California State University, Long Beach

Multivariate Analysis of State Crime Data in 1985

STAT 550 – Multivariate Statistical Analysis: Midterm Take Home Exam

Kevin Merritt

4-11-2016

Introduction

The purpose of this report is to analyze crime data from all fifty states in 1985. In particular, I will consider the following variables: land area, population, murder, rape, robbery, assault, burglary, larceny, and auto theft. I have chosen to omit the variables state, US region number, and US division number from this analysis because they are categorical variables that explain groups without providing any explanatory information.

This report will start with a preliminary analysis that will help determine whether to use a covariance or correlation method in addition to identifying any holes in the assumptions needed to run the analysis. After the preliminary analysis, the data will be analyzed using two methods: principal component analysis and factor analysis, which will be referred to as PCA and FA, respectively. Each of these methods will be compared back to the preliminary analysis to check if my initial thoughts on the data set were accurate.

Preliminary Analysis

Covariance vs. Correlation

As we move forward, the first decision we must make is whether to use a covariance or correlation matrix approach. Both options are possible and arguments can be made for each. Looking at the covariance matrix (found in Appendix A), it is clear that it's difficult to compare the covariances to each other due to the fact that the units and size of the numbers associated with each variable are very different. For example, population and land size are going to be very large numbers while murder would be a comparatively low number. Compare this to the correlation matrix below and it becomes clear that we should use a correlation matrix, due to the fact that it allows us to compare each variable to each other based on a specific range (negative one to positive one) and eliminates the issues of each category being different in the size of the numbers. For these reasons, I have chosen to use the correlation matrix throughout my analysis of the crime data for 1985.

Pearson Correlation Coefficients, N = 50									
	land	popu	murd	rape	robb	assa	burg	larc	auto
land	1.00000	0.07188	0.24450	0.37683	-0.02054	0.16203	0.06765	0.25319	0.08236
popu	0.07188	1.00000	0.27216	0.41805	0.62324	0.42635	0.42856	0.23054	0.37589
murd	0.24450	0.27216	1.00000	0.51987	0.34106	0.81256	0.27672	0.06478	0.10983
rape	0.37683	0.41805	0.51987	1.00000	0.55144	0.69593	0.68015	0.60061	0.44070
robb	-0.02054	0.62324	0.34106	0.55144	1.00000	0.56320	0.62219	0.43618	0.61705
assa	0.16203	0.42635	0.81256	0.69593	0.56320	1.00000	0.52072	0.31670	0.33038
burg	0.06765	0.42856	0.27672	0.68015	0.62219	0.52072	1.00000	0.80110	0.70010
larc	0.25319	0.23054	0.06478	0.60061	0.43618	0.31670	0.80110	1.00000	0.55478
auto	0.08236	0.37589	0.10983	0.44070	0.61705	0.33038	0.70010	0.55478	1.00000

Can we eliminate any other variables from our analysis?

In the introduction, it was mentioned I would omit the variables state, US region number, and US division number because they are categorical variables. Similarly, if a variable does not have at least a moderate (.50) correlation with any other variables, I can choose whether to omit it as well since it will not play a major role in the analysis. Looking at our correlation matrix above, I have chosen to omit the variable land (land area) from the analysis. All other variables should be kept because they have a high correlation with at least one other variable in the analysis.

Is the normality assumption violated for any of the variables?

For the most accurate and ideal analysis of the data using PCA and FA, the variables should be normally distributed. For this reason, I ran a Shapiro-Wilk's test on each variable included in the study to test for violations of normality. If we choose to be 90 percent confident in our analysis, any p-value that is above .10 in the following tables will be considered to violate the normal distribution.

Variable: popu

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.749649	Pr < W	<0.0001

Variable: murd

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.957022	Pr < W	0.0667

Variable: rape

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.94862	Pr < W	0.0299

Variable: robb

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.815839	Pr < W	<0.0001

Variable: assa

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.964785	Pr < W	0.1410

Variable: burg

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.966393	Pr < W	0.1645

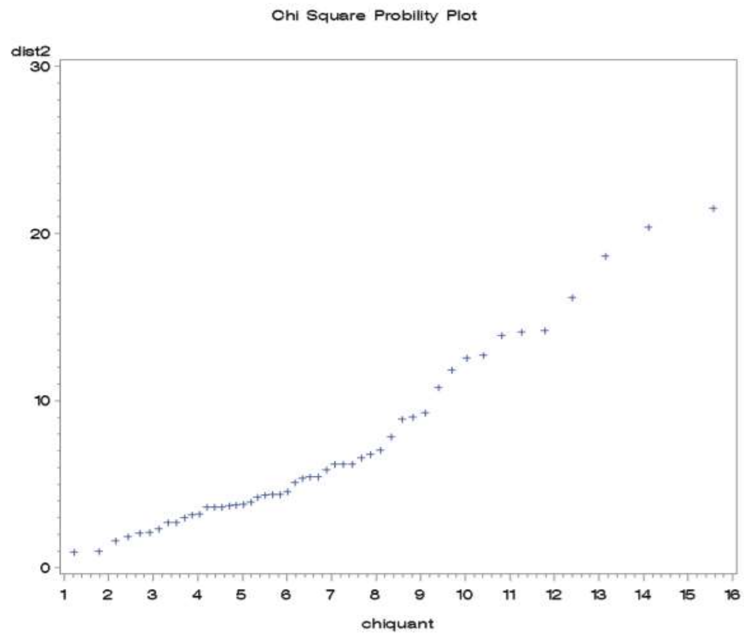
Variable: larc

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.972811	Pr < W	0.3001

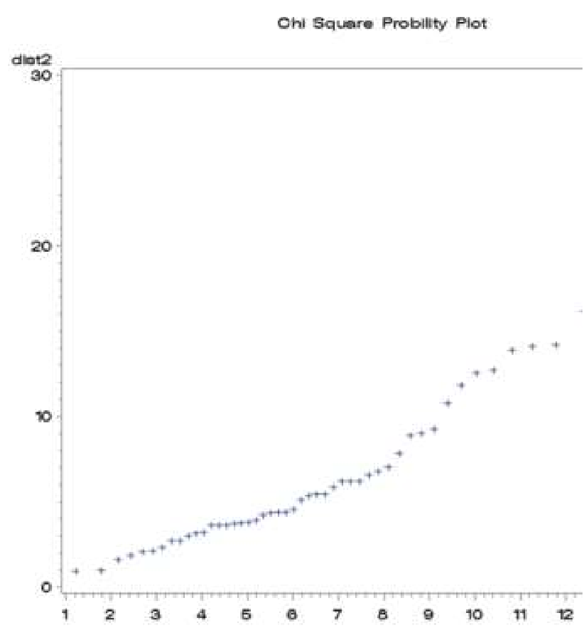
Variable: auto

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.955449	Pr < W	0.0573

Looking at the Shapiro-Wilk's test for each variable, we can see that assault (assa), burglary (burg), and larceny (larc) all have p-values that are above the alpha level of .10. Hence, these three variables violate our assumption of normality. To further test these results I can create a Chi-Square Probability plot (shown below).



Within a Chi-square Probability Plot, normality is defined by all points falling in a straight line (or close to a straight line). In our case, it does not appear to be all that straight. One option would be to use a Box-Cox Transformation to fix the normality issue; this however, can be complicated. If points to the right are "far" away from other points and not clustered, we can treat them as outliers. Therefore, we can look at the plot without the four right points. If we do this, we get the plot below.



This shows that the points do make a generally straight line and we can assume that the assumption of normality is not violated; thus, we can avoid the complication of a Box-Cox Transformation.

Possible groupings of variables

I would expect that any variables with a high correlation (.7 or higher typically) could potentially be a group. Using this threshold and the matrix below, it is easy to see that the highest correlation is between murder and assault (.81256). I would expect these two variables to be grouped together. The next highest is burglary and larceny (.80110). Burglary is also highly correlated with auto theft (.70010) so it is possible that these three variables could form a grouping.

Pearson Correlation Coefficients, N = 50
Prob > |r| under H0: Rho=0

	popu	murd	rape	robb	assa	burg	larc	auto
popu	1.00000	0.27216 0.0559	0.41805 0.0025	0.62324 <.0001	0.42635 0.0020	0.42856 0.0019	0.23054 0.1072	0.37589 0.0071
murd	0.27216 0.0559	1.00000	0.51987 0.0001	0.34106 0.0154	0.81256 <.0001	0.27672 0.0517	0.06478 0.6549	0.10983 0.4477
rape	0.41805 0.0025	0.51987 0.0001	1.00000	0.55144 <.0001	0.69593 <.0001	0.68015 <.0001	0.60061 <.0001	0.44070 0.0014
robb	0.62324 <.0001	0.34106 0.0154	0.55144 <.0001	1.00000	0.56320 <.0001	0.62219 <.0001	0.43618 0.0015	0.61705 <.0001
assa	0.42635 0.0020	0.81256 <.0001	0.69593 <.0001	0.56320 <.0001	1.00000	0.52072 0.0001	0.31670 0.0250	0.33038 0.0191
burg	0.42856 0.0019	0.27672 0.0517	0.68015 <.0001	0.62219 <.0001	0.52072 0.0001	1.00000	0.80110 <.0001	0.70010 <.0001
larc	0.23054 0.1072	0.06478 0.6549	0.60061 <.0001	0.43618 0.0015	0.31670 0.0250	0.80110 <.0001	1.00000	0.55478 <.0001
auto	0.37589 0.0071	0.10983 0.4477	0.44070 0.0014	0.61705 <.0001	0.33038 0.0191	0.70010 <.0001	0.55478 <.0001	1.00000

<u>Principal Component Analysis (PCA)</u>

Correlation or Covariance Matrix for PCA

The variables chosen for the PCA analysis are the same eight that were originally chosen for the preliminary analysis in the previous section. There are two things that I can check to determine if a covariance or correlation matrix should be used to perform PCA: the range of sample variances and the percentage each variable plays into the principal components for each matrix.

Variable	N	Mean	Std Dev	Minimum	Maximum
popu	50	4762.26	5068.98	509.00	26365.00
murd	50	6.86	3.85	0.50	15.30
rape	50	15.62	7.35	3.60	36.00
robb	50	101.51	91.19	6.50	443.30
assa	50	135.42	68.17	21.00	293.00
burg	50	930.80	361.05	286.00	1753.00
larc	50	1943.64	709.83	694.00	3550.00
auto	50	367.86	199.61	78.00	878.00

The simple statistics above show that the highest sample standard deviation is 5068.98 for population while the lowest is 3.85 for murder. Such a large range for stand deviation supports using the correlation matrix over a covariance matrix. This is confirmed by the fact that the first two principal components of the covariance analysis below are heavily weighted by a single variable. Compare this to the correlation PCA on the right, where no singular variable accounts for a majority of the principal component. For these reasons, I chose to use the correlation matrix for this analysis.

Covariance PCA

	Prin1	Prin2
popu	0.998782	-.044597
murd	0.000207	0.000161
rape	0.000611	0.005273
robb	0.011244	0.040342
assa	0.005753	0.023382
burg	0.030876	0.373759
larc	0.033123	0.914218
auto	0.014939	0.142543

Correlation PCA

	Prin1	Prin2
popu	0.298672	0.063303
murd	0.262779	0.632799
rape	0.399587	0.101583
robb	0.385958	-.040462
assa	0.371976	0.444816
burg	0.413155	-.269873
larc	0.331591	-.418809
auto	0.337778	-.370821

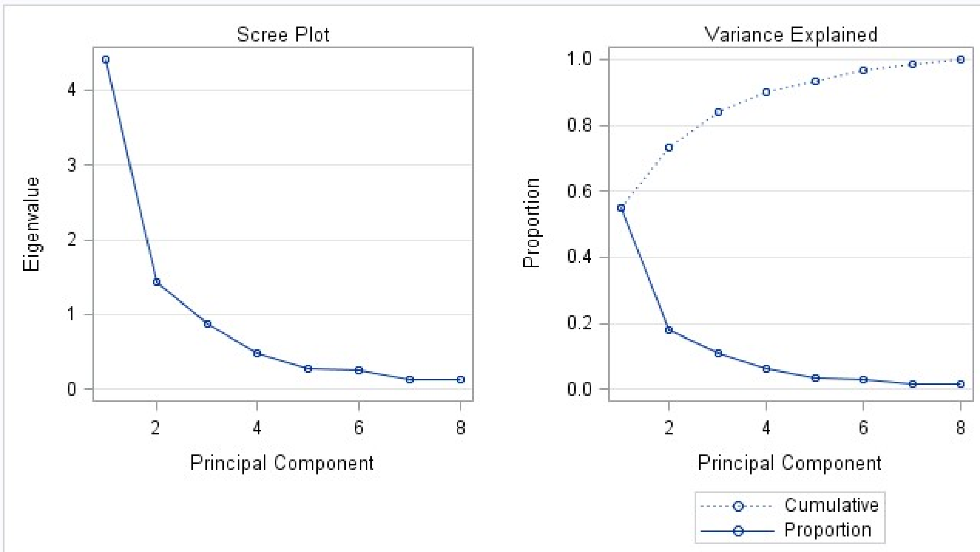
How many principal components were used?

Deciding the number of principal components to use in an analysis is a somewhat subjective matter. First, I analyzed the eigenvalues and the relative contributions they each make to the variance.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.40900906	2.97421466	0.5511	0.5511
2	1.43479441	0.55724268	0.1793	0.7305
3	0.87755173	0.39557198	0.1097	0.8402
4	0.48197976	0.20086410	0.0602	0.9004
5	0.28111566	0.03286929	0.0351	0.9356
6	0.24824637	0.11304429	0.0310	0.9666
7	0.13520207	0.00310114	0.0169	0.9835
8	0.13210094		0.0165	1.0000

The first eigenvalue has a value of about 4.41 and explains 55 percent of the variance. In my opinion, this is not enough variance to adequately describe the model; for this reason, I would add in the second eigenvalue, which would increase the cumulative variance explained to about 73 percent. I would ideally prefer at least 80 percent of the variance explained, necessitating the addition of a third eigenvalue to get to a cumulative variance explanation of 84 percent (a rather good percentage for three components). An argument could be made to include a fourth eigenvalue; however, I do not believe an increase of six percent in the variance explained is worth an extra dimension. Consider: if equally distributed, each eigenvalue would explain 12.5 percent (100%/8) of the variance. Since six percent is about half of that, it supports my decision to not add another dimension to my analysis. The scree plots below also show that a good spot to perform the cutoff would be after the third principal component. Therefore, I will use three principal components to describe the data.



What does each Principal Component represent?

Since I chose to include three principal components, I am going to focus my interpretation on the first three components (the highlighted ones) below.

The first principal component seems to be relatively evenly distributed among the variables. All have positive values, meaning each plays a positive role in the component. That said, burglary (.413155), rape (.399587), robbery (.385958), and assault (.371976) have the most effect on the principal component. It seems that the first principal component is essentially representing an average crime since:

A. no variables stick out an obscene amount when compared to the rest, and
B. Each variable has a significant positive effect on the PC.

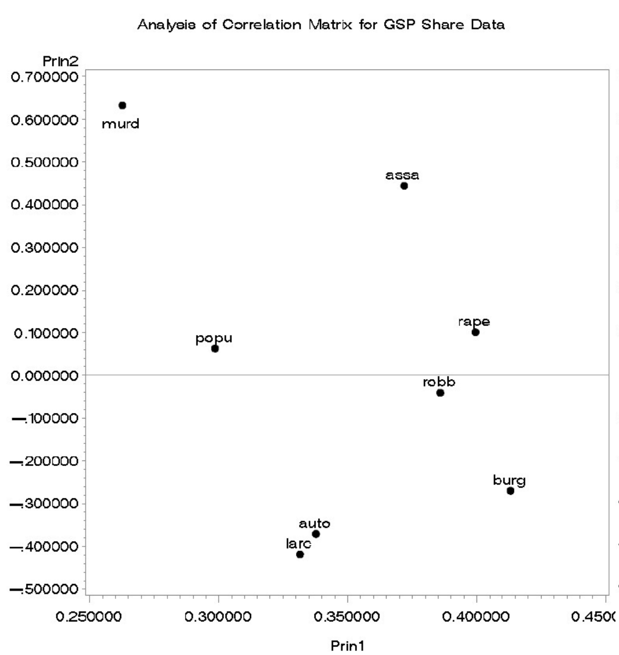
For the second PC, murder (.632799) and assault (.444816) have large positive values while larceny (-.418809) and burglary (-.269873) have large negative values. This would seem to indicate that the second principal component represents variables which involve crimes where physical damage tends to happen. This is supported by the fact the next highest component is rape (.101583).

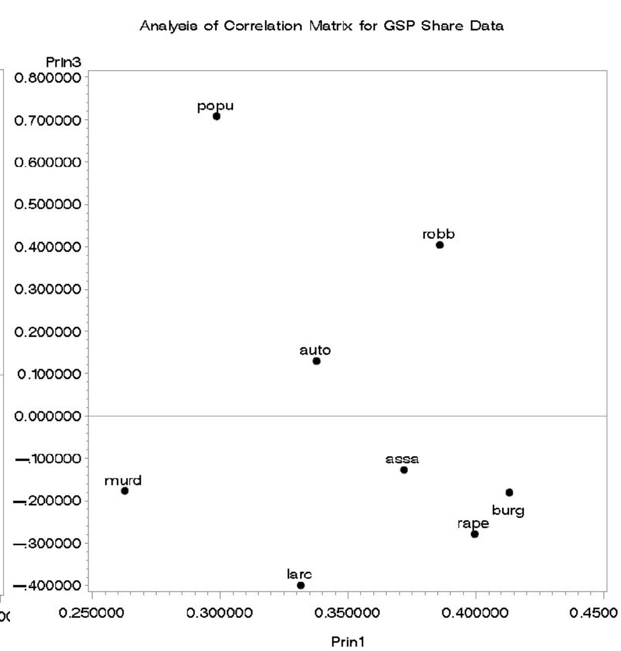
The third and final PC I chose weighs population (.708735) and robbery (.404520) the most. On the other end, larceny (-.399904) and rape (-.278409) have the largest negative effect on the PC. This would seem to indicate this PC represents crimes that involve theft that could involve harm to people in some way.

Eigenvectors								
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
popu	0.298672	0.063303	0.708735	-.489515	0.390549	0.030901	0.103880	0.024131
murd	0.262779	0.632799	-.176338	0.210056	0.217799	0.212305	0.030055	0.601566
rape	0.399587	0.101583	-.278409	-.343273	-.166217	-.767450	-.097242	0.092734
robb	0.385958	-.040462	0.404520	0.208375	-.772909	0.145646	-.035322	0.149982
assa	0.371976	0.444816	-.126847	0.109027	-.005911	0.134115	0.197686	-.760720
burg	0.413155	-.269873	-.180174	-.035528	0.194251	0.306191	-.765053	-.077167
larc	0.331591	-.418809	-.399904	-.311736	-.035494	0.356859	0.554492	0.146539
auto	0.337778	-.370821	0.130287	0.666834	0.368769	-.323560	0.213904	0.017859

Are there any variables that are grouped or outliers?

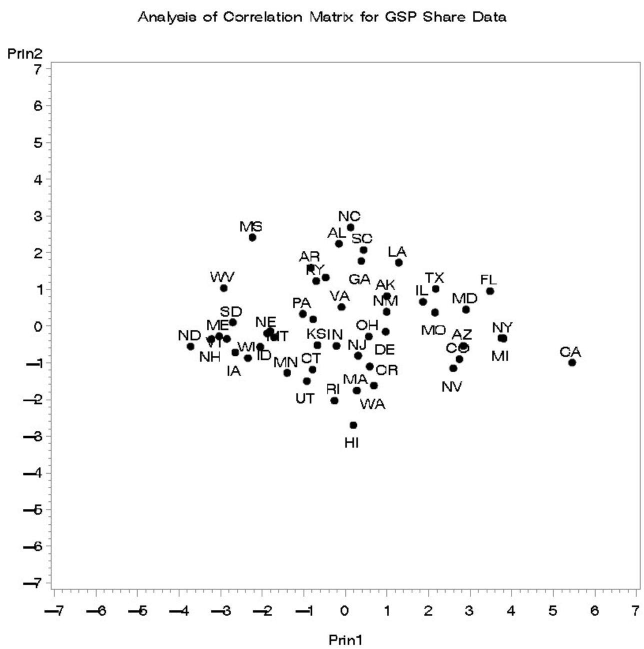
The scatterplots below show the eigenvectors of the variables plotted based on the principal components. There are three possible scatterplots: prin1 vs prin2, prin1 vs prin3, and prin2 vs prin3. Since prin1 seems to represent overall crime rates (see discussion above), I will use the two scatterplots comparing prin1 to prin2 and prin3. When comparing prin1 (overall crime) and prin2 (physical crimes), it would seem that murder is a potential outlier. However, it is not drastically distanced from the next closest data point for either PC, indicating it is most likely not an outlier. It would also appear that auto theft and larceny (mentioned as a possible grouping in preliminary analysis) are a strong grouping. In the second plot, population could be considered an outlier when it comes to prin3, but not prin1, so it should not be classified as such. Finally, it appears rape and burglary are a possible grouping.

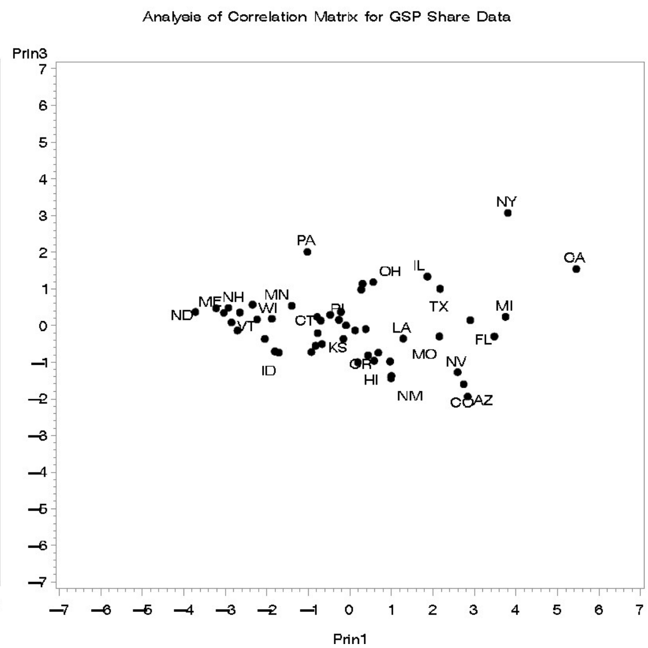




Based on PCA, where should you live in 1985?

Using scatterplots of eigenvectors by state, we can get a general feel for the quality of life in each state based on the values each PC represents (see "What does each Principal Component represent"). Again, we will use prin1 vs prin2 and prin1 vs prin3 for the same reasons listed above.





Looking at the first plot above, it appears that Mississippi, North Carolina, Alabama, and South Carolina would be the worst places for physical harm crimes. Oddly enough, all of these states can be considered part of the South East Region. This would be a good place to avoid if you're worried about these types of crimes, especially considering that Georgia is right next to South Carolina in the plot and Florida is in the top third of prin2. On the opposite side, Hawaii, Rhode Island, and Utah are examples of places you would want to live to avoid these crimes. This plot also shows that California has the most crimes overall, along with New York and Michigan. Meanwhile, North Dakota, New Hampshire, and Maine seem to have some of the lowest overall crime.

The second plot seems to indicate the highest theft states are New York, California, Pennsylvania, and Illinois, while the lowest theft states are Colorado, Arizona, and New Mexico. This plot also confirms that California, New York, and Michigan have some of the highest overall crime rates for states.

Factor Analysis

Chi-Square Test for Adequacy

In order to perform the Factor Analysis tests, I needed to determine two things:

A. If there were any common factors, and

B. Whether three factors (chosen based on our PCA analysis) was a sufficient numbers of factors for this data.

I used the maximum likelihood method to test the significance of both queries. As seen below, the p-value for no common factors is <.0001; hence, I rejected the null hypothesis and concluded that there is at least one common factor. Similarly, it is easy to see there is a p-value of .7272 when testing if three factors are sufficient. This means we do not reject the null hypothesis and conclude that three factors are sufficient. This is the same number of factors I used in my PCA analysis.

Significance Tests Based on 50 Observations			
Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	28	259.5000	<.0001
HA: At least one common factor			
H0: 3 Factors are sufficient	7	4.4457	0.7272
HA: More factors are needed			

To verify my findings, I also analyzed the residual correlations matrix. All values that are not on the diagonal are calculated by taking the difference of the true sample correlations and the correlations that are created using the three-factor solution. Basically, if three factors are an appropriate amount, we should see small values in all entries off the diagonals. This is confirmed by the small values not on the diagonal in the table below.

Residual Correlations With Uniqueness on the Diagonal								
	popu	murd	rape	robb	assa	burg	larc	auto
popu	0.52134	-0.01140	0.03846	0.01241	-0.00396	0.00783	-0.00743	-0.05682
murd	-0.01140	0.19215	-0.00307	-0.00346	-0.00001	0.01109	-0.00879	0.00699
rape	0.03846	-0.00307	0.30782	0.00343	0.00084	-0.01071	0.01368	-0.03478
robb	0.01241	-0.00346	0.00343	0.19966	0.00245	-0.01357	0.00863	0.00279
assa	-0.00396	-0.00001	0.00084	0.00245	0.09207	-0.00214	0.00133	-0.00085
burg	0.00783	0.01109	-0.01071	-0.01357	-0.00214	0.14540	-0.00160	0.04127
larc	-0.00743	-0.00879	0.01368	0.00863	0.00133	-0.00160	0.14307	-0.02360
auto	-0.05682	0.00699	-0.03478	0.00279	-0.00085	0.04127	-0.02360	0.40732

Furthermore, we can verify three factors are sufficient by noticing that the total root mean square of the off-diagonal residuals is .01838. Therefore, we can conclude that a three factor model is appropriate.

Root Mean Square Off-Diagonal Residuals: Overall = 0.01838120							
popu	murd	rape	robb	assa	burg	larc	auto
0.02705529	0.00756433	0.02074409	0.00802028	0.00205205	0.01770182	0.01168166	0.03106560

Unrotated Communalities and Factor Loadings

Next, I analyzed the communalities and factor loadings to verify that they made sense. I used the recommended approach of setting the prior communality estimates equal to the squared multiple correlation with all remaining variables for each variable. Looking at the table below, I do not see any values that exceed one (which would be an issue) or any values that are very low.

Prior Communality Estimates: SMC							
popu	murd	rape	robb	assa	burg	larc	auto
0.41969779	0.71575989	0.66466546	0.62983609	0.79816774	0.80060499	0.71306984	0.56709110

Looking at the final communality values, we can determine whether enough of the variance is explained within the eight variables that were chosen. Keep in mind communality tells us how much of the variables variation is explained by the model chosen. Ideally, all variables would be close to one. A low communality would indicate the variable's variance is not explained by the model very well. With this in mind, we can see that assault (.9079), murder (.8078), robbery (.8003), burglary (.8546), and larceny (.8569) all have variations explained well by this model. Population (.4787) is the least explained variable variance in the model. Overall, I concluded that the model explained enough of the variance within the variables chosen.

Final Communality Estimates and Variable Weights		
Total Communality: Weighted = 34.562442 Unweighted = 5.991160		
Variable	Communality	Weight
popu	0.47865611	1.9182300
murd	0.80784535	5.2042147
rape	0.69218129	3.2487526
robb	0.80034269	5.0083819
assa	0.90792963	10.8612704
burg	0.85460091	6.8773234
larc	0.85692758	6.9891435
auto	0.59267659	2.4551250

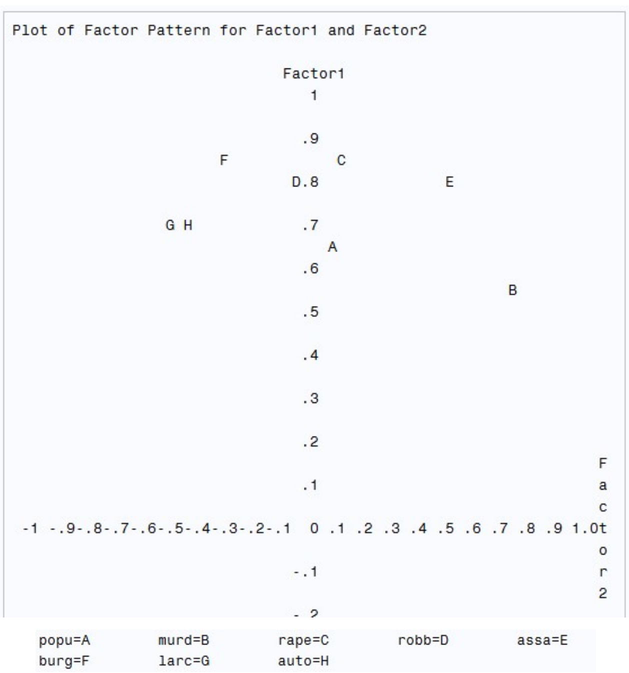
Variance Explained by Each Factor		
Factor	Weighted	Unweighted
Factor1	24.0267325	4.12000568
Factor2	8.1932161	1.27461651
Factor3	2.3424931	0.59653797

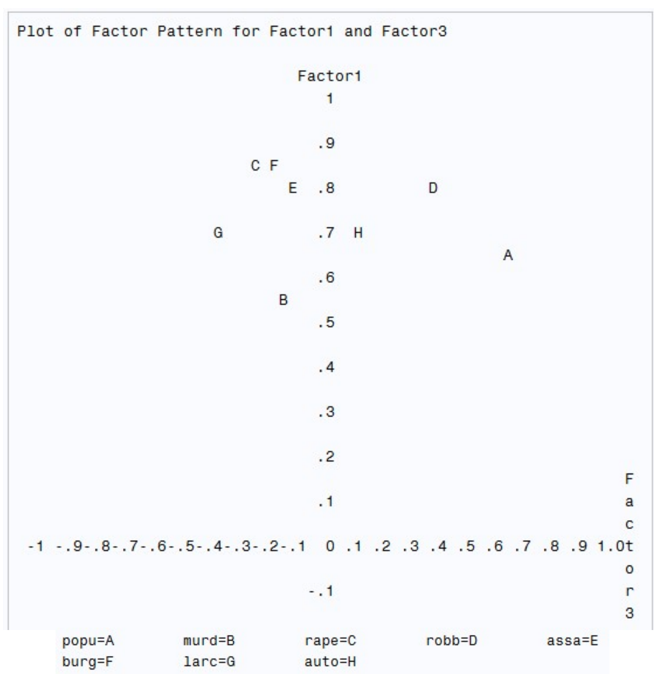
Factor Pattern			
	Factor1	Factor2	Factor3
assa	0.84527	-0.43635	0.05516
burg	0.82776	0.40992	0.03711
rape	0.82140	0.01481	0.13137
robb	0.74360	0.09360	-0.48851
larc	0.66393	0.59491	0.24941
auto	0.61183	0.39443	-0.25054
popu	0.53286	-0.00972	-0.44116
murd	0.62739	-0.63056	0.12893

The variance explained by factor 1 is three times that explained by factor 2 and twelve times that of factor 3. Similarly, factor 2 explains about four times the variance than factor 3.

Using the table on the right above, we can analyze the loading of each factor specifically. The first factor shows assault (.84527) with the highest loading and population (.53286) as the lowest loading. However, all variables in the first factor have high, positive loadings, indicating it probably represents overall crime in the states. The second factor has high positive loadings of larceny (.59491), burglary (.40992), and auto theft (.39443). The lowest loadings in factor 2 are assault (-.43635) and murder (-.63056). This indicates the second factor represents crimes involving stolen property. The third factor has high positive loadings of larceny (.24941), rape (.13137), and murder (.12893) while having high negative loadings for robbery (-.48851) and population (-.44116). The groupings within this factor make it difficult to say exactly what this factor is representing.

To determine groupings of variables, I examined both factor 2 and factor 3 vs factor 1 loading plots. I chose these two because factor 1 is almost certainly overall crime. It seems that when looking at the first two factors plotted against each other, larceny (G) and auto theft (H) (possible grouping from preliminary analysis) could be considered a grouping. Otherwise, an argument could be made that rape (C) and robbery (D) are a grouping too; however, I would not consider that to be the case. Otherwise, all other variables are rather spread out. Looking at factors 1 and 3 plotted against each other, it is very obvious that rape (C) and burglary (F) are a group. It also seems that assault (E) should be included in that group. Otherwise, the rest of the variables are spaced out.





Rotated Communalities and Factor Loadings

To examine the rotated communalities and factor loadings I chose to use the Varimax method due to the fact that it is an orthogonal rotation and won't affect the communality estimates that were produced by the unrotated factor solution. The first table below shows the transformation used for the rotation. The second table shows the variance explained by each factor. We can see that factor 1 is not nearly as heavily weighted as in the unrotated problem. In fact, it isn't even twice as much as factor 2 or factor 3. Furthermore, I looked at the loadings for each factor and found them to be considerably different than the unrotated versions. Factor 1 now has larceny (.93156), burglary (.85365) and auto theft (.70065) as its highest communalities. Meanwhile, murder (-.02168) and population (.10435) have the lowest. Factor 1 now seems to represent crimes that involve theft of some sort. Factor 2 has two very high communalities in murder (.94429) and assault (.87544) with rape (.62270) close behind. The low communalities of larceny (.08608) and auto theft (-.00698) point to the fact that factor 2 most likely explains crimes that involve physical violence of some sort. Factor 3 has high communalities in population (.88782) and robbery (.73731) and low communalities in larceny (.03978) and murder (.11885). It is hard to pinpoint exactly what factor 3 is representing in this case. One educated guess is that it represents crimes that typically take place in high population areas, which would also explain the high communality of population in the factor.

Orthogonal Transformation Matrix

	1	2	3
1	0.67630	0.53630	0.50498
2	-0.61069	0.79154	-0.02275
3	-0.41191	-0.29299	0.86283

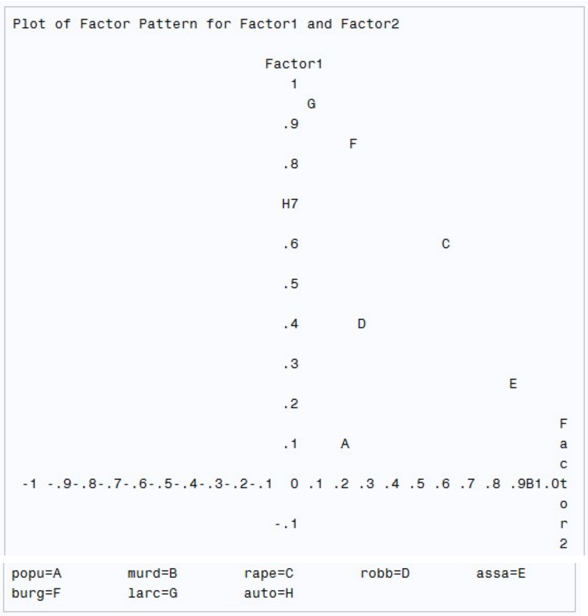
Variance Explained by Each Factor

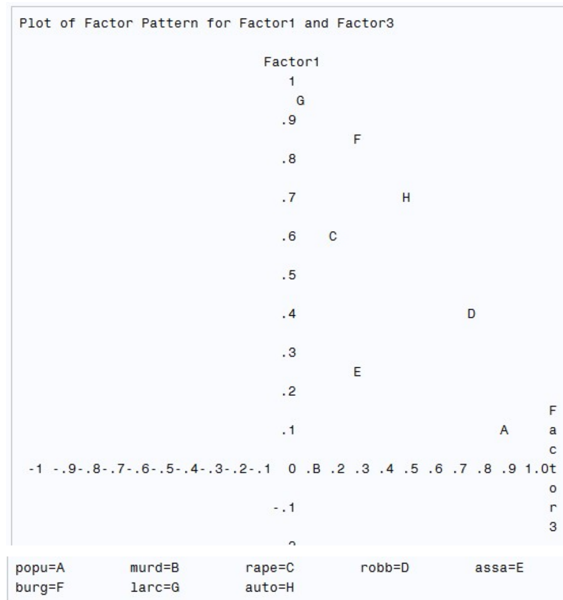
Factor1	Factor2	Factor3
2.7006090	2.2423789	1.7783674

Rotated Factor Pattern

	Factor1	Factor2	Factor3
larc	0.93156	0.08608	0.03978
burg	0.85365	0.25883	0.29980
auto	0.70065	-0.00698	0.47357
murd	-0.02168	0.94429	0.11885
assa	0.25180	0.87544	0.27977
rape	0.60057	0.62270	0.19589
popu	0.10435	0.20183	0.88782
robb	0.42160	0.28523	0.73731

Similar to the unrotated analysis above, we can use the plots of the factors against each other to check for possible groupings of variables. In both plots it appears that larceny (G) and burglary (F) (mentioned as a possible grouping from preliminary analysis) could be considered a grouping. Other than those two, it doesn't appear any of the variables are grouped.





Specific Variance

The specific variance (found by subtracting the final communality from one) tells us how much of the variance for each variable is not explained by the common factors. A low specific variance is a good thing and indicates that most of the variance is explained by the model. In this case, murder, robbery, assault, burglary, and larceny all have low specific variances. Rape has a semi-low variance and hence has a decent amount of its variance explained by the common factors. Population and auto theft have high specific variances and don't have much of their variances explained by the common factors.

$$\psi = \begin{bmatrix} .5213 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .1922 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .3078 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .1997 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .0921 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .1454 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .1431 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .4073 \end{bmatrix}$$

Final Communality Estimates and Variable Weights		
Total Communality: Weighted = 34.562442 Unweighted = 5.991160		
Variable	Communality	Weight
popu	0.47865611	1.9182300
murd	0.80784535	5.2042147
rape	0.69218129	3.2487526
robb	0.80034269	5.0083819
assa	0.90792963	10.8612704
burg	0.85460091	6.8773234
larc	0.85692758	6.9891435
auto	0.59267659	2.4551250

Conclusion

After analyzing the crime data for states in 1985, we have come to a couple of different conclusions. First, when analyzing the data based on the eight variables chosen, three factors are necessary to fully explain enough of the variance for the model. These factors, in any order, tend to describe the overall crimes, bodily harm crimes, and theft crimes. Secondly, it would seem that in both PCA and FA auto theft and larceny and rape and burglary could be considered a grouping. Rape and burglary can be verified by their high correlation (.68) but auto theft and larceny (.55) are a somewhat surprising grouping. Lastly, if determining where to live in 1985, the places with the lowest overall crime would be North Dakota, Maine, and New Hampshire. If one is simply trying to avoid theft, the best places to live would be Colorado, Arizona, and New Mexico.

Appendix A

Covariance Matrix, DF = 49

	land	popu	murd	rape	robb	assa	burg	larc	auto
land	7816047096	32214278	83176	244806	-165593	976516	2159505	15889030	1453429
popu	32214278	25694560	5309	15571	288099	147326	784325	829508	380331
murd	83176	5309	15	15	120	213	384	177	84
rape	244806	15571	15	54	370	349	1804	3133	646
robb	-165593	288099	120	370	8316	3501	20486	28235	11232
assa	976516	147326	213	349	3501	4647	12816	15325	4496
burg	2159505	784325	384	1804	20486	12816	130357	205309	50456
larc	15889030	829508	177	3133	28235	15325	205309	503858	78606
auto	1453429	380331	84	646	11232	4496	50456	78606	39844

Appendix B: Code

```
dm "output;clear;log;clear";
Title1 'Stock-Price Data (Table 8.4)';

GOptions Reset=ALL;
ODS PDF File="F:\STAT550\Take Home Midterm\Prelim.pdf"; *this will make a pdf output file;
ODS Listing Close;
GOptions NoPrompt Vsize=6 Hsize=6 Horigin=1.2 Vorigin=2.5 FText=SwissX FTitle=SwissX HText=1
HTitle=1;

/* US CRIME DATA
 This data consist of measurements of 50 states for 12 variables.
 It states for 1985 the reported number of crimes in the 50 states
 classified according to 7 categories (X4-X10)

X1: State      X2: land area (land) X3: population 1985 (popu) X4: murder (murd)
X5: rape X6: robbery (robb) X7: assault (assa) X8: burglary (burg)
X9: larceny (larc) X10: auto theft (auto) X11: US State region number (reg)
X12: US State division number (div)*/

DATA CRIME;
INPUT state $ land popu murd rape robb assa burg larc auto reg div;
```

```
DATALINES;
ME      33265  1164   1.500   7       12.600  62   562         1055    146  1    1
NH      9279   998            2               6       12.100  36   566     929     172  1   1
VT      9614   535            1.300   10.300  7.600    55   731     969     124  1    1
MA      8284   5822   3.500   12      99.500  88   1134    1531    878  1    1
RI      1212   968    3.200   3.600   78.300  120 1019    2186    859  1    1
CT      5018   3174   3.500   9.100   70.400  87  1084    1751    484  1    1
NY      49108  17783  7.900   15.500  443.300 209 1414    2025    682  1    2
NJ      7787   7562   5.700   12.900  169.400 90  1041    1689    557  1    2
PA      45308  11853  5.300   11.300  106     90  594     1001    340  1    2
OH      41330  10744  6.600   16      145.900 116 854     1944    493  2    3
IN      36185  5499   4.800   17.900  107.500 95  860     1791    429  2    3
IL      56345  11535  9.600   20.400  251.100 187 765     2028    518  2    3
MI      58527  9088   9.400   27.100  346.600 193 1571    2897    464  2    3
WI      56153  4775   2       6.700   33.100 44   539     1860    218  2    3
MN      84402  4193   2       9.700   89.100 51   802     1902    346  2    4
IA      56275  2884   1.900   6.200   28.600  48   507     1743    175  2    4
MO      69697  5029   10.700  27.400  200.800 167 1187    2074    538  2    4
ND      70703  685    0.500   6.200   6.500   21   286     1295    91   2    4
SD      77116  708    3.800   11.100  17.100  60   471     1396    94   2    4
NE      77355  1606   3       9.300   57.300  115 505     1572    292  2    4
KS      82277  2450   4.800   14.500  75.100  108 882     2302    257  2    4
DE      2044   622    7.700   18.600  105.500 196 1056    2320    559  3    5
MD      10460  4392   9.200   23.900  338.600 253 1051    2417    548  3    5
VA      40767  5706   8.400   15.400  92      143 806     1980    297  3    5
WV      24231  1936   6.200   6.700   27.300  84  389     774     92   3    5
NC      52669  6255   11.800  12.900  53      293 766     1338    169  3    5
SC      31113  3347   14.600  18.100  60.100  193 1025    1509    256  3    5
GA      58910  5976   15.300  10.100  95.800  177 900     1869    309  3    5
FL      58664  11366  12.700  22.200  186.100 277 1562    2861    397  3    5
KY      40409  3726   11.100  13.700  72.800  123 704     1212    346  3    6
TN      42144  4762   8.800   15.500  82      169 807     1025    289  3    6
AL      51705  4021   11.700  18.500  50.300  215 763     1125    223  3    6
MS      47689  2613   11.500  8.900   19      140 351     694     78   3    6
AR      53187  2359   10.100  17.100  45.600  150 885     1211    109  3    7
LA      47751  4481   11.700  23.100  140.800 238 890     1628    385  3    7
OK      69956  3301   5.900   15.600  54.900  127 841     1661    280  3    7
TX      266807 16370  11.600  21      134.100 195 1151    2183    394  3    7
MT      147046 826    3.200   10.500  22.300  75   594     1956    222  4    8
ID      83564  1005   4.600   12.300 20.500   86   674     2214    144  4    8
WY      97809  509    5.700   12.300  22      73   646     2049    165  4    8
CO      104091 3231   6.200   36      129.100 185 1381    2992    588  4    8
NM      121593 1450   9.400   21.700  66.100  196 1142    2408    392  4    8
AZ      114000 3187   9.500   27      120.200 214 1493    3550    501  4    8
UT      84899  1645   3.400   10.900  53.100  70   915     2833    316  4    8
NV      110561 936    8.800   19.600  188.400 182 1661    3044    661  4    8
WA      68138  4409   3.500   18      93.500  106 1441    2853    362  4    9
OR      97073  2687   4.600   18      102.500 132 1273    2825    333  4    9
CA      158706 26365  6.900   35.100  206.900 226 1753    3422    689  4    9
AK      591004 521    12.200  26.100  71.800  168 790     2183    551  4    9
HI      6471   1054   3.600   11.800  63.300  43   1456    3106    581  4    9
;
```

For Preliminary Analysis

```
/*Normality Test for Population*/
PROC UNIVARIATE DATA = crime NORMAL;
var popu;
RUN;

/*Normality Test for Murder*/
PROC UNIVARIATE DATA = crime NORMAL;
var murd;
RUN;

/*Normality Test for Rape*/
PROC UNIVARIATE DATA = crime NORMAL;
var rape;
RUN;
```

```
/*Normality Test for Robbery*/
PROC UNIVARIATE DATA = crime NORMAL;
var robb;
RUN;

/*Normality Test for Assault*/
PROC UNIVARIATE DATA = crime NORMAL;
var assa;
RUN;

/*Normality Test for Burglary*/
PROC UNIVARIATE DATA = crime NORMAL;
var burg;
RUN;

/*Normality Test for Larceny*/
PROC UNIVARIATE DATA = crime NORMAL;
var larc;
RUN;

/*Normality Test for Auto Theft*/
PROC UNIVARIATE DATA = crime NORMAL;
var auto;
RUN;

/*Chi-Square Probability Plot (QQ Plot)*/
proc princomp std out=pcresult;
var murd rape robb assa burg larc auto;
run;

data mahal;
set pcresult;
dist2=uss(of prin1-prin7);
run;

proc sort;
by dist2;
run;

data plotdata;
set mahal;
prb=(_n_ -.5)/51;
chiquant=cinv(prb,7);
run;

title1;
title1 bold "Chi Square Probility Plot";

proc gplot;
plot dist2*chiquant;
run;

/*Correlation and Covariance Matrix*/
proc corr data=crime cov noprob;
var land popu murd rape robb assa burg larc auto;
run;

/*Correlation Matrix Without Land Area*/
proc corr data=crime;
var popu murd rape robb assa burg larc auto;
run;
```

For Principal Component Analysis

```
/*Simple Statistics*/
proc means data=crime maxdec=2;
var popu murd rape robb assa burg larc auto;
run;

/*PC Analysis on Covariance Matrix*/
```

```
Title1 "Analysis of Correlation Matrix for GSP Share Data";
Proc PrinComp Data=crime Out=PrinComp cov; *default is PC from Correlation matrix;
 Var popu murd rape robb assa burg larc auto;
 ODS output Eigenvalues=eigenval;
 ODS output Eigenvectors=eigenvec;
Run;

/*PC Analysis on Correlation Matrix*/
Title1 "Analysis of Correlation Matrix for GSP Share Data";
Proc PrinComp Data=crime Out=PrinComp; *default is PC from Correlation matrix;
 Var popu murd rape robb assa burg larc auto;
 ODS output Eigenvalues=eigenval;
 ODS output Eigenvectors=eigenvec;
Run;

Proc Gplot data=Eigenval;
plot Eigenvalue*Number; run;

Proc Gplot data=Eigenvec;
plot Prin2*Prin1 Prin3*Prin1 Prin3*Prin2 / vref=0 href=0;
Symbol1 C=Black V=Dot I=None PointLabel=("#Variable"); run;

Proc GPlot Data=PrinComp;
 Plot Prin2*Prin1=1 Prin3*Prin1=1 Prin3*Prin2=1/  VAxis=Axis1 HAxis=Axis2 Frame;
 Axis1  Order=(-7.0 To 7.0 By 1.0);
 Axis2  Order=(-7.0 To 7.0 By 1.0);
 Symbol1 C=Black V=Dot I=None PointLabel=("#State");
Run;
Quit;
```

For Factor Analysis

```
/*Use ML level to get the Chi-Square test for adequacy*/
  PROC FACTOR METHOD=ML NFACT=3 ROTATE=VARIMAX S C EV RES REORDER DATA=crime
  SCORE OUT=SCORES2 HEYWOOD;
  VAR popu murd rape robb assa burg larc auto;
  RUN;

/*Factor Analysis for Unrotated Communalities and Factor Loadings*/
PROC FACTOR METHOD=PRINCIPAL SCREE;
 VAR popu murd rape robb assa burg larc auto;
 RUN;

 /*Factor Analysis for Rotated Communalities and Factor Loadings*/
 PROC FACTOR METHOD=prin NFACT=3 ROTATE=VARIMAX S C EV RES REORDER DATA=crime
  SCORE OUT=SCORES1 PREPLOT PLOT;
  VAR popu murd rape robb assa burg larc auto;
  RUN;
```