한국심리학회지 : 일반
Korean Journal of Psychology: General
2015, Vol. 34, No. 3, 741-767

엑셀을 이용한 비선형 자료 적합 소개

오 성 주†

서울대학교 심리학과

자료 적합은 심리물리학에서 연구자가 흔하게 마주치는 과정으로 상당한 노력과 시간이 든다. 자료 적합을 위한 상용 소프트웨어들이 많지만 처리 과정이 가려져 있거나 구입비용이 높다는 단점이 있다. 주변에서 쉽게 구할 수 있는 엑셀은 해 찾기 기능을 갖고 있으며 이를 이용해 복잡한 계산이 필요한 최적화를 수행할 수 있다. 엑셀은 스프레드시트 기반으로 자료 관찰과 그래프 그리기가 동시에 가능해 사용자가 자료 적합 과정을 직관적으로 쉽게 이해할 수 있다. 본 논문에서는 비선형 자료를 대상으로 엑셀에서 적합을 하는 과정을 단계별로 시연한다. 이를 위해, 실제 자료에 심리측정 함수로서 심리학에서 널리 사용되는 로지스틱 함수가 적용되고 최소제곱 추정과 최대우도 추정으로 적합되는 과정이 차례대로 소개된다. 끝으로 자료 적합을 통해서 구해진 자료 분포의 파라미터를 이용해 어떻게 조건간 통계적 검증을 하는지가 설명된다.

주요어 : 자료 적합, 최소제곱추정, 최대우도추정, 심리측정함수, Excel solver

* 본 원고를 세심하게 심사하시고 많은 지적을 해주신 여러 심사위원님과 부편집장님께 감사드립니다. 원고를 검토해준 이화여자대학교 김성호 교수님과 서울대학교 류대현, 이한진, 유지호군에게 감사합니다. 본 논문에 나오는 엑셀 서식 및 예제 파일이 필요하신 분은 저자에게 이메일로 연락 주십시오.

† 교신저자: 오성주, 서울대학교 사회과학대학 심리학과, (08826) 서울시 관악구 관악로 1
E-mail: songjoo@snu.ac.kr

심리물리학에서 연구자들은 독립변인과 종속변인이 연속 변인인 자료에 대해서 두 변인의 관계를 수학적으로 기술하기 위해 자료 적합을 한다. 자료 적합 과정은 매우 까다롭기 때문에 실제 실험실에서 자료 적합에 특화된 상용 소프트웨어를 사용하곤 하는데 Graphpad Prism, Mathcad, Maple, OriginLab, Sigma Plot 등 다양하다. 하지만 이 소프트웨어들에서는 계산 과정이 가려져 있어 내용을 잘 이해하기 어렵고 프로그램 구입비용이 높은 단점이 있다. 이 밖에 범용으로 개발된 MATLAB, R, C, Basic 등에서 소스 코드를 작성하여 자료 적합이 가능한데 이 역시 코드 입력 방법에 익숙해지는 데 시간과 노력이 많이 들어 초심자들이 접근하기 어렵다.

이 프로그램들의 단점을 생각해 보았을 때 Microsoft사의 엑셀(Excel)은 자료 적합에 매우 편리한 도구이다. 엑셀은 개인용 컴퓨터에 MS-Office가 설치되어 있으면 함께 들어있으므로 추가로 구하는 데 노력이 들지 않는다. 엑셀은 수치들이 들어있는 셀들이 나열되어 있는 스프레드시트 기반이어서 자료를 직접 하나 하나 편집할 수 있고, 그래프가 스프레드시트와 병렬적으로 구동되는 특징이 있다. 따라서 사용자는 특정 셀의 값을 바꿀 때 그 결과가 어떻게 그래프에 영향을 주는지 동시에 살펴볼 수 있다. 이는 매우 큰 장점인데 자료 입력과 그래프 그리기가 계열적으로 이뤄지는 MATLAB 또는 R과 비교해보면 얼마나 시각적인 이해도가 높은지 알 수 있다. 이 점은 이미 코드 쓰기에 익숙한 전문가에게 초보적인 기능으로 보일 수도 있지만 초심자에게는 매우 크게 다가온다. 이런 시각적 직관성으로 엑셀은 프로그램 실력과 통계에서 초보적인 지식을 가진 사람이어도 몇 시간이면 어려운 적합 과정을 수행할 수 있다.

엑셀을 이용하여 자료 적합을 소개한 첫 논문들은 심리학에서(Dodson, Prinzmetal, & Shimamura, 1998) 그리고 기계 공학 분야에서(John, 1998) 나온 것으로 보인다. 뒤를 이어 생물/의학 분야(Brown, 2001, 2006)에서 있었고 최근에 자연 과학 분야(Kemmer & Keller, 2010)로 이어지면서 점점 응용의 폭이 넓어지고 있다. 이와 함께, 시간이 흐르면서 엑셀에 고급 기능이 추가되고 고급 통계와 복잡한 최적화(maximization) 기능도 업그레이드되고 있어 엑셀을 이용한 자료 적합은 일시적인 인기로 끝나지 않고 지속적으로 이어질 것으로 전망된다. 하지만 아직 다른 상용 소프트웨어들과 MATLAB과 R 같은 프로그램에 비해 엑셀에 대한 평가는 매우 저평가 되어 있다. 이는 아마도 엑셀을 쉽게 구할 수 있기 때문에 간단한 계산이나 그래프 정도를 처리하는 도구 정도로 낮추어 보는 경향도 얼마간 있기 때문일 것이다.

그렇지만 실제로 실험실에서 얻은 자료를 엑셀에서 적합해보고 강의에서 교육했을 때 매우 높은 만족을 경험할 수 있다. 본 글에서는 엑셀의 최신 버전으로 적합을 단계적으로 시연하여 초심자도 쉽게 자료 적합을 해내고 이를 바탕으로 자료 적합에 대해 좀 더 깊이 있는 공부를 하도록 동기를 부여하고자 한다. 물론, 본 글에서 말하는 내용들은 이미 나와 있는 논문들과 크게 다르지 않고 연구 방법론 측면에서도 새로 기여하는 바는 거의 없을 것이다. 다만, 편재하고 있는 정보들을 종합적으로 정리하고 소개하여 자료 적합에 관심 있는 국내 학생들과 연구자들에게 길잡이가 되고자 한다.

지금까지 엑셀을 이용한 자료 적합 논문들

은 방법론적으로 크게 두 가지로 압축된다. 첫 번째는 원자료에 모델을 적용하여 적합할 때 최소제곱 추정을 이용한 시연인데 대부분의 논문들이 여기에 해당 한다(Brown, 2001, 2006; John, 1998; Kemmer & Keller, 2010; Lambert, Mytilinaios, Maitland, & Brown, 2012). 두 번째는 최대우도 추정을 이용한 자료 적합(Dodson 등 1998; Dunn, 2010)으로 엑셀로 이를 소개한 경우는 현재까지 매우 적으며 이 또한 모델의 적합도 검증(goodness of fit test)이 목적이어서 흔히 심리학 실험실에서 조건간 비교를 위한 목적의 자료 적합과 거리가 있다. 반면, MATLAB이나 R과 같은 코드를 이용하여 최대우도 추정으로 자료 적합을 배울 수 있는 논문은 쉽게 찾아볼 수 있다(예: Myung, 2003).

본 글에서는 실제 실험실에서 얻은 자료에 대해서 엑셀로 적합하는 과정을 안내한다. 자료 적합 과정에서 필연적으로 원자료에 상정된 심리측정 함수를 적용하는 과정이 필요한데 최소제곱 추정과 최대우도 추정이 흔하게 쓰이므로 이 두 방법을 어떻게 이용하는지를 차례대로 살펴본다. 먼저 전체적인 이해를 돕기 위해 자료 적합하는 과정에 대해서 간략하게 살펴보자.

자료 적합에 대한 짧은 소개

자료 적합의 필요성

가령 아주 간단한 예로, 어떤 연구자가 사람이 어느 정도 세기의 불빛을 알아 볼 수 있는지 관심 있다고 하자. 성인 10명을 모아 어두운 방에서 모니터에 불빛을 무작위로 14개 수준의 세기로 제시하면서 불빛이 보이는지 안 보이는지를 묻는 실험을 진행하였다. 불빛이 아주 약하거나 없으면 보았다는 응답은 적을 것이고 불빛이 강해지면 보았다는 응답이 높아질 것이다. 응답은 불빛의 각 세기 수준에서 100회씩 하였고 '예'라는 응답으로 응답률이 계산되었다. 그림 1은 이 실험에 대한 가상적인 결과를 보여준다. 먼저, 왼쪽의 그림에서 각 수준에 대해서 평균 '예' 응답률이 점들로만 표시된 경우를 생각해 보자. 대략적으로 불빛 세기가 점점 증가함에 따라 불빛을 보았다는 응답도 증가함을 알 수 있다. 따라서 이 연구자는 단순히 불빛 세기와 반응율의 관계 자체에 관심이 있다면 두 변인 간 회귀

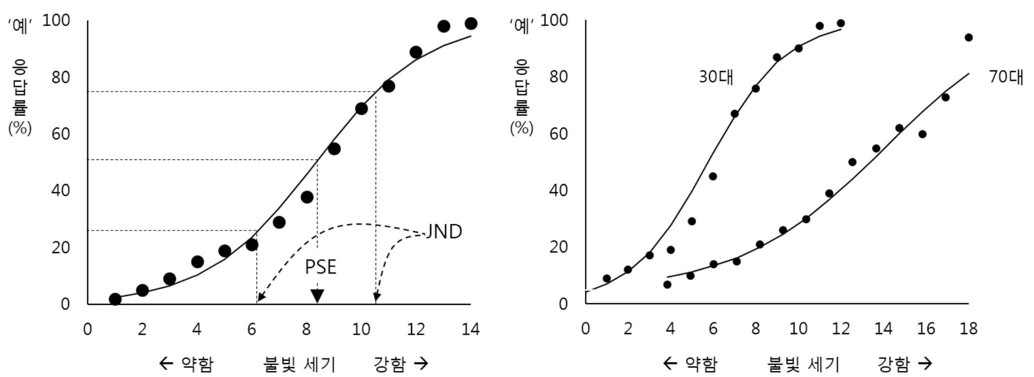


그림 1. 불빛 탐지에 대한 가상적인 실험 결과. (왼쪽) 불빛의 세기가 커질수록 보았다는 응답이 증가한다. (오른쪽) 두 연령 집단에 따른 불빛 탐지율 분포.

분석으로 통계적 검증을 하고 연구를 끝낼 것이다. 그렇지만 만일 결과 분포에 대해서 특정 모델을 적용하거나 결과를 집단간 비교를 할 목적이라면 결과 분포를 수학적으로 기술해야만 한다.

바로 이런 경우 심리측정 함수가 필요하다. 심리측정 함수(psychometric function)란 독립변수와 종속변수의 관계를 표상하는 수학적인 기술을 말하며 이 함수를 적용하는 과정을 자료적합이라 한다. 그림 1의 왼쪽에서 원자료 위에 실선으로 제시된 곡선이 심리측정 함수의 한 예이다. 심리측정 함수를 적용하는 과정은 통계학적으로 매우 까다롭고 정해진 시금석이 없기 때문에 많은 논쟁이 진행 중이며 계량심리학자들은 대부분 타당성 있는 심리측정 함수를 적용하는 방법 자체에 관심이 있다.

그렇지만 일반 실험실 연구자들은 그 보다는 심리측정 함수를 적용했을 때 얻게 되는 파라미터에 관심이 있다. 심리측정 함수를 이용하여 다양한 파라미터를 구할 수 있는데 여기에는 반응율의 50% 지점에 해당하는 역치('yes/no' 실험법에서 threshold) 또는 주관적 동등점(범주 1/범주 2' 실험법에서 PSE: point of subjective estimation), 반응율의 25%에서 75% 사이에 해당하는 자극 세기의 차이를 뜻하는 최소가지차(JND: just noticeable difference), 분포의 기울기(slope), 분포의 상한값과 하한값 등이 있다.

연구자는 이 파라미터들을 이용하여 변인들의 효과 또는 집단간 비교를 통계적으로 검증할 수 있다. 예를 들어, 만일 이 연구자는 불빛 탐지에 대한 연령 효과에 관심이 생겼다는 가정 아래, 30대와 70대 각각 10명씩을 모아 앞에서 진행했던 실험과 똑같이 실험을 하여 그림 1의 오른쪽과 같은 자료를 얻었다고 하자. 대략적으로 두 집단에서 '예' 응답이 50%르 보이는 지점을 살펴보면 30대 집단은 불빛 세기가 6 정도에서 70대 집단은 13 정도에서 형성되어 있다. 이와 더불어, 30대와 70대의 자료 분포의 기울기에서도 차이가 있는 것을 대략적으로 알 수 있다. 이런 차이를 살펴보는 데 두 집단의 평균 반응률은 비슷하므로 이를 비교하는 것은 의미가 없다. 그 보다 두 집단의 반응 분포에 대해서 자료 적합을 한 후 가령 반응이 50%를 보이는 지점을 말해주는 파라미터들을 구하고 이들을 대상으로 집단간 t 또는 F 검증을 하는 게 필요하다.

자료 적합 절차

자료 적합을 독립 변수와 종속 변수 간의

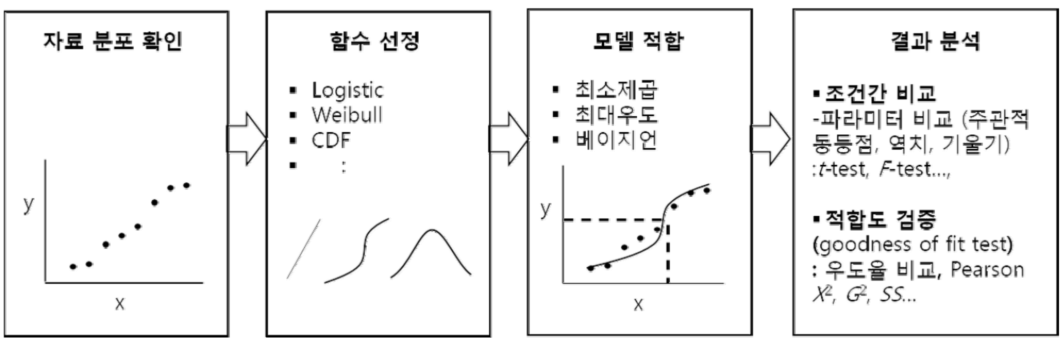


그림 2. 비선형 자료 적합의 일반적인 순서

관계에 그래프를 부여하는 과정이라는 관점에서 선형 자료는 선형 회귀식(linear regression)을 부여 받으므로 선형 적합(linear fitting)으로 비선형 자료는 곡선 회귀식(non-linear regression)을 부여 받으므로 곡선 적합(curve fitting)이라고 불리기도 한다. 그림 2는 일반적으로 비선형 자료를 적합 할 때 진행되는 순서를 보여준다.

가장 먼저 원자료를 독립변수와 종속변수에 따라 배열하여 분포를 확인한다. 둘째, 배열된 자료를 살펴보고 함수를 선택하는데 해당 분야에서 널리 사용하는 함수를 참고하기도 한다. 여기서 함수는 심리 처리 과정에 대해서 연구자가 가정한 것이므로 모델이라고 불리기도 한다. 셋째, 함수가 선택되었다면 이를 원자료에 좀 더 정밀하게 적합 하는데 여기에서 최소제곱 추정 또는 최대우도 추정을 이용하여 최적화된 적합 함수를 얻는다. 함수 적용과 적합 과정은 마치 옷을 사서 수선하는 과정과 비슷하다. 예를 들어, 졸업식에 참석할 바지를 사고 싶어 백화점이나 마트에 갔을 때, 먼저 자신의 몸에 적당히 맞는 바지를 골라야 하는데 이 과정을 함수 선정에 비유할 수 있고 바지를 구입 후 기장이나 통 등을 좀 더 정밀하게 자신의 몸에 맞추는 것을 적합 과정에 비유할 수 있을 것이다. 마지막으로 이 적합 함수를 이용해서 조건간 비교를 할 수 있는데 분포의 속성인 역치, 주관적 동등점, 기울기 등의 파라미터(parameter)들을 조건의 개인별로 계산하여 t와 F 같은 관습적인 통계 검증을 한다.

이와 함께, Logistic 또는 Weibull 등 적용된 함수 또는 모델 자체가 원자료를 얼마나 신뢰롭게 설명하는지를 살피기 위해 적합도 검증을 하기도 한다. 그렇지만, 모델 검증이 아닌 조건간 비교를 위한 가설 검증을 목적으로 하는 실험에서 적합도 검증은 꼭 필요한 과정은 아니다. 적합도 검증은 그 자체로 매우 까다로운 이론적 주제이므로 이에 관해서는 좀 더 전문적인 자료를 찾아보기를 권한다(예: Kingdom & Prins, 2010, 8장; Wichmann & Hill, 2001). 본 글에서는 원자료에 함수를 적용하는 과정과 적합 함수에서 파라미터를 계산하는 일에 대해서 집중적으로 설명하고자 한다. 본 글에서는 변수와 파라미터를 같은 의미로 사용하고자 한다.

준비 작업

해 찾기 설치

엑셀에 자료 적합을 위해 마련된 별도의 옵션이 따로 있는 것은 아니다. 다만, 엑셀이 가지고 있는 여러 기본 기능들을 조합해 적합을 수행하는 것이다. ‘해 찾기(solver)’ 기능은 가장 중요한 옵션이다. 해 찾기란 알려지지 않은 방정식의 해를 구하거나 해의 조건을 입력해주면 그 범위 안에서 반복적으로 수치를 넣어 근사 값을 내놓는 기능을 한다. 예를 들어, ‘$10 = 2x + 2$’라는 방정식이 있다고 하자. x의 값은 4라는 것이 아주 쉽게 머리에 떠오를 것이다. 그렇지만, 변수와 항이 많아지고 차수가 높아지고 지수, 로그 함수와 같은 식이 들어가면 계산기로도 풀기 어렵게 될 것이다. 이런 경우, 엑셀의 해 찾기 프로그램은 마우스로 찾아야 할 셀만 지정해주면 순식간에 해를 구해준다. 엑셀에서는 방정식을 풀어주는 기능으로 ‘목표 값 찾기’가 있다. 이는 데이터 탭 메뉴에서 가상분석 아래 메뉴에 위치해 있다. 목표 값 찾기는 변수 또는 미지의 값이 하나일 때만 그 해를 찾아준다. 반면, 해 찾기는 목표 값 찾기 기능을 넘어서서 두 개 이상

의 변수 또는 미지의 값에 대해서도 최대한 근사한 값을 찾아주는 기능도 가지고 있다. 자료 적합은 보통 찾아야 할 파라미터가 두 개 이상이므로 해 찾기를 이용해야만 한다. 엑셀에서 해 찾기를 이용하기 위해서 먼저 이 기능을 설치해야만 한다. 이 과정이 필요한 이유는 컴퓨터에 엑셀이 설치되어 있을지라도 추가 기능이어서 기본 설치에는 빠져있기 때문이다.

그림 3은 엑셀 2013 버전에서 해 찾기를 설치하는 모습을 보여준다. 상단 메뉴에서 (1) ‘파일(File)’ 버튼을 누르면 (2) ‘옵션(Options)’ 창이 뜨는데, 여기에서 (3) 추가 기능(add-ins)을 지정하고, (4) 오른쪽에서 ‘해 찾기 추가(solver add-in)’에 마우스 커서를 위치해 클릭하고 (5) 아래 ‘관리(manage)’ 항목에서 ‘Excel 추가 기능’ 옵션에서 ‘이동’ 버튼을 누른다. (6) 추가 기능 팝업 창이 뜨고, 여기에서 ‘해 찾기 기능(solver add-in)’을 선택한 후 ‘확인(OK)’ 버튼을 누른다. ‘분석 도구’를 설치하는 것도 권고한다. 분석 도구는 데이터를 통계 분석해주는 추가 기능으로 상관, 회귀, z, t, 분산분석 등이 포함된다. 잠시 프로그램이 재 설치되면서 시간이 조금 지나고 나면 엑셀 상단의 탭 메뉴 (7) ‘데이터(DATA)’ 옵션 맨 끝에 ‘데이터 분석’과 해 찾기’ 단추가 나타나 있는 것을 확인할 수 있다.

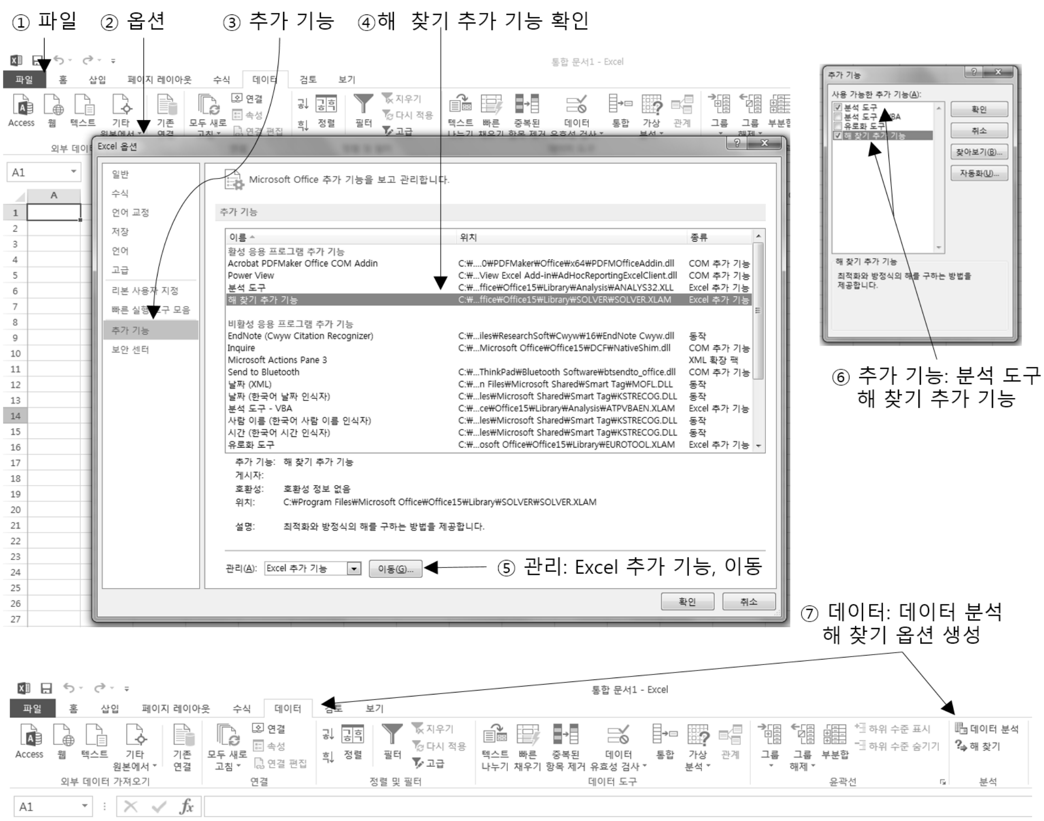


그림 3. 엑셀에서 해 찾기와 분석 도구를 추가하는 과정

엑셀의 기본 기능 익히기

엑셀로 자료 적합을 하기에 앞서 꼭 필요한 엑셀 사용의 기본 기능들을 살펴보자.

원자료 입력. 엑셀에서 원자료의 입력은 스프레트시트의 셀 단위로 이뤄진다. 셀의 명명은 세로열은 알파벳으로 가로열은 아라비아 숫자로 이뤄진다. 예를 들어, C5는 세 번째 세로열과 다섯 번째 가로열이 겹치는 셀다. 셀 하나에는 자료 하나가 입력되어야 한다. 그림 4에서, 독립변인이 X로 종속변인이 Y1, Y2, Y3로 나뉘어 임의의 원자료가 입력되어 있다. 자료 입력은 손으로 직접 할 수도 있고 이미 코딩된 값을 다른 문서에서 가져올 수도 있다.

수식입력. 그림에서, Y3는 Y1과 Y2의 평균을 내고 특정 변수 값 'a'를 곱한 값이다. 수식입력은 반드시 '=' 기호가 맨 앞에 위치해야 하며 그 뒤에 수식을 입력하고 엔터를 치면 완성된다. 엑셀의 주요 강점으로 끌어 내리기 기능이 있다. 이는 수식이 입력된 어떤 셀에 마우스를 갖다 대고 오른쪽 하단에 마우스 커서를 옮기면 커서가 십자 표시(+)로 바뀌고 이 때 마우스 왼쪽 버튼을 누른 채 밑으로 끌어 내리면 첫 번째 셀의 수식이 자동으로 복사되는데 참조 셀의 대상도 역시 함께 바뀐다. 예를 들어, Y3의 첫 번째 셀에 수식을 넣고 이를 끌어 내리기 하면 나머지 전체에 이 수식이 적용된다.

엑셀의 주요 수식 기호는 다른 프로그램에서 사용하는 기호와 비슷하다. 덧셈, 뺄셈, 곱하기, 나누기는 각각 +, -, *, / 이며, 제곱은 자판의 숫자 6 위에 있는 ^을 쓴다. 계산에서 자주 하는 실수는 곱의 기호인 *입력 여부이다. 예를 들어, x를 변수로 지정해 놓고 =4x+1을 표현할 때 4와 x 사이에 *를 반드시 넣어야 한다. 그 이외의 다양한 함수들은 도움말을 이용하여 익힐 수 있다.

이름 정의. 이름 정의는 프로그래밍에서 변수 정의와 같은 과정으로 문자로 표시된 변수 또는 파라미터가 특정 셀에 들어있는 숫자

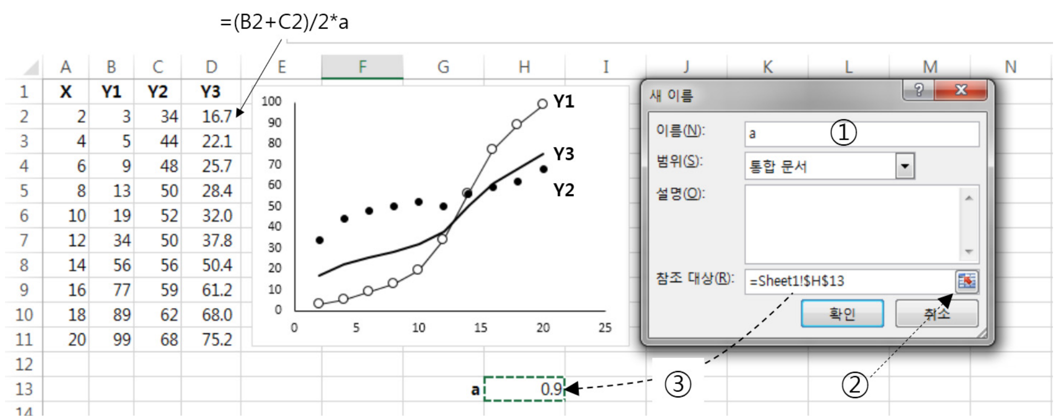


그림 4. 엑셀에서 그래프 그리기와 이름 정의하기. 그래프 형태는 '데이터 계열 서식'을 이용하여 다양하게 표현하였다. 이름 정의 순서: (1) 메뉴에서 '수식'을 활성화 시키고 이름 정의 창을 띄운다. 이름에 특정 변수 이름을 넣는다. (2) 변수의 대상이 되는 셀을 지정하는 단추를 누른다. (3) 대상 셀을 지정해 준다.

와 연결된다. 즉, 만일에 a 변수에 연결된 셀에 4가 들어 있고 b 변수에 연결된 셀에 2가 들어 있는 경우 a+b를 하면 4+2가 되어 6이 된다. 이렇게 하는 이유는 계산이 많은 경우 일일이 숫자를 넣지 않고서도 이름으로 정의된 변수 값만 바꾸면 계산이 입력 실수 없이 편리하게 이뤄지기 때문이다.

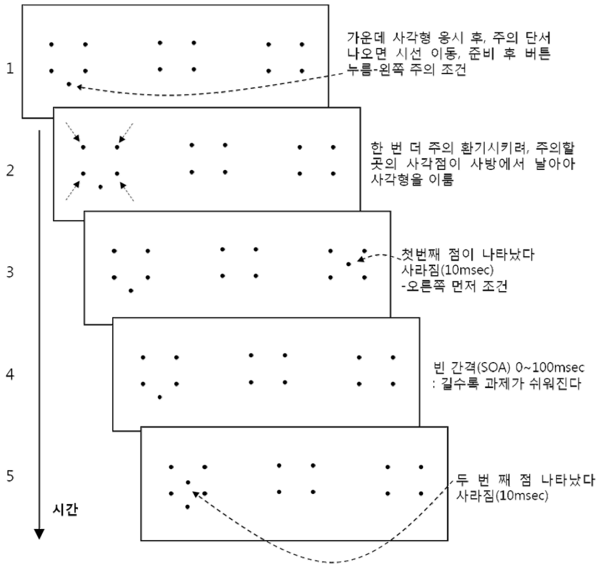
여기서 a는 0.9를 지정하였다. 이름 입력 절차는 엑셀 메뉴에서 '수식'을 클릭하고 하위 메뉴에서 '이름 정의'를 클릭한다. 그러면 그림 4의 맨 오른쪽에 보이는 옵션 창이 뜨며 여기에 이름과 대상이 되는 셀을 지정하는 것으로 이뤄진다.

그래프 그리기. 엑셀의 그래프 그리기는 메뉴에서 '입력'에서 '차트'에서 이뤄진다. 그리고 싶은 원자료의 범위를 마우스로 지정한 후 차트를 삽입한다. 그림 4에서는 X, Y1, Y2, Y3 모두를 이름과 자료를 모두 범위로 지정하였다. 만일, 원자료의 일부분만 그래프로 나타내고 싶다면, 예를 들어 Y1과 Y3만을 그리고 싶다면, Ctrl을 누른 채 마우스로 해당되는 범위를 지정한 후 그래프를 삽입한다. 그래프의 다양한 옵션을 예시하기 위해 그림 4에서 Y1은 자료 위치가 점과 선으로, Y2는 점으로, Y3는 선만으로 표시하였다. 표시 방법은 챠트에서 원하는 선 위에 마우스 왼쪽 클릭하여 지정 후 오른 클릭하여 나타난 메뉴에서 '데이터 계열 서식' 옵션을 조정하여 바꿀 수 있다. 아래 자료 적합에서 원자료는 점으로 표시하고 적합 자료는 곡선으로 표시하는데 이는 테이터 계열 서식을 변화시켜서 가능한 것이다. 본 글에서 제시된 모든 적합 곡선은 이 방법을 이용하여 표현한 것이며, 본 글의 모든 수식 그림은 엑셀에서 자료를 구하고 파워포인트로 옮겨 편집 후 그룹으로 지정하고 이를 그림으로 저장하여 한글에서 불러들인 것이다.

자료 준비

자료 적합은 독립변인과 종속변인 사이의 관계를 수학적으로 기술하는 것이므로 자료를 구성하는 독립변인과 종속변인이 연속 변인으로 표현될 수 있어야 한다. 이런 기준에 속하는 Stelmach 와 Herdman(1991)이 시행한 실험을 인용하자. 두 사건이 일어날 때, 어느 사건이 먼저이고 어느 사건이 나중에 일어났는지를 판단하는 문제는 일상에서 흔하게 일어난다. 두 사건 사이에 빈 시간이 길수록 판단이 쉽고 빈 시간이 매우 짧다면 두 사건 가운데 어느 사건이 먼저인지 판단하기 어려울 것이다. 그렇다면, 두 사건이 짧은 간격을 두고 일어날 때 주의는 어떤 역할을 할까? 직관적으로 생각해보면 두 사건의 순서를 좀 더 정밀하게 구분할 수 있을 것 같다. 정말일까? 그런데 이들의 연구에서 매우 짧은 두 사건이 연속으로 제시될 때, 나중에 일어난 사건이어도 주의를 받으면 먼저 일어난 것으로 지각되는 착시가 발견되었다.

이 실험에서 참여자들은 불빛이 어느 쪽에서 먼저 제시되었는지를 보고하는 시간 순서 판단 과제(temporal order judgment task, TOJ)를 수행하였다. 오실로스코프 모니터에 작은 점이 왼쪽 혹은 오른쪽에 차례대로 나타났다 사라지는데 참여자는 늘 가운데에 시선을 고정한 채 주어진 지시 단서에 따라 마음속으로 왼쪽 혹은 오른쪽에 주의하면서 과제를 수행하였다. 연구자들은 예비 실험을 통해서, 주의를 받은 위치의 불빛이 늦게 제시되어도 먼저 제시된 것으로 지각되는 착시를 알아냈고 본



	두 자극 제시 시간 차(msec)	오른쪽이 처음이라고 보고한 비율 %	
		오른쪽 주의 조건	왼쪽 주의 조건
왼쪽 점이 먼저 제시된 조건	-100	10	x
	-90	19	x
	-80	21	x
	-70	24	x
	-60	40	x
	-50	50	x
	-40	64	x
	-30	84	x
	-20	92	8
	-10	98	18
동시 제시 조건	0	97	16
오른쪽 점이 먼저 제시된 조건	10	99	20
	20	99	24
	30	x	26
	40	x	42
	50	x	62
	60	x	64
	70	x	78
	80	x	82
	90	x	88
	100	x	90

그림 5. (왼쪽) 실험 절차. 참여자들은 매 시행에서 가운데 네 점에 시선을 고정하고 단서에 따라 마음속으로 왼쪽 또는 오른쪽에 주의 하면서 네 점 가운데 나타난 불빛이 왼쪽이 먼저였는지 오른쪽이 먼저였는지 보고하였다. (오른쪽) 실험 결과: 오른쪽이 처음이라는 응답률을 기준으로 정리되어 있다. X는 조건이 없어 자료 없음을 뜻한다.

실험에서 이 현상이 나타나는 자극 제시의 시간 범위를 좁혀서 좀 더 면밀한 검토를 하였다. 그림 5에 제시된 것처럼 실험이 시작되면 화면 왼쪽 혹은 오른쪽에 작은 점이 10msec 정도 제시되었다 사라진다. 두 점 사이의 간격은 0~100msec 이었다. 여기서 0msec는 두 자극이 동시에 제시되었음을 뜻한다. 자극 사이의 시간적 간격은 10msec 단위로 늘어나 최고 100msec까지 길었다. 오른쪽 주의 조건의 경우 왼쪽 점이 최고 100msec 먼저 제시되었고, 오른쪽 점이 먼저인 경우는 최고 20msec 만큼만 빨랐다. 왼쪽 주의 조건은 그 반대로 오른쪽이 먼저 제시된 것은 최고 100msec이었고 왼쪽에 제시된 것은 20msec 정도까지만 빠른 것이 제시되었다.

모델 및 적합 방법 선정

그림 5의 오른쪽에 제시된 자료를 그래프로 표현하면 그림 6과 같다. 두 가지 주의 조건에서 두 자극의 제시 시간 차이가 증가함에 따라 '오른쪽이 처음'이라고 응답한 비율 역시 증가함을 알 수 있다. 이제 이 자료에 심리측정 함수 또는 모델을 적용하자. 로지스틱 함수(Logistic function)는 반응이 yes-no와 같은 이진수(binary/dichotomous) 자료 또는 범주 1과 2와 같은 범주(categorical) 자료를 적합할 때 여러 분야에 걸쳐 폭넓게 사용된다(Afifi, May, Clark, 2012). 시간 순서 판단 과제 역시 매 시행의 반응이 왼쪽 또는 오른쪽 어느 한 쪽을 고르는 자료이므로 로지스틱 함수를 적용하여 적합을 해보자. 로지스틱 함수는 S 형태를 보이는 시그모이드(sigmoid) 함수인데 그림 6의 그래프 역시 이런 형태임을 주목하라.

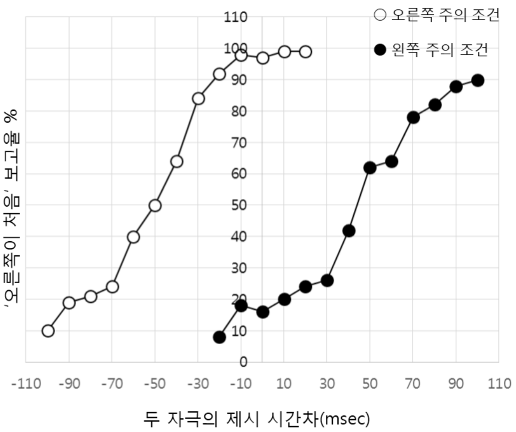


그림 6. 주의 위치에 따른 오른쪽 처음 보고 비율. 가로축에서 −는 실제로 왼쪽 불빛이 먼저 +는 오른쪽이 먼저 켜짐을 0 지점은 동시에 불빛이 켜짐을 뜻한다. 오른쪽을 주의했을 때 오른쪽이 먼저 불이 켜졌다고, 또는 왼쪽을 주의하고 왼쪽 불빛이 먼저 켜졌다고 보고하는 경향이 있었다.

실제 사용되는 로지스틱 함수는 물리학 공학 생물학 심리학 등 분야에 따라 조금씩 다르다. 그 이유는 원자료의 패턴이 다를 수 있고 관심 있는 파라미터가 달라 이를 좀 더 다루기 편리하게 부각되도록 하기 위함이다. 따라서 관련 분야에서 쓰이는 함수가 무엇인지 참고할 필요가 있다. 식 (1)은 심리학에서뿐만 아니라 다른 여러 분야에서도 널리 쓰이는 로지스틱 함수이다.

$$f(x_i|\alpha,\beta)=\frac{K}{1+\exp^{-\beta(x_i-\alpha)}} \quad (1)$$

파라미터 α와 β가 주어졌을 때 i번째 x의 값을 표현한다. 이 식에서 분자 K는 그래프가 오를 수 있는 최대 높이이다. 만일 종속변수의 원자료가 확률 분포이고 최댓값이 1인 경우에 K는 1이 되며, 원자료가 최댓값이 100인 비율이라면 이 공식에 분자가 100이 되어야 한다. α는 Y축의 K/2 지점에 대응되는 x 축의 값이며, β는 그래프의 기울기와 관련된 파라미터이다. 여기서 exp는 e로도 표현되기도 하는 오일러(Euler)의 수로 약 2.718281828. 그림 7은 로지스틱 분포에서 파라미터들이 갖는 의미를 나타낸 것이다.

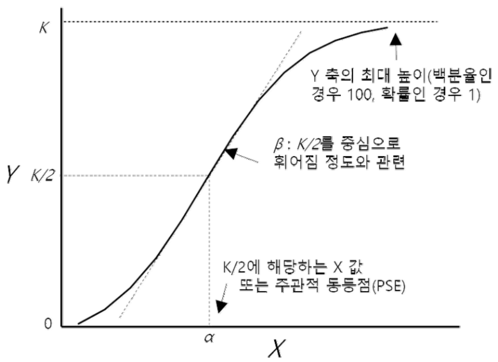


그림 7. 로지스틱 함수 그래프에서 각 파라미터의 분포 특징들.

최소제곱 추정을 이용한 적합

비선형 자료는 분포가 복잡해 보이지만 최소제곱 추정을 이용하는 경우 적합 원리는 선형자료를 적합하는 원리와 동일하다. 즉, 원자료에서 평균적인 거리가 가장 짧은 회귀 곡선을 찾는 것이 목표이다.

최소제곱합

그림 8은 원자료에 대해 최소제곱합이 가장 적은 모델 곡선을 구하는 개념을 보여준다. 이미 알고 있는 원자료의 각 지점에서부터 모르고 있는 모델의 각 지점을 구하기 위해서 모델의 지점들을 임의의 변수로 지정하고 이로부터 실제 원자료까지 거리의 제곱합이 최

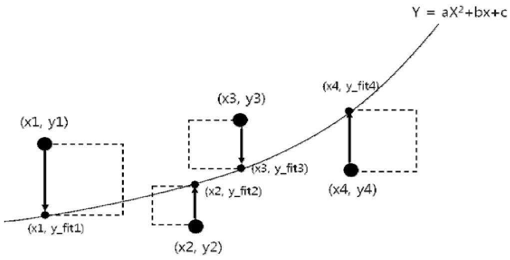


그림 8. 비선형 자료에서 제곱합을 구하는 기본 개념. 네 개의 큰 점들은 원자료 작은 점들은 적합 곡선에 위치한 지점들을 의미한다.

소가 되는 곡선을 구한다.

그림 8을 예로 들어 조금 더 자세히 설명하면, 여기에 네 개의 원자료 지점이 각각 (x1, y1), (x2, y2), (x3, y3), (x4, y4)로 표시되어 있다. 이들을 지나는 모델 곡선이 그어져 있고 원자료에서 수선을 내린 지점이 각각 (x1, y_fit1), (x2, y_fit2), (x3, y_fit3), (x4, y_fit4)으로 표시되어 있다. 여기서 모델 곡선의 자격은 각 원자료의 점들로부터 내린 수선의 길이의 총합이 가장 작아야 한다. 수학적인 편리를 위해 직선 길이의 합을 쓰는 대신 제곱해준 면적을 모두 더한 제곱합(SS, sum of squares)이 사용된다. 제곱합 공식은 각 x 값에 해당하는 y 값과 이 y값에 대한 가상적인 y_{fit} 값을 이용해 적으면 식 (2)와 같다(Σ는 누적합을 뜻한다).

$$SS = \sum_{i=1}^{n}(y_i - y\_fit_i)^2 \tag{2}$$

$$SS = \sum_{i=1}^{n}(y_i - \frac{100}{1+\exp^{-\beta(\chi_i-\alpha)}})^2 \tag{3}$$

그런데 원자료에 로지스틱 모델을 적용하고자 하므로 이 식에서 모델 값에 해당하는 항을 로지스틱 함수와 치환할 필요가 있다. 식 (3)은 이를 정리한 것이다. 이 실험의 자료는 비율이므로 분자 K는 100으로 놓는다. 최종적으로 이 최소제곱 값을 가장 작게 만드는 파라미터 α와 β를 찾아야 한다.

모델 적용

이제 로지스틱 함수를 원자료에 최소제곱 추정을 이용하여 단계별로 적합을 해보자.

원자료 입력. 원자료 값들을 엑셀의 셀들에 하나씩 입력하는 일로 그림 9의 왼쪽에 예시되어 있다. 독립변인인 두 불빛이 제시되는 시간차가 맨 첫 번째 세로열에 놓여있고 종속변인이 '오른쪽이 처음'이라는 보고를 중심으로 반응률이 계산되어 오른쪽 주의 조건과 왼쪽 주의 조건으로 나뉘어 두 번째와 세 번째 세로열에 입력되어 있다.

이름 정의. 파라미터 α와 β를 대신하여 a, b를 넣어 이름 정의하는데 이는 앞에서 로지스틱 함수를 구성하는 두 개의 파라미터a와 b에 대응한다. 적당한 위치에 셀 네 개를 정하고서 세로열에 a와 b를 적고 각 오른쪽 셀에는 임시로 '1'을 넣어주자(1이 아닌 다른 어떤 숫자를 넣어도 되고 비어 있어도 관계없다). 그 다음 각 '1'이 들어있는 셀을 마우스로 클릭 한 상태에서 엑셀 탭 메뉴의 '수식'아래 '이름 정의' 버튼을 누르고, 활성화된 창에서 먼저 '이름'에는 a를 넣고 '참조 대상'에는 1이 들어 있는 셀을 마우스로 지정해준다. 그 다음 b에 대해서도 같은 과정을 거쳐 지정해준다. 이와 함께 제곱합을 뜻하는 SS도 이름을 정의하고 참조 대상은 바로 오른쪽 셀을 지정해준다. 그림 9 아래에 세 개의 문자 a, b, SS가 각각 나와 있고 오른쪽에는 아직 시행하지는 않았지만 최종적으로 계산된 결과가 나타

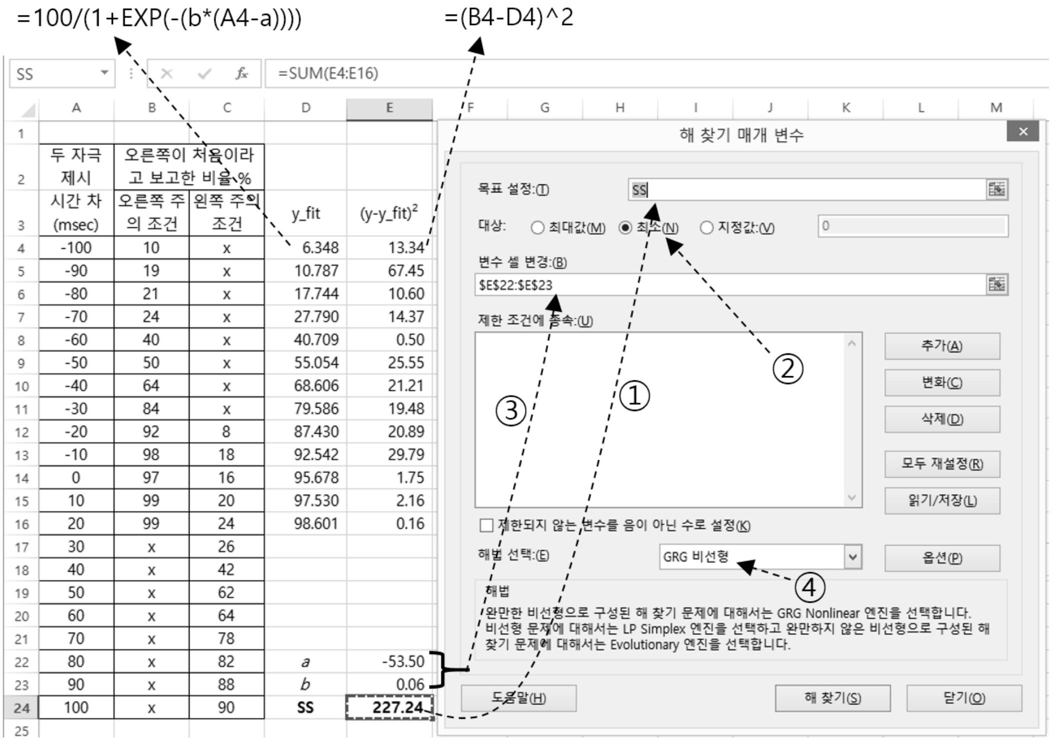


그림 9. 비선형 적합을 위해 오른쪽 주의 조건에 대해서 해 찾기를 수행한 모습. (1) 목표 설정에 SS를 넣는다. (2) 이를 최솟값이 나오도록 지정한다. (3) 변수 셀 변경란에 파라미터 a와 b의 연결 셀들을 지정해준다. (4) 해법 선택에서 GRG 비선형을 선택한다.

나 있다.

모델 적합. 모델 곡선의 x 값에 해당하는 y값을 넣어 준다. 그림 9에서 볼 때 D열의 네 번째 항에 ‘=100/(1+EXP(-(b*(A4-a))))’을 입력한 후 ‘엔터’ 버튼을 눌러 값을 확인한다. 이 셀의 서식을 마우스로 끌어 내리기 기능을 이용하여 아래 나머지 12개 셀 모두에 적용한다.

제곱합 계산. 원자료 지점과 모델 곡선 사이의 제곱합을 구한다. E열의 4번째 항에 ‘=(B4-D4)^2’를 넣는데, ^ 기호는 제곱을 뜻한다. 이 식은 B열의 4 번째 셀의 값에서 D열의 4 번째 셀의 값을 뺀 후 제곱을 하라는 뜻이다. 이를 역시 아래 모든 항목에 마우스를 끌어 적용한다. 마지막 셀에는 ‘=SUM(E4:E16)’을 넣는데, 이는 E열의 4번 째 셀부터 16번째 셀까지 모두 더하라는 뜻으로 제곱합이다.

해 찾기. 이제 제곱합이 최소가 되는 파라미터 α와 β를 찾는다. 이를 엑셀의 해 찾기 기능을 이용해 구해보자. 먼저, 탭 메뉴에서 ‘데이터’ 아래 놓인 해 찾기’ 버튼을 눌러 해 찾기 창을 활성화 시킨다. 이는 그림 9의 오른쪽과 같다. 목표 설정란에는 SS 또는 참조셀을 지정하고 대상은 최소로 지정한다. 대상을

최소로 지정한다는 뜻은 SS 값이 최대한 작은 값이 되도록 a, b 값을 정하라는 뜻이다. 다음으로 '변수 셀 변경'란에 이름 정의된 a와 b의 참조셀을 각각 지정한다. 이 셀에 있는 값은 나중에 해 찾기가 끝나면 자동으로 바뀌게 된다. '해법 선택' 옵션에는 GRG 비선형을 선택하는데 이 옵션은 비선형 자료에 대해서 해를 구할 때 사용한다. 엑셀의 해 찾기는 비선형 자료에 대해 Lasdon 등(1978)이 개발한 Generalized Reduced Gradient(GRG2) 알고리듬을 사용하는데 이 알고리듬은 각 파라미터의 범위를 정해주는 것을 원칙으로 하고 있다. 최소제곱 추정처럼 해 찾기 계산이 간단한 경우 파라미터들의 범위를 정해주지 않아도 해를 찾기도 하지만 최대우도 추정처럼 복잡한 경우 범위를 정해주어야 좀 더 빨리 해를 구할 수 있다. 끝으로, 아래에 있는 '해 찾기' 버튼을 누르면 그림 9에 나타난 것처럼 a는 -53.50으로, b는 0.06, SS는 227.24로 바뀐 것을 알 수 있다.

파라미터 α와 β의 값을 구할 때 주의할 점은 초기 값이다. 이 예에서는 1을 초기 값으로 삼았는데 초기 값이 너무 엉뚱하면 해를 구하는 데 실제보다 작은 값을 구하게 되거나 실패할 수도 있다(이에 관해서는 마지막 세션에 소개된 최적화 문제를 살펴보기를 바란다). 예를 들어, 이 예에서 b의 초기 값에 100을 넣어보면 해를 찾는 데 실패할 수 있다. 이는 해 찾기가 초기 값을 중심으로 결과 값을 추정하기 때문이다. 아쉽게도 이를 피할 수 있는 정해진 규칙이란 없다. 다만, 여러 가지 값을 넣어 결과가 같은지 또는 해가 잘 구해지는지를 반복해서 확인하는 것뿐이다. 따라서 초기 값을 임의로 넣을 때 파라미터의 값을 어림짐작해서 넣는 게 유리하다. 예를 들어, 파라미터 α는 Y축의 1/2 지점에 해당하는 X축의 값인데 원자료를 살펴보고 대략 -50 정도이므로 이를 초기 값으로 삼는다. 다음으로, β는 그래프의 기울기와 관련되는데 손으로 몇 개를 넣어가면서 제곱합의 크기를 살펴본다. 경험적으로 대략 0.05 수준에서 크게 벗어나지 않는데, 그래프를 보기 전에 SS 값의 크기를 보아 가면서 가장 작게 나타나는 값을 대략적으로 찾아 이를 초기 값으로 정한다. 이 과정을 충분히 거친 후에 해 찾기 기능을 이용하여 정밀한 값을 찾는다.

이 과정을 왼쪽 주의 조건에 대해서도 똑같이 적용하면 a는 44.62, b는 0.04 그리고 SS는 246.34로 구해진다. 이름 정의에서 같은 이름이 두 개 이상 있으면 서로 충돌하므로 다른 이름을 넣어야 한다. 이렇게 얻은 변수 값을 적용해 두 주의 조건에 해당하는 적합 선분을 그려보자. 이는 원자료의 종속변인을 이루는 각 수준의 값들은 점으로 표현하고 이와 함께 Y_fit에 해당하는 값들을 그래프에 추가하되 이번에는 점이 아닌 곡선으로 표현하고 모든 표식을 모두 지워서 구현할 수 있다. 그림 10은 원자료에서 독립변인인 제시 시간 차이를 가로축에 그리고 종속변인인 오른쪽이 '처음'이라고 보고한 비율을 세로축으로 하여 원자료와 모델 곡선을 함께 그린 것이다. 오른쪽 주의 조건과 왼쪽 주의 조건 모두 원자료와 모델이 매우 비슷하게 배열되어 있음을 알 수 있다.

파라미터 계산

이제 두 조건의 자료에 대해서 최소제곱 추정을 이용한 비선형 적합을 완성하였다. 남은 일은 통계 검증을 하기 위해 두 집단이 얼마나 큰 차이가 있는지를 살펴보는 일이다. 주

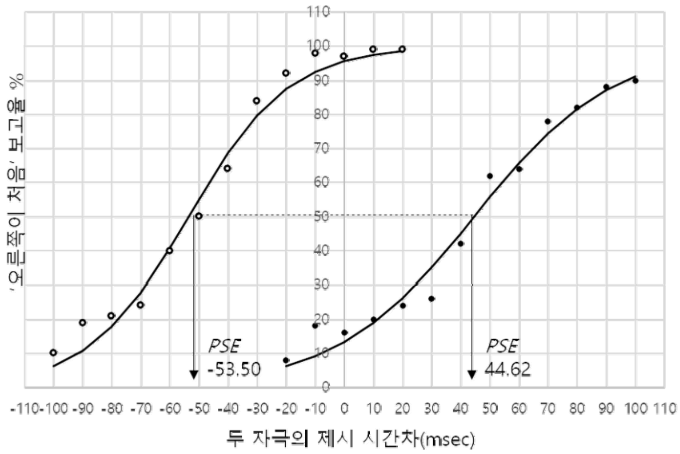


그림 10. 오른쪽 주의 조건과 왼쪽 주의 조건 두 조건의 원자료에 비선형 적합을 적용한 결과. 원자료의 값들은 점으로 표현하고, 모델 값들은 그래프 옵션에서 '윤곽선'을 지정하여 선분으로 연결하였다. 화살표는 주관적 동등점(PSE)을 가리킨다.

관적 동등점(PSE: Point of Subjective Equality)이란 자극의 지각된 크기가 같다고 여겨지는 지점을 뜻한다. 이는 일반적으로 반응 확률이 0.5인 지점 또는 반응 비율이 50%인 지점으로 삼는다. 그림 10에서, 각 집단의 주관적 동등점이 표시되어 있다. 대략적으로 두 집단에서 차이가 있지만, 통계적 검증을 위해서는 정확한 값을 구할 필요가 있다. 이 값은 표 1에 제시된 식을 이용해서 구할 수 있다. 윗변이 100이므로 아랫변을 2로 만들어 주면 Y 값이 50이 된다. 이를 위해서는, 아랫변에서, 왼쪽 항의 1을 제외한 나머지 항의 값을 1이 되도록 하는 것이다. 이를 위한 가장 간단한 방법은 어떤 수의 0의 제곱은 1이 되므로, b(x-a)의 값을 0이 되도록 하면 된다. 즉, a와 동일한 값이 바로 주관적 동등점이다. 따라서 따로 계산할 필요 없이, 그림 10에 제시된 것처럼, 오른쪽 주의 조건의 주관적 동등점은 -53.50이고, 왼쪽 조건의 주관적 동등점은 44.62이다. 그렇지만, 때때로 이 값들을 직접 계산하는 방법도 필요하다. 그 이유는 실험 목적에 따라 Y의 1/2 지점이 아닌 다른 지점에 대한 x 값을 구할 경우가 있기 때문이다. 예를 들어, 최소가지차이(JND: Just-Noticeable Difference)는 반응율 25% 지점과 75% 지점에 해당하는 x 값들의 차이를 2로 나누어서 구하는데 이런 경우 계산이 필요하다.

앞서 '해 찾기'로 찾아 낸 변수들을 표 1에 제시된 식에 적용하되 Y축의 50%에 해당하는 지점이므로 f(x)를 50으로 놓고 이 두 방정식

표 1. 주의하는 두 위치 조건에서 주관적 동등점(PSE)을 구하는 식

오른쪽 주의 조건	오른쪽 주의 조건
$50 = \dfrac{100}{1+\exp^{-0.06(x+53.50)}}$	$50 = \dfrac{100}{1+\exp^{-0.04(x-44.62)}}$

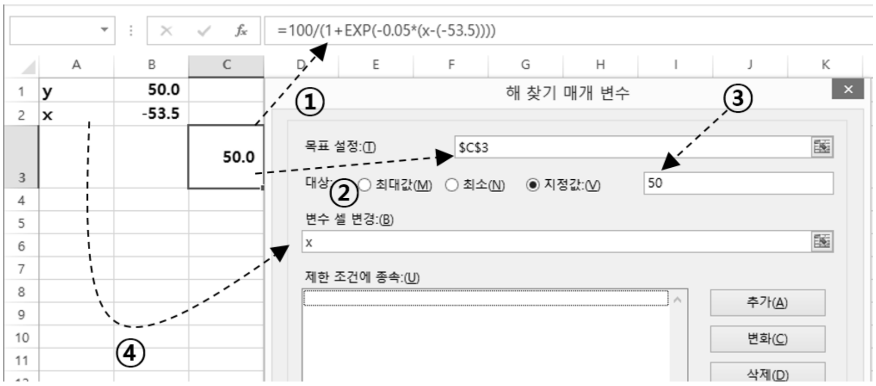


그림 11. 주관적 동등점을 찾기 위한 문제 설정과 '해 찾기' 기능 수행 절차. (1) 앞서 구한 파라미터 a와 b를 넣어 식을 적용한다. (2) 목표 셀을 지정한다. (3) 목표값인 50을 지정한다. (4) 변수 셀 변경에 x를 지정한다. 해법 선택은 GRG 비선형을 고른다.

을 만족하는 x를 구한다. 이 식을 풀 수 있는 기능은 엑셀에서 두 가지 방법이 있다. 하나는 앞에서 말한 것처럼 데이터 탭 메뉴에서 가상분석 아래에 위치한 목표값 찾기로 가능하다. 둘째, 다시 엑셀의 '해 찾기' 기능에 의존하여 풀 수 있다. 여기서도 해 찾기 기능을 통해 풀어보도록 하자. 실행 방법은 그림 11에 제시되어 있다. 먼저, x를 이름 정의 한 후 그 옆 칸을 영역으로 지정한다. 다음으로 오른쪽 주의 조건의 주관적 동등점을 구하기 위해, 표 1의 함수를 C3 셀에 넣는다. '해 찾기' 기능을 활성화하여 목표를 C3 셀로 지정해주고 목표값을 '50'으로 넣어준다. 다음으로 변화 변수를 x 값으로 지정해준다. 즉, C3의 결과 값이 50이 되도록 x 값을 조절하라는 뜻이다. 마지막으로 해 찾기 버튼을 누르면 x 값이 -53.50으로 바뀐 것을 확인할 수 있다. 이 방법을 그대로 왼쪽 주의 조건에도 수행하면 결과는 44.62 이다. 그러므로 오른쪽 조건의 주관적 동등점은 -53.50msec 이고, 왼쪽 조건은 44.62msec 이다.

최대우도 추정을 이용한 적합

전통적으로 역치를 재는 실험에서 참여자의 반응은 신호가 있는지 없는지(yes/no) 또는 신호가 어느 쪽이 더 세거나 약한지 등 두 가지 가운데 하나를 고르는 강제선택(forced choice) 과제가 주를 이루었다. 이런 실험에서 얻어진 자료는 이진수 자료(binary data)이므로 최대우도 추정을 이용하여 심리물리학 모델에 적합하기 위해서는 이항분포를 이해할 필요가 있다. 그렇지만 이항분포 정리와 확률질량함수 개념에 익숙하지 않다면 곧바로 4. 3. 모델 적용으로 건너 뛰어 그림 15에 제시된 최대우도 추정을 이용한 적합 과정을 먼저 따라해 보고, 그런 다음 아래에 서술한 확률 이론을 천천히 살펴보는 것이 더 효율적일 것이다.

이항분포의 확률질량함수와 최대우도

간단한 예로, 동전을 12번 던졌을 때 9번이 앞면이 나왔을 경우를 생각해 보자. 만일, 앞/뒤가 나올 확률이 정확히 0.5였다면 9번이나 앞면이 나오는 경우는 상당히 이례적일 것이

다. 좀 더 현실적인 가정은 동전의 한 쪽 구석이 앞면이 나오도록 유리하게 닳았을 가능성을 의심해보는 것이다. 그렇다면, 단일 시행에서 앞면이 나올 확률이 얼마나 되어야 12번 던졌을 때 9번이나 앞면이 나올 가능성이 가장 클까? 다시 말해, 어떤 모집단이 앞면이 9번이 나올 가능성이 가장 내기 쉬울까? 앞면이 나올 가능성이 0.5 이상일 것이라고 짐작은 되겠지만 정확한 값을 알기 위해서는 계산이 필요한데 이는 이항분포 함수를 이용해서 해결할 수 있다. 앞면을 H 뒷면을 T라 했을 때 12번 던져 앞면이 9번인 예로 HTHTHHHHHTHH으로 표현할 수 있다. 여기서 매 시행에서 동전의 앞면이 나올 확률을 p라 했을 때 이에 대한 확률은 곱의 법칙에 의해 이 순서대로 앞면이 나올 확률을 구하는 식은 'p×(1-p)×p×(1-p)×p×p×p×p×p(1-p)×p×p'로 표현된다.

이를 정리하면 앞면이 나올 확률이 p인 동전을 n번 던졌을 때 실제로 앞면이 나온 시행수가 k인 경우 이 사건에 대한 확률질량함수(probability mass function)는 간단하게 식 (4)로 표현할 수 있고 순서와 상관없이 9번이 앞면이 나올 확률은 식 (4)에 이항계수를 곱한 식 (5)와 같다.

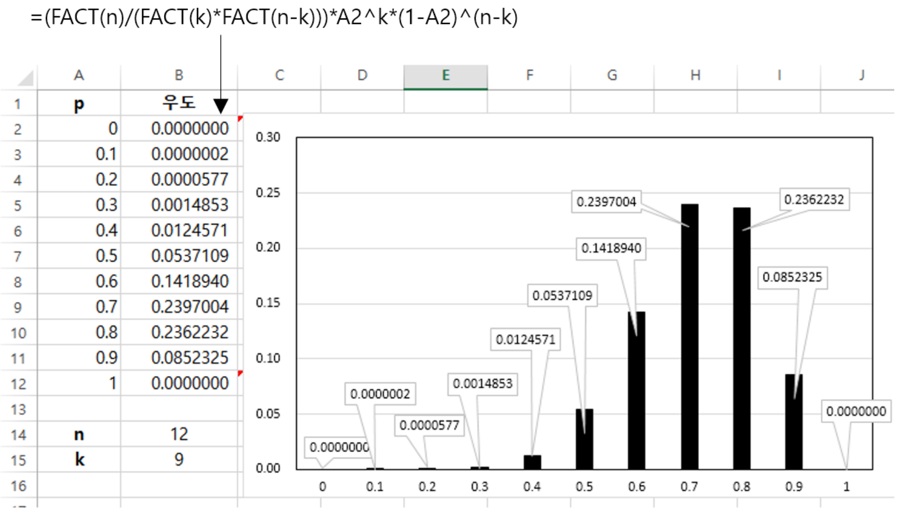
$$f(k|n,p) = p^k(1-p)^{n-k} \tag{4}$$

$$f(k|n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \tag{5}$$

그렇다면 역으로 12번 던져 9번이 앞면이 나오게 하는데 가장 높은 가능성을 가진 p 값은 얼마일까? 이것이 우도 함수(likelihood function)에 관한 문제이다. 여기서 우도 함수는 알려진 총 시행수 n과 앞면이 나온 횟수 k에서 p를 추정하므로 k가 p와 자리를 바꿔 $f(k \mid n, p)$가 아닌 $L(p \mid n, k)$로 주어진다. 비록, p는 확률은 아니지만 확률과 비슷하게 처리되므로 우도라 하고 0과 1의 사이에 놓인 값이다.

직관적인 이해를 위해, 식 (5)를 이용하여 p 값을 0부터 시작해서 0.1만큼 1까지 일일이 대입해서 그 결과를 엑셀을 이용하여 살펴보자. 결과는 그림 12 위에 나타나 있는데, 히스토그램을 살펴보면 대략 p 값이 0.7 근처일 때가 우도가 가장 높다는 것을 알 수 있다. 즉, 단일 시행으로 동전을 던졌을 때 앞면이 나올 확률이 0.7 근처인 경우에 12회 던져 9회가 앞면이 나올 가능성이 가장 높으며 이 지점이 최대우도이다. 여기서, 최대우도가 되는 지점을 알기 위해서 이항계수를 제거한 식 (4)를 적용해도 동일한 결과가 나오는데 그림 12 아래에서 확인할 수 있다. 즉, 이항계수를 넣거나 뺐을 때 우도의 값은 전체가 단조로운 변환(monotonic change)이 일어나므로 최대우도 지점을 찾는 데 어떤 식을 사용해도 문제가 되지 않는다. 따라서 최대우도를 계산할 때 계산의 편리성을 감안하여 흔히 식 (4)를 이용한다.

그러면 좀 더 정밀한 값을 계산하기 위해서 해 찾기 기능을 빌려보자. 그림 13은 이 과정을 예시한다. 그런데 이번에는 계산이 복잡해지기 때문에 p 값의 조건을 0과 1사이로 설정할 필요가 있다. p 값이 목표값에 비해서 너무 차이가 나면 해를 구하기가 어렵게 된다. 예를 들어, 0.1로 초기 값을 설정하면 해가 구해지지 않을 가능성이 있다. p의 초기 값은 0.5 이상이 나올 가능성이 크므로 이 정도로 설정하자. 해 찾기 수행 결과 p 값이 약 0.75로 구해졌다. 여기서 우도가 매우 작은데 최대우도 추정의 목적은 우도 자체의 값이 아니



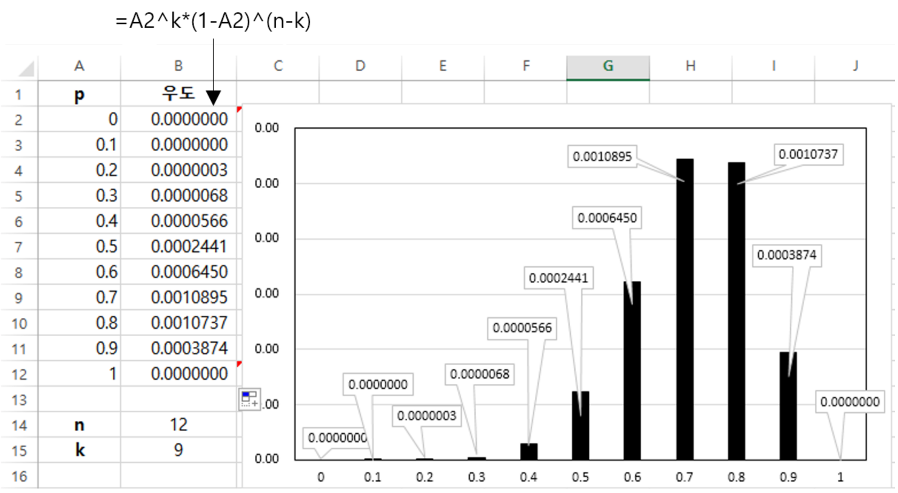


그림 12. (위) p 값을 임의로 넣었을 때 우도 값과 히스토그램. n과 k를 이름 정의한 후 각각 12와 9를 넣었고, 우도는 식 5를 따라 입력하였다. 엑셀 식에서 'FACT(n)'은 n!을 뜻한다. (아래) 우도를 식 4를 넣어 계산한 결과. 위와 비교할 때 절대 값만 다를 뿐 패턴은 비슷하다.

라 가장 큰 우도 지점 자체이므로 비록 값이 매우 작더라도 의미 있는 수치이다. 결론적으로 동전을 12번 던져 9번이 앞면이 나오게 할 가장 가능성이 높은 모집단의 최대우도 또는 이 최대우도의 파라미터는 0.75이다. 즉, 한 번 던졌을 때 앞이 나올 확률이 0.75인 동전이 12번 던졌을 때 9번 앞이 나올 사건을 낼 가능성이 가장 크다.

최대우도를 구하는 수학적인 방법을 생각해 보자. 이는 미분을 이용해 풀 수 있다. 동전을 모두 n번 던졌을 때 앞면이 나온 횟수 k에 대한 우도 함수 $L(p \mid n, k)$가 있고 이 함수가 미분 가능하다고 할 때 우도가 최대가 되는 지점은 p에 관해서 미분을 했을 때 또는 다른

그림 13. 해 찾기를 이용하여 동전을 12번 던졌을 때 앞이 9번 나온 사건에 대한 모집단의 최대우도 구하기. n, k, p를 각각 이름 정의하고, 편의상 이항계수를 제거한 식 (4)를 넣어 완성하였다. p와 최대우도의 자릿수를 셀 서식에서 소수점을 10개로 표시하도록 하였다. 해 찾기에서, p값을 0과 1사이로 입력하였다. GRG비선형으로 해법 선택하고 옵션에서 모든 해법의 제한 조건 정밀도를 소수점 10의 자리 이하로 낮게 하였다. 이 옵션은 계산이 얼마나 정밀하게 작동하도록 하는지를 정해준다. 값은 0과 1 사이에 조작되는데 0에 가까울수록 정밀한 해를 찾을 수 있지만 계산이 더 오래 걸린다.

표현으로 이 함수의 접선의 기울기가 0이 되는 지점이다. 이는 편미분의 기호를 사용하여 식 (6)으로 주어진다. 그런데 시행수가 많아지는 경우 값이 매우 작아져 계산이 어려워지는 불편이 따른다. 이를 피하기 위해 자연 로그를 취하는데 이를 로그 우도(log likelihood)라 하며 기호로 LN이다. 원자료에 자연로그를 취해도 전체가 일관되게 유지되는 단조 변환이 일어나므로 최대우도 추정 결과가 달라지지 않는다. 또한 자연로그를 취하게 되면 1보다 작은 값들은 음의 수로 바뀌고 곱셈은 덧셈으로 변환되는데 식 (4)를 우도함수로 상정하여 자연로그를 취하여 정리하면 식 (7)로 주어진다. 다시 식 (7)을 미분을 하면 식(8)로 정리된다. 결국, p = k/n이 되어 12번 가운데 9번이 앞면이 나온 사건을 낼 가능성이 가장 큰 우도는 9/12 = 0.75가 된다. 이는 그림 13에서 해 찾기 기능을 이용해서 찾은 값과 정확히 일치한다.

$$L(p|n,k) = \frac{\partial}{\partial p}[L(p|n,k)] = 0,\ \frac{\partial}{\partial p}[\ln L(p|n,k)] = 0 \quad (6)$$

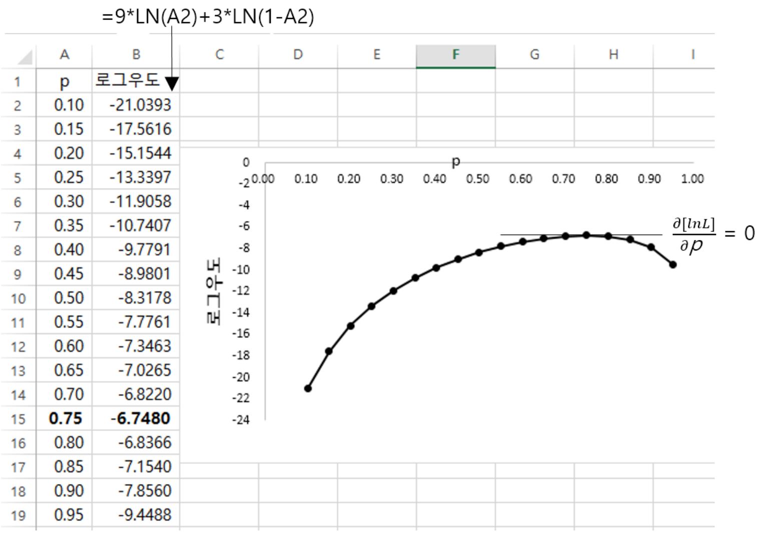
$$\frac{\partial}{\partial p}[k \cdot \ln(p) + (n-k) \cdot \ln(1-p)] = 0 \quad (7)$$

$$\frac{k}{p} - \frac{n-k}{1-p} = 0,\ p = \frac{k}{n} \quad (8)$$

그림 14는 지금까지 설명한 내용을 직관적으로 이해하기 쉽도록 엑셀 그래프를 이용하여 표현한 것으로 동전을 12회 던졌을 때 9회가 앞면이 나온 사건에 대한 우도 함수 분포에 로그를 취한 로그우도 분포를 보여주고 있다. 이 그림에서도 p 값이 대략 0.75일 때 우도가 가장 높은 값임을 알 수 있다. 로그를 취하기 전과 후의 값들은 수식을 바꾸어서 살펴보기를 권한다.

결합우도

앞에서 동전 던지기 사건에서 최대우도 추정의 예를 살펴보았다. 그런데 여기서 다루고 있는 주의에 따른 시간 순서 판단 과제에서 얻은 자료는 13개 수준에서 독립적인 지점이 있으며 이에 따라 13개의 우도 함수가 존재한다. 그러므로 이 모든 우도들을 아우르는 우도 개념이 있는데 이를 결합 우도(joint likelihood)라 한다.



	A	B
1	p	로그우도
2	0.10	-21.0393
3	0.15	-17.5616
4	0.20	-15.1544
5	0.25	-13.3397
6	0.30	-11.9058
7	0.35	-10.7407
8	0.40	-9.7791
9	0.45	-8.9801
10	0.50	-8.3178
11	0.55	-7.7761
12	0.60	-7.3463
13	0.65	-7.0265
14	0.70	-6.8220
15	**0.75**	**-6.7480**
16	0.80	-6.8366
17	0.85	-7.1540
18	0.90	-7.8560
19	0.95	-9.4488

그림 14. 동전을 12회 던졌을 때 9회가 앞면이 나온 사건 확률에 대한 우도함수. (식) 4를 적용하여 p 값을 0.1부터 0.95까지 0.5씩 증가 시켜 표현하였다. 로그우도가 최대인 지점이 p 값이 대략 0.75이며, 이 때 접선의 기울기가 0이 된다.

좀 더 자세히 설명하면, 동전의 예에서 12번 던져 9번이 앞면이 나왔다면 이에 대한 확률은 9/12로 0.75이다. 그렇지만, 시간 순서 판단 과제에서 13개의 실험 조건이 있었고 오른쪽이 먼저라는 반응 기준으로 볼 때, 각 조건을 100번의 시행으로 가정하여 비율을 확률로 바꾸면 0.1, 0.19, 0.21, 0.24, 0.4, 0.5, 0.64, 0.84, 0.92, 0.98, 0.97, 0.99, 0.99 이다. 이 13개 자료 각각의 최대우도를 고려하되 이번에는 이들 모두가 동시에 고려된 우도를 찾되 값이 최대한 높아야 한다. 결합 우도는 다시 확률의 곱의 법칙에 의해서 모든 우도들을 곱한 값으로 식 (9)로 주어진다. 여기서 π는 누적곱의 기호이며 나머지 식은 식 (4)와 같다는 것을 알 수 있다.

앞에서 설명하였듯이, 시행수와 조건들이 많아지면서 곱의 계산에서 어려움이 발생한다. 즉, 각 수준에서 시행수를 고려한 우도를 구하고 또 모든 수준의 우도들을 누적곱을 하는 과정에서 값의 크기가 급격하게 작아지는데 컴퓨터로 계산하기 어려운 수준에 이른다. 따라서 계산의 편리를 위해 각 조건에서 최대우도를 구하여 자연로그로 먼저 변환한다. 이때 각 조건들의 값을 모두 곱해야 하는 누적

$$f(p|n,k)=\prod_{i=1}^{n}p^{k}(1-p)^{n-k} \tag{9}$$

$$\ln(f(p|n,k))=\sum_{i=1}^{n}\ln(p^{k}(1-p)^{n-k}) \tag{10}$$

$$\ln(f(\alpha,\beta|n,k))=\sum_{i=1}^{n}\ln[(\frac{1}{1+\exp^{-\beta(\chi_i-\alpha)}})^{k}(1-\frac{1}{1+\exp^{-\beta(\chi_i-\alpha)}})^{n-k}] \tag{11}$$

곱에서 모두 더해야 하는 누적합으로 바뀌어 식 (10)으로 주어진다.

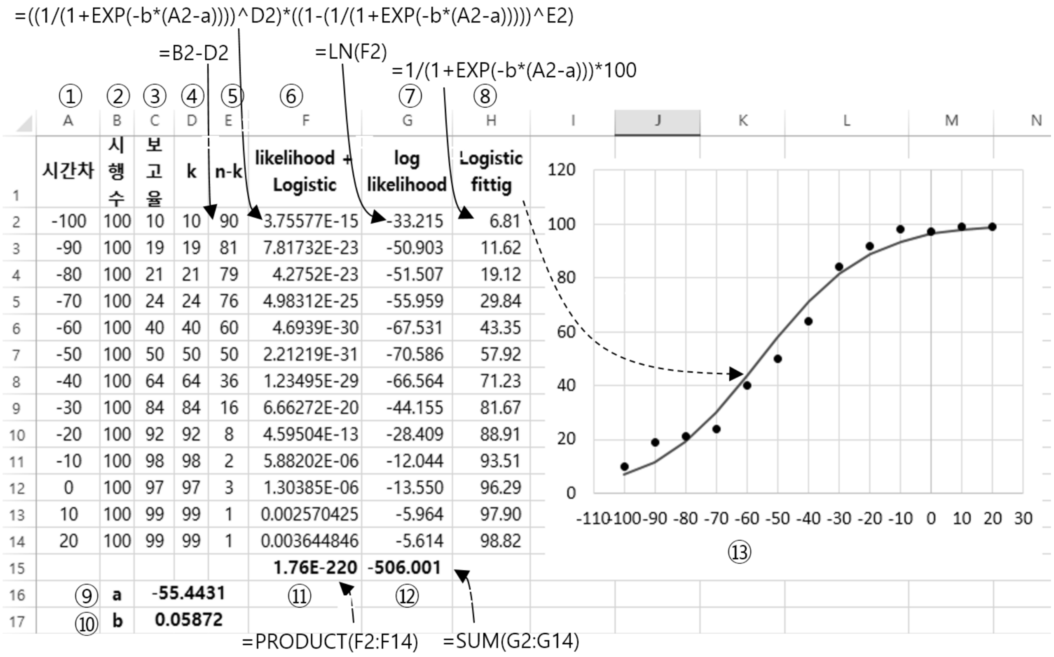
또 한 가지 중요한 절차는 각 수준의 자료지점이 상정된 로지스틱 함수를 따르므로 이를 반영해야 한다는 것이다. 이 과정은 앞에서 최소제곱 추정으로 자료 적합했던 것과 비슷하다. 이를 위해서는 단순히 식 (10)에서 p를 로지스틱 함수와 치환하는 것으로 해결되며 알고 있는 총 시행수 n과 오른쪽이 먼저라는 반응 수 k에서 파라미터 α와 β를 구하는 것이므로 이는 식 (11)로 주어진다.

주목할 점으로 이번에는 구해야 할 파라미터가 p 하나에서 α와 β 두 개로 늘어난다. 이는 앞에서 미분을 통해서 단순히 계산되는 상황과 다르다. 이런 경우 자료 형태가 고도로 비선형적으로 되어 수학적으로 풀기가 매우 까다로운 것으로 알려져 있다. 이 문제를 컴퓨터 프로그램으로 해결하기 위해 다양한 알고리듬들이 고안되었는데 모두 어림법(heuristic)으로 이뤄진다 즉, 임의의 초기 값을 넣고 그 결과 값이 얼마나 원하는 조건에 근접하는지를 비교해가면서 원하는 조건에 가장 근접한 해를 구하는 것이다. 엑셀의 해 찾기는 고안된 여러 방법 가운데 하나의 알고리듬을 사용하는 것이다.

모델 적용

지금까지 설명한 과정을 엑셀을 통해서 실제로 적합을 단계별로 수행해보자. 그림 15는 엑셀에서 오른쪽 주의 조건의 자료에 대해서 최대우도를 보이는 파라미터 α와 β를 구하고 이를 로지스틱 함수에 넣어 원자료를 적합한 결과를 보여준다. 이를 위해서는 먼저, 원자료와 우도 계산을 위한 기본 자료를 열1-5번처럼 입력한다. 다음으로, 9번과 10번에 나와 있는 것처럼 파라미터 α와 β를 이름을 정의한다. 최대우도는 열6의 13개의 값을 모두 곱한 11번의 값이지만, 계산이 복잡해지므로 자연로그(LN)를 취하여 모두 합한 12번을 해 찾기에서 목표 설정 값으로 설정하며 대상 옵션에는 최댓값에 표시를 한다. 파라미터 α와 β에 해당하는 셀들을 변수 셀 변경란에 입력하여 구한다. 이때도 역시 α는 y축의 50에 해당하는 x 축의 값이므로 대략 -50을 넣고, β 값을 조절해 가면서 12번 셀의 값이 최대가 되는 지점을 골라 적당히 입력한다.

그런데 여기서 주의해야 할 점으로 로그우도 값이 엑셀이 계산할 수 있는 범위가 넘으면 '#NUM!' 에러가 뜰 수 있다. 예를 들어, α 값에 －50을 넣고 β 값을 1을 넣으면 이런 오류가 뜰 수 있다. 이를 피하는 방법은 이 오류가 나오지 않을 때까지 수를 줄여가는 것이다. 0.5를 넣으면 누적곱이 비로소 오류에서 벗어나는 것을 알 수 있다. 만일 0을 넣으면 계산은 가능하지만 그래프가 편평해지는 것을 알 수 있다. 따라서 β 값은 0보다 크고 0.5 사이에 있을 수 있다. 이렇게 적합 그래프가 원자료에 최대한 비슷한 지점까지 두 파라미터의 값을 찾은 다음 마지막으로 해 찾기 기능을 사용하여 최적 값을 찾는다. 최대우도 추정을 이용한 적합은 최소제곱 추정을 이용한 적합에 비해서 숫자들이 매우 작은 값이기 때문에 좀 더 정밀하게 해의 범위를 정해줄 필요가 있다. 제약조건 설정에 대략 a<=0, a>=-100, b<=0.5, b>=0으로 해주고 옵션에서 모든 해법의 제한 조건 정밀도를 소수점 10의 자리 이하로 낮게 설정해준다. 적합을 한 결과 파라미터 α는 -55.44, 파라미터 β는 0.0587로 계산되었다. 한편, 왼쪽 주의 조건의 자료에 대해서도 똑같은 과정을 거치면 파라



시간차	시행수	보고율	k	n-k	likelihood + Logistic	log likelihood	Logistic fittig
-100	100	10	10	90	3.75577E-15	-33.215	6.81
-90	100	19	19	81	7.81732E-23	-50.903	11.62
-80	100	21	21	79	4.2752E-23	-51.507	19.12
-70	100	24	24	76	4.98312E-25	-55.959	29.84
-60	100	40	40	60	4.6939E-30	-67.531	43.35
-50	100	50	50	50	2.21219E-31	-70.586	57.92
-40	100	64	64	36	1.23495E-29	-66.564	71.23
-30	100	84	84	16	6.66272E-20	-44.155	81.67
-20	100	92	92	8	4.59504E-13	-28.409	88.91
-10	100	98	98	2	5.88202E-06	-12.044	93.51
0	100	97	97	3	1.30385E-06	-13.550	96.29
10	100	99	99	1	0.002570425	-5.964	97.90
20	100	99	99	1	0.003644846	-5.614	98.82
					1.76E-220	**-506.001**	
	a	**-55.4431**					
	b	**0.05872**					

그림 15. 오른쪽 주의 조건에 대한 적합 과정. (1) 종속 변인인 시간차이다. (2) 총 시행수로 100회를 상정하였다. (3) 오른쪽이 먼저라는 보고 비율이다. (4) 보고 횟수이며 이는 총 시행수가 100회이므로 보고율과 같다. (5) 총 시행수에서 오른쪽 보고 시행수를 뺀 값이다. (6)식 (11)에서 자연로그를 취하기 전의 값으로 p 대신에 로지스틱 함수를 넣은 값이다. (7)(6)의 값을 로그 우도로 변환한 값이다(LN은 자연로그로 변환하는 엑셀 함수이다) (8) 해 찾기로 찾아낸 파라미터를 적용하여 로지스틱 함수를 적용한 값이다. (9) 파라미터 α이다. (10) 파라미터 β 값이다. (11) 우도 값들의 누적곱으로 엑셀 함수는 PRODUCT이다. (12) 로그우도 값들의 누적합 (13) 12번의 로그우도의 합이 최댓값을 되도록 하는 파라미터 α와 β 값을 구하여 원자료에 적합한 결과 그래프로 8번열에 있는 값들이다. 원자료와 적합된 자료를 모두 그래프로 그린 후 적합된 자료에 해당하는 그래프에서 표식은 지우고 곡선을 남겨서 완성되었다.

미터 α는 43.72, 파라미터 β는 0.04로 계산된다. 이들 값은 앞서 최소제곱법으로 추정한 결과와 매우 근소한 차이가 남을 알 수 있다.

정리

조건간 비교 검증

본 논문에서 다루고 있는 시간순서 판단 과제는 주의 위치에 따라 시간 순서 지각이 다른지를 살펴보고자 하였다. 즉, 영가설은 '시간 순서 지각은 주의 위치와 무관하다'는 것이다. 따라서 이 영가설을 검증하기 위해서는 오른쪽 주의 조건과 왼쪽 주의 조건에서 시간 판단율이 통계적으로 다른지를 살펴보아야 한다. 앞에서 원자료를 로지스틱 함수에 적용하였고 최소제곱 추정과 최대우도 추정을 이용하여 적합을 하였다. 이에 따라, 분포의 특징을 말해주는 파라미터를 구할 수 있었다. 여기서 연구자들의 주요 관심은 불빛이 왼쪽과 오른쪽에서 보였는지 정확히 구분이 되지 않

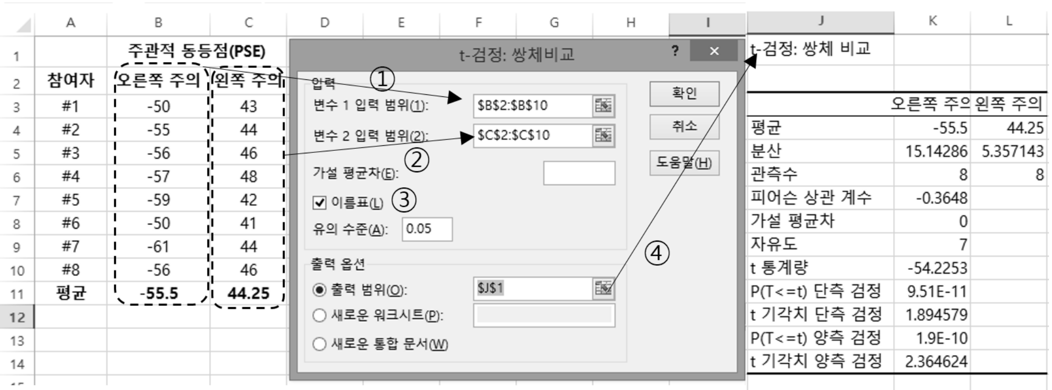


그림 16. 임의로 넣은 두 주의 조건 자료에 대한 *t*-test 결과. 입력 범위 1에는 오른쪽 주의 조건을 범위 2에는 왼쪽 주의 조건을 넣어 준다. 이름표는 각 조건의 제목을 뜻한다. 출력범위는 결과가 나타날 위치를 지정하는 것으로 여기서는 J1을 선정하였다.

는 또는 동등하게 보이는 주관적 동등점(PSE)이었다. 반응이 둘 중 하나를 고르는 이진수 자료이고 통계적 검증이 확률적으로 좀 더 타당하도록 최대우도 추정을 이용한 주관적 동등점 계산이 권해진다. 이를 위해 이 실험에 참여한 개인들의 원자료를 분석하여 주관적 동등점을 구한다. 이 실험은 참여자내 반복 실험이었으므로 두 조건에 대해서 짝진 *t*-test 또는 반복 *F*-test를 적용하여 통계적 검증을 하는 것으로 마무리 한다.

Stelmach와 Herdman의 연구 결과를 통계적으로 검증하는 방법을 시연하고자 이 논문에서 제시된 집단간 평균을 참고하여 오른쪽 주의 조건과 왼쪽 주의 조건의 값들을 임의로 만들어 보았다. 이 자료에서 8명이 실험에 참여한 것으로 가정되었다. 그림 16에는 이 자료와 이 자료에 대해서 *t*-test를 실시한 결과를 보여준다. 결과 분석은 엑셀의 ‘데이터’ 탭 아래에 놓인 ‘분석 도구’에서 ‘t-검정: 쌍체비교’를 이용하였다. 분석 도구는 해 찾기와 마찬가지로 엑셀의 추가 기능이므로 엑셀의 옵션 메뉴에서 따로 추가하여야 이용할 수 있다. 통계 분석 결과 값이 오른쪽에 나타나 있고, $t(7) = -54.2$, $p < .001$로 정리된다. 따라서 주의에 의해 시간 순서 판단에 차이가 없다는 영가설은 기각된다.

때때로 어떤 참여자들의 원자료는 적합을 하지 못할 정도로 예외적인 경우가 있다. 보통 강제선택 과제에서 참여자가 어느 한 쪽 반응만 일방적으로 하여 주관적 동등점 또는 기울기가 다른 참여자들에 비해서 크게 다르기 때문이다. 정말로 자극이 구분되지 않아서일 수도 있고 실험에 나태하게 참여해서일 수도 있다. 이런 경우 자료 적합이 되지 않는 이유를 분포적인 설명을 하고 해당 참여자의 자료를 제외해야 한다.

연구에 따라 주관적 동등점이 아닌 분포의 기울기 분석이 관심인 경우도 있을 것이다. 예를 들어, 젊은 성인에 비해 노인은 기울기가 더 완만하여 시간차가 훨씬 큰 지점부터 착시를 경험할 수지도 모른다는 가설을 검증하고 싶다면 젊은 성인과 노인의 반응 분포에서 기울기를 비교해야 할 것이다.

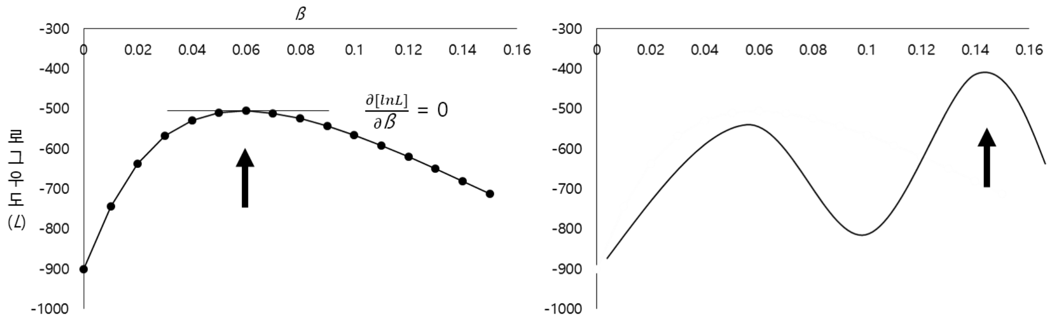


그림 17. (왼쪽) 파라미터 α를 -55.44로 고정한 채 β 값을 0.01씩 바꾸면서 로그우도의 누적합을 구한 모습으로 전형적인 볼록 함수를 보인다. 화살표는 누적합이 최대가 되는 지점이다. 이 지점은 미분했을 때 접선의 기울기가 0이 되는 지점이다. (오른쪽) 누적합의 분포가 봉우리를 두 개 이상 갖는 가상적인 비볼록 함수의 예. 이 경우 접선의 기울기가 0이 되는 지점이 3 개가 존재한다. 초기 값에 따라 국지 해에 탐색이 멈추어 화살표가 가리키는 전역 해를 찾지 못할 수 있다.

최적화 문제

어떤 문제의 해를 구하고자 할 때 이 문제를 구성하는 파라미터가 두 개 이상인 경우 이론적으로 두 파라미터의 값을 조절해가면서 원하는 해에 최대한 가까이 다가가는 전략을 취하는데, 이를 최적화 문제(optimization problem)라 한다. 자료 적합 과정에서 흔히 최적화 문제가 발생하는데, 앞서 최소제곱 추정과 최대우도 추정을 이용한 자료 적합 과정도 이에 해당한다.

최대우도 추정을 이용하여 해를 구한 앞의 예를 들어 설명해보자. 이 예에서 최적화 문제 상황을 다시 기술하면, 로그우도를 모두 더한 값이 최대가 되도록 하는 파라미터 α와 β를 구하는 것이다. 이 문제 상황을 그림을 통해 이해해 보자. 이 예에서, 파라미터 α를 -55.44로 고정한 채 β 값을 조금씩 바꿀 때 로그우도의 누적합의 변화를 살펴보자. 그림 17의 왼쪽은 α를 -55.44로 고정한 채 β 값을 0.01씩 바꿀 때 얻어지는 로그우도의 누적합을 예시한다. 여기서 볼 수 있듯이 최댓값은 β 값이 0.06 근처에 있다. 최적화 문제를 해결하는 과정은 두 파라미터의 초기 값을 중심으로 체계적으로 파라미터의 값을 바꾸면서 로그우도의 누적합을 살펴보는 과정을 반복하게 된다. 만일 증가하는 방향이 이전 시행에 비해 바뀌게 된다면 여기서 탐색을 멈출 것이다.

그런데, 만일 얻어진 자료가 그림 오른쪽처럼 봉우리가 여러 개인 비볼록함수를 보인다고 가정해보자. 이 공간에서는 값의 방향이 갑자기 바뀌는 지점을 해라고 가정하는 알고리듬 입장에서 가상의 해가 동시에 두 개 이상이 될 수 있고, 모든 해 공간에서 값이 최대인 전역 최댓값(global maximum)과 그 다음으로 최댓값을 가지는 국지 최댓값(local maximum)으로 구분할 수 있다. 미분 개념을 빌어 설명하면 볼록함수의 경우 봉우리가 하나뿐이므로 접선의 기울기가 0이 되는 지점도 하나가 되어 전역 최댓값과 국지 최댓값이 일치하여 혼동이 없다. 그렇지만 그림 오른쪽처럼 전역 최댓값과 국지 최댓값이 같지 않은 경우 즉, 접선의 기울기가 0이 되는 지점이

여러 개인 경우 국지 최댓값에서 해 찾기 탐색을 멈출 수도 있다. 따라서 이런 경우 초기 값이 매우 중요한 역할을 한다. 그림 17의 오른쪽 그림처럼 우도 함수가 형성되고 만일 초기 값이 0.02로 한다면 탐색은 0.06 근처에서 멈출 것이다. 반면, 운이 좋아서 0.12로 정했다면 전역 최댓값을 만나게 될 것이다. 현재까지 최적화 문제를 완벽하게 해결할 수 있는 방법은 없는 것으로 알려져 있다(Myung, 2003). 이 문제를 푸는 일반적인 방법은 일종의 어림법으로, 간단히 할 수 있는 방법으로 임의로 초기 값을 넣어 가면서 구하고자 하는 값이 얼마까지 커지는지를 대략적으로 살펴보면서 초기 값을 설정하는 것이다. 다른 방법은 엑셀 해 찾기에서 GRG 비선형 옵션에서 ‘Multistart’ 단추를 활성화하는 일이다. 이는 초기 값을 여러 개를 상정하여 탐색하는 것으로 시간이 더 소요되는 단점이 있다.

다양한 심리측정 함수

본 글에서 예로 들고 있는 자료처럼 누적분포 자료를 적합할 때 흔히 사용되는 측정 함수는 Logistic 이외에 Weibull, Gaussian, Gumbel, Power 등 다양하며 표 2에 엑셀 수식과 함께 제시되어 있다. 이 표에 제시된 함수의 분포는 X 축의 값이 증가함에 따라 Y 축의 값도 증가한다. 하지만 심리측정 함수는 이런 형태만 있는 것은 아니며 원자료의 분포에 따라 매우 다양하다. 이에 따라 적합 함수 역시 복잡해지거나 파라미터들의 범위가 달라지기도 한다. 더욱이 분포의 봉우리가 두 개 이상인 경우도 많다(예: Brown, 2006). 그렇지만 분포

표 2. 자주 사용되는 심리측정 함수. 이 함수들은 누적 분포에 대한 것이며 그렇지 않은 경우 파라미터의 범위가 달라질 수 있다.

함수	수식*	엑셀 수식*
Weibull	$1-\exp[-(\frac{x}{\alpha})^{\beta}]$ 0≤χ<+∝, 0≤α<+∝, 0≤β<+∝	1-EXP(-((χ/α)^β))
Cumulative Normal (Gaussian, CDF)	$\frac{1}{2}\left[1+erf\left(\frac{x-\alpha}{\beta\sqrt{2}}\right)\right]$ -∝<χ<+∝, -∝<α<+∝, 0≤β<+∝	0.5*(1+ERF((χ-α)/(β*SQRT(2)))
Gumbel (Log-Weibull)	$1-\exp[-\exp(\frac{(x-\alpha)}{\beta})]$ -∝<χ<+∝, -∝<α<+∝, 0≤β<+∝	1-EXP(-EXP((χ-α)/β))
Power	αx^{β} -∝<χ<+∝, -∝<α<+∝, 0≤β<+∝	a*χ^β

* 이 수식은 확률을 기준으로 삼아 최댓값이 1이다. 따라서 원자료가 비율인 경우 이 수식 전체에 100을 곱해야 한다. 각 함수들은 다양한 변형들이 있을 수 있다. Cumulative normal 함수에서 *erf*는 오차함수(error function)로 불리며, s 형태의 확률누적 분포를 이룬다: $erf(x)=\frac{2}{\sqrt{\pi}}\int_0^x \exp^{-t^2}dt$

가 다르다고 해서 적합 원리도 다른 것은 아니다. 대부분의 자료 적합은 본 글에서 소개한 원리를 따른다. 약간의 수고를 더하면 엑셀을 통해서 충분히 적합이 가능하므로 끈기있게 마주하기를 제안 한다.

맺음말

이상으로 자료 적합하는 과정을 엑셀의 해 찾기 기능을 이용하여 시범하였다. 최소제곱 추정법은 직관적으로 이해하기 쉽다는 장점이 있지만 원자료가 등간척도 또는 비율척도인 연속 값을 가질 때에만 의미가 있다. 또한, 각 변수의 수준에서 확률을 고려하지 않기 때문에 모든 수준에서 수행을 동등한 것으로 보는 특징이 있다(Cunningham & Wallraven, 2012, p. 353). 그렇지만, 심리학 실험실에서 주로 얻는 자료는 각 수준에서 얻어지는 수행이 확률적으로 같지 않는 경우가 많다. 그림 15에서 로그우도 값의 절댓값을 살펴보면 반응률이 50% 지점에서 최고를 보이고 0%와 100% 방향으로 갈수록 작아지는 것을 알 수 있다. 이는 최대우도 적합에서 이 값들의 합이 가장 크게 하는 파라미터를 구하므로 50% 지점의 반응률이 양 극단 수준의 반응률보다 더 중요하게 처리된다는 뜻이다. 이에 반해, 최소제곱 적합에서는 수행률에 관계없이 원자료를 일률적으로 처리한다. 따라서 최대우도 추정에서는 사건이 일어나는 확률론이 반영되기 때문에 추론적 통계 검증을 위해서는 최소제곱 추정에 비해 통계적으로 더 타당한 방법일 것이다.

참고문헌

Afifi, A., May, S., & Clark, V. A. (2012). *Practical multivariate analysis*. Taylor & Francis Group: Florida.

Brown, A. M. (2001). A step-by-step guide to non-linear regression analysis of experimental data using a Microsoft Excel spreadsheet. *Computer Methods in Biomedicine*, *65*, 191-200.

Brown, A. M. (2006). A non-linear regression analysis program for describing electrophysiological data with multiple functions using Microsoft Excel. *Computer Methods and Programs in Biomedicine*, *82*, 51-57.

Cunningham, D. E., & Wallraven, C. (2012). Experimental Design: From user studies to Psychophysics. Taylor & Francis Group: Florida.

Dodson, C. S., Prinzmetal, W., & Shimamura, A. P. (1998). Using Excel to estimate parameters from observed data: An example from source memory data. *Behavior Research Methods, Instruments, & Computers*, *30*(3), 517-526.

Dunn, J. C. (2010). How to fit models of recognition memory data using maximum likelihood. *International Journal of Psychological Research*, *3*(1), 140-149.

John, E. G. (1998). Simplified curve fitting using spreadsheet add-ins. *International Journal of Engineering Education*, *14*, 375-380.

Kemmer, G., & Keller, S. (2010). Nonlinear least-squares data fitting in Excel spreadsheets. *Nature Protocol*, *5*, 267-281.

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A practical introduction*. Academic

Press: London, UK.

Lambert, R. J. W., Mytilinaios, I., Maitland, L., & Brown, A. M. (2012). Monte Carlo simulation of parameter confidence intervals for non-linear regression analysis of biological data using Microsoft Excel. *Computer Methods and Programs in Biomedicine*, *107*, 155-163.

Lasdon, L. S., Waren, A. D., Jain, A., & Ratner, M. (1978). Design and testing of a generalized reduced gradient code for nonlinear programming. *ACM Transactions on Mathematical Software*, *4*(1), 34-50.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90-100.

Stelmach, L. B., & Herdman, C. M. (1991). Directed attention and perception of temporal order. *Journal of Experimental Psychology: Human Perception and Performance, 17*(2), 539-550.

Wichmann, F., & Hill, J. N. (2001). The psychometric function: Ⅰ. Fitting, sampling, and goodness fit. *Perception & Psychophysics*, *63*(8), 1293-1313.

1차원고접수 : 2015. 05. 12.
수정원고접수 : 2015. 07. 03.
최종게재결정 : 2015. 08. 20.

한국심리학회지 : 일반
Korean Journal of Psychology: General
2015, Vol. 34, No. 3, 741-767

An introduction to non-linear data fitting using the Microsoft Excel

Songjoo Oh

Department of Psychology, Seoul National University

Psychologists working on psychometric data often struggle with the process of data fitting which requires advanced knowledge about programming and mathematics. Although some commercial softwares reduce investigators effort to perform the process, still most of them are far from easy tools that a beginner tries them without hesitation. In contrast, the Microsoft Excel provides intuitive ways to fit data and calculate the parameters. In this paper, the processes of data fittings are demonstrated step by step using the Excel solver. Specifically, in fitting non-linear data, the least squares estimation and maximum likelihood estimation are introduced and the processes are compared to understand the difference. Finally, it was discussed how to statistically test the parameters between groups that were obtained from the data fitting.

Key words : data fitting, least squares estimation, maximum likelihood estimation, Psychometric functions, Excel solver