

Global Report

교통 관련 학술문헌 검증을 통한 통계적 적합성 및 신뢰성에 대한 선택모델의 유효성 검사

1. 개요
2. 과학 및 공학의 신뢰성에 대한 위기
3. 유효성검사를 위한 데이터 분할 및
재샘플링 방법_부트 트래핑 방법
4. 성능 측정
5. 현장에서의 개선된 유효성검사 지향점 및
결론

1. 개요

2005년 Ioannidis 등은 자연과학에서 입증된 재현성(reproducibility)의 부족 문제를 조명한바 있으며 [1], 그 후 몇 년 동안 과학에서 "재현성 위기(reproducibility crisis)"는 대중 매체에서 헤드라인을 장식하였다[2]. 특히 운송산업에서 이러한 재현성에 대한 근본적인 우려가 있다. 운송산업 분야는 일반적으로 쉽게 재시도 할 수 있는 실험에 의존할 수 없으며, 관측연구를 통한 합리적인 결과가 유용성을 확보하기 위해서는 허용된 범주 내에서 시간과 공간을 통해 일반화해야 하기 때문이다. 이를 위해 관찰연구에서 정보를 추출하는 데 사용되는 모델의 유효성에 더 많은 주의를 기울여야 한다고 주장한다. 이 논문에서 제시한 제안은 운송 커뮤니티가 모델의 성과와 정책 관련성을 평가할 때 통계적 적합성에 지나치게 의존한다는 것이다. 예를 들어, 이동성에 대한 관찰연구가 일반화된 모델의 추정 운송공간에서 일반화되어 있다는 것이다. 이에 대해 분석전문가들은 버스를 이용하는 경우 버스를 기다리는 데 따른 상대적인 번거로움을 정량화한 데이터 세트를 사용하여 운송모드 선택을 결정하는 노력이 필요하다고 강조한다. 이러한 버스 대기시간은 선형회귀 또는 이산적인 통계 모델링의 경우, 승마보다 3배 정도 불편하다는 주장이 있다. 분석전문가들은 R-squared or pseudo-rho-squared과 같은 유사 연구를 기반으로 강력한 적합성 통계와 주요 변수에 대한 계수의 통계적 유의성을 보여주는 통계(계수가 0이 아닌)에서 정확성을 인지하고 있다. 분석전문가가 새로운(혹은 동일한) 샘플링을 통해 이러한 결과를 시도하거나 복제하는 것은 매우 드문 일이다.

2. 과학 및 공학의 신뢰성에 대한 위기

2016년 Nature지에서 자연재현성에 관한 독자의 설문조사 결과를 발표한바 있다. 설문결과, 연구원의 70% 이상이 다른 과학자의 실험을 시도하고 재현하지 못했으며, 절반 이상이 자신의 실험을 재현하지 못한 것으로 나타났다. 아울러 응답자의 52%는 상당한 재현성 위기가 있다고 응답했다[3]. 이러한 인식은 연구의 긍정적인 예측치를 추정한 2005년 Ioannidis의 주장을 반영하는 것이다. Ioannidis는 대부분의 기존 연구결과에 대한 통계데이터(연구설계 정의, 결과 및 방법에 대한 유효성 등)가 부족하여 효력이 낮을 가능성이 있다는 것을 주장하였다. Ioannidis에 의한 연구의 근본적인 관심사는 자연과학의 실험적인 연구결과가 운송산업 연구분야와 관련성이 있다. 일반적으로 학술적인 교통연구의 목적은 교통정책 설계 및 구현을 더 잘 알리기 위해 운송산업과 관련된 인간의 행동을 더 잘 이해하고 궁극적으로 예측하는 것이다. 그러나 운송산업 분야에서 실험을 수행하는 것은 비용이 많이 들거나 수요자의 운송에 방해가 되는 경우가 많다. 예를 들어, 자가용 승용차에서 대중교통으로의 모달 전환을 평가하기 위해 새로운 지하철 노선이 건설되는 실험을 실행하는 것을 고려해 보자. 실험경제학 응용프로그램의 몇 가지 예를 보면(2004년 Holguín-Veras 등) 교통연구의 대부분은 고전적인 방법으로 가설 테스트를 어렵게 하는 단면적인 관측연구에 의존하고 있음을 알 수 있다[4].

모델의 유효성 검사가 부족하면 모델이 추정 데이터에는 잘 맞지만, 이 추정 데이터집합 외부에서는 제대로 수행되지 않는 모델(즉, 오버 피팅) 위험이 증가함으로써 모델은 데이터신호 대신 잡음만 남게 된다. 이처럼 관찰연구에 의한 한계에도 불구하고, 모델 기반 학술연구에 따른 정책 영향평가는 거의 수행되지 않고 있다. 이러한 모델과 그들로부터 파생된 정책적 권장사항이 얼마나 옳고 얼마나 잘못된 측면이 있는지에 대한 피드백이 필요하다. 이러한 문제는 결과의 유효성을 평가하기 위해 분석수단에 통합할 필요성을 강하게 강조된다. 이 연구에서는 학술적 문헌이 얼마나 적합한지, 또는 통계적인 모델 검증에 지나치게 의존하고 있음을 보여준다.

3. 유효성검사를 위한 데이터 분할 및 재샘플링 방법_부트 트래핑 방법

Bradley Efron은 교차검증방법의 몇 가지 제한사항 중 일부를 해결하기 위해 부트 트래핑

(bootstrapping methods) 유효성 검사방법을 제안했다. 1-1-out 교차검증은 예측정확도의 거의 편견없는 추정을 제공하지만, 특히 샘플 크기가 작을 때, 부트 스트래핑 방법이 더 효율적인 것으로 나타난 반면, 종종 수용할 수 없는 높은 가변성을 나타내기도 한다[5][6]. 이 연구에서는 부트 스트래핑의 보다 광범위한 추정을 요약하여 간략하게 검증하고자 한다. 이 아이디어는 크기 n 의 데이터 집합을 감안할 때 부트스트랩 샘플은 원래 데이터 집합에서 n 인스턴스를 임의로 샘플링하고 이 샘플에서 모델을 추정하여 생성되는 간단한 개념을 갖고 있다. 이 샘플에 대한 예측오차 추정치는 모델을 원래 샘플에 적용하여 얻을 수 있다. 이 프로세스를 B 번 반복하여 이를 통해 간단한 부트스트랩 평균 예측오차 추정치를 (식 1)과 같이 얻을 수 있다.

$$BS_{simple} = \frac{1}{B} \sum_{b=1}^B \sum_{n=1}^N Q[y_n, \hat{y}_n^b] / N \dots\dots\dots (식 1)$$

또 다른 추정치는 부트스트랩을 그대로 하는 추정치로서 (식 2)와 같이 얻을 수 있다.

$$BS_{loo} = \frac{1}{N} \sum_{n=1}^N \sum_{b \in C_n} Q[y_n, \hat{y}_n^b] / B_n \dots\dots\dots (식 2)$$

여기서, B 부트스트랩 샘플 세트를 감안할 때, C_n 은 인스턴스 n 을 포함하지 않는 부트 스트래핑 샘플의 하위집합이며, B_n 은 C_n 의 크기를 의미한다. 이 추정치는 1-1-out 교차검증 추정치의 최신버전으로 간주될 수 있으며, 보다 효율적이면서 상향된 것으로 나타났다. 보다 개선된 방법은 명백한 예측오차(모델을 추정하는 데 사용된 동일한 샘플에서 추정되는 예측오차)에 대한 부트스트랩 추정치를 얻고 이를 수정하는 것이다. 이 추정치의 가중치는 부트스트랩 샘플에 존재할 인스턴스의 확률이 0.632라는 사실에서 비롯된 것에서 1-1-out 부트스트랩 추정치의 상향편향을 완화시킬 수 있다[7].

4. 성능 측정

교통 관련 문헌에서 1982년 Gunn과 Bates, 같은 해 Koppelman과 Wilmot은 개별 선택 모델에 대한 자세한 검증방법을 제시한바 있다. 이러한 연구를 출발시점으로 하여 운송산업 관련 연구문헌에 보고된 주요 성과 측정값을 간략하게 검토한다. 여기서 논의된 조치 목록은 완전하지는 않지만 검토된 연구에서 내부 및 외부 유효성검사를 위해 보고된 대부분의 조치를 다루고 있다는 점에서 포괄적이다. 모델의 예측능력을 평가하는 가장 간단한 방법 중 하나는 예측된 시장점유율을 관찰된 주식과 비교하는 것이다. 매우 간단하고 이해하기 쉽지만, 이 방법은 예측과 관측 사이의 합의적 수준을 평가하기 위한 정량적 측정을 제공하지는 않는다. 절대적/상대적 대응책과 통과/실패 조치로 분류하여 문헌에 보고된 모델 유효성 검사 성능 측정값의 정의를 요약하면 다음과 같다.

- 상대적 측정값은 모델 선택에 유용하지만 예측정확도를 절대적으로 측정하지는 않는다.
- 절대적 측정값은 다른 한편으로는 연구자가 평가할 수 있는 벤치마크 값을 생성하는 데 사용할 수 있다.

유사한 연구에 대한 이러한 모델의 성능이 절대적인 측정값을 사용하는 경우에도 모델성능은 선택 설정 크기 및 대체항목에 대한 기본 시장점유율과 같은 요인에 해당한다는 점에 유의해야 한다. 따라서, 다른 연구결과와의 비교는 이러한 한계의 명확한 이해로 해석되어야 하며, 모델우월성의 절대지표로 이해해서는 안된다.

5. 현장에서의 개선된 유효성검사 지향점 및 결론

대부분의 연구자들은 여행수요 분석의 목적이 효과적인 정책평가를 돕기 위해 유효한 예측을 하는 것이다. 하지만 대부분의 연구는 일종의 정책적 의미를 보고하는 것이었고, 이 연구에서는 교통산업 연구목적과 현재의 연구관행 사이에 명백한 단절이 있음을 보여주었다. 연구자들 사이에서 검증을 위해 독립적인 샘플을 생산하는 것은 엄청난 비용이 들지만, 가능한 한 연구자들은 이러한 데이터를 생성하거나 가정용 여행 관련 설문조사와 같은 기존 정기 설문조사를 사용하려고 노력할 필요가 있다. 이러한 데이터가 없는 경우에는 데이터 분할 및 부트 트래핑 방법은 저렴하며 필요한 노력의 관점에서 관련 정책을 추론하는 것과 크게 다르지 않아야 한다. 결과의 신뢰성과 통계 모델을 기반으로 추론의 신뢰성을 높이기 위해 학계의 모델을 평가하기 위해서는 더 강력한 기준이 필요하다.

이 연구에서는 현장에서 유효성 검사 관행을 개선하기 위해서는 i) 유효성검사를 의무화할 필요가 있으며, ii) 벤치마크 데이터셋을 공유할 필요가 있으며, iii) 검증연구를 장려할 필요가 있으며, iv) 명확한 보고 가이드라인을 작성하고 시행할 필요가 있다는 권장사항을 제시했다. 이 연구에서는 2004년과 2008년 사이에 발표된 문헌에서 운송산업 분야의 유효성검사 관행을 검토했다. 연구의 92%가 적합성 통계의 장점을 보고하고 있으며, 64.6%가 일종의 정책 관련 추론분석을 보고했지만, 검증 보고 비율은 18.1%에 달하는 것으로 나타났다. 이는 관측(비 실험적) 연구에 대한 운송 산업 연구 분야의 의존성을 감안할 때 검증 성능보다는 적합성 측정에 대한 의존도가 현명하지 않다는 것이다. 이를 통해 모델 검증이 학술 저널에서 모델 보고 및 동료 검토의 협상불가능한 부분이어야 한다고 주장하고 사용 가능한 리소스를 제공하는 유효성 검사 방법을 선택하는 간단한 추론을 제안했다. 아울러 더 나은 검증 관행을 촉진하고 가이드 라인을 보고하는 도구와 지식을 제공하기 위한 적절한 인센티브의 필요성을 인식하고, 기존 모델과 이론의 검증에 초점을 맞춘 제출 저널의 격려, 그리고 새로운 이론적으로 혁신적인 모델을 제시하였다. 이 연구결과는 국내 대도시의 도심, 고속도로 및 자동차 전용 도로에서의 교통 관련 학술연구 자료로 활용할 수 있을 것이다.

[참고문헌]

- [1] Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2 (8), 696–701. <https://doi.org/10.1371/journal.pmed.0020124>. Irfan, M., et al., 2018. Policy implications of work-trip mode choice using econometric modeling. J. Transport. Eng., Part A: Systems 144 (8), 04018035.
<<https://doi.org/10.1061/jtepbs.0000158>>
- [2] Baker, M., Penny, D., 2016. Is there a reproducibility crisis? Nature 533 (7604), 452–454.
<<https://doi.org/10.1038/533452A>>
- [3] Baker, M., Penny, D., 2016. Is there a reproducibility crisis? Nature 533 (7604), 452–454.
<<https://doi.org/10.1038/533452A>>
- [4] Holguín-Veras, J., Thorson, E., Ozbay, K., 2004. ‘Preliminary results of experimental economics application to urban goods modeling research’, Transportation Research Record. Journal of the Transportation Research Board 1873 (1), 9–16.
- [5] Efron, B., 1983. ‘Estimating the error rate of a prediction Rule : improvement on cross-validation’. J. Am. Stat. Assoc. 78 (382), 316–331.
- [6] Efron, B., Tibshirani, R., 1995. Cross-Validation and the Bootstrap : Estimating the Error Rate of a Prediction Rule. Division of Biostatistics, Stanford University.
- [7] Steyerberg, E.W., et al., 2001. ‘Internal validation of predictive models : efficiency of some

procedures for logistic regression analysis'. J. Clin. Epidemiol. 54, 774–781.

<원문제목> The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature

<원문출처> Journal of Choice Modelling 38 (2021) 100257