G000F8I-2021-144

데이터 예측을 위한 통계적 방법 비교 및 활용

건강보험심사평가원
HEALTH INSURANCE REVIEW & ASSESSMENT SERVICE

G000F8I-2021-144

데이터 예측을 위한 통계적 방법 비교 및 활용

연구진

연구책임자 신현철 부연구위원

주 제 어 미래예측, 예측방법, 시계열, 머신러닝, 신경망, HIRA, Health Insurance Review & Assessment, Machine learning, Forecasting

건강보험심사평가원
HEALTH INSURANCE REVIEW & ASSESSMENT SERVICE

머리말

보건 데이터 예측은 정책효과 평가, 건강보험 재정추계, 급여비 이상징후 탐지 등 보건의료 분야 영역에서 다양하게 활용되고 있다.

이와 더불어, 예측모형을 통해 산출된 근거는 보건정책 개발 및 의사결정과정에 중요한 역할을 하고 있다. 따라서 예측의 과정은 과학적이고 체계적이어야 하며, 정확성 측도는 예측방법의 중요한 선택의 기준이 되어야 한다.

이번 연구는 보건의료 영역에서 적용 중인 데이터 예측방법 및 최신 예측기법에 대한 조사와 통계적 예측방법별 특성을 비교하고, 보건 자료 특성에 적합한 예측방법을 제시하고자 하였다. 이를 위해서, 문헌검토 및 최신 예측방법론 조사, 자료 특성별 예측성능 비교 분석을 수행하고, 최종적으로 종합하여, 자료 특성에 적합한 예측방법을 제안하였다. 예측방법 검토 대상에 회귀모형, 시계열 모형, 머신러닝(신경망)을 포함하였고, 건강보험 청구자료를 이용하여 예측성능을 비교 분석하였다. 본 연구 결과는 예측작업을 수행하는 과정에서, 정교한 예측치 산출에 필요한 예측방법을 선택하는 데 도움이 될 것으로 기대한다.

끝으로, 본 보고서의 내용은 연구자의 개인적인 의견으로, 건강보험심사평가원의 공식적인 견해가 아님을 밝혀 둔다.

2021년 12월

건강보험심사평가원 원장 김 선 민

건강보험심사평가원 심사평가연구소장 이 진 용

목 차

표 목 차

그 림 목 차

요 약

제1장 서론

1. 연구 배경

○ 보건의료 예측모형은 정책효과 평가, 건강보험 재정 추계, 이상징후 탐지 등 보건의료 영역에서 다빈도로 활용되고 있음

○ 예측모형을 통해 산출된 근거를 바탕으로 보건정책 개발과 의사결정이 이루어지고 있음

- 예측의 과정은 과학적이고 체계적이어야 하며, 높은 정확성이 요구 됨

○ 보건의료 환경과 자료 특성을 고려한 통계방법론 선택을 위한 검토가 필요함

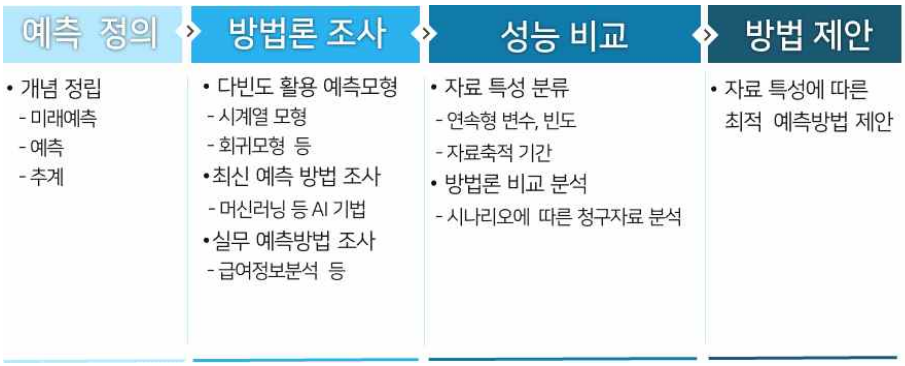
2. 연구 목적

○ 본 연구는 보건의료 영역에서 적용 중인 데이터 예측방법 및 최신 예측기법에 대한 조사와 통계적 예측방법론의 특성을 비교하여, 보건자료 특성에 적합한 통계방법을 제시하는 것을 주된 목적으로 하며, 세부적인 목적은 다음과 같음

- 첫째, 데이터 예측의 개념을 정립하고, 유형 분류를 시도하며,
- 둘째, 예측을 위한 통계방법론을 조사하고 예측 모형 간 성능을 비교 분석하여
- 셋째, 자료 특성에 적합한 통계방법을 제시하고자 함.

3. 연구 내용 및 방법

○ 본 연구는 크게 데이터 예측의 개념 정의, 유형별 예측방법론 조사, 예측성능 비교 분석, 자료 특성에 따른 적합한 예측방법 제안 등으로 구성됨



[요약 그림 1] 주요 연구 내용

○ 데이터 예측의 개념을 정립하고, 유형별 개념에 관련 연구를 검토함. 또한, 미래예측, 일반 예측, 추계 예측 등 개별 예측의 개념을 비교 검토함.

○ 예측을 위한 통계적 예측방법론을 조사 검토함. 심사평가 업무 등 보건의료에 적용되는 방법론을 조사하였고, 연구 분야에서 다빈도로 활용되고 있는 예측방법론을 검토함

○ 마지막으로, 데이터 예측을 위한 방법론 간 성능을 비교 분석함. 자료특성에 따른 시나리오를 구성하고, 데이터 분석을 수행하여 예측방법의 성능을 비교 분석 함. 분석 결과에 따라 자료 특성 및 유형별 적합한 예측방법을 제시함.

제2장 예측방법 개요

1. 예측 관련 업무 적용 현황

가. 급여정보분석 업무

○ 급여정보분석시스템을 이용하여, 사전적 지출관리 및 합리적 의료이용 실현을 위해 의료이용 상시 모니터링 업무를 지원하고 있음. 해당 업무는 급여항목 이상감지, 주제별 변동 분석, 기관별 변동 분석 등이 있고, 이 중에서 급여항목 이상감지 기능은 예측기법을 적용하여 운영함. 급여항목 전체를 대상으로 각 분류 단위별 예측진료비와 진료비 증감률을 산출하고, 이상치 초과 여부를 판단

함. 급여항목 전체는 행위(5단코드), 치료재료(중분류코드), 약제(일반명코드) 등 2만 3천여 개 항목으로 구성되어 있음

나. 정책효과 연구

○ 보건의료 분야 영역을 보면, 요양급여비용 자율점검제도 도입 효과 측정, 의료급여 혁신 재정절감 효과 측정, 외래약제 적정성 평가 가감지급 도입 효과 등의 연구 사례와 같이 통계적 예측방법론에 기반하여 예측값 산출을 수행한 정책효과 연구가 다수 있음

2. 예측개념

○ 본 연구에서는 아래의 3가지 예측유형 중에서, 미래 예측 개념을 중심으로 적합한 예측방법론을 조사·검토하고, 이들 방법론 간 예측성능을 비교 분석함

가. 일반 예측

○ 경험과 지식을 기반으로 구체적이고 명시적인 값을 진술하는 것임. 일반적으로 주어진 범위 안에서 확률적으로 값을 제시함

- 예측요인으로 사용되는 어떤 한 변수 또는 변수집단에 근거하여 다른 목표변수의 값을 추론하는 통계적 방법으로, 예측방정식을 구하여 예측변수의 값을 대입함으로써 예측되는 목표변수의 값을 구함

나. 미래 예측

○ 현재의 과학적 지식에 근거하여 실현 개연성이 가장 높은 진술을 연구자가 제시하는 것임. 과거와 현재 데이터의 추이를 분석하여 미래시점의 값을 제시함

○ 미래예측의 방법에는 시계열분석(Times series analysis), 회귀분석, 머신러닝(신경망) 등이 있음

다. 추계 예측

○ 일련의 가정들이 충족 될 때 실현 되는 값에 대한 조건적 진술임. 즉, 미래의 변동추세에 관한 일련의 가정에 근거하여 제시하는 조건적 전망을 의미함

3. 예측방법 동향

가. 자료특성,유형에 따른 예측방법

○ 통계적 예측 분석 방법에서 필요한 자료 개수는 정해진 기준이 없음. 경험적으로, 시계열 예측에서 필요한 자료(수)는 계절성 확인을 위해 12개가 필요하고, 또는 추세 경향 및 반복 여부 확인을 위해 24개의 자료가 필요하다고 알려짐. 일부 보고서에서 제시하는 필요자료의 수는 30개~50개 수준 임

○ 단변량 예측기법의 경우, 과거 시계열 추이 및 계절성 특성이 예측방법에 적용 가능한 지수평활법, ARMA, ARIMA 모형이 있음

○ 외부영향요인을 고려한 예측모형에는 개입효과모형(전이함수 모형), ARIMAX 등이 있음

○ 빅데이터 활용이 활발해지고 인공지능 기술이 발전하면서 머신러닝(신경망)을 이용한 예측방법이 다수 활용됨. 머신러닝(신경망)을 이용한 예측기법은 연속형 자료 또는 카운트형 자료 대상으로 적용가능하고, 회귀유형 예측 문제는 지도학습 형태로 접근할 수 있음. 활용 가능한 자료개수에 따라 과소 적합 또는 과대 적합 문제가 발생할 수 있음

○ 다수의 예측방법 중에서 최적의 방법 1개를 선택하기보다는, 개별 예측방법들을 수행하고, 이를 종합하여 중앙값 또는 평균값을 제시하는 방법을 사용하기도 함

나. 다빈도 예측방법

○ 다수의 연구에서 이용되는 예측기법은 시계열분석이며, 그 다음으로 머신러닝 등 인공지능 기법, 회귀모형 등이 활용되고 있음

- 머신러닝 기법의 예측방법은 최근의 빅데이터 이용 활성화 추세에 맞추어 점점 사용 횟수가 늘고 있음
- 분석 대상 자료유형은 연속형 변수인 경우가 많았고, 카운트형 변수인 경우는 그 횟수가 적었음

제3장 예측방법

1. 회귀 모형(다항 추세모형)

○ 시계열 자료를 분석하기 위해 많이 사용되는 방법은 관측치 값을 시간의 함수로 표현하는 방법임. 회귀모형(다항 추세모형)은 독립변수를 시간의 함수로 표현하고, 종속변수와 독립변수들 사이의 상호관련성을 함수 형태로 나타냄.

○ 회귀모형에서 모수 추정은 최소제곱법을 이용하고, 모형의 적합도 및 기본 전제조건에 대한 검토는 잔차 분석을 실시함

2. 시계열 모형

가. 개요

○ 과거 행태가 미래에도 그대로 지속된다는 대전제에 기초하여 시계열을 따라 제시된 과거 자료로부터 추세나 경향(일정한 패턴:규칙성/시계열변동)을 파악하여 미래의 관찰값을 예측하는 방법임

나. ARMA

○ 시계열 과정에서 종속변수의 과거 값들과 오차항의 과거 값들로써 현재의 시계열 값을 설명하는 방식으로, 자기회귀과정과 이동평균과정을 동시에 포함하는 확률과정임. 자기회귀이동평균과정이라고도 함

- 자기회귀과정은 현재 관측값들을 과거 관측값들의 함수 형태로 표현할 수 있다고 가정함
- 이동평균과정은 현재의 시계열 설명을 위해 오차항의 과거 값들을 모형에 포함하는 방식임

다. ARIMA

○ 보건의료 분야에서 시계열 과정은 추세를 갖고 증가하거나, 분산의 정도가 변화하는 특성을 보이는 데, 이것은 비정상시계열이고 분석이론 적용이 어려움

○ 시계열분석은 정상시계열 특성을 가정하고 있으므로, 비정상성을 보이는 경우 로그변환 또는 차분을 통해 시계열을 정상화한 이후 분석방법을 적용함

○ 즉, ARIMA 모형 적합과정은 변환 또는 차분을 통해 정상화 과정을 수행한 후 적절한 ARMA(자기회귀이동평균과정) 모형을 적합 하는 것임

라. ARIMAX

○ ARIMA 모형에 독립변수까지 고려하여 외생변수를 포함한 자기회귀누적이동평균 모형을 말함

○ ARIMAX 모형에 독립변수들의 차수를 결정하는 것은 어려운 과정일 수 있으며, 지체차수를 정하기 위해 종속 시계열과 독립시계열 간의 교차상관계수를 구하고, 이를 통해 후보 차수들을 정하게 됨

마. 지수평활법

○ 시계열에 변화가 발생하는 경우, 과거의 모든 자료에 동일한 비중을 부여하는 대신에 최근의 변화시점에 가까운 자료에 큰 비중을 두는 예측방법임

바. 자기회귀오차모형

○ 일반 회귀모형과 다르게 오차 항이 서로 독립되지 않고, 자기상관성이 존재하는 모형임

3. 머신러닝(신경망) 모형

가. 개요

○ 머신러닝(신경망) 모형은 주어진 자료를 컴퓨터로 학습하여 명시적으로 정의되지 않은 패턴을 결과로 만들어 내는 방식임

○ 사전에 예측문제 환경에 대해 주어진 지식(이론, 구조) 없이 주어진 알고리즘을 통해 데이터에 담긴 정보를 추출해 내는 학습방법임

나. 방법론

○ 머신러닝 모형 종류에는 신경망 외에, 기저벡터머신, 확률밀도 분포 추정법이 있고, 패턴인식 문제 해결 또는 특정점 학습을 위해 많은 수의 신경층을 갖는 딥러닝 기술 등이 있음

○ 미래 예측의 방법으로 머신러닝(신경망)을 검토하였고, 본 연구에서는 SciKit-learn을 이용하여 예측성능을 비교 분석함

○ SciKit-learn은 파이썬 머신러닝 라이브러리 중에서 가장 많이 사용되는 툴임. 실제 사용 시에 다양한 알고리즘 개발을 위해 편리한 API를 제공해 주고, 실전 환경에서 검증된 우수한 라이브러리 임

4. 일반화선형 모형(카운트형 자료)

○ 카운트형 자료 대상 예측모델 분석에서 분포가정에 대한 조건이 성립하지 않는 경우, 분석자료에 로그변환을 취하거나 분포가정 조건에 감마분포, 로그노말 분포 등을 고려하여 모델선택, 모델적합, 해석 등 분석을 수행할 수 있는 방법임

○ 일반화선형모형은 종속변수에 영향을 주는 한 개 이상의 독립변수 효과를 추정하는 기존의 선형모형을 일반화한 모형으로 연속형 변수와 카운트형 변수에 대한 모든 분석이 가능함

- 카운트형 변수 분석에 적합한 일반화선형모형으로 기본적인 모델에 포아송이 있으며, 이외에 음이항 분포 등 여러 가지 분포가 있음

제4장 건강보험 청구자료를 활용한 사례분석

1. 분석 방법

○ 사례분석은 초음파항목 진료비 및 건강보험 총진료비 청구자료 대상으로 함

○ 자료유형을 구분하기 위해 연속형 변수에 진료비를, 카운트형 변수에 실시횟수로 분리하여 분석함

○ 비교 대상 예측방법은 회귀모형(다항 추세모형), 시계열모형, 머신러닝(신경망)이었고, 예측성능 측도로 평균절대백분위오차(MAPE)와 평균 제곱근 예측오차(RMSE)를 사용함

- 각 예측모형별 3개월 또는 3개년의 예측값을 생성한 후 실제 관찰값과 비교하여 예측 오차를 산출함

○ 통계분석 툴은 SAS EG와 파이썬, 쥬피터 노트북, SciKit-learn library를 이용함

2. 분석 대상

○ 자료유형별, 축적기간별 분석 대상의 특징은 아래표와 같음. 초음파 청구자료는 진료월 기준 2018년 4월부터 2021년 3월(심사결정기준: 2018년 4월~2021년 5월)까지의 자료를 이용하였고, 건강보험 총진료비는 1990년부터 2019년까지의 자료를 이용하였음

〈요약 표 1〉 분석 대상별 자료 특성

	분석 대상	
	초음파 청구자료	건강보험 전체 청구자료
모형구축 자료	- 12개월('18.04~ '19.03) - 24개월('18.04~ '20.03) - 30개월('18.04~ '20.09)	-27개년 (1990년~2016년)
연속형/카운트형 (수집단위)	-연 속 형: 초음파진료비(월별) -카운트형: 초음파시행횟수(월별)	-심사결정진료비(연도별)
독립변수		-소비자물가지수(연도별) -1인당 GDP(국민총생산, 연도별) -시간*시간(연도)
예측기간	-단기예측: 3개월	-단기예측: 3개년

○ **연속형 예측값 산출에는 회귀모형(다항 추세모형), 시계열(자기회귀오차, 지수평활, ARIMA, ARIMAX), 머신러닝(신경망)을 이용하였고, 카운트형 예측값 산출에는 일반화선형모형, 머신러닝(신경망)기법을 이용함**

〈요약 표 2〉 자료유형별 예측방법

구 분		초음파 청구자료	건강보험 전체 청구자료
연속형	회귀모형	-다항 추세모형	
	시계열모형	-자기회귀오차 -지수평활 -ARIMA	-ARIMAX
	머신러닝 (신경망)	-머신러닝(신경망)	-머신러닝(신경망)
카운트형	일반화선형모형 (GLM)	-GLM 모형 (음이항 분포)	
	머신러닝 (신경망)	-머신러닝(신경망)	

3. 초음파 진료비 사례분석

가. 개요

○ 건강보험 청구자료를 살펴보면 보건의료 정책 영향에 따라 청구 경향이 크게 변함. 이러한 자료 특성이 반영된 초음파(보장성 강화 항목) 진료비 청구자료를 대상으로 통계적 예측방법 성능을 비교 분석함

나. 분석 결과

○ 연속형 자료(초음파 진료비) 분석 결과 머신러닝(신경망) 기법의 예측성능이 가장 우수하였고, 자기회귀오차 및 ARIMA 방법도 우수하였음

○ 카운트형 자료(초음파 시행횟수) 분석 결과, 머신러닝(신경망) 기법의 예측성능이 가장 우수하였고, 일반화 선형모형 방법도 우수하였음

4. 건강보험 총진료비 사례분석

가. 개요

○ 건강보험 지출 예측은 건강보험에 영향을 주는 요인(독립변수)을 고려하고, 시계열적인 특성을 감안해야 함. 즉 외부요인의 동태적인 특성을 반영하여 예측모형을 설정해야 함

나. 분석 결과

○ 외부요인을 고려하여 예측방법의 성능을 비교 분석한 결과, ARIMAX 방법의 예측성능이 머신러닝(신경망) 기법보다 우수하였음

○ 총진료비 변수의 로그변환 이후 ARIMAX모형의 예측성능이 향상된 것을 확인하였고, 이를 통해 비용변수 예측모형 적합 시 변수변환에 대한 필요성이 확인됨

제5장 고찰 및 결론

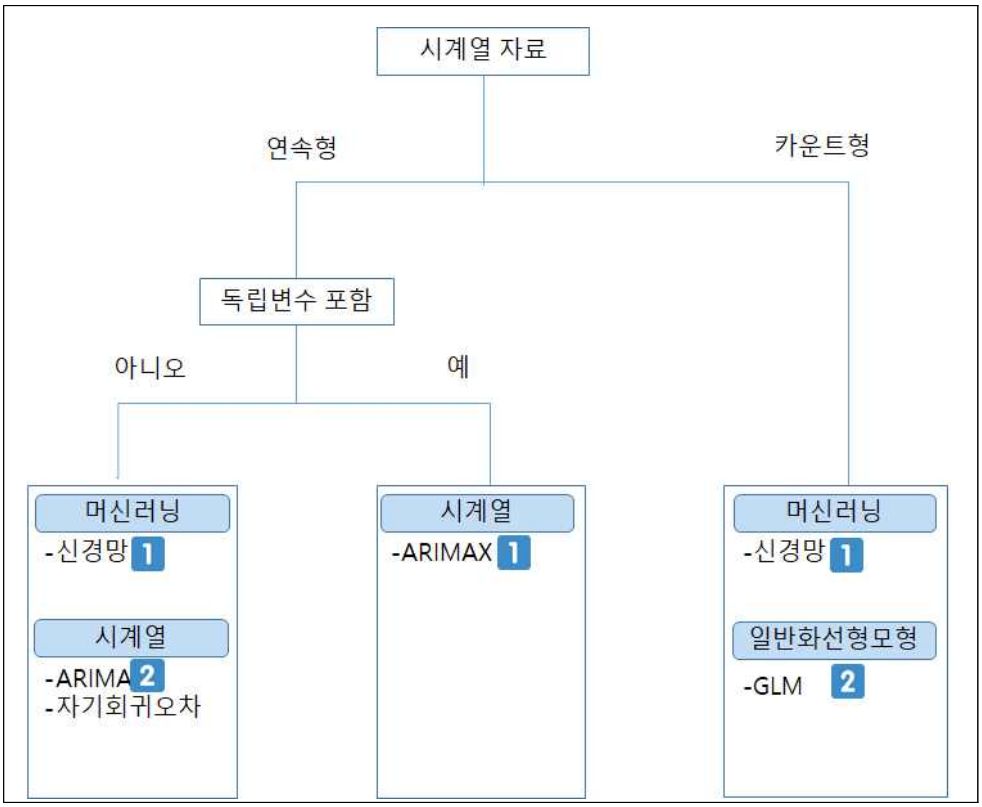
1. 고찰

○ 이번 연구는 보건의료 분야에서 활용되고 있는 예측방법을 조사하고, 자료유형 및 외부변수 유무에 따른 적절한 예측방법을 제시하였음

○ 다빈도로 활용되는 예측방법은 시계열 모형이었고, 최근에는 머신러닝 기법 등 인공지능 방법을 활용한 예측도 실시됨. 본 연구에서는 회귀모형, 시계열 모형, 머신러닝(시계열) 기법을 통계적 예측방법을 후보목록으로 선정하고, 이들 간의 성능을 비교 분석함.

○ 단변량 예측에서 머신러닝(신경망)기법의 성능이 가장 우수함을 확인하였고, 자기회귀모형 및 ARIMA 모형도 성능이 우수함을 확인함. 예측과정에서 30개 이상의 자료 개수에서 최적의 성능을 나타냈으며, 머신러닝 기법에서 자료개수에 비례하여 예측성능이 향상됨을 확인함

○ 결론적으로, 본 연구는 자료유형별 통계적 예측방법 선택과 관련한 제안을 아래와 같이 제시하였음



[요약 그림 2] 자료유형별 통계적 예측방법 제안

2. 결론

○ 보건의료 정책의 효과평가 및 성과·측정 관리 분야에서 이용되는 예측값은 종종 보건의료 정책의 근거자료로 활용되기 때문에 정교하고 논리적인 예측방법 사용이 중요함

○ 본 연구에서는 자료유형 및 자료 개수 등 에 따른 예측방법을 비교·분석하고, 상황에 맞는 적합한 예측방법을 제시함

○ 합리적인 건강보험 정책 설계·수행을 비롯하여, 안정적인 급여비 관리를 위해 미래 시점의 정교한 예측치 산출과 최적의 통계적 예측방법 적용이 필요함. 이를 위해서는

- 첫째, 본 연구에서 제안한 자료유형별 적합한 예측방법을 선택하여 올바르게

적용하는 것이 중요함. 개별 상황에 적합한 예측방법을 이용하여 정책효과 평가 및 정책목표 설정 과정에서 정교한 예측치를 제시할 수 있어야 함

- 둘째, 다양한 자료유형에 따른 예측방법을 제안하기 위해 단순한 사례 형태가 아닌 자료 개수의 연속적 변화에 따라 각 예측방법 성능의 변화수준을 보여주는 그래프로 제시할 필요가 있음
- 셋째, 예측방법의 활용 범위를 확대하고 최신 기법을 습득해야 함. 건강보험 정책의 효과평가를 위한 근거를 산출할 뿐 만 아니라, 상시 모니터링 업무에 예측기능을 탑재하여 이상치를 감지하고 미래 예측작업을 강화해야 함. 또한, 예측성능이 우수하고, 기술이 빠르게 발전하는 인공지능(머신러닝 등) 기법을 업무에 적극 활용해야 함

제1장 서 론

1. 연구 배경

보건의료 예측모형은 정책효과 평가, 건강보험 재정 추계, 이상징후 탐지 등 보건의료 영역에서 다빈도로 활용되는 통계적 접근 방법이다. 특히, 예측모형을 통해 산출된 근거를 바탕으로 보건정책 개발과 의사결정이 이루어지기 때문에, 예측의 과정은 과학적이어야 하고, 예측모형은 높은 정확성이 요구된다.

보건의료 예측을 위한 다양한 통계방법론이 개발·활용되고 있으나, 보건의료 환경과 자료특성을 고려하여 통계방법을 선택하여야 한다. 따라서 통계 방법론을 비교함으로써 자료 특성을 잘 반영하는 예측방법론을 제시할 필요가 있다. 보건 분야에서는 보장성 강화 등 보건 정책 개입이 수시로 발생하고, 비교적 단기간 자료를 이용한 예측모형을 적용한다는 점에서 자료 특성을 반영한 통계방법론에 대한 검토가 필요하다.

2. 연구 목적

본 연구는 보건의료 영역에서 적용 중인 데이터 예측방법 및 최신 예측기법에 대한 조사와 예측 통계방법론의 특성을 비교하여 보건자료 특성에 적합한 통계방법을 제시하는 것을 주된 목적으로 하며, 세부 목적은 다음과 같다.

첫째, 데이터 예측의 개념을 정립하고 유형을 분류하며,

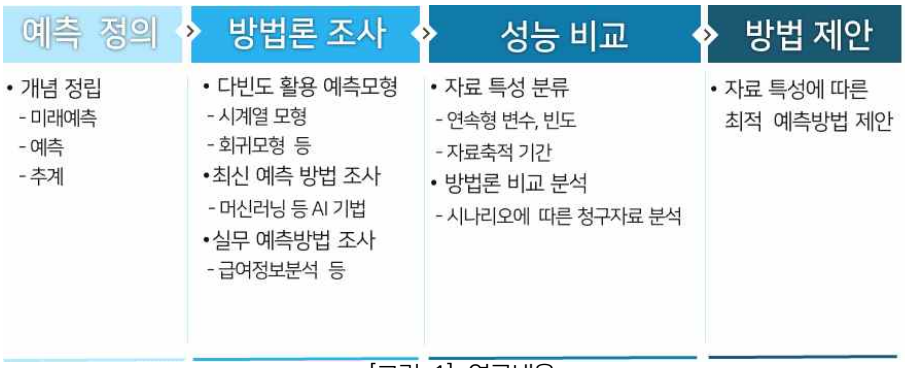
둘째, 예측을 위한 통계방법론을 조사하고 예측 모형 간 성능을 비교 분석하여,

셋째, 자료 특성에 적합한 통계방법을 제시하고자 한다.

3. 연구 내용 및 방법

본 연구는 크게 세 부분으로 구성된다.

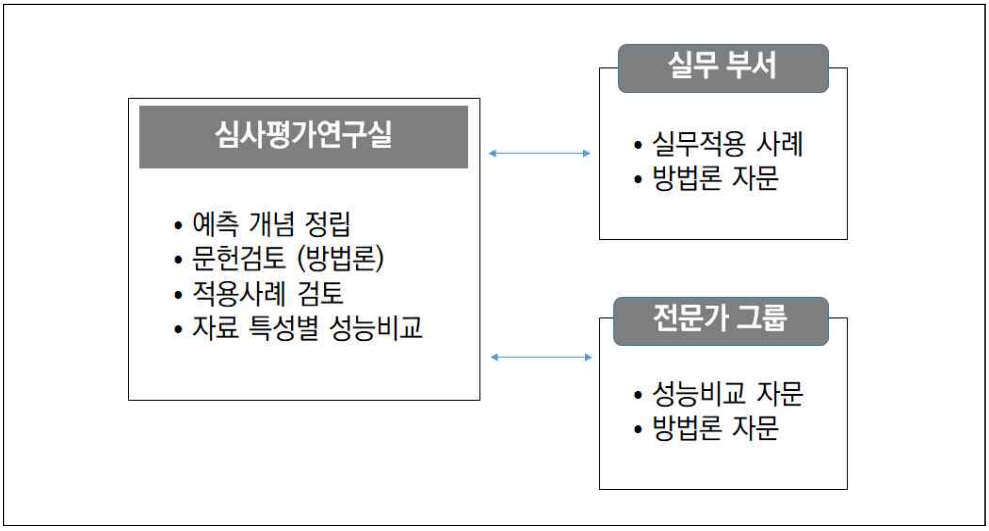
먼저, 데이터 예측의 개념을 정립하고, 유형별 분류에 대한 관련 연구를 검토하였다. 또한, 미래 예측, 일반 예측, 인구추계 등 개별 영역별 예측의 개념을 비교 검토하였다.



[그림 1] 연구내용

두 번째, 데이터 예측을 위한 통계방법론을 조사 및 검토하였다. 심사평가 업무 등 보건의료에 적용되는 방법론을 조사하고, 연구 분야에서 다빈도로 활용되고 있는 예측방법론을 검토하였다.

마지막으로, 데이터 예측을 위한 방법론 간 성능을 비교 분석하였다. 자료특성에 따른 시나리오를 구성하고, 데이터 분석을 통해 최적의 예측방법을 제시하였다. 자료특성은 예측변수의 특성, 외부요인 존재 여부 등에 따라 유형을 구분하였으며, 건강보험 청구자료를 이용하여 사례분석을 수행하였다.


심사평가연구실
• 예측 개념 정립
• 문헌검토 (방법론)
• 적용사례 검토
• 자료 특성별 성능비교
실무 부서
• 실무적용 사례
• 방법론 자문
전문가 그룹
• 성능비교 자문
• 방법론 자문

[그림 2] 연구수행 체계

제2장 예측방법 개요

1. 예측 관련 업무 적용 현황

가. 급여정보분석 업무

급여정보분석 시스템은 보건의료 정책지원, 급여기준 개선, 합리적 의료이용 등 업무에 필요한 분석·예측 정보를 즉시 이용할 수 있도록 의료이용 모니터링 및 예측, 이상감지 등의 분석서비스를 제공하고 있다. 본 시스템은 분석기법을 다양화하고 인공지능 기법을 도입함으로써 예측정확도를 향상시켰고, 코드별 단위까지 예측하는 기능과 이상감지 기능을 구현하여 촘촘한 모니터링을 가능케 하였다. 본 시스템의 각 기능별로 적용된 예측기법은 다음과 같다.

〈표 1〉 예측모형 적용 현황

분석대상		자료 축적기간	분석기법
전체급여	행위(5단코드) 약제(일반명코드) 치료재료(8단코드) 상급,종합병원 병원급 의원급 약국	4개월이하	월별 접수데이터기준 진료데이터 비중
		5~12개월	지수평활법
		12~24개월	시계열(계절성 비반영)
		24개월	시계열(계절성 반영)
		분류유형별 총진료비 상위100	딥러닝(LSTM)
보장성 강화	고시항목	1~2개월	월별 접수데이터 기준 진료데이터 비중에 따른 비율 증가
		3~7개월 미만	유사항목 패턴(k-means) & 지수평활법
		7~12개월미만	지수평활법
		12~36개월미만	시계열
		36개월 이상	딥러닝(LSTM)
MRI/초음파	전체 MRI	전기간	다중회귀분석, 시계열분석
	전체 초음파		
	보장성 MRI	전기간	시계열
	보장성 초음파		
노인/만성/ 4대중증질환	노인	전기간	다중회귀분석/ 시계열분석
	만성(약국,한방제외)		
	4대중증(약국제외)		

본 시스템에 적용된 예측모형 기법은 시계열, 다중회귀분석, 지수평활법, K-means, 딥러닝(LSTM) 등이다. 시계열 기법은 시간적 변동에 따른 관측 데이터에 의해 그 변동 원인 파악 및 예측을 위한 분석 방법이고, 다중회귀분석은 독립변수와 결과가 되는 종속변수와의 회귀모델에 대한 분석방법으로 데이터의 변동에 영향을 주는 변수들 간의 인과관계를 통계적으로 추정하는 방법이다. 그리고, 지수평활법은 모든 시계열 데이터를 사용하여 평균값을 구하고 시간의 흐름에 따라 최근 시계열에 더 많은 가중치를 부여하는 예측방법이다. 마지막으로 K-means 방법은 비지도학습의 방법으로 레이블이 없는 데이터를 K개의 군집으로 묶어주는 알고리즘이고, LSTM은 장기·단기로 학습에 필요한 데이터를 더 오래 이용함으로써 신경망의 성능을 향상시킨 모델로 각 고시항목 등에 적용하여 예측값을 추정하는 방식이다.

급여정보분석시스템은 사전적 지출관리 및 합리적 의료이용 실현을 위해 의료이용 상시 모니터링 업무를 지원하고 있다. 해당 업무는 급여항목 이상감지, 주제별 변동 분석, 기관별 변동 분석 등 있고, 이 중에서 급여항목 이상감지 기능은 예측기법을 적용하여 운영하고 있다. 급여항목 전체를 대상으로 각 분류 단위별 예측진료비와 진료비 증감률을 산출하고, 이상치 초과 여부를 판단한다. 급여항목 전체는 행위(5단코드), 치료재료(중분류코드), 약제(일반명코드) 등 2만 3천여 개 항목으로 구성되어 있다.

나. 정책효과 연구

요양급여비용 자율점검제도 도입 효과 측정, 의료급여 혁신 재정절감 효과 측정, 외래약제 적정성 평가 가감지급 도입 효과 등과 같이 다수의 연구 사례에서 미래 예측값 기반의 정책효과를 산출하고 있다.

이성우 외 연구를 보면, 자율점검제도 도입 효과를 산출하기 위해서 제도시행 이전의 경향을 바탕으로 시행 이후의 예측 진료비를 도출하고, 이후에 실제 진료비와 비교하여 간접적 예방금액을 산출하였다. 이와 더불어, 다른 유사행위로의 대체 청구 가능성(풍선효과)을 고려한 간접적 예방금액의 산출을 시도하였다. 이 연구에서 사용한 예측방법은 시계열 분석으로, ARIMA 모형

적합을 통해서 모형의 식별, 모수추정, 모형의 타당성 검토 및 진단을 수행하였다. 최적의 시계열 분석법의 선정기준은 AIC통계량 기준을 활용하였다.

노상윤 연구를 보면, 의료급여 제도 시행 이후의 재정절감 효과를 산출하기 위해 시계열 분석법을 이용하였다. 의료급여대상자를 1종 기초수급대상, 1종 타법적용자, 2종 기초수급대상 등 3개 그룹으로 구분하여 분석을 수행하였다. 제도 시행 이전의 과거 시계열을 이용하여 모형을 설정하고, 이후 기간에서의 예측치를 계산하였다. 재정절감 효과는 실적치와 예측치의 차이를 비교하여 산출하였다. 또한 노상윤은 의료급여 재정 규모 예측 연구를 수행하였다. 해당 연구에서는 2002년 1월부터 2007년 12월까지 72개월의 자료를 이용하여 ARMA 모형을 적합하였고, 2012년까지 5개년의 중기예측 급여비를 산출하였다.

김지애는 외래약제 적정성 평가 가감지급 모형 개선 연구에서 약제처방률 절감효과를 분석하였다. 분석 방법으로는 구간회귀분석과, ARIMA 분석을 이용하였다. 분석에는 2014년 7월 기준으로 전후 24개월 자료를 이용하였다.

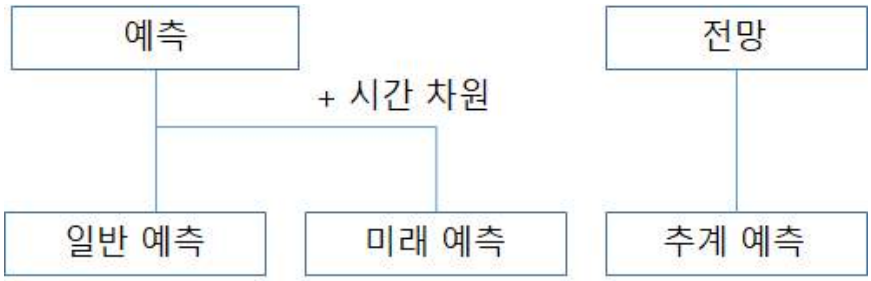
성병찬은 사회보장재정추계 기반 연구를 수행하면서 건강보험지출 추계작업을 수행하였다. 해당 연구에서는 변수의 과거 추세 및 독립변수의 영향이 동시에 모형이 가능한 자기회귀차분이동평균모형(ARIMAX)을 적용하였다. 연구자는 모형 적합을 위해 잔차에 대한 백색잡음 검토(포맨토우 검정), 잔차의 자기상관함수 및 부분자기상관함수 검토 및 단위근 검정을 실시하였다. 최종 모형 선택을 위하여 AIC 정보량, MSE, MAE의 측도를 기준으로 활용하였다.

정형선은 국민의료비 미래추계구축방안 연구에서 미래의료비 추계를 위해서 경상의료비 총계치의 단기 및 장기 추이 전망 작업을 수행하였다. 예측의 정확성에 비중을 두는 단계예측을 위해 향후 10년간 의료비 총계치를 예측하였고, 장기예측에는 인구구성의 변화를 반영한 조성법을 기본으로 50년 이상 추계치를 산출하였다. 단기예측에 이용된 분석방법으로 입력변수를 가진 ARIMA 모형을 검토하였고, 그 대상에는 ARIMAX 모형, 전이함수 모형, 개입모형 등 동적회귀모형을 포함하였다. 모형식별 및 추정, 진단하는 과정에는 표본자기상관함수, 표본부분자기상관함수, ACF, PACF, AIC 등이 활용되었다. 그리고, 결정요인에 의한 추계작업을 위해 OECD 보건데이터에 실려 있는 국가

단위 데이터가 패널구조를 가지는 점을 활용하여 확률효과모형에 기반한 국가 단위의 의료비 총계치의 결정모형을 구축함으로써 미래추계에 이용하였다.

2. 예측 개념

미래 예측(forecasting)은 과거와 현재의 데이터 추이를 분석하여 미래시점의 값을 예측하는 과정이며, 일반적으로 추세 분석을 통해 수행할 수 있다. 예를 들어, 특정 미래 날짜에 목표변수의 추정 값이 미래 예측값일 수 있다. 반면에, 예측(prediction)은 비슷해 보이지만 더 일반적인 용어로 사용된다. 즉, 확률적 분포에 기반한 예측에서 시간 차원을 추가하면, 미래 예측이고, 시간 차원을 고려하지 않으면, 일반 예측으로 분류할 수 있다. 또한, 일련의 주어진 조건들에 따라 조건적 미래 전망이 가능한데, 이러한 예측은 추계 예측이라 할 수 있다.



가. 일반 예측

1) 개요

경험과 지식을 기반으로 구체적이고 명시적인 값을 진술하는 것이다. 일반적으로 주어진 범위 안에서 확률적으로 값을 제시하게 된다.

즉, 예측인자로 사용되는 어떤 한 변수 또는 변수집단에 근거해서 다른 목표변수의 값을 추론하는 통계적 방법이다. 일반적으로 회귀분석법을 이용하여 방정식을 구하고, 식에 예측변수 값을 대입하여 예측되는 목표변수의 값을 구한다.

이때의 예측은 동일 상황에서 다시 자료를 수집하는 경우, 목표변수들이 어떤 특정한 값으로 나타날지에 대하여 예측변수를 예측방정식에 대입하여 짐작할 수 있다는 의미이다. 이때 목표변수의 예측치가 실제값에 얼마나 가까운가에 관심이 있다. 그러나 시간차원에서 미래의 상황을 예측한다는 의미는 아니다.

2) 방법론

일반 예측의 방법론에는 로지스틱 회귀분석, 의사결정나무, 회귀분석, 머신러닝(회귀유형, 분류유형) 등이 있다.

나. 미래 예측

1) 개념

현재의 과학적 지식에 근거하여 미래에 실현 개연성이 가장 높은 진술을 연구자가 제시하는 것이다. 과거와 현재 데이터의 추이를 분석하여 미래시점의 값을 제시하게 된다. 예를 들어 인구 분야에 미래 예측개념을 적용하면, 미래 특정 시점의 인구를 실제값으로 기대하면서 추정하는 작업을 수행할 수 있다.

2) 방법론

미래 예측의 방법론에는 시계열분석(Time series analysis), 회귀분석, 머신러닝(회귀유형), 딥러닝 기법 등이 있다.

다. 추계 예측

1) 개념

일련의 가정들이 충족 될 때 실현 되는 값에 대한 조건적 진술이다. 즉, 미래의 변동추세에 관한 일련의 가정에 근거하여 제시하는 조건적 전망을 의미한다. 방법론에는 조성법, 계량경제모형 등이 있다. 인구추계를 예를 들어 살펴보면, 인구추계는 인구 일반 예측과 구별된다. 인구 센서스가 진행된 기간 사이의 특정 시점 인구를 제시하기 위해 여러 가지 자료나 방법을 동원하여 계산하는 것은

인구 일반 예측에 해당한다. 하지만, 인구 센서스 이후의 미래시점의 인구 추정은 추계에 해당한다. 인구추계는 단순한 예측이 아니라 미래 장기간의 인구 연령구조와 규모 변화를 보여주기 위한 것이다.

2) 방법론

인구 추계 분야에서 사용된 사례를 기준으로 정리하고자 한다. 우선, 수학적 접근 방법론이 있다. 인구와 다른 요인의 관계가 고려되지 않는 단순 인구증가율에 기초하고 있으며, 증가율 계산 방식에 따라 로지스틱 증가율, 자연대수 증가율, 지수증가율 방법 등으로 구분할 수 있다.

경제적 접근 방법론이 있다. 경제 상황의 변화를 고려하여 인구를 추계하는 방법이다. 경제 상황에 따라 출생, 사망, 국제이동 등 인구변동요인이 변화하여 궁극적으로 인구가 변화하게 되는 연계성을 기초로 추계를 수행한다.

코호트 요인 접근 방법론이 있다. 동시 출생 집단에서 매년 발생하는 출생, 사망 및 국제이동을 계산하여 장래의 인구를 구하는 방법이다.

본 연구에서는 3가지 유형의 예측개념 중에서, 미래 예측 개념을 중심으로 적합한 예측방법론을 조사·검토하고, 이들 방법론 간 예측성능을 비교 분석하였다.

3. 예측방법 동향

가. 자료 특성, 유형에 따른 예측방법

보건의료 및 경제 등 다양한 분야에서, 데이터 예측을 위해 어떠한 통계적 방법이 활용되고 있는지 조사하였다. 이를 위하여 논문, 연구보고서 등을 검토하고, 외부전문가 자문을 통해 데이터 예측에 필요한 조건 및 방법에 대한 의견을 청취하였다. 이를 토대로 정리한 내용을 적어보면 아래와 같다.

자료 개수 관련하여, 통계적 예측 분석 방법에서 필요한 자료 개수는 정해진 기준이 없다. 예를 들어, 시계열 예측에서 필요한 자료(수)는 계절성 확인을 위해 12개가 필요하고, 또는 추세 경향 및 반복 여부 확인을 위해 24개의 자료가 필요하다고 언급된다. 일부 교과서에 제시하는 자료의 수는 30개~50개 이상이다. 다만, 자료의 개수가 그 이하라도 모형 적합 시도는 가능하지만 정교성(precision)의 문제가 발생한다.

연속형과 카운트형 자료 대상 간의 예측작업은 접근 방식이 달라야 한다. 다만 정교한 분석이 필요한 경우가 아니라면, 연속형 자료 분석방법을 카운트형에 준용해도 가능할 수 있다. 예를 들어 시계열 분석은 연속형 변수에 일반적으로 적용하고 있는데, 카운트형 변수 대상에서도, 자료의 개수가 큰 경우 적용이 가능하다고 판단된다. 다만, 정교한 분석이 필요한 경우, Poisson , Negative Binomial 등 일반화선형회귀분석 방법 등 을 고려해야 한다.

그리고, 외부변수가 없는 단변량 예측기법의 경우, 과거 시계열 추이 및 계절성 특성이 예측방법에 적용할 수 있는 지수평활법, ARMA, ARIMA 모형 등이 있다. 외부 정책개입 등 영향을 고려하기 위한 예측모형에는 개입효과모형(전이함수 모형)등을 이용할 수 있다. 만일 독립변수가 모형에 포함되어 예측을 시도하는 경우에는, 동시에 독립변수 예측값도 생성해야 한다. 외부 독립변수가 추가된 모형으로는 자기회귀오차모형, ARIMAX 모형 등이 있다. 다수의 독립변수를 고려해야 하는 경우, 예측모형에 포함하기 위한 독립변수 선정과 독립변수의 시차 결정을 위해 추가적인 상세분석이 필요하다.

최근에는, 다수의 예측방법 중에서 최적의 방법을 1개를 선택하기보다는, 개별

예측방법들을 수행하고, 이를 종합하여 중앙값 또는 평균값을 제시하는 방법을 사용하기도 한다.

예측방법의 예측정확도 비교기준에는 평균 절대 예측오차, 평균 제곱근 예측오차(RMSE), 평균 절대백분율 예측오차(MAPE)가 있고, 이중에서 평균 절대백분율 예측오차가 주로 활용된다. 평균절대백분율 예측오차 이용시 주의사항으로 관찰값이 0을 많이 포함하는 간헐적 데이터 인 경우, 분모값이 작아지면서 크게 변동하는 현상이 나타날 수 있다는 것이다.

예측모형 구축은 인과관계에 따른 이론적 접근과 상관관계에 근거한 축약형 형태로 구분할 수 있다. 예측방법 선택 시에는 직관적 해석과 예측력 높은 결과를 원하고 있으며, 이 기준으로 모델성능을 평가하고 있다. 하지만, 이 기준은 서로 상충할 수 있으며, 자료 개수와 예측기간 등에 따라 예측모델의 성능이 변할 수 있다. 이론적 모델은 직관적인 설명이 가능하지만, 실제 예측성능 측면에서는 영향력 높은 변수만을 포함한 간결한 예측모델이 더 우수하다고 알려져 있다.

빅데이터 활용 용이성과 인공지능 기술의 발전에 따라 머신러닝을 이용한 예측방법이 많이 활용되고 있다. 머신러닝(신경망)을 이용한 예측기법은 연속형 자료 또는 카운트형 자료 대상으로 적용가능하다. 예측유형의 문제는 지도학습 형태로 접근할 수 있다는 점과, 이용 가능한 자료개수에 따라 과소 적합 또는 과대 적합 문제가 발생할 수 있다는 점이다.

머신러닝은 분석 데이터를 훈련세트(training set)와 검증세트(validation set)로 구분하여 분석하며, 모형(신경망) 적합과정에서 추정해야 할 모수가 많아, 활용 가능한 자료 개수가 많을수록 유리한 예측방법이다. 머신러닝의 성능측도로 이용되는 기준은 연속형 자료의 경우 RMSE(평균제곱근예측오차), MAPE (평균절대백분위오차) 등이 있고, 범주형 자료의 경우 HIT rate, 정분류율, 특이도, ROC(Receiver operating characteristics, 반응자 작용특성) 커브 등이 있다. 카운트형 자료인 경우, 사전에 통계분포를 Poisson으로 가정하고, 모수추정 과정에 손실함수를 우도함수로 설정하여 수행하면 효율적인 추정이 가능하다. 연속형 자료인 회귀형 추정은 손실함수를 최소자승함수로 설정하여 진행하면 된다.

머신러닝(신경망)을 수행하기 위한 툴로서 SciKit-learn library를 이용한 회귀모형 예측, 분류 예측 모형이 있다. SciKit-learn 라이브러리는 일반적으로 파이썬 기반 환경에서 사용되며, 여러 환경에서 검증된 것으로 알려져 있다.

또한 심도 있는 머신러닝(딥러닝)을 수행하기 위해 필요한 library로는 Pytorch, Tensorflow, Keras 등이 있다. 특히, 딥러닝은 인공신경망의 계층적 수준을 이용하여 비선형 접근방식을 택하지만, 머신러닝(신경망)은 데이터분석을 선형적으로 처리하는 수준으로 알려져 있다. 머신러닝(딥러닝)은 데이터를 처리하는 과정에서 인간 두뇌 신경 경로를 모방하여 그것을 의사결정, 물체감지, 음성인식, 언어 번역에 사용한다. 이때, 딥러닝은 사람의 감독이나 개입 없이 구조화되지 않은 데이터와 라벨이 부착되지 않은 데이터로부터 정보를 습득하게 된다.

Library	장점	단점
Pytorch	-유연성/디버깅 능력 우수, 짧은 훈련기간 -데이터 병렬 영역에서 파이썬을 통한 네이티브 지원으로 비동기 실행이 가능하고 최적의 성능 구현	-디버깅 시 제한된 시각화 제공 -훈련모델을 이식할 때 프레임워크 제공 안함. 별도로 백엔드 서버 필요
Tensorflow	-문서화된 프레임워크 및 풍부한 훈련 모델, -훈련된 모델을 이식할 때, 텐서플로우 서버 프레임워크 지원 -대규모 데이터 세트 작업하는 경우, 객체 탐지작업 및 우수한 기능과 성능이 필요한 경우, 연구자들이 선호	-특정 디바이스에서 분산된 훈련을 허가하는 과정 등 모든 내역을 수작업으로 코딩해야 하므로 최적화 과정이 어려움
Keras	-모델을 신속하게 제작, 훈련 및 평가할 수 있는 플러그 앤 플레이 프레임워크 제공 -텐서플로우 위에서 운용되는 고도의 신경망 라이브러리 임 -다양한 이식 옵션과 용이한 모델 내보내기 가능 -재사용 가능한 코드와 튜토리얼 접근성 좋음 -소형 데이터 세트, 신속한 시제품 제작, 다중 백엔드 지원 등 우수	-속도가 느리고, 낮은 성능을 보임

나. 다빈도 예측방법

보건의료 분야 외에도 산업, 금융, 직업, 환경 등 다양한 분야에서 예측 모형을 이용하고 있다. 개별 목적에 따라서 미래의 예측값을 도출하여 목표값을 설정하거나, 예측값에 대한 영향 요인을 파악하는 등 그 목적이 다르다.

다수의 연구에서 이용되는 예측기법은 시계열분석이었고, 그 다음으로 머신러닝 등 인공지능 기법, 회귀모형 등이 활용되고 있다. 머신러닝 기법의 예측방법은 최근의 빅데이터 이용 활성화 추세에 맞추어 점점 이용 횟수가 늘고 있다. 그리고, 분석 대상 자료유형은 연속형 변수인 경우가 많았으며, 카운트형 변수인 경우는 드물었다.

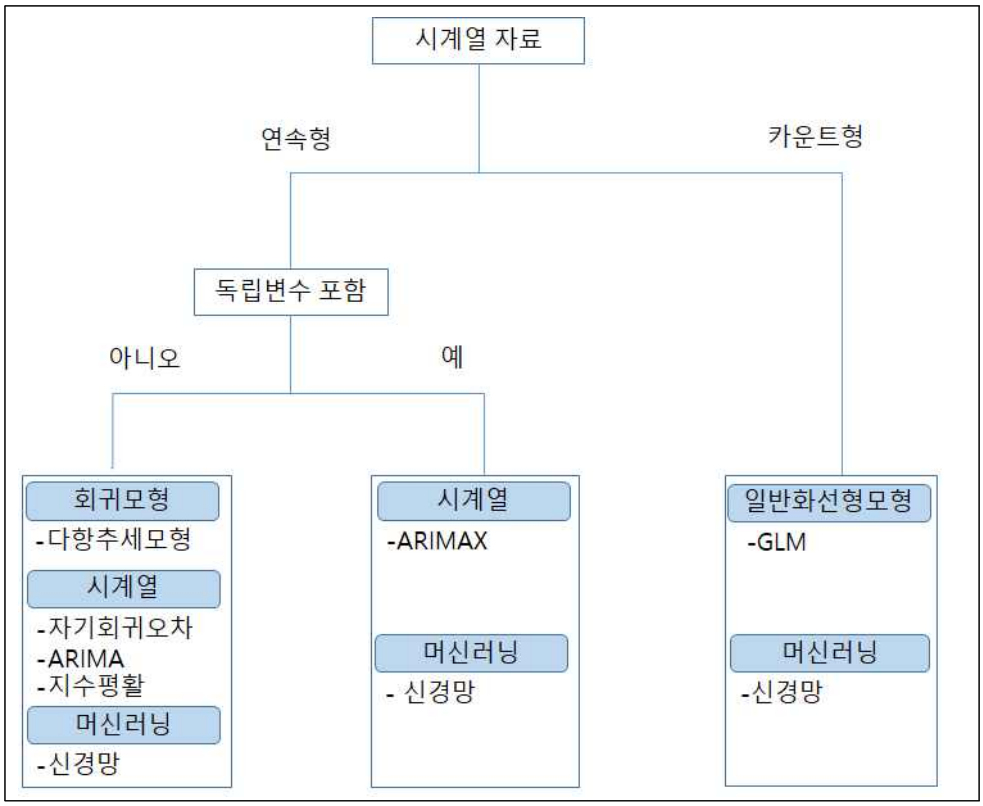
지금까지 조사된 결과를 바탕으로, 회귀모형, 시계열, 머신러닝을 이용한 예측방법을 비교한 결과를 제시하면 아래 표와 같다.

〈표 2〉 시계열 자료에서 예측방법별 특성 비교

	회귀모형	시계열 모형	머신러닝(신경망)
자료 개수	-30개 * 관찰값(Y)의 정규분포 충족조건	-30개 ~ 50개 *관찰값(Y)의 정규분포 충족조건	-훈련세트와 검증세트로 분리 가능한 자료개수
독립 변수	-관찰값과 독립변수의 관계식으로 모형구성 -관찰값을 예측하기 위해 독립변수 예측값 필요	-독립변수 없이 관찰값의 자체 시계열로 예측 가능 (ARIMA,지수평활법 등) -독립변수가 포함된, ARIMAX 모형의 경우 독립변수의 시계열추이 값이 필요	-관찰값과 독립변수의 값으로 구성된 자료가 입력 값으로 주어지고, 블랙박스 형태의 튜닝 수행 -관찰값 예측을 위한 독립변수 예측 필요
변수 유형	-연속형 -카운트형인 경우 일반화선형모형 (예:포아송) 적용가능	-연속형 -카운트형이라도 관찰개수 (N)가 충분하면, 연속형으로 분석 가능	-연속형 -카운트형
성능 측도	-MSE(평균제곱오차) -R^2 (결정계수) -AIC, BIC, SBC	-MSE(평균제곱오차) -MAPE(평균절대백분율 오차) -AIC, BIC, SBC	-MSE(평균제곱오차) -MAPE(평균절대백분율 오차) -정분류율 -ROC

세부 모형	-단순선형회귀 -다중선형회귀	-지수평활법 -자기회귀오차모형 -ARIMA -ARIMAX	-회귀문제유형 (선형회귀,가우시안프로세스) -분류문제유형 (로지스틱회귀, 신경망)
S/W	SAS, R	SAS, R	R, 파이썬 Library (keras, scikit-learn)
해석 가능	-독립변수와 관찰값의 관계를 명확히 규명	-독립변수와 관찰값의 관계를 명확히 규명	-관찰값이 생성되는 과정설명이 어려움
재현 가능	-모형을 수식으로 기술 가능하고, 재현 가능	-모형을 수식으로 기술 가능하고, 재현 가능	-분석대상 자료가 바뀌면 훈련세트도 변경되어, 모형식이 상이함

본 연구는 자료유형 및 독립변수 유무 등 개별 상황에 따른 적절한 예측방법을 제시하고자 한다. 지금까지 조사된 예측방법을 기초로 개별 상황에서 선택 가능한 방법을 분류하면 아래와 같다.



[그림 3] 자료유형별 통계적 예측방법 분류

제3장 예측방법

1. 회귀모형(다항 추세모형)

가. 개요

다항 추세모형은 회귀모형의 특별한 경우이므로, 회귀모형 중심으로 분석방법을 설명하고자 한다. 회귀모형은 예측하고자 하는 종속변수가 독립변수와 선형관계가 있다고 기본 가정을 한다. 즉, 추세모형은 회귀모형의 기본가정을 이용한 시계열 자료 분석 방법으로, 관측치를 시간(독립변수)의 함수로 표현 한다.

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p + \epsilon_t$$

추세모형의 모형 적합 단계는, 회귀모형의 분석절차를 따라 수행하면 된다. 즉, 회귀모형의 기본조건인 등분산성, 정규성, 독립성 조건을 점검하고, 이 조건들을 위반하는 경우, 이를 해결하기 위해 자료변환 작업, 예를 들어 로그변환, 제곱근변환 등 분산안정화 방법 등을 시도할 수 있다. 최종 모형이 선택되면, 관측치를 이용하여 모형의 모수를 추정하고, 추정된 회귀모형을 이용하여 예측작업을 수행한다. 모형 진단에는 잔차 분석이 이용되는 데, 모형의 적합도 및 기본 가정 조건 등을 점검한다.

나. 방법론

회귀모형 분석 방법은 두 개 이상의 변수들 $Y, X_1, X_2, \cdots, X_p$사이의 상호관련성을 아래 수식과 같이 표현한 모형에 의해 분석하는 방법을 일컫는다.

$$Y_t = f(X_{t\,i}; \beta_i) + \epsilon_t\ ,\ t = 1,2,3,\cdots,n;\ i = 1,\cdots,p$$

여기서, Y_t는 종속변수로서, X_{ti}들의 함수관계로서 예측되는 변수이고, X_t는 p개의 독립변수, 혹은 설명변수라 하고, β는 추정해야 할 모수이다. 회귀모형에서 아래첨자 t는 시간을 의미한다. 그리고 f의 형태는 선형함수가 많이 쓰이고, 기본 가정은 오차 ϵ_t가 정규분포(평균이 0, 분산이 σ_ϵ^2)를 따르고, 서로 독립이라는

것이다.

일반적으로 이용되는 회귀모형은 선형회귀 형태로 아래와 같이 표현된다.

선형회귀모형: $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + \epsilon_t$

여기서, 모수 β를 추정하기 위한 최소제곱법은 오차제곱합 ($\sum \epsilon_t^2$)을 최소로 만들어 주는 β 를 찾아가는 방식이다. 이때, $\epsilon_t = Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t} - \cdots$로 계산된다.

설정된 회귀모형이 주어진 자료를 설명하는 정도를 판단하기 위해 잔차분석을 실시한다. 일반적으로 오차의 추정값으로 잔차를 통계량으로 이용하는데, 오차항에 대한 기본 가정을 검토할 때 분석 대상이 된다.

오차항 점검은 잔차에 대한 산포를 그래프로 나타내어 확인한다. 오차항의 분산이 일정하지 않고, 시간이 흐름에 따라 시간에 비례하여 변화하는 모습이 나타나면 등분산 가정에 위배되는 것으로 판단하고, 자료를 변환하고 모형적합을 다시 시도한다. 그리고, 오차항 들의 독립성 가정에 대한 위배 여부 확인하고자, 자기상관성 테스트인 Durbin -Watson(DW)을 수행한다.

2. 시계열 모형[1)]

가. 개요

과거의 행태가 미래에도 그대로 지속된다는 가정에 기초하여 시계열형태로 제시된 과거 자료로부터 추세나 경향(일정한 패턴:규칙성/시계열변동)을 파악하여 미래의 관찰값을 예측하는 방법을 말한다. 시계열 모형에서 가장 널리 사용되는 2가지 접근 방식은 ARIMA 모형과 지수평활법이다. ARIMA 모형은 자료에 나타나는 자기상관을 설명하는 데 기반하고, 지수평활 모형은 자료의 추세와 계절성에 기초하여 표현하는 데 목적이 있다.

나. ARMA

1) 정의

시계열 ARMA 모형 분석의 기본 전제는 정상성 조건이다. 시계열 자료가 정상성 조건을 만족한다는 것은 해당 시계열이 관측된 시간에 무관하다는 것이다. 즉, 이러한 관점에서 추세나 계절성이 나타나는 시계열은 정상시계열이 아니다. 추세와 계절성이 존재하면, 서로 다른 관측시간에 시계열 값에 영향을 줄 수 있기 때문이다. 예를 들어, 자기상관이 없는 백색잡음 시계열(평균이 0이고 분산이 일정한 정규분포)은 정상성을 나타내는 대표적 시계열이다.

ARMA 모형의 기본 개념은 현재 시계열이 종속변수의 과거값과 오차항의 과거값을 이용하여 설명하는 것이다. ARMA모형은 AR모형과 MA 모형이 결합된 모형이고, AR모형(자기회귀과정모형)과 MA 모형(이동평균모형)을 각각 기술해 보면 다음과 같다.

회귀모형(다항 추세모형)에서 독립변수의 선형조합을 이용하여 종속변수를 예측하는 것과 같이, 자기회귀과정(AR)은 현재의 상태가 과거의 상태에 연관된다면 현재 관측값을 과거 관측값들의 함수 형태로 표현할 수 있다고 가정한다. 즉,

1) 시계열모형의 이론적 내용은 조신섭 외 SAS/ETS를 이용한 시계열 분석을 참고하였음.

$$Y_t = f(Y_{t-1}, Y_{t-2}, \cdots) + \epsilon_t$$

의 관계를 만족할 때 자기회귀과정(AR, autoregressive process)라고 한다. f 의 형태는 선형함수가 많이 쓰이고, ϵ_t는 평균이 0이고, 분산 σ_ϵ^2을 갖는 정규분포이고, 자기상관이 없는 시계열(백색잡음)로 가정한다. 일반 정상시계열에서 평균이 μ인 p-차 자기회귀과정, AR(p)과정은

$$Y_t - \mu = \sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu) + \epsilon_t$$

로 표현된다.

그리고, 이동평균과정(MA, Moving average process)은 현재의 시계열을 설명하기 위해 오차항(백색잡음)의 과거값을 모형에 포함하는 방식이고, 다음과 같이 표현할 수 있다.

$$Y_t = \mu + \epsilon_t - \theta \epsilon_{t-1}$$

이고, 모든 t에 대하여, $E(Y_t) = \mu$ 이다. 하지만 ϵ_t의 값은 실제 관측되는 값이 아니므로, 보통 생각하는 회귀모형의 선형 결합 유형과는 다르다. Y_t의 값은 과거 여러 개 오차항 들의 가중 평균으로 생각할 수 있다.

시계열 과정에서 평균과 분산은 시간 t와 무관하고, 오직 시차 k만의 함수이므로 이 과정은 정상시계열(확률과정)이 된다. 이 형태의 모형을 따르는 확률과정을 이동평균과정(MA)라고 한다.

$$Y_t - \mu = \epsilon_t + \psi_1 \epsilon_{t-1} + \cdots + \psi_q \epsilon_{t-q}$$

즉, 처음 q개의 ψ_j들은 $\psi_1 = -\theta_1, \psi_2 = -\theta_2, \cdots \psi_q = -\theta_q$ 이다

이제, ARMA모형의 설정과정과 사용 이유에 대해 설명하고자 한다. 시계열 자료를 자기회귀모형이나 이동평균모형 중 하나로만 설명하려면 p의 값, 또는 q의 값이 너무 커질 수 있다. 그리고 추정해야 할 모수 개수가 많아지면서 추정과정 효율성이 떨어지고 해석이 어려워진다. 만일, 자기회귀부분과 이동평균부분을 동시에 사용한다면 추정해야 할 모수의 개수가 줄어든다. 자기회귀 과정과 이동평균 과정을 동시에 포함하는 확률과정을 자기회귀이동평균과정이라 하고, 이 과정을 ARMA(p,q)

라고 표현한다.

$$Y_t = \delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots \theta_q \epsilon_{t-q}$$

ARMA모형 적합단계에서 모형 식별(모형차수 p, q)을 하기 위해 자기상관함수를 이용한다. 시계열 자료는 현재의 상태가 과거 및 미래의 상태와 밀접하게 연관되어 있으므로, 시간의 흐름에 따른 독립이 아니다. 분석에서는 시간에 따른 상관정도를 나타내기 위해서 다음과 같은 자기상관함수 또는 자기상관계수(ACF)를 사용한다. 이때, 기본가정으로 표본수가 충분히 커지면 자기상관계수(ACF, autocorrelation function)는 근사적으로 정규분포를 따른다. 즉 이를 이용하면 상관계수가 0인지 여부를 검정할 수 있게 된다. 즉, ACF를 이용하면 주어진 시계열자료가 시간에 따라 독립(백색잡음)인지 아닌지, 독립이 아니라면 시간에 따라 어떻게(시차 선택) 연관되어 있는지를 파악하는 데 유용하다. 결론적으로, 위 사실을 이용하여 이동평균과정(MA)의 차수 q를 결정하게 된다.

두 변수의 순수한 상관관계를 구하기 위해서, 다른 요인의 영향을 제거한 후, 자기상관계수를 계산하는데, 이를 부분자기상관계수라고 한다. 그리고, 부분자기상관계수가 그려진 그래프를 살펴보면, 주어진 시계열자료가 어느 확률과정의 모형으로부터 생성된 것인지를 짐작할 수 있다. 예를 들어, $Y_t, Y_{t+1}, \cdots, Y_{t+k-1}, Y_{t+k}$가 관측되었을 때, k시차만큼 떨어진 Y_t와 Y_{t+k}의 진정한 상관관계를 구하기 위해 시계열분석에서는 부분자기상관계수 PACF(partial autocorrelation function)를 이용한다. 즉, PACF 값은 Y_t와 Y_{t+k}에서 $Y_{t+1}, Y_{t+2}, \cdots, Y_{t+k-1}$의 의 효과를 제거한 후의 상관계수를 의미한다. 이를 이용하면 자기회귀모형(AR)의 차수 p를 결정할 수 있다.

ARMA과정의 이론적인 ACF와 PACF는 모두 절단되지 않고 감소하지만, ACF 또는 PACF가 특정 차수에서 절단되는 경우(더 이상 상관관계가 없는 경우), 모형에 포함하게 될 차수, p와 q가 결정된다. 이러한 특성들을 활용하여, AR(p)과정, MA(q)과정, ARMA(p,q)과정의 모형식별을 위하여 ACF와 PACF의 형태를 활용하게 된다.

〈표 3〉 정상시계열 과정의 ACF와 PACF의 특징

	ACF	PACF
AR(p)	지수적으로 감소하거나 소멸하는 사인함수 형태	p시차 이후에는 0으로의 절단형태
MA(q)	q시차 이후에는 0으로의 절단 형태	지수적으로 감소하거나 소멸하는 사인함수 형태
ARMA(p,q)	시차(q-p)이후에 지수적으로 줄어들거나 소멸하는 사인함수 형태	시차(p-q)이후에 지수적으로 줄어들거나 소멸하는 사인함수 형태

다. ARIMA

일반적으로 사회과학분야, 보건의료 분야에서 접하는 시계열은 정상시계열이 아닌 경우가 대다수이다. 즉, 시간이 흐름에 따라 관측값이 시계열 추세를 갖고 증가하거나 분산 정도가 변화하는 등 정상성 특성을 만족하지 못하는 경우가 있다. 이와 같이, 시계열이 추세를 갖거나, 시계열의 수준이 시간대에 따라 다르거나, 계절성을 보이거나, 시계열의 분산이 시간에 비례하여 변화하는 경우가 비정상성에 해당된다.

시계열분석 이론들은 대부분 정상성을 가정하고 있으므로, 시계열에 비정상성이 나타난 경우, 로그변환 또는 추세분석이나 차분을 통한 추세 제거 방법을 이용하여 정상시계열로 전환한 후 정상시계열 분석 방법을 적용해야 한다.

시계열이 증가하는 추세를 보이는 경우, 이는 비정상시계열이므로, 추세를 제거해야 주어야 한다. 만일 추세의 증가하는 경향이 시간의 규칙적인 함수의 형태로 표현할 수 있으면 결정적 추세라 하고, 규칙적인 함수로 표현하기 어려우면 확률적 추세라고 한다.

결정적 추세의 경우 차분을 통해 정상화하기 보다는 시간을 독립변수로 하여 추세모형을 적합시켜 주면 된다. 즉 결정적 추세를 갖는 시계열은 추세를 모형화하여 제거하면, 정상시계열을 따르게 된다. 시계열을 $Y_t = T_t + S_t + I_t$ 라고 하면, T_t(추세)와 S_t(계절성)는 최소제곱법 또는 평활법과 같은 추정법으로 추정할 수 있다. 추세와 계절성은 확률적인 프로세스가 아니므로 분산이 0이고, 평균($\hat{I}_t = Y_t - \hat{T}_t - \hat{S}_t$)이 0인 정상시계열이 된다.

확률적 추세가 존재하는 경우에는 정상시계열로 변환해야 한다. 이 경우 차분을 이용하면 정상시계열로 적용 가능하다. 비정상과정(확률적 추세)의 존재 여부를 확인하는 방법으로 Fuller와 Dickey의 단위근 검정방법이 있다. 이를 통해 확률적 추세의 존재여부를 확인할 수 있다. 차분을 통한 정상화 시계열은 자기회귀누적이동평균 모형으로 표현할 수 있다.

즉, 자기회귀누적이동평균 모형 ARIMA(p,d,q)는 d-차분된 시계열 $W_t = (1-B)^d Z_t$가 평균 수준이 μ로 일정하고, 정상시계열 ARMA(p,q)을 따르게 되는 경우를 일컫는다.

차분과 관련하여 차수 d는 시계열 그림과 SACF(sample autocorrelation function)를 참고하여 결정하는데, 시계열이 추세를 갖고 있는 경우, 또는 SACF가 서서히 감소하는 경우에 차분이 필요하다. 예를 들어, 결정적 추세가 아니면서, 몇 개의 추세가 연속해서 나타나는 비정상시계열이 있다면, 차분(1-B)을 취하여 ARIMA 모형으로 분석을 시도할 수 있다.

계절형 자료가 아닌 대다수의 시계열 차분의 차수는 0,1,2,이면 충분하고, 계절형 자료의 경우도 계절 주기와 동일한 차수의 차분을 실시하면 좋다. 또한 차분이 이루어진 경우, 모형에 상수항을 포함할지 여부도 함께 고려해야 한다.

ARIMA 모형을 적합하기 위해서는 우선 시계열이 정상적인지를 확인해야 한다. 시계열의 정상성 여부를 확인하려면 특별한 추세, 계절성을 체크하고, 수준과 변동폭이 일정한지 점검한다. 만일 시계열이 정상적이 아니라면, 변환 또는 차분을 통해 정상화 과정을 수행한다.

모형 적합과정에서는 ARMA(p,q) 모형 중에서 가장 적합한 모형 하나를 선정한다. 즉 시계열을 가장 잘 설명할 수 있는 차수p 와 차수q 값을 선택한다. 이 단계에서의 주의사항은 p와 q를 크게 잡으면 추정해야 할 모수가 증가하므로 비효율적이다.

그리고, p와 q값이 선택되면, 여러 가지 추정법을 이용하여 AR-모수$\phi_1, \phi_2, \cdots, \phi_p$와 MA-모수$\theta_1, \theta_2, \cdots, \theta_q$ 및 시계열의 평균값 μ 와 오차항 백색잡음의 분산 σ_ϵ^2을 추정한다. 다음으로 모형의 진단과정은 주로, 잔차 분석을 이용해 수행하고, 독립성 검정, 정규성 검정을 진행한다. 정규성 검정은 정규확률그림 또는

Jarque-Bera검정통계량을 이용할 수도 있다. 그리고, 백색잡음 가정을 확인하려면 잔차의 시계열 그림을 그려보거나, 잔차의 표본자기상관계수(RSACF; residual SACF)가 0인지 유의성 검정 결과도 확인한다.

라. ARIMAX

분석 대상이 되는 시계열을 설명하기 위해서 현재의 관측값과 과거 관측값, 오차항의 현재와 과거의 값들만을 이용하는 모형이 단변량 시계열분석이다. 반면에, 분석 대상이 되는 시계열과 밀접한 관련성이 있는 시계열 변수가 있는 경우에 이를 모형에 독립변수(입력변수)로 포함하여 모형을 개선할 수 있다. 단변량 시계열인 ARIMA 모형에 추가적인 외부 영향을 고려할 수 있는 모형으로 전이함수 모형이나 동태적 회귀모형개념과 연결하여 인용되거나 설명되기도 한다. 즉, ARIMAX모형은 ARIMA모형에 독립변수를 외생변수로 포함하는 자기회귀이동평균모형(autoregressive moving average model with exogenous inputs model, ARIMA(p,d,q,r))을 의미한다. 이때, p, d, q는 각각 ARIMA 모형의 차수, r은 독립변수의 지체차수(lagged order)를 의미한다. 이 모형에는 독립변수가 시계열 형태로 결합하여 장기 예측력이 우수하다는 장점이 있다.

일반적으로 ARIMAX 모형은 자기회귀오차모형보다 더 일반적인 모형이고, 변수 선택 결과에 따라 모형 적합이나 분석 이후의 결과에 많은 차이가 발생할 수 있다.

ARIMAX 모형적합에서 잔차분석은 중요한 단계이다. 백색잡음 검토, 잔차의 자기상관함수(ACF, autocorrelaton coefficent function) 및 부분자기상관함수(partial autocorrelation coefficeint function) 검토, 단위근 검정 등의 모형적절성 진단 절차가 있다. 백색잡음의 경우는 포맨토우 검정을 이용한다. 차분에 대한 필요성을 검토하기 위해 단위근 검정을 실시하는데, ADF (Augmented Dickey Fuller) 검정 또는 KPSS (Kwiatkowski-Phillips-Schmidt-Shin) 검정을 통하여 확인한다.

제안된 모형이 수립되면, 최적의 모형을 선택해야 하는데 이때에는 정보량 기준 또는 예측오차 기준을 고려하여 선택한다. 정보량 기준 유형에는 AIC(Akaike information criterion), SBC(Schwarz bayesian information criterion) 이

있고, 예측오차기준에는 MSE(mean squared errors), MAE(mean absolute errors)가 있다.

이번 연구에서 적용 예정인 ARIMAX(p,d,q,r) 모형은 아래와 같은 형식으로 표현할 수 있다. 원시 자료 y_{0t} 에 대하여 $y_t = (1-B)^d y_{0t}$는 정상성이 만족되는 종속시계열이라고 가정한다.

$$\phi(B)y_t = \mu + \psi(B)x_t + \theta(B)z_t$$

여기서 $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$는 시차연산자(B)를 이용하여 종속시계열의 과거 값이 영향을 포함할 수 있으며, $\theta(B) = 1 - \theta_1 B - \cdots - \theta_p B^q$는 백색잡음 z_t가 서로 상관될 수 있는 유연성 있는 구조를 허용한다. 외생적 변수인 x_t가 종속변수인 y_t와 지체되어 영향력을 가질 수 있도록 그 계수들의 구조가 $\psi(B) = 1 - \psi_1 B - \cdots - \psi_r B^r$와 같다

ARIMAX 모형에서 독립변수들의 지체 차수(r)을 결정하는 것은 어려운 과정일 수 있다. 지체차수(r)을 정하기 위해서, 독립 변수시계열과 종속 시계열간의 교차상관계수를 구하고, 이를 통해 지체차수를 결정하는 것과 관련하여, 후보 차수들을 선택할 수 있다.

마. 지수평활법

시간추세 모형을 가정하여 예측값을 구하는 방법으로, 추세 모수가 서서히 변화한다고 가정하고 과거의 관측값들을 지수적으로 가중 평균하여 예측을 시도한다. 시계열에 따라 변화가 발생하는 경우에는 과거의 모든 자료에 동일한 비중을 부여하는 것보다 최근의 변화 시점에 가까운 자료에 큰 비중을 두는 예측법이 더욱 합리적이라 할 수 있다.

지수평활법은 최신 자료 활용으로 시계열 자료 변화에 쉽게 대처할 수 있으며, 계산법이 쉽고, 과거의 많은 자료 저장이 필요 없다는 것이 특징이다. 그리고, 지수평활법 종류에는 단순지수평활법, 이중지수평활법, 삼중지수평활법, 윈터스의 계절지수평활법 등이 있다.

단순지수평활법 모형은 다음과 같이 표현할 수 있다.

$$Y_t = \beta_{0,t} + \epsilon_t$$

여기서 ϵ_t는 서로 독립이고, 평균이 0이고, 분산이 σ_ϵ^2을 가지며, $\beta_{0,t}$는 시간에 따라 변화하는 모수이다. 위 모형을 부연 설명하면, 미시적으로 동일한 평균수준을 유지하지만, 거시적으로는 시간대별로 평균수준이 변화하고 있다. 즉, 평균수준 $\beta_{0,t}$는 고정되어 있지 않고, 변화하는 시스템이 반영되어 그 수준이 변화한다. 즉, 새로운 관측 자료가 갱신될 때마다, 변화의 정보를 반영한 $\beta_{0,t}$의 추정값이 갱신되어야 한다. 즉 $Y_n(=\beta_{0,n})$의 예측값 추정치는 $\omega\sum_{j=0}^{\infty}(1-\omega)^j Y_{n-j}$로 표현할 수 있다. 이는 시점 n까지 자료들의 가중평균 값이다. 그리고, 최근에 관측된 자료값들의 가중치들은 각각 $\omega, \omega(1-\omega), w(1-\omega)^2 \cdots$ 이므로 과거의 값일수록 가중치가 작아지는 형태이다. 즉 최근의 관측치 값들일수록 가중값은 지수적으로 증가하여 $Y_n(=\beta_{0,n})$의 예측값 추정에 많이 반영된다.

단순지수평활법의 장점은 예측의 갱신이 쉽다는 것이다. 새로운 관측값이 추가될 때마다 현재 시점까지의 예측값과 최근의 관측값을 결합하여 예측값을 추정할 수 있다.

이중지수평활법은 시계열이 선형추세에 따라 증가하는 경우인데, 모형으로

표현하면 다음과 같다.

$$Y_t = \beta_{0,t} + \beta_{1,t}t + \epsilon_t$$

이는 미시적으로는 동일한 추세를 갖지만 전체적으로 시간대별로 선형추세를 따라 변화하는 특징을 가진다.

삼중지수평활법은 시계열이 시간대별로 2차 추세 모형을 따라 변화하는 특징을 가진다

$$Y_t = \beta_{0,t} + \beta_{1,t}t + \beta_{2,t}t^2 + \epsilon_t$$

계절성분과 같이 일정한 주기를 바탕으로 변화하는 시계열인 경우 계절지수 평활법을 이용하여 표현할 수 있다.

시간의 흐름에 따라 시계열의 평균수준이 변화하지만 그 변동 폭이 시간의 흐름에 관계없이 일정한 경우에는 가법계절모형을 이용하고, 분산이 시간의 흐름에 따라 점차 커지는 경우에는 승법계절모형을 이용한다.

가법계절모형은 시간의 흐름에 관계없이 시계열의 변동 폭이 동일한 경우에 사용되며, 일반적인 윈터스의 가법계절모형은 시점 n+l에서 다음과 같다.

$$Y_{n+l} = T_{n+l} + S_{n+l} + I_{n+l}$$

여기서 T_{n+l}은 추세성분, S_{n+l}은 계절주기 s를 갖는 계절성분, 그리고 I_{n+l}은 오차항으로서 불규칙 성분에 해당한다.

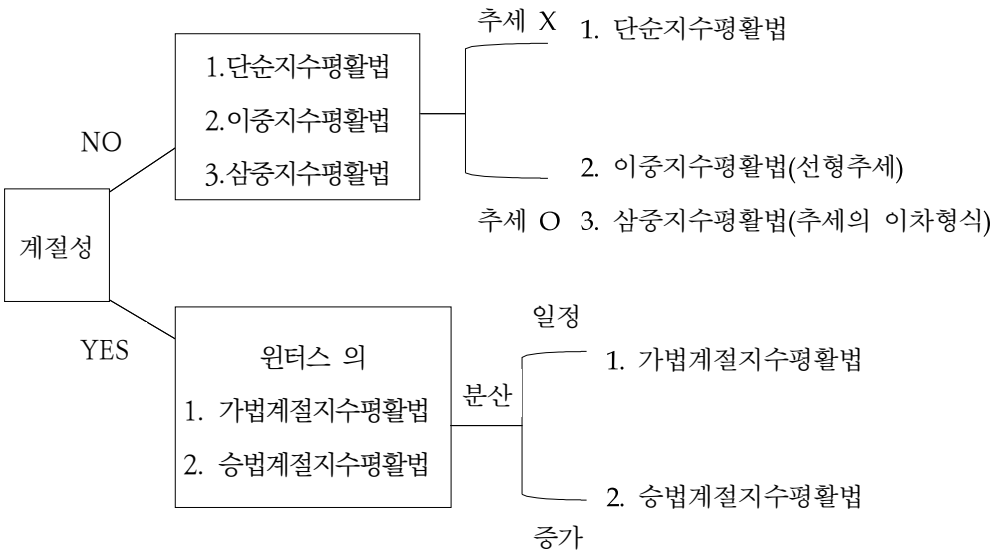
윈터스의 가법계절지수 평활법은 시계열이 추세성분과 계절성분 및 불규칙성분들의 개별값을 평활법에 의해 추정한 후 이를 합성하여 예측값을 구하는 방법이다.

승법계절지수 평활법은 추세와 비례하여 시계열 변동 폭과 계절주기의 폭이 변화할 때 사용하는 방법이다. 즉, 추세성분과 계절성분의 곱에 의해 시계열을 표현한다. n+l시점에서 예측값은 아래와 같이 표현된다.

$$\begin{aligned} Y_{n+l} &= T_{n+l}S_{n+l} + I_{n+l} \\ &= (T_n + \beta_{l,n}l)S_{n+l} + I_{n+l} \end{aligned}$$

여기서 T_{n+l}은 추세성분, S_{n+l}은 계절주기 s를 갖는 계절성분, 그리고 I_{n+l}은

오차항으로서 불규칙 성분에 해당한다. 윈터스의 승법계절지수평활법은 추세성분과 계절성분의 곱에 불규칙 성분이 더해진 것으로 가정하고, 개별 성분들을 평활법에 의해 추정한 후 이를 활용하여 예측값을 구하는 방법이다



[그림 4] 자료특성에 따른 지수평활유형 선택방법

바. 자기회귀오차모형

시계열 자료 분석 시 오차항이 시간에 따라 자기 상관을 갖거나, 이분산 형태를 보이는 경우가 많다. 이때 회귀모형을 적용을 위한 기본 전제는 오차항의 독립이다. 이 전제가 만족하지 않아도, 최소 제곱법을 이용하여 모수추정이 가능하지만, 모수추정은 효율성이 떨어지고 예측값의 정확성 검정에 오류가 발생할 수 있다. 반면에 시계열 자료의 속성상 시간의 흐름에 따라 과거의 측정된 값에 영향을 받아 자기상관성이 존재할 수 있다. 이를 감안하여, 회귀모형을 개선한다면, 예측의 편의가 발생되는 요인을 제거할 수 있고, 예측의 성능을 높일 수 있다.

잔차의 정규성 검정을 위한 Shapiro-Wilks 검정을 시행한 다음, 회귀모형을 적합하여 잔차의 자기상관성여부 검정은 Durbin-Watson(DW) 검정을 수행한다

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + .. + \beta_p X_{tp} + \epsilon_t,\ t = 1,2,\cdots,n$$

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + ... + \phi_k \epsilon_{t-k} + a_t$$

이고, a_t들은 서로 독립이다.

자기회귀오차모형이 일반회귀모형과 다른 점은 오차항인 ϵ_t들이 회귀모형처럼 서로 독립이 아니고 자기상관성이 존재한다는 것이다. 즉, 오차항이 자기회귀과정 AR(k)를 따르는 특징이 있다.

3. 머신러닝(신경망) 모형

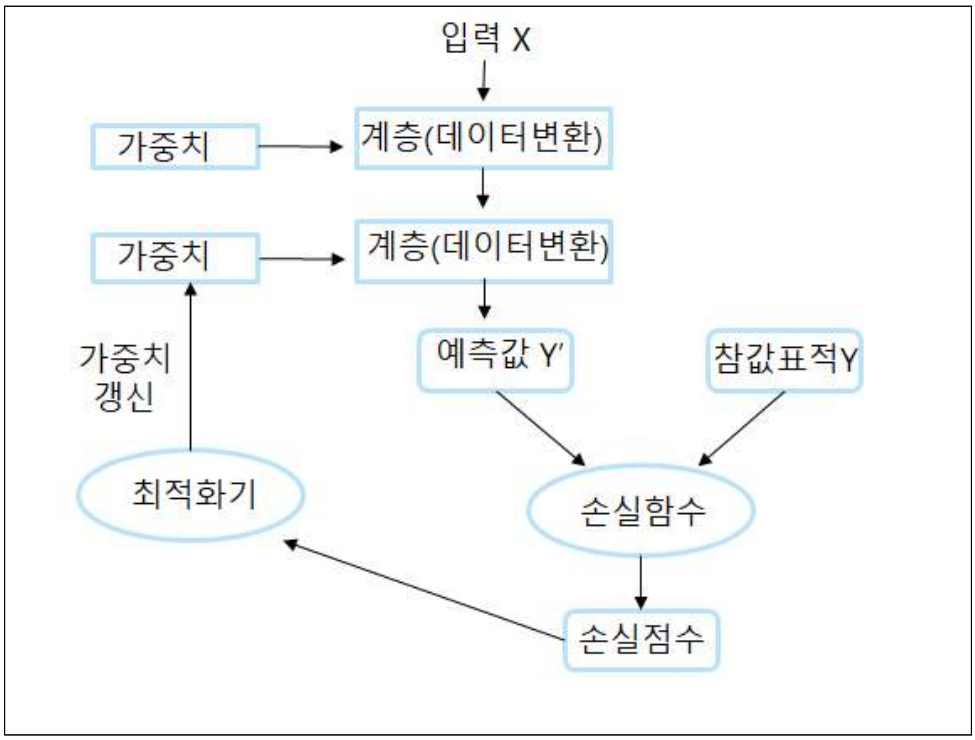
가. 개요

머신러닝(신경망) 모형은 주어진 자료를 컴퓨터로 학습하여 명시적으로 정의되지 않은 패턴을 결과로 만들어 내는 방식이다.

빅데이터를 이용한 머신러닝 예측법은 축약형 접근법으로 데이터를 활용하여 유용한 정보를 추출하게 된다. 머신러닝은 예측문제 환경에 대한 사전적인 이론적·구조적 지식 없이 주어진 알고리즘을 통해 데이터에 담긴 정보를 추출해 내는 학습 방법으로 다양한 구조의 데이터를 분석할 수 있다. 알고리즘을 통해 데이터를 학습하고 예측변수와 관련 있는 정보를 추출함으로써 회귀분석이 가능한 프레임을 제공하기도 한다.

궁극적으로 머신러닝 기법은 빅데이터에 축적된 정보를 찾아내 이용할 수 있는 틀을 제공하며, 검증자료에 대한 결과는 쉽게 도출할 수 있으나, 분석 결과의 직관적 해석은 어려울 수 있다. 그리고 머신러닝 기법은 변수 간의 관련된 사전정보 없이 빅데이터에 담긴 과거 정보를 알고리즘에 따라 분석하게 되는데, 작업자의 조정 없이 단순한 기계학습을 수행할 경우 잘못된 결론이 도출될 수도 있다. 즉, 예측은 가늠할 수 없는 미래 변수의 행동변화나 충격을 모델 설계에 반영할 수 없는 한계가 있으므로, 작업자의 통찰력을 반영한 정성적인 평가도 중요하다.

머신러닝의 예측 모델에 대한 평가는 이론적 적합성뿐 아니라 예측력에 대한 상대적 비교 형태로 이루어지는데, 예측모델링의 경우 변수 예측력을 제고하는 방법론을 선택하기도 한다. 예측력에 대한 평가는 실증 분석 결과에 기반한 결과론적 평가로 이루어지는데, 문제 환경(자료 개수, 예측기간)에 따라 예측 결과가 다르게 나타나므로, 하나의 모델을 선택하기 보다는 각 모델에 상응하는 가중치를 부여하여 예측값들을 종합하는 방법도 쓰이고 있다.



[그림 5] 머신러닝- 회귀형 문제 학습 개념도

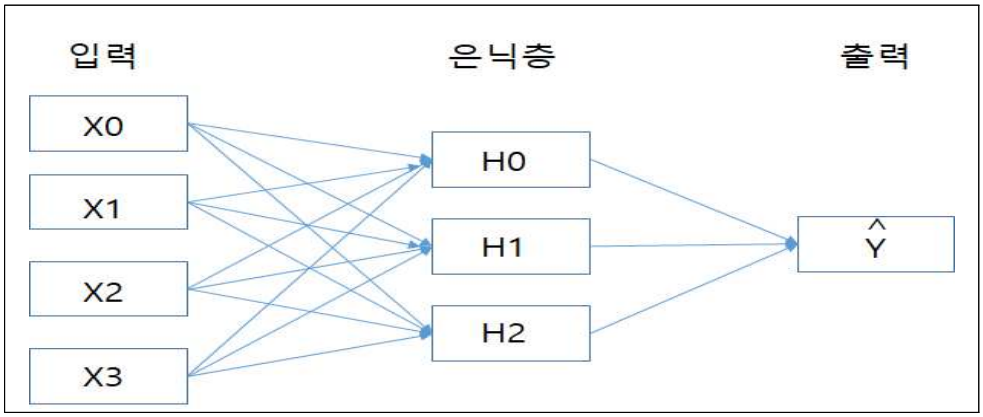
그리고, 미래 추정값을 산출할 수 있으나 실제값과의 비교가 불가능하므로, 예측력 평가는 현재의 주어진 데이터에 기초해서 수행된다. 즉 현재 활용 가능한 자료를 모델 적합용과 예측력 평가용으로 나누어 모델의 예측력을 평가한다.

나. 방법론

기계학습을 통해 해결 가능한 문제는 데이터의 클래스를 구분해야 하는 패턴인식 문제와 연속적인 값을 추정해야 하는 회귀 또는 함수 근사화 문제로 나눌 수 있다. 또한 학습 방법에 따라 지도학습, 비지도학습, 반지도학습으로 구분할 수 있다. 지도학습이란 학습데이터의 분류값이나 출력값을 알고 이에 대한 피드백을 통해 학습하는 방법이다. 비지도학습은 학습데이터의 분류값이나 출력값을 활용하지 않고, 데이터 유형이나 클러스터 밀도 등을 추정하는 방법을

일컫는다. 그리고, 반지도학습은 분류값이나 출력값을 아는 데이터와 모르는 데이터를 함께 사용하는 경우를 말한다. 결과에 대한 피드백만 주어지고 정확한 분류값이나 출력값은 주어지지 않는 강화학습으로 구분 할 수 있다.

머신러닝 기법 중 하나인 신경망기법은 사람의 뇌 구조를 모방한 기계학습 방법으로 계량적 접근을 통해 결정경계를 찾는 방법이다.



[그림 6] 머신러닝(신경망) 구조도

사람의 뇌에는 10^{11}개의 뉴런이 있고 이 뉴런들 사이에 10^{15}개의 시냅스 연결이 존재한다. 1943년에 McGulloch와 Pitts에 의해 제안된 McCulloch-Pitts neuron은 가중치가 곱해진 입력값들의 합을 계산하여 그 합이 임계값을 넘으면, 1 아니면 0을 출력하는 인공뉴런이다. 1957년에 Rosenblatt는 이 인공뉴런의 개념을 기반으로 퍼셉트론이라고 불리는 뉴런의 입력값에 곱해지는 가중치를 학습하는 인공신경망 모델을 만들었다. 그러나 퍼셉트론은 다양한 문제에 적용할 수 있을 것으로 기대되었지만, XOR과 같이 선형으로 분리되지 않는 문제는 해결할 수 없다는 점이 지적되었다. 이 문제는 다층 신경회로망을 사용하면 풀리게 되는데, 다층구조를 사용하게 되면서 몇 개의 은닉층을 사용할 것인지, 몇 개의 은닉뉴런을 사용할 것인지, 어떤 활성함수를 이용할 것인지 등 신경회로망의 구조를 결정해야 한다. 여기서 한 층의 은닉층만 사용해도 문제를 해결할 수 있다고 알려져 있는데, 이는

국소적으로 한계가 있고, 구간별로 연속인 활성함수가 비다항식 형태인 한 층의 은닉 뉴런으로 어떤 연속적인 함수 학습할 수 있다는 이론에 근거한다. (Universal approximation theorem)

머신러닝(신경망)을 구축할 때 고려할 점은 편향-분산 트레이드 오프 문제이다. 기계학습 모델의 성능 평가 기준에는 이용되는 평균제곱오차는 편향과 분산으로 구성되어 있다. 편향은 모델을 학습하는 데 학습 데이터를 얼마나 유연하게 받아들일 것인가에 대한 지표로 학습데이터를 유연하게 고려하지 못하면 기계학습 모델이 제대로 학습되지 않는 과소학습 문제가 발생한다. 즉 편향값은 데이터에 따라 모델이 정확도가 어떻게 변하는지를 특정하는 기준이 된다. 분산은 학습데이터에 대한 모델의 민감도를 나타내는 지표로 높은 분산값은 학습데이터에 포함된 노이즈까지 기계학습 모델이 학습했다는 의미가 되고, 과도학습 문제로 이어지게 된다. 편향과 분산은 서로 반비례하고 있으므로 필연적으로 트레이드-오프가 발생한다.

머신러닝 모델 종류에는 머신러닝(신경망)외에, 기저벡터머신, 확률밀도 분포 추정법이 있고, 패턴인식 문제 학습을 위해 많은 수의 신경층을 갖고 모델을 구성하는 딥러닝 기술이 있다.

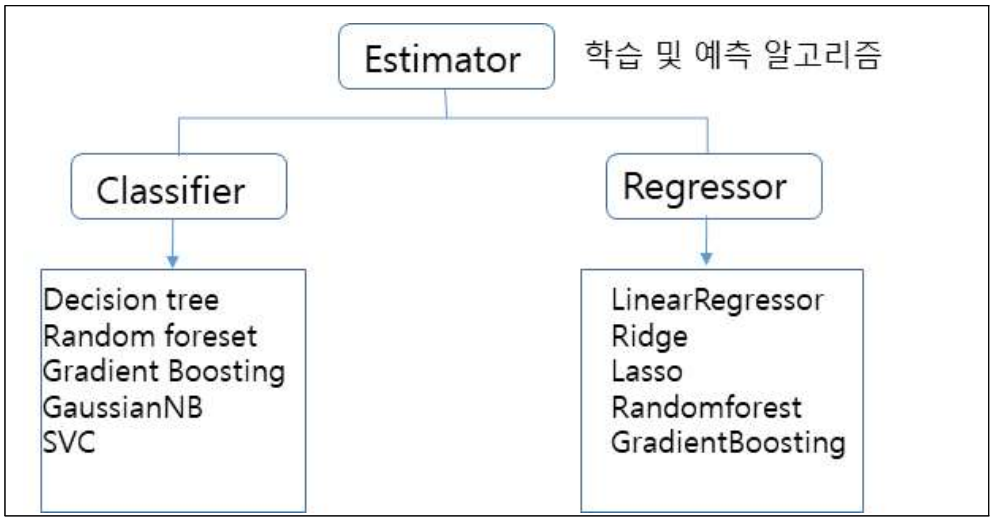
본 연구에서는 미래 예측방법의 가능성을 비교하기 위해서, 머신러닝(신경망)을 선정하여 분석 방법 및 성능을 검토하였고, 머신러닝 툴 SciKit-learn library를 이용하여 미래예측을 수행하고자 하였다. 만일 단순 머신러닝 이외의 딥러닝을 수행하고자 한다면, Keras, Tensorflow, Pytorch 등 툴을 활용해야 한다.

SciKit-learn 은 MLP(mutilayer perceptron) 출력레이어로 구성되어 있고, 손실함수 형태는 회귀문제 유형에서 제곱오차, 분류형 문제유형에서 교차엔트로피 함수가 사용된다. 각 계층에 대해서 서로 다른 활성화 함수 및 가중치 설정 등과 같은 매개변수의 미세한 조정이 불가능하지만, 본 라이브러리는 사용하기가 쉽고, 실용적인 기능을 다수 제공하고 있다.

Scikit-learn은 파이썬 머신러닝 라이브러리 중에서 가장 많이 사용되는 라이브러리이다. 사용 시에 매우 다양한 알고리즘을 개발하는 데에 편리한

API를 제공해 준다. 오랜 기간 실전 환경에서 검증 되었으며, 여러 환경에서 이용되는 일반적인 라이브러리이다.

Scikit-learn은 MLP 모델 학습을 위해 fit() 함수와 학습된 모델을 예측하기 위해 predict()함수를 제공한다. 분류 알고리즘 구현을 위해 classifier 클래스를, 회귀알고리즘을 구현한 Regressor 클래스를 제공하고 있다. 클래스 classifier와 regressor를 estimator라고 하고, 이 estimator가 지도학습의 모든 알고리즘을 구현한 클래스이다.



[그림 7] SciKit-learn library의 목적별 기능 분류

Scikit-learn의 주요 모델과 기능을 요약하면 다음과 같다

〈표 4〉 SciKit-learn의 핵심 기능

분류	모듈명	기능
예제 데이터	sklearn.datasets	Sciket-learn에 내장된 예제 데이터 세트
피쳐 처리	sklearn.preprocessing	데이터 전처리에 필요한 다양한 가공 기능
	sklearn .feature_selection	알고리즘에 큰 영향을 미치는 피처를 우선 순위로 selection작업을 수행하는 기능
	sklearn .feature_extraction	텍스트 데이터나 이미지 데이터의 백터화된 피처를 추출하는데 사용
피쳐 처리 차원축소	sklearn .decomposition	차원 축소와 관련된 알고리즘을 지원 -PCA, NMF, Truncated SVD 등을 통해서 차원 축소 기능 수행 가능
데이터 분리, 검증 패러미터 튜닝	sklearn .model_selection	교차 검증을 위한 학습용/테스트용 분리, 그리드 서치로 최적 파라미터 추출 등의 API 제공
평가	sklearn.metrics	분류, 회귀, 클러스터링, 페어와이즈에 대한 다양한 성능 측정 방법 제공 예시) Accuracy, Precision, Recall, ROC-AUC 등
ML 알고리즘	sklearn.ensemble	앙상블 알고리즘 제공 랜덤 포레스트, 에이다 부스트, 그래디언트 부스팅 등
	sklearn.linear_mode	선형 회귀, 릿지, 라쏘 및 로지스틱 회귀 등 회귀 관련 알고리즘 지원
	sklearn.naive_bayes	나이브 베이즈 알고리즘 제공 -가우시안 NB, 다항 분포 NB 등
	sklearn.neighbors	최근접 이웃 알고리즘 제공(K-NN 등)
	sklearn.svm	서포트 벡터 머신 알고리즘 제공
	sklearn.tree	의사 결정 트리 알고리즘 제공
	sklearn.cluster	비지도 클러스터링 알고리즘 제공 (K-평균, 계층형, DBSCAN 등)
	sklear.neural_network	분류, 회귀, RBM(확률분포 학습에 기반한 신경망) 알고리즘 제공
유틸리티	sklearn.pipeline	피처 처리 등 변환과 ML 알고리즘 학습, 예측 등을 함께 묶어서 실행할 수 있는 유틸리티를 제공

자료원: 사이킷런으로 시작하는 머신러닝(https://jaaamj.tistory.com/16) 수정·보완

4. 일반화선형 모형(카운트형 자료)

카운트형 자료 대상으로 예측모형 적용하여 분석을 진행할 때, 분포 가정에 대한 조건이 성립하지 않는 경우가 다수 존재한다. 이때, 분석 데이터에 로그를 취하여 분석하거나, 자료 분포 조건에 정규분포가 아닌 감마분포, 로그노말 분포 등을 이용하기도 한다. 즉, 분석대상 데이터의 종류에 따라 모델 선택, 모델 적합, 해석 등 적용방법이 달라 질 수 있다. 이와 같이 다양한 환경을 고려하여 데이터를 분석할 수 있는 방법이 일반화선형모형의 적용이다.

일반화선형모형(Generalized Linear Model)은 종속변수에 영향을 주는 한 개 이상의 독립변수 효과를 추정할 수 있는 기존의 선형모형을 일반화한 모형이다. 일반화 선형모형을 구성하는 성분은 확률요소, 선형예측자 성분, 연결함수이다. 확률요소는 종속변수의 분포를 의미하고, 선형예측자 성분은 독립변수(X)를 규정하여, 확률분포의 평균인 종속변수 Y의 기댓값 $E(Y) = \mu$를 나타낸다. 즉 종속변수는 독립변수 수준에 따라 다양하게 나타나게 된다. 연결함수는 확률요소(E(y))와 선형예측자 성분(βX)을 연결하는 역할을 한다. 일반화선형모형은 연결함수로서 항등함수, 로그 함수 등을 포함하여 여러 가지 연결함수를 적용할 수 있다.

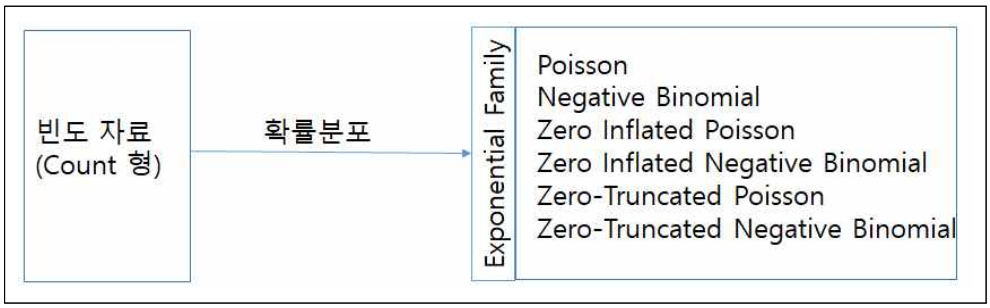
$$E(Y) = \mu$$

$$g(\mu) = \beta X$$

여기서, g는 연결함수, E(Y)는 확률요소, βX는 선형예측자 성분이다.

기존의 선형 회귀모형은 확률요소에 대한 정규분포를 가정하였고, 연결함수는 항등함수를 사용하고 있다.

일반화 선형모형을 이용하게 되면, 연속형 변수와 카운트형 변수를 모두 분석할 수 있다. 본 보고서에서는 카운트형 변수에 대해서 기술하고자 한다. 비연속적인 변수 중 병원 방문 횟수는 0번, 1번, 2번, ...처럼 셀 수 있는 값으로 즉 0과 양의 정수 범주에 속하게 된다.



[그림 8] 카운트형 자료에 대한 분석모형

빈도자료(count형)를 분석하는 적합한 일반화선형모형(Generalized Linear Model)으로 가장 기초적인 모델인 Poisson(포아송)이 있으며, 이밖에 Negative Binomial(음이항 분포) 등 여러 가지 분포가 있다. 포아송 모델은 빈도변수를 분석하기 위해 만들어진 초기의 모델이며 자주 이용되는 모형이다. 다만 평균과 분산이 같아야 한다는 조건 때문에 실제 상황에 적용하는 데 한계가 있기도 한다. 종속변수의 평균과 분산의 크기를 비교하여 분산이 평균보다 큰 경우, 과대산포 되었다는 전제하에 포아송 대신 음이항 분포를 사용하게 된다. 관측되지 않은 이질성 때문에 과대산포 문제가 발생할 수 있으며, 모델에서 중요변수를 누락 했을 때 과대산포의 문제가 발생할 수도 있다. 이 경우 중요변수들을 모델에 포함하고 포아송 분포를 사용해도 타당하다.

과대산포의 원인으로 확인해야 할 부분은 이상치(Outlier) 존재 여부이다. 이상치로 인해 과대산포 문제가 발생할 수 있으므로, 사전에 이상치를 탐색하여 제거하는 것이 최적의 모델 적합에 필요한 사전 작업으로 고려되어야 한다.

일반적이지 않지만, 빈도변수(Count형 변수) 분석에 선형회귀모형(ordinary least squares)을 적용하기도 한다. 빈도변수에 대한 로그변환을 취한 후 회귀분석을 시행하게 되는데, 이 경우에 0값으로 인한 자료 손실, 편의된 추정 등 여러 이슈 사항이 발생할 수 있다.

제4장 건강보험 청구자료를 활용한 사례분석

1. 분석 방법

정책 영향으로 진료비 청구 변화가 크게 나타나는 보장성 항목(초음파)과 건강보험 총진료비 청구자료 대상으로 분석 자료를 생성하였다. 이 자료는 DW(데이터웨어하우스)를 이용해 구축하였으며, 월별 및 연도별 자료 형태로 수집되었다. 외부변수 관련 자료 수집은 통계청의 KOSIS를 이용하였고, 시간변수는 제곱값을 생성하여 모형개발에 활용하였다.

분석 자료를 활용한 예측방법별 성능을 비교하기 위하여, 연속형 변수(진료비)와 카운트형 변수(실시횟수)로 구분하였다. 비교 대상의 예측방법은 회귀모형, 시계열모형, 머신러닝(신경망)이고, 성능측도는 MAPE(평균절대백분위 예측오차), RMSE(평균 제곱근 예측오차)를 이용하였다. 최적의 예측모형을 적합한 후, 각 예측모형별 3개월 또는 3개년의 예측값을 생성한 후, 실제 관찰값과 비교하여 예측오차를 계산하였다.

통계분석 수행은 SAS EG와 SciKit-learn library를 사용하였다. SAS EG는 대용량 자료 분석이 가능하고, 회귀모형, 시계열모형 등 다양한 통계분석과 데이터 가공이 우수한 통계 분석 툴이다. 그리고, SciKit-learn은 머신러닝(신경망) 등 인공지능 구현에 필요한 분석 툴로서, 이를 이용하기 위해서 파이썬 및 쥬피터 노트북 등 분석 환경을 새로 구축하였다.

2. 분석 대상

가. 자료유형

자료유형별, 축적기간별 분석 대상의 특징을 기술하면 아래 표와 같다. 초음파 청구자료는 심사결정기준 2018년 1월부터 2020년 5월까지의 자료를 포함하였다. 분석을 위해 진료월 기준으로 자료를 재정렬하였고, 진료비 및

시행횟수를 월 단위로 합산하여 산출하였다. 그리고, 건강보험 전체 청구자료는 1990년부터 2019년까지의 심사결정기준으로 구축하였고, 소비자물가지수, 1인당 GDP 변수도 동일기간을 적용하여 구축하였다.

〈표 5〉 분석 대상별 자료 특성

	분석 대상	
	초음파 청구자료	건강보험 전체 청구자료
모형구축 자료	- 12개월('18.04~ '19.03) - 24개월('18.04~ '20.03) - 30개월('18.04~ '20.09)	-27개년 (1990년 ~ 2016년)
연속형/카운트형 (수집단위)	-연 속 형: 초음파진료비(월별) -카운트형: 초음파시행횟수(월별)	-심사결정진료비(연도별)
독립변수		-소비자물가지수(연도별) -1인당 GDP(국민총생산, 연도별) -시간*시간(연도)
예측기간	-단기예측: 3개월	-단기예측: 3개년

나. 예측방법

예측모형별 연속형 예측값 산출에는 회귀모형(다항 추세모형), 시계열(자기회귀오차, 지수평활, ARIMA, ARIMAX), 머신러닝(신경망)을 이용하였고, 카운트형 예측값 산출에는 일반화선형 모형 및 머신러닝(신경망)을 이용하였다.

카운트형 자료는 정규분포 가정을 벗어나기 때문에 poissson(포아송 분포)이나, negative binomial(음이항 분포) 형태를 주로 이용하게 된다. 이 경우, 과대산포 형태가 나타났는지 검증한 후, 적합한 자료 분포를 선정해야 한다. 본 분석(초음파 시행횟수)에서는 과대산포 분포형태가 확인되어, negative binomial 분포를 카운트형 자료 분석에 적용하였다. 적합 분포 검정 과정 절차에는 proc genmod의 옵션(dist=negbin scale=0 noscale)을 이용하였다.

독립변수(외부요인)를 포함한 ARIMAX와 머신러닝(신경망)의 경우 로그 변환을 통한 예측모형과 일반모형 2가지를 적합한 후 비교·검토하였다.

분석모형에서 후보모형 탐색 및 적합모형 선정을 위해 SAS 툴이 제공하는 proc hpfdiagnose와 proc hpgenselect를 이용하였다. 최근의 통계분석 소프트웨어(SAS)는 최적의 모형 제안을 위해 여러 모형에 대한 분석·비교 과정을 표시한 후, 최종 모형을 제안해 주는 자동화 기능을 탑재하고 있다.

머신러닝(신경망)을 이용한 예측모형 추정에는 가용 데이터에 대한 훈련세트와 검증세트 분할작업이 필요하다. 본 분석에서는 훈련세트와 검증세트 자료개수 비율을 8대 2로 설정하였다.

〈표 6〉 자료유형별 예측방법

구 분		초음파 청구자료	건강보험 전체 청구자료
연속형	회귀모형	-다항 추세모형	
	시계열모형	-자기회귀오차 -지수평활 -ARIMA	-ARIMAX
	머신러닝 (신경망)	-머신러닝(신경망)	-머신러닝(신경망)
카운트형	일반화선형모형 (GLM)	-GLM 모형 (음이항 분포)	
	머신러닝 (신경망)	-머신러닝(신경망)	

다. 예측 정확도 측도

예측력의 성능을 판정하기 위해 아래와 같이 2가지 측도를 이용하였다. 정확한 예측이 목적이라도, 어떤 측도를 사용하느냐에 따라 최종 선택되는 예측방법이 달라질 수 있다. 본 보고서에 이용한 예측정확도 측도를 기술해 보면

아래와 같다.

만일 현재까지 관측된 자료가 $\{Y_t, t=1, \cdots, n\}$이 있고, 시점 t 에서의 1시차 이후의 예측값을 $\hat{Y}_t(1)$이라고 한다면, 1시차 후의 예측오차는

$$\hat{e}_t(1) = Y_{t+1} - \hat{Y}_t(1)$$

라고 정의된다. 위의 예측오차를 이용하여 2가지 측도를 정의하면 다음과 같다.

1) 평균 제곱근 예측오차 RMSE(root mean square prediction error)

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{m} \hat{e}_{n-1+t}(1)^2}{m}}$$

여기서, m은 총 예측횟수를 나타낸다.

2) 평균절대백분위예측오차 MAPE(mean absolute percentage prediction)

$$\text{MAPE} = \frac{100}{m} \sum_{t=1}^{m} \left| \frac{Y_{n+t} - \hat{Y}_{n+t-1}(1)}{Y_{n+t}} \right|$$

여기서, m은 총 예측횟수를 나타낸다.

라. 예측모형 적합도 측도

주어진 관찰값을 토대로 예측모형을 선정한 경우, 해당 모형이 관찰값들을 어느 정도로 우수하게 설명할 수 있는지를 평가할 수 있다, 이를 위한 평가측도로 모형적합도 지표인 R^2 점수와 AIC를 이용하고자 한다.

1) R^2 (결정계수)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

이는 관측치로부터 추정된 회귀선이 원 자료를 얼마나 설명해 주고 있는가를 나타내는 지표이다. 여기서, SSR은 추정되는 회귀선으로 설명되는 부분이며, SST는 총변동을 나타낸다. 따라서 R^2 값이 1에 가까울수록 설명력이 높다고

판단할 수 있다.

2) AIC (Akaike Information Criterion)

$$AIC = -2\ \ln(L) + 2k$$

여기서 −2 ln(L)은 모형의 적합도를 나타내는 우도함수와 관련된 값이고, k는 모형에 포함 변수의 개수를 의미한다. 변수의 개수가 많은 모형은 적합도 측면에 유리하게 되는 것을 상쇄시키기 위해 변수 개수가 증가할수록 페널티를 부여하는 방식으로 구성되었다. 따라서 모형의 적합도를 높이고, 변수의 개수 k를 최소화 하는 최적의 모델을 찾기 위한 목적으로 AIC 값이 작을수록 적합도가 높은 모형이라고 판단할 수 있다.

3. 초음파 급여비 사례분석

가. 개요

건강보험 청구자료는 보건의료 정책 시행 여부에 따라 청구경향이 크게 변화한다. 이러한 자료 특성이 반영된 초음파 급여비 청구자료를 대상으로 통계적 예측방법을 적용하여 성능을 비교 분석하고자 하였다.

보건복지부는 건강보험 보장성 강화대책('17년 8월)의 후속 조치로 2018년 4월부터 상복부 초음파 보험적용을 전면 확대하였다. 간, 담낭, 담도, 비장, 췌장, 질환이 있거나 의심되어 진료 의사의 의학적 판단에 따라 시행한 경우 보험급여 대상이 된다. 상복부 진단초음파 검사(정밀, 일반)와 단순초음파 검사가 이에 해당된다. 질환 증상이 의심되거나, 최초 검사 후 경과관찰이 필요하여 시행한 경우 필수급여로 적용한다. 이외에 초음파 검사 이후 특별한 증상변화나 이상이 없는 데 추가검사를 하는 경우는 본인부담률을 높여 선별급여(예비급여)로 적용한다.

상복부 초음파 급여화로 인한 재정 소요는 연간 2400억 원으로 예상되며, 향후 불필요한 초음파 검사가 증가하지 않도록 의료기관 적정성 평가를 실시하고, 노후 중고장비 등 질 낮은 장비에 대한 관리를 강화하는 계획이

추진될 예정이다.

보장성 강화 로드맵에 따라, 상복부 초음파 보험적용을 시작으로 단계적으로 모든 초음파 검사에 대해 보험 적용을 확대할 계획이다.

이에 따라, 초음파 검사 급여비용의 급증에 대비하여 이용량 모니터링을 강화하고, 미리 급여비를 예측하여 건강보험 재정을 안정적으로 관리해야 한다.

〈표 7〉 보험적용 전후 환자부담금 변화

구분		의원	병원	종합병원	상급종합
보험적용이전	최소~최대	4~10만 원	5~12만 원	8~16만 원	10~20만 원
	평균	6만1000원	8만4000원	10만4000원	15만9000원
보험적용이후	외래	2만8600원	3만6000원	4만6900원	5만8500원
	입원	1만9100원	1만8000원	1만8700원	1만9500원

자료원: 대한초음파의학회

〈표 8〉 보장성 강화 주요 정책, '18년~'20년

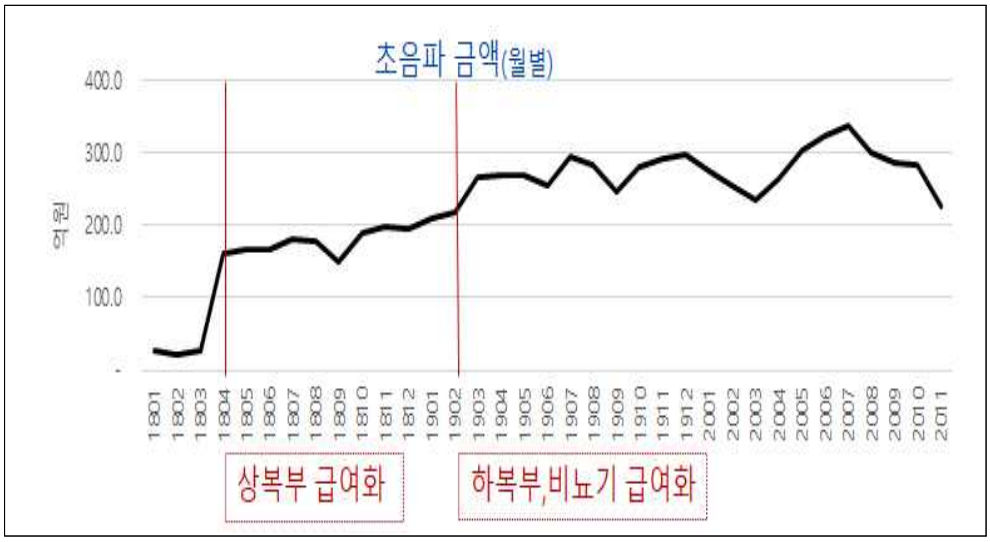
진료 시점	'18년				'19년				'20년			
	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q
상급병실 급여화			(2 · 3인실)									
						상급병실 손실보상-신생아/소아중환자실 전담전문의 가산						
노 인			틀니 본인부담률▼									
	외래정액제개선											
여 성	난임시술						난임시술 기준확대					
응급수가												
초음파		상복부 초음파 급여화			하복부비뇨기 초음파 급여화		남성생식기 초음파 급여화		여성생식기초음파 급여화			
정신질환			정신치료 수가개편▲									

본 분석에 이용된 초음파 데이터의 수가코드는 〈표 9〉와 같다. 분석대상은 진료월 기준 2018년 4월부터 2021년 3월까지(심사결정월 기준: 2021년 5월까지)의 자료들이다. 자료 개수기준에 따라 예측모델의 성능을 분석하기 위해서 12개월, 24개월, 30개월로 구분하여 자료세트를 구축하였다. 연속형변수 분석에는 초음파 진료비 금액을 사용하였고, 카운트형 변수 분석에는 초음파 시행횟수를 이용하였다.

〈표 9〉 초음파 수가코드 목록

분류번호	코드	분류
나-944		복부
		가. 복부초음파
		(1) 간, 담낭, 담도, 비장, 췌장
	EB441	(가) 일반
	EB442	(나) 정밀
	EB443	(2) 충수
	EB444	(3) 소장, 대장
	EB445	(4) 서혜부
	EB446	(5) 직장, 항문
	EB447	주. 항문 초음파만 시행한 경우
		나. 비뇨기계 초음파
	EB448	(1) 신장, 부신, 방광
	EB449	(2)신장, 부신
	EB450	(3) 방광

보장성 강화정책이후 초음파 청구양상을 보면, 상복부 급여화(2018년 4월) 이후 급격히 증가한 것으로 나타났다. 청구 급여비는 서서히 상승하였고, 하복부 급여화(2019년 2월) 이후 다시 증가한 것으로 나타났다. 이후에는 상승과 하강을 반복하면서 변동 폭이 넓어진 것으로 보인다.



[그림 9] 초음파 진료비 청구 추이

나. 분석 결과

연속형 자료 대상 예측방법별 성능을 비교 분석한 결과, 머신러닝(신경망) 기법이 가장 우수하였고, 자기회귀오차 및 ARIMA 방법이 우수한 것으로 나타났다. 머신러닝(신경망) 기법은 자료 개수가 많아질수록 예측오차가 작아지는 경향을 보였으며, 자기회귀오차, 지수평활 및 ARIMA의 경우에는 일부를 제외하고 자료 개수가 많아지면 예측오차가 감소하였다.

카운트형 자료에서도 머신러닝(신경망)기법이 우수한 예측성능을 보여주었고, 일반화선형모형(GLM)의 경우 일부 경우를 제외하곤, 자료의 개수가 많아질수록 예측오차가 작아지는 경향이 확인되었다.

〈표 10〉 예측방법별 평균 절대백분위 예측오차(MAPE, %)

예측모형		성능측도	자료 개수		
			12개월	24개월	30개월
연속형	회귀모형	MAPE	29.52	36.37	33.09
	자기회귀 오차	MAPE	7.95	17.36	5.29
	지수평활	MAPE	6.47	16.13	6.01
	ARIMA	MAPE	7.68	17.65	5.67
	머신러닝 (신경망)	MAPE	6.79	6.03	3.35
카운트형	GLM 음이항분포	MAPE	8.11	5.07	11.40
	머신러닝 (신경망)	MAPE	6.59	6.40	5.01

〈표 11〉 예측방법별 평균 제곱근 예측오차(RMSE)

예측모형		성능측도	자료 개수		
			12개월	24개월	30개월
연속형	회귀모형	RMSE	8,173,902,772	10,700,386,946	10,628,624,739
	자기회귀 오차	RMSE	2,527,292,740	5,759,925,582	2,113,416,514
	지수평활	RMSE	2,136,896,587	5,471,534,711	2,356,972,957
	ARIMA	RMSE	2,452,902,595	5,856,316,857	2,240,336,100
	머신러닝 (신경망)	RMSE	2,238,135,033	2,118,312,695	1,549,916,122
카운트형	GLM 음이항분포	RMSE	38,534	31,026	61,538
	머신러닝 (신경망)	RMSE	33,001	35,001	28,877

예측에 이용된 예측모형별 적합도를 보면, 동일한 자료 개수 조건에서 지수평활 및 ARIMA 모형의 적합도가 우수하였다. 머신러닝(신경망)의 적합도는 훈련세트와 검증세트의 배분 비율에 따라 변동이 발생하므로 타 모델과의 비교는 적절하지 않다.

〈표 12〉 예측방법별 모형 적합도

예측모형		적합도	자료 개수		
			12개월	24개월	30개월
연속형	회귀모형	R^2	0.8610	0.8793	0.8883
	자기회귀 오차	AIC	879.6754	1164.7559	1457.2492
	지수평활	AIC	514.5254	1036.6869	1298.0271
	ARIMA	AIC	504.0421	1056.7130	1334.2510
	머신러닝 (신경망)	R^2	-0.3028	0.5478	0.3738
카운트형	GLM 음이항분포	AIC	281.4395	573.3925	717.8336
	머신러닝 (신경망)	R^2	-141.2848	0.8176	0.6973

주) 머신러닝(신경망) 모형에서 R^2 값이 음수이면, 적합 모형이 데이터에 맞지 않음을 의미하는데, 이 경우 예측성능은 일반평균을 예측값으로 산출하는 것보다 더 나쁨.

4. 건강보험 총진료비 사례분석

가. 개요

한국의 보건의료지출은 미래에 급격하게 증가할 것으로 예상되며, 특히, 건강보험은 보건의료지출에서 차지하는 비중이 크므로, 재정지속성 측면에서 관리해야 한다. 이를 위해 미래의 건강보험 지출을 예측하고, 그 예측치를 목표로 하여 지속적으로 관리해야 한다.

건강보험 지출 예측은 건강보험에 영향을 주는 요인(독립변수)을 고려하고, 시계열적인 특성을 감안하여야 한다. 즉 외부요인의 동태적인 특성을 반영한 장단기 예측모형설정이 필요하다.

본 분석은 1990년도부터 2016년까지 건강보험 심사결정진료비를 대상으로 하였다. 심사결정진료비에 영향을 주는 요인으로 시간, 소비자물가지수, 1인당 GDP(국내총생산)를 포함하였다.

적절한 시계열 모형을 수립하기 위해서 로그변환을 자주 사용하는데, 본 분석에서도 모형의 적합도 수준을 높이기 위해 변수변환을 시도하였다. 분석모형으로 ARIMAX 와 머신러닝(신경망)을 이용하였고, 성능측도는 MAPE를 이용하였다.

나. 분석 결과

연속형 자료 대상으로 외부요인을 고려한 예측방법의 성능을 비교 분석한 결과, ARIMAX의 성능이 머신러닝(신경망)기법보다 우수한 것으로 나타났다. 다만 머신러닝의 경우, 외부요인에 포함되는 변수의 개수가 많아질수록 성능 개선이 이루어지는 것으로 나타났다. 따라서, 예측하고자 하는 대상의 독립변수(영향요인) 정보가 많아질수록 예측성능이 향상될 것으로 기대된다.

로그변환 적용 이후 예측방법을 보면, ARIMAX에서 로그모형의 예측성능이 우수한 것으로 확인되었고, 머신러닝(신경망)의 경우에는 예측성능이 저하되는 것으로 확인되었다.

〈표 13〉 예측방법별 평균 절대 백분위 예측오차(MAPE)

예측모형		성능측도	독립변수	
			GDP	시간(t^2), GDP
ARIMAX	일반모형	MAPE	4.10	
	로그모형	MAPE	3.45	
머신러닝 (신경망)	일반모형	MAPE	30.08	14.50
	로그모형	MAPE	100.00	87.73

〈표 14〉 예측방법별 평균 제곱근 예측오차(RMSE)

예측모형		성능측도	독립변수	
			GDP	시간(t^2), GDP
ARIMAX	일반모형	RMSE	4,335,909,631	
	로그모형	RMSE	2,944,627,190	
머신러닝 (신경망)	일반모형	RMSE	19,620,241,199	7,198,530,378
	로그모형	RMSE	78,057,956,511	69,685,758,461

모형 적합도 기준에서 비교해 보면, ARIMAX 방법은 로그모형의 적합도가 우수한 것으로 확인되었고, 머신러닝(신경망) 방법은 일반모형의 적합도가 우수한 것으로 확인되었다.

〈표 15〉 예측방법별 모형 적합도

예측모형		적합도	독립변수	
			GDP	시간(t^2), GDP
ARIMAX	일반모형	AIC	1,160.2340	
	로그모형	AIC	-54.2573	
머신러닝 (신경망)	일반모형	R^2	0.8934	0.9968
	로그모형	R2	-37.9755	0.1080

주) 머신러닝(신경망) 모형에서 R^2 값이 음수이면, 적합 모형이 데이터에 맞지 않음을 의미하고, 이 경우 예측성능은 일반평균을 예측값으로 산출하는 것보다 더 나쁨.

〈표 16〉 분석대상 변수별 자료 구축, 1990~2019년

	연도	건강보험 심사진료비(천원)	소비자물가지수	1인당GDP(달러)
1	1990	2,941,905,848	40.564	6,608
2	1991	3,200,995,792	44.350	7,634
3	1992	3,731,063,384	47.106	8,125
4	1993	4,348,201,320	49.367	8,884
5	1994	4,897,171,778	52.460	10,383
6	1995	6,144,220,034	54.811	12,569
7	1996	7,623,960,900	57.510	13,398
8	1997	8,803,894,793	60.063	12,401
9	1998	9,964,955,253	64.576	8,297
10	1999	11,705,694,586	65.101	10,667
11	2000	13,140,959,342	66.572	12,261
12	2001	17,819,469,989	69.279	11,563
13	2002	19,060,635,812	71.193	13,164
14	2003	20,533,558,654	73.695	14,669
15	2004	22,355,887,341	76.341	16,506
16	2005	24,796,775,728	78.444	19,399
17	2006	28,557,969,399	80.202	21,727
18	2007	32,258,974,677	82.235	24,088
19	2008	35,036,562,324	86.079	21,340
20	2009	39,429,565,295	88.452	19,152
21	2010	43,657,027,651	91.051	23,083
22	2011	46,076,036,175	94.717	25,100
23	2012	48,234,935,369	96.789	25,458
24	2013	50,742,582,327	98.048	27,178
25	2014	54,527,451,220	99.298	29,242
26	2015	58,017,032,863	100.00	28,724
27	2016	64,662,332,221	100.97	29,287
28	2017	69,627,144,460	102.93	31,605
29	2018	77,914,125,431	104.45	33,429
30	2019	85,793,843,234	104.85	31,838

제5장 고찰 및 결론

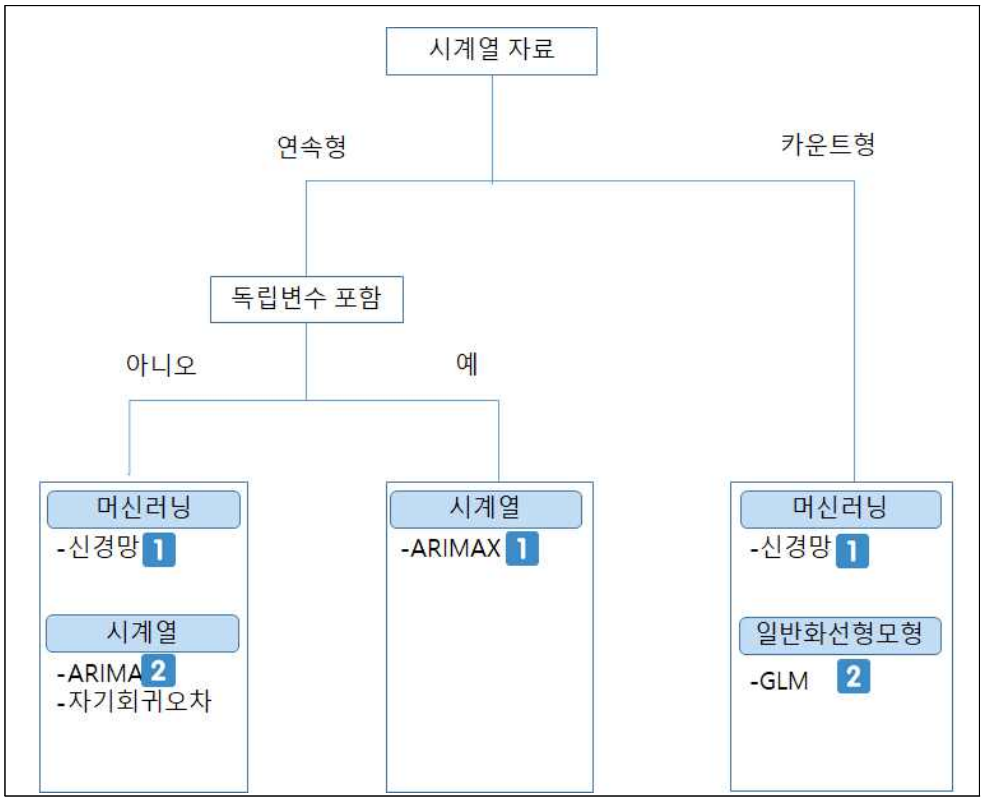
1. 고찰

가. 연구 요약

본 연구는 보건의료 분야에서 활용되고 있는 예측방법을 조사하고, 자료유형 및 외부변수 유무에 따른 적절한 예측방법을 제시하고자 하였다. 다양한 환경에서의 예측치 산출값은 보건의료 정책 효과평가 및 목표치 선정 근거자료로 활용되므로 정교한 예측치 산출을 위한 방법론 개발 및 활용이 중요하다.

통계적 예측방법 검토 과정은 문헌 및 연구보고서 조사를 토대로 다빈도 예측방법 조사, 예측방법의 업무활용 현황 조사, 그리고 건강보험 청구자료 분석을 통한 예측방법별 성능 비교 등의 순서로 진행되었다. 다빈도로 활용되는 예측방법은 시계열 모형이었고, 최근에 머신러닝 기법 등 인공지능 방법을 활용한 예측도 빈번하게 이용되었다. 본 연구는 회귀모형, 시계열 모형, 머신러닝(시계열) 기법을 통계적 예측방법 후보 목록으로 선정하고, 이들 간의 성능을 비교 분석하였다. 통계 및 인공지능 분야의 전문가 의견을 수렴하고, 실제 업무 담당자 조언을 반영하여 자료유형별 분석시나리오를 작성하였고, 이를 기반으로 예측성능 비교 분석을 수행하였다. 단변량 예측에서 머신러닝(신경망)기법 성능이 가장 우수하였고, 자기회귀모형 및 ARIMA 모형도 성능이 비교적 우수하였다. 자료 개수가 30개 이상인 자료에서 최적의 성능을 나타냈으며, 특히, 머신러닝 기법에서 자료 개수에 비례하여 예측성능이 향상되었다.

결론적으로, 본 연구는 자료유형별 통계적 예측방법 선택과 관련한 제안을 아래와 같이 제시하고자 한다.



[그림 10] 자료유형별 통계적 예측방법 제안

나. 연구 고찰

기존 사례에서 머신러닝 기법이 적용된 연구는 대부분 분류형 문제를 해결하기 위한 것이었다. 본 연구는 미래시점의 값을 예측하기 위한 회귀형 문제를 다룸으로써 기존 연구와의 차별화를 시도하였다. 예측방법으로 머신러닝 기법과 시계열 모형, 회귀모형 등 통계적 방법 간의 예측성능을 비교·분석하였다. 기존 연구에서 시계열 모형 예측에 필요한 자료 개수를 30~50개로 제안하고 있는데, 본 연구에서도 12개, 24개, 30개 자료에서 성능을 분석한 결과, 30개 수준에서 최적의 성능이 나타남을 확인하였다. 머신러닝의 경우에도 자료의 개수가 증가할수록 예측성능이 향상됨을 검증하였다. 결론적으로, 연속형 변수의 단변량 예측에 필요한 자료 개수는 30개 이상이어야 함을 제안 한다.

카운트형 자료의 경우에도 머신러닝 기법의 예측력이 일반화선형모형보다 우수하였고, 자료 개수 30개 수준에서 최적의 성능이 나타남을 확인하였다.

최종 예측값은 최적의 통계적 예측방법 1개를 선택하여 산출할 수도 있지만, 또 다른 방식으로 개별 예측모형의 예측값을 각각 구하고 이들의 평균값(중앙값)을 계산하여 최종 예측값으로 활용할 수도 있다. 이 방식은 여러 상황이 반영된 안정적 예측값 산출이 가능하다는 장점이 있다.

외부요인을 고려한 예측은 ARIMAX 방법이 머신러닝(신경망) 모형보다 우수한 것으로 확인되었다. 단변량 예측에서 우수한 성능을 나타낸 머신러닝 기법이 열세를 보였는데, 이는 머신러닝 학습에 이용된 자료 개수(22개)가 적고, 반면에 추정해야 할 모수가 많아 모델 구축과정에서 효율성이 떨어진 것으로 추정된다.

본 연구의 사례분석에 활용된 건강보험 청구자료 분석 결과는 타 분석자료 대상으로 확대 해석하기에 한계점이 존재한다. 즉 예측성능이 우수하다고 확인된 예측방법이 타 자료에서도 그대로 유지될지는 알 수 없다. 하지만, 예측방법 적용 대상이 건강보험 자료로 제한된다면, 최적의 예측방법 선택 과정에 도움이 될 수 있다.

2. 결론

보건의료 정책의 효과평가 및 성과·측정 관리 분야에서 예측방법은 중요한 역할을 한다. 예측값은 보건의료 정책의 근거(목표)자료로 활용되기 때문에 정확하고, 정교해야 하며, 논리적인 과정 속에서 산출되어야 한다. 본 연구에서는 자료유형 및 자료 개수 등에 따른 예측방법을 비교 분석하고, 상황에 맞는 적합한 예측방법을 제시하였다.

합리적인 건강보험 정책 설계와 수행, 그리고 안정적인 급여비 관리를 위해, 미래 시점의 정교한 예측치 산출과정에는 최적의 통계적 방법이 필요하다. 이를 위해서는 첫째, 본 연구에서 제안한 자료유형별 적합한 예측방법을 선택하고 올바르게

사용하는 것이 중요하다. 또, 개별 상황에 적합한 예측방법을 사용하여 정책효과 평가 및 정책목표 설정 과정에 정교한 예측치를 제시할 수 있어야 한다. 둘째, 다양한 자료유형에 따른 예측방법을 제시하려면, 단순한 사례 형태가 아닌 자료 개수의 연속적 변화에 따라 각 예측방법 성능의 변화 수준을 보여주는 그래프를 제시할 필요가 있다. 또한, 분석자료 특성에 따라 예측성능의 변동이 있을 수 있기 때문에 대표적인 자료 (예시: 초음파 급여비)를 설정하여 적용 분야 및 대상을 제시할 필요가 있다. 셋째, 예측방법의 활용 범위를 확대하고 및 최신 기법을 습득해야 한다. 건강보험 정책의 효과평가를 위한 근거를 산출할 뿐 아니라, 상시 모니터링 업무에 예측기능을 응용하여 이상치 감지 및 미래 예측작업을 강화할 필요가 있다. 또한, 기술 발전이 빠르고 성능이 우수한 최신 AI기법(머신러닝 등)을 예측 업무에 적극 도입해야 하겠다.

참고문헌

강창구. 시계열 예측기법에 대한 비교 분석. Quarterly National Accounts. 2006
건강보험심사평가원. 급여정보분석시스템 사용자 매뉴얼, 2021
권신혜. 기계학습 기반의 영화 흥행예측 방법 비교: 인공신경망과 의사결정나무를. 아시아퍼시픽 저널. 2017.
길지은. 인공지능 딥러닝을 이용한 갑상선 초음파에서의 갑상선암의 재발 예측. 영상의학회지. 2019.
김동숙. 감염병 이상징후 감지시스템 모형 개발 방안. 건강보험심사평가원. 2016.
김용익. 전이함수모형을 이용한 국민의료비 예측. 보건행정학회지. 2003.
김지애. 외래약제 적정성 평가 가감지급 모형 개선 연구. 건강보험심사평가원. 2017.
김진섭. 시계열 분해 데이터를 이용한 LSTM기법 기반 항공기 수리부속 수요예측방안 연구. 경영과학. 2020.
김충영. 이동통신서비스 해지고객 예측모형의 비교 분석에 관한 연구. 경영정보학연구. 2002.
김한상. 의료이용 모니터링 지표개발 연구 보고서. 건강보험심사평가원. 2020.
김한상. 대형병원 중증도 중심 의료이용 분석 및 모니터링 체계 마련 연구. 건강보험심사평가원. 2020.
노상윤. ARMA모형을 이용한 의료급여 제도 혁신에 따른 지방재정 절감효과 분석. 한국지방재정논집. 2008.
노상윤. 의료급여 진료비추계모형과 향후 5개년 급여비지출기준선 전망. 한국재정정책학회. 2008.
문성은. 기계학습 및 딥러닝 기술동향. 머신러닝. 2016.
박기영. 머신 러닝을 이용한 경제분석. 한국경제학보. 2019.
박성배. 효과적 수요 예측 방법과 사례 삼성경제연구소. 2012.
박일수. 건강보험 중장기 재정전망 연구 국민건강보험공단. 2010.

박일수. 미래환경변화에 따른 건강보험 중장기 재정추계 연구. 국민건강보험공단. 2011.
배재용. 보건의료인력의 효율적 관리와 수급체계를 위한 의료수요예측 모형 개발. 보건사회연구원. 2019.
선정연. 이상치 탐색을 위한 통계적 방법과 활용 방안 연구 보고서. 2019. 건강보험심사평가원.
성병찬. 사회보장 재정추계 기반강화 연구. 보건사회연구원. 2014.
성병찬. Forecasting Drug Expenditure with Transfer Function Model. 한국응용통계학회. 2018.
성병찬. Arimax 모형을 적용한 건강보험지출 장기전망. 보건경제와 정책연구. 2015.
성진옥. 시계열 자료를 이용한 응급의료센터 수요예측 모델링. 아주대학교. 2010.
송재호,허향진. 시계열을 활용한 제주지역 관광객 수요 예측: 예측 모델간 비교와 유치 목표치 설정. 산경논집. 2002.
신창훈. 항만물동량 예측력 제고를 위한 ARIMA 및 인공신경망 모형들의 비교연구. 한국항해항만학회지. 2011.
오미애. 기계학습 기반 이상탐지 기법 연구 -보건사회 분야를 중심으로. 보건사회연구원. 2018.
오미애. 기계학습 기반 사회보장 빅데이터 분석 및 예측모형 연구. 보건사회연구원. 2017.
오종민,신현수,신예술,정형철. 시계열 분석을 활용한 서울시 미세먼지 예측. Journal of The Korean Data Analysis. 2017.
우해봉. 장래인구추계의 방법과 현황. 국민연금연구원. 2008.
우해봉. 인구추계 방법론의 현황과 평가. 보건사회연구원. 2018.
이삼식. 인구예측모형 국제비교 연구. 보건사회연구원. 2013.
이성경. 빅데이터 시대의 예측도구, Machine Learning 의 올바른 활용법. posri 이슈리포트. 2018.
이성우. 요양급여비용 자율점검제도 발전방안 연구 보고서. 건강보험심사평가원. 2021.
이성천. 의사인력 수급 추계 방법 서울대 보건대학원. 2004.
이태진. 건강보험 중장기 재정 추계 방법론 연구. 국회예산정책처. 2020.
이현진. 응급의료체계 운영방안 도출을 위한 시공간적 응급의료수요 예측모형 비교 연구. 2016년 추계산업공학회 발표자료. 2016.
임달오. 로지스틱 모형에 의한 고가 및 특수의료장비의 중장기 수요예측(2013).

보건산업 브리프(확산모형). 2013.
임달오. 보건산업시장 예측모형 개발 연구-의료서비스 산업규모 예측. 한국보건사업진흥원. 2012.
정용찬. 예측방법론 고찰을 통한 방송시장 전망. kisdi 이슈리포트. 2009.
정지윤 백내장 수술건수 추이예측 분석. 보건행정학회지. 2020.
정진용. ARIMA모형 신경망과 웨이블릿 분석을 사용한 시계열 예측방법 비교 연구. 연세대 석사논문. 2000.
정찬미. 영화 관객 수 예측을 위한 기계학습 기법의 성능 평가 연구. 한국전자거래학회지. 2020.
정형선, 신정우, 이준협, 정완교. 국민의료비 미래추계 구축방안. 보건복지부. 2015.
정혜린. A review of artificial intelligence based demand forecasting techniques. 한국응용통계학회. 2019.
조대헌. 카운트 데이터 기반 공간 군집 분석 연구의 동향과 방법론적 이슈. 대한지리학회지. 2013.
조민호. 마케팅 데이터를 대상으로 중요 통계예측 기법의 정확성에 대한 비교 연구. 한국전자통신학회논문지. 2019.
조상섭. 통신서비스산업 예측모형 예측력 비교 분석. 전자통신동향분석. 2002.
지대욱,박상아, 공인택,신광섭. 수요예측을 통한 다빈도 구매상품의 적정재고 수준 결정 모형 개발:공항 면세점 사례. 한국빅데이터학회지. 2020.
지한나, 김명석. 응급의료센터의 hourly data를 이용한 내원환자 수용예측. 한국지능정보시스템학회 학술대회논문집, 2013.
최필선. 머신러닝 기법을 이용한 대졸자 취업예측 모형. 직업능력개발연구. 2018.
홍동숙. AI기반 개인사업자 업종별 부도율 예측에 관한 연구. 한국신용정보원. 2020.
황선영. 월별시계열 자료 분석기법들의 비교 분석 및 변동성 모형을 이용한 단기예측력 개선 방안. Quarterly National Accounts. 2005.
Bruce H. Andrews. Building arima and arimax models for predicting long-term disability benefit application rates in the public/private sectors. University of Southern Maine 2013
John C. Chambers How to Choose the right forecasting technique. Harvard Business Review 1971
Leshno,M., Ya, V., Pinkus, A., and Schocken,S. Multilayer feed forward

networks with a nonpolynomial activation function can approximate any function. Neural Networks, Vol.6(6) PP.861-867.

Rob J Hyndman and George Athanasopoulos. Forecating: Principles and Practives Monash University, Australia 2018

SAIGAL S. Performance comparison of time series data using predictive data mining techniques. Advances in information mining 2012

ABSTRACT

Comparison and utilization of statistical methods for data forecasting.

Healthcare data forecasting is used in a variety of fields, including policy evaluation, health insurance financial estimation, and the detection of unusual claims symptoms.

Furthermore, the evidence generated by the predictive model is critical in the development of health policies and decision-making processes.

As a result, the prediction process must be scientific and systematic, and accuracy measures should be a key selection criterion for forecasting methods.

By examining the data prediction methods used in the health care field as well as the most recent forecasting techniques, we attempted to suggest a forecasting method suitable for the characteristics of healthcare data in this study.

To that end, a literature review, investigations on the most recent forecasting methodologies, and a comparative analysis of forecasting performance based on data type were performed. And the results of the analysis were synthesized to suggest a forecasting method appropriate for the type of healthcare data.

The subject of forecasting method review included regression models, time series models, and machine learning. The data types were separated into continuous variables and count variables, and data sets of 12, 24, and 30 sizes were created.

Health insurance claim data was used to compare forecasting performance, and SAS and Python were used as analysis tools.

According to the findings, the machine learning forecasting method performed best for both continuous and count data types, while the ARIMA, time series analysis method performed reasonably well for continuous variables. Forecasting performance improved as the number of data points increased. In this study, we recommend a sample size of at least 30 subjects.

This research is anticipated to help in the selection of an appropriate forecasting method for performing complex prediction tasks.

부 록

부록 1. 예측사례 문헌 검토 - 시계열

제목(목적)	분석방법	핵심변수	변수유형
ARIMAX 모형을 적용한 건강보험 지출 장기전망	• ARIMA 모형에 독립변수까지 고려하는 '외생변수를 포함한 자기회귀 이동평균' 모형 적용 (모형의 차수는 p d q r) • 교차상관계수 • 후보모형 검토 ・ARIMAX ・전이함수 모형 ・AUTOREG 모형 • 잔차검정 ・백색잡음,포트맨토 검정 ・차분, 단위근검정 ・ADF검정 • 모형 선택기준 ・AIC, SBC ・MSE, MAE	• 건강보험 지출자체 시계열 (연속형)	• 제도가 안정화된 이후 2000~2010년까지의 11년 자료이용 • 독립변수의 동태적 특성 고려 필요 • 잔차 요인을 기존 인구 및 소득, 기타요인 구분 유형에서 65세 이상 인구 상대지출 비중, 고가 의료장비, 의약품 연구개발비, 1인당 GDP, 인구 고령화율 구분 등 상세화 • 변수 로그변환, 차분
Forecasting drug expenditure with transfer function model	• 구간별 자기회귀오차모형 (segmented autoregressive error model) *구간별 추세(a*t), 시차간 선택을 위한 후진제거법, 잔차에 대한 포트맨토검정, AIC / BIC 기준 모형선택 • 전이함수 모형 ・시계열 변수 정상화 ARIMA모형화, 사전백색화, 교차 상관성 검토, 전이함수 차수 결정, 잔차모형화 및 잔차검정	• 약품비 (연속형)	• 2007년 1월~2016년 12월 • (독립변수)약가일괄인하 시행, 의약품사용자수, 노인환자비중
미래 환경변화에 따른 건강보험 중장기 재정추계	• AR, MA, ARMA 과정을 따른 오차항을 지닌 회귀모형	• 1인당 급여비 (종별,	• 독립변수 ・실질 GDP, 1/2/3분기 더미, t(분기)

제목(목적)	분석방법	핵심변수	변수유형
연구		연령별, 입원/외래)	• 자료는 2002년 1월부터 2011년 3월까지 급여비 • 2012년 이후 예측 시뮬레이션은 수가동결, 2%, 3%인상 가정함
백내장 수술건수 추이 예측 분석	• 삼중지수평활법 • ARIMA 모형 • 예측모형 평가지표 · RMSE, MAPE, MASE	• 백내장 수술 건수 (빈도수)	• 2006년부터 2018년까지 13개년도 백내장 수술 건수 · 예측:2019-2021년까지 추이 예측
시계열을 활용한 제주지역 관광객 수요 예측: 예측 모델간 비교와 유치목표치 설정	• 추세분석법 • 윈터스 지수평활법 (승법,가법) • 계절형 ARIMA	• 미래 관광객 수요예측 (2000년 ~2005년)	• 1960년~1999년간 연간 관광객 통계와 1976년 1월~1999년 12월의 24년간 월별 관광객
일반화선형모형을 이용한 자동차보험 요율상대도 산출방법 연구	• 일반화선형모형 (사고액 또는 교통사고 빈도, 병원방문 횟수) · 건수변수 (사고빈도) * 포아송, negative binomial, zero inflated poisson · 연속변수(사고심도,사고액) * 가우시안, log-normal, gamma, inverse gaussian 등	• 사고건수 • 사고심도 (사고당 손해액) • 보험료	• 독립변수 · 성, 연령, 가입경력, 법규위반, 할인할증률
감염병 이상징후 감지시스템 모형 개발 방안	• 해열제/소염제 · log winters method · seasonal dummy 모형 · log ARIMA(2,0,0)(1,0,0)s · linear trend with searsonal terms · AIC, SBC, SSE 기준 모형 선정	• 약제처방 건수	• 2014~2015년 의료기관에서 청구한 환자 진료내역 및 처방 자료 · 상병코드, 연령 및 지역 구분 • (주요감염병) 인플루엔자, HCV, 메르스, 지카바이러스, 바이러스 장감염
수요예측을 통한 다빈도 구매 상품의 적정재고 수준 결정 모형	• 계절형 ARIMA 모형 (중단기 시계열 예측모형) • 선형회귀방정식 (출국객 대비 판매량)	• 월별/일별 출국객수	• 월별 출국객수 ('13.1~ '17.12) • 일자별 출국객수 ('18.3~ '18.5)

제목(목적)	분석방법	핵심변수	변수유형
개발	• 모형식별 • 1차 차분, 계절차분 • ACF, PACF 형태파악 • BIC, R^2, RMSE, MAPE를 이용한 모형 선정		
시계열 분석을 활용한 서울시 미세먼지 예측	• 가변수 회귀분석 • 잔차의 자기상관여부 (더빈-왓슨 통계량) • 계절형 ARIMA • AIC, SBC 이용 모형선정 • R^2, RMSE, MSE 활용	• 미세먼지 평균농도 PM10	• 15개년도 자료 (2001~2015)
로지스틱 모형에 의한 고가 및 특수의료장비의 중장기 수요예측	• 로지스틱 성장모형 • 시간에 따른 누적 수요추이 조사자료) • 데이터가 부족한 상황에도 파라미터 안정적 추정가능	• 5종의 대표적인 고가 및 특수 의료장비	• 12개년도 자료 (2001~2012) • OECD health data (1984~2000)

부록 2. 예측사례 문헌 검토 - 머신러닝

제목(목적)	분석방법	핵심변수	변수의 기본유형
영화 관객 수 예측을 위한 기계학습 기법의 성능 평가 연구	(분류 기반) - Random Forest classifier - Support Vector Machine (회귀모형 기반) - Random Forest Regressor - k-NN Regressor	-종속변수: 관객수 (개봉후 3주차 누적) -입력변수: 배우1의 영향력점수, 감독/제작사/배급사의 영향력점수, 스크린수, 상영횟수,1주차2주차 관객수	영화의 관객수 -1)연속형 종속변수를 위한 회귀모형 기반 기계학습모형과 2)범주형 종속변수 대상의 분류기반 기계학습 모형
항만물동량 예측력 제고를 위한 ARMA 및 인공신경망 모형 비교	- ARIMA - Artificial Neural network(Genetic Algorithm 적용) - Hybrid model	-항만 컨테이너 물동량 (수입, 수출 물동량)	-1991년~2006년까지 월별자료(192개)
머신러닝기법을 이용한 대졸자 취업예측모형	-랜덤포레스트 기법	-대졸자의 취업여부 -취업의 질 (정규직 / 비정규직)	-대졸자 15,000 명을 조사한 GOMS 2014 데이터 -인구통계학적특성, 가족특성,졸업대학, 대학생활,외국어,취업지원프로그램,교육훈련, 자격증, 정부 고용정책 참여 등 96개 변수
기계학습 기반의 영화흥행 예측 방법 비교: 인공신경망과 의사결정나무 중심으로	- 인공신경망 기법 - 의사결정나무 기법	-전국 총 관객수 -감독파워, 배우파워,제작국가, 장르, 관람등급, 러닝타임, 스크린 수, 전문가평가, 관객평가, 첫 주 전국 주말 관객수	-2004년 1월부터 2014년 12월까지 기간 중 1100편 영화 수집 -Weka ver3.8 활용

부록 3. 머신러닝 소프트웨어 사용 환경

구 분		설 명
S/W 종류	파이썬 (Python)	- 플랫폼 독립적이기 때문에 다양한 플랫폼에서 사용 가능하고 또한 기본 제공되는 라이브러리가 매우 많음. -인터프리터식 동적 타이핑(Dynamically typed) 대화형 언어이고, 작성한 프로그램을 바로 실행할 수 있어 사용자가 쉽게 결과 확인 가능.
	R 스튜디오	- 통계 컴퓨팅, 그래픽스를 위한 프로그래밍 언어인 R을 위한 자유-오픈 소스 통합 개발 환경임.
	쥬피터노트북	- IPython으로 부터 파생된 주피터 노트북은 여러 개의 언어를 통한 인터랙티브 실행 환경을 지원함.(Julia, Python, R)
설치 방법	파이썬 및 R	- 다음의 설치화일을 윈도우 컴퓨터에 설치. . python-3.8.9-amd64 . R-4.1.0-win
	쥬피터노트북	- 윈도우즈 PowerShell을 실행 후 다음 명령어 수행. .pip install notebook *연결오류 문제발생시 아래와 같이 환경 설정 후 다시 실행. . pip --trusted-host pypi.org --trusted-host files.pythonhosted.org install <라이브러리>
	R 과 쥬피터 노트북 연결	- R studio를 실행 후, 다음 명령어 수행. . install.packages('IRkernel') . IRkernel::installspec() . IRkernel::installspec(user = FALSE)
분석 환경	쥬피터노트북 구동	- 윈도우즈 PowerShell을 실행 후 다음 명령어 수행. . jupyter notebook
	쥬피터노트북 환경 설정	- 윈도우즈 PowerShell을 실행 후 다음 명령어 수행. . jupyter notebook --generate-config
	쥬피터노트북 실행오류	- 컴퓨터 재부팅시 *.html 파일이 삭제된 경우 다음 조치 수행. . /Lib/site-packages/notebook/template/*.html 관련 백업본을 미리 만든 후, 동일 폴더에 복사 붙여넣기 수행.
	라이브러리설치	- 윈도우즈 PowerShell을 실행 후 다음 명령어 수행. . pip install 라이브러리

데이터 예측을 위한 통계적 방법 비교 및 활용

발 행 일 : 2021년 12월
발 행 인 : 김선민
편 집 인 : 이진용
발 행 처 : 건강보험심사평가원 심사평가연구소
강원도 원주시 혁신로 60(반곡동)
대표전화 : 1644-2000
홈페이지 : www.hira.or.kr