연구논문/작품 최종보고서

2018 학년도 제 2 학기

제목 : Lasso Regression을 이용한 지역 경제 성장과 비만율의 상관관계 분석

길은규 (2012314166)

2018 년 11 월 6 일

지도교수: 김 응 모 서명

계획(10)	주제(20)	개념(20)	상세(30)	보고서(20)	총점(100)

* 지도교수가 평가결과 기재

■ 요약

본 연구에서는 Lasso Regression을 기반으로 하여 특정 지역의 경제 성장에 따른 비만율을 예측한다. 이를 위해 3단계로 나누어 연구를 진행한다. 연구를 진행하기에 앞서, 본 연구에서는 다음과 같은 가정을 따른다. 첫째, 특정 지역의 성장률은 연속적으로 변하며 일정한 흐름을 가진다. 즉, 수집된 과거의 데이터를 이용하여 다음의 성장률을 예측할 수 있다. 둘째, 지역의 성장률과 비만율 사이에는 특정 관계가 존재한다. 앞의 가정을 기반으로, 3단계로 나누어 연구를 진행한다. 1단계에서는 지역 성장을 대변할 수 있는 가상의 GDP 수치를 여러 지표를 통해 구한다. 2단계에서는 가상의 GDP 수치와 비만율 데이터를 이용하여, 학습 모델을 만들고 학습시킨다. 3단계에서는 1단계의 데이터를 이용하여 앞으로의 성장을 예측하고 학습 모델에 적용하여 비만율을 예측한다.

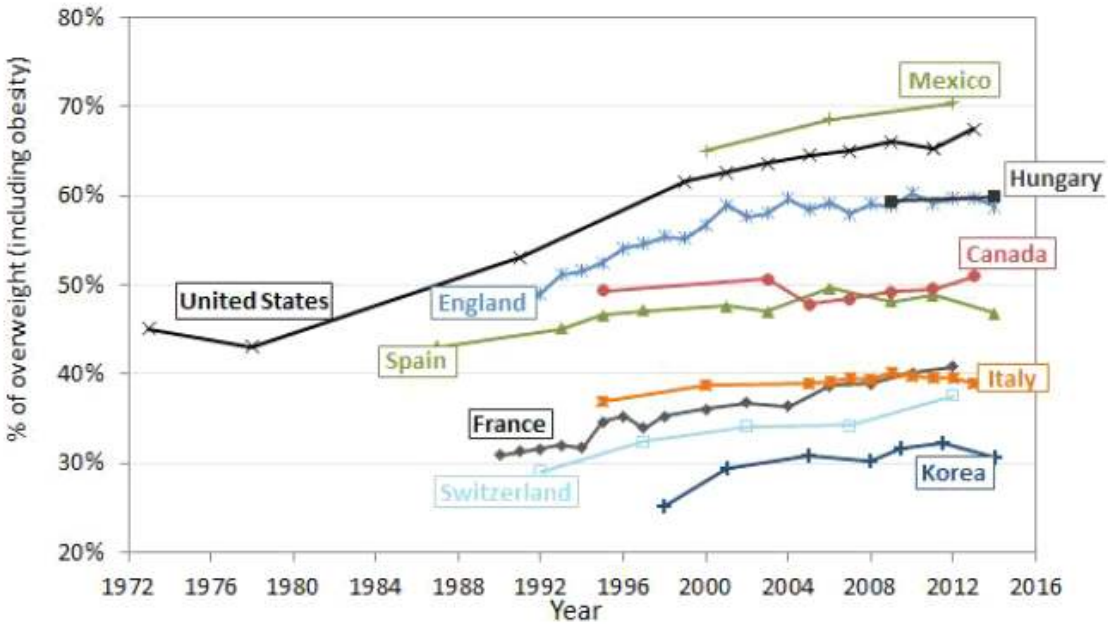
본 연구의 데이터는 학습(training) 데이터와 실험(test) 데이터로 구성된다. 학습데이터는 국내의 8도중 하나인 강원도의 데이터를 이용하며, 실험 데이터는 강원도 내의 도시 중 강릉과 원주의 데이터를 이용한다. 본 연구의 평가 비교 대상으로는 과거의 흐름을 반영하는 최소자승법 예측 기법을 선정하여 비교한다. 강릉의 경우, 본 연구에서 제안한 방법과 최소자승법 결과의 오차율 평균은 1.22%로, 최소자승법을 통해 예측된 결과와 크게 다르지 않음을 알 수 있다. 따라서 이를 통해, 본 연구에서 제안하는 방법이 과거의 흐름을 기반으로 작성됨을 확인할 수 있다. 하지만, 최소자승법을 이용하여 예측하는 방법은 단순하게 비만율 하나의 요소에 대한 흐름만을 보는 것이며, 실제로는 비만율에 여러 요소가 복합적으로 작용하기 때문에, 여러 요소의 흐름을 기반으로 예측하는 본 연구의 예측 방법이 유의미한 방법이라 여겨진다.

본 연구는 특정 지역의 데이터를 이용하여 진행되지만, 향후 타 지역으로도 확대될 수 있다. 또한, 수집된 데이터의 양이 늘어나고 예측 방법이 보완된다면, 보다 정확한 미래의 비만율 예측이 가능할 것이라고 여겨진다. 또한, 국내뿐만 아니라 국제사회에서도 큰 문제인 비만율을 보다 정확하게 예측할 수 있다면, 이를 이용하여 문제를 해결하고 대비할 수 있을 것이라 여겨진다.

■ 서론

21세기가 되면서 세계는 이전과 비교 할 수 없을 정도로 많은 분야에서 발전을 거듭해 왔다. 특히 과학기술이 이전과는 비교할 수 없을 정도로 발달함에 따라 여러 다른 분야에서의 발전을 촉진시켰다. 이에 따라 이전에는 없던 많은 일자리가 생겨났고 사람들의 삶은 점점 더 풍족해 졌다.

이 변화는 우리나라에도 똑같이 적용되었다. 과거의 우리나라가 먹고 살기에 급급했다면 21세기에 들어서면서 먹을 음식이 없어 굶는 경우는 거의 보기 힘들어 졌다. 사람들은 먹고 사는 문제가 해결되니 사치품이나 여가 및 문화생활에 관심을 가지게 되며 이전과는 비교할 수 없는 행복한 삶을 누릴 수 있게 되었다[1].

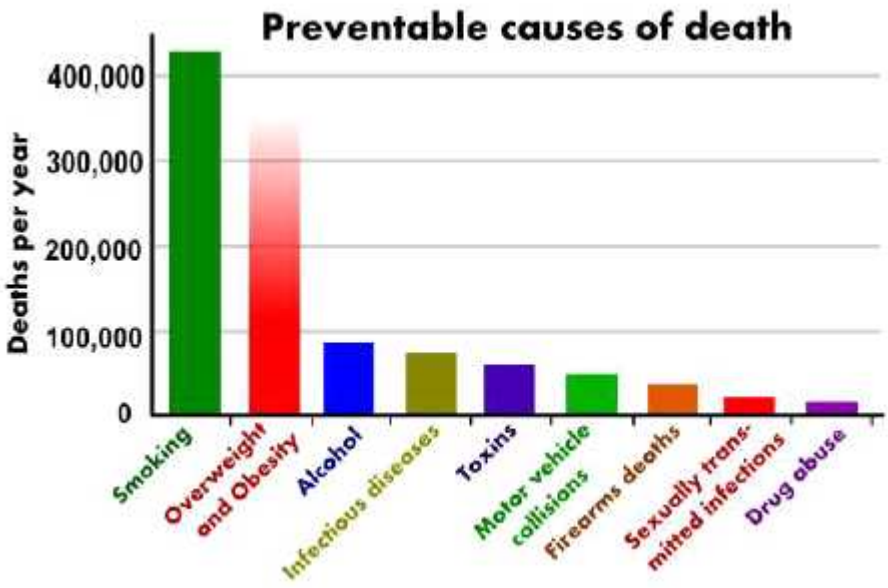


[그림 1] 주요 OECD 국가들의 연도별 과체중률(비만 포함) 변화

하지만 이러한 발전이 긍정적인 변화만을 초래한 것은 아니다. 사람들이 경제적으로 편하고 안정적인 삶을 누리게 됨에 따라 기존에 필요한 것 보다 많은 것을 요구하게 되었다. 특히 이 변화는 '먹는 것'에서 두드러져 왔다. 이전과는 달리 삶이 풍족해 짐에 따라 필요이상의 음식을 섭취하게 되었다. 이는 광고 또는 부모, 친구 등을 통한 사회적인 원인에 의해서 나타나기도 한다. 즉, 인간의 기본적인 욕구들 중 '하나인 살기위해 먹는 것'이 아니라 그것을 뛰어넘어 정신적인 만족 또는 다른 사회적 원인에 의해 필요 이상으로 음식을 먹는 시대가 온 것이다[2]. 이는 사람들의 체형과 건강에 직접적인 영향을 주게 되었다. [그림 1]에서 시간이 지남에 따라 사람들의 생활이 풍족해 지면서 과체중률이 늘어나는 것을 확인 할 수 있다[3]. 이 변화는 서구문화권에서 더 두드러진다. 근대화가 우리나라보다 더 빨리 일어난 서구 문화권 국가들의 사람들은 더 빨리 바쁘지만 풍족한 삶을 영유하게 되었고 이에 따라 급격한 체형변화를 맞이하게 되었다. 또한, 국내의 상황 또한 이와 유사한 모습을 보인다.

앞에서 언급한 서구문화권처럼 극단적이지는 않지만 짧은 시간에 비교적 상당한 수준의 경제성장을 이루게 되었고 이에 따라 먹는 음식의 양도 달라졌으며 이는 직접적인 체형 변화를 초래했다.

이처럼 국제사회에서 비만은 심각하게 다루어 지는 문제점들 중 하나이다. 사람들이 필요 이상의 음식을 섭취함으로서 평균 체중은 점점 늘어가고 고도비만에 해당되는 사람들도 급격하게 증가했다. 하지만 더 큰 문제는 외형적 변화에 그치는 것이 아니라 이 증상이 다양한 합병증을 유발하여 건강을 위협한다는 점이다. 이 합병증들은 건강에 굉장히 치명적이며 심각한 경우에는 목숨을 위협하기도 한다.



[그림 2] 미국의 연간 예방 가능한 사망 원인별 사망자 수

[그림 2]에서 볼 수 있다시피 비만은 여러 원인들을 제치고 사망원인들 중 상위권에 위치해 있다[4]. 비만은 절대 가볍게 다루어야 할 문제가 아닌 사람의 목숨과 직결되는 중요한 문제로 작용되고 있다.

본 연구에 앞서 연구 대상을 한정하는 것이 선행되어야 한다. 도시마다 성장정도가 크게 다르기 때문에 단순히 대한민국 전체에 관해 조사하는 것은 큰 의미가 없으며 또한 이미 발전이 충분히 된 도시들에 대해 연구하는 것은 예상한 결과를 내기에는 적절하지 않다. 또한 광역시와 특별시 그리고 8개의 도에 대해서는 경제 성장이 여러 통계수치에 의해 구체적인 수치로 나타나지만, 내부 도시에 대해서는 각각의 경제 성장 수치를 얻기 힘들다. 따라서 이를 대변할 여러 지표들을 통해 특정 도시의 경제 성장 수치를 얻는다.

이를 위해서는 도시의 성장와 어느 정도 인과관계를 갖는 지표들을 찾는 것이 중요하다. 국내의 공공기관인 통계청에서 이와 관련해 다양한 공공데이터들을 찾을 수가 있다. 하지만 앞서 말한 것과 같이, 경제성장의 제일 직접적인 지표로 볼 수 있는 1인당 소득 즉, GDP에 관한 데이터가 특별시, 광역시 및 도 단위로만 제공된다. 따라서 GDP이외에 경제성장을 대변할만한 다른 지표들을 이용하여 경제 성장 지표를 생성하고 비만율과의 관계를 학습한다.

그리고 이를 이용하여, 미래의 비만율을 예측하는 것을 본 연구의 목표로 한다.

따라서 본 연구는 다음과 같이 진행한다. 우선 GDP와 제공되는 여러 지표를 비교하여 경제성장을 대변할 지표를 선정한다. 그 다음으로는, 선정된 지표와 측정된 비만율을 이용하여 Lasso regression을 기반으로 학습 모델을 구축한다. 마지막으로 특정 도시의 데이터를 이용하여 비만율을 예측한다. 이 예측된 비만율의 정확도를 측정하기 위해 최소자승법을 이용한 결과와 비교한다.

바로 다음의 '관련 연구' 에서는 연구에서 사용된 여러 기술들에 대해 언급하고 간단한 부연 설명을 할 것이다. 그 다음 장인 '제안 작품 소개' 에서는 어떻게 연구를 진행했는지 자세하게 설명하고 바로 다음 장인 '구현 및 결과 분석' 에서는 구체적인 결과 데이터들을 살펴본 후 이 데이터들이 연구를 진행하기 전에 설정했던 목표와 부합하는지를 살펴 볼 것이다. 마무리 부분인 '결론 및 소감' 에서는 구현 결과를 정리하고 연구를 진행하며 느꼈던 점들을 서술할 것이다.

■ 관련연구

1. 머신러닝(Machine Learning)

머신러닝은 이름에서 알 수 있듯이 기계를 즉, 컴퓨터를 인간처럼 학습시켜 스스로 규칙을 형성하는 방법이다. 주로 통계적인 접근방법을 사용하며 인간이 하는 추론 방식과 매우 유사하며 강력하다. Tom M. Mitchell의 'Machine Learning'에서는 기계학습을 어떠한 작업(Task)에 대해 꾸준한 경험(Experience)을 통하여 그 T에 대한 성능(Performance)을 높이는 것이라고 정의한다[5]. 즉, 데이터를 의미하는 E가 가장 중요하며 좋은 품질의 데이터를 많이 가지고 있다면 보다 높은 성능을 끌어낼 수 있다는 것을 의미한다.

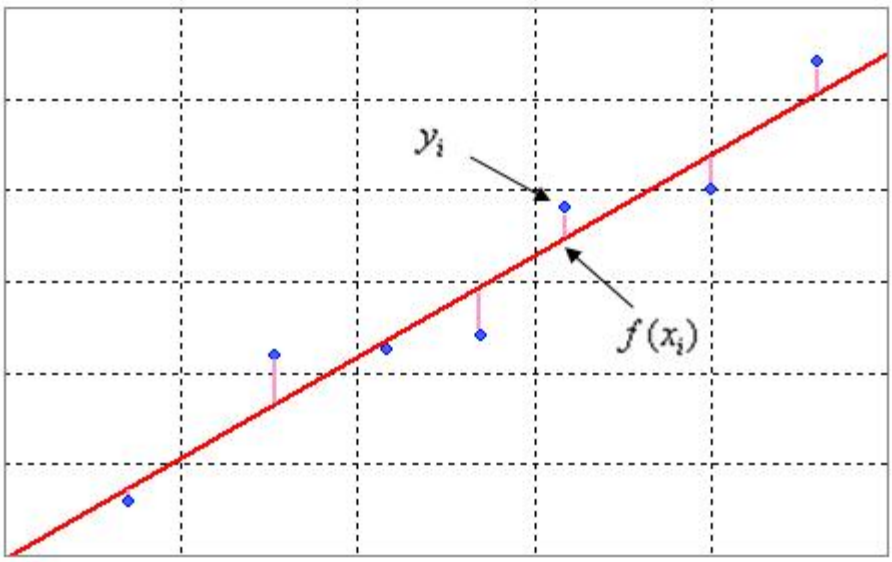
1-1. 회기분석(Regression Analysis)

통계학에서 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석방법이다. 특히 비선형 회기 분석 방법은 공학에서 뿐만 아니라 수학 또는 사회과학 분야에서 널리 사용되는 기법이다[6]. 회기분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과관계의 모델링 등의 통계적 예측에 이용될 수 있다.

1-2. 최소자승법(Method of Least Squares)

일반적으로 어떤 실험을 행할 때, 변량 x (독립변수 Independent Variable)를 변경하며, 그에 따른 실험값 y (종속변수 Dependent Variable)의 쌍 (x, y)을 얻는다. 실험을 N회 반복하여 (x_1, y_1), (x_2, y_2), … (x_n, y_n)의 데이터를 확보했다고 하자. 이 수많은 데이터들이 일정한 규칙성을 갖지 못한다면, 이 실험은 아무런 의미를 갖지 못한다. 따라서 데이터들의 유용성을 판단하기 위해서 가장 먼저 해야 할 작업은, 두 변수 간에 상관관계 여부 확인과 어떤 상관관계를 갖는지 찾는 것이다. 또한, 변수 간의 상관관계를 함수로 표현할 수 있다면, 변수 간의 규칙성이 있음을 의미한다.

수학적으로, N회 측정한 측정값 $y_1, y_2, ..., y_n$이 어떤 다른 측정값 $x_1, x_2, ..., x_n$ 의 함수라고 추정할 수 있을 때, 측정값 y_i와 함수값 $f(x_i)$의 차이를 제곱한 것의 합이 최소가 되도록 하는 함수 f(x)를 구하는 것이 최소자승법의 원리이다. 이렇게 해서 구해진 함수 y=f(x)는 이 측정값들의 관계에 가장 적합한 함수라고 할 수 있다.



[그림 3] 최소자승법의 이해를 위한 그래프

위 [그림 3]에서 표시된 각 점들은 측정값 (x_i,y_i)이고, 직선 $(x_i,f(x_i))$은 최소자승법을 사용해 구한, 측정값들의 분포를 가장 잘 나타내는 일차함수이다. 즉, 이 함수는 (측정값-함수값)^{2}의 총합(오차의 총합)이 최소가 되는 직선이다.

최소자승법은 Microsoft Office Excel에서 제공하는 추세선 기능을 통해 쉽게 이용할 수 있다. 본 연구에서는 이 기능을 통해 데이터에 해당하는 함수를 구하여 미래의 데이터를 예측한다.

2. 파이썬 라이브러리 (Python Libraries)

2-1. pandas, matplotlib, seaborn, numpy

파이썬에서는 matrix연산과 다양한 시각화를 지원 등 데이터 분석에 필요한 기능을 제공하는 라이브러리로 pandas, matplotlib, seaborn, numpy가 있다. pandas는 DataFrame, Series와 같은 도구를 통해 데이터 객체를 이용해서 데이터를 쉽게 가공하며, 평균, 분산, 최대, 최소 등을 쉽게 연산할 수 있다. 그 외에도 변수사이의 연관성, 그룹, 선택, 조인 등의 다양한 함수를 통해 matrix를 효율적으로 쉽게 가공할 수 있다. matplotlib은 데이터의 분포 및 패턴을 차트(chart)나 플롯(plot)로 시각화해준다. seaborn은 matplotlib를 기반으로 하며 통계 수치들을 다양한 방법으로 시각화시키는 데 사용된다. numpy는 고성능 수치해석 및 통계관련기능을 구현해준다.

2-2. scikit-learn

파이썬으로 구현된 scikit-learn은 기계 학습 오픈소스 라이브러리 중 하나이다. scikit-learn

의 장점은 라이브러리 외적으로는 scikit 스택을 사용하고 있기 때문에 다른 라이브러리와의 호환성이 좋다. 또한, 내적으로는 통일된 인터페이스를 가지고 있기 때문에 매우 간단하게 여러 기법을 적용할 수 있어 쉽고 빠르게 최상의 결과를 얻을 수 있다. 본 연구에서는 해당 라이브러리의 Lasso regression을 이용하여 미래의 데이터를 예측한다.

■ 제안 작품 소개

본 연구에서는 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.

1. 경제 성장 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 경우 GDP를 직접적인 지표로 선정할 수 있지만, 이는 특별시, 광역시 및 도 단위로만 제공되기 때문에, 본 연구에서 초점으로 하는 특정 도시에 적용이 불가능하다. 따라서 본 연구에서는 도시별로 제공되며 경제성장을 대변할 수 있는 지표를 선정하여 이를 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.

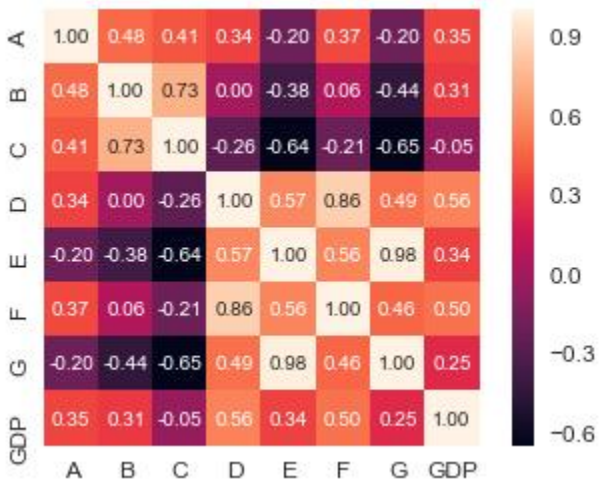
경제성장을 대변하는 지표를 선정하기 위해, 우선 도 단위의 1인당 GDP와 [표 1]의 지표 간의 상관관계를 분석한다. 상관관계를 분석을 위하여, 파이썬 라이브러리 matplot을 이용하였으며, 데이터의 범위는 2008년부터 2015년이다.

1) http://kosis.kr

[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: 0은 각 도의 이름을 나타내며, 편의상 1~8의 숫자로 대체한다. Unnammed: 1은 년도를 의미한다.



[그림 4] 지표 데이터와 GDP의 관련성을 표현한 히트맵

[그림 4]는 지표 데이터와 GDP 간의 상관관계를 표현한 히트맵이다. [그림 4]와 각 칸의 숫자는 피어슨 상관계수이며, [표 3]과 같은 관계를 의미한다. [그림 4]의 제일 마지막 행을 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

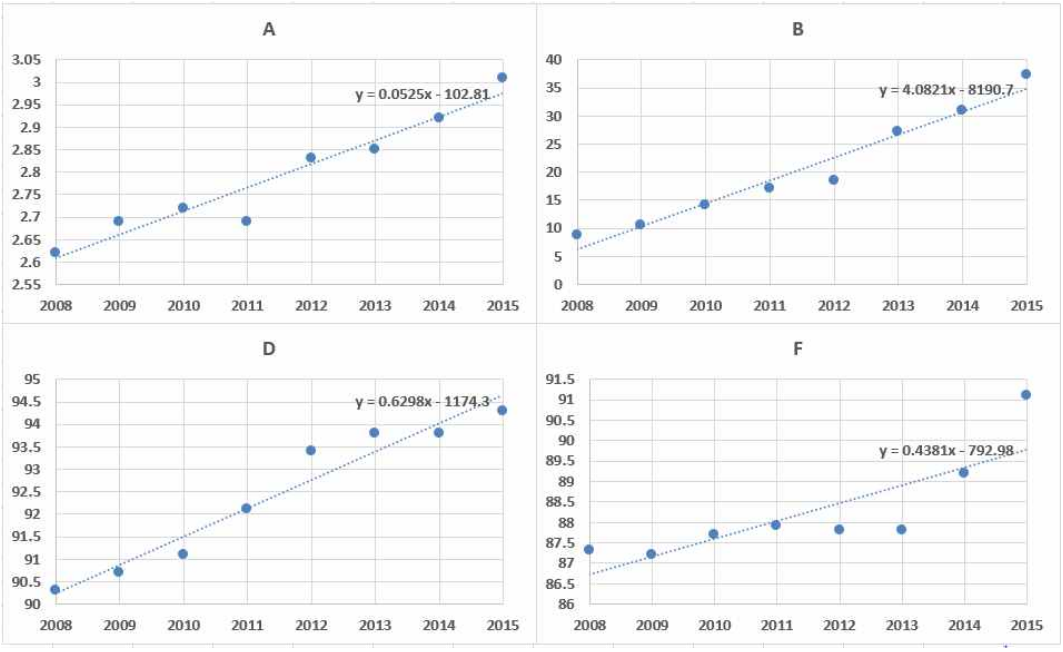
[그림 4]와 [표 3]에 따르면 A, B, D, E, F 지표가 강한 양의 상관관계를 가짐을 알 수 있다. 하지만, 본 연구는 선형회귀 방식으로 학습 모델을 구축하였기 때문에, 비율을 나타내는 다른 지표와 달리, 수를 나타내는 E 지표의 경우, 차이가 많이 발생하여, 사업체수를 나타내는 E데이터가 학습데이터와 실험데이터 사이의 차이가 너무 커서 제대로 예측되지 않는 현상이 발생하므로 E데이터는 학습 및 실험데이터에서 제외한다. 따라서 본 연구에서는 A, B, D, F 4개의 지표를 경제성장 대변 지표로 선정하였다.

2. Lasso Regression 학습 모델 구축

본 연구에서는 특정 도시에 초점을 맞추어 경제 성장에 따른 비만율을 예측한다. 따라서 학습 모델을 구축할 때, 해당 도시가 속해 있는 도를 선정하여 학습 데이터를 생성한다. 본 연구에서는 데이터가 수집되는 시기에 성장률이 높은 경향이 띄는 도시에 초점을 맞추었기 때문에, 강원도를 선정하였으며, 데이터는 2008년부터 2015년 범위의 데이터를 이용하였다. 실험 데이터로는 강원도 내부의 도시 중, 강릉과 원주를 선정하여 두 도시의 데이터를 이용하여 연구를 진행한다.

3. 모델 평가를 위한 최소자승법 평가 모델 구축

다음으로 본 연구의 학습 모델 평가를 위한 평가 모델을 구축한다. 평가 모델은 최소자승법을 이용하여 구축하였다. 추세선을 기반으로 두 변수의 관계를 이용하여 미래의 데이터를 예측한다.



[그림 5] 강원도 내부 도시 강릉의 미래 지표 데이터 예측

[그림 5]는 앞 단계에서 선별된 성장지표(A, B, D, F) 데이터를 이용하여, 추세선과 상관관계 함수를 구한 그림이다. 이 함수를 통해, 연도별 데이터 변화량을 분석하여 다음의 데이터를 예측할 수 있다. 본 연구의 평가 모델은 각 성장지표별로 추세선을 이용하여 다음의 데이터를 예측하고, 예측된 데이터를 이용하여 비만율을 구하는 방식을 취한다.

■ 구현 및 결과분석

[표 4] 강원도의 7개 지표와 GDP 연간 데이터

강원도	A	B	C	D	E	F	G	GDP
2008	2.08	11.8	9.28	85.5	117150	76.4	361	11194
2009	2.14	15.2	10.77	86.1	117569	77	353	11908
2010	2.18	17.45	10.65	86.5	118266	81.3	353	12499
2011	2.17	19.2	10.41	87.5	121273	83.1	353	12775
2012	2.2	20.28	10.85	87.9	125192	84.1	352	13304
2013	2.25	23.86	11.8	88.6	129403	85	351	13600
2014	2.3	25.58	12.82	88.8	133314	85.6	351	14454
2015	2.35	26.52	13.29	89.6	133517	86	351	15142

이전 장에서 살펴보았다시피 본 연구에서는 처음에 선택했던 7개의 지표들 중 4개만을 선별하여 사용한다. [표 4]는 강원도에 대한 기존의 7개 지표와 GDP의 연간데이터이다. 파이썬의 matplotlib 라이브러리를 통해, 7개의 지표 중 GDP를 대신하여 경제성장을 대변할 수 있는지 여부를 확인하였다. 그 결과, GDP와 연관성을 가지면서 선형으로 데이터를 예측하기에 적합한 A, B, D, F 지표를 선정하였다.

[표 5] 강원도의 선별된 4개의 지표와 전년도 비만율 데이터

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

강원도를 최종 학습 데이터로 선정한 후, 본 연구에서 제안한 학습 모델을 구축한다. [표 5]는 학습 모델을 위해, 전처리가 완료된 학습 데이터 표이다. 데이터 전처리 단계에서는, 경제 성장을 대변하는 A, B, D, F 지표 데이터와 전년도의 비만율 데이터 PreObesity를 생성한다. 현년도의 비만율을 예측할 때에는 전년도의 비만율를 기반으로 예측되어야 하기 때문에 즉, Time Locality를 반영해야 하기 때문에 학습 데이터를 생성할 때에 전년도 비만율을 의미하는 PreObesity를 넣는다.

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

[표 6] 강릉의 예측된 데이터

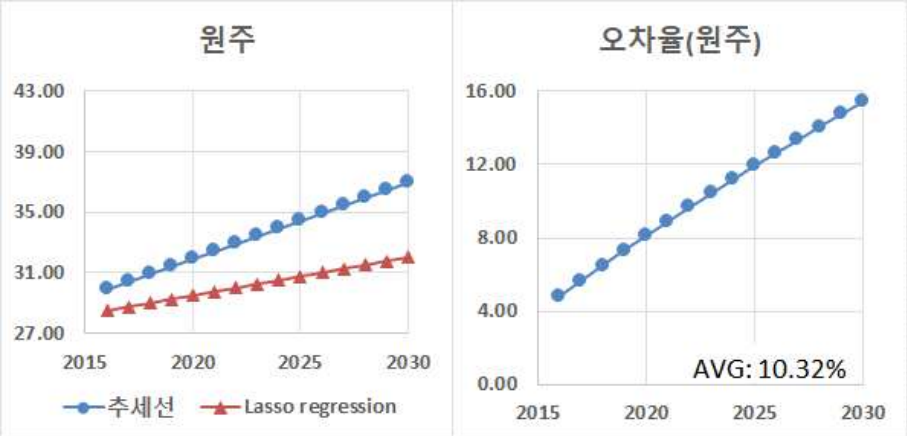
	A	B	D	F	Obesity
2008	2.62	8.7	90.3	87.3	23.2
2009	2.69	10.58	90.7	87.2	27.1
2010	2.72	14.19	91.1	87.7	24.3
2011	2.69	17.01	92.1	87.9	25.4
2012	2.83	18.37	93.4	87.8	25.7
2013	2.85	27.21	93.8	87.8	29.3
2014	2.92	31.05	93.8	89.2	28.2
2015	3.01	37.29	94.3	91.1	29.2
2016	3.03	38.81	95.38	90.23	29.86
2017	3.08	42.90	96.01	90.67	30.61
2018	3.13	46.98	96.64	91.11	31.36
2019	3.19	51.06	97.27	91.54	32.10
2020	3.24	55.14	97.90	91.98	32.85
2021	3.29	59.22	98.53	92.42	33.60
2022	3.35	63.31	99.16	92.86	34.35
2023	3.40	67.39	99.79	93.30	35.09
2024	3.45	71.47	100	93.73	35.84
2025	3.50	75.55	100	94.17	36.59
2026	3.55	79.63	100	94.61	37.34
2027	3.61	83.72	100	95.05	38.09
2028	3.66	87.80	100	95.49	38.83
2029	3.71	91.88	100	95.92	39.58
2030	3.77	95.96	100	96.36	40.33

[표 6]은 강릉의 지표 및 비만율을 예측한 결과표이다. 추세선을 이용한 최소자승법 평가 모델로 예측된 결과와 본 연구에서 제안한 학습 모델로 예측된 결과를 비교하여 학습 모델을 평가 한 결과는 다음과 같다.

[그림 6] 강릉의 비만율과 오차율

[그림 6]은 2016년부터 2030년까지의 강릉의 비만율을 본 연구의 학습 모델과 평가 모델을 통해 예측한 결과와 그 오차율이다. 왼쪽 그래프를 통해 본 연구방법을 통해 예측한 비

만율 그래프와 과거의 경향을 따르는 추세선 그래프가 큰 차이가 없음을 알 수 있다. 오른쪽 그래프를 통해 연도별 오차율을 확인할 수 있으며 평균 오차율은 1.22%이다. 이는 미래의 비만율이 과거의 비만율의 경향을 반영한다는 것을 의미하며 지표 데이터의 경향 또한 이와 유사함을 보여준다.



[그림 7] 원주의 비만율과 오차율

[그림 7]은 2016년부터 2030년까지의 원주의 비만율을 본 연구의 학습 모델과 평가 모델을 통해 예측한 결과와 그 오차율이다. 왼쪽 그래프를 보면 본 연구방법을 통해 예측된 비만율 그래프와 추세선의 차이가 강릉의 경우보다 큼을 알 수 있다. 또한, 오른쪽 그래프를 통해 오차율이 해마다 크게 증가하며 평균은 10.32%로 다소 높음을 알 수 있다. 즉, 특정지역의 성장흐름과 단순한 비만율의 흐름은 비슷하지만 일치하지 않는다. 이는 비만율의 흐름만으로는 미래의 비만율을 예측하는 데 사용할 수 없으며 정확한 비만율 예측을 위해서는 여러 가지 요소를 고려해야함을 의미한다. 따라서 본 연구에서 사용한 방법이 유의미함을 나타낸다.

■ 결론 및 소감

1. 결론

본 연구에서는 Lasso Regression을 기반으로 하여 지역의 경제 성장에 따른 미래의 비만율을 예측하는 방안을 제시한다. 특정지역의 성장률은 연속적으로 변하며 일정한 흐름을 가지고, 이 성장률이 비만율과 특정 관계를 가진다는 가정 하에 3단계로 나누어 연구를 진행한다. 1단계에서는 GDP수치를 대변할 수 있는 여러 지표를 선별하여 가상의 GDP를 구한다. 2단계에서는 이 가상의 GDP수치와 비만율 데이터를 학습모델을 통해 학습시킨다. 마지막 3단계에서는 미래의 성장을 예측하고 학습 모델에 적용시켜 미래의 비만율을 예측한다.

본 연구의 데이터는 크게 학습데이터와 실험데이터로 이루어진다. 학습데이터는 국내의 8도중 하나인 강원도의 데이터를 이용하고 실험데이터는 강원도 내의 도시 중 강릉과 원주의 데이터를 이용한다. 이 데이터들을 최소자승법을 이용하여 다음의 지표 데이터 및 비만율을 계산한다. 예측된 지표 데이터를 본 연구방법을 통해 미래의 비만율을 예측하는 데 사용한다. 이 비만율 데이터를 본 연구의 평가 비교 대상인 최소자승법을 통해 예측된 비만율과의 오차율을 구한다. 강릉의 경우 1.22%로 미래의 비만율이 현재까지 비만율의 흐름을 따르며 특정지역의 성장 또한 경우에 따라 유사한 흐름을 보임을 확인할 수 있다. 하지만 원주의 경우 10.32%로 강릉의 경우와는 달리 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.

본 연구는 강원도의 강릉, 원주를 대상으로 진행되었지만, 더 많은 데이터가 수집되고 예측방법이 개선된다면 다른 지역에도 적용할 수 있을 것이며, 보다 정확하게 미래의 비만율을 예측할 수 있을 것이다. 또한 국내뿐만 아니라 다른 나라의 도시에도 이를 적용함으로서 국제사회의 큰 문제인 비만을 효과적으로 대처할 수 있을 것이라 생각한다.

2. 소감

지금까지 <Lasso Regression을 이용한 지역 경제 성장과 비만율의 상관관계 분석> 라는 주제로 연구를 진행하였다. 연구를 위해 도시의 경제성장을 대변할 수 있는 여러 지표들을 선정하여 연구를 진행하였으며 나름 성공적으로 비만율을 예측하였다고 생각한다. 하지만 빅데이터의 중요성이 대두된 지 오래되지 않았기 때문에 데이터가 아직 충분히 축척되어 있지 않아서 정확히 예측하는 데 한계가 있었다. 또한 본 연구방법을 통해 예측된 데이터를 비교할만한 데이터가 존재하지 않아 가상의 데이터와 비교했다는 점이 아쉽다.

이미 현대 사회는 필요 이상으로 많이 먹을 뿐만 아니라 균형 잡히지 않은 식습관이 만연해 있다. 이로 인해 사람들의 평균 체중은 점점 늘어가며 특히 업무로 인해 바쁜 현대인들이 이를 간과하는 것은 치명적인 결과를 가져올 수도 있다. 현재 우리나라는 과거의 어느 때 보다 발전하고 있으며 이에 따라 비만인 사람들의 수도 점점 더 많아 질 것이다. 본 연구에서 사용한 방법들이 추후에 더 나은 방법으로 개선되어 이러한 문제를 효과적으로 대처할 수 있기를 기대해 본다.

■ 참고문헌

[1] 이경호, "21 세기 한국사회 전망/[복지분야] 21 세기 사회에 대비한 보건복지분야의 주요 정책과제." 한국행정연구 9,1 (2000)

[2] 이명해, "비만의 원인과 평가에 관한 고찰." 부산여자전문대학 논문집 17 (1995)

[3] Prevalence of overweight http://www.oecd.org/health/obesity-update.htm/

[4] Obesity in America http://www.centralohiobariatrics.com/obesity-in-america/

[5] Tom M. Mitchell, Machine Learning (McGraw-Hill International Editions Computer Science Series) 1st Edition (McGraw-Hill,1997)

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

■ 부록

1. Lasso regression을 이용한 비만율 예측 코드(python)

```
from sklearn import linear_model


# 강원도 데이터 학습
X = [
        [2.14, 15.2, 86.1, 77, 25.8],
        [2.18, 17.45, 86.5, 81.3, 27.1],
        [2.17, 19.2, 87.5, 83.1, 27],
        [2.2, 20.28, 87.9, 84.1, 27.6],
        [2.25, 23.86, 88.6, 85, 27.7],
        [2.3, 25.58, 88.8, 85.6, 27.8],
        [2.35, 26.52, 89.6, 86, 28.5]
]

Y = [27.1, 27, 27.6, 27.7, 27.8, 28.5, 30.2]

# Lasso Regression 함수 실행

clf = linear_model.Lasso()

clf.fit(X, Y)

print(clf.predict([
        # 강릉, 원주 데이터 입력
]))
```

2. 성장 지표와 GDP의 관계를 표현하는 히트맵 출력 코드(python)

```
%matplotlib inline
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

#데이터 입력
se = pd.read_csv(r'data_revised.csv')
se

#히트맵 구조 작성
cols = ['A','B','C','D','E','F','G','GDP']
cm = np.corrcoef(se[cols].values.T)
sns.set(font_scale=1.2)

#히트맵 출력
hm = sns.heatmap(cm, cbar = True, annot = True, square = True, fmt = '.2f',
annot_kws = {'size':10}, yticklabels=cols, xticklabels=cols)

plt.show()
```