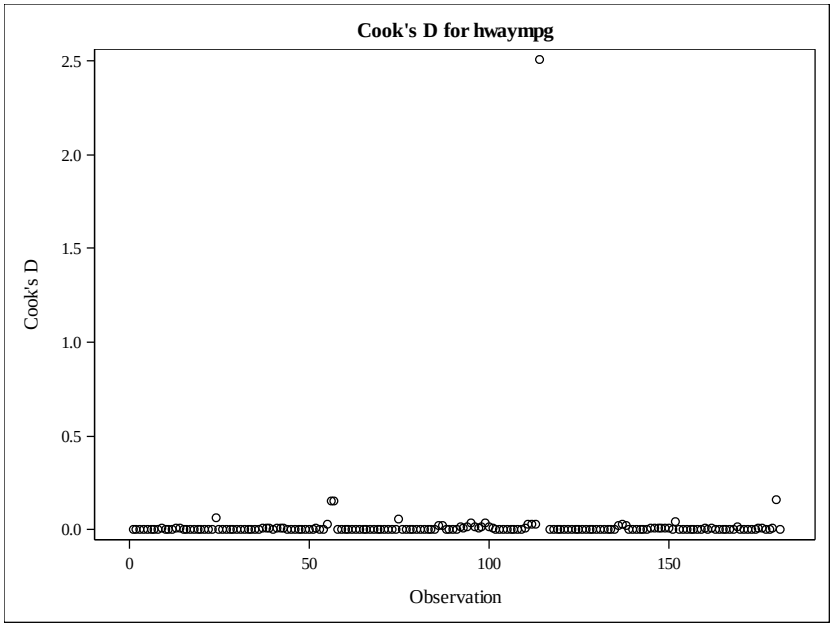
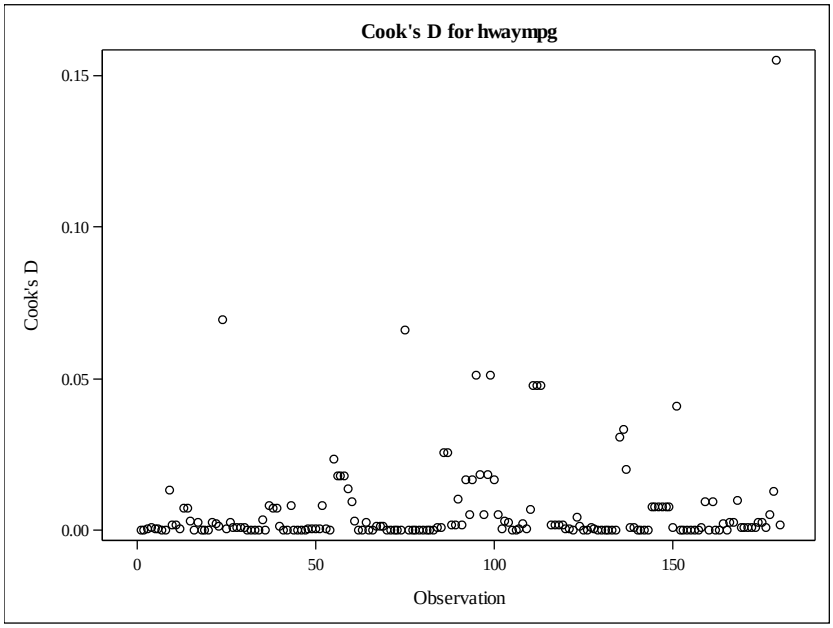
Exercise 1

(a) From an inspection of the Cook's distances for all of the data, we see there is one extremely influential point with a Cook's distance of about 2.5. This point should be removed.



After removing that point, the Cook's distances look much better. There is still one point that is at least twice as large as the next closest Cook's distance, so we will remove that point, too. After removing that point, the Cook's distances (shown in the results for part b) look good, so we will proceed with those two points removed.



(b) Results for the model after removing the points in part a follow.

Model Information		
Data Set	WORK.PDIAGNOSTICS2	Predicted Values and Diagnostic Statistics
Distribution	Poisson	
Link Function	Log	
Dependent Variable	hwaympg	

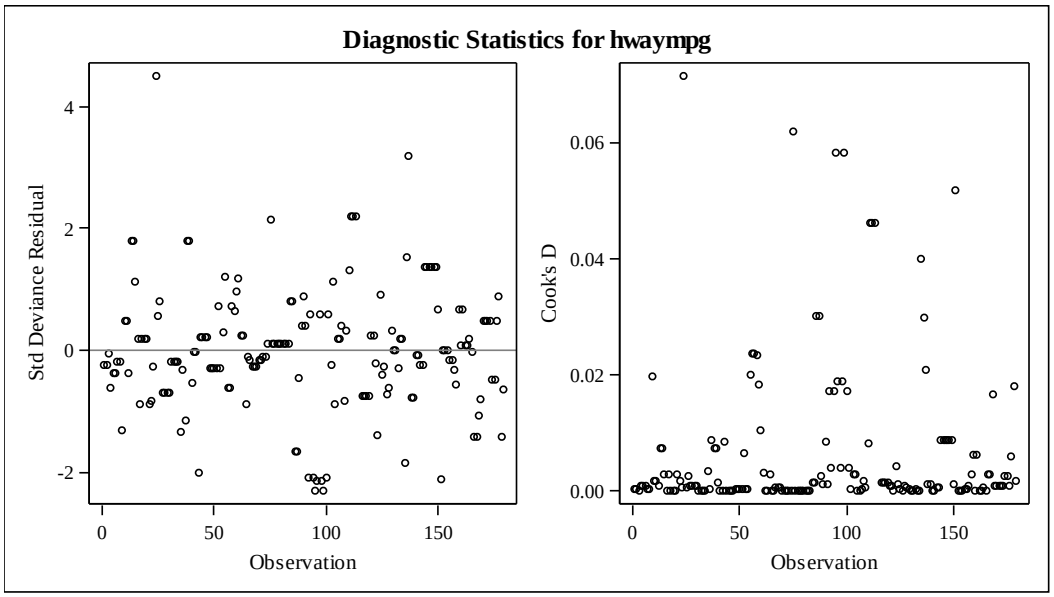
Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	170	47.4158	0.2789
Scaled Deviance	170	170.0000	1.0000
Pearson Chi-Square	170	48.1458	0.2832
Scaled Pearson X2	170	172.6171	1.0154
Log Likelihood		49166.6937	
Full Log Likelihood		-489.8755	
AIC (smaller is better)		993.7509	
AICC (smaller is better)		994.4136	
BIC (smaller is better)		1015.9840	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	4.1358	0.0825	3.9740	4.2976	2510.38	<.0001
fuel	diesel	1	0.1182	0.0272	0.0648	0.1715	18.83	<.0001
fuel	gas	0	0.0000	0.0000	0.0000	0.0000	.	.
drive	fwd	1	0.0620	0.0205	0.0218	0.1021	9.15	0.0025
drive	rwd	0	0.0000	0.0000	0.0000	0.0000	.	.
hp		1	-0.0034	0.0005	-0.0044	-0.0025	54.06	<.0001
enginesize		1	-0.0023	0.0006	-0.0034	-0.0011	15.29	<.0001
cylinders	eight	1	-0.1191	0.0808	-0.2776	0.0394	2.17	0.1407
cylinders	four	1	-0.1398	0.0362	-0.2108	-0.0689	14.91	0.0001
cylinders	six	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	0.5281	0.0000	0.5281	0.5281		

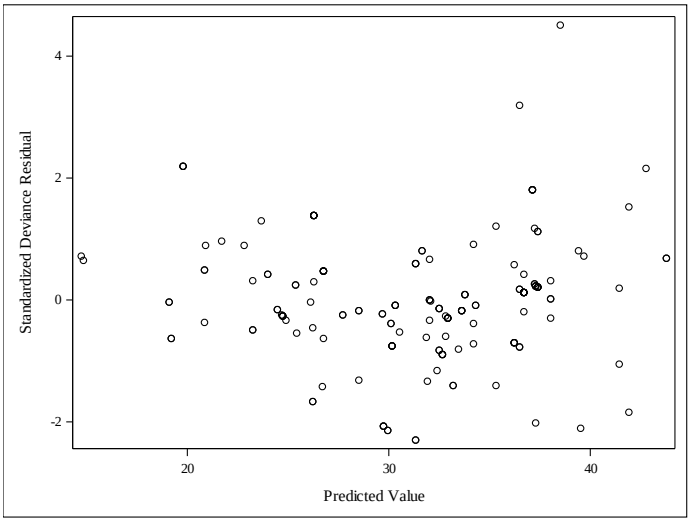
Note: The scale parameter was estimated by the square root of DEVIANCE/DOF.

LR Statistics For Type 1 Analysis							
Source	**Deviance**	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
Intercept	254.4251						
fuel	233.9554	1	170	73.39	<.0001	73.39	<.0001
drive	132.1692	1	170	364.93	<.0001	364.93	<.0001
hp	57.6923	1	170	267.02	<.0001	267.02	<.0001
enginesize	52.4623	1	170	18.75	<.0001	18.75	<.0001
cylinders	47.4158	2	170	9.05	0.0002	18.09	0.0001

LR Statistics For Type 3 Analysis						
Source	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
fuel	1	170	18.50	<.0001	18.50	<.0001
drive	1	170	9.18	0.0028	9.18	0.0024
hp	1	170	55.38	<.0001	55.38	<.0001
enginesize	1	170	15.38	0.0001	15.38	<.0001
cylinders	2	170	9.05	0.0002	18.09	0.0001



There are no major concerns with the residuals. Looking at the standardized deviance residuals, we see one point with value over 4. While it is a bit larger than the rest, it is not of great concern. Plotting the standardized deviance residuals against predicted value, we also see no major trends to concern us. The residuals are fairly flat and balanced around 0 aside from the point with a value a bit over 4.



(c) This model and the model from Homework 4 are pretty similar. The terms are all still significant in the type 1 and type 3 analyses. Looking at the F statistics p-values, we can see that degree of significance has changed slightly for the terms, but the terms are all still highly significant.

The AIC values are fairly comparable as well. We cannot compare the goodness of fit numbers directly because the models are based on slightly different data sets with slightly different scale estimates, but they should be better or at least not wildly different for this model and that is the case.

When looking at the parameter estimates, we see that the parameters that were significant before are again significant and the directions have remained the same (positive relationships have remained positive, and negative relationships have remained negative). The magnitudes and level of significance, have changed a bit. This model expects a smaller difference between diesel and gas powered cars, a slightly smaller difference between front wheel and rear wheel drive cars, a slightly larger decrease in fuel efficiency as horsepower increases, a slightly smaller decrease as engine size increases, and a slightly smaller difference between four and six cylinder vehicles.

Specifically, we have the following findings:

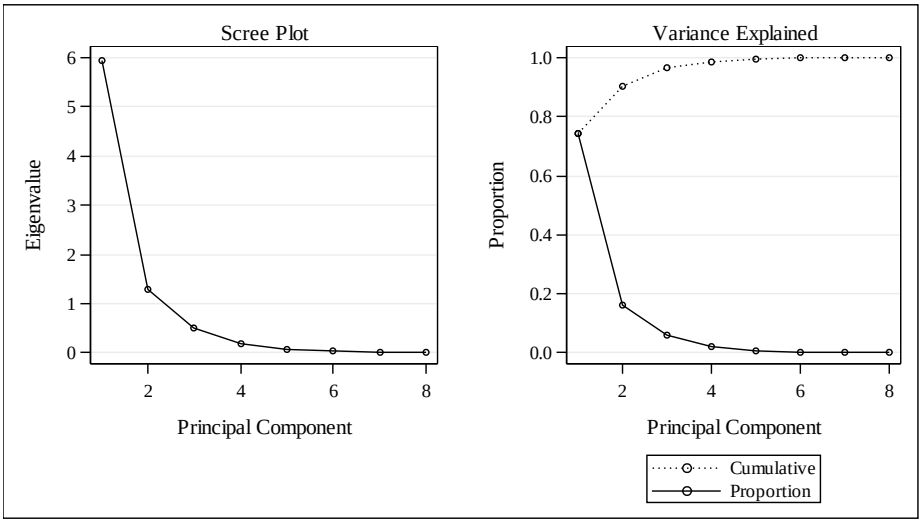
- Diesel vehicles are now expected to have 1.1255 times the highway fuel efficiency of gas cars, as opposed to 1.15 times in the previous model.
- We expect front wheel drive cars to be 6.4% more fuel efficient than rear wheel drive vehicles on the highway, as compared to 7% more efficient in the previous model.
- A unit increase in horsepower now corresponds to an expected 0.34% drop in fuel efficiency on the highway as compared to a 0.22% decrease in the previous model.
- A one cc increase in engine size corresponds to an expected 0.23% decrease in highway fuel efficiency as compared to a 0.3% decrease in the previous model.
- Four cylinder vehicles are expected to be 13% less fuel efficient than six cylinders on the highway as compared to 11.8% less fuel efficient in the previous model.

Exercise 2

(a) Results for the correlation-based principal compoent analysis follow. We can see that nearly 75% of the variation in life expectancy can be retained with just the first component. To keep 95% of the variation, we will want to keep the first 3 components, which together describe 96.42% of the variation in the life expectancy data. Based on the average eigenvalue criterion, we would keep 2 components,since component 2 has an eigenvalue above the average value of 1 and component 3's eigenvalue is less than 1. Based on the scree plot, 1 component looks like a good choice, though a reasonnable argument could also be made for 2.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	5.94144668	4.65905350	0.7427	0.7427
2	1.28239318	0.79301828	0.1603	0.9030
3	0.48937490	0.30423168	0.0612	0.9642
4	0.18514322	0.12603780	0.0231	0.9873
5	0.05910542	0.03410292	0.0074	0.9947
6	0.02500250	0.01274064	0.0031	0.9978
7	0.01226186	0.00698962	0.0015	0.9993
8	0.00527224		0.0007	1.0000

Eigenvectors								
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**
m0	0.347818	-.449633	-.056320	0.264173	0.239790	0.070815	0.398146	-.618756
m25	0.390784	-.105387	-.293761	-.338034	-.018157	-.764498	0.141734	0.175338
m50	0.361456	0.160789	-.550984	-.405466	0.268432	0.533468	-.129792	0.044947
m75	0.281595	0.548747	-.321171	0.702558	-.092664	-.114589	-.020881	0.026256
f0	0.340720	-.459231	0.174862	0.325149	0.186222	0.163229	-.085124	0.684215
f25	0.399307	-.120272	0.210117	-.026579	-.273319	-.060065	-.771617	-.327916
f50	0.390885	0.130771	0.261837	-.173179	-.674043	0.274370	0.441935	0.081482
f75	0.296711	0.466141	0.600429	-.146752	0.545722	-.085312	0.078509	-.037009



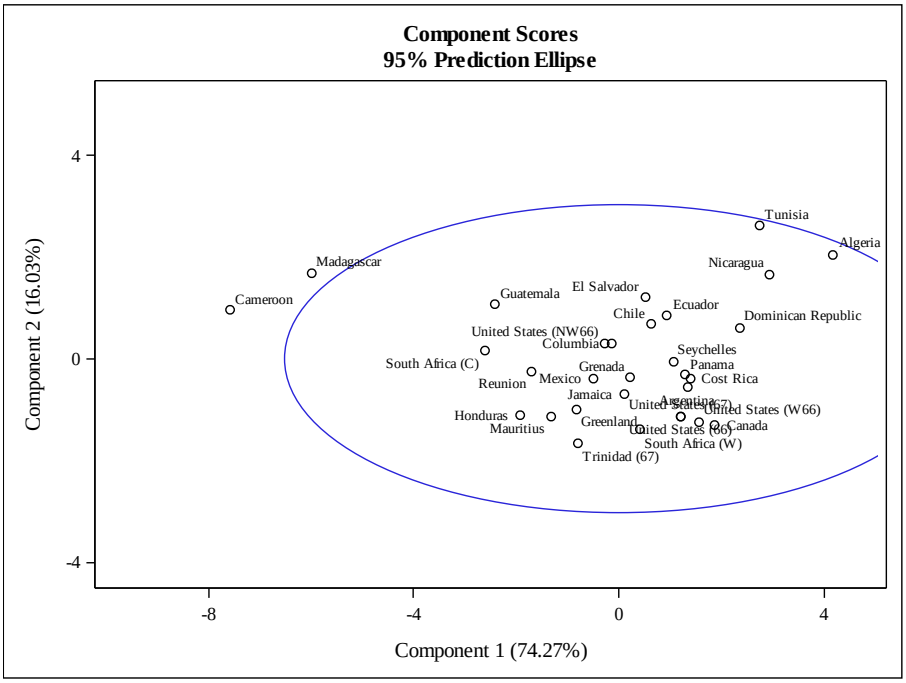
(b) To interpret the three retained components, we look at the eigenvectors for the first three principal components.

For component 1, we see that all coefficients are positive and fairly similar in magnitude. This appears to be a general longevity feature, as it increases as any of the life expectancies increase and the amount of increase is fairly similar across age groups and genders.

For component 2, we see that there is little difference across genders—the values for male expectancies are pretty similar to female expectancies at each time point. There is, however, a noticeable difference in magnitudes and directions. The age 0 coefficients are very negative, the age 75 coefficients ae very positive, and the coefficients increase in between. This appears to be a contrast of life expectancies for younger individuals and life expectancies for older individuals.

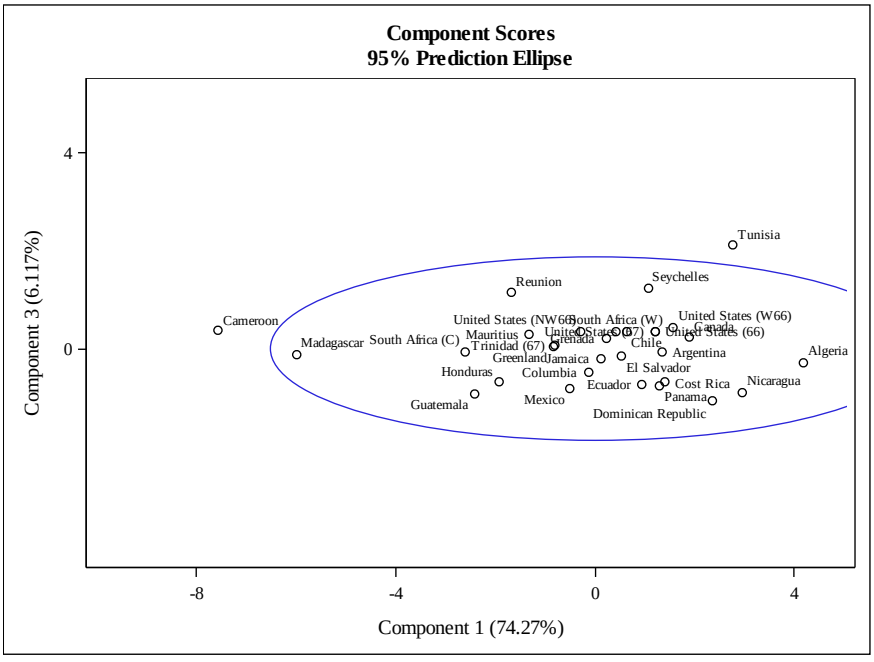
In component 3, we see a gender contrast—female coefficients are all positive and male coefficients are all negative. In the female expectancies, the age 75 expectancy plays the largest role and in males the age 50 expectancy plays the largest role. Male infant expectancy has the smallest contribution and the other terms have moderate contributions. Countries with bigger gaps between life expectancies for males and females should have larger magnitudes for this component. When females are expected to live a lot longer relative to males, the component will have a large positive value. Large negative will be found when the life expectancy gap between males and females is smaller than on average (from looking at the data, we can see that females live longer in general, so negative values will generally correspond to smaller differences rather than males living longer).

(c) Score plots for the 3 components follow. We could also include a heat map version showing all 3 components at once is also included, but we can see everything we need from the pairwise plots.

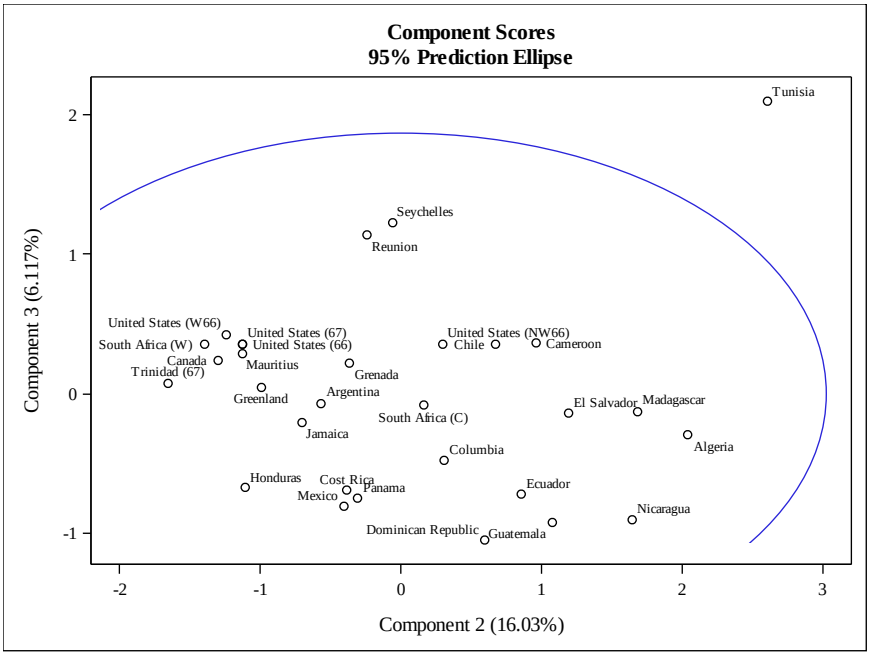


The most extreme negative values for component 1 are for Cameroon and Madagascar, indicating much lower than average life expectancy in those countries. Algeria has the highest value, indicating longer than expected life expectancies in that country. Tunisia and Nicaragua also have fairly high life expectancies.

We see that Tunisia has the highest value of component 2, indicating bigger gaps between life expectancy for the old and the young with the old living longer on average and the young living shorter on average. This may be an indication or higher infant mortality rates or other conditions (such as wars) in which the young are more likely to die. The others are not as extreme in general, but we will note that countries with more negative values may tend to be safer countries for infants and the young.



For component 3, we again see Tunisia having largest positive value. This indicates the largest gap between female and male life expectancy, so at the time the data was gathered females were expected to outlive males by more years on average in Tunisia than in other areas surveyed.



Exercise 3

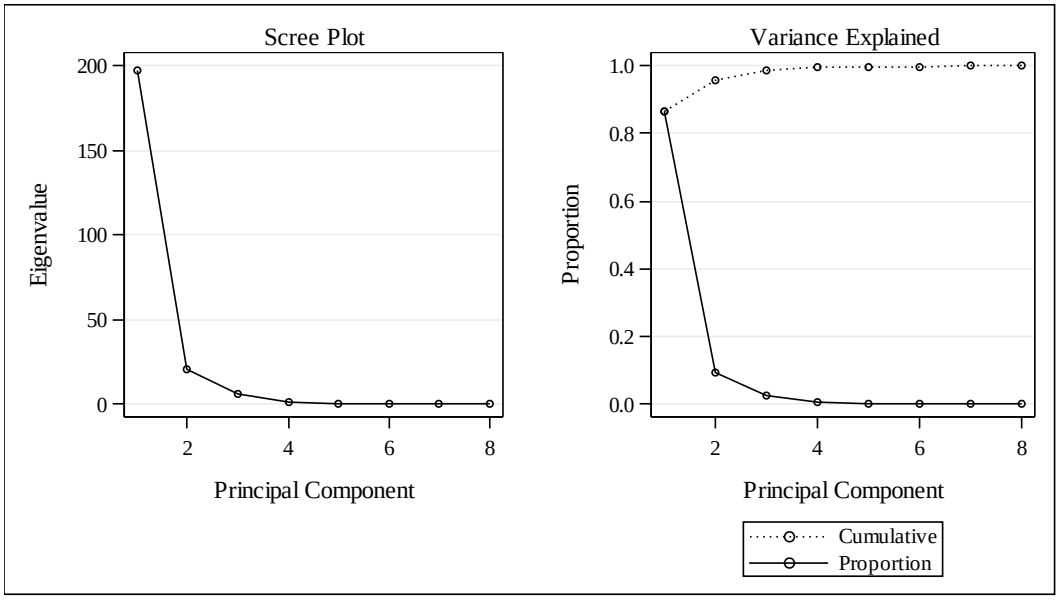
(a) Now we complete the analysis using the covariance matrix instead. In this analysis, variables with larger variances will represent a larger percentage of the overall variation in the data and thus have reater importance in the resulting components. To retain 95% of the variation, we now need only 2 components.

For the average eigenvalue, the avergae will no longer be 1. We need to calculate it by dividing the total variation by the number of components. This gives an average value of a little over 28, thus we would only choose 1 component based on the average eigenvalue. One component also looks like a good choice based on the scree plot.

Total Variance	227.60344828

Eigenvalues of the Covariance Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	197.032700	176.005433	0.8657	0.8657
2	21.027267	14.603637	0.0924	0.9581
3	6.423630	4.910791	0.0282	0.9863
4	1.512839	0.897419	0.0066	0.9929
5	0.615420	0.203749	0.0027	0.9956
6	0.411671	0.041135	0.0018	0.9975
7	0.370536	0.161151	0.0016	0.9991
8	0.209386		0.0009	1.0000

Eigenvectors								
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**
m0	0.559892	-.279593	-.286797	0.217325	0.055380	0.376995	0.552213	-.169514
m25	0.313497	0.335787	-.433098	-.417351	0.047088	0.428576	-.436897	0.224582
m50	0.178754	0.430092	-.469488	0.118258	0.277424	-.666179	0.151562	-.070245
m75	0.061080	0.334985	-.026077	0.796713	-.364288	0.171683	-.293134	0.022657
f0	0.621426	-.394043	0.278060	0.089408	0.098456	-.348542	-.407758	0.275329
f25	0.344342	0.263122	0.302291	-.274945	-.346217	-.085830	-.077049	-.715578
f50	0.199011	0.364858	0.299587	-.168841	-.397001	-.071087	0.468774	0.571491
f75	0.090607	0.392754	0.500624	0.133246	0.705603	0.258893	0.057728	-.030352



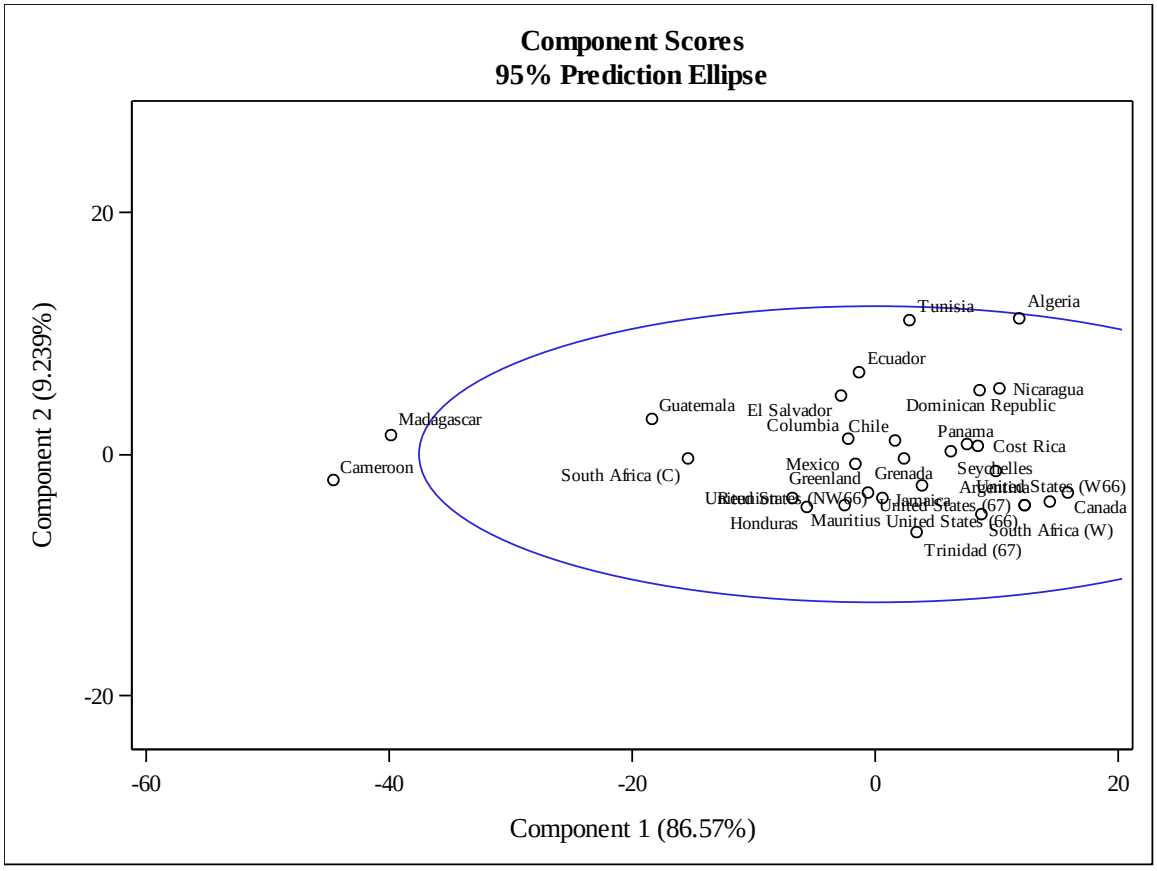
(b) To interpret the two retained components, we again look at their eigenvectors.

For component 1, we see that all coefficients are positive and the male and female values are fairly similar, but the magnitude decreases with age. This again appears to be a general longevity feature, but now there are large coefficients for the variable with greater variability in the data— the variables for younger ages.

For component 2, again we see fairly similar values across genders with a slightly larger difference for age 0. The age 0 coefficients are negative and the other coefficient are positive with relatively similar magnitudes. This component seems to be contrasting infant or child mortality rates with mortality rates at other ages.

(c) We can look at a single score plot to investigate extreme observations. As before, we see that Cameroon and Madagascar have the most negative principal component 1 scores, indicating much lower than average general life expectancy.

Looking at component 2, Algeria and Tunisia are the most extreme with the largest positive values. This indicates a bigger gap in life expectancy for adults and infants. The positive values likely indicated higher infant or child mortality rates and/or higher than typical life expectancy for those who reached adulthood in these two countries.



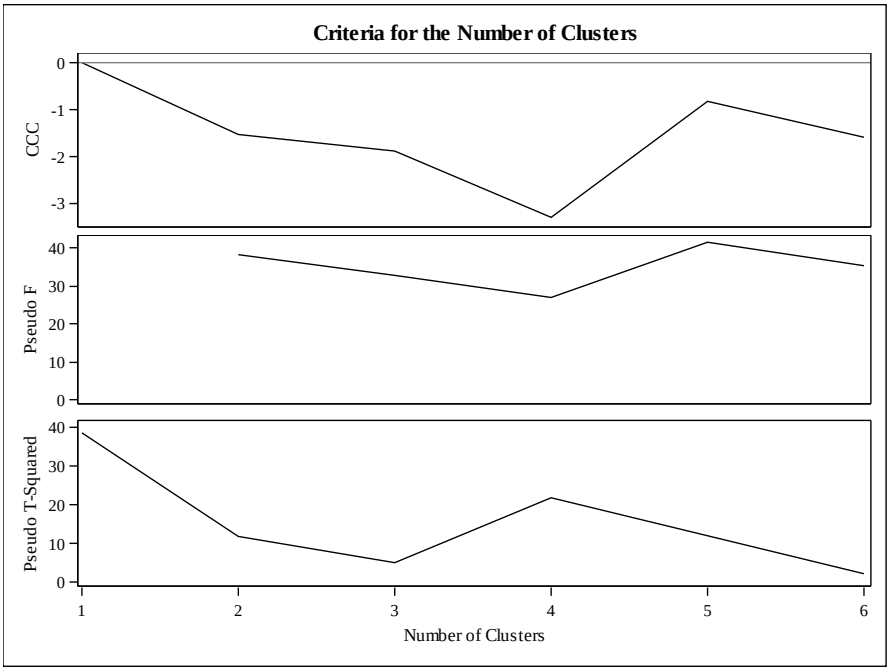
Exercise 4

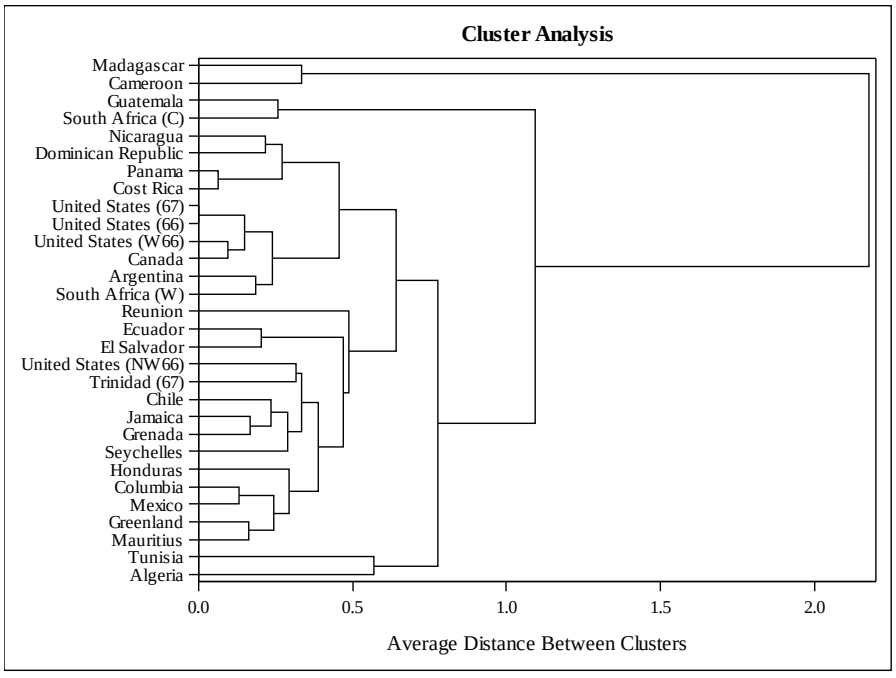
The cluster history is not necessary here, but including it for the last few clusters allows us to read the test statistic values from the table. The table also wraps to additional lines because it has some many columns.

Cluster History								
Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-Square**	**R-Square**	**Approximate Expected R-Square**	**Cubic Clustering Criterion**	**Pseudo F Statistic**
10	Cameroon	Madagascar	2	0.0039	.955	.	.	46.9
9	CL13	CL11	11	0.0167	.938	.	.	39.8
8	CL18	CL15	10	0.0281	.910	.	.	31.8
7	CL9	CL21	13	0.0185	.892	.	.	31.5
6	CL7	Reunion	14	0.0110	.881	.908	-1.6	35.4
5	Algeria	Tunisia	2	0.0111	.869	.886	-.84	41.6
4	CL6	CL8	24	0.1115	.758	.853	-3.3	27.1
3	CL5	CL4	26	0.0491	.709	.800	-1.9	32.8
2	CL3	CL16	28	0.1302	.579	.669	-1.5	38.4
1	CL2	CL10	30	0.5786	.000	.000	0.00	.

Cluster History					
Number of Clusters	**Clusters Joined**		**Pseudo t-Squared**	**Norm RMS Distance**	**Tie**
10	Cameroon	Madagascar	.	0.3347	
9	CL13	CL11	6.1	0.3906	
8	CL18	CL15	17.3	0.4564	
7	CL9	CL21	4.7	0.4724	
6	CL7	Reunion	2.2	0.4904	
5	Algeria	Tunisia	.	0.5683	
4	CL6	CL8	21.6	0.6416	
3	CL5	CL4	5.0	0.7761	
2	CL3	CL16	11.8	1.0917	
1	CL2	CL10	38.4	2.1768	

(a) From the dendrogram, we definitely want at least 3 and 4 or 5 look like reasonable choices. CCC and Pseudo F both show local peaks at 5, indicating 5 as a good choice based on those two measures. The Pseudo T-Squared plot is less conclusive. Might consider 2, 5 or 6 based on that plot. Taking each of these measures into account, we would choose 5 clusters for our data.





(b) We can do a basic means analysis by cluster to compare the values of components 1 and 2 across clusters.

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
Prin1	10	10.8372596	2.7562227	7.5251386	15.7794184
Prin2	10	-0.9469166	3.9170318	-4.9814433	5.4271811

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
Prin1	14	-0.3566516	3.6599506	-6.7392341	6.2604373
Prin2	14	-1.0731093	3.7200909	-6.5032348	6.7931621

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
Prin1	2	-16.8603797	2.0796819	-18.3309368	-15.3898225
Prin2	2	1.2757631	2.2595340	-0.3219687	2.8734949

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
Prin1	2	-42.2005069	3.3704697	-44.5837889	-39.8172250
Prin2	2	-0.2465638	2.5606625	-2.0572256	1.5640980

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
Prin1	2	7.3711497	6.3665808	2.8692973	11.8730022
Prin2	2	11.2171483	0.1343312	11.1221619	11.3121348

We see that component 1 is highest in cluster 1 and pretty high in cluster 5. These two clusters have greater life expectancies in general. Conversely, cluster 4 has a very nagative average component 1 and the values in cluster 3 are pretty negative, too. These two clusters have shorter life expectancies in general, with cluster 4 being the shortest.

Looking at component 2, cluster 5's large values indicate high infant/child morality and/or very high life expectancy for those who live to adulthood. The other clusters have much smaller values of component 2 on average, indicating smaller gaps between infant and adult life expectancies on average. However, there are pretty wide ranges of values in clusters 1 and 2.

Taken together, we can draw the following inferences about each cluster:

- Cluster 1 has long life expectancies and small to moderate contrast between infant and adult life expectancy.
- Cluster 2 has fairly typical overall life expectancies (mean near 0) and small to moderate contrast between infant and adult life expectancy.
- Cluster 3 countries have low overall life expectancy compared to the average, and roughly average contrast in infant and adult life expectancy.
- Cluster 4 countries have very low overall life expectancy compared to the average, and roughly average contrast in infant and adult life expectancy.
- Cluster 5 countries have fairly high overall life expectancy compared to the average, and a big contrast between infant and adult life expectancy.

(c) A scatter plot showing the clusters and principal components 1 and 2 values follows.

We can see that Madagascar and Cameroon are the countries in cluster 4. These two countries have low life expectancy in general.

Tunisia and Algeria are in cluster 5 and show a bigger difference in infant and adult life expectancy than the typical country. Their people as a whole were expected to live a bit longer than the average, though.

Guatemala and South Africa (C) (presumably standing for people of color in South Africa at the time) are in the cluster with lower than average life expectancies overall by higher than the life expectancies of Madagascar and Cameroon).

Observations in cluster 1 with high overall life expectancy and generally low contrast of infant and adult mortality rates include some countries which were more developed or segments of countries that had more privilege (e.g. the white population in South Africa) at the time. These countries or segments may have had better access to quality health care and had generally higher quality of life at the time.

Cluster 2 is average with respect to both overall life expectancy and contrast between infant and adult life expectancy. The variability within the cluster is also high with respect to both of those features. It is possible that some of these countries are just roughly average. It's also possible that some of these countries or segments within countries experienced a wider variety of health services and quality of life in general at the time the data was collected, leading to longer and shorter life expectancies averaging out in those countries or segments.