http://www.jsebs.org
pISSN 2288-3908
eISSN 2765-3846

The Journal of Society for e-Business Studies
Vol.27, No.1, February 2022, pp.81-94
https://doi.org/10.7838/jsebs.2022.27.1.081

순서형 프로빗 모델을 활용한 정성적 평가 결과의 객관성 확보방안

Securing Objectivity of Qualitative Assessment Results using Ordered Probit Model

정민규(Minkyu Jeong)*, 강윤철(Yuncheol Kang)**
김남철(Namchul Kim)***, 정광헌(Kwanghun Chung)****

초 록

서비스 산업에서 널리 사용되는 설문조사 형태의 정성적 평가방법은 서비스 품질을 평가하는 주요한 평가 수단으로 활용되는데, 이때 설문조사 결과에 응답자의 주관적 판단이 개입될 수 있는 여지가 존재한다. 본 연구에서는 이러한 정성적 평가 결과에 포함될 수 있는 주관적 판단을 최대한 배제하고 객관성을 확보하기 위한 방안을 제시한다. 특히 같은 종류의 정성적 평가도구가 여러 사람에 의해 반복적으로 사용되는 상황에서의 객관성 확보 방안에 대해 연구하였다. 이를 위해, Ordered Probit 모델 및 제3자 평가 결과를 함께 활용하여 주관성 개입 여부를 판단하고, 주관성을 보정한 후, 최종 결과에 대해 통계적으로 검증하는 절차를 수행하였다. 본 연구에서 분석한 어플리케이션은 의료서비스 분야이며, 특히 특정 의료서비스 제공 환경에서 해당 서비스 제공자가 자가 진단한 실제 평가 결과를 본 연구에서 제안하는 방법론을 통해 어떻게 객관성이 확보될 수 있는지 구체적으로 설명하였다.

ABSTRACT

In the service sectors, the qualitative evaluation method in the form of a survey is widely used as a major assessment tool to evaluate the quality of service. However, the results obtained from a survey can involve the subjective judgment of the respondent. In this study, we propose a method to secure objectivity by excluding subjectivity that may be included in the qualitative evaluation results. In particular, we deal with a situation where the same type of qualitative evaluation tool is used repeatedly by several service providers. To this end, by utilizing both the Ordered Probit model and third-party evaluation results, we determine whether subjectivity is involved in the results. After correcting subjectivity, the final results are obtained through statistical analysis. The application analyzed in this study is the medical service area. With the actual evaluation results supplied by the service providers, we explain how objectivity can be secured from the assessment data by applying our proposed approach.

키워드 : 순서형 프로빗, 주관적 평가, 객관성
Ordered Probit, Subjective Assessment, Objectivity

This work was supported by 2019 Hongik University Research Fund.
* First Author, Department of Industrial and Systems Engineering, KAIST(minkyujeong@kaist.ac.kr)
** Co-Author, College of Business Administration, Ewha Womans University(yckang@ewha.ac.kr)
*** Co-Author, 365mc Networks(namchul.kim@365mc.com)
**** Corresponding Author, College of Business Administration, Hongik University(khchung@hongik.ac.kr)
Received: 2022-02-03, Review completed: 2022-02-11, Accepted: 2022-02-17

1. 서 론

일반적으로 서비스 평가에서 주로 사용하는 설문조사에서는 해당 서비스를 제공받은 고객 본인의 경험을 토대로 설문이 수행된다. 해당 설문의 목적은 대부분 제공한 서비스의 품질을 측정하고, 고객들의 설문 조사 결과를 바탕으로 더 나은 품질의 서비스를 제공하고자 하는 것이다. 이때, 고객들의 설문조사 결과에는 제공받은 서비스에 대한 결과 외에도 추가적으로 서비스와 상관없는 고객의 주관적 성향이 개입될 수 있다. 예를 들어, 설문조사 시 응답자의 개인 성향에 따라 Likert Scale 기준으로 주로 극단적인 값(Extreme Response)을 선택하는 경향이 있거나, 혹은 주로 중간값(Midpoint Response)만을 선택하는 경향이 있는 게 대표적이다[11]. 이러한 개인의 주관적 성향은 설문조사 결과를 왜곡하며, 잘못된 서비스 품질 측정으로 이어질 수 있는 위험이 있다. 설문조사 결과의 주관성 개입 문제는 현실적으로 피하기 힘든 부분일 수 있으나, 연구자 입장에서는 최대한 주관성이 배제된 결과가 언제나 바람직하다. 특히, 해당 평가 도구의 결과가 차후 다른 분석의 입력 값으로 사용될 경우 주관성 개입으로 인한 왜곡문제가 계속 전파될 수 있는 위험이 존재한다.

한편, 의료 현장에서 사용되는 다양한 평가 도구(Diagnostic Assessment)는 의료서비스 제공자가 환자의 상태를 진단하거나, 제공된 의료서비스의 결과를 측정하는데 널리 사용되고 있다. Mental Healthcare의 예를 들면, 치매 질환의 정도를 측정하는데 MMSE(Mini Mental State Exam)와 같은 평가 도구를 사용하여 환자의 치매 질환 상태를 측정한다[3]. 수술 후 자기 평가(Self-assessment)의 경우 집도 된 수술이 전반적으로 어떠한지, 그리고 수술 예후의 결과가 어떻게 될지에 대해 의료서비스 제공자들(예를 들면, 집도의)의 정성적 기준에 근거하여 해당 수술 서비스를 평가한다. 이때, 이러한 평가 도구들 역시 앞서 언급한 설문조사의 부류에 속하며, 설문조사 결과가 태생적으로 갖게 되는 개인의 주관적 경향성 개입 문제가 발생할 수 있다. 예를 들면, 집도의 개인의 평가 성향에 따라 자신의 수술 결과를 낙관적으로 혹은 보수적으로 평가하여 실제 수술 결과를 왜곡하는 문제가 발생할 수 있다.

본 연구에서는 위에서 기술한 응답의 주관성 개입 문제, 즉 주관적 판단이 개입되는 평가(Assessment)의 경우 해당 결과값을 어떻게 최대한 객관적인 시각에서 해석할 수 있는지에 대한 방법을 모색해 본다. 특히, 하나의 평가 도구가 여러 사람에 의해 반복적으로 사용되는 경우가 존재할 수 있는데, 이 상황에서 결과값의 객관화 방법을 제안하고자 한다. 구체적으로는 여러 사람이 평가한 데이터를 모아 분포를 만들고, 이를 통해 개별 평가 응답 내용의 주관적 척도를 구하여 보정하는 방식을 고려한다. 평가한 데이터를 모아 분포를 만들고, 이를 통해 객관화하는 모델로 Ordered Probit 모델을 활용하였으며[1], 관련 통계 검정을 통해 해당 모델의 타당성을 검증하였다. 적용 예로, Medical Staff 들의 수술 후 결과에 대한 평가 데이터를 이용하여 Ordered Probit 모델을 만들고, 비슷한 상황에서 수술한 환자의 경우에 각 medical staff 별로 결과값이 어떻게 달라질 수 있는지에 대해 계산하고, 주관적 성향의 개입 여부와 이를 보정한 후의 결과에 대해 분석 및 검증하였다.

본 논문의 구성은 다음과 같다. 먼저, 제2장에서는 주관적 평가에 대한 분석 이론과 본 연구에서 주로 사용하게 될 Ordered Probit 모델에 대해 기술한다. 제3장에서는 본 연구에서 활용된 예인 수술 프로세스 및 평가 프로세스에 대해 설명하고, 이에 적합한 모델을 Ordered Probit 방법론을 활용하여 모델링한다. 다음으로 제4장에서는 해당 모델을 실제 데이터를 활용하여 개발하고, 모델에서 사용되는 파라미터에 대한 의미를 살펴본 뒤, 몇 가지 통계 검정에 기초하여 해당 모델을 검증하는 작업을 수행한다. 마지막으로, 제5장에서는 본 연구결과가 나타내는 의미를 기술하고, 한계점 및 추후 연구 방향에 대해 논의한다.

2. 기존 방법론 연구

주관적 평가에 관한 객관성 보정에 대한 대부분의 기존 이론은 크게 두 가지로 나눌 수 있다. 우선 객관성 극대화를 위해 심사자들의 평가 문항 자체를 보완하여, 최대한 객관화된 결과를 도출할 수 있도록 만드는 방법이 있다[10]. Item Response Theory(문항 반응 이론)이라고도 불리는 이러한 방법론은 응답자의 각 문항에 대한 응답 특성을 대상으로, 실제 반응해야 하는 응답(진실값)과 실제로 반응한 응답 간의 차이를 확률분포를 활용하여 설명한다. 다른 방법론으로는 평가가 완료된 결과 데이터를 활용하여 한 심사자 평가들 간 신뢰도 및 심사자들의 평가들 간의 신뢰도를 측정하여, 이를 활용하여 객관성을 보정하는 사후 방법론이 있다. 특히, 스포츠 분야에서 심판의 정성적 평가가 필요한 부분(예시: 체조, 피겨스케이팅 등)에서 널리 연구되고 있으며, 한 심사자 내 신뢰도 및 심사자들 간의 신뢰도 등을 측정한다. 다음으로 다른 심사자들과의 결과를 비교하여 주관성이 개입된 척도를 측정하는 Pearson의 적률상관계수 혹은 유목내 상관계수(Inter-class Correlation Coefficient; ICC)가 활용될 수 있다[5]. 다만, 이 경우 주로 여러 명의 평가자들이 한 피험자(Subject)를 대상으로 한 연구가 대부분을 차지하며, 이와는 대조적으로 한 명의 평가자가 여러 명의 피험자를 대상으로 평가를 한 경우 어떻게 객관성을 보장할 것인지에 대한 부분은 연구가 부족한 상황이다. 일반적으로 한 명의 평가자가 오랜 시간 동안 다양한 피험자를 평가하는 경우, 시간이 지나면서 일정 수준 이상의 평가의 객관성은 확보될 수 있으나, 평가 결과에 대한 객관적인 피드백이 존재하지 않는 경우는 평가의 객관성이 보정되지 않은 채로 평가자의 습관처럼 잘못된 방향으로 굳어질 수 있는 위험이 있다.

이때, 평가자들이 여러 명 있는 경우, 해당 응답들을 수행한 환경 자체의 유사성을 이용하여 평가자들의 성향을 유추해 볼 수 있을 것이며, 이를 통해 객관성을 보정해볼 수 있는 시도를 할 수 있을 것이다.

위에서 기술한 문제를 해결하기 위해 본 연구에서는 Ordered Probit 모델을 활용한 방법론을 제시하고자 한다. 본래, Ordered Probit 모델은 종속 변수가 Likert scale 같이 범주 간 서열을 갖는 순서형 척도일 때, 이를 분석하기 위해 사용하는 모델이다. 종속 변수가 명목형 척도이고 순서를 갖지 않는 이항 종속 변수(Y = 1 또는 Y = 0)인 경우에는 Probit 모델이나 Logit 모델을 이용하여 분석이 가능하지만[2], 종속 변수가 3개 이상의 값을 갖는 순서형 척도

인 경우에는 적합하지 않다. 또한 일반적으로 많이 사용되는 다항회귀분석은 종속 변수 값들 사이의 차이를 동일한 것으로 인식하여 분석함으로써, 종속 변수 간 서열 관계를 고려하지 못하고 오류를 범할 수 있다는 한계가 있다[4].

그러므로 종속 변수가 명목형 변수이면서 서열을 갖는 값들로 이루어진 순서형 척도인 경우, Probit 모델과 Logit 모델을 응용한 Ordered Probit 모델이나 Ordered Logit 모델을 이용한다. 이때 직접적으로 관측할 수는 없지만 종속 변수를 결정하는데 영향을 미치는 내재된 잠재 변수를 갖는다고 가정하는데, 이 잠재 변수를 특정 분포를 갖는 연결 함수를 통해 추론한다. Ordered Logit 모델은 잠재 변수의 오차항의 분포가 로지스틱 분포를 따른다고 가정하고, Ordered Probit 모델의 경우 정규분포를 따른다고 가정한다는 점에서 차이가 있다. 일반적으로 오차항의 확률분포를 정규분포로 가정하는 경우가 많아 Ordered Probit 모델을 사용하는 것이 바람직하다고 할 수 있다[6]. 이러한 이유로 본 연구에서도 Ordered Probit 모델을 이용하여 수술에 대한 의사의 평가에서 주관성을 고려한 예측 모델을 만들고 분석을 수행하였다.

3. 평가 프로세스 및 Ordered Probit 모델

3.1 수술 및 환자에 대한 평가 프로세스

본 연구에서는 의사의 수술 전/후 평가 및 수술 1일 후 간호사의 환자 상태 체크 과정에서 발생하는 데이터와 Ordered Probit 방법론을 활용해 모델링을 수행하였다. 분석 대상 의료서비스는 지방흡입수술 서비스이며 복부, 팔, 허벅지 등의 지방을 캐뉼라는 관을 통해 흡입하는 수술이다. 집도의 인터뷰를 통해 수술 특성 상 집도의의 술기(Clinical Skill)가 수술 결과에 중요한 영향을 미칠 수 있음을 파악하였다. 본 절에서는 모델링에 사용하는 데이터가 어떤 과정을 거쳐 생성되는지에 대해, 각 평가 주체들이 어떤 항목을 어느 시점에 어떤 척도로 평가하는지를 기준으로 기술하였다.

먼저 <Table 1>에 제시한 바와 같이 의사는 수술 전 환자와의 상담 및 진료 과정에서 환자의 지방 두께를 초음파 검사를 통해 cm 단위로 측정하고, 환자의 피부 탄력도에 대해 본인의 기준으로 보통 이하, 보통, 양호, 좋음, 아주 좋음의 순서로 1점부터 5점까지의 점수로 평가한

〈Table 1〉 Surgeon's Checklist Before and After Surgery

Item	Content	Scale	Evaluation Time
Fat thickness	Patient's fat thickness	cm	Before surgery
Skin elasticity	Patient's skin elasticity	5-point scale	
Abnormal stroke	Presence or absence of abnormal strikes	Yes or No	After surgery
Volume reduction	Amount of fat reduced	5-point scale	
Line correction	Degree of line correction	5-point scale	
Surface regularity	Degree of regularity on the surface of the surgical site	5-point scale	
Total result	Evaluation of surgery performed	5-point scale	

〈Table 2〉 Nurse's Checklist After One Day of Surgery

Item	Description	Scale	Evaluation Criteria
Sickness	Degree of pain the patient feels	10-point scale	Degree to which the patient talks
Edema	Degree of edema	5-point scale	Degree to which the recorder judges
Bruise	Degree of bruise	5-point scale	
Unevenness	Degree of unevenness of the surgery area	5-point scale	
Line change	Degree of line change	5-point scale	
Satisfaction	Satisfaction with surgery	5-point scale	

다. 이어서 수술 직후, 본인의 수술에 대해 다섯 가지 항목을 직접 평가한다. 먼저 Abnormal Stroke는 수술 과정에서 환자가 비정상적인 Stroke가 있었는지 여부를 Yes or No의 항목으로 체크한 뒤, 지방 용적 감소 정도에 대해 아주 적음, 적음, 보통, 많음, 아주 많음의 순서로 1점부터 5점까지의 점수로 평가한다. 연이어 라인 교정 정도, 피부 표면의 울퉁불퉁함 정도에 대해 보통 이하, 보통, 양호, 좋음, 아주 좋음의 순서로 1점부터 5점 사이의 점수를 매긴다. 그 후 최종적으로 수술 전반에 대해 보통 이하, 보통, 양호, 좋음, 아주 좋음의 순서로 1점부터 5점까지의 점수로 평가한다.

다음으로 수술 1일 후 환자의 상태에 대해 간호사가 체크하는 과정에 대해 <Table 2>에 기술하였다.

<Table 2>에 제시한 바와 같이 간호사는 수술 1일 후 환자의 상태에 대해 다섯 가지 항목을 확인한다. 그 중 첫 번째 항목인 통증은 환자가 주관적으로 애기하는 통증의 정도로, 그 정도를 좀더 세분화하여 파악하기 위해 가장 아플 때를 10으로 했을 때 1점부터 10점까지의 점수로 기록하는 값이다. 이어서 간호사는 환자의 상태에 대해 평가 및 기록한다. 먼저 부종이 얼마나 발생했는지, 멍이 어느 정도로 나타나는지, 환자 피부의 울퉁불퉁한 정도에 대해 아주 심함, 심함, 보통, 약간 있음, 없음의 순서로 1점부터 5점까지의 점수로 평가한다. 이와 함께 수술 부위의 라인 변화 정도에 대해 보통 이하, 보통, 양호, 좋음, 아주 좋음의 순서로 1점부터 5점까지의 점수로 평가한다. 마지막으로 수술에 대한 환자의 만족도를 보통 이하, 보통, 양호, 좋음, 아주 좋음의 순서로 1점부터 5점까지의 점수로 기록한다.

3.2 Ordered Probit 모델

앞서 기술한 바와 같이 종속 변수가 연속적인 수치형이 아닌 경우, 직접적으로 관측되지는 않지만 종속 변수에 영향을 미치는 잠재 변수가 기저에 존재한다고 가정한다. Ordered Probit 모델에서 이러한 잠재 변수 y^*는 다음과 같이 나타낼 수 있다.

$$y^* = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0,1). \qquad (1)$$

식 (1)에서 X_i는 독립 변수(설명 변수)를 의미하고 β는 추정 계수를 의미한다. ϵ_i은 오차항을 나타내며 Ordered Probit 모델에서는 이 오차항이 표준정규분포를 따른다고 가정한다. 본

연구에서는 의사가 본인이 집도한 수술에 대해 5점 척도로 평가하는 Total result 변수를 종속 변수 y로 설정하였다. 그리고 의사가 수술에 대한 평가 점수를 산정할 때 영향을 미치지만, 겉으로 관측될 수 없는 의사의 주관성과 같은 요소를 잠재 변수 y^*로 설정하였다. 이때, 독립 변수로는 의사와 간호사의 평가항목인 지방 두께, 피부 탄력도, Abnormal stroke 발생 여부, 지방 용적 감소 정도, 라인 교정 정도, 피부 표면의 울퉁불퉁함 정도, 통증의 정도, 부종의 정도, 멍의 정도, 균일함, 라인 변화 정도 등의 항목들을 사용하였다. 간호사와 같이 제3자의 평가 항목을 독립 변수로 포함시켜 모델을 만들고 이를 통해 의사의 주관성을 보정하는 것이 본 실험의 목적이다. 최종적으로 의사의 평가 점수 y와, 의사가 y의 점수로 평가하도록 영향을 미치는 잠재 변수 y^*는 다음과 같은 관계를 갖는다.

$$y=\begin{cases}1 & if \quad y^* < t_1,\\ 2 & if \quad t_1 \le y^* < t_2,\\ 3 & if \quad t_2 \le y^* < t_3,\\ 4 & if \quad t_3 \le y^* < t_4,\\ 5 & if \quad y^* \ge t_4.\end{cases} \qquad (2)$$

식 (2)에서 t_i는 각 독립 변수에 대한 추정 계수 β와 함께 추정되는 한계값(Threshold)으로 종속 변수의 범주 개수-1개를 갖는다[7]. 식 (1)에서 각 독립 변수의 추정 계수 β를 이용하여 잠재 변수 y^*의 값을 계산할 수 있고, 이 y^*의 값이 추정된 t_i의 값을 경계로 갖는 구간 중 어느 구간에 속하는지에 따라 종속 변수 y의 값이 정해지는 것이다. 예를 들어 Ordered Probit 모델에서 t_3가 40.0412, t_4가 42.5641로 각각 계산되고, 식 (1)에 따라 계산된 y^*의 값이 41.4043이라면 식 (2)에 의해 종속 변수인 y의 값을 4로 예측할 수 있다. 즉, 의사가 수술에 대해 4점의 점수로 평가할 것이라는 예측값을 얻을 수 있는 것이다.

본 연구에서는 R에서 MASS 패키지의 polr() 함수를 이용하여 한계값 t_i의 값을 구하였다. 그리고 앞서 언급한 바와 같이 간호사와 같은 제3자가 체크한 환자 상태 역시 고려하여 주관성을 보정하고, 수술에 대한 평가 예측값을 구한 뒤 이를 기존에 의사가 평가한 값과 비교 및 검증하였다.

4. 모델 구축 및 실험 검증

4.1 연구 가설

본 연구에서 검증하고자 하는 연구 가설은 다음과 같다.

“간호사와 같은 제3자의 평가를 고려하여 Ordered Probit 모델을 생성하고, 이를 이용하여 예측한 수술에 대한 평가값과 환자의 만족도 간의 상관관계 값은, 기존에 의사가 주관적으로 수술 결과에 대해 평가한 값과 환자의 만족도 간의 상관관계 값보다 양의 방향으로 높게 나올 것이다.”

이는 수술에 대한 의사의 평가 과정에서 의사 본인의 주관성이 개입되어 실제 환자의 만족도와는 동떨어진 방향으로 평가하게 될 경우 의사의 평가 결과와 환자 만족도 간의 연관성은, 제3자의 평가를 통해 보정한 수술 결과 값과 환자 만족도 간의 연관성 수준보다 상대적으로 낮게 측정될 것이다 라는 의미로도 해석될 수 있다. 그리고 이러한 두 상황의 상관관계 값 비교를 통해 제3자의 평가를 고려한 Ordered

Probit 모델이 객관성 보정에 도움을 줄 수 있음을 밝히고자 한다.

4.2 데이터 전처리 및 실험 환경 구축

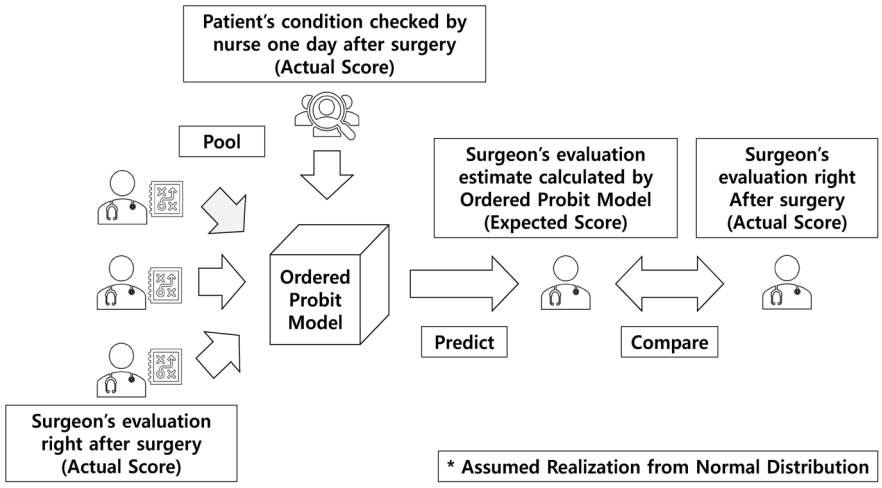
본 연구의 실험에서는 2017년 6월부터 2018년 4월까지 수술을 받은 환자 413명에 대한 데이터를 사용하였다. 데이터 전처리 과정은 다음과 같다. 먼저 의사의 수술 전/후 환자 상태 및 수술에 대한 평가와 수술 1일 후 환자 상태에 대한 간호사의 평가 및 환자의 만족도 데이터를 수집하였다. 또한 의사나 간호사의 평가 항목에서 Likert scale로 측정하는 값들이 실제 데이터 상으로는 각 응답의 수준별로 1 또는 0의 이진 변수 값으로 기록되어 있어, 이를 모두 1부터 10 또는 1부터 5까지의 정수로 변환하는 작업을 수행하였다. 그 후 결측 값이 있거나 중복 기록된 항목들을 제외시켰고, 최종적으로 413건의 데이터를 실험에 사용하였다.

실험에 사용한 소프트웨어 및 컴퓨팅 환경은 다음과 같다. 모든 실험 및 분석은 R version 3.6.3을 Windows 10 64bit환경에서 구동하여 수행하였다. 분석을 위해 활용한 MASS 패키지는 7.3-51.5 버전(2019-12-20)을 사용하였다.

4.3 분석 프로세스

전체 분석 과정을 정리하면 <Figure 1>과 같다.

본 연구에서 실험 및 분석 과정은 다음과 같이 진행하였다. 먼저 의사가 본인의 수술에 대해 평가한 데이터들을 모아 전처리 후 Ordered Probit 모델을 만드는 단계를 수행하였다. 이때, 주관성을 보정하기 위해 수술 1일 후 간호사, 즉 제3자가 환자의 상태를 체크하면서 기록 및 평가한 항목들 역시 모델에 포함하여 모델링하였다. 그 후 기존에 의사가 수술에 대해 평가한 값과 Ordered Probit 모델로 주관성을 보정하여 수술에 대해 예측한 평가값 사이에 유의미한 차이가 발생하는지 비교하였다. 이때, 유의



〈Figure 1〉 Analysis Process Framework

미한 차이가 나타날 경우, 모델링 결과에 대한 검증을 위해 환자의 만족도 값을 검증 변수로 삼아 검증 작업을 수행하게 된다. 검증 과정에서 Correlation test를 이용하여 기존에 의사가 평가한 값과 Ordered Probit 모델로 예측한 평가값 중, 주관성을 보정하여 Ordered Probit 모델로 예측한 값이 환자의 만족도와 비교하여 상대적으로 더 높은 상관관계가 있는지 여부를 판단하였다.

4.4 모델 실험

모델링을 위해 사용할 Data set을 구성하는 변수를 정의하기 위해, 다음의 <Table 3>과 같이 실제 업무에서 평가자가 평가 시 사용하던 항목들에 변수 이름을 부여하고 값의 유형을 정리하였다.

<Table 3>의 변수들 중 LINE_CORRECTION과 SURFACE_REGULARITY는 두 독립 변수 간 상관관계가 높아 종속 변수인 TOTAL_RESULT와 상대적으로 더 작은 상관 계수 값을 갖는 LINE_CORRECTION을 독립 변수에서 먼저 제외시켰다. 또한 SURFACE_REGULARITY와 UNEVEN은 평가 시기와 주체가 다르기 때문에 독립성을 갖는다고 볼 수 있으므로 독립변수에 같이 포함하였으나 UNEVEN은 유의미하지 않은 변수로 판별되어 제외시켰다. 이외에도 여러 차례 모델링을 수행하면서 변수의 p-value를 확인하여 유의미하지 않은 변수들을 독립 변수에서 제거하였고, 이렇게 변수 선택 과정을 거쳐 최종적으로 다음의 <Table 4>와 같은 변수들을 모델링에 사용하였다.

Ordered Probit 모델링 결과 최종적으로 독립변수 VOLUME_REDUCTION, SURFACE_REGULARITY, BRUISE는 모두 유의수준 10%하에서 모두 유의미한 변수임을 확인할 수 있고, 추정된 한계값 역시 p-value가 매우 작아 유의미한 값으로 추정되었음을 확인할 수 있다.

〈Table 3〉 List of Variables

Variable	Variable Name	Item name for Evaluation	Types of Values
Independent variable	FAT_THICKNESS	Fat Thickness	Continuous real number
	SKIN_ELASTICITY	Skin Elasticity	5-point scale
	ABNORMAL_STROKE	Abnormal Stroke	0 or 1
	VOLUME_REDUCTION	Volume Reduction	5-point scale
	LINE_CORRECTION	Line Correction	5-point scale
	SURFACE_REGULARITY	Surface Regularity	5-point scale
	SICK	Sickness	10-point scale
	BRUISE	Bruise	5-point scale
	EDEMA	Edema	5-point scale
	LINE_CHANGE	Line Change	5-point scale
	UNEVEN	Unevenness	5-point scale
Dependent variable	TOTAL_RESULT	Total Result	5-point scale
Validation variable	SATISFY	Satisfaction	5-point scale

〈Table 4〉 Ordered Probit Modeling Results

Independent Variable	Estimated Coefficient	S.E.	t-statistic	p-value
VOLUME_REDUCTION	1.0167	0.1945	5.2274	$p < 0.0001$
SURFACE_REGULARITY	4.0761	0.2845	14.3268	$p < 0.0001$
BRUISE	0.3006	0.1633	1.8409	0.0664
Threshold	Estimated Coefficient	S.E.	t-statistic	p-value
t_1	12.9351	1.3969	9.2599	$p < 0.0001$
t_2	17.7628	1.6078	11.0477	$p < 0.0001$
t_3	22.5988	1.8742	12.0581	$p < 0.0001$
Residual Deviance	139.8121			
AIC	151.8121			

이를 통해 수술 결과에 대한 평가값이 결정되는데 영향을 미치는 잠재 변수 $TOTAL_RESULT^*$에 대해 다음과 같은 식을 도출할 수 있다.

$$\begin{aligned} TOTAL\_RESULT^* &= 1.0167 \times VOLUME\_REDUCTION \\ &+ 4.0761 \times SURFACE\_REGULARITY \\ &+ 0.3006 \times BRUISE + \epsilon_i, \\ &\epsilon_i \sim N(0,1). \end{aligned} \tag{3}$$

위 식에서 VOLUME_REDUCTION의 값이 한 단위 증가할 때마다 잠재 변수 $TOTAL_RESULT^*$를 1.0167만큼 증가시키며, SURFACE_REGULARITY의 경우 한 단위당 $TOTAL_RESULT^*$를 4.0761만큼 증가시키고, BRUISE는 0.3006만큼 증가시키는 것을 확인할 수 있다. 이때 제3자인 간호사가 평가한 BRUISE 항목이 잠재 변수 값에 영향을 미침으로써 의사의 주관성을 보정하는 역할을 한다.

한편, Ordered Probit 모델의 연결함수를 이용해 종속 변수 값에 대한 확률을 구하고 이를 역이용하여 각 종속 변수가 나오는 구간에 대한 한계값을 계산할 수 있다. 실험에 사용한 Data set에서 종속 변수 TOTAL_RESULT가 갖는 값의 구간이 2에서 5까지 4개의 범주로 나타나 다음의 <Table 5>와 같이 한계값도 t_1부터 t_3까지 계산되었다.

<Table 5>에서 식 (3)의 결과로 계산된 $TOTAL_RESULT^*$의 값이 12.9351보다 작다면 종속 변수인 TOTAL_RESULT는 2점이 되는 것이고, 마찬가지로 $TOTAL_RESULT^*$이 12.9351 이상이고 17.7628 미만이면 TOTAL_

〈Table 5〉 Threshold Interval for TOTAL_RESULT

Dependent Variable(TOTAL_RESULT)	Threshold Interval
2	$TOTAL\_RESULT^* < 12.9351$
3	$12.9351 \leq TOTAL\_RESULT^* < 17.7628$
4	$17.7628 \leq TOTAL\_RESULT^* < 22.5988$
5	$22.5988 \leq TOTAL\_RESULT^*$

RESULT는 3점이 되며, $TOTAL_RESULT^{*}$이 17.7628 이상이고 22.5988 미만이면 4점, $TOTAL_RESULT^{*}$이 22.5988 이상이면 5점이 된다.

4.5 모델 검증

앞서 '4.1 연구 가설'에서 Ordered Probit 모델로 예측한 수술에 대한 평가값과 수술에 대한 환자의 만족도 간의 상관관계가, 수술에 대한 의사의 주관적 평가값과 수술에 대한 환자의 만족도 간의 상관관계보다 양의 방향으로 높게 나올 것이라 가정하였다. 본 장에서는 이 가설을 검증하기 위해, 먼저 두 결과 값 사이에 유의미한 차이가 발생했는지를 확인하고 Correlation test를 통해 환자의 만족도와의 상관관계를 비교하였다.

4.5.1 Wilcoxon-test

본 절에서는 의사가 본인의 수술에 대해 기존에 평가한 결과와 Ordered Probit 모델을 이용해 예측한 결과 사이에 유의미한 차이가 있는지를 확인하였다. 이때, 수술에 대한 평가 점수는 서열 척도에 해당하고 이 항목에 대해 제3자의 개입 전후에 대한 차이를 비교하는 것이므로, 서열 척도 값에 대해 독립적이지 않은 두 쌍의 차이를 분석하는 Wilcoxon-test를 수행하였다[9, 12].

Wilcoxon-test 수행 결과, p-value가 0.0164로 계산되어 유의수준 2% 내에서 의사가 기존에 평가한 값과 Ordered Probit 모델을 이용하여 예측한 값 사이에 유의미한 차이가 있음을 확인할 수 있었다. 즉, 간호사와 같은 제3자의 평가를 반영함으로 인해 수술에 대한 기존 평가 값과 예측 값 사이에 차이가 발생한다는 것을 의미한다. 하지만 아직 의사의 주관성이 보정된 예측 값이 환자의 만족도에 상응하는 방향으로 도출되었는지 알 수 없으므로 추가적인 검증을 시행하였다.

4.5.2 상관관계 비교

의사가 기존에 수술에 대해 평가한 값과 Ordered Probit 모델로 예측한 평가값 중 어떤 것이 더 환자의 만족도와 높은 상관관계를 갖는지 검증하기 위해 Correlation test를 수행하였고, 결과는 다음의 <Table 6>과 같다. 이때 순서형 자료에 대한 상관계수로써 Spearman 상관계수를 이용하였다[8].

Correlation test 결과, 기존에 의사가 수술에 대해 평가한 값과 환자의 만족도 간의 Spearman 상관계수는 0.1368이고 p-value는 0.0054로 유의수준 0.6% 내에서 유의미함을 확인할 수 있었고, Ordered Probit 모델로 예측한 평가값과 환자의 만족도 간의 Spearman 상관계수는

〈Table 6〉 Results of Correlation Test

Correlation Test	rho	p-value
Correlation between surgeon's existing evaluation and patient's satisfaction	0.1368	0.0054
Correlation between evaluation value predicted by Ordered Probit model and patient's satisfaction	0.1618	0.0010

0.1618이고 p-value는 0.0010으로 유의수준 0.1%내에서 유의미함을 확인할 수 있었다. 이 과정에서 Ordered Probit 모델로 예측한 평가값과 환자의 만족도 사이의 상관계수가 양의 방향으로 0.03 정도 높아져, 주관성을 보정한 후의 결과가 환자의 만족도와 더 높은 상관관계를 갖는 것을 볼 수 있었다. 이를 통해 앞서 '4.1 연구 가설'에서 설정한 "간호사와 같은 제3자의 평가를 고려하여 생성한 Ordered Probit 모델에서 예측한 수술에 대한 평가값과 수술 만족도 간의 상관계수는, 기존에 의사가 주관적으로 수술 결과에 대해 평가한 값과 수술 만족도 간의 상관계수보다 양의 방향으로 높게 나올 것이다."라는 주장이 타당한 것을 확인할 수 있다.

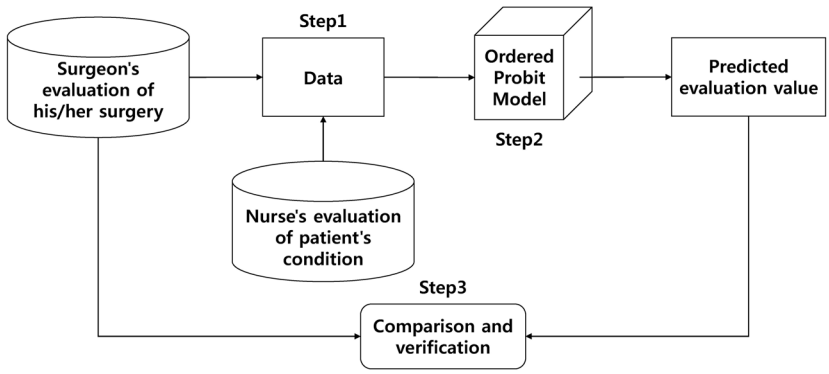
4.6 결과 해석 및 토의

본 연구에서는 Ordered Probit 모델을 사용하여 집도의가 시행한 수술에 대해 평가하는 과정에서 개입될 수 있는 주관성을 보정하는 연구를 수행하였다. 전체적인 분석 과정은 다음의 <Figure 2>와 같이 세 단계로 나누어 볼 수 있다.

먼저 첫 번째 단계에서는 평가 데이터를 추출하고 분석에 사용할 수 있도록 정제하였다. 분석에 사용된 데이터로 수술 정보, 수술에 대한 평가, 수술 후 환자 상태에 대한 평가 정보들을 활용하였다. 이후 결측 값이 있거나 및 중복값이 있는 데이터를 제외하고, 이진 변수로 표현된 값들을 변환하는 작업 등을 통해 최종 분석 대상 데이터 셋을 구축하였다.

두 번째 단계에서는 구축된 데이터 셋을 이용하여 Ordered Probit 결과를 도출하고 이 결과에 기반하여, 주관성을 보정한 평가 결과값을 예측하였다. 이는 모델에서 사용된 독립 변수 값에 의해 결정되는 잠재 변수가 한계 값의 어느 범위에 속해 있는지에 따라 결정되었다.

마지막으로 보정된 결과값이 실제로 기존 의사의 평가 값보다 환자의 만족도에 더 부합한 값이라고 할 수 있는지 검증하기 위해 환자의 만족도값과의 상관계수 비교 분석을 수행하였다. 그 결과 기존에 의사가 내린 평가와 Ordered Probit 모델로 예측한 평가 값 사이에 유의미한 차이가 발생하였고, Ordered Probit 모델을 이용해 예측한 평가와 환자의 만족도 사이의 상관계



〈Figure 2〉 Experimental Process

수가 의사의 기존 평가와 환자의 만족도 사이의 상관계수 보다 더 증가한 것을 확인할 수 있었다. 이를 통해 주관성을 보정한 예측 값이 환자의 만족도와 더 부합한 값이라는 것을 알 수 있었다.

참고로 Correlation test를 통한 검증과정에서 사용한 Spearman 순위상관계수는 데이터의 자유도가 증가함에 따라 유의 수준을 만족하는 계수의 값이 감소하는 경향을 보인다[10]. 그러므로 본 실험 결과를 검증하는 과정에서 계산된 상관계수의 값 0.1368과 0.1618은 413건의 데이터에서 비롯되는 자유도의 크기를 고려했을 때, 적합한 수준으로 측정된 것으로 판단된다. 즉 두 상관계수 값이 상대적인 상관계수 값의 차이와 p-value를 보았을 때, 상당히 낮은 유의수준(p-value < 0.01) 하에서 양의 방향으로 상승한 모습을 보이므로 본 결과는 통계적으로 유의미하다고 결론 내릴 수 있다.

5. 결 론

서비스의 품질 평가를 위해 널리 사용되는 설문조사에서 응답자의 주관성 개입은 흔히 나타나는 현상이나, 이는 분석결과를 왜곡하고 잘못된 의사결정을 야기한다. 특히, 하나의 평가 도구가 여러 사람에 의해 반복적으로 사용되는 경우, 평가결과를 객관화하는 방법에 대해서는 아직 많은 연구가 진행되지 않은 상황이다.

본 연구에서는 Ordered Probit 모델을 이용하여 평가결과의 객관성을 확보하는 방안을 제시하였다. 이를 위해, 특정 병원에서 수술 후 집도의가 직접 평가한 수술 결과 데이터를 수집 및 분석하였고, 이를 기반으로 Ordered Probit 모델을 통해 결과에 대한 객관성을 설명할 수 있는 방법을 제시하였다. 본 연구에서 제안한 Ordered Probit 모델을 이용해 예측한 값과 의사가 기존에 평가한 값 사이에 차이가 있는지를 Wilcoxon-test를 통해 확인한 결과, 유의미한 차이가 있음을 확인하였고 주관성이 어느정도 개입되어 있음을 입증하였다. 다음으로, Correlation test를 수행하여 Spearman 상관계수를 분석한 결과, 간호사와 제3자의 평가를 고려하여 생성한 Ordered Probit 모델에서 예측한 수술에 대한 평가값이 환자가 느끼는 수술 만족도와 더 높은 상관관계를 갖는 것을 볼 수 있었다. 이를 통해 본 연구에서 제시한 방법들이 설문조사에서 응답자의 주관성을 보정하고 평가결과의 객관성을 확보하는 데 기여할 수 있음을 확인하였다.

한편, 본 연구에서는 객관성 확보에 대한 검증 지표로 수술에 대한 환자의 만족도라는 항목을 사용할 수 있는 상황이었지만, 다양한 상황에서 실시되는 일반적인 설문조사에서는 어떤 기준을 객관성을 판단하는 척도로 삼을 수 있을지에 대한 추가적인 논의가 필요할 것으로 보인다.

References

[1] Aitchison, J. and Silvey, S. D., “The generalization of probit analysis to the case of multiple responses,” Biometrika, Vol. 44, No. 1/2, pp. 131-40, 1957.

[2] Aldrich, J. H. and Nelson, F. D., Linear Probability, Logit, and Probit Models, No.

45, Sage, 1984.

[3] Folstein, M. F., Robins, L. N., and Helzer, J. E., “The mini-mental state examination,” Archives of General Psychiatry, Vol. 40, No. 7, pp. 812-812, 1983.

[4] Ju, M., “Probit and ordered probit analysis and its application,” Journal of Governmental Studies, Vol. 6, No. 1, pp. 24-49, 2000.

[5] Kang, S., “How to estimate inter-rater reliability?,” Korean Society For Measurement And Evaluation In Physical Education And Sports Science, Vol. 13, No. 1, pp. 1-8, 2011.

[6] Lee, G., Kang, K., and Rho, J., “Development of bicycle level of service model from the user’s perspectiveusing ordered probit model,” Journal of Korean Institute of Intelligent Transport Systems, Vol 8, No. 2, pp. 108-117, 2009.

[7] Mckelvey, R. D. and Zavoina, W., “A statistical model for the analysis of ordinal level dependent variables,” Journal of Mathematical Sociology, Vol. 4, No. 1, pp. 103-120, 1975.

[8] Myers, J. L., Well, A. D., and Lorch Jr, R. F., Research Design and Statistical Analysis, Routledge, 2013.

[9] Park, J. and Kim, S., “The significance test on the AHP-based alternative evaluation: An application of non-parametric statistical method,” The Journal of Society for e-Business Studies, Vol. 22, No. 1, pp. 15-35, 2017.

[10] Reeve, B. B. and Fayers, P., “Applying item response theory modeling for evaluating questionnaire item and scale properties,” Assessing quality of life in clinical trials: Methods of practice, Vol. 2, pp. 55-73, 2005.

[11] Van der Linden, W. J. and Hambleton, R. K., Handbook of modern item response theory, Springer Science & Business Media, 2013.

[12] Wilcoxon, F., Individual Comparisons by Ranking Methods, pp. 196-202, Breakthroughs in statistics, Springer, New York, 1992.

저 자 소 개

정민규 (E-mail: minkyujeong@kaist.ac.kr)
2020년 홍익대학교 경영학과, 산업공학과 (학사)
2020년~현재 KAIST 산업및시스템공학과 석사과정 재학 중
관심분야 Deep Learning, Machine Learning, Natural Language Processing, XR

강윤철 (E-mail: yckang@ewha.ac.kr)
2002년 KAIST 산업공학과 (학사)
2004년 서울대학교 산업공학과 (석사)
2014년 Pennsylvania State University, Industrial and Manufacturing Engineering (Ph. D.)
2004년~2007년 LG CNS
2007년~2008년 서울대학교 자동화 연구소
2014년~2016년 Research Associate, Pennsylvania State University
2016년~2020년 홍익대학교 산업공학과 조교수
2020년~현재 이화여자대학교 경영대학 조교수
관심분야 Decision-making under uncertainties

김남철 (E-mail: namchul.kim@365mc.com)
2014년 경희대학교 (의학박사)
2006년~현재 ㈜365mc 네트웍스 대표이사
2007년~현재 부산 365mc 병원 대표원장
관심분야 인공지능 기술을 활용한 비만관리

정광헌 (E-mail: khchung@hongik.ac.kr)
1997년 서울대학교 산업공학과 (학사)
1999년 서울대학교 산업공학과 (석사)
2010년 University of Florida, Industrial & Systems Engineering (Ph.D.)
1999년~2004년 삼성 SDS 정보기술연구소 선임연구원
2010년~2011년 Post-Doctoral Fellow, Center for Operations Research and Econometrics(CORE), Belgium
2011년~현재 홍익대학교 경영학부 교수
관심분야 Optimization Theory and Algorithms, Application of Management Science Techniques to Networks, Supply Chain, Energy Systems, Healthcare Service