R을 이용한 패널자료분석 : OECD 국가의 자동차 휘발유 소비량 패널모형에 적용

박 범 조*

요 약

본 연구는 계량경제학에서 매우 중요한 연구주제로 다루어지고 있는 패널자료분석과 관련된 기존 문헌의 서베이를 통해 고정효과모형(fixed effect model)이나 확률효과모형(random effect model)과 같은 선형패널모형과 추정방법을 소개하고 선형패널모형에서 고정효과를 검정할 수 있는 F-검정법과 확률효과를 검정할 수 있는 LM 검정법과 Hausman 검정법에 대해 설명한다.

그리고 경제분야의 전문가들에게는 잘 알려져 있지 않지만 전문성과 사용의 편리성이라는 기준에 의해 패널자료분석에 가장 적합한 소프트웨어 중 하나인 R 소프트웨어의 plm (Croissant and Millo, 2008) 패키지를 소개하고 Baltagi and Griffin(1983)의 자동차 휘발유 소비량에 대한 패널회귀모형을 최신 자료를 이용하여 실증적으로 분석해봄으로써 응용경제분야의 전문가들에게 R 소프트웨어를 이용한 패널자료분석의 이해와 활용성을 향상시키고자 한다.

핵심주제어 : 패널자료분석, R 소프트웨어, 고정효과모형, 확률효과모형, LM 검정, Hausman 검정.

Ⅰ. 서 론

패널자료분석(panel data analysis)은 계량경제학에서 매우 중요한 분야로서 지속적인 발전을 해오고 있다.[1] 특정 변수를 개체(units)별로 시간의 흐름에 따라 여러 시점에서 관측한 종

* 단국대학교 상경대학 경제학과 교수, E-mail: bjpark@dankook.ac.kr
〈논문 투고일〉 2012-02-02 〈논문 수정일〉 2012-02-29 〈게재 확정일〉 2012-03-13

적자료(longitudinal data)의 한 형태인 패널자료에 대한 분석은 순수 시계열자료분석이나 횡단면자료분석보다 많은 관측값을 이용할 수 있기 때문에 효율적인 추정량을 제공해주며 설명변수 사이의 다중공선성을 완화해주고 개인이나 가계, 기업 등과 같은 미시적 단위별 자료수집과정에서 발생할 수 있는 자료 편의(data bias)를 줄여준다는 통계적인 잇점을 갖는다.

최근에는 시계열자료분석과 횡단면자료분석의 장점을 극대화할 수 있는 계량모형의 발전으로 인해 패널자료분석의 추가적인 장점이 더욱 강조되고 있다. Hsiao(2003, 2007), Baltagi (2001) 등에 따르면 우선 패널자료분석은 개체들의 관찰되지 않는 이질성(heterogeneity)과 시간효과(time effect)를 고려할 수 있다. 그리고 시간의 흐름에 무관한 횡단면자료분석에서는 분석이 어려운 변수들의 동태적 관계(dynamic relationship)를 분석할 수 있으며 시계열자료분석이나 횡단면자료분석에서 다룰 수 없는 복잡한 행태 모형을 설정하거나 검정할 수 있다.

이런 패널자료분석이 가능할 수 있도록 사회과학 분야에서 다양한 패널자료들이 제공되고 있으며 경제학 분야의 대표적인 패널자료로는 여러 개인그룹의 노동시장활동 및 생활사건(life events)에 대한 정보를 수집하기 위해 1966년부터 조사가 시작된 미국의 National Longitudinal Surveys(NLS)(http://www.bls.gov/nls)과 미국의 미시간 대학 사회연구소에서 1968년부터 미국 전역에서 추출한 6,000가구와 개인 15,000명에 대한 고용, 소득, 인적자본 변수 등의 다양한 경제정보를 매년 수집해온 Panel Study of Income Dynamics(PSID) (http://psidonline.isr.umich.edu) 등이 있다. 우리나라의 경우에는 패널조사의 기간이 길지 않지만 1993년부터 1998년까지 대우경제연구소에서 PSID 패널자료와 유사하게 가구 및 개인의 소득, 소비실태, 주택에 관한 변수들에 대한 정보를 수집한 한국가구패널조사(Korea Household Panel Study : KHPS)가 있으며 1998년 이후에는 한국노동연구원에서 한국노동패널조사(KLIPS : Korean Labor and Income Panel Study) (http://www.kli.re.kr/klips/ko/main/main.jsp)를 수행하고 있다.

본 연구는 패널자료분석을 위한 선형패널모형과 추론 방법에 대해 설명하면서 통계분석을 위해 폭 넓게 활용되고 있는 R 소프트웨어[2](http://www.r-project.org)와 패널분석을 위한

1) 패널자료분석과 관련된 기존 연구에 대해 자세한 내용을 알고자 하는 독자는 Chamberlain(1984), Baltagi(2001), Wooldridge(2002), Hsiao(2003, 2007) 등을 참고하기 바람.

2) R 소프트웨어는 뉴질랜드 Auckland 대학의 Ross Ihaka와 Robert Gentleman에 의해 통계계산과 그래프 분석을 위해 개발된 프로그래밍 언어로 효율적인 자료 처리와 저장, 행렬 및 배열 형태의 연산, 자료분석을 위한 유용한 도구 모임의 제공, 그래프 분석을 위한 편리성 등의 장점을 갖는다. 특히 R 소프트웨어는 벨 연구소(Bell Lab.)의 Becker, Chambers, and Wilks(1988)에 의해 개발되어 통계학 분야에서 광범위하게 사용되어 왔던 S 언어와 유사한 언어구조를 갖기 때문에 통계분석과 관련된 매우 다양한 패키지들이 제공되어 그 활용성이 매우 높다.

패키지 plm(Croissant and Millo, 2008)을 이용하여 Baltagi and Griffin(1983)의 자동차 휘발유 소비량에 대한 패널회귀모형을 추정하고 검정해봄으로써 R 소프트웨어에 익숙하지 않은 응용계량경제학자와 응용경제 분야의 전문가들에게 R 소프트웨어를 이용한 패널자료분석의 이해와 활용성을 증진시키고자 한다.[3)]

II. 선형패널모형

순수 시계열과정과 개별단위의 횡단면자료에서 발생할 수 있는 오차항을 시간효과와 개별효과로 통제할 수 있는 일반적인 선형패널모형은 다음과 같이 간단하게 표기할 수 있다.

$$y_{it} = x_{it}\beta + \varepsilon_{it},\ \ i = 1,\cdots,N; t = 1,\cdots,T \tag{2.1}$$

여기서 x_{it}는 상수항을 포함하는 설명변수 행렬이며 $\varepsilon_{it} = \delta_i + u_{it}$로 가정한다. 즉, 오차항은 개별특성을 나타내는 변수 δ_i와 평균이 0이며 고정 분산을 갖는 독립적 확률오차항 u_{it}으로 구성된다. 만일 개체불변 시간특성변수 η_t를 오차항에 추가로 포함한다면 오차항은 $\varepsilon_{it} = \delta_i + \eta_t + u_{it}$이 되며(Balestra and Nerlove, 1966), 이 모형을 양방향(two-ways) 패널자료모형이라고도 한다. 이 경우 δ_i는 절편에 존재하는 개체별 차이를 나타내며 절편계수(β_0)와 δ_i의 합은 개별효과를 나타내고 η_t는 시간 흐름에 따른 절편의 차이를 나타내며 절편계수(β_0)와 η_t의 합은 시간효과를 나타낸다.

1. 고정효과모형(fixed effect model)

식 (2.1)에서 개별특성을 나타내는 변수 δ_i는 시간의 흐름에 관계없이 고정되어 비확률적이지만 개체별로 다를 수 있다고 가정하면 고정효과모형은 다음과 같이 표현될 수 있다.

$$y_{it} = x_{it}\beta + \delta_i + u_{it},\ \ i = 1,\cdots,N; t = 1,\cdots,T \tag{2.2}$$

이 고정효과모형을 (2.1)의 회귀식을 이용하여 일반최소제곱(OLS) 추정량으로 추정하는 경

3) 경제학 분야의 전문가들은 패널자료분석을 위해 LIMDEP, EViews, SAS, SHAZAM, STATA 등과 같이 사용하기 쉬운 소프트웨어를 일반적으로 이용하고 있다.

우 δ_i를 포함하게 되는 오차항 ε_{it}이 설명변수와 상관될 가능성이 높아 오차항이 설명변수와 독립이라는 기본가정이 위배되어 일반최소제곱 추정치는 편의를 가질 뿐만 아니라 일치추정치가 아닌 문제가 발생하게 된다. 따라서 이 모형을 추정하기 위하여 최소제곱더미변수(least squares dummy variable: LSDV)모형과 그룹내(within group)모형이 널리 활용된다.

❒ LSDV 모형

고정효과모형을 추정하기 위해 개별특성을 반영한 N-1개의 더미변수를 추가한 다음의 LSDV 모형으로 전환한다.

$$y_{it} = x_{it}\beta + \sum_{h=2}^{N} \gamma_k D_{it}^h + u_{it},\ i = 1,\cdots,N; t = 1,\cdots,T \tag{2.3}$$

여기서 D_{it}^h (만일 $h = i$이면 $D_{it}^h = 1$, 아니면 $D_{it}^h = 0$)는 개체 h의 특성을 고려하기 위한 더미변수이다.

하지만 이 LSDV 모형은 개체 수가 많은 경우 N-1개의 더미변수 도입으로 인해 자유도 문제가 발생하거나 다중공선성의 가능성이 존재한다. 또한 성별, 인종과 같은 시간불변변수가 모형에 설명변수로 포함되는 경우 시간불변변수가 종속변수에 미치는 영향을 탐지하지 못하는 문제가 있다.

❒ 그룹내(within group)모형

한편 고정효과모형을 추정하기 위한 보다 일반적인 방법은 변수를 시간에 대한 각 개체의 평균값으로부터의 편차로 구한 그룹내모형을 고려하는 것이다.

$$y_{it} - y_i^m = (x_{it} - x_i^m)\beta + (u_{it} - u_i^m),\ i = 1,\cdots,N \tag{2.4}$$

여기서 $y_i^m = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $x_i^m = \frac{1}{T}\sum_{t=1}^{T} x_{it}$, $u_i^m = \frac{1}{T}\sum_{t=1}^{T} u_{it}$이며 이와 같이 변수를 시간에 대한 각 개체의 평균값으로부터의 편차로 변형하면 개별특성을 나타내는 δ_i가 사라지게 된다. 이 그룹내모형은 LSDV모형과 달리 더미변수 도입으로 인한 자유도 문제나 다중공선성의 발생 가능성이 완화되지만 여전히 시간불변변수의 영향을 탐지할 수 없다. 그룹내모형을 최소제곱추정을 하게 되면 다음과 같은 고정효과 추정량을 얻게 되며 독립변수의 강외생성(strict

exogeneity)($E[(x_{it}-x_i^m)'u_{it}]=0$) 가정하에 고정효과 추정량은 일치추정량이 된다.

$$\beta_F=\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-x_i^m)'(x_{it}-x_i^m)\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-x_i^m)'(y_{it}-y_i^m)\right) \tag{2.5}$$

2. 고정효과 검정

수집한 자료에서 특정 개체 h에 대한 개별효과를 알아보기 위해서는 LSDV모형에서 표준 유의수준을 이용하여 계수 γ_k에 대한 유의성 검정을 수행해볼 수 있다. 하지만 모든 개체별 개별효과가 반영된 고정효과를 탐지하기 위해서는 δ_i가 모두 동일하게 0이라는(즉, 그룹간 상수항이 모두 동일함) 귀무가설을 설정하고 이에 대한 F-검정을 수행할 수 있다.

$$F_{(N-1,NT-N-K)}=\frac{(RSS_P-RSS_F)/(N-1)}{RSS_F/(NT-N-K)} \tag{2.6}$$

여기서 RSS_F는 고정효과모형의 잔차제곱의 합, RSS_P는 합동(pooled)모형의 잔차제곱의 합, 그리고 K는 모형에 포함된 설명변수의 수를 나타낸다.

3. 확률효과모형(random effect model)

확률효과모형은 고정효과모형과는 다르게 식 (2.1)에서 개별특성을 나타내는 변수 δ_i가 확률적이며 다음과 같은 분포를 갖는다고 가정한다.

$$\delta_i \sim i.i.d.N(0,\sigma_\delta^2), \qquad i=1,\cdots,N \tag{2.7}$$

추가적으로 δ_i는 u_{it}와 독립이라고 가정한다. 즉, 이 모형의 오차항 $\varepsilon_{it}=\delta_i+u_{it}$은 다음 구조를 갖는다.

$$E(\varepsilon_{it})=E(\delta_i+u_{it})=0 \tag{2.8}$$

$$E(\varepsilon_{it}^2)=E\left((\delta_i+u_{it})^2\right)=\sigma_\delta^2+\sigma_u^2 \tag{2.9}$$

$$E(\varepsilon_{it}\varepsilon_{il})=E((\delta_i+u_{it})(\delta_i+u_{il}))=\sigma_\delta^2 \ \text{ for } t\neq l \tag{2.10}$$

$$E(\varepsilon_{it}\varepsilon_{jt})=E((\delta_i+u_{it})(\delta_j+u_{jt}))=0 \ \text{ for } i\neq j \tag{2.11}$$

따라서 개체 i에 대한 오차항 $\varepsilon_i = [\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{iT}]'$의 공분산 행렬($T \times T$)은

$$\Omega = \begin{bmatrix} \sigma_\delta^2 + \sigma_u^2 & \sigma_\delta^2 & \cdots & \sigma_\delta^2 \\ \sigma_\delta^2 & \sigma_\delta^2 + \sigma_u^2 & \cdots & \sigma_\delta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\delta^2 & \sigma_\delta^2 & \cdots & \sigma_\delta^2 + \sigma_u^2 \end{bmatrix} = \sigma_\delta^2 + \sigma_u^2 I_T \quad (2.12)$$

이며 I_T는 $T \times T$ 항등행렬이다. 그리고 NT 원소를 갖는 오차항 벡터 $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N]$를 정의하면 $E(\boldsymbol{\varepsilon}) = 0, E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \Omega \otimes I_N$이 된다. 여기서 $\otimes$는 크로네커 곱(Kronecker product), I_N는 $N \times N$ 항등행렬을 나타낸다.

확률효과모형은 다음의 일반최소제곱(generalized least squares : GLS)에 의해 추정될 수 있다.

$$\beta_R = [x'(\Omega^{-1} \otimes I_N)x]^{-1} x'(\Omega^{-1} \otimes I_N)y \quad (2.13)$$

여기서 $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$, $y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$, $x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iT} \end{bmatrix}$이다.

4. 확률효과 검정

분석하려는 패널자료에 실제로 확률효과가 존재하여 확률효과모형이 일반 회귀모형과 통계적 차이를 나타내는지, 고정효과모형과 확률효과모형 사이에 유의미한 차이가 있는지 등을 검증하기 위해 Breusch and Pagan(1980)의 라그랑지 승수(Lagrange multiplier) 검정과 Hausman (1978) 검정 등을 수행할 수 있다.

❒ LM 검정

Breusch and Pagan(1980)은 합동모형의 잔차를 이용하여 확률효과 존재여부를 판단할 수 있는 LM 검정 통계량을 제안하였다. 즉, 확률효과가 없다는 귀무가설인 $H_0 : \sigma_\delta^2 = 0$[4]을 검정하기 위한 통계량은

4) 합동모형의 오차항들 간에 상관관계가 없다는 귀무가설과 동일함

$$LM = \frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\varepsilon_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it}^2} - 1\right]^2 \quad (2.14)$$

이며, 이 LM 검정 통계량은 자유도 1인 카이제곱 분포를 하게 된다.

❒ Hausman 검정

추가적으로 확률효과모형의 타당성여부를 판단하기 위해 Hausman(1978)의 모형설정 검정 방법을 적용할 수 있다. 이 검정의 기본적인 아이디어는 다음과 같다. 확률효과가 하나 이상의 설명변수들과 상관관계가 없다는 가정하에서 고정효과모형과 확률효과모형에 대한 추정량이 모두 일치성을 만족하지만 이 가정이 위배되면 고정효과모형에 대한 OLS 추정량은 일치성을 만족하는데 반해 확률효과모형에 대한 GLS 추정량은 일치성을 만족하지 못하게 되어 고정효과모형과 확률효과모형에 대한 추정치는 구조적으로 다르게 된다. 따라서 고정효과모형과 확률효과모형이 동일하다는 귀무가설을 설정하여 점근적 카이제곱 분포를 하는 다음 Hausman(1978)의 검정 통계량(HTS)을 계산하고, 만일 귀무가설이 기각되면 확률효과가 하나 이상의 설명변수들과 상관될 가능성이 높아 기각되는 것을 의미하기 때문에 확률효과모형이 타당하지 못하다고 해석할 수 있다.

$$HTS = (\beta_R - \beta_F)'[Var(\beta_R) - Var(\beta_F)]^{-1}(\beta_R - \beta_F) \quad (2.15)$$

Ⅲ. R을 이용한 선형패널모형 추정 및 검정

본 연구에서는 R 소프트웨어를 이용하여 실증분석을 수행하기 위해 Baltagi and Griffin (1983)이 OECD 국가의 자동차 한대당 휘발유 소비에 대한 분석에 사용했었던 패널자료와 모형을 최근 9년 동안의 자료로 수정하여 사용할 것이다. 구체적으로 살펴보면, 처음 이 분석에 사용되었던 자료는 1960년부터 1978년까지 OECD 18개 국가에 대한 데이터를 담고 있었으나, 새롭게 수정된 자료는 OECD 22개 국가에 대한 2001년부터 2009년까지의 데이터를 담고 있다. 회귀모형은 다음과 같다.

$$\ln G_{it} = \beta_0 + \beta_1 \ln I_{it} + \beta_2 \ln P_{it} + \beta_3 \ln C_{it} + \varepsilon_{it},\ i = 1,\cdots,22;\ t = 1,\cdots,9 \qquad (2.16)$$

여기서 G_{it}는 자동차 한 대당 석유소비량, I_{it}는 1인당 실질소득, P_{it}는 실질 석유가격, C_{it}는 1인당 자동차 수를 각각 나타낸다.[5)]

이 모형을 추정하기 위해 R 소프트웨어의 plm 패키지를 이용한다. R 소프트웨어의 plm 패키지는 R 소프트웨어 메뉴에서 package → package install을 차례로 선택하여 CRAN mirror 창이 나타나면 다운 받기 원하는 국가를 선택하고 Packages 창에서 plm을 선택하여 설치할 수 있다. 패키지 설치가 되면 R 소프트웨어에서 library("plm") 명령문에 의해 패키지를 사용할 수 있는데, Baltagi and Griffin(1983)이 분석에 사용했던 1960년부터 1978년까지의 연간 패널자료는 Console 1과 같이 입력하여 불러올 수 있다. 즉 Console 1의 (1)과 (2)는 plm 패키지를 불러온 후 패키지에 포함된 "Gasoline"패널자료를 로딩하는 명령문이며, (3)은 "Gasoline"패널자료의 변수 이름과 구조를 알아보기 위한 R 소프트웨어의 명령문이다.

▌Console 1

```
〉library("plm")                                                    (1)
〉data(Gasoline, package = "plm")                                  (2)
〉head(Gasoline)                                                   (3)
  country  year lgaspcar lincomep    lrpmg   lcarpcap
1 AUSTRIA 1960  4.173244 -6.474277 -0.3345476 -9.766840
2 AUSTRIA 1961  4.100989 -6.426006 -0.3513276 -9.608622
3 AUSTRIA 1962  4.073177 -6.407308 -0.3795177 -9.457257
4 AUSTRIA 1963  4.059509 -6.370679 -0.4142514 -9.343155
5 AUSTRIA 1964  4.037689 -6.322247 -0.4453354 -9.237739
6 AUSTRIA 1965  4.033983 -6.294668 -0.4970607 -9.123903
```

Console 2에서 (1)은 엑셀 파일의 자료를 'New.Gasoline'이라는 객체에 할당하여 사용할

5) 본 연구에 사용된 종속변수와 독립변수의 출처는 각각 다음과 같다.
종속변수 : 연간 가솔린 소비량 / 당해 연도 자동차 보유 대수로 산출.
- 연간 가솔린 소비량 : U.S. Energy Information Administration(http://www.eia.gov)
- 당해 연도 자동차 보유 대수: 세계자동차통계연보, 한국자동차공업협회(http://www.kama.or.kr)
독립변수
- GDP per capita : OECD (http://www.oecd.org)
- Cars per capita: 세계자동차통계연보, 한국자동차공업협회(http://www.kama.or.kr)
- Price of Motor Gasoline: 국제에너지기구(IEA) (http://www.iea.org)

수 있으며, (2)는 Console 1의 (3)과 마찬가지로 변수 이름과 구조를 대략적으로 알아보기 위한 명령문이다. Console 2에 입력된 명령문의 실행 결과는 다음과 같다.

▌Console 2

```
〉New.Gasoline〈-read.table("clipboard",header="T")                    (1)
〉head(New.Gasoline)                                                  (2)
  country  year lgaspcar lincomep      lrpmg      lcarpcap
1 Austria 2001 1.391594 10.07174 -0.211956362 -0.6540080
2 Austria 2002 1.514230 10.14929 -0.193584749 -0.7067798
3 Austria 2003 1.520959 10.34512 -0.005012542 -0.6942792
4 Austria 2004 1.488273 10.47523  0.163818085 -0.6878329
5 Austria 2005 1.440835 10.51575  0.249980205 -0.6834551
6 Austria 2006 1.394473 10.57190  0.313349819 -0.6778688
```

명령문을 실행한 결과 나타난 변수이름은 각각 lgaspcar=$\ln G_{it}$, lincomep=$\ln I_{it}$, lrpmg=$\ln P_{it}$, lcarpcap=$\ln C_{it}$을 의미하며, 이 변수들의 기본통계량은 summary 명령어를 이용하여 다음과 같이 계산할 수 있다.

▌Console 3

```
〉 summary(New.Gasoline)
     country          year         lgaspcar        lincomep
 Austria : 9   Min.   : 2001   Min.   : 0.6039   Min.   : 8.275
 Belgium : 9   1st Qu. : 2003   1st Qu. : 1.3461   1st Qu. : 9.808
 Czech : 9   Median : 2005   Median : 1.5548   Median : 10.361
 Denmark : 9   Mean   : 2005   Mean   : 1.6802   Mean   : 10.187
 Finland : 9   3rd Qu. : 2007   3rd Qu. : 1.9647   3rd Qu. : 10.618
 France : 9   Max.   : 2009   Max.   : 3.2250   Max.   : 11.691

      lrpmg            lcarpcap
 Min.   : -0.96758   Min.   : -1.6729
 1st Qu. : -0.01106   1st Qu. : -0.9704
 Median : 0.27041   Median : -0.7971
 Mean   : 0.22019   Mean   : -0.8740
 3rd Qu. : 0.48042   3rd Qu. : -0.7200
 Max.   : 0.80960   Max.   : -0.3955
```

다음 Console 4, 5, 6은 식 (2.16)의 회귀모형을 합동(pooling)모형, 고정효과모형의 그룹내모형, 확률효과모형으로 설정하여 추정한 결과를 각각 보여준다.

▌Console 4

```
〉 pd 〈- plm(lgaspcar ~ lincomep + lrpmg + lcarpcap, data = New.Gasoline, model = "pooling")
〉 summary(pd)
Oneway (individual) effect Pooling Model
Call:
plm(formula = lgaspcar ~ lincomep + lrpmg + lcarpcap, data = New.Gasoline,
    model = "pooling")
Balanced Panel: n=22, T=9, N=198
Residuals :
    Min. 1st Qu.  Median 3rd Qu.     Max.
-0.8540 -0.1630  0.0491  0.1790  0.6680
Coefficients :
             Estimate Std. Error t-value  Pr(〉|t|)
(Intercept) -8.204410   0.546316 -15.018 〈 2.2e-16 ***
lincomep     0.896568   0.046768  19.171 〈 2.2e-16 ***
lrpmg       -1.217263   0.063797 -19.080 〈 2.2e-16 ***
lcarpcap    -1.165840   0.113845 -10.241 〈 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Total Sum of Squares:    59.676
Residual Sum of Squares: 15.611
R-Squared      :  0.7384
     Adj. R-Squared :  0.72348
F-statistic: 182.525 on 3 and 194 DF, p-value: 〈 2.22e-16
```

▌Console 5

```
〉 fe 〈- plm(lgaspcar ~ lincomep + lrpmg + lcarpcap, data = New.Gasoline, model = "within")
〉 summary(fe)
Oneway (individual) effect Within Model
Call:
```

```
plm(formula = lgaspcar ~ lincomep + lrpmg + lcarpcap, data = New.Gasoline,
    model = "within")
Balanced Panel: n=22, T=9, N=198
Residuals :
    Min. 1st Qu.  Median 3rd Qu.     Max.
-0.3700 -0.0425  0.0121  0.0470  0.1600
Coefficients :
            Estimate Std. Error t-value  Pr(>|t|)
lincomep     0.230822   0.075935  3.0397  0.002736 **
lrpmg       -0.491354   0.070754 -6.9445 7.364e-11 ***
lcarpcap    -0.625721   0.102404 -6.1103 6.395e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Total Sum of Squares:    3.4395
Residual Sum of Squares: 1.1994
R-Squared      :  0.65127
     Adj. R-Squared :  0.56904
F-statistic: 107.696 on 3 and 173 DF, p-value: < 2.22e-16
```

▌Console 6

```
re <- plm(lgaspcar ~ lincomep + lrpmg + lcarpcap, data = New.Gasoline, model =
"random")
> summary(re)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)
Call:
plm(formula = lgaspcar ~ lincomep + lrpmg + lcarpcap, data = New.Gasoline,
    model = "random")
Balanced Panel: n=22, T=9, N=198
Effects:
                     var  std.dev share
idiosyncratic 0.006933 0.083266 0.083
individual    0.076461 0.276516 0.917
theta:  0.9001
Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.42100 -0.04970  0.00373  0.05030  0.29300
```

```
Coefficients :
                Estimate Std. Error  t-value  Pr(>|t|)
(Intercept)   -3.098308   0.719107   -4.3085  2.609e-05 ***
lincomep       0.425508   0.067190    6.3329  1.638e-09 ***
lrpmg         -0.671971   0.062468  -10.7571  < 2.2e-16 ***
lcarpcap      -0.676953   0.103857   -6.5181  6.003e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Total Sum of Squares:    4.0004
Residual Sum of Squares: 1.5015
R-Squared      :  0.62467
     Adj. R-Squared :  0.61205
F-statistic: 107.625 on 3 and 194 DF, p-value: < 2.22e-16
```

이들 모형의 추정결과를 보면 합동모형을 이용한 계수 추정치가 그룹내모형이나 확률효과 모형의 추정치와 상당히 다르게 추정됨을 알 수 있다. 이는 분석자료에 고정효과나 확률효과가 존재하고 있음을 암시한다고 판단할 수 있다. 따라서 앞 절에서 소개한 고정효과와 확률효과 검정을 R 소프트웨어의 plm 패키지를 이용하여 수행하고자 한다. 우선 고정효과 검정을 위한 F-검정 통계량은 다음과 같이 pooltest 명령어를 사용하여 수행할 수 있다. F-검정 통계량의 값이 12.3706으로 1%의 유의수준에서 δ_i가 모두 동일하게 0이라는 귀무가설을 기각함으로써 "New.Gasoline"패널자료에 통계적으로 유의미한 고정효과가 존재한다고 판단할 수 있다.

▌Console 7

```
> pooltest(lgaspcar ~ lincomep + lrpmg + lcarpcap, data = New.Gasoline, model = "within")
        F statistic
data:  lgaspcar ~ lincomep + lrpmg + lcarpcap
F = 12.3706, df1 = 63, df2 = 110, p-value < 2.2e-16
alternative hypothesis: unstability
```

또한 확률효과를 검정하기 위한 LM 검정 통계량과 Hausman의 검정 통계량 추정결과는 다음과 같다.

▌Console 8

```
〉 plmtest(pd, effect = "individual", type = "bp")
        Lagrange Multiplier Test - (Breusch-Pagan)
data:  lgaspcar ~ lincomep + lrpmg + lcarpcap
chisq = 558.6188, df = 1, p-value 〈 2.2e-16
alternative hypothesis: significant effects
〉 phtest(fe, re)
        Hausman Test
data:  lgaspcar ~ lincomep + lrpmg + lcarpcap
chisq = 16.4133, df = 3, p-value = 0.0009329
alternative hypothesis: one model is inconsistent
```

검정결과에 의하면 1%의 유의수준에서 확률효과가 없다는 귀무가설을 기각하게 되어 패널자료에 통계적으로 유의미한 확률효과가 있다. 이제 두 모형 중 어떤 모형을 선택하는게 더 바람직한지를 판단해보기 위해 추가적인 Hausman 검정을 수행할 수 있다. R 소프트웨어를 이용하여 계산한 카이제곱 검정통계량이 16.4133으로 1%의 유의수준에서 고정효과모형과 확률효과모형이 동일하다는 귀무가설을 기각하므로 패널자료의 경우 고정효과모형이 확률효과모형보다 더 타당하다고 판단할 수 있게 된다.

〈표 1〉 1960~1978 OECD 18개 국가에 대한 패널자료의 분석 결과

합동 모형(Pooling Model)					
Balanced Panel	n=18, T=19, N=342				
Coefficients		Estimate	Std. Error	t-value	Pr(〉|t|)
	(Intercept)	2.391326	0.116934	20.450	〈 2.2e-16 ***
	lincomep	0.889962	0.035806	24.855	〈 2.2e-16 ***
	lrpmg	-0.891798	0.030315	-29.418	〈 2.2e-16 ***
	lcarpcap	-0.763373	0.018608	-41.023	〈 2.2e-16 ***
R-Squared(Adj. R^2)	0.85494(0.84494)				
F-statistic	663.999 on 3 and 338 DF (p-value: 〈 2.22e-16)				
고정효과 모형(Fixed Effect Model)					
Balanced Panel	n=18, T=19, N=342				
Coefficients		Estimate	Std. Error	t-value	Pr(〉|t|)
	lincomep	0.662250	0.073386	9.0242	〈 2.2e-16 ***
	lrpmg	-0.321702	0.044099	-7.2950	2.355e-12 ***
	lcarpcap	-0.640483	0.029679	-21.5804	〈 2.2e-16 ***

R-Squared(Adj. R^2)	0.8396(0.78805)
F-statistic	560.093 on 3 and 321 DF (p-value: < 2.22e-16)

확률효과 모형(Random Effect Model)

Balanced Panel	n=18, T=19, N=342
Effects	var std.dev share idiosyncratic 0.008525 0.092330 0.182 individual 0.038238 0.195545 0.818 theta : 0.8923
Coefficients	Estimate Std. Error t-value Pr(>|t|) (Intercept) 1.996698 0.184326 10.8324 < 2.2e-16 *** lincomep 0.554986 0.059128 9.3861 < 2.2e-16 *** lrpmg -0.420389 0.039978 -10.5155 < 2.2e-16 *** lcarpcap -0.606840 0.025515 -23.7836 < 2.2e-16 ***
R-Squared(Adj. R^2)	0.82931(0.81961)
F-statistic	F-statistic: 547.4 on 3 and 338 DF (p-value: < 2.22e-16)

〈표 2〉 2001~2009 OECD 22개 국가에 대한 패널자료의 분석 결과

합동 모형(Pooling Model)

Balanced Panel	n=22, T=9, N=198
Coefficients	Estimate Std. Error t-value Pr(>|t|) (Intercept) -8.204410 0.546316 -15.018 < 2.2e-16 *** lincomep 0.896568 0.046768 19.171 < 2.2e-16 *** lrpmg -1.217263 0.063797 -19.080 < 2.2e-16 *** lcarpcap -1.165840 0.113845 -10.241 < 2.2e-16 ***
R-Squared(Adj. R^2)	0.7384(0.72348)
F-statistic	182.525 on 3 and 194 DF (p-value: < 2.22e-16)

고정효과 모형(Fixed Effect Model)

Balanced Panel	n=22, T=9, N=198
Coefficients	Estimate Std. Error t-value Pr(>|t|) lincomep 0.230822 0.075935 3.0397 0.002736 ** lrpmg -0.491354 0.070754 -6.9445 7.364e-11 *** lcarpcap -0.625721 0.102404 -6.1103 6.395e-09 ***
R-Squared(Adj. R^2)	0.65127(0.56904)
F-statistic	107.696 on 3 and 173 DF (p-value: < 2.22e-16)

확률효과 모형(Random Effect Model)	
Balanced Panel	n=22, T=9, N=198
Effects	var std.dev share idiosyncratic 0.006933 0.083266 0.083 individual 0.076461 0.276516 0.917 theta : 0.9001
Coefficients	Estimate Std. Error t-value Pr(>|t|) (Intercept) -3.098308 0.719107 -4.3085 2.609e-05 *** lincomep 0.425508 0.067190 6.3329 1.638e-09 *** lrpmg -0.671971 0.062468 -10.7571 < 2.2e-16 *** lcarpcap -0.676953 0.103857 -6.5181 6.003e-10 ***
R-Squared(Adj. R^2)	0.62467(0.61205)
F-statistic	107.625 on 3 and 194 DF (p-value: < 2.22e-16)

두 기간으로 나누어 수행한 실증분석결과가 〈표 1〉과 〈표 2〉에 기록되어 있으며, 이를 종합해보면 두 기간 모두 OECD국가의 자동차 한대당 휘발유 소비가 1인당 실질소득, 실질 석유가격, 1인당 자동차 수에 의해 통계적으로 매우 유의하게 영향을 받는 것으로 나타난다. 하지만 분석기간에 관계없이 개체들의 관찰되지 않는 이질성과 시간효과를 전혀 고려하지 않고 분석한 합동모형의 추정치는 이들 효과를 반영한 고정효과모형이나 확률효과모형의 추정치와 상당한 차이를 나타낸다. 결과적으로 적합한 패널모형을 선택하여 올바른 계량기법에 의해 실증분석을 수행한다는 것은 패널자료를 이용한 경험적 연구에서 매우 중요한 일임을 알 수 있다.

V. 결 론

본 연구에서는 기존 문헌을 기반으로 패널자료분석을 위해 발전되어 온 선형패널모형과 추정방법을 소개하고 선형패널모형에서 고정효과 및 확률효과를 검정할 수 있는 검정방법에 대해 설명하였으며 경제분야의 전문가들에게는 잘 알려져 있지 않지만 패널자료분석에 가장 적합한 소프트웨어 중 하나인 R 소프트웨어의 plm(Croissant and Millo, 2008) 패키지를 이용하여 Baltagi and Griffin(1983)의 자동차 휘발유 소비량에 대한 패널회귀모형을 최신 자료를 이용하여 추정하고 실증적으로 분석하였다. 따라서 본 연구의 목적은 패널자료분석과 관

련된 독창적인 계량기법이나 실증분석을 통해 새로운 현상을 발견하기보다 응용계량경제학자 및 전문가들에게 패널자료분석에 필요한 기존의 계량기법을 소개하고 R 소프트웨어의 활용성을 높이는데 있다.

본 연구에서는 일반적 선형패널모형에 대한 계량기법에 대해 다루고 있지만 향후 비선형 패널모형에 대한 계량기법, 설명변수에 시차 종속변수를 포함하는 보다 동적인 패널모형을 추정하기 위한 GMM(generalized method of moments) 추정량의 적용, 비모수 패널모형, 베이지안 패널모형 등에 대한 이론적 소개와 함께 프로그래밍 기법에 대한 추가적인 설명이 요구된다.

참고문헌

Balestra, P. and M. Nerlove. 1966. Pooling Cross-section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas. *Econometrica*, 34, 585-612.

Baltagi, B. H. 2001. Econometric Analysis of Panel Data, second edition. John Wiley and Sons, New York.

Baltagi, B. H. and J. M. Griffin. 1983. Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures. *European Economic Review*, 22(2), 117-137.

Breusch, T. and A. Pagan. 1980. The Lagrange Multiplier Test and Its Applications to Model Specication in Econometrics. *Review of Economic Studies*, 47, 239-253.

Richard A. Becker, John M. Chambers and Allan R. Wilks (1988), The New S Language. Chapman & Hall, New York.

Chamberlain, G. 1984. Panel data. In Z. Griliches and M. Intriligator, eds., Handbook of Econometrics, Vol.2, 1247-1318. North Holland, Amsterdam.

Croissant Y. and G. Millo. 2008. Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software*, 27(2), 1-43.

Hausman, J. 1978. Specication Tests in Econometrics. *Econometrica*, 46, 1251-1271.

Hsiao, C. 2003. Analysis of Panel Data, second edition. Cambridege, UK: Cambridege University.

Hsiao, C. 2007. Panel Data Analysis-Advantages and Challenges. *TEST*, 16, 1-22.

Wooldridge, J. 2002. Econometric Analysis of Cross-Section and Panel Data. MIT press.

Panel Data Analysis Using R Software: Its application to the Panel Model for Gasoline Demand in the OECD

Park, Beum-Jo*

ABSTRACT

This paper comprehensively surveys econometric methods for panel data analysis that is at watershed of time-series data analysis and cross-section data analysis, and thus it is widely known that panel data analysis is an important field in econometrics. In particular, this paper introduces fixed effect model and random effect model, and explains estimation methods for the models and test statistics such as F test statistic, LM test statistic, and Hausman test statistic.

Although R software is not well known in economics, it should be one of the appropriate statistical packages for analyzing panel data. Thus, this paper introduces plm(Croissant and Millo, 2008) package in R software and applies it to an empirical study on motor gasoline consumption per auto in OECD countries with a Baltagi and Griffin's(1983) panel model. This application might help experts in empirical economics to improve their understanding of panel data analysis.

Key Words : Panel Data Analysis, R Software, Fixed Effect Model, Random Effect Model, LM Test, Hausman Test.

* Professor, Dept. of Economics, Dankook University