국가 연구개발 투자의 결정요인 분석
: Panel Regression과 Regression Tree 비교

장한수*, 이경재**, 홍정석***

Ⅰ. 문제 제기

한국을 비롯한 선진국에서 연구개발(R&D) 투자의 적정 수준에 대한 의문은 지속적으로 제기된 문제이다. 특히, R&D 투자에 의하여 연구성과, 연구잠재력, 나아가 경제성장을 담보할 수 있는가에 대한 논의는(Bravo-Ortega and Marin, 2011; Sandu, 2010) 쉽게 결론지을 수 없는 학술적 도전과제이다.

한편, Zastrow(2016)는 한국의 GDP 대비 R&D투자 수준은 세계 최고이지만, 연구개발인력, 논문 게재건 등은 그에 상응하는 수준에 미치지 못한다는 견해를 보였다. 그 견해에 대한 학술적 타당성뿐만 아니라, 국가 R&D 투자가 GDP로 대비되는 경제적 축적이나 경제시스템(생산성)과 어떠한 연관성이 있는지 검토할 필요가 있다. 이와 관련하여 Panel 자료를 이용한 회귀분석 (Coccia, 2009; Coccia, 2012), 구조방정식에 의한 분석(김인자, 2015), CGE 모형에 의한 분석(과학기술정책연구원, 2012) 등 다양한 연구가 수행되었다.

연구개발과 경제 관련 지표를 이용한 적정 연구개발 투자 또는 그 효과성 추정에 관한 연구의 어려움 중 하나는 패널(Panel) 자료 수집과 통계분석에 있다. 즉, 각 국가별 다양한 지표에 대한 균형자료를 얻기가 용이하지 않다. 또한, 패널 회귀분석 적용에 있어서 고정(fixed)/임의(random) 효과 식별이 필요하다. 더욱이, 시계열 회귀분석을 위한 안전성(stationary) 검토 외에도 이분산성(heteroskedasticity) 검토 등도 필요하다(Torres-Reyna, 2010). 환언하면 자료 확보의 어려움, 회귀분석을 위한 기본 검토와 패널자료로써 통계적 분석 등의 단계가 필요하다.

본 논문에서는 회귀식 추정에 의한 패널 자료분석에 대한 대안으로 회귀나무(Regression Tree) 방법론을 적용하여 R&D 투자의 결정요인을 판별하고자 한다. 본 논문의 구성은 II장에서 회귀나무 방법론에 대한 기본 개념과 패널 자료 분석이 가능한 회귀나무에 대하여 살펴본다. 또한, 분석에 적용되는 데이터와 기초 통계를 검토한다. III장에서는 회귀나무 분석 결과를 살펴보고 IV장에서 결론을 도출한다.

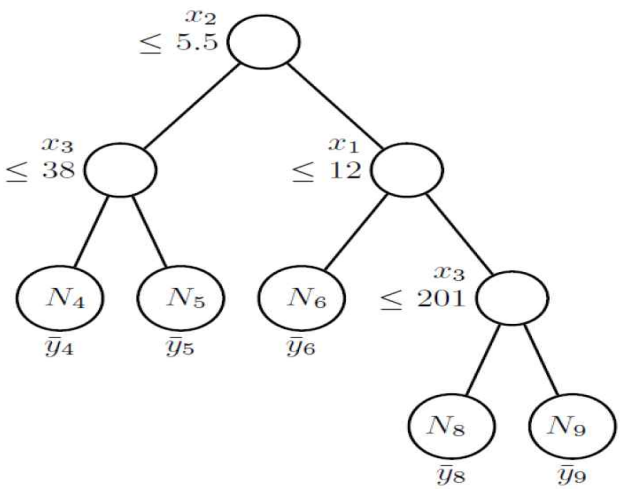
* 장한수, 국가핵융합연구소 선임연구원, 02-589-2810, jjang@kistep.re.kr, jjang@nfri.kr
** 이경재, 한국과학기술기획평가원 연구위원
*** 홍정석, 한국과학기술기획평가원 연구위원

II. 분석 방법론 및 변수

1. 분석 방법론

1) 회귀나무의 기본 개념

회귀나무는 독립변수로 이루어진 공간을 재귀적으로 분할하고 해당 영역에서 종속변수의 최선의 예측값을 찾고자 하는 비모수적 방법론이다. 나무구조를 형성할 때 변수선택 과정은 독립변수 공간 분할과 직결되는 요소이므로 매우 중요하다고 할 수 있다. 회귀나무를 형성하기 위해서는 적절한 변수를 선택한 뒤 이를 기준으로 데이터를 나누고 나뉘어진 각 독립변수 공간에서 종속변수와 독립변수와의 관계성을 찾아가는 과정을 거치게 된다. 이러한 과정을 정리하면 (그림 1)과 같다. (그림 1)은 반복 이분할 (binary recursive partitioning) 과정을 통해 구현된 회귀나무를 나타내고 있다. 그림에서 원으로 표시된 것이 노드 (node)라고 불리는 것으로서 일정 기준에 따라 분할된 독립변수 공간으로 생각할 수 있다. 즉, 분할의 기준이 되는 변수에 따라 데이터가 분류되는 곳이다. 실선으로 표시된 것이 가지(branch)라고 불리는 것으로서 각 단계에서 조건에 따라 하위 노드로 데이터를 나누어 가는 과정을 나타낸다. 이러한 분할 과정이 반복적으로 실행되면서 전체 나무의 모습을 이루게 된다. 최상위 노드는 모든 학습샘플 (learning sample)을 포함하고 있으며, 각 분기점이 되는 노드에서 분기변수 (split variable)로 선택된 설명변수 (x_1, x_2, x_3, ...)의 값에 따라 분기가 반복되는 단계를 거쳐 최종 노드 (N_4, N_5, N_6, N_8, N_9)에 이르게 된다. 최종 노드 아래에 표시된 값은 회귀나무 모형에 따른 예측값이다. 만약 가장 기본적인 상수항 회귀나무 모형을 적합했다면 예측값은 각 최종 노드에 위치한 표본의 평균값 ($\bar{y}_4$, $\bar{y}_5$, $\bar{y}_6$, $\bar{y}_8$, $\bar{y}_9$)이 된다. 즉, 각 노드에 위치한 종속변수들의 평균치를 예측값으로 하는 상수항 모형이다(장영재, 2014).



(그림 1) 회귀나무 분석 결과 예(장영재, 2014)

2) 패널 자료 분석을 위한 회귀나무 방법론

패널 연구는 종단연구 (longitudinal study)의 특별한 한 가지 형태라고 할 수 있다. 종단연구는 연구의 대상 집단을 어느 시점에서 표집하여 이들을 대상으로 오랜 기간에 걸쳐 반복적으로 관찰함으로써 시간의 흐름에 따라 각종 변인들의 변화 상태를 파악하는 연구방법이다. 관측치의 특성이 시간의 흐름에 따라 변화하는 것을 추적하고 관찰하는 연구 방법이다. Sela and Simonoff(2012)는 관측치 간의 차이점을 임의효과 (random effects)로 간주하고 이를 주효과인 고정효과 (fixed effects)와 함께 고려하는 혼합모형을 구축한 뒤 나무모형과 접목하여 추정하는 방법을 제안하였다. 추정을 위하여는 EM 알고리즘을 사용하였으며 이러한 점에서 RE-EM 나무모형이라 명명하였다(장영재, 2014).

해당 방법론은 R의 REEMtree 패키지(Sela and Simonoff, 2011)를 이용하여 분석이 가능하다. 연구개발 투자와 관련하여 회귀나무를 적용한 사례는 김동근 외(2012)가 있으나, 패널 자료를 사용하지는 않았다.

2. 변수와 기초통계

1) 변수 선택

연구개발 투자의 적정 수준 판단을 위하여 회귀나무 방법론을 적용할 패널 변수는 다음과 같다. 연구개발 활동의 가장 큰 목적 중 하나가 국가적 부의 축적이라고 보고 이를 대표할 종속변수는 GDP 관련 지표를 선택하되, GDP 지표의 상대적 비율을 나타낸 지표를 활용한다. 독립변수는 국가적 부의 축적에 영향을 미치는 연구개발 활동 관련 지표라고 볼 수 있는 GDP 대비 연구개발 투자 비율, 1인당 연구개발 투자액, 고용인구 1천 명당 연구개발인력, 노동생산성 등을 적용한다.

2) 변수별 기초통계

위에서 언급한 패널 변수는는 OECD(2016)에서 제공하는 26개 국가의 연구개발 관련 시계열 지표(2000~2013년)를 이용하였다. 본 논문에 활용한 지표는 i국가의 2010년 대비 t년도 GDP 비율(PI_{it}), GDP 대비 연구개발 투자비율($XGDP_{it}$), 1인당 연구개발투자액($XPOP_{it}$), 고용인구 1천 명당 연구개발인력($RSXEM_{it}$), 노동생산성($GDPHRS_{it}$) 등이다. 각 지표의 기초통계는 <표 1>과 같다.

<표 1> 변수별 기초통계

	종속변수	독립변수			
	PI_{it}	$XGDP_{it}$	$XPOP_{it}$	$RSXEM_{it}$	$GDPHRS_{it}$
최소값	0.2217	0.3533	0.03559	0.06207	1.417
제1사분위수	0.8594	1.0575	0.21755	0.49296	2.755
중앙값	0.9451	1.621	0.54275	0.69219	4.248
평균	0.9184	1.7328	0.56879	0.7019	4.079
제3사분위수	1.0007	2.3802	0.83367	0.87536	5.292
최대값	1.3082	4.1485	1.48995	1.76781	8.171

III. 실증분석 결과

1. 패널 회귀에 의한 추정

회귀나무에 의한 분석에 앞서 패널 회귀식 추정으로 모형의 고정/임의 효과와 그에 따른 추정 계수를 살펴본다. 추정식은 식(1)과 같다. 분석은 R을 활용하였으며 추정 결과는 <표 2>와 같다.

$$PI_{it} = \beta_0 + \beta_1 XGDP_{it} + \beta_2 XPOP_{it} + \beta_3 RSXEM_{it} + \beta_4 GDPHRS_{it} + \epsilon_{it} \quad 식(1)$$

<표 2> 패널 회귀에 의한 추정 결과

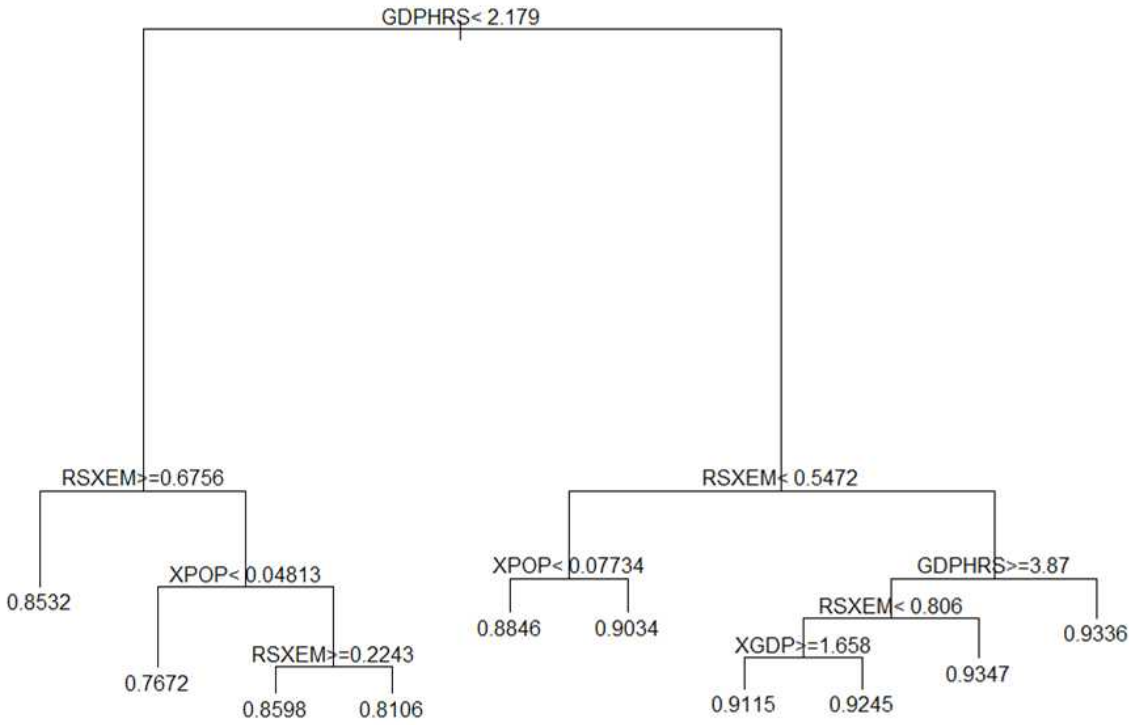
		β_1	β_2	β_3	β_4	R^2 / p값
OLS 모형	추정계수	**-0.062**	**0.253**	0.042	0.002	0.2086 / 2.2×10^{-16}
	t값	-2.355	3.753	1.052	0.289	
고정 모형	추정계수	0.038	0.073	-0.106	**0.289**	0.4625 / 2.22×10^{-16}
	t값	0.8932	0.9769	-1.2150	9.8045	
임의 모형	추정계수	**-0.098**	**0.342**	0.124	0.009	0.2357 / 2.22×10^{-16}
	t값	-2.9420	4.5112	1.9453	0.7433	
F검정(F값, p값)		F=11.313, p값<2.2×10^{-16}				
Hausman검정 (χ^2, p값)		χ^2=147.76, p값<2.2×10^{-16}				

※ 굵은 글씨는 통계적으로 유의한 추정치를 나타냄

<표 2>의 결과에 나타난 바와 같이 OLS보다는 고정 모형이 통계적으로 유의하고, 고정 모형이 통계적으로 타당한 모형임을 보여준다. 고정모형의 R2값도 다른 모형보다 비교적 높다. 특이한 점은 다른 모형에서는 $XGDP_{it}$나 $XPOP_{it}$가 GDP 성장에 유의한 것으로 나타나지만, 통계적으로 유의한 고정모형에서는 $GDPHRS_{it}$만 통계적 유의성을 보인다.

2. 회귀나무에 의한 추정

회귀나무에 의한 추정 결과는 (그림 2)와 같다. 회귀분석의 결과와는 다르게 $GDPHRS_{it}$의 영향이 가장 크며 $RSXEM_{it}$의 영향이 두 번째로 나타났다.



(그림 22) REEMtree 방법론에 의한 추정 결과

IV. 결론

한국의 연구개발 투자 수준이 세계적으로 성장하였지만, 성과는 그에 미치지 못한다는 비판이 제기되었다. 본 논문은 국가연구개발 투자의 적정 수준을 판별하는 요인이 무엇인지 추정하고자 하였다. 이에 대한 정량적 검증을 위하여 기존의 패널회귀의 난제에 대한 대안으로 회귀나무 방법론을 적용하였다.

추정 결과 회귀분석과는 다르게 국가의 생산성 수준이 국가의 부에 가장 연관이 큰 것으로 나타났다. 이는 연구개발에 자원이 투입되더라도 국가의 생산성으로 대표되는 전체 시스템 측면의 효율성이 더 중요하게 작용한다고 해석할 수 있다. 즉, 한구긍ㄴ 연구개발투자로 인하여 시스템 측면의 효율성 향상에도 기여할 수 있는 방안을 고려하여야 한다.

참고 문헌

과학기술정책연구원 (2012), 「연구개발투자의 경제적 효과 평가 및 예측모형 개발」, 정책연구 2012-08

김동근 · 천영돈 · 김성규 · 이윤빈 · 황지호 · 김용수 (2012), "회귀분석 및 의사결정나무 분석을 통한 R&D 연구비 추정에 관한 연구", 「산업경영시스템학회지」, 35(4) : 73-82.

김인자 (2015), 「연구개발(R&D)활동이 GDP에 미치는 영향 분석: 과학기술논문과 특허의 매개를 통하여」, 서울: 한국과학기술기획평가원 연구보고 2016-059.

장영재 (2014), "회귀나무 모형을 이용한 패널데이터 분석", 「한국데이터정보과학회지」, 25(6) : 1253-1262.

Bravo-Ortega, Claudio and Marin, Alvaro (2011), "R&D and Productivity: A Two Way Avenue?", *World Development*, 39(7): 1090 - 1107.

Coccia, Mario (2012), "Political economy of R&D to support the modern competitiveness of nations and determinants of economic optimization and inertia", *Technovation*, 32: 370-379.

Coccia, Mario (2009), "What is the optimal rate of R&D investment to maximize productivity growth?" *Technological Forecasting and Social Change*, 76(3):

Croissant, Yves and Millo, Giovanni (2008), "Panel Data Econometrics in R: The plm Package". *Journal of Statistical Software*, 27(2), http://www.jstatsoft.org/v27/i02/

Finch, Holmes (2015), "Recursive Partitioning in the Presence of Multilevel Data", *General Linear Model Journal*, 41(2): 30-44

Fu, Wei and Simonoff, Jeffrey (2015), "Unbiased regression trees for longitudinal and clustered data", *Computational Statistics and Data Analysis*, 88: 53-74.

Fu, W. and Simonoff, J.S. (2014), "Unbiased Regression Trees for Longitudinal Data" Available at SSRN: http://ssrn.com/abstract=2399976

OECD (2016), Main Science and Technology Indicators 2015

Ripley, Brian (2016), "tree: Classification and Regression Trees", R package version 1.0-37, https://CRAN.R-project.org/package=tree

Sandu, Steliana (2010), "The Optimal Rate of R&D Expenditures in GDP? Between Theory and Practice", *Constatin Brancusi University of Targu Jiu Annals - Economy Series*.

Sela, Rebecca J. (2009) "RE-EM Trees: A New Data Mining Approach to Longitudinal Data", *Conference on Quantitative Social Science Research Using R*

Sela, Rebecca and Simonoff, Jeffrey (2011), "REEMtree: Regression Trees with Random Effects", R package version 0.90.3.

Sela, R.J. and Simonoff, J.S. (2012), "RE-EM Trees: A Data Mining Approach for Longitudinal and Clustered Data", *Machine Learning*, 86: 169-207.

Simonoff, Jeffrey S. (2014), "Regression Trees for Longitudinal and Clustered Data

Based on Mixed Eects Models: Methods, Applications, and Extensions", www2.ims.nus.edu.sg/Programs/014swclass/files/simonoff.pdf

Simonoff, J.S. (2013), "Regression Tree-Based Diagnostics for Linear Multilevel Models", *Statistical Modelling*, 13: 459-480.

Simonoff, Jeffrey S. (2011), "Regression tree-based diagnostics for linear multilevel models", *The Fourth Erich L. Lehmann Symposium*

Torres-Reyna, Oscar (2010), "Getting Started in Fixed/Random Effects Models using R", http://dss.princeton.edu/training/

Wickham, Hadley (2016), "readxl: Read Excel Files", R package version 0.1.1., https://CRAN.R-project.org/package=readxl

Zastrow, Mark (2016), "South Korea's Nobel dream", 「Nature」, 534: 20-23