http://dx.doi.org/10.11627/jkise.2012.35.4.73

pISSN 2005-0461
eISSN 2287-7975

A Study on Estimation of R&D Research Funds by Linear Regression and Decision Tree Analysis

Dong-Guen Kim* · Youngdon Cheon** · Sungkyu Kim**
Yoon Been Lee* · Ji Ho Hwang* · Yong Soo Kim***†

*Korea Institute of Science and Technology Evaluation and Planning
**Department of Industrial and Management Engineering, Graduate School, Kyonggi University
***Department of Industrial and Management Engineering, Kyonggi University

회귀분석 및 의사결정나무 분석을 통한 R&D 연구비 추정에 관한 연구

김동근* · 천영돈** · 김성규** · 이윤빈* · 황지호* · 김용수***†

*한국과학기술기획평가원
**경기대학교 대학원 산업경영공학과
***경기대학교 산업경영공학과

Currently, R&D investment of government is increased dramatically. However, the budget of the government is different depending on the size of ministry and priorities, and then it is difficult to obtain consensus on the budget. They did not establish decision support systems to evaluate and execute R&D budget. In this paper, we analyze factors affecting research funds by linear regression and decision tree analysis in order to increase investment efficiency in national research project. Moreover, we suggested strategies that budget is estimated reasonably.

Keywords : R&D, Research Funds, Decision Tree Analysis, Linear Regression Analysis

1. 서 론

1.1 연구의 필요성

연구개발예산이 빠르게 증가함에 따라 연구개발사업에 대한 효율적인 투자의 필요성이 지속적으로 제기되고 있다. 그러나, 정부 R&D 투자의 전략성과 효율성 제고를 위한 정부 연구개발예산의 주요 이슈분석과 제도 개선 연구는 충분하지 못한 상황이다. 또한, 최근 R&D 예산의 지속적인 증가에 비해, 재정투자의 효율성 제고를 위한 심층 분석은 부족한 실정이다.

현재 정부 R&D 예산편성 시스템은 데이터 요구사항에 대하여 정규화 과정을 통하여 논리적 데이터 모델을 설계하고, 각 엔티티에 대하여 주요 속성 및 식별자를 표현하는 방식으로 작성되고 있다. 그러나, 예산 편성과정에서 과제당 연구비 규모에 대한 통계적인 지원 시스템은 예산편성 시스템에서 지원되고 있지 않다.

Received 31 July 2012; Accepted 27 September 2012
† Corresponding Author : kimys@kgu.ac.kr

또한, 집행과정에 대한 자료의 경우, 대부분의 자료가 인건비에 집중되고 있어 연구개발 단계, 연구분야 등 다른 요인들에 대한 분석은 부족한 상황이다. 즉, 연구개발 예산을 집행하는 과정에서 특정 요인들에 따라 연구비 규모의 차이가 크게 나타나는 것을 알 수 있다. 예를 들어, 총 연구비에 대하여 집행하였을 경우, 연구분야 및 참여 연구자 규모에 따라 5000만 원, 5억 원, 10억 원 등 연구비의 차이가 발생한다. 이러한 차이가 나타나는 원인과 연구비 규모를 결정하는 요인을 분석할 필요가 있다.

현재는 객관적인 연구비 산정을 위한 통계 자료 및 시스템이 없어, 과거의 유사한 연구를 바탕으로 인원 및 기간을 파악함으로써, 기존의 결과를 활용한다. 이러한 과정에서 정부의 R&D 투자예산이 증가하였음에도 불구하고, 연구비 산정 시 정확한 단가를 추정하지 못함으로써, 효율적인 예산 집행이 이루어지고 있지 못한 것으로 판단된다. 따라서, 본 연구에서는 연구비 규모를 산정하는 통계적인 분석 지원 시스템을 개발하기 위한 가이드라인을 제시함으로써, 향후 보다 과학적인 연구비 단가추정을 통한 R&D 예산의 투자 효율성 제고를 달성할 수 있을 것이다.

1.2 연구의 목적

본 연구는 총 연구비에 대한 회귀분석 및 의사결정나무 분석을 수행함으로써, 합리적 예산편성 전략을 모색하고, 총 연구비를 의미 있게 설명하는 비용항목을 도출하여 국가 연구개발 사업 투자의 효율성을 제고하고자 한다.

본 연구는 비용의 세부구분에 따라 총 연구비를 가장 우수하게 설명하는 비용요인을 도출하고 도출된 요인을 활용하여 보다 고도화된 총 연구비의 추정을 목적으로 한다.

또한, 총 연구비를 유의미하게 설명하는 비용항목을 찾아 연구비 산정 시 연구 유형별 특성을 반영한 정확한 비용을 추정할 수 있다면 R&D 연구비를 효율적으로 집행할 수 있을 것이다.

본 연구에서는 총 연구비 규모에 영향을 미치는 유의한 변수를 파악하여, 향후 총 연구비 규모산정을 위한 분석모델 도출을 위한 기초 자료를 도출하고자 한다.

2. 관련문헌 연구

한 국가의 예산은 그 국가의 정책결정 결과를 반영한 정책문서이자 정치과정의 결과를 기록한 종합적인 정치문서이다[8]. 또한, 우리나라는 경제 위기 극복과 미래대비 성장잠재력 확충을 위해 정부 R&D 투자가 확대되고 있으며, 2009년에도 정부 R&D 예산이 전년대비 11.4%(123,437억 원)가 증가하였다[2].

그러나, 이와 같이 증가하는 정부의 R&D 예산에 따라 체계적인 예산배분을 위한 실질적인 시스템의 개발이 진행되지 않고 있다. 또한, 정부 부처별 예산의 규모와 지원 우선순위가 달라 합의점을 구하기가 매우 어렵다[3].

지금까지 수행된 연구는 공산품 등 일반적인 원가 및 단가에 관한 연구가 수행되었다. 정호진[7]의 연구에서는 거시적인 방법을 이용하여 우리나라의 과거 공급지장비 단가를 산정하고, 이를 회귀분석을 활용하여 미래의 공급지장비 단가를 추정하는 연구를 수행하여 계통 계획안별 공급지장비 평가가 가능한 접근법을 제안하였다.

어원재 외[6]는 국내에서 개발되는 무기체계에 관한 자료를 활용하여 무기체계에 대한 비용추정관계식의 개발 절차를 제시하였으며, 백종건 외[9]에서는 기존 PDM 기능에 원가기획 개념과 방법론을 추가하여 제품개발 단계에서의 원가추정에 필요한 새로운 원가계산시스템의 구조와 체계적인 원가추정방법을 제시하였다. 유일상 외[13]는 세계 주요 발사체의 개발 예산을 조사하고, NASA에서의 비용 추정 기법의 적용 현황을 조사하였으며, 발사체 추정 모델인 TRANSCOST를 활용하여 발사체의 개발비용을 추정하는 사례를 제시하였다.

또한, 송용철 외[11]에서는 토지정책수단의 주요 대상이 되는 농지에 관한 가격 추정 시 세부지역별로 발생될 수 있는 공간적 효과를 고려하지 않은 전통적 회귀모형의 한계를 극복하기 위하여 공간계량경제 접근방법으로 수도권 근교통지의 가격을 추정하였다.

이 외에도, 통계적인 원가 분석기법을 이용하여 호텔기업의 매출액과 원가 사이의 관계를 조사하고, 적용할 수 있는 통계적인 부문별 원가 추정모형을 설정하는 연구가 수행되었으며[1], 서선덕 외[10]는 비용요소에 대해 명시적인 확률분포를 고려하여, Monte Carlo 시뮬레이션 방법을 활용한 위험분석을 수행하여 효율적인 도로 투자비를 산정하는 연구를 수행하였다.

Zhu et al.[14]에서는 프로젝트에 대한 비용을 추정하기 위한 GA(genetic algorithm)과 SOFM(self-organizing feature map)을 활용한 genetic fuzzy neural network를 적용한 프로젝트 비용을 추정하는 연구를 수행하였으며, 이는 기존의 방법에서 나타나는 정확성과 예측력을 높여주었다.

또한, DiMasi et al.[5]에서는 의약품 개발 비용에 대한 새로운 추정방법을 통하여 의약품에 대한 합리적 가격에 대하여 제안하였으며, Wiesenthal et al.[12]에서는 데이터가 거의 없는 지역의 R&D 투자 수준을 저 탄소 에너지 기술에 의한 상향식 접근방법으로 구분하는 연구를 수행하였다.

이처럼 단가추정에 관한 연구는 크게는 도로투자비용 산정부터 작게는 장비 및 제품의 원가추정에 관한 연구

까지 다양하게 수행되고 있다. 이와 같이 폭 넓은 단가 추정연구에도 불구하고, R&D 연구비 산정에 관한 연구는 수행되지 않고 있으며, R&D에 관한 연구가 수행 단가추정이 아닌, 투자수준을 구분하는 연구가 진행되었다. 즉, 연구 개발에 대한 투자가 급속히 증가하고 있는데 반해, 연구비 산정 시 필요한 통계적인 시스템이 부족하여, 효율적인 연구비 단가 추정이 이루어지지 못한 것으로 판단된다.

본 연구에서는 회귀분석과 의사결정나무 분석을 통하여 연구비 산정 시 영향을 미치는 변수를 분석하고, 이 변수들을 통하여 총 연구비를 추정하기 위한 변수들을 도출하고자 한다.

3. 데이터 전처리 및 데이터 모형 수립

3.1 데이터 전처리 및 시각화

본 연구에서는 2008년부터 2010년까지 종료된 연구과제 15,454개를 대상으로 수집하였으나, 결측치 및 이상치가 많으며, 본 데이터를 통하여 얻은 결과는 정합성이 떨어지는 결과를 가져올 수 있다. 따라서, 유의미한 통계 분석을 위해 데이터 전처리 과정이 필요한 것으로 분석되었다.

또한, 본 데이터는 연구비에 관한 데이터로 일반적인 연구비의 규모에서 지나치게 벗어나는 값을 갖는 과제의 경우 빈도에 비해 분석에 높은 영향력을 나타내게 되어 보다 정확한 결과를 도출하는데 악영향을 미치게 된다. 이러한 극단치를 갖는 데이터를 제거하여 통계적으로 유의미한 결과를 도출하고자 데이터 전처리 과정을 수행하였다.

데이터 전처리 과정은 원시데이터에서 결측치를 제거한 후 시각화를 실시한다. 데이터를 시각화함으로써, 극단치의 존재여부를 확인하고, 극단치 제거과정을 수행하여 통계적으로 유의미한 결과를 도출할 수 있도록 하였다.

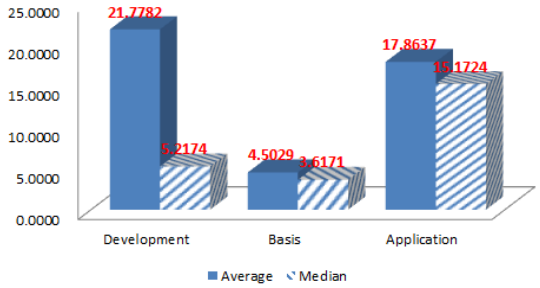
3.1.1 결측치 제거 및 시각화

수집한 총 15,454개의 데이터에서 총 연구비에 대한 결측치 254개와 '기타'로 표시되어있는 데이터 11,685개의 데이터를 제거하여 4,515개의 데이터를 산출하였다.

'기타'로 표시되어있는 데이터는 각 변수 중 '기타'로 분류된 데이터이며, '기타'로 포함된 변수로 분석을 수행할 경우 '기타'변수가 결과에 유의한 변수로 선정되므로 제거하였다.

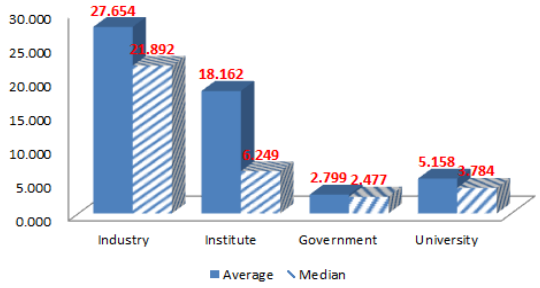
결측치를 제거한 4,515개의 데이터를 대상으로 극단치가 존재하는지를 확인하기 위하여 시각화를 실시하였으며, 시각화된 데이터를 통하여 어느 부분의 연구에서 연구비가 많이 포함되는가에 대한 기초통계량을 분석하였다.

아래의 <Figure 1>은 총 연구비를 월별 연구비로 계산하여 연구개발단계별 평균과 중위수를 그래프로 나타낸 그림으로 개발 부분의 경우, 평균이 약 2천 1백만 원인 반면, 중위수가 약 5백만 원으로 평균과 중위수의 차이가 크다. 이는 극단치가 존재한다는 것을 알 수 있다.



<Figure 1> Average and Median Based on R&D Phases

또한, <Figure 2>는 연구 수행 주체 별로 월별 총 연구비를 평균과 중위수로 나타낸 것으로 우선 연구소 변수에서 극단치가 존재하는 것을 알 수 있으며, 전체적으로는 산업체(대기업, 중소기업)의 경우가 다른 주체보다 R&D 연구 수행시 연구비를 많이 사용하는 것을 알 수 있다.



<Figure 2> Average and Median Based on Sector of Performance

위의 시각화된 자료를 통하여 결측치를 제거한 4,515개의 데이터에서 극단치가 존재하는 것을 확인 할 수 있었다. 따라서, 극단치를 제거하는 과정을 수행하였다.

3.1.2 극단치 제거

극단치 제거는 SAS 9.2 Enterprise Miner를 활용하였으며, Filter Outlier노드를 사용하였다.

본 연구에서는 중위수를 중심으로 연구를 수행하였으므로, 중위수로부터 n편차값 보다 큰 값을 제거하는 MAD (Median Absolute Deviation) 옵션을 사용하여 극단치를 제거하였다. 그 결과, 4,515개의 데이터 중 405개의 극단치

데이터가 제거되었으며, 최종 산출된 4,110개의 데이터를 이용하여 데이터 분석을 실시하였다.

3.1.3 데이터 변수 선정

본 연구에서는 이와 같이 전처리를 통하여 생성된 데이터를 기반으로 하여 본 연구에서 사용될 변수들을 다음의 <Table 1>과 같이 정리하였다.

변수는 범주형 변수와 수치형 변수로 구분하였고, 범주형 변수는 X_1은 연구 개발 단계를 나타냈으며, 연구개발 단계는 x_{1a}(개발), x_{1b}(기초), x_{1c}(응용)으로 더미변수를 생성하였다. X_2는 연구수행주체로서, 산업체, 연구소, 정부기관, 대학으로 구분하였으며, X_3변수인 분류변수 또한, BT(Bio Technology), CT(Culture Technology), ET(Environment Technology), IT(Information Technology), NT(Nano Technology), ST(Space Technology)로 구분하여, 회귀분석을 수행하기 위한 변수를 생성하였다.

연속형 변수의 경우, X_4(과제기간)은 과제를 수행하는 기간을 개월수로 나타냈으며, X_5(정부출연금)부터 X_{10}(간접비 비율)까지는 각각의 금액이 총 연구비에 해당하는 비율을 %로 구분하여 나타냈다. X_{11}(참여인원)은 연구에 참여한 인원으로, 박사, 석사, 학사 등 연구에 참여한 모든 인원의 합을 나타냈다.

<Table 1> Definition of Variables

Variable	Variables Name		Contents
Categorical variable	X_1	Research and development phases	Basis(x_{1a}) Development(x_{1b}) Application(x_{1c})
	X_2	Conduct research subject	Industry(x_{2a}), Institute(x_{2b}), Government(x_{2c}), University(x_{2d})
	X_3	Research area	BT(x_{3a}), CT(x_{3b}), ET(x_{3c}), IT(x_{3d}), NT(x_{3e}), ST(x_{3f})
Continuous variable	X_4	Period	Unit : Month
	X_5	Cost of government contribution	Cost of government contribution/ Total research fund(Unit : %)
	X_6	Labor cost	Labor cost/Total research fund (Unit : %)
	X_7	Research equipments and materials cost	Research equipments and materials cost/Total research fund(Unit : %)
	X_8	Operating cost	Operating cost/Total research fund (Unit : %)
	X_9	Cost of researches commissioned	Cost of researches commissioned/ Total research fund(Unit : %)
	X_{10}	Overhead cost	Overhead cost/Total research fund (Unit : %)
	X_{11}	Number of researchers	Number of researchers
	Y	Total research funds	Unit : One million won

3.2 데이터 분석결과

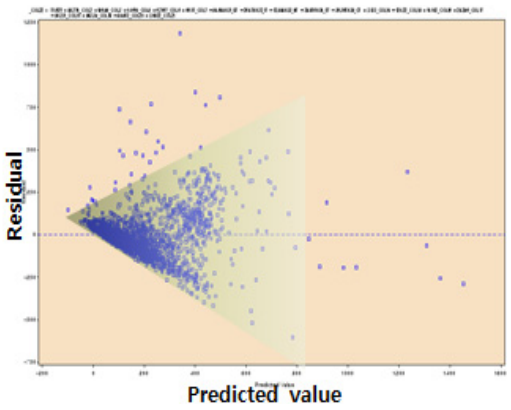
본 연구에서는 SAS 9.2 Enterprise Miner를 활용하여 회귀분석 및 의사결정나무 분석을 수행하였고, 총 연구비를 선정하는데 영향을 미치는 변수를 확인하였으며, 데이터는 학습용 데이터 70%와 평가용 데이터 30%로 분할하여 데이터 마이닝 기법 별 예측력을 평가하였다.

3.2.1 선형회귀분석

본 연구에서는 선형회귀분석을 수행하여 총 연구비에 따른 유의미한 변수를 파악하고 총 연구비를 예측한다.

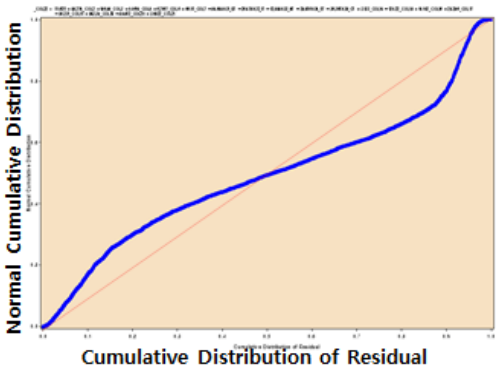
선형회귀분석을 수행하기 전 수집된 데이터를 통하여 잔차분석을 수행하였다. 잔차분석은 선형회귀분석에서 각 모수에 대한 추론은 모집단 회귀모형에 포함된 오차항에 대한 가정을 바탕으로 하나, 오차항은 관측될 수 없으므로, 일종의 추정량인 잔차를 이용하여 이 가정의 타당성을 조사하는 것이다. 회귀분석에서 오차항에 대한 가정은 등분산성과 정규성 검정을 통하여 확인하며, 잔차가 등분산성을 나타내지 않을 경우, 목표변수(Y)의 변환을 통하여 등분산성을 나타내어야 한다.

따라서, 본 연구에서는 수집된 데이터를 바탕으로 목표변수를 총 연구비로 하여 잔차분석과 정규성 검정을 수행하였다. 아래의 <Figure 3>은 목표변수를 총 연구비로 하여 잔차분석을 수행 한 결과이며, 그림에서 나타나는 바와 같이 잔차의 분산이 예측된 종속변수가 증가함에 따라 확산되는 Funnel형 분포를 보여주고 있으므로, 목표변수의 변환이 필요하다는 것을 알 수 있다.



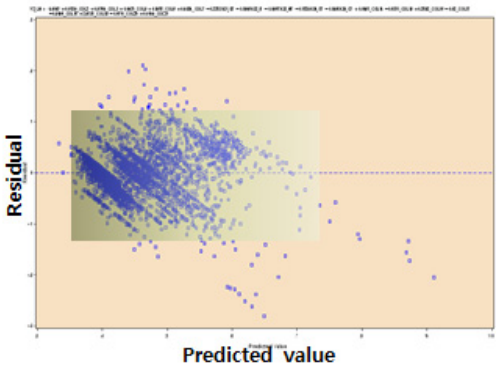
<Figure 3> Residual Analysis for 'Y = Total Research Funds'

또한, 정규성 검정을 통하여 나타난 결과가 <Figure 4>이며, 직선의 형태를 나타내어야 하나, 기존의 데이터는 직선의 형태에서 벗어난 형태를 보이고 있다.

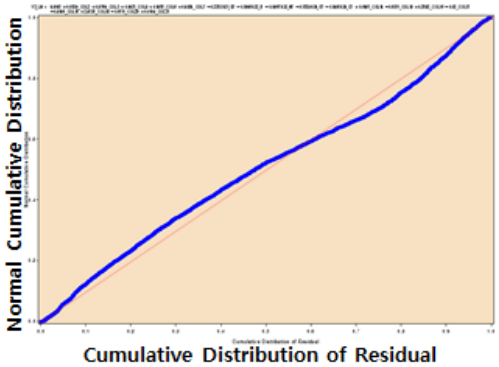


<Figure 4> Normality Test for 'Y = Total Research Funds'

Douglas CM et al.[4]에서는 Funnel형의 분포에서는 일반적으로 목표변수에 자연로그(ln)을 취하여 등분산성을 나타내도록 제안하였다. 이에 따라, 목표변수에 자연로그를 취하여 잔차분석을 실시한 결과가 <Figure 5>이다. 자연 로그를 취함으로써, 데이터의 분포가 비교적 등분산성에 근접한 것을 알 수 있다. 또한, 자연로그를 취하여 정규성 검정을 통한 결과에서도 <Figure 6>과 같이 직선에 근접하게 나타나고 있음을 확인 할 수 있다.



<Figure 5> Residual Analysis for 'Y = ln(Total Research Funds)'



<Figure 6> Normality Test for 'Y = ln(Total Research Funds)'

본 연구에서는 전체 데이터뿐만 아니라, 데이터를 연구 분야(BT, CT, ET, IT, NT, ST)로 구분하여 잔차분석을 실시하였으며, 이후 선형회귀분석 및 의사결정나무에서도 동일한 분석을 수행하였다.

목표변수를 선정된 형태로 회귀분석을 전체 데이터 및 연구 분야별로 Full model과 Stepwise로 옵션을 선택하여 분석을 수행하였으며, 두 옵션 중 우수한 결과를 나타내는 옵션으로 결과를 해석하였다. 아래의 <Table 2>은 전체 데이터에 대한 회귀분석 결과이며, 연구개발단계가 응용단계일 때 기초단계나 개발단계일 때보다 총 연구비가 약 43%가 증가하는 것을 알 수 있다. 또한, 연구수행주체는 산업체에서 연구를 수행하는 경우, 정부나 대학에서 연구를 수행할 때보다 연구비가 약 26%가 증가하는 것을 알 수 있으며, ST연구일수록 연구비가 약 12% 증가하는 것을 확인할 수 있다. 과제기간과 참여인원은 각 단위당 약 1~2% 정도 증가하는 것으로 나타났다.

<Table 2> Summary Statistics of All Possible Regression Entire Data

Parameter	Estimate	Standard Error	t-value	Pr>|t|	Multiple
Intercept	4.66	0.11	41.58	< .0001	105.41
Development(x_{1b})	-0.22	0.02	-13.05	< .0001	0.80
Basis(x_{1a})	-0.14	0.02	-8.62	< .0001	0.87
Application(x_{1c})	0.36	-	-	-	1.43
Industry(x_{2a})	0.23	0.07	3.38	0.0007	1.26
Institute(x_{2b})	0.03	0.07	0.42	0.5535	1.03
Government(x_{2c})	-0.18	0.19	-0.92	0.3552	0.84
University(x_{2d})	-0.08	-	-	-	0.92
BT(x_{3a})	-0.09	0.02	-5.31	< .0001	0.92
CT(x_{3b})	-0.05	0.05	-1.16	0.2468	0.95
ET(x_{3c})	-0.04	0.02	-1.76	0.0793	0.96
IT(x_{3d})	-0.01	0.02	-0.68	0.4978	0.99
NT(x_{3e})	0.08	0.02	4.01	< .0001	1.09
ST(x_{3f})	0.11	-	-	-	1.12
Period	0.02	0.00	34.63	< .0001	1.02
Cost of government contribution	-0.63	0.10	-6.20	< .0001	0.53
Labor cost	0.28	0.04	6.37	< .0001	1.33
Research equipments and materials cost	-0.51	0.05	-10.66	< .0001	0.60
Cost of researches commissioned	2.97	0.14	21.44	< .0001	19.58
Overhead cost	-0.68	0.13	-5.41	< .0001	0.51
Number of researchers	0.01	0.00	23.85	< .0001	1.01

또한, 전체 데이터 및 분야별 데이터에 대한 회귀식을 아래의 <Table 3>와 같이 정리하였으며, 회귀분석을 통하여 응용단계에서 연구비가 전반적으로 높고, 산업체를 주관으로 하는 연구의 총 연구비가 높게 분석된다. 반면,

연구수행주체가 정부기관 및 학교일 경우 총 연구비가 감소한다.

<Table 3> Regression Equation Based on Total Research Funds

Division	Regression equation
All	$\ln(\hat{Y}) = 4.66-0.22x_{1a}-0.14x_{1b}+0.23x_{2a}+0.33x_{2b}-0.18x_{2c}-0.09x_{3a}+\cdots+0.01x_{11}$
BT	$\ln(\hat{Y}) = 4.48-0.18x_{1a}-0.09x_{1b}+0.16_{2a}-0.14x_{2c}+0.02x_{4}-0.59x_{5}-0.41x_{6}+\cdots+0.01x_{11}$
CT	$\ln(\sqrt{\hat{Y}}) = 6.60-2.74x_{1a}-1.33x_{1b}+0.08x_{4}+26.72x_{9}+0.01x_{11}$
ET	$\ln(\hat{Y}) = 4.46+0.48x_{2a}+0.11x_{2b}-0.37x_{2c}-0.61x_{5}+\cdots+0.01x_{11}$
IT	$\ln(\hat{Y}) = 4.01-0.42x_{1a}-0.34x_{1b}+0.76x_{1c}+0.23x_{2a}-0.02x_{2b}+0.02x_{5}+0.28x_{7}\cdots+0.01x_{11}$
NT	$\ln(\hat{Y}) = 5.59-0.48x_{1a}+0.11x_{2a}+0.01_{2b}+0.02x_{4}-1.70x_{5}+0.20x_{6}+0.28x_{7}+\cdots+0.01x_{11}$
ST	$\ln(\hat{Y}) = 4.23+0.04x_{1a}-0.56x_{1b}+0.02x_{4}+7.90x_{9}+0.02x_{11}$

3.2.2 의사결정나무분석

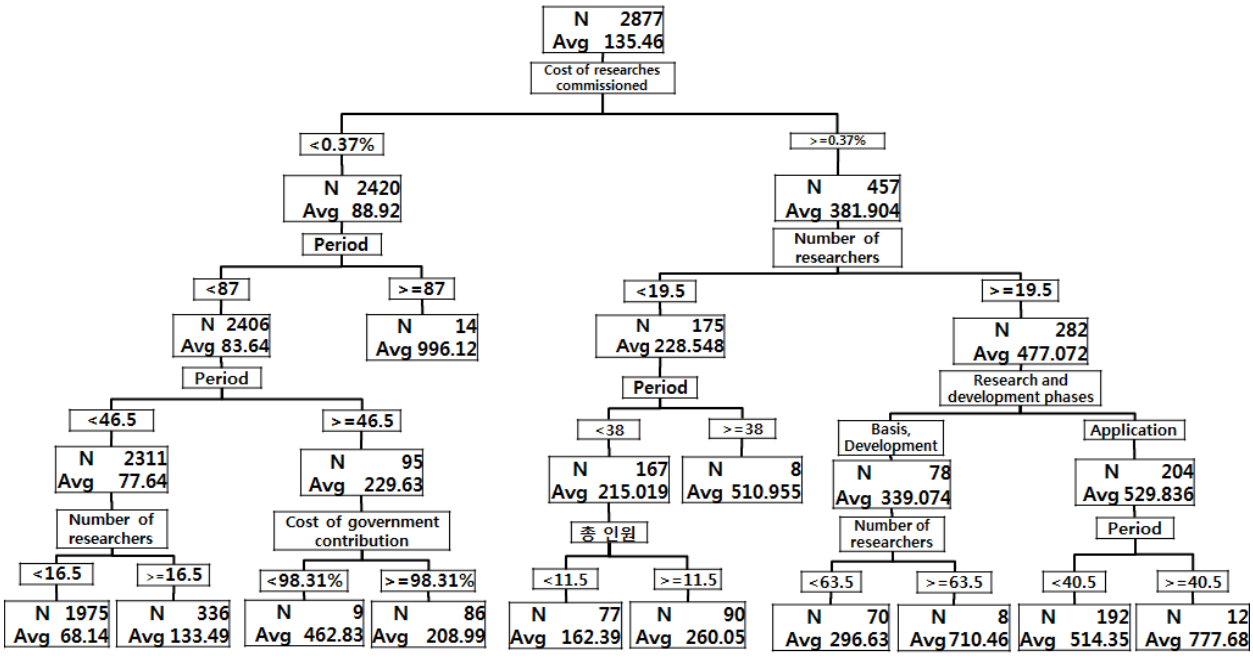
의사결정나무 분석에서는 목표변수에 변화를 취하지 않고 분석을 실시하였다.

분류기준은 자식마디 관측치 최소개수를 5, 자식마디가 형성될 때 고려될 최대의 분리개수 2, 전체 트리의 최대길이를 6단계로 설정하여 모형을 구축한 결과 평균 오차 제곱 합이 더 이상 낮아지지 않은 횟수에서 정지규칙에 의해 23개의 끝 마디를 형성하였으며, CART와 CHAID옵션 두 가지를 사용하여 분석하였다. 두 옵션 중 우수한 결과를 나타내는 옵션으로 결과를 해석하였고, 아래의 <Table 4>는 전체 데이터 및 분야별 의사결정나무분석 요약통계량을 나타냈으며, 전체 데이터의 RMSE는 83.7950이며, 결정계수는 0.8179로 위 모형은 전체 데이터의 81.79%를 설명한다고 할 수 있다. 또한, 각 분야별 평균과 RMSE, R-square를 구하였으며, CART와 CHAID분석 결과 중 RMSE가 낮은 것을 표로 정리하였다.

<Table 4> Summary Statistics of Decision Tree Analysis

Division	Option	N	Average	RMSE	R-square
All	CART	2877	135.46	83.80	0.82
BT	CHAID	1518	127.77	73.73	0.82
CT	CHAID	47	105.52	39.70	0.78
ET	CART	361	112.92	108.21	0.65
IT	CHAID	480	137.15	82.52	0.87
NT	CHAID	432	174.33	130.59	0.89
ST	CHAID	39	121.80	132.17	0.54

전체 데이터로 의사결정나무 분석을 한 결과 아래의 <Figure 7>과 같으며, 연구비의 전체 평균은 약 1억 3,546만원이고, 위탁연구비를 기준으로 첫 번째로 시작하여 과제기간과 연구에 투입되는 인원으로 하여 나머지 나무를 분류하였다.



<Figure 7> Decision Tree of Entire Data

이는 위탁연구비 비율이 총 연구비에 영향을 미치는 정도가 높음을 알 수 있으며, 연구에 참여하는 인원과 과제기간에 따라 연구비의 변동이 큰 것을 알 수 있다.

아래의 <Table 5>는 전체 데이터의 의사결정나무에 대한 요약표이다. 위탁연구비 비율이 0.37% 이상이고, 총 인원이 19.5명 이상이며, 연구개발단계가 응용, 과제기간이 40.5개월 이상일 경우의 연구일수록 총 연구비가 증가한다.

<Table 5> Summary Statistics of Decision Tree Analysis Using Entire Data

Division	Condition		Result
Cost of researches commissioned < 0.37%	Period < 87month	Period < 46.5month and Number of researchers < 16.5 person	N = 1975
			Avg = 68.41
		Period < 46.5month and Number of researchers ≥ 16.5 person	N = 336
			Avg = 133.49
		Period ≥ 46.5month and Cost of researches commissioned < 98.31%	N = 9
			Avg = 462.83
		Period ≥ 46.5month and Cost of researches commissioned ≥ 98.31%	N = 86
			Avg = 208.99
	Period ≥ 87month	-	N = 14
			Avg = 996.12
Cost of researches commissioned ≥ 0.37%	Number of researchers < 19.5 person	Period < 38month and Number of researchers < 11.5 person	N = 77
			Avg = 162.39
		Period < 38month and Number of researchers ≥ 11.5 person	N = 90
			Avg = 260.05
		Period ≥ 38 person	N = 8
			Avg = 510.96
	Number of researchers ≥ 19.5 person	Development(x_{1b}), Basis(x_{1a}) and Number of researchers < 63.5 person	N = 70
			Avg = 296.63
		Development(x_{1b}), Basis(x_{1a}) and Number of researchers ≥ 63.5 person	N = 8
			Avg = 710.46
		Application(x_{1c}) and Period < 40.5month	N = 192
			Avg = 514.35
		Application(x_{1c})and Period ≥ 40.5month	N = 12
			Avg = 777.68

또한, 회귀분석과 같이 연구 분야별 의사결정분석을 수행하였으며, 결과는 아래와 같다.

<Table 6>은 BT에 대한 분석결과로, 연구장비의 비율이 높을수록 연구비가 증가하는 것을 알 수 있으며, 응용연구이고 과제기간이 길어질수록 총 연구비가 증가한다.

<Table 6> Summary Statistics of Decision Tree Analysis Using BT Data

Division	Condition		Result
Cost of researches commissioned < 0.21%	Period < 87month	Number of researchers < 16.5 person and Period < 43 month	N = 983
			Avg = 67.27
		Number of researchers < 16.5 person and Period ≥ 43 month	N = 49
			Avg = 157.20
		Number of researchers ≥ 16.5 person and Period < 55.5 month	N = 236
			Avg = 135.29
		Number of researchers ≥ 16.5 person and Period ≥ 55.5 month	N = 15
			Avg = 289.33
	Period ≥ 87month	-	N = 6
			Avg = 1129.65
Cost of researches commissioned ≥ 0.21%	Number of researchers < 23.5 person	Number of researchers < 13.5 person and Research equipments and materials cost < 46.8%	N = 43
			Avg = 128.34
		Number of researchers < 13.5 person and Research equipments and materials cost ≥ 46.8and	N = 16
			Avg = 221.00
		Number of researchers ≥ 13.5 person	N = 53
			Avg = 258.20
	Number of researchers ≥ 23.5 person	Application(x_{1c}), Basis(x_{1a}) and Period < 40.5month	N = 89
			Avg = 486.25
		Application(x_{1c}), Basis(x_{1a}) and Period ≥ 40.5month	N = 7
			Avg = 858.67
		Development(x_{1b}) and Cost of researches commissioned < 12.5%	N = 11
			Avg = 321.49
		Development(x_{1b}) and Cost of researches commissioned ≥ 12.5%	N = 10
			Avg = 152.11

CT의 경우 아래의 <Table 7>에서와 같이 연구개발 단계가 응용단계일 경우 총 연구비가 높게 측정되며, 연구장비/재료비는 총 연구비에 유의한 영향을 받지 않은 것을 알 수 있다. 반면, CT 연구는 투입되는 인원에 따라 연구비의 차이가 난다.

<Table 7> Summary Statistics of Decision Tree Analysis Using CT Data

Division	Condition		Result
Basis (x_{1a})	Period < 34month	Number of researchers < 17 person	N = 29
			Avg = 55.69
		Number of researchers ≥ 17 person	N = 6
			Avg = 85.79
	Period ≥ 34month		N = 7
			Avg = 112.54
Application (x_{1c})	-		N = 5
			Avg = 408.4

ET의 경우, 정부에서 수행하는 연구가 많으므로, 정부출연금과의 관련성이 높으나, 정부출연금의 비율이 높을수록 연구비는 적어지는 경향을 보이고 있다. 아래의 <Table 8>은 ET 데이터의 의사결정나무에 대한 결과를 해석한 것이다.

<Table 8> Summary Statistics of Decision Tree Analysis Using ET Data

Division	Condition		Result
Cost of researches commissioned < 1.5%	Period < 46.5 month	Cost of government contribution < 60.6%	N = 5
			Avg = 408.96
		Cost of government contribution ≥ 60.6 % and Industry(x_{2a})	N = 54
			Avg = 130.37
		Cost of government contribution ≥ 60.6 % and University(x_{2d})	N = 259
			Avg = 66.01
	Period ≥ 46.5 month	-	N = 14
			Avg = 277.45
Cost of researches commissioned ≥ 1.5%	Period < 35.5 month	Cost of researches commissioned < 22.1% and Labor cost < 33%	N = 8
			Avg = 381.14
		Cost of researches commissioned < 22.1% and Labor cost ≥ 33%	N = 8
			Avg = 220.33
		Cost of researches commissioned ≥ 22.1%	N = 6
			Avg = 147.09
	Period ≥35.5 month	-	N = 8
			Avg = 644.06

<Table 9>는 IT 데이터의 의사결정나무 분석에 대한 결과를 해석한 것으로, IT 연구의 특성상 연구장비/재료비의 비율은 연구비에 큰 영향을 미치지 않았다. 반면, IT 연구의 경우 인건비의 비율이 높을수록 총 연구비가 증가하는 것을 알 수 있으며, IT의 기술 특성상 외부위탁을 요청하는 경우가 많으므로, 위탁연구비의 비율도 총 연구비에 영향을 준다.

<Table 9> Summary Statistics of Decision Tree Analysis Using IT Data

Division	Condition		Result
Development (x_{1b}), Basis(x_{1a})	Period < 52개월	Cost of researches commissioned < 2.1%	N = 383
			Avg = 68.41
		Cost of researches commissioned ≥ 2.1%	N = 336
			Avg = 133.49
		-	N = 15
			Avg = 194.6
	Period ≥ 52개월	Number of researchers < 14 person	N = 7
			Avg = 153.36
Application (x_{1c})	Number of researchers < 17.5명	Number of researchers ≥ 14 person	N = 5
			Avg = 352.99
		Cost of researches commissioned < 7.4%	N = 29
			Avg = 605.38
	Number of researchers ≥ 17.5명	Cost of researches commissioned ≥ 7.4%	N = 31
			Avg = 448.75

NT의 경우 아래의 <Table 10>에서 알 수 있듯이 IT와 유사하게 성격상 위탁 연구가 수행되어야 하므로, 위탁연구비가 총 연구비에 많은 영향을 준다. 또한, 인력이 증가할수록 총 연구비가 증가하는 것을 확인할 수 있다.

<Table 10> Summary Statistics of Decision Tree Analysis Using NT Data

Division	Condition		Result
Cost of government contribution < 89%	Number of researchers < 17.5 person	Number of researchers < 8.5 person	N = 5
			Avg = 82.12
		Number of researchers ≥ 8.5 person	N = 11
			Avg = 405.91
	Number of researchers ≥ 17.5 person	Number of researchers < 37.5 person	N = 33
			Avg = 559.63
		Number of researchers ≥ 37.5 person	N = 13
			Avg = 767.72
Cost of government contribution ≥ 89%	Cost of researches commissioned < 0.3%	Period < 46 month	N = 298
			Avg = 71.53
		Period ≥ 46month	N = 21
			Avg = 264.23
	Cost of researches commissioned ≥ 0.3%	Number of researchers < 22.5 person	N = 30
			Avg = 193.43
		Number of researchers ≥ 22.5 person	N = 21
			Avg = 443.91

아래의 <Table 11> ST의 의사결정나무분석에 대한 결과 해석으로, ST의 경우 연구 경비 및 간접비의 비율이 총 연구비에 영향을 미치고 있다. 또한, 연구에 참여하는 인원이 11명이 넘으면 연구비의 평균이 3억 원 이상이 된다. 이는 ST 연구의 경우 인건비가 총 연구비에서 매우 큰 영향을 미치고 있음을 의미한다.

<Table 11> Summary Statistics of Decision Tree Analysis Using ST Data

Division	Condition		Result
Number of researchers < 11명	Operating cost ≥ 7.7%	Period < 28 and Overhead cost < 16.6%	N = 12
			Avg = 56.97
		Period < 28 and Overhead cost ≥ 16.6%	N = 7
			Avg = 51.83
		Period ≥ 28 month	N = 8
			Avg = 74.58
	Operating cost < 7.7%	-	N = 5
			Avg = 177
Number of researchers ≥ 11명	-		N = 7
			Average = 317.43

3.2.3 기법간 분석

지금까지 선형회귀분석과 의사결정나무분석을 통하여 결과를 해석하였다. 본 장에서는 선형회귀분석의 Ad R-

Square 값과 의사결정나무분석의 R-Square 값 그리고 선형회귀분석과 의사결정나무분석의 RMSE 값을 이용하여 예측력이 우수한 방법을 찾아보고자 한다.

아래의 <Table 12>는 선형회귀분석과 의사결정나무분석간의 R-Square 값과 RMSE 값을 비교한 표를 나타낸 것으로, CT와 IT의 경우 선형회귀분석이 가장 예측력이 높은 것으로 나타났으나, 그 외 나머지 연구에서는 의사결정나무분석의 예측력이 높게 나타났다.

<Table 12> Comparison Table of R-square and RMSE

Method of analysis			Entire	BT	CT
Linear Regression	Full Model	Ad R-Square	0.65	0.64	0.85
		RMSE	355.97	137.99	39.99
	Step wise	Ad R-Square	0.65	0.64	0.86
		RMSE	355.97	139.32	**36.71***
Decision Tree	CART	R-Square	0.82	0.82	0.67
		RMSE	**83.80***	77.40	39.95
	CHAID	R-Square	0.81	0.78	0.67
		RMSE	84.52	**73.73***	39.70

Method of analysis			IT	NT	ST	ET
Linear Regression	Full Model	Ad R-Square	0.82	0.77	0.78	0.63
		RMSE	77.36	233.62	1007.09	125.73
	Step wise	Ad R-Square	0.83	0.77	0.74	0.63
		RMSE	**77.23***	239.68	184.24	124.31
Decision Tree	CART	R-Square	0.90	0.92	0.54	0.65
		RMSE	83.54	139.19	132.17	**96.40***
	CHAID	R-Square	0.87	0.89	0.54	0.63
		RMSE	82.52	**130.59***	**132.17***	108.21

4. 결론 및 향후 연구방향

선형회귀분석과 의사결정나무분석을 통하여 전반적인 결론을 도출한 결과, 대부분의 연구에서는 연구개발단계가 응용단계일 경우 총 연구비가 많았으며, 과제기간 및 참여인원이 증가할수록 총 연구비가 증가하는 추세를 나타냈다. 이는 장기 대형 과제일수록 연구비 규모가 커지는 현상에 기인하는 것으로 판단된다.

BT의 경우, 연구장비/재료비에 따른 총 연구비 변화가 큰 것으로 나타났다. 이는 해당 분야 연구에 고가의 장비 및 재료가 필요하기 때문으로 분석된다. 아울러 위탁연구비의 경우도 합동 연구가 많이 필요한 분야 특성 때문에 영향이 큰 것으로 분석된다.

CT의 경우는 인건비가 총 연구비에 유의한 영향을 끼치고 있는 것으로 나타났으며, ET의 경우, 총 연구비가 개발단계와는 관계가 적은 것으로 나타났다.

ET, IT와 NT의 경우, 위탁연구비에 따른 총 연구비 변화가 큰 것으로 나타났으며, 이는 위탁연구비를 책정하기 위해서는 일정 규모 이상의 총 연구비가 있어야 하는 현실이 반영된 것으로 판단된다.

ST의 경우, 연구인원 위탁연구비에 따른 총 연구비 변화가 큰 것으로 나타난 반면, 연구경비 및 간접비등 세부비용에 대해서는 낮은 수준의 영향을 보이고 있다.

본 연구에서는 이와 같이 전체 데이터 및 연구 분야별 데이터를 활용하여 총 연구비에 영향을 미치는 요인들 사이의 관계 추정을 위한 연구를 수행하였으며, 각 변수별 연구비에 영향을 주는 요인을 찾아냈으며, 연구비 산정에 영향을 미치는 결정요인의 상대적 중요성을 파악하고 접합한 예측 모형을 제시하여 연구비 산정 시 실질적인 도움을 주기 위하여 선형회귀분석 및 의사결정나무분석을 통하여 어떠한 변수가 연구비에 영향을 미치는지 알아보았다.

그러나, 본 연구의 데이터는 결측치가 많아 누락된 정보가 많으므로, 유의미한 통계적인 모형의 수립이 어려운 한계점을 보이고 있다.

또한, 본 연구에서는 과제기간에 영향을 보기 위하여 목표변수를 총 연구비로 두고 분석을 수행하였다. 그러나, 본 연구의 결과에서 과제기간에 따라 연구비가 증가한다는 결론은 일반적인 결론이므로, 과제기간에 영향을 주지 않도록 총 연구비를 월별 연구비로 계산하여 분석을 수행해야 할 것이다.

References

[1] An, K.H., A Study on Analysis of Cost Behavior and Cost Estimation in the Hotel Industry. *Journal of Tourism Science Society of Korea*, 2000, Vol. 15, No. 1, p 133-152.

[2] Bae, S.T., *A Study on outage Cost Assessment of Power analysis system*. Korea Institute of S&T Evaluation and Planning, 2011.

[3] Cho, H.H., Seong, J.E., Jung, B.K., and Jang, J.H., *Governance of the R&D Budget allocation*. Science and Technology Policy Institute, 2006.

[4] Douglas, C.M., Elizabeth, A.P., and Geoffrey, G.V., *Introduction to linear regression analysis* (3rded), New York : Wiley Interscience, 2001, p 138-154.

[5] DiMasi, J.A., Hansen, R.H., and Grabowski, H.G., The price of innovation : new estimates of drug development costs. *Journal of Health Economics*, 2003, Vol. 22, No. 2, p 151-185.

[6] Eo, W.J., Lee, Y.B., and Kang, S.J., Developing the R&D CER using Historical R&D Data in Korea. *Journal of Society of Korea Industrial and Systems Engineering*, 2010, Vol. 33, No. 3, p 55-62.

[7] Jung, H.J., A Study on Outage Cost Assessment of Power Supply [master's thesis]. [Jinju, Korea] : Gyeongsang National University, 2011.

[8] Kim, T.G., A Study on the Adjustment and Distribution System of National R&D project Budget [master's thesis]. [Seoul, Korea] : Chung-Ang University, 2010.

[9] Paek, J.G. and Rim, S.C., Product Cost Estimation using Integrated BOM in PDM. *Journal of Society of Korea Industrial and Systems Engineering*, 1999, Vol. 35, No. 2, p 231-242.

[10] Suh, S.D. and Kwon, K.J., Application of Monte Carlo Simulation to Efficiently Estimate Highway Investment Cost.,*Proceedings of the Korea Society for Simulation 98 Conference*, 1998, Sejong Univ, p 61-64.

[11] Song, Y.C. and Park, H.S., A Study on the Estimation of Farmland Price Using Spatial Econometrics Approach : Focused on Urban Fringe in Seoul Metropolitan Area. *Journal of Korea Research for Human Settlements*, 2012, Vol. 72, p 121-140.

[12] Wiesenthal, Y., Leduc, G., Haegeman, K., and Schwarz, H.-G., Bottom-up estimation of industrial and public R&D investment by technology in support of policy-making : The case of selected low-carbon energy technologies. *Research Policy*, 2012, Vol. 41, No. 1, p 116-131.

[13] Yoo, I.S., Seo, Y.K., Lee, J.H., and Oh, B.S., Application of Cost Estimation to Space Launch Vehicle Development Program. *Journal of Society of Korea Industrial and Systems Engineering*, 2007, Vol. 30, No. 3, p 165-173.

[14] Zhu, W.-J., Feng, W.-F., and Zhou, Y.-G., The application of genetic fuzzy neural networkin project cost estimate. *E-product E-Service and E-Entertainment (ICEEE) International Conference*, 2010, p 1-4.