Journal of the Korean Data & Information Science Society
2017, **28**(1), 131–142

http://dx.doi.org/10.7465/jkdi.2017.28.1.131
한국데이터정보과학회지

사례연구: 대구 파티마 병원 폐렴 입원 환자 수에 영향을 미치는 날씨 변수 선택

최소현[1] · 이학래[2] · 박천건[3] · 이경은[4]

[1]서울대학교병원 의학연구협력센터 · [2]한국통계정보원 · [3]경기대학교 수학과 · [4]경북대학교 통계학과

접수 2016년 12월 29일, 수정 2017년 1월 16일, 게재확정 2017년 1월 18일

요 약

매년 폐렴 입원 환자 수는 증가하는 추세이며, 국내 질환 중 입원율 1위이기도 하다. 주로 박테리아와 바이러스가 주된 원인인 폐렴은 날씨의 영향을 받기도 한다. 본 연구에서는 날씨 변수로는 습도, 일조량, 일교차, 평균온도, 미세먼지 농도를 각각 1일 전부터 27일 전까지의 총 135개 변수를 고려하였다. 날씨와 입원 환자 수에 잠재적으로 영향을 미치는 위험 요인으로 연도 효과, 휴일 효과, 계절 효과를 추가적으로 고려하였다. 벌점화 일반화 선형 모형을 이용하여 폐렴 입원 환자 수와 관련된 변수를 선택하였다.

주요용어: 라쏘, 릿지, 벌점화 일반화 선형모형, 엘라스틱 넷.

1. 서론

지구 온난화로 인해 전 세계는 날씨가 변하였고, 그에 따라 여러 가지 질병의 발생률도 증가하였다. 그 중 폐렴은 국내 질환 중 입원율 1위인 질환으로 폐렴 (pneumonia)에 대한 위험성은 계속 부각되고 있다. 겨울이 되면 폐렴환자가 증가하는 것으로 보아 날씨에 큰 영향을 받는 질환임은 부정할 수 없는 사실이다 (Yim 등, 2012; Kim 등, 2016b). 그렇다면 수 많은 날씨 변수들 중 어떤 종류의 날씨가 폐렴 발병에 영향을 끼치는지 어느 정도 잠복기를 가지는 지를 알아보기 위해 본 연구를 계획하였다. 폐렴은 감염 통로에 따라 community acquired pneumonia (CAP)와 hospital acquired pneumonia (HAP)로 나눌 수 있다. HAP는 날씨와 밀접한 관련이 없는 것으로 판단되어 CAP 환자들로 국한하여 연구를 진행하였다. 대구 파티마 병원의 폐렴으로 입원한 일별 환자 수와 날씨 자료에 대해 적절한 변수를 선택하고 일별 환자 수와의 관계를 알아 볼 것이다. 날씨 자료는 습도, 일조량, 일교차, 평균온도, 미세먼지 농도를 고려하였다 (Lieberman과 Friger, 1999). 포아송 일반화 선형 모형을 사용하였고, 이때 영향을 미치지 않는 변수로 인한 모형 과적합성과 예측 성능 저하를 피하기 위해 적절한 변수를 선택하여야 한다. 하지만 날씨 변수들은 서로 높은 상관관계를 가지기 때문에 기존의 변수 선택법을 사용하기에 무리가 따른다. 따라서 벌점화 기법을 적용한 변수 선택법을 통해 실질적으로 입원 환자 수에 영향을 미치는 변수를 선택하였다.

[1] (03080) 서울시 종로구 대학로 101, 서울대학교병원 의학연구협력센터, 연구원.
[2] (35203) 대전광역시 서구 둔산대로 117번길 18, 한국통계정보원, 연구원.
[3] (16227) 경기도 수원시 영통구 광교산로 154-42, 경기대학교 수학과, 부교수.
[4] 교신저자: (41566) 대구광역시 북구 대학로 80, 경북대학교 통계학과, 부교수. E-mail: artlee@knu.ac.kr

본 논문은 다음과 같은 구조로 되어있다. 2절에서는 본 연구에서 다룰 폐렴으로 입원한 일별 환자 수 자료와 날씨 자료의 특징과 변수 구조에 대해 알아본다. 3절에서는 모형의 설명력과 예측력 향상을 위한 벌점 기법과 본 연구에서 사용할 엘라스틱 넷 기법과 라쏘에 대해 살펴보고 4절에서는 폐렴 입원 환자 수와 날씨 자료에 적절한 모형을 선택하고 적합시킨다. 5절에서는 분석을 정리하며 본 연구의 한계점과 나아가야 할 방향을 살펴본다.

2. 자료 소개

본 논문에서 사용된 자료는 2008년부터 2012년까지의 대구 파티마 병원에 폐렴으로 입원한 일별 환자 수와 대구 지역의 날씨 자료이다. 이번 절에서 자료로부터 추출한 변수의 선정과 구조에 대해 알아본다.

2.1. 폐렴 입원 환자 수

원자료는 2008년부터 2012년까지의 대구 파티마 병원의 폐렴으로 입원한 환자 자료이다. 인구학적 특성인 성별과 나이, 그리고 입원일, 퇴원일이 있었으며, 본 연구에서는 일별 입원 입원 환자 수를 추출하여 변수로 선정하였다. 입원 환자 수의 시간적 추세를 보기 위해 그래프를 그렸으나, 관측치에 0이 많아 추세를 보기 힘들어 월별 합계로 그래프를 그렸다 (Figure 2.1). 그래프에서 점점 증가하는 추세가 보이며, 계절에 따른 차이 또한 확인할 수 있다. 평균적으로 여름에는 낮은 수치를 보이며, 겨울과 봄에는 높은 수치를 보인다. 환자의 인구학적 특성인 연령과 성별에 따른 차이는 크지 않았고, 본 연구의 목적과 맞지 않으므로 제외하였다. 퇴원날짜 또한 본 연구에 적절한 변수가 아니므로 제외하였다.

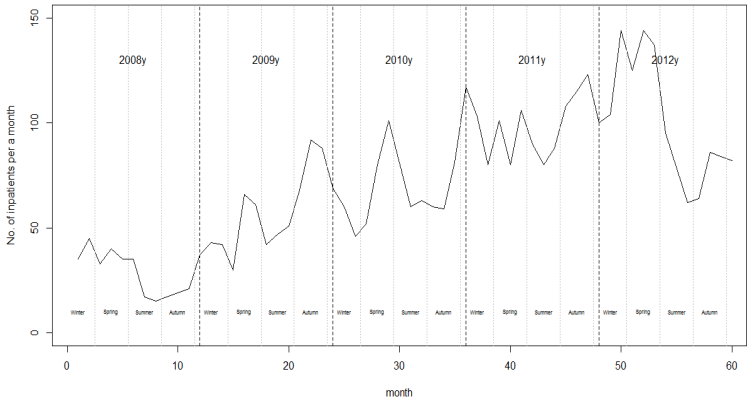

Figure 2.1 Monthly sums of hospitalized CAP patients

2.2. 날씨 변수

폐렴에 영향을 끼칠 것으로 예상되는 날씨 변수로 습도, 일조량, 일교차, 평균온도, 미세먼지 농도를 고려하였다 (Yim 등, 2012). 기간은 폐렴 입원 자료와 동일한 기간인 2008년 1월 1일부터 2012년 12월 31일까지의 대구 지역의 날씨를 추출하였다. 이 때, 일교차는 일일 최고기온과 최저기온의 차로 산출하였다. 각 날씨 변수와 입원 환자수와의 관계를 살펴보기 위하여 각 날씨 변수와 입원환자수와의 상관관계를 구하였다. Figure 2.2에서 나타나듯이 각 날씨변수와 입원환자수는 관련이 있으며 날씨변수그룹별

로 비슷한 양상을 보여주고 있다. Figure 2.3 열지도 (heatmap)에서 알 수 있듯이, 높은 상관계수를 가질수록 더 진한 붉은 색을 가지는 데, 평균온도 변수 그룹은 서로 높은 상관관계가 높고, 일교차, 미세먼지농도는 가까운 과거 시점들과 높은 상관관계를 가지는 것을 알 수 있다.

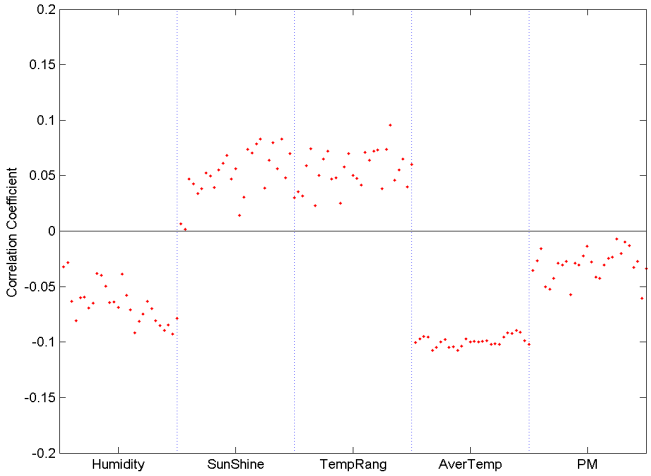

Figure 2.2 Correlation coefficient plot between response variable and weather variables

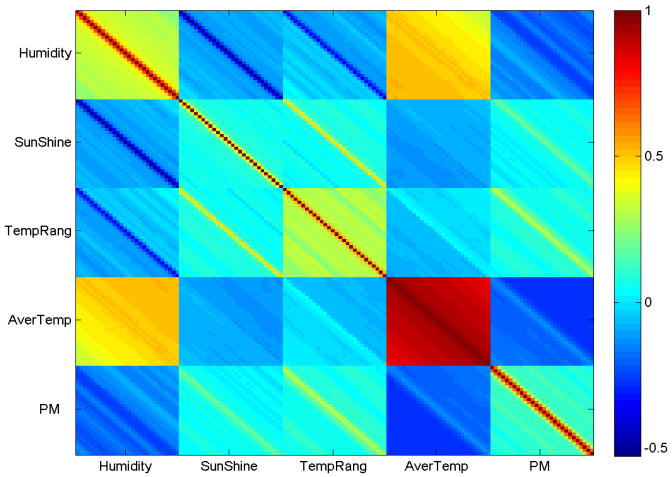

Figure 2.3 Correlation Plot among weather variables

폐렴 입원 환자 수는 지난 날씨에 영향을 받기 때문에 입원 전 4주 간의 날씨를 고려하였다. 날씨 변수는 당일을 제외하고 1일 전부터 27일 전까지 총 27일 간의 날씨 자료로 변수를 구축하였다.

$$\mathbf{x}_{\text{습도}} = (x_{\text{습도1일전}}, \ldots, x_{\text{습도27일전}})$$
$$\mathbf{x}_{\text{일조량}} = (x_{\text{일조량1일전}}, \ldots, x_{\text{일조량27일전}})$$
$$\mathbf{x}_{\text{일교차}} = (x_{\text{일교차1일전}}, \ldots, x_{\text{일교차27일전}})$$
$$\mathbf{x}_{\text{평균온도}} = (x_{\text{평균온도1일전}}, \ldots, x_{\text{평균온도27일전}})$$
$$\mathbf{x}_{\text{미세먼지}} = (x_{\text{미세먼지1일전}}, \ldots, x_{\text{미세먼지27일전}})$$
$$\Rightarrow \mathbf{x}_{\text{날씨}} = (\mathbf{x}_{\text{습도}}, \mathbf{x}_{\text{일조량}}, \mathbf{x}_{\text{일교차}}, \mathbf{x}_{\text{평균온도}}, \mathbf{x}_{\text{미세먼지}})$$

2.3. 더미 변수

폐렴 환자 수와 단순히 날씨 관계 만을 고려하기에는 무리가 따른다. 날씨는 매년 큰 변화를 보이지 않는 반면 입원 환자 수는 증가하는 추세를 보이는데, 이는 폐렴에 대한 중요성이 부각되면서 사람들의 인식의 변화도 있을 것이다. 따라서 연도별 더미 변수도 모형에 추가하였다. 기준은 가장 최근인 2012년을 기준으로 두었다. 본 연구의 자료는 '입원'에 촛점을 두었다. 폐렴의 증상은 감기와 비슷하기 때문에 감기인 줄 알고 가볍게 생각하다가 증상이 심해지면 병원을 찾았다가 폐렴으로 진단 받고 입원하는 경우가 많을 것이다. 그래서, 휴일보다는 평일에 입원할 가능성이 높다. 주말에는 내원하지 않고 다음 날인 월요일, 혹은 긴 연휴를 보내고 그 다음 날에 병원을 찾는 것을 고려하여 '휴일 다음 날' 효과 또한 모형에 추가하였다. 마지막으로 계절에 대한 더미 변수 또한 추가하였다. 날씨와 직접적인 관계가 있으나 '계절'이라는 구분이 주는 영향을 무시할 수는 없다. 계절의 구분은 12월, 1월, 2월은 겨울, 3월, 4월, 5월은 봄, 6월, 7월, 8월은 여름, 9월, 10월, 11월은 가을로 구분하였고, 입원 환자 수가 가장 낮은 '여름'을 기준으로 더미 변수를 생성하였다. 연도별 더미 변수 4개, 휴일 효과 더미 변수 1개, 계절 더미 변수 3개로 총 8개의 더미 변수를 포함하여 최종 변수를 구축 하였다.

3. 벌점화 모형

이번 절에서는 일일 폐렴 입원 환자 수 예측을 위한 엘라스틱 넷과 엘라스틱 넷의 기초인 라쏘와 능형회귀 모형에 대해 리뷰하려고 한다. 먼저 기본적인 선형 회귀 모형에 대해 고려해보자.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

여기서, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$, 각 독립변수 $\mathbf{x}_j$는 $n \times 1$ 벡터이며 (n는 표본의 수이고, p는 독립변수의 수이다), 변수선택의 공정성을 위하여, 평균이 0, 표준편차가 1로 표준화되었고, 종속변수 $\mathbf{y}$는 평균이 0으로 중심화 (centered) 되어있다. 계수 벡터 $\boldsymbol{\beta}$는 다양한 방법으로 추정될 수 있으며 가장 대표적인 방법이 잔차제곱합($= (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$)을 최소화하는 방법인 최소제곱 추정법 (least squares estimation)이다. 최소제곱추정량 (least squares estimator)은 최소선형불편추정량 (best linear unbiased estimator)이 되는 좋은 성질을 가지고 있으나, 독립변수의 수가 표본의 수보다 많을 때는 추정량의 유일성 (uniqueness)이 깨지는 문제점이 있으며, 종속변수와 관련없는 독립변수들로 인하여 모형의 복잡성이 증가하고, 예측력을 떨어뜨리기도 한다. 따라서, 관련없는 독립변수들을 선택하지 않거나 편의가 있으나 분산이 작은 추정량 (bias-variance trade off)을 고려하는 것이 모형의 단순성이나 예측력을 높일 수 있다. 최소제곱추정법의 대안 방법 중 하나는 벌점화 최소제곱법 (penalized least squares)이며 벌점화 최소제곱추정량은 다음과 같이 표현 할 수 있다:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + p_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \right) \tag{3.1}$$

여기서 $p_{\boldsymbol{\lambda}}(\cdot)$는 주어진 비음 (nonnegative)의 벌점함수이며 $\boldsymbol{\lambda}$는 조율 (regularization) 모수이다 (Kim 등, 2016a; Shim 등, 2016). $p_{\boldsymbol{\lambda}}(\cdot)$의 형태에 따라 추정량이 달라지게 된다. 능형회귀 (ridge regression;

Hoerl과 Kennard, 1970)는 l_2 벌점화 (제곱 벌점화)를 가진 최소제곱법으로 $\mathbf{X}$가 full rank가 아닌경우에도, 주어진 λ에 따라 유일한 근을 제공해주는 방법으로 추정량의 편의가 발생하는 반면 분산을 줄여 예측력을 높이는 방법이다. 하지만, 모형의 단순성 측면에서는 모든 변수를 선택하게 되는 단점을 가지고 있다. 능형회귀 추정량은 다음과 같이 표현될 수 있다:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right)$$

여기서 $\|\boldsymbol{\beta}\|^2 = \sum_{j=1}^{p} \beta_j^2$이다.

라쏘 (Lasso, Tibshirani, 1996)는 l_1 벌점화(절대값 벌점화)를 부과하여 계수를 추정하는 방법이며 라쏘 추정량은 아래와 같다:

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \right)$$

여기서 $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$. 라쏘 기법은 계수 축소 (shrinkage) 뿐만 아니라 변수 선택이 자동적으로 이루어져 예측력을 높이는 장점이 있지만, 몇 가지 상황에서 제한점을 가진다: (a) $p > n$ 일 때, 블록 최적화 기법 (convex optimization)의 특성상 변수가 최대 n개 까지만 선택할 수 있어서 변수 선택 기능이 제한되어진다. (b) 변수들간의 상관 계수가 높은 그룹이 있으면 그룹 내에서 한 개의 변수가 랜덤하게 선택되는 경향이 있다. (c) $n > p$이고, 변수들간의 높은 상관관계가 존재할 때 능형회귀 모형에서 더 좋은 예측력을 가진다 (Zou와 Hastie, 2005).

이러한 능형회귀 모형과 라쏘 모형의 단점을 서로 보완할 수 있는 엘라스틱 넷 (elastic net; Zou와 Hastie, 2005)이 제안되었는 데, 나이브 엘라스틱넷 (naive elastic net)은 l_1 벌점화와 l_2 벌점화 모두 부과하여 계수를 추정하는 기법이며 추정량을 다음과 같이 구할 수 있다:

$$\hat{\boldsymbol{\beta}}_{\text{NaiveEN}} = \arg\min_{\boldsymbol{\beta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|^2 \right).$$

나이브 엘라스틱넷 추정량은 두 개의 조율모수로 인한 이중 축소 (double shrinkage) 문제점을 가지고 있어 다음과 같이 재척도화 (rescaling)하여 엘라스틱넷 추정량을 정의하고 있다 (Zou와 Hastie, 2005):

$$\hat{\boldsymbol{\beta}}_{\text{EN}} = (1 + \lambda_2)\hat{\boldsymbol{\beta}}_{\text{NaiveEN}}.$$

엘라스틱넷 기법은 라쏘 기법과 같이 변수 선택이 자동적으로 되어 차원축소가 가능하며, 상호 연관된 변수들을 집단적으로 선택하며 계수 추정값들도 비슷하게 추정하는 경향이 있어 집단적 설명 또한 가능하다는 장점이 있다. 상호 연관된 변수들이 많이 있을 때, 변수 선택력은 라쏘 기법보다 우월한 것으로 알려져 있다.

입원환자수는 범주형 변수이므로 선형모형 대신 일반화 선형모형으로 분석하는 것이 적절하며 벌점화 일반화 선형회귀 모형은 식 (3.1)에서 $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ 대신 일반적으로 로그우도함수에 -2를 곱한 항으로 대체해서 사용한다. 벌점화 일반선형모형에서는 종속변수를 중심화 하지 않으므로 β_0가 존재하며, 벌점화는 다른 회귀계수벡터 $\boldsymbol{\beta}$ 부분에만 적용하기 때문에 식 (3.1)은 다음과 같이 변경된다:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(-2l(\beta_0, \boldsymbol{\beta}; \mathbf{y}) + p_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \right) \tag{3.2}$$

여기서 $l(\beta_0, \boldsymbol{\beta}; \mathbf{y})$는 로그우도함수이다. 포아송 회귀 모형의 로그우도함수는

$$l(\beta_0, \boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \log \mu_i - \mu_i - \log(y_i!) \right], \mu_i = \exp(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})$$

이고 로지스틱 회귀 모형의 로그우도함수는 다음과 같다:

$$l(\beta_0, \boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right], p_i = \frac{\exp(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})}.$$

벌점함수가 없는 경우는 결국 우도함수를 최대로 하는 추정량이되므로 최대우도추정량이 된다. 포아송 회귀모형에서 벌점화 회귀 계수 추정량을 구체적으로 살펴보면, 최대우도추정량 $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, 능형 회귀추정량 $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$, 라쏘 회귀추정량 $\hat{\boldsymbol{\beta}}_{\text{Lasso}}$, 나이브 엘라스틱넷 회귀추정량 $\hat{\boldsymbol{\beta}}_{\text{NaiveEN}}$은 각각 다음과 같이 표현될 수 있다:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{MLE}} &= \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left[y_i(\beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta}) - \exp\left(\beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta}\right) - \log(y_i!) \right] \right\}, \\
\hat{\boldsymbol{\beta}}_{\text{Ridge}} &= \arg\min_{\boldsymbol{\beta}} \left\{ -2\sum_{i=1}^{n} \left[y_i(\beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta}) - \exp\left(\beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta}\right) - \log(y_i!) \right] + \lambda\|\boldsymbol{\beta}\|^2 \right\}, \\
\hat{\boldsymbol{\beta}}_{\text{Lasso}} &= \arg\min_{\boldsymbol{\beta}} \left\{ -2\sum_{i=1}^{n} \left[y_i(\beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta}) - \exp\left(\beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta}\right) - \log(y_i!) \right] + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \\
\hat{\boldsymbol{\beta}}_{\text{NaiveEN}} &= \arg\min_{\boldsymbol{\beta}} \left\{ -2\sum_{i=1}^{n} \left[y_i(\beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta}) - \exp\left(\beta_0 + \mathbf{x}_i^{'}\boldsymbol{\beta}\right) - \log(y_i!) \right] + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2 \right\}.
\end{aligned}$$

또한, 로지스틱 회귀모형에서 다양한 벌점화 회귀 계수 추정량들은 식 (3.2)에서 로지스틱 모형의 로그우도함수와 여러 벌점화 함수를 이용하여 쉽게 표현할 수 있다.

4. 사례연구

종속변수는 일별 폐렴 입원 환자 수를, 독립변수로는 연도 더미 변수 4개, 계절 더미 변수 3개, 휴일 더미 변수 1개와 날씨 변수 135개로 총 143개의 독립변수를 고려한다. 독립변수들 중에서 서로 상관관계가 높은 변수군들이 존재하므로 엘라스틱넷 변수 선택법을 적용하여보고, 최소제곱추정법, 라쏘기법, 능형 기법도 같이 적용하여 변수들을 선택하려고 한다.

자료는 2008년부터 2012년까지 5년 동안의 자료를 사용하였다. 그 중 2008년 1월 1일부터 1월 27일까지의 자료는 27일 전까지의 날씨정보를 적용할 수 없어 분석에서 제외하였다. 자료 중 무작위로 뽑은 360일 (자료의 20퍼센트) 자료는 엘라스틱넷, 라쏘, 능형 회귀, 최소제곱법 모형의 테스트자료로 사용하려고 한다. 그래서, 전체 1,827일 중 처음 27일, 테스트자료 360일을 제외하고 1,440일의 자료로 분석을 실시하였다.

날씨 변화에 따른 일별 폐렴 입원 환자수에 영향을 미치는 날씨 변수 선택을 위해 포아송 일반화 선형 모형에 적합시키고, 벌점화 함수로, 라쏘 벌점화, 능형 벌점화, 엘리스틱 넷 벌점화, 무 벌점화 등을 적용하였다. 추가적으로, 일별 폐렴 환자수의 유무로 이분화하여 로지스틱 회귀모형에도 적합시켜 보았다. 연도별 환자수의 빈도수는 Table 4.1에 있다.

Table 4.1 Numbers of hospitalized patients by year

Year	Total	0	1	2	3	4	5	6	7	8	9	10	11	12
2008	318	150	105	55	17	8	4							
2009	699	73	96	84	56	26	19	9	2					
2010	860	40	89	86	78	33	22	6	6	2	1	2		
2011	1174	18	53	77	70	72	36	15	10	6	2	4	1	1
2012	1039	16	34	61	66	44	41	19	14	4	1	4	1	

벌점화 기법으로 선택된 변수는 훈련자료가 바뀔때 마다 선택된 조율모수도 달라지게 되고 결과론적으로 선택된 변수들도 달라지게 된다. 이러한 임의성을 고려하기 위해 선택 사후 추론 (post-selection inference)에 관하여 최근 많은 연구들을 하고 있다 (Hastie 등, 2015). 본 연구에서는, 붓스트랩 방법으로 1,000번 반복하여 각 벌점화 함수를 적용하여 계수 추정하였다. 즉, 훈련 데이터 1,440일 자료 중 1,440개의 자료를 임의로 복원 추출하여 붓스트랩 자료를 생성하여 계수 추정하였고 이 과정을 1,000번 반복하여 산출한 1,000개의 계수로 비모수적 $(1-\alpha)100\%$ 신뢰구간을 구하여 0을 포함하지 않으면 변

수가 유의한 것, 즉, 변수가 선택되는 것으로 하였다. Figure 4.1부터 Figure 4.4는 포아송 일반화 선형모형에 라쏘 기법, 엘라스틱넷 기법, 능형 기법, 무벌점화 기법(최대우도추정)을 각각 적용하여 유의수준을 0에서 0.5까지 변화시킴에 따라 변수 선택의 변화를 나타내었다. 비모수적 신뢰구간은 유의수준 (= 1−신뢰수준)에 따라 크기가 결정된다. 유의수준이 크면 클수록, 신뢰구간의 길이는 짧아지게 되므로 특정한 유의수준에서 신뢰구간이 0을 포함하지 않으면 더 큰 유의수준에서는 0을 포함하지 않게 된다. 즉, 특정한 유의수준에서 변수가 선택되기 시작하면 더 큰 유의수준에서는 그 변수는 반드시 선택되게 된다. 그림에서 선들이 나타나는 이유이다.

라쏘 기법과 엘라스틱넷 기법으로 선택된 변수들은 비슷한 경향을 보인다. 엘라스틱넷 기법이 조금 더 많은 변수들이 선택됨을 알 수 있다. 특별히, 엘라스틱넷 모형에서는 유의수준 0부터 0.5까지 보았을 때 습도의 경우 유의수준 0.38부터 17일 전 습도가 선택되었다. 유의수준 0.43부터 1일 전 습도가 선택되었다.일조량은 유의수준 0.19부터 22 일 전 일조량이 선택 되었고, 유의수준 0.31부터 19일 전 일조량, 유의수준 0.37부터 10일 전 일조량, 유의수준 0.41부터 1일 전, 24일 전 일조량, 유의수준 0.48에서부터 26일 전 일조량이 선택되었다. 일교차는 유의수준 0.13부터 23일 전 일교차가 선택 되었고, 유의수준 0.31부터 22일 전 일교차, 유의수준 0.38부터 3일 전 일교차가 선택되었다. 평균온도는 유의수준 0.01부터 0.5까지 선택되지 않았다. 미세먼지 농도는 유의수준 0.42부터 9일 전 미세먼지 농도가 선택 되었고, 유의수준 0.44부터 2일 전 미세먼지 농도가 선택되었다. 연도별 더미변수는 2012년 기준 2011년 변수는 선택되지 않았고, 2008년, 2009년, 2010년 변수는 유의수준 0.01부터 선택되었다. 휴일 효과 변수 또한 유의수준 0.01부터 선택되었다. 계절 더미 변수는 여름 기준 봄, 가을 변수는 유의수준 0.02부터 선택되었고, 겨울 변수는 유의수준 0.18부터 선택되었다. 능형기법과 무벌점화 기법을 통해서는 더 낮은 유의수준에서부터 변수들이 많이 선택되는 경향을 보인다. 본 연구에서 고려된 기법으로는 더미 변수들 외에는 선택력이 강하지 않음을 볼 수 있다. Table 4.2.에서 엘라스틱넷 기법으로 상대적으로 유의한 변수들의 최소유의수준 (least significant level), 추정값(붓스트랩 추정값들의 평균, Estimate), 최소유의수준에서 신뢰구간 (C.I.)를 제시하였다.

Table 4.2 Least significant level

		Least significant level	Estimate (C.I.)
Humidity	1-day lag	0.43	0.0237 (0.0002,0.0408)
	17-days lag	0.38	-0.0318 (-0.0571,-0.0013)
Sunlight	1-day lag	0.41	-0.0214 (-0.0380,-0.0004)
	10-days lag	0.37	0.0248 (0.0004,0.0437)
	19-days lag	0.31	0.0270 (0.0005,0.0498)
	22-days lag	0.19	0.0292 (0.0006,0.0578)
	24-days lag	0.41	0.0213 (0.0003,0.0367)
	26-days lag	0.48	0.0217 (0.0004,0.0365)
Diurnal	3-days lag	0.38	0.0261 (0.0005,0.0463)
Temperature	22-days lag	0.31	0.0300 (0.0002,0.0548)
Range	23-days lag	0.13	-0.0423 (-0.0847,-0.0002)
Particulate	2-days lag	0.44	0.0286 (0.0008,0.0497)
Matter	9-days lag	0.42	-0.0303 (-0.0533,-0.0005)

종속변수를 이분화하여 로지스틱 회귀모형을 적합한 경우 선택되는 변수들은 조금 다른 경향을 보인다. 일조량과 미세먼지가 다른 변수들보다는 상대적으로 강한 연관성을 보이고 있긴 하나, 선택되는 날씨 변수들찾기 어렵고, 계절 더미 변수도 봄이 여름에 비하여 상대적으로 연관성이 높긴 하나 유의수준 0.05에서 선택되는 변수가 없다. 2008년, 2009년, 2010년 연도 변수들은 작은 유의수준에서 선택되기 시작하였다. 결론적으로 환자수 유무는 일조량이나 미세먼지가 상대적으로 연관성이 높으며 환자수의 크기는 일교차, 일조량, 미세먼지 순으로 연관성이 높다고 할 수 있으나, 날씨 변수들과는 연관성이 작다라고 할 수 있다.

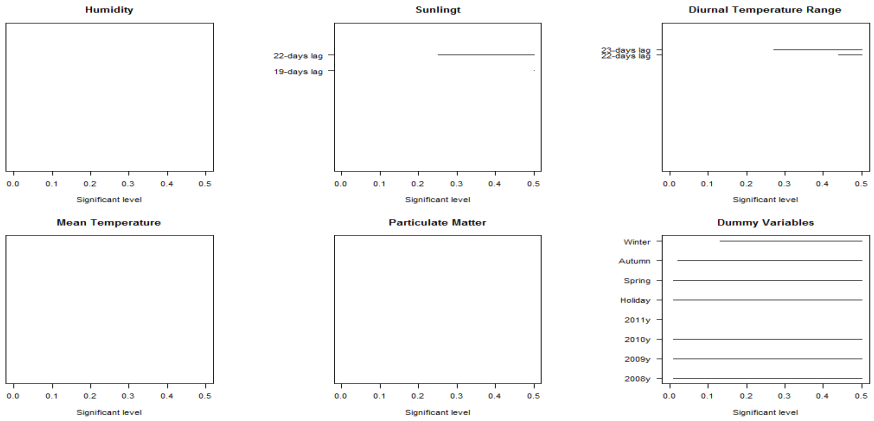

Figure 4.1 Variable selection by significance level using lasso for Poisson regression

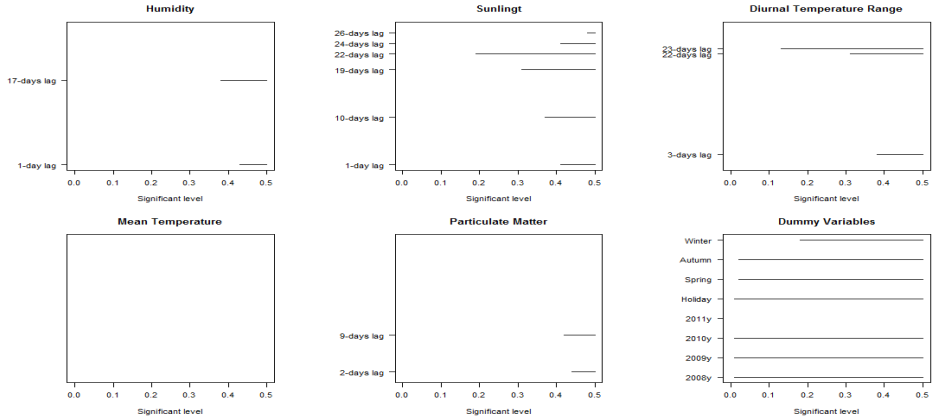

Figure 4.2 Variable selection by significance level using elastic net for Poisson regression

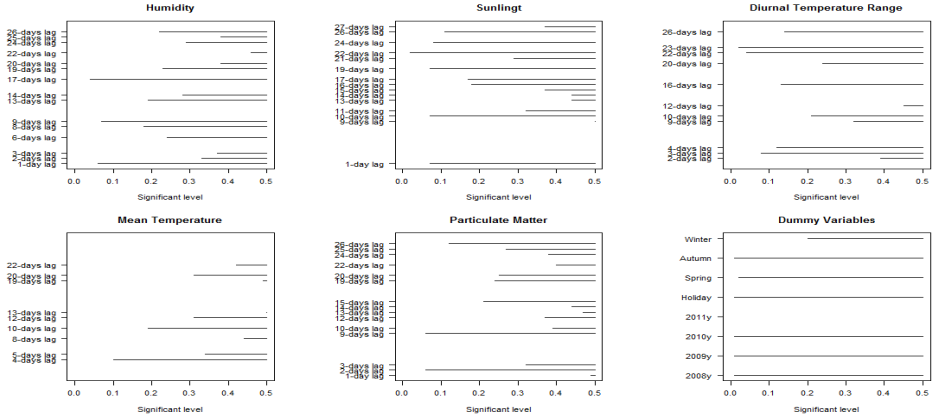

Figure 4.3 Variable selection by significance level using Ridge for Poisson regression

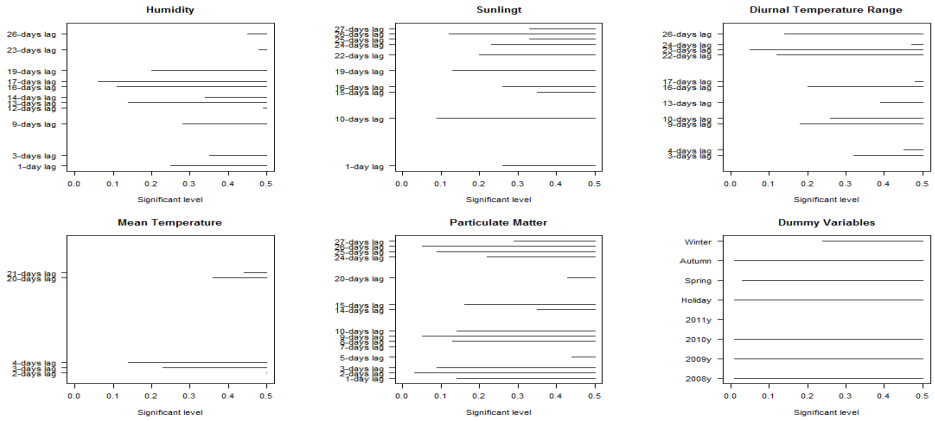

Figure 4.4 Variable selection by significance level for Poisson regression

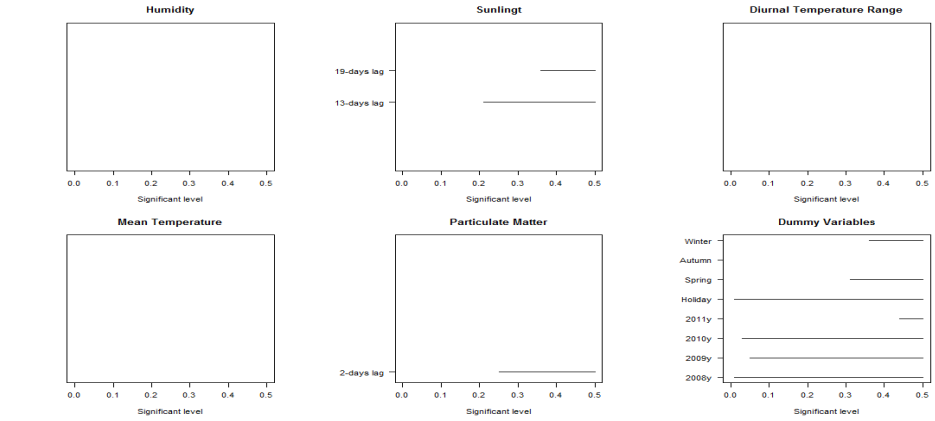

Figure 4.5 Variable selection by significance level using lasso for logistic regression

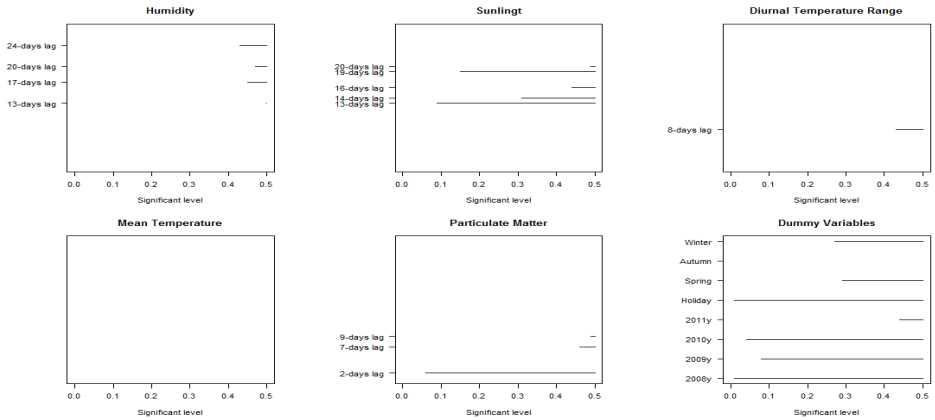

Figure 4.6 Variable selection by significance level using elastic net for logistic regression

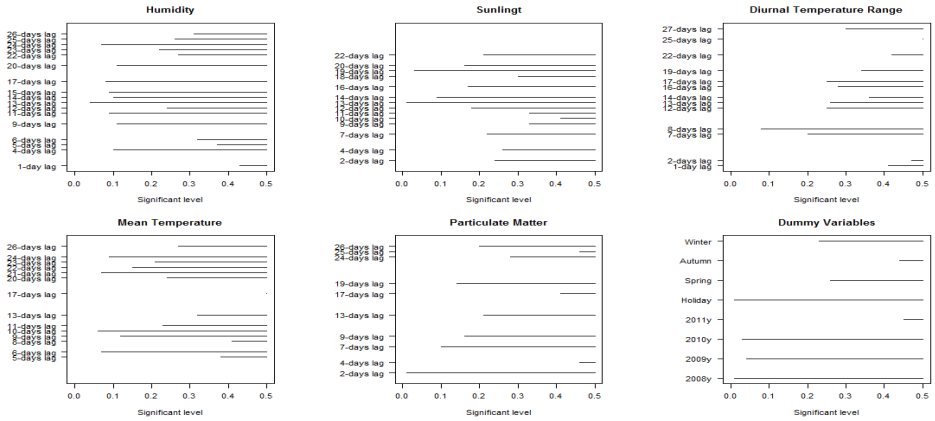

Figure 4.7 Variable selection by significance level using ridge for logistic regression

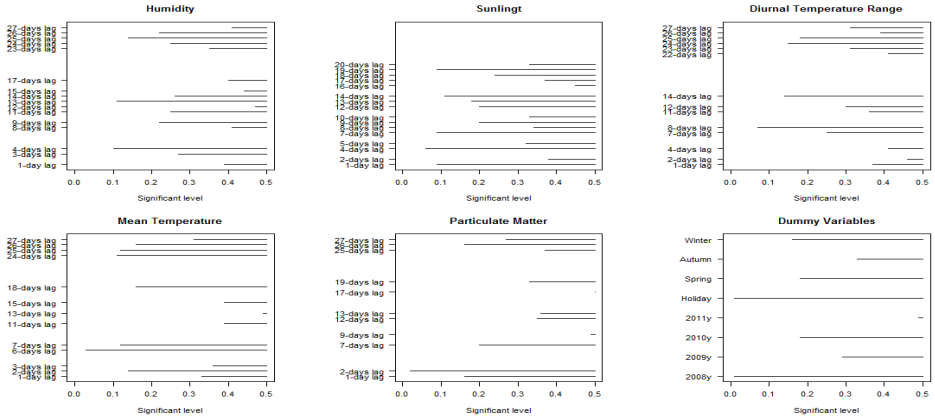

Figure 4.8 Variable selection by significance level for logistic regression

5. 결론

본 연구는 언제, 어떠한 날씨가 폐렴 발병에 영향을 미치는지 알아보기 위해 수행되었다. 포아송모형과 로지스틱 모형에 다양한 벌점화 기법을 적용하여 날씨 변수들을 선택하려고 하였으나, 분석에서 본 것 처럼 날씨 변수들 중 선택력이 강한 변수들이 없었다. 미약하지만, 환자수와 관련해서는 22일전 23일전 일교차, 19일전, 22일전 일조량 변수들이 다른 날씨 변수들에 비해 상대적으로 강한 연관성을 보이고 있다. 폐렴 환자 입원 유무와 관련해서는 3일전 일조량, 2일전 미세먼지 변수가 다른 날씨 변수들에 비해 상대적으로 강한 연관성을 보이고 있다. 자료를 특정한 한 병원에 국한하였기 때문에 날씨와 연관성을 찾기가 어려운 것으로 판단됨으로, 보험공단의 청구자료를 이용하여, 한국 전체의 폐렴 입원 환자 자료를 확보하여 날씨변수와의 연관성을 연구하는 것이 더 좋을 결과를 줄 것으로 예상된다.

References

Lim, Y., Hong, Y. and Kim. H. (2012). Effects of diurnal temperature range on cardiovascular and respiratory hospital admissions in Korea. *Science of the Total Environment*, **417**, 55-60.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*, Chapman and Hall, London.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.

Kim, B., Ha, I. D. and Lee, D. (2016a). Analysis of multi-center bladder cancer survival data using variable-selection method of multi-level frailty models. *Journal of the Korean Data & Information Science Society*, **27**, 499-510.

Kim, J., Kim, J. H., Cheong, H. K., Kim, H., Honda, Y., Ha, M., Hashizume, M., Kolam, J. and Inape, K. (2016b). Effect of climate factors on the childhood pneumonia in papua new guinea: A time-series analysis. *International Journal of Environmental Research and Public Health*, **13**, 213-228.

Lieberman, D. and Friger, M. D. (1999). Seasonal variation in hospital admissions for community-acquired pneumonia: A 5-year study. *Journal of Infection*, **39**, 134-140.

Shim. J., Bae, J. and Seok, K. (2016). Estimation and variable selection in censored regression model with smoothly clipped absolute deviation penalty. *Journal of the Korean Data & Information Science Society*, **27**, 1653-1660.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B (Methodological)*, **58**, 267-288.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society B (Methodological),* **67***, 301-320.*

Journal of the Korean Data & Information Science Society
2017, **28**(1), 131–142

http://dx.doi.org/10.7465/jkdi.2017.28.1.131
한국데이터정보과학회지

Case study: Selection of the weather variables influencing the number of pneumonia patients in Daegu Fatima Hospital

Sohyun Choi[1] · Hag Lae Lee[2] · Chungun Park[3] · Kyeong Eun Lee[4]

[1]Medical Research Collaborating Center, Seoul National University Hospital
[2]Data Management, Korea Statistical Information Institute
[3]Department of Mathematics, Kyonggi University
[4]Department of Statistics, Kyungpook National University

Received 29 December 2016, revised 16 January 2017, accepted 18 January 2017

Abstract

The number of hospital admissions for pneumonia tends to increase annually and even more, pneumonia, the fifth leading causes of death among elder adults, is one of top diseases in terms of hospitalization rate. Although mainly bacteria and viruses cause pneumonia, the weather is also related to the occurrence of pneumonia. The candidate weather variables are humidity, amount of sunshine, diurnal temperature range, daily mean temperatures and density of particles. Due to the delayed occurrence of pneumonia, lagged weather variables are also considered. Additionally, year effects, holiday effects and seasonal effects are considered. We select the related variables that influence the occurrence of pneumonia using penalized generalized linear models.

Keywords: Elastic net, lasso, penalized generalized linear model, ridge.

[1] Researcher, Medical Research Collaborating Center, Seoul National University Hospital, Seoul 03080, Korea.
[2] Researcher, Korea Statistical Information Institute, Daejeon 35203, Korea.
[3] Associate professor, Department of Mathematics, Kyonggi University, Gyeonggi-do 16227, Korea.
[4] Corresponding author: Associate professor, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. E-mail: artlee@knu.ac.kr