline{\square}_2) \text{ and } \hat{h} = \frac{1}{2}(\underline{\square}_1 - \underline{\square}_2)^T S^{-1}{}_p(\underline{\square}_1 - \underline{\square}_2).$$

If we reject the null hypothesis, implying the covariances are not equal then we would apply the quadratic discriminant function. The classification rules are as follows:

$$\text{Classify } \underline{x} \text{ into } \Pi_1 \text{ if } -\frac{1}{2}\underline{x}^T(S^{-1}{}_1 - S^{-1}{}_2)\underline{x} + (\underline{\square}_1{}^T S^{-1}{}_1 - \underline{\square}_2{}^T S^{-1}{}_2)\underline{x} \ge k,$$

$$\text{Classify } \underline{x} \text{ into } \Pi_2 \text{ if } -\frac{1}{2}\underline{x}^T(S^{-1}{}_1 - S^{-1}{}_2)\underline{x} + (\underline{\square}_1{}^T S^{-1}{}_1 - \underline{\square}_2{}^T S^{-1}{}_2)\underline{x} < k,$$

$$\text{where } k = \frac{1}{2}\ln\left(\left|\frac{S_1}{S_2}\right|\right) + \frac{1}{2}(\underline{\square}_1{}^T S^{-1}{}_1\underline{\square}_1 - \underline{\square}_2{}^T S^{-1}{}_2\underline{\square}_2).$$

After determining the classification function we must find any misclassifications of our predicted training sample. The apparent error rate (APER) measures the accuracy of the model. We use the APER to give us the percentage of observations misclassified. Our objective is to minimize the number of misclassifications; this means we aim for

small values of the APER. Another method to minimize misclassifications is to use the total probability of misclassification (TPM). The calculation of TPM involves measuring the Mahalanobis distance, Δ_p, between the two populations given by

$$(\hat{\Delta}_p)^2 = (\underline{\square}_1 - \underline{\square}_2)^T S^{-1}_p (\underline{\square}_1 - \underline{\square}_2).$$

Also,

$$\text{TPM} = \alpha = \alpha_1 + \alpha_2 = 2\Phi\left(-\frac{1}{2}\hat{\Delta}_p\right).$$

Let D_i be the Mahalanobis distance of $\underline{x}$ from Π_i given by $D_i^2 = (\underline{X} - \underline{\square}_i)^T S^{-1}_p (\underline{X} - \underline{\square}_i)$ where i=1,2. We classify $\underline{x}$ into Π_1 if D_1^2 is less than D_2^2. Otherwise we classify $\underline{x}$ into Π_2. This is known as the minimum distance classification rule. The general classification rule for the j^{th} population case is to classify $\underline{x}$ into Π_k if $D_k^2 = \text{Min}(D_1^2, D_2^2, \ldots, D_j^2)$.

Discriminant Analysis Application

Discriminant analysis classifies countries into two populations. The first step before running the analysis on Minitab is to specify a criterion. The criterion for this data set is the school life expectancy of a student. The median of the measured life expectancy corresponding to the countries, 14 years, divides the countries into two separate groups of 32 countries. Next we randomly generate 32 numbers through the Minitab software. In the same randomly generated order we assigned the countries in each group the list of numbers when the countries were in alphabetical order. Next we sorted the randomly generated numbers in ascending order for one group and descending order for the other group. We then chose the first 20 numbers from each group and created our training sample. This leaves 24 countries in the test sample. Appendix B gives a list of the training sample.

We must decide which type of discriminant function is the best predictor. So we test to check the equality of the two covariance matrices of the two groups in the training sample. We use a likelihood ratio test with a chi-square distribution. For our data set the test statistic $\frac{M}{C}$ =266.62 and the critical value $\chi^2_{190,.05} = 223$. Since $\chi^2 < \frac{M}{C}$ we reject our null and we use the quadratic discriminant function. Table 6 shows the quadratic model results of the training sample. We calculate the apparent error rate to be 0%. Thus, the model correctly classifies all of the countries in the training sample 100 percent of the time.

Table 6 : Quadratic Model

Classified into Group	True Group: Exceptional Education	True Group: Fair Education
Exceptional Education	20	0
Fair Education	0	20
Total N	20	20
N Correct	20	20
Proportion	1	1
N = 40	N correct =40	Proportion Correct = 1.000

Next we use this model to classify the countries in the test samples. We can compare our results like the results of the training samples. Appendix D shows the quadratic model results of the test samples. This model correctly classifies 18 of the 24 countries in the test sample. Thus our quadratic model was 75% efficient. This model classified Luxembourg as being the most exceptional country in education and Israel as the least exceptional in the data set.

Table 7: Linear Model

Classified into Group	True Group: Exceptional Education	True Group: Fair Education
Exceptional Education	20	0
Fair Education	0	20
Total N	20	20
N Correct	20	20
Proportion	1	1
N = 40	N correct = 40	Proportion Correct = 1.000

The linear discriminant model is still a good predictor in classifying the countries in the data set. The Table 7 gives the results of the training sample using the linear discriminant function. The apparent error rate of the linear function is 0%. The linear model classified 18 out of 24 countries into the correct populations from our test samples. The linear model was 75% efficient. The linear model classified Italy as being the most exceptional country in education and Peru as the least exceptional in the data set. Appendix C shows the linear model results of the countries in the test sample.

Conclusion

We are very satisfied with the principal component analysis. We originally started with twenty variables and we were able to reduce the dimensionality to only five principal components that explained the majority of the variability in the data. Yet the five principal components were not easily interpreted. The first two principal components were interpreted feasibly, but the remaining ones were difficult to label.

The factor analysis application succeeded in reducing the dimensionality as well. We reduced twenty variables to six factors. The six factors still explained over 60% of the variance in the data set. The factors created in the analysis were difficult to label. The factor loadings were not distinct enough to confidently interpret. Factor rotation would have helped but it was a lack of time that kept us from rerunning the test.

The discriminant analysis application gave unexpected results. There was a high percentage of misclassification in both discriminant models. Thus the discriminant analysis could have been improved. More descriptive variables and detailed data about the countries would improve the analysis. Since the criterion carries a significant amount of weight in the analysis, a stronger criterion could improve the analysis. From the analysis, we observed that many countries were classified in a way that is the opposite of the public's general assumption. Macedonia was misclassified in the quadratic model. This country rarely receives public recognition yet it is predicted to have a exceptional education system. Furthermore, the United States of America is a very powerful and worldly known nation, but it was predicted to have a fair education system.

We realize the amount of time and effort it takes to conduct a successful statistical research experiment. This experiment does not measure the I.Q. of individuals in a country. It is more of a measure of seriousness and efforts a nation puts into its education system. We learned that a countries' wealth does not determine a countries' education system. This project does not specify the depth of material covered in a nation's education system. New doors of growth and technology wait to be opened, yet education holds the key to all locks.

Acknowledgments

First and foremost we would like to thank God who is ahead of our lives for this opportunity. We would like to thank the sponsors of SUMSRI, the national Science Foundation, and the Nation Security Agency who made this research experience possible. We would also like to thank Dr. Dennis Davenport, Dr. Vasant Waikar, and their committee for directing SUMSRI. A special thanks to Dr. Waikar for teaching us the statistical techniques and advising us for our research. Another special thanks to Shelly-Ann Meade for assisting outside of class and helping with our research experience. In addition we would like to thank Tom Farmer for conducting the writing seminar and assisting with revising our research papers. We thank Bonita Porter, Dr. Ortiz, Dr. Dowling, and all of the students and faculty who participated in SUMSRI. Thanks to everyone we left out!

Bibliography

[1] Haaretz.com, *Haaretz*:Retrived June 2006 from www.haaretz.com

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

[3] Morrison, Donald F. *Multivariate Statistical Methods*. (Second Edition), MacGraw- Hill Book Company (1967)

[4] Nationmaster.com, *Nationmaster: Education Statistics*: 2003 – 2006. Retrived June 2006 from http://www.nationmaster.com/index.php

[5] Nationmaster.com, *Nationmaster: Education Statistics*: 2003-2006. Retrieved June 2006 from http://www.nationmaster.com/cat/edu-education

[6] Nces.ed.gov, *National Center for Education Statistics- Education Indicators* : Retrieved June 2006 from nces.ed.gov/pubs/eiip/eiipid40.asp

[7] Oecd.org, *Organization for Economic Co-operation and Development*:2003. Retrieved June 2006 from http://www.oecd.org/document/34/0, 2340,en 2649 37455 14152482 1 1 1 37455,00.html

[8] Oced.org, *Organization for Economic Co-operation and Development*: 2003. Retrieved June 2006 from www.oced.org

[9] USAID.org, *Global Education Database-Office of Education of the US Agency for International Development* : 2002. Retrieved June 2006 from http://qesdb.cdie.org/ged/index.html

Appendix A

Principal Components	Labels
Principal Components 1	Expected Student Performance
Principal Components 2	Teacher Importance
Principal Components 3	Parents Participation
Principal Components 4	Overall Importance of Education
Principal Components 5	Student/Teacher Relationship

Appendix B

Training Sample Population π_1	Training Sample Population π_2
Australia	Cape Verde
Austria	Croatia
Brazil	Egypt
Canada	India
China	Iran
Czech Republic	Jamaica
Estonia	Jordan
Greece	Kazakhstan
Hungary	Kuwait
Iceland	Malaysia
Ireland	Mexico
Japan	Montserrat
Korea (Rep)	Philippines
Lithuania	Russia
New Zealand	Serbia Montenegro
Norway	Thailand
Portugal	Uganda
Sweden	Ukraine
United Kingdom	United Arab Emirates
Virgin Island (British)	Venezuela

Population 1- Exceptional
Population 2- Fair(non-exceptional)

Appendix C: Linear Model Classification- Predicted Group is Discriminant Analysis. True Group is criteria. 1-Exceptional, 2-Fair.

Country	Predicted Group	True Group	Squared distance to π_1	Squared distance to π_2
Belarus*	1	2	20.136	20.791
Belgium	1	1	33.234	81.769
Bulgaria*	1	2	13.93	20.846
Chile	2	2	35.663	27.558
Cuba	2	2	34.251	25.232
Denmark	1	1	23.028	59.265
Finland	1	1	21.434	55.396
France	1	1	18.698	28.582
Germany	1	1	45.858	82.4
Israel*	2	1	39.42	26.139
Italy	1	1	9.673	26.2
Kyrgyzstan	2	2	55.514	27.767
Luxembourg*	1	2	11.984	34.519
Macedonia	2	2	31.442	17.414
Peru	2	2	34.611	17.053
Poland	1	1	21.526	40.695
Qatar	2	2	46.07	26.734
Romania*	1	2	20.026	21.148
Slovenia	1	1	42.485	53.817
South Africa	2	2	76.028	70.814
Spain	1	1	20.413	23.466
Switzerland	1	1	25.4	40.565
Trinidad and Tabago	2	2	42.222	20.754
United States*	2	1	43.374	39.182

Appendix D: Quadratic Model Classification- Predicted Group is Discriminant Analysis. True Group is criteria. 1-Exceptional, 2-Fair.

Country	Predicted Group	True Group	Squared distance to π_1	Squared distance to π_2
Belarus	2	2	710.079	283.893
Belgium	1	1	90.074	363.208
Bulgaria	2	2	356.117	33.913
Chile	2	2	956.114	78.525
Cuba	2	2	150.919	51.988
Denmark	1	1	46.071	176.837
Finland	1	1	188.612	377.038
France	1	1	207.058	1023.952
Germany	1	1	263.318	505.292
Israel*	2	1	351.213	26.07
Italy	1	1	135.25	159.318
Kyrgyzstan	2	2	1017.523	71.326
Luxembourg*	1	2	13.091	453.156
Macedonia*	1	2	83.248	99.56
Peru	2	2	499.079	114.57
Poland*	2	1	290.663	207.024
Qatar	2	2	818.333	120.446
Romania	2	2	281.847	87.685
Slovenia*	2	1	306.005	268.183
South Africa	2	2	854.098	796.988
Spain	1	1	66.52	177.657
Switzerland	1	1	201.816	485.425
Trinidad & Tabago	2	2	470.435	87.978
United States*	2	1	1141.701	265.64

Appendix E

Factors	Labels
Factor 1	Teacher/Student Relationship
Factor 2	Teacher Importance
Factor 3	School Performance
Factor 4	Overall Seriousness of Education
Factor 5	Education Time
Factor 6	Overall Expected Performance of Student