The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.1.145
29(1), 145–155

Comparison of forecasting models of disease occurrence due to the weather in elderly patients

Seonjae Lee[a] · In-Kwon Yeo[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received December 15, 2015; Revised December 23, 2015; Accepted December 23, 2015)

Abstract

In this paper, we compare forecasting models for disease occurrences in elderly patients due to the weather. For the analysis, the medical data of aged patients released from Health Insurance Review and the weather data of the Korea Meteorological Administration are weekly and regionally merged. The ARMAX model, the VARMAX model and the TSCS regression model are considered to analyze the number of weekly occurrences of some diseases attributable to climate conditions. These models are compared with MSE, MAPE, and MAE criteria.

Keywords: ARMAX model, HIRA data, meteorological data, TSCS regression model, VARMAX model

1. 서론

소득 수준의 향상, 의학 기술의 발달에 의한 수명 연장과 더불어 출생률의 저하는 인구구성에서 고령자의 비율을 높이는 원인이 되고 있다. 65세 이상 인구가 총인구의 7% 이상이면 고령화 사회, 14% 이상이면 고령 사회, 20% 이상이면 후기 고령 사회 혹은 초고령 사회라 한다. 통계청 자료에 의하면 2015년 현재 전 세계에서 65세 이상 인구는 총인구의 8.2%를 차지하고 있다고 한다. 우리나라의 65세 이상 인구 비율은 13.1%로 고령화 사회라고 할 수 있으며 곧 고령 사회가 되는 것뿐만 아니라 증가 속도가 빨라 초고령 사회로의 진입도 멀지 않은 것으로 예상된다. 따라서 얼마 남지 않은 고령 사회와 초고령 사회를 대비하기 위해 고령 인구에 대한 통계적 연구가 필요하다. 이를 통해 고령 사회에서 발생할 수 있는 문제를 해결하기 위한 선제적 정책 마련은 물론 실버산업 등에서 운용될 수 있는 새로운 패러다임을 제시할 것이라 기대된다. 이러한 맥락에서 이 논문에서는 건강보험심사평가원(이하 심평원)에서 공개한 고령환자의 의료자료를 분석하고자 한다.

우리나라 고령 환자와 관련된 연구는 Lee와 Park (2015), Lee와 Hwang (2015), Kim 등 (2015) 등을 포함하여 다양한 분야에서 이루어져 왔지만 대부분의 경우 특정 질병에 초점이 맞추어져 부분적으로 진행되어 왔다. 심평원에서 2015년 7월까지 약 3천억 건에 해당하는 보건의료 빅데이터를 공개하였다. 지금까지 심평원의 빅데이터는 건강보험 사기 예측 모형을 개발하거나, 기술통계량을 이용한 의료서비스 개편, 의학계에서의 개별적인 질병 연구 등에 사용되었으며 앞으로도 폭 넓은 연구들이 진행될 것으로

This research is supported by grants from Sookmyung Women's University (1-1503-0123).
[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Chengpa-ro 47-gil 100, Yongsan-Gu, Seoul 04310, Korea. E-mail: inkwon@sm.ac.kr

Table 2.1. Primary and secondary variables of data sets

데이터셋	연결변수	주요변수
명세서 일반내역	명세서 연결코드, 요양기관 고유번호	연령군, 성별, 질병코드, 진료과목코드, 요양개시일자, 요양만료일자, 내원일수, 심결요양급여비용총액, 심결본인부담금, 심결보험자부담금
진료내역	명세서 연결코드, 분류유형코드	분류코드, 단가, 1회투약량, 1일투약량, 1일투여량 실시횟수, 총투여일수 또는 실시횟수, 총사용량 또는 실시횟수, 금액, 가산적용금액, 일반명코드
상병	명세서 연결코드	상병진료과목코드, 청구상병기호, 청구진료과목코드
처방전 상세내역	명세서 연결코드	분류유형코드, 단가, 1회투약량, 1일투약량, 금액, 일반명코드 총투여일수 또는 실시횟수, 총사용량 또는 실시횟수
요양기관 현황	요양기관 고유번호	요양기관 종별코드, 설립구분, 시도코드, 병상수준

예상된다. 본 연구에서는 기후와 관련된 질병을 알아보고 질병 발생빈도를 예측할 수 있는 모형들에 대해 알아본다. 금융 또는 경제 시계열 자료분석에 많이 활용되고 있는 ARMAX모형, VARMAX모형, TSCS회귀모형을 의료 자료에 적용시키고자 한다. 이들 모형을 근거로 해당 고령 환자의 질병발생빈도에 대한 예측모형을 유도하고 평균제곱오차, 평균절대백분위오차, 평균절대오차를 이용하여 모형을 비교하고자 한다. 고령 환자 수에 대해 시의적절하게 예측을 할 수 있다면 보건당국이나 병원에서는 선제적으로 의료 조치나 서비스를 준비할 수 있으며 제약회사의 경우 해당 약품의 생산량 조절 등 효율적인 재고 관리를 할 수 있을 것으로 예상된다.

2. 분석자료

2.1. 심평원자료

이 논문에서 분석하고자 하는 자료는 건강보험심사평가원 고령환자데이터셋인 HIRA-APS-2012-0057과 HIRA-APS-2013-0059이다. 전체 데이터셋에는 다음과 같은 5개의 세부데이터셋으로 나누어져 관계형 데이터베이스 형태로 구성되어 있다.

1. 명세서 일반내역: 수진자의 일반적인 특성 포함.
2. 진료내역: 환자들이 외래 또는 입원 시 발생하는 진료행위와 원내처방이 된 약제정보 등에 대한 정보 포함.
3. 상병내역: 환자들의 모든 진단명 정보 포함.
4. 원외 처방전 상세내역: 원외처방으로 이루어지는 모든 약제에 대한 정보 포함.
5. 요약기관 현황: 환자가 진료를 받은 요양기관의 정보 포함.

각각의 데이터셋은 환자와 요양기관 코드가 독립적으로 할당되어 연도별로 데이터를 매칭하는 것이 불가능하도록 하여 개인정보를 보호하고 있다.

Table 2.1은 이 데이터셋의 주요 변수를 정리한 것이다. 여기서 연결변수는 데이터셋 간의 관계를 연결시켜 주는 변수로 코드화되어 있다.

이 데이터셋은 원 자료에서 개인 및 법인에 대한 정보를 제거한 후 통계학적으로 표본 추출된 2차 자료로 연도별로 의료 서비스를 이용한 모든 환자 중 20% 정도를 성별과 연령(5세 단위)에 따라 층화추출한 것이다. 연도별로 표본을 독립적으로 추출했기 때문에 환자나 요양기관에 따른 매칭이 불가능 하도록

하여 개인 정보를 보호하고 있다. 명세서 일반내역서의 수진자고유번호는 주민번호를 대체한 일련 번호로 2012년 자료의 경우 1부터 1107132까지, 2013년의 경우 1부터 1161198까지 부여되었다. 즉, 이 데이터셋에는 2012년 1107132명의 의료 자료가, 2013년 1161198명의 의료 자료가 저장되어 있다. 이는 의료 서비스를 받은 고령 환자가 2012년에는 554만명, 2013년에는 581만명 정도 되는 것을 의미한다. 실제로 건강보험에 가입한 고령 인구는 2012년에는 547만, 2013년은 574만명이 되는 것으로 조사되었다. 또한 이 데이터셋에는 2012년 38,579,420건, 2013년 40,290,746건의 명세서가 포함되어 있는데 추출률 20%를 단순 고려하면 2012년에는 약 1억9천만 건, 2013년에는 약 2억 건의 고령환자 명세서가 있는 것을 추정할 수 있다. 이 자료에는 해당연도에 요양개시일 기준으로 1년간 청구된 진료내역과 처방내역이 저장되어 있다.

2.2. 전처리 과정

기후와 질병의 관계를 분석하기 위해 2012년, 2013년 전국 16개의 시도의 90개 기상관측소에서 측정된 기상 자료를 기상청 국가기후데이터센터에서 추가적으로 다운로드 하였다. 이 논문의 목적은 환자 발생빈도에 대한 예측이기 때문에 기상정보 중 예보가 가능하고 질병 발생에 대한 설명력이 높을 것으로 생각되는 평균기온과 일교차를 분석에 사용하였다. 평균습도의 경우 평균기온과 상관관계가 높아 다중공선성의 문제가 발생하여 설명변수에서 제외시켰다.

일별 분석의 경우 변동성이 크고 월별 분석의 경우 시의성의 떨어질 뿐만 아니라 분석 자료의 크기가 작아지는 문제가 있어 이 연구에서는 환자 데이터셋과 기상 데이터셋을 주별 단위로 재구성 하였다. 또한 세부적인 분석을 위해 전국을 경기, 강원, 충청, 호남, 영남의 5개 권역으로 나누어 자료를 추가 구성하였다. 각 시도별 기상자료는 해당 관측소의 평균자료를, 권역별 기상자료는 해당 시도의 평균 자료를, 전국의 기상자료는 시도별 자료의 평균 자료를 적용하였다.

이 논문에서는 특정 질병 환자 발생건수와 기후와의 관계를 ARMAX, VARMAX, TSCS회귀모형을 이용하여 분석하였다. 이를 위해 모형에 따라 분석 자료를 다르게 구성했는데 VARMAX는 5개 권역별 질병 발생빈도와 기상변수의 전국평균의 자료를, ARMAX와 TSCS회귀모형은 5개 권역별 질병 발생빈도와 해당 권역의 기상변수의 자료로 데이터셋을 만들었다.

자료분석을 위한 세부 전처리과정은 다음과 같다.

- 심평원 자료는 명세서 기준으로 구성되어 있다. 그러므로 같은 날짜의 특정 질병에 대해 명세서가 여러 개 있을 수 있어 질병 발생빈도가 과대추정 될 수 있다. 이를 방지하기 위해 자료를 환자 기준으로 재구성하고 같은 날짜에 동일 질병의 자료가 여러개 있는 경우 하나로 통합했다. 환자별로 같은 날짜에 여러 가지 질병이 있을 수 있다.
- 심평원 자료에는 환자의 거주 지역이 포함되어 있지 않다. 기상자료와 환자가 거주하는 지역을 매칭시키기 위해 명세서에 기록된 요양기관의 시도코드를 이용하여 환자의 거주 권역을 추정하였다. 환자의 명세서가 해당연도에 한 권역에서만 발생했다면 이 환자는 해당권역에 거주하는 환자일 가능성이 높다. 만약 두 권역 이상에서 발생한 경우에는 다음과 같은 기준으로 거주 권역을 추정하였다.
  1. 상급종합병원의 진료기록만 있는 경우, 최빈 권역을 그 환자의 거주 권역으로 추정한다. 만약 동일한 빈도를 갖는 지역이 있다면 그 중 하나를 무작위로 선정한다.
  2. 상급종합병원을 제외한 나머지 병원 방문 기록 중 최빈 권역을 그 환자의 거주 권역으로 추정한다. 만약 동일한 빈도를 갖는 권역이 있다면 그 중 하나를 무작위로 선정한다.
- 고령 환자 발생빈도는 권역단위에 따라 주별로 몇 건이나 발생했는지 계산한다. 환자 데이터의 경우

Table 2.2. The estimated number and the proportion of elderly patients

권역	2012년		2013년	
	빈도	백분율	빈도	백분율
경기	600,752	54.26	484,832	41.75
강원	21,382	1.93	45,595	3.93
충청	95,342	8.61	133,500	11.50
호남	121,753	11.00	174,032	14.99
영남	267,903	24.20	323,271	27.84
전국	1,107,132	100	1,161,230	100

Table 2.3. The proportions of major diseases

2012년		2013년	
질병명	비율(%)	질병명	비율(%)
등 병증	14.52	등 병증	14.33
피부염 및 습진	10.69	피부염 및 습진	10.58
관절증	6.79	관절증	6.61
구강, 침샘 및 턱의 질환	4.51	구강, 침샘 및 턱의 질환	4.98
변형성 등 병증	4.15	당뇨병	4.18
당뇨병	4.11	변형성 등 병증	4.12
식도, 위 및 십이지장의 질환	3.72	식도, 위 및 십이지장의 질환	3.55
급성 상기도 감염	3.70	급성 상기도 감염	3.43
기타 급성 하기도 감염	2.61	기타 급성 하기도 감염	2.57
만성 하기도질환	2.24	만성 하기도질환	2.23

연도별로 전체 고령 인구가 다르고 5개 권역별 전체 고령 인구가 다르기 때문에 해당연도와 권역별 고령 건강보험가입자 수에 따라 표준화하여 각 질병에 대해 고령 인구 10만 명당 발생빈도를 도출한다.

- 기상자료도 전국 단위(VARMAX)와 권역 단위(ARMAX와 TSCS회귀모형)에 따라 주별 평균을 계산한다. 심평원 자료와 기상청 자료를 주/권역에 맞춰 병합하여 횡단면 시계열 데이터셋을 재구성한다.

Table 2.2는 위의 분류방식에 의해 분류된 권역별 환자 수와 비율이다.

2.3. 질병분류

한국표준질병사인분류(Korean Standard Classification of Diseases; KCD)는 우리나라에서 발생하는 질병 및 사망 자료를 유사성에 따라 유형화한 것으로 보건복지부, 대한의사협회 등의 자문을 받아 통계청이 작성하고 있다. 현재 시행되고 있는 6차 개정 분류는 2011년 1월 1일부터 시행된 것으로 분류체계는 대분류 22개, 중분류 267개, 소분류 2,093개, 세분류 12,603개, 세세분류 6,335개로 구성되어 있으며 현재 7차 개정작업을 진행 중이다. 이 논문에서는 중분류 기준에 따라 고령자에서 발생 빈도가 높은 상위 10개 질병을 정리하면 Table 2.3과 같다.

이 중에서 기상과 연관이 있을 것이라고 예상되는 급성 상기도 감염(이하 상기도 감염), 기타 급성 하기도 감염(이하 하기도 감염), 피부염 및 습진에 대해 분석하였다.

Figure 2.1은 위의 세 질병에 의해 발생한 주별 10만 명당 전국 환자 수와 평균기온을 그린 것이다. 이 그림에서 상기도 감염(Upper Respiratory Infection)과 하기도 감염(Lower Respiratory Infection)의

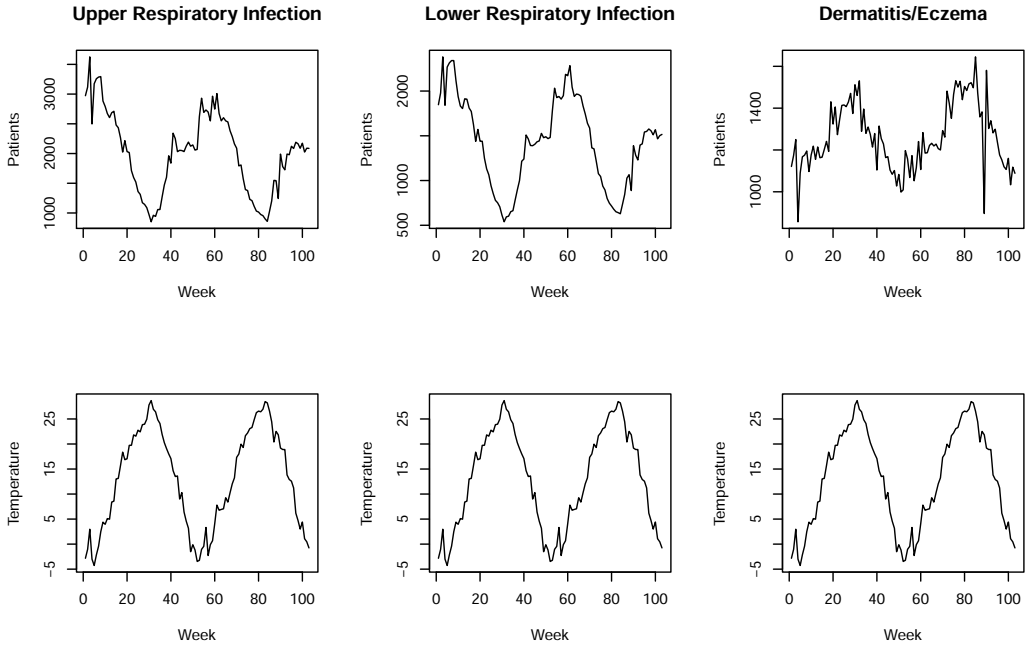

Figure 2.1. Time series plots of the number of elderly patients and temperature.

경우 호흡기 계통의 질환으로 겨울에 상대적으로 발생 빈도가 높을 것을 확인할 수 있다. 반면 피부염 및 습진(Dermatitis/Eczema)의 경우는 겨울철보다 습도가 높은 여름에 발생 빈도가 더 높은 패턴을 보이며 호흡기 계통의 질환과 상반된 패턴을 보이고 있다.

3. 분석모형

이 논문에서는 질병 발생빈도가 이전의 발생빈도와 해당 주의 기상자료에 영향을 받는다고 가정한다. 또한 연도별로 시계열 패턴은 비슷하지만 Table 2.2에서도 보는 것과 같이 빈도 수에 있어 급격한 차이를 보이고 있으므로 연도에 대한 가변수를 포함시켰다. 이에 대한 모형식을 표현하기 위해 Y_{jt}는 j번째 권역, t시점에서 관심있는 질병에 걸린 고령 환자 수라 하고, $\boldsymbol{x}_{jt} = (x_{1jt}, x_{2jt}, x_{3jt})^T$는 각각 j번째 권역 t시점에서의 평균기온, 일교차, 연도 가변수라고 한다. 기상에 따른 고령 환자의 질병 발생빈도 예측을 위해 아래와 같은 3가지 모형을 고려하였다.

3.1. ARMAX모형

ARMAX(Auto-Regressive Moving-Average model with eXogenous variables)모형은 ARMA모형에 외생변수를 추가한 일변량 시계열분석 모형으로 ARMA의 차수가 p, q이고 k개의 외생변수가 있는 경우 구조식은 다음과 같다.

$$\Phi_j(B)Y_{jt} = \sum_{i=1}^{k} \psi_{ij} x_{ijt} + \Theta_j(B)\varepsilon_t,$$

여기서 ε_t는 백색잡음, B는 후향연산자를 의미하고 $\Phi(B)$와 $\Theta(B)$는 AR과 MA 작용소로

$$\Phi_j(B) = 1 - \sum_{s=1}^{p} \phi_{js} B^s, \qquad \Theta_j(B) = 1 - \sum_{s=1}^{q} \theta_{js} B^s$$

이다. 이 모형으로 질병 발생빈도를 분석한다는 것은 각 권역에서의 환자 발생빈도를 별개의 모형으로 분석한다는 것을 의미한다. 고려된 외생변수는 평균기온, 일교차, 연도 가변수이므로 $k = 3$이고 전체 권역을 추정하는데 필요한 추정모수는 $5(p + q + 3)$개가 된다.

3.2. VARMAX모형

VARMAX(Vector ARMA model with eXogenous variables)모형은 VARMA모형에 외생변수를 포함시킨 모형으로 여러 개의 시계열을 동시에 분석할 수 있는 다변량 시계열 분석모형이다. VARMAX모형의 경우 동일한 외생변수를 사용하기 때문에 권역별 기상자료를 사용하지 못한다. 이런 이유 때문에 해당 시점의 기상자료는 전국평균을 사용했다. 외생변수 x_{it}를 t시점에서의 i번째 전국평균 기상자료와 연도 가변수라고 할 때 $\boldsymbol{x}_t = (x_{1t}, x_{2t}, x_{3t})^T$, $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{5t})^T$로 표시한다면 해당 시점의 외생변수만 영향을 주는 VARMAX$(p, 0, q)$모형의 일반적인 형태는 다음과 같다.

$$\boldsymbol{\Phi}(B)\boldsymbol{Y}_t = \boldsymbol{\Psi}\boldsymbol{x}_t + \boldsymbol{\Theta}(B)\boldsymbol{\varepsilon}_t,$$

여기서 $\boldsymbol{\Phi}(B)$와 $\boldsymbol{\Theta}(B)$는

$$\boldsymbol{\Phi}(B) = \mathbf{I} - \boldsymbol{\Phi}_1 B - \boldsymbol{\Phi}_2 B^2 - \cdots - \boldsymbol{\Phi}_p B^p,$$
$$\boldsymbol{\Theta}(B) = \mathbf{I} - \boldsymbol{\Theta}_1 B - \boldsymbol{\Theta}_2 B^2 - \cdots - \boldsymbol{\Theta}_q B^q$$

가 되는데 $\boldsymbol{\Phi}_j$와 $\boldsymbol{\Theta}_j$는 5×5 행렬이고 $\boldsymbol{\Psi}_j$는 5×3 행렬이다.

VARMAX모형의 경우 추정해야 할 모수의 수가 $5(3 + 5(p + q))$이므로 ARMAX모형보다 상대적으로 많은 모수를 추정해야 한다. 하지만 실제분석에서는 VARMAX의 모수에 대해 제약조건을 주어 추정모수의 수를 줄일 수 있다. 그러므로 일차분석을 통해 유의하지 않은 모수를 찾고 모형 추정에서 해당 모수의 값을 0이 되도록 제약조건을 주어 분석하면 된다.

3.3. TSCS Regression 모형

분석 자료는 시계열 특성을 가지면서 권역 간에 관련이 있는 형태를 가지고 있다. 이렇게 시계열과 횡단적 특성을 동시에 가지는 자료를 분석하는 모형으로 TSCS(Time Series Cross Sectional)회귀모형이 있다. TSCS회귀모형은 다음과 같이 표현할 수 있다.

$$Y_{jt} = \boldsymbol{\beta}^T \boldsymbol{z}_{jt} + u_{jt},$$

여기서 $\boldsymbol{z}_{jt}$는 이전 발생빈도와 기상자료를 의미하며 $\boldsymbol{z}_{jt} = (y_{j,t-1}, \ldots, y_{j,t-p}, x_{1jt}, x_{2jt}, x_{3jt})^T$가 된다. 이 모형에서는 자료들 간의 종속성을 오차항 u_{jt}의 구조로 모형화한다. 대표적인 모형은 다음과 같이 오차항이 2요인 효과를 갖는다고 가정한다.

$$u_{jt} = \nu_j + \epsilon_t + \varepsilon_{jt},$$

여기서 ν_j는 횡단면 변동효과(cross-sectional variant effect)를 ϵ_t는 시계열 변동효과(time series variant effect)를 의미하고 ε_{jt}는 백색잡음을 의미한다. 만약 변동효과를 랜덤효과로 가정하면 해당효과의

Table 4.1. Descriptive statistics of the number of elderly patients

질병	권역	평균	표준편차	최솟값	최댓값
상기도 감염	경기	2243.40	917.96	837.97	4845.58
	강원	1489.39	775.99	383.99	3190.79
	충청	1810.49	742.01	580.20	3606.60
	호남	1816.71	803.66	580.09	3813.61
	영남	1900.05	666.22	713.83	3387.88
하기도 감염	경기	1657.40	623.95	672.18	3159.33
	강원	876.50	404.99	229.94	1836.32
	충청	1261.40	513.82	410.36	2523.92
	호남	1198.10	537.80	365.18	2625.84
	영남	1322.99	499.32	459.65	2522.05
피부염 및 습진	경기	1409.02	313.65	787.07	2071.77
	강원	887.59	369.22	338.01	1594.29
	충청	1089.58	301.31	619.20	1750.69
	호남	1116.56	359.96	563.33	2095.27
	영남	1241.67	241.70	725.50	1808.03

분산에 대한 추론이 이루어지며 같은 시점 또는 권역별 자료들 간의 상관관계를 도출할 수 있게 된다. Fuller와 Battese (1974)는 이러한 랜덤효과에서의 모수 추론 방법을 소개하였다. 만약 변동효과가 고정효과이고 유의한 차이가 있다고 하면 각각의 평균 모수에 대한 추정이 이루어져야 하므로 상대적으로 많은 모수를 추정해야 하는 문제가 발생한다. 이와 별도로 Parks (1967)는 오차항이 다음과 같이 1차 자기회귀 구조를 가진다는 가정 하에서 모수추론 방법을 연구했다.

$$u_{jt} = \rho_j u_{j,t-1} + \varepsilon_{jt}.$$

Da Silva (1975)는 2요인 랜덤효과모형에서

$$\varepsilon_{jt} = \alpha_0 e_t + \alpha_1 e_{t-1} + \cdots + \alpha_m e_{t-m}$$

라는 가정 하에 추론 방법을 소개하였다. 여기서 e_t는 백색잡음을 의미한다. 이 논문에서는 Fuller와 Battese, Parks, Da Silva 방법을 모두 비교 모형으로 고려한다.

4. 분석결과

Table 4.1은 권역별 인구 10만 명당 환자 발생건수에 대한 주요 기술 통계값이다. 모든 질병에서 경기 지역의 평균 환자수가 많은 반면 강원 지역은 낮은 것으로 나타났다. 권역에 따라 환자 수가 현저히 차이 나는 것은 실제 발생건수에서의 차이에 의한 것일 수도 있고 의료시설의 차이에 의한 것일 수도 있다.

이 논문에서는 예측의 편의와 비교를 위해 모든 시계열 모형은 AR(1)으로 설정하고 외생변수가 추가되는 것으로 가정하였다. 실제 분석결과에서도 이 모형은 유의미한 것으로 나타났다. 또한 VARX모형의 일차분석 결과 유의수준 0.1을 기준으로 유의미하지 않은 계수를 0으로 제약시켜 분석하였다. 피부염/습진의 경우 모든 모형에서 일교차에 영향을 받지 않은 것으로 나타나 설명변수에서 제외시켰다.

모형 추정결과, 상기도 감염과 하기도 감염의 경우 ARX, VARX, Fuller-Battese의 모든 추정에서 평균기온이 내려갈수록 환자 수가 증가하고 일교차가 클수록 증가하는 것으로 나타나 이들 질병은 기상에 영향을 받고 있는 것으로 분석되었다. 또한 전 주 주의 환자수에 영향을 많이 받는 것으로 나타났으며

Table 4.2. Estimates of regression coefficients in the ARX(1) model

질병	설명변수	$\hat{y}_{1t}$	$\hat{y}_{2t}$	$\hat{y}_{3t}$	$\hat{y}_{4t}$	$\hat{y}_{5t}$
상기도 감염	절편	591.9	135.0	467.2	638.2	621.6
	평균기온	−13.41	−13.90	−20.52	−25.42	−19.80
	일교차	31.27	33.84	34.17	29.34	24.77
	연도	−257.9	387.8	271.8	334.4	98.48
	전주 빈도	0.729	0.657	0.608	0.607	0.659
하기도 감염	절편	358.0	112.9	234.1	320.9	310.5
	평균기온	−9.892	−7.908	−11.14	−13.42	−11.71
	일교차	19.95	17.61	19.97	15.26	16.38
	연도	−103.9	167.3	159.9	144.2	63.29
	전주 빈도	0.7714	0.6768	0.6947	0.7148	0.7407
피부염/습진	절편	1320.9	339.2	544.6	323.5	796.7
	평균기온	9.692	7.770	10.125	9.685	12.285
	연도	−454.0	524.8	402.1	324.8	325.0
	전주 빈도	0.1426	0.2213	0.2004	0.4435	0.0925

Table 4.3. Estimates of regression coefficients in the VARX(1) model

질병	설명변수	$\hat{y}_{1t}$	$\hat{y}_{2t}$	$\hat{y}_{3t}$	$\hat{y}_{4t}$	$\hat{y}_{5t}$
상기도 감염	절편	646.1	285.0	554.6	696.5	673.4
	평균기온	−14.49	−16.83	−22.43	−26.04	−20.64
	일교차	25.41	44.03	42.74	40.53	29.81
	연도	−278.4	293.5	157.7	163.4	0
	$y_{j,t-1}$	0.7414	0	0.0774	0	0
	$y_{2,t-1}$	0	0.4552	0	0	0
	$y_{3,t-1}$	0	0	0.4603	0	0
	$y_{4,t-1}$	0	0.4490	0.4615	0.9060	0.1614
	$y_{5,t-1}$	0	−0.3356	−0.4151	−0.3454	0.4802
하기도 감염	절편	384.7	105.3	227.6	350.2	371.6
	평균기온	−10.66	−7.10	−10.42	−14.20	−13.29
	일교차	17.84	16.47	20.96	22.22	21.31
	연도	−120.2	164.9	137.9	68.87	0
	$y_{j,t-1}$	0.7795	0	0	0	0
	$y_{2,t-1}$	0	0.5272	0	0	−0.1507
	$y_{3,t-1}$	0	0	0.4892	0	0.1911
	$y_{4,t-1}$	0	0.1351	0.2354	0.9977	0.2344
	$y_{5,t-1}$	0	0	0	−0.3092	0.4001
피부염/습진	절편	1421.6	828.2	1066.7	896.8	1020.3
	평균기온	14.20	11.13	13.14	12.11	11.67
	연도	−243.3	285.9	156.1	140.6	123.8
	$y_{j,t-1}$	0.1503	−0.3496	−0.3643	−0.3937	−0.3186
	$y_{2,t-1}$	−0.3235	0.3009	0	−0.2618	−0.2344
	$y_{3,t-1}$	0	0	0	0	0.1694
	$y_{4,t-1}$	0	0	0.2600	0.7009	0
	$y_{5,t-1}$	0	0	0	0	0.3886

Table 4.4. Estimates of regression coefficients in the TSCS regression model

질병	모형	절편	평균기온	일교차	연도	전주 빈도
상기도 감염	Fuller-Battese	266.90	−10.280	13.9900	78.91	0.8300
	Parks	198.10	−4.710	−3.0390	55.66	0.9250
	Da Silva	−75.78	6.461	0.9952	31.35	0.9820
하기도 감염	Fuller-Battese	184.80	−7.289	9.0780	47.35	0.8378
	Parks	81.76	−2.645	−0.4147	22.35	0.9512
	Da Silva	59.88	−1.958	−0.5897	18.60	0.9688
피부염/습진	Fuller-Battese	140.60	1.864	-	58.85	0.8329
	Parks	36.94	0.125	-	16.07	0.9644
	Da Silva	−49.38	4.522	-	12.25	0.9898

TSCS = time series cross sectional.

VARX의 분석에서는 경기 지역을 제외한 나머지 권역은 해당 권역뿐만 아니라 다른 권역의 전 주 환자 수에도 영향을 받고 있는 것으로 나타났다. 연도에 따른 환자 수의 경우 경기 지역만 2013년의 환자수가 2012년도에 비해 적었는데 이는 Table 2.2에서 보는 것과 같이 경기 지역의 표본 수가 줄어들어 생긴 현상으로 예상된다. 피부염/습진의 경우 모든 모형에서 평균기온이 올라갈수록 환자 수가 증가하는 것으로 조사되었다. 전 주의 환자 수는 상기도 감염이나 하기도 감염에 비해 그 영향력이 크지 않은 것으로 확인되었다. VARX모형에서의 충청 지역은 상관관계가 높은 주변 지역의 전 주 환자 수에 의해 설명되고 있다. Parks와 Da Silva방법의 경우 위의 결과와 일치하는 부분도 있으나 평균기온과 일교차의 영향력이 서로 상쇄되어, 상기도 감염에서 Da Silva방법의 평균기온과 하기도 감염에서 Parks와 Da Silva방법의 일교차에서 반대의 결과가 나오기도 했다.

위의 세 모형의 5가지 추정에 위한 예측력을 비교하기 위해 다음과 같은 기준을 고려했다.

$$\text{MSE} = \frac{1}{n-k}\sum_{j=1}^{5}\sum_{t=2}^{103}(y_{jt}-\hat{y}_{jt})^2,$$

$$\text{MAPE} = \frac{100}{n-k}\sum_{j=1}^{5}\sum_{t=2}^{103}\frac{|y_{jt}-\hat{y}_{jt}|}{y_{jt}},$$

$$\text{MAE} = \frac{1}{n-k}\sum_{j=1}^{5}\sum_{t=2}^{103}|y_{jt}-\hat{y}_{jt}|,$$

여기서 k는 추정한 모수의 개수를 의미한다.

Table 4.5는 위 기준 하에서 각 질병에 대한 모형의 예측력을 비교한 것으로 최솟값에 대해 밑줄을 표시하였다. 표에서 볼 수 있듯이 모든 결과에서 VARX모형이 좋은 것으로 나타났다. VARX모형의 경우 외생변수인 기상자료를 권역이 아닌 전국 평균을 사용하는데 우리나라의 경우 권역별로 기상변수의 차이가 크지 않아 전국 평균을 사용하더라도 문제가 없다는 것을 의미한다.

5. 결론

지금까지 기상에 영향을 받는 고령 환자의 질병 발생빈도를 예측하기 위한 모형에 대해 알아보았다. 심평원 자료와 기상청 자료를 모형에 맞게 병합하여 어떤 방식으로 데이터셋을 구성했는지에 대해 소개했다. 분석결과 상기도와 하기도 감염과 같은 호흡계통의 질환은 평균 기온이 낮을수록, 일교차가 클수록 발생 빈도가 증가하는 것으로, 피부 질환은 평균 기온이 높을수록 발생 빈도가 증가하는 것으로 분석되

Table 4.5. Comparison of forecast errors

질병	모형	MSE	MAPE	MAE
상기도 감염	ARX	53385.1	9.103	155.83
	VARX	52929.8	8.665	153.56
	Fuller-Battese	62642.1	9.119	160.42
	Parks	64686.0	8.717	156.82
	Da Silva	77909.2	10.832	184.15
하기도 감염	ARX	17843.9	8.000	93.66
	VARX	17284.0	7.680	90.98
	Fuller-Battese	21012.3	8.528	98.39
	Parks	21237.4	7.936	94.81
	Da Silva	21699.4	8.078	96.29
피부염/습진	ARX	12439.2	7.245	76.89
	VARX	10791.6	6.656	71.64
	Fuller-Battese	19684.3	8.824	95.19
	Parks	20320.3	8.403	92.61
	Da Silva	23040.5	9.086	99.78

었다. 또한 1주차 전의 발생빈도와 양의 상관관계가 있었으며 이에 대한 추정식을 유도하였다. 모형의 예측력은 VARX모형이 우수한 것으로 나타났는데 이는 기상 자료의 경우 전국 평균을 이용해도 큰 문제가 없음을 의미한다.

References

Da Silva, J. G. C. (1975). *The Analysis of Cross-Sectional Time Series Data*, Ph.D. dissertation, Department of Statistics, North Carolina State University.

Fuller, W. A. and Battese, G. E. (1974). Estimation of linear models with crossed-error structure, *Journal of Econometrics*, **2**, 67–78.

Kim, S. Y., Choi, M. O., and Han, J. T. (2015). Analysis of the elderly with Parkinson's disease, *International Journal of Welfare for the Aged*, **68**, 217–250.

Lee, J. H. and Hwang, T. Y. (2015). Oral health status and care needs of elderly patients in long-term care hospital, *Journal of Korean Society of Dental Hygiene*, **15**, 411–416.

Lee, M. H. and Park, Y. H. (2015). Factors influencing attitude toward advance directives of older cancer patients, *The Journal of Korean Academic Society of Adult Nursing*, **27**, 449–458.

Parks, R. W. (1967). Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated, *Journal of the American Statistical Association*, **62**, 500–509.

기상에 따른 고령환자의 질병 발생빈도 예측모형 비교

이선재[α] · 여인권[α,1]

[α]숙명여대 통계학과

(2015년 12월 15일 접수, 2015년 12월 23일 수정, 2015년 12월 23일 채택)

요 약

이 논문에서는 기상에 따른 고령 환자의 질병 발생빈도를 예측하는 방법을 비교한다. 분석을 위해 건강보험심사평가원의 고령 환자 의료 자료와 기상청 자료를 주별, 권역별로 병합한다. 기상에 영향을 받는 질병의 주별 발생 빈도를 ARMAX모형, VARMAX모형, TSCS회귀모형으로 분석하고 MSE, MAPE, MAE 기준으로 모형을 비교했다.

주요용어: ARMAX모형, TSCS회귀모형, VARMAX모형, 심평원자료, 기상자료

본 연구는 숙명여자대학교 교내연구비 지원에 의해 수행되었음 (과제번호 1-1503-0123).

[1]교신저자: (04310) 서울시 용산구 청파로47길 100, 숙명여대 통계학과. E-mail: inkwon@sm.ac.kr