Final Analysis Report

Submitted By Akshaya R

Introduction:

Questions to be addressed :

1. **Do developed vs. developing countries differ in regards to life expectancy and total expenditure?**

   Null Hypothesis : Developed vs Developed Countries differ in regards to life expectancy and total expenditure

   Analysis = MANOVA

   Variables used: life expectance and total expenditure as the dependent variables and developed vs. developing countries as the independent

2. **Are alcohol consumption, BMI, schooling, Hepatitis B, Polio, Diphtheria related to life expectancy and health expenditure?**

   Null Hypothesis : Alcohol consumption, BMI, schooling, Hepatitis B, Polio, Diphtheria relate to life expectancy and Total expenditure

   Analysis = MV regression

   Variables used: Dependent = life expectancy and Total expenditure; Predictor variables = alcohol consumption, BMI, schooling, vaccination rates for Hepatitis B, Polio, and Diphtheria

3. **Do the mortality rates have an association with the immunization rates ?**

   Null Hypothesis: Mortality rates have an association with the immunization rates

   Analysis: Canonical Correlation

Variables used : Var: adult mortality, infant mortality, under 5 deaths With : hep B, polio, diphtheria).

Data Description:

The dataset is taken from the Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data is from year 2000-2015 for 193 countries.

Data set Dictionary :

Variable Name	Description	Datatype	Accepts Null Values
Country	Country Name	Object	N
Year	Year	Object	N
Status	Developed or Developing	Object	N
Life Expectancy	Life expectancy in age	Object	N
Adult Mortality	Probability of dying between 15 and 60 years per 1000 population	Object	N
infant deaths	Number of infant deaths per 1000 population	Object	N
Alcohol	recorded per capita consumption(in litres)	Object	N
percentage expenditure	Expendidture on health as per GDP(%)	Object	N
Hepatitis B	Immunization coverage among 1 year old(%)	Object	N
Measles	Number of reported cases per 1000 population	Object	N
BMI	Average BMI of entire population	Object	N
under-five deaths	Number of under five deaths per 1000 population	Object	N
Polio	Immunization coverage amoung one year olds(%)	Object	N

Variable Name	**Description**	**Datatype**	**Accepts Null Values**
Total Expenditure	Government expenditure of health as a percentage of total govt. expenditure(%)	Object	N
Diphtheria	Immunization coverage amoung one year old(%)	Object	N
HIV/AIDS	Deaths per 1000 population	Object	N
GDP	per capita(USD)	Object	N
Population	population of the country	Object	N
thinness 10-19 years	Thinness amomg children from age 10-19(%)	Object	N
thinness 5-9 years	Thinness amomg children from age 5-9(%)	Object	N
Income composition of resources	Index ranging from 0-1	Object	N
Schooling	Number of years of schooling	Object	N

Analysis Methods:

1. Do developed vs. developing countries differ in regards to life expectancy and total expenditure?
   Analysis = MANOVA
   Variables used: life expectance and total expenditure as the dependent variables and developed vs. developing countries as the independent

   Reason for using Manova:

   Manova is used in this hypothesis since it is better at finding the relationship between various independent variables with multiple dependent variables , Manova Is performed using proc glm in sas which is easy to compute the values and it gives an clear interpretation about the results. The variables are taken into account to find the relationship between life expectancy and total expenditure with developed and developing countries.

2. Are alcohol consumption, BMI, schooling, and vaccination rates for Hepatitis B,

Polio, Diphtheria related to life expectancy and total expenditure?
Analysis = MV regression
Variables used: Dependent = life expectancy and total expenditure; Predictor variables = alcohol consumption, BMI, schooling, vaccination rates for Hepatitis B, Polio, and Diphtheria

Reason for using MV Regression:

MV Regression is used in this hypothesis since it is similar to manova which is better at finding the relationship between various independent variables with multiple dependent variables. It is performed using proc glm in sas which given clear interpretation from which the results can be derived. The variables are taken to check how the variables alcohol consumption, BMI, schooling, vaccination rates for Hepatitis B, Polio, and Diphtheria relate to life expectancy and total expenditure.

3. **Do the mortality rates have an association with the immunization rates ?**

Null Hypothesis: Mortality rates have an association with the immunization rates
Analysis: Canonical Correlation
Variables used : Var: adult mortality, infant mortality, under 5 deaths With : hep B, polio, diphtheria).

Reason for using Canonical Correlation:

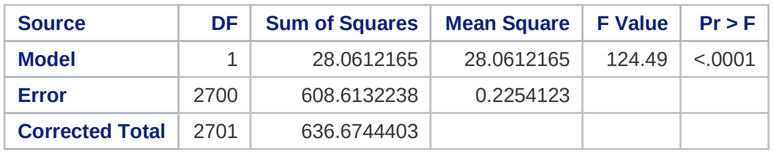
Canonical correlation analysis is used to identify and measure the associations among two sets of variables which are the dependent and independent variables stated above. Canonical correlation is appropriate in the same situations where multiple regression would be, but where are there are multiple intercorrelated outcome variables. Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that best explain the variability both within and between sets. The motive is to check how the various mortality rates relate to the immunization rates.

RESULTS:

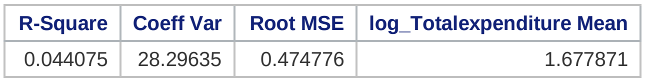
1. Do developed vs. developing countries differ in regards to life expectancy and total expenditure?
   Analysis = MANOVA
   Variables used: life expectance and total expenditure as the dependent variables and developed vs. developing countries as the independent

   The variables were standardized before performing the analysis since the data contained missing values. The mean and standard deviation of the variables are been checked.
   The result includes the interpretation before including the transformations and the results after performing the transformations. The transformation used is the log transformation in which the results does not change much. The results are shown using both QQ-plot and the Histogram.

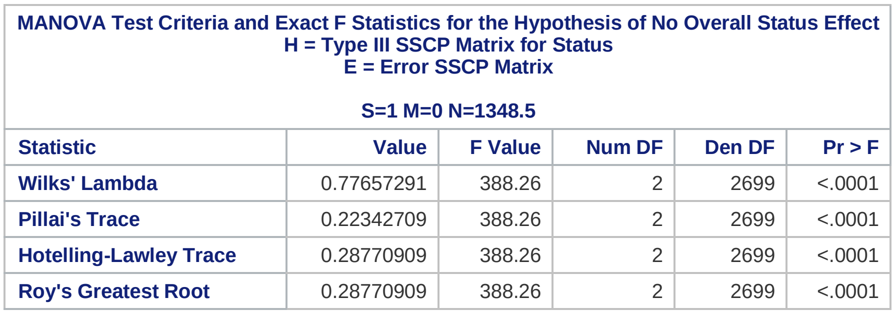
   The output are shown below:



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	28.0612165	28.0612165	124.49	<.0001
Error	2700	608.6132238	0.2254123		
Corrected Total	2701	636.6744403			



R-Square	Coeff Var	Root MSE	log_Totalexpenditure Mean
0.044075	28.29635	0.474776	1.677871



MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Status Effect
H = Type III SSCP Matrix for Status
E = Error SSCP Matrix

S=1 M=0 N=1348.5

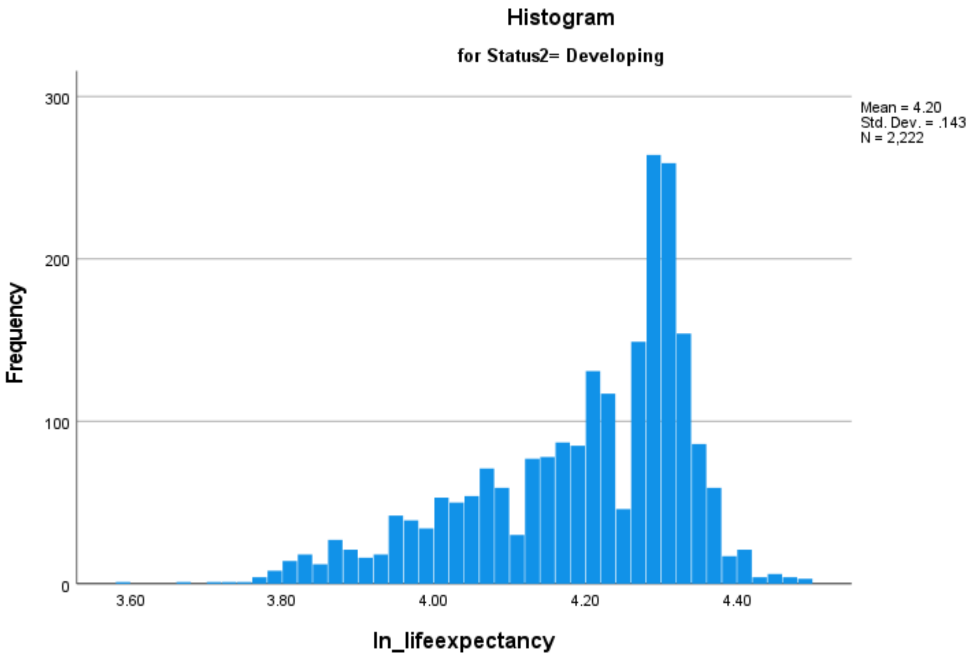
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.77657291	388.26	2	2699	<.0001
Pillai's Trace	0.22342709	388.26	2	2699	<.0001
Hotelling-Lawley Trace	0.28770909	388.26	2	2699	<.0001
Roy's Greatest Root	0.28770909	388.26	2	2699	<.0001

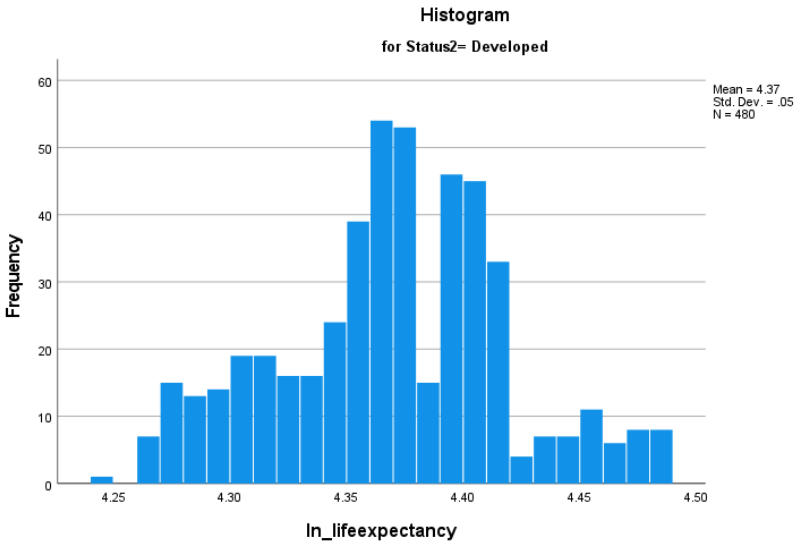
When on the Wilks Lambda and other statistic it says that the P values are significant and we reject the null hypothesis, checking on the R-Square value is not much high but checking on the plots it says that the model fits the data somewhat well.Natural log transformation of the data is been taken since it works for data where the residuals gets bigger for bigger values of the dependent variable.

Natural Log Transformed Data:

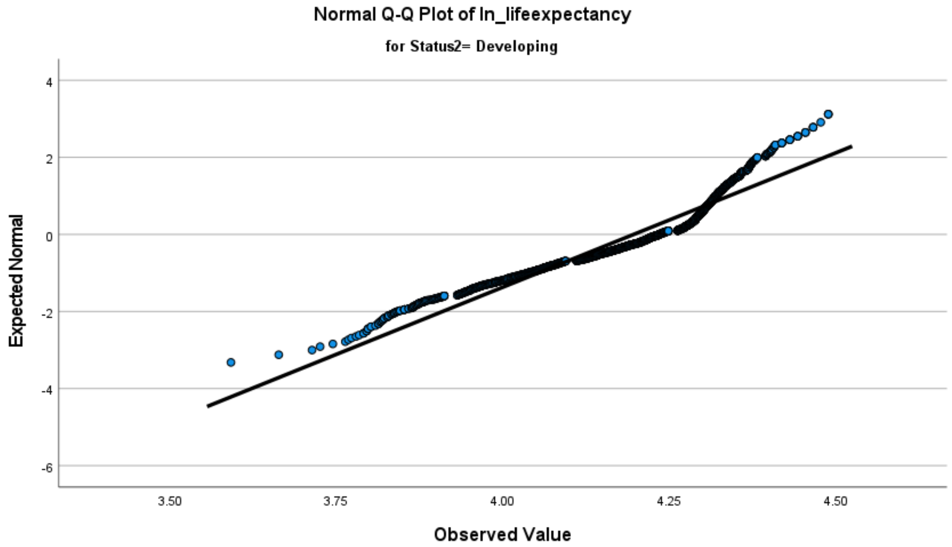
(This chart is also in the MANOVA SAS data)

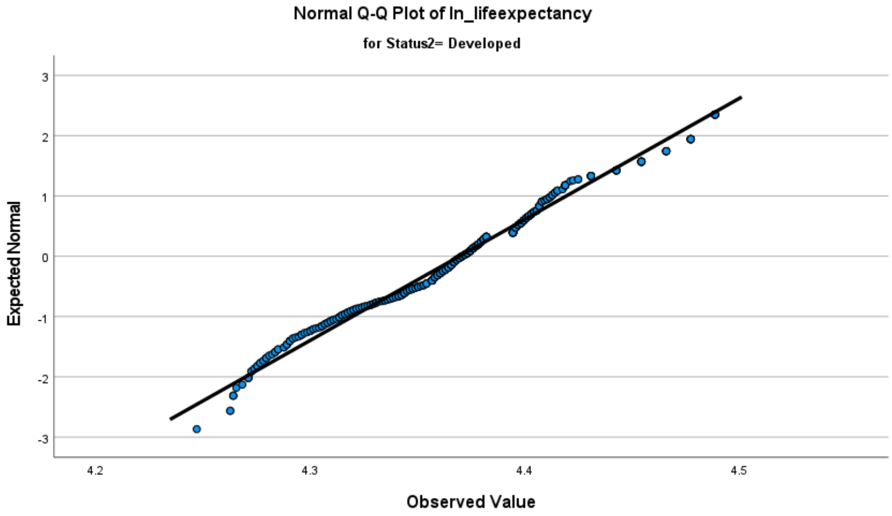
	Mean ± SD	**Variance**
Developing Countries::		
Life Expectancy (ln)	4.19 ± 0.143	0.021
% of Total Expenditure (ln)	1.63 ± 0.447	0.200
Developed Countries:		
Life Expectancy (ln)	4.37 ± 0.049	0.002
% of Total Expenditure (ln)	1.89 ± 0.585	0.343



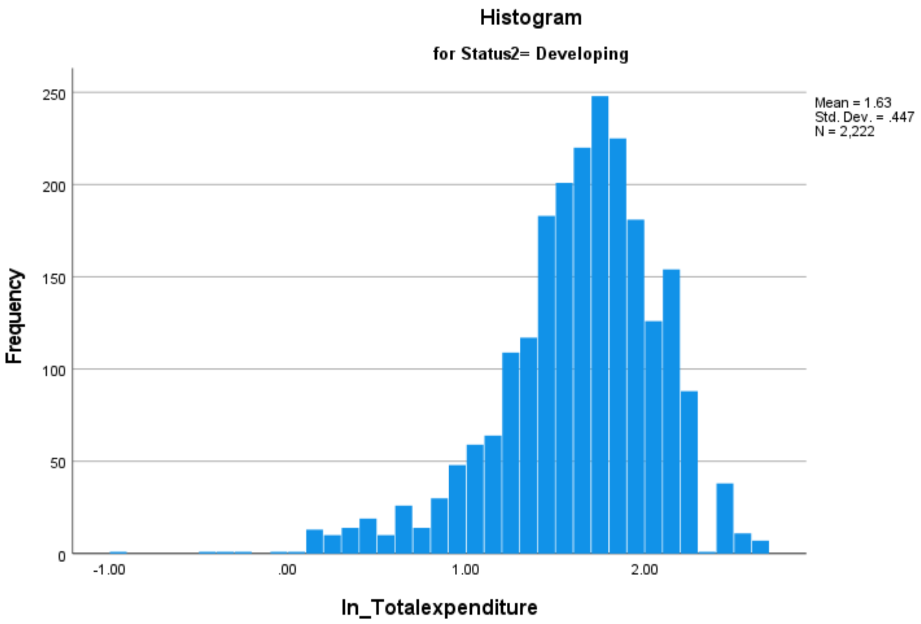


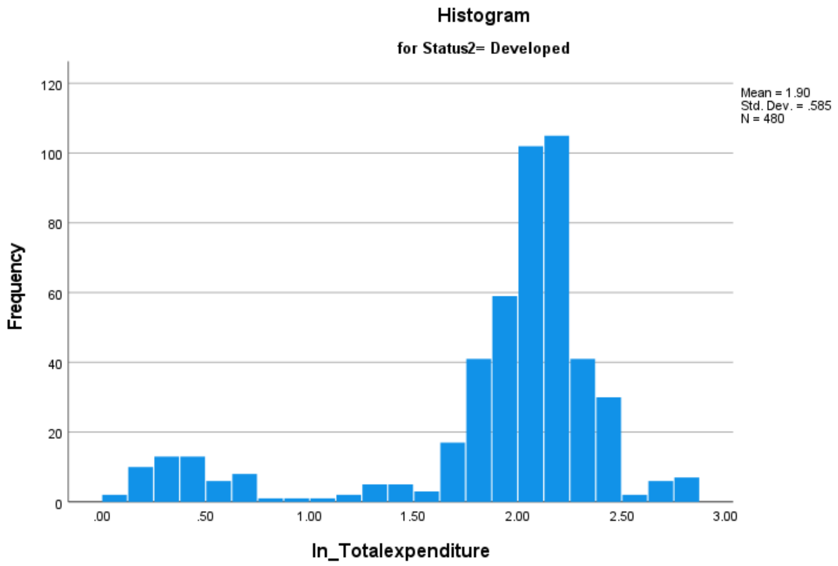
The above histogram plots on developed and developing countries for life expectancy , it could be seen that the mean value of developing countries for life expectancy is about 4.20 with the standard deviation of .142 , the number of observation taken is 2,222. The histogram shows that the life expectancy in the developing country is left skewed. The mean value of developed countries for life expectancy is about 4.37 with the standard deviation of .05 , the number of observation taken is 480 , it is bimodal which represent the maximum frequency.



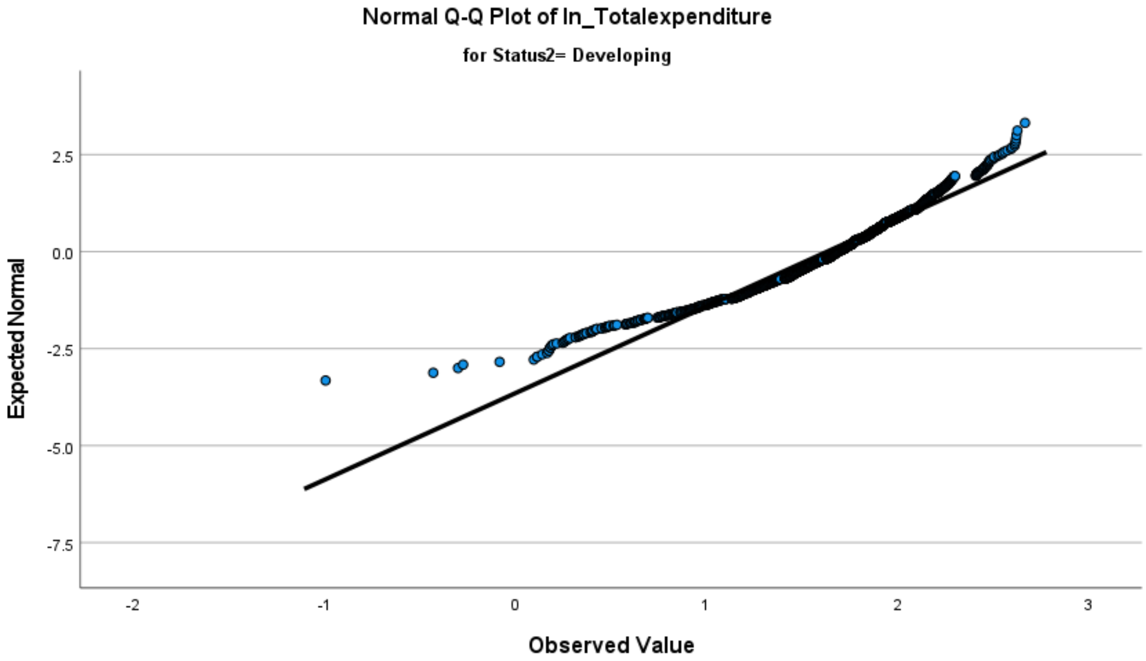


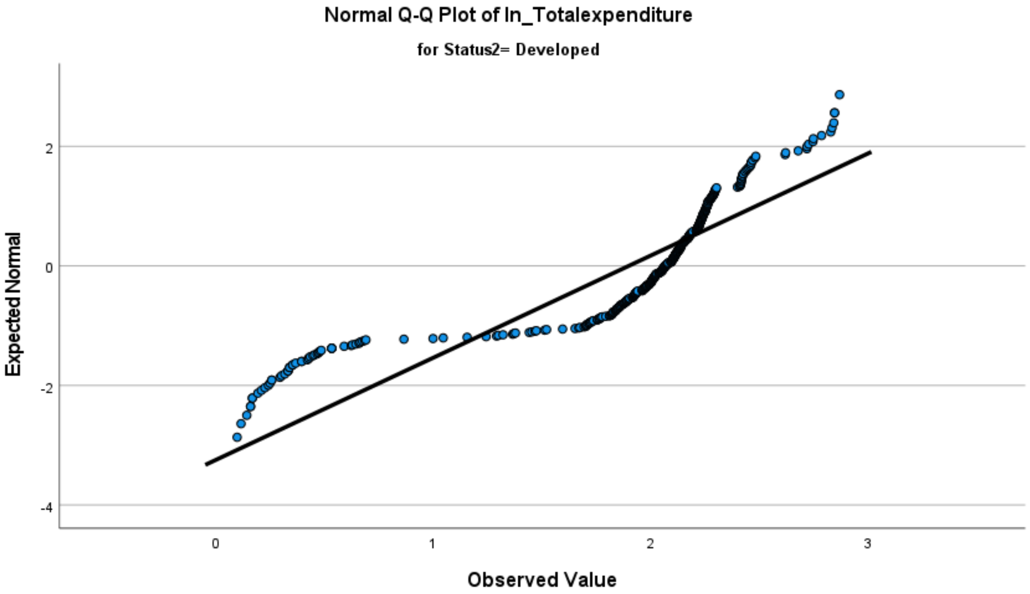
It could be seen from the above QQ plot that the life expectancy is normally distributed in both the developed countries also fit the model somewhat linear. In the developing countries it could be seen that the life expenditure is lightly tailed





The above histogram plots on developed and developing countries for Total expenditure , it could be seen that the mean value of developing countries for Total expenditure is about 1.63 with the standard deviation of .447 , the number of observation taken is 2,222 and also left sweked. The mean value of developed countries for total expenditure is about 1.90 with the standard deviation of .585 , the number of observation taken is 480 and it is a bimodal.



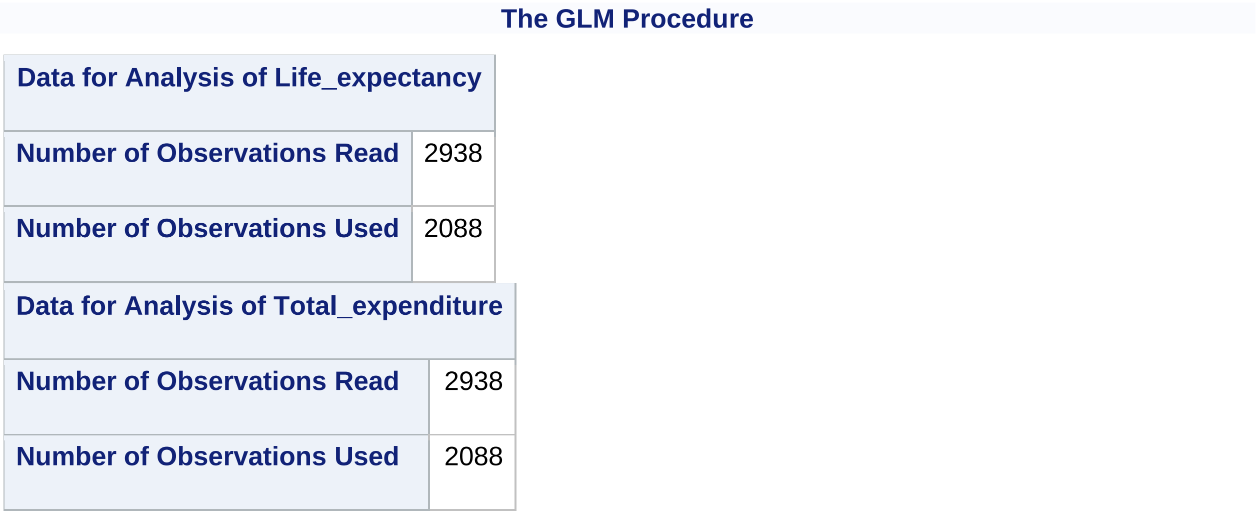


It could be seen from the above QQ plot that the total expenditure is heavily tailed in developing countries and with the developed countries the total expenditure is not normally distributed and is bimodal.

Question 2:

2. Are alcohol consumption, BMI, schooling, and vaccination rates for Hepatitis B, Polio, Diphtheria related to life expectancy and total expenditure?
   Analysis = MV regression
   Variables used: Dependent = life expectancy and total expenditure; Predictor variables = alcohol consumption, BMI, schooling, vaccination rates for Hepatitis B, Polio, and Diphtheria

The variables were standardized before performing the analysis since the data contained missing values. The mean and standard deviation of the variables are been checked. The results are shown using both QQ-plot and the Box plot. The fit diagnostic values gives the interpretation of the residuals , mean square value and the predicted mean square values which can be used for the interpretation. The R-squared values of the variables are taken into account which are ~54% which tells that the data fits good into the model. The output are shown below:

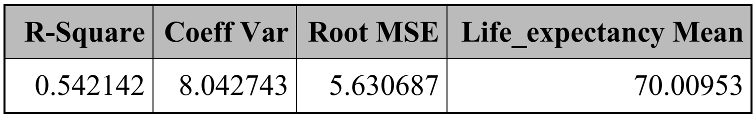


The GLM Procedure

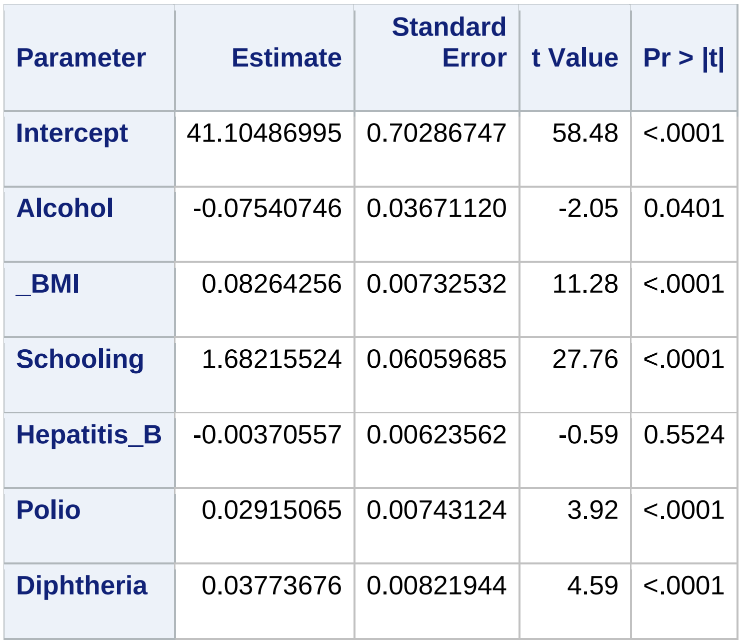
Data for Analysis of Life_expectancy	
Number of Observations Read	2938
Number of Observations Used	2088

Data for Analysis of Total_expenditure	
Number of Observations Read	2938
Number of Observations Used	2088

The below values are for the Life Expectancy

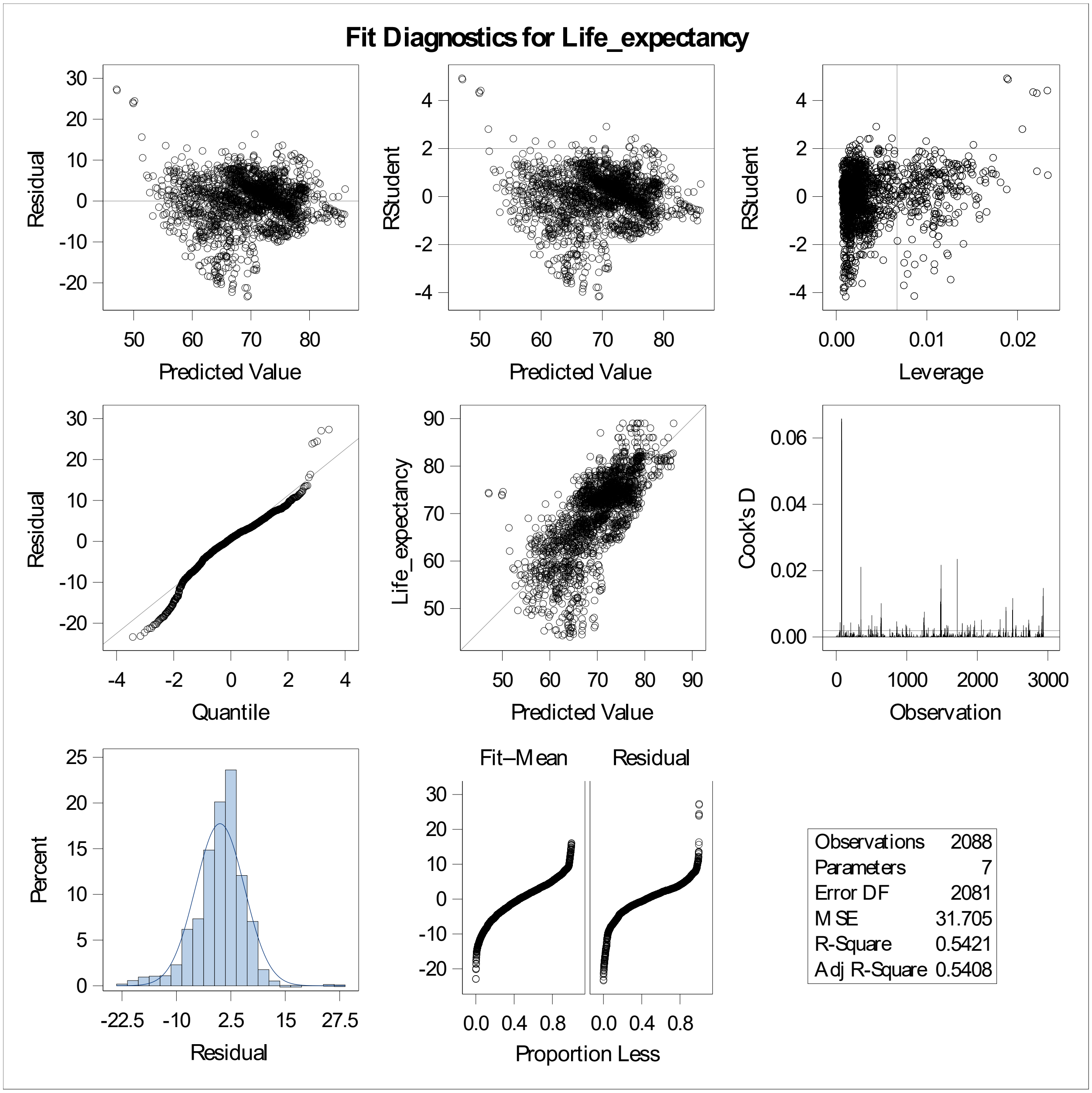


R-Square	Coeff Var	Root MSE	Life_expectancy Mean
0.542142	8.042743	5.630687	70.00953



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	41.10486995	0.70286747	58.48	<.0001
Alcohol	-0.07540746	0.03671120	-2.05	0.0401
_BMI	0.08264256	0.00732532	11.28	<.0001
Schooling	1.68215524	0.06059685	27.76	<.0001
Hepatitis_B	-0.00370557	0.00623562	-0.59	0.5524
Polio	0.02915065	0.00743124	3.92	<.0001
Diphtheria	0.03773676	0.00821944	4.59	<.0001

It could be seen that Hepatitis_B , Alcohol has the p values >0.01 which says that it has impact on the dependent variable and we fail to reject the null hypothesis whereas other independent variables value <0.01 . Since we are testing the dependent variable as a whole how it relates to the Life expectancy.



Fit Diagnostics for Life_expectancy

Observations	2088
Parameters	7
Error DF	2081
MSE	31.705
R-Square	0.5421
Adj R-Square	0.5408

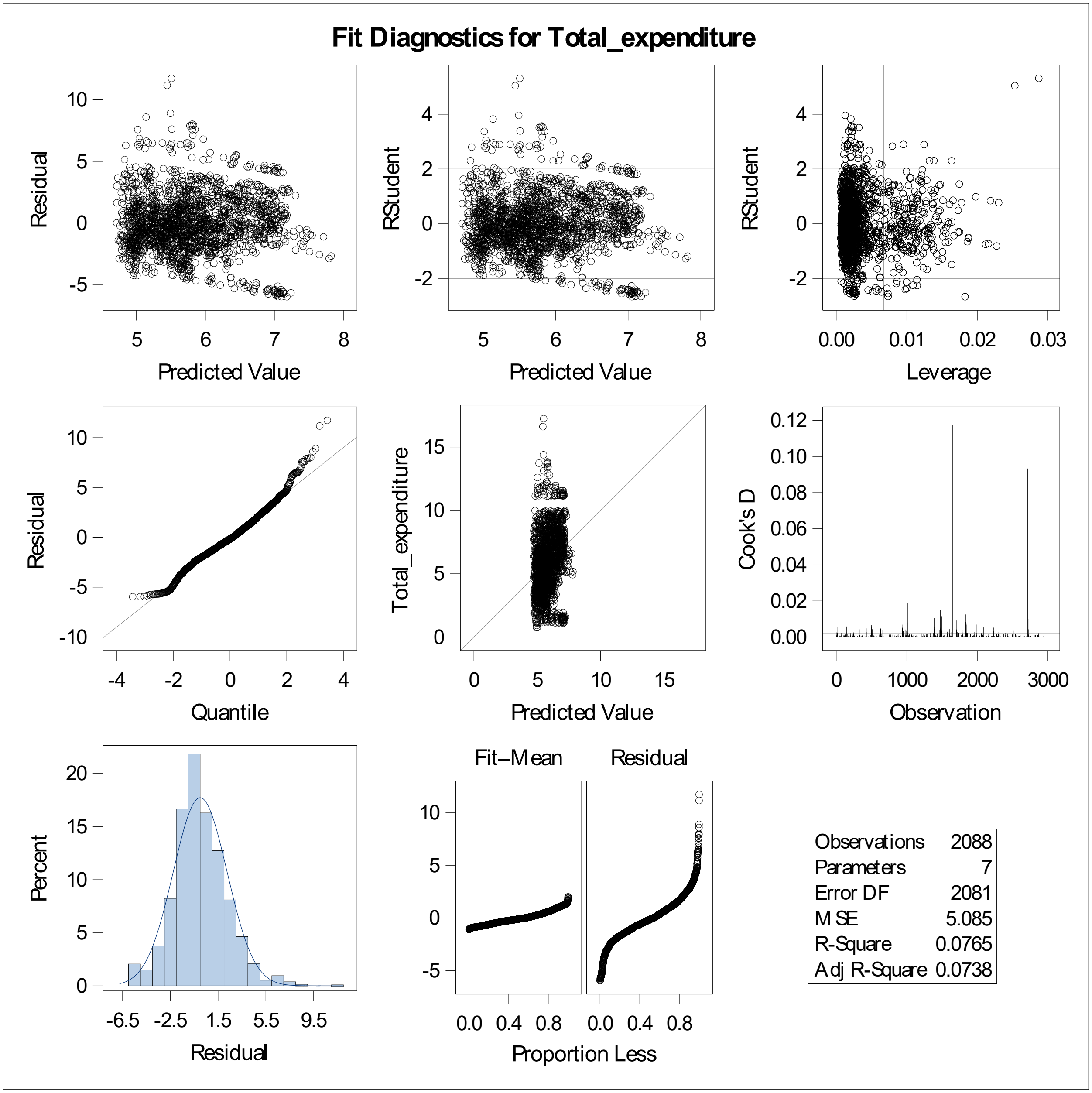
Checking on the predicted variable it could be seen that the vraibles somhow linearly fit the data and from the QQ plot it could be seen that the variables are normally distributes , checking on the residual plot there is still unexplained variation and it does not explain most of the variation among the data .

The below result is for total expenditure :

R-Square	Coeff Var	Root MSE	Total_expenditure Mean
0.076465	38.79982	2.254996	5.811873

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.474611944	0.27914687	16.03	<.0001
Alcohol	0.128325369	0.01468798	8.74	<.0001
_BMI	0.011983187	0.00288958	4.15	<.0001
Schooling	0.011019093	0.02390996	0.46	0.6449
Hepatitis_B	0.001597618	0.00249098	0.64	0.5214
Polio	-0.000693277	0.00298990	-0.23	0.8167
Diphtheria	0.001217824	0.00330746	0.37	0.7128

It could be seen that Hepatitis_B , Schooling , Polio and Diphtheria has the p values >0.01 which says that it has no significance and we fail to reject the null hypothesis whereas other independent variables value <0.01 . Since we are testing the dependent variable as a whole how it relates to the Total expenditure



Fit Diagnostics for Total_expenditure

Observations	2088
Parameters	7
Error DF	2081
MSE	5.085
R-Square	0.0765
Adj R-Square	0.0738

Checking on the predicted variable it could be seen that the variables somehow linearly fit the data and from the QQ plot it could be seen that the variables are normally distributes , checking on the residual plot there is still unexplained variation and it does not explain most of the variation among the data .

Question 3:

Canonical correlation to test the hypothesis that various mortality rates (adult mortality, infant mortality, under 5 deaths) have an association with the immunization rates (hep B, polio, diphtheria).

The canonical correlation gtest is preformed using the proc cancorr in sas the var statement is given as the variables which are tested with the variables in the with statement. Canonical correlation is appropriate in the same situations where multiple regression would be, but where are there are multiple intercorrelated outcome variables.

Results:

The CANCORR Procedure

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.353259	0.351100	0.017963	0.124792	0.1426	0.1268	0.8947	0.8947	0.86075032	40.71	9	5765.7	<.0001
2	0.124832	.	0.020204	0.015583	0.0158	0.0149	0.0993	0.9940	0.98348115	9.91	4	4740	<.0001
3	0.030830	.	0.020504	0.000951	0.0010		0.0060	1.0000	0.99904948	2.26	1	2371	0.1332

Multivariate Statistics and F Approximations

S=3 M=-0.5 N=1183.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Raw Canonical Coefficients for the WITH Variables

	W1	W2	W3
Hepatitis_B	-0.01137098	-0.050208024	-0.000156649
Polio	-0.020768268	0.0163152202	0.0513585183
Diphtheria	-0.020852657	0.0355076715	-0.049990403

Multivariate Statistics and F Approximations

S=3 M=-0.5 N=1183.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.86075032	40.71	9	5765.7	<.0001
Pillai's Trace	0.14132592	39.07	9	7113	<.0001
Hotelling-Lawley Trace	0.15936711	41.94	9	3719.5	<.0001
Roy's Greatest Root	0.14258586	112.69	3	2371	<.0001

The output starts with a sample description and then shows the general fit of the model reporting Pillai's, Helling's, Wilk's and Roy's multivariate criteria. The commonly used test is Wilk's lambda, but we find that all of these tests are significant with p<.05.

The CANCORR Procedure

Canonical Correlation Analysis

Raw Canonical Coefficients for the VAR Variables

	V1	V2	V3
Adult_Mortality	0.0048731453	-0.003577346	-0.006279064
infant_deaths	-0.047853922	0.0471523519	-0.101170959
under_five_deaths	0.0395563473	-0.029183222	0.0758467789

The raw canonical coefficients are similar to the coefficients in linear regression; they can be used to calculate the canonical scores. Only the values with the positive value would be considered significant from the above table and the process continuous for the standardized canonical coefficients.

The CANCORR Procedure

Canonical Correlation Analysis

Standardized Canonical Coefficients for the VAR Variables

	V1	V2	V3
Adult_Mortality	0.5760	-0.4228	-0.7421
infant_deaths	-5.0188	4.9452	-10.6105
under_five_deaths	5.5833	-4.1191	10.7056

Standardized Canonical Coefficients for the WITH Variables

	W1	W2	W3
Hepatitis_B	-0.2845	-1.2561	-0.0039
Polio	-0.4519	0.3550	1.1174
Diphtheria	-0.4468	0.7609	-1.0712

The CANCORR Procedure

Canonical Structure

Correlations Between the VAR Variables and Their Canonical Variables

	V1	V2	V3
Adult_Mortality	0.7216	-0.4307	-0.5420
infant_deaths	0.5925	0.8055	0.0009
under_five_deaths	0.6373	0.7685	0.0567

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables

	V1	V2	V3
Hepatitis_B	-0.2741	-0.0774	-0.0036
Polio	-0.3004	0.0238	0.0151
Diphtheria	-0.3123	0.0252	-0.0130

Correlations Between the WITH Variables and Their Canonical Variables

	W1	W2	W3
Hepatitis_B	-0.7758	-0.6200	-0.1168
Polio	-0.8503	0.1910	0.4904
Diphtheria	-0.8841	0.2015	-0.4215

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables

	W1	W2	W3
Adult_Mortality	0.2549	-0.0538	-0.0167
infant_deaths	0.2093	0.1006	0.0000
under_five_deaths	0.2251	0.0959	0.0017

When considering the canonical correlations WITHIN variables sets, the significant correlates have fairly definable patterns. When considering the mortality rates, all three variables are positively associated with the first canonical variate. While infant death and under 5 death are correlated with the first variate, they are more strongly associated (positively) with the second correlate. When considering the vaccination rates, all three are strongly negatively associated with variate one, while Hepatitis B vaccination also shows a slightly weaker, but still fairly strong association with correlate two.

Interpretation :

Question 1:

Variables are standardized before performing the analysis. The significance of the variables are interpreted using the Wilks' Lambda p-value is <0.0001, the MANOVA is significant and therefore **we reject the null hypotheisis that the developed and developing counties differ in regards to life expectancy and total expenditure.** There is a significant difference between developed and developing countries in regards to life expectancy and total expenditure. By taking the inverse ln to get meaningful representation of the means, in developed countries, the mean life expectancy is ~79 and the health expenditure is ~6.6% of total expenditure.

For developing countries mean life expectancy is ~66 and the health expenditure is ~5.1% of total governmental expenditure. Those in developed countries have a higher life expectancy and spend more of their total governmental expenditure on health.

Question 2:

Variables are standardized before performing the analysis. For the dependent variables Hepatitis B and Alcohol have negative significance on the Life Expectancy and the dependent variables Schooling , Hepatitis B , Diphtheria and Polio have a negative significance towards the total expenditure. The p values for Hepatitis B , Diphtheria and Polio is >0.01 on total expenditure which says that the total expenditure depend on Schooling, Hepatitis B, Diphtheria and Polio. Also the variables Hepatitis B and Alcohol depend on the Life expectancy. Overall it could be interpreted that **we fail to reject the null hypothesis that Alcohol consumption, BMI, schooling, Hepatitis B, Polio, Diphtheria relate to life expectancy and Total expenditure**

Question 3:

Variables were standardized prior to analysis. Three correlates were defined, but only the first two were significant. Those two correlates account for 99% of the variation in the model (the first correlate accounts for 89% of the variance by itself), so a correlation does exist between the various mortality rates and the immunization rates. **We fail to reject the hypothesis that Mortality rates have an association with the immunization rates**

When considering the canonical correlations WITHIN variables sets, the significant correlates have fairly definable patterns. When considering the mortality rates, all three variables are positively associated with the first canonical variate. While infant death and under 5 death are correlated with the first variate, they are more strongly associated (positively) with the second correlate. When considering the vaccination rates, all three are strongly negatively associated with variate one, while Hepatitis B vaccination also shows a slightly weaker, but still fairly strong association with correlate two.

When considering the canonical correlations BETWEEN variable sets, similar patterns emerge for correlate 1, although they are much weaker. Looking at the mortality rates crossed with vaccination rate correlates, there is a positive association with correlate 1. This correlate in the vaccination set had a negative association, so an increase in mortality is associated with decreased vaccination rates. When looking at the cross between vaccination rates and the mortality correlates, the same pattern emerges for correlate 1. When looking at the second correlate in both sets of cross loadings, the associations become extremely weak, so there is no real definable pattern after the first correlate. However, the first correlate accounts for almost 90% of the variation in the model, it would have the most definable associations.