Journal of the Korea Academia-Industrial cooperation Society
Vol. 20, No. 2 pp. 200-209, 2019

https://doi.org/10.5762/KAIS.2019.20.2.200
ISSN 1975-4701 / eISSN 2288-4688

결측 데이터 보정법에 의한 의사 데이터로 조정된 예측 최적화 방법

김정우
아산생명과학연구원

Predictive Optimization Adjusted With Pseudo Data From A Missing Data Imputation Technique

Jeong-Woo Kim
Asan Institute for Life Sciences

요 약 미래 값을 예측할 때, 학습 오차(training error)를 최소화하여 추정된 모형은 보통 많은 테스트 오차(test error)를 야기할 수 있다. 이것은 추정 모델이 주어진 데이터 집합에만 집중하여 발생하는 모델 복잡성에 따른 과적합(overfitting) 문제이다. 일부 정규화 및 리샘플링 방법은 이 문제를 완화하여 테스트 오차를 줄이기 위해 도입되었지만, 이 방법들 또한 주어진 데이터 집합에서만 국한 되도록 설계되었다. 본 논문에서는 테스트 오차 최소화 문제를 학습 오차 최소화 문제로 변환하여 테스트 오차를 줄이기 위한 새로운 최적화 방법을 제안한다. 이 변환을 수행하기 위해 주어진 데이터 집합에 대해 의사(pseudo) 데이터라고 하는 새로운 데이터를 추가하였다. 그리고 적절한 의사 데이터를 만들기 위해 결측 데이터 보정법의 세 가지 유형을 사용하였다. 예측 모델로서 선형회귀모형, 자기회귀모형, ridge 회귀모형을 사용하고 이 모형들에 의사 데이터 방법을 적용하였다. 또한, 의사 데이터로 조정된 최적화 방법을 활용하여 환경 데이터 및 금융 데이터에 적용한 사례를 제시하였다. 결과적으로 이 논문에서 제시된 방법은 원래의 예측 모형보다 테스트 오차를 감소시키는 것으로 나타났다.

Abstract When forecasting future values, a model estimated after minimizing training errors can yield test errors higher than the training errors. This result is the over-fitting problem caused by an increase in model complexity when the model is focused only on a given dataset. Some regularization and resampling methods have been introduced to reduce test errors by alleviating this problem but have been designed for use with only a given dataset. In this paper, we propose a new optimization approach to reduce test errors by transforming a test error minimization problem into a training error minimization problem. To carry out this transformation, we needed additional data for the given dataset, termed pseudo data. To make proper use of pseudo data, we used three types of missing data imputation techniques. As an optimization tool, we chose the least squares method and combined it with an extra pseudo data instance. Furthermore, we present the numerical results supporting our proposed approach, which resulted in less test errors than the ordinary least squares method.

Keywords : bias and variance, prediction, missing data imputation, overfitting, pseudo data, test error and training error

1. 서론

미래 값을 예측하기 위해 보통 우리는 모델을 설정하고 주어진 데이터 집합을 사용하여 모델의 계수를 추정한 이후에 미래 값을 예측한다. 대부분의 예측 모델은 이 방식에서 벗어나지 않으므로 여기에는 일반적으로 과적

*Corresponding Author : Jeong-Woo Kim(Asan Institute for Life Sciences)
Tel: +82-10-3392-2263 email: kurtjw@amc.seoul.kr
Received October 12, 2018 Revised (1st November 5, 2018, 2nd December 5, 2018, 3rd December 7, 2018)
Accepted February 1, 2019 Published February 28, 2019

합(overfitting) 문제가 개입된다. 통계학 및 기계 학습 분야에서는 과적합 문제를 해결하기 위해 계수의 개수를 제어하는 Akaike 정보 기준 [1], 하위 집합 선택[2] 또는 가장 높은 설명력을 갖는 변수를 찾는 principal component 회귀분석과 같은 방법들이 제시되었다 [3, 4]. 또한 이러한 접근법과는 달리 ridge 회귀분석[5]과 lasso 방법[6]은 모델을 최적화 하는 과정에서 매개변수 숫자에 패널티를 부여하는 방법으로 계수의 숫자를 축소하는 방안을 제시하였다. 특정 모델링과 상관없이 리샘플링 (Resampling), 부트스트랩 (Bootstrap)과 잭나이프 (Jackknife)와 같은 방법들은 동일한 데이터 집합으로부터 조금씩 다른 데이터 집합들을 얻는 방법들로 과적합 문제를 완화하는데 도움이 될 수 있다. 이런 방법들의 연장선에서, 교차 유효성(Cross validation) 방법은 주어진 데이터를 단순히 몇 개의 구간으로 나누어 상기 방법들과 유사한 성능을 보이므로 과적합 문제를 해결하는데 널리 사용되고 있다[7].

상기 언급한 방법들은 모델의 계수를 다양한 방식조정하거나 또는 데이터를 반복 추출하는 방식으로 과적합 문제를 완화할 수 있으나, 여전히 이 방법들은 주어진 데이터만을 사용하고 있다. 그러므로 우리는 주어진 데이터를 확장하여 과적합 문제를 해결하는 접근법을 생각해 볼 수 있다.

주어진 데이터를 확장하여 새로운 데이터를 생성하는 기법들은 비모수 통계학 분야에서 많이 이루어졌다. 이와 관련한 선행 연구들은 커널 밀도 함수를 추정하기 위해 주어진 데이터 집합 외의 추가 데이터를 사용하는 연구들이 많은 부분을 차지하고 있다. Cowling 등[8] 및 Cline 등[9] 은 밀도 함수의 대칭성에 기반하여 의사 (pseudo) 데이터를 생성하여 커널 밀도 함수를 추정하는데 사용하였다. 특히, 최근에는 밀도 함수의 대칭성을 활용하여 금융 데이터 또는 기후 데이터와 같은 극단치를 포함한 데이터에서의 커널 밀도 함수를 추정하는데도 활용되었다 [10. 11]

의사 데이터를 밀도 함수 추정에 적용하는 사례들 외에도, Gerlovina[12]는 의사 데이터를 부트스트래핑 방법과 결합하여 실증 (empirical) 밀도함수 추정 시의 smoothing 정도를 높였으며, Breiman[13]은 주어진 데이터를 선형결합하여 새로운 의사 데이터를 생성하고 이를 사용하여 회귀분석에서의 테스트 오차를 줄이는 방법을 제안하였다.

결측 데이터 보정법(Imputation)은 의사 데이터를 활용하는 대표적인 방법일 것이다. 보정법의 목적은 누락된 값을 적절한 값으로 대체하는 것이므로 실제 응용 연구에서 널리 사용되고 있다. Purwar and Singh[14] 는 K-means 군집화 방법을 사용하여 결측 데이터를 대체하고, 이를 통하여 당뇨병, 간염 및 유방암 예측 정확도를 개선하였다. Liu 등[15] 은 교통 량 데이터에서 결측된 데이터를 대체하기 위하여 k-최근접 이웃 알고리즘을 사용하였으며, Wu 등[16] 는 서포트 벡터 머신 방법을 사용하여 시계열 데이터에서의 결측 자료 문제를 해결하고자 하였다.

주어진 데이터를 활용하여 새로운 의사 데이터를 생성해내고 이 데이터를 활용하여 주어진 데이터에 더욱 적합한 밀도함수를 구하는 것은 통계학적 관점에서 무엇보다 중요한 문제이며 주어진 데이터에 대한 평균, 분산과 같은 기본 통계량의 정확성을 높이는데 기여할 수가 있다. 반면에, 상기 선행 연구들과 달리 의사 데이터를 활용하여 미래를 예측하는 연구는 많지가 않다. Breiman[13]이 주어진 데이터의 선형 결합을 통한 의사 데이터를 생성하고 이를 예측 문제에 적용하였으나, 주어진 데이터의 양이 많은 경우 어떠한 선형 결합방법을 사용해야 하는지에 대한 실용적 측면에서의 문제점을 내포하고 있다. 본 연구에서는 의사 데이터를 미래 값을 예측하는데 활용하였으며, 활용 방법으로써 결측 데이터 보정법의 개념을 차용하여 예측 문제에서의 의사 데이터의 실용성을 높이고자 한다.

결측 데이터 보정법은 누락된 데이터 값을 되도록 비슷한 의사 데이터로 대체하여 결측치가 없는 완벽한 데이터 집합을 얻는 데에 주로 중점을 둔다. 즉, 결측 데이터 보정법에서 의사 데이터를 사용하는 것은 완전한 데이터를 얻기 위한 부차적인 방법인 것이다. 그러나 본 연구에서는 결측 데이터 보정법의 개념을 활용하여 의사 데이터를 생성하고 이것을 회귀모형 등에 적용하여 모델의 예측정확도를 높이고자 하므로 의사 데이터의 역할은 중요도가 높다고 볼 수 있다.

또한, 결측 데이터 보정법을 사용하여 누락된 데이터를 대체하면 그 데이터는 최소자승법 등의 최적화 과정에서 변하지 않는 반면, 본 연구에서는 누락된 데이터가 결측 데이터 보정법을 통한 의사 데이터로 대체된 후, 최적화 과정에서 다시 한번 의사 데이터는 예측값으로 변환된다. 그러므로 이 방법은 기존의 일반적인 결측 데이

터 보정법과는 다르며 모델의 예측정확도를 높이기 위한 새로운 방법이라고 할 수 있다. Fig. 1은 제안된 방법의 개요를 나타내고 있다. T개의 데이터가 주어질 시, 일반적인 결측 데이터 보정법은 주어진 데이터의 중간에 결측된 데이터를 추정하는 것이라면, 이 연구에서는 $T+1$번째 의사 데이터가 생성되고 이 데이터는 최적화 과정을 통하여 잠정적인 예측값으로 변환된 후, 최종적으로 $T+1$번째 의사 데이터가 된다. 이 $T+1$번째의 의사 데이터가 예측 정확도를 높일 수 있는 이론적 근거는 2절에서, 실증 데이터를 활용한 검증은 3절에서 다루어진다.

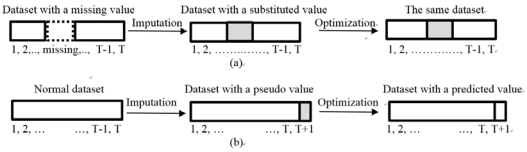

Fig. 1. (a) Missing data imputation and (b) the proposed approach with pseudo data.

2. 본론

주어진 데이터에 특정 모델을 최적화하는 문제는 두 가지 경우로 생각할 수 있다. 첫 번째는 적합(fitting)의 문제이고 두 번째는 예측의 문제이다. 전자와 목적은 학습 오차(또는 잔차)를 최소화하는 것이고 후자는 테스트 오차(또는 예측 오차)를 최소화하는 것으로 최적화의 성격이 다르다. 테스트 오차는 주로 학습 오차보다 크고 모델 복잡성이 높아짐에 따라 증가하는 경향이 있다[17].

T 개의 데이터가 주어지고 $T+1$ 번째의 값을 예측한다고 할 시에 우리는 확률분포함수가 주어지지 않으면 테스트 오차를 최소화하기보다는 주어진 데이터로부터의 학습 오차를 최소화하고 여기서 얻어진 추정모델로 $T+1$의 값을 예측한다. 이러한 경우 우리는 학습 오차를 최소화하는 모델을 추정하게 되며, 결국에는 우리는 학습 오차보다 더 높은 테스트 오차를 가지게 될 수밖에 없다. 그러므로 학습 오차 최소화 문제를 테스트 오차 최소화 문제로 최대한 가깝게 변환하면, 우리는 보다 낮은 테스트 오차를 얻을 수 있을 것이다. 이러한 변환을 가능하게 하는 것은 가상의 $T+1$ 번째의 값, 즉, 의사 데이터를 추가하므로써 가능할 것이다.

2.1 의사 데이터

테스트 오차를 최소화 문제를 학습 오차 최소화 문제로 변환하기 위해서는 미래의 $T+1$ 값이 주어지지 않은 상태에서 기존의 데이터를 활용하여 의사 데이터로 $T+1$번째 값을 대체하여야 한다.

비모수 통계학에서 커널 밀도 함수를 추정할 시에 의사 데이터를 사용하는 몇 가지의 방법들이 도입되어 있다. Cowling 등[8] 은 주어진 밀도 함수의 정의역을 벗어난 의사 데이터를 사용했다. 예를 들어 그들은 주어진 데이터의 정의역 [0, 1]의 왼쪽에 위치한 의사 데이터 $X_{(-1)}$를 밀도 함수의 대칭성을 기반으로 $X_{(-1)} = 3X_{(1)} + X_{(2)}$ 로 추정하였다(여기서 $X_{(i)}$ 는 주어진 데이터의 순서통계량). 따라서 의사 데이터는 주어진 정의역의 반대편에 있는 값이 될 수는 있지만 정의역의 경계 지점에 있는 값으로는 될 수 없다. Cline 등[9] 는 reflection 데이터 방법을 제안했다. 밀도 함수의 미분값이 정의역의 경계에서 0일 때, 이 방법은 주어진 데이터가 $\{X_1, X_2, \ldots, X_n\}$일 때 정의역 경계의 반대편 의사 데이터는 $\{-X_1, -X_2, \ldots, -X_n\}$으로 설정하는 것이다. 즉, 이 방법 또한 밀도 함수의 대칭성에 의존하고 있는 것이다.

상기 언급한 밀도 함수 추정 방식과 달리 $T+1$의 미래 값을 예측하는 모델의 문제에서는 밀도 함수의 대칭성을 활용하여 의사 데이터를 만들 수 없다. 그러므로 우리는 대칭성의 조건이 주어지지 않는 상태에서도 의사 데이터를 생성하기 위해 주어진 데이터의 특징을 잘 활용하는 방법을 활용해야 한다.

2.2 결측 데이터 보정법

테스트 오차를 줄이는데 필요한 의사 데이터를 만드는 방법은 다양할 것이다. 의사 데이터가 실제 미래 값에 가까울수록 테스트 오차는 그만큼 낮춰질 것으로 예상할 수 있다. 우리는 다양한 의사 데이터를 결측 데이터 보정법을 통해 생성할 수 있다. 결측 데이터 보정법을 크게 구분하면 결측치가 체계적으로 결측되었는지, 또는 임의적으로 결측되었는지로 구분할 수 있다[18]. 통계학적으로는 후자의 경우가 결측치 보정을 위해서 더욱 많이 연구되고 있는 분야이다. 임의적으로 결측치가 발생하는 경우에는 단순임의보정법, 즉 주어진 데이터내에서 확률적으로 임의의 데이터 한 개를 추출하여 결측치를 보완하는 방법이 손쉽게 사용될 수 있다. 하지만 시계열 자료와 같이 자료들간의 상관관계가 존재하는 경우에는 단순

임의보정법의 적용은 무리가 있을 것이다. 단순임의보정법과 달리 Hot-deck 보정법은 주어진 데이터가 입력변수와 출력변수로 이루어지고 한 개의 출력변수가 결측된 경우, 입력변수가 비슷한 데이터 쌍을 찾아내어 그 데이터 쌍의 출력변수를 결측치 대신 대체하는 방법이다[18]. 이 방법은 변수들 사이의 관계를 고려하는 장점이 있지만, 오차항이 비교적 큰 경우, 즉 입력변수와 출력변수 간의 관계가 고유하지 않은 경우에는 대체된 결측치가 실제 결측치와 전혀 다른 값으로 될 수 있는 위험을 내포하고 있다. 본 연구에서는 여러 가지 결측치 보정법 중에서 일반적인 통계학적 개념으로 접근가능한 세가지 결측치 보정법을 활용하여 의사 데이터를 생성하였다.

2.2.1 Last Observation Carried Forward

첫번째로 고려한 보정법은 결측값의 직전에 마지막 관찰된 데이터 값으로 결측값을 대체하는 Last Observation Carried Forward 방법이다. 이 보정법은 활용이 편리하면서도 데이터의 연속성을 고려한다는 장점을 가지고 있다. 그러나 이 방법은 결측값이 주어진 데이터에서 크게 벗어나지 않는다는 다소 강한 가정을 하고 있다. 하지만 금융 시계열 데이터를 다루는 연구들에서는 Last Observation Carried Forward 방법과 유사한 Martingale 프로세스 가정을 상정하고 시계열 분석을 하는 경우가 많으므로 이 연구에서도 의사 데이터를 생성하는 방법으로 활용하기로 한다.

2.2.2 평균값 보정법

평균값 보정법은 누락된 데이터 값을 주어진 데이터 집합의 평균값으로 대체하는 방법이다. 특히 평균값은 주어진 데이터가 대칭 분포일 경우에 합리적인 결측값 보정치라고 볼 수 있다. 또한 데이터의 크기가 클수록 설득력이 높은 방법이기도 하다. 하지만 누락된 데이터가 무작위로 누락되지 않으면 평균값 보정법은 모델 추정의 편의(bias)를 높일 수 있으며, 이는 샘플 크기가 적을수록 편의가 높아질 수도 있다.

2.2.3 회귀 보정법

회귀 보정법은 주어진 데이터에 선형 회귀 모델을 추정하여 이 추정된 회귀식으로 미래 값을 예측한 후 얻어진 예측값을 결측 데이터 값으로 대체한다. 이 방법은 데이터 집합이 지닌 선형 추세를 유지하는 장점을 가지므로 평균값 보정법과 마찬가지로 샘플의 크기가 모델 추정에 영향을 줄 수 있다. 또한, 회귀분석을 통한 추정치는 오차항이 없는 값이므로 주어진 데이터의 고유한 변동성을 간과하는 경향이 있다.

상기에서 설명한 방법들은 주어진 데이터 세트의 특성에 따라 다르게 사용될 수 있을 것이며, 각 방법의 장단점은 아래 표와 같이 정리할 수 있다.

Table 1. Comparison of imputation methods

보정법	장점	단점
Last Observation Carried Forward	∘ 사용의 편의성 ∘ bias가 적음	∘ 데이터의 변동성에 취약 ∘ 장기 예측은 불가
평균값 보정법	∘ 대칭분포에 유리 ∘ 대규모 데이터에 적합	∘ 소규모 데이터에는 bias 큼 ∘ 변수간 상관관계를 고려하지 못함
회귀 보정법	∘ 선형 추세의 데이터에 적합 ∘ 다변수 고려 가능	∘ 데이터의 오차항 무시 ∘ 소규모 데이터에는 bias 큼

2.3 의사 데이터를 활용한 최소자승법

상기 결측 데이터 보정법들로부터 우리는 세 가지종류의 의사 데이터를 생성할 수 있다.

우선, 주어진 데이터 집합 $S_T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)\}$를 생각해보자. 여기서 x_i 는 입력변수, y_i 는 출력변수이다.

Last Observation Carried Forward 방법에 따라서 $T+1$번째 의사 데이터(y^p_{T+1})는 아래와 같이 생성할 수 있다.

$$(x_{T+1}, y^p_{T+1}) = (x_{T+1}, y_T). \quad (1)$$

평균값 보정법에 의한 의사 데이터는

$$(x_{T+1}, y^p_{T+1}) = (x_{T+1}, \bar{y}_T),$$
$$\bar{y}_T\text{는 } y_i\text{의 평균값.} \quad (2)$$

회귀 보정법에 의한 의사 데이터는

$$(x_{T+1}, y^p_{T+1}) = (x_{T+1}, \hat{y}_T) \quad (3)$$

여기서 $\hat{y}_T$는 주어진 데이터로 추청한 선형회귀식에 x_{T+1}

을 대입하여 얻은 값이다.

본 연구에서는 상기 세가지 보정법에 의해서 얻어진 세가지 의사 데이터를 $T+1$번째에 각각 대입하여 새로운 데이터 집합을 만든다. 그러면 의사 데이터 y^p_{T+1}를 포함한 새로운 데이터 집합은 $S_{T+1} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T), (x_{T+1}, y^p_{T+1})\}$ 가 된다. 여기서 새로운 데이터 집합 S_{T+1}은 상기 세가지 의사 데이터를 사용하여 얻어졌으므로 우리는 세가지 종류의 S_{T+1}을 얻게되며, 이 세가지 보정법에 의한 세가지 종류의 S_{T+1}은 모두 후절의 실증 검증에서 고려될 것이다.

다음으로, 예측 모델 함수를 $f(x)$로 두고, 집합 S_T를 활용하여 얻은 실제 미래값 y_{T+1}에 대한 추정치를 $f(x_{T+1};S_T)$, S_{T+1}를 활용하여 얻은 추정치를 $f(x_{T+1};S_{T+1})$로 표기하자. 그러면 Ruppert and Wand[19]의 평균제곱오차(mean square error, MSE)에 대한 정리를 응용하면 우리는 아래와 같은 부등식을 얻을 수 있다.

$$\mathrm{E}[y_{T+1} - f(x_{T+1};S_{T+1})]^2 \le \mathrm{E}[y_{T+1} - f(x_{T+1};S_T)]^2 \quad (4)$$

즉, 의사 데이터를 포함한 집합 S_{T+1}를 사용할 시에 주어진 집합 S_T를 사용하는 경우보다 더 적은 테스트 오차를 얻을 수 있는 것이다.

3. 실증 검증

본 연구에서는 의사 데이터 방법의 실증 검증을 위하여, 우선 의사 데이터 방법을 세가지 예측모델들에 적용한 실험을 통하여 예측 정확도의 향상 여부를 이론적인 측면에서 검증해보았다. 그 다음으로써 이 이론적인 검증 결과를 토대로, 실제 데이터를 활용하여 의사 데이터 방법의 실질적인 예측 정확도 개선여부를 검증해보았다.

본 연구에서 예측 정확도 측정을 위한 테스트 오차는 평균제곱오차

$$\mathrm{MSE} = \mathrm{E}[y_{T+1} - f(x_{T+1})]^2 \quad (5)$$

를 사용하였다. 또한 결과표들에서 Original은 의사 데이터를 사용하기 이전의 세가지 예측 모델로부터 얻어진 테스트 오차를 뜻하며, P_ONE은 Last Observation Carried Forward 방법을 사용한 경우, P_MEAN은 평균값 보정법에 의한 의사데이터를 사용한 경우, P_LNR은 회귀 보정법을 사용한 경우를 지칭한다. ACT는 실제 미래값을 예측 모형에 대입한 경우의 잔차로 테스트 오차의 하한값이라고 볼 수 있다.

3.1 예측 모델 적용

이 절에서는 세가지 예측 모델을 사용하여, 의사 데이터 방법이 시뮬레이션을 통한 다양한 데이터 환경에서 적용가능한지를 검토해보았다.

이를 위하여, 우선 데이터의 복잡도에 따라 선형함수 모의 데이터(Model 1), 다항함수 모의 데이터(Model 2), 삼각함수 및 로그함수 모의 데이터(Model 3)를 생성하였다.

$$\begin{aligned}
&\text{Model 1} \quad y_t = 3 - 0.1x_t + e_t \\
&\text{Model 2} \quad y_t = 3 - x_t^2 + x_t^3 + e_t \\
&\text{Model 3} \quad y_t = 3 + \sin(x_t/2) + 1.5\log(|x_t|) + e_t
\end{aligned}$$

여기서 x_t는 $\{1,2,\ldots,T\}$이며 e_t는 표준정규분포 $N(0, 1)$을 따른다. 모의 데이터는 25, 50 및 100개 크기의 세가지 샘플을 사용하였으며, 각 크기의 샘플에 대해 0.1, 0.5 및 1 크기의 세가지 분산(σ^2)을 적용하였다. 몬테카를로 반복 횟수는 각 샘플마다 100회씩 시행하였으며 각 시행마다 마지막 데이터를 예측하여 테스트 오차를 계산하였다.

예측 모델로써는 선형회귀모형, 자귀회귀(Autoregressive, AR)모형, ridge 회귀모형 등 총 세가지의 회귀 모형들이 사용되었으며, 선형회귀모형과 ridge 회귀모형에서는 부트스트래핑(반복횟수 : 100, 500, 1000번) 방법을 적용하여 예측 정확도 결과에 대한 강건성을 높였다.

선형회귀모형은 모형의 단순한 가정으로 변수들의 상관관계 등을 무시하는 단점이 있으나, 많은 연구들에서 레퍼런스 모형으로 쓰이는 기본 모형이므로 예측 모델로 사용하였다. 자기회귀모형은 시계열 예측 모델링에서 필수적인 모형이므로 본 연구에서 채택하였으며, 본 연구에서는 위 세가지 모의 데이터와 별도로 시계열 모의 데이터를 생성하여 의사 데이터 방법을 검증해보았다. Ridge 회귀모형은 편향 추정치를 제공하는 단점이 있으나 추정치의 분산이 많이 줄어드는 장점이 있어서 예측의 정확도를 높일 수 있는 모형으로 본 연구에 포함하였

다. 본 연구에서 쓰인 세가지 예측 모델을 정리하면 아래 표와 같다.

Table 2. Comparison of prediction models

예측모델	특성
선형회귀모형	+ 통계분석의 기본 모형으로 사용이 편리 + 변수들간의 다중공선성 및 변수내의 상관관계 등은 고려하지 못함
자기회귀모형	+ 시계열 예측 모델의 기본 모형 + 주로 단일변수 모형에 쓰임
Ridge 회귀모형	+ 편향 추정치를 제공하는 단점이 있음 + 추정치의 분산이 적음

3.1.1 선형회귀모형

여기서는 각 함수 모델을 선형회귀모형 $y_t = a + bx_t + e_t$로 추정하고, 주어진 T개의 데이터로 미래의 값(T + 1번째)을 예측한다. 결과를 나타내는 표에서는 선형회귀 모형의 테스트 오차를 세가지 의사 데이터를 사용한 경우와 비교하였다.

Table 3은 모델 1에서의 테스트 오차 비교결과를 보여주고 있다. 모델 1은 선형이기 때문에 전반적으로 의사 데이터를 사용한 경우가 Original보다 상당히 작은 테스트 오차는 보여주지 않는다. 특히, 분산이 적을수록 테스트 오차들의 차이는 크지 않다. 반대로, 분산이 커지면 의사 데이터를 사용하는 것이 테스트 오차가 다소 줄어드는 모습을 보인다.

Table 3. Test Error Comparison: Model 1

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	0.0114	0.0115	0.0117	0.0114	0.0099
	0.5	0.2122	0.2069	0.2082	0.2122	0.1819
	1	1.3849	1.3645	1.3624	1.3849	1.1800
50	0.1	0.0125	0.0123	0.0127	0.0125	0.0116
	0.5	0.3130	0.3116	0.3167	0.3130	0.2891
	1	1.2710	1.2668	1.2571	1.2710	1.1638
100	0.1	0.0120	0.0120	0.0119	0.0120	0.0115
	0.5	0.2556	0.2541	0.2540	0.2556	0.2427
	1	0.7962	0.7937	0.7997	0.7962	0.7629

Table 4는 모델 2에 대한 테스트 오차 비교 결과이다. 모델 1과 달리 P_MEAN은 모든 경우에서 Original보다 적은 테스트 오차를 제공한다. P_ONE은 T = 50 및 σ^2 = 1의 경우를 제외하고는 양호한 성능을 나타내었다.

Table 4. Test Error Comparison: Model 2

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	2.6013	2.4323	2.3934	2.6013	1.7655
	0.5	1.2487	1.1961	1.2197	1.2487	1.0046
	1	3.2067	3.0798	2.8847	3.2067	2.227
50	0.1	1.0225	1.0193	1.0264	1.0225	0.9012
	0.5	2.6383	2.6185	2.6043	2.6383	2.0825
	1	2.8292	2.7634	2.7929	2.8292	2.4905
100	0.1	1.9331	1.9141	1.9313	1.9331	1.7482
	0.5	2.5431	2.5186	2.533	2.5431	2.2892
	1	2.1493	2.1351	2.124	2.1493	2.0379

Table 5 또한 상기 결과와 유사한 테스트 오차 결과를 보여준다. P_MEAN은 모든 경우에 Original보다 적은 테스트 오차를 보여주었고 P_ONE도 T = 100 및 σ^2 = 1의 경우를 제외하고 낮은 테스트 오차를 보였다.

Table 5. Test Error Comparison: Model 3

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	2.6452	2.5176	2.3906	2.6452	1.9954
	0.5	4.3301	4.2318	4.284	4.3301	3.7149
	1	4.124	3.7527	3.8119	4.124	3.1462
50	0.1	2.5507	2.5203	2.5452	2.5507	2.2941
	0.5	3.4352	3.4064	3.3757	3.4352	3.0713
	1	3.2236	3.2077	3.1906	3.2236	2.9861
100	0.1	2.4982	2.474	2.454	2.4982	2.3803
	0.5	2.9674	2.9303	2.9157	2.9674	2.8159
	1	3.4305	3.4189	3.4666	3.4305	3.2837

3.1.2 자기회귀모형

자기회귀모형은 1차, 2차, 3차의 차수를 사용하여 아래와 같은 함수로 모의 데이터를 생성하였다. 나머지 조건들은 선형회귀 모형의 경우와 동일하게 설정하였다.

AR(1) $y_i = 0.7y_{i-1} + e_t$,

AR(2) $y_t = 0.7y_{i-1} + 0.25y_{i-2} + e_t$,

AR(3) $y_t = 0.7y_{i-1} + 0.25y_{i-2} - 0.5y_{i-3} + e_t$,

Table 6 - 8는 자기회귀모형에 대한 테스트 오차 비교 결과를 보여주고 있다. 이전과 마찬가지로 P_MEAN은 Original 보다 낮은 테스트 오차를 많은 경우에서 보여주었다. 하지만 그 차이가 크지는 않았는데, 이것은 계수 0.7이 1과 크기가 비슷함에 따라 데이터가 비정상성

(non-stationary)을 지니게 하기 때문이다. 이러한 현상은 AR(1), AR(2), AR(3) 모두에서 관찰되었다.

Table 6. Test Error Comparison: AR(1)

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	0.0888	0.0865	0.0903	0.0888	0.0799
	0.5	0.6231	0.6248	0.6108	0.6231	0.562
	1	1.0509	1.0328	1.0504	1.0509	0.9467
50	0.1	0.093	0.0939	0.0892	0.093	0.0878
	0.5	0.4406	0.4352	0.4418	0.4406	0.419
	1	0.8965	0.9078	0.8737	0.8965	0.8511
100	0.1	0.1202	0.12	0.1205	0.1202	0.118
	0.5	0.4416	0.4406	0.4432	0.4416	0.4329
	1	0.8875	0.8848	0.8902	0.8875	0.8667

Table 7. Test Error Comparison: AR(2)

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	0.0888	0.0859	0.0932	0.0904	0.0704
	0.5	0.6231	0.4232	0.4473	0.4276	0.3476
	1	1.0509	0.9953	1.0678	1.0401	0.8092
50	0.1	0.094	0.0926	0.1022	0.094	0.0831
	0.5	0.4604	0.4578	0.4735	0.4604	0.4179
	1	1.0135	0.9924	1.0152	1.0135	0.9125
100	0.1	0.1029	0.1024	0.1029	0.1029	0.0978
	0.5	0.6349	0.6331	0.6346	0.6349	0.6071
	1	0.9669	0.9618	1.0333	0.9669	0.9202

Table 8. Test Error Comparison: AR(3)

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	0.1384	0.1293	0.1387	0.1384	0.0939
	0.5	0.5394	0.5082	0.5495	0.5394	0.3551
	1	0.7368	0.6722	0.7417	0.7368	0.5308
50	0.1	0.0999	0.0985	0.099	0.0999	0.0858
	0.5	0.5508	0.5442	0.543	0.5508	0.4764
	1	0.8266	0.8164	0.8793	0.8266	0.7219
100	0.1	0.0894	0.0887	0.0911	0.0894	0.0836
	0.5	0.4814	0.4794	0.4691	0.4814	0.4557
	1	1.1431	1.1398	1.1572	1.1431	1.0704

3.1.3 Ridge 회귀모형

상기 두 모형은 비편향(unbiased) 추정치를 제공한다. 하지만, 이 절에서는 편향 추정치인 ridge 회귀분석을 사용하였다. Ridge 회귀분석은 모형 복잡도를 조정하는 매개변수를 선형회귀모형에 추가로 사용하는 분석방법이다. 이에 따라 ridge 회귀분석은 인위적으로 추가된 매개변수로 인해 편향 추정치를 제공하나 추정치의 분산이 줄어 오차를 줄일 수 있는 장점이 있으며, 동시에 다중공선성(multicollinearity) 문제도 완화하는 장점이 있어 단순선형회귀모형의 대안으로 많이 사용된다[20].

모의 실험 데이터는 3.1.1절에서 사용된 세가지 모델을 사용하여 생성되었으며, 데이터 생성과 관련된 다른 조건도 3.1.1절과 동일하게 설정하였다.

Table 9 - 11는 ridge 회귀분석을 사용하여 테스트 오차를 비교한 결과이다. 상기 두 모형과 유사하게 P_MEAN은 가장 작은 테스트 오차를 보여주었다. 특히, 선형회귀모형에서는 모델 1의 경우 P_MEAN의 예측 정확도가 그리 높지 않았으나, ridge 회귀분석의 경우에는 낮은 테스트 오차를 보였다. 이것은 추정치의 분산을 줄여서 오차를 낮추는 ridge 회귀분석의 효과가 반영된 것으로 보인다.

Table 9. Test Error Comparison: Model 1

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	1.0324	1.023	1.0148	1.0447	0.9222
	0.5	1.1488	1.1296	1.146	1.15	1.0567
	1	1.0818	1.0757	1.0756	1.0832	1.012
50	0.1	0.9903	0.979	0.9902	0.9904	0.9553
	0.5	1.209	1.2064	1.2059	1.21	1.1638
	1	0.9942	0.9933	0.9943	0.9936	0.953
100	0.1	0.8804	0.8779	0.8799	0.8805	0.865
	0.5	0.7806	0.7771	0.7804	0.7809	0.7638
	1	1.153	1.1493	1.1529	1.1526	1.1288

Table 10. Test Error Comparison: Model 2

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	2.9391	2.8698	2.8005	3.0474	2.2881
	0.5	5.0434	4.9387	4.853	5.2174	3.1644
	1	2.9808	2.8679	2.8916	3.0293	2.4343
50	0.1	3.5036	3.4955	3.4618	3.5261	3.1319
	0.5	3.8772	3.8542	3.8681	3.8829	3.0934
	1	2.6823	2.5957	2.6541	2.6971	2.4395
100	0.1	2.9615	2.9555	2.9602	2.9612	2.804
	0.5	3.1466	3.1398	3.1251	3.1621	2.9326
	1	4.3206	4.3013	4.3018	4.3363	3.8767

Table 11. Test Error Comparison: Model 3

T	σ^2	Original	P_MEAN	P_ONE	P_LNR	ACT
25	0.1	4.1812	4.086	4.137	4.2137	3.9757
	0.5	3.279	3.0915	3.2287	3.3048	2.8957
	1	2.4907	2.4413	2.4971	2.4861	2.2873
50	0.1	4.8768	4.8864	4.8812	4.8741	4.7205
	0.5	4.2753	4.2452	4.2744	4.2791	4.1133
	1	4.5651	4.5229	4.561	4.5695	4.4479
100	0.1	4.5775	4.5861	4.5815	4.5753	4.4552
	0.5	3.1692	3.1572	3.1724	3.1672	3.0807
	1	4.0401	4.0317	4.0415	4.0385	3.9711

3.2 실제 데이터 검증

상기 분석과 같이 의사 데이터 방법이 다양한 데이터 환경에서도 적용가능하다는 것이 이론적인 측면에서 확인되었으므로, 이 절에서는 두 가지 종류의 실제 데이터를 사용하여 의사 데이터 방법의 예측 정확도 향상 여부를 실증적으로 검토하였다.

첫 번째로 선박의 중유(Heavy oil) 소비로 인한 온실가스(CO_2) 배출량 예측 문제에 의사 데이터 방법을 적용하였으며, 두 번째로는 환율(EUR/USD)에 따른 다우존스 산업평균지수(DJIA) 예측 문제를 의사 데이터를 통하여 다루어보았다.

3.2.1 온실가스 배출량 예측

선박의 중유소비량 및 온실가스 배출량의 월별 데이터는 한국석유공사와 한국기상청에서 각각 얻어졌다. 중유는 해운업계의 주요 연료이며, 이 연구에서는 선박의 중유 소비로 인한 온실가스 배출량을 예측하였다. 데이터 수집 기간은 2011년 11월에서 2015 년 12월(총 50개월)까지이며, 마지막 20개월의 온실가스 배출량을 직전까지의 데이터들을 사용하여 예측하였다.

Table. 12는 세가지 의사 데이터를 사용하여 중유 소비로 인한 온실 가스를 예측할 때의 테스트 오차를 각 모형별로 나타내고 있다.

선형회귀분석의 경우 P_MEAN이 의사 데이터를 사용하지 않은 선형회귀분석(Original)보다 테스트 오차가 적다. AR(1) 모형에서는 P_ONE이 P_MEAN보다 낮은 테스트 오차를 보였다. Ridge 회귀 분석을 사용하면 P_ONE의 테스트 오차가 P_MEAN의 경우보다 낮았다. 또한 부트스트래핑 반복 횟수가 증가함에 따라 테스트 오차도 작아지는 경향을 보였다. 모의 데이터의 경우와 마찬가지로 P_MEAN은 Original보다 낮은 테스트 오차를 보여주었으며 P_ONE의 경우도 특정 경우에 따라 낮은 테스트 오차를 나타내었다.

Table 12. Results of the Test Error Comparison: CO_2(ppm)

N	Original	P_MEAN	P_ONE	P_LNR	ACT
Linear regression					
100	22.9922	22.9787	23.1552	22.9922	21.3812
500	24.2576	24.2252	24.2585	24.2576	22.4942
1000	23.6579	23.6067	23.6928	23.6579	21.8721
AR(1) model					
50	8.2147	8.2050	7.9799	8.2147	7.7883
Ridge regression					
100	2.3641	2.3408	2.2791	2.3549	2.3499
500	2.2976	2.2404	2.1794	2.3322	2.2847
1000	2.3284	2.1902	2.1899	2.3881	2.2841

Note. N is the number of bootstrap iterations in linear regression and ridge regression, and the sample size in AR(1) model.

3.2.2 다우존스 산업평균지수 예측

두 번째 실제 데이터 검증으로서 환율(Euro to US Dollar, EUR / USD)로 미국의 다우존스 산업평균지수(Dow Jones Industrial Average, DJIA)를 예측해보았다. 2016년 10월 19일부터 12월 30일까지 일별 데이터를 사용하였으며, 마지막 20일동안의 산업평균지수가 그 전의 데이터들로부터 예측되었다. Table 13은 테스트 오차 비교 결과이며, 대부분의 경우에 P_MEAN이 가장 작은 테스트 오차를 보여주고 있다. 온실가스 예측과는 달리, Ridge 회귀분석의 경우에는 P_ONE과 P_LNR도 비교적 낮은 테스트 오차를 보여주었다. Ridge 회귀분석은 추정치의 편의를 크게 하는 대신, 분산을 적게 함으로써 전체 오차를 줄이는 방식이므로 변동성이 심한 금융 데이터의 경우에 높은 예측 정확도를 보일수도 있는 것으로 해석할 수 있을 것이다[21].

Table 13. Results of the Test Error Comparison: DJIA

N	OLS	P_MEAN	P_ONE	P_LNR	ACT
		Linear	regression		
100	53823.6	52944.2	57764.6	53823.6	49453.7
500	61479.9	60187.8	60892.0	61479.9	57235.9
1000	63846.0	63085.9	64082.9	63846.0	59526.1
AR(1) model					
50	8150.9	8129.7	8185.1	8150.9	7691.9
Ridge regression					
100	958677.7	941970.2	942547.8	955376.5	957379.0
500	963750.0	947896.4	948206.0	960299.1	962687.5
1000	939439.8	923427.6	923970.7	935821.2	937965.0

Note. N is the number of bootstrap iterations in linear regression and ridge regression, and the sample size in AR(1) model.

4. 결론

미래 값을 예측하는 방법은 주로 특정 모델을 설정하고 모델의 계수를 추정하기 위해 주어진 데이터 집합에 모델을 최적화시키는 과정으로 이루어진다. 하지만 주어진 데이터 집합만을 사용하다보면 과적합 문제가 발생하며 이는 테스트 오차를 증가시킬 수 있다. 본 연구에서는 주어진 데이터에 의사 데이터를 추가하여 미래 값을 예측할 시의 테스트 오차를 줄이는 방법을 제안하였다. 또한 적절한 의사 데이터를 만들기 위해 세가지의 결측치 보정법을 활용하였다.

제안된 방법의 검증을 위해 우선 시뮬레이션을 통한 다양한 데이터 환경 하에서 의사 데이터 방법이 적용가능한지 검토해보았으며, 이를 토대로 이 방법을 실제 데이터에 적용하여 실제적인 예측 정확도 개선여부를 확인해보았다.

세가지 의사 데이터 방법 중 평균값 보정법에 기반한 P_MEAN은 원래의 예측모형 및 다른 의사 데이터 방법보다 테스트 오차가 적은 것으로 대부분의 사례에 발견되었다. 실제 데이터를 사용한 경우에도 의사 데이터를 활용한 경우의 테스트 오차가 원래의 예측모형보다 적은 사례들이 많이 발견되었고, 모의 데이터의 경우와 마찬가지로 P_MEAN이 많은 사례에서 가장 작은 테스트 오차를 보였다. 이것은 데이터가 변동이 심하지 않을 때는 중심극한정리가 적용될 수 있다는 사실과 관련이 있다고 해석할 수 있다.

본 연구에서는 비모수 통계학에서 밀도함수의 대칭성을 기반으로 추정된 의사 데이터를 주로 밀도함수 추정에만 사용하던 것과는 다르게, 의사 데이터를 결측치 보정법을 활용하여 생성하고 이를 통해 예측 모델의 예측 정확도를 높였다는 점에서 연구의 의의가 있다고 볼 수 있다. 하지만 이 연구에서 제안된 의사 데이터를 활용하는 방법은 주어진 데이터의 크기가 커질수록 예측 정확도의 개선정도가 낮아지는 경향이 있다. 이것은 오차를 구성하는 분산이 데이터의 개수에 반비례$(1/T)$한다는 사실에 근거하고 있다. 따라서 향후에는 주어진 데이터의 부분만을 사용하는 k-최근접 이웃 알고리즘이나 k-군집화 알고리즘에 제안된 방법을 적용하면 더 높은 예측 정확도를 기대할 수 있을 것이다.

References

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in Selected papers of hirotugu akaike, ed: Springer, pp. 199-213, 1998.
DOI: https://doi.org/10.1007/978-1-4612-1694-0_15

[2] M. J. Garside, "The Best Subset in Multiple Regression Analysis," *Applied Statistics*, Vol. 14, pp. 196-200, 1965.
DOI: https://doi.org/10.2307/2985341

[3] M. G. Kendall, A course in multivariate analysis, C, Griffin, London, pp. 23-29, 1957.

[4] H. Hotelling, "The relations of the newer multivariate statistical methods to factor analysis," *British Journal of Statistical Psychology*, Vol. 10, pp. 69-79, 1957.
DOI: https://doi.org/10.1111/j.2044-8317.1957.tb00179.x

[5] A. E. Hoerl, and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, Vol. 12, No. 1, pp. 55-67, 1970.
DOI: https://doi.org/10.2307/1271436

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267-288, 1996.
DOI: https://doi.org/10.2307/41262671

[7] S. Arlot, and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, Vol. 4, pp. 40-79, 2010.
DOI: https://doi.org/10.1214/09-SS054

[8] A. Cowling, and P. Hall, "On pseudo data methods for removing boundary effects in kernel density estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 551-563, 1996.
DOI: https://doi.org/10.2307/2345893

[9] D. B. H. Cline, and J. D. Hart, "Kernel estimation of densities with discontinuities or discontinuous derivatives," Statistics: *A Journal of Theoretical and Applied Statistics*, Vol. 22, No. 1, pp. 69-84, 1991.
DOI: https://doi.org/10.1080/02331889108802286

[10] I. Gerlovina, Small Sample Inference, Doctoral dissertation, UC Berkeley, 2016. Available From: http://digitalassets.lib.berkeley.edu/etd/ucb/text/Gerlovina_berkeley_0028E_16680.pdf

[11] J. El Methni, L. Gardes, & S. Girard, "Kernel estimation of extreme regression risk measures," *Electronic journal of statistics*, Vol. 12, No. 1, pp. 359-398, 2018.
DOI: https://doi.org/10.1214/18-EJS1392

[12] M. Mudelsee, Extreme Value Time Series. *In: Climate Time Series Analysis*. Springer, pp. 217-267, 2014.
DOI: https://doi.org/10.1007/978-90-481-9482-7

[13] L. Breiman, "Using convex pseudo-data to increase prediction accuracy," *breast (Wis)*, Vol. 5, No. 2, pp. 1-18, 1998. Available From: https://statistics.berkeley.edu/sites/default/files/tech-reports/513.pdf

[14] A. Purwar, and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Systems with Applications*, Vol. 42, No. 13, pp. 5621-5631, 2015.
DOI: https://doi.org/10.1016/j.eswa.2015.02.050

[15] Z. Liu, S. Sharma, and S. Datla, "Imputation of missing traffic data during holiday periods," *Transportation Planning and Technology*, Vol. 31, No. 5, pp. 525-544, 2008.
DOI: https://doi.org/10.1080/03081060802364505

[16] S. F. Wu, C. Y. Chang, and Lee, S. J., "Time series forecasting with missing values," *In Industrial Networks and Intelligent Systems (INISCom), 2015 1st International Conference on IEEE*, pp. 151-156, 2015.
DOI: https://doi.org/10.4108/icst.iniscom.2015.258269

[17] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning, Springer, Berlin: Springer series in statistics, p. 38, 2009.

[18] D. Blend, & T. Marwala, Comparison of data imputation techniques and their impact, Available From: https://arxiv.org/ftp/arxiv/papers/0812/0812.1539.pdf

[19] D. Ruppert, and M. P. Wand, "Multivariate locally weighted least squares regression," *The annals of statistics*, pp. 1346-1370, 1994.
DOI: https://doi.org/10.1214/aos/1176325632

[20] M. R., Piña-Monarrez, "A new theory in multiple linear regression," *International Journal Of Industrial Engineering*, Vol. 18, No. 6, pp. 310-316, 2011
Available From: https://www.researchgate.net/publication/279181297A_new_theory_in_multiple_linear_regression

[21] B. Al-hnaity, and M. Abbod, "Predicting Financial Time Series Data Using Hybrid Model," *In Intelligent Systems and Applications*. Springer International Publishing, pp. 19-41, 2017.
DOI: https://doi.org/10.1007/978-3-319-33386-1_2

김 정 우(Jeong-Woo Kim) [**정회원**]

- 2005년 8월 : 고려대학교 심리학/경제학과 (심리학/경제학 학사)
- 2012년 8월 : 연세대학교 경제대학원 (경제학 석사)
- 2018년 2월 : 연세대학교 경제학과 (경제학 박사)
- 2018년 3월 ~ 현재 : 아산생명과학연구원 박사후연구원

<관심분야>
계량경제, 기계학습 등