로짓모형에 있어서 다중공선성의 영향에 관한 연구

Effects of Multicollinearity on Logit Model

류 시 균

(경기개발연구원 교통정책연구부 연구위원)

목 차

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 회귀분석과정에서 세심한 검토와 대응이 이루어진다. 그러나 함수의 구조나 계수의 추정방법 측면에서 다중회귀모형과 유사 구조를 갖고 있는 로짓모형[1]과 관련해서는 다중공선성이 크게 주목받고 있지 못하고 있다[2].

로짓모형과 관련해서 다중공선성이 크게 주목받지 못하고 있는 데에는 나름대로 몇 가지 이유가 있을 것으로 판단된다. 필자는 두 가지의 요인이 핵심적으로 작용하고 있다고 추정하고 있는데, 첫째는 다중회귀모형의 발달과정에서 다중공선성의 영향과 대응방안이 충분히 검토되었기 때문에 로짓모형과 관련해서 새삼스럽게 다중공선성을 논할 필요성을 느끼는 연구자가 많지 않다는 점이고, 둘째로는 다중공선성 문제를 완화시키기 위한 방안 가운데 하나로 설명변수의 대수변환방식이 제안되고 있고, 지수함수(Exponential Function)를 취함으로써 변수변환과정을 거치는 로짓모형에서는 다중공선성 문제가 어느 정도 완화되었을 것으로 생각하기 때문으로 추정된다. 실제로 다항로짓모형의 분산공분산행렬, $E(-\nabla V)^{-1}$에서 $-\nabla V$의 k행 j열 원소는 식(1)과 같은데(토목학회, 1995), 이는 추정해야 할 계수의 분산(또는 표준편차)이 효용함수내 설명변수들의 지수함수로써 정의된다. 즉, 회귀모형에 있어서의 다중공선성을 완화하기 위한 과정, 즉 설명변수의 대수변환이 부분적으로 이루어졌기 때문에 다중공선성으로 인한 추정계수의 신뢰도 저하는 다중회귀분석

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

2) 다중회귀모형과 로짓모형내 효용함수는 공통적으로 선형방정식의 형태를 취하며 다중회귀모형의 보편적 추정방법인 최소자승법(Ordinary Least Square, OSL)과 로짓모형의 보편적 추정방법인 최우추정법(Maximum Likelihood method, MLM) 모두 Newton-Rapson법과 같은 Gradient방식 계열의 최적화모델을 활용한다.

에서 만큼은 발생하지 않을 것이라는 추론이 가능하다.

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

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 효용함수내 k번째 설명변수 값이다.

그러나 회귀분석에서의 대수변환방식은 상관관계가 높은 두 변수 가운데 하나의 변수에 대해서만 적용하기 때문에 설명변수들의 선형방정식으로 표현되는 효용함수 전체를 대상으로 지수함수를 취하는 로짓모형에 있어서는 다중회귀모형에서와 같은 정도로 다중공선성 문제가 완화될 것으로 예상되지 않는다. 다시 말해서 다중공선성에 대한 세심한 주의를 기울이지 않으면 다중회귀모형에서 나타나는 다중공선성 문제가 로짓모형에서도 나타날 수 있을 것으로 예상된다.

본 연구는 구조화된 실험을 통해서 효용함수내 설명변수들간 다중공선성이 로짓모형과 추정된 계수들의 신뢰도에 어떠한 영향을 미치는지를 실증적으로 규명하는데 목적이 있다.

2. 선행연구 고찰

1) 로짓모형의 다중공선성에 관한 연구

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. 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.

2) 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.

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

역행렬 공식을 이용해서 정리하면, 두 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 회귀계

수의 신뢰도지표 즉, t값(=$\alpha/\sqrt{Var(\alpha)}$)이 0에 수렴해감을 의미하며, 결과적으로 회귀계수의 신뢰도는 저하된다.

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

II. 실험의 개요

1. 실험의 기본개념

본 연구의 목적은 로짓모형의 전체적 신뢰도와 추정된 계수의 값 그리고 추정된 계수의 신뢰도에 대한 다중공선성의 영향을 규명하는데 있다. 따라서 효용함수내 설명변수간 상관계수가 실험에 있어서 중요한 제어변수가 된다. 다시 말해서 효용함수를 구성하는 설명변수간 상관관계의 정도에 따라서 모형의 설명력(ρ^2), 계수의 신뢰도(t_α) 그리고 계수(α_j)의 변동을 살펴볼 것이다.

한편, 회귀분석에서는 모든 설명변수들이 하나의 회귀방정식으로 정의되지만 로짓모형에서는 설명변수들이 여러 개의 효용함수로서 정식화된다. 효용함수를 단위로 지수함수를 취하는 로짓모형에 있어서는 상관관계를 맺고 있는 변수들이 동일한 선택대안의 효용함수에 존재하는가 아니면 서로 다른 효용함수의 설명변수로 정의되는가에 따라서 다중공선성의 영향도 달라질 것으로 예상된다. 이러한 특성을 감안하기 위해서 본 연구에서는 시나리오를 설정하고, 시나리오별로 효용함수를 구축, 실험을 수행한다. 자세한 내용은 2절의 '1) 시나리오 설정'을 참조하기 바란다.

마지막으로 실험은 연구자에 의해서 인위적·구조적으로 만들어진 데이터를 활용해서 수행하게 된다. 이에 대한 상세한 내용은 3절의 '3) 로짓 데이터의 생성'을 참조하기 바란다.

2. 실험조건의 설정

본 연구에서는 이항로짓모형을 이용해서 로짓모형내 두변수간 다중공선성의 영향을 검증하고자 한다. 즉, 기본적으로 본 연구에서 활용할 효용함수의 기본형은 식(6)과 같다.

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

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 유사한 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 지적해둔다.

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 경우와 그렇지 않은 경우 역시 다중공선성의 영향은 다르게 나타날 것으로 예상된다. 이상의 두 가지 사안은 단일방정식으로 정의되는 회귀분석에서는 고려할 필요가 없지만 여러 개의 효용함수를 채용하는 로짓모형에서는 고려되어야 할 사항이라 판단된다.

마지막으로 상관관계를 갖고 있는 두 변수 중에서 오직 한 변수만이 선택에 영향을 미치는 경우와 두 변수 모두 선택행동에 영향을 미치는 경우 역시 다른 결과가 예상된다[3].

3) 회귀분석에 있어서는 상관관계를 맺고 있는 변수

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

(1) 시나리오 1

상관관계를 맺고 있는 두 설명변수가 동일한 효용함수에 포함되어 있는 경우를 검증한다.

- 시나리오 1-1 : 상관관계를 맺고 있는 두 설명변수 가운데 오직 하나의 변수만이 선택행동에 영향을 미치는 경우
- 시나리오 1-2 : 상관관계를 맺고 있는 두 설명변수 모두 선택행동에 영향을 미치는 경우

(2) 시나리오 2.

상관관계를 맺고 있는 두 설명변수가 서로 다른 효용함수에 포함된 경우를 검증한다.

- 시나리오 2-1 : 상관관계를 맺고 있는 두 설명변수 가운데 오직 하나의 설명변수만이 선택행동에 영향을 미치는 경우
- 시나리오 2-2 : 상관관계를 맺고 있는 두 설명변수 모두 선택행동에 영향을 미치는 경우

(3) 시나리오 3.

설명변수에 대해서 공통의 계수를 갖도록 모델이 구축되는 경우를 검증한다.

- 시나리오 3-1 : 상관관계를 맺고 있는 두 설명변수가 계수를 공유하는 경우
- 시나리오 3-2 : 상호독립인 두 설명변수가 계수를 공유하는 경우

2) 설명변수설정 및 효용함수의 구축

(1) 설명변수의 도입

전절에서 설정된 6개의 시나리오를 모두 검증하기 위해서는 최소한 3개의 설명변수를 도입해야 하며, 각각의 변수들 간에는 식(7)과 같은 관계가 전제되어야 한다. 즉, 설명변수 x_2와 설명변수 x_3 사이에 다중공선성이 존재하며 나머지 변수들 사이에는 상관관계가 존재하지 않는다.

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

(2) 효용함수의 구축

3개의 변수를 활용해서 전절에서 설정된 시나리오별로 다중공선성의 영향을 검증하기 위해서는 다음과 같이 4가지 유형의 효용함수가 필요하다.

$$U_1 = \alpha_0 + \alpha_1 x_1 + \epsilon_1 \quad (8)$$
$$U_2 = \alpha_2 x_2 + \alpha_3 x_3 + \epsilon_2$$

$$U_1 = \alpha_0 + \alpha_1 x_1 + a_3 x_3 + \epsilon_1 \quad (9)$$
$$U_2 = \alpha_2 x_2 + \epsilon_2$$

$$U_1 = \alpha_0 + \alpha_1 x_1 + a_2 x_3 + \epsilon_1 \quad (10)$$
$$U_2 = \alpha_2 x_2 + \epsilon_2$$

$$U_1 = \alpha_0 + \alpha_1 x_1 + a_2 x_3 + \epsilon_1 \quad (11)$$
$$U_2 = \alpha_1 x_2 + \epsilon_2$$

여기서 α_0은 대안속성 더미의 계수, $\alpha_1,\alpha_2,\alpha_3$는 각각 효용함수를 구성하는 설명변수 x_1,x_2,x_3의 계수, ϵ_1,ϵ_2는 오차항이다. 변수간 상관관계에 대해서는 $Cov(x_1,x_2) = Cov(x_1,x_3) = 0$, $Cov(x_2,x_3) \neq 0$을 가정하였기 때문에 식(8)은 시나리오 1을, 식(9)는 시나리오 2를, 그리고 식(10)과 식(11)은 시나리오 3을 검정하는데 활용된다.

3. 변수의 생성 및 실험의 수행

1) 효용함수에 대한 가정

본장의 제1절에서 언급한 바와 같이 본 연구에서는 효용함수에 대한 모든 정보를 알고 있음을 전제로 하고 있다. 이는 다시 말해서 모든 개인의 선택결과는 $P(U_1 \geq U_2)$이면 대안 1이, 그 역이면 대안 2가 선택되었음을 의미한다. 본 연구에서는 로짓모형의 계수에 대해서 다음의 값들을 가정하였다.

$$E(\alpha_0) = 0 \quad (12)$$
$$E(\alpha_1) = E(\alpha_2) = 1.0$$
$$E(\alpha_3) = \begin{cases} 1.0 & x_3\text{가 선택에 영향을 미칠 경우} \\ 0.0 & x_3\text{가 선택에 영향을 미치지 않을 경우} \end{cases}$$

의 도입 자체만으로도 회귀계수의 신뢰도가 저하된다(식(4), 식(5) 참조). 상세한 내용은 이성우 외(2005) pp284-298을 참조하기 바란다.

2) 설명변수의 생성

수치실험에 이용될 설명변수의 값들은 기본적으로 난수발생기를 통해서 생성되었으며, 다음과 같이 정규분포를 따르도록 추가적인 조작이 가해졌다[4]. x_i가 정규분포를 따르도록 한 것은 실험환경이 사회현상을 최대한 반영하도록 하기 위함이다.

$$x_1, x_2, x_3 \sim N(0,1) \tag{13}$$

한편, 식(14)와 trial-and-error방식[5]을 이용해서 x_1, x_2에 대해서는 상관관계가 0이 되도록, 그리고 x_2, x_3간에는 실험의 구조화를 위해서 상관계수(r_{x_2,x_3})가 특정 값을 갖도록 하였다. 참고적으로 각각의 변수들은 모두 1000개씩이 생성되었다(관측치가 1000개인 경우를 상정함).

$$x_3 = \theta \cdot x_2 + (1-\theta) \cdot x_k,\ x_2, x_k \sim N(0,1) \tag{14}$$

<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
...	...	...	...
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 작성하는 것이다.

한편, 효용함수를 구성하는 두 요소, 즉 확정적 효용과 확률적 효용 중에서 확정적 효용은 효용함수식 (8)~(11)과 식(12)의 계수값, 그리고 <표 1>의 설명변수의 값을 이용해서 산정이 가능하지만 확률적 효용은 아직까지는 제시되어 있지 않다[6]. 로짓모델은 ϵ_i에 대해서 최빈치(mode) η(=0), 스케일파타메타 ω(=1)인 감벨분포를 가정함으로써 도출된다. 따라서 수치실험을 보다 엄격하게 수행하기 위해서는 가정된 분포파라메타의 감벨분포를 따르는 난수들을 생성해서 각각의 데이터 행의 오차항으로 추가하고 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.

선택결과를 결정하기 위한 모든 요소들(효용함수의 형태, 관련 변수들-x_i, α_j, ϵ_k)이 결정됨에 따라 다음의 조건식을 이용해서 로짓 데이터를 작성하였다.

$$\text{선택결과} = \begin{cases} 1 & \text{if} \dfrac{\exp(U_1)}{\exp(U_1)+\exp(U_3)} \geq 0.5 \\ 2 & otherwise \end{cases} \tag{15}$$

가령 <표 2>는 시나리오 1-1, $r_{x_2,x_3} = 0.9$인 경우의 제 변수들과 선택결과를 나타내고 있다.

<표 2> 로짓 데이터의 작성 예($r_{x_2,x_3} = 0.9$)

선택결과	x_1	x_2	x_3	ϵ_1	ϵ_2
2	-0.4451	1.71464	1.90481	0.09892	0.82970
2	0.15425	2.26234	1.31009	0.71747	-1.6675
1	0.32816	-1.0499	-1.2579	-1.4334	-0.7183
...	...	...	...	...	...
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

4) 본 연구에서는 엑셀의 난수발생모듈인 RAND()를 이용해서 설명변수의 값들을 생성하였다. 단, 이 모듈이 제공하는 난수는 균일분포를 따르기 때문에 정규분포의 난수로 가공하기 위해 다음 식(중심극한정리)을 활용해서 설명변수를 생성하였다.

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

여기서, $\bar{x}$는 x_i의 평균, s_x는 x_i의 표준편차

5) 두 설명변수간 상관계수(r_{x_2,x_3})가 특정 값을 갖도록 하기 위해서는 θ값을 반복적으로 조정하는 과정이 필요하다. 식(14)는 상관계수가 0인 두 변수를 생성하는 데에도 활용된다.

6) 확정적 효용만으로 선택결과를 결정하면 로짓모형은 추정되지 않는다.

한편, 모델 추정에 사용된 각종 변수들간의 상관관계는 아래의 <표 3>과 같다. 표에서 보는 바와 같이 x_1과 x_2, x_3 사이에는 상관관계가 없는 반면 x_2와 x_3 사이에는 일정 수준의 상관관계가 존재한다. 본 연구에는 두 변수간 상관계수(r_{x_2,x_3})가 0.0에서 0.9까지 0.1씩 증가하도록 제어되고 있다. 오차항과 설명변수간 미세한 상관관계가 분석결과를 왜곡시킬 가능성은 매우 낮다고 판단된다.

<표 3> 모델에 활용된 데이터간 상관계수

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

III. 로짓모형에 대한 다중공선성의 영향

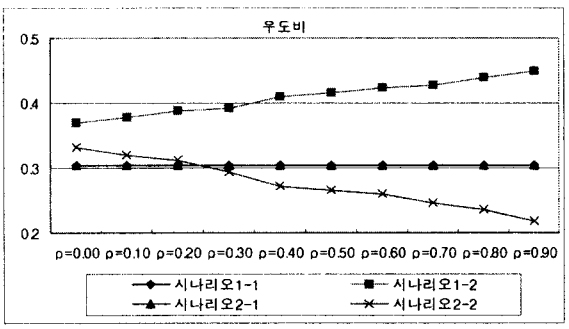
제2장의 실험을 통해서 총 60개(시나리오 6 × x_2와 x_3의 상관계수(r_{x_2,x_3})의 수준 10)의 로짓모형이 구축되었다. 본 장에서는 구축된 모델로부터 로짓모형의 적합도, 추정된 계수 그리고 계수의 신뢰도에 대한 다중공선성의 영향을 살펴본다.

1. 모델의 적합도에 대한 영향

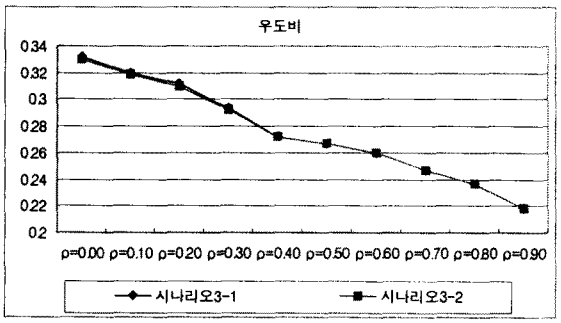
로짓모델의 적합도는 로그우도비 또는 McFadden의 결정계수라 불리는 ρ^2에 의해서 측정된다. 아래의 <그림 1>과 <그림 2>는 각각의 시나리오별, 상관계수(r_{x_2,x_3})별 우도비를 나타낸 것이다. 먼저, x_3가 효용함수에는 포함되어 있지만 선택에는 영향을 미치지 않는 시나리오 1-1과 2-1의 우도비는 x_2와 x_3간 상관의 정도나 x_2와 x_3의 상대적 위치에 관계없이 불변인 것으로 나타났다. 이와 같은 현상은 회귀분석에서도 발견되는 현상이다(Appendix의 회귀모델-시나리오1 참조).

한편, 효용함수내 모든 변수가 선택에 영향을 미치는 경우(시나리오 1-2와 시나리오 2-2)에는 상관관계를 맺고 있는 변수의 상대적 위치에 따라서 모델의 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 모델의 적합도가 감소하는 경향을 보였다. 시나리오 1-2와 2-2로 대표되는 상황은 로짓모형으로 구축된 교통수단선택모형에서 쉽게 발견할 수 있다. 가령 수단선택모형에 있어서 동일 수단의 통행요금과 통행시간은 높은 상관관계를 맺고 있으며 같은 효용함수의 설명변수로 도입된다(시나리오 1). 한편, 서로 다른 교통수단의 통행시간 사이에도 높은 상관관계가 있으며 서로 다른 효용함수의 설명변수로 도입된다(시나리오 2)

한편, 계수의 공유를 실험한 시나리오 3은 기본적으로 상관관계를 맺고 있는 두 변수(x_2, x_3)가 서로 다른 효용함수에 위치하고 있고 모든 변수가 선택에 영향을 미친다는 점에서 시나리오 2-2와 유사하며, 상관계수의 변화에 따른 우도비의 변화 패턴 역시 시나리오 2-2와 유사함을 알 수 있다.



<그림 1> 시나리오 1과 2에서의 우도비



<그림 2> 시나리오 3에서의 우도비

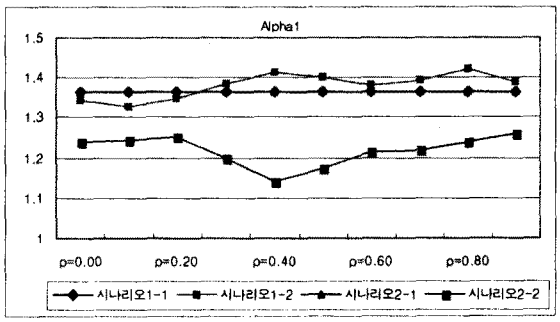
2. 추정계수에 대한 영향

설명변수의 계수는 선택행동에 대한 설명변수의 영향력으로서의 의미를 가지며, 정부정책을 통해서 제어가능한 변수인 경우에는 정책의 효과를 예측하는데 필요한 기초정보를 제공해 준다. 따라서 추정된 계수가 왜곡되어 있다면 정책효과에 대한 과대평가 또는 과소평가로 이어져 잘못된 의사결정에 이르게 된다.

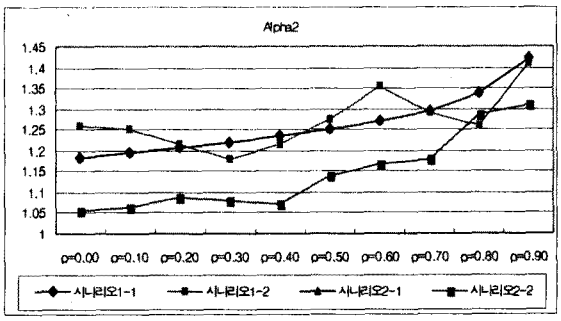
시나리오별 추정된 계수값은 아래의 <그림 3>~<그림 5>와 같다(추정된 계수 가운데 α_0는 기대치와 신뢰도가 0이므로 해석에서 제외한다). x_2, x_3와 독립인 x_1의 계수(α_1)는 시나리오나 두 변수(x_2, x_3)간 상관의 정도에 관계없이 비교적 안정된 값을 유지하고 있는 반면 x_3와 상관관계를 맺고 있는 x_2의 계수(α_2)는 x_3와의 상관관계가 높아짐에 따라 값이 커지는 경향을 보이고 있다. 이러한 현상은 모든 시나리오에서 공통적으로 나타나지만 시나리오 2-2에서 가장 급격하게 나타남을 알 수 있다.

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

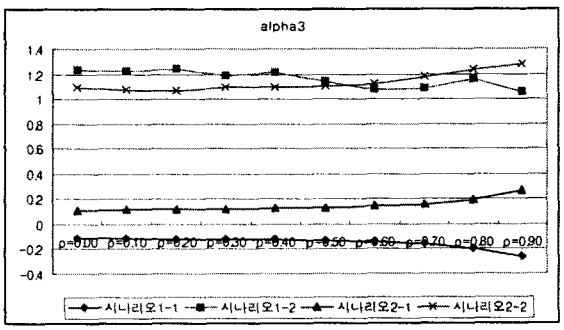
이상의 내용을 정리하면, 첫째, 선택행동에는 영향을 미치지는 않지만 기존 변수와 상관관계가 높은 변수가 로짓모형의 설명변수로 도입되면 높은 상관관계에 있는 기존변수의 선택행동에 대한 기여도는 과다하게 평가될 가능성이 크다. 둘째, 높은 상관관계의 두 설명변수가 동일한 효용함수에 포함되는 경우보다 서로 다른 효용함수의 설명변수로 포함되는 경우(시나리오 2-2)가 보다 심각한 문제를 야기한다. 셋째, 상관관계에 있는 두 변수 중에서 과다하게 평가된 변수를 규명하는 것은 현실에서는 발견하기 불가능하다. 실질적으로 시나리오 2-2에 있어서 x_2와 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 있음을 알 수 있다. 로짓모형을 이용하는 과정에서 어떤 변수가 과다 추정되었는지는 확인할 수 없다는 문제가 있다.



<그림 3> 시나리오 1과 2에서 α_1의 추정치



<그림 4> 시나리오 1과 2에서 α_2의 추정치

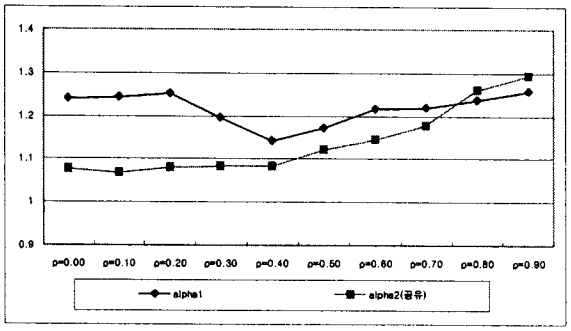


<그림 5> 시나리오 1과 2에서 α_3의 추정치

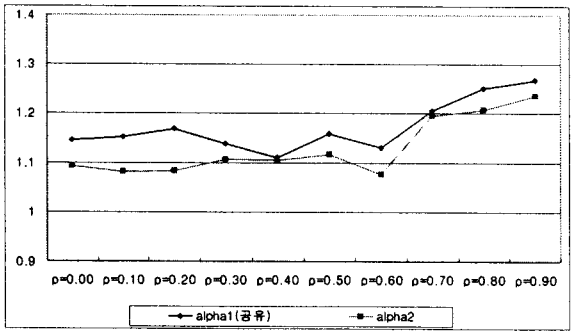
한편, 계수의 공유를 전제로 한 시나리오 3에 대한 분석결과는 아래의 <그림 6>, <그림 7>과 같다. 상관관계에 있는 두 변수(x_2, x_3)에 대해서 공통의 계수를 설정한 시나리오 3-1에서

는 두 설명변수간 상관계수가 커짐에 따라 공유된 계수(α_2)의 값도 증대되는 것을 알 수 있다. 반면 상관관계에 있는 두 변수에 대해서는 독립의 계수를, 상호 독립인 두 변수(x_1, x_2)에 대해서는 공통의 계수(α_1)를 설정한 시나리오 3-2에서는 추정된 계수 모두가 상관관계가 높아짐에 따라 커지는 경향을 보였다.

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



<그림 6> 시나리오 3-1에서 계수추정치



<그림 7> 시나리오 3-2에서 계수추정치

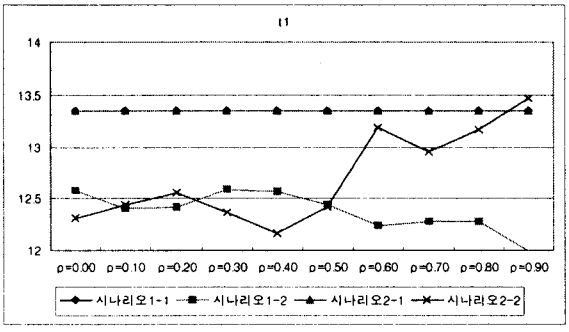
3. 추정계수의 신뢰도에 대한 영향

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 신뢰도에 어떤 형태로든 영향을 미칠 것으로 예상된다.

아래의 <그림 8>~<그림 10>은 다양한 시나리오에 있어서의 추정된 계수의 신뢰도(t통계량)를 플로팅한 것이다. <그림 8>은 α_1의 시나리오별-상관계수(r_{x_2,x_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. 단, 이러한 변동이 유의한 의미를 갖는지는 추가적인 분석이 필요한 것으로 판단된다.

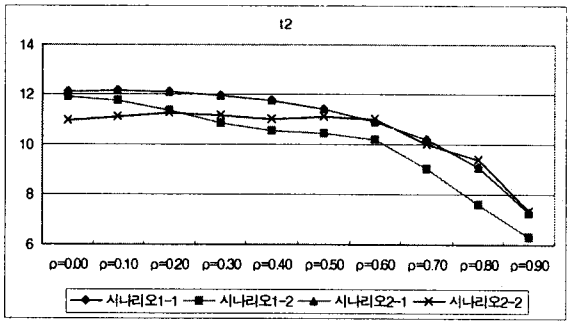
본 연구의 주요 관심대상인 α_2의 t통계량은 시나리오에 관계없이 설명변수 x_3와의 상관관계가 높아짐에 따라 급격하게 저하되는 것으로 나타났다. 다만 다중공선성이 계수의 신뢰도에 직접적으로 영향을 미치는 회귀분석보다는 지수함수를 취함으로써 상관관계가 다소 완화되는 로짓모형에서 다중공선성의 영향이 작게 나타나는 것으로 판단된다[7].

마지막으로 α_3의 신뢰도는 당해 변수가 선택행동에 영향을 미치지 않는 경우(시나리오 1-1과 2-1)에는 신뢰성 자체가 없을 뿐만 아니라 설명변수 x_2와의 상관계수가 높아져도 신뢰도는 높아지지 않음을 알 수 있다(이러한 결과는 회귀분석에서도 찾아볼 수 있다. Appendix의 시나리오 1 참조).

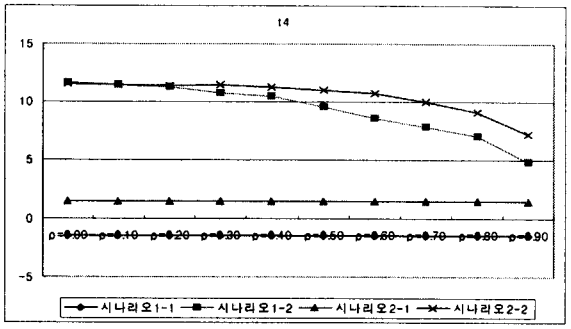


<그림 8> 시나리오 1과 2에서 α_1의 t통계량

7) Appendix의 시나리오 1, 2의 경우 두 변수간 상관계수가 0.9일 때의 α_2의 t통계량은 상관관계가 없을 때의 t통계량의 40% 수준이나 로짓모형에 대한 시나리오 1과 2에서 상관계수가 0.9일 때의 α_2의 t통계량은 상관관계가 없을 때의 t통계량의 60% 수준임



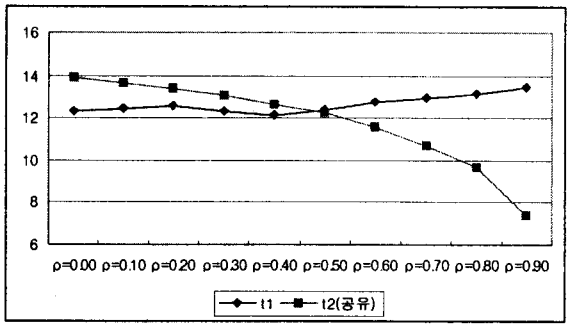
<그림 9> 시나리오 1과 2에서 α_2의 t통계량



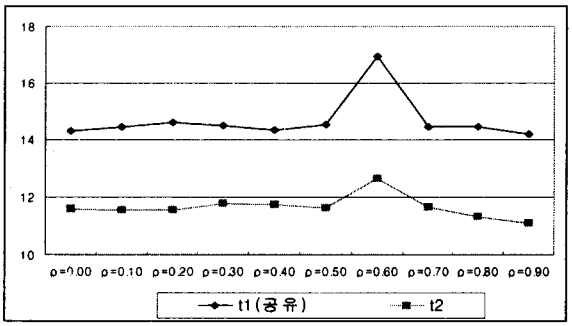
<그림 10> 시나리오 1과 2에서 α_3의 t통계량

계수를 공유하는 경우에 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 3-1에서 공통 계수 α_2의 신뢰도는 상관관계가 높아짐에 따라 급격히 저하되는 반면 두 변수(x_2,x_3)와 독립이고 계수 역시 독립적으로 취하도록 한 α_1의 신뢰도는 미세하나마 향상되는 경향을 보였다.

독립인 두 변수(x_1,x_2)를 대상으로 공통계수를 취한 시나리오 3-2에서는 일부 지점(ρ=0.6)에서 특이점이 출현하기는 하였으나 시뮬레이션이라고 하는 특수성을 감안하면 전반적으로 신뢰도는 안정되어 있다고 할 수 있다.



<그림 11> 시나리오 3-1의 계수의 t통계량



<그림 12> 시나리오 3-2의 계수의 t통계량

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 설명변수가 추가될 경우 모델의 적합도(R^2)가 개선되지만 로짓모형에서는 설명변수의 추가를 통해서 모델의 적합도가 저하될 수도 있음이 규명되었다.

둘째, 유사변수에 대해서 계수를 공유하도록 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 있다.

넷째, 다중공선성은 상관관계를 맺고 있는 변수들(또는 계수들)의 신뢰도를 저하시키는 역할을 하며 그 정도는 회귀분석의 경우에 비해서는 작지만 절대적 측면에서 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

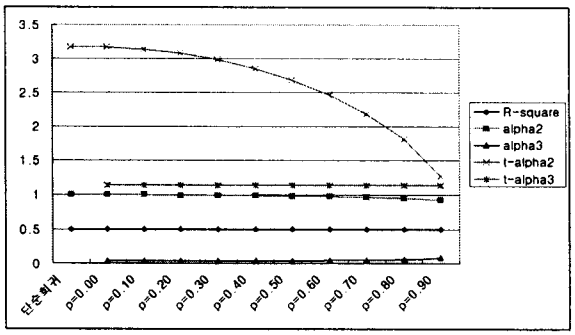
절대 다수는 상당히 높은 상관관계를 맺고 있는 변수들을 설명변수로 활용하고 있으며, 결과적으로 추정된 모형에는 본 연구에서 발견된 다중공선성의 영향이 작용하고 있을 것으로 판단된다.

향후 연구과제로 다른 형태의 선택모형, 가령 네스티드 로짓모형에 있어서의 다중공선성의 영향을 규명하는 것도 필요한 작업이라 판단된다. 또한, 로짓모형에 있어서 다중공선성 문제를 완화하기 위한 방안에 대한 연구도 반드시 수행되어야 할 것으로 판단된다. 가령, 통행시간이라는 변수 대신 통행속도변수를 활용하는 방법 등이 고려될 수 있을 것이다.

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.

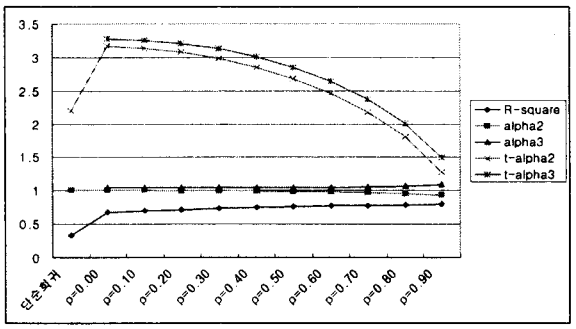


주1 : α_2의 t통계량은 실제 값의 1/10로 환산된 값임
주2 : 단순회귀는 x_3를 설명변수로 포함하지 않음
주3 : y는 $y_i = x_{2i} + \epsilon_i$를 이용해서 생성함

<A-그림 1> 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.



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

<A-그림 2> 시나리오 2의 회귀분석결과

참고문헌

1. 유지성, 오창수(2004), "현대통계학", 박영사
2. 이성우, 민성희, 박지영, 윤성도(2005), "로짓・프로빗모형 응용", 박영사
3. 土木學會(1995), "非集計モデルの理論と實際, 丸善(株)
4. Juan de Dios Ortúzar, David Hensher, Sergio Jara-Diaz(1998), " Travel Behaviour Research : Updating the State of Play", Pergamon
5. Moshe Ben-Akiva, Steven R. Lerman(1987), Discrete Schoice Analysis : "Theory and application to travel demand", MIT Press
6. Norbert Oppenheim(1994), "Urban Travel Demand Modeling from individual choice to general Equilibrium", John Willy & Sons, Inc.
7. Simon P. Washington, Matthew G. Karlaftis, Fred L. Mannering(2003), "Statistical and Econometric Methods for Transportation Data Analysis, Chapman & Hall/CRC
8. Tommy Gärling, Tomas Laitila, Kerstin Westin(1998), "Theoretical Foundations of Travel Choice Modeling", Pergamon