BIOAGE

건강지수의 정의와 이를 이용한 생체나이 측정

한대룡, 백우현

1. 배경 및 방향 제시

우리가 조사한 바에 따르면 현재 생체나이는 각종 신체검사를 통해 측정한 결과를 바탕으로 해당 나이와 비슷한 연령대 사람들의 평균값을 활용하여 그보다 좋으면 생체나이가 젊게 계산되고, 좋지 못하면 더 나이가 든 것으로 계산을 한다. 그러나 우리의 데이터에서도 볼 수 있듯이 연령 별 건강검진 데이터가 균등하지 않으며, 18세와 80세 주변의 관측치는 다른 연령보다 현저히 적음을 알 수 있다. 이 말은 즉, 이와 같은 방법으로 생체나이를 계산하게 되면 이 연령대의 사람은 다른 연령보다 부정확한 값을 얻을 가능성이 크다는 것을 의미한다. 또한, 각 나이대별로 다르게 계산이 되기에 절대적인 건강함의 정도를 나타내기 힘들다는 한계를 지닌다.

따라서 우리는 정상수치와의 거리를 기준으로 좀 더 정확하고 일관적인 방법으로 생체나이를 계산하고자 했으며, 동시에 기하적/시각적으로도 쉽게 표현할 수 있는 방법을 고안하였다. 또한 계층 군집분석을 통해 모든 변수를 같은 비중으로 상정하고 거리를 계산했을 때 보다 좀 더 합리적으로 접근할 수 있었으며 이를 바탕으로 그 사람의 전반적인 건강상태를 나타내는 건강 수치를 생성하였고, 이를 이용하여 최종적으로 생체나이를 계산하였다.

이 분석을 진행하면서 크게 두 가지 논점이 존재하였는데, 이는 다음과 같으며 이 논점에 대한 우리의 생각을 제시하면서 분석을 진행해보려 한다.

1. 혈액, 폐기능, 면역, 종양, 요 검사 등의 항목을 각각 분류하여 고려해야 할까?

2. 20세와 80세 두 명의 사람이 만일 모든 건강검진 항목의 수치가 동일할 경우 이 두 사람의 생체나이는 같은 것일까?

2. 데이터 전처리

2-1. 결측치 처리

생체나이 예측 분석방향은 완전정상 벡터로부터의 거리에 기초하였기에 우리의 데이터에는 결측치를 대체하거나 삭제할 필요가 있었다. 결측치 처리 방향은 크게 두 단계로 진행됐는데, 다음과 같다. 처음으로 결측치가 전체 obs 3만 개 중 1만 개 이상(30%)인 변수를 삭제해주었다. 해당 변수와 각 변수 별 결측값의 수는 아래 그림과 같은데 이와 같이 너무 많은 결측값을 가진 변수를 대체하기에는 예측의 정확성 감소는 물론 오히려 부정확한 결론이 나올 것이 우려되어 삭제를 하였다. 그런데 여기서 PSA와 CA125 항목 역시 결측치가 10,000개 이상이었으나 조사해보니 PSA(전립선 특이항원)는 남자에게만, CA125(난소암)는 여자에게만 있는 항목임을 알게 되었고, 이 변수는 삭제를 하지 않고 남자와 여자의 경우를 분리하여 생체나이를 분석하기로 하였다.

WAIST	HIP	CPK	DIRBIL	HBA1C	IRON	TIBC	UIBC	AMYLASE	WBC
22213	28788	25989	20398	17063	26807	26841	26849	25177	14993
ESR	PDW	T3	CRP						
11330	28739	21824	20880						

Figure 1 삭제 변수

위 변수들을 삭제한 이후 나머지 변수들 중 결측값이 수는 모두 7000개 이하였는데 해당 결측값들은 변수가 아닌 obs(관측치)를 삭제해 주었다. Mice같은 패키지를 이용하여 다중대체 혹은 평균대체를 할 수도 있었으나 대체가 아닌 삭제를 한 이유는 다음과 같다. 먼저 생체나이를 계산하는 새로운 방안을 고안하는 것이기에 우리가 구한 생체나이가 올바른 결과라는 것을 평가할 마땅한 척도가 존재하지 않았다. 따라서 대체값을 이용하기보단 최대한 데이터의 그 특성을 살리는 것이 맞다고 판단을 하였다.

그리고 무엇보다도 결측값이 있는 관측치를 삭제한 결과 남녀 각각 대략 8000 ~ 9000개의 관측치가 남았는데 이것만으로도 이 분석을 진행하기에 충분한 양의 데이터라고 판단을 하였다. 앞서 말했듯이 우리의 분석 방식은 각각의 연령을 구분하는 것이 아니기에 모든 데이터를 한 번에 사용할 수 있었기에 이 정도 데이터의 크기면 충분하다고 생각을 했다. 아래의 그림은 남/여 데이터의 결측값 전처리 과정후의 히스토그램을 비교해 봄으로써 전체적인 분포가 크게 변하지 않았음을 시각적으로 확인을 하였다.

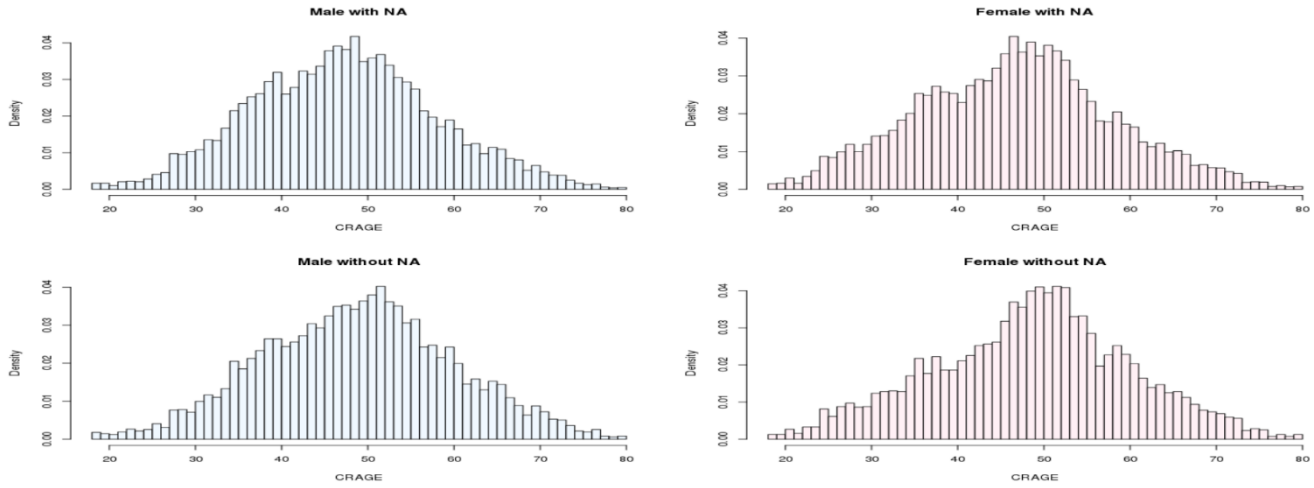

Figure 2 결측치 처리 전후 히스토그램(좌: 남자, 우: 여자)

2-2. 이상치(Outlier) 제거

일반적인 이상치 처리 방법으론 사분위수를 이용한 여러 방법이 존재하는데, 의료데이터의 특성상 일반적인 이상치 처리를 적용하면 안 된다고 판단을 하였다. 그 이유는 건강검진 데이터의 경우 정상의 경우가 대다수이고, 비정상의 값은 굉장히 극소수인 상당히 비균형적인 특성을 지닌다. 관상동맥 데이터 역시 마찬가지로 질병에 걸린 환자의 수가 3%미만이었는데 단순히 값이 작고 크다라는 기준으로 이상치를 제거할 경우 오히려 우리에게 정말 중요한 데이터를 잃는 결과를 초래할 것이라고 판단하였다.

따라서 데이터 전처리 과정에서는 이상치 처리는 각각 변수의 Boxplot을 일일이 살펴본 후 아래의 그림처럼 BMI가 45이상(이는 키 150cm, 몸무게 100kg을 의미)와 같이 직관적으로 결측 오류라고 판단되는 관측치만 삭제를 해주었으며 남자의 경우 14개, 여자의 경우 17개의 관측치를 삭제를 하였다. 설령 이 값이 실제 값이라 하더라도 이 값으로 인해 우리가 만든 건강지수가 왼쪽으로 지나치게 쏠리는 현상이 발생하기에 이를 해결하기 위해서는 삭제를 하는 것이 맞다고 판단하였다. 대신 건강지수를 만든 뒤 그 분포를 확인 후 그 수치가 극단적으로 높은 값을 삭제하는 방식으로 이상치의 악영향을 방지할 수 있을 것이라 생각했다.

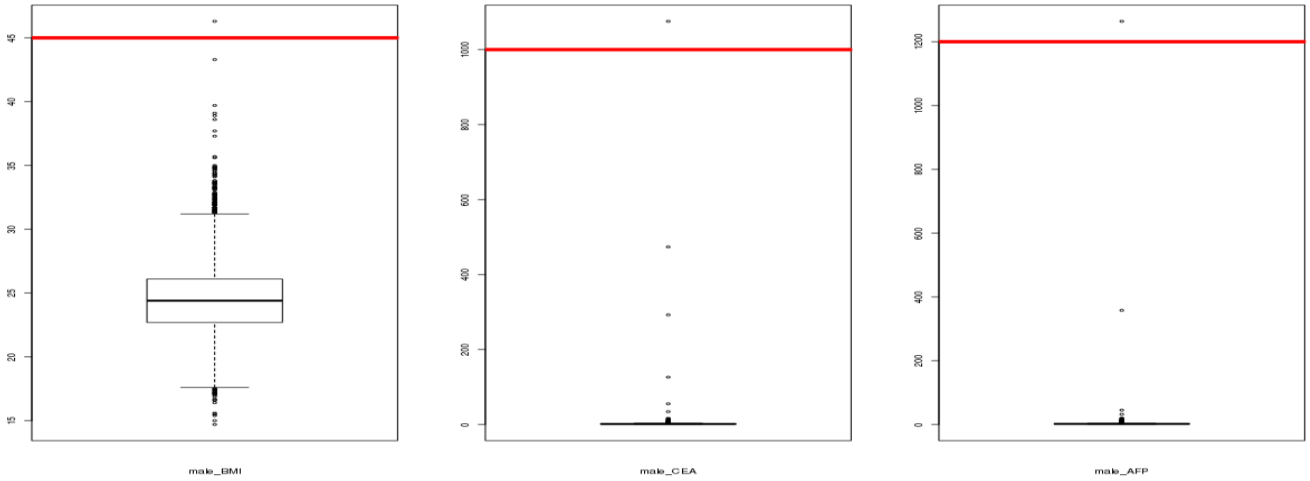

Figure 3 BMI, CEA, AFP Boxplot

3. 분석 과정

3.1. 건강지수: 정상범위와의 격차

총 78개의 건강검진 변수가 존재하였는데 이 변수를 원래의 상태로 사용하기에는 무리가 있었다. 그 이유는 예를 들어 BMI의 정상 범위가 20 ~ 25라고 가정할 경우 BMI가 19인 사람과 26인 사람 모두 정상이 아닌 사람으로 고려가 되어야 하기 때문이다. 따라서 우리는 각각의 변수를 그대로 사용한 것이 아니라 변수 설명에 있는 각 변수 별 정상수치 범위를 참고하여 정상수치와의 거리로 값을 수정해주었다. (참고로 정상수치가 표시 되어있지 않은 변수들은 일반적인 경우 측정되는 임상참고치를 정상수치로 대체하여 참가했다.) 이는 사람이 얼마나 건강한지, 다시 말해 얼마나 완전 정상수치로부터 떨어져있는지를 수치상으로 표현한 것이다. 즉, 0을 기준으로 값을 재배치해주었기 때문에 숫자가 클수록 건강이 안 좋은 것을 의미하며 이와 같은 과정을 진행함으로써 후에 선형회귀에 데이터를 적합할 수 있는 기반을 마련하였다.

추가적으로 조사한 바에 따르면 FVCPP는 그 자체 변수만으로 건강함의 정도를 판단하는 것이 아니라 FEV1_FVC의 변수와 같이 고려함으로써 건강의 상태를 나타내기에 이 두 변수를 통합하여 fvc라는 새로운 변수를 생성하였다.

각 obs별로 변수의 값이 정상수치 범위로부터 떨어져 있는 거리를 계산하여 삽입해준 후 변수 별로 scale이 상당히 다양하기 때문에 정상수치와의 거리를 일정하게 통일시켜줄 필요가 있었다. 따라서 계산된 거리를 각 변수별로 최대/최솟값 기준 0~1로 Scaling을 해주었다. (MinMax Scaling) 아래는 계산과정의 단순한 예시이다. (*'변수이름 [정상수치범위]'*로 표시) 해당 과정을 남성데이터와 여성데이터에 각각 적용시켜주었다.

기존데이터

BAID	SBP [0~120]	DBP[0~80]	BMI[18.5~25]	FEV1_FVC[0.7이상]
BA00102	124	83	21.1	0.683
BA00109	131	85	25.6	0.865
BA00110	133	90	24.0	0.784

정상범위와의 거리

BAID	SBP [0~120]	DBP[0~80]	BMI[18.5~25]	FEV1_FVC[0.7이상]
BA00102	4	3	0 (정상)	0.017
BA00109	11	5	0.6	0 (정상)
BA00110	13	10	0 (정상)	0 (정상)

변수별로 MinMax Scaling (0 ~ 1)

BAID	SBP [0~120]	DBP[0~80]	BMI[18.5~25]	FEV1_FVC[0.7이상]
BA00102	0.0444	0.0667	0 (정상)	0.0651
BA00109	0.1222	0.1111	0.024	0 (정상)
BA00110	0.1444	0.2222	0 (정상)	0 (정상)

Numeric 변수는 위와 같은 과정으로 정상으로부터의 거리를 쉽게 계산할 수 있었다. 문제는 HBSAG, RF와 같은 factor형 변수였는데 결론부터 말하자면 위와 같은 방식으로 처리를 하였다. 즉, 변수별로 정상인 level을 기준으로 정상 level 이면 0, 정상보다 한 단계 위면 1, 두 단위 위면 2로 데이터를 바꾸고, 이것 역시 MinMax Scaling을 진행하였다. 예를 들어, UBLD(요잠혈)은 Level이 Negative, Trace, Positive(1,2,3)의 5개로, 정상기준은 Negative(음성)이다. 따라서 Negative는 0, Trace는 1, Positive 3 (양성 3)은 4로 설정을 하였고 이를 Scaling 함으로써 0 ~ 1사이의 값으로 맞춰주었다. (0, 0.25, 0.5, 0.75, 1)

분석의 초기에는 factor 변수로 위와 같은 방식으로 처리하지 않고 factor 상태에서 "Gower" method를 사용하여 뒤의 군집분석에 사용될 거리를 구하려 했다. 그러나 곧 이 방식은 우리 데이터 적합하지 않다고 판단했는데, 우리 factor 변수의 경우 levels에 분명히 '순위'가 매겨진 형태가 존재했기 때문이다. 다시 말해, 우리는 factor 변수 내 각 level들 간의 거리를 동일하게 보는 것이 아니라 정상수치를 기준으로 더욱 부정적인 level은 거리가 더 멀어져야 하는 것이다. 위의 UBLD(요잠혈) 변수를 다시 보자면 Negative(음성) 정상으로부터 Positive 3

은 Positive 2보다, Positive 2는 Positive 1보다 더욱 좋지 않은 상태임을 알 수 있다. 따라서 이를 고려하기 위해선 factor 변수 역시 numerice 변수와 마찬가지로 정상과의 거리로 바꾼 후 정상수치(원점)과의 거리를 구할 필요가 있었다.

위의 모든 전처리 과정이 끝나면 각 obs는 전체 변수 개수만큼의 차원으로 벡터화 시킬 수 있다. 정상수치 범위 내에 있다면 해당 변수의 값은 0이므로 모든 변수에서 정상이라면 그 점은 완전정상(완전건강)인 포인트고 이는 기하적으로 표현하면 좌표의 원점에 해당할 것이다. 따라서 정상범위와의 거리로 값을 표시한 후에는 해당 obs가 원점으로부터 얼마나 떨어져 있는지, 즉 distance를 구하여 완전 정상수치와의 거리를 구하는 것이 가능해졌다.

. 아래는 지금까지의 과정과 Distance 아이디어를 보다 쉽게 이해할 수 있도록 시각화한 자료이다. X, Y, Z 축에는 각각 GLU, LDL, MCH가 들어가 있다. 각 값들은 위의 Reshape 과정을 거쳐 정상범위로부터 거리를 구한다음 0~1로 스케일링 되었다. 원점에는 모든 수치가 정상인 완전정상 Point가 있으며 원점으로부터 멀어질수록 색이 붉게 변하도록 설정했다. 하얀색 원점(0,0,0)으로부터의 거리가 멀수록 건강이 나빠지는 것이다. 시각화를 위한 샘플은 전체 데이터 중 300개를 랜덤 추출하여 사용했다.

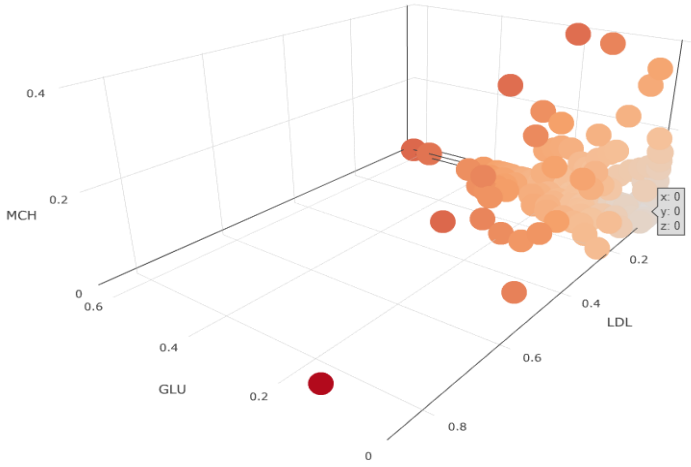

Figure 4 distance 시각화

3.2. 군집분석을 통한 그룹 생성

위의 과정에서 데이터 전처리를 모두 마무리하였고, 이제 그 데이터를 활용하는 일만 남았다. 즉, 각각의 obs가 원점으로부터 거리가 얼마나 떨어져 있는지를 계산하여 그 값을 더하면 그것이 바로 그 사람의 건강상태를 나타내는 지표이다. 그런데 한 가지 더 고려할 점이 존재했는데 그것은 앞서 말한 논점1과 관련된 부분이다. 각각의 변수를 독립적으로 보는 것은 인체의 메커니즘을 전혀 고려하지 못한 방법이라고 생각을 하였는데 우리가 신체 내부에 대한 모든 것을 알 순 없지만 상당히 밀접하고, 연관이 있는 하나의 유기체라는 것은 누구나 아는 사실이기 때문이다. 그렇다면 혈액 검사는 혈액 검사끼리, 폐기능은 폐기능 검사끼리 구분한 후 정상수치로부터 거리를 구하면 될까? 이에 대한 대답 역시 "No"이다. 앞서 말했듯이 인체 내부의 메커니즘을 우리가 다 알 수는 없고 실제로 혈액검사 A가 면역 검사 B와 상당 수 연관이 있을 수도 있다. 이를 고려하지 않고 정상수치를

구하게 되면 A와 B를 각각 독립적으로 구한 거리는 A와 B를 함께 묶어 구한 거리보다 클 수 밖에 없고 이는 생체나이가 그만큼 중첩되어 증가된다는 것을 의미한다.

A	B	Distance
c(0,0), c(1,1)	c(0,0), c(1,1)	$2\sqrt{2}$

A, B	Distance
c(0,0,0,0), c(1,1,1,1)	2

따라서 우리는 변수를 각각의 항목으로 볼 것이 아니라 연관이 있는 변수끼리 묶어서 볼 필요성을 느끼게 되었고, 계층 군집 분석을 통해 이를 구현하기로 하였다. 일반적인 클러스터링은 관측치끼리의 거리를 바탕으로 가까운 거리의 관측치는 같은 클러스터로 먼 거리의 관측치는 다른 클러스터로 구분을 해주는 기법이다. 그러나, 우리는 이 분석에서 관측치를 구분하는 것이 아니라 연관이 있는 변수를 구분하고 싶은 것이었기에 다음과 같은 방법을 이용하였다. 먼저 일반적인 클러스터링처럼 dist 함수의 Euclidean 메소드를 이용하여 관측치끼리의 거리를 구하였다. 우리가 앞선 전처리 과정에서 정상인 0을 기준으로 건강이 악화되면 숫자가 커지도록 전처리를 해주었기에 Euclidean 거리르 사용하여도 무방하다고 판단하였다. 그리고 다음의 과정이 중요한데 hclust의 경우 row 단위로 클러스터링을 진행하게 된다. 즉 dist를 그냥 적합시킬 경우 이는 변수간의 클러스터가 아닌 관측치간의 클러스터가 되는 것이다. 따라서 우리는 그냥 dist값이 아닌 이에 transpose를 취함으로써 row에는 변수가, column에는 각각의 관측치 값이 들어가도록 설정을 해주었고, 이를 통해 계층 군집을 나눔으로써 변수간의 클러스터를 설정할 수 있었다. 아래의 그림은 남자를 대상으로 계층 군집을 진행한 후 볼 수 있는 덴드로그램을 시각화한 것이다.

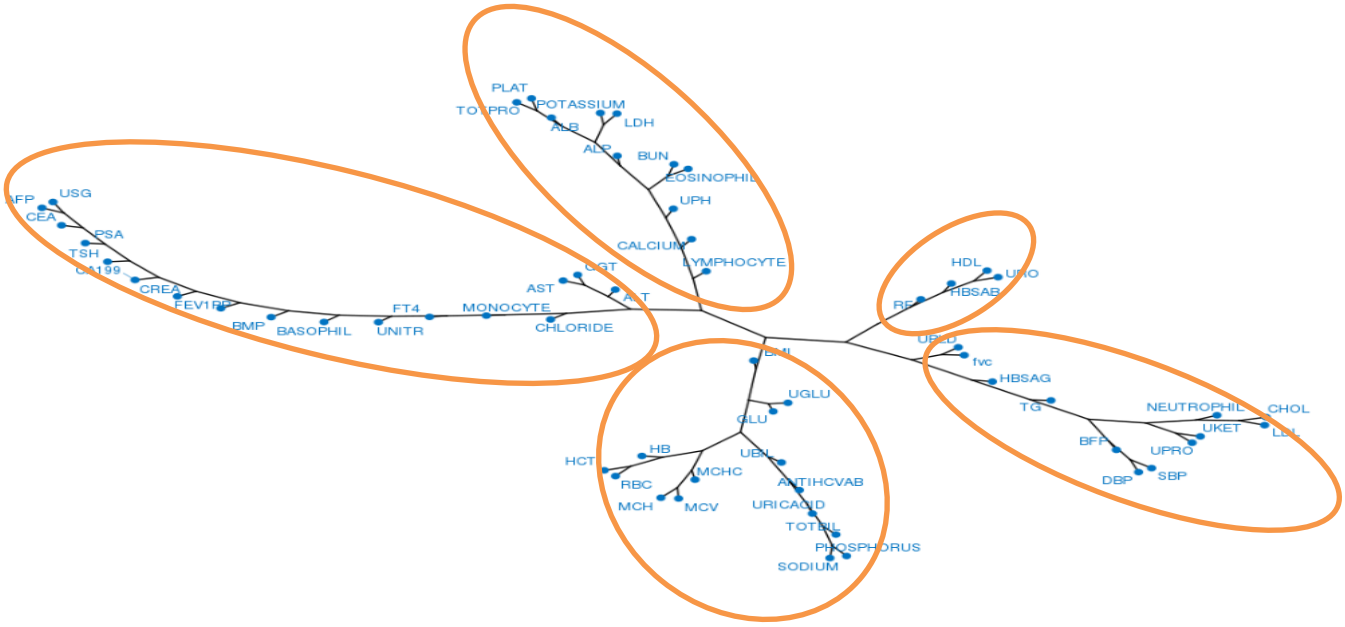

Figure 5 남자 계층 군집 덴드로그램

여기서 흥미로운 점은 우리가 혈액 검사, 종양 검사, 면역 검사 등의 변수들을 따로 구분하지 않아도 대체로 이 변수들이 모여 있다는 점이었다. 이는 대체로 검사 유형에 따라 검출되는 노화의 정도가 비슷함을 의미한다. 덴드로그램에 따라 그룹을 5개로 분류를 하였는데 그룹1과 혈액과 갑상선 기능, 그룹2는 혈액, 그룹3은 BMI, GLU(공복혈당), UGLU(요당)과 같은 체내 당 수치, 그룹 4은 요검사의 항목들이 대체로 모여있었다. 허나 우리는 이보다 더욱 나아 다른 검사일지라도 연관이 있는 변수들을 함께 묶어줌으로써 우리가 알지 못한 신체의 메커니즘을 조금이라도 반영하기 위해 노력하였다. 여자의 경우 역시 아래 그림과 같이 대체로 비슷한 양상을 띄웠으며, 마찬가지로 그룹을 5개로 나누어 분석을 진행하였다.

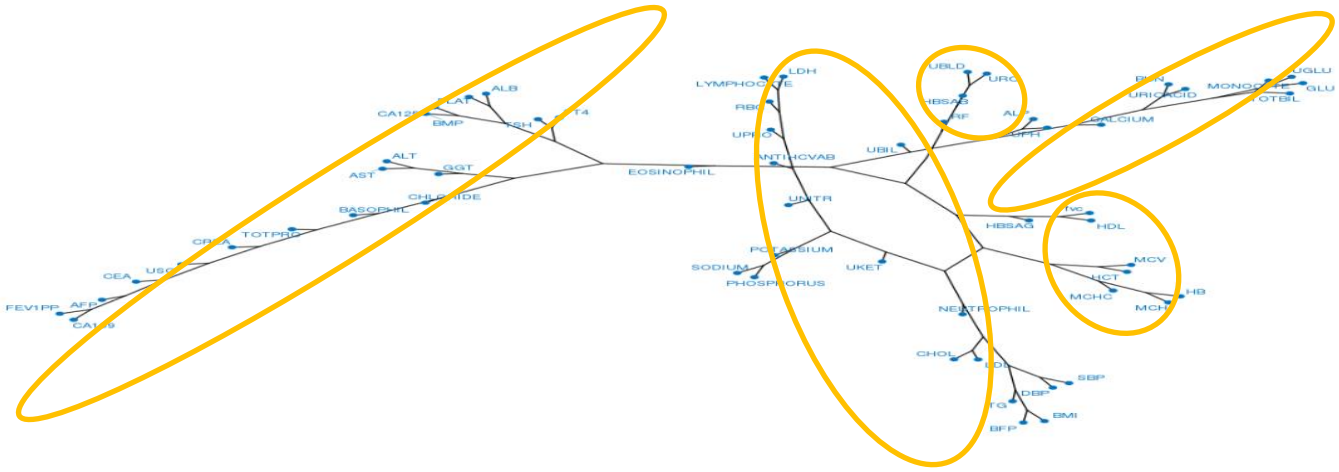

Figure 6 여자 계층 군집 덴드로그램

3.3. 건강지수 생성

이제 각각의 그룹으로 나누어 원점으로부터의 거리를 각각 구한 뒤 그 합을 계산하면 된다. 그리고 이것이 바로 그 사람의 전반적인 건강상태를 나타내는 건강지표인 것이며 그 값이 클수록 건강 상태가 좋지 않음을 의미한다. 우리의 데이터 상으론 모든 평가에서 0점, 즉 모든 변수가 정상 범위 내에 속한 사람은 없었으며 남자의 경우 0.244~6.473, 여자의 경우는 0.08~5.884의 분포를 따른다. 그러나 우리가 정상을 0을 기준으로 비정상의 거리를 구하였기에 그 분포가 상당히 왼쪽으로 쏠린 비균형적인 분포였다. 따라서 이를 일반화하기엔 무리가 있다고 판단을 하였고, 상위 1%의 값은 이상치라고 판단을 하여 이 값들을 삭제해주었고 이를 -10 ~ 10의 값으로 스케일링을 진행하였다.

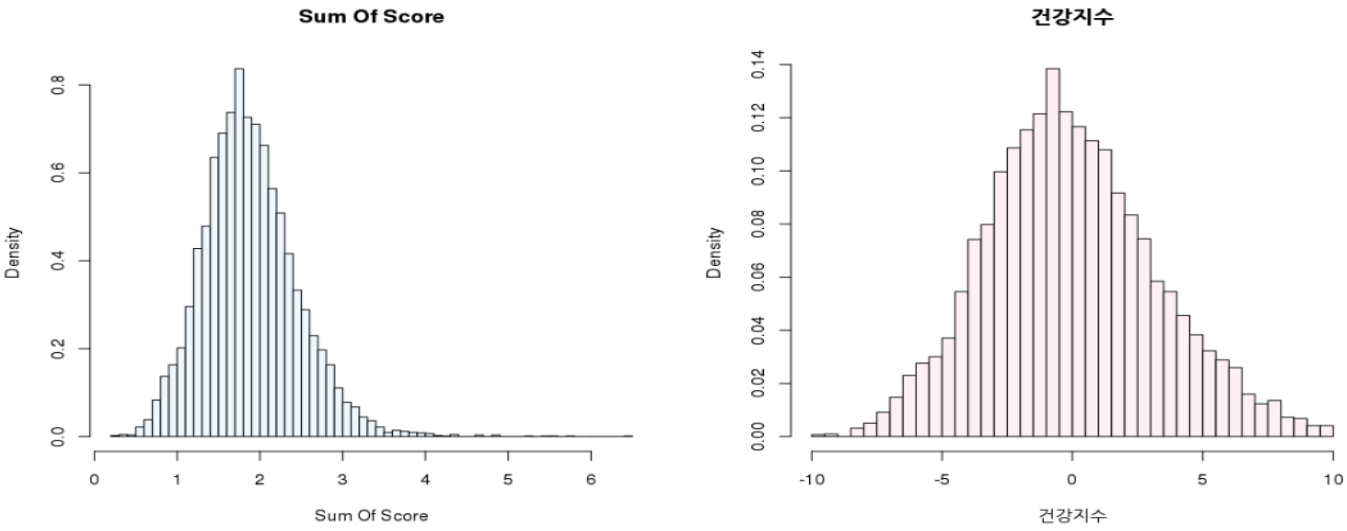

Figure 7 건강지수 스케일링(-10~10)

-10~10에 스케일링을 한 이유는 약간의 우리의 주관이 반영된 것이며 앞서 말한 두 번째 논점과 관련이 깊다. 현재 건강검진 데이터에는 대략 70개의 변수가 존재하지만 우리가 생각하기에 이 변수들이 인간의 노화의 일부분을 나타낼 뿐이지 모든 노화를 나타낼 순 없다고 판단을 하였다. 즉, 18세와 80세 두 명의 건강검진 수치가 모두 같더라도 이는 건강검진 내에서 파악된 변수 상에서만 수치가 같은 것뿐이지 이것만으로 두 사람의 생체나이를 같다고 단정지을 수 없다고 생각을 하였다. 따라서, 생체나이를 구할 때 절대적인 나이를 생성하는 것이 아니라 본인의 나이에서 건강이 좋으면 나이를 감소시키고, 건강이 좋지 않으면 나이를 증가시키는 방식으로 생체나이를 계산하는 것이 합리적이라고 판단을 하였고 우리가 생각하기에 생체 나이는 일반적으로 본인 나이의 +10에서 -10 사이라고 생각을 하여 스케일링을 진행한 것이다.

3.4. 선형회귀모델

이제 우리는 -10~10의 값을 갖는 건강지수라는 새로운 변수를 생성하였다. 그리고 이를 원래의 데이터에 추가를 하여 건강지수를 반응변수로, 그 외 건강검진 변수를 종속변수로 하는 선형회귀 모델을 만들 수 있다. 마지막에 건강 지수를 다시 회귀식에 적합한 이유는 다음과 같다.

1. 우리가 만든 생체 지수를 나이로 사용할 경우 생체나이에서 도출될 수 있는 값이 -10 ~ 10으로 한정되어 버린다. 이 역시 생체나이를 구하는 하나의 방안이 될 수는 있지만 실제로 본인의 나이보다 지나치게 건강하거나, 그 반대의 경우 생체 나이가 그 이상으로 차이가 날 수 있기 때문에 이는 문제가 있다고 판단을 하였다. 즉, 회귀모형을 적합함으로써 변수의 값이 지나치게 높거나 낮다면 -10 이하 혹은 10 이상의 값을 가질 수 있게 된다.

2. 대략 70개의 변수 중 인체의 노화와는 무관한 변수가 존재할 것이라 판단을 하였다. 즉 회귀 모델에 적합을 함으로써 그 안에 유의미한 변수의 영향력은 높이고, 무의미한 변수의 영향력은 낮게 해줌으로써 좀 더 명확한 생체나이의 기준을 제시하고자 하였다.

4. 결과 해석 및 시뮬레이션

Figure 8 estimate 상위 5개 변수

순위	변수	estimate
1	BMP	7.48
2	UNITR	7.22
3	FEV1PP	7.06
4	FT4	6.56
5	BASOPHIL	6.50

남자와 여자 각각 선형 회귀 모형에 적합을 시켰는데 처음에 적합 시 estimate 값이 NA가 뜨는 값이 있었다. 이는 다중공선성을 의미한다고 생각을 하여 선형 회귀의 단계적 방법(stepwise selection)을 변수를 제거해주었고 이렇게 하여 남자의 경우 58개, 여자의 경우 59개 변수를 사용한 선형회귀식 모델이 만들어졌다. 건강지표의 악화정도를 설명변수로, 건강지수를 반응변수로 설정을 하였기에 estimate 값이 큰 변수는 전반적인 건강상태와 관련이 있다고 해석이 가능한데 상위 5개를 추출한 결과는 옆의 그림과 같다. 반대로 estimate 값이 작은 변수는 전반적인 건강상태에 영향을 끼치기보단 특정 한 부위에 국한된 질병일 가능성이 높음을 의미한다.

그렇다면 지금까지의 알고리즘이 실제 상황에서는 어떻게 작동될 수 있고 사람들에게 어떻게 보여지는지 시뮬레이션 해볼 수 있겠다. 우선 해당 알고리즘이 작동되는 '생체나이 분석기계'를 가정하고 사람들이 검진받은 결과를 기계에 입력하는 것에서 시작된다. 다음 페이지의 그림에서 보이는 것처럼 A,B,C,D,E 다섯 명의 사람들이 자신들의 건강 검진결과를 기계에 입력했다고 생각해보자. 최초 데이터에 제공된 상태처럼, 다섯 명의 사람들은 건강검진 70여개의 변수들을 입력하게 된다. 물론 기계는 모델링될 때 제거된 변수들 (위의 결측치 처리 부분 참조)을 제거한 상태로 수치 Reshape 및 생체 Score 계산을 하게 된다.

BAID	CRAGE	SBP	DBP	BMP	BFP	BMI	FEV1PP
A	62	127	75	74.9	19.3	21.2	117
B	53	128	66	67.2	27.0	25.4	88
C	38	134	84	67.1	27.1	27.5	110
D	61	132	78	73.0	21.2	22.0	83
E	46	115	74	68.2	26.2	25.5	89

Data Reshape (정상수치로부터의 거리)

BAID	CRAGE	SBP	DBP	BMP	BFP	BMI	FEV1PP
A	62	0.08641975	0.00000000	0	0.00000000	0.00000000	0
B	53	0.09876543	0.00000000	0	0.19830028	0.02185792	0
C	38	0.17283951	0.05970149	0	0.20113314	0.13661202	0
D	61	0.14814815	0.00000000	0	0.03399433	0.00000000	0
E	46	0.00000000	0.00000000	0	0.17563739	0.02732240	0

군집별 계산 (원점으로부터의 거리)

```
      Group1    Group2    Group3    Group4    Group5
A 0.00000000 0.24038462 0.1087568 0.08641975 0.5270463
B 0.03870968 0.00000000 0.3811701 0.34766045 0.6167668
C 0.05585375 0.00000000 0.2682863 1.15060519 1.1180340
D 0.00000000 0.07692308 0.4135547 0.52259304 0.8112823
E 0.00678196 0.02711586 0.1596139 0.26617380 0.5121969
> mean(group1(test))> mean(group2(test)) > mean(group3(test)) > mean(group4(test))> mean(group5(test))
[1] 0.02090346     [1] 0.06699471     [1] 0.1534961     [1] 0.497129     [1] 1.125894
```

Bioage 산출 회귀식에 Fitting

```
        A         B         C         D         E
-4.863927 -1.776311  4.564618 -0.620937 -5.494829
```

NAME	A	B	C	D	E
AGE	62	53	38	61	46
Bioage	**57.136**	**51.224**	**42.565**	**60.379**	**40.505**

여기서 검사자들은 도출되는 자신의 생체나이를 알 수 있을 뿐 아니라 자신이 정상수준을 벗어나게 측정된 부분을 세부적으로 파악할 수 있다. 예를 들어 B와 C는 Group1에서 일반적인 평균치(0.02)보다 높은 수치를 보였으므로 ALT, GGT 등의 수치에서 이상을 보였을 수 있다. 따라서 간 검사를 추가적으로 받아보는 것이 추천된다. 또한 5명중 4명이 (B,C,D,E)가 Group 3에서 평균치(0.1535)를 초과했는데 이를 통해 BMI, GLU, UGLU 등에서 이상을 보였을 수 있다. 따라서 당뇨, 고혈당 검사 등을 추가로 받는 것이 추천된다. 이처럼 세부적인 건강도 파악을 통해 추가적인 검사가 추천되며 최종적으로는 자신들의 실제 나이를 감안한 생체나이를 확인하게 된다.

5. 의의 및 한계

생체나이의 범위를 -10에서 10 라고 정한 것은 우리의 어느 정도의 주관이 반영된 부분이라 볼 수 있다. 이 부분에 있어 고민을 많이 해봤으나 합리적인 기준을 구할 마땅한 방법은 존재하지 않았고, 우리의 주관대로 분석을 진행할 수 밖에 없었다. 그러나 앞서 말했듯이 이 값을 직접 사용한 것이 아닌 선형회귀의 반응변수로 설정하였다는 점에서 생체나이가 반드시 -10에서 10값을 가지는 것은 아니며, 특정 변수가 크게 증거하거나 혹은 감소한다면 생체 나이 역시 그에 맞게 10이상 혹은 10이하의 값을 가질 수 있기에 그런 한계점을 조금은 극복한 것이라 판단한다.

우리의 생체 나이 모델은 혈관, 면역, 갑상선과 같은 각각의 분야를 각각의 나이로 계산하기보단 신체의 메커니즘을 반영하여 노화의 정도가 중첩되지 않도록 설계를 하였다. 이로써 더욱 정확하고 체계적인 생체 나이를 계산할 수 있을 것이라 예상된다. 또한 단순히 생체나이가 얼마인지만 알려주는 것이 아니고, 각 건강 분야 중 어느 부분에서 내 건강지수가 나쁘고 추가검사 및 건강관리를 집중적으로 해야 하는지 구체적으로 제시한다. 따라서 보다 정확하게 자신의 건강함의 정도를 알 수 있다. 추가적으로 연령을 구분하지 않았다는 것 역시 큰 장점으로 작용하는데 앞서 말했듯이 우리의 데이터 역시 마찬가지로 18세와 80세 주변 연령대 사람의 표본은 굉장히 적은 것을 알 수 있다. 이 말은 이 연령을 구분하여 생체 나이를 구할 경우 이 연령대는 표본의 수가 적기 때문에 과소적합의 문제가 발생할 수 있고 이는 곧 옳지 않은 생체 나이 계산으로 이어질 수 있다. 반면 우리의 모델은 각 변수의 정상 수준에서의 거리를 사용하였기에 연령이 다를지라도 각각의 정상 기준으로부터 거리를 각각 구해주면 되며, 심지어 PSA와 CA125와 같이 남성에만 있는 변수 혹은 여성에게만 있는 변수를 이용하지 않는다면 성별이 다르더라도 모두 통합하여 사용이 가능하다. 즉, 같은 관측치라도 이전보다 우리는 더 많은 관측치를 이용하는 효과를 얻을 수 있는 것이며 기본정보와 직결된 건강 검진 데이터의 특성 상 데이터를 생성하는 것부터 유지 보수하는 것이 상당히 까다로운데 우리의 생체 나이 계산 방안은 이런 어려움에 큰 도움이 될 것이라 생각한다.