한국의료패널 자료와 R을 이용한
노인의 의료비를 결정하는데 영향을 미치는 요인 분석

20160716 정보통계학과 박영빈

의료서비스에 대한 수요는 환자들이 의료서비스를 받은 후 건강이 개선될 때 얻는 만족감의 증가에서 발생한다. 이는 평균적으로 건강수준이 낮은 고령층일수록 의료서비스에 대한 수요와 의료비 지출이 높아진다는 것이라고 말할 수 있다. 따라서 인구 고령화가 심화될수록 국민의 의료비 지출이 늘어날 수 있음을 쉽게 예상할 수 있다. 이와 관련하여 본 논문은 개인의 의료비를 결정하는 데에 영향을 미치는 요인은 무엇인지 알아보고자 한다. 데이터 불러오기, 데이터 전처리, 그리고 다중회귀분석으로 모델링하는 모든 과정을 통계 패키지 'R'을 이용해 진행했다.

이 연구의 목적은 단순히 개인이 1년 동안 지출하는 의료비에 어떤 요인들이 영향을 미치는지 파악하는 데에 그치지 않는다. 의료데이터가 개인정보로 인식되면서 보안의 중요성과 우려가 커지고 있다. 한국의료패널 데이터는 자료 활용 동의서를 작성해야 데이터를 얻을 수 있다는데 의미가 있다. 그리고 코로나19와 4차 산업혁명 시대로 헬스케어 분야에 변화가 일어나고 있다. 의료서비스가 치료가 아닌 예방 중심으로 변할 것이며, 개인 맞춤형 서비스가 보편화 되면 우리나라 의료서비스의 만족도가 올라갈 것이다.

Ⅰ. 서론

1. 연구배경 및 수행절차

최근 50년간 우리나라의 고령화 속도가 경제협력개발기구(OECD) 37개국 중 가장 빠르다는 분석이 나왔다. 전국경제인연합회 산하 한국경제연구원은 1970년~2018년 OECD 통계를 분석한 '저출산·고령화 추세 국제비교와 정책시사점' 분석을 통해 이와 같이 밝혔다. 1970년~2018년 우리나라의 고령화비율 연평균 증가율은 3.3%로 OECD에서 가장 높았다. 우리나라는 2000년 고령화사회(고령인구 비중 7% 이상)로 진입한 이후 18년만인 2018년 고령사회(고령인구 비중 14% 이상)가 됐다. 이런 추세라면 2026년에 초고령사회(고령인구 비중 20% 이상) 진입이 유력하다고 OECD는 예상했다.

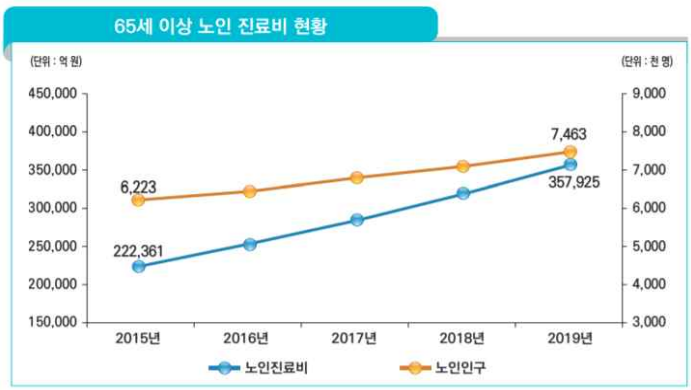
<그림 1.1.1> 우리나라 고령화비율 및 OECD 순위 추이 (자료 : 한국경제연구원)

또한 유병장수 시대가 도래 하면서 노인 인구와 1인당 노인 의료비가 매년 늘고 있다. 노인층의 연소득과 노후생활비를 고려했을 때 경제적 빈곤을 가중시키는 요인으로 작용한다는 분석이다. 건강보험심사평가원에 따르면 지난 2019년 기준 우리나라 65세 이상 인구의 1인당 평균진료비는 연간 491만 원으로 집계됐다. 이는 전체 인구 1인당 평균진료비 168만 원의 약 2.9배에 달한다. 노인인구 증가는 노인진료비 증가로 이어져 2019년 노인진료비는 35조 7,925억 원으로 2015년과 비교하면 1.6배 증가했음을 알 수 있다.

<그림 1.1.2> 우리나라 65세 이상 노인 진료비 현황
(자료: 2019 건강보험통계연보)

구분	2015년	2016년	2017년	2018년	2019년
전체인구(천 명)	50,490	50,763	50,941	51,072	51,391
65세 이상 인구 (천 명)(비율, %)	6,223 (12.3)	6,445 (12.7)	6,806 (13.4)	7,092 (13.9)	7,463 (14.5)
65세 이상 진료비 (억 원)(증가율, %)	222,361 (11.4)	252,692 (13.6)	283,247 (12.1)	318,235 (12.4)	357,925 (12.5)
노인 1인당 연평균 진료비 (천 원)	3,620	3,983	4,255	4,568	4,910
전체 1인당 연평균 진료비 (천 원)	1,149	1,275	1,391	1,528	1,681

<그림 1.1.3> 우리나라 65세 이상 노인 진료비 현황
(자료: 2019 건강보험통계연보)

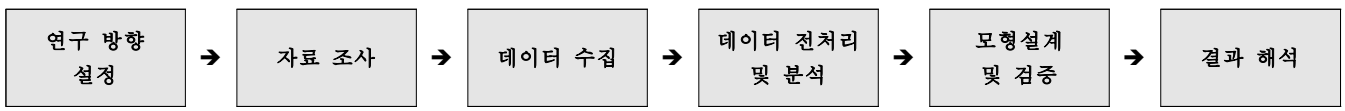


이처럼 급속한 고령화 사회의 진전, 기대수명, 만성질환의 증가 등으로 의료비는 앞으로도 더욱 증가할 것으로 보인다. 그러므로 급증하는 의료비는 장기적으로 경제사회에 부담이 될 것이다. 이와 같은 문제로 정부에서는 노인과 취약계층의 의료비 부담을 줄인다는 목표로 일명 '문재인 케어'로 불리는 건강보험 보장성 강화대책이 시행되고 있다. 효율적인 정책개발은 미래에 국민의료비가 안정화되고 향후 얼마나 증가할지를 예측하는 작업일 것이다.

연구절차는 Kaggle에서 제공하는 데이터를 통해 연구의 방향성을 잡아 이와 관련된 우리나라의 현황을 알 수 있는 통계자료를 찾고 유사연구를 조사한다. '한국의료패널'에서 제공하는 데이터 셋을 이용해 데이터를 가공하고 전처리하여 분석하기 용이하게 바꾼다. 그 다음으로 R을 통해 다중회귀분석을 실시하여 모형을 설계, 검증한 후 결과를 해석하는 과정으로 연구를 마무리 한다.

다중회귀분석은 종속변수가 1개, 독립변수가 2개 이상인 경우를 분석대상으로 하는 회귀분석 방법 중 하나로 독립변수의 변화에 의해 종속변수가 어떻게 변화하는지를 검증하는 분석 방법이다. 본 논문은 '개인지출의료비'라는 종속변수 1개, 그리고 종속변수에 영향을 줄 수 있는 2개 이상의 독립변수를 사용했다. 따라서 다중회귀분석이 사용하기 적합하다고 판단했다.

<그림 1.1.4> 연구방법 순서도



연구목표는 다음과 같다. R을 이용해서 의료비 예측 회귀모델을 만들어 유의하고 합리적으로 설명이 가능한지 판단한다. 그리고 한국의료패널 데이터가 갖는 의미를 찾는 것이다. 개인 정보 유출 문제가 일어나면서 이에 대한 우려와 보안유지의 중요성이 커지고 있다. 여기서 한국의료패널은 어떤 보안 시스템 하에서 자료를 제공하는지 알아보는 것이다.

2. 사전조사

1) Kaggle - 'Insurance Premium Prediction'

가장 먼저 실시한 사전조사는 Kaggle의 'Insurance Premium Prediction' 데이터 셋이다. Kaggle에서 제공하는 데이터의 레코드 수는 1,338개, 칼럼 수는 7개로 '나이', '성별', 'BMI', '자녀수(부양가족 수)', '흡연여부', '거주지', '의료비'로 구성되어 있다. 그리고 이것을 R로 분석한 네이버 블로그를 찾을 수 있었다. 작성자가 직접 분석한 과정을 게시글로 보여주고 있다. 먼저 그는 각 데이터를 탐색하여 R에서 다중선형회귀분석을 실시했다. 모델 성능은 회귀모형의 유의성, 회귀 계수의 유의성, 수정된 결정계수 총 3가지를 두고 판단했다. 유의성 판단 시, 유의수준을 0.05로 선정하여 연구를 진행했다. 마지막으로 수정된 결정계수를 통해 모형의 설명력을 판단했다.

또한 그는 모델을 개선하는 방안에 대해서도 생각했다. 그의 개선 방안은 총 3가지이다. 첫 번째는 연령에 따른 의료비 지출은 일정하지 않으므로 연령에 대한 비선형 항목(age^2)을 추가했다. 따라서 높은 차수 항을 회귀 모델에 추가했다. 두 번째는 BMI이다. BMI 수치를 30(비만 단계) 이상과 미만으로 나누어 더미변수로 전환했다. 세 번째는 흡연과 비만의 상호작용을 가정하여 이를 회귀분석 변수에 추가했다.

개선된 모델의 성능을 평가한 결과 회귀모형은 그 전과 동일하게 통계적으로 유의하다고 나타났다. 또한 각 변수의 통계적 유의성을 판단하는데 비선형 항목인 age^2변수가 통계적으로 유의함을 알 수 있었고 BMI(30)와 흡연자 사이의 상호작용은 엄청난 영향을 미치고 있음을 알 수 있었다. 흡연만 했을 때는 $13,404인데, 여기에 BMI가 30이상인 경우 추가적으로 $19,180 의료비 지출이 발생함을 알 수 있었다. 수정된 결정계수는 이전에는 75%였지만 개선된 모델의 경우 87%로 설명력이 증가했다.

2) 유사 연구 및 보고서

한국의료패널 데이터를 사용해서 연구한 유사 논문에서는 어떤 데이터와 어떤 변수를 사용했는지 조사했다. 참고한 논문 중, '세대별 의료비 지출에 영향을 미치는 요인 분석(황연희, 2011)'은 '한국의료패널 연간통합 데이터 베타버전 1.1.1' 과 '2009년 한국의료패널 본조사 베타버전 1.1.1'을 활용했다. 의료비에 영향을 미치는 요인을 인구학적 변수, 사회경제적 변수, 건강상태 및 건강행위를 나타내는 변수로 나누었다.

인구학적 변수에는 성별, 교육수준, 세대구성 변수를 사용했다. 우리나라 노인들의 가구 유형은 점점 노인 부부가구 또는 단독가구로 유형이 축소되는 경향을 보기 위하여 노인 단독가구, 부부가구, 그리고 노인 부부와 자녀가 함께 사는 가구, 기타 가구로 범주화했다.

사회경제적 변수는 의료보장 형태, 경제활동 여부, 소득 수준, 거주 지역, 주택소유 여부를 변수로 사용했다. 의료보장 형태는 직장가입자와 지역가입자로 구성된 건강보험가입자와 의료급여 1종, 2종, 국가 유공자 등의 의료급여수급자로 구분했다. 그리고 미가입자, 자격 상실자와 급여가 정지된 사람은 분석에서 제외했다. 소득 수준은 각 가구의 해당 가구원의 1년 동안 근로소득의 합과 자산소득, 사회보험을 포함한 모든 가구의 소득을 통해 가구당 소득을 구했다. 이를 6개의 구간으로 범주화하고 그것을 각 해당 가구의 가구원에 적용하여 분석했다. 1,000만원 단위로 최저소득은 1,000만 원 미만, 최고소득은 5,000만 원 이상이다. 거주 지역은 서울, 경기의 수도권에 거주하는 사람과 그렇지 않은 사람으로, 주택소유 여부는 자가 주택을 소유했는지 여부에 따라 이분했다.

건강상태 및 건강행위를 나타내는 변수로는 만성질환의 수, 흡연, 음주, 장애 여부, 우울 경험 여부 변수를 사용했다. 지금까지 살면서 총 5갑(100개피) 이상의 담배를 피운 사람을 흡연자로 보고, 그렇지 않은 사람은 비흡연자로 정의했다. 음주 여부 변수는 지난 일 년 간 한 달에 한번 이상 술을 마신 사람 중에서 소주 7잔(맥주 5캔), 여성의 경우 소주 6잔(맥주 3캔)을 과음의 기준으로 한 달에 1회 이상 과음한 사람과 그렇지 않은 사람으로 구분했다. 우울 경험 여부는 최근 1년 동안 2주 이상 연속으로 일상생활에 지장이 있을 정도로 많이 슬펐거나 불행하다고 느낀 적이 있는 지 여부의 변수이다.

그리고 노인 연령에 따라 베이비 붐 세대(1955년생~1963년생), 준고령 세대(1944년생~1954년생), 고령 세대(1943년생)로 세대를 구분했다. 이유는 응급, 입원, 외래 의료서비스 이용 및 의료비 지출에 영향을 미치는 요인을 분석하고 세대별 의료비 지출 관리를 위한 시사점을 모색해 보고자하기 위함이었다. 해당 연구에서 쓰인 독립변수는 14개이다. 의료비는 서비스 이용 시 수납한 금액으로 정의했다.

베이비붐 세대에는 성별, 의료보장형태, 소득수준, 만성질환수의 의료이용 및 의료비 지출에 영향을 미치는 요인으로 나타났고, 준고령 세대에서는 세대구성이 중요하게 의료이용 및 의료비 지출에 영향을 미쳤다. 마지막으로 고령세대에서는 교육수준과 의료보장형태, 그리고 만성질환의수가 의료이용 및 의료비 지출에 영향을

미치는 요인인 것으로 나타났다. 결과적으로 세대별로 소득수준과 만성질환의수가 공통적으로 의료이용 및 의료비 지출에 영향을 미치는 요인으로 작용했고, 의료 이용행태 별로는 의료보장형태와 만성질환 수가 의료이용 및 의료비 지출에 영향을 미치는 것으로 나타났다.

필자의 연구에서 의료서비스에 영향을 미치는 요인은 주로 만성질환개수와 신제활동제한 변수였다. 해당 연구와 세대별로 나누었다는 것은 다르지만 건강상태 및 건강행위에 대한 요인이 영향을 미친다는 공통점이 있다. 해당 연구는 세대별로 세 가지 의료이용과 의료비 지출 영향 요인을 분석한 것이 필자의 분석과 차이점이다.

'한국의료패널로 본 한국 노인들의 의료이용 및 의료비 지출(황연희, 2011)'은 '한국의료패널 2008년 통합 데이터 베타버전 1.1.1'과 '한국의료패널 2009년도 베타버전1.1.1'을 사용했다. 해당 연구는 '인구·경제적 요인'과 '건강상태 및 건강행위'에 따른 요인'으로 나누어서 분석했다. 그리고 각 요인별로 의료서비스를 이용한 인원수, 빈도, 전체 평균의료비를 파악했다. 앞선 연구(세대별 분석)와 차이점은 분석 방법이다. 세대별로 분석한 연구에서는 회귀분석으로 의료비에 영향을 미치는 요인을 분석했다.

인구·경제적 요인에는 성별, 나이, 세대구성, 의료보장, 경제활동, 소득수준, 주택소유가 있다. 나이를 65세부터 85세 이상까지 5년을 단위로 총 5개의 범주로 나누었다. 세대구성은 단독가구, 부부가구, 부부자녀가구, 기타가구이다. 의료보장은 건강보험 가입자와 의료급여 수급자로 나누었다. 경제활동은 하는 사람과 그렇지 않은 사람으로 구분된다. 소득수준은 '소득 없음'을 포함해서 1,000만 원 단위로 6구간으로 나누었으며 총 7개 범주로 구성했다. 주택소유는 자가를 갖고 있는 사람과 전세, 월세 등 2 범주로 나누었다.

건강행위 및 건강상태에 따른 변수에는 만성질환개수, 흡연, 음주, 장애, 신체활동여부, 주관적 건강상태로 6개로 구성되었다. 만성질환개수는 0개, 1개, ..., 5개 이상으로 총 6개의 범주로 구성된다. 흡연과 장애, 신체활동 제한은 '그렇다'와 '그렇지 않다'로 이분되었다. 흡연은 지금까지 총 5갑(100개피)이상의 담배를 피운 노인을 흡연자, 그렇지 않으면 비흡연자로 정의했다. 신체활동에서 제한이 있다는 것을 현재 건강상의 문제나 신체 혹은 정신적 장애로 일상생활 및 사회활동에 제한을 받고 있다고 정의했다. 주관적 건강 상태는 0~100점으로 이루어진 연속형 변수이지만 해당 연구에서는 네 구간으로 나누고 모름/무응답을 포함해 총 5개 범주이다. 이 문항은 2013년까지 조사에만 존재했고 이후에는 수치가 아닌 1(매우 좋음)~5(매우 나쁨), -1(해당사항 없음), -9(무응답/응답거절)로 나눈 문항을 사용하고 있다.

의료비는 응급, 입원, 외래 의료이용 시 지출한 본인부담금과 처방약값을 더한 비용을 기준으로 했다. 따라서 총 사용한 변수의 개수는 12개이다.

인구·경제적 요인에서 의료이용 경험이 있는 의료이용자의 연평균 의료비는 남성이 여성보다 높았다. 연령별로는 65~60세까지의 노인이 의료이용을 가장 많이 했으나, 평균 의료비 지출은 오히려 70~75세에서 가장 높게 나타났다. 의료이용을 한 노인은 부부가구에서 가장 많았으나 의료비 지출은 부부 및 자녀가 함께 사는 경우가 연평균 94만 원으로 가장 높았다. 필자의 연구에는 의료이용을 한 세대구성은 부부가구에서 가장 많은 것은 응급, 입원, 외래 모두 동일하나 의료비 지출을 많이 한 세대는 의료서비스마다 결과가 달랐다.

건강행위 및 건강상태에 따른 요인에서 만성질환의 수가 많을수록 전체 노인 중에서 의료이용을 한 노인의 비중이 큰 것을 알 수 있으며, 만성질환이 6개 이상인 노인의 평균 의료비는 120만원으로 가장 많았다. 필자는 만성질환수를 '0개', '1개', ..., '5개' 이상으로 나누었고, '5개 이상'인 범주에서 의료이용을 가장 많이 했다는 결과가 동일했다. 흡연유무의 경우 비흡연자가 흡연자보다 의료이용과 의료비 지출을 더 많이 한다는 결과가 나왔다. 이는 흡연유무에 따른 의료이용의 차이를 흡연으로 인해 발생한 질병의 치료를 위한 의료이용 보다는 건강에 대한 인식의 차이를 대리하는 변수로 해석했다. 필자도 같은 결과를 얻었으며 해석이 타당하다고 판단해서 이를 인용했다. 신체활동에 제한이 있는 노인과 제한이 없는 노인의 평균 의료비 지출이 약 33만 원 정도 차이가 나므로 신체활동 제한 여부는 의료이용과 높은 상관관계가 있다고 볼 수 있다. 필자의 연구도 비슷한 결과이다. 입원의료비에 가장 큰 영향을 미치는 요인에서 '신체활동제한여부' 변수가 독립변수들 중에서 15%로 가장 큰 영향력을 차지했다.

'한국의료패널로 본 의료이용 및 본인부담의료비 지출(정영호, 2011)'은 한국의료패널의 2008년 상반기 조사, 2008년 하반기 조사, 2009년 본 조사의 데이터를 결합하여 2008년 통합자료를 구축했다.

앞서 언급한 황연희(2011)의 연구와 정영호(2011)의 연구에 차이점이 있다면 황연희(2011)의 연구에서는 65세 이상 노인의 요인별 의료이용과 의료비 지출 현황에 대한 분석과 의료기관 종별 응급, 입원, 외래 이용률, 진단코드별 응급, 입원, 외래 의료이용 실태를 조사했다. 정영호(2011)의 연구는 황연희(2011)의 분석 내용에 입원 시 간병인에 따른 지출액, 가구소득에 따른 가구당, 가구원 당 지출의료비 등을 추가해서 분석했다. 그리

고 2008년과 2009년의 현황을 파악했다. 황연희(2011)는 응급, 입원, 외래별 평균 의료비를 각각 조사하지 않고, 각 변수에 대한 의료이용자 전체의 의료비를 조사했다. 반면 정영호(2011)는 응급, 입원, 외래별 각 변수에 대한 2008년, 2009년의 평균 의료비를 조사했다. 독립변수는 성별, 연령, 혼인상태, 장애여부, 경제활동, 가구소득, 총 6개의 변수이다. 의료비는 법정본인부담과 비급여를 합한 비용으로 정의했다.

나이를 0세부터 80세 이상까지 10년 단위로 나누고 65세 미만과 65세 이상으로도 나누어 총 11개의 범주를 만들었다.. 혼인상태는 유배우자와 배우자없음, 의료보장은 건강보험, 의료급여, 기타로 나누었다. 장애여부는 있음과 없음, 경제활동은 있음, 없음, 14세 이하로 분류했다. 가구소득은 999만 원 이하와 1,000만 원에서 5,000만 원까지 1천만 원 기준으로 4등분하고 5,000만 원 이상까지 총 6개 범주로 나누었다.

해당 연구와 필자의 연구의 차이점은 경제활동여부와 장애여부에 있었다. 정영호(2011)의 결과는 경제활동을 하는 사람들이 적은 차이로 의료비 지출을 더 많이 하지만 필자의 결과는 경제활동을 하지 않는 사람이 의료비 지출을 더 많이 했다. 정영호(2011)의 결과는 장애가 없는 사람들이 의료비를 더 지출한다는 결과가 나왔지만 필자는 장애가 있는 사람이 지출을 더 많이 한다는 결과가 나왔다. 그리고 정영호(2011)는 전 연령을 조사했기 때문에 경제활동 변수에 '14세 이하' 범주고 있지만 필자는 65세 이상에 초점을 맞추었기에 필요 없는 범주이다.

가장 최근 자료인 '2016년 한국의료패널 기초분석보고서(Ⅰ)'(이하 '보고서Ⅰ')와 '2016년 한국의료패널 기초분석보고서(Ⅱ)'(이하 '보고서Ⅱ')에는 여러 변수들 간 빈도와 비율을 분석하면서 2016년도의 각 유형에 따른 현황을 보여주고 있다. 보고서Ⅰ은 주로 의료비 지출을 여러 요인들로 나누어서 분석을 했다. 의료비를 지출하는 이유를 '보건의료서비스', '의약품', '보건의료용품', '기타'로 나누어서 비율로 표현했다. 다음에는 응급 이용, 입원 이용, 외래 이용으로 나누어 각 의료이용 형태에 따른 이용 횟수와 지출비 등을 분석했다. 또한 만성질환과 중증질환의 의료비 지출에 대한 분석을 실시하여 각 질병에 따른 의료비 특성을 65세 이상 노인을 기준으로 분석하거나, 각 질병에 해당하는 환자 비율 추이 등을 나타냈다.

보고서Ⅱ의 분석 내용은 다음과 같다. '가구와 가구원 특성'을 가구원 수/세대 구성 추이를 보여주며 2016년의 세대 구성/가구원 수/경제활동 등을 분석했다. '질병 이환 현황 및 유병 상태'에서는 '질병 현황', '주요 만성질환 유병 상태', 그리고 '8개 각 질병에 대한 연도별 유병률 추이'를 성별·연령별, 성별·질환 개수별 등으로 나누어 나타냈다. '의료이용과 의료비 지출'에서는 가구(원)에 따른 의료 이용률, 이용횟수, 의료비 지출 추이에 대해서 분석했으며, 8개 각 질병에 대한 의료비와 복용률 등을 나타냈다. '건강 형태와 건강 수준'에서 주요 건강 형태를 '현재 흡연율', '고위험 음주율', '중증도 이상 신체활동' 총 3가지로 선정했고 성별로 3가지 행동 실천을 연도별로 나타냈다. 또한 복합적 건강 위험 행동에 대해 분석했는데, 흡연, 고위험 음주, 신체활동 기준으로 이에 해당되는 사람의 비율을 나타냈다.

'보고서Ⅰ/Ⅱ'와 본 논문에 차이점이 있다면, '보고서Ⅰ/Ⅱ'는 여러 변수들 간의 빈도와 비율을 파악한 결과를 시각화하여 차이점을 해석하는 것이 보고서의 주를 이룬다. 따라서 '데이터 탐색'에 대한 보고서라 해도 무방할 것 같다. 또한 각 분석 주제에 대한 결과가 서로 영향을 주지 않는 독립된 분석이라고 할 수 있다. 다시 말해 '질병 이환 현황 및 유병 상태'에 대한 분석결과와 해석이 그 다음 주제인 '건강 형태와 건강 수준'의 분석을 해석하는 데에 아무런 영향을 주지 않는다. 반면 본 본문은 '데이터 탐색'이 연구순서의 하나의 과정이며 모든 연구과정이 연결되어있다고 할 수 있다. 예를 들어 <그림 4> '연구방법 순서도'의 '데이터 전처리 및 분석'에서 실수를 하거나 빠뜨린 내용이 있다면 그 영향이 해당 단계에서 그치는 것이 아니라 다음 순서와 최종 결과에도 영향을 미칠 수 있다.

<표 1.2.1> 유사 연구와 본 논문의 차이점

	네이버 블로그	참고문헌Ⅰ	참고문헌Ⅱ	참고문헌Ⅲ	보고서Ⅰ	보고서Ⅱ	본 논문
분석 데이터	Kaggle	한국의료패널	한국의료패널	한국의료패널	한국의료패널	한국의료패널	한국의료패널
사용 프로그램	R	언급X	언급X	언급X	언급X	언급X	R
조사 대상자	전 연령	65세 이상 노인	65세 이상 노인	65세 이상 노인	전 연령	전 연령	전 연령
EDA 방식	변수 1개 빈도/비율	변수 1개 빈도/비율 (응급.입원.외래별)	변수 1개 빈도/비율	변수 1개 빈도/비율	여러 변수 간 빈도/비율	여러 변수 간 빈도/비율	변수 1개 빈도/비율
연구방식	회귀 모델 추정	회귀 모델 추정	EDA	EDA	EDA	EDA	회귀 모델 추정

	(EDA 포함)	(EDA 포함)					(EDA 포함)
연구과정의 연결성	○	○	×	×	×	×	○

※ 참고문헌Ⅰ: 세대별 의료비 지출에 영향을 미치는 요인 분석(황연희, 2011)
※ 참고문헌Ⅱ: 한국의료패널로 본 한국 노인들의 의료이용 및 의료비 지출(황연희, 2011)
※ 참고문헌Ⅲ: 한국의료패널로 본 의료이용 및 본인부담의료비 지출(정영호, 2011)
※ 보고서Ⅰ: 2016년 한국의료패널 기초분석보고서(Ⅰ)
※ 보고서Ⅱ: 2016년 한국의료패널 기초분석보고서(Ⅱ)

<표 1.2.2> 변수 설정

변수		참고문헌Ⅰ	참고문헌Ⅱ	참고문헌Ⅲ	본 논문	비고
인구학적 변수	성별	O	O	O	O	
	출생년도	O				
	나이		O	O	O	전처리
	교육수준	O				
	혼인상태			O		
	세대구성	O	O		O	
사회경제적 변수	의료보장형태	O	O		O	
	경제활동 여부	O	O	O	O	
	소득 수준	O	O	O		
	거주 지역	O				
	주택소유 여부	O	O		O	
건강상태 및 건강행위를 나타내는 변수	만성질환의 수	O	O		O	
	흡연	O	O		O	
	음주	O	O		O	
	장애여부	O	O	O	O	
	우울경험여부	O			O	
	신체활동제한 여부		O		O	
	주관적 건강상태 여부		O			

※ 참고문헌Ⅰ: 세대별 의료비 지출에 영향을 미치는 요인 분석(황연희, 2011)
※ 참고문헌Ⅱ: 한국의료패널로 본 한국 노인들의 의료이용 및 의료비 지출(황연희, 2011)
※ 참고문헌Ⅲ: 한국의료패널로 본 의료이용 및 본인부담의료비 지출(정영호, 2011)
※ 보고서Ⅰ/Ⅱ는 한국의료패널에서 작성한 기초 보고서로 모든 변수를 분석했으므로 해당 표에 기입하지 않음

3) 사전조사 후 기존 연구방향 재설정

앞서 언급한 사전조사 후 분석 방법을 바꾸었다. <표 1.2.1>의 '본 논문'에 조사 대상자를 '전 연령'으로 하였다. 하지만 연구배경에서 고령화 속도와 의료비 부담, 노인 의료비 증가와 같은 내용을 언급했으므로 시의성 있는 연구를 위해 65세 이상 노인을 기준으로 연구를 진행하는 것으로 바꾸었다. 또한 기존의 EDA는 단순히 변수 1개에 대한 빈도와 비율을 파악했다면 수정한 EDA 방식은 응급, 입원, 외래별 의료이용횟수와 비율 그리고 세 가지 의료서비스에 대한 평균 의료 지출비용을 산출했다. 마지막으로 이 세 가지 의료서비스의 의료비 지출에 영향을 미치는 요인에는 무엇이 있는지 다중회귀분석을 진행했다.

Ⅱ. 대상 데이터 및 변수 소개

1. 작업 수행 환경

분석은 개인 노트북으로 진행했다. 노트북 프로세서는 Intel(R) Core(TM) i3-6100U CPU @ 2.30GHz이며, 로컬 디스트(C:) 용량은 222GB, 설치된 메모리(RAM)는 4GB, 시스템 종류는 64비트 운영체제, x64 기반 프로세서이다.

R은 오픈소스 소프트웨어로 자유롭게 무료로 사용가능한 통계 분석 도구이고, 다양한 패키지와 최신 분석 기법을 이용할 수 있다. 게다가 데이터 분석의 중요성과 함께 데이터 시각화의 중요성도 커지고 있다. R은 이 두 가지 기능에 최적화된 도구이며 기업 내 데이터 분석 직무에서 필수적으로 사용하는 분석 도구 중 하나이다. 따라서 R이 이번 논문에 사용하기에 적합하다고 판단했다. 필자는 R Studio를 사용했다. R Studio는 R을 사용하기 편리하게 만들어주는 IDE 소프트웨어로 다양한 부가 기능을 활용해 데이터를 효율적으로 분석할 수

있다. IDE(Integrated Development Environment)란 코딩, 파일 관리, 배포 등 프로그래밍에 필요한 다양한 작업을 수행할 수 있는 소프트웨어이다. 사용 중인 R Studio는 2021년 04월 19일 기준으로 가장 최신 버전인 R 4.0.5 버전을 사용하고 있다.

R Studio에서 데이터 분석을 시작하기 전에 먼저 프로젝트(Project)를 만들었다. 데이터 분석을 하다 보면 수많은 파일을 활용하는데, 프로젝트 기능을 이용하면 이런 파일들을 효율적으로 관리할 수 있다. 또한 파일들을 프로젝트 폴더별로 관리하면 편리하다. 프로젝트 명은 'Term Project'라고 했다. 그리고 install.packages("패키지명")을 통해 작업에 필요한 패키지를 설치하고 library("패키지명")로 패키지를 불러온다. 분석에 사용한 패키지는 dplyr, haven, ggplot2, car, corrplot, lm.beta, dlooker 총 7개이다. 각각 전처리 패키지, SAS, SPSS 등 통계 프로그램 데이터를 불러오는 패키지, 시각화 패키지, 회귀분석과 관련된 패키지, 상관행렬을 시각화하는 패키지, 표준화계수를 구하는 패키지, 데이터 품질 진단 패키지를 의미한다.

2. 데이터 수집 과정 및 대상 데이터 소개

Kaggle의 'Insurance Premium Prediction' 데이터 셋을 통해 수집하고자 하는 데이터 형식을 잡았다. 본래 Kaggle의 데이터를 활용하여 했지만 3년 전 외국 자료이며 본 논문은 우리나라의 자료를 활용해야 의미가 있다고 생각하여 Kaggle의 데이터 셋과 유사한 자료를 수집하고자 했다.

사전 조사 중, '한국의료패널 자료'를 활용한 유사논문을 참고했다. 유사 논문에서 사용한 데이터가 Kaggle과 비슷한 데이터 셋을 갖고 있기 때문에 사용하기 좋다고 판단했다. 따라서 '한국의료패널 자료'를 활용하였다. '한국의료패널 조사'는 의료이용 형태와 의료비 지출 규모에 관한 정보뿐만 아니라 의료이용 및 의료비 지출에 영향을 미치는 요인들을 포괄적이고 심층적으로 분석할 수 있는 패널 데이터를 구축하는데 목적을 두고 있다. 전국에서 7000여 가구를 선정해 2006년부터 매년 추적 조사한 자료로, 경제활동, 생활실태, 복지욕구 등 천여 변수로 구성되어 있다.

'한국의료패널'에서 제공하는 데이터는 '한국의료패널 자료 활용 동의서'를 작성하여 담당자에게 이메일 혹은 팩스로 전송해야 얻을 수 있다. 제공되는 데이터 파일 형식은 SAS, SPSS, STATA이다. 필자도 이러한 절차를 통해 SPSS 데이터 파일을 이메일을 통해 얻을 수 있었다. 제공받은 데이터 셋은 '한국의료패널 2008년~2017년 연간데이터(Version 1.6)'이고 총 145개의 데이터 셋으로 이루어져 있다. 그 중에서 필자는 가구사항(T17HH), 가구원사항(T17IND), 만성질환(T17CD), 성인가구원대상 부가조사(T17APPEN), 응급서비스 이용(T17ER), 입원서비스 이용(T17IN), 외래서비스 이용(T17OU) 데이터(총 7개)를 이용했고, 가장 최근인 2017년도 데이터 셋을 활용했다. 사용할 각 파일에 대한 특징은 아래 표와 같다.

<표 2.1.1> 각 파일의 내용 및 특징

파일명		내용	특징
T17IND	가구원사항	가구원 정보, 경제활동상태와 같은 정보	용량: 13.5MB 레코드 수: 17,184개 칼럼 수: 92개
T17HH	가구사항	가구정보, 주거사항, 가구 소득 및 지출, 의료관련지출과 같은 정보	용량: 3.13MB 레코드 수: 6,408개 칼럼 수: 64개
T17APPEN	성인가구원대상 부가조사	흡연, 음주, 정신적/육체적 건강과 같은 정보	용량: 10.4MB 레코드 수: 13,460개 칼럼 수: 97개
T17CD	만성질환	만성질환의 여부, 의약품이용과 같은 정보	용량: 35MB 레코드 수: 139,227개 칼럼 수: 22개
T17ER	응급서비스 이용	응급 의료서비스 이용에 대한 정보	용량: 1.12MB 레코드 수: 2,123개 칼럼 수: 48개
T17IN	입원서비스 이용	입원 의료서비스 이용에 대한 정보	용량: 2.34MB 레코드 수: 3,521개 칼럼 수: 66개
T17OU	외래서비스 이용	외래 의료서비스 이용에 대한 정보	용량: 262.67MB 레코드 수: 301,540개 칼럼 수: 65개

<표 2.1.2> IND 데이터 셋 모습

데이터 셋 구성	변수 및 설문문항	문항 구성
C3 성별 / C4_0 출생연도 / C7 혼인상태 / C8 교육수준 2 1940 3 3 1 1953 1 51 2 1958 1 44 1 1996 5 42 1 1955 1 44 2 1956 1 16 1 1993 5 42 1 1949 1 52	C3(성별) 설문문항: “OOO(가구원 이름)님은 성별이 어떻게 되십니까?”	(1) 남 (2) 여
	C4_0(출생년도) 설문문항: “OOO(가구원 이름)님의 생년월일이 어떻게 되십니까?”	유효 (-9) 모름/무응답
	C7(혼인상태) 설문문항: “OOO(가구원 이름)님은 혼인상태가 어떻게 되십니까?”	(1) 혼인 중(사실혼 포함) (2) 별거(이혼전제) (3) 사별 또는 실종 (4) 이혼 (5) 없음 (-1) 해당사항 없음 (-6) 설문대상 아님
	C8(교육수준) 설문문항: “OOO(응답자 이름)님은 학교를 어디까지 다니셨습니까? 혹은 다니고 계십니까?”	(1) 미취학(만7세이하) (2) 무학(문자해독불가) (3) 무학(문자해독가능) (11)~(16) 초등학교 1학년~6학년 (21)~(26) 중학교 1학년~3학년 (31)~(36) 고등학교 1학년~3학년 (41)~(46) 대학교 1학년~6학년 (51) 대학원 석사 (52) 대학원 박사

※ R의 view() 함수를 이용해 캡처한 모습이다. 변수명과 뜻은 한국의료패널 데이터 관리 부서를 통해 알 수 있다. (044-287-8121)

<표 2.1.3> HH 데이터 셋 모습

데이터 셋 구성	변수 및 설문문항	문항 구성
B6 주거형태 / B7 주택소유여부 / B8 자가의 경우 현시세 / B9 전세보증금 1 5 -1 -1 1 5 -1 -1 4 1 47000 -1 3 2 -1 20000 3 1 31000 -1 3 1 25000 -1 3 1 20000 -1 1 2 -1 6000	B6(주거형태) 설문문항: “귀하께서 살고 계신 집은 어떤 형태입니까?”	(1) 단독주택 (2) 다세대주택 (3) 연립주택 (4) 일반(임대)아파트 (5) 영구임대아파트 (6) 영업용건물 내 거주 (7) 오피스텔 (8) 기타
	B7(주택소유여부) 설문문항: “귀하께서 살고 계신 집은 자가입니까? 전세나 월세입니까?”	(1) 자가 (2) 전세 (3) 월세 (4) 무상(관사, 사택 등) (5) 기타(부모의 명의로 된 주택에 거주하는 경우) (-9) 모름/ 무응답
	B8(자가의 경우 현시세) 단위: ( )만원 설문문항: “귀하께서 살고 계신 집의 현시세는 대략 어느 정도입니까?”	유효 (-1) 해당사항 없음 (-9) 모름/무응답
	B9(전세보증금) 단위: ( )만원 설문문항: “귀하께서 살고 계신 집의 전세 보증금은 얼마입니까?”	유효 (-1) 해당사항 없음 (-9) 모름/무응답

<표 2.1.4> APPEN 데이터 셋 모습

데이터 셋 구성	변수 및 설문문항	문항 구성
	S17(음주여부) 설문문항: “최근 1년 동안 얼마나 자주 술을 드셨습니까?”	(1) 평생 마시지 않음 (2) 최근 1년간 금주 (3) 월 1회 미만 (4) 월 1회 (5) 월 2~3회 (6) 주 1회

S17 음주여부	S17_0 음주 시작시기	S18_1 금주 시작시기(년)	S19_1 금주 시작시기(월)		
4	45	-1	-1		(7) 주 2~3회 (8) 거의 매일 (-9) 모름/무응답
1	-1	-1	-1	S17_0(음주 시작시기) 만 ( )세" 설문문항: "처음으로 술 1잔을 모두 마셔본 적은 언제입니까?" (제사, 차례 때 몇 모금 마셔본 것은 제외합니다.)	유효 (-9) 모름/무응답 (-1) 해당사항 없음
1	-1	-1	-1	S18_1(금주 시작시기(년)) 설문문항: "( )년"	유효 (-9) 모름/무응답 (-1) 해당사항 없음
1	-1	-1	-1	S19_1(금주 시각시기(월)) 설문문항: "( )월"	유효 (-9) 모름/무응답 (-1) 해당사항 없음
3	17	-1	-1		
3	19	-1	-1		
7	18	-1	-1		
3	20	-1	-1		
4	36	-1	-1		
4	16	-1	-1		

<표 2.1.5> CD 데이터 셋 모습

데이터 셋 구성	변수 및 설문문항	문항 구성
CD2 만성질환 확인 / CD3_2 의사진단여부 / CD3_1 진단시기(연도) / CD3 진단시기(연령) 2 -1 -6 -6 2 -1 -6 -6 1 1 2008 -1 1 1 1990 -1 2 -1 -6 -6 2 -1 -6 -6 2 -1 -6 -6 1 1 2014 -1	CD2(만성질환 확인) 설문문항: "다음의 질환을 앓았거나, 앓고 있습니까?"	(1) 예 (2) 아니오 (3) 신규/누락 추가 (4) 완치 (5) 더 심각한 질환으로 악화 (6) 타질환과의 중복됨 (-9) 모름/무응답
	CD3_2(의사진단여부) 설문문항: "OOO(질환명)은 의사진단을 받았습니까?"	(1) 예 (2) 아니오 (-1) 해당사항 없음
	CD3_1(진단시기(연도)) 설문문항: "진단받은 시기는 언제입니까?" (최초 진단 시기)	유효 (-1) 해당사항 없음 (-6) 설문대상 아님 (-9) 모름/무응답
	CD3(진단시기(연령)) 설문문항: "진단받은 시기는 언제입니까?" (최초 진단 시기)	유효 (-1) 해당사항 없음 (-6) 설문대상 아님 (-9) 모름/무응답

<표 2.1.6> ER 데이터 셋 모습

데이터 셋 구성	변수 및 설문문항	문항 구성
ER33 처방약값 / ER33_1 약 구매일 / ER34 입원연계 / ER35 만족도 -1 -1 3 3 -1 -1 1 3 -1 -1 3 3 3300 0 3 3 -1 -1 3 3 -1 -1 3 3 -1 -1 3 3 -1 -1 1 3 -1 -1 3 2	ER33(처방약값) 설문문항: "처방전으로 약국에서 처방약을 구매하였다면 얼마를 지불하였습니까?"	유효 (91) 무료 (0) 의료 급여자 무료 (-1) 해당사항 없음 (-9) 모름/무응답
	ER33_1(약 구매일) 설문문항: "약국에서 처방약을 구입하는 일자는 진료일자와 동일하였습니까?"	유효 (-1) 해당사항 없음 (-9) 모름/무응답
	ER34(입원연계) 설문문항: "응급실 방문 후에 입원실로 이동하셨습니까?"	(0) 사망 (1) 응급실 후 입원 (2) 타병원 이동 (3) 귀가 (4) 조사현재 입원중
	ER35(만족도) 설문문항: "이번에 응급실 이용과 관련하여 전반적으로 어느 정도로 만족하셨습니까?"	(1) 매우 불만족 (2) 불만족 (3) 만족 (4) 매우 만족 (-9) 모름/무응답

<표 2.1.7> IN 데이터 셋 모습

데이터 셋 구성	변수 및 설문문항	문항 구성
	IN15(선택기준) 설문문항: "입원하셨던 00병(의)원을 선택한 가장 중요한 이유는 무엇이었습니까?"	(1) 의료진이 우수하다고 생각되어서 (2) 장비, 시설 및 병원 환경이 뛰어나서

IN15 선택기준	IN16 입원결정	IN17 입원 대기여부	IN18 대기일수		
6	1	2	-1		
5	1	8	-1		
5	1	2	-1		
8	1	2	-1		
1	1	1	6		
1	1	2	-1		
6	1	8	-1		

	(3) 의료진들이 친절해서 (4) 비용이 저렴해서 (5) 병원이 가까워서 (6) 다니던 병원이어서 (7) 다른 의료기관으로부터의 이송 의뢰 (8) 기타 (-9) 모름/무응답
IN16(입원결정) 설문문항: “입원결정에 가장 중요한 역할을 한 사람은 누구입니까?”	(1) 의료진 (2) 본인(환자 자신) (3) 기타: 가족, 지인 등 (4) 가해자, 타인, 비혈연
IN17(입원 대기여부) 설문문항: “방문 당일 날 입원을 해야 하는 상황이었으나 입원하지 못하고 기다리셨습니까?”	(1) 예 (2) 아니오 (8) 비해당(응급실을 통한 입원) (-9) 모름/무응답
IN18(대기일수) 설문문항: “기다리셨다면 며칠을 기다리셨습니까?”	유효 (-1) 해당사항 없음(기다리지 않음) (-9) 모름/무응답

<표 2.1.8> OU 데이터 셋 모습

데이터 셋 구성	변수 및 설문문항	문항 구성
	OU35_1(약구매일) 설문문항: “약국에서 처방약을 구입하는 일자는 진료일자와 동일하였습니까?”	유효 (0) 당일 처방약 구매 (-1) 해당사항 없음 (-9) 모름/무응답
	OU36(교통수단) 설문문항: “이 병(의)원 이용 시 사용한 가장 주된 교통수단은 무엇입니까?”	(1) 자기차량 (2) 택시 (3) 대중교통 (4) 기차/비행기 (5) 도보, 자전거 (6) 기타 (7) 오토바이, 경운기 (8) 기타(무료 셔틀버스 등) (-9) 모름/무응답
	OU37(교통시간) 설문문항: “집(또는 일터)에서 병원까지 몇 분이 소요되었습니까?”	유효 (0) 0분 (-9) 모름/무응답
	OU38(교통비) 설문문항: “이 병(의)원을 이용할 때의 편도교통비는 얼마였습니까?” (동반자 포함)	유효 (0) 도보, 자전거 등 (77) 자가 차량의 경우 (91) 무료(도보나 가자 차량 외) (-9) 모름/무응답

OU35_1 약구매일	OU36 교통수단	OU37 교통시간	OU38 교통비
-1	3	75	2500
-1	3	75	2500
0	3	30	1200
0	3	30	1200
0	3	30	1200
0	3	20	1200
0	3	20	1200
0	3	20	1200
0	3	20	1200

3. 대상 변수 소개

의료서비스 이용 시 발생하는 의료비는 성별, 나이, 세대구성 등 인구사회학적 특성과 주택소유, 경제활동여부와 같은 경제적 특성의 영향을 받는다. 또한 의료비는 건강상태와 건강행위에도 관련이 있다. 노령화가 지속될수록 건강수준이 저하되어 낙상 등의 사고위험이 높아지고, 만성질환의 발생률이 높아진다. 따라서 독립변수들을 ‘인구학적 변수’, ‘사회경제적 변수’, ‘건강상태 및 건강행위를 나타내는 변수 3가지 범주로 분류했다. 3개의 유사 연구를 참고해서 본 연구에서는 독립변수 12개와 종속변수(개인지출의료비) 1개를 사용했다. 세부적으로 인구학적 요인에는 성별, 출생년도, 세대구성, 사회경제적 변수에는 의료보장형태, 경제활동여부, 주택소유여부 그리고 건강상태 및 건강행위를 나타내는 변수에는 만성질환개수, 흡연여부, 음주여부, 장애여부, 신체활동제한, 우울경험이 있다.

우선 종속변수이다. 기존에는 IND 데이터에 있는 ‘개인지출의료비(I_ MEDICAL EXP1)’를 종속변수로 정했다. 하지만 I_ MEDICAL EXP1 변수는 응급, 입원, 외래별로 나눈 값이 아닌 전체를 더한 값이므로 세 가지 의료서비스에서 각각 발생한 지출은 알 수 없다. HH 데이터에 약품, 건강식품, 의료기기 구매액 변수가 있지

만 이것은 부가지출이기에 지출금액으로 정의하지 못한다. 유사한 변수가 있는지 찾아본 결과 응급, 입원, 외래서비스에 대한 데이터 셋인 ER, IN, OU에서 '총진료비'와 '처방약값' 변수가 있었다. 결국 총진료비와 처방약값을 더한 값을 개인지출의료비로 정의했다. 각각의 데이터에 총진료비 변수와 처방약값이 존재했기 때문에 총 3개의 데이터 셋을 분석에 사용하는 것이 적합하다고 판단했다. '총진료비' 변수는 '급여'와 '비급여'를 모두 더한 값이다. 보험 적용이 되는 진료를 '급여'라고 하고, 보험적용이 안 되는 진료를 '비급여'라고 한다. 급여는 100% 공단에서 부담하는 것이 아닌 공단에서 부담하는 공단부담금(보험료)과 본인이 직접 부담하는 본인부담금이 있다. 반면에 비급여는 보험적용이 안되기 때문에 100% 자비부담이다.

<표 2.2.1> 개인지출 의료비 변수 변경 과정

기존 지출 의료비 변수			수정 후 지출 의료비 변수	
데이터 명	변수 명		데이터 명	변수 명(총진료비 + 처방약값)
ind	I_MEDICALEXP (개인지출 의료비)	➔	er(응급)	ER26_5 + ER33
			In(입원)	IN35_6 + IN37
			ou(외래)	OU29_7 + OU35

인구학적 요인에서 성별과 출생년도는 개인 정보이다. 총 145개 데이터 셋에서 IND 데이터는 가구원 정보 즉, 개별 정보가 있는 데이터이다. 그래서 개개인에 대한 정보를 알 수 있는 IND 데이터에서 성별(C3)과 출생년도 변수(C4_0)를 사용했다. 나이 변수는 존재하지 않아 출생년도 변수를 전처리해서 새로 만들었다. 세대구성(B3)은 가구에 대한 정보이다. 일반적으로 우리나라 노인들의 가구 유형은 점점 노인 부부가구 또는 단독가구로 유형이 축소되는 경향을 보이는 것으로 알려져 있다. 이러한 사회현상이 의료비 지출에 미치는 영향을 보기 위하여 노인 단독가구와 부부가구, 그리고 노인 부부와 자녀가 함께 사는 가구, 기타 가구로 범주화해서 분석에 이용했다. 세대구성 변수는 가구정보가 있는 HH 데이터 셋에서 얻었다.

사회경제적 요인에서는 경제적 여건에 따른 지출 의료비의 영향을 알고자 했다. 의료보장형태(C11), 경제활동유무(C24)는 개별 정보라고 판단했다. 한 가구 안에서 건강보험 가입 유무와 경제활동 여부는 가구 내 가구원마다 다르기 때문이다. 따라서 의료보장형태와 경제활동여부 변수는 가구원의 정보를 알 수 있는 IND 데이터 셋에서 얻었다. 의료보장형태에서 미가입자와 급여정지는 건강보험가입자와 의료급여수급자 둘 다 해당하지 않으므로 분석에서 제외하고 건강보험가입자와 의료급여수급자, 두 가지로 분류했다. 주택여부(B6)에 대한 변수는 한 개인이 자가를 갖고 있느냐에 대한 여부를 묻는 변수이기 때문에 IND에서 얻으려고 했으나 주거관련 정보는 HH에 있었다. HH는 가구정보 뿐만 아니라 주거사항, 가구 소득 및 지출에 대한 정보 등이 포함된다. 유사 논문에서는 가구당 소득을 구했다. 소득변수는 HH에 있다. 유사 연구에서 이것을 6개의 구간으로 범주화하고 그것을 각 해당 가구의 가구원에 적용했다. 하지만 가구원마다 소득 수준과 소득 활동 여부가 전부 다르므로 가구의 소득 수준이 가구원의 소득 수준이라고 하기 에는 무리가 있다고 판단해서 본 연구에서 사용하지 않았다.

건강상태 및 건강행위에 대한 요인에서 만성질환여부(CD2)는 만성질환 여부에 대한 정보를 알려주는 CD 데이터 셋에 있다. 흡연, 음주, 우울감, 신체활동제한(SH117)은 성인가구원대상 부가조사인 APPEN 데이터 셋에서 찾았다. 흡연, 음주, 정신적/육체적 건강과 같은 정보가 있기 때문이다. 흡연여부 변수는 '세대별 의료비 지출에 영향을 미치는 요인 분석(황연희)'에서 사용한 변수(S1)를 사용하려 했으나 2009년 이후 사라진 문항이다. 따라서 흡연자의 기준을 현재 담배를 피우는지 아닌지를 묻는 변수, (S2)를 이용했다. 음주여부는 최근 한 달간 한 번의 술좌석에서 남성은 소주 7잔(맥주 5캔) 이상, 여성은 소주 6잔(맥주4캔) 이상이면 과음이라고 정의하는 변수(S22)를 이용했다. 우울감(S44)은 최근 1년간 2주 이상 연속으로 일상생활에 지장이 있을 정도로 많이 슬펐거나 불행하다고 느낀 사람인가를 묻는 문항을 이용했다. 음주여부와 신체활동제한은 유사 연구에서 사용했기에 신뢰가 있다. 주관적 건강상태 변수는 유사 연구에서 사용했지만 본 연구에서 사용하지 않았다. 그 이유는 해당 변수가 2013년도 이후에는 수치형이 아닌 범주형으로 제공돼서 이것을 쓰기에 충분한 정보를 제공하지 못한다고 판단했기 때문이다. 그래서 다른 변수를 탐색하던 도중 장애여부에 따라 의료비 지출이 차이가 있을 것이라고 판단했다. 또한 정영호(2011)의 연구에서 장애여부 변수가 쓰였기 때문에 장애여부 변수(C14)를 사용했다. 장애여부 변수는 IND 데이터에 있다. 장애 관련 변수는 '장애 종류', '장애등급 판정 유무', '장애등급 등록', '장애등급'으로 총 4가지가 있다. 그 중에서 '장애 종류'와 '장애등급' 변수에서 '해당사항 없음'

항목 비고란에 '비 장애인'이라고 명시되었다. 그래서 간편한 전처리를 위해 비교적 적은 항목으로 구성된 '장애등급(S14)'변수를 이용해서 '장애 있음'과 '장애 없음'으로 이분했다.

Ⅲ. 대상 데이터 품질 진단 및 데이터 전처리

이 장은 데이터 품질 진단을 해서 사용할 7개의 원본 데이터의 상태를 확인하고 데이터를 가공하는 EDA 과정과 결과를 보여준다. EDA(Exploratory data analysis: 탐색적 자료 분석, 이하 '데이터 탐색' 또는 'EDA')란, 데이터를 분석하기 전에 그래프나 통계적인 방법으로 자료를 보다 직관적으로 바라보는 과정을 말한다. EDA에서 파악할 것은 65세 이상 노인들의 응급, 입원, 외래별 의료서비스 이용 횟수의 비율과 여러 요인에 따른 의료서비스 이용현황 및 평균 의료비 지출을 탐색하는 것이다.

R에서 SPSS파일을 불러들이는 read_spss() 함수를 이용하여 사용할 7개의 데이터를 만들었다. 아래의 표를 통해 이해할 수 있다.

<표 2.3.1> spss 데이터 불러오기(데이터 명은 임의로 지정)

가공 전(SPSS)		가공 후(R)
T17IND.sav		ind
T17HH.sav		hh
T17APPEN.sav	read_spss()	append
T17CD.sav	➔	cd
T17ER.sav		er
T17IN.sav		in
T17OU.sav		ou

1. 데이터 품질 진단

사용하려는 데이터가 어떤 상태인지 품질 검증을 하는 과정이 필요하다. 데이터 품질에 문제가 있으면 데이터를 수정하거나 다시 획득해야 한다. 연구에서 사용한 한국의료패널 데이터는 이미 검증을 거친 데이터일 가능성이 높으므로 특별한 이상이 없을 것이다. 하지만 데이터의 각 변수가 어떻게 구성되었고 결측값이 몇 개인지 확인해서 결측값을 제거하거나 다른 값으로 대체해야 한다. 따라서 연구에 사용한 총 7개의 데이터 셋에서 사용한 변수들만 품질 진달을 수행했다. 데이터 품질 관리를 위한 패키지는 'dlookr' 패키지이다. dlooker 패키지는 데이터를 진단하고 데이터 분석 또는 보정해야 할 변수를 찾는 데 사용할 수 없는 변수를 선택하는 것을 목표로 한다. 먼저 diagnose() 함수로 변수에 대한 기본 진단을 수행했다. 변수명과 변수 타입, 결측값 개수, 비율 등을 알 수 있다. 다음은 diagnose_numeric() 함수로 수치형 변수를 진단했다. 변수들의 기초 통계량과 음수값, 이상치 등을 확인할 수 있다.

diagnose()의 결과에 'missing_count'가 있다. 이것은 결측값의 개수를 의미하는데 한국의료패널 데이터는 결측값을 'NA'로 지정하지 않고 음수값으로 표시한다. 예를 들면 -1은 '해당사항 없음', -9는 '모름/무응답'이다.(음수값이 곧 결측값이다.) 따라서 missing_count 값이 항상 0이 나온다. 결과의 0값은 결측값이 없다를 의미하는 것이 아니므로 유의해야한다. 코드북을 참고하여 설문의 보기문항내용이 어떻게 구성되어 있는지 반드시 파악해야 한다. 한국의료패널 데이터는 보기문항 내용을 수치로 바꾸어서 제공하기 때문에 높은 보기 문항과 음수로 표현된 문항이 이상치가 될 수 있고, 기통통계량을 파악하는 것은 의미가 없다. 따라서 본 연구에서는 이상치와 기초통계량은 고려하지 않았다. 문항 분류와 결측치 처리에 대한 내용은 '데이터 전처리'에서 소개하므로 이 장에서는 언급하지 않는다.

<표 > IND 데이터 품질 진단

변수 명	변수 타입	범주의 개수	결측값 개수
성별(C3)	numeric	2	0
출생년도(C4_0)	numeric	-	0

의료보장형태(C11)	numeric	10	0	
경제활동여부(C24)	numeric	3	2,150	(-6) 비해당(만 14세 이하)
				(-9) 모름/무응답
장애여부(C14)	numeric	7	16,163	(-1) 해당사항 없음(비장애인)

IND 데이터에서 사용한 총 5개의 변수를 품질 진단한 결과로 성별, 출생년도, 의료보장형태에서 결측값은 없었다. 경제활동여부에서 결측값에 해당되는 사람은 모두 만 14세 이하였다. 이후에 65세 이상과 미만을 나눌 때 제외된다. 장애여부에서 비장애인에 응답한 사람은 변수 전처리할 때 '장애 없음'으로 분류한다.

<표 > HH 데이터 품질 진단

변수 명	변수 타입	범주의 개수	결측값 개수
세대구성(B3)	numeric	20	0
주택소유여부(B7)	numeric	5	0

HH 데이터에서 사용한 두 개의 변수를 품질 진단한 결과, 세대구성 변수는 총 20개의 범주로 구성되어 있다. 전처리 과정에서 단독, 부부, 부부자녀, 기타, 총 4개의 범주로 구성했다. 주택소유여부는 전처리 과정에서 자가가 있는 사람과 없는 사람으로 나누었다. 두 변수 모두 결측값은 없다.

<표 > APPEN 데이터 품질 진단

변수 명	변수 타입	범주의 개수	결측값 개수	
S2(흡연여부)	numeric	5	3	(-9) 모름/무응답
				(-1) 해당사항 없음
S22(음주여부)	numeric	9	4,702	(-9) 모름/무응답
				(-1) 해당사항 없음
S44(우울감)	numeric	4	527	(-9) 모름/무응답
				(-1) 해당사항 없음
SH117(신체활동제한)	numeric	4	12,982	(-1) 비해당
				(-6) 설문대상 아님

흡연여부 변수의 결측값 개수는 총 3개인데 모두 -9에 응답한 사람들이다. 음주여부에서 -9에 응답한 사람은 4명이고, 4,698명은 -1에 응답했다. 우울감에서 -9에 응답한 사람은 3명, 우울하지 않다고 응답한 사람은 524명이다. 신체활동에 제한을 느끼지 않는 사람은 -1에 응답했다. -6(설문대상 아님)은 65세 미만에 해당되는 사람으로 나이를 65세 이상과 미만을 기준으로 나누었을 때 제외된다.

<표 > CD 데이터 품질 진단

변수 명	변수 타입	범주의 개수	결측값 개수
만성질환여부(CD2)	numeric	5	0

연구에서는 '만성질환개수'라는 변수를 이용했지만 원본 데이터의 모습은 만성질환의 여부에 대한 변수이다. 만성질환여부에서 만성질환개수로 전처리하는 과정은 '사전 데이터 전처리' 장에서 다룬다. 총 7개의 문항으로 구성되어있지만 -5(더 심각한 질환으로 악화)와 -9(모름/무응답)에 응답한 사람이 없어서 범주의 개수가 5로 표시되었다.

<표 > ER 데이터 품질 진단

변수 명	변수 타입	최댓값	최솟값	결측값 개수	
총진료비(ER26_5)	numeric	14,019,720	3,000	684	(-9) 모름/무응답
					(-1) 해당사항 없음

처방약값(ER33)	numeric	68,100	500	1,954	(-9) 모름/무응답
					(-1) 해당사항 없음

응급서비스의 총진료비의 최댓값은 14,019,720원, 최솟값은 3,000원이다. 결측값 개수는 684이다. -1은 상세금액이 확인되지 않음을 의미한다. 처방약값의 최솟값은 500원, 최댓값은 68,100원이다. 결측값 개수는 1,954개인데 모두 -1에 응답한 사람들이다. -1은 약국에 방문하지 않은 것을 의미한다.

<표 > IN 데이터 품질 진단

변수 명	변수 타입	최댓값	최솟값	결측값 개수	
총진료비(IN35_6)	numeric	50,595,450	1,002	536	(-9) 모름/무응답
					(-1) 해당사항 없음
처방약값(IN37)	numeric	110,640	500	3,456	(-9) 모름/무응답
					(-1) 해당사항 없음

입원서비스의 총진료비의 최댓값은 50,595,450원, 최솟값은 1,002원이다. 결측값 개수는 536이다. -1은 상세금액이 확인되지 않음을 의미한다. 처방약값의 최댓값은 110,640원, 최솟값은 500원이다. 결측값 개수는 3,456개인데 모두 -1에 응답한 사람들이다. -1은 약국에 방문하지 않은 것을 의미한다.

<표 > OU 데이터 품질 진단

변수 명	변수 타입	최댓값	최솟값	결측값 개수	
총진료비(OU29_7)	numeric	10,133,411	70	70,175	(-9) 모름/무응답
					(-1) 해당사항 없음
처방약값(OU35)	numeric	705,380	500	132,284	(-9) 모름/무응답
					(-1) 해당사항 없음

외래서비스의 총진료비의 최댓값은 10,133,411원, 최솟값은 70원이다. 결측값 개수는 70,175이다. -1은 상세금액이 확인되지 않음을 의미한다. 처방약값의 최댓값은 705,380원, 최솟값은 500원이다. 결측값 개수는 132,284개 이다. -1은 약국에 방문하지 않은 것을 의미한다.

2. 데이터 전처리

처음부터 데이터가 본인이 원하는 모습으로 정리되어 있다면 분석이 훨씬 수월하고 단기간에 분석 목표에 도달할 것이다. 하지만 대부분의 데이터는 그렇지 않다. 따라서 분석자가 자신의 목표에 맞게 데이터를 가공해야한다. 이 작업을 '전처리'라 하고, 이 절에서는 그 과정을 다룬다.

1) 65세 이상 노인의 의료서비스별 의료이용 횟수 및 비율

한국의료패널 기관의 '데이터관리' 부서 담당자와 전화하여, ind, hh, appen 데이터와 달리 cd, er, in, ou 데이터는 앞의 3개의 데이터와 구성이 다르다는 것을 알았다. 총 7개의 데이터를 한 번에 병합하면 오류가 발생하고 잘못된 데이터가 만들어진다고 한다. 따라서 cd, er, in ,ou 데이터를 이용하기 위해서 각각의 데이터에서 필요한 변수들만 추출하여 합쳐야 한다. select() 함수를 이용하여 데이터 분석 시 사용할 변수들만 추출하여 별도의 새로운 데이터 'new_cd', 'new_er', 'new_in', 'new_ou'를 만들었다.

PIDWON 변수는 가구원 고유번호이다. cd1_1 변수에는 한국표준질병분류(KCD)에 따라 부여된 만성질환 코드가 있고, CD2 변수는 만성질환 여부 변수이다. ERCOUNT, INCOUNT, OUCOUNT 변수는 각각 응급실 이용횟수, 입원 이용횟수, 외래 이용횟수를 나타낸다.

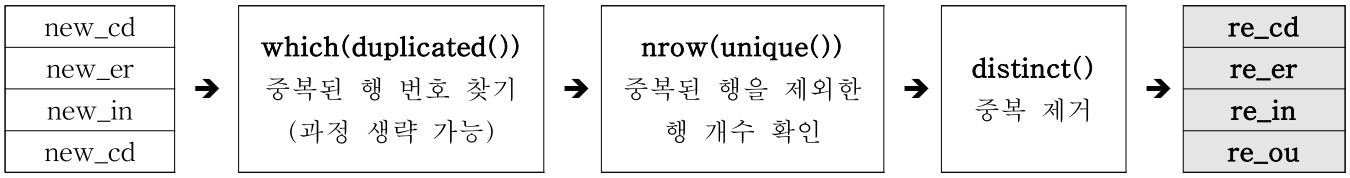
<표 2.3.2> cd, er, in, ou 데이터 전처리

기존 데이터 명	select()	새로운 데이터 명	사용할 변수
cd	➔	**new_cd**	PIDWON, cd1_1, CD2

er		**new_er**	PIDWON, ERCOUNT
in		**new_in**	PIDWON, INCOUNT
ou		**new_cd**	PIDWON, OUCOUNT

데이터 구조상, 한 사람이 여러 질문에 응답할 수 있기에 중복된 행이 존재할 수 있다. 그래서 이를 확인하고 제거하는 과정이 필요하다. which(duplicated()) 함수를 이용해 중복된 행 번호를 찾고 nrow(unique()) 함수를 이용해 중복된 행을 제외한 나머지 행 개수를 알 수 있다. 이 과정에서 중복된 행이 있으면 distinct() 함수로 제거한다. which(duplicated()), nrow(unique()) 함수는 중복 행의 유무를 찾기 위한 과정이라고 볼 수 있다. 이 과정으로 're_cd', 're_er', 're_in', 're_ou' 데이터를 얻었다.

<표 2.3.3> 중복 행 제거 과정



re_er, re_in, re_ou는 원하는 대로 데이터가 이루어져있지만 re_cd 데이터는 원하는 형태가 아니다. 마지막으로 re_cd 데이터를 원하는 모습으로 가공하는 과정이 필요하다. re_cd는 PIDWON(가구원 고유번호), cd1_1(만설질환 코드), CD2(만성질환 여부), 총 3개의 변수들로 구성되어있다. 같은 사람에게 어떤 만성질환이 있는지에 대한 데이터이므로 한 사람에 대해 여러 데이터가 존재한다. 하지만 필자가 원하는 데이터 모습은 한 사람이 몇 개의 만성질환을 갖고 있냐는 것이다.

먼저 CD 변수는 1(예), 2(아니오), 3(신규/누락 추가), 4(완치), 6(타질환과 중복됨)으로 총 다섯 개의 값이 있는데 2와 4는 만성질환이 없음, 그리고 1, 3, 6은 만성질환이 있음으로 해석할 수 있다. ifelse() 함수를 이용하여 2, 4는 0(없음)으로 1, 3, 6은 1(있음)로 치환했다. 그리고 파생변수를 만드는 mutate() 함수를 이용해 0(없음)과 1(있음)값을 갖고 있는 새로운 chg_cd 변수를 만들었다.

마지막으로 가구원별 만성질환 개수를 알아야한다. group_by() 함수를 이용해 가구원별로 그룹화하고, 요약통계량을 구해주는 summarise() 함수와 sum() 함수를 이용해 만성질환 개수를 나타내는 count 변수를 만들었다. 최종적으로 re_cd 데이터는 PIDWON 변수와 개인이 보유한 만성질환 개수를 나타내는 count 변수로만 이루어지도록 select() 함수를 이용했다. 아래에 이해하기 쉽도록 re_cd의 전처리 과정을 표 형태로 정리했다.

<표 2.3.4> re_cd 전처리 과정 (예시)

PIDWON	cd1_1	CD2	→	chg_cd	→	count
1000102	고혈압	1(예)		1	**group_by()**	
1000102	당뇨	2(아니오)	**mutate()**	0	+	
1000102	식도염	3(신규/누락 추가)	+	1	**summarise()**	3
1000102	고지혈증	4(완치)	**ifelse()**	0	+	
1000102	섭식장애	6(타질환과 중복)		1	**sum()**	
1000404	…	…	…	…	…	…

원래 re_cd의 데이터 구성

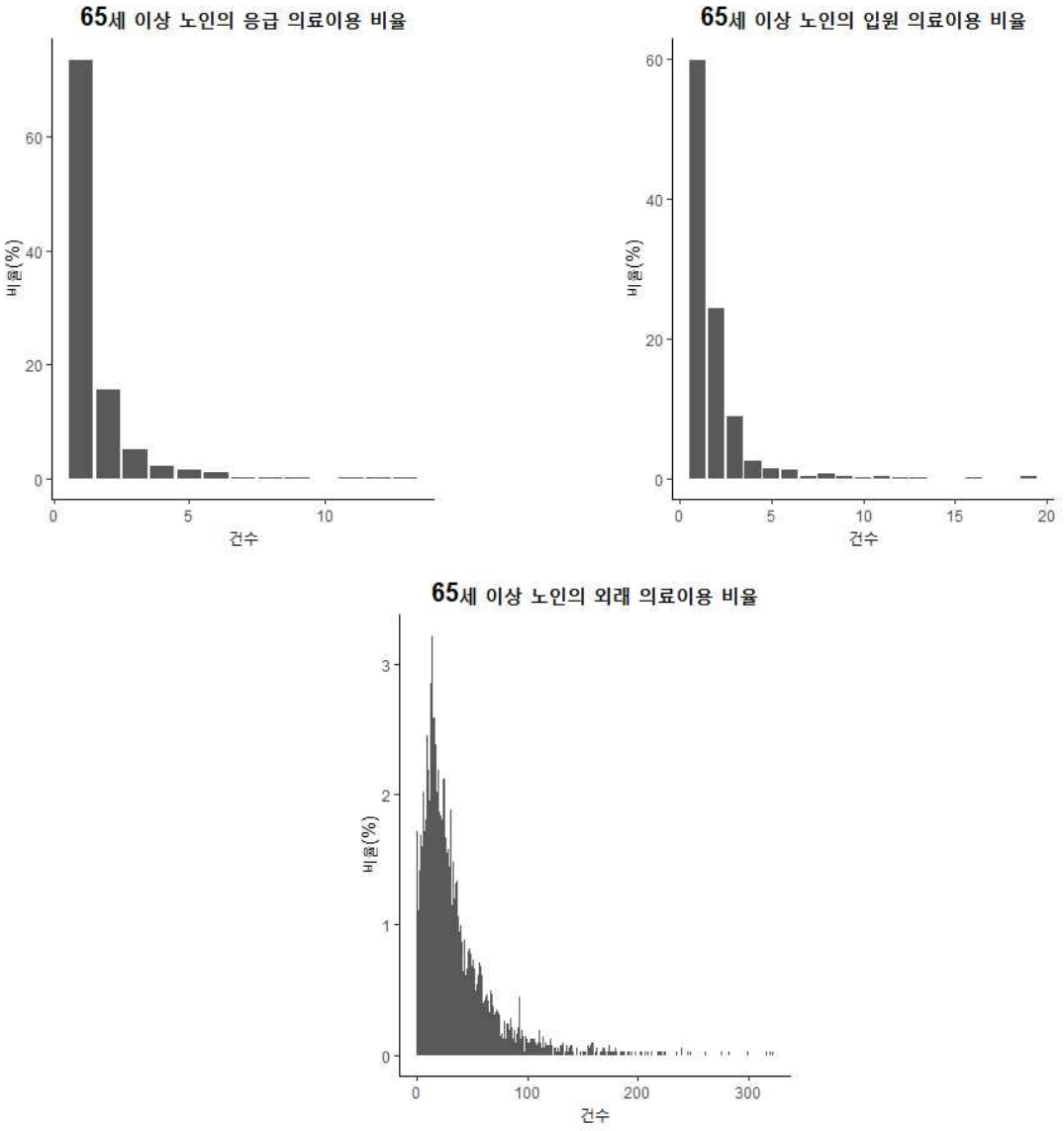
65세 이상 노인들을 기준으로 분석하는 것으로 연구 방향을 변경했다. 2017년도 자료를 이용했으므로 '2017 - 출생년도 + 1' 수식으로 2017년도 사람들의 나이를 알 수 있는 '나이' 변수를 만들었다. 그리고 '연령대' 변수를 만들어 65세 이상인 사람을 '65세 이상', 그렇지 않은 사람을 '65세 미만'으로 분류했다.

앞서 전처리한 데이터들을 하나로 합치는 과정이 필요하다. R 내장함수인 'merge()' 함수를 사용한다. 이 함수는 여러 데이터를 한 번에 결합하는 것이 아닌 두 개의 데이터를 동일한 열을 기준으로 결합한다. 함수 안에 'by' 매개변수에는 두 데이터에 공통적으로 들어있는 칼럼 명을 넣는다. 기준 칼럼을 무엇으로 정하느냐에

대해서는 한국의료패널 기관의 '데이터 관리' 부서에서 해답을 얻었다. 각 데이터에 'HHID' 변수와 'PIDWON' 변수가 있다. 데이터를 결합할 때 'HHID' 변수는 같은 집(가구)을 기준으로 삼을 때 이용하고, 'PIDWON' 변수는 같은 사람을 기준으로 삼을 때 이용한다.

ind와 er, in, ou를 각각 PIDWON 변수를 기준으로 merge() 함수를 이용하여 총 3개의 데이터를 만든다. 그리고 filter() 함수를 이용해 65세 이상인 노인만 추출하고 select() 함수로 PIDWON 변수와 ERCOUNT(응급이용 횟수), INCOUNT(입원이용 횟수), OUCOUNT(외래이용 횟수) 변수만 추출한다. 이렇게 해서 ind_er_65, ind_in_65, ind_ou_65 데이터를 얻었다. 의료이용 횟수를 파악하기 위해 COUNT 변수별로 그룹을 지어 summarise() 함수와 n() 함수를 이용해서 각 이용횟수에 해당하는 사람들의 수를 구한다. 예들 들어, 1회 이용한 사람은 n_1명, 2회 이용한 사람은 n_2명 이런 식으로 말이다. 그리고 mutate() 함수로 각 이용횟수에 해당하는 사람들의 비율을 구한다. 여기서 얻은 데이터 명을 각각 'ind_er_65_pct', 'ind_in_65_pct', 'ind_ou_65_pct'라고 지었다.(pct는 percent를 의미한다.)

<표 3.1.1> 응급, 입원, 외래별 노인 의료이용 비율 그래프



ggplot() 함수를 통해 응급, 입원, 외래 의료이용 횟수에 대한 막대그래프를 작성했다. 65세 이상 노인이 2017년 1년 동안 이용한 응급 의료이용 횟수에서 1건인 노인이 73.59%로 가장 많고 최대 응급 의료이용은 13회이다. 입원 의료이용은 1년에 1회 입원한 노인이 59.82%로 가장 많고, 최대 19회이다. 외래 의료이용은 14건을 이용한 노인이 전체의 3.22%를 차지하고, 최대로 외래 의료이용을 한 노인은 322건으로 나타났다.

2) 65세 이상 노인의 여러 요인에 따른 의료이용 현황과 의료비 지출 현황

65세 이상 노인들의 인구학적 요인, 사회경제적 요인, 그리고 건강행위 및 건상상태에 따른 의료이용과 의료비 지출 현황을 파악했다. 그리고 독립변수들이 모인 데이터(b1)와 응급, 입원, 외래에 따른 개인지출 의료비 즉, 종속변수 데이터(dt_er, dt_in, dt_ou)를 가공하는 두 가지 과정으로 나누어 소개했다.

(1) 독립변수 데이터

'ind', 'hh', 'appen', 're_cd' 데이터를 하나로 병합했다. ind과 hh은 'HHID' 변수를 기준으로, appen와 re_cd는 'PIDWON' 변수를 기준으로 결합했다. 하나로 합친 'a3' 원본 데이터는 복구해야 할 상황에 대비해 그대로 두고 'raw_data'라는 복사본을 만들었다. 최종적으로 얻은 raw_data의 레코드 수는 13,460개, 칼럼 수는 254개이다. 이를 도식화하면 다음 표와 같다.

<표 3.2.1> 데이터 병합

ind	hh	appen	re_cd	복사본
merge(**ind, hh**, by = c("HHID"))				
↓				
a1				a3
	merge(a1, **appen**, by = c("PIDWON"))			
	↓			↓
	a2			
		merge(a2, **re_cd**, by = c("PIDWON"))		**raw_data**
		↓		
		a3(원본 데이터)		

영어로 구성된 칼럼 명을 쉽게 알아보기 위해 11개 변수 명을 한글로 수정했다. rename() 함수를 이용하였으며 아래 표에서 확인할 수 있다. 이 과정에서 얻은 데이터를 'dt'라고 이름 지었다. 그리고 select() 함수를 통해 사용할 변수만 추출했다. 여기서 얻은 데이터를 'b1'이라고 했다.

<표 3.2.2> 변수 명 변경

이전 변수 명		이후 변수 명
C3		성별
B3		세대구성
C11		의료보장형태
C24		경제활동여부
B7	**rename()**	주택소유여부
count	➔	만성질환개수
S2		흡연여부
S22		음주여부
C14		장애여부
SH117		신체활동제한
S44		우울경험

※ 출생년도 변수에서 얻은 '연령대' 변수는 Ⅱ-3-(5)에서 이미 전처리과정을 거쳤으므로 이 장에서는 수행 안 함

※ 만성질환개수 변수는 Ⅱ-3-(4)에서 최종적으로 얻은 count 변수를 사용함

<표 3.2.3> raw_data 전처리

raw_data	➔	**rename()** 변수 명 변경	➔	dt	➔	**select()** 필요한 변수만 추출	➔	**b1** **최종 독립변수 데이터**

다음은 각 변수의 범주를 나누고 결측값을 제거했다. 남성은 1, 여성은 0으로, 세대구성은 단독가구, 부부가구, 부부+자녀, 기타, 총 4개의 범주로 분류했다. 경제활동을 하는 사람(1)은 그대로 1로 두었고, 그렇지 않은 나머지 값들은 0(경제활동 안함)으로 처리했다. 모름/무응답(-9)에 해당하는 사람은 없었다.

<표 3.2.4> 경제활동유무 변수 전처리

경제활동유무 전처리 전	경제활동유무 전처리 후
(1) 예	1: 경제활동 함
(2) 아니오 (-6) 비해당(만14세이하) (-9) 모름/무응답	0: 경제활동 안함

의료보장 형태는 직장가입자와 지역가입자와 건강보험에 가입되어 있으면서 특례자인 사람(국가유공자), 차상위 경감 대상자를 건강보험가입자로 분류했다. 차상위 경감 대상자는 희귀난치성, 중증질환자, 만성질환자, 18세 미만인 자 중 소득인정액[1] 기준과 부양의무자[2] 기준을 모두 갖춘 사람을 말한다. 여기서 소득인정액 기준은 기준세대[3]의 소득인정액이 중위소득[4]의 50% 이하를 말하여 부양의무자 기준은 차상위 본인부담경감을 적용 받고자 하는 사람이 부양의무자가 없거나 있어도 부양능력이 없거나 부양 받을 수 없는 자를 의미한다. 의료급여수급자에는 의료급여 1종[5], 2종[6]이 포함된다. 나머지 미가입자, 자격 상실자와 급여가 정지된 사람, 국가유공자이지만 건강보험에 가입하지 않은 사람은 둘 중 하나에 포함되지 않으므로 분석에서 제외했다. 주택소유 여부는 자가를 소유한 사람과 전세, 월세 등인 사람으로 이분했다. 모름/무응답(-9)에 해당되는 사람은 없었다. 만성질환개수 변수 '0개', '1개', '2개', '3개', '4개', '5개 이상'으로 분류했다.

<표 3.2.5> 의료보장형태 변수 전처리

의료보장형태 전처리 전	의료보장형태 전처리 후
(1) 공무원, 교직원 건강보험 (2) 직장 건강보험 (3) 지역 건강보험 (6) 건강보험가입 + 특례자 (10) 건강보험 + 차상위 경감 대상자	1: 건강보험가입자
(4) 의료급여 1종 (5) 의료급여 2종	0: 의료급여수급자
(7) 국가유공자 특례(건강보험 가입하지 않음) (8) 미가입(외국국적) (9) 건강보험체납 - 급여정지	NA: 급여정지/미가입

흡연여부는 현재 담배를 피우는가를 기준으로 분류했다. 과거에는 피웠지만 현재는 피우지 않는 사람과 피운 적이 없음에 응답한 사람은 흡연안함(0), 반대로 매일 또는 가끔 피우거나 모름에 응답한 사람은 흡연함(1)으로 분류하였다. 음주여부는 지난 일 년 간 한 달에 한번 이상 술을 마신 사람 중에서 남성의 경우는 소주 7잔(맥주 5캔), 여성의 경우 소주 6잔(맥주 3캔)을 과음의 기준으로 하여 한 달에 1회 이상 과음한 사람과 그렇지 않은 사람으로 구분했다. 우울 경험 여부는 최근 1년 동안 2주 이상 연속으로 일상생활에 지장이 있을 정도로 많이 슬펐거나 불행하다고 느낀 적이 있는 지를 기준으로 구분했다. 흡연여부, 음주여부, 우울경험의 -9(모름/무응답)는 NA로 지정하여 분석에서 제외했다.

장애여부는 원래 장애등급 변수로 '1급', '2급', ..., '6급', '비등록 장애인', '해당사항 없음'으로 구성된다. '해당사항 없음'은 비장애인이므로 장애 없음(0), 반대인 경우에는 장애 있음(1)으로 구분했다. 신체활동제한 변수는 성인가구원대상 부가조사(T17APPEN)의 설문 문항 중 하나로 '옷 입기, 세수하기, 목욕하기, 식사하기, 침상에

1) 소득인정액: 기준세대의 소득과 재산을 더한 액수
2) 부양의무자: 가족을 부양할 의무가 있는 사람
3) 기준세대: 희귀난치성, 중증질환자, 만성질환자, 18세 미만의 아동이 속한 세대
4) 중위소득: 우리나라의 소득을 가구별로 일렬로 세웠을 때 중간에 위치하는 소득
5) 의료급여 1종: 근로능력이 없거나, 보건복지부 장관이 근로가 곤란하다고 인정한 자, 국민기초생활보장법에서 정한 보장시설에서 급여를 받고 있는 자
6) 생계·의료·주거·교육급여 수급자 중 의료급여 1종에 해당되지 않는 자

서 일어나기, 화장실 사용, 대소변 조절 등의 활동에서 당신의 장애나 신체 및 정신적 문제로 인한 다른 사람의 도움이 3개월 이상 계속 필요할 것이라고 생각하십니까?'에 대한 질문이다. 제한이 있는 사람은 제한 있음(1), 제한이 없는 사람은 제한 없음(0)으로 나누었다.

<표 3.2.6> 독립변수의 정의

구분		의미
인구학적 변수	성별	남자: 1 여자: 0
	나이	2017 - 출생년도 + 1
	세대구성	1: 단독가구 2: 부부가구 3: 부부+자녀 4: 기타
사회경제적 변수	의료보장형태	1: 건강보험가입자 0: 의료급여수급자 NA: 결측값
	경제활동여부	1: 결제활동 함 0: 경제활동 안함
	주택소유여부	1: 자가 0: 전세, 월세 등
건강상태 및 건강행위를 나타내는 변수	만성질환개수	0개 / 1개 / 2개 / 3개 / 4개 / 5개 이상
	흡연여부	1: 흡연함 0: 흡연안함
	음주여부(과음)	1: 월 1회 이상 0: 월 1회 미만
	장애여부	1: 장애 있음 0: 장애없음
	신체활동제한	1: 제한 있음 0: 제한 없음
	우울경험	1: 우울감 경험 0: 우울감 경험 안함

(2) 종속변수 데이터(응급, 입원, 외래 개인지출 의료비)

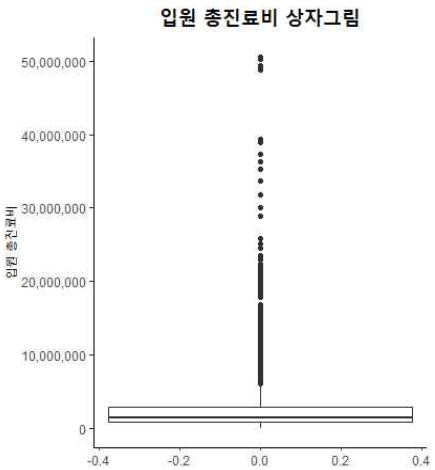
총진료비 변수의 문항은 유효, -1(해당사항 없음), -9(모름/무응답)로 구성된다. '해당사항 없음'은 상세 금액이 확인되지 않은 경우이다. 총진료비가 -9로 표시된 행들의 공통점은 건보부담금, 법정본인부담금, 비급여 항목 모두 -9로 처리되었고, 수납금액만 표시되었다는 것이다. 수납금액이란, 법정본인부담금과 비급여를 더한 값이다. 총진료비의 -9를 결측값으로 처리하고 결측값 대체 과정을 수행하는 것보다 수납금액을 그대로 사용하는 것이 상대적으로 정확한 값이 나올 것이라고 판단했다. 따라서 -9(모름/무응답)를 해당 행의 수납금액과 동일하게 바꾸었고, -1(해당사항 없음)은 결측값(NA)으로 처리했다.

처방약값의 문항 구성은 유효, 91(무료), 0(의료급여자, 건보무료검진), -1(해당사항 없음), -9(모름/무응답)이다. 여기에서의 -1(해당사항 없음)은 총진료비의 -1(해당사항 없음)과 다르다. 처방약값의 -1은 약국에 방문하지 않음을 의미한다. 때문에 91(무료)과 -1(해당사항 없음)을 0으로 바꾸었으며, -9(모름/무응답)는 결측값으로 처리하고 결측값 대체 과정을 수행했다.

결측값을 대체할 값으로는 평균과 중앙값 중에 중앙값을 사용했다. 평균은 모든 자료를 이용한 값으로 자료의 대표성을 갖고 있지만 극단값이 존재하는 경우 대표성을 잃을 수 있다. 반면, 중앙값은 평균보다 대표성은 떨어지지만 극단값이 있을 때 평균보다 상대적으로 자료의 특징을 잘 나타낼 수 있다. 세 의료서비스의 총진료비와 처방약값 모두 극단값이 다수 존재하므로 결측값을 중앙값으로 대체하는 것이 적합하다고 판단했다.

개인별 연간 응급, 의료, 외래 지출의료비 데이터를 각각 'dt_er', 'dt_in', 'dt_ou' 라고 설정했다. 아래 표에 종속변수 데이터 전처리 과정을 정리했다.

<그림 3.2.7> 입원 총 진료비 상자그림



※ 위 사진은 결측값을 제외한 입원 총진료비의 상자그림이다. 점으로 표시된 것이 극단값인데 이 값들을 포함한 평균을 대체값으로 하면 자료의 대표성을 잃을 수 있다. 따라서 중앙값으로 대체하는 것이 적합하다고 판단했다.

<표 3.2.8> 총진료비 변수 전처리

전처리 전(기존 문항)	전처리 후
유효	이전과 동일
해당사항 없음(-1)	NA → 중앙값
모름/무응답(-9)	동일 행의 수납금액

<표 3.2.9> 처방약값 변수 전처리

전처리 전(기존 문항)	전처리 후
유효	이전과 동일
무료(91)	0
의료급여자/건보무료검진(0)	이전과 동일
해당사항 없음(-1)	0
모름/무응답(-9)	NA → 중앙값

<표 3.2.10> 종속변수 데이터 전처리 과정

원본 데이터		복사본 생성		처방약값/총진료비 전처리		개인지출 의료비 산출		최종
er	➔	mof_er	➔	ifelse() + 결측값 대체	➔	group_by() + summarise() + sum()	➔	dt_er
in		mof_in						dt_in
ou		mof_ou						dt_ou

(3) 각 요인별 의료 이용 현황과 의료비 지출 현황

가공한 독립변수 데이터 'b1'과 종속변수 데이터 'dt_er', 'dt_in' ,'dt_ou'를 PIDWOM 변수 기준으로 결합했다. 최종적으로 생성된 데이터를 각각 b1_er, b1_in, b1_ou라고 설정했다. 이 데이터를 이용해서 응급, 입원, 외래별 의료이용 현황과 평균의료비 지출을 분석하고 해석했다.

<표 3.2.11> 65세 이상 노인의 인구·경제적 요인과 건강행위 및 건강상태에 따른 응급 의료이용 및 의료비 지출(의료비 단위: 원)

변수		응급		
		명	%	개인지출 의료비
성별	1: 남자	222	41.9	381,011
	0: 여자	308	58.1	272,335
세대구성	1: 단독가구	107	20.2	276,895
	2: 부부가구	259	48.9	363,974
	3: 부부+자녀	53	10	353,645
	4: 기타	111	20.9	232,643
의료보장형태	1: 건강보험가입자	486	91.7	318,969
	0: 의료급여수급자	44	8.3	305,560
경제활동여부	1: 경제활동 함	160	30.2	254,554
	0: 경제활동 안 함	370	69.8	345,230

주택소유여부	1: 자가	385	72.6	318,512
	0: 전세, 월세 등	145	27.4	316,113
만성질환개수	0개	13	2.45	336,374
	1개	37	6.98	202,030
	2개	55	10.38	254,963
	3개	83	15.66	453,620
	4개	85	16.04	239,645
	5개 이상	257	48.49	329,076
흡연여부	1: 흡연	48	9.06	243,130
	0: 흡연안함	482	90.94	325,298
음주여부	1: 월 1회 이상	31	5.85	217,571
(과음)	0: 월 1회 미만	499	94.15	324,086
장애여부	1: 장애있음	104	19.6	449,455
	0: 장애없음	426	80.4	285,729
신체활동제한	1: 제한있음	91	17.19	478,063
	0: 제한없음	439	82.89	284,647
우울경험	1: 우울감 경험	54	10.2	276,456
	0: 우울감 경험안함	476	89.8	322,553
계		530	100.0	

① 응급

남성의 연평균 응급 의료비는 38만 원, 여성은 27만 원으로 남성이 더 지출한다. 일반적으로 우리나라 노인들의 가구 유형은 점점 노인부부가구 또는 단독가구로 유형이 축소되는 경향을 보이는 것으로 알려져 있다. 세대 구성에 따른 의료이용 형태를 보면 의료이용을 한 노인은 부부가구에서 가장 높았고, 의료비 지출은 부부가구와 부부 및 자녀가 36만 원과 35만 원으로 높았다. 건강보험 가입자의 연평균 응급실 의료비는 약 32만 원 정도로 의료급여 수급자의 30만 원보다 높지만 큰 차이는 없다. 노인의 절반 이상은 경제활동을 하지 않으며, 경제활동을 하지 않는 노인이 경제활동을 하는 노인보다 의료이용을 더 많이 한다. 평균 의료비도 경제활동을 하지 않는 노인은 약 34 만 원 정도로 경제활동을 하는 노인의 의료비인 25만 원보다 의료비의 지출이 더 많다.

빈도와 비율 분석결과 의외인 점이 있었다. 그것은 바로 주택소유여부이다. 의료이용자 기준으로 자가를 갖고 있는 사람이 그렇지 않은 사람보다 45.2% 더 많았다. 2020년과 2021년인 현재, 주택 값과 땅 값이 상당히 상승했다는 것을 뉴스와 기사를 통해 심심치 않게 볼 수 있다. 과거에는 부동산 매매 문제가 오늘날처럼 심각한 문제는 아니었을 지라도 자가를 소유하는 사람들보다 그렇지 않은 사람들이 훨씬 많을 것이라 생각했다. 때문에 이러한 결과는 예상 밖이라고 생각했다. 만성질환개수에 따른 응급 의료이용 현황 또한 특이하다. 보유한 만성질환개수가 많을수록 의료이용횟수와 지출 의료비가 많을 것 같지만 응급실을 이용한 사람들의 경우에는 만성질환을 3개 갖고 있는 사람의 의료비 지출이 더 많다. 하지만 응급실 이용 비율은 만성질환을 5개 이상 보유한 사람들의 비율이 더 높다.

비흡연자는 흡연자보다 의료이용을 더 많이 경험하며, 연평균 의료비도 더 많은데, 이는 흡연유무에 따른 의료이용의 차이를 흡연으로 인해 발생한 질병의 치료를 위한 의료이용 보다는 건강에 대한 인식의 차이를 대리하는 변수로 해석할 수 있다. 음주여부 결과도 흡연여부와 마찬가지로 건강에 대한 인식의 차이를 보여준다고 해석할 수 있다. 장애를 갖고 있는 사람이 지출한 의료비는 장애가 있는 사람보다 약 16만 원 더 많았다. 신체활동에 제한이 있는 노인과 제한이 없는 노인의 평균 의료비 지출이 약 20만 원 차이가 나므로 신체활동 제한 여부는 의료이용과 높은 상관관계가 있다고 볼 수 있다. 우울경험에 대해서는 우울증을 경험한 사람이 응급실 이용 지출비가 약 5만 원 더 많았다.

<표 3.2.12> 65세 이상 노인의 인구·경제적 요인과 건강행위 및 건강상태에 따른 입원 의료이용 및 의료비 지출(의료비 단위: 원)

변수		입원		
		명	%	개인지출 의료비
성별	1: 남자	375	40.5	5,455,995
	0: 여자	551	59.5	5,708,217

세대구성	1: 단독가구	225	24.3	4,733,263
	2: 부부가구	448	48.4	5,979,719
	3: 부부+자녀	100	10.8	5,733,612
	4: 기타	153	16.5	5,712,197
의료보장형태	1: 건강보험가입자	852	92.01	5,665,181
	0: 의료급여수급자	74	7.99	4,925,565
경제활동여부	1: 경제활동 함	284	30.7	4,567,047
	0: 경제활동 안 함	642	69.3	6,065,708
주택소유여부	1: 자가	676	73.0	5,627,269
	0: 전세, 월세 등	250	27.0	5,548,767
만성질환개수	0개	18	1.94	2,332,897
	1개	58	6.26	5,400,255
	2개	84	9.06	5,496,768
	3개	139	14.99	4,097,266
	4개	143	15.43	5,352,323
	5개 이상	484	52.32	6,279,727
흡연여부	1: 흡연	84	9.07	5,523,329
	0: 흡연안함	842	90.93	5,614,330
음주여부	1: 월 1회 이상	47	5.08	3,870,454
(과음)	0: 월 1회 미만	879	94.92	5,698,879
장애여부	1: 장애있음	161	17.4	5,857,065
	0: 장애없음	765	82.6	5,553,252
신체활동제한	1: 제한있음	147	15.9	8,726,001
	0: 제한없음	779	84.1	5,017,334
우울경험	1: 우울감 경험	95	10.3	6,674,159
	0: 우울감 경험안함	831	89.7	5,483,972
계		926	100.0	

② 입원

여성의 연평균 입원 의료비는 약 570만 원, 남성의 연평균 입원 의료비는 약 545만 원으로 여성이 남성보다 약 25만 원 더 지출한다. 세대구성은 응급과 마찬가지로 부부가구의 이용률이 더 높으며 의료비 지출은 부부가구, 부부 및 자녀, 기타, 단독가구 순으로 높다. 건강보험 가입자의 연평균 입원 의료비는 약 566만 원 정도로 의료급여 수급자의 492만 원보다 약 74만 원 많다. 노인의 절반 이상은 경제활동을 하지 않으며, 경제활동을 하지 않는 노인이 경제활동을 하는 노인보다 의료이용을 더 많이 한다. 평균 의료비도 경제활동을 하지 않는 노인은 약 607만 원으로 경제활동을 하는 노인의 의료비인 456만 원보다 의료비의 지출이 더 많다. 자가를 소유한 사람은 소유하지 않은 사람보다 약 8만 원을 더 지출한다.

그리고 보유한 만성질환이 많을수록 의료비 지출이 더 많으며, 위의 응급 의료비 지출과 다르게 보유한 만성질환개수가 3개인 사람의 의료비 지출이 만성질환을 1개, 2개 갖고 있는 사람보다 적었다. 흡연여부와 음주여부에 대한 결과 해석은 앞선 응급 이용 현황과 마찬가지로 건강에 대한 인식차이로 해석할 수 있다. 신체활동에 제한이 있는 노인과 제한이 없는 노인의 평균 의료비 지출이 약 370만 원 정도 차이가 나므로 신체활동 제한 여부는 의료이용과 높은 상관관계가 있다고 볼 수 있다.

<표 3.2.13> 65세 이상 노인의 인구·경제적 요인과 건강행위 및 건강상태에 따른 외래 의료이용 및 의료비 지출(의료비 단위: 원)

변수		외래		
		명	%	개인지출 의료비
성별	1: 남자	1,753	41.9	1,518,837
	0: 여자	2,432	58.1	1,574,756
세대구성	1: 단독가구	919	22.0	1,645,583
	2: 부부가구	2,104	50.3	1,602,275
	3: 부부+자녀	476	11.4	1,369,795
	4: 기타	686	16.4	1,394,796
의료보장형태	1: 건강보험가입자	3,937	94.07	1,556,751

	0: 의료급여수급자	248	5.93	1,465,318
경제활동여부	1: 경제활동 함	1,518	36.3	1,501,929
	0: 경제활동 안 함	2,667	63.7	1,579,452
주택소유여부	1: 자가	3,087	73.8	1,590,965
	0: 전세, 월세 등	1,098	26.2	1,439,908
만성질환개수	0개	152	3.63	457,425
	1개	356	8.51	754,742
	2개	574	13.72	996,812
	3개	706	16.87	1,323,331
	4개	663	15.84	1,506,697
	5개 이상	1,734	41.43	2,104,227
흡연여부	1: 흡연	408	9.75	1,383,343
	0: 흡연안함	3,777	90.25	1,569,480
음주여부	1: 월 1회 이상	335	8.0	1,153,927
(과음)	0: 월 1회 미만	3,850	92.0	1,585,912
장애여부	1: 장애있음	591	14.1	1,888,828
	0: 장애없음	3,594	85.9	1,495,835
신체활동제한	1: 제한있음	361	8.63	1,741,031
	0: 제한없음	3,824	91.37	1,533,425
우울경험	1: 우울감 경험	312	7.46	1,907,021
	0: 우울감 경험안함	3,873	92.54	1,522,680
계		4,185	100.0	

③ 외래

여성의 연평균 외래진료비는 157만 원, 남성은 152만 원으로 큰 차이는 없지만 여성이 더 많이 지출한다. 세대구성에서 단독가구가 약 164만 원으로 가장 많이 지출하는 것을 알 수 있다. 건강보험 가입자의 연평균 외래 의료비는 약 155만 원으로 의료급여 수급자의 146만 원보다 약 9만 원 많다. 노인의 절반 이상은 경제활동을 하지 않으며, 경제활동을 하지 않는 노인이 경제활동을 하는 노인보다 의료이용을 더 많이 한다. 연평균 외래진료비도 경제활동을 하지 않는 노인은 약 158만 원으로 경제활동을 하는 노인의 의료비인 105만 원보다 의료비 지출이 더 많다.

보유한 만성질환개수가 많을수록 외래진료 이용 비율과 지출 의료비 모두 많은 것을 알 수 있다. 흡연여부와 음주여부는 앞서 언급한 두 의료서비스와 같은 결과를 보이고 있다. 흡연을 하지 않는 사람과 음주를 하지 않는 사람들 모두 그렇지 않은 사람들보다 지출 의료비가 각각 약 18만 원, 약 43만 원 높다. 장애가 있는 사람, 또는 장애가 있는 사람은 그렇지 않은 사람보다 약 40만 원 더 지출한다. 또한 신체활동에 제한이 있는 사람이 그렇지 않은 사람보다 약 21만 원 더 지출하므로 외래 진료에 있어서 장애여부, 신체활동제한 여부, 우울경험 변수는 외래진료 이용과 높은 상관관계가 높다고 할 수 있다.

④ 세 가지 의료서비스 분석 결과의 공통점

세대구성 변수에서 부부와 자녀가 함께 사는 가구는 의료이용 횟수와 지출하는 의료비가 다른 세대구성보다 높을 것이라 예상했지만 세 가지 결과 모두 그렇지 않고 오히려 의료이용회수가 가장 낮거나 지출하는 의료비는 부부가구와 비슷하게 가장 높은 의료비를 지출한다는 결과를 얻었다. 건강보험 가입자가 의료급여 수급자보다 더 많은 의료비를 지출했으며 65세 이상 노인들의 절반 이상이 경제활동을 하지 않는 것으로 나타났다. 그리고 경제활동을 하지 않는 노인들의 의료비 지출이 더 높았다. 또한 앞서 언급했듯이 자가를 소유하고 있는 노인들의 수가 많음을 알았으며 자가를 소유한 노인들이 의료비 지출을 더 많이 한다는 것을 알 수 있다. 이것은 자가를 소유하고 있다는 것이 상대적으로 부유하다고 생각했고 건강관리에 더 많은 투자를 할 수 있을 것이라 추측했다. 특이한 점은 흡연을 하지 않거나 음주를 하지 않은 사람의 지출 의료비가 더 높다는 것이다. 필자는 이것을 흡연과 음주로 인해 발생한 질병의 치료를 위한 의료이용 보다는 건강에 대한 인식의 차이를 대리하는 변수로 해석했다. 마지막으로 신체활동에 제한이 있으면 그렇지 않은 사람들보다 더 많은 의료비를 지출한다는 결과를 봤을 때 신체활동 제한 여부는 의료이용과 높은 상관관계가 있다고 해석했다.

Ⅳ. 분석 결과

1. 더미변수 생성

선형회귀분석은 종속변수와 독립변수 모두 연속형 변수이어야 한다. 하지만 독립변수에 범주형 변수가 있다면 그 독립변수를 '더미변수'로 변환 즉, '연속형 변수'스럽게 만들어서 회귀분석에 이용해야한다. 세대구성 변수는 1(단독세대), 2(부부가구), 3(부부+자녀), 4(기타)로 이루어진 범주형 변수인데, R에서는 연속형 변수로 인식한다. 이를 방지하기 위해서 0과 1로만 이루어진 '더미변수'로 변환했다.

더미변수는 원래 범주형 변수의 범수 개수보다 1개 적게 만들어진다. 예를 들어 원래 변수가 성별(남, 녀)이라면 남성여부 또는 여성여부 둘 중 하나만 만든다.(범주의 개수 2개, 더미변수 1개) 더미변수로 만들어지지 않고 생략되는 범주는 기준이 되는 값이라고 이해할 수 있다. 이를 참조항목(Reference group)이라고 한다. 즉, 더미변수화해서 회귀분석에 투입되지 않는 항목이다. 필자는 세대구성에서 '4(기타)'를 참조항목으로 지정했다. 더미변수까지 포함된 데이터 명을 응급, 입원, 외래 각각 'lm_er', 'lm_in', 'lm_ou' 라고 지정했다.

<표 4.1.1> 세대 구성 더미변수 생성

세대 구성		더미_단독	더미_부부	더미_부부자녀
1(단독가구)		1	0	0
2(부부가구)	➔	0	1	0
3(부부+자녀)		0	0	1
4(기타)		0	0	0

2. 응급 의료비 지출 회귀분석 결과

변수	비표준화 계수		표준화 계수	t	*p-value*	*VIF*
	B	SE	β			
(상수)	-402,457	352,011	-	-1.143	.253	-
성별	115,344	60,862	.082	1.895	.059·	1.021
나이	8,186	4,575	.080	1.789	.074·	1.086
장애여부	123,009	76,549	.071	1.607	.109	1.046
신체활동제한	127,495	83,052	.070	1.535	.125	1.110
F(p)	3.75(.005***)					
adj.R^2	.020					
Durbin-Watson	1.952					

·p<.1, *p<.05, **p<.01, ***p<.001

Reference group : 세대구성*기타

<그림 4.2.1> 응급 데이터(lm_er) 독립변수 상관행렬 히트맵

인구·경제적 요인과 건강상태 및 건강행위에 따른 요인(이하 건강 요인)들 중 어떤 요인이 65세 이상 노인들의 응급 의료비 지출에 영향을 미치는지 알아보기 위해 다중회귀분석을 실시했다. 종속변수에 영향력을 미치는 최소한의 변수들만 파악하기 위해 변수 선택 방법은 단계 선택법을 이용했다.

응급 의료비 지출에 영향을 미치는 변수의 영향력을 파악하기 위한 회귀분석에 앞서, 독립변수들의 상관관계를 살펴보았다. 이는 다중공선성을 확인하기 위해서이다. 다중공선성이란 회귀 분석에서 사용된 모형의 일부 독립변수가 다른 변수와 상관 정도가 높아, 데이터 분석 시 부정적인 영향을 미치는 현상을 말한다. 두 변수의 상관이 높다는 것은 같은 변수를 회귀분석에 두 번 넣은 것으로 적합한 회귀모형을 찾기 어렵다. 따라서 상관계수와 분산팽창지수를 살펴보았다. 분산팽창지수(*VIF: Variance Inflation Factor, 이하 VIF*)는 '분산이 팽창된 지수'라는 뜻이며 이는 회귀계수의(모형의 기울기) 표준오차(*S.E*)가 팽창된다는 말과 같다. 표준오차(*S.E*)가 커지면 회귀계수의 유의성을 판단할 때 구하는 $t\text{-}value(=\frac{\text{추정된 회귀계수}}{\text{회귀계수의 표준오차}})$가 작아져 회귀계수의 유의성을 판단하기에 어려워진다. 모든 변수든지 *t-value*의 절댓값이 커지면 회귀계수가 유의하다는 결과가 나오기 때문이다. 이론적으로 *VIF*가 10 이상이면 다중공선성을 의심한다.

<그림 4.2.1>에서 독립변수들 간의 상관관계가 0.60수준 이하(가장 높은 상관은 더미_부부와 더미_단독으로 상관계수가 -0.49로 나타남)로 다중공선성 문제에 크게 유의하지 않았다. 또한 각각의 *VIF*가 10보다 훨씬 작으므로 공선성에 문제는 없다고 할 수 없기에 모든 변수를 독립변수로 하여 다중회귀분석을 실시하였다.

분석결과, 인구·경제적 요인에서는 성별과 나이 변수가, 건강 요인에서는 장애여부와 신체활동제한 변수가 응급 의료비 지출에 유의한 영향을 미친다. 결과표에 없는 변수는 응급 의료비 지출과 유의한 관련성이 없었다. 수정된 결정계수(adj.R^2)는 모델의 분산 설명력이다. 만들어진 모델(즉, 독립변수)이 얼마나 데이터를 잘 설명하는지를 의미한다. 수정된 결정계수(adj.R^2)가 0.02로 추정된 회귀모형의 설명력은 2%라고 할 수 있다. 이는 설명력이 매우 낮다고 할 수 있다. 하지만 이 값만으로 회귀모델 전체를 판단하기는 무리이다. 그리고 일반적으로 사람을 대상으로 한 데이터는 천차만별이므로 높은 R^2을 기대하기는 힘들다. 필자는 하나의 값에 집중하기보다는 전체 모형이 합리적으로 설명이 가능한지를 보기위해 노력했다.

남성은 여성보다 응급 의료비 지출을 115,344원 더 지출하고, 나이가 1살 증가할수록 8,186원 더 지출하는 것으로 나타났다. 또한 장애가 있는 사람은 그렇지 않은 사람보다 123,009원을 더 지출하며, 신체활동에 제한이 있는 사람은 그렇지 않은 사람보다 127,495원을 더 지출하는 것으로 나타났다. 하지만 어떤 성별이 의료비를 더 지출하는지에 대한 결과는 각 연구마다 다르다. 한국의료패널 자료를 이용한 연구도 어떤 연도의 데이터를 사용했는가에 따라 다르다. 본 연구는 응급 측면에서 남성이 더 많이 지출한다는 결과가 나왔지만 절대적인 것이 아님을 밝힌다. 유의한 4개의 변수들 중에서 성별(8.2%)이 응급 의료비 지출에 가장 큰 영향을 미

치고 그 다음은 나이(8.0%), 장애여부(7.1%), 신체활동제한(7.0%) 순이다. 성별이 나이보다 영향력이 크다는 결과를 얻었지만 수치상으로 큰 차이가 없다. 또한 일반적으로 성별보다 나이의 증가가 의료비 지출에 더 큰 영향을 주기에 본 결과에 대한 영향력은 확신하기 어렵다고 판단했다.

본 회귀모델은 인구·경제적 요인에서 합리적인 설명을 할 수 없지만, 건강 요인에서 장애여부와 신체활동이 응급 의료비 지출과 연관성이 있다는 것에서 결과의 정확성을 이야기할 수 있을 것이다.

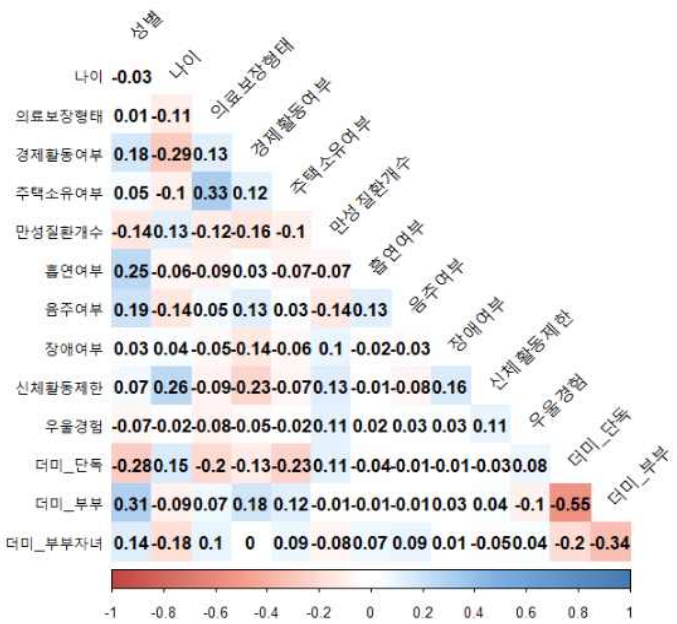
3. 입원 의료비 지출 회귀분석 결과

변수	비표준화 계수		표준화 계수	t	*p-value*	*VIF*
	B	SE	β			
(상수)	4,662,604	631,184	-	7.387	.000***	-
더미_단독	-1,348,340	621,097	-.071	-2.171	.030*	1.029
경제활동여부	-893,872	593,846	-.051	-1.505	.133	1.088
만성질환개수	209,514	100,679	.069	2.081	.038*	1.047
신체활동제한	3,209,098	742,862	.144	4.320	.000***	1.069
$F(p)$	9.288(.000***)					
adj.R^2	0.035					
Durbin-Watson	2.076					

·p<.1, *p<.05, **p<.01, ***p<.001

Reference group : 세대구성*기타

<그림 4.3.1> 입원 데이터(lm_in) 독립변수 상관행렬 히트맵



인구·경제적 요인과 건강 요인들 중 어떤 요인이 65세 이상 노인들의 입원 의료비 지출에 영향을 미치는지 알아보기 위해 다중회귀분석을 실시했다. 변수 선택 방법은 단계 선택법을 이용했다.

회귀분석에 앞서, 독립변수들의 상관관계와 *VIF*를 살펴보았다. <그림 4.3.1>에서 알 수 있듯이 독립변수들 간의 상관관계가 0.60수준 이하(가장 높은 상관은 더미_부부와 더미_단독으로 상관계수가 -0.55로 나타남)로 다중공선성 문제에 크게 유의하지 않았다. 또한 각각의 *VIF*가 10보다 훨씬 작으므로 공선성에 문제는 없기에 모든 변수를 독립변수로 하여 다중회귀분석을 실시하였다.

분석결과, 인구·경제적 요인에서는 세대구성 중 단독가구와 경제활동여부 변수가, 건강 요인에서는 만성질환개수와 신체활동제한 변수가 입원 의료비 지출에 유의한 영향을 미친다. 결과표에 없는 변수는 입원 의료비 지출과 유의한 관련성이 없었다. 수정된 결정계수(adj.R^2)가 0.035로 해당 모델은 데이터를 설명하기에 다소 낮은 설명력을 갖는다고 할 수 있다. 2인 이상인 세대는 단독세대보다 입원 의료비를 1,348,340원 더 지출하는

것으로 나타났다. 가구원수가 많을수록 지출하는 입원 의료비가 많다고 생각할 수 있지만 어디까지나 추측이다. 정확한 설명이 가능하려면 세대구성에서 2인 이상에 해당되는 부부가구, 부부자녀자구, 기타가구가 유의한 변수로 나타나야하기 때문이다. 경제활동을 하지 않는 사람은 경제활동을 하는 사람들보다 893,872원 더 지출하는 것으로 나타났다. 또한 보유한 만성질환 개수가 1개씩 증가할 때마다 209,514원을 더 지출하며, 신체활동에 제한이 있는 사람은 그렇지 않은 사람보다 무려 3,209,098원 더 지출하는 것으로 나타났다. 이것은 신체활동이 다른 요인들보다 입원 의료비 지출에 가장 큰 영향을 미친다는 것을 시사한다. 유의한 4개의 변수들 중에서 신체활동제한(14%)이 입원 의료비 지출에 가장 큰 영향을 미치고 그 다음은 단독가구(7.1%), 만성질환개수(6.9%), 경제활동여부(5.1%) 순이다. 신체활동제한이 다른 변수들보다 두 배 이상의 영향을 미치는 것은 설득력이 있다. 다만 만성질환개수 변수가 단독가구보다 영향을 덜 미친다는 것은 일반적으로 보이지 않고 수치로도 큰 차이가 없다.

본 회귀모델에서 신체활동여부와 만성질환개수가 입원 의료비에 유의한 영향을 준다는 것은 합리적인 설명이 가능해 보인다. 하지만 표준화계수로 어떤 요인이 더 영향을 미치는가는 수치로만 판단하기 어렵다.

4. 외래 의료비 지출 회귀분석 결과

변수	비표준화 계수		표준화 계수	t	*p-value*	*VIF*
	B	SE	β			
(상수)	1,731,950	382,725	-	4.525	.000***	-
성별	171,452	66,406	.042	2.582	.010**	1.230
나이	-26,269	4,721	-.086	-5.564	.000***	1.094
더미_단독	222,257	88,468	.046	2.512	.012*	1.536
더미_부부	150,211	70,767	.037	2.123	.034*	1.434
의료보장형태	433,213	134,092	.051	3.231	.001**	1.148
주택소유여부	190,211	71,770	.042	2.650	.008**	1.142
만성질환개수	239,420	12,431	.299	19.259	.000***	1.111
음주여부(과음)	-335,693	115,592	-.045	-2.904	.004**	1.127
장애여부	269,272	86,135	.047	3.126	.002**	1.030
우울경험	195,618	114,070	.026	1.715	.086·	1.028
F(p)	44.52(.000***)					
adj.R^2	.094					
Durbin-Watson	1.940					

·p<.1, *p<.05, **p<.01, ***p<.001

Reference group : 세대구성*기타

<그림 4.4.1> 외래 데이터(lm_ou) 독립변수 상관행렬 히트맵

인구·경제적 요인과 건강 요인들 중 어떤 요인이 65세 이상 노인들의 외래 의료비 지출에 영향을 미치는지 알아보기 위해 다중회귀분석을 실시했다. 변수 선택 방법은 단계 선택법을 이용했다.

회귀분석에 앞서, 독립변수들의 상관관계와 *VIF*를 살펴보았다. <그림 4.4.1>에서 알 수 있듯이 독립변수들 간의 상관관계가 0.60수준 이하(가장 높은 상관은 더미_부부와 더미_단독으로 상관계수가 -0.53로 나타남)로 다중공선성 문제에 크게 유의하지 않았다. 또한 각각의 *VIF*가 10보다 훨씬 작으므로 공선성에 문제는 없다고 할 수 있다. 따라서 모든 변수를 독립변수로 하여 다중회귀분석을 실시하였다.

분석결과, 인구·경제적 요인에서는 성별, 나이, 세대구성 중 단독가구, 부부가구, 의료보장형태, 그리고 주택소유여부 변수가 입원 의료비 지출에 유의한 영향을 미친다는 결과를 얻었다. 건강 요인에서는 만성질환개수, 음주여부, 장애여부, 우울경험 변수가 입원 의료비 지출에 유의한 영향을 미친다. 결과표에 없는 변수는 외래 의료비 지출과 유의한 관련성이 없었다. 수정된 결정계수(adj.R^2)가 0.094로 추정된 회귀모형의 설명력, 즉 독립변수가 종속변수를 9.4% 정도 설명한다. 다소 낮은 설명력이지만 응급과 입원의 회귀모델보다 설명력이 높음을 알 수 있다. 이는 아마도 노인(뿐만 아니라 모든 연령이)들이 세 가지 의료서비스 중에서 외래진료를 가장 많이 이용하므로 외래진료에 대한 데이터가 많이 저장되어 있어 응급과 입원보다 상대적으로 적합한 회귀모델이 만들어진 것으로 추정할 수 있다.(응급 데이터 레코드 수: 530개, 입원 데이터 레코드 수: 927개, 외래 데이터 레코드 수: 4,185개)

남성은 여성보다 외래 의료비를 171,452원 더 지출하고, 나이가 1살 증가할수록 26,269원 덜 지출하는 것으로 나타났다. 나이가 증가할수록 의료비를 더 지출하는 것이 일반적이다. 하지만 본 회귀모델에서는 반대의 결과가 나왔다. 전처리 과정에 문제 여부를 보았지만 발견되지 않았다. 단독가구와 부부가 함께 사는 사람들은 각각 222,257원, 150,211원을 더 지출하는 것으로 나타났다. 건강보험에 가입한 사람은 의료급여를 받는 사람들보다 433,213원을 더 지출하는 것으로 나타났다. 자가를 갖고 있는 사람은 그렇지 않은 사람들보다 190,211원을 더 지출한다. **Ⅲ-2)-(3)**에서 밝힌 바와 같이 자가를 갖고 있다는 것은 상대적으로 부유하다는 것으로 건강관리에 더 투자할 수 있음을 나타내는 간접적인 지표라고 할 수 있다. 또한 보유한 만성질환 개수가 1개씩 증가할 때마다 239,420원 더 지출하며, 과음을 하는 사람은 과음하지 않는 사람들보다 335,693원을 덜 지출한고 나타났다. **Ⅲ-2)-(3)**에서 밝힌 것처럼 과음으로 인해 발생한 질병의 치료를 위한 의료이용 보다는 건강에 대한 인식의 차이를 대리하는 변수로 해석할 수 있다. 과음을 하지 않는 사람은 과음하는 사람보다 상대적으로 건강에 대한 인식이 높아 외래 이용과 건강 검진 등으로 의료 지출비가 더 높다는 것이다. 그리고 장애가 있는 사람은 장애가 없는 사람보다 269,272원 더 지출하고, 우울증을 경험한 사람은 그렇지 않은 사람들보다 195,618원 더 지출한다.

유의한 10개의 변수들 중에서 만성질환개수 변수가 무려 29.9%로 외래 의료비 지출에 가장 큰 영향을 미치고 그 다음은 나이(8.6%), 의료보장형태(5.1%), 장애여부(4.7%) 순이다. 만성질환이 의료비 지출에 큰 영향을

준다는 것은 설득력이 있다. 종속변수에 대한 영향력만 봤을 때 나이를 인구·경제적 요인으로만 국한할 것이 아니라 나이가 증가하면 신체건강에 직접적인 영향을 미치므로 의료비 지출에 두 번째로 영향을 준다는 결과를 받아들일 수 있다.

본 회귀모델은 나이 변수와 세대구성 변수에서 합리적인 설명이 어렵다. 세대구성원 수가 많을수록 의료비 지출이 증가할 것이라 생각했지만 단독가구와 부부가구만 유의한 결과가 나와서 전체 모델이 유의한가에 대한 의문점이 있다.

Ⅴ. 결론

이상으로 R과 한국의료패널 데이터를 활용하여 세 가지 의료서비스에 대한 노인의 의료비 지출에 영향을 미치는 요인을 살펴보았다. 주로 노인의 의료비에 영향을 미치는 요인은 인구·경제적 요인보다는 건강상태 및 건강행위에 따른 요인이었다. 그 중에서 만성질환개수, 장애여부, 그리고 신체활동제한이 의료비 지출에 많은 영향을 주고 있었다. 외래 의료비 지출에 대한 회귀모델을 보면, 건강보험에 가입한 사람이 의료급여 수급자보다 의료비 지출을 더 많이 하고 있었다. 이러한 결과는 한국의 의료보장제도인 전 국민 대상의 건강보험정책과 의료급여 제도가 비교적 잘 시행되어 노인의 소득이나 경제활동으로 인한 의료서비스 이용불균형을 어느 정도 해결하고 있다고 볼 수 있다.

연구에 사용한 한국의료패널 데이터는 데이터 활용 동의서를 작성한 후 얻을 수 있다. 의료 데이터도 개인정보로 취급되면서 이에 대한 보안의 중요성과 우려가 커지고 있다. 국민의 건강을 보장한다는 명분으로 의료데이터를 활용하지만 진화하는 기술 보안 범죄가 법과 제도를 앞설 수 있기 때문이다. 현재 우리나라에서 보건의료데이터는 건강검진과 노인 등 동일집단에 한해 일부 개방하거나 공익적 연구에만 제한적으로 활용된다. 한국의료패널은 연구자의 소속, 활용목적, 성명 등의 동의서를 요구하여 보안 시스템 하에 자료를 주고받는 것에 의미가 있다.

의료비 지출에 대한 정책과 이를 점검하는 것은 사회공동체와 국가의 몫이지만 빠져나가는 의료비를 직접 관리하는 것은 사회와 국가가 아닌 개인의 몫이다. 개인은 본인의 의료비 지출 상태를 정확하게 알고 이에 따라 재정지출을 안정적으로 관리해야 할 것이다. 가장 먼저 본인의 건강상태를 파악하는 것이 우선이다. 보건복지부에 따르면 2020년 기준 일반 건강검진, 의료급여 생애전환기검진은 모든 검진 비용을 국민건강보험공단에서 전액 부담한다고 한다. 일반건강검진의 대상자는 지역가입자의 경우 세대주와 만 20세 이상 세대원, 피부양자의 경우 만 20세 이상 대상자, 직장가입자의 경우 비사무직 전체와 격년제 실시에 따른 사무직 대상자, 그리고 의료급여 수급자의 경우 만 19세~만 64세 대상자이다. 국가에서 제공하는 건강복지혜택을 통해 청년기부터 건강상태를 점검하고 예방하여 미래의 불필요한 지출을 막을 수 있을 것이다.

코로나19로 인해 헬스케어 분야에서 변화가 일어나고 있다. 사회적 거리두기의 중요성이 커지면서 병원도 불필요한 접촉을 피하고 분산하기 위한 거리두기 정책을 중요시하고 있다. 그 중 하나가 원격의료이다. 이로 인해 병원의 기능이 치료 중심에서 예방 중심으로 바뀌는 전환점이 만들어질 것이다. 또한 4차 산업혁명으로 digital transformation이 가속화되어 의료분야와 빅데이터, AI와 같은 IT 분야가 융합하여 헬스케어의 패러다임을 바꾸는데 기여할 것이다. 해외 사례를 보면, 미국의 'Mercy Virtual'은 최초의 원격의료 거점센터이다. 원격의료와 다양한 디지털 의료 서비스를 뜻하는 가상의료 서비스를 통해 재입원율이 50% 하락했으며 35% 정도 재원기간이 단축되었다. 이처럼 각 새로운 요소들이 의료 산업에 합쳐져서 새로운 패러다임에 맞는, 적응하는 구조로 병원의 모습이 변할 것이다. 인공지능, 웨어러블 기기, 의사 및 의료기관과의 원격 커뮤니케이션은 의료서비스 제공과 인간 건강상태를 모니터링하는 데 있어서 점진적으로 익숙한 도구가 되고 있다. businesswire[7]의 통계자료에 따르면 2019년 가장 주목할 만한 디지털 헬스 기술은 AI(42.0%)이고 앱(27.0%)이 그 뒤를 이었다.

그리고 일차의료의 변화를 주목할 만하다. 바로 '개인 맞춤형 의료서비스'이다. 의학서비스 + 건강위험인자의 관리 + 사회적인서비스(돌봄, 주거, 영양 등)가 합친 통합서비스로 개인화된 치료를 위해 질병 경로를 추적할 수 있다. 이러한 추세는 환자, 의사 및 기타 의료 전문가 간의 상호작용을 근본적으로 바꿀 것이다. 다시 말해

7) 뉴스 매체, 금융 시장, 공개 시스템, 투자자, 정보 웹 사이트, 데이터베이스, 블로거, 소셜 네트워크 및 기타 대상에게 전 세계 수천 개의 회사 및 조직에서 전체 텍스트 보도 자료를 배포하는 미국 회사

서 의료서비스의 새로운 소통의 장이 열릴 것이다. 기대할 수 있는 효과는 불필요한 의료비를 지출하지 않고 예측 가능한 질병을 초기에 예방하여 의료서비스의 만족도가 증가하는 것이다. 의료서비스가 더 이상 고비용 저효율이 아니라 중비용 고효율로 인식이 변화할 것이다.

본 연구의 한계점은 다음과 같다. 첫째, 의료서비스의 개선점을 제안하는데 한계가 있다. 연구 결과를 보고 의료정책의 문제점 혹은 개선해야할 부분을 구체적으로 밝히고 제시하는 것은 필자로서 추상적인 제안만 가능할 뿐이고, 능력 범위를 벗어나기 때문이다.

둘째, 몇 개의 변수에서 보편적으로 알고 있거나 일반적인 사실과 다른 결과를 얻었다. 그 중 하나는 나이의 증가와 외래 의료비 증가 간의 유의성이다. 일반적으로 나이가 증가할수록 신체 기능이 저하되고 만성질병에 걸릴 가능성이 높아져 나이가 의료비 지출을 증가시키는 데에 큰 영향을 미칠 것이라 예상했다. 하지만 그렇지 않았다. 나이가 증가할수록 지출하는 외래 의료비가 감소한다는 결과가 나왔다. 향후 연구에서 인구·경제적 요인만 독립변수로 하여 진행할 필요가 있다. 또 다른 측면으로 의료비 지출에 대한 연령의 영향은 전 연령에 걸쳐 일정하지 않다. 고령일수록 지출해야 할 의료비가 더 증가한다는 가정을 할 수 있다. 20대에서 30대로 연령이 올라갈 때의 차이와 50대에서 60대, 더 나아가 70대로 접할 때의 차이는 같은 증가폭이지만 의료비 지출의 차이는 크게 날 수 있다. 즉, 기존의 연령 변수는 전형적인 회귀 방정식인 '$y=\alpha+\beta x$' 형태를 따르지만, 고 연령층일수록 증가폭이 커지는 비선형 관계를 설명하기 위해서는 높은 차수 항을 회귀 모델에 추가해 다항식으로 만드는 방법이 있다. 따라서 연령에 관한 변수는 다음과 같이 모델링을 할 수 있을 것이다. '$y=\alpha+\beta_1 x+\beta_2 x^2$' 또한 세대구성에서 유의한 결과를 얻지 못했다. 단독가구보다 부부가구일 때, 부부가구 보다 부부자녀 가구일 때 지출의료비가 증가할 것이라고 예상했다. 하지만 더미변수로 변환한 4개('기타' 포함)의 변수 모두 유의한 결과가 아닌 일부만 유의한 결과가 나왔다. 세대구성 변수는 본래 21개의 문항으로 구성되었다. 분석의 편의를 위해 총 4개의 범주로 나누었고, '단독', '부부', '부부자녀'에 해당되지 않는 나머지 항목은 '기타' 항목으로 통일시켰다. 향후 연구에서는 좀 더 세분화된 기준으로 세대구성을 범주화할 필요가 있다.

셋째, 본 연구결과의 표준화계수로 각 독립변수가 종속변수에 얼마나 영향을 미치는지 수치로 비교하는 것에 한계가 있다. 일반적으로 알고 있는 사실과 다른 결과가 있었다. 그리고 표준화계수가 서로 큰 차이가 없는 변수들은 종속변수에 미치는 영향력에 우위를 판단할 수 없었다. 두 번째, 세 번째 한계점을 보면 도출된 회귀모델이 과연 유의한지, 합리적인 설명이 가능한지의 의문 때문에 결과에 대한 아쉬움이 있다.

넷째, 본 연구에서는 각 변수가 개별적으로 결과에 기여한 것만 고려했다. 하지만 어떤 변수는 각 특징이 결합되어 종속 변수에 더 큰 영향을 미칠 수 있다. 만성질환개수 변수와 신체활동제한 변수는 각자 의료비 지출을 증가시키는 결과가 나왔다. 하지만 이들이 결합된 영향이 각각의 합보다 줄 수 있는 영향이 크다고 가정할 수 있다. 추후 연구에는 상호작용 관계를 추가해서 진행하면 좀 더 적합한 회귀모델을 얻을 수 있을 것이다. 이 부분에서는 도메인 지식이 바탕이 된다면 모델의 성능을 개선시킬 수 있을 것이다.

다섯째, 개인지출 의료비의 범위가 한정되었다. 본 연구에서는 개인지출 의료비를 총진료비(건보부담금 + 법정본인부담금 + 비급여)와 처방약값을 합한 금액으로 정의했다. 노인 의료비 지출현황을 보다 정확하게 이해하기 위해서는 수납금액 뿐만 아니라 간병비, 교통비, 앰뷸런스 비용 등 간접의료비를 포함하여 의료비의 범위를 확대한 분석이 요구된다.

참 고 문 헌

[1] 데이터마케팅 공부방, “R 프로그래밍) 다중선형 회귀분석 사례(의료비 예측 모델링) (3)”, 2019. 7. 8, <https://blog.naver.com/bestinall/221580078436>

[2] 김영우, “Do it! 쉽게 배우는 R 데이터 분석”, 이지스퍼블리싱, 2020. 3. 31, pp. 125-173

[3] 김양균, ZDNet Korea, “보건의료 빅데이터 악용 위험이 활용 가치보다 클 수 있다”, 2021. 4. 30, <https://zdnet.co.kr/view/?no=20210430160554>

[4] 서울아산병원, 정신건강이야기(노년기 우울증에 대하여), <http://psy.amc.seoul.kr/asan/depts/psy/K/bbsDetail.do?menuId=862&contentId=213594>

[5] 한국과학기술단체총연합회, “제23차 한국과총-의학한림원-과학기술한림원 온라인 공동포럼 COVID-19와 우리나라 보건의료의 미래”, YouTube, 2021. 4. 9, <https://www.youtube.com/watch?v=U9ycjK3nQIA&t=6764s>

[6] 건강보험심사평가원, 국민건강보험공단, “2019년 건강보험통계연보”, 2020, p. 65.

[7] 통계청 통계개발원, “한국의 사회동향2014”, 2014. 12. 18, pp. 202-203.

[8] 한국경제연구원, “한국, 연평균 저출산·고령화 속도 OECD 37개국 중 가장 빨라”, 2021. 3. 3, pp. 2-3.

[9] 한국의료패널, “2016년 한국의료패널 기초분석 보고서(Ⅰ) - 가계부담의료비 구성, 의료이용, 민간의료보험, 주요 질병별 의료비 분석”, 2018. 12.

[10] 한국의료패널, “2016년 한국의료패널 기초분석 보고서(Ⅱ) - 질병 이환, 만성질환, 건강 행태와 건강 수준”, 2018. 12.

[11] 이윤환·강임옥·유창훈·장후선 외, “노인의료비 지출 현황 분석”, 제3회 한국의료패널 학술대회, 서울, 서울시여성가족재단, 2011. 12. 1, pp. 2-21.

[12] 황연희, “세대별 의료비 지출에 영향을 미치는 요인 분석”, 제3회 한국의료패널 학술대회, 서울, 서울시여성가족재단, 2011. 12. 1, pp. 437-450.

[13] 송태민, “앤더슨 행동모형을 이용한 노년기 외래의료서비스 이용에 대한 스트레스 취약요인의 매개효과 분석”, 보건사회연구 제33권 제1호, 2013. 3. 1, p. 569.

[14] 이현숙, 염영희, “패널데이터를 이용한 노인의 의료서비스 이용, 의료비 지출, 건강성과의 발달궤적 및 연량차에 관한”, 보건사회연구 제37권 제2호, 2017. 6. 30, p. 317-319

[15] 정영호, “한국의료패널로 본 의료이용 및 본인부담 의료비 지출”, 보건복지포럼 2011년 9월 통권 제179호, 2011. 9. 1, pp. 64-74.

[16] 황연희, “한국의료패널로 본 한국 노인들의 의료이용 및 의료비 지출”, 보건복지포럼 2011년 12월 통권 제182호, 학술저널, 2011. 12. 1, pp. 51-59.

[17] 신정우, “가계 의료비 지출의 결정 요인 분석”, 연세대학교 대학원 석사학위논문, 2007. 7. pp. 36-46.

[18] businesswire, “CES 2020 Digital Health: What to Expect from 2019’s Most Impactful Health Tech Trends”, 2020, <https://www.businesswire.com/news/home/20200106005636/en/CES-2020-Digital-Health>

[19] Choonghyun Ryu, Data quality diagnosis, 2021, https://cran.r-project.org/web/packages/dlookr/vignettes/diagonosis.html

부 록 1. 데이터 품질 진단 코드

```
# install.packages("dlookr")
# install.packages("dplyr")
library(dlookr)
library(dplyr)

# APPEN(성인가구원 부가조사) 진단 [흡연여부(S2), 음주여부(S22), 우울감(S44), 신체활동제한(SH117)]
diagnose(appen, S2, S22, S44, SH117)
diagnose_numeric(appen, S2, S22, S44, SH117)

# IND(가구원) 진단 [성별(C3), 출생년도(C4_0), 의료보장형태(C11), 경제활동여부(C24), 장애여부(C14)]
diagnose(appen, S2, S22, S44, SH117)
diagnose_numeric(ind, C3, C4_0, C11, C24, C14)

# HH(가구) 진단 [세대구성(B3), 경제활동여부(B7)]
diagnose(hh, B3, B7)
diagnose_numeric(hh, B3, B7)

# CD(만성질환) 진단 [만성질환여부(CD2)]
diagnose(cd, CD2)
diagnose_numeric(cd, CD2)

# ER(응급서비스) 진단 [총진료비(ER26_5), 처방약값(ER33)]
diagnose(er, ER26_5, ER33)
diagnose_numeric(er, ER26_5, ER33)

# IN(입원서비스) 진단 [총진료비(IN35_6), 처방약값(IN37)]
diagnose(In, IN35_6, IN37)
diagnose_numeric(In, IN35_6,IN37)

# ER(외래서비스) 진단 [총진료비(OU29_7), 처방약값(OU35)]
diagnose(ou, OU29_7, OU35)
diagnose_numeric(ou, OU29_7, OU35)
```

부 록 2. 데이터 품질 진단 보고서

데이터 품질 진단 보고서는 diagnose_repport()를 사용하였으며 다음과 같이 7개의 진단 보고서를 생성하였다.

구분	진단 변수	진단 결과
부록 2-1	C3(성별) C4_0(출생년도) C11(의료보장형태) C14(장애여부) C24(경제활동여부)	성별, 출생년도, 의료보장형태 변수는 결측값 없음 경제활동여부의 결측값은 -6(비해당_만 14세 미만)으로 총 2,150개 존재 장애여부의 결측값은 -1(비장애인)로 총 16,163개 존재
부록 2-2	B3(세대구성) B7(주택소유여부)	두 변수 모두 결측값 없음
부록 2-3	S2(흡연여부) S22(음주여부) S44(우울감) SH117(신체활동제한)	흡연, 음주, 우울감 변수의 결측값은 각각 3개, 4개, 3개 존재 신체활동제한의 결측값은 -1(비해당)은 5,827개, -6(설문대상 아님_만 65세 미만)은 7,155개 존재
부록 2-4	CD2(만성질환여부)	범주 개수 총 5개이며 결측값 없음
부록 2-5	ER26_5(총진료비) ER33(처방약값)	총진료비 최댓값 14,019,720원, 최솟값 3,000원, 결측값 684개 처방약값 최댓값 68,100원, 최솟값 500원, 결측값 1,954개
부록 2-6	IN35_6(총진료비) IN37(처방약값)	총진료비 최댓값은 50,595,450원, 최솟값 1,002원, 결측값 536개 처방약값 최댓값은 110,640원, 최솟값 500원, 결측값 3,456개
부록 2-7	OU29_7(총진료비) OU35(처방약값)	총진료비 최댓값은 10,133,411원, 최솟값 70원, 결측값 개수 70,175개 처방약값 최댓값은 705,380원, 최솟값 500원, 결측값 개수 132,284개

[1] 부록 2-1. Data Quality Diagnosis Report(ind).pdf
[2] 부록 2-2. Data Quality Diagnosis Report(hh).pdf
[3] 부록 2-3. Data Quality Diagnosis Report(appen).pdf
[4] 부록 2-4. Data Quality Diagnosis Report(cd).pdf
[5] 부록 2-5. Data Quality Diagnosis Report(er).pdf
[6] 부록 2-6. Data Quality Diagnosis Report(in).pdf
[7] 부록 2-7. Data Quality Diagnosis Report(ou).pdf