Life Expectancy Analysis

using different classification methods

Weixiong Junyan Deng

Instructor: Dr. Kim Sung

May 2016

1. Introduction

Human census shows that people in different countries have different expected lifetime. In this paper, we are trying to categorize countries based on the life expectancy of each country as an indicator. We define three groups based on the life expectancy (i.e., good = life expectancy greater or equal than 75, bad = life expectancy less than 64, and medium otherwise). Under this criteria, we are going to use discriminant analysis and logistic regression to build a best model for future prediction with given input variables.

2. Data Description

The data is collected and maintained by The World Bank. We extracted the variables that had most of the countries information and exclude variables with more than 50% missing data. Thus, the dataset consists of 233 observations, in which each observation represents a country, and 11 variables. Table 1 contains detailed information on the variables in the dataset.

Table 2.1. Overview of Variables

Variable	Type	Description	Units
Country	Descriptive	Name of country	None
Code	Descriptive	Abbreviation for the country	None
FertRate	Input	Number of children that would be born to a woman.	Integer
Measles	Input	Percentage of children ages 12-23 months who received vaccinations before 12 months or at any time before the survey.	Percent
Sanitation	Input	Percentage of the population using improved sanitation facilities.	Percent
WaterSource	Input	Percentage of the population using an improved drinking water source such as piped water, public taps, protected springs, etc.	Percent
InternetUsers	Input	Individuals who have used the Internet (from any location) in the last 12 months. (per 100 people)	Proportion
Cellphone	Input	Mobile cellular telephone subscriptions (per 100 people)	Proportion
MortRate	Input	the probability per 1,000 that a newborn baby will die before reaching age five	Proportion
Population	Input	Total population	Integer
tuberculosis	Input	Estimated number of new and relapse tuberculosis cases arising in a given year, expressed as the rate per 100,000 population.	Proportion
unemployment	Input	The labor force that is without work but available for and seeking employment.	Percent
LifeExpectancy	Output	The number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.	Integer

Table 2.2 Summary Statistic of Variables

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
FertRate	180	2.84927	1.41650	512.86801	1.21000	7.59900
Measles	180	87.41004	13.50040	15734	22.00000	99.00000
Sanitation	180	71.56351	29.49313	12881	6.70000	100.00000
WaterSource	180	88.01366	15.22538	15842	31.70000	100.00000
InternetUsers	180	42.30406	28.84239	7615	0	98.16000
Cellphone	180	104.61857	38.72059	18831	6.38600	218.43029
MortRate	180	33.94962	33.90800	6111	2.00000	162.20000
Population	180	77751899	273041872	1.39953E10	54944	2264058207
tuberculosis	180	122.83344	160.37899	22110	0.71000	852.00000
unemployment	180	8.31237	5.20911	1496	0.30000	27.90000
LifeExpectancy	180	70.80754	8.36529	12745	48.93473	83.58780

3. **Methodology**

In the preliminary step, we will use cluster analysis, correlation matrix, and principal component to explore the nature of our dataset. In cluster analysis, we are going to look the number of grouping in our dataset based on similarities of the variables (excluding target variable-life expectancy) using average linkage. And then we look at the correlation matrix to examine how each variable is correlated to one another and exclude the uncorrelated variables before running principal component analysis. Finally, we use principal component analysis that will allow us to see if we can reduce our number of variables into few independent principal components.

After preliminary analysis, we separate the data into two parts,70 percent to be training data and 30 percent to be validation data in order to build and test our classification model for life expectancy of a country. We use two methods for

classification of life expectancy, discriminant analysis and logistic regression, into one of the three groups: good, medium, or bad life expectancy. At last, we will calculate and compare the misclassification rate of each method to determine its predictive power.

4. Preliminary Analysis

4.1 Correlation analysis

Table 4.1 Corelation matirx

	FertRate	Measles	Sanitation	WaterSource	InternetUsers	Cellphone	MortRate	Population	tuberculosis	unemployment	LifeExpectancy
FertRate	1.00000	-0.58183 <.0001	-0.81837 <.0001	-0.77754 <.0001	-0.74410 <.0001	-0.53256 <.0001	0.87490 <.0001	-0.07014 0.3494	0.43896 <.0001	-0.05515 0.4621	-0.82811 <.0001
Measles	-0.58183 <.0001	1.00000	0.57768 <.0001	0.55891 <.0001	0.48643 <.0001	0.42167 <.0001	-0.62005 <.0001	0.00079 0.9916	-0.34333 <.0001	0.03695 0.6224	0.56683 <.0001
Sanitation	-0.81837 <.0001	0.57768 <.0001	1.00000	0.78015 <.0001	0.78411 <.0001	0.57817 <.0001	-0.83588 <.0001	-0.05395 0.4719	-0.58697 <.0001	0.10239 0.1714	0.84371 <.0001
WaterSource	-0.77754 <.0001	0.55891 <.0001	0.78015 <.0001	1.00000	0.69331 <.0001	0.56583 <.0001	-0.81364 <.0001	0.04329 0.5639	-0.49576 <.0001	0.07596 0.3108	0.76189 <.0001
InternetUsers	-0.74410 <.0001	0.48643 <.0001	0.78411 <.0001	0.69331 <.0001	1.00000	0.59972 <.0001	-0.75375 <.0001	-0.03189 0.6709	-0.56625 <.0001	0.11728 0.1169	0.80429 <.0001
Cellphone	-0.53256 <.0001	0.42167 <.0001	0.57817 <.0001	0.56583 <.0001	0.59972 <.0001	1.00000	-0.54302 <.0001	-0.06052 0.4197	-0.32488 <.0001	0.01270 0.8656	0.51447 <.0001
MortRate	0.87490 <.0001	-0.62005 <.0001	-0.83588 <.0001	-0.81364 <.0001	-0.75375 <.0001	-0.54302 <.0001	1.00000	0.00491 0.9478	0.55253 <.0001	-0.04677 0.5330	-0.91330 <.0001
Population	-0.07014 0.3494	0.00079 0.9916	-0.05395 0.4719	0.04329 0.5639	-0.03189 0.6709	-0.06052 0.4197	0.00491 0.9478	1.00000	0.02564 0.7327	-0.01071 0.8865	0.01396 0.8524
tuberculosis	0.43896 <.0001	-0.34333 <.0001	-0.58697 <.0001	-0.49576 <.0001	-0.56625 <.0001	-0.32488 <.0001	0.55253 <.0001	0.02564 0.7327	1.00000	-0.10866 0.1465	-0.68531 <.0001
unemployment	-0.05515 0.4621	0.03695 0.6224	0.10239 0.1714	0.07596 0.3108	0.11728 0.1169	0.01270 0.8656	-0.04677 0.5330	-0.01071 0.8865	-0.10866 0.1465	1.00000	0.07302 0.3300
LifeExpectancy	-0.82811 <.0001	0.56683 <.0001	0.84371 <.0001	0.76189 <.0001	0.80429 <.0001	0.51447 <.0001	-0.91330 <.0001	0.01396 0.8524	-0.68531 <.0001	0.07302 0.3300	1.00000

Correlation analysis can show us a preliminary relationship between all variables. Based on the correlation matrix (Table 4.1), the population and unemployment variables are not included in our classification analysis since they are not related to other variables. we can find out that population and unemployment are not related to other variables. In addition, we can group the variables sanitation, measles, vaccination, and water source together because they have high correlation to one another.

4.2 Cluster Analysis

Cluster Analysis is an unsupervised analysis in which there is no target variable included in the analysis and it focuses on grouping countries based on their similarities (i.e. Fert Rate, measles, sanitation, etc.) But, cluster analysis fails to find out a reasonable division only on the variable life expectancy. Therefore, we use all the variables for clustering. Figure 4.2.1, Pseudo-F statistics vs number of cluster, shows the first peak at 3 indicating that the best number for clustering is three. This result matches our objective (i.e., country's expectancy of life: good, medium and bad). However, cluster analysis obtained all the similarity from all the variables.

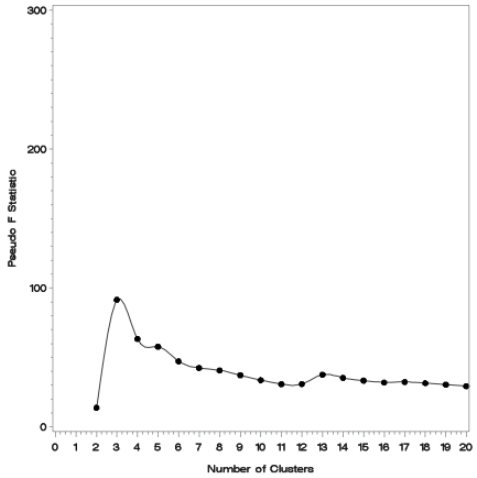

Fig. 4.2.1 Pseudo-F statistic vs. number of cluster

Table 4.2.2 Summary Statistic of each cluster.

CLUSTER=1

Analysis Variable : LifeExpectancy				
N	Mean	Std Dev	Minimum	Maximum
121	75.5139555	4.4303771	63.9656585	83.5878049

CLUSTER=2

Analysis Variable : LifeExpectancy				
N	Mean	Std Dev	Minimum	Maximum
55	60.3876982	5.2036484	48.9347317	70.0746829

CLUSTER=3

Analysis Variable : LifeExpectancy				
N	Mean	Std Dev	Minimum	Maximum
4	71.7114106	4.2223501	68.0138049	75.7822683

Our interest is only in longitude. We use the lower boundary of cluster1 and upper boundary of cluster 3, which is also the mean of cluster1, to form the interval of medium state. So we can categorize countries into three categories: good (index=2), life

expectancy greater or equal than 75, bad(index=0), life expectancy less than 64, and medium (index=1) otherwise.

4.3 Principal Component Analysis.

In the principal component analysis, we are trying to reduce the number of variables into principal components which are independent on each other. From the correlation matrix (see table 4.1), we notice that both the variables population and unemployment are uncorrelated (p-value greater than .05), therefore we do not include them into principal component analysis.

Table 4.3.1 Eigenvalues of the Correlation Matrix and Principal Component 1

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	5.40090391	4.68692644	0.6751	0.6751
2	0.71397747	0.10262893	0.0892	0.7644
3	0.61134854	0.11176841	0.0764	0.8408
4	0.49958013	0.20986725	0.0624	0.9032
5	0.28971287	0.08379482	0.0362	0.9394
6	0.20591805	0.03762283	0.0257	0.9652
7	0.16829522	0.05803141	0.0210	0.9862
8	0.11026381		0.0138	1.0000

	Prin1
FertRate	-.385666
Measles	0.298368
Sanitation	0.395939
WaterSource	0.377992
InternetUsers	0.372638
Cellphone	0.296332
MortRate	-.398960
tuberculosis	-.277043

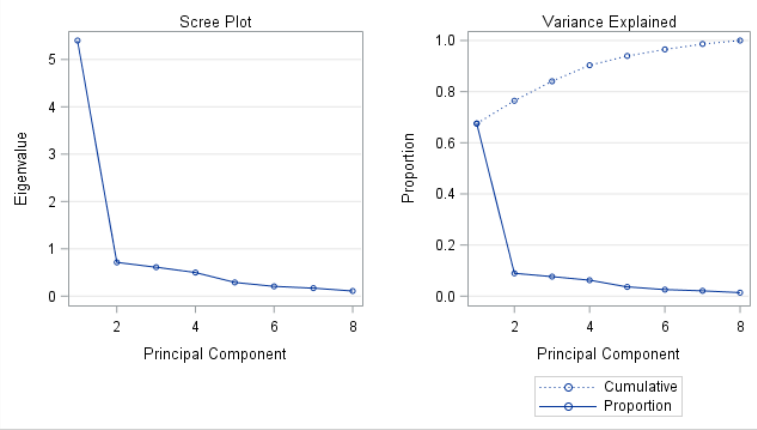

Figure 4.3.1 Scree Plot

From Table 4.3.1, we can find out that only the eigenvalue of prin1 is greater than 1 and it accounts for 68 percent of the total variation in our data. Thus, we take prin1 as our major component. Prin1 have equal weights on all of the variables that represent the overall information of all variables. In prin1, the variables Measles, Sanitation, Water source, Internet, and cellphone are positive while Fert Rate, Mort Rate, and tuberculosis are negative. Measles vaccination, sanitation facility and water source represent health resources that are directly positive related life longevity. It is conceivable that healthy people tend to live longer. Internet and cell phone represent high tech resources that represent the improvement of society. It makes sense that the variables Measles, Sanitation, Water source, Internet, and cellphone are group together as positive since high tech resources and health resources are related in some way. For example, people can get information about health through internet and cell phone. On the other hand, Fertile rate, Infant mort rate and tuberculosis are negative in prin1. It makes sense that poor countries tend to have high fertility rate because of lack of Moreover, more people would share the limited sources so that one person would get smaller portions of the sources. High infant mortality rate indicates bad sanitation condition and lack of nutrition. Additionally, tuberculosis is a severe disease and is the first killer in some poor country where the sanitation condition is bad. Hence, they have negative effect to prin1.Prin1 is overall description of all variables.

5 Classification Results

5.1 Discrimination Classification Using Original Variables

We run discrimination in SAS with stepwise selection with entry level to 0.4 and stay level to 0.05. Also, we include the uncorrelated variables, unemployment and population, into the model. In Table 5.1.1, the final model has only Mort rate, Internet users, Cellphone and tuberculosis.

Table 5.1.1 Discrimination Stepwise Selection

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	MortRate		0.7579	192.55	<.0001	0.24208069	<.0001	0.37895966	<.0001
2	2	InternetUsers		0.3448	32.10	<.0001	0.15861727	<.0001	0.55094989	<.0001
3	3	Cellphone		0.0844	5.58	0.0048	0.14522787	<.0001	0.57821309	<.0001
4	4	tuberculosis		0.0564	3.59	0.0306	0.13703063	<.0001	0.58531893	<.0001
5	5	Population		0.0447	2.79	0.0657	0.13090021	<.0001	0.59814148	<.0001
6	4		Population	0.0447	2.79	0.0657				

Then we run the model again with the significant variables and we get the parameters of Statistical Likelihood function for different groups after calculating three likelihood functions respectively. The likelihood functions are shown below,

L0= - 40.19829+0.37061*Sanitation+0.02190*tuberculosis
+0.23326*InternetUsers+0.72252*Mortrate

L1= - 26.16463+0.44593*Sanitation+0.01526*tuberculosis
+0.11563*InternetUsers+0.48469*Mortrate

L2= - 31.20009+0.45058*Sanitation+0.01253*tuberculosis
+0.21249*InternetUsers+0.46791*Mortrate

Table 5.1.2 Parameter value of Statistical Likely hood Function

Linear Discriminant Function for index			
Variable	0	1	2
Constant	-40.19829	-26.16463	-31.20009
Sanitation	0.37061	0.44593	0.45058
tuberculosis	0.02190	0.01526	0.01253
InternetUsers	0.23326	0.11563	0.21249
MortRate	0.72252	0.48469	0.46791

The observation groups according to the largest likelihood value among L0, L1 and L2. For example, if L2 is greater than L0 and L1, the observation should be categorized into group 2 which is in the good category (index = 2).

The model has 83.97% accuracy for training data and 85.19 % for validation data.

Table 5.1.3 Classification summary for training data(left) and for validation data(right)

Number of Observations and Percent Classified into index				
From index	0	1	2	Total
0	25 96.15	1 3.85	0 0.00	26 100.00
1	5 9.09	38 69.09	12 21.82	55 100.00
2	0 0.00	6 13.33	39 86.67	45 100.00
Total	30 23.81	45 35.71	51 40.48	126 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for index				
	0	1	2	Total
Rate	0.0385	0.3091	0.1333	0.1603
Priors	0.3333	0.3333	0.3333	

Table of index by pred				
	pred			
index	0	1	2	Total
0	14 25.93 93.33 93.33	1 1.85 6.67 6.67	0 0.00 0.00 0.00	15 27.78
1	1 1.85 5.26 6.67	13 24.07 68.42 86.67	5 9.26 26.32 20.83	19 35.19
2	0 0.00 0.00 0.00	1 1.85 5.00 6.67	19 35.19 95.00 79.17	20 37.04
Total	15 27.78	15 27.78	24 44.44	54 100.00

In order to compare the model estimate with real categorization, we generate more data to visualize them in plots. From stepwise summary, Figure 5.1.1, we can find out that the F value of Mort rate is the largest and the category “Internet users” is the second largest. We generate extra 2809 data on variables of Moret rate and Internet users. First, we obtain the gap number, that is,

Inc= (Max of the variable - Min of the variable)/50.

Secondly, generate a sequence from the difference between Min and Inc to the sum of Max and Inc by Inc, for More rate and Internet uses, respectively. We use this new dataset as test data and combine it with the original training dataset to run discrimination again.

For model estimation, in figure 5.1.5, group bad is in the red part. Countries in this group are at high infant mortality rate and only related to Mort rate. Group bad and group good are at relatively low infant mortality rate. The only difference between them is that group bad has lower internet coverage and that group good has higher internet coverage.

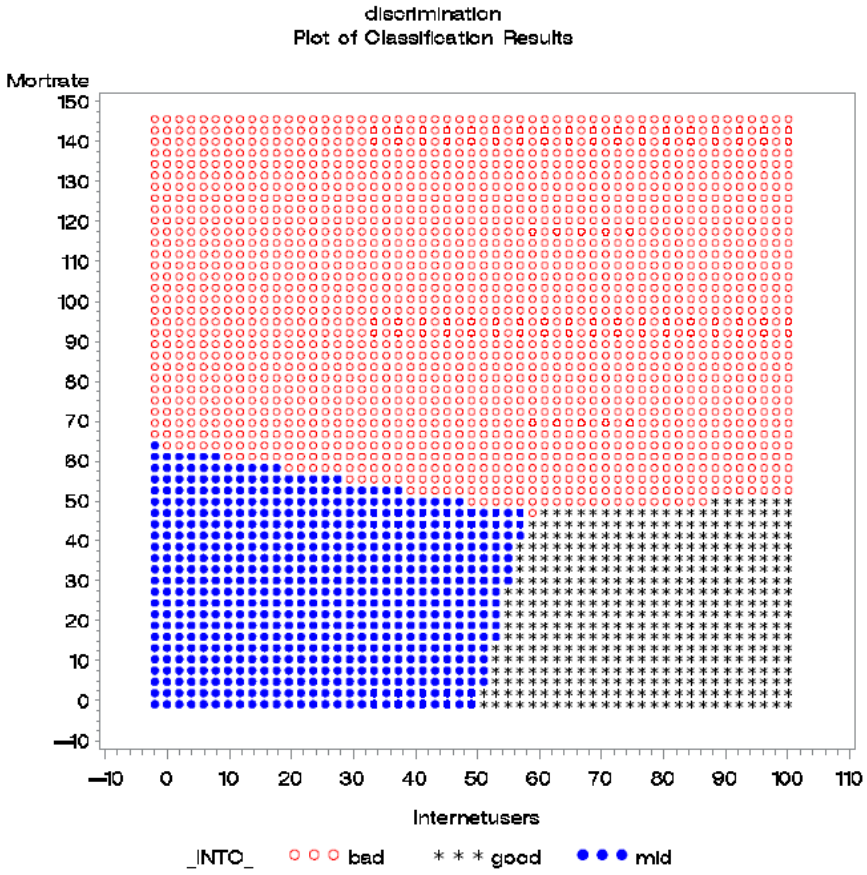

Figure 5.1.5 Estimate of categorization for the model on Mor trate and Internet users

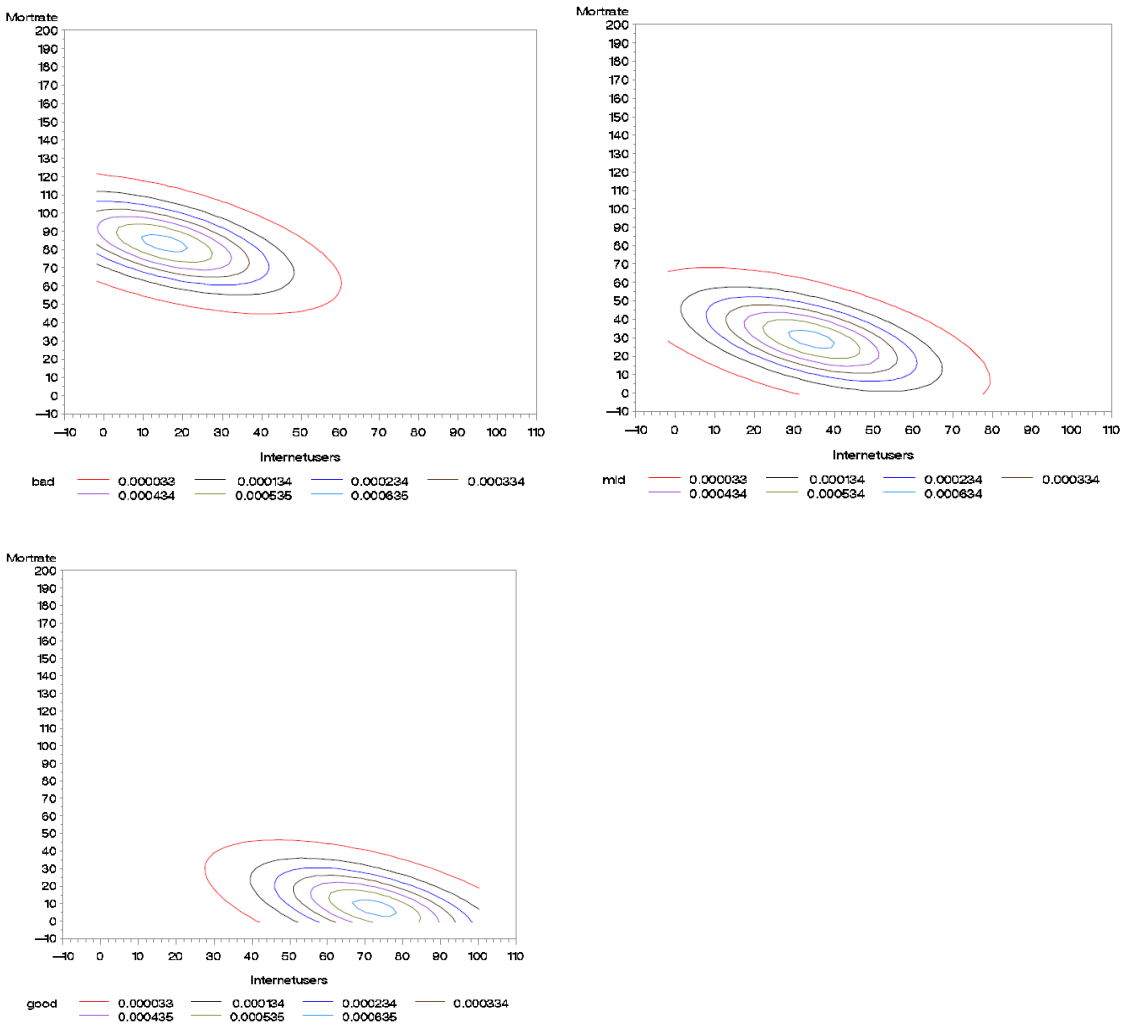

Figure 5.1.6 Estimated Densities for groups, bad, medium, good.

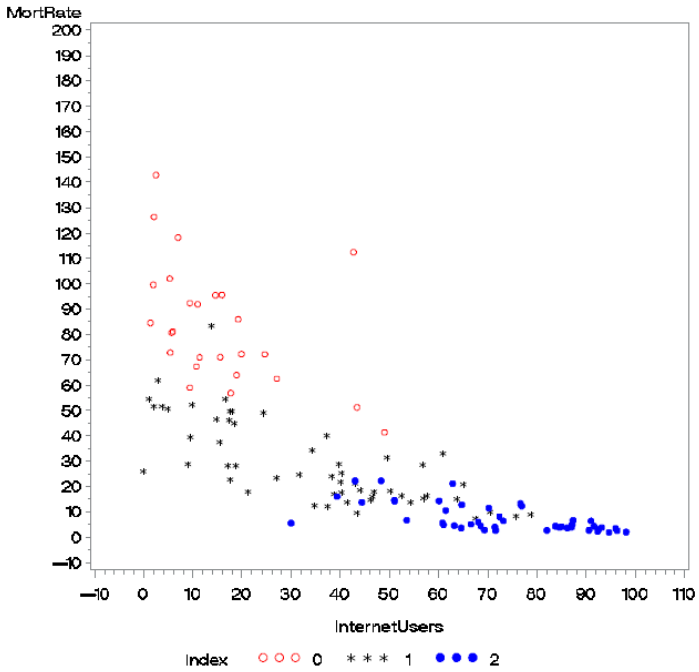

Figure 5.1.7 Categorization plot on Mort rate and Internet users

On the other hand, for real classification, in Figure 5.1.7, Categorization plot on Mort rate and Internet users, we can find out that Mort rate greater than 55 in 1000 can be identified as bad (index =0). Only two points in this group their mort rate is less than 55 in 1000. Moreover, observations in this group, their value of Internet users is less than 30%. High infant mort rate and low coverage of internet indicate countries in this group are low developed and lack of modern medication facility and high-tech factory. New born babies are so weak that they can easily get attack by disease. They need more care and nutrition to survive. These countries are short of health condition to guarantee long living. The group centers at Mort rate 85 and Internet users 15(see contour plot). Look into the observation in classification we can find out that most of these countries are poor countries.

Mort rate less than 55 in 1000 and Internet users rate less than 40% should be grouped into medium (index = 1) statistically. Countries in this group, their infant mortality rate and internet coverage are medium. The group centers at Mort rate 30 and Internet users 35(see contour plot). Look into the observation in classification we can find out that most of these countries are developing countries.

Countries in the last group are good. Their infant mortality is low, less than 30 in 1000. And they have high internet coverage, greater than 60%. Countries in this group have good healthy source and high-tech facility. The group centers at Mort rate 10 and Internet users 75(see contour plot). Most of countries in this group are well developed counties.

Group medium and Group good overlap between 10 and 30 in Mort rate and between 40 and 60 in Internet users. This overlap includes countries of upper level in medium group

and of lower level in good group. This is counterpart of the overlap between upper level of developing counties and lower level of developed countries.

5.2 Discrimination Classification Using Principal Component

From section 4.3, Principal Component Analysis, we know that the first principal component is the major component whose eigenvalue is greater than 1. In this section, we apply principal component as input variables to fit discrimination model. We include the variables of Unemployment and Population into analysis. Using stepwise selection, setting entrance significant level to be 0.4 and stay significant level to be 0.05. Unemployment does not enter because its p value is greater than 0.4.

Table 5.2.1 Stepwise Selection Summary

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	Prin1		0.7177	156.39	<.0001	0.28225252	<.0001	0.35887374	<.0001
2	2	Population		0.0334	2.11	0.1259	0.27282439	<.0001	0.37549977	<.0001
3	1		Population	0.0334	2.11	0.1259				

Only Prin1 is left for the final model. The Linear Likelihood functions are below,

L0= - 3.60887-2.20305*Prin1
L1= - 0.00144+0.04395*Prin1
L2= - 1.39445+1.36943*Prin1

Table 5.2.2 Statistical Likelihood Function for Index

Linear Discriminant Function for index			
Variable	0	1	2
Constant	-3.60887	-0.00144	-1.39445
Prin1	-2.20305	0.04395	1.36943

Calculate the Linear Discriminant Function for different indexes and compare them. The observation should be grouped into the largest result where its function belongs to. For example, if L1 is the largest one among L0, L1 and L2, then the observation should be assigned to group medium(index=1).

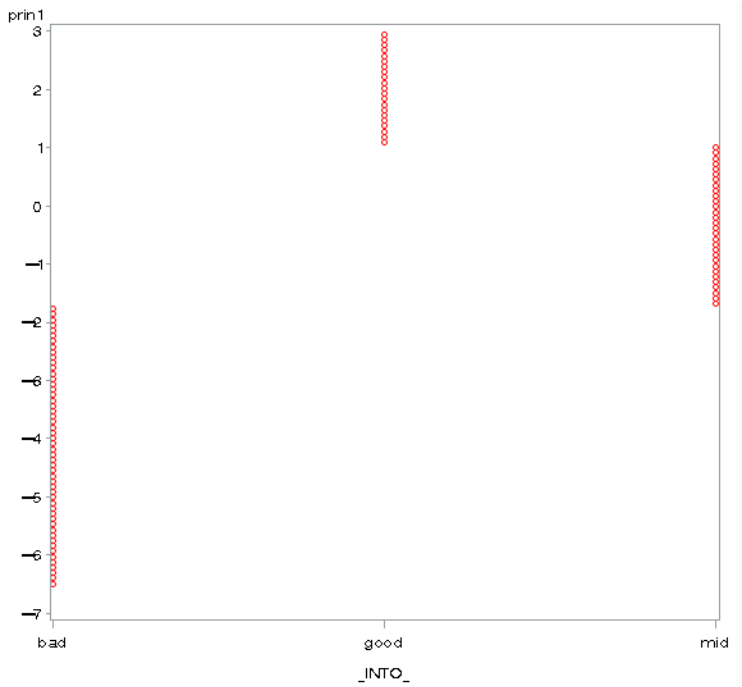

Figure 5.2.1 Estimate of Likelihood Function for different groups

From Figure 5.2.1, we can find out that prin1 less than -1.7 will be grouped into bad. This group has bad health sources and lack of high tech facilities. People in these countries live shorter than other countries, less than 64 years old. Prin1 larger than 1 will be grouped in to good, more than 75 years old. The countries in this group have good sanitation sources and high-tech facilities. People in these countries live longer than people in other groups. Prin1 greater than -1.7 but less than 1 will be grouped into medium. Countries in this group have medium level of health sources and high-tech facilities. People in these countries live between 64 and 75 years old.

The model has 79.46% accuracy in training data and 85.19% in validation data.

Table 5.2.2 Classification summary for training data(left) and validation data(right)

Number of Observations and Percent Classified into index				
From index	0	1	2	Total
0	24 92.31	2 7.69	0 0.00	26 100.00
1	8 14.55	29 52.73	18 32.73	55 100.00
2	0 0.00	3 6.67	42 93.33	45 100.00
Total	32 25.40	34 26.98	60 47.62	126 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for index				
	0	1	2	Total
Rate	0.0769	0.4727	0.0667	0.2054
Priors	0.3333	0.3333	0.3333	

Table of index by pred				
	pred			
index	0	1	2	Total
0	14 25.93 93.33 93.33	1 1.85 6.67 6.67	0 0.00 0.00 0.00	15 27.78
1	1 1.85 5.26 6.67	13 24.07 68.42 86.67	5 9.26 26.32 20.83	19 35.19
2	0 0.00 0.00 0.00	1 1.85 5.00 6.67	19 35.19 95.00 79.17	20 37.04
Total	15 27.78	15 27.78	24 44.44	54 100.00

5.3 Logistic Regression Analysis Using Original Variables

Logistic regression is a method of classification when the response variable is categorical. The method relies on calculating the posterior probability of each country with the given input variables and then classify it to the category or group with the highest posterior probability. For example, we define p1=g(Y=0|x), p2= g(Y=1|x), and p3= g(Y=2|x) as shown below. If P1>P2 AND P1>P3 then we classify to bad category(index=0). In our dataset, we have a response variable (life expectancy) with three levels:0=bad, 1=medium, and 2=good.

We run backward elimination to extract the significant variables from our original dataset with all the variables and we arrive at the final model with only the variables MortRate, Tuberculosis, and unemployment.

Table 5.3.1 Summary of Backward Elimination

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	InternetUsers	1	9	0.0343	0.8531
2	Measles	1	8	0.1019	0.7495
3	Population	1	7	0.1444	0.7040
4	WaterSource	1	6	0.5481	0.4591
5	Cellphone	1	5	1.7719	0.1831
6	Sanitation	1	4	1.4444	0.2294
7	FertRate	1	3	3.1984	0.0737

Table 5.3.2 MLE estimates of parameters.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0	1	-14.3838	2.7735	26.8965	<.0001
Intercept	1	1	-2.1323	0.7382	8.3428	0.0039
MortRate		1	0.2202	0.0399	30.5041	<.0001
tuberculosis		1	0.00762	0.00275	7.7001	0.0055
unemployment		1	-0.1453	0.0709	4.2002	0.0404

We fit the model using Maximum likelihood estimates for the parameter in Table 5.3.2. The fitted model is below,

$$\text{logit}(\Pr(\text{index}=0)) = -14.3838 + 0.2202 \times \text{MortRate} + 0.00762 \times \text{tuberculosis} - 0.1453 \times \text{unemployment}.$$

$$\text{Logit}(\Pr(\text{index}<=1)) = -2.1323 + 0.2202 \times \text{MortRate} + 0.00762 \times \text{tuberculosis} - 0.1453 \times \text{unemployment}$$

Then we can calculate the probability for different groups. Their probability is below,

P1=Pr(index=0|x) = **1**/(**1**+exp(-logit(pr(index=0))))

=**1**/(**1**+exp(-(-**14.3838**+**0.2202***MortRate+**0.00762***tuberculosis-**0.1453***unemployment)))

P2=Pr(index=1|x) = **1**/(**1**+exp(-logit(pr(index<=1)))) - **1**/(**1**+exp(-logit(pr(index=0))))

= **1**/(**1**+exp(-(-**2.1323**+**0.2202***MortRate+**0.00762***tuberculosis-**0.1453***unemployment)))

- **1**/(**1**+exp(-(-**14.3838**+**0.2202***MortRate+**0.00762***tuberculosis-**0.1453***unemployment)))

P3=Pr(index=2|x)=1-p1-p2

We will group the observation based on the largest probability. For example, we calculate p1, p2, and p3 for an observation. If p1 is the largest one, then we group the observation into group bad(index=0).

The model has 88.89% accuracy in training data and 79.63% in validation data.

Table 5.3.3 Classification summary for training data(left) and validation data(right)

Table of index by PREDICT				
	PREDICT			
index	0	1	2	Total
0	22 17.46 100.00 95.65	0 0.00 0.00 0.00	0 0.00 0.00 0.00	22 17.46
1	1 0.79 1.79 4.35	48 38.10 85.71 88.89	7 5.56 12.50 14.29	56 44.44
2	0 0.00 0.00 0.00	6 4.76 12.50 11.11	42 33.33 87.50 85.71	48 38.10
Total	23 18.25	54 42.86	49 38.89	126 100.00

Table of index by PREDICT				
	PREDICT			
index	0	1	2	Total
0	12 22.22 80.00 85.71	3 5.56 20.00 18.75	0 0.00 0.00 0.00	15 27.78
1	2 3.70 10.53 14.29	12 22.22 63.16 75.00	5 9.26 26.32 20.83	19 35.19
2	0 0.00 0.00 0.00	1 1.85 5.00 6.25	19 35.19 95.00 79.17	20 37.04
Total	14 25.93	16 29.63	24 44.44	54 100.00

5.4 Logistic Regression Analysis Using Principal Component

In this section, we use principal component to run logistic regression. We run backward selection of principal component 1 and the two variables, population and unemployment, which are not included in principal component, to see whether it generates a different model with better predictive power. The final model has only prin1.

Table 5.4.1 Summary of Backward selection.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	Population	1	2	0.1313	0.7170
2	unemployment	1	1	2.2304	0.1353

Table 5.4.2 MLE parameter estimates for logistic function.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0	1	-4.2059	0.7677	30.0148	<.0001
Intercept	1	1	2.7091	0.5269	26.4399	<.0001
Prin1		1	-1.8331	0.2856	41.1892	<.0001

We use MLE to estimate the parameter for the model. The fitted model is below,

Logit(Pr(index=0))= -**4.2059**-**1.8331***prin1

Logit(pr(index<=1))= **2.7091**-**1.8331***prin1

Then we can calculate the probability for different groups. Their probability is below,

P1=Pr(index=0|x) = **1**/(**1**+exp(-logit(pr(index=0))))

=**1**/(**1**+exp(-(-**4.2059**-**1.8331***prin1)))

P2=Pr(index=1|x) = **1**/(**1**+exp(-logit(pr(index<=1)))) - **1**/(**1**+exp(-logit(pr(index=0))))

=**1**/(**1**+exp(-(**2.7091**-**1.8331***prin1)))- **1**/(**1**+exp(-(-**4.2059**-**1.8331***prin1)))

P3=Pr(index=2|x) =1-p1-p2

We will group the observation based on the largest probability. For example, we calculate p1, p2, and p3 for an observation. If p1 is the largest one, then we group the observation into group bad(index=0).

The model has 80.16% accuracy in training data and 79.63% in validation data.

Table 5.4.3 Classification summary for training data(left) and validation data(right)

Table of index by PREDICT				
	PREDICT			
index	0	1	2	Total
0	22 17.46 84.62 84.62	4 3.17 15.38 8.00	0 0.00 0.00 0.00	26 20.63
1	4 3.17 7.27 15.38	40 31.75 72.73 80.00	11 8.73 20.00 22.00	55 43.65
2	0 0.00 0.00 0.00	6 4.76 13.33 12.00	39 30.95 86.67 78.00	45 35.71
Total	26 20.63	50 39.68	50 39.68	126 100.00

Table of index by PREDICT				
	PREDICT			
index	0	1	2	Total
0	13 24.07 86.67 92.86	2 3.70 13.33 9.09	0 0.00 0.00 0.00	15 27.78
1	1 1.85 5.26 7.14	15 27.78 78.95 68.18	3 5.56 15.79 16.67	19 35.19
2	0 0.00 0.00 0.00	5 9.26 25.00 22.73	15 27.78 75.00 83.33	20 37.04
Total	14 25.93	22 40.74	18 33.33	54 100.00

6 Conclusion

In the exploratory phase, we learned three things: first, cluster analysis provided us with information about the three possible clusters in our dataset based on their similarities (input variables) and their means for creating the index for the classification step. Second, correlation matrix allows us to exclude the uncorrelated variables, population and unemployment for principal component analysis. Lastly, principal component analysis reduces the eight variables to one principal component, which accounts for 68 percent of the total variation.

We build four models to determine the "best" model with the smallest misclassification rate. The results are as follow:

1. Discrimination method has 83.97% accuracy in training data and 85.19% in validation data using all variables to fit the model. It has 79.46% accuracy in training data and 85.19% in validation data using only the first principal component.
2. Logistic regression has 88.89% accuracy in training data and 79.63% in validation data using all variable to fit the model. It has 80.16% accuracy in training data and 79.63% in validation data using only the first principal component.

We conclude that the model built by discrimination using the first principal component is the best although it performs not good at training data. Moreover, using principal component can reduce the variables into only one variables. This can increase the efficiency for discrimination. But overall, all of the models have good predictive power.

From the classification analysis above, we conclude that the countries with good health sources and high-tech facilities can sufficiently reduce the fatal disease, such as

tuberculosis. People in these countries tend to live longer. On the other hand, the countries with poor health sources and lack of high tech facilities may be hard to handle lethal sickness. People in these countries tend to live shorter.

Appendix 1

Sas code

```
proc import datafile="F:\course\550\project\good_data_revisedtest.csv"
     out=life_expect
     dbms=csv
     replace;
     getnames=yes;
run;

/*Removing missing data*/
DATA life_expect1;
SET life_expect;
CHK=FertRate+Measles+Sanitation+WaterSource+InternetUsers+Cellphone+MortRate+Population+t
uberculosis+unemployment+lifeExpectancy;
 IF CHK=. THEN DELETE;  DROP CHK;
 /*get corr matrix for preliminary analysis*/
proc corr data=life_expect1;
run;



PROC CLUSTER DATA=life_expect1 S STANDARD METHOD=average
     PSEUDO OUTTREE=TREELIFE;
 VAR FertRate Measles Sanitation WaterSource InternetUsers Cellphone MortRate Population
tuberculosis  unemployment lifeExpectancy;
 ID Code;
 RUN;

 PROC GPLOT DATA=TREELIFE;
 PLOT _PSF_*_NCL_= 1 /HAXIS=0 TO 20 BY 1 VAxis=0 to 300 by 10 VAxis=Axis1;
 Axis1 Label=(A=90)
 order=0 to 300 by 100;
 Symbol1 C=Black V=Dot I=SplineS;
 RUN; quit;

PROC TREE DATA=TREELIFE OUT=TREEOUTLIFE NCLUSTERS=3 VAxis=Axis1;
 COPY FertRate Measles Sanitation WaterSource InternetUsers Cellphone MortRate Population
tuberculosis unemployment lifeExpectancy;
 ID Code;
 Axis1 Label=(A=90);
 RUN;

 PROC SORT DATA=TREEOUTLIFE; BY CLUSTER;
 proc print data=treeoutlife;
 by cluster;
 run;
 proc means data=treeoutlife;
 by cluster;
 var lifeExpectancy;
 run;
 /*from corr matrix, we find out that population and unemployment are not related to
other varaibles(p-value>0.05)*/
/*They should not be included in PCA and FA*/
proc princomp data=life_expect1 out=prin;
var FertRate--mortrate tuberculosis ;
run;
```

```
 /*from above we can define tree groups. good is >=75, medium is 64<=
lifeExpectancy<75,bad is <64*/
data rawdat;
set prin;
if lifeExpectancy>=75 then index=2;
if lifeExpectancy <75 and lifeExpectancy>=64 then index=1;
if lifeExpectancy<64 then index=0;
if index=2 then Cat="good";
if index=1 then Cat="mid";
if index=0 then cat="bad";
run;
/*partition the data into training group and test group*/
proc surveyselect  data=rawdat out=split  rate=0.7 outall;
run;
proc sql;
create table life as select * from split where selected=1;
quit;
proc sql;
create table test as select * from split where selected=0;
quit;

/*classify using orginal variables,without reducing the varaibles.*/
title1 "discrimination";
proc discrim data= life crossvalidate mahalanobis;
class index;
var FertRate--unemployment;
run;
title2 'Stepwise Seletion';
PROC STEPDISC DATA=life METHOD=Stepwise SLE=.40 SLS=.05;
 CLASS index;
 VAR FertRate--unemployment;
 RUN;
/* using only significant variables*/
 proc discrim data=life crossvalidate mahalanobis;
class index;
var Sanitation tuberculosis InternetUsers MortRate;
run;
/*validate discrimination model,using Maximum Likelyhood function*/
data test1;
set test;
L0=-
40.19829+0.37061*Sanitation+0.02190*tuberculosis+0.23326*InternetUsers+0.72252*Mortrate;
L1=-
26.16463+0.44593*Sanitation+0.01526*tuberculosis+0.11563*InternetUsers+0.48469*Mortrate;
L2=-
31.20009+0.45058*Sanitation+0.01253*tuberculosis+0.21249*InternetUsers+0.46791*Mortrate;
run;
data test1;
set test1;
if L0>L1 and L0>L2 then pred=0;
if L1>L0 and L1>L2 then pred=1;
if L2>L0 and L2>L1 then pred=2;
run;
/*we can use this table to calculate missclassification*/
PROC FREQ;
 TABLES index*pred;
 RUN;

/*plot main effect and contour*/
/*from Stepdisc,F values of Mortrate and Internetusers are the laregest two.Plot them*/
Proc Means Data=life NoPrint;
 Var Mortrate Internetusers;
 Output Out=life_m Min=MinM MinI Max=MaxM MaxI;
```

```
Run;

Data Plotlife;
 If _N_=1 Then Set life_m;
 IncM=(MaxM-MinM)/50;
 IncI=(MaxI-MinI)/50;
 Do Mortrate = (MinM-IncM) To (MaxM+IncM) By IncM;
   Do Internetusers = (MinI-IncI) To (MaxI+IncI) By IncI;
     Output;
     Keep Mortrate Internetusers;
   End;
 End;
 Stop;
Run;

Proc DISCRIM Data=life
             Testdata=Plotlife TestOut=PlotP TestOutD=PlotD;
   Class cat;

Run;

/*Mortrate and Internetusers are the most important in stepwise selection*/
 PROC GPLOT DATA=life;
 plot Mortrate*Internetusers = Index/ HAxis=Axis1 VAxis=Axis2;
 Axis1  Order=(-10 To 110 By 10);
 Axis2  Order=(-10To 200 By 10);
 Symbol1 V=circle H=0.7 I=None C=RED;
 Symbol2 V=Star H=0.7 I=None C=BLACK;
 symbol3 V=dot H=0.7 I=None C=BLUE;
run; quit;

Title2 'Plot of Estimated Densities';
Proc GContour Data=PlotD;
 Title3 "life";
plot Mortrate*Internetusers = bad/ HAxis=Axis1 VAxis=Axis2;
  Axis1  Order=(-10 To 110 By 10);
 Axis2  Order=(-10To 200 By 10);
Run; quit;

Proc GContour Data=PlotD;
 Title3 "life";
plot Mortrate*Internetusers = mid/ HAxis=Axis1 VAxis=Axis2;
  Axis1  Order=(-10 To 110 By 10);
 Axis2  Order=(-10To 200 By 10);
Run; quit;

Proc GContour Data=PlotD;
 Title3 "life";
plot Mortrate*Internetusers = good/ HAxis=Axis1 VAxis=Axis2;
  Axis1  Order=(-10 To 110 By 10);
 Axis2  Order=(-10To 200 By 10);
Run; quit;

Title2 'Plot of Classification Results';
Proc GPlot Data=PlotP;
 Plot Mortrate*Internetusers=_Into_;
Symbol1 V=circle H=0.7 I=None C=RED;
 Symbol2 V=Star H=0.7 I=None C=BLACK;
 symbol3 V=dot H=0.7 I=None C=BLUE;
Run; quit;
```

```
TITLE2 'logistic';
PROC LOGISTIC DATA=life;
 MODEL index =FertRate-- unemployment/SELECTION=BACKWARD SLSTAY=.05;
 OUTPUT OUT=PDICTS  PREDICTED=PHAT;

DATA ONE;
 SET PDICTS;
 IF _LEVEL_=0 THEN P1=PHAT;

 IF _LEVEL_=0;

DATA TWO;
 SET PDICTS;
 IF _LEVEL_=1 THEN P2=PHAT;
 IF _LEVEL_=1;

DATA THREE; DROP _LEVEL_;
 MERGE ONE TWO;
 P2=P2-P1;
 P3=1-P1-P2;
RUN;

DATA FINAL; SET THREE;
 IF P1>P2 AND P1>P3 THEN PREDICT=0;
 IF P2>P1 AND P2>P3 THEN PREDICT=1;
 IF P3>P1 AND P3>P2 THEN PREDICT=2;
 run;

 PROC PRINT data=final;

VAR index P1 P2 P3 PREDICT;

PROC FREQ;
 TABLES index*PREDICT;
 RUN;


/*validate logistic model*/
data test2;
set test;
ph1=1/(1+exp(-(-14.3838+0.2202*MortRate+0.00762*tuberculosis-0.1453*unemployment)));
ph2=1/(1+exp(-(-2.1323+0.2202*MortRate+0.00762*tuberculosis-0.1453*unemployment)));
p1=ph1;
p2=ph2-ph1;
p3=1-p1-p2;
run;
DATA test2; SET test2;
 IF P1>P2 AND P1>P3 THEN PREDICT=0;
 IF P2>P1 AND P2>P3 THEN PREDICT=1;
 IF P3>P1 AND P3>P2 THEN PREDICT=2;
 run;

 PROC PRINT data=test2;

VAR index P1 P2 P3 PREDICT;

PROC FREQ;
```

```
 TABLES index*PREDICT;
 RUN;



/*discrimination using principal component method to reduce variables*/
 /*only the eigen value of prin1 is greater than 1,we choose prin1*/
proc stepdisc data=life method=stepwise sle=0.4 sls=0.05;
class index;
var prin1 population unemployment;
run;
/*from above we can find out that population and unemployment are insiginificant*/
proc discrim data=life crossvalidate mahalanobis;
class index;
var prin1;
run;
/*reduce the variables, but missclassification rate very close.*/


/*validate discrimination model,using Maximum Likelyhood function*/
/*this can be found on table Linear Discriminant Function for index*/
data test2;
set test;
L0=-3.60887-2.20305*prin1;
L1=-0.00144+0.04395*prin1;
L2=-1.39445+1.36943*prin1;
if L0>L1 and L0>L2 then pred=0;
if L1>L0 and L1>L2 then pred=1;
if L2>L0 and L2>L1 then pred=2;
run;
/*calcalate missclassification*/
PROC FREQ data=test2;
 TABLES index*pred;
 RUN;


TITLE2 'logistic,using principal componient ';
PROC LOGISTIC DATA=life;
 MODEL index =prin1 population unemployment/SELECTION=BACKWARD SLSTAY=.05;
 OUTPUT OUT=PDICTS  PREDICTED=PHAT;

DATA ONE;
 SET PDICTS;
 IF _LEVEL_=0 THEN P1=PHAT;

 IF _LEVEL_=0;

DATA TWO;
 SET PDICTS;
 IF _LEVEL_=1 THEN P2=PHAT;
 IF _LEVEL_=1;

DATA THREE; DROP _LEVEL_;
 MERGE ONE TWO;
 P2=P2-P1;
 P3=1-P1-P2;
RUN;

DATA FINAL; SET THREE;
 IF P1>P2 AND P1>P3 THEN PREDICT=0;
 IF P2>P1 AND P2>P3 THEN PREDICT=1;
 IF P3>P1 AND P3>P2 THEN PREDICT=2;
```

```
 run;

 PROC PRINT data=final;

VAR index P1 P2 P3 PREDICT;

PROC FREQ;
 TABLES index*PREDICT;
 RUN;
 /*validate logistic model*/
data test2;
set test;
ph1=1/(1+exp(-(-4.2059-1.8331*prin1)));
ph2=1/(1+exp(-(2.7091-1.8331*prin1)));
p1=ph1;
p2=ph2-ph1;
p3=1-p1-p2;
run;
DATA test2; SET test2;
 IF P1>P2 AND P1>P3 THEN PREDICT=0;
 IF P2>P1 AND P2>P3 THEN PREDICT=1;
 IF P3>P1 AND P3>P2 THEN PREDICT=2;
 run;

 PROC PRINT data=test2;

VAR index P1 P2 P3 PREDICT;

PROC FREQ;
 TABLES index*PREDICT;
 RUN;
```

Appendix 2

Classification Table

index=0

Obs	LifeExpectancy	Country
1	60.37446341	Afghanistan
2	52.26687805	Angola
3	59.51058537	Benin
4	58.58846341	Burkina Faso
5	56.69202439	Burundi
6	55.4927561	Cameroon
7	51.55580488	Chad
8	63.25685366	Comoros
9	58.65919512	Congo, Dem. Rep.
10	62.31114634	Congo, Rep.
11	51.55958537	Cote d'Ivoire
12	62.0155122	Djibouti
13	57.64704878	Equatorial Guinea
14	63.66302439	Eritrea
15	60.22843902	Gambia, The
16	61.31163415	Ghana
17	58.73343902	Guinea
18	55.16004878	Guinea-Bissau
19	62.74743902	Haiti
20	61.57636585	Kenya
21	49.70058537	Lesotho
22	60.83441463	Liberia
23	62.72163415	Malawi
24	57.98626829	Mali
25	63.01658537	Mauritania
26	55.02595122	Mozambique
27	61.4584878	Niger
28	52.75426829	Nigeria
29	62.60692683	Papua New Guinea
30	63.96565854	Rwanda
31	50.87878049	Sierra Leone
32	55.35480488	Somalia
33	57.18212195	South Africa
34	55.68221951	South Sudan
35	58.55893507	Sub-Saharan Africa
36	63.45853659	Sudan
37	48.93473171	Swaziland
38	59.65580488	Togo
39	58.46641463	Uganda
40	60.04704878	Zambia
41	57.49831707	Zimbabwe

index=1

Obs	LifeExpectancy	Country
42	74.80809756	Algeria
43	70.76321951	Azerbaijan
44	71.62590244	Bangladesh
45	72.97560976	Belarus
46	70.07743902	Belize
47	69.4712439	Bhutan
48	68.344	Bolivia
49	64.4292439	Botswana
50	74.40187805	Brazil
51	73.147	Cabo Verde
52	68.21229268	Cambodia
53	73.99314634	Colombia
54	73.50002439	Dominican Republic
55	74.92852462	East Asia & Pacific
56	71.12170732	Egypt, Arab Rep.
57	72.75456098	El Salvador
58	64.03502439	Ethiopia
59	70.08912195	Fiji
60	64.38339024	Gabon
61	74.66863415	Georgia
62	73.36631707	Grenada
63	71.72241463	Guatemala
64	66.40641463	Guyana
65	73.13570732	Honduras
66	68.01380488	India
67	68.8884878	Indonesia
68	69.39968293	Iraq
69	74.05214634	Jordan
70	71.62	Kazakhstan
71	65.95168293	Kiribati
72	70.07468293	Korea, Dem. People?
73	74.58502439	Kuwait
74	70.40243902	Kyrgyz Republic
75	66.11736585	Lao PDR
76	74.94206018	Latin America & Car
77	74.18780488	Latvia
78	73.96585366	Lithuania
79	65.08560976	Madagascar
80	74.71829268	Malaysia
81	74.19439024	Mauritius
82	72.81665963	Middle East & North
83	71.45587805	Moldova
84	69.46390244	Mongolia
85	74.01609756	Morocco
86	65.85785366	Myanmar

V

86	65.85785366	Myanmar
87	64.68019512	Namibia
88	69.60468293	Nepal
89	74.81014634	Nicaragua
90	66.18336585	Pakistan
91	72.92170732	Paraguay
92	74.52553659	Peru
93	68.26563415	Philippines
94	70.36585366	Russian Federation
95	73.51182927	Samoa
96	66.38460976	Sao Tome and Princi
97	74.33721951	Saudi Arabia
98	66.37258537	Senegal
99	73.22926829	Seychelles
100	67.93080488	Solomon Islands
101	68.1210446	South Asia
102	74.79480488	Sri Lanka
103	71.15143902	Suriname
104	70.07102439	Syrian Arab Republi
105	69.59797561	Tajikistan
106	64.94390244	Tanzania
107	74.42202439	Thailand
108	68.25914634	Timor-Leste
109	72.79219512	Tonga
110	70.44056098	Trinidad and Tobago
111	74.14390244	Tunisia
112	65.59853659	Turkmenistan
113	71.18658537	Ukraine
114	71.91831707	Vanuatu
115	74.23619512	Venezuela, RB

index=2

Obs	LifeExpectancy	Country
116	77.83046341	Albania
117	76.15860976	Argentina
118	82.25121951	Australia
119	81.33658537	Austria
120	75.23365854	Bahamas, The
121	76.68326829	Bahrain
122	75.49641463	Barbados
123	80.58780488	Belgium
124	76.4332439	Bosnia and Herzegov
125	75.40731707	Bulgaria
126	81.95660976	Canada
127	81.49619512	Chile
128	75.78226829	China
129	79.40270732	Costa Rica
130	77.32926829	Croatia
131	79.39082927	Cuba
132	80.13156098	Cyprus
133	78.27560976	Czech Republic
134	80.54878049	Denmark
135	76.59	Dominica
136	75.8724878	Ecuador
137	77.23902439	Estonia
138	76.89311163	Europe & Central As
139	81.12926829	Finland
140	82.37317073	France
141	80.84390244	Germany
142	81.28536585	Greece
143	75.87317073	Hungary
144	82.06097561	Iceland
145	75.38931707	Iran, Islamic Rep.
146	81.15365854	Ireland
147	82.15365854	Israel
148	82.6902439	Italy
149	75.6535122	Jamaica
150	83.58780488	Japan
151	79.37309756	Lebanon
152	82.20731707	Luxembourg
153	75.34239024	Macedonia, FYR
154	76.77282927	Maldives
155	81.74634146	Malta
156	76.72185366	Mexico
157	76.18070732	Montenegro
158	81.30487805	Netherlands