■ 論 文 ■

로짓모형에 있어서 다중공선성의 영향에 관한 연구

Effects of Multicollinearity in Logit Model

류 시 균
(경기개발연구원 연구위원)

목 차

Key Words : 로짓모델, 다중공선성, 수치실험, 로짓모형의 신뢰도, 로짓모형의 적합도
Logit Model, Multicollinearity, Numerical Experiments, Reliability of Logit, Fitness of Logit

요 약

비확률변수간 선형관계로 정의되는 다중공선성은 설명변수간 선형방정식으로 표현되는 회귀모형의 신뢰도를 저하시키기 때문에 회귀모형의 구축과정에서는 세심한 검토와 대응이 이루어진다. 본 연구에서는 구조화된 수치실험을 통해서 로짓모형에 대한 다중공선성의 영향을 규명하였다. 효용함수를 구성하는 설명변수들간 상관관계의 정도에 따라서 추정된 모형의 적합도 지표와 계수의 신뢰도 지표가 어떻게 변동하는 지를 추적함으로써 다음과 같은 시사점을 확인할 수 있었다. 첫째, 설명변수의 추가를 통해서 모델의 적합도 개선이 가능한 회귀모형과 달리, 로짓모형에서는 효용함수에 설명변수를 추가하는 경우 로짓모형의 적합도가 개선될 수도, 역으로 저하될 수도 있음이 확인되었다. 둘째, 공통의 계수를 갖도록 모델을 구성하면 제네릭 변수간 상관관계가 높아짐에 따라 모델의 적합도가 저하됨을 확인하였다. 셋째, 설명변수간 상관관계가 높은 경우 선택행동에 대한 설명변수의 기여도가 과대평가될 가능성을 확인하였다. 넷째, 설명변수간 상관관계가 높으면 추정된 계수의 신뢰도가 저하됨을 확인하였다. 결론적으로 본 연구를 통해서 그동안 로짓모형의 구축과정에서는 주목받지 못했던 다중공선성이 실제로는 세심한 배려와 적절한 대응을 통해서 제어되어야 함이 규명되었다.

This research aims to explore the effects of multicollinearity on the reliability and goodness of fit of logit model. To investigate the effects of multicollinearity on the multinominal logit model, numerical experiments are performed. The exploratory variables(attributes of utility functions) which have a certain degree of correlations from (rho=) 0.0 to (rho=) 0.9 are generated and rho-squares and t-statistics which are the indices of goodness of fit and reliability of logit model are traced.

From the well designed numerical experiments, following findings are validated : 1) When a new exploratory variable is added, some of rho-squares increase while the others decrease. 2) The higher relations between generic variables lead a logit model worse with respect to goodness of fit. 3) Multicollinearity has a tendency to produce over-evaluated parameters. 4) The reliability of the estimated parameter has a tendency to decrease when the correlations between attributes are high. These results suggest that we have to examine the existence of multicollinearity and perform the proper treatments to diminish multicollinearity when we develop logit model.

Ⅰ. 서론

1. 연구의 배경 및 목적

다중공선성(Multicollinearity)이란 비확률변수들(Non-Stochastic Variables) 사이에 선형관계가 존재하는 경우를 지칭한다(이성우 등, 2005). 다중공선성은 다중회귀모형에서 추정계수의 분산을 증대시켜 결과적으로 설명변수의 신뢰도를 저하시키기 때문에 회귀모형의 구축과정에서 세심한 검토와 대응이 이루어진다. 그러나 함수의 구조나 계수의 추정방법 측면에서 다중회귀모형과 유사한 구조를 갖고 있는 로짓모형[1]과 관련해서는 다중공선성의 영향이 크게 주목받고 있지 못한 것이 사실이다[2].

로짓모형과 관련해서 다중공선성이 크게 주목받지 못한 데에는 나름대로 몇 가지 이유가 있을 것으로 추정된다. 필자는 여기에는 두 가지의 요인이 핵심적으로 작용하고 있다고 추론하고 있는데, 첫째는 다중회귀모형의 발달과정에서 다중공선성의 영향과 대응방안이 충분히 검토되었기 때문에 로짓모형과 관련해서 새삼스럽게 다중공선성을 논할 필요성을 느끼는 연구자가 많지 않다는 것이고, 둘째로는 다중회귀분석에서 다중공선성을 완화시키는 방안 가운데 하나로 설명변수의 대수변환방식이 제안되고 있기 때문에 지수함수(Exponential Function)를 통해서 변수변환과정을 거치는 로짓모형에서는 다중공선성의 문제가 어느 정도 완화되었을 것으로 생각하기 때문으로 추정된다. 실제로 다항로짓모형의 분산공분산 행렬, $E(-\nabla V)^{-1}$에서 $-\nabla V$의 k행 l열 원소는 식(1)과 같은데(토목학회, 1995), 이는 추정해야 할 계수의 분산(또는 표준편차)이 효용함수내 설명변수들의 지수함수로 정의됨에 따라 회귀모형에 있어서의 다중공선성을 완화하기 위한 과정, 즉 설명변수의 대수변환이 부분적으로 수행되었다고 해석하는 것도 무리가 아니다. 결과적으로 설명변수들간 선형관계(다중공선성)로 인한 추정계수의 신뢰도 저하는 다중회귀분석에서 만큼은 발생하지 않을 것이라는 추론은 일면 타당해 보인다.

$$-\sum_{n=1}^{N}\sum_{i\in A_n} P_{\in}(x_{ink}-\sum_{j\in A_n} x_{jnk}p_{jn})(x_{inl}-\sum_{j\in A_n} x_{jnl}p_{jn}) \qquad (1)$$

여기서, N은 표본의 수, i, j는 개인 n의 대안집합 A_n에 포함되어 있는 선택대안, $P_{in}(=\frac{\exp(V_{in})}{\sum_{j\in A_n}\exp(V_{jn})})$은 개인 n이 대안 i를 선택할 확률, x_{jnk}는 개인 n의 선택대안 j의 k번째 (효용함수내)설명변수 값이다.

그러나 회귀분석에서의 대수변환방식은 상관관계가 높은 두 변수 중에서 하나의 변수에 대해서만 적용해야 하기 때문에 설명변수들의 선형방정식으로 표현된 효용함수 전체를 대상으로 지수함수를 취하는 로짓모형에서는 다중회귀모형에서와 같은 정도로 다중공선성이 완화될 것으로 예상되지는 않는다. 다시 말해서 다중공선성에 대한 세심한 주의를 기울이지 않으면 다중회귀모형에서 나타나는 다중공선성의 문제가 로짓모형에서도 나타날 수 있을 것으로 예상된다.

본 연구는 구조화된 실험을 통해서 효용함수내 설명변수들간 다중공선성이 로짓모형과 추정된 계수들의 신뢰도에 어떠한 영향을 미치는가를 실증적으로 규명하는데 목적이 있다. 부연하자면 추정된 모형의 신뢰도를 평가하는데 있어서 다중공선성의 영향이 고려되어야 함을 주장하기 위해서 본 연구는 기획되었다.

2. 선행연구 고찰

1) 로짓모형에 관한 연구동향

로짓모델에 관한 국내연구의 대부분은 사례연구에 집중되어 있는 것으로 파악되고 있다. 국내 저널에 수록된 논문 가운데 박상준・김성수(2007)는 네스티드 로짓모형을 이용해서 승용차 보유대수와 차종선택행태를 규명하였으며 박규영・이수범(2006)은 지방부 4차로 국도상에서의 보행자교통사고 발생확률을 역시 로짓모형을 이용해서 산정하였다. 서상언・정진혁・김순관(2006)은 일반인과 고령자의 통행특성 차이를 분석하기 위해 네스티드 로짓모형을 이용한 활동 스케줄링 모형을 구축하였으며, 이영인・황준환(2001)은 간선도로 돌발상황 검지 모형으로서 로짓모형을 활용하였다. 김강수(2002)는 시뮬레이션 방법을 활용한 화물수단선택모형을 제안하였으며, SP데이터에 내재된 편의를 제거하기 위한 또

1) 본 연구에서는 다항로짓모형 이외의 이산선택모형, 가령 네스티드 로짓모형이나 프로빗모형 등에 대해서는 논외로 한다.

2) 다중회귀모형과 로짓모형내 효용함수는 모두 선형방정식 형태를 취하는 것이 일반적이며 다중회귀모형의 보편적 추정방법인 최소자승법(Ordinary Least Square, OSL)과 로짓모형의 보편적 추정방법인 최우추정법(Maximum Likelihood method, MLM)은 모두 Newton-Rapson법과 같은 Gradient방식 계열의 최적화기법이 활용된다.

다른 연구로서 김강수·조혜진(2004)은 SP순위자료를 활용한 순위 로짓모형의 오차발생 메카니즘을 규명하고 개선방안을 제시하였다. 이상의 두 논문은 국내 논문으로서는 드물게 로짓모형의 신뢰도 개선방안을 검토하였다는 점에서 의의를 찾을 수 있다.

한편 국외에서는 로짓모형의 개량을 목적으로 한 연구가 셀 수 없을 정도로 많이 수행되어 왔고 지금도 수많은 연구자들이 로짓모형의 개량을 위한 연구에 매진하고 있다. 특히, 본 연구에서 검토한 다항로짓모형의 개량을 목적으로 한 연구논문들을 유형별로 구분하면 다항로짓모형의 구축과정에서 도입된 가정에서 파생되는 문제점을 완화하고자 하는 연구, 연속변량을 선택대안으로 도입하기 위한 연구(Ben-Akiva et al., 1981), 계층간 동질성 가정을 완화하기 위한 연구(Cardel et al., 1980), SP데이터에 내재된 bias 문제를 해소하기 위한 연구(Morikawa, 1989) 등이 있다. 물론 본고에서 제시한 모델개량 연구분야에 포함되지 않는 로짓모형 관련 연구들도 많지만 모든 연구성과를 모두 망라해서 기술하는 것은 본 연구의 취지에 부합되지 않는다고 판단된다.

다항로짓모형에 내재된 가장 대표적인 가정은 효용함수의 오차항이 서로 독립이면서 동일한, 즉 IID(Identical and Independent Distribution)의 감벨분포를 따른다고 하는 가정으로서 다항로짓모형의 가장 치명적인 약점으로 지적되고 있는 IIA문제와 등분산성 가정에 의해서 야기되는 현실묘사력의 저하문제를 야기한다. 오차항의 독립성 가정에서 야기되는 IIA문제를 완화하기 위해 Nested Logit모델이나 Probit 모델이 이미 오래 전부터 대안적 방안으로서 개발되어 활용되고 있으며(토목학회, 1995; Ben-Akiva et al., 1987), 오차항의 등분산성 가정을 완화하기 위한 연구로는 William et al.(2006), Frank et al.(2005), Karthik et al. (2003), Karthik et al.(2005) 등이 있다. 본 연구의 목적은 효용함수의 설명변수간 상관관계가 다항로짓모델의 적합도와 추정된 계수의 신뢰도에 미치는 영향을 규명하는데 있으며, 이와 같은 연구사례는 발견되지 않았다.

2) 로짓모형과 다중공선성

로짓모형과 관련된 연구의 흐름에 있어서 다중공선성 문제는 주요 관심분야는 아닌 것으로 판단된다. 로짓모형을 전문적으로 다루고 있는 몇몇 텍스트에서 조차 다중공선성에 관한 내용을 찾아보기 어렵다는 사실이 필자의 추론을 반증하고 있다(토목학회 1995, Ben-Akiva et al. 1987, Washington et al. 2003, Ortúzar et al. 1998, Gärling et al. 1998, Oppenheim 1994) 로짓모형을 전문적으로 다루고 있는 국내 저서 가운데 이성우 등(2005)에서 다중공선성에 관한 상세한 내용이 발견되지만 로짓모형 보다는 다중회귀분석과 관련된 내용으로 일관하고 있어 본 연구에서 참조하기에는 한계가 있다.

로짓모형과 유사한 형태를 취하고 있는 로지스틱 회귀모형에 관한 자료에서는 비교적 다중공선성에 관해서 상세하게 다루고 있는데, NC State University의 통계학 강의노트는 로짓모형에 대한 다중공선성의 영향을 부분적으로 추론하는데 유용하다 판단된다[3]. 상기 자료에 의하면 "로지스틱회귀모형에 있어서 설명변수간 상관관계가 커지면 회귀계수의 표준편차는 커지지만 회귀계수의 값 자체는 변하지 않는다"라고 기술되어 있다. 다시 말해서 회귀계수의 신뢰도는 저하되지만 추정량 자체는 보편적인 다중회귀모형에서와 같이 과대 또는 과소추정되지 않는다는 것이다.

로짓모형에 있어서의 다중공선성의 영향이 로지스틱 회귀모형에 대한 다중공선성의 영향과 유사한 형태로만 작용한다면 다중공선성으로 인해서 로짓모형의 설명변수를 과대 또는 과소평가하는 오류는 발생하지 않으며, 결과적으로 추정된 계수의 신뢰도가 논리적 관점에서 과다하게 낮게 도출된 경우 데이터의 결함을 점검하기에 앞서 다중공선성의 영향을 검증함으로써 모형 추정을 경제적으로 수행할 수 있게 된다.

3) 다중공선성의 해석적 고찰

선형방정식으로 정의된 일반적 다중회귀모형에서 추정된 회귀계수의 분산공분산행렬은 다음과 같이 주어진다(유지성 외, 2004).

$$Var(\alpha) = \sigma^2 (XX')^{-1} = \sigma^2 (\Sigma x_{ij} x_{ik})^{-1} \quad (2)$$

여기서 α는 회귀계수벡터, σ^2는 모분산벡터, X는 설명변수벡터, $x_{ij(k)}$은 i번째 관측치의 $j(k)$번째 설명변수값이다. 이해를 쉽게 하기 위해서 설명변수가 2개인 경우를 대상으로 논의를 전개해 보자. 위 식은 다음과 같이 전개된다.

3) (http://www2.chass.ncsu.edu/garson/PA765/logistic.htm)

$$Var(\alpha) = \sigma^2 \begin{pmatrix} \Sigma x_{i1}^2 & \Sigma x_{i1}x_{i2} \\ \Sigma x_{i2}x_{i1} & \Sigma x_{i2}^2 \end{pmatrix}^{-1} \quad (3)$$

추정될 두개의 회귀계수에 대한 분산은 다음과 같이 정리된다.

$$Var(\alpha_1) = \sigma_1^2 \frac{\Sigma x_{i2}^2}{\Sigma x_{i1}^2 \Sigma x_{i2}^2 - (\Sigma x_{i1}x_{i2})^2} \quad (4)$$

$$Var(\alpha_2) = \sigma_2^2 \frac{\Sigma x_{i1}^2}{\Sigma x_{i1}^2 \Sigma x_{i2}^2 - (\Sigma x_{i1}x_{i2})^2} \quad (5)$$

두 변수(x_1, x_2)의 상관계수(r)는 $\frac{\Sigma x_{i1}x_{i2}}{\sqrt{\Sigma x_{i1}^2 \Sigma x_{i2}^2}}$이므로 상관계수가 1 또는 -1에 근접할수록 식(4)와 식(5)의 분모는 0에 수렴하고 결과적으로 회귀계수의 분산은 무한대로 커진다. 이는 회귀계수의 신뢰도지표 즉, t값($= \alpha / \sqrt{Var(\alpha)}$)이 0에 수렴해감을 의미하며, 결과적으로 회귀계수의 신뢰도는 저하된다.

그러나 로짓모형에 있어서 추정될 계수의 분산공분산 행렬은 식(1)에서 보는 바와 같이 회귀모형의 그것에 비해서 매우 복잡한 구조를 갖고 있고 또한 지수함수를 취하고 있어 선형방정식으로 표현되는 보편적 단순회귀모형처럼 함수를 통해서 단순하게 해석적으로 고찰하기에는 어려움이 있다. 본 연구가 구조화된 수치실험을 통해서 로짓모형에 대한 다중공선성의 영향을 검증하게 된 이면에는 회귀모형과 같이 해석적인 방법으로는 다중공선성의 영향을 파악하는데 한계가 있기 때문이다.

Ⅱ. 실험의 개요

1. 실험의 기본개념

본 연구의 목적은 다중공선성이 로짓모형의 전체적 신뢰도와 추정된 계수의 값 그리고 추정된 계수의 신뢰도에 미치는 영향을 규명하는데 있다. 따라서 효용함수 내 설명변수간 상관계수가 실험에 있어서 중요한 제어변수가 된다. 분산분석의 관점에서 기술하자면 효용함수를 구성하는 설명변수간 상관관계가 인자(因子)가 되고 상관계수의 값이 수준(Level)이 된다. 즉, 효용함수를 구성하는 설명변수간 상관계수의 값의 변화에 따라서 모형의 설명력(ρ^2), 계수의 신뢰도(t_α) 그리고 계수값(α_j)의 변동을 살펴볼 것이다.

한편, 회귀분석에서는 고려되는 모든 설명변수들이 하나의 회귀방정식에 포함되지만 여러 개의 효용함수를 필요로 하는 다항로짓모형에서는 설명변수들이 여러 개의 효용함수로 분산되어 배치된다. 효용함수를 단위로 지수함수를 취하는 로짓모형에 있어서는 상관관계를 맺고 있는 변수들이 동일한 선택대안의 효용함수에 존재하는가 아니면 서로 다른 효용함수의 설명변수로 존재하는가에 따라서 다중공선성의 영향도 다를 것으로 예상된다. 회귀분석과는 다른 로짓모형 특유의 구조적 특성을 다중공선성과 연계해서 분석하기 위해서 본 연구에서는 몇 가지 시나리오를 설정하고, 시나리오별로 효용함수를 구축해서 실험을 수행한다. 자세한 내용은 2절의 '1) 시나리오 설정'을 참조하기 바란다.

마지막으로 실험은 연구자에 의해서 인위적·구조적으로 만들어진 데이터를 활용해서 수행하게 된다. 좀 더 부연해서 설명하자면, 수치실험에서 활용될 효용함수의 구조, 설명변수의 값, 오차항의 값, 선택결과 및 설명변수의 계수 등, 모형과 관련된 모든 요소들은 사전에 연구자에 의해서 인위적으로 만들어져 있다는 것이다. 실험의 구조와 관련해서는 3절의 '3) 로짓 데이터의 생성'을 참조하면 이해가 쉬울 것이다.

2. 실험조건의 설정

본 연구에서는 이항로짓모형을 이용해서 로짓모형내 두 변수간 다중공선성의 영향을 검증하고자 한다. 기본적으로 다중공선성이란 다수의 설명변수들 사이에 선형관계가 존재하는 경우로 정의되지만 두 변수간의 상관관계에 따른 영향분석결과를 통해서 다수의 변수들 사이에 선형관계가 존재하는 경우를 추론할 수 있고, 더욱이 본 연구에서는 하나의 방정식으로 정의되는 회귀모형과 달리 여러 개의 효용함수로 정의되는 로짓모델 특유의 구조적 특성에 따른 다중공선성의 영향을 분석하고자 하기 때문에 지면제약도 고려해서 여러 변수들간 선형관계가 존재하는 경우에 대한 영향 검증은 차기 연구과제로 남겨두고자 한다. 기본적으로 본 연구에서 활용할 효용함수의 기본형은 식(6)과 같다.

$$U_1 = V_1 + \epsilon_1 \quad (6)$$
$$U_2 = V_2 + \epsilon_2$$

여기서 U_1, U_2는 대안 1, 2의 효용이고, V_1, V_2는 각각 대안 1, 2의 관측가능한 효용성분, ϵ_1, ϵ_2는 관측불가능한 효용성분이다.

한편, 다중공선성에 대한 영향을 이항로짓모형으로 검증하더라도 분산공분산행렬의 구조가 동일한 다항로짓모형으로 확장해서 해석하는 데에는 무리가 없을 것으로 판단된다. 단, 본 연구를 통해서 도출된 시사점을 모델의 분산공분산행렬의 구조가 다른 모형(가령 네스티드 로짓모형)으로 확장해서 해석하는 것은 옳지 않음을 지적해 둔다.

1) 시나리오 설정

전절에서 언급한 바와 같이 로짓모형은 하나의 방정식으로 정의되는 회귀모형과 달리 다수의 효용함수로 구성된다. 상관관계를 맺고 있는 설명변수들이 동일한 효용함수내에 포함되어 있는가 아니면 서로 다른 효용함수에 분산되어 있는가에 따라서 다중공선성의 영향은 다를 것으로 예상된다. 또한, 상관관계에 있는 두 변수들에 대해서 공통의 계수를 전제로 하는 경우(generic variable)와 그렇지 않은 경우 역시 다중공선성의 영향은 다르게 나타날 것으로 예상된다. 이상의 두 가지 사안은 단일방정식을 근간으로 하는 회귀분석에서는 고려할 필요가 없지만 여러 개의 효용함수를 채용하는 로짓모형에서는 고려되어야 할 사항이라 판단된다.

마지막으로 상관관계를 맺고 있는 두 변수 중에서 오직 한 변수만이 선택에 영향을 미치는 경우에 있어서의 다중공선성의 영향은 두 변수 모두 선택행동에 영향을 미치는 경우와는 다를 것으로 추정된다[4]. 회귀분석에서는 종속변수에 영향을 미치지 않는 변수라 하더라도 타 설명변수와 상관관계만 있다면 상관관계에 있는 변수들의 신뢰도 저하와 기여도의 과대평가 또는 과소평가를 야기한다[5]. 그러나 회귀모형처럼 연속변량의 종속변수가 존재하지 않는 로짓모형에서는 선택행동에 영향을 미치지 않는 변수가 효용함수의 설명변수로서 도입되었을 경우 타 변수들의 기여도와 신뢰도에 어떤 형태의 영향을 미칠 것인 지는 단정하기가 쉽지 않다. 즉, 이러한 변수들의 영향을 검토하는 것도 의미 있는 일이라 판단된다.

본 연구에서는 이상의 사안들을 고려하기 위해서 다음과 같이 3개의 기본 시나리오와 2개의 서브시나리오를 구축해서 다중공선성의 영향을 검증한다.

(1) 시나리오 1.

상관관계를 맺고 있는 두 설명변수가 동일한 효용함수에 포함되어 있는 경우를 검증한다. 단, 상관관계를 맺고 있는 두 변수 가운데 한 변수만이 선택행동에 영향을 미치는 경우와 두 변수 모두 선택행동에 영향을 미치는 경우를 구분하기 위해서 각각을 서브 시나리오로 구성한다. 즉,

- 시나리오 1-1 : 상관관계를 맺고 있는 두 설명변수가 동일한 선택대안의 효용함수에 포함되어 있으면서 오직 한 변수만이 선택행동에 영향을 미치는 경우
- 시나리오 1-2 : 상관관계를 맺고 있는 두 설명변수가 동일한 선택대안의 효용함수에 포함되어 있으면서 두 변수 모두 선택행동에 영향을 미치는 경우

(2) 시나리오 2.

상관관계를 맺고 있는 두 설명변수가 서로 다른 효용함수에 포함된 경우를 검증한다. 또한 마찬가지로 상관관계를 맺고 있는 두 설명변수 가운데 하나의 변수만이 선택행동에 영향을 미치는 경우와 두 설명변수 모두 선택행동에 영향을 미치는 경우를 구분하기 위해서 각각을 서브시나리오로 구성한다. 즉,

- 시나리오 2-1 : 상관관계를 맺고 있는 두 설명변수가 서로 다른 효용함수에 포함되어 있으면서 오직 하나의 설명변수만이 선택행동에 영향을 미치는 경우
- 시나리오 2-2 : 상관관계를 맺고 있는 두 설명변수가 서로 다른 효용함수에 포함되어 있으면서 두 설명변수 모두 선택행동에 영향을 미치는 경우

(3) 시나리오 3.

상관관계를 맺고 있는 두 설명변수가 서로 다른 효용함수에 포함되어 있고, 두 설명변수 모두 선택행동에 영

4) 설명변수가 선택행동에 영향을 미친다는 것은 추정된 계수가 통계적으로 유의하다는 의미와 같다. 역으로 설명변수가 선택행동에 영향을 미치지 않는다는 것은 추정된 계수가 통계적으로 유의하지 않다는 의미와 같다.

5) 회귀계수의 신뢰도 저하는 식(4) 또는 식(5)에 의해서 자명한데, 기여도의 과대 또는 과소평가는 별도의 설명이 필요하다. 상세한 내용은 이성우 외(2005) pp284-298을 참조하기 바란다.

향을 미침을 전제로 설명변수의 계수를 공유(generic variable)하는 경우를 검증한다.

- 시나리오 3 : 상관관계를 맺고 있는 두 설명변수가 계수를 공유하는 경우

2) 설명변수설정 및 효용함수의 구축

(1) 설명변수의 도입

전절에서 설정된 5개의 시나리오를 모두 검증하기 위해서 최소한 3개의 설명변수를 도입해야 하며, 각각의 변수들 간에는 식(7)과 같은 관계가 전제되어야 한다. 즉, 설명변수 x_2와 설명변수 x_3 사이에 상관관계(다중공선성)가 존재하며 나머지 변수들 사이에는 상관관계가 존재하지 않는다.

$$Cov(x_2,x_3)\neq 0,\ \ Cov(x_1,x_2)=Cov(x_1,x_3)=0 \qquad (7)$$

(2) 효용함수의 구축

3개의 변수를 활용해서 전절에서 설정된 시나리오별로 다중공선성의 영향을 검증하기 위해서는 다음과 같이 3가지 유형의 효용함수가 도입되어야 한다.

$$\begin{aligned} U_1 &= \alpha_0+\alpha_1 x_1+\epsilon_1 \\ U_2 &= \alpha_2 x_2+\alpha_3 x_3+\epsilon_2 \end{aligned} \qquad (8)$$

$$\begin{aligned} U_1 &= \alpha_0+\alpha_1 x_1+a_3 x_3+\epsilon_1 \\ U_2 &= \alpha_2 x_2+\epsilon_2 \end{aligned} \qquad (9)$$

$$\begin{aligned} U_1 &= \alpha_0+\alpha_1 x_1+a_2 x_3+\epsilon_1 \\ U_2 &= \alpha_2 x_2+\epsilon_2 \end{aligned} \qquad (10)$$

여기서 α_0은 대안속성 더미의 계수, α_1, α_2, α_3는 각각 효용함수를 구성하는 설명변수 x_1, x_2, x_3의 계수, ϵ_1, ϵ_2는 오차항이다. 변수간 관계에 대해서 $Cov(x_1,x_2)=Cov(x_1,x_3)=0$, $Cov(x_2,x_3)\neq 0$을 전제로 하였기 때문에 식(8)은 시나리오 1을, 식(9)는 시나리오 2를, 그리고 식(10)은 시나리오 3을 검정하는데 활용된다[6].

3. 변수의 생성 및 실험의 수행

1) 효용함수에 대한 가정

본장의 제1절에서 언급한 바와 같이 본 연구에서는 효용함수에 대한 모든 정보(효용함수의 구조, 설명변수의 값과 계수, 그리고 개개인의 관측불가능한 효용)를 알고 있음을 전제로 하고 있다. 이는 다시 말해서 모든 개인의 선택결과는 $P(U_1\geq U_2)$이면 대안 1이, 그 역이면 대안 2가 선택됨을 의미한다. 본 연구에서는 로짓모형의 계수에 대해서 다음의 값들을 사전적으로 부여해서 활용한다.

$$\begin{aligned} &\alpha_0 = 0 \\ &\alpha_1=\alpha_2=1.0 \\ &\alpha_3=\begin{cases} 1.0 & x_3\text{가 선택에 영향을 미치는 경우} \\ 0.0 & x_3\text{가 선택에 영향을 미치지 않을 경우} \end{cases} \end{aligned} \qquad (11)$$

2) 설명변수의 생성

수치실험에 이용될 설명변수의 값들은 기본적으로 난수발생기를 통해서 생성되었으며, 다음과 같은 분포를 갖도록 추가적인 조작이 가해졌다[7]. x_i를 정규분포하도록 조작한 것은 사회현상과 실험환경이 최대한 유사성을 갖도록 하기 위함이다[8].

$$x_1,x_2,x_3\sim N(0,1) \qquad (12)$$

난수발생기를 통해서 만들어진 세 개의 설명변수간에는 이론적으로는 상관관계가 존재해서는 안 되지만 현실적으로는 미약하나마 상관관계가 존재하게 된다. $Cov(x_1,x_2)=0$의 실험조건이 부가되어 있기 때문에 두 변수간 상관관계를 제거하는 추가적 과정이 필요하며, 나아가, x_2와 x_3간

6) 식(10)은 상관관계를 맺고 있는 두 설명변수(x_2, x_3)에 대해서 α_2로 계수를 공유(generic variable)시킨 경우이다.

7) 본 연구에서는 엑셀의 난수발생모듈인 RAND()를 활용해서 설명변수의 값들을 생성하였다. 단, 이 모듈이 제공하는 난수는 균일분포를 따르기 때문에 정규분포의 난수로 가공하기 위해서 추가적인 가공작업이 수행되었는데, 구체적으로는 (중심극한정리를 활용해서) 6개의 균일분포하는 난수(y_i)를 발생시켜 더한 다음 아래의 식을 활용해서 평균 0, 분산 1이 되도록 중심이동과 스케일 조정을 하였다.

$$\hat{x_i}=\frac{(x_i-\overline{x})}{s_x},\ x_i=\sum_{n=1}^{6} y_i,\ y\sim U(0,1)$$

8) 정규분포를 따르는 실제 관측된 설명변수를 (평균=0, 분산=1이 되도록) 표준화하면 평균=0, 분산=1의 난수와 동일한 형태를 갖게 된다.

상관계수(r_{x_2,x_3})는 실험의 구조화를 위해서 특정 값에 고정시켜야 하기 때문에 별도의 추가적인 작업을 필요로 한다. 본 연구에서는 식(13)과 trial- and-error방식[9]을 이용해서 두 설명변수의 상관계수가 특정 값을 갖도록 보정하였다. 참고적으로 각각의 변수들은 모두 1,000개씩이 생성되었다(본 실험에서는 관측치가 1,000개인 경우를 상정하고 있다).

$$x_3 = \theta \cdot x_2 + (1-\theta) \cdot x_k,\ x_2, x_k \sim N(0,1) \qquad (13)$$

〈표 1〉 설명변수의 생성예($r_{x_2,x_3} = 0.9$)

연번	x_1	x_2	x_3
1	-0.44516	1.714645	1.904817
2	0.154254	2.262345	1.310092
3	0.328168	-1.04991	-1.25795
4	-0.2731	-0.23662	-0.08727
5	0.763221	-0.36839	-0.00568
…	…	…	…
1000	1.264257	0.880586	0.261225
평균	0.0000	0.0000	0.0000
분산	1.0000	1.0000	1.0000

3) 로짓 데이타의 생성

로짓모형을 추정하기 위해서는 기본적으로 설명변수와 더불어 선택결과가 주어져야 한다. 본장 제3절의 '1) 효용함수에 대한 가정'에서는 선택결과의 작성방법에 대해서 개략적으로 기술하고 있다. 즉, $P(U_1 \geq U_2)$이면 대안 1을, 그 역이면 대안 2가 선택되도록 선택결과를 생성하는 것이다.

한편, 효용함수를 구성하는 두 요소, 즉 확정적 효용과 확률적 효용 중에서 확정적 효용은 효용함수식(8)~(10)과 식(11)의 계수값, 그리고 〈표 1〉의 설명변수의 값을 이용해서 산정이 가능하지만 확률적 효용은 아직까지는 제시되어 있지 않다[10]. 로짓모델은 ϵ_i에 대해서 최빈치(mode) $\eta(=0)$, 스케일파타메타 $\omega(=1)$인 감벨분포를 가정함으로써 도출된다. 따라서 수치실험을 보다 엄격하게 수행하기 위해서는 가정된 분포파라메타의 감벨분포를 따르는 난수들을 생성해서 각각의 데이터 행에 추가해 주고 선택결과를 결정해야 한다. 그러나 생성이 용이한 평균 0, 분산1의 정규분포를 따르는 난수를 생성해서 사용해도 결과물의 해석에는 큰 영향을 미치지 않을 것으로 판단됨에 따라 본 연구에서는 정규난수를 사용해서 효용함수내 확률적 효용값들을 생성하였다[11].

선택결과를 결정하기 위한 모든 요소들(효용함수의 형태와 관련 변수들-x_i, α_j, ϵ_k)이 모두 결정됨에 다음의 조건식을 이용해서 로짓 데이터를 작성한다.

$$\text{선택결과} = \begin{cases} 1 & \text{if} \dfrac{\exp(U_1)}{\exp(U_1)+\exp(U_3)} \geq 0.5 \\ 2 & otherwise \end{cases} \qquad (14)$$

〈표 2〉는 시나리오 1-1, $r_{x_2,x_3} = 0.9$인 경우의 각종 변수들과 선택결과를 나타내고 있다.

〈표 2〉 로짓 데이터의 작성예($r_{x_2,x_3} = 0.9$)

연번	선택결과	x_1	x_2	x_3	ϵ_1	ϵ_2
1	2	-0.4451	1.71464	1.90481	0.09892	0.82970
2	2	0.15425	2.26234	1.31009	0.71747	-1.6675
3	1	0.32816	-1.0499	-1.2579	-1.4334	-0.7183
4	2	-0.2731	-0.2366	-0.0872	0.21899	0.28863
5	1	0.76322	-0.3683	-0.0056	0.18466	0.74825
6	1	-1.7337	-0.5161	-0.8720	0.26609	-0.9353
…	…	…	…	…	…	…
1000	1	1.26425	0.88058	0.26122	1.66837	-1.2200
평균		0.0000	0.0000	0.0000	0.0000	0.0000
표준편차		1.0000	1.0000	1.0000	1.0000	1.0000

한편, 모델 추정에 사용된 각종 변수들간 상관관계는 〈표 3〉과 같다. 표에서 보는 바와 같이 x_1과 x_2, x_3사이의 상관계수는 0에 근사되어 있음을 알 수 있다. 한편, x_2와 x_3 사이에는 일정 수준의 상관관계가 존재함을 가정하고 있는데 두 변수간 상관계수(r_{x_2,x_3})가 0.0에서 0.9까지 0.1씩 증가하도록 제어되고 있다. 오차항과 x_i간에는 미세한 상관관계가 있지만 이러한 수준의 상관관계가 분석결과의 해석을 왜곡시킬 가능성은 매우 낮다고 판단된다.

9) 두 설명변수간 상관계수(r_{x_2,x_3})가 특정 값을 갖도록 하기 위해서는 θ값을 반복적으로 조정하는 과정이 필요하다.

10) 확정적 효용만으로 선택확률을 계산하고 산정된 확률값을 기준으로 선택된 대안을 결정해서 로짓데이터를 작성하면 로짓모형은 추정되지 않는다.

11) 土木學會(1995)는 효용함수의 오차항이 정규분포하고 있을 것으로 추론하고, 다만 분포의 유사성 그리고 모형의 추정 및 활용의 용이성을 이유로 감벨분포를 가정, 로짓모형을 유도하고 있다. 본 논문에서도 정규분포를 따르도록 생성된 오차항에 대해서 감벨분포를 가정하고 있기 때문에 로짓모델의 이론 체계에 위배되지 않는다. 또한 모든 시나리오에 대해서 오차항 만큼은 동일한 값들이 활용되고 있기 때문에 오차항으로 인해서 모델의 적합도, 계수의 값, 그리고 계수의 신뢰도가 변동하는 경우는 발생하지 않는다.

〈표 3〉 모델에 활용된 데이터간 상관계수

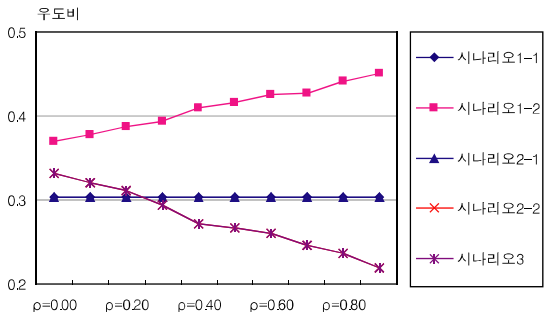
		x_1	x_2
x_1		1.0000	
x_2		-0.0001	1.0000
x_3	$r_{x_2x_3}$=0.000	0.0000	0.0000
	$r_{x_2x_3}$=0.100	0.0000	0.1002
	$r_{x_2x_3}$=0.200	0.0000	0.2000
	$r_{x_2x_3}$=0.300	0.0000	0.3003
	$r_{x_2x_3}$=0.400	0.0000	0.4001
	$r_{x_2x_3}$=0.500	0.0000	0.5001
	$r_{x_2x_3}$=0.600	-0.0001	0.6001
	$r_{x_2x_3}$=0.700	-0.0001	0.7000
	$r_{x_2x_3}$=0.800	-0.0001	0.8002
	$r_{x_2x_3}$=0.900	-0.0001	0.9002
ϵ_1		-0.0231	-0.0165
ϵ_2		0.0015	0.0011

Ⅲ. 로짓모형에 대한 다중공선성의 영향

제2장의 실험을 통해서 총 50개(시나리오 5 × x_2와 x_3의 상관계수($r_{x_2x_3}$)의 수준 10)의 로짓모형이 구축되었다. 본 장에서는 구축된 모델로부터 로짓모형의 적합도, 추정된 계수 그리고 계수의 신뢰도에 대한 다중공선성의 영향을 살펴본다.

1. 모델의 적합도에 대한 영향

로짓모델의 적합도는 로그우도비 또는 McFadden의 결정계수라 불리는 ρ^2에 의해서 측정된다. 〈그림 1〉과 〈그림 2〉는 각각의 시나리오별, 상관계수($r_{x_2x_3}$)별 우도비를 나타낸 것이다. 먼저, x_3가 효용함수에는 포함되어 있지만 선택에는 영향을 미치지 않는 시나리오 1-1과 2-1의 우도비는 x_2와 x_3간 상관계수($r_{x_2x_3}$)의 값이나 x_2에 대한 x_3의 상대적 위치에 관계없이 불변인 것으로 나타났다. 이는 회귀분석에서 나타나는 현상과 유사한데, 가령 회귀분석에서도 종속변수의 변동에 영향을 미치는 설명변수(본 연구에서는 x_2)와 통계적으로는 유의하지 않지만 종속변수와 단지 상관관계만을 갖는 설명변수(본 연구에서는 x_3)를 회귀모형에 포함시키더라도 회귀모델의 적합도(R^2)는 크게 변화하지 않기 때문이다(Appendix의 회귀모델-시나리오 1 참조).



〈그림 1〉 시나리오별 우도비

한편, 보편적인 경우 즉, 모든 변수가 선택에 영향을 미치는 경우(시나리오 1-2와 시나리오 2-2)에는 상관관계를 맺고 있는 변수의 상대적 위치에 따라서 적합도에 대한 다중공선성의 영향이 다르게 나타나는 것을 알 수 있다. 상관관계를 맺고 있는 두 설명변수가 동일한 효용함수에 포함되어 있는 시나리오 1-2의 경우에는 두 설명변수간 상관계수가 높아짐에 따라 모델의 적합도가 향상된 반면 서로 다른 효용함수로 분산되어 있는 경우에는 적합도가 감소하는 경향을 보였다. 시나리오 1-2와 2-2는 교통분야의 선택모형에서 보편적으로 발견할 수 있는 상황을 대변하고 있다. 가령 장거리 통행의 수단선택모형에 있어서 통행요금과 통행시간은 높은 상관관계를 맺고 있으며 동일한 효용함수의 설명변수로 도입된다(시나리오 1의 경우). 또한, 수단선택모형에 있어서 승용차의 통행시간과 대중교통수단의 통행시간은 높은 상관관계를 맺고 있으며 서로 다른 효용함수의 설명변수로 도입된다(시나리오 2의 경우)

한편, 계수의 공유를 전제로 한 시나리오 3은 기본적으로 두 변수(x_2, x_3)의 상대적 위치(서로 다른 효용함수의 설명변수로 도입된 점)와 선택행동에 대한 x_3의 영향 측면에서 시나리오 2-2와 매우 유사하며, 상관계수의 변화에 따른 우도비의 변화 패턴 역시 시나리오 2-2와 유사함을 알 수 있다. 결론적으로 상관관계가 높은 설명변수가 서로 다른 효용함수의 설명변수로 도입되면 독립적으로 계수를 취하든 공통적으로 계수를 취하던(generic variable) 모델의 적합도에 대한 다중공선성의 영향은 동일하게 나타나는 것을 알 수 있다.

2. 추정계수에 대한 영향

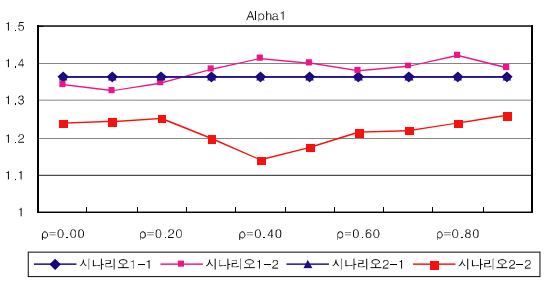
로짓모델에 있어서 설명변수의 계수는 선택행동에 대

한 설명변수의 영향력 또는 기여도로서의 의미를 가지며, 설명변수가 정부정책을 통해서 제어가능한 변수인 경우에는 정책의 효과를 사전에 예측하는데 있어서 필요한 기초적인 정보를 제공해 준다. 따라서 추정된 계수가 왜곡되어 있다면 정책효과에 대한 과대평가 또는 과소평가로 이어져 잘못된 의사결정에 이르게 한다.

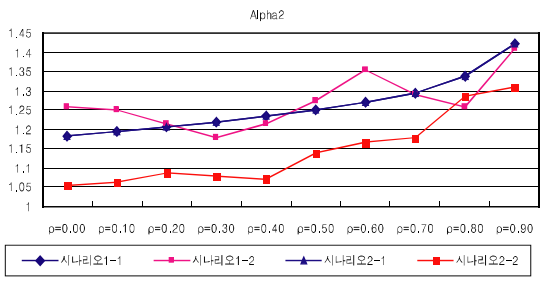
시나리오별 추정된 계수값은 〈그림 2〉~〈그림 4〉와 같다[12]. 설명변수 x_3가 선택행동에 영향을 미치지 않음을 전제로 한 시나리오 1-1과 2-1에서의 추정된 계수값들은 유사한 변동패턴을 보이고 있음을 알 수 있다. x_2 및 x_3와 독립인 설명변수 x_1의 계수(α_1)는 시나리오나 두 변수(x_2, x_3)간 상관의 정도에 관계없이 비교적 안정된 값을 유지하고 있는 반면 x_3와 상관관계를 맺고 있는 설명변수 x_2의 계수(α_2)는 x_3와의 상관관계가 높아짐에 따라 값이 커지는 경향을 보이고 있다. 이러한 현상은 시나리오 2-2(x_3가 선택행동에 영향을 미치고 동일한 효용함수에 포함된 경우)에서 보다 급격하게 발생하고 시나리오 1-2(x_3가 선택행동에 영향을 미치고 서로 다른 효용함수에 포함된 경우)에서 가장 완만하게 나타나고 있다. 특이한 점은 x_3가 선택행동에 영향을 미치지 않음에도 불구하고(시나리오 1-1과 2-1의 경우) 단순히 효용함수에 포함되었다는 점만으로도 상관관계를 맺고 있는 설명변수 x_2의 계수(α_2)가 증가한다는 점이다.

한편, x_3의 계수(α_3)값은 시나리오에 관계없이 비교적 로짓 데이터 구축과정에서 설정한 기대치와 유사한 값을 취하고 있다.

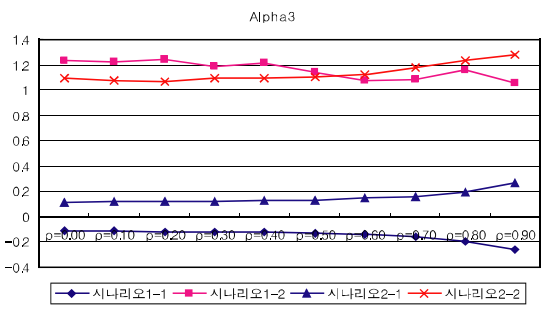
이상의 내용을 정리하면, 첫째, 선택행동에는 영향을 미치지는 않지만(통계적으로는 유의하지 않지만) 로짓모형내 다른 변수와 상관관계가 높은 변수가 로짓모형의 설명변수로 도입되면 높은 상관관계에 있는 기존변수의 선택행동에 대한 기여도는 과대평가될 가능성이 높다. 둘째, 통계적으로 유의하면서 높은 상관관계에 있는 두 설명변수가 동일한 효용함수에 포함되는 경우(시나리오 1-2)보다 서로 다른 효용함수의 설명변수로 포함되는 경우(시나리오 2-2)가 보다 심각한 문제를 야기한다. 셋째, 그럼에도 불구하고 상관관계에 있는 두 변수 중에서 과다하게 평가된 변수를 규명하는 것은 현실에서는 발견하기 어렵다. 실질적으로 시나리오 2-2에 있어서 x_2와



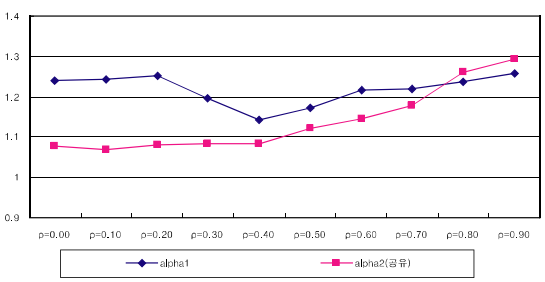
〈그림 2〉 시나리오 1과 2에서 α_1의 추정치



〈그림 3〉 시나리오 1과 2에서 α_2의 추정치



〈그림 4〉 시나리오 1과 2에서 α_3의 추정치



〈그림 5〉 시나리오 3에서 계수추정치

12) 로짓모델의 추정을 위한 데이터의 생성과정에서 α_0(더미변수의 계수)는 0의 값을 부여하였기 때문에(식 11 참조) 계수의 값과 신뢰도의 기대치는 모두 0이고, 추정결과 또한 기대치와 같아 해석에서 제외한다.

x_3는 모두 선택행동에 영향을 미치고 서로 다른 효용함수의 설명변수로 도입되어 있다. 그러나 α_3는 비교적 안정되어 있는 반면 α_2는 상관관계가 높아짐에 따라 계수값이 증가하고 있어 다중공선성의 영향이 상관관계를 맺고 있는 두 변수 중에서 한쪽 변수에만 작용하고 있고, 실제 로짓모형을 이용하는 과정에서 어떤 변수가 과다추정되었는지는 확인할 수 없다는 문제가 있다.

한편, 계수의 공유를 전제로 한 시나리오 3에 대한 분석결과는 〈그림 5〉와 같다(시나리오 3에서는 설명변수 x_3가 선택행동에 영향을 미친다는 점에 유의할 것).

통계적으로 유의하면서 서로 다른 효용함수에 포함되어 있는 상관관계에 있는 두 변수(x_2, x_3)에 대해 공통의 계수(generic variable)를 취하도록 했을 경우 두 설명변수간 상관계수가 커짐에 따라 공유된 계수(α_2)의 값도 증대되는 것을 알 수 있다.

로짓 데이터의 생성과정에서 가정된 계수들의 값들이 1.0임을 감안하면 추정된 계수들은 다소 설명변수를 과대하게 평가하는 경향이 있다고 판단되며, 결론적으로 상관관계가 높은 변수들이 효용함수에 도입되면 선택행동에 대한 설명변수의 기여도는 과대평가될 가능성이 높다고 할 수 있다.

3. 추정계수의 신뢰도에 대한 영향

회귀분석에 있어서 다중공선성의 가장 큰 문제는 추정된 회귀계수의 과대 또는 과소평가문제와 더불어 추정된 계수의 신뢰도를 급격하게 저하시킨다는 점이다(Appendix의 내용 참조). 로짓모형에 있어서도 추정된 계수의 분산공분산행렬이 비록 지수함수의 형태이기는 하지만 변수간 상관관계의 영향을 받게 되어 있어 추정된 계수의 신뢰도를 저하시킬 것으로 예상된다.

〈그림 6〉~〈그림 8〉은 시나리오별 추정된 계수의 신뢰도(t-통계량)를 플로팅한 것이다. 〈그림 6〉은 상관관계에 있는 두 변수(x_2, x_3)와는 독립인 설명변수 x_1의 계수(α_1)의 시나리오별-상관계수($r_{x_2x_3}$)의 수준별 t통계량을 나타내고 있는데, 설명변수 x_3이 선택행동에 영향을 미치지 않는 시나리오 1-1과 2-1에서는 불변인 반면 설명변수 x_3이 선택행동에 영향을 미치는 시나리오 1-2와 2-2에서는 신뢰도가 다소 변동하고 있다. 설명변수 x_3와 x_1이 동일한 효용함수에 포함되어 있는 시나리오 2-2의 경우에는 소폭이기는 하지만 신뢰도가 상승하는 반면 서로 다른 효용함수에 분산되어 있는 시나리오 1-2에서는 소폭 하강하는 경향을 보이고 있다. 단, 변동의 진폭이 크지 않고 상관계수의 크기에 대해서 일관성도 보이고 있지 않아 유의한 의미를 갖고 있다고 판단하기는 어렵다.

본 연구의 주요 관심대상인 설명변수 x_2의 계수인 α_2의 t통계량은 시나리오에 관계없이 설명변수 x_3와의 상관관계가 높아짐에 따라 급격하게 저하되는 것으로 나타났다. 다만 다중공선성이 계수의 신뢰도에 직접적으로 영향을 미치는 회귀분석보다는 지수함수를 취함으로써 상관관계가 다소 완화되는 로짓모형이 다중공선성의 영향을 작게 받는 것으로 판단된다[13].

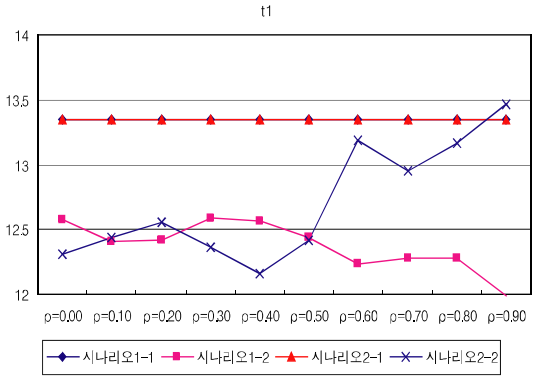
마지막으로 설명변수 x_3의 계수인 α_3의 신뢰도는 당해 변수가 선택행동에 영향을 미치지 않는 경우(시나리오 1-1과 2-1)에는 x_2와의 상관관계의 정도에 관계없이 신뢰성 자체가 없는 것으로 나타났다(이러한 결과는 회귀분석에서도 찾아볼 수 있다. Appendix의 시나리오 1 참조).

한편, 계수를 공유하는 경우(generic variabel)에 있어서의 추정된 계수의 신뢰도는 보다 세심한 해석이 필요한 것으로 판단된다. 먼저, 상관관계에 있는 두 변수(x_2, x_3)를 대상으로 공통의 계수를 적용한 시나리오 3에서 공통 계수 α_2의 신뢰도는 두 변수간 상관관계가 높아짐에 따라 급격히 저하되는 것으로 나타난 반면 두 변수(x_2, x_3)와 독립이고 계수 역시 독립적으로 취하도록 한 설명변수 x_1의 계수(α_1)의 신뢰도는 미세하나마 향상되는 경향을 보였다.

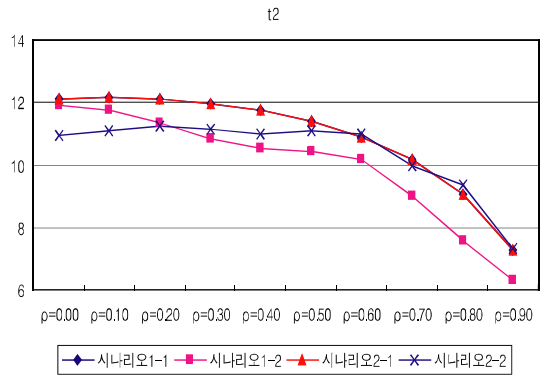
상관관계를 맺고 있는 두 변수(x_2, x_3)와 독립인 변수 x_1의 계수의 신뢰도가 상관관계를 맺고 있는 두 변수간 상관관계가 높아짐에 따라서 증대되는 이유에 대해서는 다음과 같은 추론이 가능할 것으로 판단된다. 먼저, 상관관계를 맺고 있는 두 변수(x_2, x_3)에 대해서 공통의 계수값(α_2)을 갖도록 모델이 구축되어 있기 때문에 모델에서는 실질적으로는 두 변수(x_2, x_3)의 차(-)가 하나의 변수로 작용하게 된다[14]. 실험조건에서 설명변수 x_1은 상관관계를 맺고 있는 두 변수(x_2, x_3)와 상관관계가 없

13) Appendix의 시나리오 1, 2의 경우 두 변수간 상관계수가 0.9일 때의 α_2의 t통계량은 상관관계가 없을 때의 t통계량의 40% 수준이지만 로짓모형에 대한 시나리오 1, 2에서 상관계수가 0.9일 때의 α_2의 t통계량은 상관관계가 없을 때의 t통계량의 60% 수준이다. 이로써 로짓모형이 회귀모형에 비해서 추정된 계수의 신뢰도 측면에서 다중공선성의 영향을 작게 받는다고 판단할 수 있다.

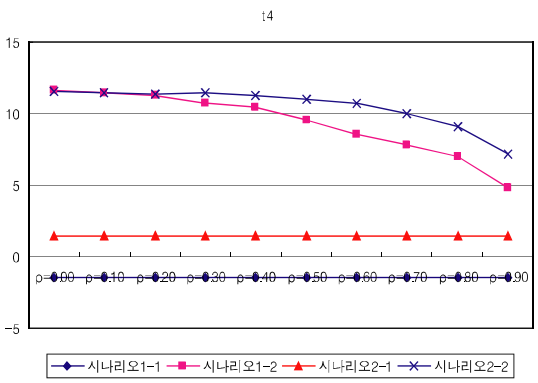
14) 다항로짓모델에서 대안 i의 선택확률은



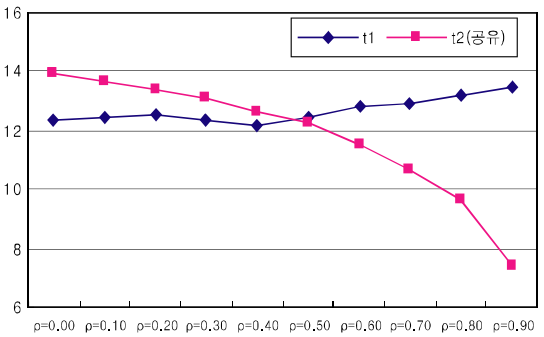
〈그림 6〉 시나리오 1과 2에서 α_1의 t통계량



〈그림 7〉 시나리오 1과 2에서 α_2의 t통계량



〈그림 8〉 시나리오 1과 2에서 α_3의 t통계량



〈그림 9〉 시나리오 3의 계수의 t통계량

음(식(7) 참조)을 전제로 하였지만 현실적으로 수치실험의 수행과정에서 상관관계가 완전히 제거된 변수를 생성하는 것은 불가능하다[15]. 따라서 설명변수 x_1과 두 변수(x_2, x_3)의 차(-)라고 하는 새로운 변수간에도 미약한 상관관계라 존재할 것으로 추정되며, 두 변수(x_2, x_3)간 상관관계가 높아짐에 따라서 변수 x_1과 두 변수간 차(-)와의 상관관계가 낮아지면서 결과적으로 x_1의 계수의 신뢰도가 증대되었을 것으로 추정된다[16]

4. 실험결과의 종합

실험결과를 종합하면 〈표 4〉와 같다. 로짓모형에 있어서 다중공선성은 회귀분석에서와 같이 단적으로 나쁘다고만은 규정할 수 없는 결과가 얻어졌다. 가령, 상관관계에 있는 두 변수가 동일한 효용함수의 설명변수로 도입될 경우 모델의 적합도는 개선될 수 있음을 분석결과

〈표 4〉 변수간 상관관계의 증가에 따른 영향

	모델의 적합도	상관관계에 있는 변수의 계수		독립인 변수의 계수	
		계수의 값	신뢰도	계수의 값	신뢰도
시나리오 1	개선됨	증가와 불변이 공존	저하	불변	미세한 진동
시나리오 2	저하됨	증가와 불변이 공존	저하	불변	미세한 진동
시나리오 3	저하됨	증가	저하	불변	미세한 증가

$P_i = 1/(1+\sum_{j \neq i \in A} \exp(V_j - V_i))$이므로 공통의 계수를 취하는 변수는 두 변수의 차가 하나의 변수로 작동한다.

15) 〈표 3〉은 본 연구의 수치실험을 위해서 생성된 데이터간 상관관계를 나타내고 있는데 x_1과 x_2, x_1과 x_3의 사이에는 미약하게나마 상관관계가 존재함을 확인할 수 있다.

16) 수치실험에 이용된 데이터를 이상의 추론을 검증한 결과, $r_{x_2,x_3)}$가 0.0일 때의 r_{x_1,x_2-x_3}의 상관계수는 $r_{x_2,x_3)}$가 0.9일 때의 약 4배 수준임이 확인되었다.

는 보여주고 있다.

한편, 상관관계에 있는 설명변수의 계수는 어느 변수의 계수인지는 특정할 수 없지만 과대추정되고 있음을 보여주고 있다. 특히, 상관관계에 있는 변수를 대상으로 공통의 계수를 취하도록 하는 경우(generic variable)에 있어서는 추정된 계수는 과대추정되어 있다고 볼 수 있다.

Ⅳ. 결론

본 연구에서는 로짓모형에 대한 다중공선성의 영향을 구조화된 수치실험을 통해서 실증적으로 규명하였다. 본 연구를 통해서 얻어진 시사점들을 정리하면 다음과 같이 요약될 수 있다.

첫째, 회귀분석에서는 설명변수의 추가를 통해서 모델의 전반적인 적합도(R^2)를 제고할 수 있지만 로짓모형에서는 설명변수의 추가를 통해서 모델의 적합도가 개선될 수도, 역으로 저하될 수도 있음이 규명되었다.

둘째, 유사변수에 대해서 계수를 공유(generic varialbe)하도록 모델을 구성하면 두 변수간 상관관계가 높아짐에 따라 모델의 적합도가 저하되는 경향이 있음을 확인하였다.

셋째, 다중공선성이 설명변수의 계수값에 미치는 영향은 크지는 않지만 높은 상관관계를 맺고 있는 변수들 중에는 선택행동에 대한 기여도가 과대평가될 가능성이 있음을 확인하였다.

넷째, 다중공선성은 상관관계를 맺고 있는 변수들(또는 계수들)의 신뢰도를 저하시키는 역할을 하며 그 정도는 회귀분석의 경우에 비해서는 작지만 절대적 측면에서는 결코 무시할 수 없는 수준임을 확인하였다. 또한, 상관관계가 높은 변수들을 대상으로 공통의 계수를 적용하는 경우에 공유된 계수의 신뢰도는 상관관계가 높아짐에 따라서 저하됨을 확인하였다.

로짓모형은 교통수요추정과정에서 보편적으로 이용되고 있는 모형 가운데 하나로 특히 4단계 교통수요추정모형에 있어서 가장 대표적인 수단선택모형으로 활용되고 있다. 지금까지 발표된 로짓모형에 관한 연구성과물이나 실제 교통수요추정모형으로 활용되고 있는 모형들의 절대 다수는 상당히 높은 상관관계를 맺고 있는 변수들을 설명변수로 활용하고 있다. 가령, 통행시간변수와 통행비용변수를 이용한 수단선택모형은 상관관계가 높은 변수들이 설명변수로 도입된 대표적인 사례라 할 수 있다. 수단선택모형을 존쌍별로 추정하지 않고 모든 존쌍을 대상으로 공통 모형으로 추정하기 때문에 당연히 통행시간과 통행비용 사이에는 비교적 높은 상관관계가 내재되어 있으며, 결과적으로 추정된 모형에는 본 연구에서 검증된 다중공선성의 영향이 작용하고 있을 것으로 예상된다.

회귀분석에 대해서는 다중공선성에 대한 다양한 대처방법들이 제안되어 있지만 로짓모형과 관련해서는 다중공선성의 영향 자체를 심각하게 생각하고 있지 않으며, 따라서 다중공선성의 대처방안에 대한 연구도 전혀 이루어지고 있지 않다. 본 논문의 학술적 의의는 로짓모형에 대해서 다중공선성의 영향을 검토한 최초의 시도라고 하는 점에 있다고 판단되며, 다만, 실제 관측된 데이터가 아닌 인위적으로 작성된 데이터를 통해서 로짓모형의 다중공선성이 검토되었다고 하는 실증적 차원에서의 한계점은 추가적 연구를 통해서 보완되어야 할 것으로 판단된다. 다시 말해서 본 연구결과에 근거해서 장래교통수요예측에 활용되고 있는 기 구축된 모형에 문제점이 많을 것으로 단정해서는 아니되며, 변수간 다중공선성의 존재여부 확인과 다중공선성을 제거한 상태에서의 모형 개량가능성을 검증할 필요가 있다는 것이다.

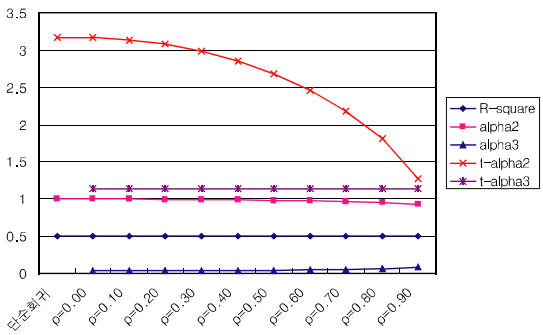
또한, 본 연구의 수치실험에서 활용된 데이터에 대한 조건과 실제 관측을 통해서 수집된 테이터의 조건이 일치하지 않는다면 로짓모형에 대한 다중공선성의 영향이 다르게 나타날 가능성을 배제할 수 없다. 대표적인 예로서 본 연구에서는 설명변수가 정규분포를 따른다는 가정하에 다중공선성의 영향이 검토되었는데, 실제 관측된 데이터가 정규분포하지 않는다면 실제 데이터를 토대로 구축된 로짓모델에 있어서의 다중공선성의 영향이 다르게 나타날 가능성도 있다는 것이다. 이러한 부분들이 구조화된 수치실험을 통해서 로짓모델에 대한 다중공선성의 영향을 검토한 본 연구의 한계점이라 할 수 있다.

따라서 향후 연구과제로서 첫째, 실제 관측된 데이터를 이용해서 로짓모형의 다중공선성을 검증하는 작업이 필요한 것으로 판단된다. 또한, 다른 형태의 선택모형, 가령 프로빗모형이나 네스티드 로짓모형에 있어서의 다중공선성의 영향을 규명하는 것도 필요한 작업이라 판단된다. 그리고 로짓모형에 있어서 다중공선성 문제를 완화하기 위한 방안에 대한 연구도 반드시 수행되어야 할 것이다. 가령, 특정 교통수단의 통행시간이라는 변수는 다른 교통수단의 통행시간이나 교통수단별 통행비용과 상관계수가 높은 변수 가운데 하나인데 이러한 변수를 존간 거리로 나누어서 속도변수로 변환시킨다거나 하는 방안의 효용성도 검증할만한 연구분야라 판단된다.

Appendix : 회귀모형에서의 다중공선성

1. 시나리오 1

본고에서 설정한 시나리오 X-1과 마찬가지로 종속변수 y의 변동은 설명변수 x_2에 의해서만 결정되며 설명변수 x_3는 x_2와 상관관계만 있을 뿐 종속변수의 변동에는 영향을 미치지는 않는 경우이다. 두 변수(x_2, x_3)의 상관관계에 따른 적합도, 회귀계수, 계수의 신뢰도는 아래의 그래프와 같다. 설명변수 x_2의 회귀계수인 α_2의 t통계량를 제외한 모든 통계량(R^2, α_2, α_3, α_3의 t통계량)은 두 변수간 상관계수($r_{x_2x_3}$)에 대해서 거의 변화가 없는 반면 α_2의 t통계량은 급격하게 저하되는 것을 알 수 있다.

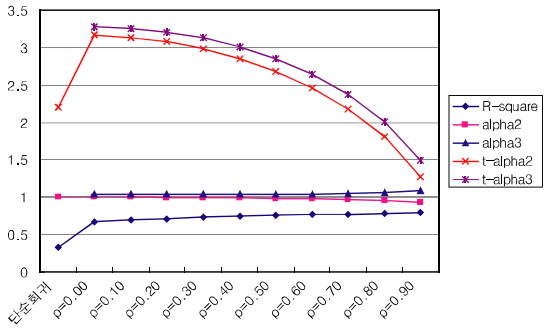


주1 : α_2의 t통계량은 실제 값의 1/10로 환산된 값임
주2 : 단순회귀는 x_3를 설명변수로 포함하지 않음
주3 : y는 $y_i = x_{2i} + x_{3i} + \epsilon_i$를 이용해서 생성함

〈그림 10〉 시나리오 1의 회귀분석결과

2. 시나리오 2

본고에서 설정한 시나리오 X-2와 마찬가지로 종속변수 y의 변동은 설명변수 x_2와 x_3에 의해서 영향을 받는 경우이다. 두 변수(x_2, x_3)의 상관관계에 따른 적합도, 회귀계수, 계수의 신뢰도는 아래의 그래프와 같다. 시나리오 1과는 다른 형태의 결과물이 얻어졌는데, 먼저, 모델의 적합도(R^2)가 상관관계가 높아지면서 증가하는 것을 알 수 있다. 또한 단순회귀모형에서는 설명하지 못했던 종속변수의 변동을 x_3가 설명변수로서 도입됨에 따라 설명이 가능해져 단순회귀에 비해 x_2의 회귀계수의 t통계량이 급격히 개선되었다는 점도 주목할 필요가 있다(단순회귀와 $\rho = 0.0$의 t-alpha2를 주목). 그러나 두 변수(x_2, x_3)의 상관관계가 높아짐에 따라 두 변수의 신뢰도가 모두 급격하게 저하되는 점에 유의할 필요가 있다. 특히, 종속변수와 오차항의 단순합에 의해서 생성되었다는 점(〈그림 11〉 주3 참조)을 감안하면 단순히 상관계수만 높아졌다는 사실 만으로 회귀계수의 신뢰도가 저하되는 현상은 논리적인 관점에서는 납득하기 쉽지 않다.



주1 : α_1과 α_2의 t통계량은 실제 값의 1/10임
주2 : 단순회귀는 x_3를 설명변수로 포함하지 않음
주3 : y는 $y_i = x_{2i} + x_{3i} + \epsilon_i$를 이용해서 생성함

〈그림 11〉 시나리오 2의 회귀분석결과

참고문헌

1. 김강수(2002), "SP 화물수단선택을 위한 Inherent Random Heterogeneity 로짓 모형 연구", 대한교통학회지, 제20권 제3호, 대한교통학회, pp.83~92.
2. 김강수 · 조혜진(2004), "SP 순위 자료별 오차를 고려하는 순위로짓 모형 추정에 관한 연구", 대한교통학회지, 제22권 제6호, 대한교통학회, pp.197~206.
3. 박규영 · 이수범(2006), "보행자사고확률모형을 이용한 도로안전시설물의 효과도 추정 (4차로 일반국도를 대상으로)", 대한교통학회지, 제24권 제4호, 대한교통학회, pp.55~65.
4. 박상준 · 김성수(2007), "승용차 보유대수와 차종선택에 대한 네스티드로짓모형의 추정", 대한교통학회지, 제25권 제1호, 대한교통학회, pp.133~141.
5. 서상언 · 정진혁 · 김순관(2006), "활동 스케줄 분석을 통한 고령자의 통행특성과 통행행태에 관한 연구", 대한교통학회지, 제24권5호, 대한교통학회, pp.89~108.
6. 유지성, 오창수(2004), "현대통계학", 박영사.
7. 이성우, 민성희, 박지영, 윤성도(2005), "로짓 · 프

로빗모형 응용", 박영사.
8. 土木學會(1995), "非集計モデルの理論と實際,丸善(株).
9. Ben-Akiva, M., Steven R. Lerman(1987), Discrete Schoice Analysis : "Theory and application to travel demand", MIT Press.
10. Ben-Akiva, M., Watanatada(1981), "Application of a Continuous Choice Logit Model. In Structural Analysis of Discrete Data with Econometric Applications", MIT Press.
11. Cardell, N.S., F.C. Dunbar(1980), " Measuring the Societal Impacts of Automobile Sownsizing", Transportation Research A 14,
12. Frank S. Koppelman, Vaneet Sethi(2005), "Incorporating variance and covariance heterogeneity in the Generalized Nested Logit model: an application to modeling long distance travel choice behavior", Transportation Research Part B 39.
13. Juan de Dios Ortúzar, David Hensher, Sergio Jara-Diaz(1998), "Travel Behaviour Research : Updating the State of Play", Pergamon.
14. Karthik K. Srinivasan, Sudhakar R. Athuru (2005), "Analysis of within- household effects and between-household differences in maintenance activity allocation", Transportation, Volume 32, Number 5.
15. Karthik K. Srinivasan, Hani S. Mahmassani (2003), "Analyzing heterogeneity and unobserved structural effects in route-switching behavior under ATIS: a dynamic kernel logit formulation", Transportation Research Part B 37.
16. Norbert Oppenheim(1994), "Urban Travel Demand Modeling from individual choice to general Equilibrium", John Willy & Sons, Inc.
17. Simon P. Washington, Matthew G. Karlaftis, Fred L. Mannering(2003), Statistical and Econometric Methods for Transportation Data Analysis, Chapman & Hall/CRC.
18. Tommy Gärling, Tomas Laitila, Kerstin Westin (1998), "Theoretical Foundations of Travel Choice Modeling", Pergamon.
19. William H. Greene, David A. Hensher, John Rose(2006), "Accounting for heterogeneity in the variance of unobserved effects in mixed logit models", Transportation Research Part B 40.

♧ 주 작 성 자 : 류시균
♧ 교 신 저 자 : 류시균
♧ 논문투고일 : 2007. 5. 19
♧ 논문심사일 : 2007. 7. 30 (1차)
2007. 9. 5 (2차)
2007. 11. 29 (3차)
♧ 심사판정일 : 2007. 11. 29
♧ 반론접수기한 : 2008. 6. 30