Lasso Regression을 이용한 지역 경제 성장과 비만율의 상관관계 분석

길은규*, 오수진**, 김응모*
*성균관대학교 소프트웨어대학
**성균관대학교 정보통신대학
e-mail : kasta1255@gmail.com, bgbanana4@gmail.com, ukim@skku.edu

Analysis of the relationship between regional economic growth and obesity by using Lasso Regression

Eungyu Kil*, Sujin OH**, Ung-Mo Kim*
*College of Software, Sungkyunkwan University
**College of Information and Communication Engineering, Sungkyunkwan University

요 약

본 연구에서는 Lasso Regression을 기반으로 하여 지역 경제 성장과 비만율을 예측한다. 연구는 3단계로 나누어 진행된다. 우선 지역성장을 대변할 수 있는 대변할 수 있는 가상의 GDP 수치를 구한다. 그 다음 가상의 GDP수치와 비만율 데이터를 이용하여 학습모델을 만든다. 마지막으로 이전의 데이터를 이용하여 앞으로의 성장을 예측하고 학습모델에 적용하여 비만율을 예측한다.

본 연구의 데이터는 학습데이터와 실험데이터를 구성된다. 학습데이터로는 국내의 8도 중 하나인 강원도의 데이터를 이용하며 실험데이터로는 강릉과 원주의 데이터를 이용한다. 평가 비교 대상으로는 과거의 흐름을 반영하는 최소자승법 예측기법을 선정하여 비교한다. 연구 결과 강릉의 경우 비교 데이터와의 오차율 평균은 1.22%로 큰 차이가 없음을 알 수 있다. 따라서 본 연구에서 제안하는 방법이 과거의 흐름을 기반으로 작성됨을 알 수 있다. 하지만 단순히 과거의 흐름만을 통해 예측하는 것은 여러 요소가 복합적으로 작용하는 비만율 예측에 알맞지 않기 때문에 본 연구 방법이 유의미하다고 여겨진다.

1. 서론

1.1 연구 배경

21세기가 되면서 세계는 이전과 비교 할 수 없을 정도로 많은 분야에서 발전을 거듭해 왔다. 하지만 이러한 발전이 긍정적인 변화만을 초래한 것은 아니다. 이전과는 달리 삶이 풍족해 짐에 따라 더 이상 '살기 위해 먹는 것'이 아니라 그것을 뛰어넘어 정신적인 만족 또는 사회적 원인에 의해 필요 이상으로 음식을 먹게 되는 경우가 많아졌다 [1]. 결과적으로 비만인구는 점차 증가했으며 이는 국제사회의 큰 문제가 되었다. 하지만 더 큰 문제는 비만이 외형적 변화에 그치는 것이 아니라 다양한 합병증을 유발하여 건강을 위협한다는 점이다. 이 합병증들은 건강에 굉장히 치명적이며 심각한 경우에는 목숨을 위협하기도 한다. 따라서 비만은 더 이상 가볍게 다루어야할 문제가 아니게 되었다.

1.2 연구 목적

본 연구에서는 GDP를 대변할 수 있는 지표들을 선정하여 특정지역의 경제성장을 분석하고 이와 비만율과의 관계를 학습시킨다. 이를 이용하여 미래의 비만율을 예측하는 것을 본 연구의 목표로 한다.

1.3 Overview

2장에서는 연구에 사용된 여러 기술들에 대해 설명한다. 3장은 데이터수집 및 전처리 과정을 설명한다. 4장에서는 학습 모델과 평가모델을 어떻게 구성했는지 설명한다. 5장에서는 연구 결과를 알아보고 이를 분석한다. 6장에서는 연구 결과를 통해 결론을 도출한다.

2. 관련 연구

2.1 회기분석

통계학에서 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 방법이다. 특히 비선형 회기 분석 방법은 공학에서 뿐만 아니라 수학 또는 사회과학분야에서 널리 사용되는 기법이다 [2]. 회기분석은 시간에 따라 변화하

는 데이터나 어떤 영향 가설적 실험, 인과간계의 모델링 등의 통계적 예측에 사용될 수 있다.

2.2 최소자승법

일반적으로 어떤 실험을 할 때, 독립변수 x를 변경하며, 그에 따른 종속변수 y의 쌍 (x,y)를 얻는다. 이러한 실험을 여러번 반복하여 얻은 수많은 데이터들이 일정한 규칙성을 갖지 못한다면 이 실험은 아무런 의미를 갖지 못한다. 따라서 데이터의 유용성을 판단하기 위해 가장 먼저 할 일은 두 변수간의 상관관계 여부 확인과 어떤 성관관계를 갖는지 찾는 것이다. 또한, 변수 간의 상관관계를 함수로 표현할 수 있다면 변수간의 규칙성이 있음을 의미한다.

2.3 파이썬 라이브러리

파이썬에서는 여러 라이브러리를 통해 matrix연산과 다양한 시각화 방법을 제공한다. 여러 도구를 통해 데이터를 쉽게 가공하며 평균, 분산, 최대, 최소 등을 쉽게 연산할 수 있다. 또한 변수사이의 연관성, 그룹, 선택, 조인 등의 다양한 함수를 통해 matrix를 효율적으로 쉽게 가공할 수 있다. 뿐만 아니라 데이터의 분포 및 패턴을 차트(chart)나 플롯(plot)으로 나타낼 수 있으며 통계수치들을 다양한 방법으로 시각화 시킬 수 있다. 이러한 라이브러리로는 pandas, matplotlib, seaborn, numpy가 있다.

머신러닝 또한 파이썬 라이브러리를 이용하여 편리하게 구현가능하다. 본 연구에서는 Scikit-learn의 Lasso Regression을 이용한다.

3. 데이터 수집 및 전처리

본 연구에서 제안하는 분석기법은 두 가지 사항을 가정한다. 첫째, 특정지역의 성장률은 연속적으로 변하며 일정한 흐름을 가진다. 즉, 수집된 과거의 데이터를 이용하여 다음의 성장률을 예측할 수 있다. 둘째, 지역의 성장률과 비만율 사이에는 특정관계가 존재한다.

3.1. 경제 지표 선정

<표 1> 도시별 제공 지표

NUM	지표	keyword
1	인구 천명당 의료기관 종사 의사 수	A
2	인구 십만명당 사회복지시설 수	B
3	인구 십만명당 문화기반시설 수	C
4	상수도보급률	D
5	사업체수	E
6	하수도보급률	F
7	초등학교 수	G

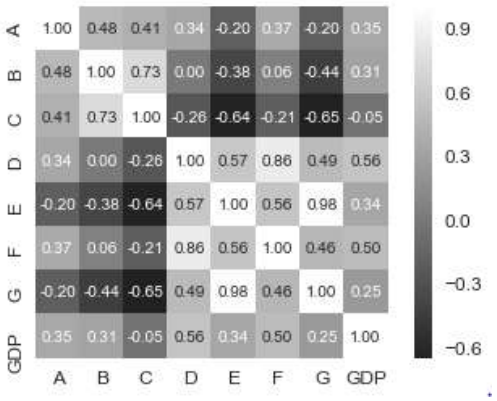
경제 성장률의 직접적인 지표인 GDP가 특별시, 광역시 및 도 단위로만 제공되기 때문에, 본 연구에서는 도시별로 제공되며 경제성장을 대변할 수 있는 지표를 선정하여 비만율 예측에 이용한다. 도시별로 제공되는 지표는 국가통계포털을 통해 제공 받았으며[1] 이는 <표 1>과 같다. <표 1>에서와 같이 각 지표는 앞으로는 keyword로 지칭된다.

경제성장을 대변하는 지표를 선정하기 위해, 우선도 단위의 1인당 GDP와 <표 1>의 지표간의 상관관계를 분석한다. 상관관계를 분석을 위하여, 파이썬 라이브러리 matplot을 이용하였으며, 데이터의 범위는 2008년부터 2015년이다.

<표 2> GDP와 <표 1>의 지표 비교를 위해 작성된 표 일부

Unnamed: 0	Unnamed: 1	A	B	C	D	E	F	G	GDP
1	2008	1.79	6.46	2.73	94.4	651428	88.1	1094	12592
1	2009	1.85	9.19	3.05	95.3	660008	89.9	1114	12697
1	2010	1.88	11.81	2.96	95.7	687022	90.6	1145	13592
1	2011	1.95	11.72	3.17	96.4	720851	91.3	1159	14303
1	2012	1.99	12.87	3.36	96.9	751108	92.7	1176	14773
1	2013	2.06	13.37	3.61	97.5	773216	93.4	1187	15384
1	2014	2.09	14.22	3.89	97.6	810260	93.7	1195	16131
1	2015	2.13	14.65	3.93	97.9	827983	94.0	1213	17130
2	2008	2.08	11.80	9.28	85.5	117150	76.4	361	11194
2	2009	2.14	15.20	10.77	86.1	117569	77.0	353	11908
2	2010	2.18	17.45	10.65	86.5	118266	81.3	353	12499
2	2011	2.17	19.20	10.41	87.5	121273	83.1	353	12775
2	2012	2.20	20.28	10.85	87.9	125192	84.1	352	13304

<표 2>는 <표 1>의 지표와 GDP 비교를 위해 작성한 표의 일부이다. *Unnammed: 0*은 각 도의 이름을 나타내며, 편의상 1~8의 숫자로 대체한다. *Unnammed: 1*은 년도를 의미한다.



(그림 1) 7개의 지표 데이터와 GDP의 상관관계를 표현한 히트맵

1) http://kosis.kr

(그림 1)은 지표 데이터와 GDP 간의 상관관계를 표현한 히트맵이다. (그림 1)의 제일 마지막 행과 <표 3>을 통해, 지표데이터와 GDP의 상관관계를 확인 할 수 있다.

<표 3> 피어슨 상관계수

-1.0 ≤ r ≤ -0.7	매우 강한 음의 상관관계
-0.7 < r ≤ -0.3	강한 음의 상관관계
-0.3 < r ≤ -0.1	약한 음의 상관관계
-0.1 < r ≤ 0.1	상관관계 없음
0.1 < r ≤ 0.3	약한 양의 상관관계
0.3 < r ≤ 0.7	강한 양의 상관관계
0.7 < r ≤ 1.0	매우 강한 양의 상관관계

(그림 1)과 <표 3>에 따르면 A, B, D, E, F 지표가 강한 양의 상관관계를 가짐을 알 수 있다. 하지만, 비율을 나타내는 다른 지표와 달리, 사업체수를 나타내는 E데이터가 학습데이터와 실험데이터 사이의 차이가 너무 커서 학습 및 실험데이터에서 제외한다. 따라서 본 연구에서는 A, B, D, F 4개의 지표를 경제성장 대변하는 지표로 선정하였다.

3.2. 학습데이터

본 연구의 학습모델은 경제 성장을 대변하는 A, B, D, F 지표 데이터뿐만 아니라 Time Locality를 반영해야 한다. 따라서 학습데이터를 생성할 때에 전년도 비만율을 의미하는 PreObesity를 포함한다.

4. 모델구축

4.1. 학습모델 구축

본 연구에서는 특정 도시에 초점을 맞추어 경제성장에 따른 비만율을 예측한다. 본 연구에서는 데이터가 수집되는 시기에 성장률이 높은 경향이 띄는 강원도를 선정하였으며, 2008년부터 2015년 범위의 데이터를 이용하였다. 실험데이터로는 강원도 내부의 도시 중, 강릉과 원주를 선정하여 두 도시의 데이터를 이용하여 연구를 진행한다.

4.2. 평가모델 구축

다음으로 본 연구의 학습 모델 평가를 위한 평가 모델을 구축한다. 평가 모델은 최소자승법을 이용하여 구축하였다.

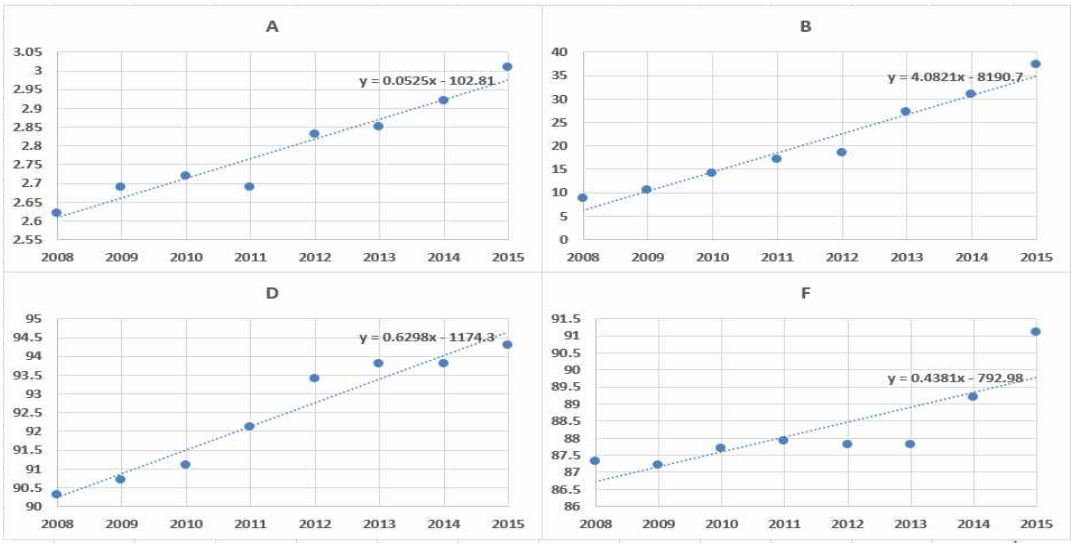
(그림 2)는 3장에서 선별된 성장지표(A, B, D, F) 데이터를 이용하여, 추세선과 상관관계 함수를 구한 그림이다. 추세선을 이용하여 다음에 올 데이터를 예측하고, 예측된 데이터를 이용하여 비만율을 구하는 방식을 취한다.

5. 결과 및 분석

<표 5> 강원도의 최종 학습데이터

강원도	A	B	D	F	PreObesity
2009	2.14	15.2	86.1	77	25.8
2010	2.18	17.45	86.5	81.3	27.1
2011	2.17	19.2	87.5	83.1	27
2012	2.2	20.28	87.9	84.1	27.6
2013	2.25	23.86	88.6	85	27.7
2014	2.3	25.58	88.8	85.6	27.8
2015	2.35	26.52	89.6	86	28.5

<표 5>는 최종 학습데이터로 선정된 강원도의 데이터이다. 이 데이터를 본 연구에서 제안한 학습 모델에 학습시킨다.



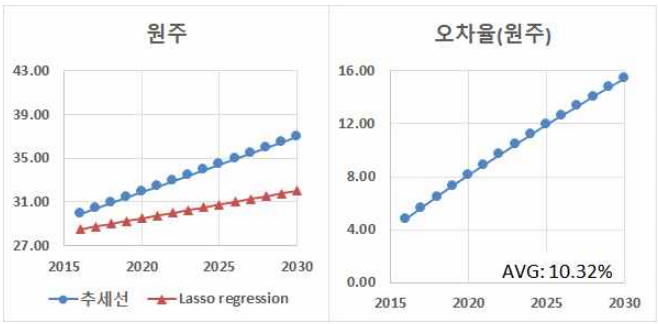
(그림 2) 강원도 내부 도시 강릉의 미래 지표 데이터 예측

다음으로, 강원도의 학습데이터를 이용하여 학습된 학습 모델을 이용하여 강릉과 원주의 비만율을 예측하고 평가 모델과 비교한다.

추세선을 이용한 최소자승법 평가 모델로 예측된 결과와 본 연구에서 제안한 학습 모델로 예측된 결과를 비교하여 학습 모델을 평가 한 결과는 다음과 같다.

(그림 3) 강릉의 비만율과 오차율

(그림 3)은 2016년부터 2030년까지의 강릉의 비만율을 본 연구의 학습 모델과 평가 모델을 통해 예측한 결과와 그 오차율이다. 왼쪽 그래프를 통해 본 연구에서 제안한 비만율 그래프와 과거의 경향을 따르는 추세선 그래프가 큰 차이가 없음을 알 수 있다. 또한 평균 오차율이 1.22%의 작은 값을 가지기 때문에 미래의 비만율이 과거의 비만율의 경향이 유사함을 의미한다.



(그림 4) 원주의 비만율과 오차율

(그림 4)은 2016년부터 2030년까지의 원주의 비만율을 본 연구의 학습 모델과 평가 모델을 통해 예측한 결과와 그 오차율이다. 이를 통해 본 연구방법을 통해 예측된 비만율 그래프와 추세선의 차이가 강릉의 경우보다 크며 오차율 평균 또한 10.32%로 다소 높음을 알 수 있다. 즉, 특정지역의 성장흐름과 단순한 비만율의 흐름은 비슷하지만 일치하지 않는다. 이는 비만율의 흐름만으로는 미래의 비만율을 예측할 수 없으며 정확한 비만율 예측을 위해서는 여러 가지 요소를 고려해야함을 의미하며 본 연구에서 사용한 방법이 유의미함을 나타낸다.

6. 결론

본 연구에서는 Lasso Regression을 기반으로 특정지역의 경제 성장에 따른 미래의 비만율을 예측하는 방안을 제시한다. 특정지역의 성장률은 연속적으로 변하며 일정한 흐름을 가지고, 이 성장률이 비만율과 특정 관계를 가진다는 가정 하에 연구를 진행한다. 선별된 여러 지표를 통해 얻어진 가상의 GDP 수치와 비만율 데이터를 학습모델을 통해 학습시킨 후 성장을 예측하여 학습 모델에 적용시키고 미래의 비만율을 예측한다.

본 연구의 학습데이터는 국내의 8도중 하나인 강원도의 데이터를 이용하고 실험데이터는 강릉과 원주의 데이터를 이용한다. 이 데이터들을 최소자승법을 통해 다음의 지표 데이터를 계산한다. 이 데이터를 본 연구방법을 통해 미래의 비만율을 예측하는데 사용한다. 이 예측된 데이터를 본 연구의 평가 비교 대상인 최소자승법을 통해 예측된 비만율과의 오차율을 구한다. 강릉의 경우 1.22%로 미래의 비만율이 현재까지 비만율의 흐름을 따르며 성장 또한 유사한 흐름을 보임을 확인할 수 있다. 하지만 원주의 경우 10.32%로 다소 차이가 있다. 이는 비만율의 흐름이 특정지역의 성장의 흐름과는 차이가 있으며 비만율의 흐름만으로 미래의 비만율을 예측할 수 없다는 것을 의미한다. 따라서 특정지역의 성장을 토대로 비만율을 예측하는 본 연구방법은 유의미하다.

7.참고문헌

[1] 이명해, “비만의 원인과 평가에 관한 고찰.” 부산여자전문대학 논문집 17 (1995)

[2] Huang, Hsin-Hsiung, Su-Yun Huang, "Nonlinear Regression Analysis." International Encyclopedia of Education (2010)