The Korean Journal of Applied Statistics
2024, Vol. 37, No. 1, 13–38

DOI: https://doi.org/10.5351/KJAS.2024.37.1.013
Print ISSN1225-066X / Online ISSN 2383-5818

Real-time private consumption prediction using big data

Seung Jun Shin[a], Beomseok Seo[1,b]

[a]Department of Statistics, Korea University;
[b]Office of Economic Modeling and Policy Analysis, Bank of Korea

Abstract

As economic uncertainties have increased recently due to COVID-19, there is a growing need to quickly grasp private consumption trends that directly reflect the economic situation of private economic entities. This study proposes a method of estimating private consumption in real-time by comprehensively utilizing big data as well as existing macroeconomic indicators. In particular, it is intended to improve the accuracy of private consumption estimation by comparing and analyzing various machine learning methods that are capable of fitting ultra-high-dimensional big data. As a result of the empirical analysis, it has been demonstrated that when the number of covariates including big data is large, variables can be selected in advance and used for model fit to improve private consumption prediction performance. In addition, as the inclusion of big data greatly improves the predictive performance of private consumption after COVID-19, the benefit of big data that reflects new information in a timely manner has been shown to increase when economic uncertainty is high.

Keywords: kernel quantile regression, MIDAS, variable selection, economic forecasting

1. 서론

국민소득통계(이하 GDP)는 국가의 경제상황을 종합적으로 보여주는 경제지표로서 경제정책의 방향성을 설정하는 중요한 근거로 사용되고 있다. 국내 GDP 통계는 한국은행에서 분기 및 연간 통계를 속보, 잠정, 확정으로 구분하여 발표하고 있다. 대부분의 국가통계와 마찬가지로 GDP 통계는 조사기준 시점과 공표시점 간의 차이가 발생하는 공표지연 문제가 발생한다. 한국은행에서는 이를 줄이기 위한 많은 노력을 기울이고 있으나, 통계생산을 위한 기초자료의 최초 가용시점, 정확한 통계를 생산하기 위해 소요되는 물리적인 시간 등을 고려할 때, 공표시점을 앞당기는 것은 현실적으로 한계가 있다. 공표시차의 발생은 비단 우리나라만의 문제는 아니며, 미국 등을 포함한 선진국 등에서도 GDP 통계의 공표시차 제약을 보완하기 위해 많은 노력을 기울이고 있다. 이러한 노력의 일환으로 GDP 통계와 상관관계가 높으면서 보다 빠르게 수집이 가능한 자료들을 바탕으로 계량 모형을 구축하여 당분기 GDP를 실시간으로 추정하는 방법(이하 nowcasting)이 오래전부터 활발하게 연구되어 왔다 (Bańbura 등, 2013; Kim과 Kim, 2016). 특히, 최근과 같이 경제 불확실성이 높은 상황에서는 시의성이 있는 정책의 실행과 평가가 무엇보다 중요하므로, 정확한 nowcasting의 필요성은 어느 때보다 크다 하겠다.

This work was supported by the Bank of Korea.
[1]Corresponding author: Office of Economic Modeling and Policy Analysis, Bank of Korea, 39 Namdaemun-ro, Jung-gu, Seoul 04531, Korea. E-mail: bsseo@bok.or.kr

Published 29 February 2024 / journal homepage: http://kjas.or.kr

GDP를 구성하는 다양한 항목 중 민간소비는 국민들이 체감하는 경제상황을 직접적으로 보여준다는 점에서 정부의 정책 결정과 평가에 중요한 지표로 활용된다. 그러나 GDP 총액과 달리 민간소비를 실시간으로 추정하는 방법에 대한 연구는 최근까지 상대적으로 많이 이루어지지 않고 있다. 특히, 미국 경제를 중심으로 한 해외 연구에서는 중앙은행, 통계청 등을 중심으로 민간소비를 최신 nowcasting 모형으로 예측하고자 하는 연구들이 점차 확대되고 있으나 (Chen과 Hood, 2021), 국내에서는 저자가 파악한 바로는 아직 이러한 연구가 시도된 바가 없다. 이는 GDP 총액 nowcasting의 경우 GDP 통계보다 공표주기가 빠른 다른 거시경제지표를 이용하여 모형을 구성하는 것이 용이하지만, GDP의 구성항목인 민간소비의 경우 활용 가능한 기초자료가 적은 점이 국내 연구의 발전을 더디게 한 요인으로 작용했기 때문임을 유추할 수 있다. 그러나 최근 4차산업의 발전으로 국내에서도 민간소비와 밀접한 관련이 있는 빅데이터에 대한 접근성이 크게 증대되고있다. 대표적인 예로 신용카드 소비액, 신문기사 텍스트 데이터, 이동량 통계 등을 들 수 있는데, 이러한 빅데이터를 적극 활용하면 민간소비 nowcasting의 예측력을 보다 향상시킬 수 있을 것임을 예상해 볼 수 있다.

이에 본 연구에서는 보다 정확한 민간소비 nowcasting을 위해 빅데이터를 활용할 수 있는 기계학습 방법의 사용을 제안하고, 이를 실증적으로 비교 분석하였다. 본 연구의 의의는 크게 세 가지로 나눌 수 있다. 첫째, GDP 총액이 아닌 민간소비에 대한 nowcasting을 다루었다는 점이다. 전술하였다시피 민간소비의 경우 실시간으로 활용가능한 카드 빅데이터, 신문기사 텍스트 데이터 등과 더욱 밀접한 연관을 지니고 있어, 이러한 빅데이터를 민간소비 예측에 적극 활용하고자 하였다. 둘째, 본 연구에서는 다양한 기계학습 방법을 활용한 nowcasting 방법을 제안하고 그 성능을 실증적으로 비교 분석하였다. nowcasting에서 널리 쓰이는 MIDAS-회귀모형 (Ghysels 등, 2004)을 다양한 기계학습 알고리즘을 적용할 수 있는 형태로 확장하였다. 마지막으로, 본 연구에서는 다양한 종류의 빅데이터뿐만 아니라 기존에 널리 쓰이는 거시경제지표에 대한 자료도 nowcasting에 모두 활용한 고차원 모형의 활용을 제안하였다. 차원의 저주(curse of dimensionality)를 최소화하고 모형의 성능을 최적화하기 위해 변수선별(feature screening)의 활용을 제안하고 그 효용을 확인해 보았다. 그 결과 코로나19 팬데믹으로 인해 경제 불확실성이 높은 최근의 상황에서는 변수선별이 nowcasting의 성능에 큰 영향을 미치는 것을 확인하였다.

본 논고의 구성은 다음과 같다. 2절에서는 연구 배경과 nowcasting에 관련된 선행연구들에 대해 살펴보았다. 3절에서는 본 연구에서 제안하는 민간소비 nowcasting 방법에 대해 자세히 소개하였다. MIDAS-회귀모형을 바탕으로 다양한 기계학습 방법론을 제안하고 추가적인 변수선별에 대한 내용도 상세히 기술하였다. 4절에서는 제안된 방법을 실제자료에 적용하고 그 결과를 심층적으로 분석하였다. 마지막으로 본 연구의 결론과 시사점을 5절에서 제시하였다.

2. 배경 및 관련 연구 리뷰

2.1. 연구 배경

Nowcasting은 장기전망을 의미하는 forecasting과 대비되는 개념으로, GDP와 같이 공표시차가 발생하는 경우 다양한 연관 정보를 활용하여 공표시점 이전에 빠르게 추정치를 계산해 내는 것을 의미한다. 국내 GDP의 경우 해당 분기 종료 이후 약 28일 이내에 속보치를 발표하고 분기 잠정치 발표까지는 약 2달의 시차가 발생한다 (Figure 1 참조).

GDP nowcasting은 GDP보다 공표주기가 빠르거나 일찍 발표되는 다양한 관련 자료를 공변량으로 활용하여 GDP를 실시간으로 예측한다. 따라서 GDP에 활용가능한 공변량 계열은 통상적으로 GDP보다 빠른 주기를 가지는 시계열 자료이다. 이러한 관점에서 GDP nowcasting의 통계/계량적 방법론의 핵심은 자료별 관측주기가 다르기 때문에 발생하는 혼합주기(mixed frequency)와 자료의 가용시점이 다르기 때문에 발생하는 문제(ragged-edge problem)를 어떻게 해결하는가에 있다. 혼합주기 문제를 해결하는 가장 간단한 방법은 고주기 자료에서 나오는 다수의 값들을 하나의 값으로 결합하는 방법이다. 2.2절에서 소개한 교량방정식(bridge

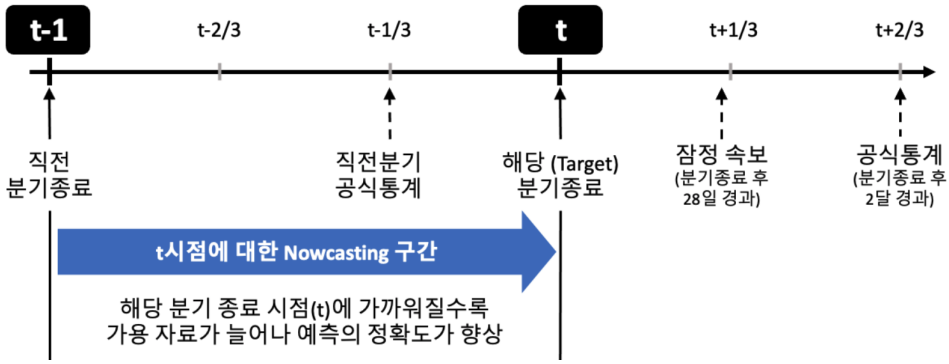


Figure 1: *GDP nowcasting: Nowcasting GDP at the end of a quarter (t) is possible at various points in time. As the end of the quarter (t) approaches, the number of available data for nowcasting increases, making the forecast at point t the most accurate. For example, two months before the end of the quarter ($t-2/3$), advance estimates for the previous quarter ($t-1$) can be used, and one month before ($t-1/3$), preliminary statistics can be utilized.*

equation)이나 MIDAS-회귀모형(mixed data sampling regression model) 등에서 활용하는 방법으로 빠르고 효율적이지만 정보의 손실이 발생하는 방법이다. 다른 대안으로 혼합주기 문제를 결측으로 인식하여 풀어내는 방법이 있다. 이 경우, 자료의 분포함수를 직접 추정하는 방식으로 접근하면 자연스럽게 혼합주기 문제를 해결할 수 있다. 뿐만 아니라 자료의 서로 다른 가용시점(ragged-edge) 문제 역시 결측의 한 형태로 볼 수 있기 때문에 유사한 방법으로 처리가 가능하다는 장점이 있다. GDP nowcasting에 가장 널리 쓰이는 동적인자모형(dynamic factor model) 등이 이에 해당한다.

한편, nowcasting에서 빅데이터의 활용이 점점 늘어나고 있는 것이 최근의 추세이다. 자료 수집주기가 매우 짧고 공표시차가 거의 없다고 할 수 있는 빅데이터의 특성상, 이들 자료를 잘 활용한다면 nowcasting의 성능을 크게 향상시킬 수 있다. 이러한 빅데이터의 활용을 위해 알고리즘 효율성이 높은 기계학습 방법론이 많이 이용되고 있다 (Richardson 등, 2018; Chen 등, 2019; Babii 등, 2022; Lee 등, 2022). 빅데이터는 사용할 수 있는 정보의 양(quantity)이 많고 수집 속도가 빠른 장점이 있는 반면, 관찰자료이기 때문에 정보의 질이 통상적인 조사자료에 비해 떨어진다. 이러한 경우, 변수선택은 필수적이다. 모형에 사용되는 변수의 개수가 늘어날수록 추정해야 하는 모수의 개수가 증가하기 때문에 해당 변수가 nowcasting에 필요한 (양과 질 모든 측면에서) 충분한 수준의 정보를 보유하고 있지 않다면 오히려 예측성능이 저하되는 역효과가 발생하게 된다.

본 연구에서는 기존 거시경제지표 자료와 빅데이터를 모두 활용한 기계학습 기반 고차원 nowcasting 방법론에 대해 분석하였다. 이를 위해 고차원 변수를 효율적으로 다룰 수 있는 다양한 알고리즘을 제안하고 그 성능을 실증적으로 비교평가하였다.

2.2. 선행연구

GDP nowcasting을 위한 방법은 크게 결합모형(joint model)과 부분모형(partial model) 두 가지로 나눌 수 있다. 결합모형은 종속변수 시계열 Y_t와 공변량 시계열 X_t의 결합분포$(X_t, Y_t) \sim \mathcal{P}_{X,Y}$를 추정하는 방법으로, 동적인자모형(dynamic factor model) (Giannone 등, 2008)이 그 대표적인 예이다. nowcasting을 위한 결합모형은 상태-공간 표현형식(state-space representation) (Harvey, 1990)을 활용하여 각 시계열을 공통성분(common component)과 고유성분(iodiosyncratic component)으로 직교분해한다. 이 때 공통성분은 모든 시계열이 공유하는 공통의 부분을 의미하며, 이를 바탕으로 종속변수 시계열의 값을 예측한다. 결합모형은 분포가정에 의존한 방법으로 통상적인 최대가능도추정이나 베이지안 방법으로 모형을 적합할 수 있다. 특히 Kalman filter

나 EM 알고리즘 등이 널리 활용된다. 결합모형은 종속변수와 공변량의 결합분포를 추정하므로 혼합주기와 서로 다른 가용시점 문제(ragged-edge problem)를 결측(missing)의 형태로 인식하여 자연스럽게 해결할 수 있다. 뿐만 아니라, 결합모형은 추정된 결합분포로부터 각 공변량별로 예측치를 계산해 낼 수 있어 이를 바탕으로 종속변수에 미치는 영향을 정량적으로 평가할 수 있다. 이러한 장점 때문에 결합모형의 대표적인 예인 동적인자모형이 GDP nowcasting에서 가장 널리쓰이고 있다 (미국, Giannone 등 (2008); 뉴질랜드 Matheson (2010); 노르웨이, Luciani와 Ricci (2013); 프랑스, Barhoumi 등 (2010); 캐나다, Chernis과 Sekkel (2017) 등).

이와는 달리 부분모형은 공변량이 주어졌을 때, 종속변수의 조건부 분포 $Y_t \mid X_t \sim \mathcal{P}_{Y|X}$를 추정하는 방법으로 교량방정식(bridge equation)과 MIDAS-회귀(MIDAS-regression) (Ghysels 등, 2004)가 그 대표적인 예이다. 종속변수 Y_t와 Y_t보다 m배 빠른 고주기 공변량 X_t간의 교량방정식은 다음과 같다.

$$Y_t = \alpha + \beta \bar{X}_t + \epsilon_t,$$

여기서 $\bar{X}_t = m^{-1} \sum_{j=1}^{m} X_{t-(j-1)/m}$이다. 즉, 교량방정식은 고주기 공변량의 구간별 평균값을 활용하여 종속변수의 주기와 일치시킴으로써 혼합주기의 문제를 해결한다. MIDAS-회귀는 교량방정식을 확장한 형태로서, 평균값 계산에 활용되는 가중치를 자료로부터 학습시켜 구하는 방법이다(자세한 내용은 3.1절 참조). 부분모형은 ragged-edge 문제를 해결하기 위해 결측값 대체(missing imputation)를 해주어야 한다. 만약 그렇지 않으면 정보의 손실이 발생하여 모형의 성능이 떨어질 수 있다는 단점이 있다. 또한 결합모형과 비교하여 종속변수와 공변량간의 연관성을 정량화하는데 어려움이 발생할 수 있다. 하지만 부분모형은 결합모형에 비해 단순한 가정을 활용하기 때문에 데이터의 크기가 증가하더라도 효율적인 계산이 가능하다.

최근 GDP nowcasting에도 새로운 방법이 많이 활용되고 있는 추세이다. 특히 빅데이터의 활용이 주목받고 있다. 신용카드나 POS (point of sales) 데이터, 혹은 스마트 폰에서 수집된 GPS 데이터 등을 효과적으로 활용하면 GDP nowcasting의 성능을 크게 향상시킬 수 있다 (Seo, 2017; Moriwaki, 2019). 빅데이터는 자료의 생성 직후 신속하게 수집이 가능하여 현재시점의 경제상황을 지체없이 확인할 수 있다. 이와 같은 이유로 다양한 국가에서 빅데이터에 기반한 nowcasting 방법론을 개발하고자 노력하고 있다 (Barnett 등, 2016; Galbraith와 Tkacz, 2018; Kim과 Swanson, 2018; Raju와 Balakrishnan, 2019; Aastveit 등, 2020, 등). Nowcasting에서 빅데이터를 활용하는 데 가장 큰 어려움은 계산의 효율성 확보이다. 동적인자모형을 비롯한 결합모형의 경우, 대용량 빅데이터를 활용하는 데 계산상의 어려움이 있으며, 이러한 단점을 보완하기 위한 연구가 이루어지고 있지만 그 성과가 제한적이다 (Chan과 Jeliazkov, 2009; Delle Monache와 Petrella, 2019). 이러한 관점에서 빅데이터에 기반한 nowcasting 문제에서 부분모형의 활용가치는 매우 크다 하겠다 (Babii 등, 2022).

3. 방법론

3.1. MIDAS-회귀모형

본 연구에서는 빅데이터 활용에 용이한 부분모형을 고려하였다. 특히 최근 Babii 등 (2022)이 제안한 고차원 MIDAS-회귀모형을 활용하였다. 종속변수 시계열 y_t와 K개의 공변량 시계열 $x_{t,k}, k = 1, 2, \ldots, K$에 대한 혼합주기 회귀모형(mixed frequency regression)은 다음과 같이 정의된다.

$$\phi(L; J) y_t = \rho_0 + \sum_{k=1}^{K} \psi\left(L^{\frac{1}{m_k}}; \boldsymbol{\theta}_k\right) x_{t,k} + \epsilon_t, \quad t = 1, 2, \ldots, T, \tag{3.1}$$

여기서 ϵ_t는 랜덤오차, $\phi(L; J) y_t = (I - \rho_1 L - \rho_2 L^2 - \cdots - \rho_J L^J) y_t$는 저주기 종속변수 y_t에 대한 시차-다항식(lag-polynomial)을 나타낸다.

우변에 있는 $\psi(L^{1/m_k}; \boldsymbol{\theta}_k) x_{t,k} = m_k^{-1} \sum_{j=1}^{m_k} \theta_{j,k} x_{t-(j-1)/m_k,k}$는 종속변수보다 주기가 m_k배 빠른 공변량 $x_{t,k}$에 대한 시차-다항식을 의미하며, $\boldsymbol{\theta}_k = (\theta_{1,k}, \ldots, \theta_{m_k,k})^T, k = 1, \ldots, K$는 회귀계수벡터를 나타낸다.

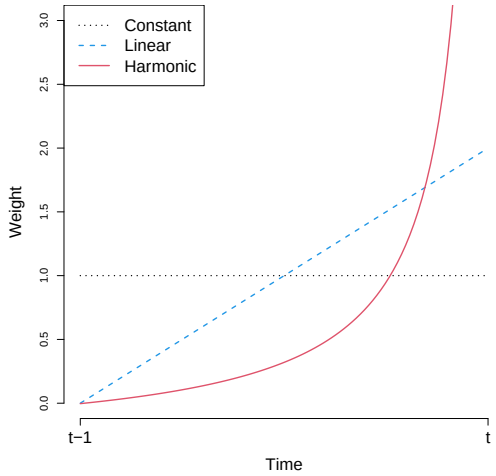


Figure 2: *Weight function: All three functions determine the weights of high-frequency observations observed in the interval* $(t-1, t]$. *The constant function provides a simple average, while the other two functions assign higher weights as it gets closer to point* t, *calculating the average to align with the cycle.*

혼합주기 문제를 해결하기 위한 MIDAS (Ghysels 등, 2004) 아이디어를 (3.1)에 적용하면 회귀계수 $\boldsymbol{\theta}_k$를 다음과 같이 d개의 모수 $\boldsymbol{\beta}_k = (\beta_{1,k}, \ldots, \beta_{d,k})^T$에 의해 조절되는 가중치함수 $w(\cdot)$를 통해 표현할 수 있다.

$$\theta_{j,k} = w\left(\frac{j-1}{m_k}; \boldsymbol{\beta}_k\right) = \sum_{d=1}^{D} \beta_{k,d} v_d\left(\frac{j-1}{m_k}\right) = \boldsymbol{\beta}_k^T \mathbf{v}_{j,k}, \quad j = 1, \ldots, m_k, \tag{3.2}$$

여기서 $\{v_d(\cdot), d = 1, \ldots, D\}$는 사용자가 특정할 수 있는 가중치함수의 기저함수셋을 의미하며 $\mathbf{v}_{j,k} = \{v_d((j-1)/m_k), d = 1, \ldots, D\}^T$이다. 즉, D가 증가할수록 표현할 수 있는 가중치의 형태가 복잡해지므로 모형의 복잡도가 증가한다. 본 연구에서는 Bajari 등 (2015)가 제안한 3차 다항식(legendre polynomial) 대신 Figure 2와 같은 세 가지 형태의 서로 다른 기저함수를 고려하였다. 본 연구에서는 고려하는 공변량의 개수가 매우 많아 가중치 함수의 복잡도를 높이는 것이 모형의 분산을 크게 만들기 때문에 오히려 nowcasting의 성능을 저하시키는 것을 확인하였다. 한가지 대안으로, 최적의 가중치함수를 찾기 위해 교차-타당성 검증(cross validation)과 같은 방법을 고려할 수 있으며, 이는 후속연구를 통해 살펴볼 가치가 있는 것으로 판단한다.

이제, (3.2)를 (3.1)에 대입하면 다음과 같은 선형모형의 형식으로 (3.1)을 재표현할 수 있다.

$$y_t = \mathbf{b}^T \mathbf{z}_t + \epsilon_t, \quad t = 1, 2, \ldots, T, \tag{3.3}$$

여기서 $\mathbf{b} = \{\rho_1, \ldots, \rho_L, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T\}$이고 $\mathbf{z}_t = (1, y_{t-1}, \ldots, y_{t-L}, m_1^{-1} \sum_{j=1}^{m_1} \mathbf{v}_{j,1}^T \cdot x_{j,1}, \ldots, m_k^{-1} \sum_{j=1}^{m_k} \mathbf{v}_{j,K}^T \cdot x_{j,K})^T$이다. Babii 등 (2022)에서는 (3.3)의 모수를 추정하기 위해 벌점화 최소제곱법(penalized least square estimation)을 활용할 것을 제안하였다.

3.2. 모형 확장과 기계학습의 활용

한편, 기계학습에서는 y_t와 $\mathbf{z}_t$간의 관계가 반드시 단순한 선형함수일 필요가 없다. 따라서 다음과 같은 모형의 확장을 고려할 수 있으며 이는 nowcasting 문제를 다양한 기계학습 알고리즘을 통해 해결할 수 있는 연결고리

를 제공한다.

$$y_t = f(\mathbf{z}_t) + \epsilon_t, \quad t = 1, 2, \ldots, T. \tag{3.4}$$

이제, (3.4) 하에서 다양한 기계학습 방법을 적용하여 비선형 함수를 포함한 다양한 회귀함수 f를 추정할 수 있다. 가장 간단한 방법으로는 Babii 등 (2022)이 처음 제안한 바와 같이 f에 대한 선형성을 가정한 뒤 벌점화를 통해 고차원 모수를 추정하는 방법이다. 이 때 사용할 수 있는 대표적인 벌점함수로는 LASSO (Tibshirani, 1996), elastic net (Zou와 Hastie, 2005), 그리고 SCAD (smoothly clipped absolute deviance) (Fan과 Li, 2001) 등이 있다. 특히 SCAD 벌점함수는 LASSO에서 발생하는 편향을 줄임으로써 근사적으로 더 좋은 추정량을 제공한다는 것이 알려져 있다. 다만, SCAD 벌점함수는 비볼록함수이므로 LASSO에 비해 계산이 좀더 어렵다고 알려져 있으나, 좌표별 강하법(coordinate descent algorithm) 등을 활용하면 대용량 자료에 대해서도 효율적 계산이 가능하다 (Breheny와 Huang, 2011). 본 연구에서는 f를 선형함수로 가정한 뒤, LASSO와 SCAD 벌점함수를 적용한 모형을 활용하였다.

선형 모형을 비선형으로 확장하는 데 가장 중요한 것은 차원의 저주를 해결하는 것인데, 차원의 저주를 피할 수 있는 대표적인 방법으로 다음과 같은 단일지표모형(single index model)을 고려할 수 있다.

$$y_t = g\left(\boldsymbol{\beta}^T \mathbf{z}_t\right) + \epsilon_t, \quad t = 1, 2, \ldots, T, \tag{3.5}$$

즉, 종속변수와 다차원 공변량의 연관성에 대한 정보를 모두 포함하고 있는 단일지표 모수 $\boldsymbol{\beta}$를 우선 추정한 뒤 1차원으로 축약된 공변량 $\boldsymbol{\beta}^T \mathbf{z}_t$와 y_t의 관계를 나타내는 회귀함수 g를 비모수적으로 추정하는 것이다. 본 연구에서는 단일지표모형 (3.5)에서 $\boldsymbol{\beta}$를 추정하기 위해 충분차원축약(sufficient dimension reduction; SDR) (Li, 2018)을 활용하였으며, 차원축약 후 1차원 회귀함수를 적합하기 위해서는 스플라인(spline) 방법을 활용하였다. 기존의 연구에서는 차원축소의 방법으로 주성분분석(principal component analysis)을 많이 활용하였지만 (Marcellino와 Schumacher, 2010; Kim과 Swanson, 2018), nowcasting과 같은 지도학습 문제에서는 충분차원축약이 주성분분석보다 더 좋은 성능을 나타내는 것으로 알려져 있다.

또한 비선형 모형 (3.4)를 직접 추정하는 방법도 다양하게 존재한다. 예를 들어 나무 모형이나 나무모형을 기반으로 앙상블 방법을 활용할 수 있으며, 본 연구에서는 가장 대표적인 앙상블 방법 중 하나인 랜덤 포레스트를 고려하였다. 이 외에도 평활-스플라인 모형이나 인공신경모형도 대안으로 고려할 수 있다.

비선형 모형을 직접 추정하는 방법으로, 본 연구에서는 커널분위수회귀(kernel quantile regression; KQR)도 고려하였다. 커널을 활용하면 공변량이 많은 고차원 상황에서 회귀함수를 효율적으로 추정할 수 있다는 것이 잘 알려져 있다 (Zhang, 2002). 이러한 점에 착안하여 커널분위수회귀 (Li 등, 2007)를 이용하여 (3.4)를 추정하였다. 커널분위수회귀는 다음과 같은 최적화 문제를 통해 (3.4)의 회귀함수 f를 추정한다:

$$\hat{f} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_i - f(\mathbf{z}_i)\right) + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}_K}, \tag{3.6}$$

여기서 $\rho_\tau(u) = u(\tau - \mathbb{1}\{u \geq 0\})$는 $100 \times \tau\%$ 번째 분위수 함수를 추정하기 위한 체크 손실함수이며, $\mathcal{H}_K$는 주어진 커널함수 $K(\mathbf{z}, \mathbf{z}')$에 의해 생성되어진 재생커널힐버트공간(reproducing kernel Hilbert space)을 나타낸다. 위 모형은 조율모수 $\lambda > 0$를 통해 모형복잡도와 적합도 사이에 균형점을 찾아 최적의 성능을 구현할 수 있다. 주어진 최적화 문제는 함수공간상에 정의된 무한차원의 문제로 보이지만, 대표자 정리(representer theorem) 등에 의해 유한차원의 최적화 문제로 재표현할 수 있어 표본으로부터 계산이 매우 용이하다. 특히 커널을 이용할 경우 모수의 개수가 공변량의 개수가 아닌 표본의 크기(종속변수의 크기)에 의존하게 되어 공변량의 수가 많은 고차원 모형에 특히 강점을 가진다. 본 연구에서는 $\tau = 0.5$로 설정하고 중위수 회귀를 고려하였다. 중위수 회귀 대신 통상적인 제곱 손실함수를 이용할 수 있으나, 본 연구에서는 추정의 강건성을 위해 분위수회귀를 고려하였다. 본 연구에서 활용된 기계학습 방법들에 관한 보다 자세한 설명은 부록에 자세히 기술하였다.

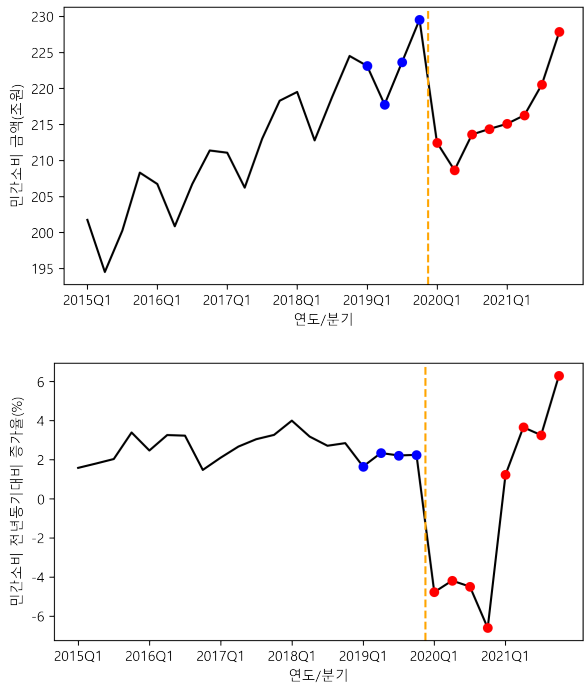


Figure 3: *Amount of private consumption and year-over-year growth rate from the first quarter of 2015 to the third quarter of 2021.*

3.3. 변수선별

고빈도 데이터 등 다양한 종류의 빅데이터를 활용한 GDP nowcasting은 고차원 문제이다. 즉, 주어진 데이터의 크기에 비해 공변량의 개수가 많다. 이 경우 차원의 저주가 발생하는데, 이를 해결하기 위해서 기계학습에서는 희박성 가정(sparsicty assumption)을 활용한다. 즉, 많은 공변량 중 소수의 공변량만이 종속변수에 영향을 준다는 것이다. 희박성 가정하에서 LASSO나 SCAD 등의 벌점함수를 이용하면 변수선택을 통해 추정량의 성능이 향상된다는 사실이 잘 알려져 있다 (Hastie 등, 2015). 그런데 공변량의 개수가 월등히 많은 초고차원(ultrahigh-dimensional) 상황이 되면 추정해야 할 모수가 주어진 정보에 비해 지나치게 많아져 벌점화를 활용하여 변수선택을 하더라도 모형의 성능이 현저히 떨어진다는 것 또한 알려져 있다 (Fan과 Lv, 2008). 이를 해결하기 위한 방법으로 Fan 등 (2009)는 종속변수와 개별 공변량 간의 주변 연관성(marginal association)을 정량화하고 이를 기반으로 한 변수선별(feature screening)의 활용을 제안하였다.

변수선별의 핵심은 공변량 간의 상호작용은 배제하고, 종속변수와의 주변 연관성만을 바탕으로 빠르게 종속변수에 영향을 주는 신호변수를 선별하는 것이다. 변수선별 후 남은 변수들 중에는 종속변수에 영향을 주는 변수도 있고 그렇지 않은 변수도 있다. Fan과 Lv (2008)는 주변 연관성을 이용한 변수선별이 신호변수를 제거할 확률이 0으로 수렴함을 보였으며, 이를 확실선별성질(sure screening property)이라고 명하였다. 확실선별성질은 주변 연관성을 기반으로 한 변수선별의 이론적 정당성을 제공하여 이후 다양한 형태의 변수선별 방법이 제안될 수 있는 기반을 마련하였다 (Fan 등, 2011; Zhu 등, 2011; Li 등, 2012; He 등, 2013; Chang 등, 2013, 등).

Nowcasting을 위한 변수선별을 위해 종속변수 y_t와 주어진 k 번째 공변량 $\mathbf{x}_k$간에 다음과 같은 주변 모형

Table 1: Summary of covariates used in private consumption nowcasting

Category1	Category2	Period	Variables
Macro-economic data	Financial markets (10)	Monthly	주가, 수신금리, 대출금리 등
	Price (7)	Monthly	생산자물가지수, 유가, 기대인플레이션, 소비자물가지수 등
	Consumption (41)	Monthly	소비자 동향조사, 소매업태별 판매액지수, 경제심리지수 등
	Production(5)	Monthly	산업생산지수
	Real-estate(12)	Monthly	월세가격, 민간부문 미분양 건수
	Recruitment (6)	Monthly	규모별 고용현황
	Business cycle (3)	Monthly	경기종합지수 (선행, 동행, 후행)
	Labor (24)	Monthly	연령대 별 경제활동인구, 비경제활동인구, 실업자, 취업자
Big data	Payment (91)	Monthly	소비 유형별 신용카드 소비액 / 업종별 카드 사용금액 및 건수
	Text (4)	Daily	뉴스심리지수, 경제불확실성지수, 경제뉴스 텍스트에서 추출한 물가 지표
	High-freq. (5)	Daily	영화매출액, 수출입, 고속도로통행량, 전력사용량 등

(marginal model)을 고려할 수 있다.

$$y_t = \mu_k + m_k^{-1} \sum_{j=1}^{m_k} x_{j,k} \mathbf{v}_{j,k}^T \boldsymbol{\beta}_k + u_t, \quad t = 1, \ldots, T, \tag{3.7}$$

여기서, μ_k는 주변모형의 상수항을 나타낸다. 위의 주변모형 (3.7)에 근거하여 가장 최근의 종속변수 값을 가장 잘 추정하는 변수를 선별할 수 있다. 즉, 각 공변량별로 (3.7)에 근거한 검증오차를 계산한 뒤 검증오차가 가장 작은 값 순으로 변수를 선별하는 것이다. 이 때 (3.7)의 모형적합을 위해 다양한 방법을 활용할 수 있다. 본 연구에서는 변수선별의 안정성을 높이기 위하여 가중평균회귀, 가중중위수회귀, 그리고 간편화된 형태의 직교회귀를 이용하여 세 가지 검증오차를 계산한 뒤, 그 평균에 근거하여 변수를 선별하였다. 가중치는 최근 관찰치에 더 많은 중요도를 부여하기 위해 검증 시점과의 차이에 반비례하도록 설정하였다. 그리고 간편화된 직교회귀라 함은 데이터의 기울기(s.d(y_t)/s.d($x_{k,t}$))를 바탕으로 회귀계수를 추정하는 방법으로 측정오차(measurement error)가 있는 경우에 활용하는 직교회귀(orthogonal regression)의 아이디어를 응용한 것이다.

4. 실증 분석

4.1. 데이터

예측의 목표가 되는 종속변수는 GDP 구성항목 중 민간소비로 설정하고, 실질 민간소비 전년동기대비 증가율을 y_t로 고려하였다. 분석을 위해서 2000년 1분기부터 21년 3분기까지 약 20년간의 분기 데이터를 활용하였다. 여기서 민간소비는 연구 시점에서 가장 완전한 정보를 바탕으로 추계된 수치를 ground-truth로 가정하고자 2000년부터 2020년까지는 확정치를 그리고 2021년은 잠정치를 이용하였다. Figure 3은 2015년 1분기부터 21년 3분기 실질 민간소비 금액과 전년동기대비 증가율을 나타낸다. 본 연구에서는 마지막 11개 분기의 전년동기대비 증가율을 nowcasting하여 실제값과 비교해봄으로써 그 성능을 평가하였다. 11개 분기의 예측을 위해서는 해당 분기 종료시점까지 가용할 수 있는 모든 데이터를 활용하였다. 특히 2020년 초에 발생한 코로나19 팬데믹을 기점으로 이전 4개 분기와 이후 7개 분기의 nowcasting 성능을 비교해 봄으로써 경제 불확실성이 증가한 상황에서 빅데이터의 활용이 nowcasting에 어떠한 영향을 미치는지 확인해 보았다.

민간소비에 영향을 미치는 공변량은 매우 광범위하다. 통상적인 GDP nowcasting에 사용되어 온 거시경제지표 및 통계청에서 주기적으로 발표하는 경제관련 조사자료부터 최근 민간 소비지표 추정에 흔히 활용되는 신용카드 빅데이터까지 여러가지 공변량들을 활용할 수 있다. 본 연구에서는 한국은행과 통계청 데이터베이스에서 제공하는 다양한 공변량을 최대한 고려하여 총 281개의 후보 변수를 수집하였다(전체 공변량 변수

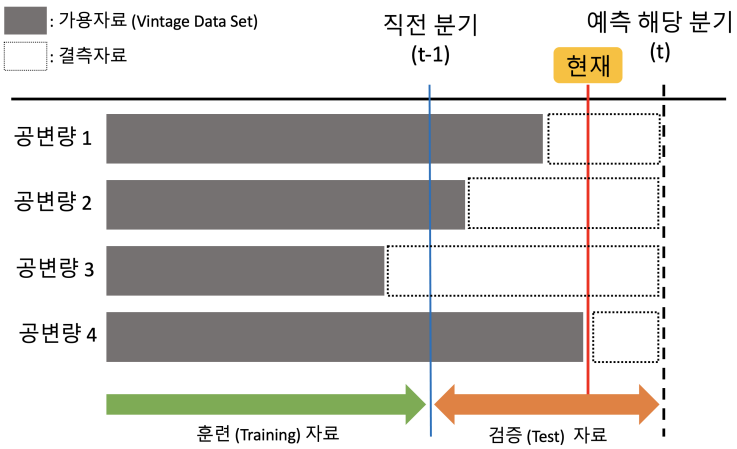


Figure 4: *Construction of vintage datasets.*

목록은 부록 참조). 이 중 각 변수들이 제공하는 정보들간의 중복성, 공표주기 및 공표지연 등으로 인한 가용성 등을 고려하여 최종적으로 208개의 변수를 이용한 모형을 고려하였다. 총 208개의 변수를 그 특성에 따라 11개의 범주로 나누었고, 크게는 통상적인 거시경제데이터와 빅데이터로 구분하여 정리하였다 (Table 1 참조). 빅데이터는 자료의 수집빈도가 높고 공표지연이 없어 nowcasting에 유용한 정보를 제공할 것으로 기대하였다.

종속변수를 기준으로 모형적합에 활용할 수 있는 표본의 크기는 최대 90개를 넘지 못한다. 더욱이 공변량의 경우 이용가능 시점이 더 최근인 경우도 많기 때문에 경우에 따라 30개 미만의 데이터를 바탕으로 200여개의 공변량을 포함하는 고차원 모형을 적합시켜야 하므로, 이는 전형적인 고차원 혹은 초고차원 문제가 된다. 특히 MIADS 모형의 가중치 함수의 복잡도를 높이기 위해 여러개의 기저함수를 활용하게 되면 추정해야 하는 모수가 크게 늘어날 뿐만 아니라, 추가적인 표본이 수집되는 속도에 비해 활용가능한 변수의 개수가 더욱 빠르게 증가할 것으로 예상된다. 따라서 의미있는 공변량을 신속 정확하게 선별하는 것이 매우 중요할 것으로 예상할 수 있다.

4.2. 빈티지 데이터셋 구축과 결측값 예측

Nowcasting의 핵심은 예측 대상분기와 관련된 정보를 최대한 많이 활용하는 것이다. 하지만 공변량으로 사용되는 많은 경제지표들도 공표지연의 문제가 발생하고 지연 기간도 다르기 때문에 현재 시점에서 실제로 활용가능한 데이터, 즉 빈티지 데이터셋(vintage data set)을 구성하는 것이 필요하다 (Figure 4 참조).

MIDAS를 활용하는 경우, 결측값을 대체하지 않더라도 모형을 적합시키는 것이 불가능한 것은 아니지만, 본 연구에서는 nowcasting의 성능을 최적화하기 위해 추가적인 결측치 대체를 실시하였다. 이는 코로나19 팬데믹 상황과 같이 경제의 불확실성이 큰 경우에는 공표지연 등으로 결측이 발생한 단기간 동안 경제상황이 급변할 수 있고, 결측치를 무시하는 경우 예측의 정확도가 현저히 떨어질 수 있기 때문이다. 본 연구에서는 가중회귀대체법(weighted regression imputation)을 활용하였다. 즉 결측자료의 시점에 가까운 값들일수록 상대적으로 많은 가중치를 주고 회귀모형을 적합시켜 그 예측치로 대체하는 방법이다. 이 때, 대체의 정확도와 안정성을 높이기 위해 1차부터 3차 다항회귀까지 세 가지 모형을 적합시켜 각 모형의 예측치 평균으로 대체하는 방법을 활용하였다. 기존 연구 등에서는 자기회귀모형을 결측치 대체 모형으로 많이 사용하였다. 하지만, 경제의 불확실성이 높은 경우에 자기회귀모형은 직전 관찰치의 영향을 지나치게 많이 받아 그 활용이 제한적

Table 2: Summary of variables selected for each nowcasting point

	Big data			Macroeconomic data							
	Payment	Text	High-freq	Finan. markets	Price	Consump.	Product.	Real-estate	Recruit.	Busi. Cycle	Labor
2019/1	8	1	0	0	1	12	0	0	1	2	5
2019/2	9	1	1	0	0	13	0	1	3	0	2
2019/3	8	0	1	1	1	12	1	1	3	1	1
2019/4	13	0	1	1	1	10	1	0	1	0	2
2020/1	10	2	0	3	0	7	4	0	0	1	3
2020/2	4	0	1	2	0	15	0	3	2	0	3
2020/3	5	0	1	3	1	11	0	1	1	1	6
2020/4	19	0	1	0	0	3	2	0	2	0	3
2021/1	13	0	1	0	1	7	0	0	2	0	6
2021/2	5	0	0	4	2	14	0	0	1	1	3
2021/3	7	1	2	1	1	11	0	0	0	2	5

일 것으로 판단하였다.

4.3. 변수선별

빈티지 데이터셋을 구축하고 결측치를 대체한 뒤, 3.3절에서 제안된 방법을 바탕으로 매분기 30개의 변수를 선별하여 모형을 적합하였다. 선별하는 변수의 개수는 통상적으로 표본의 크기를 넘지 않게 설정하는 것이 좋다고 알려져 있다 (Fan과 Lv, 2008). Table 2는 각 예측 시기별로 선별된 변수들의 구성을 나타낸 표이다. 신용카드 빅데이터와 소비자 동향조사나 소매업태별 판매액지수 등을 포함하는 소비관련 지표들이 민간소비 nowcasting에 유용한 정보를 제공한다는 사실을 확인하였다. 한 가지 주목할 점은 두 종류의 변수들이 차지하는 상대적 비중이 코로나19 경과와 함께 경제 불확실성 정도에 따라 변화하였다는 점이다. 경제 불확실성이 높은 시기(20년 1분기 및 21년 1분기 전후)에는 신용카드 빅데이터가 nowcasting에 더 유용한 것으로 나타난 반면, 동 시기를 제외한 다른 시기에서는 전통적인 소비 관련 지표들이 더 중요한 역할을 하는 것을 확인하였다. 전통적인 거시경제지표의 경우 공표주기가 빠르지 않고 공표지연 역시 불가피하므로 경제상황이 급변하는 상황에서 민간소비 상황을 반영하는 데 어려움이 있기 때문으로 판단된다.

4.4. Nowcasting 성능 비교

이제 3절에서 소개한 다양한 방법을 적용하여 민간소비 nowcasting을 실시하고 그 성능을 평가해 보았다. 우선 가장 기본적인 모형으로 종속변수만을 활용한 1차 자기상관회귀모형(AR(1))을 고려하였다. 기계학습 방법으로는 LASSO와 SCAD 벌점화를 이용한 선형회귀모형(LASSO, SCAD), 커널분위수회귀(KQR), 충분차원축약에 기반한 단일지표모형(SDR), 랜덤 포레스트(RF) 그리고 이 모든 추정량을 결합한 앙상블 방법(각 방법의 추정치 단순 평균)을 고려하였다. 실제 연구과정에서는 elastic net (Zou와 Hastie, 2005) 등과 같은 다른 벌점함수도 적용해 보았으나, 예측성능의 측면에서 LASSO와 SCAD가 가장 안정적인 성능을 보이는 것을 확인하였다. 단일지표모형 추정을 위한 충분차원축약은 최소제곱추정량을 활용하였으며, 차원축약 이후 함수추정에는 자연 3차 스플라인 회귀(cubic natural spline regression)를 이용하였다. 커널분위수회귀에서는 $\tau = 0.5$로 설정하고 선형커널을 이용하였다. 가우시안 커널(Gaussian kernel)도 적용하였으나 선형커널에 비해 예측성능이 좋지 않음을 확인하였다 (5절 참조). 이 외에도 LASSO의 비선형 확장인 COSSO (Lin과 Zhang, 2006), 인공신경망모형, L_2-부스팅 등도 적용해 보았으나 해당 방법들은 모형의 복잡도가 지나치게 높아서

Table 3: Comparison of MAE among the proposed models (Relative efficiency compared to the AR(1) model): It can be observed that the performance of nowcasting significantly improves regardless of the training methods applied when variable selection is performed

• In the case where variable selection is not performed.							
		LASSO	SCAD	KQR	SDR	RF	결합
$L = 0$	Entire period	0.844	0.964	0.622	0.908	1.077	0.713
	Before COVID	0.235	0.321	0.120	0.211	0.187	0.199
	After COVID	1.196	1.336	0.912	1.311	1.591	1.009
$L = 1$	Entire period	0.779	1.014	0.574	0.797	1.065	0.714
	Before COVID	0.222	0.207	0.107	0.114	0.165	0.134
	After COVID	1.101	1.481	0.844	1.192	1.585	1.050
• In the case where variable selection is performed.							
		LASSO	SCAD	KQR	SDR	RF	결합
$L = 0$	Entire period	0.431	0.395	0.396	0.426	0.693	0.413
	Before COVID	0.131	0.152	0.129	0.185	0.042	0.109
	After COVID	0.605	0.535	0.549	0.565	1.068	0.588
$L = 1$	Entire period	0.523	0.656	0.431	0.402	0.671	0.447
	Before COVID	0.102	0.102	0.056	0.016	0.116	0.056
	After COVID	0.766	0.976	0.648	0.624	0.992	0.673

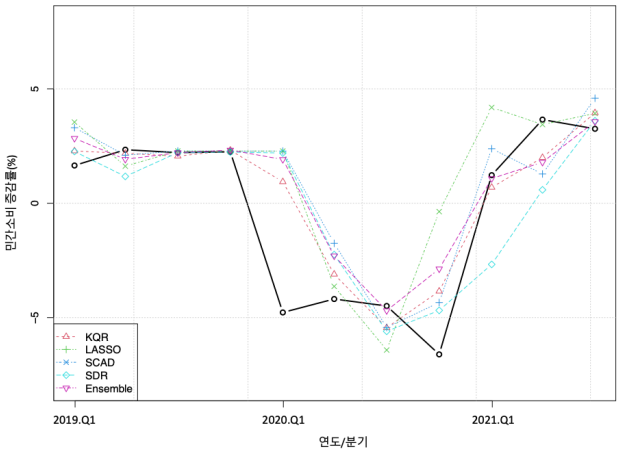
과적합이 일어나거나 초기치에 지나치게 민감하게 반응하여 예측성능이 크게 떨어지는 것을 확인하였다. 마지막으로, 각 기계학습모형의 적합에 필요한 조율모수는 교차검증(cross-validation)을 통해 추정하였다. 즉, 시계열자료의 특성을 반영하여 자료를 분할하도록 수정하였다.

본 실증분석에서는 t 시점(혹은 분기)의 민간소비를 t 시점에 nowcasting하는 상황('end of quarter' 시나리오)을 설정하였다. 2절에서 설명한 바와 같이 nowcasting 시점은 t 시점 이전에도 가능하지만 t 시점에 가까울수록 공변량의 가용정보가 많기 때문에 보다 정확한 예측이 가능해진다. 모형의 성능을 평가하기 위해서 평균절대오차(mean absolute deviance)을 고려하였다. 예측하고자 하는 분기의 관찰값을 y_t, t시점에서 사용가능한 공변량 정보를 최대한 이용하여 nowcasting한 값을 $\hat{y}_t$라 하면, 평균절대오차는 다음과 같이 절대오차의 평균으로 계산된다.

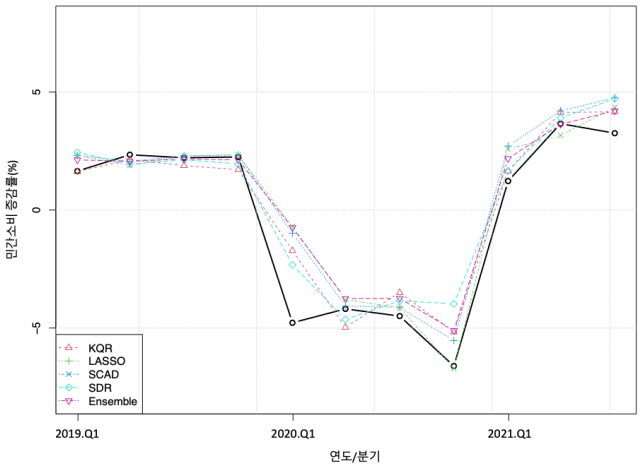
$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|\,.$$

본 연구에서 검증의 대상이 되는 분기는 총 11개 이므로 $T = 11$로 하여 전체 MAE를 계산하였다. 나아가 코로나19 발생 이전의 4개 분기와 이후의 7개 분기를 구분하여 각각의 MAE를 계산하여 코로나19 전후의 nowcasting 성능도 비교해보았다. Table 3은 본 연구에서 제안한 다양한 방법의 예측 성능을 기본모형인 1차 자기상관회귀모형과의 MAE값에 대한 비로 나타낸 것이다. 즉, 해당 값이 0에 가까울수록 좋은 성능의 의미하며, 1이 넘는 경우 기본모형에 비해 성능이 떨어짐을 의미한다. 기본모형인 1차 자기상관회귀모형(AR(1))의 경우, 코로나19 이전의 안정적인 경제상황에서는 어느 정도의 예측력을 보여주고 있지만, 코로나19 발생 이후에는 예측력이 현저하게 떨어진다. 이는 이전 분기의 관찰치에 의존하는 자기상관회귀모형의 특성상 자연스러운 결과로 해석된다.

본 연구에서 제안한 MIADS-회귀모형에는 1차 자기상관 항을 넣은 경우($L = 1$)와 그렇지 않은 경우($L = 0$)를 모두 고려하였다. 변수선별을 하지 않은 경우, 동 모형은 기본모형인 1차 자기상관모형에 비해서는 향상된 성능을 보이지만 코로나19 이후 분기의 민간소비 증가율 nowcasting에는 어려움을 보인다. 동



(a) In the case where variable selection is not performed



(b) In the case where variable selection is performed

Figure 5: *Comparison of private consumption forecasts between models without an AR(1) term* ($L = 0$)*: The effect of variable selection can be observed.*

모형은 SCAD와 같은 변수선택을 반영하는 경우에도 예측성능이 좋지 않은 것으로 나타난다. 통상적인 회귀모형의 경우, 종속변수와 공변량 간의 전체적인 상관관계를 통해 예측치를 추정하게 되는데, nowcasting의 경우 데이터 전체구간의 상관성 보다는 예측 대상과 가장 가까운 가장자리 값들의 연관성이 상대적으로 훨씬 중요하다. 이러한 점을 고려하지 않고 다수의 변수를 포함하면, 종속변수의 전반적인 트렌드는 잘 예측하지만, nowcasting의 대상이 되는 바깥쪽 끝부분의 값은 제대로 예측하지 못하는 문제가 발생한다. Table 3에서 변수선별을 한 경우의 전체 기간 예측성능을 살펴보면, 선형모형인 SCAD나 선형커널분위수회귀(KQR)가 비선형 모형인 랜덤포레스트(RF)나 충분차원축소 기반 방법(SDR)보다 좋은 성능을 보이는데, 이 역시 주어

진 데이터의 바깥쪽 부분을 예측하는 nowcasting의 특성상 모형의 복잡도가 지나치게 높은 경우 과적합 등의 문제로 인하여 예측성능이 저하되기 때문이다. 특히 코로나19 이후의 상황처럼 데이터의 변동성이 커지는 경우, 비선형 모형을 이용한 nowcasting의 예측오차가 높아지는 것을 확인할 수 있다. 한편, 변수선별을 실시한 뒤 선별된 변수만으로 모형적합을 한 경우 변수선별을 실시하지 않은 경우에 비해 nowcasting의 성능이 크게 향상됨을 확인하였다. 특히 모형의 복잡도가 높지 않은 SCAD나 커널분위수회귀가 안정적인 예측성능을 보여주었으며, 커널분위수회귀가 가장 좋은 성능을 보여주었다. 커널분위수회귀의 경우, SCAD와는 달리 추가적인 변수선택을 하지 않지만, 절대 손실함수를 사용하기 때문에 데이터의 바깥쪽 끝부분을 예측하는 nowcasting 문제에서 좀 더 안정적인 성능을 보여주는 것으로 판단된다. 실제 분위수회귀는 중위수 뿐만 아니라 다양한 수준의 분위수를 예측치로 활용할 수 있다는 장점이 있어, 다양한 계량경제 모형에 널리 활용되고 있다 (Koenker와 Hallock, 2001). 충분차원축소 기반의 단일지표 모형은 변수선별을 하지 않았을 때는 성능이 좋지 못했지만, 변수선별 후에는 그 성능이 크게 개선된 것을 확인할 수 있다. 이는 변수선별을 통해 모형의 복잡도를 크게 줄여 비모수적 추정방법의 안정성을 향상시킨 것이 예측성능을 개선하였기 때문으로 보인다. 실증분석에 활용한 모든 모형을 결합하여 예측하는 것도 매우 우수한 성능을 보이고 있는 사실에도 주목할 필요가 있다. 결합 모형의 경우, 앙상블 효과로 인해 개별모형이 가진 편향의 효과가 감소하고 예측값의 분산이 줄어들기 때문이다. 예측의 정확성의 관점에서 본다면 nowcasting에서 결합모형의 활용 가능성을 좀 더 면밀히 검토해볼 가치 있다고 판단된다.

마지막으로, Figure 5는 변수선별에 따른 예측 결과를 AR(1) 항이 없는 ($L = 0$) 모형 하에서 비교한 것이다(AR(1)항이 있는 모형에 대한 결과는 부록 참조). 코로나19 이전의 4분기 동안에는 모든 방법이 안정적인 성능을 보이고 있지만, 코로나19 이후의 시기에는 변수선별에 기반한 방법이 월등히 우수한 성능을 보이고 있다.

5. 결론 및 제언

본 연구에서는 빅데이터 기반 민간소비의 nowcasting 문제를 다루었다. nowcasting은 2000년대 이후 매우 활발히 연구되어 온 분야지만, 최근 빅데이터의 출현과 기계학습 방법론의 비약적인 발전으로 더욱 다양하고 새로운 시도들이 이루어지고 있다. 민간소비의 경우, 국민들의 경제활동 결과를 직접적으로 반영하므로 다양한 종류의 빅데이터를 활용하여 nowcasting 성능을 높일 필요가 있다. 본 연구에서는 대용량 빅데이터 활용이 용이한 부분모형에 다양한 기계학습 방법론을 적용하여 nowcasting을 실시하고 그 활용 가능성을 확인하였다.

본 연구 결과가 시사하는 내용을 요약하면 다음과 같이 세 가지로 정리할 수 있다. 첫째, nowcasting을 위한 변수선별의 중요성이다. 데이터 관련 기술의 발전으로 nowcasting에 활용할 수 있는 변수의 개수는 점차 증가하는 추세이다. 특히 빅데이터의 활용이 더 대중화되면 분석에 포함될 가능성이 있는 후보변수들이 크게 증가할 것이다. 하지만 변수, 혹은 그에 해당하는 정보의 증가가 반드시 모형의 성능을 향상시키는 방향으로 작용하는 것은 아니다. 특히 GDP nowcasting과 같이 모형적합에 활용할 수 있는 표본의 수가 충분하지 않은 경우에는 예측력 향상에 도움이 되는 변수를 상황에 맞게 적절히 선별하는 것이 매우 중요하다. 본 연구에서는 초고차원 기계학습에서 널리 활용되는 방법인 변수선별방법의 활용을 제안하고 그 효과를 확인하였다. 하지만 변수선별이 반드시 데이터 기반 알고리즘에 의해서만 가능한 것은 아니다. GDP nowcasting과 같이 오랜 시간 연구되어 온 주제의 경우, 해당 분야의 전문가 의견과 경제이론 등을 바탕으로 1차적으로 변수를 선별하는 것이 가능하다. 나아가 이러한 경험적 변수선별과 데이터 기반 변수선별을 적절한 방법으로 결합하면 nowcasting의 성능을 추가적으로 개선할 수 있을 것으로 예상한다.

둘째, 본 연구를 통해 민간소비 nowcasting을 위한 빅데이터의 활용 가능성을 확인하였다. 통상적인 거시경제지표 혹은 조사자료 등은 대부분 국가에서 관리하는 공식통계이므로 공표주기가 월별인 경우가 많으며 공표지연이 발생한다. 따라서 국민들이 체감하는 경제상황을 적시에 파악하기에는 한계가 있다. 따라서 민간

기업으로부터 수집이 가능한 신용카드 사용액 등의 자료와 뉴스 기사나 포털 검색 정보 등을 바탕으로 추출한 텍스트 자료 등 빅데이터를 활용하면 민간소비 nowcasting의 성능을 향상시킬 수 있을 것으로 예상된다 (Lee와 Hwang, 2016). 본 연구에서는 신용카드 데이터, 경제 뉴스기사 기반의 텍스트 자료 등을 포함한 빅데이터를 민간소비 nowcasting에 활용해 보았다. 그 결과, 코로나19 팬데믹 사태와 같이 경제 불확실성이 높은 상황에서는 빅데이터 활용가치가 상대적으로 높아진다는 것을 확인하였다.

마지막으로, 빅데이터를 활용한 nowcasting에서 교량방정식이나 MIDAS 모형과 같은 부분모형의 활용 가능성을 제고하였다. 현재 GDP nowcasting에서 가장 널리 쓰이는 방법은 동적인자모형 같은 결합모형에 근거한 방법이다. 결합모형을 활용하면 nowcasting에서 발생하는 혼합주기나 ragged-edge 문제를 손쉽게 해결할 수 있을 뿐만 아니라, 모형의 해석이나 공변량 예측과 같은 보조적인 정보의 획득이 용이하기 때문이다. 하지만 대용량 빅데이터를 활용하기 위해서는 알고리즘의 계산효율성(computational efficiency)과 확장성(scalability)이 매우 중요하다. 이러한 측면에서 본다면, 빅데이터를 활용함에 있어 부분모형을 활용하는 것이 동적인자모형과 같은 결합모형에 비해 더 나은 선택이 될 수 있을 것이다.

한편, 본 연구의 한계점으로 다음과 같은 부분을 지적할 수 있다. 본 연구에서는 nowcasting을 위해 기계학습 방법론을 적용하였는데, 이는 시계열 방법론에 근거한 전통적인 계량경제모형과는 차이가 있다. 특히 시계열자료가 가지는 특성을 제대로 모형에 반영하지 않았다는 비판을 받을 수 있으며, 이는 타당한 지적이라 하겠다. 그럼에도 불구하고, 본 연구에서는 빅데이터를 보다 효율적으로 활용하기 위해 통상적인 시계열 방법론에서 활용하는 정상성 등의 모형가정을 고려하지 않았다. 그렇기 때문에 본 연구를 통해 제안된 방법에 시계열자료의 특성을 세밀하게 고려하여 반영할 수 있도록 개선한다면 nowcasting 성능을 더욱 향상시킬 수 있을 것으로 기대한다. 또한, 본 연구는 실증적 분석에 주안점을 두었으며, 제안된 방법론들과 관련된 이론적 내용에 대해서는 자세히 다루지 못하였다. 이후, 추가적인 연구를 통해 실증적 결과를 뒷받침 할 수 있는 이론적 결과를 유도할 수 있을 것으로 기대한다.

Appendix A: 연구에 활용된 기계학습 방법론 추가 설명

– SCAD 벌점화 선형회귀

주어진 자료 $(y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^p$에 대해 다음과 같은 선형모형을 고려하자.

$$y_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right),$$

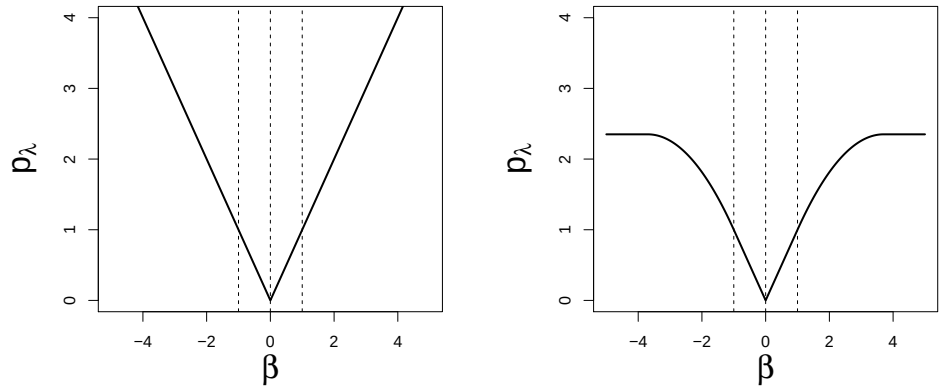
이 때 공변량의 개수 p가 많은 경우 다음과 같은 벌점화 회귀를 사용하여 회귀계수 $\boldsymbol{\beta}$를 추정할 수 있다.

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \boldsymbol{\beta}^T \mathbf{x}_i\right)^2 + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right),$$

여기서 $p_\lambda(\theta)$는 벌점함수를 나타내며, 대표적인 예로는 LASSO $p_\lambda(\theta) = \lambda\theta$와 능형 벌점함수 $p_\lambda(\theta) = (\lambda/2)\theta^2$ 등이 있다. 특히 LASSO 벌점항은 추정과 동시에 변수선택도 가능하여 다양한 분야에 널리 활용된다. 하지만 LASSO 추정량은 편향이 크다는 사실이 알려져 있으며 이로 인해 추정의 효율이 떨어지게 된다. 이러한 단점을 개선하기 위한 방법으로 다음의 SCAD 벌점함수가 개발되었다 (Fan과 Li, 2001). SCAD 벌점함수는 다음과 같이 도함수의 형태로 정의되며, LASSO와 달리 회귀계수의 값이 0에서 멀어지면 추가적인 벌점을 부여하지 않아 편향을 줄일 수 있다 (Figure A.1 참조).

$$p'_\lambda(\theta) = \lambda\left\{\mathbb{1}(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda}\mathbb{1}(\theta > \lambda)\right\}, \quad \theta > 0,$$

여기서 a는 2보다 큰 상수로 통상적으로 2.7을 사용한다. 다음은 LASSO와 SCAD 벌점함수를 비교한 그림이다. SCAD 벌점항이 더 좋은 추정량을 제공하지만 목적함수를 비볼록(non-convex)하게 만들어 그 계산이 더



(a) LASSO (b) SCAD

Figure A.1: *Comparison of LASSO and SCAD penalty terms: LASSO increases the penalty term as it moves further away from 0, whereas SCAD reduces bias by not imposing additional penalties once the distance from 0 exceeds a certain level.*

어려워진다. 하지만 다양한 계산 알고리즘이 개발되었으며, 적어도 제곱손실함수를 이용한 회귀문제에서는 LASSO만큼이나 그 계산이 간단하다 (Breheny와 Huang, 2011).

– 충분차원축소를 이용한 단일지표 모형의 추정

단일지표모형 (3.5)에서 $\boldsymbol{\beta}$를 추정하기 위한 충분차원축소는 다음과 같은 가정에서 출발한다:

$$E\left(Y \mid \mathbf{X}\right) = E\left(Y \mid \boldsymbol{\beta}^T\mathbf{X}\right), \tag{A.1}$$

즉, (A.1)의 가정하에 $\boldsymbol{\beta}$를 찾게 되면 p차원 공변량 $\mathbf{X}$를 1차원 공변량 $\boldsymbol{\beta}^T\mathbf{X}$로 정보의 손실없이 축약할 수 있게 된다. 여기서 (A.1)의 가정을 만족하는 $\boldsymbol{\beta}$는 유일하지 않지만, $\boldsymbol{\beta}$로 생성되는 공간은 유일하기 때문에 충분차원축약에서는 $\boldsymbol{\beta}$가 아닌 $\boldsymbol{\beta}$로 생성되는 공간, 즉 span($\boldsymbol{\beta}$)를 추정하는 것을 목표로 한다. span($\boldsymbol{\beta}$)를 추정하는 방법은 매우 다양하지만, 가장 간단한 방법은 최소제곱추정량을 이용하는 것이다. 즉, $\hat{\boldsymbol{\beta}}$를 주어진 자료 $(y_i, \mathbf{x}_i) \in \mathbb{R}\times\mathbb{R}^p$로 부터 구한 최소제곱추정량이라 하면 다음이 성립한다.

$$\hat{\boldsymbol{\beta}} \in \text{span}\left(\boldsymbol{\beta}\right),$$

따라서 최소제곱추정법을 통해 우선 $\hat{\boldsymbol{\beta}}$를 계산한 뒤 $\mathbf{z}_i = \mathbf{x}_i^T\boldsymbol{\beta}$와 y_i사이의 관계를 비모수적 방법(예, 스플라인 회귀 등)으로 추정하면 단일지표모형의 추정을 완료할 수 있다. 단, 본 연구에서는 공변량의 개수가 많기 때문에 최소제곱추정량을 바로 구할 수가 없으며, 대안으로 SCAD 벌점화 추정량을 구하여 변수선택을 진행한 뒤, 선택된 변수들만을 바탕으로 최소제곱 추정량을 계산하는 방법을 활용하였다.

– 커널 분위수회귀

커널분위수회귀 (3.6)를 적합하기 위해서 대표자 정리(representer theorem)를 이용할 수 있다. 대표자 정리는 (3.6)의 해가 다음과 같은 형태로 주어진다는 것을 보장한다는 내용이다.

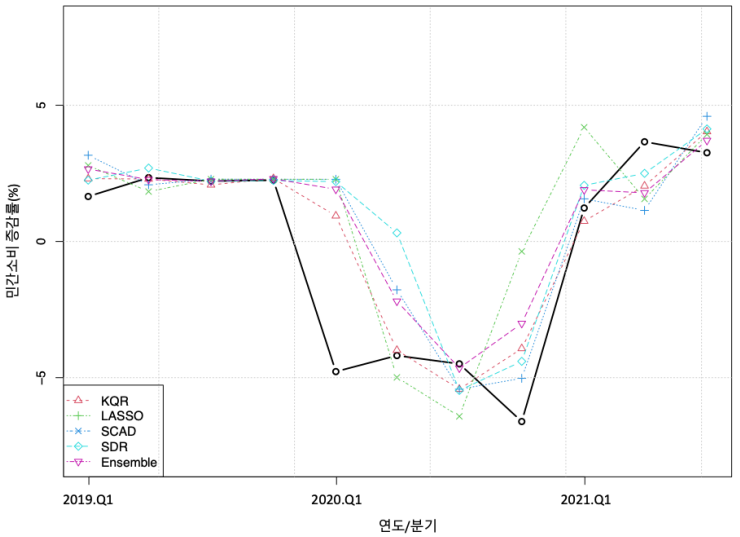
$$f(\mathbf{z}) = \alpha_0 + \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{z}, \mathbf{z}_i). \tag{A.1}$$

이제, (A.1)를 (3.6)에 대입하면 다음과 같은 최적화 문제로 표현할 수 있다.

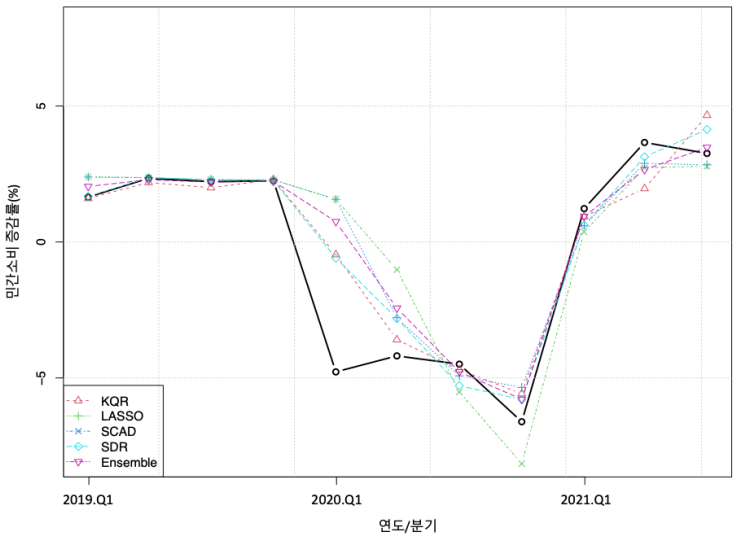
$$\min_{\alpha_0, \alpha} \sum_{i=1}^{n} \rho_\tau \left(y_t - \alpha_0 - \sum_{j=1}^{n} y_j \alpha_j K\left(\mathbf{z}_i, \mathbf{z}_j\right) \right) + \frac{\lambda}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K\left(\mathbf{z}_i, \mathbf{z}_j\right).$$

위의 문제는 통상적인 이차계획법(quadratic programming)의 형태로 표현할 수 있음이 잘 알려져 있다 (Li 등, 2007). 또한 추정해야하는 모수의 개수가 p가 아닌 표본의 크기 n에 의존하는 것을 알 수 있으며, 공변량 $\mathbf{z}$에 대한 정보가 커널함수 K를 통해서만 전달되기 때문에 $\mathbf{z}$가 고차원인 경우에도 계산량을 증가시키지 않는다는 장점이 있다.

Appendix B: AR(1)항을 포함한 ($L = 0$) 모형 예측 결과



(a) In the case where variable selection is not performed



(b) In the case where variable selection is performed

Figure B.1: *Comparison of private consumption forecasts between models with an AR(1) term* ($L = 0$)*: The effect of variable selection can be observed.*

Appendix C: 공변량 전체 목록

Table C.1: Complete list of covariates

설명	원_지수	주기	시작시점	공표지연	분석포함
환율					
원/미국달러(매매기준율)	원자료	일	2000-01-04	1	O
원/일본엔	원자료	일	2000-01-04	1	O
원/유로	원자료	일	2000-01-04	1	O
고빈도데이터					
동행종합지수	지수	월	2012-01-31	0	X
GDP(성장률)	지수	분기	2012-03-31	2	X
최대전력사용량	원자료	일	2012-01-01	1	O
개인신용카드사용액	원자료	일	2016-01-01	1	O
수출액	원자료	일	2012-01-01	1	O
수입액	원자료	일	2012-01-01	1	O
영화매출액	원자료	일	2012-01-01	1	O
고속도로통행량	원자료	일	2012-01-01	1	O
통화금융통계					
M2(광의통화,평잔)(십억원)	원자료	월	2000-01-31	2	X
LF(평잔)(십억원)	원자료	월	2000-01-31	2	X
카드데이터					
식료품 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
식료품 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
의류 및 신발 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
의류 및 신발 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
주택 및 연료 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
주택 및 연료 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
가구, 가사 비품 및 기타 생활용품 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
가구, 가사 비품 및 기타 생활용품 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
보건 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
보건 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
개인운송장비 구매 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
개인운송장비 구매 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
개인운송장비 운영 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
개인운송장비 운영 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
여객운송서비스 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
여객운송서비스 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
상품운송서비스 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
상품운송서비스 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
정보통신 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
정보통신 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
오락용품 및 애완동물 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
오락용품 및 애완동물 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
오락문화서비스 및 여행 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
오락문화서비스 및 여행 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
교육서비스 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
교육서비스 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
일반음식점 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O

일반음식점 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
주점 등 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
주점 등 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
숙박서비스 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
숙박서비스 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
보험 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
보험 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
개인미용 및 기타 서비스 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
개인미용 및 기타 서비스 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
백화점 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
백화점 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
할인점 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
할인점 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
편의점 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
편의점 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
홈쇼핑 및 인터넷판매 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
홈쇼핑 및 인터넷판매 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
면세점 및 기타 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
면세점 및 기타 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
상품권+세금공과금 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
상품권+세금공과금 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
기타 기간중 카드 사용 금액	원자료	월	2015-11-30	0	O
기타 기간중 카드 사용 건수	원자료	월	2015-11-30	0	O
경기종합지수					
선행종합지수	지수	월	2000-01-31	0	O
동행종합지수	지수	월	2000-01-31	0	O
후행종합지수	지수	월	2000-01-31	0	O
경제심리지수					
경제심리지수	지수	월	2003-01-31	0	O
경제활동인구조사-연령별					
15 - 19세_경제활동인구 (천명)	원자료	월	2000-01-31	1	O
15 - 19세_비경제활동인구 (천명)	원자료	월	2000-01-31	1	O
15 - 19세_실업자 (천명)	원자료	월	2000-01-31	1	O
15 - 19세_취업자 (천명)	원자료	월	2000-01-31	1	O
20 - 29세_경제활동인구 (천명)	원자료	월	2000-01-31	1	O
20 - 29세_비경제활동인구 (천명)	원자료	월	2000-01-31	1	O
20 - 29세_실업자 (천명)	원자료	월	2000-01-31	1	O
20 - 29세_취업자 (천명)	원자료	월	2000-01-31	1	O
30 - 39세_경제활동인구 (천명)	원자료	월	2000-01-31	1	O
30 - 39세_비경제활동인구 (천명)	원자료	월	2000-01-31	1	O
30 - 39세_실업자 (천명)	원자료	월	2000-01-31	1	O
30 - 39세_취업자 (천명)	원자료	월	2000-01-31	1	O
40 - 49세_경제활동인구 (천명)	원자료	월	2000-01-31	1	O
40 - 49세_비경제활동인구 (천명)	원자료	월	2000-01-31	1	O
40 - 49세_실업자 (천명)	원자료	월	2000-01-31	1	O
40 - 49세_취업자 (천명)	원자료	월	2000-01-31	1	O
50 - 59세_경제활동인구 (천명)	원자료	월	2000-01-31	1	O
50 - 59세_비경제활동인구 (천명)	원자료	월	2000-01-31	1	O

50 - 59세_실업자 (천명)	원자료	월	2000-01-31	1	O
50 - 59세_취업자 (천명)	원자료	월	2000-01-31	1	O
60세이상_경제활동인구 (천명)	원자료	월	2000-01-31	1	O
60세이상_비경제활동인구 (천명)	원자료	월	2000-01-31	1	O
60세이상_실업자 (천명)	원자료	월	2000-01-31	1	O
60세이상_취업자 (천명)	원자료	월	2000-01-31	1	O
미분양주택현황보고-규모별미분양					
공공부문미분양	원자료	월	2007-01-31	1	O
민간부문미분양	원자료	월	2007-01-31	1	O
전국주택가격동향조사-규모별월세가격지수					
아파트전국_규모40㎡이하	지수	월	2017-11-30	1	O
아파트전국_규모40㎡ 60㎡	지수	월	2015-06-30	1	O
아파트전국_규모60㎡ 85㎡	지수	월	2015-06-30	1	O
아파트전국_규모85㎡ 102㎡	지수	월	2015-06-30	1	O
아파트전국_규모102㎡ 135㎡	지수	월	2015-06-30	1	O
아파트전국_규모135㎡초과	지수	월	2015-06-30	1	O
연립전국_규모40㎡이하	지수	월	2017-11-30	1	O
연립전국_규모40㎡ 60㎡	지수	월	2015-06-30	1	O
연립전국_규모60㎡ 85㎡	지수	월	2015-06-30	1	O
연립전국_규모85㎡초과	지수	월	2015-06-30	1	O
사업체노동력조사-고용부문					
300인미만_근로자_상용 (명)	원자료	월	2009-06-30	1	O
300인미만_근로자_임시일용 (명)	원자료	월	2009-06-30	1	O
300인미만_종사자_기타 (명)	원자료	월	2009-06-30	1	O
300인이상_근로자_상용 (명)	원자료	월	2009-06-30	1	O
300인이상_근로자_임시일용 (명)	원자료	월	2009-06-30	1	O
300인이상_종사자_기타 (명)	원자료	월	2009-06-30	1	O
통화금융통계-대출금리					
기업대출(연리%)	원자료	월	2000-01-31	0	O
가계대출(연리%)	원자료	월	2000-01-31	0	O
공공및기타부문대출(연리%)	원자료	월	2000-01-31	0	O
생산자물가조사					
생산자물가지수	지수	월	2000-01-31	1	O
소비자동향조사					
현재생활형편	지수	월	2009-12-31	0	O
현재경기판단	지수	월	2009-12-31	0	O
생활형편전망	지수	월	2009-12-31	0	O
향후경기전망	지수	월	2009-12-31	0	O
취업기회전망	지수	월	2009-12-31	0	O
금리수준전망	지수	월	2009-12-31	0	O
가계수입전망	지수	월	2009-12-31	0	O
소비지출전망	지수	월	2009-12-31	0	O
내구재지출전망	지수	월	2009-12-31	0	O
의류비지출전망	지수	월	2009-12-31	0	O
외식비지출전망	지수	월	2009-12-31	0	O
여행비지출전망	지수	월	2009-12-31	0	O
교육비지출전망	지수	월	2009-12-31	0	O
의료,보건비지출전망	지수	월	2009-12-31	0	O

교양,오락,문화생활비지출전망	지수	월	2009-12-31	0	O
교통비및통신비지출전망	지수	월	2009-12-31	0	O
주거비지출전망	지수	월	2013-01-31	0	O
현재가계저축	지수	월	2009-12-31	0	O
가계저축전망	지수	월	2009-12-31	0	O
현재가계부채	지수	월	2009-12-31	0	O
가계부채전망	지수	월	2009-12-31	0	O
물가수준전망(1년후)	지수	월	2009-12-31	0	O
주택가격전망	지수	월	2013-01-31	0	O
임금수준전망	지수	월	2013-01-31	0	O
소비자심리지수	지수	월	2009-12-31	0	O
소비자동향조사-기대인플레이션					
물가인식	지수	월	2013-01-31	0	O
기대인플레이션율	지수	월	2002-02-28	0	O
소비자물가조사					
소비자물가지수	지수	월	2000-01-31	1	O
통화금융통계-수신금리					
순수저축성예금(연리%)	원자료	월	2000-01-31	0	O
시장형금융상품(연리%)	원자료	월	2000-01-31	0	O
외국인국제이동					
입국자	원자료	월	2013-01-31	6	X
출국자	원자료	월	2013-01-31	6	X
전산업생산지수					
전산업생산지수(농림어업 제외)	지수	월	2000-01-31	1	O
광공업생산지수	지수	월	2000-01-31	1	O
건설업생산지수	지수	월	2000-01-31	1	O
서비스업생산지수	지수	월	2000-01-31	1	O
공공행정생산지수	지수	월	2000-01-31	1	O
주민등록총인구					
주민등록총인구	원자료	월	2011-01-31	1	X
석유제품가격통계-주유소 평균 판매가격					
보통휘발유	원자료	월	2000-01-31	1	O
실내등유	원자료	월	2000-01-31	1	O
자동차용 경유	원자료	월	2000-01-31	1	O
소비 유형별 전국 신용카드 소비					
가구/가전-가구	원자료	월	2009-12-31	0	O
가구/가전-가전제품/정보통신기기	원자료	월	2009-12-31	0	O
공과금/개인 및 전문 서비스	원자료	월	2009-12-31	0	O
교육	원자료	월	2009-12-31	0	O
금융/보험	원자료	월	2009-12-31	0	O
기타	원자료	월	2009-12-31	0	O
숙박/음식-숙박	원자료	월	2009-12-31	0	O
숙박/음식-음식점	원자료	월	2009-12-31	0	O
식료품-건강보조식품	원자료	월	2009-12-31	0	O
식료품-일반식료품	원자료	월	2009-12-31	0	O
여행/교통-대중교통	원자료	월	2009-12-31	0	O
여행/교통-여행사/자동차임대	원자료	월	2009-12-31	0	O
여행/교통-항공사	원자료	월	2009-12-31	0	O

연료	원자료	월	2009-12-31	0	O
오락/문화-서적/문구	원자료	월	2009-12-31	0	O
오락/문화-스포츠/오락/여가	원자료	월	2009-12-31	0	O
의료/보건-일반병의원/기타의료기관	원자료	월	2009-12-31	0	O
의료/보건-종합병원	원자료	월	2009-12-31	0	O
의류/잡화-복식잡화	원자료	월	2009-12-31	0	O
의류/잡화-시계/귀금속/안경	원자료	월	2009-12-31	0	O
의류/잡화-의복/직물	원자료	월	2009-12-31	0	O
의류/잡화-화장품	원자료	월	2009-12-31	0	O
자동차-국산자동차신품	원자료	월	2009-12-31	0	O
자동차-기타운송수단	원자료	월	2009-12-31	0	O
자동차-자동차 부품 및 정비	원자료	월	2009-12-31	0	O
전자상거래/통신판매	원자료	월	2009-12-31	0	O
종합소매-대형마트/유통전문점	원자료	월	2009-12-31	0	O
종합소매-면세점	원자료	월	2009-12-31	0	O
종합소매-백화점	원자료	월	2009-12-31	0	O
종합소매-슈퍼마켓	원자료	월	2009-12-31	0	O
종합소매-편의점	원자료	월	2009-12-31	0	O
주가지수					
코스피	원자료	월	2000-02-28	1	O
코스닥	원자료	월	2003-12-31	1	O
소매업태별 판매액지수					
백화점	지수	월	2010-01-31	1	O
대형마트	지수	월	2010-01-31	1	O
면세점	지수	월	2010-01-31	1	O
체인 슈퍼마켓	지수	월	2010-01-31	1	O
일반 슈퍼마켓 및 잡화점	지수	월	2010-01-31	1	O
편의점	지수	월	2010-01-31	1	O
승용차 소매점	지수	월	2010-01-31	1	O
연료 소매점	지수	월	2010-01-31	1	O
가전,컴퓨터,통신기기 소매점	지수	월	2010-01-31	1	O
의복,신발,가방 소매점	지수	월	2015-01-31	1	O
음식,가정,문화상품 소매점	지수	월	2015-01-31	1	O
의약품,화장품,기타상품 소매점	지수	월	2015-01-31	1	O
인터넷 쇼핑	지수	월	2010-01-31	1	O
홈쇼핑	지수	월	2010-01-31	1	O
방문 및 배달 소매점	지수	월	2010-01-31	1	O
텍스트					
경제불확실성지수	지수	일	2005-01-15	0	O
뉴스심리지수	지수	일	2005-01-03	0	O
경제뉴스텍스트에서 추출한 소비지표	지수	일	2005-01-01	0	O
경제뉴스텍스트에서 추출한 물가지표	지수	일	2005-01-08	0	O

References

Aastveit KA, Fastbø, TM, Granziera E, Paulsen KS, and Torstensen KN (2020). Nowcasting norwegian household consumption with debit card transaction data, *Norges Bank Working Paper*, **2020**, 1–33.

Babii A, Ghysels E, and Striaukas J (2022). Machine learning time series regressions with an application to nowcasting, *Journal of Business & Economic Statistics*, **40**, 1094–1106.

Bajari P, Nekipelov D, Ryan SP, and Yang M (2015). Machine learning methods for demand estimation, *American Economic Review*, **105**, 481–485.

Bańbura M, Giannone D, Modugno M, and Reichlin L (2013). Now-casting and the real-time data flow, *ECB Working Paper*, **1564**, 1–55.

Barhoumi K, Darné O, and Ferrara L (2010). Are disaggregate data useful for factor analysis in forecasting French GDP?, *Journal of Forecasting*, **29**, 132–144.

Barnett W, Chauvet M, Leiva-Leon D, and Su L (2016). *Nowcasting Nominal gdp with the Credit-card Augmented Divisia Monetary Aggregates*, University of Kansas, Department of Economics, Lawrence, Kansas, 1–76.

Breheny P and Huang J (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, *The Annals of Applied Statistics*, **5**, 232–253.

Chan JC and Jeliazkov I (2009). Efficient simulation and integrated likelihood estimation in state space models, *International Journal of Mathematical Modelling and Numerical Optimisation*, **1**, 101–120.

Chang J, Tang CY, and Wu Y (2013). Marginal empirical likelihood and sure independence feature screening, *Annals of Statistics*, **41**, 2123–2148.

Chen B and Hood K (2020). Nowcasting of Advance Estimates of Personal Consumption of Services in the US National Accounts: Individual Versus Forecasting Combination Approach, *BEA Working Paper*, **2021**, 1–37.

Chen JC, Dunn A, Hood K, Driessen A, and Batch A (2019). Off to the races: A comparison of machine learning and alternative data for predicting economic indicators, In Katharine G. Abraham, Ron S. Jarmin, Brian Moyer, and Matthew D. Shapiro (Eds), *Big Data for 21st Century Economic Statistics* (pp. 373–402), University of Chicago Press, Chicago, USA.

Chernis T and Sekkel R (2017). A dynamic factor model for nowcasting Canadian GDP growth, *Empirical Economics*, **53**, 217–234.

Delle Monache D and Petrella I (2019). Efficient matrix approach for classical inference in state space models, *Economics Letters*, **181**, 22–27.

Fan J, Feng Y, and Song R (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models, *Journal of the American Statistical Association*, **106**, 544–557.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan J and Lv J (2008). Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B*, **70**, 849–911.

Fan J, Samworth R, and Wu Y (2009). Ultrahigh dimensional feature selection: Beyond the linear model, *The Journal of Machine Learning Research*, **10**, 2013–2038.

Galbraith JW and Tkacz G (2018). Nowcasting with payments system data, *International Journal of Forecasting*, **34**, 366–376.

Ghysels E, Santa-Clara P, and Valkanov R (2004). The MIDAS touch: Mixed data sampling regression models,

UCLA: Finance, **2004**, 1–32.

Giannone D, Reichlin L, and Small D (2008). Nowcasting: The real-time informational content of macroeconomic data, *Journal of Monetary Economics*, **55**, 665–676.

Harvey AC (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, UK.

Hastie T, Tibshirani R, and Wainwright M (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC press, Boca Raton, FL.

He X, Wang L, and Hong HG (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, *The Annals of Statistics*, **41**, 342–369.

Kim C and Kim H (2016). Study on Nowcasting of GDP Growth Rates., *BOK National Account Review*, **2016**, 1–23.

Kim HH and Swanson NR (2018). Methods for backcasting, nowcasting and forecasting using factor-MIDAS: With an application to Korean GDP, *Journal of Forecasting*, **37**, 281–302.

Koenker R and Hallock KF (2001). Quantile regression, *Journal of Economic Perspectives*, **15**, 143–156.

Lee K and Hwang S (2016). Development of economic indicators using big data: Creation of Naver search economic index and review of its usefulness, *BOK Economic Analysis*, **20**, 1–38.

Lee H, Choi D, Kim Y, and Huh J (2022). Development of a real-time economic forecast system for the current quarter (GDP nowcasting) using digital new technologies, *BOK Issue Note*, **2022**.

Li B (2018). *Sufficient Dimension Reduction: Methods and Applications with R*, Chapman and Hall/CRC, Boca Raton, FL.

Li R, Zhong W, and Zhu L (2012). Feature screening via distance correlation learning, *Journal of the American Statistical Association*, **107**, 1129–1139.

Li Y, Liu Y, and Zhu J (2007). Quantile regression in reproducing kernel Hilbert spaces, *Journal of the American Statistical Association*, **102**, 255–268.

Lin Y and Zhang HH (2006). Component selection and smoothing in multivariate nonparametric regression, *The Annals of Statistics*, **34**, 2272–2297.

Luciani M and Ricci L (2013). Nowcasting norway, *International Journal of Central Banking*, **10**, 215–248.

Marcellino M and Schumacher C (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP, *Oxford Bulletin of Economics and Statistics*, **72**, 518–550.

Matheson TD (2010). An analysis of the informational content of New Zealand data releases: The importance of business opinion surveys, *Economic Modelling*, **27**, 304–314.

Moriwaki D (2019). Nowcasting unemployment rates with smartphone GPS data, In Chiara Renso, Stan Matwin, Konstantinos Tserpes (Eds), *International Workshop on Multiple-Aspect Analysis of Semantic Trajectories* (pp. 21–33), Springer, Würzburg, Germany.

Raju S and Balakrishnan M (2019). Nowcasting economic activity in India using payment systems data, *Journal of Payments Strategy & Systems*, **13**, 72–81.

Richardson A, Mulder T, and Vehbi T (2019). Nowcasting New Zealand GDP using machine learning algorithms, *IFC Bulletins chapters in Bank for International Settlements*, **50**, 1–15, Available from: https://ideas.repec.org/h/bis/bisifc/50-15.html

Seo J (2017). Study on the feasibility of using credit card information for estimating domestic household income and consumption expenditure, *Journal of The Korean Data Analysis Society*, **19**, 403–412.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society:*

Series B (Methodological), **58**, 267–288.

Zhang T (2002). Covering number bounds of certain regularized linear function classes, *Journal of Machine Learning Research*, **2**, 527–550.

Zhu LP, Li L, Li R, and Zhu LX (2011). Model-free feature screening for ultrahigh-dimensional data, *Journal of the American Statistical Association*, **106**, 1464–1475.

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

Received May 19, 2023; Revised August 21, 2023; Accepted August 24, 2023

빅데이터를 이용한 실시간 민간소비 예측

신승준[a], 서범석[1,b]

[a]고려대학교 통계학과; [b]한국은행 경제모형실

요 약

최근 코로나19 등으로 경제 불확실성이 확대됨에 따라 민간 경제주체의 경제상황을 직접적으로 반영하는 민간소비 동향을 신속히 파악할 필요성이 높아지고 있다. 이에 본 연구는 기존 거시경제지표 뿐만 아니라 빅데이터를 종합적으로 활용하여 민간소비를 실시간으로 추정(nowcasting)하는 방법을 제안하였다. 특히 초고차원 빅데이터의 적합을 위해 활용 가능한 다양한 기계학습 방법론을 비교분석하여 민간소비 추정의 정확도를 향상시키고자 하였다. 실증 분석 결과, 빅데이터를 비롯한 가용 공변량의 수가 많은 경우에는 변수를 미리 선별하여 모형적합에 활용하는 것이 민간소비 예측 성능을 향상시킬 수 있음을 확인하였다. 또한 코로나19 이후 빅데이터의 반영이 민간소비 예측 성능을 더욱 크게 향상시킴에 따라 경제 불확실성이 높은 상황일수록 새로운 정보를 적시에 반영할 수 있는 고빈도 빅데이터의 활용가치가 높은 것으로 판단된다.

주요용어: 커널분위수회귀, MIDAS, 변수선별, 경제전망

본 연구는 한국은행의 연구용역지원을 받아 수행되었습니다.

[1]교신저자: (04531) 서울특별시 중구 남대문로 39 (남대문로3가), 한국은행 경제모형실. E-mail: bsseo@bok.or.kr