기획논문

프로빗과 순차적 프로빗 분석에 대한 이해와 적용

주 미 영

사회과학에서 사용되는 다양한 종류의 계량적 분석모델 중 가장 흔히 사용되는 회귀분석은 종속변수가 질적 변수일 경우 독립변수들의 효과의 크기를 심각하게 잘못 평가할 수 있다. 회귀분석은 기본 가정을 위반할 경우 다른 적절한 기법으로 대체되어야 한다. 이 글은 종속변수가 두 개 이상의 가치를 지니는 범주적인 이산 변수일 경우 사용하기 적합한 모델 중의 하나인 프로빗과 순차적 프로빗을 소개하고 있다. 프로빗은 종속변수가 양분변수(0, 1)일 경우 그리고 순차적 프로빗은 다분적(0, 1, 2, 3, …)인 서열변수일 경우 사용할 수 있다. 이 글에서는 두 모델의 소개와 응용을 위해서 프로빗을 사용하는 선택방정식과 순차적 방정식을 사용하는 실제 방정식으로 구성된 2단계 방정식 모델이 제시되고 있다.

1. 머리말

우리는 사회현상을 설명하기 위해서 어떤 특정한 상황을 설명해 주는 일련의 독립변수와 다양한 상황을 위해 예측되어져야 할 종속변수 사이의 인과적 관계를 모색하여 이를 일반화하려고 한다. 결국, 이를 위해서는 자료수집을 통한 통계학적인 추론이 사용된다. 사회과학 분야에서 다양한 종류의 계량적 분석 모델들이 사용되어 왔으나, 그 중에서도 회귀분석(*regression*)이나 교차표분석(*crosstabulation*) 등의 사용이 가장 눈에 띄게 증가해 왔다. 사회과학에 포함되더라도 연구분야와 연구 목적의 차이에 따라 주로 사용되는 방법에는 큰 차이가 나타난다.

그 중 많이 사용되는 회귀분석의 경우 연구 목적에 실제로 제대로 적용되었는가에 대해서는 많은 의문점이 제기될 수 있다. 회귀분석이

주미영은
미국 미시간주립대학교에서 정치학 박사학위를 받고 현재 아태평화재단 선임연구위원으로 재직중이다.

myju@chollian.net

흔히 사용되는 표준적인 통계학적 도구로서의 역할을 해 온 것을 시인할 수 있지만, 동시에 너무 지나치게 많이 남용되어 온 것도 인정해야 한다. 그 이유는 이 모델이 다원적 분석이 가능하고, 비교적 이해하기가 쉽기 때문이다. 그럼에도 불구하고, 회귀분석을 사용할 경우 필요한 여러 가정 중 어떤 하나라도 만족되지 않는다면, 아주 심각하고 불합리한 결과가 발생된다는 것을 감안하여야 한다.

이 글에서는 많은 문제점 중에서도 특히 회귀분석의 사용을 위한 여러 전제 조건들의 하나인, 즉 종속변수가 연속적 변수이어야 한다는 조건이[1] 맞지 않는 경우에 사용될 수 있는 방법 중의 하나를 소개하려고 한다. 종속변수가 질적 변수인 경우에 선형회귀분석(*linear regression analysis*)은 독립변수들의 효과의 크기를 심각하게 잘못 평가할 수 있다. 회귀분석은 모든 연구의 만병통치약이 아니라 단지 어떤 제약된 가정하에서만 타당한 것일 뿐이다. 기본 가정을 위반할 때에는 종속 변수의 성질에 따라 회기분석이 아닌 다른 적절한 기법으로 대체되어야 한다. 실질적이고 이론적인 관심에 부적당한 통계방법의 선택은 분명히 가치 없는 연구를 낳을 수 있기 때문이다.

사회과학의 많은 분야, 즉 정치학, 경제학, 행정학 등의 주요 관심은 인간의 선택행위를 이해하는 것이다. 모델이나 가설은 결정과정의 본질에 따라 형성되며, 관찰된 행위로 평가된다. 우리는 사회과학분야에서 행태학적인 연구를 위해 사용할 자료를 수집할 경우, 모든 자료들이 정확하게 양적(*quantitative*)이고 연속적인(*continuous*) 자료만이 수집된다고 볼 수 없다. 특히, 종속변수로서의 변수의 범위가 경우에 따라서는 제한되어 있을 수도 있고, 때로는 변수가 질적인(*qualitative*) 성격을 띠는 경우도 상당히 많다. 종속변수가 두 개 이상의 가치를 지니는 범주적(*categorical*) 변수일 경우가 바로 프로빗 사용에 적합한 종속 변수라 할 수 있다. 예를 들면, 직업의 종류, 선거 행태, 물건 구매, 대학의 선택, 정당의 선택 등으로 결국 이런 종속변수는 각 개인, 대행기관, 또는 단체의 행동이나 선호도에 대한 관찰을 의미한다. 정치학에서는 주로 선거참여의 유무와 선호정당 및 선

호하는 후보, 경제학에서는 상품의 구입 혹은 매매시의 선택과 계약 체결, 그리고 사회학에서는 지원대학의 선정, 가족계획 등과 같이 실제로 많은 관찰이 질적이며 비연속적인 경우로 나타난다.

이 글에서는 범주형 변수로서의 이산적 종속변수를 위한 모델로, 종속변수가 양분적(*dichotomous*)인 경우[2]에는 프로빗(*probit*)을, 그리고 종속변수가 세 개 이상의 서열변수일 경우에는 순차적 프로빗(*ordered probit*)을 사용하는 2단계 방정식(*Two-equation*) 모델을 사용해 보려고 한다. 이와 함께 소개되어질 모델들을 이용한 사례들과 더불어 나온 결과들을 어떻게 해석하는가에 관하여 설명하려고 한다.

2. 프로빗 모델을 위한 자료의 성격과 적합한 통계적 응용

계량적 연구방법을 사용하는 연구에서는 종속변수의 측정수준에 따라 통계기법의 선택이 정해진다. 일반적으로 종속변수가 등간변수(*interval variable*)일 경우에는 회귀분석이 적합하고, 서열 혹은 명목 변수의 경우는 프로빗 혹은 로짓(*logit*) 모델이나 판별분석(*Discriminant analysis*)을 사용하는 것이 바람직하다고 할 수 있다. 하지만 판별분석의 경우 사용될 독립변수는 반드시 등간이나 비율변수를 사용해야 하는 제약이 있다.

프로빗 모델은 원래 종속변수가 이원적 값을 갖는 경우를 위해 개발되었으나, 그후 다원적 값을 갖는 경우(Aitchison & Silvey, 1957; Mckellvey & Zavoina, 1969; Cnudd, 1971)까지 확장되었다. 이와 함께 이를 응용한 다양한 방법이 등장하고 있지만, 아직은 다른 방법에 비해 그다지 많이 사용되고 있는 편은 아니다. 하지만 특히 경제학 분야의 학자들이 대표적으로 프로빗 모델에 대해 상세히 설명하고 사용의 예를 제시하여 이 모델의 확장에 상당히 기여하고 있다. 비선형 모델로서 프로빗과 로짓은 정치학자들에게도 양분적 종속변수를 갖는 모델을 추정할 경우 OLS가 가지고 있는 효율성과 규격(*specification*) 문제를 극복하기 위해 선호되고 있다.

선형회귀분석과 같은 방법은 연구자로 하여금 본래 불연속적인 개념을 가진 변수를 연속적으로 정의를 내리거나 또 그와 반대가 되는 상황을 갖게 하곤 했다. 따라서 측정의 오류나 편견이 발생하고 심각한 양의 정보를 잃게 된다. 이런 문제들을 극복하기 위해서, 종속변수가 '연속 변수'가 아니거나 또는 서열적인 '이산변수'인 경우에는 주로 프로빗이나 로짓 등의 방법들이 사용되는 것이 바람직하다. 물론 이에 앞서 연구의 목적이 무엇인가도 중요하고, 연구에 쓰이는 독립변수들의 성질, 즉 독립변수가 연속변수인가 이산변수인가도 살펴보아야 한다.

종속변수가 이산적이며 범주형일 경우에 흔히 사용되는 비선형(*nonlinear*) 모델의 대표적인 것들이 바로 프로빗과 로짓이다.[3] 이 둘은 특히 설명되어질 변수가 양분 변수(0, 1)일 경우에 주로 사용된다. 이와 같은 경우는 주로 어떤 성향의 유무와 선택의 유무를 의미한다. 또한 이것에서 응용된 순차적 프로빗은 종속변수가 2개 이상의 범주형 변수이면서도 서열(*ordinal*) 변수일 경우에 사용되는데, 이런 순차적 프로빗은 행위의 강도, 효과와 선택의 선호도 등을 설명하기에 적합하다.

종속변수가 양분 변수라 가정하면, 실제의 가치는 0과 1만을 갖게 되는데, 회귀 직선을 사용한다면, 예측치가 (0, 1)를 넘어선 곳에서도 존재할 수 있는 문제점이 발생된다. 이 문제의 명백한 해결책은 모든 독립변수들(X_{is})의 값에도 불구하고 Y_i의 예측치가 (0, 1)의 간격 내에 놓여지게 하기 위한 방법을 사용하여 원래의 모델을 전환시키는 것이다. 이 전환 과정에서 가장 필요한 일은 전 범위의 가치를 가질 수 있는 X_i 값을 0과 1 사이의 범위에서만 가능한 값을 갖는 확률로 전환시키는 것이다. 결국 누적확률함수(*Cumulative Probability Function*)를 사용해야 되며, X_i에 의해 결정되는 이론적(하지만 실제로 측정되지 않는) 지표 Z_i가 존재함을 가정해야 한다. 지표 Z_i는 무작위적이고(*random*), 정상 분포(*normal distribution*)를 갖는 연속 변수로 간주된다. 종속변수의 예측치는 주어진 독립변수들의 가치하에서 각 사례

가 각 범주에 속할 확률로서 해석될 수 있다.

프로빗과 로짓은 우도(尤度, *likelihood*)를 이용한 통계적인 추론을 위한 일반적인 모델이기 때문에, 우리는 이것을 바탕으로 관찰된 자료를 가지고 관찰되지 않은 우리의 관심사를 추론할 수 있다. 이 모델들에서는 선형회귀분석에서 해석이 가능한 모수 추정치(*parameter estimates*)와 모델의 적합도(*goodness-of-fit*)의 값이 제한적으로 사용된다. 일반 회귀분석에서 다상관결정계수(R^2)가 1이면 완벽한 모델의 적합도를 의미한다. 즉 종속변수가 회귀 곡선에 의해 정확하게 설명된다고 말할 수 있지만, 종속변수가 단지 두 개의 값을 갖는 경우에는 그러한 적합도란 본질적으로 불가능할 수밖에 없다. 최대우도추정(*Maximum Likelihood Estimation*, MLE) 방법에 의해 산출되는 우도비지수(*likelihood ratio index*)와 정확하게 예측된 사례들의 비율을 가지고 모델의 적합성 여부를 평가한다. 최대우도추정은 보통 최소제곱(OLS)과는 다른데, 그 이유는 최대우도추정은 작은 표본의 경우 불편성(*unbiasedness*), 효율성(*efficiency*), 정상성(*normality*)이 없게 된다. 하지만 큰 표본의 경우에는 점근적(*asymptotic*) 성향을 보이기 때문에 위의 특성들이 나타난다. 최대우도는 계산하기가 쉽고 빠르며 점근적으로 유효하다.

3. 프로빗과 순차적 프로빗으로 구성된 2단계 방정식 모델

이 글에서 예로서 소개하는 모델은 단순한 프로빗이 아닌 두 개의 방정식, 즉 프로빗(*probit*)과 순차적 프로빗(*ordered probit*)으로 구성된 2단계 방정식 모델이다. 이 방법의 사용이 어떤 경우에 적합하며, 어떤 특성을 지니고 있으며, 나온 결과들을 어떤 방법으로 어떻게 해석하여야 하는가를 실제의 구체적인 예를 사용하여 전개하고자 한다.

1) 2단계 방정식 모델의 소개

2단계 방정식 모델은 흔히 표본선택을 위한 2단 모델(*two-stage model*

with sample selection)이라고도 알려져 있으며, 중도절단 표본(*censored sample*)을 사용하는 경우의 논리와 같다. 여기에서는 프로빗과 순차적 프로빗을 두 개의 방정식으로 사용하는데, 이것들은 앞서 말했듯이 종속변수가 이산변수이거나 명목변수일 경우에 사용되는 일종의 변형된 형태의 회귀분석이다. 프로빗은 종속변수가 양분변수(0, 1)일 경우에, 그리고 순차적 프로빗은 종속변수가 순차적 다분(*polychotomous*)변수(0, 1, 2, 3, …)일 경우에 사용된다. 2단계 방정식 모델은 두 개의 각기 다른 단계(선택 방정식과 실제 방정식)를 통해 구체적으로 하고자 하는 연구 목적이 있다. 이 모델을 구성하는 두 가지 과정에서 사용되는 방법은 종속변수의 특징을 감안하여 연구자가 적절하게 선택하여야 한다.

(1) 선택방정식 (Selection Equation)

첫 단계로서 선택(*selection*) 과정은 우리가 궁극적으로 하고자 하는 연구를 위해 수집한 총 자료 중에서 필요한 사례(*cases*)만 선택할 때에 발생하는 위험부담률(λ_i = Hazard rate)을 얻기 위한 과정이라고 할 수 있다. 이 방정식에서의 종속변수는 양분 변수이기 때문에 베르누이 분포(*Bernoulli distribution*)를 기초로 하는 프로빗 모델을 사용한다.

$$Y_{1i} = \alpha + \sum \beta_k X_{ki} + u_i$$
$$u_i \sim N(0, \delta^2 I)$$

$$Y_{1i} = \begin{cases} 0 & \text{만일 } Y_i^* \leq \mu \\ 1 & \text{만일 } Y_i^* > \mu \end{cases}$$

관찰되어지지 않은 임의의 변수 Y_i^*는 단지 하나의 분계점($\mu = 0$)을 가지며, 다음의 방정식을 만족시킨다.

$$Y_i^* = \sum \beta_k X_{ki} + u_i$$
$$u_i \sim N(0, \delta^2 I)$$

결국, $Y_{1i}=1$의 가치를 얻는 함수는 임의의 변수 Y_i^*의 누적분포함수(*Cumulative Distribution Function*, CDF)와 같다.

$$\begin{aligned} P(Y_{1i}=1) &= P(Y_i^*>0) = P(u_i > -\Sigma\beta_k X_{ki}) \\ &= P(u_i < \Sigma\beta_k X_{ki}) = P(u_i < Z_i) \\ &= F(Z_i) \\ &= 1/\sqrt{2\pi} \int_{-\infty}^{Zi} e^{-t^2/2} dt \end{aligned}$$

지수 Z_i의 추정을 얻기 위해서는 역누적분포함수인 $F^{-1}(P_i)$가 적용될 수 있다.

(2) 실제 방정식(Substantive Equation)

이 방정식은 선택 방정식을 통해 선택된 사례만을 가지고 실제로 독립적으로 하고자 하는 연구에 초점을 둔다. 관찰되는 변수, Y_{2i}는 서열변수이며 이산변수의 특성을 갖는다. 결국, 회귀분석이 아닌 순차적 프로빗이 사용되는 것이 바람직하다.

$$\begin{aligned} Y_{2i} &= \alpha + \Sigma\beta_k X_{ki} + u_i \qquad u_i \sim N(0, \delta^2 I) \\ Y_i^* &= \Sigma\beta_{jk} X_{jki} + u_i \qquad u_i \sim N(0, \delta^2 I) \end{aligned}$$

$$y_{ij} = \begin{cases} 1 \text{ 만일 } \mu_{j-1} < Y_i^* \le \mu_j \\ 0 \text{ 만일 그 이외의 경우} \end{cases}$$

$$\begin{cases} i = 1, \cdots, n \\ j = 1, \cdots, M \\ M = \text{범주의 크기(숫자)} \end{cases}$$

$$Y_{2i} = \begin{cases} 1 \text{ 만일 } Y_i^* < 0 \\ 2 \text{ 만일 } 0 \le Y_i^* < \mu_2 \\ 3 \text{ 만일 } \mu_2 \le Y_i^* < \mu_3 \\ \cdots \end{cases}$$

관찰되지 않는 임의의 변수 Y_i^*는 연속변수로서 정상분포의 형태를 가지며, $\mu_1 = 0$ 과 함께 M-1 개의 분계점 매개변수(μ_1, μ_2, μ_3, μ_4, …)를 갖는다(Mckelvey and Zavoina, 1975; Maddala, 1983; King, 1989). 따라서 종속변수가 세 개의 범주를 가지고 있을 경우에는 $\mu_1 = 0$을 포함하여 두 개의 분계점(μ_1, μ_2)이 존재하게 된다.

2) 적용된 모델의 실례

사용되는 실례의 목적은 행정부 수반으로서의 정치적 지도자들의 재임기간(*duration of office*)을 가지고 그들의 성공성 여부를 살펴보는 것으로, '정치적 성공'의 개념은 '정치적 리더십의 유지'로 간주된다. 정치적 성공 또는 행정부 수반으로서의 직책 유지는 단순히 정치 지도자의 임무 수행 능력이라는 차원에서 그의 직책의 임기로 설명되어진다. 행정부 수반으로서 오래 머물렀던 지도자일수록, 더욱 더 성공적이라고 설명된다. 비록 한 행정부 수반의 정치적 수명이 이론적으로 정부의 안정성 또는 지구(*durability*)에 의해 설명될지라도, 직접적으로는 정부의 존속 차원이 아닌 정치지도자 개인의 생존성(*survival*)의 문제로 귀결된다.

정치적 리더십에 대한 연구가 사회배경이론(*social background theory*), 성격(또는 기술)이론(*personality or skill theory*), 환경이론(*situation theory*), 그리고 상호작용이론(*interaction theory*) 등을 기초로 행해져 왔으나 실제로 경험적이고 비교적인 견지에서 연구 분석된 것은 극히 찾아보기 어렵다. 지도자들의 사회적 배경 요소인 그들의 연령, 출생 지역, 교육의 정도 등과 지도자가 되기 전까지의 경력 요소인 정치적 경험의 유무, 장기적인 외국 경험, 정치적 이념성, 직업, 지도자가 된 방법 등은 지도자들의 행태적 패턴 또는 정책적 선호를 결정짓는 중요한 역할을 할 뿐 아니라, 정치체제가 필요로 하는 요구사항과 기능적인 관계를 갖는다.

성격 이론 또는 기술 이론은 어떤 특별한 성격적이고 기술적인 특성들이 지도자들의 효율성과 권력 추구 성향에 큰 영향이 있다고 설

명한다. 이와 더불어, 지도자들이 권력을 얻을 수 있도록 기회를 줄 수 있는 환경적인 조건들 — 국가의 나이, 인구밀도, 동질성, 문자 해득률, 정당 체제, 경제적 자유, 정치적 자유 등 — 또한 그들의 임무 수행력 또는 성공성과 관계가 있다는 것이 환경 이론의 중심 내용이다. 이런 개별적인 이론들과 달리 최근에는 정치적 리더십이 개인들의 차이뿐만 아니라 환경적인 차이의 영향도 함께 받는다고 설명하는 상호작용이론이 있다.

연구의 대상이 되는 지도자들은 1945년에서 1991년 사이에, 147개국에서 대통령으로서, 수상으로서, 또는 공산당 서기장으로서의 임무를 수행했던 자들이며,[4] 1991년 당시에 임무를 수행하고 있던 자들도 포함된다. 997명의 지도자들이 연구의 대상에 속하나, 이들 중에는 순수한 군주체제하에서 단지 국가를 상징하는 허수아비격의 지도자들은 제외되어 있으며, 또 집단 지배체제하의 지도자들도 제외되었다.[5] 하지만 짧게 임기를 마치기는 했지만, 임시(*stop-gap* 또는 *interim*) 지도자들은 포함되었다. 또한 997명 중에는 자료 수집을 하던 당시에 재임중이던 자들이 115명 있었지만 자료의 손실을 막기 위해서, 그들의 체류 기간이 3년 미만에 이르는 자들은 제외시켜 총 984명의 지도자가 연구 분석의 대상이 된다.[6]

지도자들의 체류 기간은 측정상 지속기간(*duration*)의 의미를 갖는데, 재임 기간의 측정단위는 '연'(*year*)으로, 그 기간이 6개월 미만일 경우는 0년으로, 6개월 이상이면 1년으로 기록된다. 예를 들면, 1년 3개월은 1년으로, 2년 8개월은 3년으로 측정된다. 이와 같이 재임기간의 성격이 이산적인 성격을 보이는 이유도 있지만, 성공성의 예측이라는 연구의 목적을 위하여 이를 범주화시켜 종속변수로 사용하고자 한다.[7]

재임기간의 빈도분포 형태를 살펴보면, 단기에 속할수록 빈도가 높아지고 장기로 갈수록 빈도수가 줄어드는 지수함수(*exponential function*)를 갖는다. 이와 함께 2년이라는 시점에서의 빈도수의 감소가 현저히 나타나기 때문에, 지도자들의 생존성의 구분을 이를 기점으로

단기와 장기로 나눌 수 있다. 한편, 생존성의 강도를 위한 범주 역시 빈도 분포에서 나타나는 특성과 함께 지도자들의 임기에 대한 제도적인 특성[8]을 고려하여 세 개로 나누기로 한다.

지도자들의 성공성을 설명하는 독립변수들은 이미 언급한 것처럼 그들의 사회적 배경, 경력 배경, 그리고 환경적인 조건들이 사용된다. 연구에서 사용될 자료 중에서 특히 지도자들에 관한 배경 조건들은 주로 객관적인 견지에서 인물들을 서술하는 전기, 인명사전, 신문에 실리는 사망기사(*obituary*)와 역사사전들로부터 수집되었다.[9] 각각의 이론들의 설명력을 비교하기 위해서는 각 이론을 기초로 한 2단계 모델이 필요하지만 여기에서는 상호작용이론을 기초로 한 2단계 모델만 소개하기로 하겠다.

2단계 모델은 두 방정식으로 구성되는데, 첫째는 지도자들의 체류기간(정권 유지 기간)이 단기(비성공적)인가, 장기(성공적)인가를 예측할 수 있는, 즉 생존성(*survival*)을 위한 선택 방정식, 둘째는 만일 어떤 지도자가 장기적인 체류를 했다면, 즉 성공적이라면, 성공성의 강도(생존성의 강도 또는 권력 유지의 강도)를 예측하기 위한 실제 방정식이다. 두 방정식은 각기 다른 체류 기간의 범주화를 통해 다음과 같이 설명된다.

(1) 생존성을 위한 선택방정식

선택 방정식에서의 종속변수의 값은 양분인 값, 즉 (0, 1) 인데 Y_{1i} (생존성)은 체류기간이 0~2년이면 0의 값을, 3년 이상이면 1의 값을 갖는다. 이 방정식에서 0은 단기체류를 의미하며, 1은 장기체류라고 해석한다. 독립변수들은 5개의 연속변수(연령, 동질성, 인구밀도, 문자해득률, 경제적 자유)와 27개의 모조변수(*dummy variable*)로 구성된다.[10] 방정식의 구체적인 형태는 다음과 같다.

$$\text{생존성} = \alpha + \beta_1 \text{연령} + \beta_2 \text{대학출신} + \beta_3 \text{도시출신} + \beta_4 \text{정치법조} + \beta_5 \text{공무원} + \beta_6 \text{군인} + \beta_7 \text{교수언론기술} + \beta_8 \text{정당} + \beta_9 \text{정당}$$

의회 + β_{10} 정당의회행정 + β_{11} 해외유학 + β_{12} 해외근무 + β_{13} 유학해외근무 + β_{14} 비헌법적 + β_{15} 자유적 + β_{16} 중도적 + β_{17} 보수적 + β_{18} 1960년 이전 + β_{19}대통령제 + β_{20} 단일정당 + β_{21} 복수정당 + β_{22} 동질성 + β_{23} 인구밀도 + β_{24} 문자해득률 + β_{25} 부분자유 + β_{26} 완전자유 + β_{27} 경제자유 + β_{28} 아시아 + β_{29} 아프리카 + β_{30} 중동 + β_{31} 유럽북미 + β_{32} 라틴아메리카 + u_i

선택의 단계는 두 가지의 중요 기능을 갖고 있다. 우선 단순하게 지도자들이 단기 체류인가, 장기 체류인가를 구분하는 기능이 있다. 그리고 다음 단계, 즉 실제 방정식에서 쓰여질 각각의 선택된 지도자들($Y_i > 0$)을 위한 위험 부담률(λ_i = Hazard rate)의 값을 얻기 위한 기능을 갖고 있다.[11] 위험부담률은 표준점수(Z-score)와 유사한 성격을 갖는 함수인데 다음과 같은 특징을 갖는다. 첫째, 각 사례가 생존성의 강도(Y_{2i})를 설명하는 자료 내에 포함될 확률을 의미한다. 둘째, 하나의 사례가 '생존성의 강도'를 위한 연구 자료에 포함될 가능성이 낮을수록 위험 부담률은 더욱 더 커진다고 볼 수 있다.

(2) 생존성의 강도를 위한 실제방정식

이 방정식은 실제의 연구 목적을 위해 사용되는 것으로, 생존성의 강도(Y_{2i})가 삼분 변수로 형태로 나타나는 종속변수가 된다. 장기적 생존의 범주에 속하는 581명의 지도자들만 다시 3개의 범주로 나뉘어져 분석된다. 즉, 357명은 1(3~6년)로, 131명은 2(7~12년)로, 그리고 93명은 3(13년+)으로 명시된다. 앞서 선택방정식과의 차이점은 모든 독립변수와 함께, 선택방정식에서 구해진 각 사례에 대한 위험부담률(*hazard rate*, HAZ)[12]이 실제 방정식에서는 하나의 독립 변수로서의 역할을 한다. 따라서 실제 방정식은 아래와 같다.

생존성의 강도 = 연령 + β_2 대학출신 + β_3 도시출신 + β_4 정치법조 + β_5 공무원 + β_6 군인 + β_7 교수언론기술 + β_8 정당 + β_9 정당의회 + β_{10} 정당의회행정 + β_{11} 해외유학 + β_{12} 해

외근무 + β_{13} 유학해외근무 + β_{14} 비헌법적 + β_{15} 자유적 + β_{16} 중도적 + β_{17} 보수적 + β_{18} 1960년 이전 + β_{19} 대통령제 + β_{20} 단일정당 + β_{21} 복수정당 + β_{22} 동질성 + β_{20} 인구밀도 + β_{24} 문자해득률 + β_{25} 부분자유 + β_{26} 완전자유 + β_{27} 경제자유 + β_{28} 아시아 + β_{29} 아프리카 + β_{30} 중동 + β_{31} 유럽북미 + β_{32} 라틴아메리카 + HAZ + μ + u_i

(3) 결과의 이해 방법

프로빗과 순차적 프로빗의 두 경우 모두 최대우도추정에 의하여 평가된다. 추정계수(*estimated coefficients*)는 각 독립 변수의 변화에 대한 효과를 설명하는 데 사용된다. 다만, 이 계수들은 회귀분석에서의 계수와 같이 간단하게 해석되지 않는다.[13] 왜냐하면 종속변수의 척도 혹은 '측정단위'가 불확실할 경우 기울기 혹은 추정계수는 독립변수의 한 단위 변화에 대한 종속변수의 변량으로서 해석될 수 없기 때문이다. 추정계수는 역누적분포함수에서 독립변수가 변화할 때 나타나는 효과를 설명한다. 하지만 이 계수가 회귀분석에서처럼 직설적으로 해석될 수 없는 이유는 확률에서의 증가는 최초의 확률에 의존하기 때문에 모든 독립변수들의 초기값과 계수에 따라 변화되기 때문이다. 모델 내에서 독립변수들의 영향력은 직접적으로 비교될 수 없다. 이 문제를 극복하기 위해서, 다시 말하면 독립변수들간의 비교를 위해서 표준화된 계수(*Standardized* β: βi)를 계산하는 방법도 있다.[14]

그럼에도 불구하고, 각 계수들은 독립 변수들과 종속 변수 사이의 관계에 대해 변화의 효과와 방향을 설명한다. 각 독립 변수들을 위한 가설들은 t 값으로 검증한다.[15] 계수의 기호는 변화의 방향을 의미하며, 크기는 F(Zi)에서 Zi에서의 누적분포함수의 경사를 반영한다. 누적분포함수의 경사가 클수록 독립변수의 변화가 주는 영향력은 더욱 커진다. 중요성 평가의 견지에서 비교적 큰 계수값(절대값)을 갖는 변수는 종속변수에 큰 영향을 준다고 할 수 있다. 즉, 0 혹은 1을 가질 추정 확률에 잠재적으로 큰 영향을 미치기 때문이다. 각 독립변수에

대한 추정계수는 독립변수와 종속변수 사이의 관계를 설명하고 있다.

2단계 방정식 모델에서 선택 방정식의 사용 목적은 다음 단계인 실제방정식을 위한 사례선택이고 이로서 각 사례의 위험부담률을 얻기 위한 과정이라고 하더라도, 각 독립변수가 종속변수에 미치는 영향에 대하여 분석해 볼 수 있으며, 다음 단계에서는 그 영향력이 어떻게 변화되었는가를 비교해 볼 수 있다. 〈표 1〉에서 각 독립변수의 계수에 대한 유의성을 검토해 보면, 각 모델에서 어떤 독립변수들이 생존성과 생존성의 강도를 구체적으로 설명하고 있는가에 대해서 알 수 있으며, 두 모델의 결과를 비교해 보면 독립변수들의 설명력 변화를 찾아볼 수 있다. 대체로 정치지도자가 행정부 수반이 될 당시의 연령이 낮을수록, 교육 수준이 낮을수록, 보수적인 이념적 성향을 지닐수록, 국민들의 문자해득률의 수준이 높을수록 또 해외경험이 많을수록 생존성이 높아지는 것을 알 수 있다. 하지만 정치지도자의 정치적 경험, 정치적·사회적 권리의 복합체로서의 시민의 자유, 경제활동의 자유는 단기체류냐 장기체류냐를 결정짓는 데 영향력을 행사하지만 생존성의 강도에는 전혀 영향을 주지 않는다. 반면, 생존성 자체에는 전혀 영향을 주지 않았던 요인들, 즉 지도자가 되기 이전의 직업이라든가 국가의 나이와 정당체제 등은 생존성의 강도를 설명하고 있음을 알 수 있다.

모델의 적합도를 위한 측정으로는, OLS 모델에서 추정된 다상관결정계수(R^2)가 아닌 우도비가 전반적인 모델의 중요성을 결정한다.[16] 우도비는 다음과 같이 계산된다.

$$\Delta = L(\theta')/L(\theta)$$
$$\Delta^* = -2 \log \Delta = -2\,[L^*(\theta') - L^*(\theta)]$$
$$= -2\,(LLR_0 - LLR_1)$$

LLR_0는 단지 상수(α)만 존재할 경우에 어떤 한 점 θ'에서 평가되는 제한된 모델을 위한 우도 함수의 log값을 의미하며, LLR_1은 모든

계수가 존재할 경우에 어떤 한 점 θ에서 평가되는 전체의 모델의 우도함수의 log값을 말한다. Δ^*은 계수의 숫자만큼의 자유도(*degrees of freedom*)를 갖는 χ^2 분포의 값과 비교하는 것에 의하여 검증될 수 있다. 우도비는 다른 모델과의 비교를 통해 통계학적인 유의성을 판단할 수 있지만, Y_i에서의 변량을 설명하는 모델의 전반적인 능력에 대한 직접적인 지표는 되지 않는다. 우도비가 높을수록, 추정이 더욱 잘된 것으로 해석될 수 있다.

R^2는 양분적 종속변수를 사용할 경우에도 가능할 수는 있지만 바람직하지는 않다. 회귀선 주변의 잔차 혹은 평균으로부터 종속변수의 편차를 관찰할 수 없기 때문에 순차적 프로빗 분석에서 R^2는 단지 실제의 R^2의 추정값에 불과하다(Mckelvey and Zavoina, 1975:112). 하지만 다양한 유사 R^2가 모델의 적합도 측정을 위해 소개되고 있다.[17] 헤이글과 미첼(Haglez and Mitchell, 1992:774)에 따르면, 알드리치-넬슨(Aldrich-Nelson)의 측정인 유사 $R^2 = -2\ LLR/(N-2LLR)$의 사용이 가장 바람직하다. 왜냐하면 계산상 용이하고 작은 표준오차와 OLS R^2와 회귀했을 때 더 높은 R^2 값을 갖기 때문이다. 게다가 알드리치-넬슨(Aldrich-Nelson)의 유사 R^2에 대한 수정치도 소개되었다. 하지만 양분적 프로빗에서는 사용이 가능할 수 있으나 순차적 프로빗에서는 사용되지 못한다.

모델의 적합도를 볼 수 있는 또 다른 방법은 종속 변수의 실제 가치와 예측치를 가지고 하나의 성공 분할표(*success table*)를 만들어, 얼마만큼의 비율의 사례들이 정확하게 예측되었는가를 살펴보는 방법이다. 프로빗의 경우는 2×2 형태의 분할표가 만들어지며, 순차적 프로빗은 3×3 형태의 분할표가 만들어진다.[18]

상호연관이론을 배경으로 한 2단계 방정식 모델의 결과를 〈표 1〉을 통해, 통계적 검증 가설의 기각 또는 채택의 방법을 이용하면, 두 방정식에서 종속변수와 유의한(*significant*) 관계를 갖는 독립변수들을 찾아볼 수 있다. 또한 모델의 적합도 측정치를 보면, 지도자들의 생존성의 경우 모델에서 사용된 설명 변수들로써 66.55%의 정확한 예측

을 할 수 있음을 보이며, 그들의 생존성 강도의 경우는 64.74%의 정확한 예측을 할 수 있다는 결과를 보이고 있다. 이와 함께 전체적인 모델에 대한 평가는 우도비의 값을 갖고 유의 여부를 검증할 수 있는데, 2개의 방정식에서, 포함된 독립변수들로 생존성과 그의 강도를 설명하는 것이 가능하다는 것을 의미한다. 각 방정식에서 어떤 범주가 어떻게 예측이 됐는가는 성공 분할표인 〈표 2〉로 설명할 수 있다.

선택 방정식의 경우, 성공적 예측이 66.55%이지만, 실제로 단기체류의 경우(37.5%)보다 장기체류 경우(84.1%)의 예측률이 훨씬 높다. 단기체류는 실제보다 0.2% 감소(37.7% → 37.5%)된 예측을 장기체류는 실제보다 21.8% 증가(62.3% → 84.1%)된 예측률을 보인다. 한편, 실제 방정식의 경우 범주 1(3~6년)은 실제보다 32% 증가(60.3% → 92.3%)된 예측률을, 범주 3(13년+)은 28.8% 증가(16.8% → 45.6%)된 예측률을 보이는 반면, 범주 2(7~12년)의 경우 오히려 실제보다 16.4% 감소(22.9% → 6.5%)된 예측률을 갖는다.

이와 더불어 모델의 적합도를 위한 다른 측정치와는 다르지만, 오차의 비례적 감소(*proportional reduction of error*, PRE)도 때때로 이용될 수 있다. 이것은 굿맨과 크루스칼의 램더(Goodman & Kruskal λ)와 같은 의미를 지닌 통계로서 두 범주 가운데 어느 한 범주의 값에 기초하여 다른 한 범주를 예측하는 입장에서 두 범주간의 상관관계를 산출한 값이다.[19] 이는 가트만의 예측계수(Guttman's coefficient of predictability)라고도 불리는데, 추가변수의 도입으로 오차의 크기를 얼마나 감소시킬 수 있는가에 기초하여 두 변수간의 상관관계를 나타내는 것을 의미한다. 오차의 감소가 크면 클수록, 종속변수가 독립변수를 사용하여 나온 예측률이 훨씬 높아지고, 변수 사이의 관계가 더욱 강해짐을 의미한다.

표 1

상호연관이론 모델의 계수 추정치와 모델의 적합도

선택 방정식				실제 방정식			
종속변수 : 생존성 (단기, 장기)				종속 변수 : 생존성의 강도			
변수	MLE	S. E.	t-score	변수	MLE	S. E.	t-score
상수	0. 300	0. 598	0. 503	상수	6. 005	2. 703	2. 221**
지도자의 연령	-0. 014	0. 005	-2. 915***	지도자의 연령	-0. 071	0. 026	-2. 718**
대학 출신	-0. 304	0. 157	-1. 935*	대학 출신	-1. 050	0. 570	-1. 842*
도시 출신	-0. 194	0. 010	-2. 004**	도시 출신	-0. 791	0. 355	-2. 228**
정치 · 법조계	-0. 524	0. 284	-1. 845*	정치 · 법조계	-1. 822	0. 992	-1. 838*
공무원	-0. 459	0. 307	-1. 495	공무원	-2. 226	0. 925	-2. 407***
군인	-0. 063	0. 309	-0. 205	군인	-0. 249	0. 321	-0. 777
교수 · 언론 · 기술	-0. 364	0. 278	-1. 308	교수 · 언론 · 기술	-1. 453	0. 722	-2. 011**
정당/의회/행정	0. 369	0. 292	1. 264	정당/의회/행정	0. 985	0. 710	1. 387
정당 · 의회	0. 499	0. 309	1. 614	정당 · 의회	1. 244	0. 921	1. 350
정당 · 의회 · 행정	0. 591	0. 312	1. 894*	정당 · 의회 · 행정	1. 515	1. 066	1. 421
유학 경험	0. 232	0. 120	1. 930*	유학 경험	1. 085	0. 443	2. 449**
해외 직장	0. 083	0. 136	0. 613	해외 직장	0. 484	0. 237	2. 307**
유학 · 해외 직장	0. 275	0. 154	1. 783*	유학 · 해외 직장	1. 077	0. 520	2. 072**
비헌법적 방법	-0. 151	0. 156	-0. 965	비헌법적 방법	-0. 431	0. 332	-1. 296
자유적 이념.	0. 270	0. 209	1. 292	자유적 이념	0. 826	0. 554	1. 491
중도적 이념	-0. 068	0. 266	-0. 255	중도적 이념	0. 187	0. 396	0. 471
보수적 이념	0. 477	0. 200	2. 384**	보수적 이념	1. 768	0. 900	1. 963**
1960년 이전 독립	-0. 210	0. 135	-1. 549	1960년 이전 독립	-1. 003	0. 402	-2. 496**
대통령제	0. 694	0. 138	5. 041***	대통령제	2. 194	1. 265	1. 735*
단일정당제	0. 083	0. 243	0. 341	단일정당제	0. 865	0. 352	2. 457**
복수정당제	-0. 260	0. 222	-1. 169	복수정당제	-0. 933	0. 547	-1. 707*
국가의 동질성	9. 7e-05	2. 7e-04	0. 356	국가의 동질성	1. 6e-04	3. 5e-04	0. 466
인구 밀도	-6. 5e-06	3. 0e-04	-0. 022	인구 밀도	4. 2e-04	3. 5e-04	1. 205
문자해득률	0. 007	0. 003	2. 659**	문자해득률	0. 027	0. 013	2. 025**
자유권 행사(부분)	0. 073	0. 177	0. 411	자유권 행사(부분)	0. 305	0. 246	1. 243
자유권 행사(완전)	0. 731	0. 242	3. 027***	자유권 행사(완전)	1. 956	1. 367	1. 431
경제활동의 자유	-0. 115	0. 696	-1. 654*	경제활동의 자유	-0. 366	0. 225	-1. 622
아시아	0. 311	0. 295	1. 054	아시아	0. 774	0. 632	1. 225
아프리카	0. 105	0. 306	0. 343	아프리카	0. 152	0. 376	0. 404
중동지역	-0. 019	0. 294	-0. 064	중동지역	-0. 342	0. 346	-0. 988
유럽 · 북아메리카	0. 014	0. 257	0. 055	유럽 · 북아메리카	0. 101	0. 300	0. 338
라틴아메리카	0. 174	0. 281	0. 617	라틴아메리카	-0. 374	0. 428	-0. 874
				HAZ(위험부담률)	-4. 258	2. 503	-1. 701
				μ	0. 905	0. 073	12. 441***
정확한 예측		66. 55%		정확한 예측		64. 74%	
-2×LLR		165. 20***		-2×LLR		207. 52***	
총계(N)		861		총계(N)		536	

* $p < 0.10$. ** $p < 0.05$, *** $p < 0.01$

표 2
상호연관이론에서의 예측에 대한 성공 분할표

생존성 (선택 방정식)

실제

예측		0 (단기 체류)	1 (장기 체류)	줄(N) 합계(%)
	0 (단기 체류)	122 37.5	85 15.9	207 24.0
	1 (장기 체류)	203 62.5	451 84.1	654 76.0
	칸(N) 합계(%)	325 37.7	536 62.3	861 100.0

생존성의 강도 (실제 방정식)

실제

예측		1 (3년~6년)	2 (7년~12년)	3 (13년+)	줄(N) 합계(%)
	1 (3~6년)	298 92.3	93 75.6	41 45.6	432 80.6
	2 (7~12년)	7 2.2	8 6.5	8 8.9	23 4.3
	3 (13년+)	18 5.6	22 17.9	41 45.6	81 15.1
	칸(N) 합계(%)	323 60.3	2123 22.9	90 16.8	536 100.0

4. 맺음말

이 글은 대부분의 사회과학 분야에서 가장 큰 장애요인이 되어 왔던 질적 자료를 통계적 모델에 적절하게 사용할 수 있음을 보여주기 위해 씌어졌다. 계량적인 연구에서 사용되는 변수들에 대한 자료는 조정변수의 역할을 하는 모조(*dummy*)변수를 제외하고는 모든 변수가 연속적인 성격을 갖는 것으로 간주된다. 하지만 실제로 우리가 연구를 위해 수집하거나 사용하는 자료의 성격을 면밀히 살펴볼 때, 그의 사용으로 인하여 연구의 결과가 왜곡되게 나올 수 있다. 질적 자료를

사용하여 분석을 하는 방법 — 빈도수, 그래프, 교차분석 등의 기본적인 통계처리방법 — 이 존재하지만 이런 묘사적이고 서술적인 해석을 넘어서는 설명력과 예측력을 보일 수 있는 연구방법이 필요하다.

우리가 흔히 사용하고 있는 회귀분석에서 비선형적인 문제를 해결하려는 방법으로 종속변수나 독립변수를 로그 변화(*logarithmic transformation*)하여 사용하는 경우가 있지만, 이는 비선형회귀를 운영시키기 위하여 선형회귀를 위한 컴퓨터 프로그램에서 '기교'(*tricking*)를 부릴 수 있는 유일한 방법일 뿐이다. 하지만 이것과는 아주 다른 문제로 무엇보다도 종속변수가 이산변수일 경우에는 회귀분석 사용에 필요한 조건에 위반되는 사항이 너무 많아 왜곡된 결과가 초래된다. 이산적 종속변수라는 특성을 고려할 때 주로 프로빗이나 로짓, 그리고 판별분석이 사용될 수 있는데, 각각의 모델이 지니고 있는 사용조건을 고려하여 연구의 적합한 모델 선택을 하는 일이 무엇보다도 중요하다.

이 글에서 소개한 연구방법인 프로빗 모델은 이산적 종속변수를 위한 모델이기는 하지만 회귀분석과는 성격을 달리한다. 이와 함께 선택방정식과 실제방정식으로 구성된 2단계 방정식 모델의 사용은 반드시 이산적 종속변수일 경우에만 사용되는 것은 아닐지라도, 프로빗과 순차적 프로빗을 함께 사용할 수 있는 사례를 보여주기 위해서이다. 선택방정식으로서의 프로빗은 사건이나 경험의 유무 혹은 선택의 차이를 양분적 값(0, 1)으로 표시할 수 있는 경우, 즉 선거참여 대 선거비참여, 상품의 구매 대 상품의 비구매, 폭력의 유무, 전쟁의 유무 등으로 표본 자체를 단체화(*grouping*)할 수 있을 때 가능하다. 반면, 순차적 프로빗은 다분적 값(0, 1, 2, 3, …)으로서 순차적인 의미를 주는 선택, 선호, 강도 등에 관련된 연구에 적합하다. 예를 들면, 상품에 대한 만족의 정도나 호감도, 후보자의 지지성향, 폭력의 강도 등이 이에 속한다. 다시 말하자면, 순차적 프로빗은 일단 사건이 존재할 경우에만 분석의 의의를 가질 수 있다고 할 수도 있다.

이 글에서 2단계 방정식 모델은 선택방정식으로서 프로빗은 표본을

중도절단하는 도구로서의 역할을 담당하면서도 독립변수들의 영향이나 모델의 적합도를 평가할 수 있으나, 대부분의 2단계 방정식 모델에서는 선택 방정식의 역할이란 다음 단계의 연구를 위해 표본을 걸러내는 데 불과하다. 실제방정식에 제공해 줄 위험부담률을 위해 선택방정식이 사용되는 것이다.

마지막으로, 프로빗과 순차적 프로빗의 분석시 중요한 문제는 회귀분석에서 나온 결과와 아주 유사한 통계들이 존재하지만 회귀분석에서와 같은 해석을 하지 못한다는 것이다. 특히 전체 모델의 적합도에 대한 평가에 유의하여야 하는데, 기본적으로 프로빗 모델을 사용할 경우에는 연구의 목적 자체를 사건 내지 행위에 대한 예측의 성공에 초점을 두는 것이 바람직하다.

■ 주

1) 정확하게 말하면, 문제의 전제 조건은 회귀직선 분석의 다섯 가지 전제 조건들 중에서, 종속변수와 일련의 독립변수들과 관계가 직선형 또는 선형적(*linear*)이어야 한다는 전제이다. 종속변수가 이산적(*discrete*) 변수일 경우 발생하는 문제는 다음과 같다.
   i) u_i가 정규분포가 아닌 이산적 분포를 갖는다.
   ii) u_i는 독립변수에 따라 체계적으로 변하게 된다. 즉, $E(X_i u_i) = 0$이라는 조건에 위반된다.
   iii) u_i가 이분산적(*heteroskedastic*)이다. $E(u_i u_j) = \delta^2 I$가 위반된다.
   iv) $E(u_i) = 0$의 조건이 위반된다. 다시 말하면, 잔차(*residual*)가 항상 +이거나 −이다.
2) 모집단이나 표본의 개체를 두 개의 군으로 분할하는 것을 의미한다.
3) $Pr(Y_i = 1) = F(X, \beta)$, $Pr(Y_i = 0) = 1 - F(X, \beta)$의 함수는 선형확률모델(*Linear probability model*: LPM), 프로빗 모델, 로짓 모델로 나눌 수 있다. Probit은 probability unit의 약자이고, Logit은 logistic probability unit을 의미한다. 때로는 probit을 nomit이라고 부르기도 한다(Aldrich and Nelson, 1984:37). 선형확률모델은 함수가 선형일 것으로 추정하는 모델이기 때문에 잔차의 분산, $Var(u_i)$은 X에 관계없이 일정한 것이 아니라 X에 따라 변화하며, $F(X, \beta)$도 1을 넘거나 0보다 작을 수 있다. 따라서 확률이 0~1의 범위 내에 있는 모델을 필요로 한다. 이를 만족시켜주는 것이 바로 프로빗과 로짓이다. 두 모델의 함수는 다음과 같다.

   프로빗 함수 : $P(Y_i = 1) = F(X, \beta) = \int_{-\infty}^{Zi} 1/\sqrt{2\pi}\ e^{-t^2/2}\ dt \equiv \Phi(\beta' X)$

로짓 함수 : $P(Y_i = 1) = F(X, \beta) = \exp(\beta'X)/[1+\exp(\beta')]$

두 모델의 누적분포함수는 그 형태에서는 큰 차이를 보이지 않으나, 두 곡선의 양끝의 두께에서 약간의 차이를 보일 뿐이다. 또한 결과의 분석에 있어서도, 로짓의 계수값(β)은 프로빗의 계수값의 1.8138($=\pi/\sqrt{3}$)배이다.

4) 만일 어떤 지도자가 선거 후 재선이 되어 계속 연임 할 경우는 하나의 사례로 간주되며, 한 번의 직책 임무 완료 후에 시간적 차이를 두고 다시 등장할 경우는 각기 다른 사례로 인정한다.

5) 군주제의 정치체제의 예는 1973년 이전의 아프가니스탄, 부탄, 브루나이, 바레인, 쿠웨이트, 모로코, 사우디 아라비아, 스와질란드, 오만, 요르단, 카타르, 아랍에미리트연합국 등이 있는데, 이들 국가 내에서 존재하는 수상들이란 바로 왕위계승에 따른 후계자들이다. 그리고 집단 체제로는 1955년~1965년 사이의 우루과이, 스위스, 그리고 1980년 이후의 유고슬라비아 등이 있다.

6) 자료수집의 최종 시기(1991년 말)에 재임기간이 2년 미만인 지도자는 13명이다.

7) 종속변수인 지도자들의 체류기간을 지속기간(*duration*)인 연속변수로 사용한 OLS 모델의 결과는 〈부록 1〉에 있다. 결과를 살펴보면, 독립변수들의 유의성 검증에서는 약간의 차이가 있긴 하여도 프로빗 모델의 결과와 유사하다. 하지만 모델의 적합도에서는 큰 차이가 있는데, OLS 모델에서 전체 모델의 적합도를 평가하는 R^2가 .20으로 비교적 낮은 값을 보이고 있어 설명력이 낮다고 할 수 있다. 이 밖에도 OLS의 경우 모델의 적합도를 평가하는 수단은 R^2 이외에는 다른 것이 없기 때문에 이 연구의 목적을 위해서 OLS 모델은 적합하지 않다.

8) 헌법적으로 재임과 연임 등에 관하여, 제한을 두고 있는 국가들에는 1988년 이후의 브라질, 1979 이후의 에콰도르, 1986년 이후의 과테말라, 1985년 이후의 온두라스와 대부분의 중남미 국가들, 1988년 이후의 한국, 미국, 탄자니아, 대만, 라이베리아 등이 있다. 이들은 재선이 전혀 허용되지 않거나, 재선이 허용되더라도 한 번 또는 두 번의 임기를 건너 뛴 후 가능하거나, 계속된 연임까지만 허용되는 경우들이다. 하지만 이에 속하는 사례들은 전체 대상을 기준으로 해서 볼 때, 비교적 적은 비율(997명 중 19명: 1.9%)을 보여주기 때문에 이런 헌법적 제한은 이 연구에서 고려하지 않기로 한다.

9) 예를 들면, *Current Biography, 1940~1991*. New York: H. W. Wilson Co. (*New York Times Biographical Edition Services, 1980~1991*, *The Annual Obituary, 1980~1991* (Chicago and London: St. James Press), *Who's Who in the World* (Chicago: Marquis Who's Who. Inc.) 등이 있고, 각각의 지역적인 구분에 따라 제작된 전기사전들이 존재한다.

10) 이들 중에서 정치적 경험과 해외 경험은 조작화(*operationalization*)를 위해서 다음과 같은 과정을 따른다. 정당, 의회, 행정 중에서 한 곳에서만 경험이 있을 경우, 이들 중에서 두 곳(정당·의회, 정당·행정, 의회·행정)에서 정치적 경험이 있는 경우와 세 곳 모두에서 정치적 경험을 한 경우를 구분한다. 해외 경험의 경우도, 유학 경험과 해외 근무의 두 개의 영역 중에서 단독적인 경우와 두 개 모두를 포함한 경우로 나눈다.

11) 위험부담률(λ_i)은 역밀스 비율(Inverse Mill's ratio) (Kennedy, 1992:246)라고 불리기도 하는데 이에 대한 설명은 다음과 같다.

$E(Y_i \mid X_i \text{ in sample}) = E(Y_i \mid X_i, BX_i + u_i > 0)$

$$= \beta' X_i + E(e_i \mid u_i > -BX_i)$$
$$= \beta' X_i + \theta \{\varphi(Z_i)/[1-\Phi(Z_i)]\}$$
$$= \beta' X_i + \theta \lambda_i$$

각 사례가 표본에 포함될 위험부담률은 우선 각 사례에 대한 $Z_i = -BX_i$의 추정된 평가치를 얻은 후, $\lambda_i = \varphi(Z_i)/[1-\Phi(Z_i)]$의 공식에 따라 계산된다. $\varphi(Z_i)$와 $\Phi(Z_i)$는 각각 표준정규분포의 밀도함수와 분포함수를 의미한다.

12) 위험부담률은 제1단계인 선택방정식 프로빗 모델에서 각 사례에 대해 계산된 수치들을 자료로서 저장하여 제2단계인 실제방정식인 순차적 프로빗 모델에서는 하나의 독립변수로 간주하여 사용하면 된다. 그 이유는 앞의 각주에서 설명되었듯이 선택된 표본으로 구성된 실제방정식에서 β와 θ를 추정하기 위해서는 λ_i가 독립변수로 사용되어야 하기 때문이다.

13) 각 독립변수에서, 한 단위의 표준편차의 증가 또는 감소는 그 독립변수를 제외한 다른 모든 변수가 각각의 평균치($Z_i = 0.0$, 확률 = 0.5)를 갖고 있다는 가정하에서, $Pr(Y=1)$에 영향을 준다.

14) 표준화된 계수는 회귀분석에서와 마찬가지로 $\dot{\beta_i} = \beta_i(\sigma_i/\sigma_j)$의 공식을 사용하여 계산한다.

15) t값은 각 독립변수의 유의성 즉, $H_0 : \beta_k = 0$에 대한 검증을 할 때 사용되는 기준으로 OLS 회귀분석에서와 마찬가지로 행해진다.

16) 모든 독립변수들이 포함된 전체 모델에 대한 $H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$을 검증한다.

17) Mckelvey-Zavoina(1976), Aldrich-Nelson(1984), Dhrymes(1986) 등이 Hagle and Mitchell(1992)의 연구에서 소개되고 있다.

18) 양분적 프로빗의 경우 $\hat{P} \geq 0.5$이고 $Y_{1i} = 1$ 혹은 $\hat{Pi} \leq 0.5$이고 $Y_{1i} = 0$일 경우는 성공적인 예측이 된다. 반면, 순차적 프로빗의 경우 Y_{2i}가 3개의 범주를 가지고 있기 때문에 다음과 같은 예측의 법칙을 따른다.
$P_1 = Pr(Y_{2i} = 1)$, $P_2 = Pr(Y_{2i} = 2)$, $P_3 = Pr(Y_{2i} = 3)$
만일 $[(P_1 \geq P_2) \ \& \ (P_1 \geq P_3)]$ 라면, 예측 = 1.
만일 $[(P_2 > P_1) \ \& \ (P_2 \geq P_3)]$ 라면, 예측 = 2.
만일 $[(P_3 > P_1) \ \& \ (P_3 > P_2)]$ 라면, 예측 = 3.

19) 오차의 비례적 감소(PRE)는 다음의 공식에 의해 계산된다.

$$PRE(\%) = 100 \times \frac{\%\ \text{correctly predicted} - \%\ \text{modal}}{100 - \%\ \text{modal}}$$

이것은 선택된 기본 값(해당 범주에서 가장 큰 값)에 대해 모델이 얼마만큼 향상되는가를 측정하는 것이지 일반적인 모델의 수행력을 의미하는 것이 아니다.

■ 참고문헌

Albert, J., S. Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Data." *Journal of the American Statistical Association* 88: 669~679.

Aldrich, John H., Charles F. Cnudde. 1975. "Probing the Bounds of Conventional Wisdom: A Comparison of Regression, Probit, and Discriminant Analysis." *American Journal of Political Science* 19(3) : 571~608.

Aldrich, John H., Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models.* Beverly Hills: Sage Publication.

Amemiya, Takeshi. 1975. "Qualitative Response Models." *Annals of Economic and Social Measurement* 4(3) : 363~372.

Bertschek, Irene, and Michael Lechner. 1997. "Convenient Estimators for the Panel Probit Model." *Journal of Econometrics* 87: 329~371.

Dubin, Jeffrey A., R. Douglas Rivers. 1990. *SST, Version 2.1.* Dubin/Rivers Research.

Eliason, Scott R. 1993. *Maximum Likelihood Estimation: Logic and Practice.* A Sage University Paper.

Greene, William H. 1992. *LIMDEP, Version 6.0: User's Manual and Reference Guide.* Bellport, N.Y.: Econometric Software.

Griffiths, W. E., R. C. Hill, and P. J. Pope. 1987. "Small Sample Properties of Probit Model Estimators." *Journal of the American Statistical Association* 82: 929~937.

Hagle, Timothy M., and Glenn E. Michell II. 1992. "Goodness-of fit Measure for Probit and Logit." *American Journal of Political Science* 36(3) : 762~784.

Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample selection and Limited Dependent Variables and a Sample Estimator for such Models." *Annals of Economic and Social Measurement* 5: 475~492.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 153~161.

Kaplan, David, and Richard L. Venezky. 1994. "Literacy and Voting Behavior: A Bivariate Probit Model with Sample Selection." *Social Science Research* 23: 350~367.

Kennedy, Peter. 1992. *A Guide to Econometrics*, 3rd edition. Cambridge, MA: The MIT Press.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* Cambridge: Cambridge University Press.

Lee, Lung-Fei. 1982. "Some Approaches to the Correction of Selectivity Bias." *Review of Economic Studies* 49: 355~372.

Maddala, G. S. 1983. *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Mckelvy, Richard, and William Zavoina. 1975. "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." *Journal of Mathematical Sociology* 4: 103~120.

Pindyck, Robert S., and Daniel L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* 4th edition. Boston, MA: Irwin McGraw-Hill.

Van de Ven, W. P. M. M., and B. M. S. Van Pragg. 1981. "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection." *Journal of Econometrics* 17: 229~252.

Yatchew, A., and Z. Griliche. 1985. "Specification Error in Probit Models." *Review of Economics and Statistics* 18: 134~139.

■ 부록 1

아래 결과는 지도자의 체류기간을 연속적으로 간주하여 OLS 회귀모델을 사용해 나온 것이다. 종속변수인 체류기간의 범위는 최소 0에서 최대 44이다.

Ordinary Least Squares Estimation			
종속변수: 지도자의 체류기간(DURATION)			
변수	계수	S. E.	t-statistics
상수	10. 13656	2. 54477	3. 98330***
지도자의 연령	-0. 11623	0. 02114	-5. 49769***
대학출신	-1. 08637	0. 64495	-1. 68444*
도시출신	-1. 04769	0. 41331	-2. 53484**
정치 · 법조계	-3. 38199	1. 13111	-2. 98997***
공무원	-4. 27806	1. 24275	-3. 44242***
군인	-1. 50243	1. 23204	-1. 21946
교수 · 언론 · 기술	-3. 31387	1. 10553	-2. 99753***
정당/의회/행정	0. 40314	1. 30247	0. 30952
정당 · 의회	0. 85451	1. 37807	0. 62008
정당 · 의회 · 행정	0. 81863	1. 38737	0. 59006
유학경험	1. 23355	0. 51625	2. 38945**
해외직장	0. 59810	0. 58392	1. 02428
유학 · 해외직장	1. 50344	0. 65585	2. 29236**
비헌법적 방법	0. 33858	0. 66742	0. 50729
자유적 이념	2. 08950	0. 90527	2. 30814**
중도적 이념	1. 75370	1. 15409	1. 51955
보수적 이념	2. 92347	0. 86769	3. 36928***
1960년 이전 독립	-1. 49812	0. 57361	-2. 61174***
대통령제	1. 90623	0. 56955	3. 34689***
단일정당제	3. 27892	1. 04722	3. 13107***
복수정당제	-0. 09300	0. 96138	-0. 09673
인구밀도	0. 00078	0. 00094	0. 83106
국가의 동질성	-0. 00044	0. 00108	-0. 40554
문자해득률	0. 02031	0. 01151	1. 76524*
경제활동의 자유	-0. 33774	0. 28799	-1. 17277
자유권행사(부분)	0. 66352	0. 75779	0. 87560
자유권행사(완전)	1. 09754	1. 01599	1. 08027
아시아	0. 99248	1. 23797	0. 80170
아프리카	0. 09237	1. 28693	0. 07178
중동지역	-0. 44012	1. 25672	-0. 35021
유럽 · 북아메리카	1. 37779	1. 09892	1. 25377
라틴아메리카	-0. 36371	1. 18041	-0. 30812
Number of Observations	871	R-squared	0. 20412
Corrected R-squared	0. 17373	Sum of Squared Residuals	27714. 8
Standard Error of the Regression	5. 75087	Durbin-WatsonStatistic	1. 85944
Mean of Dependent Variable	5. 25373		
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$			

■ 부록 2 Probit을 위한 컴퓨터 프로그램에 대한 소개

일반적으로 사회과학을 위해서 개발된 많은 프로그램들에는 probit 모델과 logit 모델을 위한 부분이 존재한다. 하지만 이 모델들을 사용하여 얻고자 하는 결과는 연구자에 따라 요구되는 수준을 달리한다. 우리가 흔히 쉽게 접할 수 있는 SPSS와 SAS 등은 가장 기본적으로 필요한 단순한 결과만을 보여주는 경향이 있다. 보다 깊이 있는 분석을 위해서는 연구자의 목적을 충족시키는 프로그램의 사용이 필요하다. 이 글에서 사용된 프로빗/순차적 프로빗(*probit/ordered probit*) 모델은 SST(*Statistical Software Tools*, ver. 2.1)라는 프로그램과 함께, 중도 절단된(*censored*) 표본이나 절단된 표본 등 제한된 종속(*limited dependent*) 변수를 사용하는 모델들을 위한 LIMDEP (ver. 6.0)이라는 프로그램을 사용하였다. 전자의 프로그램은 사용이 쉽고, 필요한 통계적 수치를 간단하게 얻을 수 있는 반면에, 분석을 위해 사용되는 자료의 크기가 어느 정도 제한되는 단점이 있고, 그래프 처리에 한계가 있다. 이와는 달리, LIMDEP은 사용이 다소 어렵지만, 자료의 크기와 그래프 처리와 함께 얻고자 하는 통계적 결과 또한 아주 상세하게 제공하고 있다. 이와 함께 SHAZAM 역시 프로빗 분석을 제공하고 있다. 이들 프로그램은 실제로 사회과학자들이 사용하기에 다소 어려울 수 있으므로 반드시 프로그램을 위한 사용법을 담은 안내서를 갖추고 있어야 한다.

프로빗 모델을 사용하기에 적합한 프로그램 패키지에 대한 정보는 다음과 같다.

1. Dubin, Jeffrey A. and R. Douglas Rivers. 1990. SST, Version 2.1. Dubin/Rivers Research.
2. Greene, William H. 1992. LIMDEP, Version 6.0. Econometric Software, Inc.
3. White, K. J. 1993. SHAZAM User's Reference Manual Version 7.0. McGraw-Hill, New York.

2. Probit and Ordered Probit Analysis and Its Application

Mee-young Ju

Research on social science has been conducted based on various kinds of empirical statistical analyses. The OLS regression analysis that has been most popularly used is liable to misestimate the effects of independent variables seriously when its dependent variable is discrete. If at least one of the basic assumptions in regression is violated, the model should be replaced with a proper model. A probit and an ordered probit would be recommended for a categorical or discrete dependent variable with more than two values. Whereas probit is applied for the dichotomous(0, 1) dependent variable, ordered probit is for the polychotomous(0, 1, 2, ···) dependent variable. In this paper, a two-equation model, which is composed of a probit selection equation as well as an ordered probit substantive equation, is used to estimate its parameters. The results suggest that the model has not only explanatory but also predictive value.

3. Applying Formal Theory in Public Administration

Sang-mook Kim

Formal theory is to explicitly state assumptions which can logically and consistently capture the important aspects of the situation under study, and to deductively derive conclusions which proceed from the assumptions. It is also called as formal mathematical theory because it is expressed mathematically and solved using the techniques of mathematics. This paper is introduced formal theory