Multivariate Statistics

-final project-

201811530 통계학과 임도현

Content Page

다변량통계학(1) Term Project

201811530 통계학과 임도현

1. Goal of Research

고객의 정보에 따른 의료보험비가 주어진 데이터를 분석해보고자 한다. 다변량 통계학 강의를 수강하며 배웠던 다변량 자료 분석 기법을 이번 데이터에 적용해보고 유의미한 결과를 도출하는 방법들을 습득하는 것이 이번 프로젝트의 목표이다. 이번 분석은 의료보험비와 고객들의 정보가 어떤 관계를 갖는지에 초점을 두고 분석한다.

2. Data Description

데이터 출처 : https://www.kaggle.com/datasets/mirichoi0218/insurance
(Book : Machine Learning with R by Brett Lantz)

데이터 샘플 예시)

	age	sex	bmi	children	smoker	region	charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622

1338 rows * 7 columns

변수설명
- age : 고객의 나이 (이산형 변수)
- sex : 고객의 성별 (명목형 변수)
- bmi : 고객의 bmi지수, 즉 비만 지수 (연속형 변수)
- children : 고객의 자녀 수 (이산형 변수)
- smoker : 고객의 흡연 여부 (명목형 변수)
- region : 고객의 거주 지역 (명목형 변수)
- charges : 고객의 의료보험비 (연속형 변수)

변수에 대한 자세한 탐색은 MDA 과정에서 실시한다.

3. Data Analysis

1) MDA(Multivariate Data Analysis)

데이터에 결측치가 존재하면 분석에 있어 문제가 생길 수 있으므로, 결측치의 존재여부를 먼저 확인한다.

<각 변수별 결측치의 수>

age	sex	bmi	children	smoker	region	charges
0	0	0	0	0	0	0

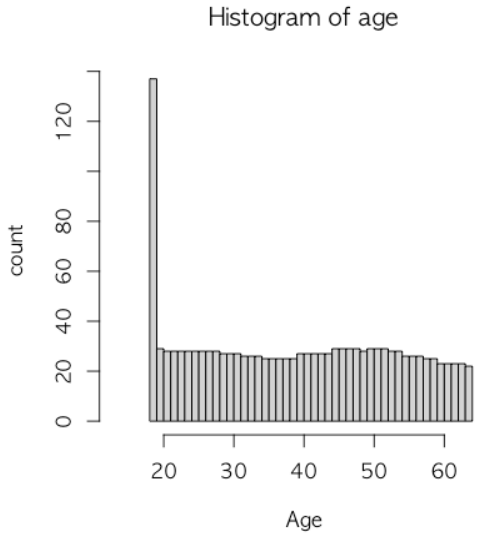
결측치는 존재하지 않는다. 이제 데이터의 변수들을 구체적으로 관찰해보자. 이를 위해 시각화를 활용한다.

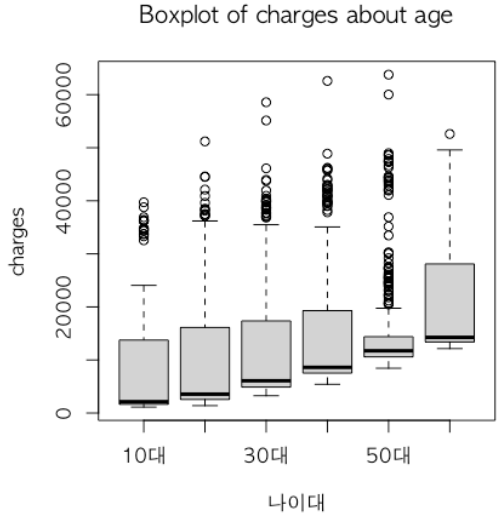
- **Age(고객의 나이)**

<고객 나이의 요약통계량>

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.00	27.00	39.00	39.21	51.00	64.00

최연소 고객의 나이가 18세이고, 최고령 고객의 나이가 64세임을 알 수 있다. 조금 더 자세한 정보를 위해 나이 변수의 분포와 고객들의 나이와 의료보험비가 어떤 관계를 갖는지를 시각화 해보자.





주어진 데이터에 10대 후반의 고객들이 다른 나이대의 고객들에 비해 상당히 많이 존재함을 알 수 있다. 또한 평균적으로 나이가 많을수록 의료보험비의 인상이 있다. 주목할 점은 50대의 의료보험비의 분산이 비교적 다른 그룹에 비해 작으며 60대의 의료보험비는 이상치가 거의 존재하지 않음을 알 수 있다.

- **Sex(고객의 성별)**

<성별에 따른 고객 수>

female	male
662	676

남녀의 수가 골고루 분포되어 있음을 알 수 있다. 성별과 의료비의 관계를 Boxplot으로 확인해보자.

의료보험비의 분산이 남성이 여성에 비해 크게 나타난다. 또한 여성과 남성의 의료보험비는 평균적으로는 비슷하지만, 상위값이 남자들이 더 많이 분포되어 있는 것을 알 수 있다. 즉 남성들이 여성에 비해 의료보험비를 더 많이 내는 것으로 해석 가능하다. 건강상태와 의료보험비는 매우 큰 관계를 갖는다. 우리의 데이터에서는 bmi지수와 흡연여부가 건강상태에 관련된 지표이므로, 두 개의 변수의 남녀 비율을 확인해보자.

<흡연자의 수>

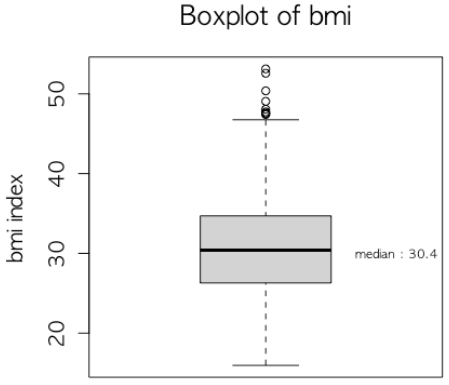
```
female   male
   115    159
```

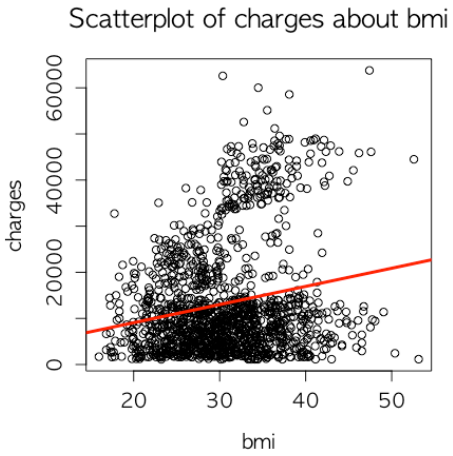
<평균 BMI지수 보다 높은 BMI지수를 갖는 고객의 수>

```
female   male
   306    340
```

흡연자의 수가 여자보다 남자가 더 많이 존재한다. 흡연자의 경우 비흡연자 보다 건강상태가 나쁘게 측정될 가능성이 높으므로, 남성의 의료보험비가 더 비싼 경향을 보인다고 추측할 수 있다. 비만으로 분류될 수 있는 비율은 남자가 더 많이 존재한다. 비만이 합병증을 유발하기 때문에 건강상태가 나쁘게 측정될 가능성이 높으므로, 남성의 의료비가 더 비싼 경향을 보인다고 추측할 수 있다.

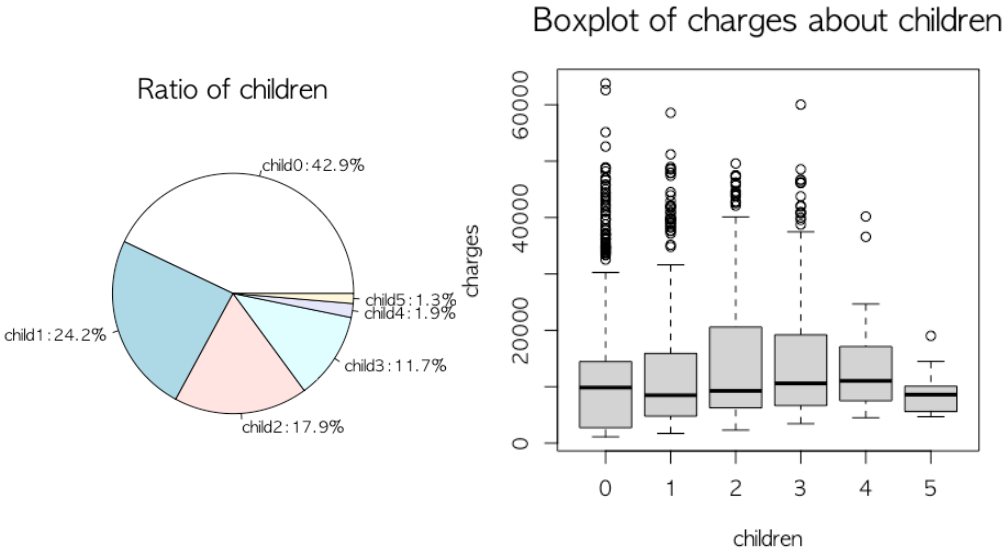
- BMI(고객의 BMI 지수)





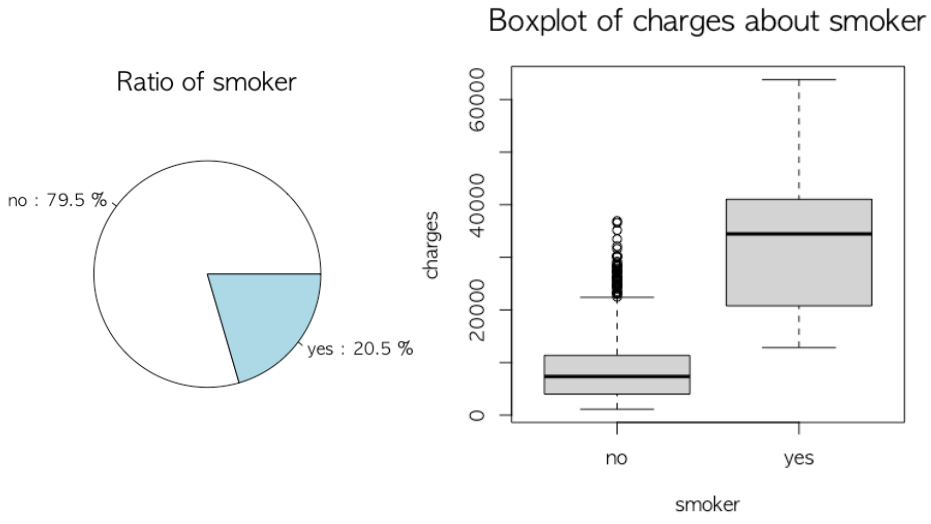
중위수가 30.4로 고객들의 BMI 지수가 꽤 높아 보인다. 또한 이상치로 간주되는 초고도비만의 고객들도 존재한다. 또한 bmi 지수가 높아질수록 의료보험비가 높아짐을 알 수 있다. 비만이 각종 합병증을 유발하므로 의료보험비가 높게 측정되는 것으로 추측할 수 있다.

- Children(고객의 자녀 수)



자녀를 가지지 않은 비율이 약 43%로 가장 많이 있음을 알 수 있다. 아마도 10대 후반의 고객의 수가 다른 나이에 비해서 많기 때문이라고 유추할 수 있다. 자녀의 수에 따른 의료비의 분명한 관계는 찾아보기 힘들다.

- Smoker(고객의 흡연 여부)



비흡연자의 비율이 80%로 매우 높은걸 알 수 있다. 흡연자의 경우 의료보험비가 비흡연자에

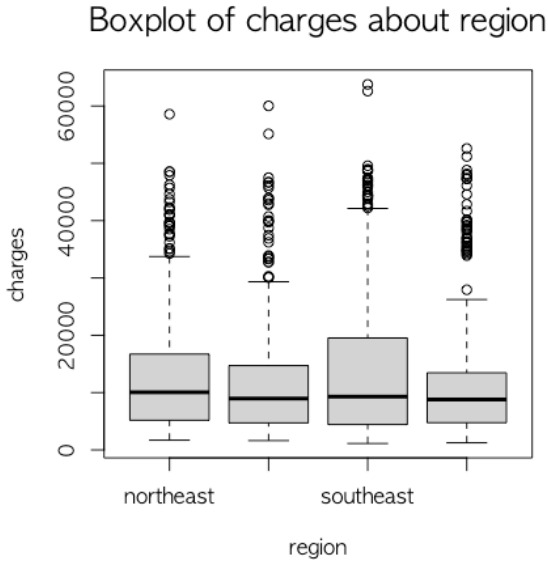
비해 매우 높게 측정됨을 알 수 있다. 매우 직관적인 결과이다.

- Region(고객의 거주 지역)

<거주지역에 따른 고객의 수>

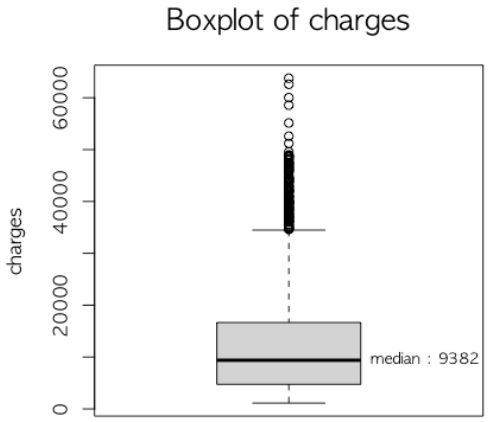
northeast	northwest	southeast	southwest
324	325	364	325

북동, 북서, 남동, 남서 지방의 수가 골고루 분포되어 있음을 알 수 있다. 고객의 거주 지역별 의료보험비의 관계를 확인해보자.



지역별로 의료보험비가 어떤 관계를 갖는지 명확한 해석은 어려워 보인다.

- Charges(고객의 의료보험비)



의료비의 중위수가 약 9382정도임을 알 수 있다. 하지만 이상치가 매우 많이 존재하며, 그래

프의 y축 스케일이 매우 차이 나는 것을 보면 고객별로 의료비가 매우 상이함을 알 수 있다.

본 분석의 목표에 알맞게 의료보험비와 다른 변수들 간의 상관관계를 살펴보자.

<연속형 변수들의 상관계수 행렬>

	age	bmi	children	charges
age	1.000	0.109	0.042	0.299
bmi	0.109	1.000	0.013	0.198
children	0.042	0.013	1.000	0.068
charges	0.299	0.198	0.068	1.000

의료비에 가장 큰 영향을 미치는 변수는 나이임을 알 수 있다. 0.299의 양의 상관계수를 갖는다. 다음으로 범주형 자료들의 독립성을 확인해보자.

<범주형 자료들의 Chi-square independence test 결과>

```
Number of cases in table: 1338
Number of factors: 4
Test for independence of all factors:
        Chisq = 93.74, df = 85, p-value = 0.2422
        Chi-squared approximation may be incorrect
```

chi-sqaure independence test 결과 p-value가 0.2422로서 범주형 자료들끼리의 독립성이 존재한다는 귀무가설을 기각하기 어려워 보인다. 즉 범주형 자료들은 서로가 독립성을 만족한다고 결론 지을 수 있다. 다음으로 관심 있는 변수인 의료보험비에 따른 범주형 자료들의 독립성을 검정해보자.

<범주형 자료들의 의료보험비에 따른 Chi-square independence test 결과>

```
Call: xtabs(formula = charges ~ sex + children + smoker + region, data = X)
Number of cases in table: 17755825
Number of factors: 4
Test for independence of all factors:
        Chisq = 3079385, df = 85, p-value = 0
```

의료비에 따른 범주형 변수들의 독립성 검정에서 p-value가 0에 매우 가까우므로 귀무가설인 독립성 존재를 기각할 수 있다. 즉 의료보험비에 대하여 범주형 변수들은 종속적인 관계를 갖음을 알 수 있다.

이제 변수들의 다변량 정규성을 확인해보자. Mardia normality test를 검정에 사용한다.

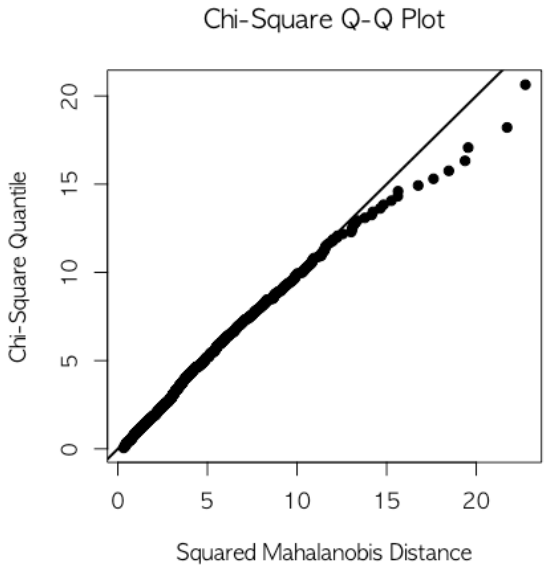
<Mardia Multivariate Normality Test 결과>

	Test	Statistic	p value	Result
1	Mardia Skewness	1007.10236610341	1.19383926312384e-200	NO
2	Mardia Kurtosis	0.622126891577761	0.533858433653274	YES
3	MVN	<NA>	<NA>	NO

Mardia normality test 결과 첨도(Kurtosis)부분에서는 정규성을 만족하지만 왜도

(Skewness)부분에서 p-value가 매우 낮아 정규성을 만족하지 않아 결론적으로 다변량 정규성을 만족하지 못함을 알 수 있다.

<Q-Q plot 시각화 결과>



Q-Q plot상으로 Mahalanobis Distance가 큰 값들이 정규성을 갖지 못하는 것에 영향을 주는 것 처럼 보인다.

2) PCA(Principal Component Analysis)

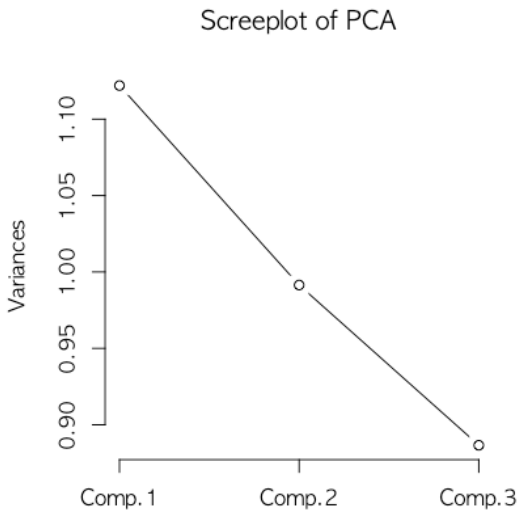
먼저, 이번 프로젝트에서는 상관행렬을 주성분 분석에 사용한다. 이유는 고객 정보의 데이터의 변수들이 각각 단위가 다르므로, 공분산 행렬을 주성분 분석에 사용하면 분산의 차이가 커질 수 있기 때문에 단위의 차이가 존재할 때 유용한 상관행렬을 사용한다. 또한 마찬가지의 이유로 주성분 점수 행렬을 구성하기 위해서 표준화 자료행렬을 사용한다.

다음으로 의료보험비 데이터는 연속형 자료와 범수형 자료가 공존하기 때문에 연속형 자료에 대해서만 주성분 분석을 시행하기로 한다. 그리고 정보에 대한 의료보험비를 분석하기 위해 주성분 분석에 의료보험비 변수는 제외하기로 한다.

마지막으로 개체 수(고객의 수)가 많기 때문에 Biplot등의 시각화에 어려움을 겪는다. 따라서 시각화 과정에는 데이터 중 일부(약 10%)만 추출하여 시각화한다. 축소된 데이터의 추출은 계통표집법으로 행의 인덱스를 10단위로 절단하여 추출한다.(예시 : 1, 11, 21, ..., 1331)

<상관행렬을 이용한 연속형 자료들의 주성분 분석 결과 요약>

```
Importance of components:
                          Comp.1    Comp.2    Comp.3
Standard deviation     1.0592327 0.9957106 0.9415873
Proportion of Variance 0.3739913 0.3304799 0.2955289
Cumulative Proportion  0.3739913 0.7044711 1.0000000
```



상관행렬을 이용한 연속형 자료들의 주성분 분석 결과를 살펴 보았을 때, 주성분을 2개로 선택하면 설명력이 약 70%로서 좋은 결과를 보인다고 할 수 있다. 하지만 Scree Plot을 살펴보면, Elbow point가 딱히 존재하지 않는다. 이유는 각각의 설명력이 모두 30% 즈음으로 비슷한 설명력을 갖기 때문이다. 우선은 70%의 설명력을 갖는 주성분의 개수 2개를 채택하기로 한다. 이후에 주성분 분석을 통한 결과를 더 살펴보자.

<주성분 점수(PC scores) 결과 샘플>

	Comp.1	Comp.2
1	-1.5722980	0.5685205
11	-1.4564303	0.5050779
21	1.3151967	1.2773425
31	-0.6005244	1.0349169
41	-1.4650409	0.5214845
51	-0.7949422	1.0133433
61	0.3222037	-1.6429160
71	-1.5151841	0.4308523
81	-0.8330553	-0.1942967
91	-0.2335404	1.2560534

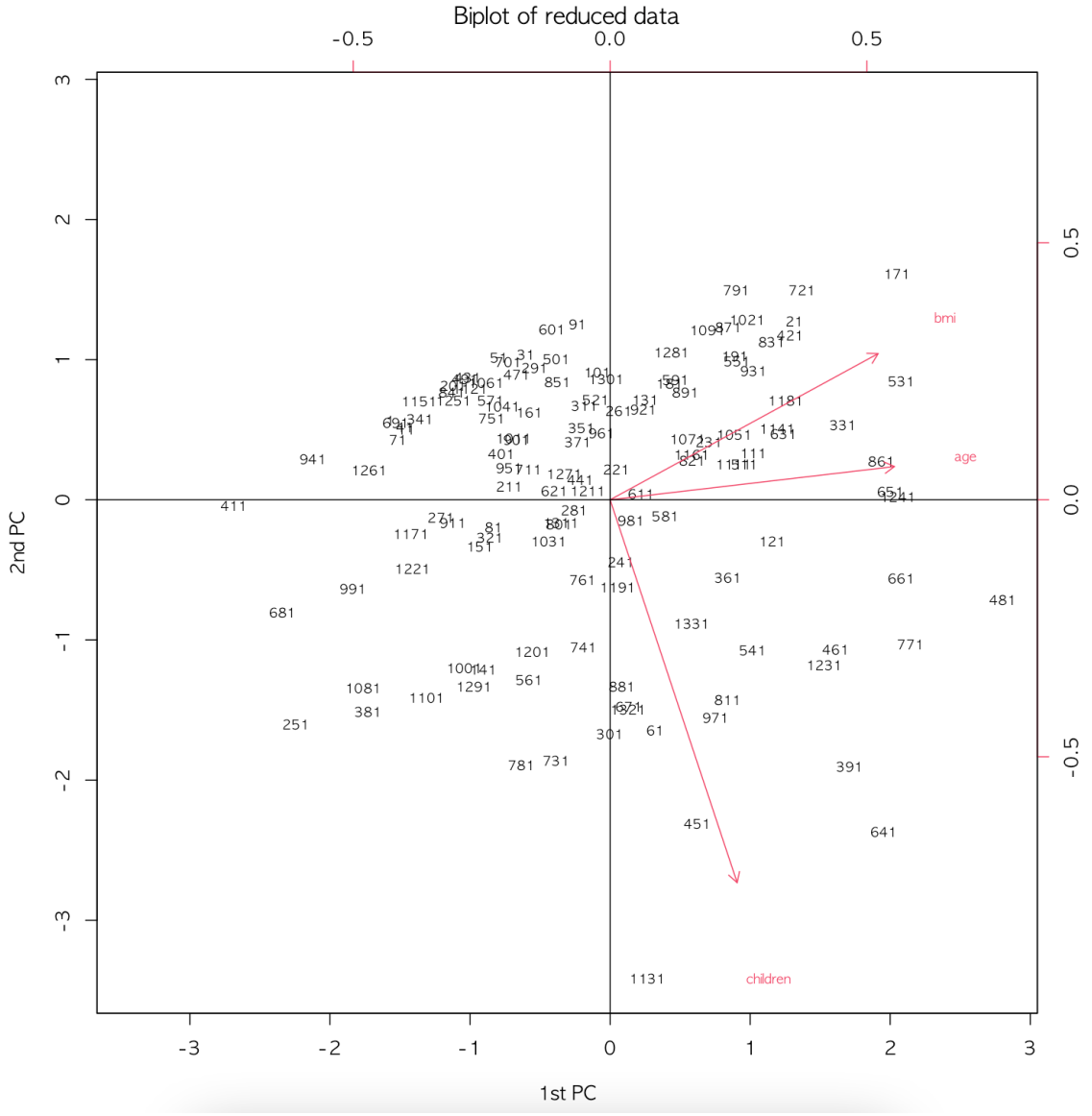
<주성분 계수(PC loadings) 결과>

	Comp.1	Comp.2
age	0.6920284	0.08086901
bmi	0.6523048	0.35555968
children	0.3091848	-0.93114849

주성분 점수의 경우 데이터가 많으므로 모든 데이터의 값을 문서에서 확인하기는 어렵다. 따라서 차후에 Biplot을 활용하여 자세히 분석해보기로 하고 주성분 계수를 분석해보자.

제1주성분을 결정하는 데에 있어 모든 변수가 양의 방향의 영향을 주었다. 특히 Age(고객의 나이)변수와 BMI(고객의 BMI지수)변수가 큰 영향을 주었다. 주목할 점은 Children(고객의 자녀수)는 다른 변수들에 비해서는 특출한 영향을 주지 않았다는 것을 알 수 있다. 앞서 MDA 과정에서 고객의 자녀보유수와 의료보험비의 관계를 찾지 못하였고, 나이와 의료보험비 또는 BMI지수와 의료보험비의 관계를 쉽게 찾을 수 있었는데 주성분 계수가 이를 증명한다고 볼 수 있다.
제2주성분을 결정하는 데에는 BMI지수가 영향을 주었지만 큰 영향을 주었다고 보긴 어렵다. 특히 나이변수의 경우에는 거의 영향이 없다라고 해도 무방하다. 주목할 점은 나머지 변수들

은 양의 방향으로 큰 영향을 주지 못하였지만 자녀보유수 변수의 경우는 매우 큰 크기의 음의 방향의 영향을 주었음을 알 수 있다. 더욱 자세한 분석은 Biplot을 통해 분석해보자.



Biplot을 통해 확인할 수 있는 결과들을 살펴보자. 우선 Age(나이)변수가 제1주성분 축과 거의 평행에 가까운 각도를 이룬다. 앞서 제1 주성분 계수를 결정하는 데에 있어 나이 변수가 큰 양의 방향의 역할을 했음을 확인하였고 이것이 Biplot에 나타난 것 이라고 할 수 있다. Children(자녀 보유수)변수는 나이 변수보다는 덜 하지만 제2주성분 축과 평행에 가까운 각도를 이룬다. 이 또한 앞서 자녀 보유수 변수의 제2 주성분 계수가 음의 방향으로 매우 높게 나왔으므로 설명된다고 할 수 있다.

제 1사분면에 있는 고객들은 대체로 자녀보유수는 적고 BMI지수가 높고 고령자임을 알 수 있다. 즉 연속형 변수에 한해 건강상태가 다른 고객들에 비해서는 좋지 않을 고객들이다. 제 1사분면을 가장 대표할 수 있는 171번 고객, 531번 고객, 721번 고객의 정보를 확인해보자.

<제 1사분면의 고객(171번, 531번, 721번 고객)>

	age	sex	bmi	children	smoker	region	charges
171	63	male	41.47	0	no	southeast	13405.39
531	57	male	42.13	1	yes	southeast	48675.52
721	51	female	40.66	0	no	northeast	9875.68

예상대로 BMI지수가 높고 고령자이며 자녀보유수가 적은 고객임을 알 수 있다. 특히 MDA과정에서 분석했듯이 1사분면에 존재하는 흡연자인 531번 고객은 매우 의료보험비가 높게 측정됨을 알 수 있다.

제 2사분면에 있는 고객들은 1사분면과 마찬가지로 자녀보유수가 적고 BMI지수가 낮고 젊은 고객층임을 알 수 있다. 즉 연속형 변수에 한해 건강상태가 다른 고객들에 비해서 좋은 고객들이다. 제 2사분면을 가장 대표할 수 있는 1151번 고객, 341번 고객, 691번 고객의 정보를 확인해보자.

<제 2사분면의 고객(1151번, 341번, 691번 고객)>

	age	sex	bmi	children	smoker	region	charges
1151	18	female	30.305	0	no	northeast	2203.736
341	24	female	27.600	0	no	southwest	18955.220
691	21	male	27.360	0	no	northeast	2104.113

예상대로 BMI지수가 낮고 젊은 층이며 자녀보유수가 적은 고객임을 알 수 있다. MDA과정에서 분석했듯이 이 고객들은 의료비가 매우 낮게 측정된다. 하지만 주목할 점은 341번 고객의 경우에는 흡연자도 아닌 여성고객임에도 불구하고 의료보험비가 높은 편이다. 현재 주어진 상황에서는 아마 다른 외적 변수에 의해서 의료보험비가 측정되었다고 할 수 있다.

제 3사분면에 있는 고객들은 제1 사분면의 특성과는 반대이다. 자녀보유수가 많고 BMI지수가 낮으며 젊은 층의 고객들이다. 제3 사분면을 대표할 수 있는 251번, 381번, 1081번 고객들을 확인해보자.

<제 3사분면의 고객(251번, 381번, 1081번 고객)>

	age	sex	bmi	children	smoker	region	charges
251	18	male	17.290	2	yes	northeast	12829.46
381	27	female	17.955	2	yes	northeast	15006.58
1081	18	male	21.780	2	no	southeast	11884.05

예상대로 BMI지수가 낮고 젊은 층이며 자녀보유수가 많은 고객임을 알 수 있다.

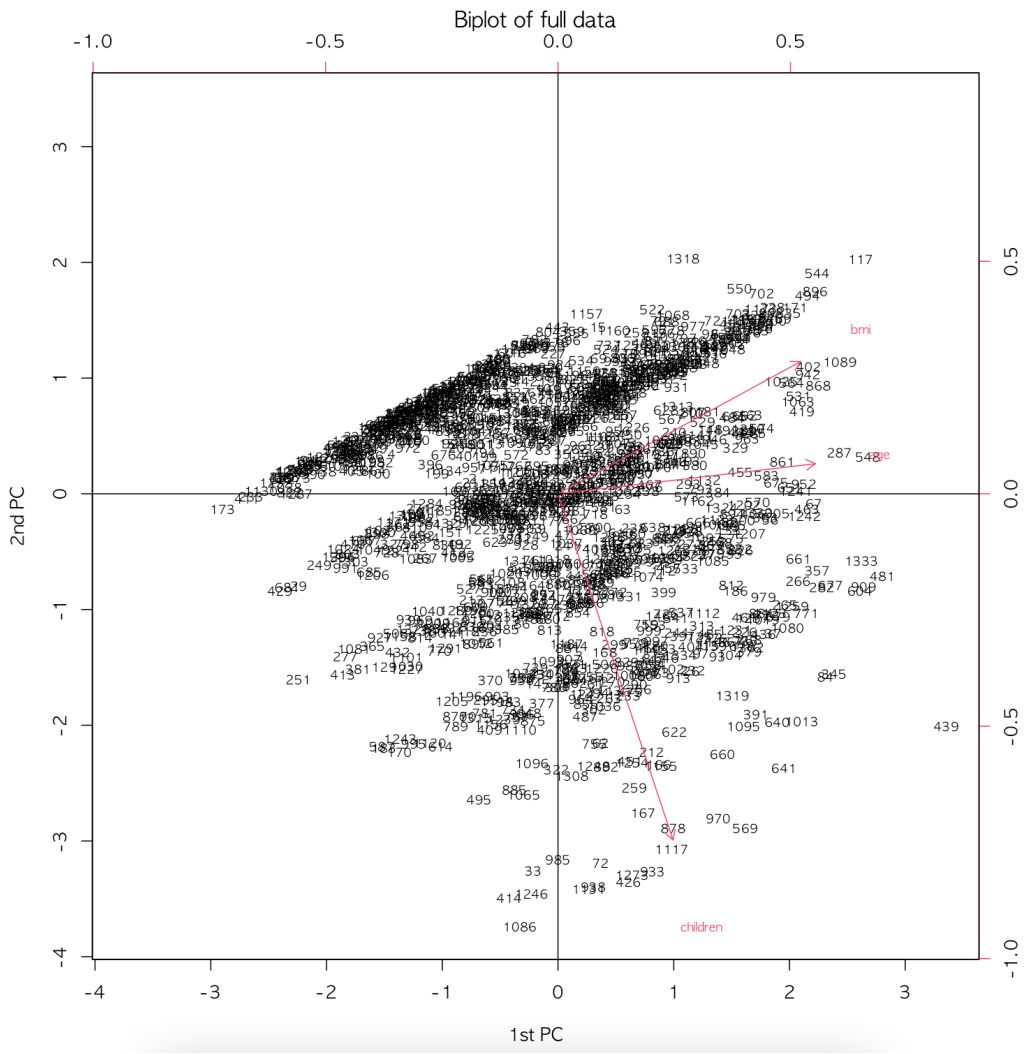
제 4사분면에 있는 고객들은 제2 사분면의 특성과는 반대이다. 자녀보유수가 많고 BMI지수가 높고 고령층의 고객들이다. 제4 사분면을 대표할 수 있는 1231번, 391번, 771번 고객을 확인해보자.

<제 4사분면의 고객(1231번, 391번, 771번 고객)>

	age	sex	bmi	children	smoker	region	charges
1231	52	male	34.485	3	yes	northwest	60021.40
391	48	male	35.625	4	no	northeast	10736.87
771	61	male	36.100	3	no	southwest	27941.29

예상대로 BMI지수가 대체로 높고 고령층에 자녀보유수가 많다. MDA 과정에서 분석했듯이 의료보험비가 다른 고객층들에 비해 매우 높고 특히 흡연자인 1231번 고객의 경우에는 의료보험비가 상당히 높게 측정된 것을 알 수 있다.

PCA 과정에서 고객층들을 여러 군집으로 나눌 수 있다. 직관적인 기준으로는 제1 사분면, 제2 사분면, 제3 사분면, 제4 사분면으로 총 4개의 군집으로 나눌 수 있다. 또한 PCA 과정에서 기본적으로 BMI지수와 나이가 의료보험비에 영향을 미치는 것을 알 수 있었고, 계중에서도 흡연 여부가 의료보험비를 폭발적으로 높게 측정하는 데에 영향을 주는 것을 알 수 있었다.
마지막으로 계통표집법으로 추출하지 않은 전체 데이터에 대한 Biplot을 첨부하고 PCA 분석을 마친다.



3)FA(Factor Analysis)

FA, PCA 모두 차원축소의 목적을 포함하는 분석기법으로서 동일하지만 PCA의 주성분(Principal Component)는 변수들의 선형결합으로 변수들과의 선형적인 관계로 구성되어 새로운 성분을 생성하는 것이 아닌 기존 변수들의 연관성을 표현하는 기법이라고 할 수 있다. 이와 다르게 FA의 인자(Factor)는 변수에서의 연관성을 가리키며 모든 변수를 인자의 적재로 생각하여 공통된 인자를 추출하여 기존과 다른 새로운 변수를 생성하는 기법이다. 우리는 PCFA 자체의 분석 결과 보다는 PCA와 FA의 차이에 중점을 두고 분석을 진행한다.

우리의 데이터는 MDA과정에서 보였듯이 다변량 정규성을 만족하지 않으므로 다변량 정규성을 가정하는 MLFA(Maximum-Likelihood Factor Anlaysis)기법을 제외하고 PCFA(Principal Component Factor Analysis)기법을 사용하여 인자분석을 진행한다. 또한 해석의 편리성을 위해 “varimax” 인자회전을 적용하여 인자분석을 진행한다. 마지막으로 PCFA에서 인자의 개수의 경우, 설명력을 기준으로 선택할 수 있는데 이는 PCA에서의 설명력과 동일하므로 PCFA에서 인자의 개수는 2개로 결정하기로 한다. 따라서 인자개수를 정하는 과정은 생략한다.
인자점수의 경우 데이터가 많아 모든 데이터의 값을 문서에서 확인하기 어렵다. 따라서 차후에 Biplot을 활용하여 자세히 분석해보기로 하고 인자적재값을 PCA와 비교하여 분석해보자.

<주성분 계수(PC loadings) 결과>

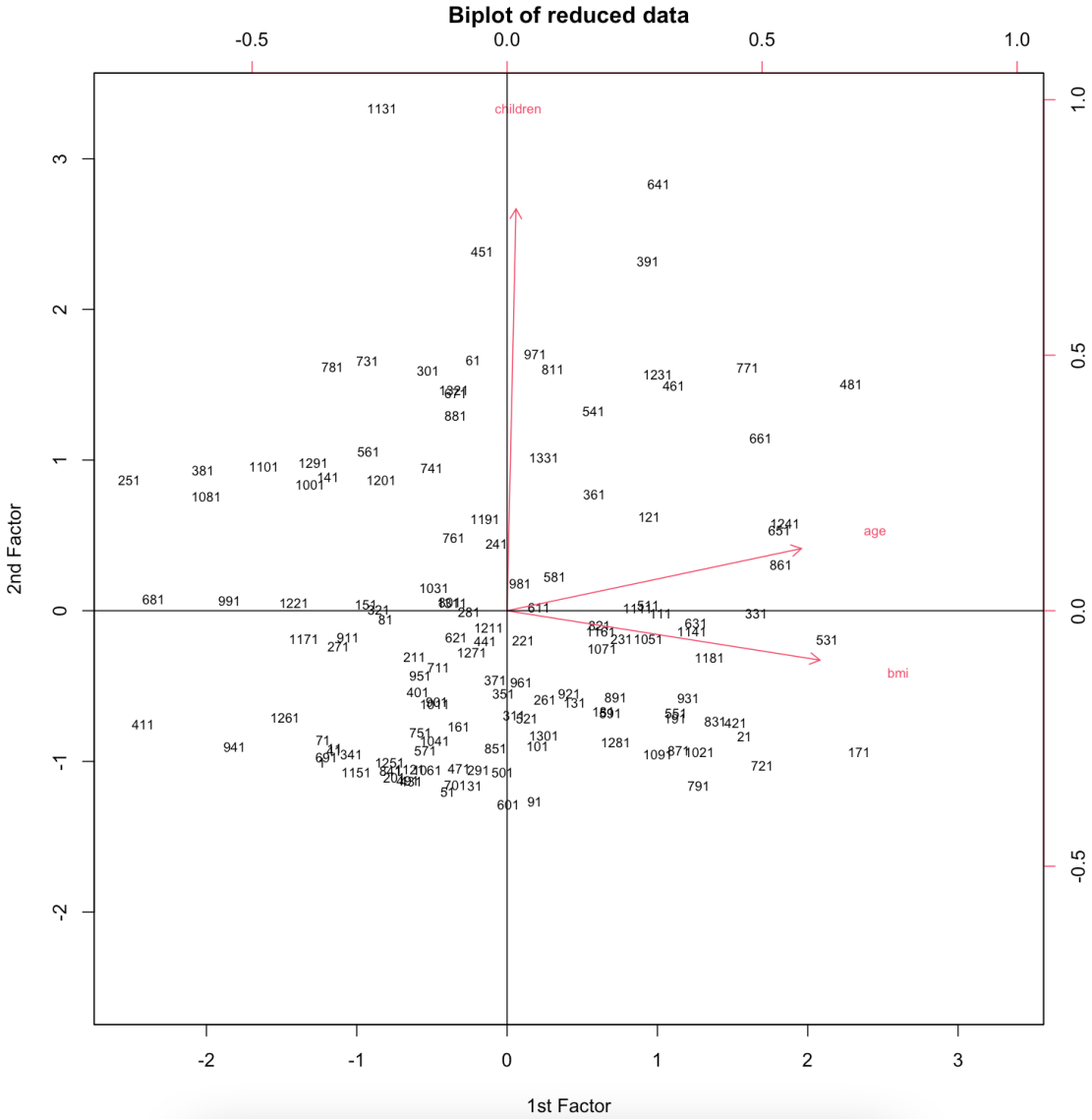
	Comp.1	Comp.2
age	0.6920284	0.08086901
bmi	0.6523048	0.35555968
children	0.3091848	-0.93114849

<인자적재값(Factor loadings) 결과>

	RC1	RC2
age	0.72156912	0.1521142
bmi	0.76689665	-0.1208783
children	0.02199432	0.9830498

제1 공통인자를 결정하는 데에 있어 모든 변수가 양의 방향의 영향을 주었다. 특히 Age(고객의 나이)변수와 BMI(고객의 BMI지수)변수가 큰 영향을 주었다. Children(고객의 자녀수)는 다른 변수들에 비해서는 특출한 영향을 주지 않았다는 것을 알 수 있다. 이는 앞서 PCA 과정에서 주성분 계수를 분석한 것과 비슷한 결과이다. 하지만 FA와 PCA의 차이를 알 수 있는데 그것은 바로 자녀 보유수 변수가 PCA에서 보다 매우 작은 영향을 주었다는 것이다. (PCA에서는 다른 변수와 FA만큼의 차이를 보여주지는 않았음.) 이는 결국에 분석가에 있어 더욱 더 직관적인 분석을 가능하게 해준다.

제2 공통인자를 결정하는 데에는 고객의 BMI지수가 유일하게 음의 방향으로 영향을 주었다. 하지만 큰 영향을 주었다고 보긴 어렵다. 고객의 나이 변수는 양의 방향을 주었지만 큰 영향을 주었다고 보긴 어렵다. 자녀보유수 변수가 압도적으로 제2 공통인자에 양의 방향의 영향을 주었다. 주목할 점은 제2 공통인자의 분석은 PCA와 FA의 결과가 상이하다는 것이다. 자녀보유수가 매우 큰 영향을 준 것은 동일하지만 각각이 양의 방향/음의 방향으로서 반대이며 PCA에서는 age/bmi 변수가 같은 양방향의 영향을 주었지만 PCFA에서는 age/children 변수가 같은 양방향의 영향을 주었다. 더욱 자세한 분석은 Biplot을 통해 분석해보자. 이번 분석은 앞서 언급하였듯 PCA와의 차이점에 초점을 두고 분석한다.

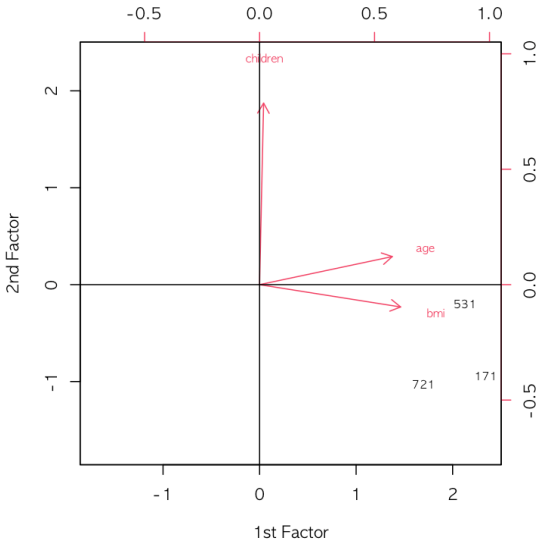


Biplot을 통해 확인할 수 있는 결과들을 살펴보자. 우선 Age(나이)변수와 BMI(bmi지수)변수가 제1주성분 축과 평행에 가까운 각도를 이룬다. 앞서 제1 주성분 계수를 결정하는 데에 있어 나이 변수와 bmi지수가 가 큰 양의 방향의 역할을 했음을 확인하였고 이것이 Biplot에 나타난 것 이라고 할 수 있다. PCA에서와는 반대로, bmi지수는 제2 주성분에대해 나이 변수와 반대인 음의 방향으로 작용을 한다. Children(자녀 보유수)변수는 제2주성분 축과 매우 평행에 가까운 각도를 이룬다. 이 또한 앞서 자녀 보유수 변수의 제2 주성분 계수가 음의 방향으로 매우 높게 나왔으므로 설명된다고 할 수 있다.

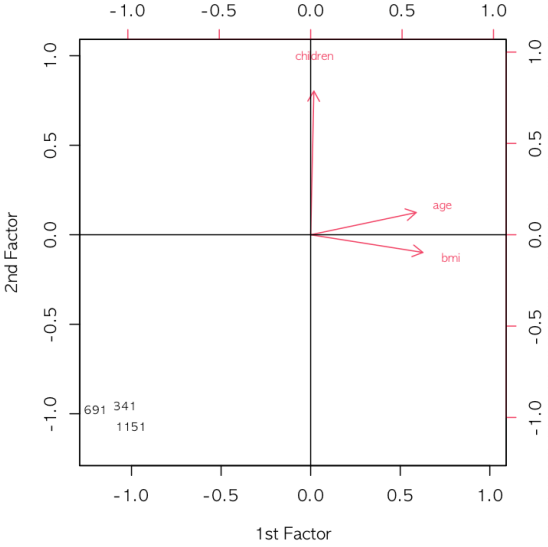
전체적으로 PCA의 결과보다 해석이 용이하다. 이는 인자분석에서 인자회전이 주는 해석력 측면의 장점이 여실히 들어나는 결과라고 할 수 있다. 또한 인자적재값들이 이루는 각도를 보았을 때, 나이와 bmi변수간의 유사성을 파악할 수 있고 자녀 보유수와 나머지 변수들 간의 이질성을 확인할 수 있다.

PCA에서 각 군집의 대표를 나타냈던 고객들이 PCFA의 결과에서는 어떻게 다른지에 대해 확인해보자.

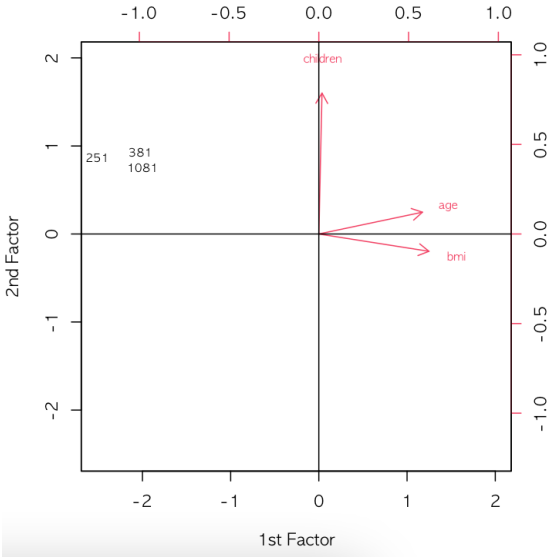
<제 1사분면 PCA → PCFA>



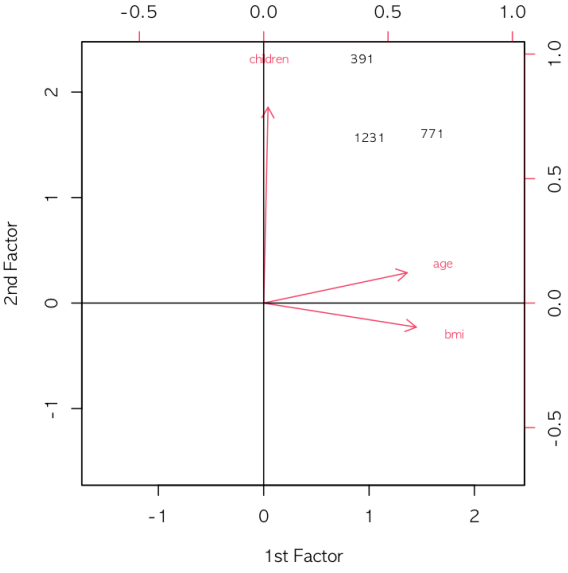
<제 2사분면 PCA → PCFA>



<제 3사분면 PCA → PCFA>

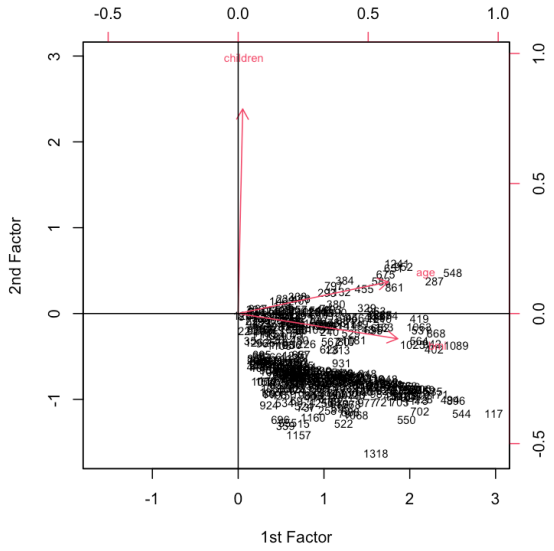


<제 4사분면 PCA → PCFA>

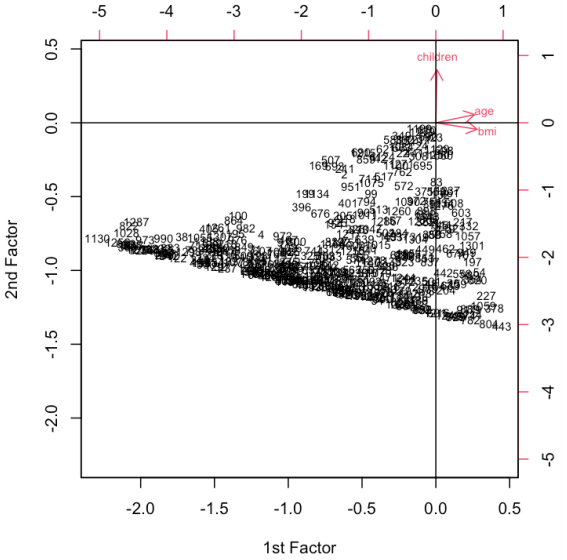


사분면의 위치가 조금씩 달라지긴 했지만 결국에는 PCA에서 이루었던 군집들이 그대로 이루어져 있음을 알 수 있다. 즉 군집의 특성은 PCA와 PCFA가 거의 동일하다. 하지만 이들은 각 군집을 대표하는 고객들이므로, 군집의 경계에 애매하게 서있던 고객들의 경우에는 PCA에서의 군집결과와 PCFA에서의 군집결과가 상이할 수 있다. 이제 PCA에서 사분면의 특성을 기준으로 구분지었던 군집의 고객들이 PCFA에서는 어떻게 분류되었는지를 확인해보자.

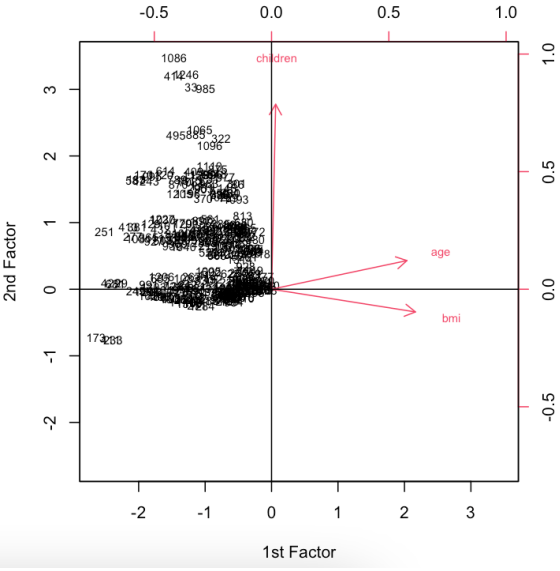
<제 1사분면 PCA → PCFA>



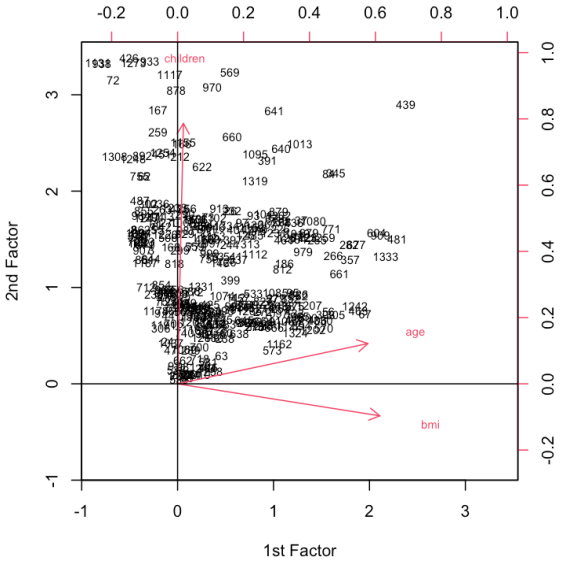
<제 2사분면 PCA → PCFA>



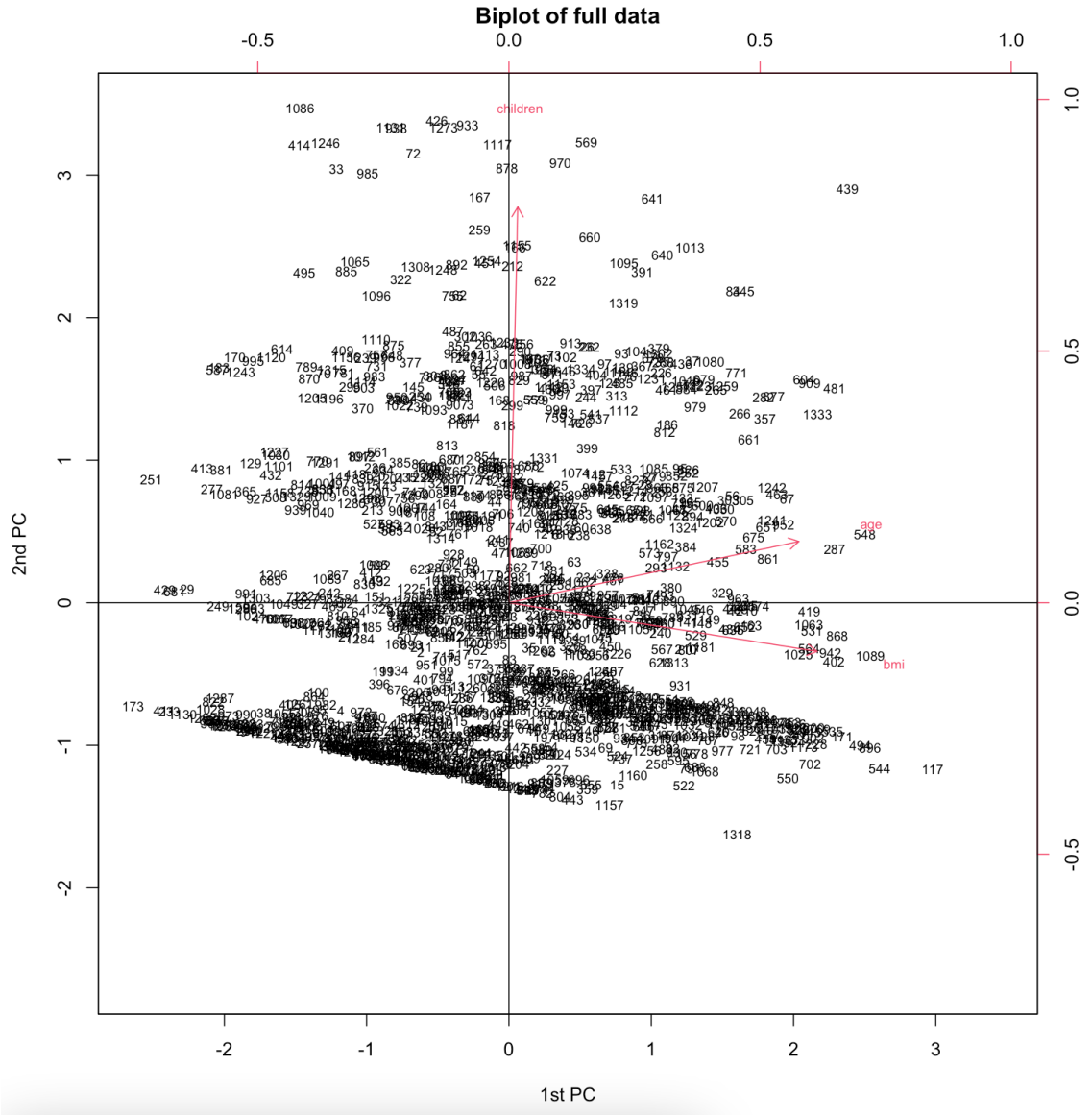
<제 3사분면 PCA → PCFA>



<제 4사분면 PCA → PCFA>



예상했던 대로 PCA에서의 고객군집과 PCFA에서의 고객 군집이 상이하다. 하지만 전체적으로 보았을 때는 대체로 겹치는 고객들이 다수 보이는 것을 알 수 있다. 마지막으로 계통표집법으로 추출하지 않은 전체 데이터에 대한 Biplot을 첨부하고 PCFA 분석을 마친다.

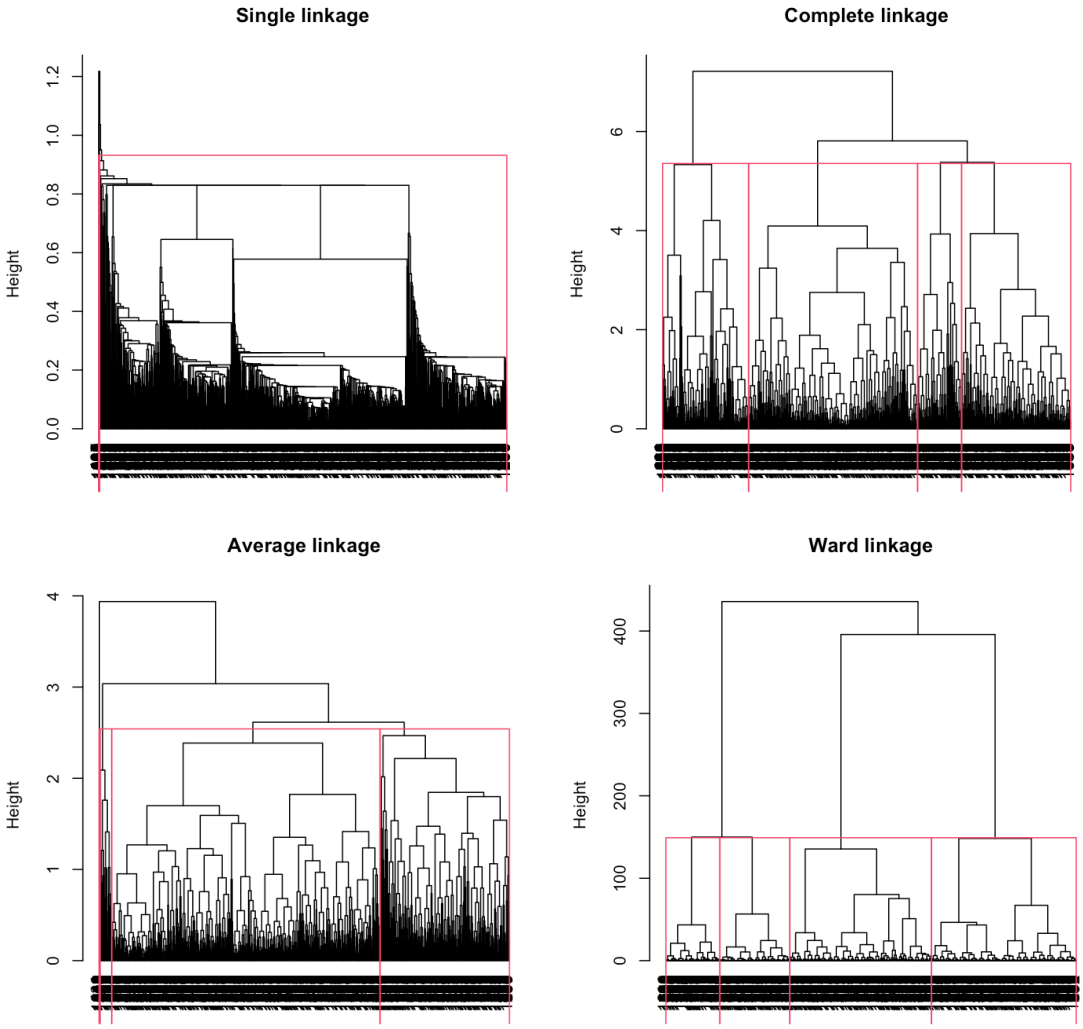


4)CA(Cluster Analysis)

군집분석에서는 계층적/비계층적 군집화 기법들로 군집화를 진행한 후 각 군집들이 고객들의 정보에 대해 어떤 차이가 있으며 결론적으로 고객들의 정보가 의료보험비와 어떤 관계를 가지는지, 즉 군집간의 의료보험비 차이를 살펴보려 한다.

군집기법으로는 계층적군집화(Hierarchical Clustering)과 비계층적 군집화 방법인 K-means Clustering을 사용한다. PCA와 PCFA에서의 군집과 비교 분석을 위하여 두 군집 기법 모두 군집의 개수를 4개로 설정한다. 또한 군집화에 사용될 기준 거리로는 표준화 유클리드 거리를 사용하며 K-means clustering에는 표준화 행렬을 사용한다.

우선 계층적 군집화에서 덴드로그램을 시각화 한 후 가장 적절히 군집이 이루어지는 연결 기법을 선택해보자.



각 연결법의 덴드로그램 결과를 살펴보면 Single linkage와 Average linkage method의 경우에는 절대 적절한 군집화가 이루어졌다고 보기는 힘들다. 특정 군집에만 고객들이 분포되어 있기 때문이다. Comple linkage와 Ward linkage method의 경우는 적절한 군집화가 이루어졌다고 할 수 있는데, 정보의 손실을 줄이는 특성을 갖는 Ward linkage를 계층적 군집화 연결기법으로 선택한다. 이 군집화 결과를 “hclust.result~” 라는 변수에 저장한 후 k-means clustering을 실시해보자.

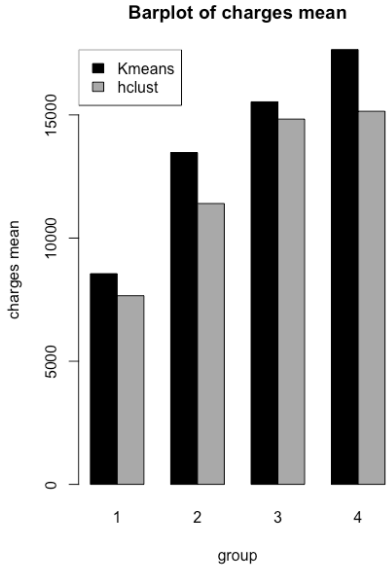
<K-means clustering 결과 요약>

```
K-means clustering with 4 clusters of sizes 265, 326, 418, 329
```

군집화가 잘 진행되어졌음을 알 수 있다. 각각의 군집이 1~4로 군집화 되었는데 이 숫자는 Hierarchical/K-means이 다르므로 각 군집의 의료보험비 평균을 구한뒤 의료보험비 평균이 가장 낮은 군집을 군집1, 의료보험비 평균이 가장 높은 군집을 군집4로 하여 순서화 시킨 뒤 시각화하면 다음과 같다.

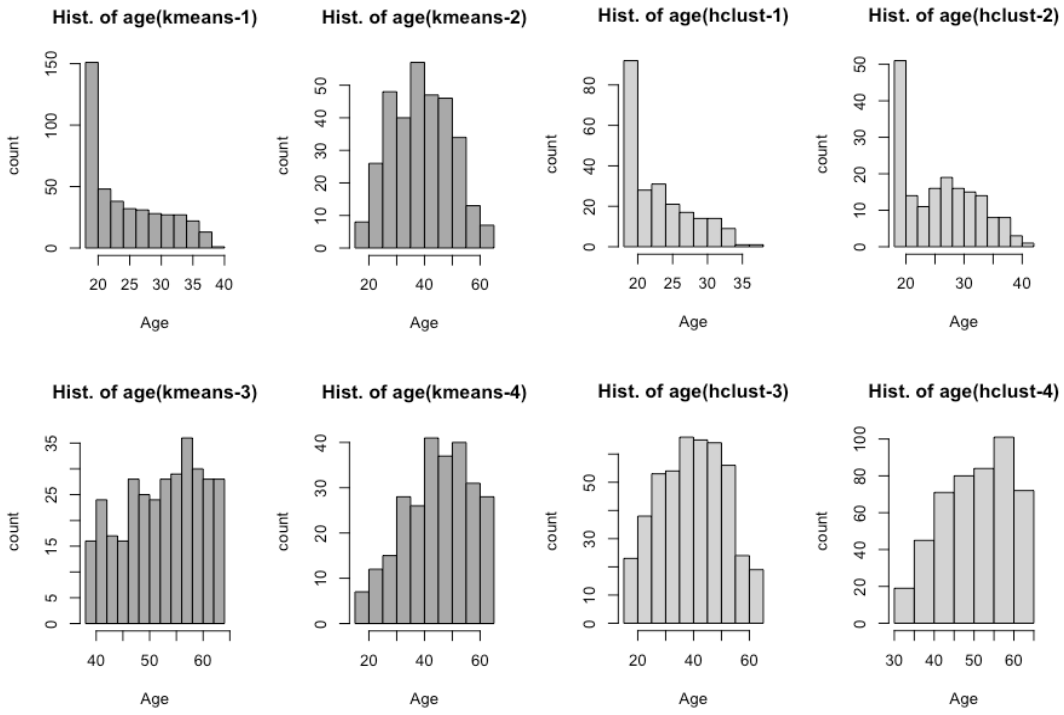
<군집화를 통한 각 군집별 의료보험비 평균>

	group	Kmeans	hclust
1	1	8556.143	7659.185
2	2	13472.080	11402.777
3	3	15530.831	14831.623
4	4	17652.135	15149.221



군집화 결과로 K-means clustering 군집들이 계층적 군집화의 군집들보다 의료보험비 평균이 높음을 알 수 있다. 그럼 각 군집별 고객들의 정보가 어떠한 차이를 보이는지 두 군집화 기법을 통한 결과를 비교하며 시각화를 통해 확인해보자. 이때 군집화 과정에서 사용된 변수들에 대해서만 비교 분석해본다. 또한 군집1이 의료보험비가 낮고 4가 의료보험비가 높음은 후에 따로 언급하지 않는다.

-Age(고객의 나이)

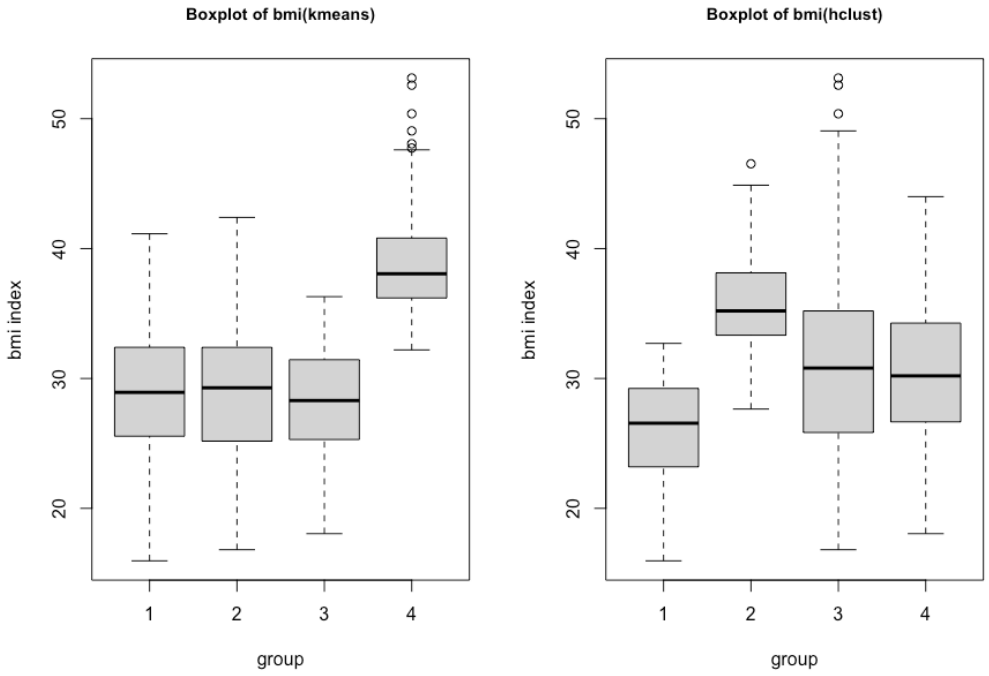


<Age 변수에 대한 분석표>

	group1	group2	group3	group4
kmeans	젊은층	중년층	중~고령층	고령층
hclust	젊은층	젊은층~중년층	중~고령층	고령층

k-means clustering은 나이별로 고객층을 세밀하게 나누었지만 hierarchical clustering은 군집1과 군집2를 젊은층 으로만 구성된 군집과 젊은층~중년층으로 나뉜 군집으로 나누었다. MDA과정에서 확인하였듯이 고령층일수록 의료보험비가 높게 측정되었음을 확인할 수 있다.

-BMI(고객의 bmi지수)

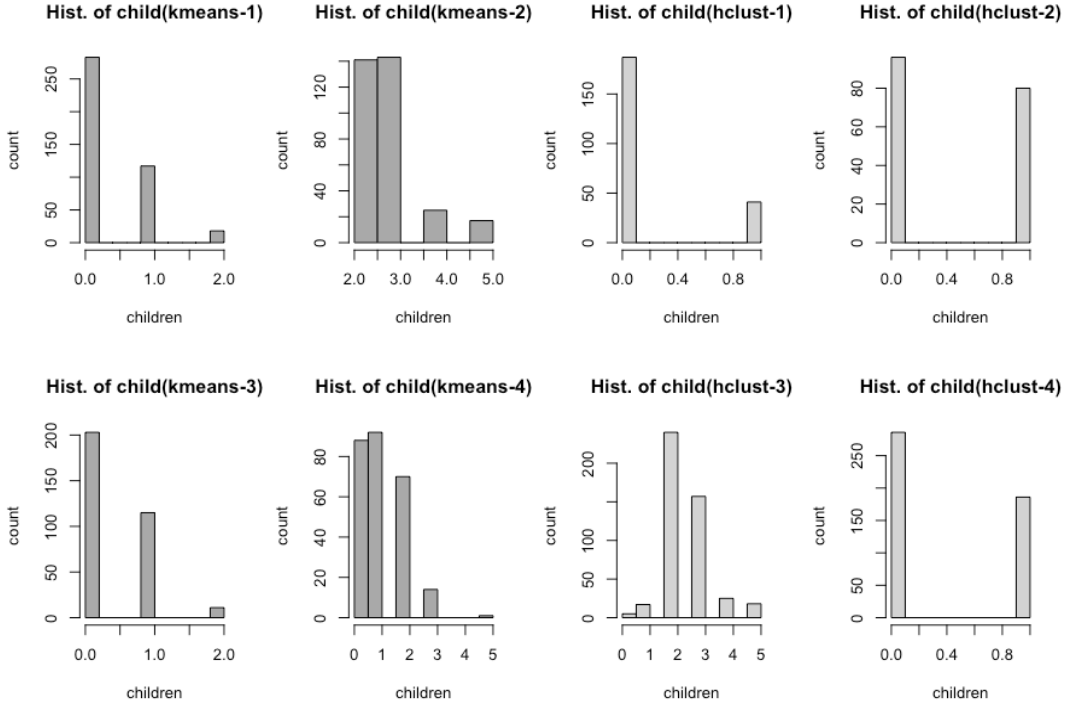


<BMI 변수에 대한 분석표>

	group1	group2	group3	group4
kmeans	정상	정상	정상	비만
hclust	저체중~정상	정상~비만	정상	정상

k-means clustering은 의료보험비가 높게 측정된 군집4에서만 bmi지수가 높게 분류되었고, hierarchical clustering은 군집별로 차이는 있지만 대개 정상범주로 bmi지수에 의한 군집화가 극명하게 된것처럼 보이진 않는다.

-Children(고객의 자녀 보유 수)



<children 변수에 대한 분석표>

	group1	group2	group3	group4
kmeans	저출산	다자녀	저출산	평범(현실적)
hclust	저출산	저출산	평범(현실적)	저출산

k-means clustering은 다자녀 가구로 구성된 군집이 존재하고, hierarchical clustering은 다자녀 가구로 구성된 군집이 존재하지 않는다는 차이점이 있다. MDA에서 자녀 보유수는 의료보험비 측정에 큰 영향을 주지 않음을 알 수 있었는데, 마찬가지로 group1~group4로 가면서 자녀 보유수와 의료보험비에 대한 관계를 찾기는 힘들다.

5)Conclusion

지금까지 MDA, PCA, FA, CA 차례로 분석과정을 지나며 의료보험비와 고객정보들의 관계에 대해서 알아보았다. 표로 종합해서 요약해보자.

<고객의 의료보험비 측정에 대한 각 변수들의 관계>

	선형 관계	설명
age(나이)	비례 관계	고령층일수록 의료보험비가 높게 측정되는 경향이 있다.
sex(성별)	애매	여성일수록 의료보험비가 낮게 측정되는 경향이 있지만 이는 성별 때문이 아닌 성별이 다른 변수에 종속되기 때문에 나타나는 결과이다.
bmi(bmi지수)	비례 관계	BMI지수가 높을수록 의료보험비가 높게 측정되는 경향이 있다.
children(자녀보유수)	없음	자녀보유수와 의료보험비 측정간의 관계는 찾아보기 힘들었다.
smoker(흡연여부)	비례 관계	흡연자의 경우 의료보험비가 높게 측정되는 경향이 있다. 특히 다른 변수들과 결합하였을 때 효과가 배로 늘어난다. (예: 고령층인데 흡연자)
region(거주 지역)	없음	거주지역과 의료보험비 측정간의 관계는 찾아보기 힘들었다.

4. CODE

1)MDA(Multivariate Data Analysis)

```
# Check Data
data = read.csv("insurance.csv", header=TRUE)
X = data
head(X)
dim(X)

colSums(is.na(X)) # 결측치 존재 X

# Check $Age
summary(X$age)
par(family="AppleGothic")
hist(X$age, main="Histogram of age", xlab="Age", ylab="count", xlim=c(min(X$age)-5, max(X$age)+3), breaks=length(unique(X$age)))

X2 = X[,c("age", "charges")]

for (i in 1:6){
  X2[(X2$age >= i*10) & (X2$age < (i+1)*10), "age"] = sprintf("%s0대", i)
}
boxplot(X2$charges~X2$age, xlab="나이대", ylab="charges", main="Boxplot of charges about age")

# Check $sex
table(X$sex)

boxplot(X$charges~X$sex, xlab="sex", ylab="charges", main="Boxplot of charges about sex", cex.lab=1.3, cex.main=1.5, cex.axis=1.5)

table(X[X$smoker=="yes", "sex"])
table(X[X$bmi > mean(X$bmi), "sex"])

# Check $bmi
boxplot(X$bmi, main="Boxplot of bmi", ylab="bmi index", cex.main=1.5, cex.lab=1.2, cex.axis=1.2)
text(x=1.4, y=30, labels=paste("median :", round(median(X$bmi), 3)), cex=0.7)
```

```
plot(X$charges~X$bmi, xlab="bmi", ylab="charges", main="Scatterplot of charges about bmi", cex.lab=1.2, cex.main=1.5, cex.axis=1.2)
abline(lm(X$charges ~ X$bmi), col="red", lwd=3)

# Check $children
tot.count.chil = sum(table(X$children))
obs.chil = paste("child", c("0","1","2","3","4","5"), sep="")
ratio.child = 0
for (i in 1:6){
  ratio.child[i] = 100*round(table(X$children)[i]/tot.count.chil, 3)
  }

pie(table(X$children), main="Ratio of children", labels=paste(obs.chil,":", ratio.child, "%", sep=""), cex=0.9, cex.main=1.5)

boxplot(X$charges~X$children, xlab="children", ylab="charges", main="Boxplot of charges about children", cex.lab=1, cex.main=1.5, cex.axis=1)

# Check $smoker
ratio.smoker = 100*round(table(X$smoker)/sum(table(X$smoker)), 3)
pie(table(X$smoker), main="Ratio of smoker", labels=paste(rownames(ratio.smoker), ":", ratio.smoker, "%"), cex=1, cex.main=1.5)

boxplot(X$charges~X$smoker, xlab="smoker", ylab="charges", main="Boxplot of charges about smoker", cex.lab=1, cex.main=1.5, cex.axis=1)

# Check$region
table(X$region)

boxplot(X$charges~X$region, xlab="region", ylab="charges", main="Boxplot of charges about region", cex.lab=1, cex.main=1.5, cex.axis=1)

# Check $charges
boxplot(X$charges, main="Boxplot of charges", ylab="charges", cex.main=1.5, cex.lab=1)
text(x=1.4, y=10000, paste("median :", labels=round(median(X$charges)), 3), cex=0.8)

# Check the correlationships between variables and charges
cont.columns = c("age", "bmi", "children", "charges")
round(cor(X[,cont.columns]), 3)
```

```
# Check the indepedence of categorical variables
cat.columns = c("sex", "children", "smoker", "region")
summary(table(X[, cat.columns]))

# Check the indepedence of categorical variables & charges
table = xtabs(charges~sex+children+smoker+region, data=X)
summary(table)

# Check the normality
library(MVN)
mvn(X[,cont.columns], mvnTest="mardia",
 multivariatePlot="qq")$multivariateNormality
```

2)PCA(Principal Component Analysis)

```
pca = princomp(X[, c("age", "bmi", "children")], cor=T)
summary(pca, loadings=T)
screeplot(pca, type="lines", main="Screeplot of PCA")

G = pca$scores[, 1:2] # pc scores
H = pca$loadings[,1:2][,1:2] # pc loadings
rownames(G) = rownames(X)
G[1:10,]
H

reduce.index = seq(1, nrow(X), 10) # we will use 10% of data
biplot(G[reduce.index,], H, xlab="1st PC", ylab="2nd PC", main="Biplot of reduced
data", cex=0.7)
abline(v=0, h=0)

pca.1 = c(171, 531, 721) ; pca.2 = c(1151, 341, 691) ; pca.3 = c(251, 381, 1081) ;
pca.4 = c(1231, 391, 771)
X[pca1, ] # 제1 사분면
X[pca2, ] # 제2 사분면
X[pca3, ] # 제3 사분면
X[pca4, ] # 제4 사분면

biplot(G, H, xlab="1st PC", ylab="2nd PC", main="Biplot of full data", cex=0.7)
abline(v=0, h=0)
```

3)FA(Factor Analysis)

```
Z = scale(X[,c("age", "bmi", "children")], scale=T)

library(psych)
pcfa = principal(Z, nfactors=2, rotate="varimax")

G.pcfa = pcfa$scores[, 1:2] # pcf scores
H.pcfa = pcfa$loadings[,1:2][,1:2] # pcf loadings
rownames(G.pcfa) = rownames(X)
G.pcfa[1:10,]
H.pcfa

people = list(pca.result1, pca.result2, pca.result3, pca.result4)
par(mfrow=c(2,2))
for (person in people){
  biplot(G.pcfa[person, ], H.pcfa, xlab="1st Factor", ylab="2nd Factor", cex=0.7)
  abline(v=0, h=0)
}

biplot(G.pcfa, H.pcfa, xlab="1st PC", ylab="2nd PC", main="Biplot of full data",
cex=0.7)
abline(v=0, h=0)
```

4)CA(Cluster Analysis)

```
ds = as.matrix(dist(Z, method="euclidean")) ; ds = as.dist(ds)

par(mfrow=c(2,2))
#single linkage
single = hclust(ds, method="single")
plot(single, main="Single linkage", hang=-1)
rect.hclust(single, k=4)

#complete linkage
complete = hclust(ds, method="complete")
plot(complete, main="Complete linkage", hang=-1)
rect.hclust(complete, k=4)

#average linkage
average = hclust(ds, method="average")
```

```
plot(average, main="Average linkage", hang=-1)
rect.hclust(average, k=4)

#ward linkage
ward = hclust(ds, method="ward.D") # we will choose ward-linkage
plot(ward, main="Ward linkage", hang=-1)
rect.hclust(ward, k=4)

hclust = cutree(ward, k=4)
hclust.result1 = which(hclust==1)
hclust.result2 = which(hclust==2)
hclust.result3 = which(hclust==3)
hclust.result4 = which(hclust==4)

kmeans.clust = kmeans(Z, 4, nstart=1000)
kmeans = kmeans.clust$cluster
order = order(aggregate(X["charges"], by=list(kmeans), FUN=mean)[["charges"]])

kmeans.index1 = which(kmeans==order[1])
kmeans.index2 = which(kmeans==order[2])
kmeans.index3 = which(kmeans==order[3])
kmeans.index4 = which(kmeans==order[4])

kmeans = replace(kmeans, kmeans.index1, 1)
kmeans = replace(kmeans, kmeans.index2, 2)
kmeans = replace(kmeans, kmeans.index3, 3)
kmeans = replace(kmeans, kmeans.index4, 4)

# save index
kmeans.result1 = which(kmeans==1)
kmeans.result2 = which(kmeans==2)
kmeans.result3 = which(kmeans==3)
kmeans.result4 = which(kmeans==4)

kmeans.charges = aggregate(X["charges"], by=list(kmeans), FUN=mean)[["charges"]]
hclust.charges = aggregate(X["charges"], by=list(hclust), FUN=mean)[["charges"]]

clusters = data.frame(group=c(1, 2, 3, 4), Kmeans=kmeans.charges,
hclust=hclust.charges)
rownames(clusters) = clusters$group
clusters
```

```
barplot(t(clusters[,c("Kmeans", "hclust")]), beside=TRUE, main="Barplot of charges mean", xlab="group", ylab="charges mean", col=c("black","darkgray"))
legend("topleft", legend=c("Kmeans", "hclust"), fill=c("black", "darkgray"))

# Check age
par(mfrow=c(2,2))
hist(X[kmeans.result1, "age"], main="Hist. of age(kmeans-1)", xlab="Age", ylab="count", col="darkgray")
hist(X[kmeans.result2, "age"], main="Hist. of age(kmeans-2)", xlab="Age", ylab="count", col="darkgray")
hist(X[kmeans.result3, "age"], main="Hist. of age(kmeans-3)", xlab="Age", ylab="count", col="darkgray")
hist(X[kmeans.result4, "age"], main="Hist. of age(kmeans-4)", xlab="Age", ylab="count", col="darkgray")

hist(X[hclust.result1, "age"], main="Hist. of age(hclust-1)", xlab="Age", ylab="count")
hist(X[hclust.result2, "age"], main="Hist. of age(hclust-2)", xlab="Age", ylab="count")
hist(X[hclust.result3, "age"], main="Hist. of age(hclust-3)", xlab="Age", ylab="count")
hist(X[hclust.result4, "age"], main="Hist. of age(hclust-4)", xlab="Age", ylab="count")

# Check bmi
par(mfrow=c(1,1))
bmi.data = X
bmi.data["kclust"] = 0 ; bmi.data["hclust"] = 0
bmi.data[kmeans.result1, "kclust"] = 1 ; bmi.data[kmeans.result2, "kclust"] = 2
bmi.data[kmeans.result3, "kclust"] = 3 ; bmi.data[kmeans.result4, "kclust"] = 4
bmi.data[hclust.result1, "hclust"] = 1 ; bmi.data[hclust.result2, "hclust"] = 2
bmi.data[hclust.result3, "hclust"] = 3 ; bmi.data[hclust.result4, "hclust"] = 4

boxplot(bmi.data$bmi~bmi.data$kclust, main="Boxplot of bmi(kmeans)", ylab="bmi index", xlab="group", cex.main=0.9)
boxplot(bmi.data$bmi~bmi.data$hclust, main="Boxplot of bmi(hclust)", ylab="bmi index", xlab="group", cex.main=0.9)

# Check children
par(mfrow=c(2,2))
hist(X[kmeans.result1, "children"], main="Hist. of child(kmeans-1)", xlab="children", ylab="count", col="darkgray")
hist(X[kmeans.result2, "children"], main="Hist. of child(kmeans-2)", xlab="children", ylab="count", col="darkgray")
```

```
hist(X[kmeans.result3, "children"], main="Hist. of child(kmeans-3)", xlab="children", ylab="count", col="darkgray")
hist(X[kmeans.result4, "children"], main="Hist. of child(kmeans-4)", xlab="children", ylab="count", col="darkgray")

hist(X[hclust.result1, "children"], main="Hist. of child(hclust-1)", xlab="children", ylab="count")
hist(X[hclust.result2, "children"], main="Hist. of child(hclust-2)", xlab="children", ylab="count")
hist(X[hclust.result3, "children"], main="Hist. of child(hclust-3)", xlab="children", ylab="count")
hist(X[hclust.result4, "children"], main="Hist. of child(hclust-4)", xlab="children", ylab="count")
```

5. REFERENCE

- R과 함께하는 다변량 자료분석(최용석 저)