한국심리학회지: 일반
Korean Journal of Psychology: General
2021, Vol. 40, No. 4, 389-413
http://dx.doi.org/10.22257/kjp.2021.12.40.4.389

빅 데이터와 기계 학습의 시대 심리학 연구 모형의 평가 원칙과 방법*

이 태 헌†

중앙대학교 심리학과

본 논문에서는 계량 심리학 분야에서 지난 수 십 년 동안 꾸준히 논의가 진행되어 왔던 모형 추정과 평가의 원칙을 심리학 연구자들에게 소개하는 것을 목적으로 한다. 계량 심리학 분야에서 진행된 논의의 핵심은 1) 후보 모형들은 참 모형(true model)이 아니라 근사 모형(approximating model)이며, 2) 데이터 크기가 무한히 커지더라도 참 모형과 근사 모형 간 불일치는 사라지는 것은 아니기 때문에, 3) 여러 후보 모형 중 참 모형과의 불일치가 가장 낮은 것으로 추정되는 근사 모형을 선정하는 것이 바람직하다는 것이다. 이러한 모형 선정의 원리는 4차 산업 혁명의 시대, 여러 학문 분야에 걸쳐 그 영역을 확장하고 있는 기계 학습(machine learning) 분야에서 채택하고 있는 모형 평가의 원칙과 동일함을 설명하였다. 즉, 기계 학습 분야에서는 훈련(training) 과정에 노출되지 않았던 새로운 사례에서 보이는 모형의 성능인 일반화 혹은 예측 오차(generalization or prediction error)를 추정함으로써 모형을 선정하는데, 이는 계량 심리학 분야에서 근사모형과 참모형의 불일치 추정량인 총체적 오차(overall discrepancy)를 추정함으로써 모형을 선정해야 한다는 원리와 동일함을 설명하였다. 본 논문의 두 번째 목적은, 이러한 모형 선정의 원칙에 대한 이해를 바탕으로, 현재 심리학 분야에서 주어진 데이터에 대한 "철저한" 분석 관행이 초래하는 과적합(overfitting) 문제와 그 해결 방안을 논의하는 데 있다. 특히, 기계 학습 분야에서 가정 널리 사용되고 있으며, 계량 심리학 분야에서도 오래전부터 논의가 되어온(Mosier, 1951) 교차-타당성 입증법(cross-validation)을 일반화 오차의 추정량이라는 관점에서 소개하고 사용을 당부하였다.

주요어 : 과적합, 일반화 오차, 훈련 오차, 교차-타당성 입증법, 편향-분산 균형

---

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2020R1H1A1102581).

† 교신저자: 이태헌, 중앙대학교 사회과학대학 심리학과 부교수, (155-756) 서울시 동작구 흑석로 84
Tel: 02-820-5124, E-mail: lee0267@cau.ac.kr

심리학 연구를 비롯한 행동 과학, 나아가 과학 전반에서 통계적 추론(statistical inference)이 이루어지는 과정은 1) 모형의 지정(model specification), 2) 모형의 추정(model estimation), 3) 모형 평가(model assessment)[1] 등을 아우르는 모형-구축 과정(model-building process)[2]으로 요약할 수 있다(Chatfield, 1995). 이러한 모형-구축 과정을 구조방정식 모형(structural equation modeling)을 예로 들어 설명하면 다음과 같다.

통계적 추론의 목적

구조방정식모형을 사용하는 연구자들은 심리적 구성개념 혹은 요인들 간의 관계(예: 선형, 비선형, 매개, 조절 관계 등)를 이해하고 설명하는 데 관심이 있다.

모형의 지정

이러한 목적을 달성하기 위해서는 요인 점수와 문항 점수의 분포 형태(예: 정규 분포 혹은 이항 분포 등) 등에 대한 통계적 가정, 요인의 개수, 요인들 간의 관계, 개별 문항과 요인 간의 관계, 및 기타 연구자의 가설을 분포의 형태, 수식(equations) 그리고 미지수(unknowns)의 형태로 모형에 명시해야 한다. 이러한 과정, 즉, 확률 변수의 분포와 관계에 대한 통계적 가정과 연구자의 가설을 수식과 미지수 등의 형태로 모형에 명시하는 과정을 모형의 지정[3]라고 한다.

모형의 추정 및 추정 방법

모형에서 지정된 미지수는 파라미터(parameter)라고도 부르며, 데이터를 기반으로 그 값을 추산할 수 있다. 이러한 과정, 즉, 주어진 데이터를 기반으로 모형에 지정된 파라미터의 값을 추산하는 과정을 모형의 추정(model estimation)[4]이라고 부른다.

파라미터를 추정하는 방법(estimation method)은 일반적으로 관찰된 데이터와 모형 간의 불일치를 정의하는 손실 함수(loss function)를 정하고, 손실 함수를 최소화하는 파라미터 값을 추정치로 사용한다. 구조방정식모형을 비롯한 많은 모형에서 음의 로그-가능도 함수(negative log-likelihood function)라는 손실 함수를 이용하여 관찰된 데이터와 모형 간 불일치를 최소화하는 값을 파라미터 추정치로 정한다. 추정 결과 요인 간 상관(factor correlations), 요인 간 경로계수(path coefficient) 등에 대한 추정치를 얻을 수 있다.

모형 평가

구조방정식 등의 모형에 대한 추정이 완료되면, 추정된 모형-기반 (model-implied)[5] 문항 간 상관 패턴과 관찰된 자료로부터 계산된 문항 간 상관 패턴을 비교하여 불일치 혹은 유사성 정도를 평가하는 등의 방법으로 추정된 모형의 타당성을 가늠해 볼 수 있다. 이러한

1) 혹은 모형 점검(model checking)이라고도 한다.
2) 혹은 줄여서 모형화/모델링(modeling)이라 부를 수도 있다.
3) 혹은 모형의 특정화 혹은 명세화라고도 부른다.
4) 혹은 모형의 적합(model fitting)이라고 부르며, 본 원고에서는 심리학자들의 이해를 돕고, 불필요하게 논의를 확장하지 않기 위해 모수적 통계 추론에 초점을 맞춘다.
5) 모형이 예측하는(model-predicted)과 동일한 의미이다.

과정은 모형 평가에 해당한다. 다양한 방법을 통해 모형 평가를 실시할 수 있는데, 요인분석과 구조방정식 등의 잠재변인을 측정하고 관계를 분석하는 분야에서는 주로 RMSEA, CFI, TLI, SRMR 등의 적합도 지수(goodness-of-fit index)를 사용하며(Hu & Bentler, 1999), 잠재계층 분석에서는 AIC, BIC 등의 정보 지수를 사용하는 경향이 있다(Nylund et al, 2007). 회귀 분석에서는 주로 R^2 또는 조정된(adjusted) R^2을 사용한다.

심리학이라는 학문 분야에서 데이터의 수집과 통계적 분석, 그 결과에 대한 통계적 추론이 차지하는 위치는 매우 높다. 학부 시절부터 대학원 학위 과정을 마칠 때까지, 다양한 기초 및 고급 통계 모형을 배우고, 통계 분석 소프트웨어에서 모형을 지정하고 추정하는 방법, 그리고 모형을 평가하는 방법 등, 모델-구축(model-building) 과정 전반에 대해 체계적인 훈련을 받는다. 다년간의 훈련을 거쳐 비로소 *p*-value, 신뢰구간, 효과 크기, 표준 오차, 모형 적합도 등의 생경했던 용어에 대한 어색함과 불편함을 극복하고, 자신의 연구 가설을 검증하기 위해 수집된 데이터를 분석하고 분석 결과에 대한 통계적 추론 및 해석을 담은 논문까지 출판할 수 있게 된다. 심리학 연구자뿐만 아니라 모든 연구자들은 자신이 내 놓은 연구 결과가 후속 연구의 기반이 됨으로써 지식의 진보에 기여하기를 바라고 믿는다. 이러한 이유로 통계 전공자가 아닌 심리학 연구자로서 데이터 분석과 통계적 추론 과정에 대한 훈련을 기꺼이 감내하는 것이 아니겠는가.

심리학 연구자들의 이러한 노력에 비추어, 역설적이게도, 그리고 안타깝게도, 심리학은 최근 재현 위기(replication crisis) 논란의 중심에 서 있었으며, 그 원인과 해결책을 다양한 각도에서 모색하고 있다(Klein et al., 2014; 2018; Simmons et al. 2011). 본 논문의 저자는, 계량 심리학자로서, 현재 심리학자들이 "철저한"(thorough) 데이터 분석 과정[6]으로 간주하고 많은 연구에서 채택하고 있는 통계적 추론 절차 역시 재현 위기를 초래한 원인 중 하나일 수 있다는 진단(Lubke & Campbel, 2016)에 동의할 수밖에 없다.

본 논문에서는 먼저 계량 심리학 분야와 기계 학습 분야에서 지난 수십 년 동안 꾸준히 논의가 진행되어왔던 모형 추정과 평가의 원칙을 심리학자들에게 소개하고자 한다. 이를 통해 현재 많은 심리학자들이 채택하고 있는 "철저한" 데이터 분석을 통한 통계적 추론은 모형 평가의 원칙에 비추어 무엇이 문제이며 어떻게 개선할 수 있을지 논의하고자 한다.

기계 학습의 정의와 목적

최근 데이터 분석을 둘러싼 환경의 변화는 여러 학문 분야에서 빅-데이터와 기계 학습이

6) 이후 관련 섹션에서 더 자세히 논의하겠지만, "철저한" 데이터 분석 과정에서 "철저한"이란 단어는, 마치 "한 사람"의 행동과 심리를 이해하기 위해 행동 관찰 결과를 철두철미하게 분석하는 심리학자의 모습을 연상하며 사용하였다. 그래서, "철저한" 데이터 분석이란, 통계적으로 유의한 결과를 얻거나 자료와 모형의 적합도를 높이기 위한 목적으로 "주어진 데이터"를 여러 가지 다양한 방식을 동원하여 빈틈없이 밑바닥까지 파헤치듯 분석하는 방식을 가리키기 위해 사용하였다.

주도하고 있다고 해도 과언이 아니다(Kim, 2019). 기계 학습 분야에서의 학습(learning)이란, 컴퓨터 프로그램이 주어진 과제를 수행하는 데 필요한 경험(experience)을 통해 그 성능(performance)이 향상되는 것을 의미한다(Mitchell, 1997). 컴퓨터 프로그램이 학습하는 과제(task)의 종류는 매우 다양한데, 대표적으로 분류(classification)와 예측(regression)은 심리학 연구자들에게도 매우 익숙한 과제이다. 기계 번역(machine translation)이나 신용 카드 사기 탐지(fraud detection) 등의 과제도 최근 일상에서 쉽게 접할 수 있는 기계학습 과제이다. 학습 과정에서 프로그램은 주어진 데이터 세트(dataset)에 담긴 사례들(examples)을 경험(experience)한다. 여러 사례들에 대한 경험을 통해 데이터 세트를 생성한 확률 분포(data generating probability distribution)를 추정함으로써 과제 수행 능력의 향상을 꾀하는 것으로 이해할 수 있다.

학습 과정에서 추정하고자 하는 확률 분포의 종류에 따라, 혹은 프로그램이 경험하는 사례의 유형에 따라 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 나누기도 한다. 하지만 이러한 구분은 과제에 따라 상호 배타적으로 엄밀하게 구분되기는 어려우며, 혼재된 경우가 많다. 예를 들면, 군집 분석(clustering) 혹은 주성분 분석(principle component analysis) 등은 비지도 학습, 그리고 회귀 분석(regression analysis)은 지도 학습의 예가 될 수 있으나, 주성분 회귀 분석(principle component regression analysis) 등은 두 종류의 학습이 혼합된 경우라 할 수 있다.

기계 학습 프로그램의 성능 평가는 학습 과정에서는 경험하지 않았던 새로운 사례(previously unseen dataset)를 대상으로 진행하는 것이 일반적이다. 왜냐하면, 학습한 사례에 국한되지 않고, 실제 세상에서 마주하게 될 새로운 사례에 대해서도 우수한 성능을 보이는, 즉 보다 높은 예측 성능을 보이는 프로그램을 찾는 것이 기계 학습의 목적이기 때문이다. 학습 과정에서 경험한 사례들을 대상으로 성능을 평가하는 것은 프로그램의 암기(memorizing) 능력을 평가하는 것에 비유할 수 있을 것이다. 학생들이 교육 장면에서 배운 내용을 암기하는 것에 그치지 않고, 실생활에서도 적절하고 올바르게 적용할 수 있을 때 성공적인 학습으로 판단하는 것과 일맥상통한다고 할 수 있다.

컴퓨터 프로그램이 학습을 한다는 표현에 익숙하지 않은 심리학 연구자들을 위해 본 논문에서는 컴퓨터 프로그램 혹은 알고리즘이라는 용어 대신, 확률 모형(probability model) 혹은 줄여서 모형(model)이라는 용어를 사용하기로 하였다. 이들 용어에 대한 엄밀한 구분[7]은 본 논문의 논지에서 벗어나기 때문에 동일한 의미를 지니는 것으로 간주하고, 심리학 연구자들에게 가장 익숙하다고 판단한 확률 모형 혹은 모형(model)을 사용하여 논의를 진행하였다. 따라서 독자들은 컴퓨터 프로그램이나 알고리즘을 자신에게 익숙한 모형, 예를 들어 회귀 분석 모형이나 요인분석 모형, 잠재계층 모형 혹은 구조방정식 모형 등으로 간주하면 기계 학습 분야에서 진행되는 논의를 이해하는 데 도움이 되리라 생각한다.

7) 보다 엄밀한 구분을 원하는 독자는 다음 블로그를 참고해도 좋을 듯하다.
https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/

기계 학습 분야에서의 모형 평가: 일반화 오차

앞서, 기계 학습 모형의 성능 평가(assessment)는 학습 과정에서는 경험하지 않았던 새로운 사례를 대상으로 진행하다고 언급하였다. 이는 기계 학습의 목적은 일반화 가능한 모형을 선정하는 것이기 때문이다. 일반화(generalization)란 모형의 성능(performance)이 학습 과정에서 경험하지 못한 새로운 사례에까지 이어지는 정도라 정의할 수 있다(Hastie, et al., 2009). 모형의 일반화 정도(ability to generalize)는 일반화 오차(generalization error)를 이용하여 측정할 수 있다. 개별 사례(example)의 집합(set)으로 이루어진 데이터 세트를 D, k번째 확률 모형과 파라미터를 각각 M_k와 θ_k, k번째 확률 모형으로부터 데이터 세트 D를 관찰할 확률을 $p(D|\theta_k, M_k)$라고 하면, M_k의 일반화 오차는 아래와 같이 정의될 수 있다.

$$\epsilon_X^k = -E_Y\left[\log p(Y|\hat{\theta}_k^X, M_k)\,\middle|\, X\right]$$

여기서 X와 Y는 동일한 (참)분포 f로부터 얻은 서로 독립인 두 세트의 데이터를 나타내며[8], $\hat{\theta}_k^X$는 훈련 세트(training set)라고 부르는 X에 포함된 사례 $\{x_1, \ldots, x_n\}$를 기반으로 추정된 모형[9] M_k의 파라미터 값을 나타낸다.

일반화 오차에 대한 위 정의에 따르면[10], 모형 M_k가 Y에 포함된 새로운 사례에서 보이는 성능은, 추정된 모형으로부터 새로운 사례가 발생할 로그-확률로 측정하며, 가능한 모든 새로운 사례로부터 얻은 로그-확률에 대한 기댓값을 구함으로써 일반화 오차 ϵ_X^k를 구할 수 있게 된다. 추정된 모형으로부터 새로운 사례를 얻을 로그-확률 기댓값이 높을수록, 일반화 오차 ϵ_X^k는 낮아지고, 그 반대의 경우 일반화 오차 ϵ_X^k는 높아진다. 만약, 연구자가 여러 개의 후보 모형(예: $k = 1, 2, 3$)을 대상으로 성능을 비교한다면, 그 중 일반화 오차가 가장 낮은 모형을 선정하고 후속 분석을 진행하는 것이 가장 합리적일 것이다.

---

8) 엄밀하게 정의하면, X와 Y는 서로 독립이며 (참)분포 f를 따르는 확률 변수(random variable)를 나타내지만, 독자의 이해를 돕기 위해 확률 변수의 관찰값(데이터)으로 간주하고 논의를 진행하였다.

9) 여기서 모형을 추정한다는 것은, X를 기반으로 모형에 지정된 파라미터를 추정한다는 것을 의미하며, $\hat{\theta}_k^X$가 추정된 파라미터를 나타낸다. 예를 들어 구조방정식 모형을 추정한다는 것은 주어진 데이터 X를 기반으로 모형에 지정된 파라미터인 요인계수(factor loading)와 경로계수(path coefficient) 등을 추정한다는 것을 의미하며, 이 때 $\hat{\theta}_k^X$은 추정된 경로 및 요인 계수를 나타낸다.

10) 일반화 오차는 예측 오차(prediction error) 혹은 시험 오차(test error)라고도 부른다. 일반화 오차를 정의할 때 반드시 로그-확률 함수를 사용해야 하는 것은 아니며, 새로운 사례에 대한 관찰값과 모형 기반 추정값 사이의 차이를 이용하여(예: 차이의 제곱 혹은 차이의 절대값 등) 일반화 오차를 정의할 수도 있다. 그러나, 본 논문에서는 심리학 연구에서 가장 많이 사용되고 있는 모형 추정 방법인 최대-가능도 추정법과 직접적인 관련이 있는 로그-확률 함수를 사용하여 일반화 오차를 정의하였다. 정보 이론(information theory) 분야에서는 이렇게 정의된 일반화 오차를 교차-엔트로피(cross-entropy)라고 부른다.

수리 및 계량 심리학 분야에서의 모형 평가: 총체적 오차

심리학 분야에서는 수리 및 계량 심리학(mathematical & quantitative psychology) 분야를 중심으로 일반화 가능한 모형을 선정하기 위한 통계적 기준 마련에 대한 논의가 오래전부터 진행되었다(Zucchini, 2000; Cudeck & Henly, 1991; MacCallum & Tucker, 1991; Linhart & Zucchini, 1986). 본 섹션에서는 수리 및 계량 심리학 분야에서 진행된 논의를 소개하고 기계 학습 분야에서의 용어와 비교하여 설명하였다.

참분포, 참모형, 근사 모형

먼저, 데이터를 생성한 확률 분포(data generating probability distribution)를 $f(\theta_0)$라고 표기할 수 있으며, 이 때 θ_0는 분포를 결정하는 파라미터를 나타낸다[11]. 본 논문에서는 참분포(true distribution) 혹은 참모형(true model)이라는 용어 역시 데이터를 생성한 확률 분포를 가리키는데 사용하였다. 앞서 언급하였듯이, 기계 학습에서 뿐만 아니라 심리학 분야에서도 데이터를 생성한 참분포 혹은 참모형을 추정하는 것은 여러 과제를 수행하는 데 밑바탕이 되므로 그 자체로도 매우 중요한 학습 과제라 할 수 있다. 현실에서는 참모형을 알고 있는 경우는 없기 때문에 후보가 되는 모형을 추정하고 그 성능을 평가하는 것이 일반적이다. 후보가 되는 모형은 근사 모형(approximating model)이라고 부르며, k번째 근사 모형은 $g_k(\theta_k)$로 나타낼 수 있다.

총체적 불일치와 일반화 오차

일반화 오차는 수리/계량 심리학 분야에서는 총체적 불일치(overall discrepancy, OD)라는 개념으로 논의되어 왔다. 근사 모형 $g_k(\theta_k)$의 파라미터를 훈련 세트 $X=\{x_1, x_2, ..., x_n\}$를 기반으로 추정할 수 있는데, 이 때 추정된 파라미터를 $\hat{\theta}_k^X$라고 하면, 근사 모형 $g_k(\hat{\theta}_k^X)$이 참분포 f로부터 얻은 새로운 사례 세트 Y에서 보이는 성능을 쿨벡-라이블러 거리(Kullback-Leibler distance; 이하 K-L 거리)[12]를 이용하여 정의하면 아래와 같다.

$$\Delta_X(\theta_0, \hat{\theta}_k^X) = E_Y\{\log f(Y, \theta_0)|X\} - E_Y\{\log g_k(Y, \hat{\theta}_k^X)|X\}$$

이렇게 정의된 거리 $\Delta_X(\theta_0, \hat{\theta}_k^X)$를 총체적 불일치라고 부르며 기계 학습 분야에서의 일반화 오차 ϵ_X^k와 동일한 개념임을 알 수 있다[13]. 여러 후보 근사 모형 $g_k(\theta_k)$

11) 일반적으로, 참분포 f를 결정하기 위한 파라미터 θ_0가 반드시 필요한 것은 아니지만, 독자들의 이해를 돕기 위해 참분포가 θ_0에 의해 결정되는 경우를 상정하고 논의를 진행하였다. 참분포가 평균이 μ_0이고 분산이 σ_0^2인 정규분포인 경우, $\theta_0=(\mu_0, \sigma_0^2)$가 되며, 예를 들어 평균과 분산이 0과 1로 결정되면 f가 결정된다. 참분포가 감마 분포인 경우 $\theta_0=(\alpha, \beta)$이며, α와 β의 값에 따라 평균, 분산, 왜도 첨도 등이 결정되어 참분포의 모양을 결정한다.

12) 쿨벡-라이블러 발산(Kullback-Leibler divergence)이라고도 한다.

13) $E_Y\{\log f(Y, \theta_0)|X\}$는 참분포에 의해 결정이

$(k = 1, 2, ..., K)$ 중에서 총체적 불일치 $\Delta_X(\theta_{0,}\hat{\theta}_k^X)$가 가장 낮은 모형을 선정하는 것이 가장 합리적이라 할 수 있다.

수리/계량 심리학 분야에서는 총체적 불일치가 다시 근사 오차(error of approximation)와 추정 오차(error due to estimation) 두 부분으로 나뉠 수 있다는 것이 잘 알려져 있다(Bozdogan, 2000; Cudeck & Henly, 1991; Browne & Cudeck, 1989). 기계 학습 분야에 익숙한 독자를 위해 결론부터 언급하면, 근사 오차와 추정 오차는 각각 편향 오차와 분산 오차에 대응되는 개념으로 이해할 수 있다(Chung et al., 1996; Arlot et al. 2001).

근사 오차와 추정 오차

먼저 근사 오차(편향 오차)는 참분포와 근사 모형간 불일치 정도를 나타낸다. Y는 참분포/참모형 f를 따르는 확률 변수, θ_0는 참분포/참모형 f를 결정짓는 파라미터, g_k는 θ_k를 파라미터로 갖는 k번째 근사모형(approximaing model)이라고 할 때, g_k와 f의 불일치 정도는 K-L 거리를 이용하여 아래와 같이 표현될 수 있다.

$$\Delta(\theta_0, \theta_k) = E_Y\{\log f(Y, \theta_0)\} - E_Y\{\log g_k(Y, \theta_k)\}$$

여기서 θ_{k0}를 참분포와의 거리를 최소화하는 근사모형 파라미터 값이라고 할 때, $\Delta(\theta_0, \theta_k)$는 $\theta_k = \theta_{k0}$ 일 때, 최소가 된다.

근사 오차(편향 오차)는 참모형과 근사모형 간의 불일치 정도를 반영하며, 대체로 수리 및 계량 심리학 분야에서는 근사 오차(error of approximation)로, 통계학 및 기계학습 분야에서의 편향 오차(bias error)로 불리는 경향이 있다. 위에서 정의된 근사 오차는 항상 양의 값을 가지며, 근사 모형과 참 모형이 일치(여기서는 $\theta_0 = \theta_{k0}$)하는 경우에만 0의 값을 가질 수 있다. 또한 근사 오차는 모집단 수준에서 정의되므로 표집 과정에 영향을 받지 않는다. 즉, 표본의 크기가 커진다고 해서 근사 오차가 줄어드는 것은 아니다[14]. 근사 오차 혹은 편향 오차는 보다 복잡하고 유연한 모형[15]에서 더 낮은 경향이 있다고 알려져 있다.

추정 오차와 분산 오차

다음으로 추정 오차(error due to estimation)는 아래와 같이 정의될 수 있다.

되는 상수(constant)이므로 $\Delta_X(\theta_{0,}\hat{\theta}_k^X) \propto -E_Y\{\log g_k(Y, \hat{\theta}_k^X)|X\}$임을 쉽게 알 수 있다.

14) 구조방정식 분야에 익숙한 독자에게 매우 친숙한 RMSEA라는 지수가 근사 오차 추정치에 기반한 모형적합도 지수이다(Browne & Cudeck, 1993). RMSEA는 root mean squared error of approximation의 약자인데, 여기에서 error of approximation이 바로 근사 오차를 가리킨다.

15) 모형의 복잡성(model complexity) 혹은 유연성(flexibility)는 다양한 방식으로 정의될 수 있지만, 가장 쉽게 이해할 수 있는 방법은, 모형에 지정된 파라미터의 수가 더 많고 비선형 관계를 가지며 서로 연관성이 높을수록 모형의 복잡도가 커지는 것으로 정의한다(Bozdogan, 2000; Preacher, 2006).

$$\Delta_X(\theta_{k0},\hat{\theta}_k^X) = E_Y\{\log g_k(Y,\theta_{k0})\} - E_Y\{\log g_k(Y,\hat{\theta}_k^X)|X\}$$

이는 훈련 세트 $X=\{x_1, x_2,...,x_n\}$를 기반으로 추정한 근사 모형과 참분포를 기반으로 추정된 근사 모형이 새로운 자료 Y에서 보이는 성능의 차이를 반영한다. 이러한 오차는 파라미터를 추정하는 것으로 인해 발생하는 오차라는 의미를 반영하여 수리/계량 심리학 분야에서는 추정 오차라고 부르며, 기계 학습 분야에서의 분산 오차(variance error)에 대응되는 개념이다(Chung et al., 1996; Arlot et al., 2010).

심리학 연구자들이 사용하는 대부분의 모형에서 훈련 세트 $X=\{x_1, x_2,...,x_n\}$의 크기가 커질수록 $\hat{\theta}_k^X$은 θ_{k0}에 근접한다는 가정이 성립하므로[16], 표본이 커질수록 추정 오차(분산 오차)는 줄어드는 경향이 있다. 다만, 주어진 훈련 세트의 크기에 비해 모형이 지나치게 복잡할 경우, 파라미터 추정치의 변동성이 증가하는 경향을 보이는데, 이러한 특성을 반영하여 기계 학습 분야에서는 분산 오차라 부르는 경향이 있다. 표본의 크기가 커질수록 파라미터 추정치의 변동성은 줄어드는 경향이 있으므로 데이터 훈련 크기가 커질수록 분산 오차 혹은 추정 오차는 줄어드는 경향이 있다.

표본 불일치와 훈련 오차

표본 불일치는 추정된 근사 모형이 훈련 세트에 포함된 사례의 특성을 설명하는 데 보이는 성능을 가리킨다. 이는 훈련 세트 $\{x_1, \dots, x_n\}$를 기반으로 추정된 근사 모형으로부터 훈련 데이터에 포함된 사례를 관찰할 로그-확률을 이용해 아래와 같이 정의할 수 있다.

$$s_k^X = -\frac{1}{n}\sum_{i=1}^{n}\{\log g_k(x_i,\hat{\theta}_k^X)\}$$

표본 불일치는 기계 학습 분야에서는 훈련 오차(training error), 그리고 수리/계량 심리학 분야에서는 표본 불일치(sample discrepancy)라고 부르는 경향이 있다.

모형 평가의 원칙: 근사 오차와 추정 오차의 균형

수리/계량 심리학 분야에서 진행되어 온 논의에 따르면, 표본 불일치 혹은 훈련 오차가 아니라 총체적 오차가 가장 낮은 모형을 선정하는 것이 바람직하다는 것을 쉽게 이해할 수 있다. 표본 불일치가 더 낮은 모형이라고 해서 총체적 불일치 역시 더 낮다는 보장은 없다. 총체적 불일치는 근사 오차뿐만 아니라, 추정 오차에 의해서도 영향을 받는데, 모형이 복잡해질수록 근사 오차는 줄어들지만 추정 오차는 커지는 경향이 있어 무조건 모형을 더 복잡하게 만드는 것은 총체적 오차를 줄이는 데 도움이 되지 않는다. 따라서, 근사 오차와 추정 오차의 균형을 맞출 수 있는 적절한 수준의 복잡도를 가진 모형을 선정하는 것이 결국 총체적 불일치를 최소화할 수 있는 방법이다. 이와 같은 모형 선정 원칙을 기계 학습 분야에서는 편향-분산 균형(bias-variance trade-off)

16) 일치 추정량(consistent estimator) 가정이라고 한다.

라고 부르며, 모형이 복잡해질수록 편향 오차는 감소하지만 분산 오차는 증가하는 경향을 감안하여 두 종류의 오차가 균형을 이루어 일반화 오차를 최소화하는 수준에서 모형의 복잡도를 결정하는 원칙이다. 결국, 기계 학습과 계량 심리학 분야의 모형 평가는 용어만 다를 뿐 그 원칙은 동일함을 알 수 있다.

심리학 분야의 "철저한" 데이터 분석을 통한 통계적 추론

현재 심리학 분야의 연구자들은 앞서 언급한 모형의 지정, 추정, 그리고 평가라는 모형-구축 과정을 충실히 따르고 있으며, 사용자 중심의 통계 처리 소프트웨어의 발전(예: SPSS, Mplus, R/RStudio, Jamovi 등)으로 대용량 자료와 복잡한 모형을 대상으로 모델-구축 과정을 편리하고 신속하게 진행할 수 있게 되었다. 모델-구축 과정이 편리하고 신속해졌다는 것에 동의하기 어려운 독자도 있을 수 있으나, 이는 과거에 비하면 그렇게 되었다는 뜻이다. 심리학자들에게는 매우 잘 알려진 스크리-검사(scree-test)를 제안하였던 Catell(1966)은 자신의 논문에서 "백여 차례의 요인 분석"을 "30여년"동안 진행해 본 경험을 밝히며 논의를 진행하고 있다. 2021년이 막바지로 치닫고 있는 현재, 심리학 연구자가 백여 차례의 요인 분석을 실시하는 데 요구되는 시간과 노력을 생각해 본다면, 현재 우리가 사용하고 있는 컴퓨터의 계산능력과 통계 프로그램의 편리함을 쉽게 미루어 짐작해 볼 수 있다.

예를 들면, 수많은 측정 변수들을 대상으로 실시한 상관 분석 결과를 기초로 주요 예측변수를 추려 내고, 이들을 대상으로 매개 관계나 상호 작용에 대한 후속 분석을 진행하는 것은 적어도 통계 소프트웨어 상에서 결과를 얻는 데에는 그리 오랜 시간이 필요하지 않다. 이 과정에서 이상값(outliers; Lee et al., 2015; Bollen & Arminger, 1991; Yuan & Zhong, 2008)이나 결측치(missing values; Lee & Shi, 2021; Enders & Mansolf, 2018)를 적절히 처리한 뒤 재분석을 실시하거나 여러 하위 집단을 대상으로 분석을 반복하는 것 역시 그 결과를 신속하게 얻을 수 있게 되었다. 심지어 데이터가 수집되는 동안 분석을 실시하여 연구 가설과 합치되는 결과를 얻는 순간 데이터 수집을 중단하기도[17] 할 만큼 매우 신속한 데이터 분석이 가능해졌다. 특히 Quasi-Newton's method(Press et al., 2007) 혹은 EM-알고리즘(Dempster et al., 1977) 등은 거의 대부분의 통계 소프트웨어에 구현됨으로써 통계적 추론에서 여러 가지 장점을 지닌 것으로 알려진 최대-가능도 추정법(maximum likelihood estimation)을 손쉽게 사용할 수 있게 되었다. 즉, "좋은" 추정 방법으로 알려진 최대-가능도 추정법을 사용하여 모형에 지정된 파라미터를 추정하고 그 과정에서 데이터 특성이나 변수들 간의 관

17) 유연 중단 규칙(flexible termination rule)이라고도 불리는 이러한 방법은 일부 독자들에게는 새롭고 매력적인 방법으로 보일 가능성이 있으나, 데이터 수집 과정에서 얻은 (초기의) 일부 자료를 분석한 결과가 통계적으로 유의하더라도 데이터 수집을 중단해서는 안된다. 데이터 크기는 검정력 분석을 바탕으로 결정되는 연구 설계의 일부이며, 수집 과정에서 얻은 일부 데이터를 분석한 결과(interim data analysis)가 통계적으로 유의하더라도, 수집이 완료되어 얻은 전체 데이터를 분석한 결과 역시 반드시 통계적으로 유의하다는 보장은 없기 때문이다(Simmons et al., 2011).

계에 대한 탐색적 분석 결과를 반영하여 수정한 모형을 다시 평가할 수 있는 "철저한" (thorough) 데이터 분석 과정을 신속하고 편리하게 진행할 수 있게 되었다.

문제점

이러한 "철저한" 데이터 분석 과정이 신속하게 진행될 수 있는 환경이 형성됨으로써 문제점도 발생하게 되었다. 즉, HARKing[18](Kerr, 1988), 혹은 *p*-hacking(Wasserstein, et al., 2019), 그리고 QRP(questionable research practice; Simmons et al., 2011; Wiggins & Christopherson, 2019) 등으로 알려진 잘못된 통계 분석 관행이 고착화되고 제 1 종 오류가 증가하는 등의 부작용이 증가하게 되었다. 그렇다면, 구체적으로 어떤 문제가 왜 발생하는지에 대한 이해가 필요하며, 해결 방안은 없는지에 대한 논의도 필요하다. 다음 섹션에서는 최대 로그-가능도 추정방법을 중심으로 이에 대한 논의를 진행하였다.

최대 로그-가능도 추정법과 과적합

컴퓨터 계산 능력의 향상과 통계 분석 소프트웨어의 발전과 더불어 현재 심리학을 비롯한 사회과학 분야에서 모형 추정과 평가를 위해 가장 많이 사용되고 있는 최적화 방법 중에 하나가 최대 로그-가능도 추정법이다. 최대 로그-가능도 추정법은 주어진 데이터에 포함된 사례 $X=\{x_{1,}\cdots,x_n\}$가 발생할 로그-확률을 최대화하는 혹은 반대로 음의 로그-확률(negative log-probability)를 최소화하는 값을 파라미터 추정치로 결정하고, 이에 근거해 모형을 평가하는 방법이다. 이렇게 얻은 추정치를 최대 로그-가능도 추정치(maximum likelihood estimate; MLE)라고 부르며 아래와 같이 표기할 수 있다.

$$\tilde{\theta}_k^{(mle)} = \arg\min\left\{ -\frac{1}{n}\sum_{i=1}^{n}\{\log p(x_i|\theta_k, M_k)\}\right\}$$

이때 음의 로그-가능도 최소값은 아래와 같이 표기할 수 있다.

$$s_k^{(mle)} = -\frac{1}{n}\sum_{i=1}^{n}\left\{\log p\left(x_i|\tilde{\theta}_k^{(mle)}, M_k\right)\right\}$$

즉, $s_k^{(mle)}$는, 앞서 논의하였던 일반화 오차 ϵ_X^k혹은 총체적 불일치 $\Delta_X(\theta_{0,}\hat{\theta}_k^X)$와 달리, 주어진 데이터 $\{x_{1,}\cdots,x_n\}$를 기반으로 추정된 M_k가 다시 $\{x_{1,}\cdots,x_n\}$에서 보이는 성능을 나타내는 것을 알 수 있다. 즉, $s_k^{(mle)}$는 추정된 모형이 학습 과정에서 경험한 사례에서 보이는 성능을 측정하며, 기계 학습 분야의 훈련 오차(training error), 계량 심리학 분야의 표본 불일치(sample discrepancy)를 가리키는 것임을 알 수 있다. 따라서, 최대 로그-가능도 추정법은, 일반화 오차가 아니라, 훈련 오차를 최소화하는 추정법이라 할 수 있다.

훈련 오차를 최소화하는 모형이라고 해서

18) Hypothesizing after research results are known의 약자로서, 주어진 데이터에 대한 탐색적 분석 결과에 맞추어 원래의 연구 가설을 수정하거나 새로운 가설을 생성해 내는 것을 비판적으로 일컫는 용어이다. 문제는 이렇게 자료를 기반으로 도출된 가설(data-driven hypothesis)을 마치 이론적으로 도출된 가설(theory-driven hypothesis)인 것처럼 다루는 것에서 발생한다.

반드시 일반화 오차를 최소화하는 모형은 아니다. 즉, $s_1^{(mle)} < s_2^{(mle)}$라고 해서 반드시 $\epsilon_X^1 < \epsilon_X^2$를 의미하는 것은 아니라는 것이다. 이는 $s_k^{(mle)}$가 모형을 추정할 때 사용한 사례 $\{x_1,\ldots,x_n\}$를 재사용하여 모형의 성능을 평가한 결과인 반면, ϵ_X^k는 새로운 사례 $\{y_1,\ldots,y_n\}$를 기반으로 모형의 성능을 평가한 결과이기 때문이다.

주어진 사례에만 존재하는 노이즈까지 학습함으로써 새로운 사례가 주어졌을 때 모형의 성능이 오히려 하락하는 현상을 과적합(overfitting)이라고 한다. 달리 말하면, 어떤 모형이 학습 과정에서 경험한 사례를 설명하는 데 지나치게 적합하여 해당 모형의 설명 방식이 새로운 사례에는 일반화되지 못하는 경우 과적합 문제(overfitting problem)가 발생했다고 한다. 두 근사 모형 M_1과 M_2에 대하여, $s_1 < s_2$이지만, $\epsilon_X^1 > \epsilon_X^2$인 경우 M_1은 과적합 모형(overfitted model)으로 이해할 수 있다.

암기 능력, 과적합, 그리고 재현 위기

일반화 오차와 훈련 오차 간 이러한 관계는 직관적 수준에서도 쉽게 이해할 수 있다. 즉, 학습 과정에서 경험한 사례를 재사용하여 모형을 평가하는 것은 암기(memorizing) 결과를 테스트하는 것에 비유할 수 있고, 반면 학습 과정에서 경험하지 못한 새로운 사례를 사용하여 모형의 성능을 평가하는 것은 일반화 능력(ability to generalize)의 습득 결과를 테스트하는 것에 비유할 수 있다. 암기 테스트에서 최고 점수를 보인 학생이라고 해서 일반화 능력 테스트 점수 또한 최고라는 법은 없다. 암기 테스트에서 높은 점수를 받기 위해서는 학습 과정에서 경험한 사례들로부터 주요 핵심 패턴(pattern) 혹은 신호(signal)뿐만 아니라, 해당 사례에만 존재하는 특유의 속성들(peculiar properties)까지 모두 암기하는 것이 유리하다. 학습 중 경험한 사례에만 존재하는 특수한 속성(들)은 새로운 사례에서는 다시 등장하지 않는 노이즈(noise)로서, 학습 사례에 대한 과도한 학습, 즉 노이즈까지 암기하는 것은 일반화 능력 테스트에서 높은 점수를 받는 데 도움이 되지 않는다. 오히려 핵심적인 주요 패턴과 신호를 파악하는 데 방해가 될 수도 있다. 따라서, 일반화 능력 테스트에서 더 높은 점수를 얻기 위해서는 학습 사례가 보이는 모든 속성들에 대한 과도한 암기 보다는, 주요 핵심 특징들을 파악하는데 주력하는 것이 훨씬 더 유리하다.

이러한 점에 비추어 본다면, 훈련 오차만을 줄이는 것을 목표로 추정된 과적합 모형을 기반으로 한 통계적 추론은 연구 결과의 재현 가능성을 낮추는 주요 원인 중 한가지가 된다는 점을 부인하기 어렵다. 재현 가능한 연구 결과를 얻기 위해서는 주어진 사례에 대한 훈련 오차를 줄이는 것뿐만 아니라, 새로운 사례에 대해서도 우수한 성능을 보이는, 즉, 일반화 오차 혹은 예측 오차가 낮은 모형에 기반하여 통계적 추론을 진행해야 할 것이다. 즉, 주어진 데이터를 "철저하게" 분석하는 것 대신, 과적합 문제에서 보다 자유롭고 일반화 오차가 낮은 모형을 선정할 수 있도록 해 주는 방법을 사용하는 것이 요구된다.

일반화(예측) 오차 추정 방법

일반화 오차 ϵ_X^k는 (참)분포 f로부터 얻을 수 있는 모든 가능한 새로운 사례를 이용하여야만 얻을 수 있는 이론적인 값이다. 그러나 현실적으로 연구자가 수집한 데이터 세트 X와 별개로 새로운 사례 Y를 추가적으로 그것도 무한한 크기의 사례를 수집한다는 것은 불가능에 가깝다. 따라서 일반화 오차는 처음 수집된 데이터 세트 X를 기반으로 추정해야 하는데, 사례가 매우 풍족한(data-rich) 이상적인 경우 다음과 같은 3-단계 전략을 사용할 수 있다(Hastie, et al., 2009).

- (데이터 분할) 먼저 데이터에 포함된 사례를 세 세트로 나누어, 하나는 훈련 세트(training set, X_C), 다른 하나는 타당화 세트(validation set, X_V), 그리고 나머지는 시험 세트(test set, X_T)로 명명한다.

- (모형의 추정) 다음으로, 훈련 세트에 있는 사례를 이용하여 모형(의 파라미터)을 추정하고, 타당화 세트를 이용하여 추정된 모형의 성능을 평가한다. 이때, 타당화 세트에서의 성능을 향상시키기 위해 특정 파라미터 값을 조정(tuning) 혹은 갱신(update)[19]할 수 있다. 보통 여러 개의 후보 모형[20]중에서 타당화 세트에서의 성능이 가장 우수한 모형을 선정한다.

- (모형의 평가) 일반적으로 타당화 세트에서 계산된 모형의 성능은 과대 추정되는 경향이 있다. 이는 타당화 세트에 있는 사례[21]들을 이용하여 일반화 오차를 낮추는 방향으로 모형의 파라미터 값을 갱신하였기 때문에 발생하는 문제이다. 따라서, 모형의 일반화 오차는 학습 과정에서 노출된 적이 없었던 사례의 집합인 시험 세트를 이용하여 추정하는 것이 보다 정확한 결과를 산출한다.

일반화 오차를 추정하기 위한 이러한 접근은 데이터가 매우 풍족할 경우 가능한 방법이며, 빅-데이터의 시대에 활용 가능성이 더 높아진 방법이라 할 수 있다. 기계 학습 분야에서는 현재 이러한 3-단계 전략을 거의 모든 모형의 성능 평가에 적용하고 있다고 하여도 과언이 아니다. 하지만 현실에서는 여전히 데이터의 크기가 이상적 수준에 접근할 만큼 풍족한 경우는 드물기 때문에, 주어진 데이터가 무한히 커지는 상황을 이론적으로 가정하는 대표본 근사법(asymptotic approximation)을 사용하거나 (예: AIC, Akaike, 1973; 1974) 혹은 주어진 데이터를 보다 효율적으로 재사용(efficient reuse)하는 방법(예: 교차-타당화, Arlot et al., 2010) 등으로 일반화 오차를 추정한다

19) 이러한 목적으로 즉, 타당화 세트에서 보이는 모형의 성능을 향상시키기 위해 그 값을 조정하는 파라미터는 훈련 세트를 이용하여 추정 혹은 최적화하는 파라미터와 구분하여 조정 파라미터(tuning parameter) 혹은 초파라미터(hyperparameter)라고 부른다.

20) 예를 들어, 요인의 개수가 3, 4, 5개인 모형, 두 집단 간 경로계수가 동일한 모형과 다른 모형, 경로계수가 +0.3인 모형과 -0.1인 모형 등의 여러 후보 모형이 있을 수 있다.

21) 기계학습에서 사례, 예, 보기(example)라는 용어는 심리학 연구자들에게 익숙한 p개의 변수를 측정한 크기가 n인 자료에서 i번째 사례(case), $X_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})'$를 가리킨다.

(Hastie, et al., 2009). 다음 섹션에서는 데이터의 크기가 이상적인 수준에 미치지 못할 경우 사용할 수 있는 일반화 오차 추정법인 교차-타당성 입증법(cross-validation)과 AIC (Akaike Information Criterion)를 일반화 오차 추정량의 관점에서 소개하였다.

교차-타당성 입증법

교차-타당성 입증 방법(Arlot, et al., 2010)은 일반화 오차를 추정하기 위해 기계 학습 분야에서 가장 널리 사용되는 방법이라 할 수 있다. 여기서는 일반화 오차를 추정하고 최소화하는 모형을 선정하기 위해 현재 가장 널리 사용되는 *k-fold* 교차-타당화 입증법을 중심으로 논의를 진행하고자 한다. 여기서 *k-fold* 라는 용어는 데이터 세트 X를 무선 할당을 통해 크기가 동일한 혹은 유사한 k개의 부분(fold)으로 나눈다는 의미를 담고 있다.

먼저, $k=2$인 경우, 크기가 n인 데이터 $X=\{x_1,...,x_n\}$를 무선 할당을 통해 두 부분으로 나누어(i.e., 2-*fold*), 하나는 근사 모형의 파라미터를 추정하는 데 사용하고, 나머지 하나는 추정된 근사 모형이 새로운 데이터에서 보이는 성능, 즉 일반화 오차를 입증하는 데 사용한다. 근사 모형의 파라미터를 추정하는 데 사용한 데이터를 훈련 세트(training set)라고 부르고, 추정된 근사 모형의 성능을 평가하기 위해 사용된 나머지 데이터를 타당화 세트(validation set)라고 부르며, 그 크기는 각각 n_c와 n_v로 나타내고 $n=n_c+n_v$의 관계가 성립한다. 대개의 경우 대체로 n_c와 n_v는 동일하거나 비슷하도록 한다. 이 과정을 공식으로 나타내면 다음과 같다.

$$\hat{\Delta}_{X_c}\left(\hat{\theta}_k^{X_c}\right)=\frac{1}{n_v}\sum_{i=1}^{n_v}\left\{\log g_k\left(x_i^v,\hat{\theta}_k^{X_c}\right)\right\}$$

여기서 $\hat{\theta}_k^{X_c}$는 훈련 세트를 기초로 추정한 모형의 파라미터를 나타내고, x_i^v는 타당화 세트에 있는 새로운 사례를 나타낸다. 따라서, $\hat{\Delta}_{X_c}\left(\hat{\theta}_k^{X_c}\right)$는 추정된 모형으로부터 새로운 사례가 발생할 평균 로그 확률로서 일반화 오차를 추정하는 추정량으로 사용될 수 있다. 이 방법의 문제점은 주어진 데이터 세트 X에 비해 X_C와 X_V의 크기는 반으로 줄어들어 모형 추정의 정확성이 낮아지고 결국 추정된 일반화 오차의 정확성에도 부정적인 영향을 미칠 수 있다는 점이다.

이러한 점을 보완하기 위해, 데이터를 두 부분으로 나누되, $n_C=n-1$, $n_v=1$로 하여 교차-타당화를 진행하는 방법이 제안되었다(Stone, 1974). 이 방법에서는 개별 사례를 제외한 나머지 사례들을 훈련 세트로 사용하고, 따로 떼어 놓은 하나의 사례를 타당화 세트로 간주하기 때문에 leave-one-out (LOO) 교차-타당도 검증법(LOOCV)으로 알려져 있다(Stone, 1977; Stone, 1974; Allen, 1974; Geisser, 1975). LOOCV를 공식으로 나타내면 아래와 같다.

$$\hat{\Delta}\left(\hat{\theta}_k^{(-i)}\right)=\frac{1}{n}\sum_{i=1}^{n}\left[\log g_k\left(x_i,\hat{\theta}_k^{(-i)}\right)\right]$$

여기서 $\hat{\theta}_k^{(-i)}$는 i번째 사례를 제외한 나머지 $n-1$개의 사례들을 기반으로 추정된 파라미터를 나타낸다. 이러한 방법의 문제점은 모형의 추정을 n번 반복해야 하므로 모형이 복잡하고 n이 매우 클 경우 계산량의 부담이

매우 커지는 문제가 발생한다. 또한 일반화 오차의 추정량으로서 분산 오차가 매우 크다는 단점이 있다.

현재 기계 학습 분야에서 가장 널리 사용되고 있는 교차-타당도 검증법은 $k=5$ 또는 $k=10$을 사용한다(James et al., 2013; Kuhn et al., 2013; Hastie et al. 2009; Geisser, 1975; Breiman et al., 1984). 예를 들어, $k=5$인 경우, 전체 자료를 무선 할당을 통해 크기가 유사한 5개의 부분으로 나눈 뒤, 그 중 한 부분을 타당화 세트로 사용하고 나머지는 훈련 세트로 사용하는 방법이다. 이 때, $k=5$인 경우, 총 다섯 개의 타당화 세트로부터 얻은 일반화 오차 추정치를 평균함으로써 최종 일반화 오차 추정치를 계산한다. $k=K$인 일반적인 경우의 *k*-fold 교차-타당화 방법을 공식으로 나타내면 다음과 같다.

$$\hat{\Delta}\left(\hat{\theta}_k^{(-F_j)}\right)=\frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{n_{v(j)}}\sum_{i\in F_j}\left\{\log g_k\left(x_i,\hat{\theta}_k^{(-F_j)}\right)\right\}\right]$$

여기서 F_j는 j번째 *fold*, $n_v(j)$는 F_j의 크기를 나타내며, $\hat{\theta}_k^{(-F_j)}$는 F_j를 제외한 나머지 부분을 훈련 세트로 사용한 파라미터 추정치를 나타낸다. 아래첨자 v는 타당화 세트로 사용된다는 의미로 사용되었다.

Akaike Information Criterion (AIC)[22)]

많은 심리학 연구자들에게도 AIC(Akaike, 1973; 1974)는 매우 익숙한 지수이며 모형 비교를 위해 사용하는 방법도 잘 알려져 있다. 그러나, AIC 역시 일반화 오차에 대한 추정량(estimator)으로서 제안되었으며, 그 목표값(target quantity)이 정확히 무엇인지를 알고 있는 심리학 연구자는 많지 않다. AIC가 추정하고자 하는 목표값은 아래와 같이 표현될 수 있다(Kuha, 2004).

$$T_{AIC}=-E_X\left\{E_Y\left[\log p(Y|\hat{\theta}_k^X,M_k)|X\right]\right\}$$

위의 목표값이 나타내는 것을 단계별로 풀어서 설명하면 다음과 같다. 먼저, 주어진 훈련 세트 X를 기초로 모형의 파라미터를 추정($\hat{\theta}_k^X$)하고, 추정된 모형으로부터 새로운 데이터 세트 Y를 관찰할 로그 확률의 기댓값을 구한다. 이 기댓값은 일반화 오차를 나타낸다. 이렇게 얻은 일반화 오차는 새로운 훈련 세트 X를 얻을 때마다 다시 계산을 할 수 있는데, 가능한 모든 훈련 세트를 대상으로 계산한 일반화 오차 분포의 평균이 바로 T_{AIC} 이다. 따라서, AIC가 추정하고자 하는 목표값은 일반화 오차의 기댓값이라 할 수 있다.

AIC는, 교차-타당도 접근법과는 달리, 주어

22) 많은 심리학 연구자들은 AIC뿐만 아니라 BIC 지수에 대해서도 익숙할 것이다. AIC와 마찬가지로 BIC 역시 훈련 오차, 표본 크기, 모형에 지정된 파라미터의 개수를 이용하여 계산이 가능하다. 따라서, BIC는 일반화 오차를 추정하는 또 하나의 공식으로 이해할 수도 있고 실제 그러한 방식으로 사용되고 있다. 그러나 보다 정확하게 이야기하면 BIC가 추정하는 목표값(target quantity)은 (로그) 베이즈 요인(Bayes Factor; Schwarz, 1978; Raftery, 1995)이며, 새로운 사례 Y에 대한 예측 정확도를 계산하는 개념은 부재하다는 논의(Kuha, 2004)에 기초하여, 본 논문에서는 다루지 않기로 하였다. AIC와 BIC를 다각도에서 종합적으로 비교한 Vrieze (2012)와 Burnham & Anderson(2004)을 참고하기 바란다.

진 데이터 세트 전체를 기초로 일반화 오차를 추정할 수 있도록 해 주는 추정량이다. 우리가 잘 알고 있듯이, AIC는 음의 로그-가능도 최솟값으로 표현되는 훈련 오차와 모형에 지정된 파라미터의 개수만 알면 쉽게 구할 수 있다.

AIC 역시 일반화 오차를 추정하기 위한 추정량으로서 편향 오차와 분산 오차를 피할 수 없으며, 편향-분산 균형의 관점에서 AIC보다 우수한 추정량을 도출하기 위한 다양한 시도가 있어 왔다(Takeuchi, 1976; Hurvish & Tsi, 1989; Burnham & Anderson, 1998; Bozdogan, 2000). 그러나 이들 지수는 모두 주어진 데이터 세트를 분할하지 않고 전체 자료를 대상으로 얻은 훈련 오차를 교정(correction)함으로써 일반화 오차(의 기댓값)를 추정한다는 공통점이 있다. 즉, AIC 및 관련 지수들은 교차-타당화 검증법과 일반화 오차를 추정한다는 목표는 동일하지만, 주어진 데이터를 활용하는 방법에 있어서 차이를 보인다고 할 수 있다.

심리학 연구 모형의 평가: 논의 그리고 당부

본 논문에서는 수리 및 계량 심리학 분야에서 논의되어 왔던 모형 추정 및 평가의 원칙을 소개하고, 기계 학습 분야에서의 논의와 관련지어 설명하였다. 이를 통해 심리학 연구 모형의 평가 원칙과 기계 학습 모형의 평가 원칙은 예측 오차 혹은 일반화 오차를 기준으로 한다는 점에서 서로 다르지 않음을 설명하였다. 또한 심리학 연구자들이 훈련 데이터를 암기하는 결과를 가져오는 "철저한" 데이터 분석 관행에서 벗어나, 일반화 오차를 직접 평가할 수 있는 두 가지 방법을 소개하였다. 일반화 오차 혹은 총체적 불일치를 최소화 하는 모형을 선정하기 위한 통계적 방법을 모형 선정 기준 (model selection criterion) 혹은 줄여서 기준(criterion)이라 부르기도 한다. 지금까지 통계학 및 수리/계량 심리학 분야에서는 다양한 종류의 모형 선정 기준이 제안되어 왔는데, 대부분은 앞서 언급하였던 이상적인 상황에서의 3-단계 전략을 주어진 모델링 환경에서 어떻게 구현할 것인가에 따라 구분된다고 할 수 있다. 그 중 기계 학습 분야에서 가장 널리 사용되는 방법인 교차-타당성 입증법(Miller et al., 2016; Chapman et al., 2016; Hastie, et al., 2009)과 AIC를 일반화 오차의 추정량이라는 관점에서 소개하였다.

본 논문에서는 또한, 빅-데이터와 기계 학습으로 요약되는 연구 환경의 변화를 맞이하여, 기계 학습 분야에서 정립된 모형 평가 기준인 일반화 오차 혹은 예측 오차에 대해 상술하였고, 수리 및 계량 심리학 분야에서 진행되어왔던 관련 개념과 비교하였다. 빅-데이터의 시대 심리학 분야에도 급격히 늘어나고 있는 기계 학습 모형과 심리적 구성 개념을 다루는 구조방정식과 같은 잠재 변인 모형의 평가 기준은 과거에도 동일했고 지금도 동일하다. 기계 학습의 일반화 오차, 예측 오차, 시험 오차와 수리/계량 심리학의 총체적 오차는 동일한 개념이다. 기계 학습 분야 서적 맨 앞부분에서 항상 등장하는 편향-분산 균형(bias-variance trade-off)는 수리/계량 심리학 분야에서 강조해 왔던 모형 평가의 원칙, 즉 모형이 복잡해질수록 근사오차는 줄어들지만 추정오차는 증가할 수 있으므로, 간명하고(parsimonious) 해석 가능한(interpretable) 모형을

선정하는 것이 바람직하다는 원칙과 동일한 개념이다. 이러한 논의는 모형 평가 기준에 대한 이해뿐만 아니라, 기계 학습 분야의 생소한 용어를 이해하는 데에도 도움이 되었길 기대해 본다.

탐색과 확인

현재 많은 심리학자들이 모형을 추정하고 평가하여 최종 모형을 선정하는 과정을 보면, 상당히 철저하게 데이터를 분석하면서도 또한 매우 "기계적인"(mechanical) 결정을 내리는 경우가 많다. 회귀분석을 예로 들면, 추정된 모형의 결정 계수(coefficient of determination)과 회귀 계수의 통계적 유의미성을 확인한 뒤, 유의미한 회귀계수를 중심으로 설명 변수와 종속 변수 간 관계를 설명한다. 이 과정에서 이상값(outliers)를 제거하기도 하고 그렇지 않기도 한다. 하위 집단별 분석을 통해 유의미한 회귀 계수를 발견하기도 하고, 이에 따라 상호작용 분석을 실시하기도 한다. 구조방정식 모형 역시 다양한 적합도 지수를 이용하여 지수별로 마련된 기준값과 비교하여 모형이 적합하다고 판단(Hu et al., 1999; Shi et al., 2019) 되면 해당 모형에 지정된 파라미터 추정치를 결과를 해석한다. 그런데, 이렇듯 "철저한" 분석 과정을 거치는 것에 무슨 문제가 있다는 것이며, 왜 HARKing, p-hacking, 또는 QRP 등의 불명예스러운 이름을 붙이는 것인가?

"철저한" 데이터 분석을 통한 통계적 추론이 여러 가지 불명예스러운 이름으로 불리고 있는 이유는 통계적으로 유의미한 결과(예: p-value < .05)를 얻을 목적으로 남용되어왔기 때문이다. 즉, 주어진 데이터의 통계적 특징을 여러 가지 방식을 동원하여 탐색한 후 그 결과를 기반으로 도출한 가설을 마치 이론적으로 도출한 것처럼 간주하고, 다시 주어진 데이터를 재사용하여 가설을 검정하는 것은 잘못된 관행이다(Wiggins & Christopherson, 2019). 이러한 관행은 탐색적 분석(exploratoration)과 확인적 분석(confirmation)을 혼동함으로써 발생하며 결과적으로 제 1 종 오류 확률을 증가시키는 문제를 가져온다.

탐색적 데이터 분석(Tukey, 1977)은 통계적 추론 과정에서 매우 중요한 위치를 차지하며, 현대 데이터 과학(data science) 분야에서는 그 유용성이 더욱 커지고 있다. 그러나 데이터 탐색의 목적과 확인적 분석의 목적은 구분되어야 한다. Exploratory Data Analysis(Tukey 1977)의 저자인 Tukey의 비유에 따르면, 탐색적 분석은 수사(detective)의 성격을 보이는 반면, 확인적 분석은 재판(judicial)의 성격을 가진다. 탐색적 분석으로부터 얻은 통찰은 반드시 새로운 데이터에서 확인되어야 한다(Shrout & Rodgers, 2018).

조건부 추론

계량 심리학자의 관점에서 본 심리학 분야 재현 위기 논란은 모형의 과적합 문제와 밀접한 관련이 있다. 이 문제는 주어진 데이터에 대한 "철저한" 분석을 통해 최종 선정된 모형이 참모형이라는 암묵적 가정을 바탕으로 *p*-value, 신뢰 구간, 효과 크기 등에 대한 후속 추론을 진행하는 일명 조건부 추론(conditional inference)의 문제로도 알려져 있다(Chatfield, 1995; Lubke & Campbell, 2016; Roca & Yarkoni, 2020). 조건부 추론에서는 추정된 모형이 참모

형이 아닐 가능성은 원천적으로 배제되므로, 후속 추론은 지나치게 낙관적인 경향을 보인다. 예를 들어, 단순 선형 회귀 분석의 회귀계수가 유의미하고 표준화 회귀 계수가 0.3일 경우, 설명 변수와 종속 변수간 관계는 중간 정도의 효과 크기가 존재하는 것으로 판단한다. 그러나 이는 단순 선형 회귀 분석이 참모형이 아닐 경우 모형 추정 결과가 달라질 수 있다는 가능성을 배제한 채 도출한 얻은 결과이며, 따라서 새로운 데이터를 분석하는 후속 연구에서는 동일한 결과를 얻지 못할 수 있다는 가능성을 배제한 매우 낙관적인 해석이다. 최근 계량 심리학 분야에서는 회귀분석(Agler et al., 2020; Jones et al., 2016; Waller et al., 2008)과 구조방정식 모델링(Lee et al., 2018; Pek et al., 2018; Prendez et al., 2019)등의 분야에서 훈련 오차를 최소화하는 방식의 추론이 초래할 수 있는 과적합 문제와 조건부 추론의 위험성을 예증함으로써 현재 심리학 분야에서 자리 잡고 있는 통계적 추론 과정에 대한 개선의 필요성을 논의하였다. 그럼에도 불구하고 수리/계량 심리학 분야에서 꾸준히 지속된 논의는 심리학 연구자들에게 큰 관심을 끌지 못한 것이 사실이다. 이는 현재에도 많은 심리학 연구자들은 주어진 데이터를 기반으로 훈련 오차를 최소화하는, 그래서 과적합 문제에서 자유롭지 못한 모형과 씨름하고 있음을 의미한다.

파라미터 불확실성과 모형 평가

사실, 수리/계량 심리학 분야에서는 주어진 데이터의 "철처한" 분석 혹은 조건부 추론의 관행이 가져올 수 있는 문제점과 해결 방안에 대해 오래전부터 논의를 진행해 왔다. 특히, Myung et al (2000)은 과적합 문제 방지의 필요성과 일반화 가능한 모형 선정을 위한 다양한 방법을 다각도에서 논의하였다. 이러한 논의에서는 공통적으로 모형에 지정된 파라미터는 미지수라는 점을 인정하고 모형의 성능에 대한 평가를 진행한다. 즉, 주어진 훈련 세트로부터 추정된 파라미터 값만을 이용하는 것이 아니라, 파라미터 공간(parameter space)에서 얻을 수 있는 가능한 모든 값들을 이용하여[23] 모형을 평가한다. 이와 같은 관점에서 모형이란, 파라미터 추정치의 분포를 통합적으로 고려한 분포의 모임(a family of distributions)으로 정의하는 것이 더 정확할 수 있다. 그렇다면, 더 복잡한 모형은 더 많은 종류/범위의, 더 다양한 파라미터 값을 가질 수 있고 따라서, 새로운 자료와 불일치 정도가 큰 사례를 생성할 확률도 높아진다. 반면 보다 단순한 모형은 파라미터의 수가 더 적고, 개별 파라미터가 가질 수 있는 값의 범위도 상대적으로 더 좁지만, 모형이 생성하는 사례와 새로운 자료와 불일치 정도는 더 낮을 수 있다. 추정된 모형, 더 정확하게는, 추정된 분포의 모임으로부터 생성된 "좋은" 사례들과 "나쁜" 사례들의 균형이 이루어지는 지점에서 적절한 수준의 모형 복잡도가 정해진다고 이해할 수 있다. 본 논문에서는 후보가 되는 근사 모형의 복잡도 수준을 일반화 오차라는 측면에서 추정할 수 있는 방법으로서 교차-타당도 입증법과 AIC를 소개하였다.

추정 방법의 한계와 연구자의 역할

23) 파라미터 추정치의 표집 분포를 이용하거나, 사전 혹은 사후 확률 분포를 이용하는 등 다양한 방법이 있을 수 있다.

교차-타당도 검증법의 기본적인 아이디어는, 연구자가 수집한 데이터 세트(X)를 추정 표본(calibration sample, X_C)과 타당화 표본(validation sample, X_V)으로 나누어, 추정 표본은 모형을 적합 혹은 추정하는 데 사용하고, 타당화 표본은 모형의 성능을 평가하는 데 활용하는 것이다(Browne, 2000; Arlot et al. 2010). 이러한 아이디어는 빅-데이터 시대 일반화 오차를 추정하는 데 활용 가능한 3-단계 전략과 직접적으로 연관됨을 알 수 있다. 그러나 한 가지 유의할 점은, 교차-타당성 검증법 (혹은 AIC) 역시 일반화 오차에 대한 추정량으로서 편향 오차와 분산 오차를 가진다. 예를 들면, LOOCV는 편향 오차는 낮으나 분산 오차가 큰 경향이 있고, 반대로 k-fold 교차-타당성 검증법은 편향 더 높으나 분산이 더 낮은 경향을 보인다. 편향-분산 균형의 관점에서 $k = 5$ 혹은 $k = 10$을 사용하는 것이 일반적이며 이 방법 외에도 일반화 오차를 보다 정확하게 추정하기 위해 보다 개선된 교차-타당도 검증 방법이 지속적으로 제안되고 있다(Krstajic et al., 2014; Kuhn & Johnson, 2013; Arlot et al., 2010; Cauley et al., 2010; Varma & Simon, 2006). 그러나 여러 가지 방법들의 기저에 깔린 아이디어와 그 목적은 동일하다. 즉, 훈련 세트에만 과도하게 적합한 모형이 아니라 새로운 사례에 대해서도 우수한 성능을 보이는 모형을 선정할 수 있는 기준을 마련하는 것을 목표로 주어진 데이터 세트를 효율적으로 분할하여 재사용하는 것이 교차-타당도 검증법의 기본적인 아이디어이다.

본 논문에서 일반화 오차 추정량으로서 소개한 AIC, 교차-타당도 검증법 등은 다른 추정량과 마찬가지로 그 추정의 정확성을 주어진 데이터가 제공하는 정보량에 의존할 수밖에 없다. 이는 결과적으로 추정된 모형 성능과 모형별 성능의 순위는 일반화 오차의 참값 순위와 다를 가능성도 배제할 수 없다는 것을 의미한다. 다만, 이러한 문제가 발생할 가능성은 데이터의 크기가 커질수록, 후보 모형의 수가 적을수록 줄어든다. 이는 다시 말하면, 빅-데이터의 시대, 보다 안정적인 추정 결과를 얻을 수 있는 가능성은 높아졌지만, 후보가 되는 근사 모형의 수가 지나치게 많을 경우 교차-타당도 검증법과 같은 모형 선정 기준만을 사용한 "자동적"이고 "기계적" 판단은 일반화 오차가 가장 낮은 후보 모형이 선정되지 못하는 선정 편향(selection bias)을 피할 수 없게 된다(Chatfield, 1995; Bozdogan, 2000). 따라서, 심리학 연구자들은 해당 분야의 사전 연구 결과나 이론적 근거를 바탕으로 소수의 그럴듯한 후보 모형을 추려내는 데 시간과 노력을 들여야 할 것이다.

예측과 설명

기계 학습 모형은 미래의 결과를 예측하는 데 유용한 예측 모형(predictive modeling)으로 잘 알려져 있고, 심리학 분야의 일각에서도 연구 문제 자체를 설명 중심에서 예측 중심으로 변경할 때가 왔다고 주장하기도 한다(Yarkoni & Westfall, 2017). 물론 인간의 성격, 태도, 행동, 정서 등을 예측하는 시스템을 구축하기 위한 예측 중심의 연구 문제를 설정하고, 이러한 종속 변수들에 대한 예측 정확성(predictive accuracy)이라는 관점에서 주요 변수를 선정하고 최종 모형을 결정하여 예측에 활용하는 것도 매우 중요하다. 마찬가지로, 구성 개념들이 어떠한 기제를 통해 서로 유기적으

로 연결되어 있는지를 경험적으로 확인함으로써 인간의 행동과 태도를 설명하고자 하는 연구 문제 역시 매우 중요하며 앞으로도 그러할 것이다. 따라서, 인간 행동과 정서에 대해 간명하고 해석 가능한 설명을 원하는 심리학 연구자는 현재 사용하고 있는 모형을 버리고 예측 모형 사용자로 전향(convert)할 필요는 전혀 없다고 생각한다. 다만, 설명 중심의 연구 주제를 모형화(modeling)하고 해당 모형을 평가할 때에도, 훈련 데이터에서뿐만 아니라 새로운 데이터, 그리고 미래의 데이터에서도 타당한 설명이 재현될 수 있도록 그에 걸맞는 모형 평가 방법을 사용하는 것이 바람직할 것이다.

온고지신(溫故知新)

마지막으로 훈련 오차가 아니라 일반화 오차를 최소화 하는 것, 그래서 재현 가능한 연구 결과를 얻는 것이 더욱 중요하며, 이를 위해서 교차-타당성 검증법을 사용할 것을 제안하는 논의가 심리학 분야에서도 이미 오래전부터 진행되어 왔다는 점을 언급하고자 한다(Mosier, 1951; Wherry, 1951; 1975). 특히, Mosier(1951)는 Symposium: The need and means of cross-validation. Problems and designs of cross-validation이라는 논문에서 회귀 계수 추정치의 유효성(effectiveness)은 독립적으로 얻은 새로운 표본에서 결정되어야 한다는 것을 분명히 밝히는 것으로 논의를 시작하고 있다. 출판 년도를 보면 적어도 1950년대에 이미 일반화 가능한 결과를 얻기 위한 방법론에 대한 논의가 심리학 분야에서 진행이 된 것을 알 수 있으며 그 방법론 또한 현재 기계 학습 분야에서 활용되고 있는 교차-타당화 방법과 기본 원리에 있어 차이가 없다. 이후에도 회귀 분석과 구조방정식 분야에 교차-타당도 검증법을 적용하기 위한 논의가 계량 심리학 분야에서 지속적으로 이루어져 왔다(Cudeck & Browne, 1983; Browne & Cudeck, 1989; Browne, 2000). 현재 심리학 연구에서 잘 보고되지 않지만, 회귀 분석 모형의 평가 지표인 Mallow's C_p(Mallow, 1973; Browne, 2000), 요인분석과 구조방정식 모형의 평가 지표인 ECVI (expected cross-validation index; Browne & Cudeck, 1989) 등이 그러한 노력의 결과물이다.

심리학 분야에서 아주 오래전 진행되었던 모형 평가 기준에 대한 논의(Mosier, 1951, p.5)의 원문을 인용하면서 본 논문을 마무리하고자 한다.

"The term "cross-validation" is often loosely applied to any one of several distinct, though closely related, experimental designs.....if the combining weights of a set of predictors have been determined from the statistics of one sample, the effectiveness of the predictor-composite must be determined on a separate, independent sample..."

온고지신(溫故知新)이라는 구절이 참으로 새삼스럽다.

참고문헌

Agler, R. A., & De Boeck, P. (2020). Factors associated with sensitive regression weights: A fungible parameter approach. Behavior research methods, 52(1), 207-22. https://doi.org/10.3758/s13428-019-01220-6

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., Csaki, F. (Eds.), 2nd International Symposium on Information Theory. Akademia Kiado: Budapest.

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. technometrics, 16(1), 125-127. https://doi.org/10.1080/00401706.1974.10489157

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79. https://doi.org/10.1214/09-SS054

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. Wadsworth Statistics/ Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. Journal of mathematical psychology, 44(1), 62-91. https://doi.org/10.1006/jmps.1999.1277

Browne, M. W. (2000). Cross-validation methods. Journal of mathematical psychology, 44(1), 108-132. https://doi.org/10.1006/jmps.1999.1279

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. Multivariate behavioral research, 24(4), 445-455. https://doi.org/10.1207/s15327906mbr2404_4

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136-162). Newbury Park, CA: Sage.

Burnham, K. P., & Anderson, D. R. (1998). Model Selection and Inference: A Practical Information-Theoretic Approach. New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research, 33(2), 261-304. https://doi.org/10.1177/0049124104268644

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate behavioral research, 1(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. Psychological methods, 21(4), 603. https://doi.org/10.1037/met0000088

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society: Series A (Statistics in Society), 158(3), 419-444.. https://doi.org/10.2307/2983440

Chung, H. Y., Lee, K. W., & Koo, J. Y. (1996). A note on bootstrap model selection criterion. Statistics & probability letters, 26(1), 35-41.

https://doi.org/10.1016/0167-7152(94)00249-5

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. Multivariate Behavioral Research, 18(2), 147-167. https://doi.org/10.1207/s15327906mbr1802_2

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. Psychological Bulletin, 109(3), 512-519. https://doi.org/10.1037/0033-2909.109.3.512

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structura equation models with multiply imputed data. Psychological Methods, 23(1), 76 – 93. https://doi.org/10.1037/met0000102

Geisser, S. (1975). The predictive sample reuse method with applications. Journal of the American statistical Association, 70(350), 320-328. https://doi.org/10.1080/01621459.1979.10481632

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal, 6(1), 1-55. https://doi.org/10.1080/10705519909540118

Hurvich, C. M., 6 Tsai, C. L. (1989). Regression and time series model selection in small samples. Biometrika, 76, 297-307. https://doi.org/10.1093/biomet/76.2.297

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning.. New York: springer.

Jones, J. A., & Waller, N. G. (2016). Fungible weights in logistic regression. Psychological Methods, 21(2), 241-260. https://doi.org/10.1037/met0000060

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. Personality and social psychology review, 2(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Kim, C. (2019). Studying psychology using big data, Korean Journal of Psychology: General 38(4), 519-548. http://dx.doi.org/10.22257/kjp.2019.12.38.4.519

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Sowden, W. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science, 1(4), 443-490. https://doi.org/10.1177/2515245918810225

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... & Nosek, B. (2014). Data from investigating variation in replicability: A “many labs” replication

project. Journal of Open Psychology Data, 2(1). http://doi.org/10.5334/jopd.ad

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of cheminformatics, 6(1), 1-15. https://doi.org/10.1186/1758-2946-6-10

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. Sociological methods & research, 33(2), 188-229. https://doi.org/10.1177/0049124103262065

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York: Springer.

Lee, T., & MacCallum, R. C. (2015). Parameter influence in structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal, 22(1), 102-114.. https://doi.org/10.1080/10705511.2014.935255

Lee, T., MacCallum, R. C., & Browne, M. W. (2018). Fungible parameter estimates in structural equation modeling. Psychological Methods, 23(1), 58-75. https://doi.org/10.1037/met0000130

Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. Psychological Methods, 26(4), 466-485. https://doi.org/10.1037/met0000381

Lubke, G. H., & Campbell, I. (2016). Inference based on the best-fitting model can contribute to the replication crisis: Assessing model selection uncertainty using a bootstrap approach. Structural equation modeling: a multidisciplinary journal, 23(4), 479-490. https://doi.org/10.1080/10705511.2016.1141355

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. Psychological Bulletin, 109(3), 502-511. https://doi.org/10.1037/0033-2909.109.3.502

Mallow, C. L. (1973). Some comments on Cp. *Technometrics*, 28, 313-319.

Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. Psychological Methods, 21(4), 583-602. https://doi.org/10.1037/met0000087

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill, New York.

Mosier, C. I. (1951). Symposium: The need and means of cross-validation. I. Problems and designs of cross-validation. Educational and Psychological Measurement, 11(1), 5-11. https://doi.org/10.1177/001316445101100101

Linhart, H. & Zucchini, W. (1986). Model selection. New York: Wiley.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). GUEST EDITORS'INTRODUCTION: special issue on model selection.Journal of mathematical psychology, 44(1), 1-2. https://doi.org/10.1006/jmps.1999.1273

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural equation modeling: A multidisciplinary Journal, 14(4), 535-569.

https://doi.org/10.1080/10705510701575396

Pek, J., & Wu, H. (2018). Parameter uncertainty in structural equation models: Confidence sets and fungible estimates. Psychological Methods, 23(4), 635-653. http://dx.doi.org/10.1037/met0000163

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. Psychological methods, 17(1), 1.

Prendez, J. Y., & Harring, J. R. (2019). Measuring Parameter Uncertainty by Identifying Fungible Estimates in SEM. Structural Equation Modeling: A Multidisciplinary Journal, 26(6), 893-904. https://doi.org/10.1080/10705511.2019.1608550

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). Numerical Recipes with Source Code CD-ROM 3rd Edition: The Art of Scientific Computing. Cambridge University Press.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. Sociological Methodology, 25, 111-163. https://doi.org/10.2307/271063

Rocca, R., & Yarkoni, T. (2020, November 12). Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction, https://doi.org/10.31234/osf.io/e437b

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464. http://www.jstor.org/stable/2958889

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. Educational and psychological measurement, 79(2), 310-334. https://doi.org/10.1177/0013164418783530

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science, 22(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. Annual review of psychology, 69, 487-510. https://doi.org/10.1146/annurev-psych-122216-011845

Stone, M. (1974). Cross validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological), 36(2), 111-133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Stone, M. (1977). An asymptotic equivalence of choice of model by cross validation and Akaike's criterion. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 44-47. https://doi.org/10.1111/j.2517-6161.1977.tb01603.x

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. Suri-Kagaku (Mathematical Sciences) 153 12-18.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley..

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for

model selection. BMC bioinformatics, 7(1), 1-8. https://doi.org/10.1186/1471-2105-7-91

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological Methods, 17(2), 228-243. https://doi.org/10.1037/a0027127

Waller, N. G. (2008). Fungible weights in multiple regression. Psychometrika, 73(4), 691-703. https://doi.org/10.1007/S11336-008-9066-Z

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". https://doi.org/10.1080/00031305.2019.1583913

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. Journal of Theoretical and Philosophical Psychology, 39(4), 202-217. https://doi.org/10.1037/teo0000137

Wherry, R. J. (1951). IV. Comparison of cross-validation with statistical inference of betas and multiple R from a single sample. Educational and Psychological Measurement, 11(1), 23-28. https://doi.org/10.1177/001316445101100104

Wherry, R. J. (1975). Underprediction from overfitting: 45 years of shrinkage. Personnel Psychology, 28(1), 1-18. https://doi.org/10.1111/j.1744-6570.1975.tb00387.x

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100-1122. https://doi.org/10.1177/1745691617693393

Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. Sociological Methodology, 38, 329-368. https://doi.org/10.1111/j.1467-9531.2008.00198.x

Zucchini, W. (2000). An introduction to model selection. Journal of mathematical psychology, 44(1), 41-61. https://doi.org/10.1006/jmps.1999.1276

1차원고접수 : 2021. 11. 06.
최종게재결정 : 2021. 11. 26.

한국심리학회지: 일반
Korean Journal of Psychology: General
2021, Vol. 40, No. 4, 389-413
http://dx.doi.org/10.22257/kjp.2021.12.40.4.389

Principles and methods for model assessment in psychological research in the era of big-data and machine learning

Taehun Lee

Department of Psychology, Chung-Ang University

The objective of the present article is to explain principles of estimation and assessment for statistical models in psychological research. The principles have indeed been actively discussed over the past few decades in the field of mathematical and quantitative psychology. The essence of the discussion is as follows: 1) candidate models are to be considered not the true model but approximating models, 2) discrepancy between a candidate model and the true model will not disappear even in the population, and therefore 3) it would be best to select the approximating model exhibiting the smallest discrepancy with the true model. The discrepancy between the true model and a candidate model estimated in the sample has been referred to as overall discrepancy in quantitative psychology. In the field of machine learning, models are assessed in light of the extent to which performance of a model is generalizable to the new unseen samples, without being limited to the training samples. In machine learning, a model's ability to generalize is referred to as the generalization error or prediction error. The present article elucidates the point that the principle of model assessment based on overall discrepancy advocated in quantitative psychology is identical to the model assessment principle based on generalization/prediction error firmly adopted in machine learning. Another objective of the present article is to help readers appreciate the fact that questionable data analytic practices widely tolerated in psychology, such as HARKing (Kerr, 1998) and QRP (Simmons et al., 2011), have been likely causes of the problem known as overfitting in individual studies, which in turn, have collectively resulted in the recent debates over replication crisis in psychology. As a remedy against the questionable practices, this article reintroduces cross-validation methods, whose initial discussion dates back at least to the 1950s in psychology (Mosier, 1951), by couching them in terms of estimators of the generalization/prediction error in the hope of reducing the overfitting problems in psychological research.

Key words : overfitting, generalization error, training error, cross-validation, bias-variance tradeoff