Journal of Korea Port Economic Association
Vol.37, No.1, 2021, pp.197-215.
https://doi.org/10.38121/kpea.2021.03.37.1.197

릿지 회귀와 라쏘 회귀 모형에 의한 부산 전략산업의 지역경제 효과에 대한 머신러닝 예측

이재득*

Machine Learning Prediction of Economic Effects of Busan's Strategic Industry through Ridge Regression and Lasso Regression

Yi, Chae-Deug

Abstract

This paper analyzes the machine learning predictions of the economic effects of Busan's strategic industries on the employment and income using the Ridge Regression and Lasso Regression models with regulation terms. According to the Ridge estimation and Lasso estimation models of employment, the intelligence information service industry such as the service platform, contents, and smart finance industries and the global tourism industry such as MICE and specialized tourism are predicted to influence on the employment in order. However, the Ridge and Lasso regression model show that the future transportation machine industry does not significantly increase the employment and income since it is the primitive investment industry. The Ridge estimation models of the income show that the intelligence information service industry and global tourism industry are also predicted to influence on the income in order. According to the Lasso estimation models of income, four strategic industries such as the life care, smart maritime, the intelligence machine, and clean tech industry do not influence the income. Furthermore, the future transportation machine industry may influence the income negatively since it is the primitive investment industry. Thus, we have to select the appropriate economic objectives and priorities of industrial policies.

Key words: Strategic Industry, Ridge Regression, Lasso Regression, Employment, Income

▷ 논문접수: 2021. 03. 16. ▷ 심사완료: 2021. 03. 29. ▷ 게재확정: 2021. 03. 29.

* 부산대학교 무역학부 교수, givethanks@pusan.ac.kr

Ⅰ. 서 론

현재는 4차 산업혁명과 인공지능(Artificial Intelligence; AI)의 시대이며, 기계는 인간과 같은 지능을 갖추고 생산 시설이 자동화되고 정보통신 기술의 발달로 대부분의 단순 노동집약적인 작업들이 기계와 시스템에 의해 대체되어 고용과 소득이 감소할 수도 있다. 그러나 인공지능의 발달은 다른 한편으로 새로운 업무영역을 창출하여 새로운 고용을 만들어 낼 수도 있고 소득을 증가시킬 수도 있다.

부산시는 1980년대 이래 침체된 지역경제의 활성화와 낙후된 부산지역 산업 및 경제구조의 고도화를 위해 2019년 부산의 지역경제를 성장시키고 지속가능한 미래를 창출하기 위한 제 5차 7대 전략산업을 새롭게 재선정하고 산업 경쟁력을 강화하기 위해 많은 노력을 하고 있다. 그리하여 부산은 2020년 현재 4차 산업혁명과 AI 시대에서 신성장산업혹은 전략산업 활성화에 힘쓰고 있다.

그러나 부산 지역경제 활성화와 지역산업의 경쟁력을 제고시키기 위해서는 먼저 부산경제와 산업경쟁력에 대한 진단과 좀 더 정확한 예측과 추정이 필요하다. 물론 몇몇 기존 연구들이 있지만 주로 전통적인 구조적 모형이나 시계열 계량기법들을 사용하고 있다. 물론 부산지역 경제에 대한 기존 연구들은 나름대로 공헌한 점이 있으나 단순히 전통적 구조적 모형은 모형설정의 오류가 있을 수 있으며, 전통적 시계열 계량경제 기법에 의한 분석에 의한 과잉적합(overfitting) 등의 한계가 있다. 특히 전통적인 시계열 예측 모형은 계절 변동과 추세가 잘 반영되어 있고, 자료의 패턴이 반복되는 경우에는 예측력이 높게 나올 수 있으나, 최근 같이 전세계적으로 코로나가 발생하여 세계경기가 불규칙하고 변동성이 큰 경우에는 예측력이 많이 떨어져, 예측에 있어서는 정확도가 많이 낮아지는 한계점을 가진다.

그러므로 현재와 같은 불확실한 경제상황에서는 이들 전통적 계량경제 모형에 전통적 모형의 설정에 의한 추정은 모형의 설정의 오류와 추정방법의 한계로 인하여, 예측오차가 더 많이 생길 수 있으며, 또한 심각한 과잉적합의 문제가 발생하여 자칫 과장되거나 혹은 잘못된 추정과 예측을 유도하기도 한다. 그리하여 이러한 전통적 계량모형에 의한 추정과 예측의 한계점을 보완하고 극복하기 위하여 최근 머신러닝(Machine Learning) 등의 기법을 경제분석에서 외국 일부분의 학자들이 시도하고 있다.

우리나라에서도 머신러닝에 의한 경제학적 분석을 시도하고 있지만, 경제학적 분석에서 개괄적인 동향과 소개만 있을 뿐, 특히 경제 진단과 예측 측면에서 다양한 머신러닝 기법에 의한 전문적이고 심층적으로 분석한 연구는 거의 없다. 물론 단편적인 머신러닝 기법들이 물류, 항만 등에서 일부 사용되고 있으나, 특히 지역차원의 경제분석에 있어 머신러닝 기법은 거의 소개도 안 되어 있고, 머신러닝에 의한 분석은 전혀 없는 실정이다.

그러나 AI시대 경제학과 지역경제 분석에서는 경제 분석, 지역경제의 신성장 혹은 전략 산업과 그에 따른 경제 정책 수요 예측 등과 같은 분야에서 최근 소개되고 도입되어 활용되고 있는 실정이며, 향후 경제분석에서 좀 더 정확한 경제진단과 예측의 정확성을 위해 점점 더 넓어지게 될 것이다.

부산시 역시 부산지역의 경제 활성화를 위해 많은 노력을 경주하고 있다. 이를 위해서 부산시는 전략산업의 선정 및 육성을 통한 지역경제 성장과 발전을 위한 여러 가지 정책지원을 하고 있으나, 효과적이고 효율적인 전략산업의 선정과 육성을 위해서는 먼저 이러한 전략산업들의 지역경제에 미치는 영향에 대한 엄밀한 추정과 예측이 필요하다. 그리고 이러한 경제분석을 통해서 전략산업 선정과 지역경제 진단과 추정, 그리고 그에 따른 새로운 해석 등을 한다는 것은 향후 유효한 경제정책을 위

해서 매우 중요한 과제이다.

그리하여 현재 인공지능의 시대에 걸맞는 기계학습들인 머신러닝 분석기법들을 활용하여 전략산업들이 소득과 성장 등에 어떤 영향을 미치는지 좀 더 정확한 예측과 추정을 해볼 필요가 있다. 이를 위해 본 연구에서는 2019년에 선정된 부산의 7대 전략산업들인 글로벌관광, 라이프케어, 미래수송기기, 스마트해양, 지능정보서비스, 지능형기계, 그리고 클린테크 산업들을 7개 산업들을 독립변수로 선택하였고, 부산의 지역경제 활성화 지표에서 가장 중요한 고용과 소득을 각각 종속변수로 채택하여 추정하였다. 7대 산업들은 22개 세부 유망 산업분야로 구성되어 있는데 이 모든 자료를 중심으로 부산시 전략산업의 부산경제 즉 부산의 소득과 고용에 대한 영향을 예측하여, 모형방법과 모형들의 인자들에 의해 그 모형들의 예측력을 비교해보고자 한다.

따라서 본 연구는 AI 시대 머신러닝 기법을 사용하여 2019년에 선정된 부산시의 7대 전략산업들의 22개 분야의 세부산업에 대해 과도적합 문제를 극복하기 위하여 규제항을 추가하여 최근 도입되고 있는 머신러닝 기법을 도입하여 릿지 회귀와 라쏘 회귀 추정예측 모형을 설정하여 비교분석한다. 그리하여 부산 지역경제 활성화를 위한 전략산업을 중심으로 부산 지역경제, 특히 소득과 고용에 대한 예측 모형을 도입하여 아래와 같이 초점을 가지고 분석한다.

첫째, 2019년 다시 새롭게 선정된 7대 전략산업을 중심으로 부산 지역경제 산업들의 부산 지역경제에 있어 중요한 변수들인 부산의 고용과 소득에 미치는 영향을 먼저 규제항을 고려하지 않는 전통적인 계량방법인 통상적 최소자승법(OLS)을 사용하여 추정해본다.

둘째, 기존의 선형회귀분석에 규제항을 도입한 AI 지도 학습의 대표적 학습 알고리즘들인 릿지 회귀(Ridge Regression), 라쏘 회귀(Lasso Regression) 회귀분석 모형 등을 통하여 부산의 전략산업들의 소득과 성장에 대한 효과를 추정한다. 그리하여 기존의 전통적 계량경제분석에서 사용하는 선형분석에서 표출되는 과잉적합(Overfitting) 문제 등을 분석하고 그 결정계수와 예측력 등을 비교한다. 이를 위해 모형들의 결정계수들과 대표적인 예측 정확도 검증 방법인 평균제곱근오차(RMSE: Root Mean Squared Error)를 이용한다.

셋째, 본 연구는 특히 규제항을 고려하지 않는 전통적인 계량방법인 통상적 최소자승법(OLS)에 의한 과잉적합 문제를 해결하고 예측성을 높이기 위한 이러한 릿지 회귀분석 혹은 라쏘 회귀분석 모형에 머신러닝 기법에 의한 지역경제 분석은 연구가 거의 없는 생소한 연구이다. 따라서 기존 계량경제학적인 예측을 중심으로 한 기존연구들과 그 연구방법들에서 뚜렷한 차이점을 보인다.

넷째, 지역경제 차원에서 규제항을 추가한 머신러닝 기법에 의한 릿지 회귀 예측과 라쏘 회귀 예측 기법 등을 처음으로 사용하여, 향후 부산의 지역경제 활성화를 위한 부산의 전략산업들의 선정과 효과를 분석한다. 그리하여 본 연구는 향후 효과적이고 효율적인 전략산업 선정과 육성에 대한 정책적인 함의를 제시하는데 기여할 수 있을 것이다. 나아가 향후 부산경제 분석에 있어 과잉적합이나 오류를 줄이기 위해 위해서 규제항(Regulation Term)을 도입한 머신러닝 기법을 사용하는 후속연구들을 유발할 수 있을 것이다.

본 연구의 구성은 I장의 서론에 이어, II장에서는 선행연구를 살펴보고, III장에서는 예측모형을 살펴본다. IV장에서는 실증결과를 분석하며, V에서는 결론으로 맺는다.

II. 선행연구

경제정책의 목표는 결국 경제성장을 통한 소득과 고용의 증대로 귀착된다. 이를 위해서 부산시 역시

전략산업들을 선정하고 그 육성과 지원을 통해서 지역 경제성장과 고용창출과 소득을 증가시키려고 하고 있다. 물론 4-5년 마다 부산의 전략산업 선정의 변동과 자료수집의 한계 등이 분명히 존재하기 때문에 쉽지 않은 연구과제이다.

그럼에도 불구하고 전략산업이 지역 경제의 성장과 소득에 대한 미치는 영향과 파급효과에 대한 예측을 한다는 것은 중요한 연구과제이다. 물론 전통적 계량방법에 의한 분석은 이재득 · 윤진영(2018) 등이 있었으나, 인공지능 시대 좀 더 예측력이 높은 최근 머신러닝과 딥러닝 기법을 전문적인 지역 차원의 경제분석에서 릿지 회귀분석과 라쏘 회귀분석을 통한 머신러닝 기법을 이용한 추정과 예측 연구는 그 중요성에도 불구하고 전혀 없는 실정이다. 그리고 머신러닝에 대한 기존 선행연구를 물류 측면에서 살펴보면, Ding et al.(2019)은 2002년부터 2014년까지 중국 닝보항과 윈저우항의 연간 컨테이너 물동량 자료를 바탕으로 SVM(Support Vector Machine)과 인공신경망 모형, SVM과 인공신경망을 결합한 추정 모형을 활용하였다.

외국의 금융 관련 머신러닝에 의한 연구를 보면, Chakraborty and Joseph (2017)는 중앙은행에서 머신러닝에 대한 연구를 하였다. Naecker and Peysakhovich (2017)은 리스크의 애매모호한 행동모델을 평가하기 위한 머신러닝에 대한 연구를 하였다. Géron (2017)은 Scikit 학습과 머신러닝 그리고 텐서플로(Tensor Flow)에 대한 연구를 하였다. Lopez de Prado (2018)은 머신러닝에 의한 금융을 연구하였다. Kreif and DiazOrdaz (2019)은 정책평가에 대한 머신러닝 연구를 하였다.

머신러닝과 규제화(Regularization)에 대한 경제학 관련 연구를 보면, Tibshirani (1996)는 라쏘 분석을 통한 회귀 변수들의 축소(Shrinkage)와 선택(Selection)에 관한 연구를 하였다. Zou and Hastie (2005)는 규제화와 변수선택에 대한 연구를 하였다. Schlkopf and Smola (2001)는 서포트 벡터 머신과 규제화에 대한 커널연구를 하였다. Schapire and Freund (2014)는 부스팅(Boosting)에 대한 알고리즘과 머신러닝 학습에 대한 연구를 하였다.

Agrawal et al. (2018)은 단순한 인공지능 경제학으로 예측에 대한 연구를 하였다. Athey (2017, 2019)는 빅데이트 자료와 정책의 문제, 그리고 인공지능 경제학에 있어 머신러닝의 경제학에 대한 충격에 대한 소개를 하였다. Chalfin et al. (2016)는 머신러닝을 이용한 생산성과 인간자본에 대한 연구를 하였고, Jean et al. (2016)은 빈곤을 예측하는 데 있어 머신러닝 기법을 사용하였다. Mullainathan and Spiess (2017)은 응용계량적인 분석으로 머신러닝에 대한 연구를 하였다. Acemoglu and Restrepo (2019)는 미국 노동시장에 있어 로봇과 직업 대체에 관한 연구를 하였다.

그러나 머신러닝 기법들은 공학과 물류 등과는 달리 상대적으로 경제분석 측면에서는 아직 생소하고 초기적인 연구가 주를 이루고 있다. 특히 우리나라에서 비구조적 데이터를 이용한 텍스트 마이닝 기법을 이용한 연구들은 조금 있으나, 머신러닝을 이용한 경제학적 분석은 거의 없는 실정이다. 박기영과 고정원 (2019)은 머신러닝을 이용한 경제 분석에 대한 동향을 간략히 소개하였고, Kim (2020)은 거시경제와 금융시장 분석을 심층적인 머신러닝의 딥러닝 기법의 적용가능성을 연구하였다. 그러나 우리나라에서 여러 가지 머신러닝 기법을 이용하여 경제학과 지역경제 측면에서 분석한 것은 아직 없는 실정이다.

이와 같이 본 연구는 부산의 스마트해양산업 등 전략산업들이 부산경제 특히 부산의 고용과 소득에 어떤 영향을 미치는지 규제항을 도입한 모형을 가지고 예측한다. 그리하여 우리나라 뿐만 아니라 지역경제차원에서 머신러닝을 이용하여 한 릿지 회귀분석과 라쏘 회귀분석은 없기 때문에 최초의 연구가 될 것이다.

III. 경제예측 회귀 모형

1. 통상 최소자승법(OLS) 회귀모형 추정

전통적인 계량경제학적인 분석은 통상적인 최소자승법(OLS)에 의해 이루어 진다. 그리하여 통상적인 오차항에 대한 i.i.d 가정하에서 종속변수 y값은 y = f(x)= ax + b 형태로 설정되며 독립변수들은 일반적인 X들에 의해 설정된다. 그 때 이러한 선형모델의 독립변수들의 종속변수에 대한 회귀분석에 의해 추정되는 y값과 실제 값들에 대한 차이를 나타내는 잔차가 최소가 되는 추정회귀 계수를 구하는 것인데, 그 전형적인 선형회귀 모델은 다음과 같다. 즉 실제값과 예측값의 차이의 제곱의 합을 최소화 시키는 추정계수 들인 X 값들의 추정치인 β값들을 다음과 같이 구하고, 그 추정의 손실 혹은 비용함수(L)는 MSE에 의해 다음과 같이 정의된다.

$$\widehat{\beta}_0, \widehat{\beta} = \underset{\beta_0, \beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2$$

$$L(\beta) = MSE(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

2. 릿지 회귀(Ridge Regression)와 라쏘 회귀 (Lasso Regression) 모형 추정

그러나 위의 선형 회귀모델 추정은 모형의 오류와 오차의 성질 등에 의해 과잉적합의 심각한 오류가 발생할 수가 있는데, 이것을 극복하기 위하여, 중요한 변수를 선정하고 중요하지 않은 변수를 버리는 작업을 머신러닝 기법에서는 독립변수 혹은 특성치들의 선택(Feature selection)이라 하며, 중요하지 않은 변수에 해당하는 가중치 혹은 추정 계수의 크기를 제어하는 수축(shrinkage) 방법을 사용한다.

그리하여 다음과 같이 비용함수(L)를 최소화시키는 대표적인 회귀모델로는 계수의 제곱을 최소화시키는 L2 규제 방법인 릿지(Ridge) 회귀 방법과 계수의 절대값을 줄이는 L1 규제 방법인 라쏘(Least Absolute Shrinkage and Selection Operator :Lasso) 회귀 방법을 사용한다. 릿지(Ridge) 회귀 방법과 라쏘(Lasso) 회귀 방법은 과잉적합을 규제하기 위해서, 그 통제는 규제항(regulation term)의 크기에 따라서 결정된다. 여기서 규제항의 모 수(regularization parameter)인 값이 높아지면 추정계수는 0에 가까워져 모형은 단순화 되고, α값이 낮아질수록 계수가 커져 표준 선형회귀모형과 같아진다.

(1) Ridge Regression Model:

$$L(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 + \alpha \sum_{i=1}^{n} \beta_i^2$$

(2) Lasso Regression Model:

$$L(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 + \alpha \sum_{i=1}^{n} |\beta_i|$$

릿지(Ridge)와 라쏘(Lasso) 회귀모형에서 회귀계수(β)를 추정하는 방법은 아래 식들에서 나타나 있듯이 β값이 s 혹은 s2 이라는 특정한 임계값(threshold)보다 작을 때 다음 식들에 나타난 것과 같이 릿지(Ridge) 혹은 라쏘(Lasso) 회귀모형의 비용함수(L)을 최소로 만들어 주는 β값을 구하는 것이다.

$$\hat{\beta}^{Ridge} = Argmin(\beta)\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}X_{ij}\beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \le s^2$$

$$\hat{\beta}^{Lasso} = Argmin(\beta)\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}X_{ij}\beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \le s$$

위 식들에서 만약 s가 클 경우에는 일반적인 선형회귀식과 비슷하거나 같은 모델을 만들 것이다. 릿지(Ridge) 회귀모형의 단점은 회귀계수가 완전히 0이 되지는 않는다는 것인데, 이 경우 모형의 해석이 어렵다는 것이다. 규제항이 없는 기존의 방법인 변수 선택을 하면, 선택변수에 따라 모형의 변화가 심하고 자유도가 떨어진다. 하지만 라쏘(Lasso) 회귀모형에서는 s가 작아질 경우에는 변수들의 수가 축소된(shrunk) 라쏘(Lasso) 모델이 만들어 져서, 중요하지 않은 독립변수의 추정계수들은 0이 되어 사라질 것이다. 이는 릿지 회귀(Ridge)와 변수선택 각각의 장단점을 모두 고려한 것이다. 따라서 중요한 독립변수들만의 선택이 가능해진다.

여기서 s를 결정하는 것은 결국 변수가 많을 때 회귀모델에 있어서 다중공선성이 생길 수 있기 때문에 몇 개의 독립변수를 사용할 것인가의 문제로 귀착된다. 횡단면 유효성(cross validation) 기법으로 좋은 s를 선택하는 방법이 있지만, s의 값이 변함에 따라 라쏘(Lasso) 모델이 달라지기 때문에 추정계수 값들이 달라질 것이다. 일반적으로 s값이 커질수록 독립변수의 개수가 늘어나지만, 라쏘(Lasso) 회귀분석은 독립변수가 축소되어 릿지(Ridge) 회귀분석보다 상대적으로 설명력과 예측력이 낮아질 수 있다.

IV. 실증결과 분석

1. 부산의 전략산업 선정

부산시는 신성장산업을 육성하기 위해 부산시 의회의 승인을 받아 부산시의회 조례에서 전략산업을 공식적으로 선정하였다. 부산시는 〈표 1〉에서 나타난 것과 같이 2019년 3월~7월 기간 동안 부산의 전략산업 마스터플랜 육성계획을 수립하였고 〈표 1〉에서와 같이 제 5차 7개의 전략산업들과 22개의 세부 전략산업 분야를 새롭게 선정하였다.

표 1. 부산의 5차 7대 전략산업 범위와 22개 분야

구 분	주요 내용(범위 및 분야)
스마트해양산업	친환경스마트선박, 항만물류, 해양바이오, 수산가공
지능형기계산업	정밀기계, 스마트팩토리(생산자동화), 하이테크소재, 로봇
미래수송기기산업	자율주행차, 항공, 드론
글로벌관광산업	MICE, 특화관광(의료, 뷰티, 해양레저)
지능정보서비스산업	서비스플랫폼(ICBM), 콘텐츠(영상, AR, VR, 게임), 스마트금융(핀테크, 블록체인)
라이프케어산업	스마트헬스케어(의료기기, 의료서비스, 고령친화용품, 방사선의학과), 리빙케어(기능성식품, 항노화, 화장품), 라이프스타일(디자인, 신발, 패션의류)
클린테크산업	에너지시스템(태양광, 풍력, 수력), 에너지저장 및 서비스(ESS, 스마트그리드, 파워 반도체), 환경대원전해체, 기후변화)

자료 : 2019년 부산시

2. 릿지 회귀모형와 라쏘 회귀모형 분석

본 절에서는 2019년 선정된 부산의 제 5차 7대 전략산업들의 22개 세부 산업 분야를 중심으로, 머신러닝 기법의 부산의 소득과 고용에 대한 예측 모형을 도입한다. 이를 위해서 2019년 부산시에서 선

정한 5차의 7대 전략산업들인 글로벌관광(a), 라이프케어(b), 미래수송기기(c), 스마트해양(d), 지능정보서비스(e), 지능형기계(f), 클린테크(g) 산업들을 독립변수로 선택하였고, 먼저 부산의 고용을 종속변수로 채택하여 변수들 모두 로그값을 취하여 추정하였다.

부산에서 선정한 전략산업들이 지역경제의 가장 중요한 변수들인 소득과 고용에 어떻게 영향을 미치는지 분석하기 위하여 전통적 선형 계량경제 모델인 통상적 최소자승법(OLS)을 먼저 사용한 후, 머신러닝 추정방식인 릿지 회귀 (Ridge Regression) 분석과 라쏘 회귀(Lasso Regression)분석 방법으로 추정하고 예측력을 비교한다.

본 연구에서는 사용한 머신러닝 기법들에 의한 추정과 예측은 모두 머신러닝 회귀분석에서 많이 사용하는 파이썬(Python) 프로그램을 이용하여 코드를 개별로 작성하여 사용하였다. 그리고 머신러닝 기법을 사용하여 각각의 예측모형을 평가하기 위하여 본 연구에서는 주로 학습용 데이터 세트(train dataset)와 평가용 데이터 세트(test dataset)를 70%와 30%로 각각 나누었다. 그리고 파이썬 프로그램의 회귀분석의 전처리 단계에서 일반적으로 변수들을 표준화하기 위한 여러 가지 함수를 사용하였다.

1) 전략산업의 고용에 대한 예측 분석

본 절에서는 먼저 고용을 종속변수로 채택한 후, 설명변수들인 7개의 각각의 전략산업들이 지역경제의 고용 혹은 고용자수에 어떻게 영향을 미치는지 분석하기 위하여 전통적 선형 계량경제 모델인 통상적 최소자승법(OLS)과 머신러닝 회귀 추정방식인 릿지 회귀(Ridge Regression) 분석과 라쏘 회귀(Lasso Regression) 분석으로 추정하여 비교한다.

(1) OLS 추정과 예측 평가

〈표 2〉에서 나타난 것과 같이, 본 절에서는 먼저 7개의 전략산업 부문이 각각 취업자의 고용에 어떻게 영향을 미치는지 분석하기 위하여 통상 최소자승 회귀모형(OLS regression)을 취하여 7개 각 전략산업들에 대한 설명변수들과 종속변수인 고용에 대한 로그를 취하여 추정하였다. 〈표 2〉는 설명변수들과 종속변수를 모두 로그를 취하여 OLS 회귀모형으로 추정한 결과를 나타낸다. 추정결과 제일 마지막에 나와 있는 F값과 F값에 대한 확률값 [Prob (F-statistic)]은 추정모형의 유효성이 없다는 귀무가설에 대한 F 통계량과 검정결과를 나타낸다.

그리하여 먼저 〈표 2〉에서 전체적인 모형의 설명력과 유효성을 보면, 이 회귀모형의 조정된 결정계수는 0.953로 아주 높은 것으로 나타났다. F-통계량으로 볼 때, F 값은 높고 확률값은 0.05보다 낮아 추정모형이 유효하다는 것을 나타낸다. 자기상관성을 측정하는 Durbin-Watson 통계량은 1.509으로 나타나 뚜렷한 자기 상관성 문제는 나타나지 않았다. 그리고 Jarque-Bera (JB)는 0.722로 각각 나타나 5% 유의수준에서 정규분포하고 있는 것으로 나타났다.

〈표 2〉에 나타나 있듯이 OLS 모형으로 추정한 결과, 라이프케어 산업, 스마트 해양산업, 지능형기계 산업과 클린테크 산업 등은 고용자수에 5% 유의수준에서 유의한 영향을 미치고 있지 않는 것으로 나타났다. 그러나 아직 미래를 위한 초기 유망산업 투자 단계인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기 산업은 5% 유의수준에서 부의 영향을 유의적으로 미치고 있는 반면, MICE, 특화관광으로 구성된 글로벌관광산업과 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스 산업은 5% 유의수준에서 고용자수를 정의 방향으로 유의하게 증가시키고 있는 것으로 나타났다.

따라서 부산시는 현재 지역특화 우위가 있고 고용을 증가시키는 글로벌관광산업과 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스 산업들과 현재에는 고용을 증가시키지 못하지만,

자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기 산업 등의 미래 투자 유망산업을 동시에 전략산업들로 선정하고 있다.

따라서 부산시는 경제정책의 우선순위를 현재 고용을 증가시키는 것을 원하는지 혹은 당장은 고용을 증가시키지 않지만 미래의 고용을 증가시키는 것을 원하는지 먼저 결정하여 현재 경쟁력이 있는 산업에 투자할 것인지 아니면 유망산업을 육성하기 위하여 투자할 것인지 선택하여 분간해서 투자할 필요가 있다.

의 결정계수보다 조금 낮게 나타났다. 그리하여 모형이 조금 과잉적합의 가능성이 나타났다. 그리고 선형 회귀모형의 오차의 정도를 평가하기 위하여 구한 평균 자승오차(MSE)의 제곱근(RMSE)의 값은 0.344로 나타났다.

표 2. 고용에 대한 통상최소자승법 모형 추정

	R-squared:			0.965		
	Adj. R-squared:			0.953		
	F-statistic:			85.57		
	Prob (F-statistic):			1.77e-14		
전략산업	coef	std err	t	P〉|t|	[0.025	0.975]
절편	-0.9001	2.796	-0.322	0.751	-6.698	4.898
글로벌관광	2.7320	0.947	2.884	0.009	0.767	4.697
라이프케어	-2.2143	1.363	-1.624	0.119	-5.042	0.613
미래수송기기	-1.5449	0.589	-2.624	0.016	-2.766	-0.324
스마트해양	-0.1085	0.425	-0.255	0.801	-0.990	0.773
지능정보서비스	2.1270	0.225	9.445	0.000	1.660	2.594
지능형기계	-0.3377	0.509	-0.664	0.514	-1.393	0.717
클린테크	0.1526	1.040	0.147	0.885	-2.005	2.310
	Durbin-Watson:			1.509		
	Jarque-Bera (JB):			0.722		

한편, 머신러닝에 의한 추정 모형의 성능을 평가하기 위해서는 주로 평가용 데이터 세트(test dataset)를 가지고 판단하지만, 학습용 데이터 세트(train dataset)를 통하여 학습이 잘되었는지, 과잉 또는 과소 적합의 유무를 판단하기 위하여 두 데이터 세트로 구한 결정계수를 아래와 같이 제시한다.

그리하여 머신러닝모형으로 추정한 결과, 〈표 3〉에서 나타난 것과 같이 학습용 데이터 세트(train dataset)의 결정계수가 0.965, 평가용 데이터 세트(test dataset)의 결정계수가 0.930으로 나타나, 평가용 데이터 세트의 결정계수가 학습용 데이터 세터

표 3. 전략산업의 고용에 대한 통상최소자승 모형의 예측 평가

학습용 데이터 세트 결정계수	0.965
평가용 데이터 세트 결정계수	0.930
RMSE	0.344

(2) 릿지 회귀모형(Ridge Regression Model) 추정과 예측 평가

앞에서 구한 표준적인 통상 최소자승법(OLS)에

의한 선형 회귀모형에서 발생하는 과도적합(overfitting) 문제를 극복하기 위하여 이제 통상적 선형회귀모형에 규제항이 포함된 릿지 선형 회귀모형을 가지고 고용에 대한 7대 전략산업들을 독립변수로 두고 종속변수인 고용에 대해 회귀분석을 하였다.

릿지 회귀분석은 릿지 선형회귀 함수에서 모형의 규제 강도를 결정하는 인자인 α가 중요한데, 기본 설정값은 α=1 이다. α값이 높으면 β 추정계수는 0으로 가까이 접근하고 학습용 세트의 성능은 약해지고 일반화 경향은 높아진다.α값이 너무 낮으면 β 계수에 대한 제약이 낮아져서 표준적인 OLS 모형과 비슷해진다. 그리하여 릿지 회귀 분석으로 구한 추정에 의한 릿지 회귀모형에 대한 평가가 다음 〈표 4〉에서 나타난 것과 같이 규제의 강도를 나타내는 α값에 따라 그 학습용과 평가용의 데이터 세트들의 각각의 결정계수들과 RMSE 값들이 다르게 나타났다.

표 4. 고용에 대한 릿지 회귀모형의 규제강도 α에 따른 모형 적합성과 예측

α = 0.1 Max_iter 1000	학습용 데이터 세트 결정계수	0.962
	평가용 데이터 세트 결정계수	0.933
	RMSE	0.337
α = 0.2 Max_iter 1000	학습용 데이터 세트 결정계수	0.917
	평가용 데이터 세트 결정계수	0.952
	RMSE	0.679
α = 1 Max_iter 1000	학습용 데이터 세트 결정계수	0.541
	평가용 데이터 세트 결정계수	0.731
	RMSE	0.675
α = 10 Max_iter 1000	학습용 데이터 세트 결정계수	0.541
	평가용 데이터 세트 결정계수	0.731
	RMSE	1.072

그리하여 먼저α=0.1일 때 학습용 데이터 세트의 결정계수 0.962, 평가용 데이터 세트의 결정계수 0.933로 나타나 약간의 과대적합이 나타날 가능성은 있지만, 모형의 오차의 정도를 측정하는 RMSE가 0.337로 가장 낮게 나타나고 있다.α=0.2일 때 학습용 데이터 세트의 결정계수 0.917, 평가용 데이터 세트의 결정계수 0.952로 나타나 α=0.1일 때보다 약간의 과소적합이 나타날 가능성은 상당히 감소하였지만, 모형의 오차의 정도를 측정하는 RMSE는 0.679로 다소 높게 나타나고 있다.

릿지 회귀모형의 규제항의 기본 설정값인α=1일 때, 학습용 데이터 세트의 결정계수 0.541, 평가용 데이터 세트의 결정계수 0.731로 나타나 약간의 과소적합이 나타날 가능성도 좀 더 높고, 모형의 오차의 정도를 측정하는 RMSE가 0.675로 역시 다소 높게 나타나 좋은 모형은 아닌 것으로 나타났다.

마지막으로 α=10으로 높을 때, 학습용 데이터 세트의 결정계수 0.541, 평가용 데이터 세트의 결정계수 0.731로 나타나 α=1일 때와 같이 약간의 과소적합이 나타날 가능성이 다소 있고, 모형의 오차의 정도를 측정하는 RMSE가 1.072로 높게 나타나고 있다.

따라서 이 4 가지 릿지 회귀모형 중 α=0.1일 때 평가용 데이터 세트의 결정계수도 0.933로 가장 높고 RMSE가 가장 작아 가장 좋은 것으로 나타났다. 한편, 릿지 회귀모형이 α=0.1일 때 가장 좋게 나타났으므로, 그 때의 평가용 데이터 세트에 대한 예측값과 릿지 회귀모형에서의 7개의 가중치인 추정계수는 아래 〈표 5〉와 같이 나타났다.

〈표 5〉에서는 릿지 회귀모형의 절편과 추정계수들이 나와 있는데 추정계수들을 보면 역시 OLS 추정결과와 동일하게 부산의 7대 전략산업들이 고용에 정의 영향 혹은 부의 영향을 미치고 있는 것으로 나타났지만,α=0.1일 때 릿지 회귀모형의 RMSE가 0.337로 나타나 OLS로 추정할 때의 RMSE인 0.344 보다 낮아 예측력이 더 좋게 나타났다.

이와 같이 릿지 회귀모형의 결과를 보면, 서비스 플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스 산업의 추정계수가 1.914로 양의 가중치가 가장 높게 나타났고, MICE, 특화관광으로 구성된 글로벌관광산업의 가중치를 나타내는 추정계수가 1.698로 각각 나타나 고용을 증가시키는데 기여하고 있다. 그러나 초기투자 단계인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기산업의 가중치인 추정계수는 -2.161로 나타나 현재 고용을 감소시키는 것으로 나타났다. 그 외 다른 4개의 전략산업들의 가중치는 상대적으로 낮은 것으로 나타났다.

지고, 모형의 가장 중요한 특성이 무엇인지 밝혀준다.

그리하여 표준적인 일반화 최소자승법에 의해 구한 선형 회귀모형인 OLS 모형에 규제항이 포함된 라쏘 선형 회귀모형을 가지고 고용에 대한 7대 전략산업들을 독립변수로 두고 회귀분석을 하였다. 라쏘 회귀 분석으로 구한 결과, 그 모형에 대한 평가가 다음 〈표 6〉과 같이 규제의 강도를 나타내는 α값에 따라 그 결정계수들과 RMSE 값들이 다르게 나타났다.

표 5. 릿지 회귀모형의 고용에 대한 전략산업의 추정계수

변수	절편	글로벌 관광	라이프 케어	미래 수송기기	스마트 해양	지능정보 서비스	지능형 기계	클린 테크
가중치 (추정계수)	-4.159	1.698	-0.535	-2.161	-0.128	1.914	0.04	0.133

3) 라쏘 회귀모형(Lasso Regression Model) 추정과 예측 평가

라쏘 회귀분석 역시 비용(손실)함수에 규제항을 추가하여 과잉적합을 방지할 수 있는 선형 회귀모형이지만, 릿지 회귀모형과는 좀 다르다. 라쏘 회귀모형에서는 가중치의 절대값의 합을 최소화 하는 제약조건을 사용하는데 L1 규제라고 하기도 한다.

그러나 릿지 회귀모형과 마찬가지로 규제의 강도를 결정하는 인자인α값이 높아질수록 β계수는 0으로 가까이 접근하고 학습용 세트의 성능은 약해지고 모형은 단순화 되어 진다. 반대로 α값이 낮으면 β 계수에 대한 제약이 낮아져서 계수의 영향력이 높아져서 표준적인 OLS 모형과 비슷해진다.

라쏘 회귀모형의 중요한 특징은 가중치α를 0까지 축소시켜, 가중치가 0인 독립변수를 모형에서 제외시켜 모형의 독립변수의 수를 감소시킬 수 있다. 이와 같이 라쏘 모형을 사용하면 모형을 단순화시켜 분산이 감소되어 과잉적합이 될 가능성이 낮아

표 6. 고용에 대한 Lasso Regression 규제강도 α에 따른 모형 적합성과 예측

α = 0.001	학습용 데이터 세트 결정계수	0.959
	평가용 데이터 세트 결정계수	0.950
	RMSE	0.291
α = 0.01	학습용 데이터 세트 결정계수	0.961
	평가용 데이터 세트 결정계수	0.945
	RMSE	0.304
α = 0.1	학습용 데이터 세트 결정계수	0.794
	평가용 데이터 세트 결정계수	0.966
	RMSE	0.239
α = 0.2	학습용 데이터 세트 결정계수	0.606
	평가용 데이터 세트 결정계수	0.825
	RMSE	0.544

〈표 6〉에서 나타나 있듯이, 먼저 라쏘 회귀모형에서α=0.001일 때, 학습용 데이터 세트의 결정계수 0.959, 평가용 데이터 세트의 결정계수 0.950로 나타났으며, 모형의 오차의 정도를 측정하는 RMSE도 비교적 낮은 0.291로로 나타나고 있다.α=0.01일 때 학습용 데이터 세트의 결정계수 0.961, 평가용 데이터 세트의 결정계수 0.945로 나타나 α=0.001일 때보다 약간의 과소적합이 나타날 가능성이 있으며, RMSE가 0.304로 다소 높게 나타나고 있다.

그러나 규제항인 α=0.1일 때 학습용 데이터 세트의 결정계수는 0.794로 다소 낮았으나, 평가용 데이터 세트의 결정계수는 0.966로 높게 나타났으며, 모형의 오차의 정도를 측정하는 RMSE가 0.239로 가장 낮게 나타나, 평가용 데이터 세트를 예측하는데 있어 좋은 모형으로 나타났다.

마지막으로α=0.2일 때는 학습용 데이터 세트의 결정계수 0.606, 평가용 데이터 세트의 결정계수 0.825로 나타나 약간의 과소적합이 나타날 가능성이 있으며, 모형의 오차의 정도를 측정하는 RMSE가 0.544로 높게 나타나고 있다.

그러므로 위의 〈표 6〉에서 나타나 있듯이, 라쏘 회귀모형 중 α=0.1일 때 모형의 적합성이 가장 좋게 나타났고, 그 때의 라쏘 회귀모형을 이용한 추정에서 절편을 제외하고는 7개의 독립변수들인 전략산업들 중에서 가중치가 0인 독립변수들의 추정계수 4개를 제외하고 그 가중치가 0이 아닌 추정계수들 3개가 나타나 있다.

그리하여 위에서 유의한 독립변수들만 가지고 추정한 결과가 〈표 7〉에서 나타나 있다. 글로벌관광산업의 추정계수는 0.177, 미래수송기기의 추정계수는 −1.067, 그리고 지능정보서비스산업의 추정계수는 1.464로 각각 나타났으며, 이 추정계수들 중 지능형 정보서비스산업의 가중치가 가장 높게 나타났다.

따라서 OLS 추정결과와 마찬가지로 미래를 위한 초기투자 단계인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기 산업은 고용에 부의 영향을 미치고 있는 반면, MICE, 특화관광으로 구성된 글로벌관광산업과 서비스플랫폼, 콘텐츠, 스마트금융 산업으로 이루어진 지능정보서비스 산업은 고용을 증가시키는 정의 영향을 미치고 있는 것으로 나타났다.

그렇지만 라쏘 회귀모형의 추정결과에서, 다른 4개의 부산의 전략산업들인 라이프케어 산업, 스마트 해양산업, 지능형기계 산업과 클린테크 산업 등의 추정계수들은 0으로 나타나, 부산의 고용자수에 유의한 영향을 미치고 있지 않는 것으로 나타나서 앞에서 살펴 본 OLS 추정결과를 재확인 해주고 있다.

그러나α=0.1일 때 라쏘 회귀모형의 결과에서 나타난 RMSE가 0.239로 낮아서 릿지 회귀모형의 결과에서 구한 RMSE 0.337보다 낮게 나타났다. 그리고 OLS로 추정할 때 구한 RMSE 0.344일 때보다도 더 낮아 라쏘 회귀분석에서 구한 고용에 대한 예측력이 릿지 회귀모형이나 전통적인 계량경제방법 보다 더 좋게 나타났다. 그리하여 부산시도 전통적인 OLS 계량경제방법이 아닌 규제항을 포함한 릿지 회귀분석이나 라쏘 회귀분석 등에 의한 머신러닝 예측기법을 통하여 부산의 전략산업의 선정과 전략산업이 부산의 고용에 미치는 영향을 분석할 필요가 있다.

표 7. 라쏘 회귀모형의 고용에 대한 전략산업의 가중치(추정계수)

변수	절편	글로벌 관광	라이프 케어	미래 수송기기	스마트 해양	지능정보 서비스	지능형 기계	클린 테크
가중치 (추정계수)	1.024	0.177	0	-1.067	0	1.464	0	0

2) 전략산업의 소득에 대한 효과 예측

(1) OLS 추정과 예측 평가

본 절에서는 2019년 제 5차 전략산업들에 선정된 부산의 7개의 전략산업 부문이 각각 소득에 어떻게 영향을 미치는지 분석하기 위하여 7개 각 전략산업들을 설명변수들로 삼고 그 종속변수인 부산의 소득에 대한 추정을 하였다.

그리하여 〈표 8〉에 나와 있는 대로, 먼저 OLS 회귀모형으로 추정한 결과, 라이프케어 산업, 스마트 해양산업, 지능형기계 산업과 클린테크 산업 등이 앞에서 분석한 고용자수에 대한 영향과 마찬가지로, 소득에 5% 유의수준에서 유의한 영향을 미치고 있지 않는 것으로 나타났다. 그러나 아직 미래를 위한 초기투자 단계 산업인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기 산업은 5% 유의수준에서 부의 영향을 유의적으로 미치고 있다. 그러나, MICE, 특화관광으로 구성된 글로벌관광 산업과 서비스플랫폼, 콘텐츠, 스마트 금융산업으로 이루어진 지능정보서비스 산업은 5% 유의수준에서 소득의 증대에 유의한 영향을 미치고 있는 것으로 나타났다.

이 회귀모형의 조정된 결정계수는 0.964로 아주 높은 것으로 나타났고, 자기상관성을 측정하는 Durbin-Watson 통계량은 1.504으로 나타나 뚜렷한 자기 상관성 문제는 나타나지 않았다. 그리고 Prob(Omnibus)는 0.938, Prob(JB)는 0.943로 각각 나타나 5% 유의수준에서 정규분포하고 있다는 것을 기각할 수 없는 것으로 나타났다.

그리고 본 절에서는 먼저 부산의 소득에 대한 전략산업들의 영향을 분석하기 위하여 최소자승 선형회귀 모형을 설정하여 파이썬 프로그램으로 평가하기 위하여 학습용 데이터 세트와 평가용 데이터 세트를 70%와 30%로 각각 나누었다. 그리고 앞에서와 마찬가지로 전처리 단계에서 각 변수들을 표준화하기 위한 함수를 사용하였다.

그리하여 파이썬의 머신러닝 기법으로 추정한 결과, 〈표 9〉에서 나타난 것과 같이 학습용 데이터 세트의 결정계수가 0.974, 평가용 데이터 세트의 결정계수가 0.935로 나타나, 학습용 데이터 세터의 결정계수가 평가용 데이터 세트의 결정계수보다 조금 높게 나타나, 소득에 대한 7대 전략산업의 영향을 평가하는데 표준적인 OLS 추정모형이 조금 과잉적합이 될 가능성이 나타났다. 그리고 통상적 최

표 8. 전략산업과 OLS 회귀모형의 소득 추정

		R-squared:		0.972		
		Adj. R-squared:		0.964		
		F-statistic:		110.9		
		Prob (F-statistic):		1.14e-15		
소득	coef	std err	t	P〉|t|	[0.025	0.975]
절편	2.4562	0.131	18.760	0.000	2.185	2.728
글로벌관광	0.1273	0.044	2.870	0.009	0.035	0.219
라이프케어	-0.0985	0.064	-1.542	0.137	-0.231	0.034
미래수송기기	-0.0959	0.028	-3.476	0.002	-0.153	-0.039
스마트해양	-0.0068	0.020	-0.341	0.737	-0.048	0.035
지능정보서비스	0.1093	0.011	10.363	0.000	0.087	0.131
지능형기계	-0.0118	0.024	-0.496	0.625	-0.061	0.038
클린테크	0.0224	0.049	0.460	0.650	0.079	0.123
Omnibus:	0.938	Durbin-Watson:	1.504			
Prob(Omnibus):	0.626	Jarque-Bera (JB):	0.943			
		Prob(JB):	0.624			

소자승법(OLS)의 오차의 정도를 평가하기 위하여 구한 평균 자승오차(MSE)의 제곱근(RMSE)의 값은 0.357로 나타났다.

〈표 9〉 전략산업의 소득에 대한 통상최소자승 모형의 예측 평가

학습용 데이터 세트 결정계수	0.974
평가용 데이터 세트 결정계수	0.935
RMSE	0.357

(2) 소득에 대한 릿지 회귀모형 예측과 평가

일반적인 통상최소자승법에 의한 예측의 과잉적합을 피하기 위하여 규제항이 포함된 릿지(Ridge) 선형 회귀모형을 가지고 종속변수인 소득에 대해서 7대 전략산업들을 독립변수로 삼고 회귀분석을 한 결과가 〈표 10〉에 나타나 있다.

소득에 대한 릿지 회귀분석 모형도 앞에서와 같이, 규제의 강도를 나타내는 인자인 α값에 따라 그 결정계수들과 RMSE 값들이 다르게 나타났다. 먼저 α=0.01일 때, 학습용 데이터 세트의 결정계수

표 10. 규제강도 α 값에 따른 소득에 대한 릿지 회귀 예측

α = 0.01	학습용 데이터 세트 결정계수	0.926
	평가용 데이터 세트 결정계수	0.963
	RMSE	0.358
α = 0.1	학습용 데이터 세트 결정계수	0.926
	평가용 데이터 세트 결정계수	0.963
	RMSE	0.349
α = 0.2	학습용 데이터 세트 결정계수	0.969
	평가용 데이터 세트 결정계수	0.945
	RMSE	0.336
α = 1	학습용 데이터 세트 결정계수	0.926
	평가용 데이터 세트 결정계수	0.963
	RMSE	0.276
α = 10	학습용 데이터 세트 결정계수	0.567
	평가용 데이터 세트 결정계수	0.751
	RMSE	0.713

0.926, 평가용 데이터 세트의 결정계수 0.963로 나타나 약간의 과소적합이 나타날 가능성은 있지만, 모형의 오차의 정도를 측정하는 RMSE가 0.358로 나타나고 있다.

α=0.1일 때, 학습용 데이터 세트의 결정계수 0.926, 평가용 데이터 세트의 결정계수 0.963으로 나타나 α=0.01일 때와 같이 약간의 과소적합이 나타날 가능성은 있지만, 모형의 오차의 정도를 측정하는 RMSE가 0.349로 조금 더 낮게 나타나고 있다. α=0.2일 때 학습용 데이터 세트의 결정계수 0.969, 평가용 데이터 세트의 결정계수 0.945로 나타나 과잉적합이 나타날 가능성은 낮게 나타났고, 모형의 오차의 정도를 측정하는 RMSE가 0.336으로 낮게 나타났다.

규제항의 기본 설정값인 α=1일 때, 학습용 데이터 세트의 결정계수 0.926, 평가용 데이터 세트의 결정계수 0.963로 나타나 약간의 과소적합이 나타날 가능성이 좀 더 높게 나타난 반면, 모형의 오차의 정도를 측정하는 RMSE는 0.276으로 가장 낮게 나타났다. 마지막으로 α=10으로 높을 때, 학습용 데이터 세트의 결정계수 0.567, 평가용 데이터 세트의 결정계수 0.751로 나타나 α=1일 때와 같이 약간의 과소적합이 나타날 가능성이 다소 있고, 모형의 오차의 정도를 측정하는 RMSE가 0.713으로 비교적 높게 나타나고 있다. 그리하여 〈표 10〉에서 나타난 것과 같이, 전략산업들의 소득에 대한 5 가지의 릿지 회귀모형으로 예측한 것 중에서 α=1일 때 평가용 데이터 세트의 결정계수도 0.963로 높게 나타났고 RMSE가 0.276로 작게 나타나 좋은 모형인 것으로 나타났다.

한편, 릿지 회귀모형의 추정에서 =0.1일 경우에 추정된 회귀모형의 결과를 보면, 평가용 데이터 세트에 대한 7개의 전략산업의 가중치인 추정계수들이 〈표 11〉에서와 같이 나타났다.

트금융산업으로 이루어진 지능정보서비스 산업의 추정계수가 1.794로 양의 추정계수가 가장 높게 나타났고, 그 다음에는 MICE, 특화관광으로 구성된 글로벌관광산업의 가중치를 나타내는 추정계수가 0.909로 각각 나타나 소득을 증가시키는데 이러한 전략산업들이 기여하는 것으로 예측된다.

그러나 역시 초기투자 단계인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기산업의 가중치인 추정계수는 −1.936으로 나타나 소득에 부의 영향을 미치는 것으로 나타났다. 그 외 다른 4개의 전략산업들의 추정계수가 라이프케어산업 −0.204로 나타나 소득을 증가시키는데 부의 영향을 미치고 있다.

반면, 스마트해양산업 0.055, 지능형기계산업 0.119, 그리고 클린테크산업 0.135 등으로 각각 나타나 소득을 증가시키는데 정의 영향을 미치는 것으로 나타났다.

표 11. 릿지 회귀모형의 소득에 대한 전략산업의 추정계수

변수	절편	글로벌 관광	라이프 케어	미래 수송기기	스마트 해양	지능정보 서비스	지능형 기계	클린 테크
가중치 (추정계수)	7.684	0.909	-0.204	-1.936	0.055	1.794	0.119	0.135

릿지 회귀모형의 추정계수들을 보면 역시 OLS 추정결과와 동일하게 부산의 7대 전략산업들이 소득에 정의 영향 혹은 부의 영향을 미치고 있는 것으로 나타났고,α=0.1일 때 릿지 회귀모형의 RMSE가 0.349로 나타나 OLS로 추정할 때의 RMSE가 0.357일 때보다 낮아 규제항이 포함된 릿지 회귀모형의 예측력이 OLS 추정모형보다 더 좋은 것으로 나타났다.

그리하여 전략산업의 소득에 대한 릿지 회귀모형의 추정계수들을 보면, 서비스플랫폼, 콘텐츠, 스마

그러나 이들의 소득에 미치는 영향은 상대적으로 가중치들이 낮게 나타나, 소득에 미치는 영향은 크지 않은 것으로 나타났다.

(3) 소득에 대한 라쏘 회귀모형 예측과 평가

이제 표준적인 선형 회귀모형인 OLS 추정모형에 규제항이 포함된 라쏘 선형 회귀모형을 가지고 소득을 종속변수로 삼고 7대 전략산업들을 독립변수로 두고 회귀분석을 하였다. 라쏘 회귀 분석으로

구한 결과, 그 모형에 대한 평가가 다음 〈표 12〉와 같이 규제의 강도를 나타내는α값에 따라 그 결정계수들과 RMSE 값들이 다르게 나타났다.

표 12. 규제강도 α값에 따른 소득에 대한 라쏘 회귀 예측

α = 0.001	학습용 데이터 세트 결정계수	0.967
	평가용 데이터 세트 결정계수	0.954
	RMSE	0.310
α = 0.01	학습용 데이터 세트 결정계수	0.971
	평가용 데이터 세트 결정계수	0.953
	RMSE	0.310
α = 0.1	학습용 데이터 세트 결정계수	0.833
	평가용 데이터 세트 결정계수	0.976
	RMSE	0.222
α = 0.2	학습용 데이터 세트 결정계수	0.647
	평가용 데이터 세트 결정계수	0.886
	RMSE	0.482
α = 1	학습용 데이터 세트 결정계수	0.214
	평가용 데이터 세트 결정계수	0.316
	RMSE	1.182

그리하여 라쏘 회귀모형에 있어 먼저α=0.001으로 아주 낮을 때, 학습용 데이터 세트의 결정계수 0.967, 평가용 데이터 세트의 결정계수 0.954로 나타나 과잉적합이 나타날 가능성이 아주 낮게 나타났고, 모형의 오차의 정도를 측정하는 RMSE가 0.310으로 나타나고 있다.α=0.01일 때 학습용 데이터 세트의 결정계수 0.971, 평가용 데이터 세트의 결정계수 0.953으로 나타나 α=0.001일 때와 같이 과잉적합이 나타날 가능성은 낮게 나타났고, 모형의 오차의 정도를 측정하는 RMSE는 0.310으로 나타났다.

라쏘 회귀모형에 있어α=0.1일 때는 학습용 데이터 세트의 결정계수 0.833, 평가용 데이터 세트의 결정계수 0.976으로 나타나, 과소적합이 나타날 가능성은 낮게 나타났지만, 모형의 오차의 정도를 측정하는 RMSE가 0.222로 가장 낮게 나타났다.

마지막으로 α=0.2일 때 학습용 데이터 세트의 결정계수 0.647, 평가용 데이터 세트의 결정계수 0.886으로 나타나, 과소적합이 나타날 가능성이 가장 크게 나타났고, 모형의 오차의 정도를 측정하는 RMSE가 0.482로 나타났다.

〈표 12〉에서 나타나 있듯이 라쏘 회귀모형 중 α =0.1일 때 평가용 데이터 세트의 결정계수도 0.976로 높게 나타났고 RMSE가 0.222로 가장 작아 가장 좋은 모형으로 나타났다. 그리하여 α=0.1일 때의 평가용 데이터 세트에 대한 예측값과 라쏘 회귀모형에서의 7개의 전략산업의 가중치가 〈표 13〉에 나타나 있다. 먼저 라쏘 모형에서 가중치가 0에 가까운 전략산업들은 앞에서와 마찬가지로 라이프케어, 스마트해양, 지능형기계, 클린테크산업 등 4개로 나타나 부산의 소득에 유의한 영향을 주고 있지 않는 것으로 나타났다.

표 13. 라쏘 회귀모형의 소득에 대한 전략산업의 추정계수

산업	절편	글로벌 관광	라이프 케어	미래 수송기기	스마트 해양	지능정보 서비스	지능형 기계	클린 테크
추정계수	10.392	0.349	0.000	-1.333	0. 000	1.669	0.000	0.000

그러나 그 외 전략산업들의 가중치인 추정계수들을 보면 역시 OLS 추정결과와 동일하게 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스 산업의 추정계수가 1.669로 양의 가중치가 가장 높게 나타났고, 그 다음에는 MICE, 특화관광으로 구성된 글로벌관광산업의 가중치를 나타내는 추정계수가 다소 낮지만 0.349로 각각 나타나 소득을 정의 방향으로 증가시키고 있다. 그러나 역시 초기투자 단계인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기산업의 가중치인 추정계수는 −1.333으로 나타나 고용에 부의 영향을 주는 것으로 나타났다.

이와 같이 위에서 α=0.1일 때 라쏘 회귀분석에서 RMSE가 0.222로 나타났다. 이를 릿지 회귀 추정모형의 RMSE 0.276과 혹은 OLS로 추정모형 RMSE 0.357일 때와 비교해보면, 라쏘 회귀모형의 RMSE가 가장 낮게 나타나 라쏘 회귀분석에 의한 전략산업의 소득에 대한 예측력이 릿지 회귀모형이나 전통적인 OLS모형보다 보다 더 좋게 나타났다.

그리하여 전략산업을 선정하고 전략산업들이 부산의 경제 특히 고용과 소득에 미치는 영향을 분석할 때는 전통적인 계량경제방법이 아닌 릿지 회귀분석이나 라쏘 회귀분석 모형 등을 도입하여 좀 더 정확히 추정하여 예측할 수 있는 분석이 필요하다.

V. 결 론

현재 각국 정부와 지방정부들은 AI 시대와 제 4차 산업혁명, 그리고 코로나 질병의 여파로 인하로 모두 새로운 경제환경을 극복하려는 어려움에 직면해 있다. 부산도 2019년 제 5차 스마트해양산업을 비롯한 7대 전략산업을 새롭게 선정하고 산업 경쟁력을 강화하기 위해 많은 노력을 하고 있다.

그러나 부산 지역경제 활성화를 위한 전략산업의 선정과 효율적인 지원과 육성을 위해서는 먼저 부산시의 전략산업의 부산경제에 대한 좀 더 정확한 예측과 추정이 필요하다. 그러나 전통적 계량경제모형은 최근 불확실한 경제환경의 변동에 따라 심각한 과잉적합 문제가 발생하여 전략산업들의 부산경제에 대한 효과의 추정과 예측을 잘못 유도 할 수 있다.

그리하여 본 연구는 2019년에 새롭게 선정된 부산시의 전략산업들이 부산경제에 미치는 영향을 분석할 때, 전통적인 통상최승자승법에 의한 추정뿐만 아니라, 과잉적합 문제를 극복하여 좀 더 정확한 예측을 위하여 규제항을 도입한 릿지 회귀분석과 라쏘 회귀분석 등을 사용하여 부산의 7대 전략산업들이 부산의 지역경제 특히 소득과 고용에 미치는 영향 등을 예측하여 비교하였다.

이러한 기법에 의한 경제예측은 외국에서는 초기단계지만, 우리나라에서는 아직 경제분석에서 거의 도입되지 않았다. 그리하여 본 연구 결과를 종합해서 요약해보면 다음과 같다.

첫째, 고용에 대한 스마트해양산업을 비롯한 전략산업들의 영향을 예측하기 위하여 규제항을 도입한 릿지 회귀모형에 의해 추정계수들을 살펴보면, 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스 산업의 추정계수가 양으로 가장 높게 나타났고, MICE, 특화관광으로 구성된 글로벌관광산업 역시 고용을 증가시키는데 기여하고 있다. 그러나 초기투자 단계인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기산업은 현재 고용을 증가시키는 것으로 나타나지 않았다.

둘째, 전략산업의 고용에 대한 효과를 규제항을 도입한 라쏘 회귀모형으로 살펴보면, 초기투자 단계인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기산업은 고용에 당장 정의 영향을 미치고 있지 않고 있는 반면, MICE, 특화관광으로 구성된 글로벌관광산업과 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스 산업은 고용을 증가시키고 있는 것으로 나타났다.

셋째, 스마트해양산업을 비롯한 전략산업의 소득

에 대한 릿지 회귀모형의 추정계수들을 보면 부산의 7대 전략산업들은 소득에 정의 영향 혹은 부의 영향을 미치고 있는 것으로 나타났다. 특히 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스산업의 추정계수가 가장 높게 나타났고, 그 다음에는 MICE, 특화관광으로 구성된 글로벌관광산업의 추정계수가 양의 값으로 나타났으므로 이들 산업들은 소득을 증가시키는데 기여하고 있다.

넷째, 소득에 대한 라쏘 회귀모형에서 라이프케어, 스마트해양, 지능형기계, 클린테크산업 등 4 개 전략산업들은 소득에 유의한 영향을 주고 있지 않다. 그러나 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스산업과 MICE, 특화관광으로 구성된 글로벌관광산업 등은 3 개의 전략산업들은 소득을 정의 방향으로 증가시키고 있다.

그리하여 본 연구결과의 정책적 의미를 살펴보면 다음과 같다. 먼저 부산지역의 5차 7대 전략산업들의 선정과 육성지원을 위해서는 부산의 스마트해양산업을 비롯한 전략산업들의 지역경제 특히 소득과 고용에 대한 영향을 좀 더 정확히 예측하기 위하여 기존의 전통적인 통상적 최소자승모형에 의한 예측보다 과잉적합과 모형설정의 오류를 줄이기 위해 라쏘 회귀분석과 릿지 회귀분석을 도입하여 추정할 필요가 있다.

그러므로 향후 부산시는 전략산업들의 부산 경제에 있어서 가장 중요한 소득과 고용에 대한 효과를 좀 더 정확하고 체계적으로 분석하기 위해 스마트해양산업을 비롯한 전략산업들을 좀 더 효율적이고 효과적으로 지원·육성하는 경제정책과 산업정책을 통해 부산의 미래 경쟁력과 지역경제를 활성화할 수 있도록 해야 할 것이다.

그러나 연구의 완성도를 좀 더 높이기 위해 향후 좀 더 긴 시간동안 부산뿐만 아니라 다른 광역시도들의 전략산업들에 대한 자료들을 수집하여 종합적으로 분석하면 좀 더 엄밀한 분석이 되겠지만, 향후 과제로 남긴다.

참고문헌

박기영 · 고정원 (2019), 머신러닝을 이용한 경제분석, 한국경제학보, 제 26권 제 2호, pp.367-408.

이재득 · 윤진영(2018), 부산지역 신성장산업의 경쟁력 분석, 한국은행 부산본부.

이재득 · 이영우(2020), 인공지능 시대 머신러닝 분석을 이용한 부산지역 경제예측과 전략산업, 한국은행 부산본부.

Acemoglu, Daron, and Pascual Restrepo (2019), "Robots and Jobs: Evidence from US Labor Markets," Journal of Political Economy,

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018), Prediction Machines: The Simple Economics of Artificial Intelligence, Harvard Business Review Press.

Athey, Susan (2017), "Beyond Prediction: Using Big Data for Policy Problems," Science, 355(6324), February 3, 483-485.

Athey, Susan (2019), "The Impact of Machine Learning on Economics,"In The Economics of Artificial Intelligence: An Agenda, 1 edition., 507-547, National Bureau of Economic Research Conference Report, University of Chicago Press.

Chakraborty, C. and A. Joseph (2017), "Machine Learning at Central Banks," Bank of Eng-land Staff Working Paper, No. 674.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan (2016), "Productivity and Selection of Human Capital with Machine Learning,"American Economic Review, 106(5), May, 124-127.

Ding, M. J., Zhang, S. Z., Zhong, H. D., Wu, Y. H., & Zhang, L. B. (2019). "A Prediction Model of the Sum of Container Based on Combined BP Neural Network and SVM", Journal of Information Processing Systems, Vol. 15, No. 2, pp.305-319.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016), "Combining Satellite Imagery and Machine Learning to Predict Poverty," Science, 353(6301), August 19, 790-794.

Kim, S., (2020), " Macroeconomic and Financial Market Analyses and Predictions through Deep Learning, BOK Working Paper No. 2020-18, Bank of Korea,

Lopez de Prado, M. (2018), Advances in Financial Machine Learning, New York, NY, USA: Wiley.

Mullainathan, Sendhil, and Jann Spiess (2017), "Machine Learning: An Applied Econometric Approach," Journal of Economic Perspectives, 31(2), May, 87-106.

Naecker, Jeffrey, and Alexander Peysakhovich (2017), "Using Methods from Machine Learning to Evaluate Behavioral Models of Choice under Risk and Ambiguity", Journal of Economic Behavior & Organization, 133, January, 373-384.

Schapire, Robert E., and Yoav Freund (2014), Boosting: Foundations and Algorithms, Adaptive Computation and Machine Learning Series, The MIT Press.

Schlkopf, Bernhard, and Alexander J. Smola (2001), Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, 1 edition. Adaptive Computation and Machine Learning Series, The MIT Press.

Tibshirani, Robert (1996), "Regression Shrinkage and Selection Via the Lasso," Journal of the Royal Statistical Society: Series B (Methodological), 58(1), January, 267-288.

Zou, Hui, and Trevor Hastie (2005), "Regularization and Variable Selection via the Elastic Net," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), April, 301-320.

릿지 회귀와 라쏘 회귀 모형에 의한 부산 전략산업의 지역경제 효과에 대한 머신러닝 예측

이재득

국문요약

본 연구는 규제항을 도입한 릿지 회귀분석과 라쏘 회귀분석을 사용하여 부산 전략산업의 지역경제에 미치는 효과를 특히 고용과 소득에 대한 영향을 중심으로 머신러닝 기법으로 예측하고 분석하였다. 주요 연구결과는 다음과 같다. 첫째, 고용에 대한 전략산업들의 영향을 릿지 회귀모형과 라쏘 회귀모형으로 추정해보면, 전략산업 가운데 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스산업과 MICE, 특화관광으로 구성된 글로벌관광산업의 순으로 고용을 증가시키는데 기여하고 있다. 둘째, 릿지 회귀모형과 라쏘 회귀모형에 의하면 초기투자 단계인 자율주행차, 항공, 드론 산업으로 이루어진 미래수송기기산업은 고용과 소득을 유의하게 증가시키지 않는 것으로 나타났다. 셋째, 전략산업의 소득에 대한 릿지 회귀모형의 추정계수들을 보면, 지능정보서비스산업과 글로벌관광산업의 순으로 부산지역의 소득을 증가시키고 있다. 넷째, 라쏘 회귀모형에서 라이프케어, 스마트해양, 지능형기계, 클린테크산업 등 4개의 전략산업들은 소득에 유의한 영향을 주고 있지 않는 반면, 지능정보서비스산업과 글로벌관광산업 등 2개의 전략산업들은 소득을 증가시키고 있으나, 장기 투자 산업인 미래수송기기산업은 현재 지역경제와 소득에 부의 영향을 줄 수 있는 것으로 나타났다. 그리하여 전략산업을 선정하고 육성하는데 있어, 부산지역 경제목표와 정책 우선순위를 먼저 설정할 필요가 있다는 점을 시사한다.

주제어: 전략산업, 릿지 회귀모형, 라쏘 회귀모형, 소득과 고용