질병 진단 개수에 대한 알고리즘

정보통계학전공
권소연, 박혜인, 오채은, 최지민

요약

본 분석은 질병관리청의 2022년 국민건강영양조사를 활용하여 질병 진단 빈도에 어떤 변수들이 영향을 미치는지 분석하였다. 연관성 분석을 진행하던 중 다중공선성 문제를 발견하였으며 차원 축소를 이용해 문제를 해결하였다. 모든 변수가 포함된 자료에 Lasso 회귀분석을 적용한 모형과 요인분석 진행 후 요인 생성 변수가 포함된 자료에 단계적 선택법을 적용한 모형을 성능 평가 지표를 이용한 비교로 가장 적절한 모형을 결정하였다. 최종 모형으로 요인분석 진행 후 단계적 선택법을 적용한 포아송 회귀모형이 설명변수의 개수가 가장 적어 적절하다고 판단하였다. 분석 결과 '식습관별 연령대', '요당', '체질량지수', '맥압', '부모 질병 진단율', '착석 시간' 순으로 질병 빈도율에 영향력이 높음을 확인하였다. 모델의 질병 예측 알고리즘은 개인의 생활 양식, 유전 등의 요인을 바탕으로 개인의 건강 위험도를 예측함으로써, 현재 생활 습관 상태를 토대로 질병에 걸릴 위험 여부를 판단할 수 있으며 체계적으로 질병을 관리하는 효과를 불러올 수 있을 것이다.

1. 서론

현재, 의학기술의 발전과 영양 식품의 발달로 인해 고령자의 사망률이 줄어들었고, 인간의 평균 기대수명이 늘어났다. 세계질병부담연구 2021 보고서에 따르면 2050년 전 세계 기대수명은 2022년에 비해 평균적으로 약 5년 증가하는 추세를 보였다(이채린, 2024). 이러한 결과에 따라 인간의 영양 상태 개선과 질병 관리는 필수적인 요소로 자리 잡았다. 따라서 질병 예방과 운동 등 꾸준하고 전문적인 방식의 건강관리와 건강한 몸을 유지하며 가속화되는 노화를 늦추자는 '저속노화'는 하나의 소비 트렌드로 대두되고 있다(임성원, 2024).

하지만 올바른 식습관 성립과 질병 관리의 중요성을 강조하는 사회 분위기에도 불구하고, 건강한 생활 습관을 유지하지 못하여 암, 당뇨, 혈관질환 등의 비전염성 질병에 걸리는 사람들이 늘고 있다. 2022년 세계보건기구(WHO)는 전 세계적으로 질병에 의해 사망하는 사람의 75%가 암이나 뇌졸중과 같은 비전염성 질병(NCD)에 걸렸다고 발표한 바 있다(윤은숙, 2022). 특히 대한민국의 비감염성 질환 발병률은 높은 것으로 알려져 있다. 질병관리본부가 2018년 발간한 '2018 만성질환 현황과 이슈'에 따르면 만성질환은 전체 사망의 80.8%를 차지하고 있으며 사망 원인 1위에 해당한 바 있다(신종학, 2019).

비감염성 질병 예방의 핵심은 꾸준한 관리이다. 하지만 현대 사회에서 개인이 본인의 생활 습관을 모두 파악하여 시의적절하게 질병을 예방하기란 쉬운 일이 아니다. 따라서 개인이 생활 습관을 파악하고 결과적으로 질병에 걸릴 위험이 있는지에 대한 여부를 파악할 수 있는 통계적 지침이 마련되어야 한다. 이에 연구진들은 개인의 비감염성 질병 발병 확률에 대한 예측을 연구 목적으로 하고 기존의 통계 자료를 분석하여 사람들의 비감염성 질병 발병과 질병 발병 원인에 해당하는 다양한 요소들의 상관관계에 대해 하나의 알고리즘으로 표현하였다.

본 분석에서는 질병관리청의 국민건강영양조사 원시 자료를 이용하여 질병에 영향을 주는 요인들을 각각 분석하고, 질병과 이러한 요인의 상관관계에 대해 다각도로 탐구하였다. 또한, 이러한 분석을 기반으로 질병 예측 알고리즘을 제시하였다. 이를 통해 비감염성 질병에 걸리지 않은 사람들이 본인의 현재 생활 습관 상태를 토대로 질병에 걸릴 위험이 있는지에 대한 여부를 파악할 수 있을 것이다. 또한, 비감염성 질병에 걸린 사람들이 질병 예측 알고리즘을 활용해 체계적으로 질병을 관리하는 효과를 불러올 수 있을 것이다.

2. 본론

2.1 데이터 설명

질병관리청의 2022년도 국민건강영양조사 원시자료 기본 DB 영역 데이터를 활용하였다. 국민건강영양조사는 1995년 제정된 국민건강증진법 제16조에 근거하여 시행된 전국 규모의 건강 및 영양조사이며, 2007년부터 매년 진행되고 있다. 데이터는 국민의 건강 수준, 건강행태, 식품 및 영양 섭취 실태에 관한 정보 6,250개를 포함하고 있다. 653개 변수 중 분석과 관련된 32개를 선정하여 분석을 진행하였다. 분석에 포함한 변수는 <표1>에 정리하였다.

<표1> 변수 설명

자료유형	변수명	설명		
반응변수	dg_sum	질병 진단빈도	0~6	
범주형	city	시도	0: 비수도권	1:수도권
	sex3	성별·임신여부	0. 남자 1. 여자&임신O	2.여자&임신X
	marri_2	결혼상태	1.유배우자, 동거 3.사별	2. 별거 및 이혼 4. 미혼
	npins	보험가입여부	1.가입	2.미가입
	BH1	건강검진 여부	1.검진	2.미검진
순서형	stress	스트레스	1. 거의 안 느낌 3. 많이 느낌	2. 조금 느낌 4. 대단히 많이 느낌
	smoke_second	간접흡연 노출 정도	0. 노출 안 됨 1. 조금 노출	2. 많이 노출
	BE3_31	걷기 운동 일수	1. 운동 안 함 2. 1일 3. 2일 4. 3일 8.7일(매일)	5. 4일 6. 5일 7. 6일
	BE5_1	근력 운동 일수	1. 운동 안 함 2. 1일 3. 2일 4. 3일	2.1일 3. 2일 5. 4일 6. 5일 이상
	fh_sum	부모 질병 진단율	0. 질병 없음 3. 질병 3개 4. 질병 4개	1. 질병 1개 2. 질병 2개 5. 질병 5개
	uriglu	요당	0. 음성 2. 양성 +,++,+++	1. 미량+- 3. 양성 ++++
	HE_Uket2	요케톤	0. 음성 1. 양성 + 2. 양성 ++	3. 양성 +++
	HE_Ubld	요잠혈	0. 음성 2. 양성 + 3. 양성 ++	1. 미량+- 4. 양성 +++
연속형	eatout	외식횟수	1~7	
	ainc	소득	17~1500	
	EC_wht_23	주당 평균 근로시간	0~999	
	weightch	체중변화	-9~9	
	BD.total	1년간 음주량	0~999	
	sleep	수면시간	14~999	
	smoke_sum	흡연량	0~999	
	sit	착석 시간	30~1200	
	dis_bp	맥압	19~125.50	
	HE_BMI	체질량지수	13.54~46.72	
	HE_Uph	요산도	5~9	
	HE_Usg	요비중	1.001~1.050	
	N_SUGAR	당 섭취량	0.1897~23.9426	
	age	나이	10~80	
	foodfreq	일주일간 식사 빈도	0~2	
	N_EN	에너지 섭취량(Kcal)	2.683~82.201	
	fateat	지방 섭취량	0.4243~24.7967	
	N_NA	나트륨 섭취량	0~145.22	

2.2 데이터 전처리/EDA

2.2.1 반응변수 생성

9개의 질병(고혈압, 이상지질혈증, 당뇨병, 천식, 아토피, 알레르기비염, 부비동염, 중이염, 콩팥병)을 합쳐 '질병 진단빈도'를 반응변수로 형성하였다. 범주 중 질병의 개수가 5~6개인 데이터가 적어, 이를 질병 개수 4로 합쳐 0~4까지의 빈도를 가지도록 범주를 형성하였다. 질병을 많이 진단받은 사람일수록 건강의 위험도가 높다고 판단하여, '질병 진단빈도' 변수를 반응변수로 선택하였다. 반응변수를 통해, 질병 위험도에 미치는 여러 요인을 알아보고자 한다.

2.2.2 변수 생성

2.2.2.1 수도권

'수도권'은 17개 시도에 대해서 조사된 '지역'을 이용하여 생성한 범주형 변수이다. '지역'의 서울, 인천, 경기를 수도권 범주로 두었고, 그 외 지역은 비수도권으로 두었다.

2.2.2.2 결혼상태

'결혼상태'는 (유배우자, 동거 / 유배우자, 별거 / 사별 / 이혼 / 미혼 / 무응답) 6개의 범주로 구성된다. 유배우자, 별거 범주의 빈도가 29개로 그 수가 적고, 이혼의 연장선으로 별거를 생각하여 별거와 이혼의 범주를 합쳐 (유배우자, 동거 / 사별 / 별거 및 이혼 / 미혼 / 무응답) 5개의 범주를 가진 변수로 변환하였다.

2.2.2.3 성별 및 임신 여부

'성별 및 임신 여부'는 (성별, 임신 경험 여부) 변수 2개를 합쳐 생성한 범주형 변수다. 임신은 여성의 신체 변화에 많은 변화를 주기에 반응변수에 영향을 미친다고 판단하였다. 단, 남자의 경우 임신 여부를 다룰 수 없으므로 남자와 임신 경험이 없는 여자, 임신 경험이 있는 여자 3가지 범주를 가진 '성별 및 임신 여부' 변수를 형성했다.

2.2.2.4 체중변화량

'체중변화량'은 (체중 변화 여부, 체중 감소량, 체중 증가량) 변수 3개를 합쳐 생성한 연속형 변수이다. 값이 작을수록 체중이 감소하고, 값이 클수록 체중이 증가했음을 의미한다.

2.2.2.5 수면시간

'수면시간' 변수는 (주중 수면시간, 주말 수면시간) 변수 2개를 이용해, 각 일수를 곱한 뒤 일주일 수면시간을 계산하였다. 값이 클수록 일주일 수면시간이 많음을 의미한다.

2.2.2.6 흡연량

‘흡연량’은 (평생 일반담배 흡연 여부, 현재 일반담배 흡연 여부, 하루 평균 일반담배 흡연량, 궐련형 전자담배 평생 사용 여부, 궐련형 전자담배 현재 사용 여부, 궐련형 전자담배 하루 평균 흡연량) 변수 6개를 합쳐 생성한 연속형 변수로, 하루 평균 흡연하는 담배 수를 뜻한다. 점수가 높을수록, 하루 평균 흡연량이 많음을 의미한다.

2.2.2.7 간접흡연 노출

‘간접흡연 노출’은 (가정 내 간접흡연 노출 여부, 공공장소 실내 간접흡연 노출 여부) 변수 2개를 합쳐 생성한 순서형 변수이다. 점수가 높을수록 개인의 간접흡연 노출 정도가 높음을 의미한다.

2.2.2.8 착석 시간

‘착석 시간’은 (평소 하루 앉아서 보내는 시간(시간), 평소 하루 앉아서 보내는 시간(분)) 변수 2개를 합쳐 생성한 연속형 변수로, 분 단위로 구성된다. 점수가 높을수록, 앉아서 보내는 시간이 많음을 의미한다.

2.2.2.9 부모 질병 진단율

‘부모 질병 진단율’은 고혈압, 고지혈증, 허혈성 심장질환, 뇌졸중, 당뇨병 5개의 질병에 대한 부모의 의사진단 여부를 종합하여 생성한 순서형 변수이다. 점수가 높을수록 부모의 질병 진단 빈도가 높음을 의미한다.

2.2.2.10 맥압

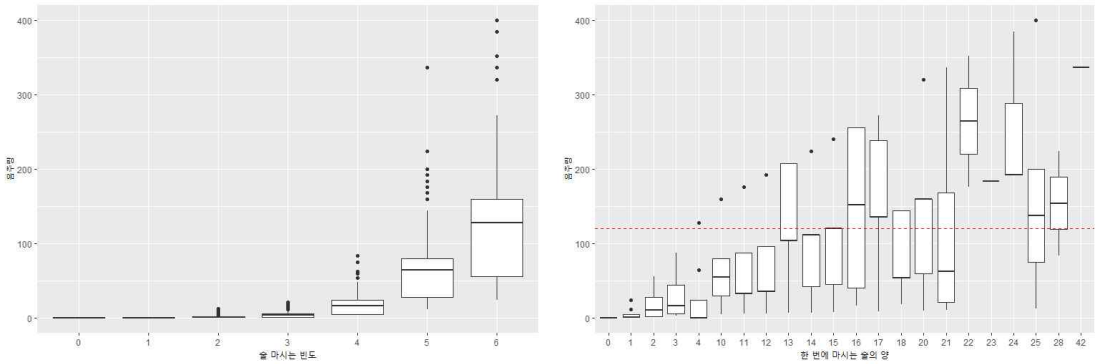
맥압은 최종 수축기 혈압과 이완기 혈압을 차감하여 환산된 수치이다. 높은 맥압은 혈관이 탄력성을 잃고 경직되어 수축기 혈압이 상승하고, 이완기 혈압이 저하되는 상태를 의미한다 (정은영, 2022). 나이에 따라 수축기, 이완기 혈압의 높고 낮음이 영향을 미치므로, 혈압 자체보다 두 혈압의 차이를 통해 동맥 건강을 판단하기로 하였다. 맥압이 높을수록 좋지 않은 건강 관리가 필요한 상태임을 의미한다.

2.2.2.11 식사 빈도

‘식사 빈도’는 1주일 중 식사 빈도를 확인하기 위해, (최근 1주 동안 아침 식사 빈도, 점심 식사 빈도, 저녁 식사 빈도) 변수 3개를 합쳐 형성한 연속형 변수이다. 숫자가 클수록 식사 빈도가 작은 관계이므로, 역코딩을 진행하였다. 점수가 클수록 식사 빈도가 높다.

2.2.2.12 한 달 음주량

한 달 동안 섭취하는 음주량을 확인하기 위해 (평생 음주 경험 여부, 음주 빈도, 1회 음주량) 변수 3개를 합쳐서 생성한 연속형 변수이다. <그림1>에서 한 달 음주량과 평생 음주 경험 여부, 음주 빈도와의 연관성을 보면 음주 빈도가 높을수록 한 달 동안 섭취하는 음주량에는 그대로 영향을 끼쳤지만, 한 번에 마시는 술의 양이 많을수록 한 달 동안 섭취하는 음주량이 많음을 의미하지는 않는다. 한 달 동안 섭취하는 음주량이 적더라도, 한 번에 마시는 술의 양은 많을 수 있다는 점을 주의해서 해석해야 한다.



<그림1> 술 마시는 빈도(좌), 한 번에 마시는 술의 양(우)과 음주량 사이 관계

2.2.2.13 스트레스

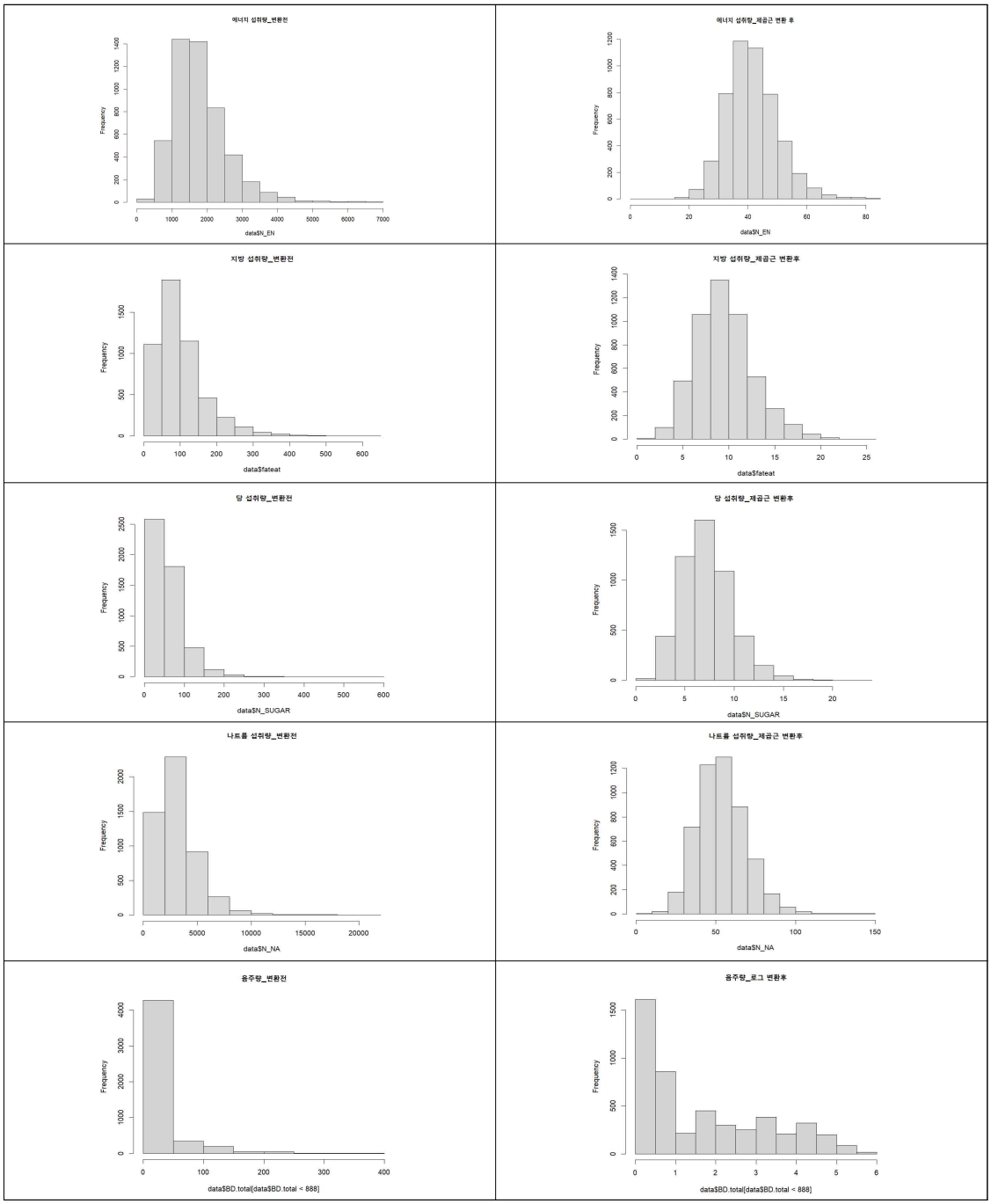
숫자가 클수록 스트레스가 작은 관계로, 역코딩을 진행했다. 점수가 높을수록 개인의 스트레스 인지 정도가 높음을 의미한다.

2.2.2.14 외식횟수

숫자가 클수록 외식횟수가 작은 관계로, 역코딩을 진행했다. 점수가 높을수록 외식횟수가 많음을 의미한다.

2.2.3 변수 변환

정규분포 형태를 따르도록 왼쪽으로 분포가 편향된 '에너지 섭취량', '지방 섭취량', '당 섭취량', '나트륨 섭취량'은 sqrt 변환하였다. '음주량'은 분포에 0의 값이 존재하여, 기존 데이터에 1을 더한 후 log 변환을 진행하였다. 에너지 섭취량, 지방 섭취량, 당 섭취량, 나트륨 섭취량, 음주량의 변환 전후 분포를 <그림2>에서 확인할 수 있다.



<그림2> 연속형 변수의 변환 전후 비교

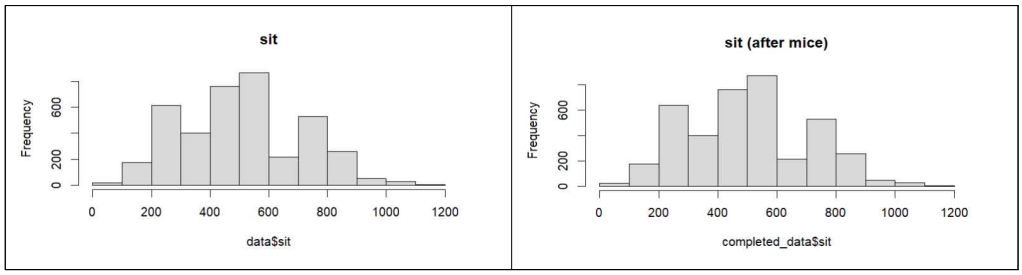
2.2.4 결측치 제거

결측치 데이터, 반응변수의 소아·청소년 범주에 해당하는 비해당 데이터, 범주형 변수의 무응답을 결측치로 판단하여 2,300개의 데이터를 제거하였다.

연속형 변수의 무응답은 제거 및 다중 대체법(MICE)으로 결측치를 처리하였다. '근로시간', '체중변화', '수면시간'의 무응답은 빈도가 적어 해당 행을 제거하였다. '착석 시간'은 무응답 빈도가 높아, 변수 간의 관계를 고려하여, 예측된 값의 분포를 기반으로 결측치를 대체하는 PMM기법을 사용하여 대체하였다. 결측치 보간 전후 분포가 크게 다르지 않음을 <표2>와 <그림3>에서 확인할 수 있다.

<표2> 착석 시간 결측치 보간 전후 기술통계량 비교

착석 시간	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
mice 전	30	360	480	517.9	660	1200
mice 후	30	360	500	519.1	660	1200



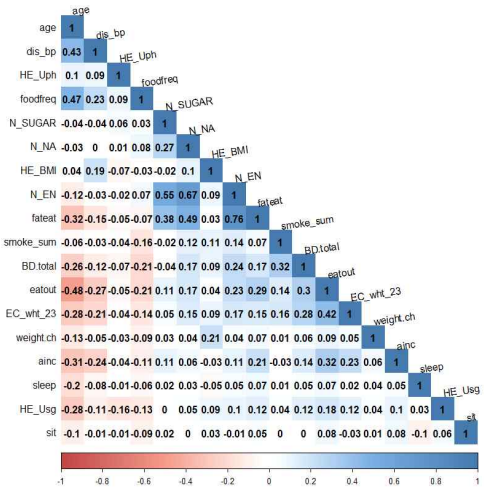
<그림3> 착석 시간 결측치 보간 전(좌) 후(우) 분포

2.3 연관성 분석

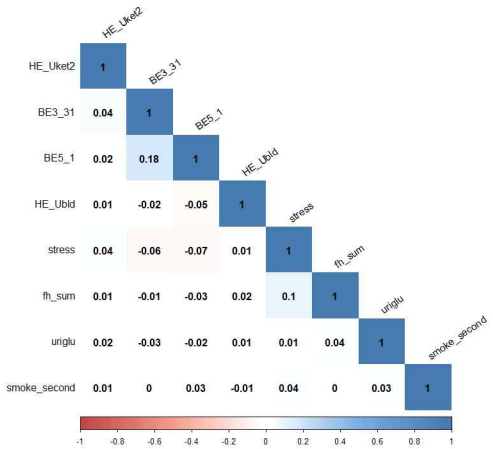
2.3.1 피어슨, 스피어만 상관분석

피어슨 상관분석으로 살펴본 연속형 변수 중 상관계수가 0.5 이상이면 다중공선성이 의심된다고 판단하였다. <그림4>에서 '에너지 섭취량'과 '지방 섭취량', '에너지 섭취량'과 '나트륨 섭취량', '에너지 섭취량'과 '당 섭취량', '나트륨 섭취량'과 '지방 섭취량', '나이'와 '외식횟수', '나이'와 '식사빈도', '나이'와 '맥압', '근로시간'과 '외식횟수'가 다중공선성이 의심되었고, 나머지 변수들은 다중공선성에 영향을 미치지 않을 것으로 보인다.

순서형 설명변수, 연속형 설명변수와 순서형 설명변수 간 스피어만 상관분석을 진행하여 상관계수가 0.4 이상인 변수를 다중공선성이 의심된다고 판단하였다. <그림5>와 <표3>에서 모든 순서형 변수들이 상관계수가 0.3 이하로 다중공선성에 영향을 미치지 않을 것으로 보인다.



<그림4> 연속형 피어슨 상관분석



<그림5> 순서형 변수 스피어만 상관분석

<표3> 연속형-순서형 스피어만 상관분석

스피어만	스트레스	간접흡연정도	걷기운동일수	근력운동일수	부모질병진단율	요당	요케톤	요잠혈
나이	-0.24	-0.05	-0.03	-0.05	-0.15	0.11	-0.14	0.04
요산도	0	0.02	0.02	-0.01	-0.04	-0.01	-0.04	-0.08
맥압	-0.14	-0.01	0.02	0.01	-0.1	0.08	-0.05	-0.01
식사 빈도	0.17	-0.01	0.04	0.06	-0.08	0.04	-0.1	0.01
체질량지수	0	0.01	-0.05	-0.03	0.05	0.11	-0.05	-0.03
당섭취량	-0.03	0.02	0.05	0.06	0.03	-0.02	-0.04	-0.03
나트륨섭취량	-0.03	0.05	-0.01	0.06	0.02	0.05	-0.05	-0.04
에너지섭취량	-0.04	0.07	0.01	0.12	0	0.03	-0.05	-0.06
지방섭취량	0.01	0.05	0.06	0.13	0.06	-0.03	0.01	-0.04
흡연량	0.08	0.05	-0.09	-0.02	0.02	0.09	0.01	-0.06
음주량	0.1	0.05	-0.01	0.09	0.06	0.01	0.06	-0.09
외식횟수	0.16	0.06	0.01	0.05	0.1	-0.05	0.07	-0.06
근로시간	0.15	0.09	-0.09	0	0.06	0	0.04	-0.07
소득	0.09	-0.02	0.02	0.08	0.17	-0.05	0.04	-0.05
체중변화	0.04	0	-0.01	-0.06	0.07	-0.1	-0.03	0.02
수면시간	-0.06	0	-0.04	0.01	-0.02	-0.02	0.01	0
요지방	0.05	0.02	0	0.05	0.04	0.19	0.12	0.04
착석 시간	0.06	0	-0.03	-0.01	0.01	0.02	0.03	0.01

2.3.2 카이제곱 검정

범주형 설명변수 간 카이제곱 검정 결과, 유의수준 1%에서 '결혼'과 '보험가입여부', '결혼'과 '건강검진여부', '결혼'과 '성별·임신여부', '보험가입여부'와 '수도권', '보험가입여부'와 '성별·임신여부', '성별·임신여부'와 '건강검진여부', '성별·임신여부'와 '수도권'은 다중공선성이 의심되었다. '건강검진여부'와 '수도권'은 다중공선성에 영향을 미치지 않을 것으로 보인다. 카이제곱 검정 결과는 <표4>에서 확인할 수 있다.

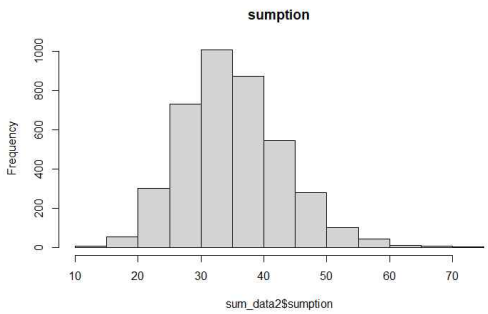
<표4> 범주형 변수 간의 카이제곱 검정

p-value	수도권	성별·임신여부	결혼상태	보험가입여부	건강검진 여부
수도권					
성별·임신여부	4.72×10^{-8}				
결혼상태	1.38×10^{-15}	$<2.23\times10^{-16}$			
보험가입여부	5.58×10^{-15}	0.0001736	$<2.23\times10^{-16}$		
건강검진 여부	1	1.88×10^{-15}	$<2.23\times10^{-16}$	1.90×10^{-6}	

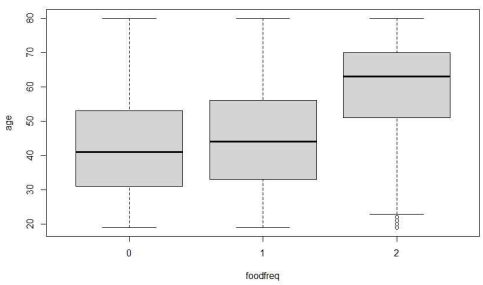
2.3.3 요인분석

차원 축소로 다중공선성 문제를 해결하기 위해 요인분석을 진행하였다. '에너지 섭취량', '지방 섭취량', '나트륨 섭취량', '나이', '식사빈도' 변수를 기반으로 두 가지 요인을 설정하여 분석을 진행하였다. 선택된 요인의 수의 충분성을 검정하는 카이제곱 검정 결과, (카이제곱 통계량=1.97, p=0.161 > 0.05)로 유의수준 5%에서 귀무가설을 기각하지 못해 2개의 요인이 충분했으며, 변수 2개로 분리되었을 때 누적 분산 비율이 64.2%로 설명력이 높았다. 따라서 (에너지 섭취량, 지방 섭취량, 나트륨 섭취량) 변수를 합쳐 영양소 섭취량 변수를 생성하고, (나이, 식사빈도) 변수를 합쳐 식습관별 연령대 변수를 생성하였다.

'영양소 섭취량' 변수는 '에너지', '지방', '나트륨 섭취량' 변수의 평균으로 형성하였다. 영양소 섭취량의 분포는 <그림6>과 같고, 점수가 높을수록 영양소 섭취량이 많음을 의미한다. '식습관별 연령대' 변수는 표준화한 '나이'와 '식사빈도'를 더해 형성하였다. <그림7>에서 '나이'와 '식사빈도'는 나이가 많을수록 식사하는 빈도도 함께 증가하는 관계라는 것을 알 수 있다.

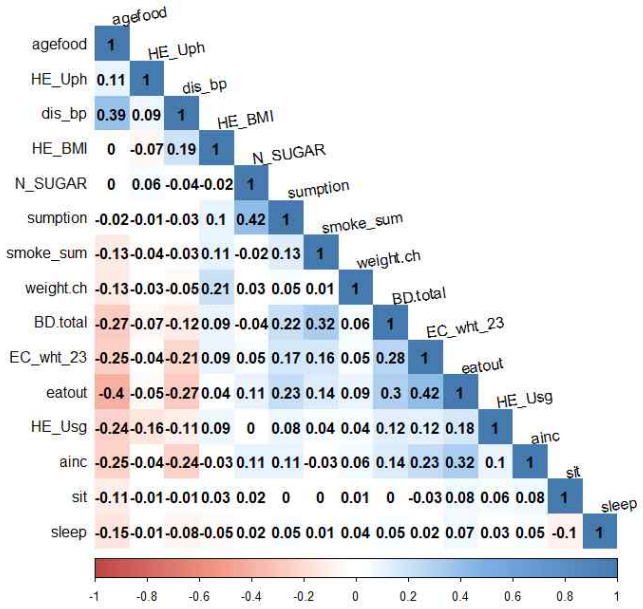


<그림6> 영양소 섭취량의 히스토그램



<그림7> 식사빈도에 따른 나이의 상자그림

요인분석 전 다중공선성이 의심되었던 연속형 변수는 ‘에너지 섭취량’, ‘지방 섭취량’, ‘나트륨 섭취량’, ‘당 섭취량’, ‘나이’, ‘외식횟수’, ‘식사빈도’, ‘맥압’, ‘근로시간’, ‘외식횟수’ 9개였다. 요인분석 후 피어슨 상관분석 결과인 <그림8>에서 ‘식습관별 연령대’와 ‘외식횟수’만 다중공선성이 의심되므로, 차원 축소로 연속형 변수의 다중공선성 문제가 감소했음을 알 수 있다.



<그림8> 연속-연속 피어슨 상관분석(요인분석 후)

2.4 최종 변수

총 3,958개의 데이터와 범주형 변수 5개, 순서형 변수 9개, 연속형 변수 15개를 사용하여 분석을 진행하였다. 이에 대한 자세한 변수 설명은 <표5>에 나타내었다.

<표5> 최종 변수 설명

자료 유형	변수명	설명		
반응변수	dg_sum	질병 진단빈도	0~4	
범주형	city	시도	0: 비수도권	1:수도권
	sex3	성별·임신여부	0. 남자 1. 여자&임신O	2.여자&임신X
	marri_2	결혼상태	1.유배우자, 동거 3.사별	2. 별거 및 이혼 4. 미혼
	npins	보험가입여부	1.가입	2.미가입
	BH1	건강검진 여부	1.검진	2.미검진
순서형	stress	스트레스	1. 거의 안 느낌 3. 많이 느낌	2. 조금 느낌 4. 대단히 많이 느낌
	smoke_second	간접흡연 노출 정도	0. 노출 안 됨 1. 조금 노출	2. 많이 노출
	BE3_31	걷기 운동 일수	1. 운동 안 함 2. 1일 3. 2일 4. 3일 8.7일(매일)	5. 4일 6. 5일 7. 6일
	BE5_1	근력 운동 일수	1. 운동 안 함 2. 1일 3. 2일 4. 3일	2.1일 3. 2일 5. 4일 6. 5일 이상
	fh_sum	부모 질병 진단율	0. 질병 없음 3. 질병 3개 4. 질병 4개	1. 질병 1개 2. 질병 2개 5. 질병 5개
	uriglu	요당	0. 음성 2. 양성 +,++,+++	1. 미량+- 3. 양성 ++++
	HE_Uket2	요케톤	0. 음성 1. 양성 + 2. 양성 ++	3. 양성 +++
	HE_Ubld	요잠혈	0. 음성 2. 양성 + 3. 양성 ++	1. 미량+- 4. 양성 +++
연속형	eatout	외식횟수	1~7	
	ainc	소득	17~1500	
	EC_wht_23	주당 평균 근로시간	0~119	
	weightch	체중변화	-3~3	
	BD.total	1년간 음주량	0~5.953	
	sleep	수면시간	14~98	
	smoke_sum	흡연량	0~60	
	sit	착석 시간	30~1200	
	dis_bp	맥압	20~117.50	
	HE_BMI	체질량지수	13.54~46.72	
	HE_Uph	요산도	5~9	
	HE_Usg	요비중	1.001~1.050	
	N_SUGAR	당 섭취량	1.711~23.943	
	agefood	연령대별 식습관	9.5~41.00	
	sumption	영양소 섭취량	10.57~73.18	

2.5 모형 적합

모델의 성능을 판단하기 위해, 학습용 데이터와 평가용 데이터 세트를 7:3의 비율로 분할하였다. 반응변수가 질병 진단빈도임을 고려하여, 모형은 포아송 회귀 모형을 선택하였다. 반응변수의 평균과 분산을 고려한 결과의 차이가 크지 않으므로 과대 산포가 아니라고 판단하여, 기본 포아송 회귀모형을 선정하였다.

이후 요인분석을 진행하여, 새로운 변수인 '식습관별 연령대 변수'와 '영양소 섭취량'을 포함한 데이터로 포아송 회귀 모형을 적합하고, 단계적 선택법을 이용하여 적합한 축소 모형과 그 결과를 비교하였다. 비교 모형으로는 요인분석 전 데이터를 기반으로 한 기존 포아송 모형과 Lasso를 이용한 축소 모형을 선정하였으며, 가장 좋은 결과의 회귀모형을 최종 모형으로 선택하였다. 모형 평가에 이용한 지표는 평균제곱오차(MSE), AIC, 잔차이탈도이다. 모형 평가 방법에 대한 세부적인 설명은 <표6>에 나타내었다.

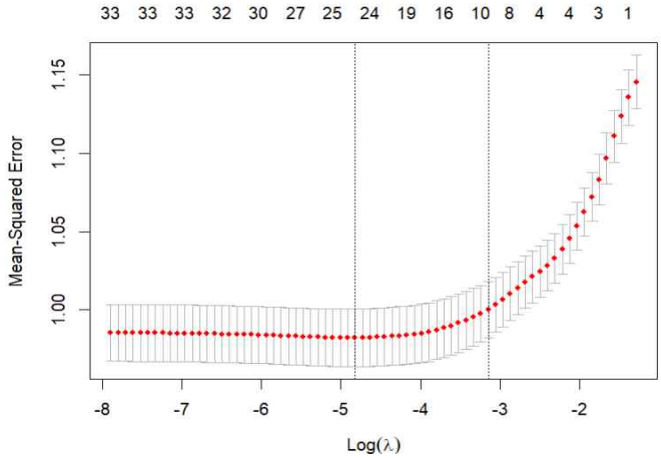
<표6> 모형 평가에 이용된 성능 평가 지표

MSE	예측값과 실제값의 차이를 제곱한 값의 평균
AIC	모델의 적합성을 의미하는 값으로 AIC가 낮을수록 적합도가 높음
잔차이탈도	모형의 예측값과 실측값이 근접하게 일치하는지 나타내는 통계량으로, 잔차이탈도가 작을수록 모형이 관측치를 잘 예측함

2.5.1 기존 포아송 회귀 모형

포아송 회귀모형은 포아송 분포를 따르는 반응변수를 예측하기 위한 모형이다. 포아송 분포란 시행 횟수는 많으나 특정 사건의 발생 확률이 아주 작은 확률 분포로서 주로 시간적, 공간적으로 발생 빈도가 낮은 사건의 발생 수를 설명한다. 질병 진단 사건의 횟수를 모델링하는 데 적합한 분포라고 판단하여, 본 연구에 포아송 회귀모형을 적합하였다.

요인분석으로 새로운 변수를 생성하기 전 '나이', '식사빈도', '에너지 섭취량', '나트륨 섭취량', '지방 섭취량'을 포함한 총 31개의 변수로 전체 포아송 회귀 모형을 적합하였고, Lasso 기법을 이용하여 축소 모형을 적합하였다. Lasso 회귀분석은 벌점회귀모형 중 하나로, 기존의 회귀모형의 손실함수인 잔차제곱합을 최소화하는 추정치를 계수들의 크기 합을 제한하는 영역 내에서 구하는 방법이다. Lasso 회귀에서의 람다는 규제의 강도를 조절하는 매개변수를 의미한다. 람다(λ)에 따라 계수 추정값이 달라지며, 계수가 정확히 0인 변수는 반응변수에 영향을 주지 않는 변수라고 판단한다(이다경, 2020).



<그림9> 로그(λ)에 따른 MSE

<그림9>에서 MSE가 가장 작을 때의 람다는 0.008047023이고, 회귀계수의 개수가 24개일 때 최적의 모델임을 알 수 있다. 최적의 람다로 변수선택을 진행한 결과, 회귀계수의 값이 0이 아닌 변수는 '임신경험이 있는 여성', '체중변화', '한 달 음주량', '사별', '당섭취량', '간접흡연 노출 정도', '나트륨 섭취량', '착석 시간', '수면시간', '맥압', '식사빈도', '비수도권', '걷기 운동 일수', '나이', '요산도', '체질량지수', '스트레스', '별거 및 이혼', '건강검진 여부', '부모 질병 진단율', '미혼', '요당', '임신경험이 없는 여성'이다. 해당 변수를 고려하여 축소모형을 만든 후 전체모형과 그 결과를 비교하였다. Lasso 회귀분석 결과로 얻어진 설명변수들의 비표준화계수의 값을 <표7>에 나타내었다.

<표7> LASSO 회귀분석 결과 비표준화계수

반응변수	항목	비표준화계수	항목	비표준화계수	항목	비표준화계수
질병 진단 빈도	임신경험있는여성	-0.05978	착석 시간	0.00026	체질량지수	0.03913
	체중변화	-0.04261	수면시간	0.003407	스트레스	0.06239
	1년간 음주량	-0.03254	맥압	0.007837	별거 및 이혼	0.08385
	사별	-0.02416	식사 빈도	0.008871	건강검진 여부	0.1074
	당섭취량	-0.0123	비수도권	0.01544	부모 질병 진단율	0.109
	간접흡연 노출정도	-0.00133	걷기 운동 일수	0.01653	미혼	0.1503
	나트륨 섭취량	-0.00104	나이	0.01778	요당	0.2059
	소득	1.37×10^{-5}	요산도	0.0387	임신경험없는여성	0.2412

<표8> 전체모형과 축소모형의 성능 평가 지표

	AIC	MSE	설명변수 개수	해석할 설명변수 개수
기본포아송	7124.7	0.9777828	31	34
lasso	7110.8	0.9704639	21	24
ANOVA	**잔차이탈도 차이=6.1223, df=10, p=0.8049**			

<표8>을 통해 전체모형과 축소모형을 비교한 결과, 모형 간 MSE, AIC 값의 차이가 크지 않았으며, 모형 간 차이를 비교한 ANOVA 결과도 유의수준 1%에서 전체모형과 축소모형이 서로 다르지 않았다. 따라서, 모형 간 차이가 크지 않다고 판단하여 최종 모형으로 설명변수 개수가 적은 LASSO를 적합한 축소 모형을 선택하였다.

2.5.2 요인분석 적용 포아송 회귀모형

요인분석 후 새로 형성한 변수 '영양소 섭취량'과 '식습관별 연령대' 변수를 포함한 총 29개의 변수로 포아송 회귀모형을 적합하였다. 해당 모형을 전체모형으로 설정하고 단계적 선택법을 이용하여 변수를 선택한 축소 모형과 그 결과를 비교하였다. 단계적 선택법은 단계별로 추가 또는 제거되는 변수의 여부를 검토하여 변수의 변화가 없을 때 중단된다.

요인분석이 이뤄진 전체모형에서 시작하여 단계적 선택법에 의해, '성별 및 임신경험', '식습관별 연령대', '요당', '체질량지수', '맥압', '부모 질병 진단율', '착석 시간', '음주량', '근로시간', '근력 운동 일수', '체중변화', '스트레스', '영양소섭취량', '걷기 운동일수', '요산도' 변수 15개가 선택되어, 이를 고려하여 축소모형을 적합하였다.

<표9> 전체모형과 축소모형의 성능 평가 지표

	AIC	MSE	설명변수 개수	해석할 설명변수 개수
요인분석	7161	0.9723545	28	31
요인분석+단계	7140.41	0.9752307	15	16
ANOVA	**잔차이탈도 차이=9.4261, df=15, p=0.8542**			

<표9>를 통해 전체모형과 축소모형의 MSE, AIC 값을 비교한 결과, 모형 간 차이가 크지 않았고 모형 간 차이를 비교하는 ANOVA 분석 결과도 유의수준 1%에서 전체모형과 축소모형이 서로 다르지 않았다. 따라서, 모형 간 차이가 크지 않다고 판단하여 최종 모형으로 설명변수 개수가 적은 단계적 선택법을 적용한 축소 모형을 선택하였다.

2.5.3 최종 모형 선정 과정

<표10>을 통해 Lasso를 적용한 축소모형과 단계적 선택법을 적용한 요인분석 축소모형을 비교하였다. ANOVA 분석 결과 유의수준 1%에서 두 모형이 서로 달랐지만, AIC와 MSE 값의 차이가 크지 않았다. 따라서, 설명변수 개수가 적으면서도 AIC와 MSE에서 Lasso 적용 축소모형과 비슷한 성능을 보이는 요인분석 축소모형이 더 좋은 모형이라고 판단하여, 요인분석 데이터를 기반으로 단계적 선택법을 적용한 포아송 회귀 모델을 최종 모형으로 결정하였다.

<표10> 최종 모형의 성능 평가 지표

	AIC	MSE	설명변수 개수	해석할 설명변수 개수
lasso	7110.8	0.9704639	21	24
요인분석+단계	7140.41	0.9752307	15	16
ANOVA	잔차이탈도 차이=45.58, 자유도=8, p=2.856×10^{-7}			

2.6 최종 모델

<표11> 최종 모형에 대한 포아송 회귀분석 결과

반응변수	항목	비표준화계수	표준편차	검정통계량	p-값	유의성
질병 진단 빈도	상수	-1.8290	0.2443	-7.487	7.04×10^{-14}	
	식습관별 연령대	0.1130	0.0147	7.694	1.42×10^{-14}	O
	요당	0.2131	0.0278	7.677	1.63×10^{-14}	O
	체질량지수	0.0390	0.0051	7.653	1.96×10^{-14}	O
	맥압	0.0102	0.0017	5.847	5×10^{-9}	O
	부모 질병 진단율	0.0938	0.0209	4.487	7.23×10^{-6}	O
	착석 시간	0.0002	0.0001	2.605	0.0092	O
	음주량	-0.0348	0.0135	-2.571	0.0101	X
	근로시간	-0.0024	0.0010	-2.447	0.0144	X
	근력 운동 일수	0.1047	0.0452	2.318	0.0204	X
	임신 경험x 여자	0.1712	0.0759	2.256	0.0241	X
	체중변화	-0.0465	0.0212	-2.197	0.028	X
	스트레스	0.0554	0.0268	2.071	0.0383	X
	영양소섭취량	-0.0054	0.0026	-2.069	0.0385	X
	임신 경험o 여자	-0.0866	0.0468	-1.852	0.064	X
	걷기 운동일수	0.0122	0.0074	1.645	0.0999	X
	요산도	0.0399	0.0243	1.64	0.101	X
MSE= 0.9752307, AIC=7140.41, Residual deviance = 3196.4 on df 2754						

최종 모형의 포아송 회귀분석 결과, 모형의 예측력은 96%, AIC는 7102.3으로 나타났다. Z 통계량 값이 크고, 추정된 회귀계수들의 p-value가 작은 변수가 반응변수에 큰 영향을 미친다고 판단하였다. '식습관별 연령대', '요당', '체질량지수', '맥압', '부모 질병 진단율', '착석 시간' 순으로 반응변수인 '질병 진단빈도'에 많은 영향을 미쳤다. '음주량', '근로시간', '근력 운동 일수', '임신 경험이 없는 여성', '체중변화', '스트레스', '영양소섭취량', '임신 경험이 있는 여성', '걷기 운동일수', '요산도'는 p값이 유의수준 0.01 이상으로, 반응변수에 영향을 미치지 못했다.

즉, 식사를 많이 하는 높은 연령일수록, 요당 수치가 높을수록, 체질량지수가 높을수록, 맥압이 클수록, 부모 질병 진단율이 높을수록, 오래 앉아 있을수록 질병에 대한 위험률이 높았다. ‘음주량’, ‘근로시간’, ‘근력 운동 일수’, ‘임신 경험이 없는 여성’, ‘체중변화’, ‘스트레스’, ‘영양소섭취량’, ‘임신 경험이 있는 여성’, ‘걷기 운동일수’, ‘요산도’는 질병에 대한 위험률에 영향을 미치지 못하였다.

3. 결론 및 기대효과

본 연구는 질병관리청의 2022년 국민건강영양조사를 활용하여 질병 진단빈도에 어떤 변수들이 영향을 미치는지에 대한 분석한 것이다. 원시데이터의 623개의 변수 중, 분석과 관계있는 변수 79개를 선택하였다. 이후 변수 전처리 과정을 통해 관련 있는 설명변수들을 통합하여 변수를 생성하였다. 이 과정에서 9개의 각 질병에 대한 질병 진단 변수를 합해, 반응변수인 ‘질병 진단 빈도’를 형성하였다. 데이터 탐색 과정을 통해 결측치, 무응답, 소아·청소년으로 인한 비해당 데이터를 제거하고, ‘착석 시간’에 대하여 MICE 기법을 적용하여 결측치를 대체하였다. 이후 상관분석과 카이제곱 검정을 통해 다중공선성이 의심되는 변수를 확인하고, 요인분석으로 에너지, 지방 나트륨 섭취량 변수를 합해 영양소 섭취량, 나이와 식사빈도를 합해 식습관별 연령대 변수를 생성하여, 변수 29개를 포함한 총 3,958개의 데이터로 분석을 진행하였다.

분석의 정밀도를 높여주기 위해서 여러 모형을 적합한 후 평가지표를 이용하여 가장 적절하다고 판단되는 모형을 최종 모형으로 선정하였다. 요인분석을 진행하기 전 전체 변수를 통하여 만들어진 포아송 회귀모형과 요인분석을 통하여 만들어진 변수를 포함하는 포아송 회귀모형을 전체 모형으로 설정하였다. 추가적인 비교를 위해 각각 LASSO 회귀분석과 단계적 변수선택법을 이용하여 축소 모형을 만들어, 전체모형과 축소 모형을 비교하였다. 모형 간 평가지표 간의 차이가 크지 않았고 ANOVA 분석 결과 모형 간 차이가 없어 설명변수 개수가 가장 적은 모형인 요인분석 데이터를 기반으로 단계적 선택법을 거친 포아송 회귀 모델을 최종 모형으로 결정하였다.

최종 모형 결과, ‘식습관별 연령대’, ‘요당’, ‘체질량지수’, ‘맥압’, ‘부모 질병 진단율’, ‘착석 시간’ 순으로 반응변수에 영향력을 미쳤다. 식사를 많이 하는 높은 연령이 질병에 대한 위험도가 가장 높았다. 식사량이 많은 건 60~80세의 특징이었다. 나이가 많을수록 대사 속도가 감소하므로, 음식물을 소화하고 처리하는 데 시간이 더 오래 걸려 식후 혈당 반응이 증가한다. 즉, 노화의 진행으로 포도당 대사장애 및 당뇨병 발생 위험이 증가하여, 질병과의 위험도가 높았다(홍은경, 2019). 요당은 소변에 포도당이 검출되는 양으로 당뇨병에 크게 영향을 미쳐 다양한 합병증을 수반하므로 질병 발생 빈도가 높았으며(유다영 외, 2022), 체질량지수(BMI)가 높을수록 비만을 의미하므로 뇌졸중, 당뇨, 신장 질환, 천식 등의 질병이 증가하여 다른 질병 발생 위험도가 높은 변수로 선택되었다(조일영 외, 2023). 맥압 수치가 높아 동맥혈관의 탄력성이 저하되는 경우에 질병 발생 위험도가 상승했으며, 부모로부터 유전되는 질병이 자녀의 질병 발생에도 영향을 미쳤고, 장기적으로 한곳에 오래 앉아 있는 경우에 질병 발생 위험도가 높았다.

본 연구의 시사점은 다음과 같다. 본 연구를 통해 개인의 질병 위험도를 판단할 때 중요하게 작용하는 요인에 대해 도출할 수 있었다. 나이가 많을수록 노화로 인한 질병 발생 가능성이 높으니, 식사의 빈도와 질을 고려하여 식습관을 개선하는 것이 건강에 좋으며, 오랜 기간

같은 자리에 앉아 있는 것은 질병 발생 확률을 높이므로 적당한 활동을 통해 신체를 자주 움직이는 것이 중요하다. 비만인 경우, 부모님의 질병 빈도가 많은 경우 건강검진을 자주 받는 것을 권한다. 또한, 건강검진 결과 소변에서 요당이 많이 검출되거나, 맥압이 높아 동맥혈관의 탄력성이 떨어진 상태는 다른 질병과도 연관이 있기에 더욱 주의해야 한다. 따라서, 모델의 알고리즘은 개인의 생활 양식, 유전 등의 요인을 바탕으로 개인의 건강 위험도를 예측함으로써, 건강 상태를 파악하고 질병 발생 확률이 높은 대상을 분류할 수 있다. 이를 통해, 질병을 조기 진단받고 예방할 수 있도록 도우며, 개인이 건강한 생활 습관을 지니도록 제안하는 데 이바지할 수 있다.

참고문헌

이채린, "2050년 전세계 기대수명 2022년 대비 약 5년 늘어", 동아사이언스, 2024.05.19 (https://m.dongascience.com/news.php?idx=65453) (접속일 : 2024.06.20.).
임성원, “운동·정신건강 관리 시대…새 소비트렌드는?”,디지털타임스, 2024.06.13., (https://www.dt.co.kr/contents.html?article_no=2024061302109963084008)(접속일 : 2024.06.20.).
윤은숙, “성인병 막는 과일 채소, 하루 적정 섭취량은?”, 건강을 위한 정직한 지식. 코메디닷컴, 2022.10.19(https://kormedi.com/1534912/)(접속일:2024.06.13.).
신종학, “우리나라 사망원인 1위 ‘비감염성질환’”, 치과신문 , 2019.01.04., (https://www.dentalnews.or.kr/news/article.html?no=23288) (접속일: 2024-06-20)
KT Word. (접근일: 2024년 6월 21일). "포아송 회귀모형". http://www.ktword.co.kr/test/view/view.php?m_temp1=870.
이다경, 김연후, 박현정.(2020).대학 졸업생의 직업만족도에 영향을 미치는 변수 탐색: 벌점회귀모형 sparse group lasso를 활용하여.아시아교육연구,21(4),1069-1097.
정은영. (2022). 한국 성인의 수면시간과 맥압과의 관련성: 국민건강영양조사 제7기 자료를 중심으로. 디지털융복합연구, 20(4), 411-418, https://doi.org/10.14400/JDC.2022.20.4.411
조유미, 이경숙. (2023). 한국 성인의 식사 빈도에 따른 심혈관대사질환 위험도와 식사 질의 매개효과 : 국민건강영양조사 제7기 자료. 중환자간호학회지, 16(2), 67-80.
결론:
조일영, 김동운, 박준상, 엄지수, 윤유용, 강태원. (2023-06-01). BMI와 각 질병의 상관관계 분석. Proceedings of KIIT Conference, 제주.
유다영, 김나리, 최가영, 이태하, 이규도, 황한정. (2022-12-01). 전이학습을 이용한 비색센서 이미지 기반 요당 식별 알고리즘 개발. Proceedings of KIIT Conference, 제주.
김선미, 이동률, 고유라, 박민선. (2011). 음식 섭취 횟수와 대사증후군의 관련성. Korean Journal of Health Promotion, 11(1), 9-17.
홍은경. "노화에 따른 포도당 대사의 변화." 대한당뇨병학회지 20권, no. 4 (2019): 215-219.