

SAS Assignment #2

Prepared By: Group #5

Joel Camacho

Mohamed Mazarul

Bradley John

The purpose of this project was to apply principal components analysis within SAS enterprise miner to determine whether air pollution is significantly related to mortality. The dataset as presented was gathered by researchers at General Motors. The dependent variable for analysis is age adjusted mortality ("Mortality"). The data include variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. Detail of these attributes is as follows.

#	Variable	Description
1	city	City ID
2	JanTemp	Mean January temperature (F)
3	JulyTemp	Mean July temperature (F)
4	RelHum	Relative Humidity
5	Rain	Annual rainfall (inches)
6	Education	Median education
7	PopDensity	Population density
8	NW	Percentage of non-whites
9	WC	Percentage of white-collar workers
10	pop	Population
11	HHSiz	Average household size
12	income	Median income
13	HCPot	HC pollution potential
14	NOxPot	Nitrous Oxide pollution potential
15	SO2Pot	Sulfur Dioxide pollution potential
16	Mortality	Age adjusted mortality

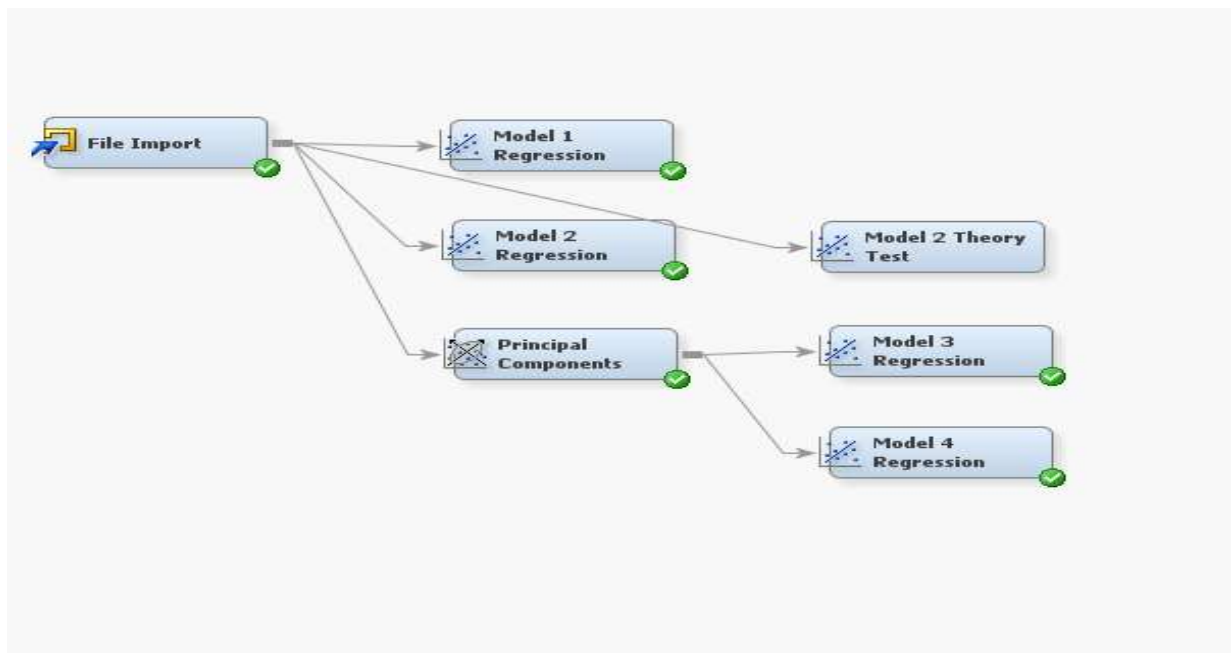
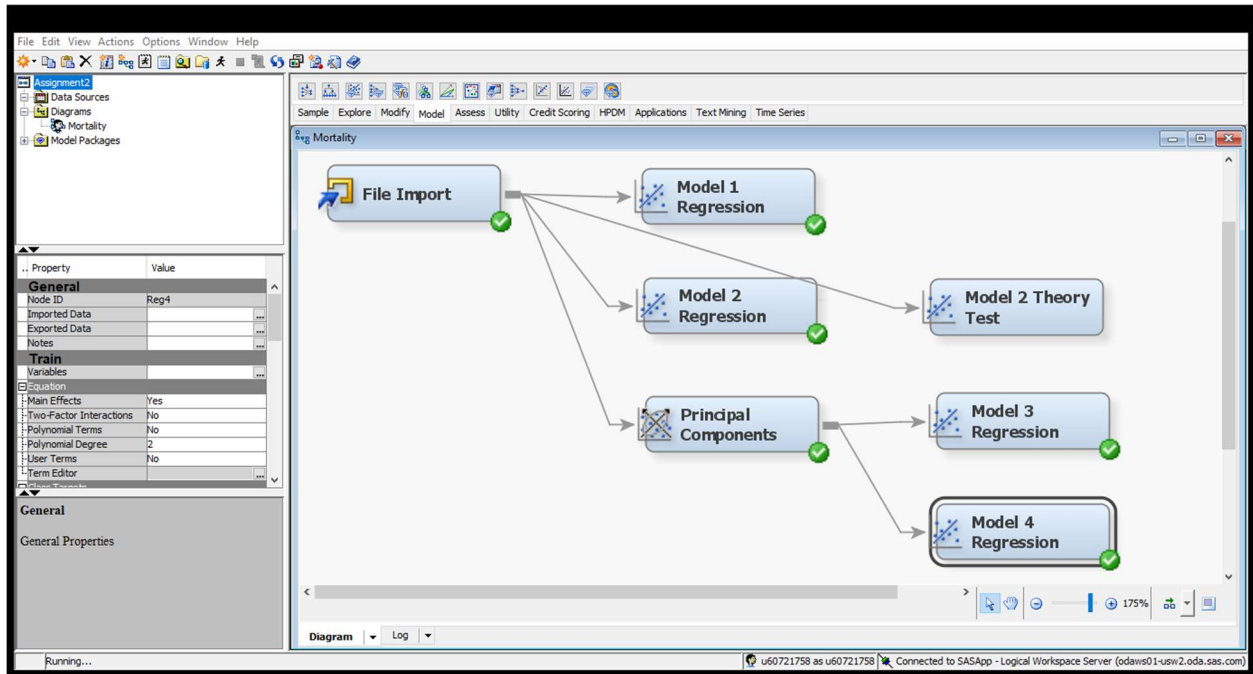
Our first step was to perform a linear regression on the data using all the variables to; 1) see how good of fit the data was in explaining the target and 2) to identify the attributes that are most significant. Our results (Appendix Table B.1) indicated that generally the model is a good fit with an r-squared of .7619 and an adjusted r-squared, given the number of attributes of .6862. The attribute indicating the percentage of non-whites was by far the most significant factor with a t value of 5.85 and associated probability of less than .0001. Only one other attribute, "JanTemp", came close to our threshold for significance. We opted to include this variable as it seemed counterintuitive that only a demographic factor would have significance when the objective is predicting mortality based upon pollution potential.

We then ran another linear regression using only the variables that were deemed significant. This was meaningful to avoid an overfitting situation given the number of variables in the training set. It appears that our decision to include "JanTemp" was not a good one as the r-squared fell to .5421 (appendix C.1) and an adjusted r-squared even lower given that we were only using two variables. We initially thought that there may be an error in our process. As a check on the process and data, we swapped "JanTemp" for a climate attribute that intuitively would seem to have more of direct connection. We choose "Rain". Choosing rain lowered the r-squared further to .4801. This corroborated our initial results by showing that even though "JanTemp" would not seem to be a logical impactor, statistically speaking, it is correct since there is a higher r-squared when using the attribute with the higher t-value and associated probability.

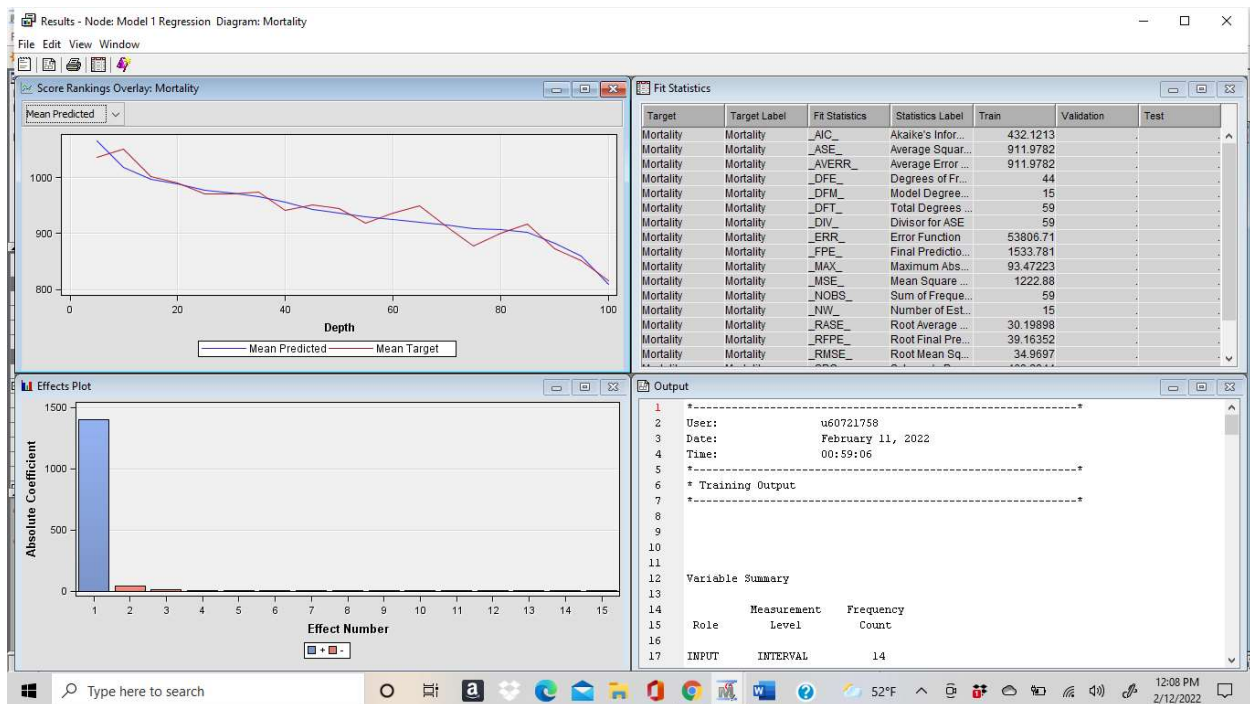
As our initial analysis seemed to indicate that there was really one, possibly two significant attributes, we performed a principal component analysis to reduce our variables and prevent an overfitting situation. The PCA analysis within Enterprise Miner identified 7 principal components (appendices D, D.1, E.1). The results of a regression analysis on the principal components presented a conundrum as the associated r-squared value (.6390) was lower when compared to the produced from the original model incorporating all attributes (Appendix B.1 & E.1). We choose to move forward and perform an additional regression analysis using only the principal components that were the most statistically significant after dimensionality reduction. Those were PC 1,2, and 6 (appendix E.1). Our concern that we may have gone to far in reducing dimensionality seems to be confirmed when we performed a regression analysis on the most significant principal component factors and the r-squared value(.5672) fell even further(appendix F.1)

Appendices:

A: Overall (from different views)








B. Model 1 Results:



B.1 – Output Only

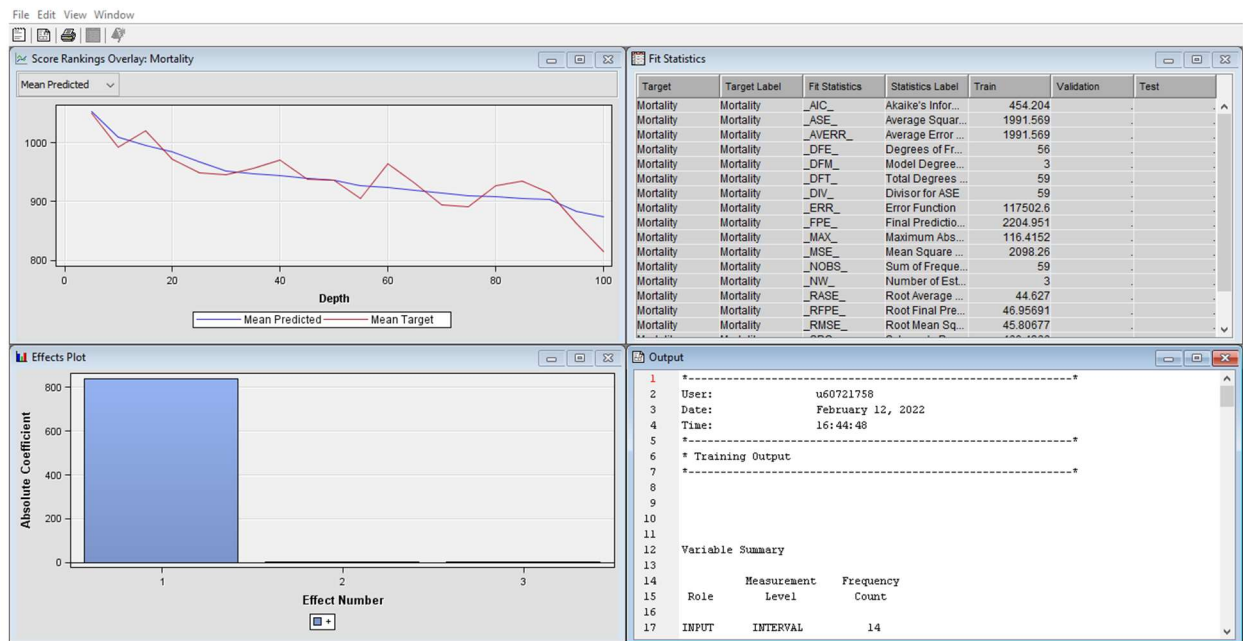
File Edit View Window



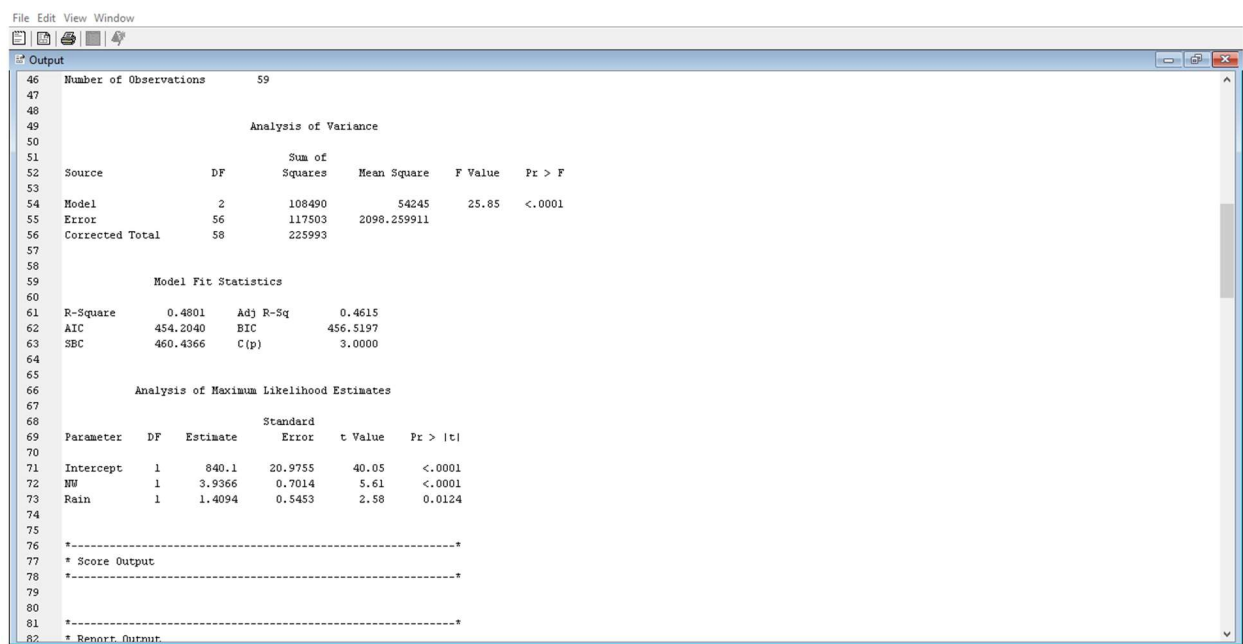
Output

50						
51						
52	Source	DF	Sum of	Mean Square	F Value	Pr > F
53			Squares			
54	Model	14	172186	12299	10.06	<.0001
55	Error	44	53807	1222.879841		
56	Corrected Total	58	225993			
57						
58						
59	Model Fit Statistics					
60						
61	R-Square	0.7619	Adj R-Sq	0.6862		
62	AIC	432.1213	BIC	444.1162		
63	SBC	463.2844	C(p)	15.0000		
64						
65						
66	Analysis of Maximum Likelihood Estimates					
67						
68						
69	Parameter	DF	Estimate	Standard	t Value	Pr > t
70				Error		
71	Intercept	1	1400.0	281.6	4.97	<.0001
72	Education	1	-11.0467	8.9980	-1.23	0.2261
73	HCPot	1	-0.6712	0.4556	-1.47	0.1478
74	HHSiz	1	-38.0416	40.2752	-0.94	0.3501
75	JanTemp	1	-1.4411	0.7638	-1.89	0.0658
76	JulyTemp	1	-2.9503	1.9351	-1.52	0.1345
77	NOxPot	1	1.1785	0.9144	1.29	0.2042
78	NW	1	5.3027	0.9064	5.85	<.0001
79	PopDensity	1	0.00472	0.00434	1.09	0.2826
80	Rain	1	0.9695	0.5851	1.66	0.1046
81	RelHum	1	0.1360	1.1508	0.12	0.9065
82	SO2Pot	1	0.0846	0.1357	0.62	0.5362
83	WC	1	-1.4923	1.2325	-1.21	0.2324
84	income	1	-0.00043	0.00129	-0.33	0.7435
85	pop	1	3.402E-6	4.116E-6	0.83	0.4129
86						

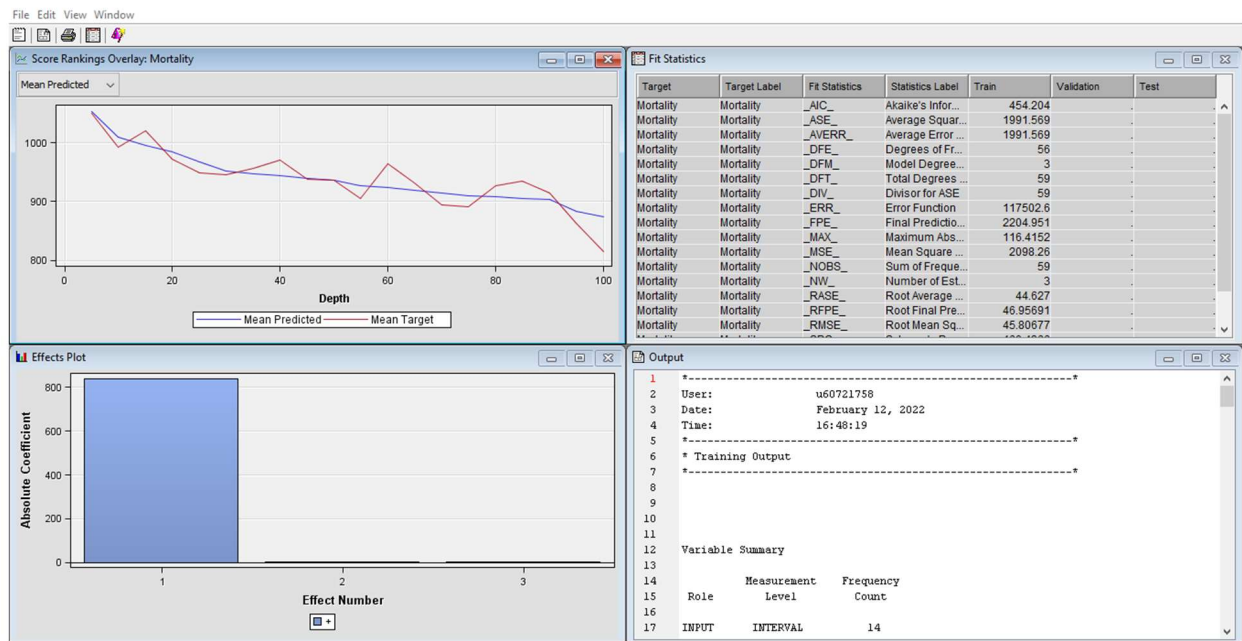
C: Model 2 Results (NW and JanTemp)



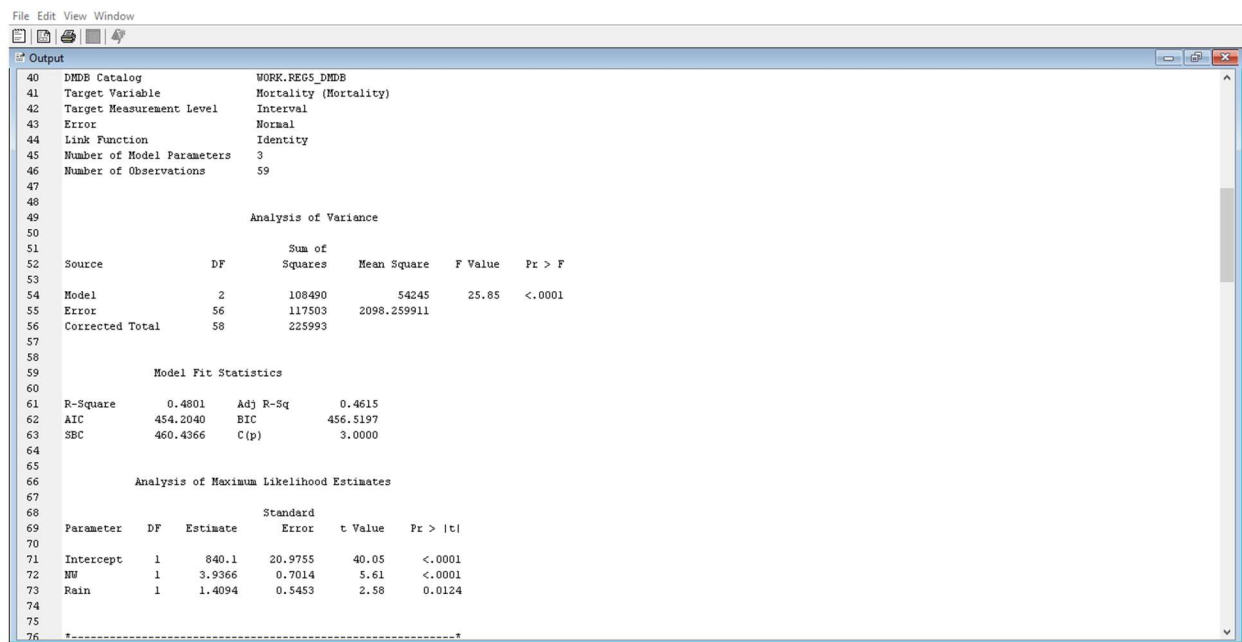
C.1 – Output



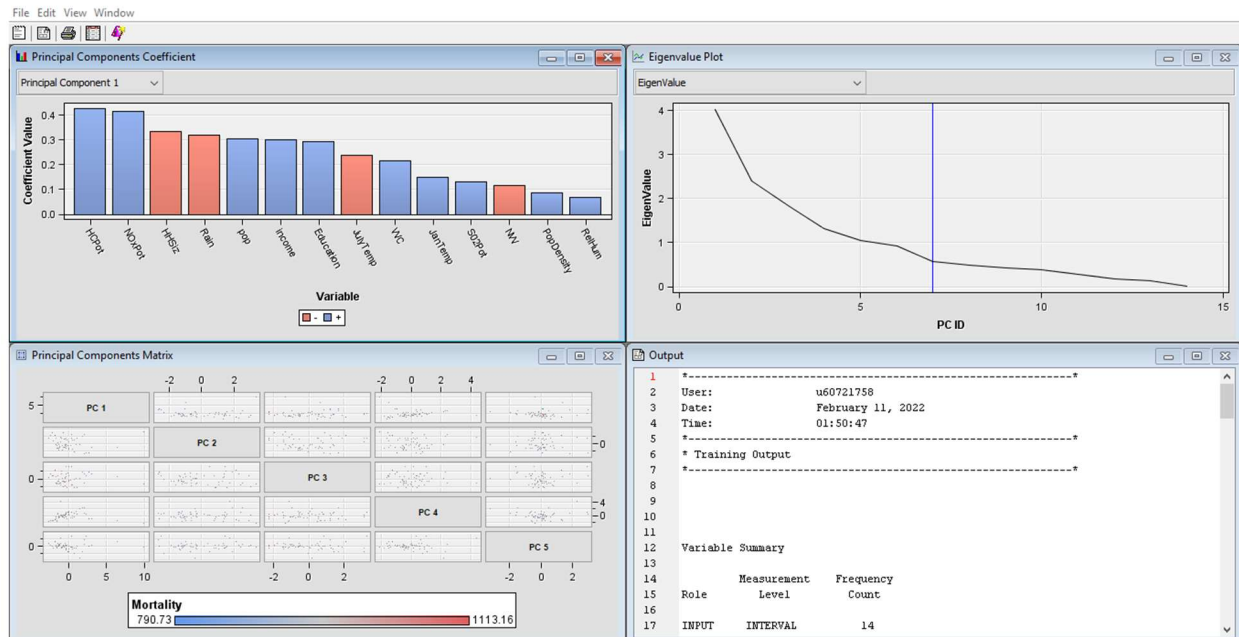
C.2 – Model 2 Theory Test (Rain and NW)



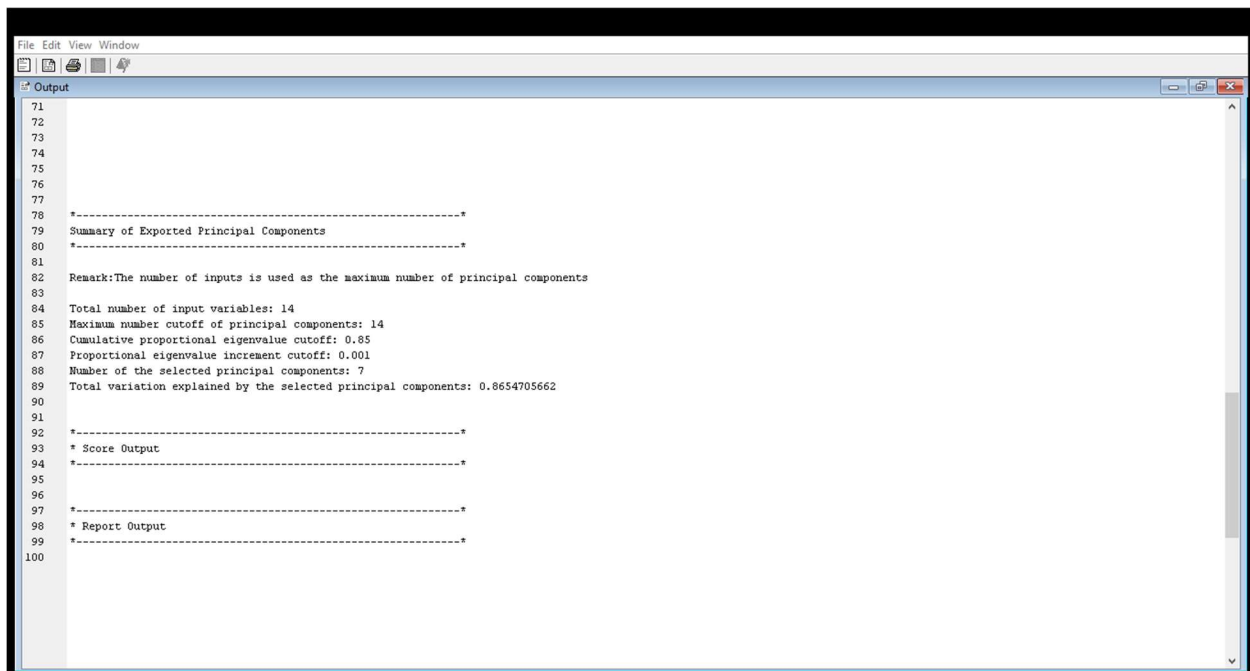
C.3 – Theory Output Only



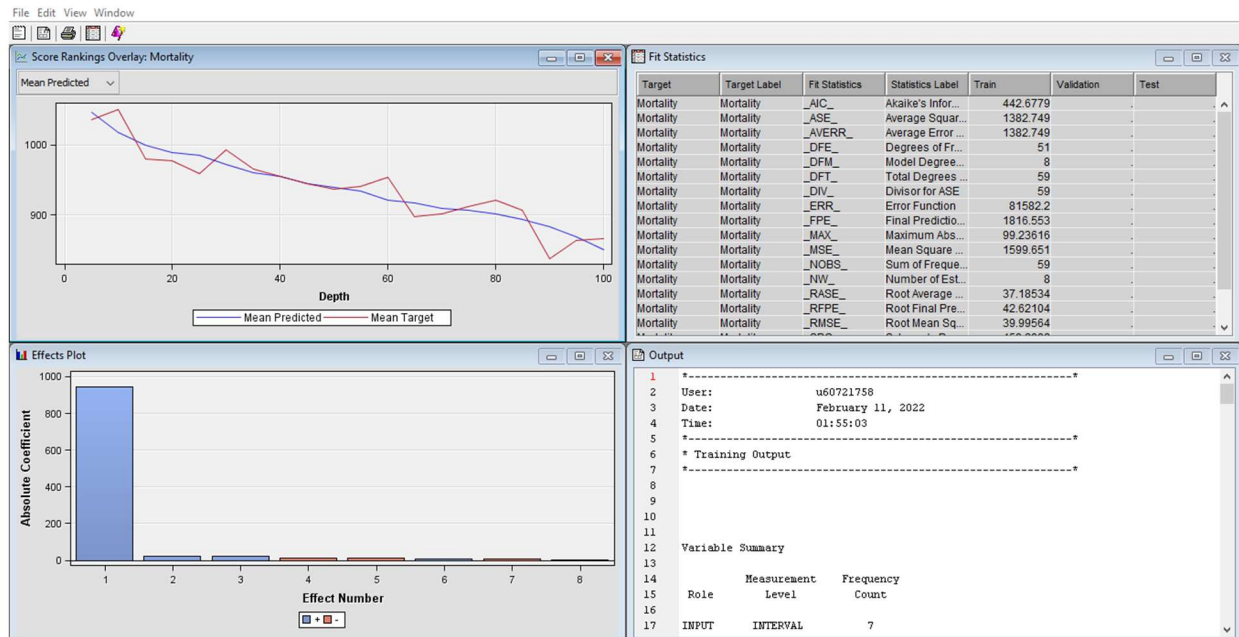
D. Principal Components Results:



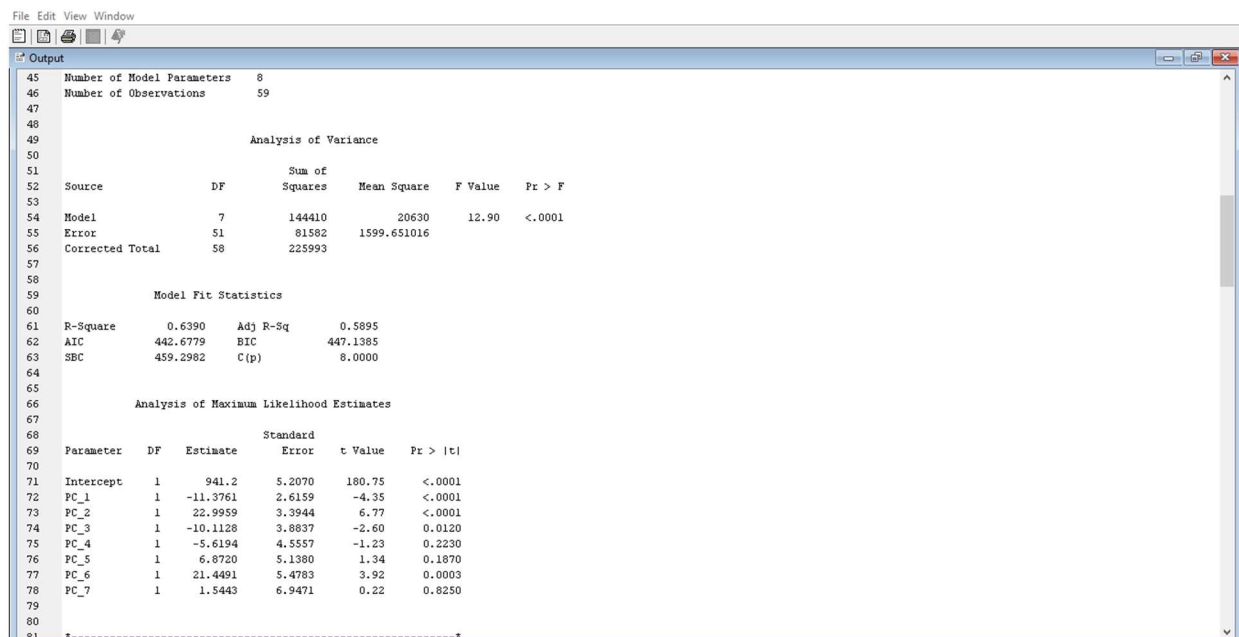
D.1 – Principal Components Output – Full out put in Adobe.



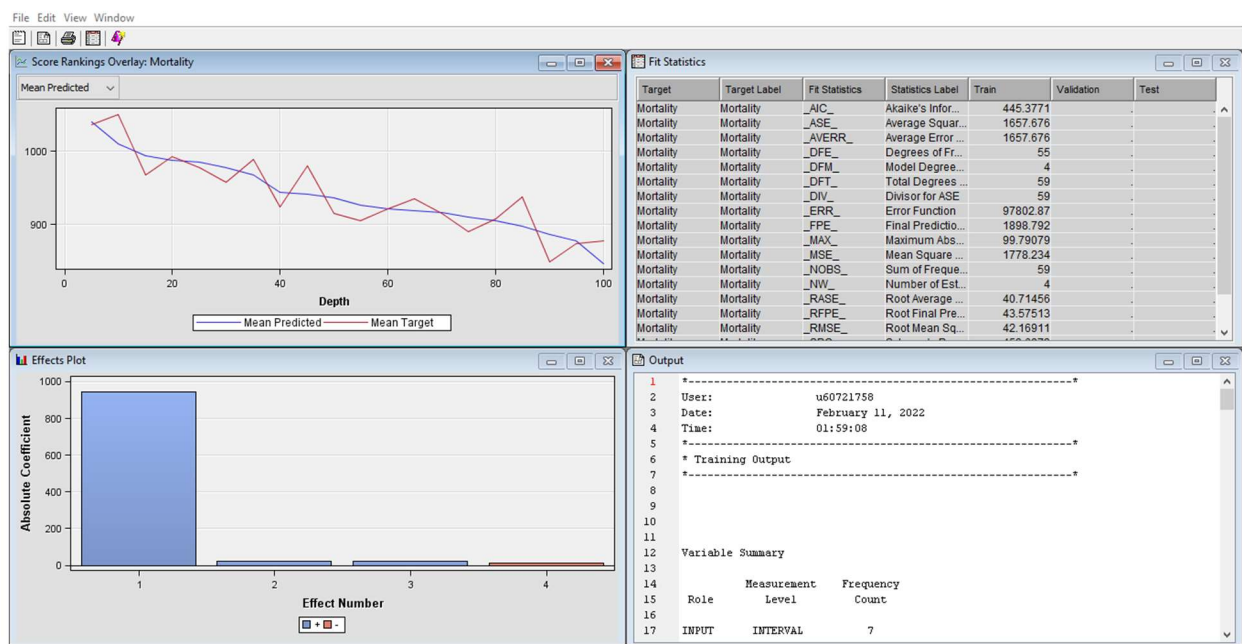
E. – Model 3 Output:



E.1 – Output Only



F – Model 4 Results:



F.1 – Output Only

