Analysis of Life Expectancy Data
Zongzhen Lee

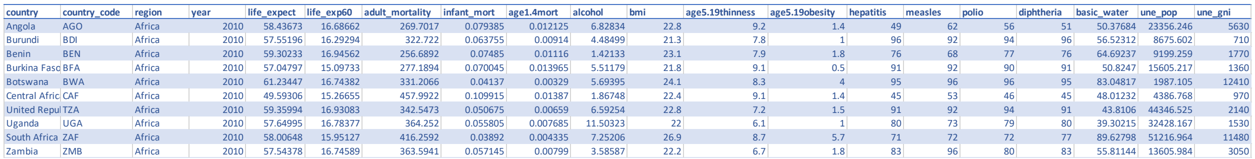
Introduction

This is a paper on performing various statistical analysis on WHO Life Expectancy data obtained from Kaggle [1]. We cleaned and modified the data into two datasets to perform two separate studies on Year 2010 Life Expectancy and Year 2015 Life Expectancy. Each modified dataset used for analysis include 20 variables and 157 observations (countries). Together, we have observations of 314 countries. The incentive to use this data for statistical analysis is because we can cluster countries based on their health performance that includes many factors listed in this dataset, and draw interesting inferences whether, for example, if alcohol consumption correlates with life expectancy or BMI correlate with GNI per capita. The following paper includes: Section 1. Discussion on our data. Section 2. Cluster analysis on the data. Section 3. PC analysis on the data. Section 4. MLE optimization. Section 5. Conclusions from the studies.

Section 1. Discussion on Data

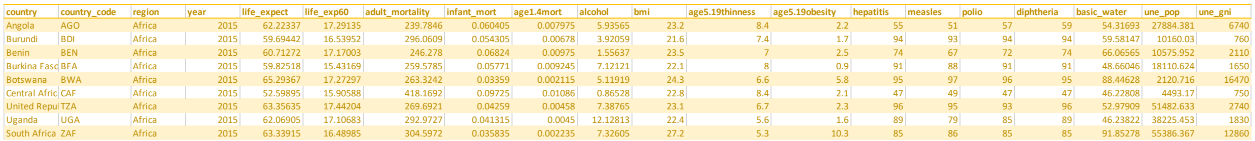
The data attached to perform analysis is after modification, selecting all variables we want to use for analysis, cleansing out the missing values, and filtered numbers at year 2010 and year 2015. Because we want to compare life expectancy of year 2010 and year 2015, we split data into two separate datasets for separate studies. Like mentioned, each dataset contains 20 variables (country name, country code, region, year, life expectancy (in years), life expectancy at age 60 (in years) , adult mortality rate (deaths per 1000 population), infant mortality rate, age 1- 4 mortality rate (in decimal probability), alcohol consumption (in liter), BMI, thinness rate of age 5 – 19 (in percentage), obesity rate of age 5 – 19 (in percentage), immunization rate of hepatitis, measles, polio, and diphtheria (in percentage), percentage of basic water usage, national population, and GNI per capita). and 157 observations (countries). There is no outliers and missing values in these data. Table 1 shows the first 10 observations of year 2010 dataset. Table 2 shows the first 10 observations of year 2015 observations. From the tables shown below, we can already kind of an improvement in most life expectancies and lower mortality rates. In the next section we will perform cluster analysis to break out more information from these data.

Table 1 includes the first 10 observations from year 2010 Life Expectancy Dataset



country	country_code	region	year	life_expect	life_exp60	adult_mortality	infant_mort	age1.4mort	alcohol	bmi	age5.19thinness	age5.19obesity	hepatitis	measles	polio	diphtheria	basic_water	une_pop	une_gni
Angola	AGO	Africa	2010	58.43673	16.68662	269.7017	0.079385	0.012125	6.82834	22.8	9.2	1.4	49	62	56	51	50.37684	23356.246	5630
Burundi	BDI	Africa	2010	57.55196	16.29294	322.722	0.063755	0.00914	4.48499	21.3	7.8	1	96	92	94	96	56.52312	8675.602	710
Benin	BEN	Africa	2010	59.30233	16.94562	256.6892	0.07485	0.01116	1.42133	23.1	7.9	1.8	76	68	77	76	64.69237	9199.259	1770
Burkina Faso	BFA	Africa	2010	57.04797	15.09733	277.1894	0.070045	0.013965	5.51179	21.8	9.1	0.5	91	92	90	91	50.8247	15605.217	1360
Botswana	BWA	Africa	2010	61.23447	16.74382	331.2066	0.04137	0.00329	5.69395	24.1	8.3	4	95	96	96	95	83.04817	1987.105	12410
Central Afric	CAF	Africa	2010	49.59306	15.26655	457.9922	0.109915	0.01387	1.86748	22.4	9.1	1.4	45	53	46	45	48.01232	4386.768	970
United Repul	TZA	Africa	2010	59.35994	16.93083	342.5473	0.050675	0.00659	6.59254	22.8	7.2	1.5	91	92	94	91	43.8106	44346.525	2140
Uganda	UGA	Africa	2010	57.64995	16.78377	364.252	0.055805	0.007685	11.50323	22	6.1	1	80	73	79	80	39.30215	32428.167	1530
South Africa	ZAF	Africa	2010	58.00648	15.95127	416.2592	0.03892	0.004335	7.25206	26.9	8.7	5.7	71	72	72	77	89.62798	51216.964	11480
Zambia	ZMB	Africa	2010	57.54378	16.74589	363.5941	0.057145	0.00799	3.58587	22.2	6.7	1.8	83	96	80	83	55.81144	13605.984	3050

Table 2 includes the first 10 observations from year 2015 Life Expectancy Dataset



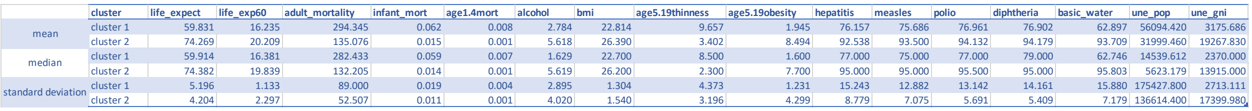
country	country_code	region	year	life_expect	life_exp60	adult_mortality	infant_mort	age1.4mort	alcohol	bmi	age5.19thinness	age5.19obesity	hepatitis	measles	polio	diphtheria	basic_water	une_pop	une_gni
Angola	AGO	Africa	2015	62.22337	17.29135	239.7846	0.060405	0.007975	5.93565	23.2	8.4	2.2	55	51	57	59	54.31693	27884.381	6740
Burundi	BDI	Africa	2015	59.69442	16.53952	296.0609	0.054305	0.00678	3.92059	21.6	7.4	1.7	94	93	94	94	59.58147	10160.03	760
Benin	BEN	Africa	2015	60.71272	17.17003	246.278	0.06824	0.00975	1.55637	23.5	7	2.5	74	67	72	74	66.06565	10575.952	2110
Burkina Faso	BFA	Africa	2015	59.82518	15.43169	259.5785	0.05771	0.009245	7.12121	22.1	8	0.9	91	88	91	91	48.66046	18110.624	1650
Botswana	BWA	Africa	2015	65.29367	17.27297	263.3242	0.03359	0.002115	5.11919	24.3	6.6	5.8	95	97	96	95	88.44628	2120.716	16470
Central Afric	CAF	Africa	2015	52.59895	15.90588	418.1692	0.09725	0.01086	0.86528	22.8	8.4	2.1	47	49	47	47	46.22808	4493.17	750
United Repul	TZA	Africa	2015	63.35635	17.44204	269.6921	0.04259	0.00458	7.38765	23.1	6.7	2.3	96	95	93	96	52.97909	51482.633	2740
Uganda	UGA	Africa	2015	62.06905	17.10683	292.9727	0.041315	0.0045	12.12813	22.4	5.6	1.6	89	79	85	89	46.23822	38225.453	1830
South Africa	ZAF	Africa	2015	63.33915	16.48985	304.5972	0.035835	0.002235	7.32605	27.2	5.3	10.3	85	86	85	85	91.85278	55386.367	12860

Section 2. Cluster Analysis

I. Summary tables:

In this section, we perform cluster analysis using k-means algorithm on year 2010 and year 2015 datasets. Beginning with Year 2010, we obtain the summary (mean, median and standard deviation of cluster 1 and cluster 2) shown in table 3. Some interesting inferences can already be drawn here. Cluster 1 tend to have lower life expectancy, higher mortality rate, lower alcohol consumption, higher rate of thinness, lower rate of obesity, lower rate of immunization rates across 4 diseases, and lower rate of basic water usage than cluster 2. We see pretty significant difference in life expectancy where cluster 1 average at 59.8 and cluster 2 average at 74.2 as shown in table 3. We can probably expect that cluster 1 contains more developing countries, while cluster 2 have more developed countries in their list. Interestingly, we see big volatility in national population and capital per capita for both clusters. They are probably not the best variables to define country's health status.

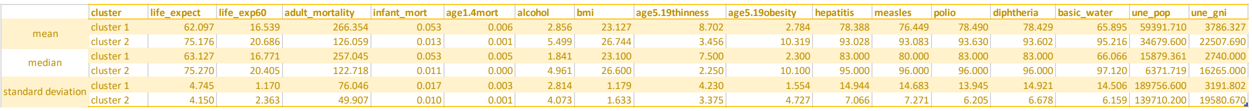
Table 3 is a summary of year 2010 Life Expectancy Dataset



	cluster	life_expect	life_exp60	adult_mortality	infant_mort	age1.4mort	alcohol	bmi	age5.19thinness	age5.19obesity	hepatitis	measles	polio	diphtheria	basic_water	une_pop	une_gni
mean	cluster 1	59.831	16.235	294.345	0.062	0.008	2.784	22.814	9.657	1.945	76.157	75.686	76.961	76.902	62.897	56094.420	3175.686
	cluster 2	74.269	20.209	135.076	0.015	0.001	5.618	26.390	3.402	8.494	92.538	93.500	94.132	94.179	93.709	31999.460	19267.830
median	cluster 1	59.914	16.381	282.433	0.059	0.007	1.629	22.700	8.500	1.600	77.000	75.000	77.000	79.000	62.746	14539.612	2370.000
	cluster 2	74.382	19.839	132.205	0.014	0.001	5.619	26.200	2.300	7.700	95.000	95.000	95.500	95.000	95.803	5623.179	13915.000
standard deviation	cluster 1	5.196	1.133	89.000	0.019	0.004	2.895	1.304	4.373	1.231	15.243	12.882	13.142	14.161	15.880	175427.800	2713.111
	cluster 2	4.204	2.297	52.507	0.011	0.001	4.020	1.540	3.196	4.299	8.779	7.075	5.691	5.409	7.179	136614.400	17399.980

Next, we examine our year 2015 dataset. Table 4 shown a summary (mean, median and standard deviation of cluster 1 and cluster 2). Like year 2010 summary, cluster 1 tend to have lower life expectancy, higher mortality rate, lower alcohol consumption, higher rate of thinness, lower rate of obesity, lower rate of immunization rates across 4 diseases, and lower rate of basic water usage than cluster 2. It is fair to conclude cluster 1 probably have more developing countries and cluster 2 have more developed countries, where life expectancy for cluster 1 averages at 62.1 and cluster 2 averages 75.2.

Table 4 is a summary of year 2015 Life Expectancy Dataset



	cluster	life_expect	life_exp60	adult_mortality	infant_mort	age1.4mort	alcohol	bmi	age5.19thinness	age5.19obesity	hepatitis	measles	polio	diphtheria	basic_water	une_pop	une_gni
mean	cluster 1	62.097	16.539	266.354	0.053	0.006	2.856	23.127	8.702	2.784	78.388	76.449	78.490	78.429	65.895	59391.710	3786.327
	cluster 2	75.176	20.686	126.059	0.013	0.001	5.499	26.744	3.456	10.319	93.028	93.083	93.630	93.602	95.216	34679.600	22507.690
median	cluster 1	63.127	16.771	257.045	0.053	0.005	1.841	23.100	7.500	2.300	83.000	80.000	83.000	83.000	66.066	15879.361	2740.000
	cluster 2	75.270	20.405	122.718	0.011	0.000	4.961	26.600	2.250	10.100	95.000	96.000	96.000	96.000	97.120	6371.719	16265.000
standard deviation	cluster 1	4.745	1.170	76.046	0.017	0.003	2.814	1.179	4.230	1.554	14.944	14.683	13.945	14.921	14.506	189756.600	3191.802
	cluster 2	4.150	2.363	49.907	0.010	0.001	4.073	1.633	3.375	4.727	7.066	7.271	6.205	6.678	6.159	139710.200	19580.670

Notice that over five years, life expectancies have increased, mortalities have decreased, alcohol consumption and BMI did not vary so much, thinness for cluster 1 have decreased quite a lot, obesity in cluster 2 have increased quite a bit, immunizations have improved some, water are more accessible, national population and GNI per capita have both increased.

II. Percent composition of regions by clusters

Now, we are interested in the percent composition of countries in each cluster by regions, which are Africa, Americas, Eastern Mediterranean, Europe, South-East Asia, and Western Pacific. In year 2010 as shown in table 5, all countries in the region of Americas and Europe are clustered as cluster 2, while 84% of African countries are clustered as cluster 1. 60% of South-East Asian countries clustered as cluster 1, and 79% of Western Pacific countries clustered as cluster 2.

Table 5 shows year 2010 Life Expectancy Dataset's percent composition of regions by clusters

Column1	Africa	Americas	Eastern Mediterranean	Europe	South-East Asia	Western Pacific
Cluster 1	86%	0%	18%	0%	60%	21%
Cluster 2	14%	100%	82%	100%	40%	79%

In year 2015 as shown in table 6, we see more African countries and South-East Asian countries are clustered in cluster 2. The rest of regions did not differ from data in 2010. The composition tables align with my hypothesis that African countries tend to be in cluster 1 where life expectancies are lower, and more developed countries such as regions of Europe and Americas are seeing higher life expectancies.

Column1	Africa	Americas	Eastern Mediterranean	Europe	South-East Asia	Western Pacific
Cluster 1	84%	0%	18%	0%	50%	21%
Cluster 2	16%	100%	82%	100%	50%	79%

III. Scatterplots between variables

Since there are 16 interval variables, we are just going to pick two sets of variables to compare. Ones that I am interested in comparing are life expectancy vs. GNI per capita, and life expectancy vs. alcohol consumption.

1. Life Expectancy at birth vs. GNI per capita

As shown in figure 1 and figure 2, countries having low life expectancy and low GNI per capita are most generally clustered in cluster 1. However, that does not infer cluster 2 has high GNI per capita and life expectancy. In fact, many countries with low GNI per capita are clustered in cluster 2. It seems like clustering is more dependent on life expectancy. Not much significant difference are shown in data from 2010 and 2015.

Figure 2 is a scatterplot of year 2010 life expectancy vs. GNI per capita

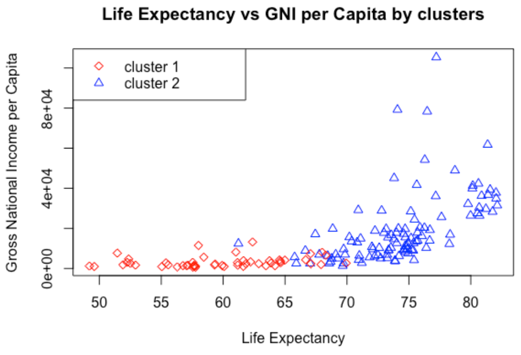
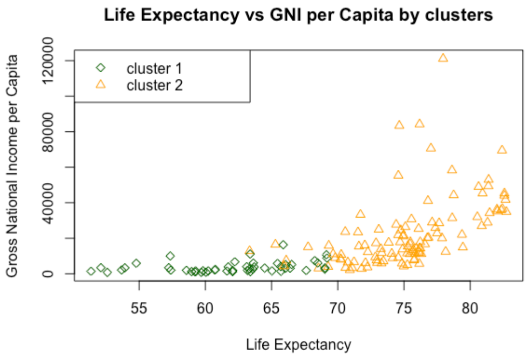


Figure 1 is a scatterplot of year 2015 life expectancy vs GNI per capita



2. Life Expectancy at birth vs. Alcohol Consumption

As shown in figure 3 and 4, alcohol consumptions do not seem to play a role in clustering. And interestingly, alcohol consumption also does not really have much of a correlation with life expectancy in both 2010 and 2015. At first, I thought maybe higher alcohol consumption can lead to lower life expectancy intuitively. However, looking at the graphs, high alcohol consumption can in fact be having high life expectancy in many countries especially in year 2010. Clustering is also not distinct in alcohol consumption. It once again shows that clustering is more dependent on life expectancy.

Figure 4 is a scatterplot of 2010 life expectancy vs. alcohol consumption

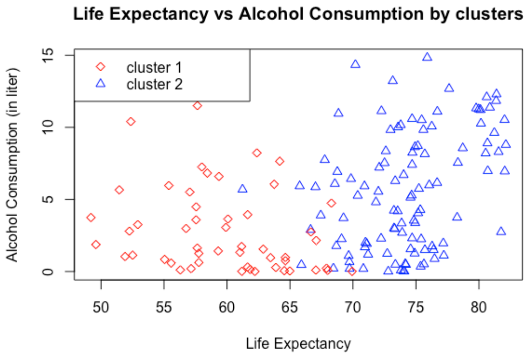
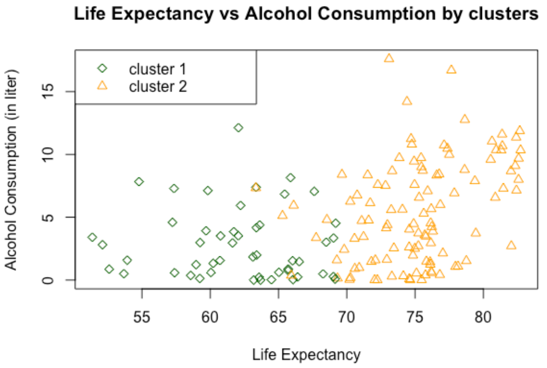


Figure 3 is a scatterplot of 2015 life expectancy vs. alcohol consumption



Section 3. PC Analysis

In this section, we perform principal component analysis on life expectancy datasets in 2010 and 2015. We will first determine the principal components, find the proportion of variance in each principal components, and retain the ones that will meet our level of variance. Then, we will interpret each principal component by our variables.

We first by finding the eigenvalues of variance covariance matrix, and compute for the cumulative proportion. Table 6 shows the cumulative proportion of variance in each PC by percentage in year 2010. Noticing that PC1 already accounts for 56% of the variability of our dataset. To achieve sufficient variance level, I think retaining at PC4 is sufficient, meeting at a variance level of 83% benchmark. Table 7 shows the cumulative proportion of variance in each PC by percentage in year 2015. Again, PC1 accounts for majority of the variability of our 2015 dataset, and PC4 is quite sufficient to meet our variance level of 84% benchmark. So for both years, we will interpret PC1-PC6 with our variables' correlations.

Table 6 is the cumulative proportion of variance in each PC from 2010

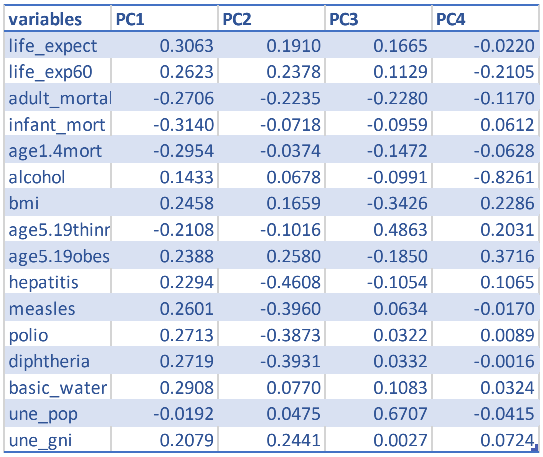
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
cumulative proportion of variance each PC (%)	56.12373	67.98562	76.61937	83.14433	87.76994	91.6995	94.20522	95.61382
	PC9	**PC10**	**PC11**	**PC12**	**PC13**	**PC14**	**PC15**	**PC16**
	96.69733	97.68633	98.51437	99.05522	99.48291	99.82739	99.98922	100

Table 7 is the cumulative proportion of variance in each PC from 2015

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
cumulative proportion of variance each PC (%)	55.68004	68.49292	76.70693	83.52433	88.20805	92.12606	94.53801	96.00022
	PC9	**PC10**	**PC11**	**PC12**	**PC13**	**PC14**	**PC15**	**PC16**
	97.17624	98.14363	98.778	99.29735	99.63391	99.90373	99.98841	100

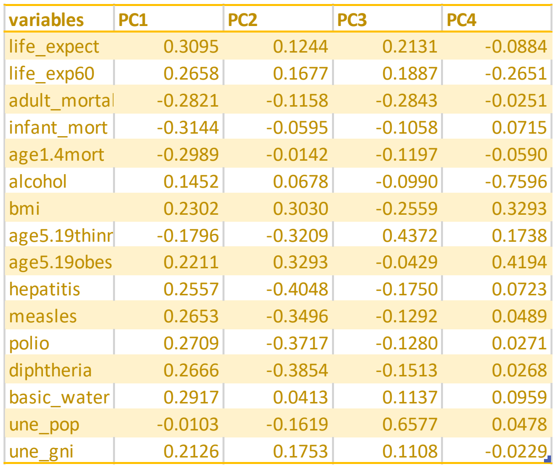
Now we interpret the Principal components in terms of the original interval variables. Table 8 shows the analysis of our PC and variables in year 2010. Figure 5 shows the direction and loading of the 16 original variables. Starting with PC1, we notice moderate positive loadings in life expectancy at birth and at 60, BMI, obesity rate, basic water usage and GNI. Moderate negative loadings in all the mortality rates. Noticing national population and alcohol are bit irrelevant in PC1, and since it accounts for over half of dataset variability, this probably is a good health measure of country. PC2 has strong negative loadings in all the rates of immunization against the four illnesses chosen. The latent variable here could be the accessibility of immunization. Looking at figure 5, we have morality rates grouped negatively, and life expectancies and some other variables related group positively, and rate of immunization one their own in another group. Interesting finding is maybe in PC4 related with positive correlation of high alcohol consumption and obesity rate. Loading in variables like national population and alcohol consumption are low, and those are expected because we have concluded in previous analysis that life expectancy do not seem to relate much to them. Year 2015 is not much different from 2010 as you can see in table 9 and figure 6. Same analysis goes for 2015.

Table 9 shows an analysis on PC v. variables in 2010



variables	PC1	PC2	PC3	PC4
life_expect	0.3063	0.1910	0.1665	-0.0220
life_exp60	0.2623	0.2378	0.1129	-0.2105
adult_mortal	-0.2706	-0.2235	-0.2280	-0.1170
infant_mort	-0.3140	-0.0718	-0.0959	0.0612
age1.4mort	-0.2954	-0.0374	-0.1472	-0.0628
alcohol	0.1433	0.0678	-0.0991	-0.8261
bmi	0.2458	0.1659	-0.3426	0.2286
age5.19thinr	-0.2108	-0.1016	0.4863	0.2031
age5.19obes	0.2388	0.2580	-0.1850	0.3716
hepatitis	0.2294	-0.4608	-0.1054	0.1065
measles	0.2601	-0.3960	0.0634	-0.0170
polio	0.2713	-0.3873	0.0322	0.0089
diphtheria	0.2719	-0.3931	0.0332	-0.0016
basic_water	0.2908	0.0770	0.1083	0.0324
une_pop	-0.0192	0.0475	0.6707	-0.0415
une_gni	0.2079	0.2441	0.0027	0.0724

Table 8 shows an analysis on PC v. variables in 2015



variables	PC1	PC2	PC3	PC4
life_expect	0.3095	0.1244	0.2131	-0.0884
life_exp60	0.2658	0.1677	0.1887	-0.2651
adult_mortal	-0.2821	-0.1158	-0.2843	-0.0251
infant_mort	-0.3144	-0.0595	-0.1058	0.0715
age1.4mort	-0.2989	-0.0142	-0.1197	-0.0590
alcohol	0.1452	0.0678	-0.0990	-0.7596
bmi	0.2302	0.3030	-0.2559	0.3293
age5.19thinr	-0.1796	-0.3209	0.4372	0.1738
age5.19obes	0.2211	0.3293	-0.0429	0.4194
hepatitis	0.2557	-0.4048	-0.1750	0.0723
measles	0.2653	-0.3496	-0.1292	0.0489
polio	0.2709	-0.3717	-0.1280	0.0271
diphtheria	0.2666	-0.3854	-0.1513	0.0268
basic_water	0.2917	0.0413	0.1137	0.0959
une_pop	-0.0103	-0.1619	0.6577	0.0478
une_gni	0.2126	0.1753	0.1108	-0.0229

Figure 6 is a loading plot from year 2010

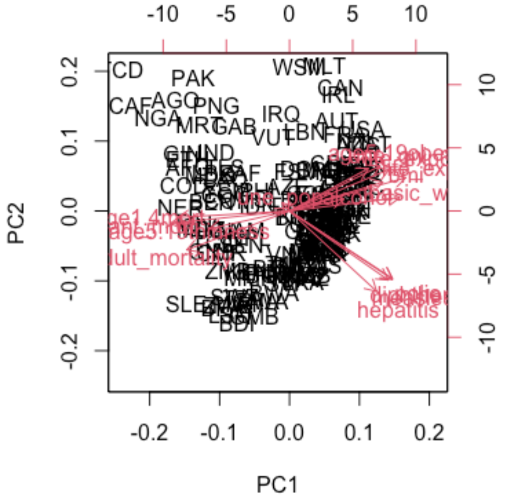
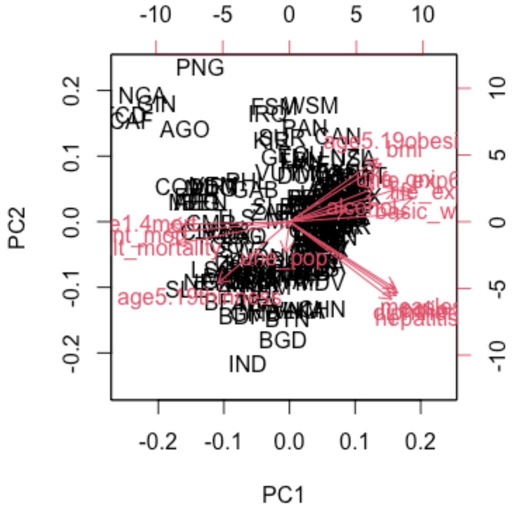


Figure 5 is a loading plot from year 2015



Section 4. MLE optimization

Variable I choose to perform MLE optimization is life expectancy at birth since our data are surrounding national life expectancy. The model I think will fit well is normal distribution because with over 157 observations, by central limit theorem, the numbers should be approximately normal. We use the Newton's algorithm provided in class to optimize our parameters mean and variance of the gaussian model. In year 2010, the optimized mean is calculated to be 69.57901years, and optimized variance is 66.13687. For confidence interval, we are 95% confident that mean lies between 68.51143 years and 70.64658 years, and variance lies between 53.85863 and 78.41511. Notice that this gaussian model has a pretty large standard deviation of about 7.4 years, so we do see a large range of spread in values. In year 2015, the optimized mean is 71.0943 years, and variance of 55.34446. For confidence interval, we are 95% confident that mean lies between 70.11770 years and 72.07089 years, and variance lies between 45.06982 and 65.61909. To compare these two years result, we see an increase in life expectancy by about a year.

Section 5. Conclusion

After different analysis, we have found some interesting findings from the life expectancy data. In cluster analysis, generally we see a higher life expectancies, lower mortalities, higher BMI, higher immunizations, higher basic water usage, and higher GNI per capita from countries in one cluster. Based on these factors, we can assume that the health of countries in this cluster is better, and as we have seen in compositions of regions in this cluster, most countries are Europeans and Americans. Meanwhile, the other cluster have lower life expectancies, higher morality rate, lower immunizations, and in proportion is contains more African countries. It goes to conclude regions do play a role in clustering countries' health. Year 2010 and year 2015 do not seem to have a very significant difference with their analysis result, but the values such as life expectancies do improved and lower mortality rates. That can also be seen in our MLE optimization when estimating for the mean and variance of our national life expectancy at birth in separate years. Through PCA, we notice combination of all the interval variable except national population and alcohol contribute to most of the variability of our data. That makes sense because national population does not seem to have much correlation to a health of a country, and alcohol, though bad for health, is probably a "rich" entertainment that isn't accessible in poorer countries. So it makes sense that those two factors did not have much loadings to our dataset. One of the significant latent variables is with immunizations of four illnesses in PC2, and that marks how accessible immunization is at the country. This dataset encompasses of interesting correlations and findings to estimating the health index of the countries around the world.

Reference:

[1] WHO National Life Expectancy: https://www.kaggle.com/mmattson/who-national-life-expectancy