1

US Crime 데이터는 1985년 미국의 각 주에서 발생한 범죄의 수를 나타낸 데이터이다. Murder, Robbery, Assault, Burglary, Larceny, Auto theft 이렇게 총 7개의 범죄가 기록되어 있고 이 변수들이 Factor Analysis(FA)의 대상이 된다. [그림1]을 보면 이 7개의 변수 사이에 correlation이 존재한다. 특히 Murder와 Assault의 correlation은 0.81로 관련이 높아 보인다. 따라서 이 변수들에 포함되어 있는 잠재적인 공통 요인을 뽑아내기 위해 Factor Analysis(FA)를 진행한다. 보다 적은 수의 변수로 축약하면서 다중공산성 문제도 어느 정도 해결할 수 있다.

먼저 요인의 개수를 결정한다. [그림2]는 Eigenvalues가 기록되어 있는 Scree Plot으로 요인의 개수를 결정하는 데 쓰이는 것 중 하나이다. 보통 이 Scree Plot이 꺾이는 지점인 elbow 부분을 요인의 개수로 정한다. 이 경우 elbow 지점이 3이므로 3개의 요인으로 변수들을 설명하기로 한다.

요인의 개수를 정한 이후 본격적으로 FA를 진행한다. 이때 모두 표준화한 변수들을 사용한다. [그림3]은 Maximum Likelihood Method로 FA를 한 결과이다. 이때 해석의 편의성을 위해서 Varimax 방법으로 회전하였다. Factor Loadings는 각각의 요인을 구성하는 수식의 계수이다. 그리고 Communality는 Factor Loading를 제곱하여 합한 값으로 공통적인 요인으로 분해되는 정도를 나타내는 지표이다. Larceny의 Communality는 0.928이다. 즉 0.928만큼의 정보가 3개의 요인에 의해 설명된다는 것이다. [그림4]은 각 요인을 구성하는 Factor Loadings의 절대값을 나타낸 그래프이다. Factor1은 Murder을 제외한 변수들로 이루어져 있다. 그리고 Factor2는 Murder와 Assault가 큰 비중을 차지한다. 마지막으로 Factor3은 Robbery, Burglary, Auto theft가 큰 비중을 차지한다. [그림5]를 보면 보다 정확한 해석을 할 수 있다. Factor2 축을 보면 Murder와 Assault가 1에 가까운 값을 가진다. [그림1]의 correlation plot에서 Murder와 Assault의 correlation 값이 0.81로 매우 큰 값을 가졌었는데 이렇게 공통 요인으로 묶으면서 다중공산성을 어느 정도 해결할 수 있다. Factor3 축을 보면 Auto theft, Robbery, Burglary가 큰 Factor Loading 값을 가진다. 그리고 Factor1 축을 보면 Larceny가 1에 가까운 값을 가진다. 따라서 Factor1은 "재산상 손해와 관련된 절도", Factor2는 "살인, 폭행 등 폭력적 범죄", Factor3은 "폭력성을 포함한 절도"라고 해석할 수 있다.

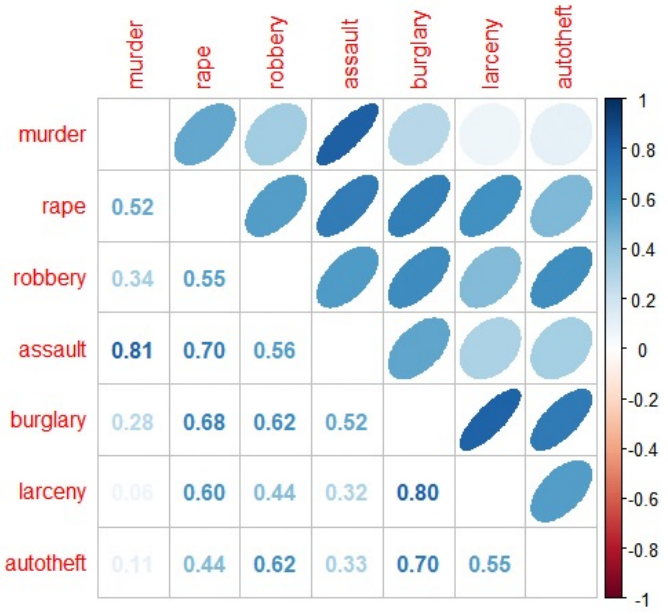
[그림3]의 표를 보면 모든 변수의 Communality가 0.6이 넘는다. 세 factor가 변수들을 잘 설명하고 있다는 뜻이다. 그런데 Rape를 제외한 모든 변수들은 0.6보다 큰 Factor Loading을 가지며, 하나의 요인에서 다른 요인보다 상대적으로 큰 loading값을 가진다. 예를 들어 Murder은 Factor2의 loading 값이 0.879이고 Larceny은 Factor1의 loading 값이 0.91이다. 이 값은 다른 요인의 loading 값보다 확연하게 큰 값이다. 하지만 Rape의 loading 값은 0.527, 0.563, 0.305로, Factor1과 Factor2에서 loading 값의 차이가 별로 없다. 따라서 Rape를 Factor1와 관련된 요소로 볼지, 혹은 Factor2와 관련된 요소로 볼지 애매하다. Rape가 0.688이라는 큰 Communality를 가지고 있지만 이러한 이유로 빼고 해석하는 것이 좋아 보인다.

 # 2

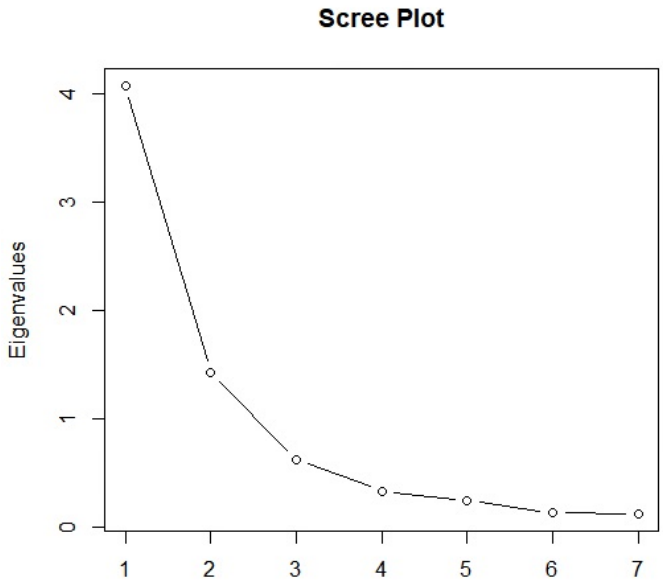
비슷한 방식으로 US health 데이터를 가지고 FA를 진행한다. US health 데이터는 미국의 주마다 사망 원인별 수를 기록한 데이터이다. Accident, Cardiovascular, Cancer, Pulmonary, Pneumonia Flu, Diabetes, Liver 이렇게 7개의 사망 원인이 있으며 이것들의 공통요인을 찾아본다. [그림6]은 이 변수들 사이의 correlation을 그린 그래프이다. 변수들 사이에 correlation이 있어 다중공산성 문제를 해결할 필요가 있어 보인다. 특히 Cardiovascular와 Cancer의 correlation 값이 0.91로 매우 큰 값이 나온다. Cardiovascular와 Cancer에 연관성이 있을 것으로 예상되며 FA 분석을 통해 확인한다. 이번에도 3개의 요인으로 FA를 한다.

이번에도 FA를 할 때 표준화된 변수를 사용하였으며 해석의 편리성을 위해 rotation을 한 후 진행하였다. [그림7]은 Maximum Likelihood Method로 FA를 한 결과이다. Acc와 pul 변수는 Communality 값이 약 0.3으로 매우 작게 나와서 제거시키는 것이 좋아 보인다. [그림8]은 각 요인을 구성하는 Factor Loading의 절댓값을 나타낸 그래프이다. Factor1은 card, canc, diab가 큰 비중을 차지한다. 그리고 Factor2는 liv가 매우 큰 비중을 차지한다. 그리고 Factor3은 pnue가 가장 큰 비중을 차지하고 canc, pul, card가 어느 정도 큰 비중을 차지한다. [그림9]를 보면 보다 정확한 해석을 할 수 있다. Factor1 축을 보면 card, canc, diab가 1에 가까운 값을 가진다. 이 세 질병의 원인은 낮은 혈당수치라고 한다. 따라서 Factor1을 "혈당수치가 낮아서 생기는 병"이라고 해석할 수 있다. 그리고 Factor2 축을 보면 liv가 1에 가까운 값을 가지고 나머지 변수들은 0에 가까운 값을 가진다. 따라서 Factor2를 "간과 관련된 질병"이라고 할 수 있다. 마지막으로 Factor3을 보면 pnue가 1에 가까운 값을 가지고 pul과 약한 양의 상관관계를 갖는다. 따라서 Factor3을 "폐와 관련된 질병"이라고 할 수 있다.

[그림10]은 Factor Score을 예측값을 나타낸 그래프이다. Factor Score은 요인별 상관성을 나타내는 지표이다. Factor Score가 높으면 해당 요인에 크게 영향을 받는 것이다. Factor1 축을 보면 RI 주가 큰 값을 가지고 있고 알레스카 주와 와이오밍 주, 유타 주, 콜로라도 주는 작은 값을 가지고 있다. RI 주는 혈당수치가 낮아 죽는 사람들이 많고 알레스카 주, 와이오밍 주, 유타 주, 콜로라도 주는 상대적으로 건강한 주라고 할 수 있다. Factor2 축을 보면 SD 주가 간 관련 질병으로 죽는 사람들이 많다. Factor3 축을 보면 캘리포니아 주, 뉴욕 주, 플로리다 주는 폐과 관련된 질병으로 죽는 사람들이 많다. 반대로 하와이 주, SD 주는 폐과 관련된 질병으로 죽는 사람이 적다. 하와이 주는 Factor2 축과 Factor3 축에서 모두 작은 값을 갖는다. 하와이 주는 다른 주에 비해 상대적으로 건강하다고 할 수 있다. 그리고 [그림10]의 가운데 그래프인 Factor2 vs Factor3을 보면 Midwest에 있는 주들이 간과 관련된 질병으로 죽는 사람이 많고 폐과 관련된 질병으로 죽는 사람이 적다. 그리고 South에 있는 주들은 간과 관련된 질병으로 죽는 사람도 적고 폐와 관련된 질병으로 죽는 사람도 적다. 미국 남부는 시골 깡촌의 이미지가 있는데 이러한 자연 환경 때문에 상대적으로 건강한 것 같다.



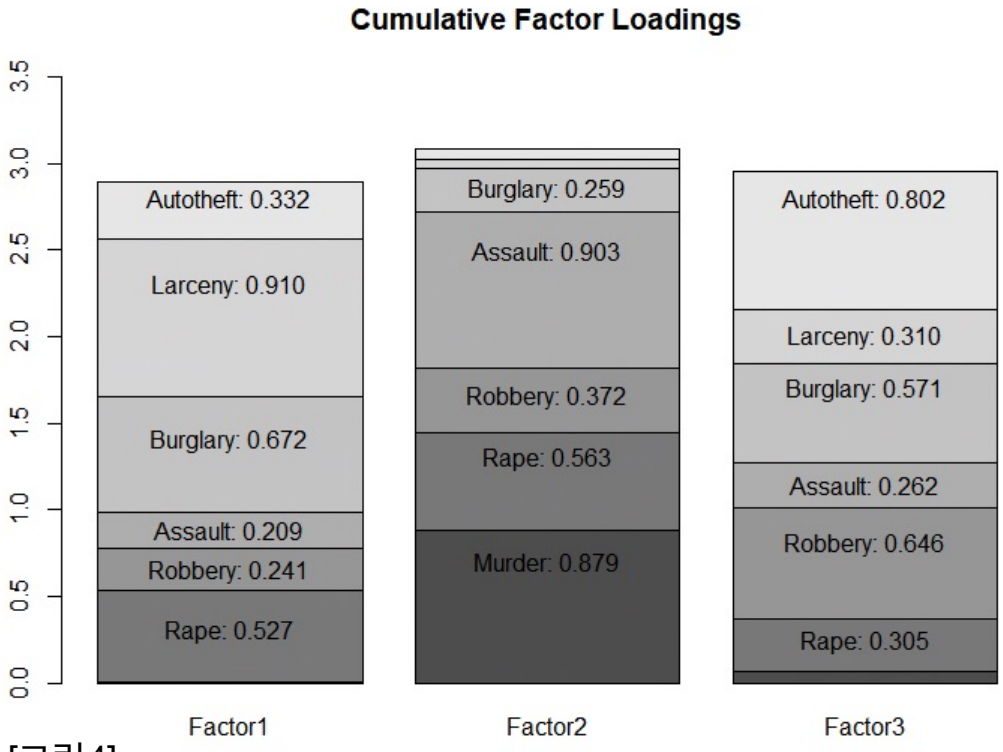
[그림1]



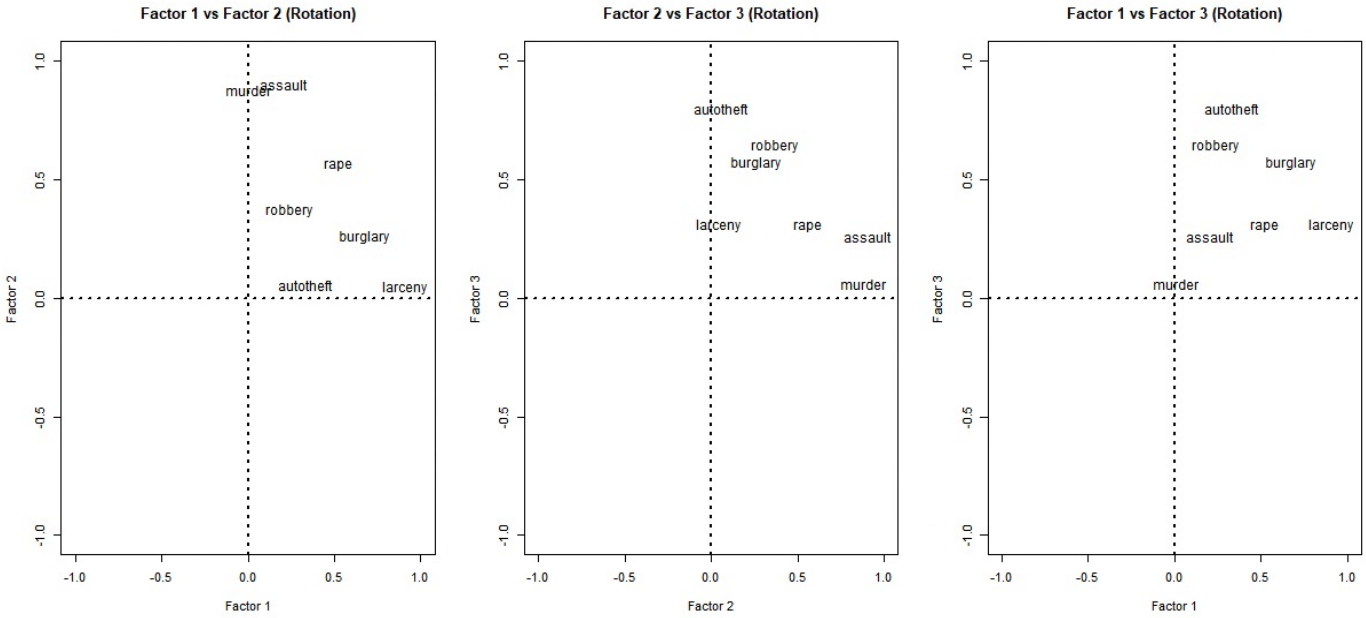
[그림2]

	Factor1	Factor2	Factor3	Communality	Specific Variances
Murder	0.006	0.879	0.063	0.777	0.223
Rape	0.527	0.563	0.305	0.688	0.312
Robbery	0.241	0.372	0.646	0.614	0.386
Assault	0.209	0.903	0.262	0.928	0.072
Burglary	0.672	0.259	0.571	0.844	0.156
Larceny	0.91	0.048	0.31	0.928	0.072
Autotheft	0.332	0.057	0.802	0.757	0.243

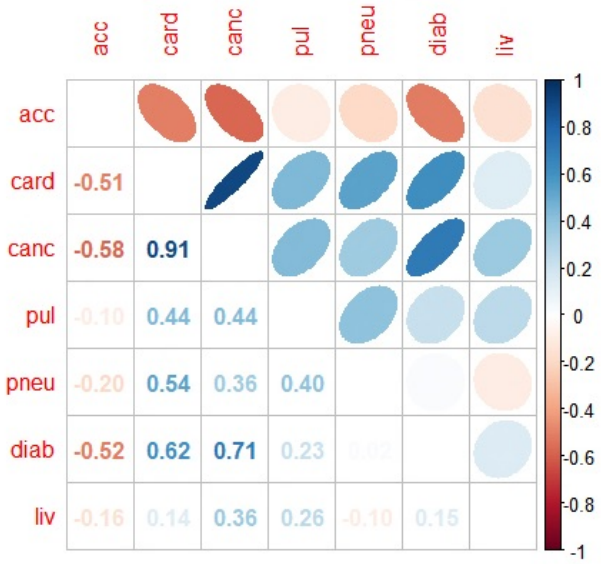
[그림3]



[그림4]



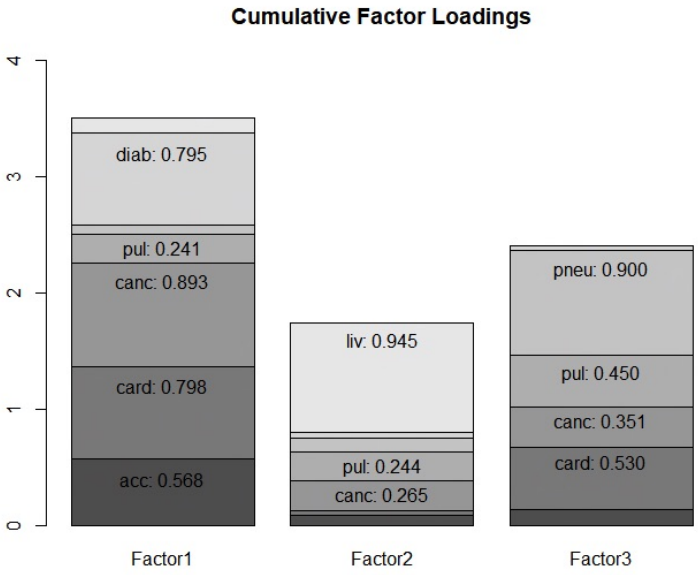
[그림5]



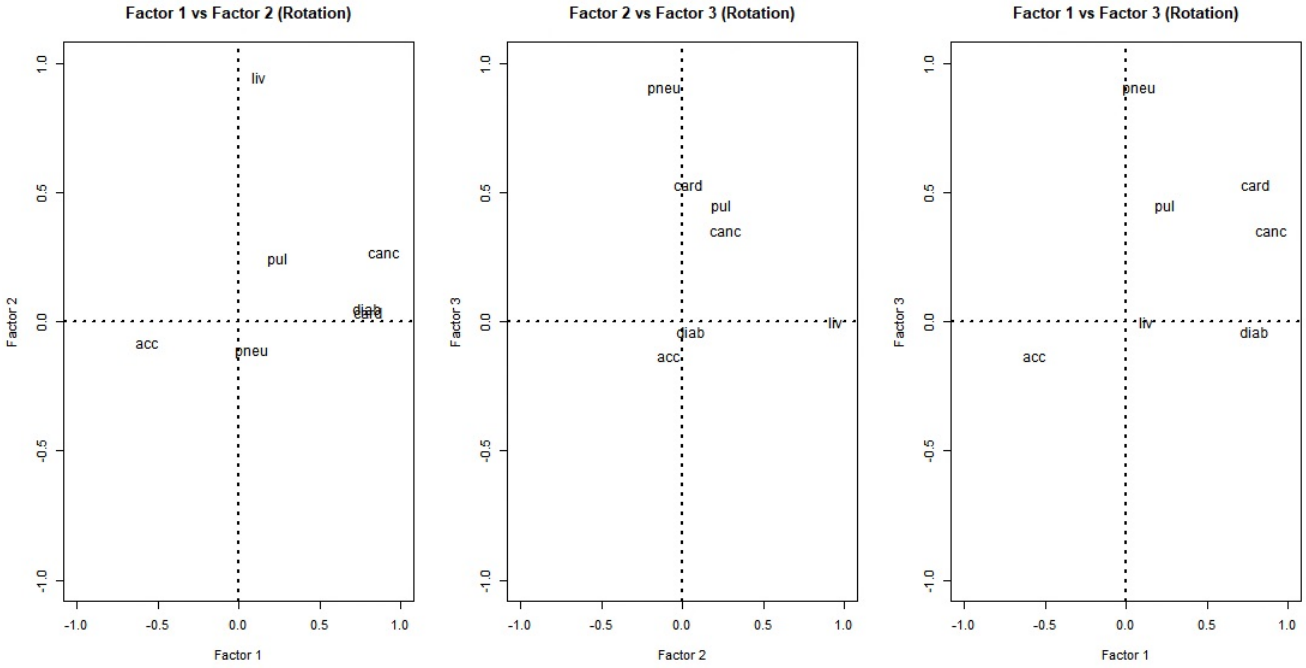
[그림6]

	Factor1	Factor2	Factor3	Communality	Specific Variances
acc	-0.568	-0.084	-0.136	0.348	0.652
card	0.798	0.037	0.53	0.919	0.081
canc	0.893	0.265	0.351	0.99	0.01
pul	0.241	0.244	0.45	0.32	0.68
pneu	0.082	-0.116	0.9	0.83	0.17
diab	0.795	0.051	-0.037	0.636	0.364
liv	0.125	0.945	0.003	0.909	0.091

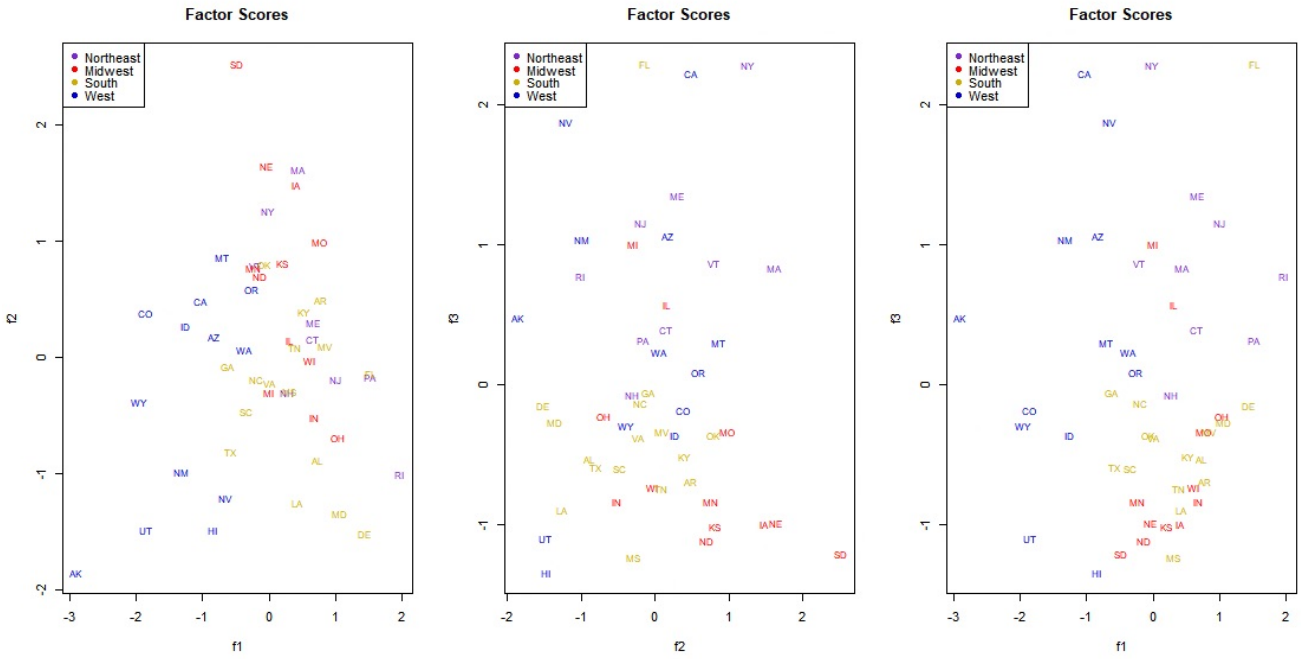
[그림7]


Cumulative Factor Loadings
diab: 0.795
pul: 0.241
canc: 0.893
card: 0.798
acc: 0.568
liv: 0.945
pul: 0.244
canc: 0.265
pneu: 0.900
pul: 0.450
canc: 0.351
card: 0.530
Factor1
Factor2
Factor3

[그림8]


Factor 1 vs Factor 2 (Rotation)
Factor 2 vs Factor 3 (Rotation)
Factor 1 vs Factor 3 (Rotation)
liv
pul
canc
acc
pneu
card
diab
Factor 1
Factor 2
Factor 3

[그림9]


Factor Scores
Northeast
Midwest
South
West
f1
f2
f3

[그림10]