Final Project Report

2017-17498 Lee SangDeok, 2020-15709 Na SeungChan

2022-12-18

1. 개요

1.1. 분석 개요

'참여자의 건강 상태'를 사용하지만 그렇다고 해서 참여자의 정확한 건강 상태를 알 필요는 없는 다양한 사회과학 연구에서, '주관적 건강인지'는 측정하는 데 시간 및 비용 문제가 크게 발생하는 '실제 건강상태'를 대리하여 각 참여자의 건강상태를 추정하는 변수로 활용된다. 그러나, 상당한 측정오차를 가질 수 있어 이와 같은 변수에 대한 추가적인 고려가 필요하다는 점이 알려져 있다. 이와 같이 주관적 건강인지를 건강 지표로 활용하는 것에 대한 문제의식이 드러난 논문으로, "주관적 건강인식은 실제 건강상태의 유효한 대리변수인가"(최요한, 보건사회연구 36권 4호, 2016)를 찾을 수 있다.

또한, 본인의 '주관적 건강인지'에 영향을 미치는 여러 가지 변수는 당뇨, 암 등 병원에서의 구체적인 진단 프로토콜을 거쳐야만 본인의 주관적 건강인지에 영향을 미칠 수 있는 종류의 변수와, 흡연 여부, 음주 여부 및 음주량, 체형인지 등 구체적인 진단 절차 없이도 본인의 판단에 의해 주관적 건강인지에 영향을 끼칠 수 있는 변수가 존재한다. 특히, 20대~30대의 경우 각종 '건강하지 않은 생활 습관'이 곧바로 고혈압 등 진단으로 이어지지 않을 수 있으며, 이에 따라 주관적 건강인지 정도가 좋지 않음에도 불구하고 실제 진단에서는 그와 같은 차이가 나지 않을 수 있을 것으로 추측된다.

따라서, 해당 연구에서는 이와 같이 주관적 건강인지에 영향을 미칠 수 있는 변수의 영향을 평가하고, 이를 기반으로 주관적 건강인지가 실제 본인의 건강상태에 대한 도구변수로 활용될 수 있는지 검증하고자 한다. 특히, 20~30대의 경우 주관적 건강인지에 각종 진단이 미치는 영향력이 낮을 것으로 추측되기에, 해당 연령대를 위주로 검증한다.

1.2. 분석 사용 자료

해당 분석을 위해, 질병관리청이 작성한 [HN19_ALL 원시자료(2019년 기본DB)]를 활용하였다. 해당 데이터는 2019년 전국 단위로 진행된 건강 및 영양 관련 실태 조사 내용을 담은 것이다.

해당 data는 질병관리청 국민건강영양조사 홈페이지(https://knhanes.kdca.go.kr/knhanes/main.do) 에서 다운로드받을 수 있는 데이터로, 해당 분석에서는 2022년 9월 13일에 해당 홈페이지에서 다운로드받은 데이터를 활용하였다. 해당 자료의 변수 설명, 조사 방법 등 구체적인 내용은 같은 사이트에서 다운로드할 수 있는 원시자료 이용지침서에 기술되어 있다.

해당 데이터는 CSV 등 R 내장 함수로 읽을 수 있는 데이터 형식이 아닌 SPSS 또는 SAS에서 처리된 파일의 형태로 배포된다. 따라서, 해당 데이터를 R에서 불러오기 위해 tidyverse의 haven 패키지를 사용하였다.

2. 데이터 분석

2.1. 분석 기초

해당 분석을 위해, modelA와 modelB의 두 가지 model을 제작하였다. modelA는 개인이 쉽게 알 수 있는 주관적 건강변수에 기반한 모델이고, modelB는 각종 유병 지표 등 객관적이 정교한 진단을 거쳐야만 확인될 수 있는 건강지표를 통한 모델이다.

우선, modelA와 modelB 각 모델의 예측 능력을 확인하는 것이 첫 번째 목표이다. modelB의 예측 능력이 좋지 않다면, 이는 각종 유병 지표 등 실제 건강 상태와 관계된 지표가 주관적 건강인지에 미치는 영향력이 제한됨을 의미한다. 이는 주관적 건강인지를 실제 건강상태의 대리변수로 활용하기 어려운 근거가 된다.

한편, modelA의 예측 능력이 좋다면 이는 주관적 건강상태 관련 지표가 실제로 주관적 건강인지에 영향을 끼치고 있다는 의미이다. 이 경우, 주관적 건강인지는 주관적 건강상태의 대표변수가 된다. 한편, modelA의 예측 능력이 좋지 않다면, 주관적 건강인지에 관한 추가적인 독립변수 조사가 필요함을 드러낸다.

또한, 해당 연구에서는 두 모델의 accuracy를 비교하고자 한다. 이를 통해, 두 모델 중 현실에 더 적합한 모델을 확인하고 추가적인 논의 가능성을 확인할 수 있을 것이다.

2.2. modelA : 주관적 건강인지 및 주관적 건강상태

이 절에서는 주관적 건강인지와 개인이 쉽게 알 수 있는 주관적 건강 관련 변수들에 기반해 변수들의 관계를 살펴본다. 개인의 생활 습관과 관련되거나, 간단한 측정으로 자신의 상황을 알 수 있어 별도의 진단 등 절차 없이도 주관적 건강인지에 영향을 끼칠 수 있는 변수와 주관적 건강인지의 관계를 살펴본다.

modelA는 주관적 건강인지(D_1_1)를 다음의 변수에 의해 예측한 모델이다.

BO1 : 주관적 체형인식 1.매우 마른 편. 2. 약간 마른 편. 3. 보통. 4. 약간 비만. 5. 매우 비만.

BP16_1 : 주중 1일 평균 수면시간

BH1 : 건강검진 수진 여부 1. 예 2. 아니요.

BD1 : 평생 음주경험 1. 술을 마셔 본 적 없음. 2. 있음.

BD1_11 : 음주빈도 1. 최근 1년간 전혀 마시지 않았다. 2. 월 1회 미만 3. 월 1회 정도 4. 월 2~4회 5. 주 4회 이상

BS1_1 : 평생 흡연 여부 1. 5갑(100개비) 미만. 2. 5갑(100개비) 이상. 3. 피운 적 없음.

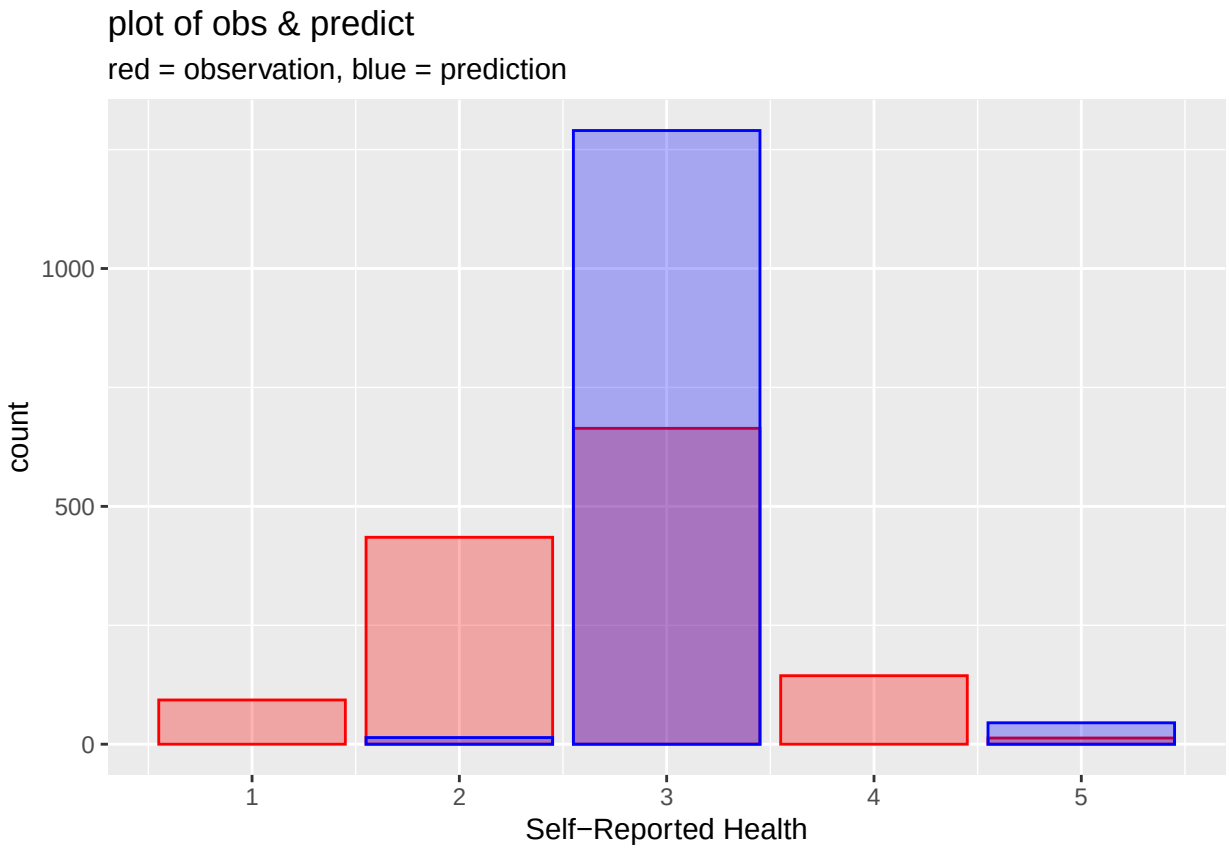
LQ4_00 : 활동제한 여부 1. 예. 2. 아니요.

단, 해당 자료에서 '음주 빈도'는 우선 해당 경험이 있는지 묻고, 그 경험이 있는 자를 한정하여 그 빈도를 묻는 방식으로 조사가 진행되었다. 이에 따라, 음주 또는 담배 경험이 없는 자를 0으로 두고, 그 밖의 자를 BD1_11(음주)의 값으로 재할당한 별도의 변수를 사용하였다.

이와 같이 선정된 변수들 중, 종속변수에 대한 설명력이 높은 독립변수를 찾아내기 위해 변수를 간소화하고 차수를 최적화하는 과정을 거친다. 이 과정에 의해 BO1, new_LQ4, BH1의 세 가지 변수에 대한 선형회귀모델을 적합하는 것이 적절함을 확인하였다.

이를 기반으로 선정된 변수들에 대해 LDA 기반 예측을 진행하고, 해당 modelA의 Accuracy를 측정하였다. Accuracy 측정을 위해 전통적인 최소제곱법을 활용하였다.

다음은 위의 예측 모델의 능력을 testing set을 통해 평가한 것이다. 이를 통해, 20~30대의 경우 쉽게 파악할 수 있는 습관 등에 의해서는 주관적 건강인지를 측정하기 어려움을 알 수 있다.



2.3. modelB : 주관적 건강인지 및 객관적 건강상태

이 절에서는 건강검진 등을 통해 측정해야 해 측정을 위해 비용과 시간이 많이 들거나, 진단되기 전까지는 자신이 해당 질병에 걸렸는지 알 수 없어 진단 전에는 주관적 건강인지에 영향을 미치지 않는 변수와 주관적 건강인지의 관계를 살펴본다.

modelB는 주관적 건강인지(D_1_1)를 다음의 변수에 의해 예측한 모델이다.

HE_HP : 고혈압 유병여부 1.정상 2.고혈압 전단계 3. 고혈압

HE_BMI : 체질량지수

HE_DM_Hba1c : 당뇨병 유병여부 1.정상 2. 전단계 3. 당뇨병

HE_HCHOL : 고콜레스테롤 0. 없음 1. 있음

HE_HTG : 고중성 지방혈증 0.없음 1.있음

HE_hepaB : B형 간염 0.음성 1. 양성

HE_hepaC : C형 간염 0.음성 1. 양성

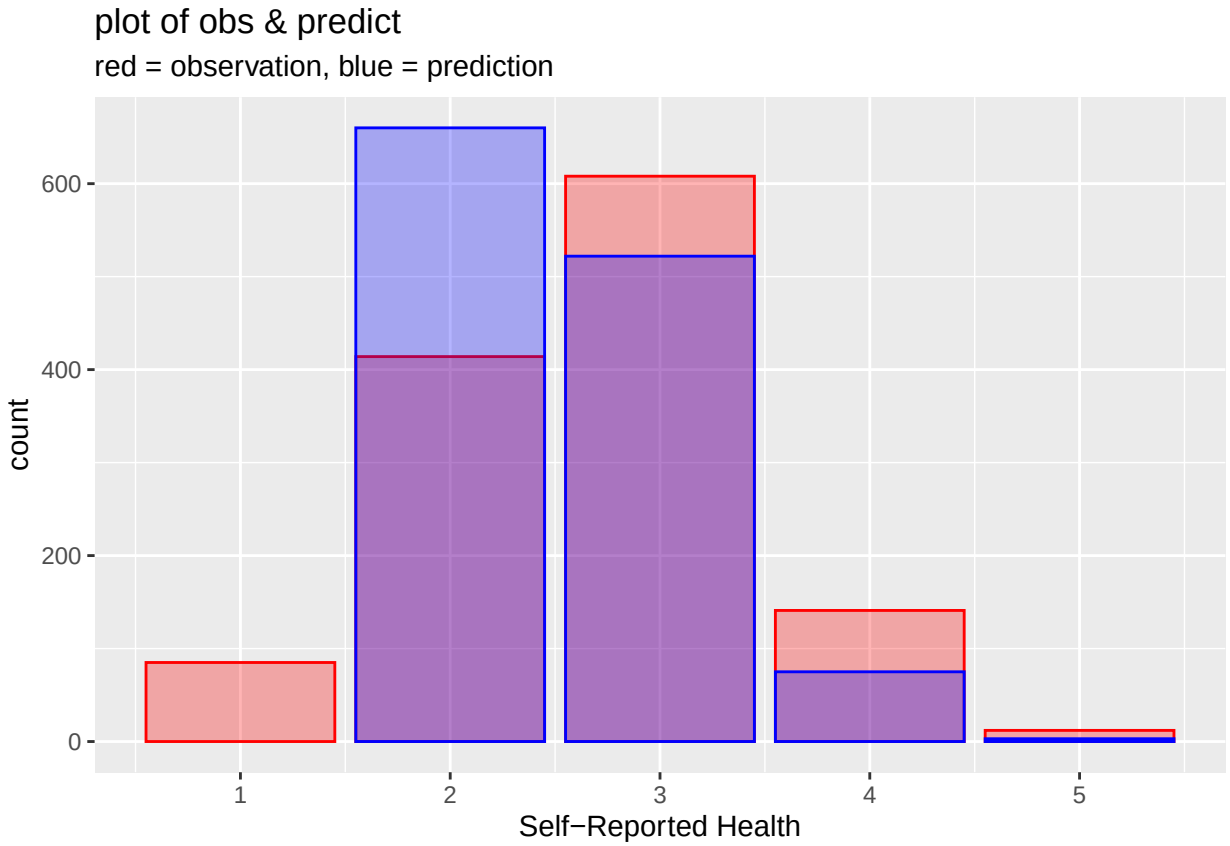
HE_anem : 빈혈 0.없음 1.있음

단, 청력 등 진단 결과 이상은 만 40세 이상에 한정되는 조사이므로 포함되지 못하였으며, 폐기능 검사의 경우 2020년 코로나19 확산 방지를 위해 중단되어 포함되지 못하였다.

이와 같이 선정된 변수들 중, 종속변수에 대한 설명력이 높은 독립변수를 찾아내기 위해 변수를 간소화하고 차수를 최적화하는 과정을 거친다. 이 과정에 의해 HE_HP, HE_BMI, HE_DM_HbA1c, HE_anem의 네 가지 변수에 대한 선형회귀모델을 적합하는 것이 적절함을 확인하였다.

이후, QDA를 기반으로 예측 모델을 적합하고, 해당 modelB의 Accuracy를 측정하였다. Accuracy 측정을 위해 전통적인 최소제곱법을 활용하였다.

다음은 위의 예측 모델의 능력을 testing set을 통해 평가한 것이다. 이를 통해, 20~30대의 경우 실제 건강상태에 의해서는 주관적 건강인지가 결정되기보다 실제 건강 상태에 비해 어느 정도 과장되어 측정된다는 것을 알 수 있다.



3. 분석 결과

3.1. 결론

1. 다음 정확도 비교에서 살펴볼 수 있는 것과 같이, 체감상 건강 상태를 기반으로 설정된 modelA의 accuracy는 modelB의 accuracy에 비해 높다. 이로 보아, 20~30대에서는 '실제 건강 상태'보다는 '체감상 건강 상태'에 의해 더 큰 영향을 받는다고 결론지을 수 있다.

```
## [1] 0.8071217
```

```
## [1] 0.8650794
```

2. 앞서 살펴본 바와 같이, 각종 연구에서는 '실제 건강 상태'를 측정하기 곤란할 때 '체감상 건강 상태'를 대신 측정하여 연구에 활용하기도 한다. 그러나, 이와 같은 모델 설정을 통해서는 정교하게 실제 건강 상태를 측정하기 어렵다는 한계가 존재하며, 이는 실제 유병 지표를 기반으로 한 예측 모델이든, 체감상 지표를 활용한 모델이든 벗어나지 못하는 한계이다. 이와 같은 한계를 보완할 추가적인 장치가 필요할 것이다.

3.2 한계점 및 제언

우선, 해당 연구에서는 20~30대를 중점으로 분석하였으나, 이와 같은 현상이 40대 이상에서도 보편적으로 발생하는지는 추가 연구가 필요한 주제이다. 앞서 살펴본 바와 같이, 20~30대에서는 주관적 건강인지에 주요 영향을

끼치는 변수가 실제로는 '실제 건강 상태'가 아닌 '체감상 건강 상태'에 가깝다. 그러나, 40대 이상에서는 이와 다른 결과가 충분히 일어날 수 있으며, 해당 연령대에서는 이 연구가 적용될 수 없다.

또한, 본 연구에서는 '주관적 건강인지'를 실제 건강 상태의 대리변수로 사용하는 것이 적합하지 않음을 확인하였으나 실제 주관적 건강인지에 영향을 미치는 것이 '주관적 건강 상태'인지는 입증되지 않았다. 즉, 주관적 건강인지가 어떤 변수에 의해 영향을 크게 받는지 예측 모델을 개발하기 위해서는 별도의 연구가 필요하다. 각 모델의 accuracy를 비교하는 과정에서 '각종 변수를 한 번에 주관적 건강인지의 독립변수로 사용'하는 모델이 사용되지 못하였는데, 이와 같은 모델을 통해 추가 연구가 가능할 것이다.

4. 참고문헌

주관적 건강인식은 실제 건강상태의 유효한 대리변수인가:주관적 건강상태(SRH)와 주관적 건강변화상태(SACH)의 비교, 최요한, 한국보건사회연구원, 보건사회연구 36[4], 2016.