가계별 식료품비 지출행위의 준모수적 분석

권 오 상*

Keywords
준모수적 모형(semiparametric models), 벌칙스플라인 회귀분석(penalized spline regression), 식품소비(food consumption)

Abstract
Household expenditures on food are analyzed with the use of the 2000 household expenditure survey. A penalized spline model, which is a semiparametric model, is used to estimate the household food expenditure functions and Engel curves. The model allows a flexible relationship between household income and expenditures on food. Results suggest households with large family size, without farm income, and employed, married, aged and well-educated household heads are more likely to consume more than others. There exit thresholds of food expenditure, and the relationship between income and food expenditure is not necessarily monotonic.

차례

1. 서론
2. 분석모형
3. 사용된 자료
4. 분석결과
5. 요약 및 결론

* 서울대학교 농경제사회학부 교수

1. 서론

사회경제적 환경이 달라지면서 식품소비행태가 어떻게 변화되는지, 그리고 소비자 가구의 특성별로 식품소비행위가 어떻게 달라지는지 등은 농업 및 식품관련 연구에 있어 매우 중요한 분석대상이라 할 것이다. 이러한 중요도를 반영하여 국내 연구자들에 의해서도 식품종류별 혹은 전체 식품에 대한 소비지출액이나 행위를 결정하는 변수들을 찾고 식품소비행위의 특성을 분석하려는 연구가 많이 시도되었다.

이러한 연구들은 김태균·사공용(1994), 사공용·김태균(1994), 사공용·최지현(1995), 서종석(1994), 조덕래·김영식(1985), 정경수·박창원(1998) 등과 같이 시계열 자료를 이용하고 국가 전체의 소비량을 나타내는 자료를 이용해 식품소비형태의 변화를 분석하려는 연구와, 이계임·김성용(2003), 이계임 외(1998, 1999, 2003), 이계임·김민정(2003) 등과 같이 가구별 횡단면자료를 이용하여 가구 특성이 식품소비행태에 미치는 영향을 분석하려는 연구로 대별할 수 있을 것이다.

시계열자료를 이용해 수요체계를 분석하는 연구는 각 식품별 수요함수를 추정할 수 있고 이를 통해 탄력성 등 수요체계의 여러 특성을 분석할 수 있으나, 국가전체 자료를 사용하기 때문에 각 소비자나 가구가 가지고 있는 사회경제적 특성이 식품소비에 어떤 영향을 미치는지를 분석하지는 못한다. 이런 점에서 가구별 횡단면자료를 이용한 식품소비행위분석이 상대적으로 더 유용할 수도 있다.

본고 역시 앞에서 소개된 이계임 외(1998, 1999, 2003) 등의 기존연구들처럼 횡단면 자료를 이용해 가구별 특성이 식료품 소비지출액에 미치는 영향을 분석하고자 한다. 그러나 이 연구는 사용하는 자료와 분석방법에 있어 기존연구와 차별화를 시도한다. 첫째, 본 연구는 기존연구들과 달리 가계소득을 명시적으로 반영하여 이를 식료품비 지출액 결정에 가장 큰 영향을 미치는 변수로 간주한다. 기존 연구들은 대개 「도시가계조사」 자료를 이용해 분석을 하였는데, 이 자료는 소득을 근로자가구에 한하여 공개하였기 때문에 이들 연구들은 소득자료를 아예 포함하지 않거나, 근로자가구만을 분석대상으로 하거나, 아니면 총지출액을 소득의 대용변수로 사용하였다. 본고가 분석에 이용하는 「가구소비실태조사」 자료는 근로자가구는 물론 자영업가구의 소득도 포함하고 있으며, 또한 농림축어업 소득이 있는 경우도 포함하기 때문에 가구유형별, 소득유형별 식료품지출액이 어떻게 다른지를 분석할 수 있다.

둘째, 기존 연구들은 회귀분석이나 수요체계방정식 추정 등을 통해 가구 특성이 소비지출액 결정에 선형 혹은 로그선형의 영향을 미친다고 가정하고 분석하였다. 그러나

경제이론의 잘 알려진 결론이 전망하는 바이지만 소득액과 식료품 소비지출액 사이에는 선형의 관계가 형성되지 않을 수가 있다. 식품소비 지출액은 낮은 소득수준에서는 소득이 늘 경우 빨리 증가하지만 높은 소득계층에서는 소득에 매우 비탄력적으로 반응할 수가 있다. 따라서 소득이 늘면 식료품소비도 지속적으로 늘어난다고 가정하는 기존의 선형방정식 분석법은 소득과 식료품 소비지출액 사이의 관계를 분석하는데 있어 큰 한계를 가진다. 본고는 가계의 여타 사회경제적 특성을 나타내는 변수들은 식료품 소비지출액과 선형의 관계를 가질 수 있지만 소득의 경우 소비지출액과 비선형/비단조적인 관계를 가질 수 있다고 보고, 함수형태에 대한 제약을 완화한 준모수적(semiparametric) 회귀분석을 통해 두 변수사이의 관계를 분석하고자 한다. 본고는 아울러 총지출액에서 식료품비 지출액이 차지하는 비중인 엥겔계수가 가구 특성에 의해 어떤 영향을 받는지도 역시 준모수적 회귀분석한다.

본고의 구성은 다음과 같다. 제2장은 분석모형에 대해 설명하며, 제3장은 분석에 사용된 자료를 소개하고 그 특성을 보여준다. 제4장은 추정결과를 설명하며, 마지막 제5장은 결과를 요약하고 결론을 내린다.

2. 분석모형

y_i를 가계 i의 식료품비 지출액이라 하고 X_i를 이 가계의 사회경제적 특성을 나타내는 벡터라 하며 ε_i를 확률변수라 하면, 이 가계의 식료품비 지출액은 $y_i = g(X_i) + \varepsilon_i$와 같이 추정되어야 한다($i = 1, \ldots, n$). 식료품비 지출액이 가계의 여러 특성, 특히 소득에 의해 어떻게 영향을 받는지를 사전에 알 수 없기 때문에 지출함수 $g(X_i)$의 함수형태를 미리 가정하지 않는 비모수적(nonparametric) 분석이 유용할 것이다.

이러한 비모수적 회귀분석을 위해서는 매우 다양한 분석법이 이용될 수 있는데, 이들 방법들은 크게 커널(kernel)함수 추정법과 준모수적 분석법이라 할 수 있는 스플라인(spline) 추정법으로 대별된다(Ruppert et al. 2003; Li and Racine 2007; Pagan and Ullah 1999). 이 중 Nadaraya-Watson 추정법으로 대표되는 커널함수 회귀분석법은 몇 가지 방향으로 개발되어 왔다. 예를 들어 최근 Li and Racine(2003)은 비교적 많은 수의 설명변수를 포함하고, 설명변수 가운데 일부는 더미변수와 같은 이산적인 변수일 경우도 포괄할 수 있는 매우 신축적인 분석기법을 개발하였고, 이들의 분석법은 본고가 사용할 자료의 성격과도 잘 부합된다. 그러나 커널함수 추정법은 추정과정에서 사

용될 최적 대역너비(bandwidth)를 구하여야 하는데, 본고의 경우처럼 설명변수가 많고 관측치도 매우 많을 경우 그 작업이 용이하지 않다[1]. 따라서 본고는 하나의 대안으로서 준모수적 분석법인 스플라인 추정법을 사용하고자 한다.

스플라인 추정법을 사용할 때 일부의 설명변수는 통상적인 회귀분석이 가정하는 바와 같이 식료품 지출액과 선형의 관계를 맺는다고 가정하고, 나머지 변수들의 경우 보다 신축적인 형태로 지출액에 영향을 미친다고 가정한다. 변수 가운데 선형의 영향을 미치는 변수들을 X_i라 하자. 각 가계가 속해있는 지역이나 가구주의 성별, 연령 등을 나타내는 더미변수나 이산변수의 경우 그 특성상 식료품 지출액과 선형관계를 맺고 있다고 가정하는 것에 큰 무리는 없을 것이며, 따라서 벡터 X_i에 포함시킬 수 있다. 반면 소득의 경우는 지출액에 선형의 영향을 미쳐 벡터 X_i의 한 구성요소가 되겠지만 추가로 비선형의 영향을 지출액에 줄 수도 있다. 따라서 소득을 z_i라 할 때 식료품 지출함수를 다음과 같이 설정할 수 있다.

$$y_i = X_i\beta + f(z_i) + \varepsilon_i \quad (1)$$
$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

이상의 모형은 설명변수의 일부는 모수적으로, 나머지 일부인 $f(z_i)$는 비모수적으로 처리하기 때문에 준모수적 분석모형이라 부를 수 있다. 스플라인 추정법은 알려지지 않은 함수 $f(z_i)$를 몇 개의 기저(basis)함수를 연결하여 근사한다. 예를 들어 다음과 같은 근사가 가능하다.

$$f(z_i) = \sum_{j=1}^{K} u_j (z_i - k_j)_+^p \quad (2)$$

식 (2)에서 k_j는 매듭(knots)이라 불리며, 그 적절한 값을 연구자가 찾아내어야 한다. $(z_i - k_j)_+$는 $z_i \geq k_j$일 경우 $z_i - k_j$와 같고 $z_i < k_j$일 경우 0이다. 기저함수의 차수 p 역시 연구자가 지정하여야 한다. 추정과정을 통해 β와 u_j들을 얻게 되고, 이를 통해 소득이 식료품 지출에 미치는 영향을 특정한 함수형태에 대한 가정이 없이도 신축적으로 근사할 수 있다.

모든 관측치를 다 포함하는 추정모형의 행렬표현을 $y = X\beta + Zu + \varepsilon$이라 하면 모형의 추정은 통상적인 회귀분석과 마찬가지로 $\| y - X\beta - Zu \|^2$을 최소화하는 β와 u를 구

[1] 본고가 사용하는 자료에 대해 Hayfield and Racine(2006)이 제공하는 np팩키지를 이용하여 최적 대역너비를 구하는 작업을 실시하였으나 72시간 이상의 작업을 하여도 최적 대역너비를 찾아내지 못하였다.

함으로서 이루어질 수 있다. 그러나 이 과정에서 추정되는 함수값이 지나치게 불안정해지는 것을 막기 위해 D를 어떤 대칭적이고 양반정인(positive semidefinite) 행렬이라 할 때 u^TDu가 어떤 상한값을 가지도록 제약하며, 따라서 실제 추정치는 $\| y-X\beta-Zu \|^2 + \lambda^{2p}u^TDu$ 를 최소화하도록 하고, 이때 λ^{2p}는 제약식에 부과되는 승수이다(Ruppert et al 2003, p. 75)[2].

이상의 모형을 추정함에 있어 연구자는 적어도 세 가지 선택을 하여야 한다. 첫 번째로 제약식에 부과되는 λ의 크기를 정하여야 하고, 두 번째로 매듭의 수인 K와 각 매듭의 위치를 선택하여야 하며, 마지막으로 기저함수의 형태와 차수 p를 선택하여야 한다.

이 세 가지 선택 가운데 추정결과에 가장 큰 영향을 미치는 것이 λ의 선택이다. λ는 흔히 평활파라메터(smoothing parameter)라 불리는데, 그 값이 0일 경우 회귀식에 제약이 부과되지 않으면서 지나치게 불안정한 함수가 추정될 수 있고, 반대로 그 값이 지나치게 클 경우 매듭과 기저함수가 차지하는 역할이 무시해도 좋을 정도로 작아지면서 추정식은 통상적인 최소자승모형이 되어버린다[3]. 따라서 적절한 λ의 값은 자료의 성격을 잘 반영하도록 선택되어야 한다. λ값을 선택하기 위해서는 잔차제곱합(residual sum of squares, RSS)을 포함하는 기준함수를 만들어 이를 최소화하는 λ를 선택하도록 하는 Mallow의 C_p기준이나 Akaike의 정보기준(Akaike's information criterion, AIC), 교차확인법(cross-validation, CV) 등을 사용할 수 있고, 이들 기법과는 달리 RSS를 이용하지 않고 모형의 추정과정에서 얻어지는 분산의 추정치를 활용하는 우도함수(likelihood function)법이 있다(Ruppert et al. 2003, pp. 112-123). 이들 방법 가운데 우도함수법이 가장 적은 계산을 통해 최적 λ를 얻게 하므로 본고는 이를 사용하기로 한다.

최우추정법은 스플라인 회귀분석법을 흔히 혼합모형(mixed model)이라 불리는 일종의 확률효과(random effect) 페널추정모형의 일종으로 변환할 수 있다는 성질을 활용한다. 즉, Brumback et al.(1999)과 Ruppert et al.(2003, pp. 108-110)이 보여준 바와 같이 이 경우 스플라인 회귀식 $y=X\beta+Zu+\varepsilon$은 혼합모형의 일종이 되고, u는 추정되는 파라메터가 아니라 그 값이 예측이 되는 일종의 확률변수로 간주할 수 있다. 이 경우 통상적인 교란항을 나타내는 ε의 공분산은 $cov(\varepsilon)=\sigma_\varepsilon^2 I_n$와 같이 가정되고, 또 다른 확률변수인 u의 공분산은 $cov(u)=\sigma_u^2 I_K$와 같이 가정된다. 아울러 두 확률변수의 평균

---

[2] 목적함수를 이렇게 설정하여 추정치를 구하는 과정을 벌칙스플라인 회귀분석(penalized spline regression)이라 부르며, 도출되는 추정량은 능형(稜形) 회귀분석(ridge regression) 추정량의 한 형태가 된다.

[3] λ가 0일 경우 따라서 과소평활(under-smoothing)이 나타나고 그 값이 무한대일 경우 과대평활(over-smoothing)이 나타난다.

은 0이고, 서로 독립이다. 혼합모형에서는 β는 고정효과(fixed effect) 파라메터라 불리고 u는 확률효과(random effect) 파라메터라 불린다.

McCulloch and Searle(2001) 등이 보여주는 바와 같이 이러한 혼합모형의 파라메터 β와 σ_ε, σ_u는 최우추정법(ML)이나 ML추정법이 가지는 자유도의 편의문제를 보정하는 제약하의 최우추정법(restricted ML, REML)을 이용해 추정하며, u의 값은 추정결과를 이용해 예측이 된다. Ruppert et al.(2003, p. 113)은 스플라인 회귀식을 혼합모형으로 설정할 경우 $\lambda=(\sigma_\varepsilon^2/\sigma_u^2)^{1/2p}$의 관계가 성립함을 보였고, 따라서 최적 λ의 값은 스플라인 회귀식을 혼합모형으로 설정·추정하여 얻는 파라메터 추정치 $\widehat{\sigma}_\varepsilon$과 $\widehat{\sigma}_u$의 값을 통해 간접적으로 결정된다.

매듭의 수와 위치, 그리고 기저함수의 형태와 차수도 C_p기준이나 교차확인절차를 통해 선정할 수 있다. 그러나 Ruppert et al.(2003, pp. 124-127)의 시뮬레이션분석에 의하면, 매듭의 수를 20개 전후로 선정할 경우 매듭의 수나 기저함수의 형태와 차수의 차이가 모형의 설명력에 큰 영향을 주지 않는 것으로 밝혀졌다. 매듭이나 기저함수의 선택은 평활파라메터의 선택과 달리 이렇게 추정결과에 미치는 영향이 상대적으로 작기 때문에 본고는 다수의 매듭과 기저함수의 조합을 모두 추정한 후 적정 매듭과 기저함수를 찾기보다는 관련 연구에서 관행적으로 사용되는 절차를 사전에 적용하고자 한다.

즉 매듭의 경우 비교적 절적한 수인 것으로 인정되는 35를 선택하고($K=35$), 그 위치는 소득의 표본을 단 한번만 나오는 수치로만 재나열하고 이 중 $k_j=\left(\frac{j+1}{K+2}\right)$번째 표본 분위수(quantile)를 찾아내어 구하였다($j=1,\ldots,K$)[4]. 이렇게 매듭의 위치를 정해주면 관측치가 분포하는 공간을 보다 균등하게 나누어 매듭을 선정되게 된다.

기저함수의 경우 식 (2)와 같이 절단된 다항식(truncated polynomial)을 사용할 수도 있으나, 몇 가지 함수형태와 차수를 시도해본 결과 역시 많이 사용되는 기저함수 가운데 하나인 방사성(radial) 기저함수 $|z_i-k_j|$를 사용하여 $f(z_i)=\sum_{j=1}^{K}u_j|z_i-k_j|^p$와 같이 설정하는 것이 적절하고, 그 차수로는 $p=3$을 선택할 경우 평활정도가 가장 적절해 보였기 때문에 이를 선택하였다.

이렇게 차수 3인 방사성 기저함수를 사용할 경우 스플라인 회귀식 $\|y-X\beta-Zu\|^2+\lambda^{2p}u^TDu$에서의 행렬 D는 $D=\begin{pmatrix}0_{p\times p} & 0_{p\times K}\\ 0_{K\times p} & (\Omega^{1/2})^T\Omega^{1/2}\end{pmatrix}$와 같이 정의된다. 단, Ω의 (l,k)

---

[4] 이렇게 매듭을 찾아내고 모형을 추정하는 모든 절차는 소프트웨어 R 2.5.1을 사용하여 이루어졌다.

원소는 $|k_l - k_k|^3$이다. 혼합모형 $y = X\beta + Zu + \varepsilon$에서의 u의 공분산을 $cov(u) = \sigma_u^2 \Omega^{-1/2}(\Omega^{-1/2})^T$와 같이 가정하면, 이 혼합모형의 최우추정시 적용되어야 하는 소위 최적선형예측(Best Linear Unbiased Prediction, BLUP)기준이 행렬 D를 사용한 스플라인 회귀식과 같다는 것이 보여진다(Brumback et al., 1999; Ruppert et al., 2003, pp. 99-100).

이러한 혼합모형에서 확률변수 u의 공분산이 비교적 복잡한 형태를 지니므로 간단한 변수변환을 통해 모형을 더 단순화시킬 수 있다. 즉, $b = \Omega^{1/2}u$와 $\tilde{Z} = Z\Omega^{-1/2}$로 변환하고, 전체 추정모형을 $Y = X\beta + \tilde{Z}b + \varepsilon$으로 변형할 수 있다. 이렇게 변형하면 $cov(b) = \sigma_b^2 I_K$인 새로운 확률변수벡터 b가 도출되고, 최종추정모형은 아래와 같다.

(3) $$Y = X\beta + \tilde{Z}b + \varepsilon, \quad cov\begin{pmatrix} b \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2 I_K & 0 \\ 0 & \sigma_\varepsilon^2 I_n \end{pmatrix}$$

3. 사용된 자료

본고를 위한 분석자료는 통계청이 2001년에 조사하고 2002년에 공표한 2000년도 기준 「가구소비실태조사」 원자료이다. 이 조사는 전국의 가구를 대상으로 연간 소득과 소비지출, 저축·부채, 가구내구재 보유현황 등 가계자산에 관한 심층조사를 통해 국민의 생활수준과 소득·소비구조를 파악하며, 기존의 「도시가계조사」와 「농·어가경제조사」가 파악하지 못하고 있는 군지역의 비농가와 1인 가구를 포함하여 시도별 가구의 소득·소비수준, 가구유형별 소비구조의 차이를 분석하고, 가구내구재 및 저축·부채 등의 보유자산을 파악하여 각종 경제정책 및 사회정책 수립의 기초자료를 제공하고자 수집되었다. 「가구소비실태조사」는 5년 주기로 조사되기로 예정된 조사자료인데 2002년에 공표된 자료는 이 이름으로 조사된 가장 최근의 자료이며, 그 이후에는 더 이상 조사가 이루어지지 않았다.

2003년 이후에는 「가계조사」가 「가구소비실태조사」를 대신하고 있고, 전자가 가구당 소비지출액을 보다 세분하여 보여주고 있다. 그러나 전자는 소득의 경우 원칙적으로 근로자 가구 소득만을 보여주는 것에 반해 「가구소비실태조사」는 사업소득과 농어업소득까지 조사를 하고 있다. 그리고 「가계조사」가 가구를 서울 소재 가구와 여타 가구로만 구분함에 반해 「가구소비실태조사」는 가구의 소재지를 보다 세분하여 보여준

다. 이 자료는 조사된지 수년이 경과하긴 하였으나, 다양한 형태의 가구별 소득자료를 포함하고 있어 이들 소득과 식료품비 지출사이의 관계를 파악하는데 유용하게 사용될 수 있고, 본고의 분석목적이 세분화된 식료품비 지출행위가 아닌 전체 식료품비 지출행위이므로 본고의 목적을 위해 유용하게 사용될 수 있다.

「가구소비실태조사」는 전국 도시 및 농어촌지역의 27,000호를 대상으로 조사가 진행되었는데 본 연구에는 중요 변수에 대한 조사가 되지 못한 가구를 제외하고 23,383가구의 조사결과가 반영되었다. 아울러 조사자체에는 농·어가도 포함되었으나, 통계청의 최종집계에서는 농·어가로 분류되는 가구는 빠졌다. 다만 농·어가가 아닌 일반가구로 분류되는 가구 가운데도 농어업소득이 있는 가구가 있는데, 이들 가구의 농림축어업소득은 사업소득에 반영되도록 하였다. 이렇게 집계되는 가구당 소득은 1년간 전체 가구원이 벌어들인 총소득으로서 경상소득과 비경상소득을 모두 포함하며, 사업소득의 경우에는 재료비, 인건비 등 제비용을 제외한 순수익만을 포함한다[5].

가계의 연간 지출은 식료품, 주거광열, 가구집기가사용품, 피복신발, 보건의료, 교육, 교양오락, 교통통신, 기타소비지출 등의 9개 소비지출과 비소비지출로 분류된다. 본고는 이 가운데 식료품 지출이 가구의 소득 및 기타 어떤 변수들에 의해 영향을 받는지를 분석하고자 한다.

본고의 분석에 포함되는 변수들의 이름과 성격, 그리고 기초통계량은 <표 1>과 같이 정리된다. 가구당 연간 식료품비 지출액은 최소 60만원에서 최대 약 9,000만원까지의 분포를 보이며, 가구소득은 최소 0에서 최대 약 16억원까지의 분포를 보인다. 전체 가구소득에서 식료품비가 차지하는 비중은 최소 1%에서 최대 69.6%까지의 분포를 보인다.

소득 외에 식료품비 지출액에 영향을 미치는 변수로 각 가구가 거주하는 지역, 가구원의 수, 세대주가 무직인 경우, 근로자인 경우, 자영업자인 경우로 구분한 직업유형, 가구 내 취업자 수, 세대주의 나이, 세대주의 성별, 세대주 교육수준, 세대주가 현재 결혼한 상태인지 아니면 미혼이거나 사별 혹은 이혼한 상태인지의 여부, 가구 수입에 농림축어업소득이 포함되는지의 여부 등이 고려되었다. 이들 변수들의 기초통계량 역시 <표 1>에 정리되어 있다. 지역변수의 경우 서울을 기준으로 하여 더미변수로 나타내었고, 가구의 직업유형은 무직을 기준으로 하여 더미변수로 나타내었다.

---

5 이렇게 총소득을 소득지표로 사용하는 대신 총소득에서 주로 조세나 연금 납입금, 보험료 납입금 등인 비소비지출을 빼준 가처분소득을 사용할 수도 있다. 이 경우 상당수 가구의 소득이 0보다 작게 잡히며, 이들 가구들의 소비지출함수가 매우 불안정한 모습을 보여주는 것으로 분석되었기 때문에 본 연구는 총소득을 소득지표로 활용한다.

표 1. 기초통계량

변수명	변수유형	최소값	평균값	표준편차	최대값
식료품비 지출액	실수 (단위:백만원)	0.6	4.28	2.16	89.72
가구소득	실수 (단위:백만원)	0	26.35	30.74	1556
엥겔계수	식료품지출/가구 총지출(%)	1	23.8	9.11	69.6
부산	더미변수	0	0.08	0.28	1
대구	더미변수	0	0.06	0.24	1
인천	더미변수	0	0.08	0.28	1
대전	더미변수	0	0.07	0.26	1
광주	더미변수	0	0.06	0.24	1
울산	더미변수	0	0.04	0.29	1
경기	더미변수	0	0.10	0.30	1
강원	더미변수	0	0.05	0.23	1
충북	더미변수	0	0.04	0.20	1
충남	더미변수	0	0.05	0.21	1
전북	더미변수	0	0.04	0.19	1
전남	더미변수	0	0.05	0.21	1
경북	더미변수	0	0.05	0.21	1
경남	더미변수	0	0.05	0.21	1
제주	더미변수	0	0.02	0.15	1
가구원 수	정수	1	3.07	1.34	10
근로자 가구	더미변수	0	0.54	0.50	1
자영업자 가구	더미변수	0	0.26	0.44	1
취업자 수	정수	0	1.30	0.84	6
세대주 나이	정수	15	45.9	13.5	94
세대주 성별	더미변수 남자=1	0	0.77	0.42	1
세대주 교육수준	순위변수	0	2.95	1.48	6
세대주 결혼여부	더미변수 결혼상태=1	0	0.75	0.43	1
농림축어업소득 여부	더미변수	0	0.26	0.16	1

주: 교육의 경우 1=초등학교, 2=중학교, 3=고등학교, 4=대학교(4년제 미만), 5=대학교(4년제), 6=대학원

4. 분석결과

본고는 제2장에서 설명한 바와 같이 추정모형을 방사성 기저함수를 사용하여 다음과 같이 설정한다.

(4) $$y_i = X_i\beta + \sum_{j=1}^{K} b_j|\widetilde{z}_i - k_j|^3 + \varepsilon_i$$

y_i는 가구당 식료품비 지출액이다. 기저함수를 이용하는 스플라인모형은 연속변수에 대해서만 적용되므로 식 (4)에서 $\widetilde{z}_i$는 설명변수 가운데 유일한 연속변수인 가구당 소득의 변형된 형태이다. 가계특성을 나타내는 소득 외의 더미변수와 정수 혹은 순위변수들은 모두 X_i에 포함되며, 소득 z_i 역시 X_i에 포함되도록 하여 지출액과 선형의 관계를 형성하는 부분도 반영하도록 한다.

식 (4)와 같은 추정모형은 제2장에서 설명한 바와 같이 혼합모형의 일종으로 전환할 수 있기 때문에 혼합모형을 추정할 수 있는 소프트웨어들을 이용해서 추정할 수 있다. 본고는 스플라인 회귀분석을 위해 특별히 고안된 Wand et al.(2005)의 R 팩키지 SemiPar 1.0을 최우추정에 사용하며, 추정결과는 <표 2> 및 <표 3>과 같다[6].

<표 2>는 식료품비 지출액과 선형 관계를 가지는 소득 및 더미변수 혹은 순위변수 등이 식료품비 지출액에 미치는 영향을 보여준다. <표 3>은 연속변수인 소득이 비선형 관계를 통해 식료품비 지출액에 대해 미치는 영향을 결정하는 b_j (j = 1,...35)의 추정치 혹은 예측치를 보여준다. <표 3>은 평활파라메터 λ의 추정치로 81.34를 보여주고 있다.

먼저 <표 2>의 효과들을 보면, 소득증가의 선형효과는 식료품비 지출을 늘리는 역할을 한다. 다만 아래에서 설명하겠지만 SemiPar는 분산의 통계적 신뢰도를 추정하는 것이 가지는 부정확성으로 인해 준모수적으로 처리되는 변수에 해당되는 파라메터의 통계적 신뢰도에 관한 정보는 제공하지 않기 때문에 소득변수의 파라메트 $\beta_{소득}$의 t-값은 계산되지 않는다.

지역효과의 경우 울산을 제외한 모든 지역이 서울에 비해서는 가구당 식료품비 지출이 적음을 보여준다. 그러나 부산과 제주의 경우 그 통계적 유의성이 없으며, 대구의 경우도 10% 유의수준에서는 서울과 차이가 없다는 가설이 기각되지 않는다. 울산의

---

[6] SemiPar는 앞 절에서 설명한 바와 같이 스플라인 회귀식을 혼합모형의 일종으로 모형화하여 추정한다. 만약 추정식을 벌칙스플라인 회귀식으로 직접 추정하고자 한다면 Hastie(2006)의 R 팩키지 gam이나 SAS의 proc GAM을 이용할 수 있다.

경우 다른 조건이 같을 경우 서울보다도 가구당 식료품비 지출액이 조금 더 높은 것으로 나타났고, 반면 전북의 경우 역시 다른 조건이 같을 경우 서울의 가구에 비해 가구당 식료품비 지출액이 연간 약 70만원 정도 더 낮은 것으로 나타났다.

가구원의 수가 많을수록 식료품비 지출액은 늘어나며, 가구원 1명이 증가할 때 연간 약 74만원이 더 늘어난다. 가구의 직업 분류상 무직인 가구와 근로자 가구 사이의 식료품비 지출액 차이는 통계적으로 의미가 없으나, 자영업자 가구는 다른 조건이 같을 경우 무직인 가구에 비해 통계적으로 의미 있는 정도로 식료품 소비지출액이 더 많다.

표 2. β의 추정치

파라메터	식료품비 지출함수			엥겔계수함수		
	추정치	t-값	p-값	추정치	t-값	p-값
상수항	1.582	8.306	0.000	22.070	21.410	0.000
소득	0.039			-0.376		
부산	-0.014	-0.347	0.728	1.133	5.248	0.000
대구	-0.074	-1.630	0.103	-0.943	-3.972	0.000
인천	-0.238	-5.766	0.000	-0.871	-4.050	0.000
대전	-0.590	-13.610	0.000	-2.459	-10.900	0.000
광주	-0.545	-11.690	0.000	-1.653	-6.816	0.000
울산	0.139	2.479	0.013	-0.614	-2.103	0.035
경기	-0.360	-9.193	0.000	-2.098	-10.310	0.000
강원	-0.652	-13.340	0.000	-2.564	-10.090	0.000
충북	-0.692	-12.980	0.000	-1.777	-6.410	0.000
충남	-0.625	-12.080	0.000	-2.351	-8.740	0.000
전북	-0.993	-18.040	0.000	-3.509	-12.270	0.000
전남	-0.511	-9.784	0.000	-1.708	-6.290	0.000
경북	-0.202	-3.934	0.000	-0.732	-2.741	0.006
경남	-0.521	-10.270	0.000	-1.649	-6.255	0.000
제주	-0.001	-0.014	0.988	-0.102	-0.295	0.767
가구원 수	0.736	73.950	0.000	1.671	32.280	0.000
근로자 가구	0.004	0.127	0.898	0.017	0.097	0.922
자영업자 가구	0.187	4.898	0.000	-0.867	-4.363	0.000
취업자 수	-0.058	-3.467	0.001	0.371	4.236	0.000
세대주 나이	0.009	10.080	0.000	0.045	9.266	0.000
세대주 성별	-0.026	-0.913	0.361	0.489	3.255	0.001
세대주 교육수준	0.112	12.700	0.000	-0.902	-19.760	0.000
세대주 결혼여부	0.116	3.688	0.000	-0.205	-1.256	0.209
농림축어업소득 여부	-0.197	-3.179	0.002	0.224	0.695	0.486
lnL	-42224.5			-80718.7		

주: SemiPar는 X_i에 포함되는 소득 파라메터의 표준편차는 제공하지 않으므로 그 t-값도 계산되지 않음

표 3. b의 추정치

파라메터	식료품비 지출함수	엥겔계수함수	파라메터	식료품비 지출함수	엥겔계수함수
b_1	-6.115e-04	0.018	b_{19}	-4.914e-04	0.008
b_2	-3.015e-03	0.025	b_{20}	-3.826e-04	0.006
b_3	-1.350e-03	0.003	b_{21}	-2.048e-04	0.004
b_4	7.877e-05	-0.006	b_{22}	-1.201e-03	0.001
b_5	1.119e-03	-0.013	b_{23}	-2.087e-03	0.001
b_6	1.417e-03	-0.011	b_{24}	-1.698e-03	0.003
b_7	1.020e-03	-0.004	b_{25}	-1.304e-03	0.011
b_8	4.022e-04	-0.007	b_{26}	-8.520e-04	0.009
b_9	9.047e-05	-0.012	b_{27}	-2.623e-04	0.012
b_{10}	2.130e-04	-0.002	b_{28}	1.149e-04	0.018
b_{11}	1.585e-04	0.006	b_{29}	-3.239e-04	0.018
b_{12}	-8.066e-05	0.006	b_{30}	-1.204e-03	0.010
b_{13}	-7.759e-04	-0.002	b_{31}	-1.317e-03	0.008
b_{14}	-1.105e-03	-0.004	b_{32}	-5.925e-04	0.022
b_{15}	-7.426e-04	-0.001	b_{33}	-1.956e-03	0.023
b_{16}	-5.650e-04	0.004	b_{34}	-6.987e-04	0.011
b_{17}	-5.597e-04	0.004	b_{35}	-5.527e-04	0.005
b_{18}	-5.122e-04	0.007	평활파라메터	81.34	54.28

가구 내 취업자 수가 많을수록 식료품비 지출액이 줄어드는데, 이는 취업자들이 직장 내에서 식사 등을 해결하는 경우가 많음을 반영한다. 새대주의 나이가 많을수록, 교육수준이 높을수록 그리고 현재 결혼한 상태일 경우가 다른 조건이 동일할 때 더 높은 식료품비 지출경향을 보여준다. 그러나 세대주의 성별차이는 식료품비 지출액에 유의한 영향을 미치지 않는다.

마지막으로 농림축어업소득이 있는 가구가 그렇지 못한 가구에 비해 다른 조건이 동일할 때 더 낮은 식료품비 지출액을 기록하고 있다. 이는 이들 가구들의 경우 식품소비량 중 일부를 자급하는 경향이 있음을 반영하는 것일 것이다.

<표 2> 및 <표 3>과 같이 추정된 결과를 가지고 소득이 식료품비 지출액에 미치는 영향을 파악할 수 있다. 우선 무엇보다도 소득변수가 식료품비 지출액을 결정하는 데 있어 통계적으로 의미가 있는지, 그리고 영향을 미친다면 통상적인 회귀분석처럼 선형의 영향을 미치는지 아니면 식 (4)의 스플라인 회귀분석이 의미하는 바와 같이 비선형의 영향을 미치는지를 검정하여야 한다. 소득이 식료품비 지출액에 영향을 미치지 않는다는 가설과 영향을 미치되 선형으로 영향을 미친다는 가설 및 그 대립가설은 각각

다음과 같이 표현할 수 있다.

(5) $H_0^{noeffect}$: $\beta_{소득}=0$, $\sigma_b^2=0$, $H_1^{noeffect}$: $\beta_{소득}\neq 0$, $\sigma_b^2>0$

(6) H_0^{linear}: $\sigma_b^2=0$, H_1^{linear}: $\sigma_b^2>0$

모형을 최우추정법을 이용해 추정하였으므로 위와 같은 가설검정에 우선적으로 사용될 수 있는 검정법이 우도비검정(likelihood ratio test)일 것이다. 우도비가 χ^2분포를 가진다는 성질은 검정대상이 되는 추정파라메터의 값이 그 값이 분포하는 경계선에 있지는 않다는 것을 전제로 하고 있다. 그러나 σ_b^2의 분포범위는 $[0,\infty)$인데, 식 (5) 및 식 (6)의 가설자체가 σ_b^2의 하방 경계선상에서 이루어지므로 위의 두 가설검정을 위해 사용되는 우도비검정은 그 정확성에 문제가 있음이 잘 알려져 있다(Pinheiro and Bates, 2000).

식 (5) 및 식 (6)의 가설검정을 위해 우도비검정의 대안으로 사용될 수 있는 것이 F-검정법이다. R_U^2, df_U를 각각 제약이 부과되지 않은 모형의 R^2와 자유도라 하고, R_R^2와 df_R를 각각 제약이 부과된 모형의 R^2와 자유도라고 하자. $df=df_R-df_U$라 할 때 Hastie and Tibshirani(1990)에 의하면 이 경우 다음과 같이 정의되는 통계량은 귀무가설하에서 자유도 df와 df_U의 F-분포를 가진다[7].

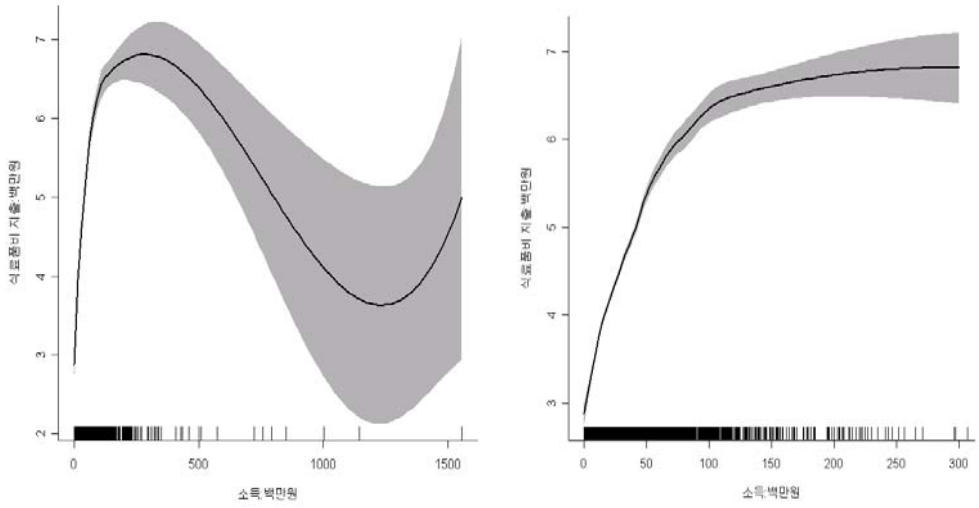
(7) $$F=\frac{R_U^2-R_R^2}{(1-R_U^2)df/df_U}$$

이상의 절차를 통해 검정하면 식 (5)의 가설을 위한 검정통계량은 348.3이고, 식 (6)의 가설 검정을 위한 검정통계량은 246.8로 계산된다. 따라서 소득이 식료품비 지출액에 영향을 미치지 않는다는 가설과 영향을 미치되 선형으로 영향을 미친다는 가설은 모두 1%의 유의수준에서도 기각된다.

---

[7] 두 모형의 자유도 차이 df는 식 (5)의 가설검정의 경우 11.76이고 식 (6)의 가설검정의 경우 10.76이다. 이 자유도의 차이 df는 변수가 어느 정도나 신축적으로 반응할 수 있도록 스플라인 회귀식이 설정되었는지를 나타낸다. 자유도가 10.76이면 소득이 식료품비 지출액에 미치는 영향을 약 10차 방정식이 근사하는 것과 유사한 정도로 스플라인 회귀식이 근사한다고 볼 수 있다. 이 자유도가 구체적으로 의미하는 바와 이를 구하는 방법에 대해서는 Ruppert et al. (2003, pp. 80-82)이 설명하고 있다.

그림 1. 가구당 소득과 식료품비 지출액



<그림 1>의 좌측 그래프는 관측된 소득자료와 식료품비 지출액의 구간 전체에서 얻어지는 두 변수 사이의 관계를 보여준다. 연간 총소득이 약 3억원일 때까지는 소득이 늘어날수록 식료품비 지출액도 늘어나지만 그 이후는 오히려 소득이 늘수록 식료품비 지출액이 감소하며, 이후 다시 증가하는 추세로 바뀐다. 그래프의 가로축 위의 음영은 각 소득수준에서의 관측치의 빈도를 보여주는데, 소득이 3억원을 넘어서는 가구의 수는 매우 적다는 것을 확인할 수 있다. 한편 그래프 주변의 회색 띠는 추정된 소비지출액의 95% 표준오차범위를 보여주는데[8], 소득이 3억원을 넘어설 경우 그 폭이 크게 확장된다. 즉 총소득이 3억원을 넘어서는 경우의 추정결과는 신뢰하기 어렵다.

이런 점을 감안하여 가구당 연간 총소득의 분포범위를 3억원 이내로 한정할 경우 <그림 1>의 우측 그래프를 얻는다. 소득이 늘어날 경우 초기에는 비교적 가파르게 식료품비 지출액이 상승하지만 그 증가속도가 둔화되며, 총소득이 연간 1억원 이상에 다다르면 소득상승이 식료품비 지출액 상승에 거의 영향을 미치지 못하는 것으로 나타난다. 즉 식료품비 지출액은 저소득층에서의 소득증가 시 상대적으로 빨리 늘어나지만 중산층 이상이라 할 수 있는 계층에서의 소득증가에는 민감하게 반응하지 않는 것으로 나타난다. 소득이 약 1,000만원 정도일 경우 소득증가액의 약 18%가 식료품지출액

[8] 추정치의 표준오차를 구하는 공식은 Ruppert et al.(2003)이나 Hastie and Tibshirani(1990)에서 얻을 수 있다.

증가로 연결되어 식료품비 지출액이 소득변화에 가장 민감하게 반응하는 것으로 나타났다.

<그림 1>의 추정결과는 식료품비 지출액은 소득증가에 따라 지속적으로 늘어나는 것이 아니라 일종의 상한이 존재함을 보여준다[9]. 소득과 식료품비 지출액 사이의 이러한 비선형, 비대칭 관계는 통상적인 선형회귀분석이나 2차방정식 형태의 회귀분석에서는 파악될 수가 없으며, 이런 점에서 본고가 사용하는 스플라인 회귀분석의 장점이 인정된다고 할 것이다.

<표 2>와 <표 3>은 가구 총지출에서 식료품비 지출액이 차지하는 비중, 즉 엥겔계수가 가구의 특성에 따라 어떻게 달라지는지를 추정한 결과도 보여주고 있다. 소득을 제외한 다른 가구특성이 엥겔계수에 미치는 영향을 보면, 먼저 지역별 차이의 경우 가구별 식료품 소비지출액의 지역별 차이와는 상당히 다른 양상을 보여주는데, 이는 식료품비 지출액은 물론 가구별 총지출액 자체가 지역별로 차이를 보이기 때문이다. 다른 조건이 동일하다면 부산의 엥겔계수 즉 식료품비지출액/총지출액이 서울보다도 더 높고, 다른 지역의 엥겔계수는 서울보다도 통계적으로 유의한 정도로 더 낮다.

가구원의 수가 많을 경우 엥겔계수가 높고, 무직이나 근로자 가구에 비해 자영업 가구의 엥겔계수가 더 낮다. 취업자의 수가 많고 세대주의 연령이 높으며, 남자일 경우 엥겔계수가 높으며, 세대주의 학력이 낮을수록 엥겔계수가 더 높다. 그러나 세대주가 현재 결혼했는지의 여부와 농림축어업소득이 있는지의 여부는 엥겔계수의 크기에 유의한 영향을 미치지 않는다.

가구당 소득이 엥겔계수에 미치는 영향은 <표 3> 및 <그림 2>와 같이 분석된다. 먼저 엥겔계수에 소득이 영향을 미치는지, 그리고 영향을 미친다면 선형의 영향인지 아니면 비선형의 영향인지 등을 식 (5) 및 식 (6)과 같이 가설검정하면 그 검정통계량이 각각 599.5와 447.2가 되어 가계소득은 엥겔계수 크기에 영향을 미치며, 아울러 비선형의 영향을 미친다고 할 수가 있다.

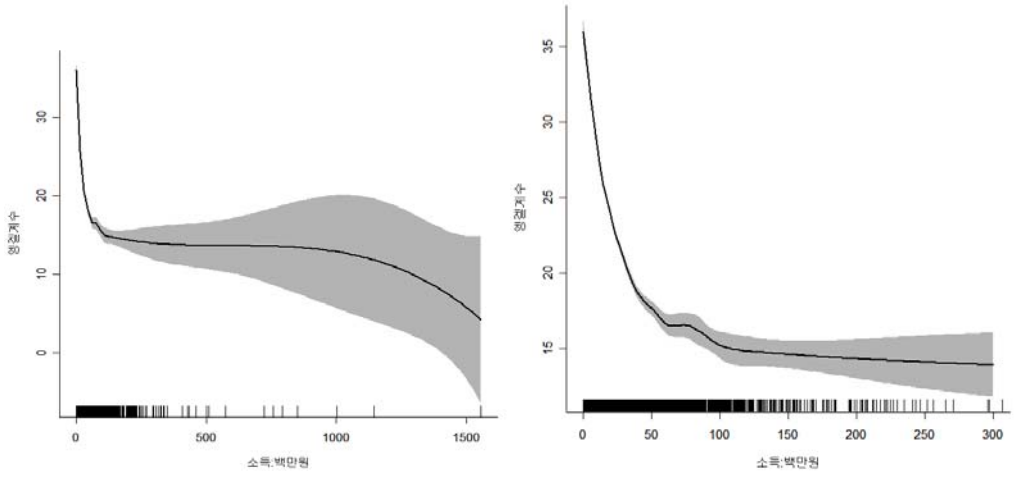
가계소득이 엥겔계수에 영향을 미치는 정도는 <그림 2>가 보여주고 있다. 경제이론이 일반적으로 예측하는 바와 같이 가구의 소득수준과 총지출액에서 식료품비가 차지하는 비중의 관계는 음의 관계이다. 그러나 그 관계는 선형관계와는 거리가 멀며, 소득수준이 약 5,000만원이 될 때까지는 소득이 늘면서 엥겔계수가 급속히 하락하지만, 이후 완만히 하락하며, 1억원 이상의 소득을 얻을 경우 소득수준이 엥겔계수에 거의 영향

---

[9] 식료품 소비지출액이 소득 외의 여타 사회경제적 특성에 의해서도 영향을 받기 때문에 이 상한은 가구별로 서로 다를 수 있다.

을 미치지 못한다. 그뿐만 아니라 연간 소득 7,000~8,000만원 수준에서도 소득수준과 관련 없이 일정한 수준의 엥겔계수가 유지되다가 8,000만원 이상의 소득이 발생하면 다시 엥겔계수가 하락하는 현상도 보여준다.

그림 2. 가구당 소득과 엥겔계수



5. 요약 및 결론

본고는 준모수적 회귀분석기법을 이용해 소득을 포함하는 가구별 사회경제적 특성이 식료품비 지출에 어떤 영향을 미치는지를 「가구소비실태조사」 자료를 이용해 분석하였다. 분석결과 소득은 식료품비 지출액과 비선형의 관계를 가지며, 소득이 낮을 경우 높은 식료품비 지출성향을 보이지만 약 1억원 이상의 소득수준에서는 소득증가가 식료품비 지출액 증가에 거의 영향을 미치지 못하는 것으로 나타났다. 엥겔계수의 경우 소득수준과 역시 비선형의 관계를 가지며, 소득이 늘수록 그 값이 하락하는 경향이 있음을 보여주었다. 그러나 엥겔계수의 경우 가구유형별로 일종의 하한값이 있어 소득이 증가한다고 해서 무한정 하락하지는 않을 것이라는 점도 발견되었다.

식료품비 지출액은 지역별로도 차이가 있으며, 가구원의 수, 직업유형, 취업자 수, 세대주의 나이, 교육수준, 결혼여부, 그리고 농림축어업소득이 있는지의 여부 등에 의해서도 영향을 받는 것으로 나타났다.

본고가 제시하는 이상의 추정결과는 소득증가, 인구의 수도권 집중, 핵가족화 등의 사회경제적 환경변화가 식품소비 지출액에 어떤 영향을 미칠지를 예측하는데 있어 나름대로 기여하리라 기대한다.

본고는 특히 소득과 식료품 소비지출액 사이에는 단순한 선형이나 로그선형과 같은 관계가 존재하는 것이 아니라 보다 신축적인 비선형의 함수관계가 존재하고, 식품소비 지출액의 가구유형별 상한도 존재할 수 있음을 보여주는 등, 준모수적 추정법이 기존의 회귀분석에 비해 보다 유용하게 사용될 수 있음을 보여주었다.

그러나 본고가 사용한 「가구소비실태조사」 자료는 식료품비 지출액을 품목별로 세분하여 집계하지 않았기 때문에 본고는 식료품비 지출액 전체가 가구특성에 따라 변하는 것만 보여줄 수 있었다는 한계를 가진다. 보다 최근에 집계된 「가계조사」 등의 자료를 활용하여 품목별로 보다 세분된 식품소비 지출함수에 대한 준모수적 혹은 비모수적 분석을 시도하는 것도 유용하리라 생각한다.

참고 문헌

김태균, 사공용. 1994. “한국의 육류수요분석에 있어서 모형의 적합성 검정: AIDS모형과 로테르담모형.”「농업경제연구」 35(2): 17-30.

사공용, 김태균. 1994. “소비의 구조적 변화와 수요함수 추정: 한국의 곡류와 미국의 육류소비를 중심으로.”「농촌경제」 17(3): 13-23.

사공용, 최지현. 1995. “소득증가에 따른 식품소비 변화분석.”「농업경제연구」 36(1): 47-62.

서종석. 1994. “Taste Change in Meat Demand.”「농업경제연구」 35(2): 51-65.

이계임, 김민정. 2003.「쌀 소비행태 분석」. 한국농촌경제연구원.

이계임, 김성용. 2003. “수산물 소비구조 분석.”「농촌경제」 26(3): 21-38.

이계임 등. 2003.「수산물 수급실태 분석과 중장기 전망에 관한 연구」. 해양수산부.

이계임, 최지현, 박준기. 1998.「과실류 소비행태에 관한 연구」. 한국농촌경제연구원.

이계임 등. 1999.「육류 소비구조의 변화와 전망」. 한국농촌경제연구원.

정경수, 박창원. 1998. “한국의 육류 수요분석.”「농업경제연구」 39(2): 63-78.

조덕래, 김영식. 1985. “준이상수요체계에 의한 동물성식품 수요분석.” 「농촌경제」 12(1): 123-134.

Brumback, B., A., D. Ruppert, and M. P. Wand. 1999. “Comment on Variable Selection and Function Estimation in Additive Nonparametric Regression Using Data-based Prior by Shively, Kohn, and Wood.” *Journal of the American Statistical Association* 94: 794-797.

Hastie T. 2006. gam: *Generalized Additive Models*. R package version 0.98.

Hastie, T. J. and R. J. Tibshirani. 1990. *Generalized Additive Models*, Chapman & Hall.

Hayfield T. and J. S. Racine. 2006. *np: Nonparametric Kernel Smoothing Methods for Mixed Datatypes*. R package version 0.12-1.

Li, Q and J. S. Racine. 2003. "Nonparametric Estimation of Distribution with Categorical and Continuous Data." *Journal of Multivariate Analysis* 86: 266-292.

Li, Q and J. S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

McCulloch, C. E. and S. R. Searle. 2001. *Generalized, Linear, and Mixed Models*, Wiley.

Pagan, A. and A. Ullah. 1999. *Nonparametric Econometrics*, Cambridge University Press.

Pinheiro, J. C. and D. M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*, Springer.

Ruppert, D., M. P. Wand and R. J. Carroll. 2003. *Semiparametric Regression*, Cambridge University Press.

Wand, M. P. et al. 2005. *SemiPar 1.0. R package.*

원고 접수일: 2007년 10월 1일
원고 심사일: 2007년 10월 22일
심사 완료일: 2008년 1월 14일