□ 研 究 論 文 □

韓國國防經營分析學會誌
第36卷 第3號, 2010. 12. 31.

다양한 데이터 특성을 고려한 무기체계 비용추정관계식 개발 연구

(A Study On Developing Weapon System CERs With Considering Various Data Characteristics)

† 정 원 일(Wonil Jung)*, 김 동 규(Dongkyu Kim)**, 강 성 진(Sungjin Kang)***

ABSTRACT

최근 국방 무기체계 획득 환경의 변화는 무기체계 획득비용의 효율적 집행이라는 측면에서 비용분석의 중요성을 더욱 강조하고 있다. 그러나 정책 및 제도적 측면에서 비용분석이 강조되고 있는 반면 비용분석을 위한 국내 기반여건은 매우 부족한 실정이다. 국내에서의 비용추정은 주로 사업초기부터 국외에서 도입한 비용추정 전산모델을 사용하고 있으나 국내 방산환경에 적합하지 않은 많은 제약사항을 가지고 있다. 이러한 이유로 최근 한국형 비용분석 전산모델을 개발하고자 하는 공감대가 형성되었으며 체계적인 연구가 현재 진행되고 있다. 따라서 본 연구에서는 한국형 비용분석 전산모델의 핵심 논리인 비용추정관계식 개발 방법과 절차를 제안하고 있다. 특히 데이터가 가지는 각각의 회귀적 한계, 즉 다중공선성, 이상치, 이분산성 등을 식별하고 이에 적합한 회귀방법을 선택함으로서 데이터의 특성을 고려한 최선의 회귀모형을 구축하는 방법 및 절차를 제안하고자 한다. 제안한 방법은 국내 포병 무기체계 연구개발 자료를 기초로 비용추정관계식 개발방법 및 절차에 대한 이론적 적용가능성을 사례를 통해 검증하였다.

ABSTRACT

Recently, the acquisition environment of the Korean defense weapon system is emphasizing more the importance of cost analysis in terms of efficient execution for defense acquisition budget. While cost analysis, however, is emphasized in law and process, its infrastructures are still insufficient We have been using computerized cost models to obtain an estimate at early phase of project. But those models have been developed by foreign companies, and so they have many limitations when using in Korean defense environment. For this reason, it began to sympathize that we need the development of the Korean version cost estimation model suitable for our defense industry environment, and now many studies are proceeding.

In this study, we suggest Cost Estimating Relationships(CERs) developing methodologies which is key logics of Korean version cost estimation model. Especially, we proposed a new CER's development process depending upon data characteristics such as, multicolinearity, outlier, small samples and heteroscedasticity. Also, we presented a case study for artillery weapon system using these methods we developed. We find that these CERs could be verified through theoretical methods.

Keywords : Regression Analysis, CER(Cost Estimation Relationship), Multicolinearity, Outlier, Heteroscedasticity

논문접수일 : 2010년 11월 9일 심사(수정)일 : 2010년 11월 19일 논문게재확정일 : 2010년 11월 29일

* 국방대학교 운영분석학과 석사과정

** 국방대학교 운영분석학과 박사과정

*** 국방대학교 운영분석학과 교수

† 교신저자

1. 서론

최근 국방 무기체계 획득 환경의 변화는 국방개혁과 더불어 국민들의 사회, 복지, 경제 등에 대한 요구 증가로 인해 국방예산의 효율적 집행을 더욱 부추기고 있다. 특히, 국방예산중 방위력개선비용은 매년 9~10조원이라는 방대한 예산이 투입되고 있다. 이는 단일 무기체계의 개발 및 양산에만 수백억원에서 수조원의 예산이 지출된다는 점을 감안할 때, 소요제기 이전단계부터 무기체계의 획득비용을 정확히 추정하는 것이 국방예산의 효율적 배분 및 통제라는 측면에서 얼마나 중요한지를 공감하고 인식하게 한다. 이에 따라 국방부 및 방위사업청에서는 무기체계 소요 및 획득관련 규정을 통해 중기전력소요 요청시부터 양산에 이르기까지 비용분석을 의무화하는 등 비용분석을 제도화하는 노력을 경주하고 있다[2, 7, 8]. 특히 최근에는 무기체계의 소요 요청단계로부터 폐기단계까지의 총소요비용(TOC : Total Ownership Cost)을 최소화하고 장비가동률을 향상시키기 위한 총수명주기체계관리(TLCCSM : Total Life Cycle Cost System Management)를 정립하고자 노력하고 있다[3].

한편 정책 및 제도적 측면에서 비용분석이 강조되고 있는 반면 비용분석을 위한 국내 기반여건은 매우 부족한 실정이다. 국내에서의 비용추정은 주로 사업초기부터 국외에서 도입한 비용추정 전산모델을 사용하고 있으나 국내 방산환경에 적합하지 않은 많은 제약사항을 가지고 있다. 이러한 이유로 최근 한국형 비용분석 전산모델 개발의 필요성에 대한 공감대가 형성되었으며, 이를 개발하기 위한 이론적, 개념적 연구가 현재 활발히 진행되고 있다.

따라서 본 연구에서는 한국의 연구개발 데이터를 사용하여 비용추정관계식 개발 방법과 절차를 제안하며, 특히 데이터가 가지는 특성, 즉 다중공선성(Multicolinearity), 이상치(Outlier), 이분산성(Heteroscedasticity) 등을 식별하고 이에 적합한 회귀방법을 선택하여 최선의 회귀모형을 구축하는 방법 및 절차를 제안하였다.

이를 위해 기존 연구고찰에서는 데이터 특성을 분석하고 식별하는 방법, 다양한 회귀분석 방법론을 간단히 소개하고 이를 기반으로 적절한 비용추정관계식을 도출하는 개발 절차를 제안한다. 마지막으로 국내 포병 무기체계 연구개발 데이터를 기초로 비용추정관계식 개발방법 및 절차에 대한 이론적 적용가능성을 사례를 통해 검증하였다.

2. 기존 문헌연구

2.1 CER 개발에 관한 기존 연구고찰

일반적인 비용추정 모델개발 절차에 관한 연구는 ISPA Parametric Estimating Handbook 2007, 미 보잉사 비용추정모델, COCOMO 2, 국방 SW 비용추정 모델 개발방법론 등에서 각각의 특성에 맞는 비용추정 모델 개발 절차를 제시하고 있다[1, 15]. 그러나 위 연구들은 주로 비용추정 모델을 개발하는데 중점을 두고 포괄적으로 기술하고 있어 데이터의 특성을 분석하고 적절한 회귀모형을 구성함으로써 비용추정모델의 핵심인 비용추정관계식을 개발하는 절차에 대해서는 자세하게 언급하지 않고 있다.

John A. Horak 외 3명은 항공기 System에 대한 사업데이터, 기술, 현재비용 등의 자료를 분석하여 비용인자를 도출하고 이를 독립변수로 하는 항공기 구성품과 장비에 대한 CER을 개발하였다. 그 결과 총 30개의 CER을 개발하였으며 이에 대한 세부내용으로 순환되는 하드웨어 CER 20개, 비순환적 기술적 하드웨어 CER 3개, 기술지원 CER 7개를 각각 개발하였다. 또한 이 연구에서 신뢰성있는 CER 구축을 위한 통계적 평가로 t-검정, R^2, SE 결과값을 제시하였다[17].

이재용 외 7명은 국내 연구개발 데이터의 부족

을 극복하기 위해 전문가 설문을 통해 유도무기체계 국방연구개발 비용추정모델을 개발하였다. 이는 전문가의 설문을 통해 비용인자의 종류, 범위, 값들을 산정하고 이를 토대로 하위체계별 연구개발 비용추정모델을 개발하였다. 비용추정관계식은 베이지안 통계분석법을 적용하였으며 모수들이 확률분포를 따른다는 가정하에 구체적인 비용분포를 도출하였다. CER 개발이 제한되는 하위체계들은 총 연구개발비용에서 하위체계들이 차지하는 일정비율로 할당하여 산출하였다[12].

백종문 외 3명은 국방 소프트웨어 비용추정분야에서 전문가를 대상으로 소프트웨어를 개발하기 위한 노력, 개발 환경 등에 대한 설문을 통해 비용인자를 식별하여 CER을 개발하였다[9].

전문가 설문을 기반으로한 CER 개발 방법은 국내 CER 개발 환경을 고려하였다는 점에서 의미가 있으나 전문가 설문을 통해 CER을 개발함에 따라 주관적 요소가 내재될 수 있어 CER에 대한 신뢰성이 떨어질 수 있다.

어원재, 이용복, 강성진은 데이터의 특성중 일부, 즉 데이터 수, 다중공선성을 고려한 비용추정관계식 개발 절차를 제시하고 이를 통해 기동무기체계 연구개발비 CER를 개발하였다[11]. 그러나 데이터의 특성을 다중공선성에만 한정함에 따라 데이터의 다양한 특성(이상치, 이분산성 등)을 고려하지 않았다는 한계를 가진다.

류민규, 이용복, 강성진은 획득비, 노후화율, 임무 특성 등 무기체계별 특성을 고려한 수리부속비 CER을 이용한 새로운 관점에서 항공기 및 장갑차 수리부속비를 추정하였다[5]. 하지만 다양한 데이터 특성을 고려하지 않고 단순히 다중선형 및 로그선형 회귀를 이용하여 CER을 개발하였다는 한계점을 가지고 있다.

따라서 본 연구는 무기체계 비용추정을 위해 수집된 데이터에서 나타나는 이상치 및 이분산성을 함께 고려하고 이에 적합한 회귀 모형을 개발하는 절차를 제시한다.

2.2 적용된 방법론에 관한 고찰

2.2.1 회귀분석에서의 데이터 특성

2.2.1.1 다중공선성

최소제곱추정법을 사용하는 다중선형회귀모형에서 독립변수간 상관관계가 높은 경우는 독립변수 행렬(X)의 한 열이 다른 한 열, 또는 다수의 열과 선형결합을 형성하게 되어 이론상 (X'X)의 역함수를 구할 수 없거나 어려울 뿐만 아니라 회귀모형의 신뢰성에도 큰 영향을 미친다.

그러나 무기체계 비용추정시 독립변수인 비용인자들(Cost Drivers)은 서로 높은 상관관계를 가지는 경우가 많다. 이처럼 다중선형회귀에서 독립변수들간 높은 상관관계가 존재하는 경우 다중공선성이 존재한다고 한다.

2.2.1.2 이상치

회귀분석에서 일부 데이터가 다른 데이터들에 비해 상당히 크거나 작은 값을 갖는 점들이 종종 생기게 되는데 이런 데이터들을 이상치라고 한다. 이상치들은 경우에 따라 회귀선을 추정하는데 큰 영향을 미칠 수 있으므로 이상치를 자료에 포함시킬 것인지 아니면 제거할 것인지를 결정하는 일은 회귀분석에서 중요한 의미를 갖는다.

2.2.1.3 이분산성

최소제곱추정량에 의한 선형회귀는 가정사항으로 오차항의 등분산을 전제하고 있다. 그러나 이러한 등분산의 가정이 성립되지 않는 경우도 있으며 오차항마다 분산이 다른 경우가 존재할 수 있다. 이를 오차항의 이분산성이라 한다. 단순 및 다중선형회귀에서 오차의 이분산성이 존재하면 선형회귀의 기본 가정사항을 위배하게 되므로 회귀식의 통계적 정확도는 떨어질 수 밖에 없다.

2.2.2 회귀분석(Regression Analysis) 방법

2.2.2.1 다중선형회귀 및 로그선형회귀

선형회귀는 독립변수와 종속변수간의 관계분석을 통해 여러 개의 변수들간에 함수관계를 규명하고자 하는 방법론으로 다음과 같은 가정사항에 근거하여 모형을 제시하고 있다[6].

첫째, 변수 x와 y사이에 존재하는 관련성은 주어진 x의 값에서 y의 기댓값을 μ_{yx}라고 할 때 $\mu_{yx} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ 와 같은 선형식으로 적절히 표현될 수 있다.

둘째, 주어진 x의 값에서 변수 y는 정규분포를 따르며 평균은 $\mu_{yx} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$로 x에 따라 변하나, 분산은 x의 값에 관계없이 일정하다 ($y \sim N(\mu_{yx}, \delta^2)$, $\varepsilon \sim N(0, \delta^2)$).

셋째, 독립변수 x는 오차없이 측정할 수 있는 수학변수이며 종속변수 y는 측정오차를 수반하는 확률변수이다. 또 y의 측정 오차들은 서로 독립이다($Cov(\varepsilon_i . \varepsilon_j)=0$, $i \neq j$).

즉, 선형회귀는 데이터가 가지는 기본적인 가정사항을 충족하는 조건하에서 유도되는 모형으로 데이터가 이러한 가정사항을 위배하게 되면 모형의 정확성 및 신뢰성을 보장할 수 없게 된다.

로그선형회귀는 회귀함수를 선형식에서 찾을 수 없는 경우 적절한 수학적 변환, 즉 독립변수와 종속변수에 로그함수를 취함으로써 비선형함수를 선형함수로 전환하여 회귀식을 구하는 방법이다. 기본적인 가정사항 및 최소제곱추정법 등에 대한 기본적인 원리는 선형회귀와 동일하다.

2.2.2.2 주성분회귀 및 능형회귀

독립변수간 다중공선성이 존재하는 경우 회귀모형의 신뢰성은 현저히 감소된다. 따라서 독립변수간 다중공선성을 제거하거나 감소시켜야하며 이를 해결하는 방법으로는 다음의 두 가지가 있다. 첫째, 상관관계가 존재하는 변수중 하나를 제거하는 방법이다. 이는 다중공선성을 완화할 수 있고 변수선택 과정에서 쉽게 적용할 수 있는 장점이 있는 반면 중요한 변수의 경우 제거할 수 없는 단점이 있다. 둘째, 주성분회귀분석, 능형회귀분석, 제한회귀분석 등과 같은 편의추정량을 사용하는 것이다[4, 14]. 이는 변수를 제거하지 않고 회귀적 방법을 활용하여 다중공선성의 문제를 해결할 수 있다는 장점이 있다.

2.2.2.3 로버스트 회귀

데이터에서 이상치가 식별되면 우선적으로 이상치가 존재하게 된 원인을 재분석하고 이를 제거할 것인지 또는 그대로 활용할 것인지를 결정하여야 한다. 이상치를 제거해도 회귀모형을 구성하는데 큰 문제가 없다면 제거하는 것이 가장 바람직하다. 그러나 데이터의 수가 적거나 이상치로 판명된 데이터가 중요하다고 판단될 경우 이상치를 포함하고 이에 대한 영향력을 감안하여 회귀모형을 구축해야 하는데 이때 사용되는 회귀방법이 로버스트 회귀이다.

2.2.2.4 가중 회귀

앞서 선형회귀모형에서 가정하였듯이 오차항은 등분산성을 가져야 한다. 그러나 이러한 등분산의 가정이 성립되지 않고, 오차항마다 분산이 다른 이분산성을 가질 수 있다. 이런 경우 가중최소제곱법(Weighted Least Square)을 사용하여 회귀모형을 구성하는 가중회귀 분석을 실시해야 한다.

2.2.2.5 Adaptive CER

Adaptive CER(Cost Estimation Relationship)은 가중회귀를 이용한 비용추정관계식으로서 가중회귀에서는 분산을 기준으로 가중치를 부여하는 반면 Adaptive CER에서는 데이터의 질과 신뢰도, 데이터의 그룹화, 기준값으로부터의 거리 등을 기준으로 가중치를 부여한다[17].

Adaptive CER은 가중치를 부여하는 방법에 따라 각 데이터가 정확도 측면에서 얼마나 신뢰할

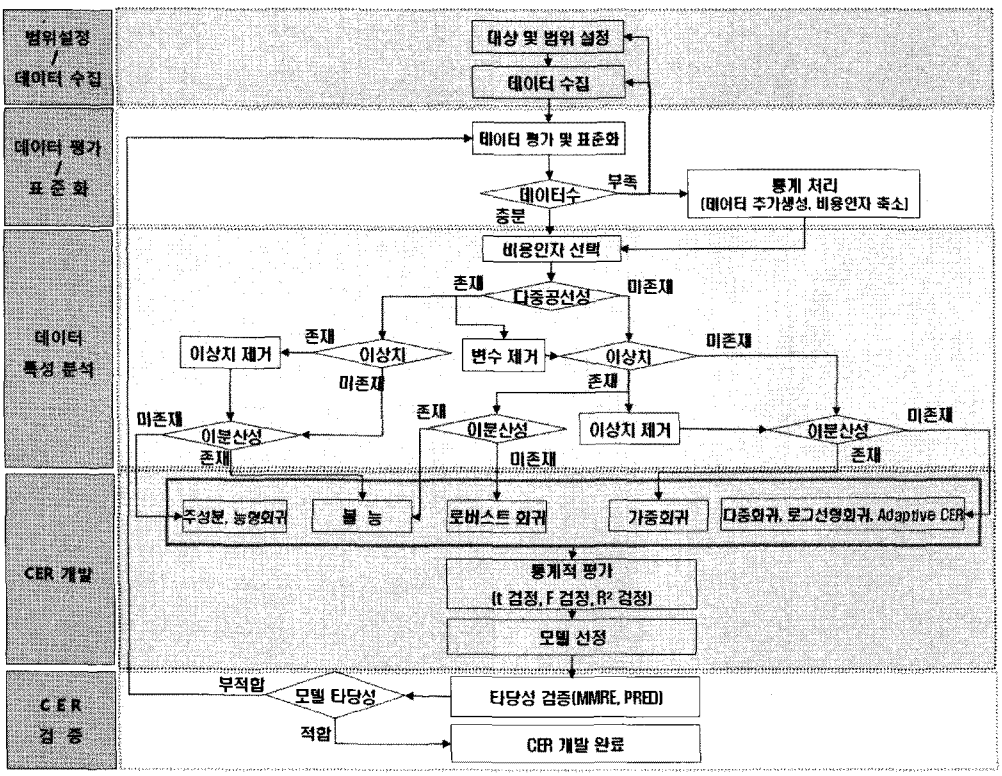
수 있는지에 따라 가중치를 부여하는 Priori Method, 데이터들이 구분된 하위 그룹들을 형성하고 있을 때 동일한 하위그룹의 데이터에 가중치를 부여함으로써 각각의 하위그룹에 대한 개별 회귀모형을 구성하는 Piecewise Method, 기준이 되는 특정값을 설정하고 실데이터가 이 값에서 얼마나 떨어져 있는지를 평가하여 가중치를 부여하는 X-Distance Method가 있다.

3. 비용추정관계식 개발 절차

본 장에서는 수집된 데이터의 다양한 특성을 기준으로 적절한 회귀방법을 찾고 실행가능한 회귀방법중 가장 적합한 비용추정관계식을 선택하는 절차를 제안하고자 한다. 제안된 비용추정관계식 개발절차는 <그림 1>과 같으며 세부적인 내용은 아래와 같다.

3.1 비용추정 대상 및 범위 설정

비용추정관계식 개발을 위한 첫 단계는 비용을 추정하려는 대상과 범위를 명확히 정의하는 것이다. 대상과 범위가 불명확하면 데이터의 동질성이 모호해 질 수 있으며, 이는 최종적으로 비용추정관계식의 정확성 및 신뢰성을 떨어드린다. 따라서 추정하려는 비용의 대상과 범위, 즉 작업세부구조(WBS : Work Breakdown Structure) 2~3 단계 수준의 부체계 또는 구성품 등 어떤 수준에서 비용을 추정할 것인지, 체계의 동질성 범위를 어디까지 정의할 것인지, 개발비, 양산비, 운영유지비 등 어떤 비용을 추정할 것인지 등에 대해 명확하게 정의하는 것이 비용추정관계식 개발의 시작이라 할 수 있다.



〈그림 1〉 비용추정관계식 개발 절차

3.2 데이터 수집

설정된 비용추정 대상 및 범위내에서 데이터를 수집하기 이전에 수집 가능한 데이터의 존재 유무를 조사하여야 한다. 선진국의 경우 비용 DB(Database)를 체계적으로 관리함으로써 데이터의 가용여부를 쉽게 확인할 수 있으나 우리군의 경우 통합된 비용 DB가 없고, 기관별로 필요한 비용데이터를 분산·관리하고 있어 데이터의 가용성 여부를 우선적으로 분석하는 것이 필요하다. 만약 데이터가 가용하지 않다면 비용추정관계식 개발은 불가능하거나 또는 비용추정 대상 및 범위를 재조정할 필요가 있다. 데이터가 가용하다고 판단되면 원시데이터, 가공데이터 등 비용과 비용인자간의 관계를 나타낼 수 있는 모든 데이터를 수집해야 한다. 또한 데이터의 신뢰도를 보장하기 위해 데이터의 출처 등에 대한 근거가 함께 수집되어야 한다.

3.3 데이터 평가 및 표준화

수집된 데이터는 우선적으로 다음의 몇 가지 사항을 평가하고 조정해야 한다. 첫째, 수집된 데이터는 명확한 근거를 가지고 있으며 신뢰할 수 있는 데이터인지를 평가해야 한다. 둘째, 수집된 데이터가 가공된 데이터라면 비용과 비용인자간의 관계를 설명할 수 있는 형태로 재가공하거나 원시데이터를 추가 수집해야 한다. 셋째, 회귀기법을 위해 요구되는 최소한의 데이터수를 만족하는지 확인해야 하며, 만족하지 못할 경우 대상 및 범위를 재조정하거나 데이터의 확률적 임의 생성, 비용인자 축소 등을 고려해야 한다.

데이터 평가가 완료되면 데이터를 표준화해야 한다. 데이터의 표준화는 데이터의 일관성을 위해 하나의 기준을 정하는 것으로서 물리적 단위, 환율, 물가상승률, 임율, 학습율, 수량 등에 대해 기준을 정하고 이 기준에 따라 데이터를 재가공해야 한다.

3.4 비용인자(Cost Driver)의 선택

비용인자는 추정하려는 대상체계의 비용을 주도하는 요인으로서 적절한 비용인자를 선택하는 것은 매우 중요하다.

비용인자를 선택하는 방법은 정성적인 방법과 정량적인 방법으로 나누어 볼 수 있다. 정성적인 방법은 비용분석가 및 체계 전문가들의 의견수렴을 통해 요인을 식별하는 방법이며, 정량적인 방법은 가능한 모든 경우의 수를 조합하여 단순 또는 다중선형회귀를 수행함으로써 회귀식을 통해 판단하는 방법이다. 특히 정량적인 방법은 비용인자의 수와 비용인자의 종류에 따른 모든 조합을 회귀식으로 구현하고 결정계수(R^2), 수정된 결정계수(R^2_{adj}), Forward/ Backward/ Stepwise Selection, C_p 등의 수치값을 상호 비교하여 비용인자를 식별한다.

3.5 데이터 특성분석

비용인자가 선택되면 데이터의 특성을 분석한다. 본 연구에서는 데이터의 특성을 다중공선성, 이상치, 이분산성의 존재여부에 따라 가장 적합한 회귀방법을 선택하는 절차를 제안한다.

3.5.1 다중공선성 식별 및 해결

비용인자가 선택되면 비용인자간 다중공선성의 존재여부를 분석한다. 다중공선성을 식별하는 방법으로 분산팽창계수(VIF : Variation Inflation Factor)와 상태지수(CI : Condition Index)가 있으며 식은 아래와 같다.

$$VIF_j = \frac{1}{1-R_j^2},\ j = 1,2,\ldots,p \qquad \text{(식 1)}$$

$$c_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}\ \ j = 0,1,2,\ldots,p \qquad \text{(식 2)}$$

일반적으로 p개의 VIF_j 중 가장 큰 값이 10을 넘거나, 상태지표가 30이상일 때 독립변수간 선형관계, 즉 다중공선성이 존재한다고 볼 수 있다[10].

다중공선성이 존재할 경우 상관관계가 높은 변수중 중요도가 낮다고 판단되는 변수를 제거하거나 이상치 및 이분산성이 존재하지 않을 경우 주성분회귀 또는 능형회귀를 수행한다.

주성분회귀(Principal Component Regression)는 Z를 X의 표준화한 행렬이라고 하고, V를 각각의 eigenvalue에 해당되는 eigenvector의 열로 이루어진 행렬이라 할 때 (식 3)과 같이 표현되며,

$$Y = ZVV'\beta + \epsilon = (ZV)(V'\beta) + \epsilon = C\gamma + \epsilon \quad \text{(식 3)}$$

최소제곱추정량은 (식 4)와 같다.

$$\hat{\gamma} = (C'C)^{-1}C'Y \quad \text{(식 4)}$$

주성분(C)들은 서로 직교(독립)이기 때문에 다중공선성이 발생하지 않는다[4].

능형회귀(Ridge Regression)는 다중공선성이 존재할 때 이를 해결하려는 대안으로 불편추정량에 편의값을 부여함으로써 다중공선성을 감쇄시키고자 하였다. 능형추정량은 (식 5)와 같다[14].

$$b(k) = (Z'Z + kI)^{-1}X'Y \quad \text{(식 5)}$$

3.5.2 이상치 식별 및 해결

다음으로 이상치의 존재여부를 판단한다.

이상치를 식별하는 방법은 다양하나 가장 보편적으로 사용되는 표준화잔차 및 표준화제외잔차 분석 방법이 있다[6].

첫째, 표준화 잔차(studentized residual)는 (식 6)과 같이 표현되며 e^*의 절대값 $|e_i^*|$가 상대적으로 과다하게 크거나, Lund가 제시한 기각치를 초과하면 이상치라고 판단할 수 있다[16].

$$e_i^* = \frac{e_i}{S(e_i)} = \frac{e_i}{S(1-h_{ii})^{\frac{1}{2}}} \quad \text{(식 6)}$$

둘째, 표준화제외잔차(studentized deleted residual)는 (식 7)와 같이 표현된다.

$$d_i^* = \frac{e_i}{S(i)\sqrt{1-h_{ii}}} \quad \text{(식 7)}$$

i번째 측정값 y_i에 대해 $|d_i^*| \geq t(n-k-2;\alpha/2)$이면 유의수준 α에서 y_i를 이상치로 판정한다.

이상치가 존재할 경우 단순히 이상치를 제거할 수 있다. 그러나 데이터의 수가 적거나 또는 이상치로 식별된 데이터가 중요하다고 판단될 경우에는 이를 제거하는데 한계가 있다.

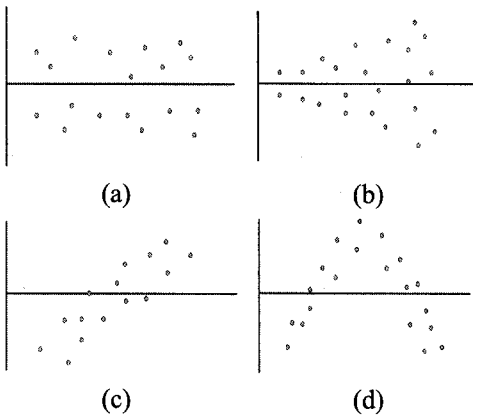
따라서 다중공선성 및 이분산성이 존재하지 않고 이상치만 존재할 경우 로버스트회귀를 수행한다. 로버스트 회귀는 M, LTS, S, MM방법이 있으며 여기서는 일반적으로 많이 사용되고 효율성이 높은 LTS와 MM을 간단히 소개한다[10].

첫째, LTS(Least Trimmed Square) 추정은 Rousseeuw가 제안한 것으로서 제곱 잔차들을 순위화하여 그 값이 큰 것들을 제외한 후 나머지 제곱잔차들의 합을 최소화하는 회귀계수를 추정하는 것이다.

둘째, MM 추정은 LST와 S 추정의 고붕괴값 추정과 M 추정의 고효율 추정을 동시에 추구한 것으로 S 추정보다 효율이 더 높다. 즉 LTS 또는 S 추정으로 계산된 척도와 회귀계수를 초기값으로 하여 최종 회귀계수를 M 추정하는 형식을 취한다.

3.5.3 이분산성 식별 및 해결

마지막으로 이분산성 존재여부를 식별한다.



〈그림 2〉 잔차의 산점도 예시

이분산성을 식별하는 일반적인 방법은 잔차의 산점도를 확인하는 것이다. <그림 2>에서 보는 바와 같이 (a)는 잔차가 불규칙하게 분포하고 있으나 (b), (c), (d)의 경우는 잔차가 일정한 패턴을 보이므로 오차항이 등분산이라고 볼 수 없다.

다중공선성 및 이상치가 없고 이분산성만이 존재할 경우 가장 적합한 회귀모형으로 가중회귀를 수행할 수 있다. 가중회귀모형은 (식 8)와 같고 추정된 회귀계수는 (식 9)와 같다.

$$\frac{y_{ij}}{w_j} = \beta_0 \frac{1}{w_j} + \beta_1 \frac{x_{1ij}}{w_j} + \ldots + \beta_n \frac{x_{nij}}{w_j} + \epsilon'_{ij} \qquad \text{(식 8)}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\Sigma w_i (x_i - \bar{x})(y_i - \bar{y})}{\Sigma w_i (x_i - \bar{x})^2} \qquad \text{(식 9)}$$

일반적으로 위에서 제시한 세 가지 데이터 특성중 두 가지 이상의 특성이 동시에 발견될 경우 이 두 가지 특성을 동시에 완화할 수 있는 적합한 회귀모형은 없다. 이 같은 경우 회귀적 기법을 사용할 수 없거나, 또는 데이터의 특성에 대한 한계를 전제로 회귀모형을 구현할 수 밖에 없다.

3.6 CER 개발

가용하다고 판단된 회귀방법들을 사용하여 구현된 회귀모형들은 t-검정, F-검정, R^2, R^2_{adj} 등과 같은 통계적 평가를 수행해야 한다. 평가값들이 일반적으로 인용되거나 또는 비용분석가가 설정한 기준을 충족하는지 확인해야 한다. 회귀모형의 후보들은 이러한 기준에 의해 일차적으로 선별되어야 하며 이후 상호 비교를 통해 가장 우수하다고 판단되는 회귀모형을 비용추정관계식으로 선택해야 한다.

그러나 가장 적합하다고 판단된 회귀방법을 사용하여 유도된 비용추정관계식 역시 예측력, 신뢰도, 정확성이 모두 월등이 높다고는 볼 수 없다. 또한 이 모형이 가장 적합한 모형이라고 단정지을 수 없다. 예를들어 다중공선성이 존재하여 주성분 회귀를 수행할 경우 회귀계수의 신뢰도는 높아지지만 R^2 또는 R^2_{adj} 값이 낮아지는 경우가 종종 있다. 따라서 비용분석가는 데이터의 제약사항을 인정하고 수행한 다른 회귀모형과 상호 비교를 통해 가장 합리적이라고 판단되는 회귀모형을 선택해야 하며 동시에 선택된 회귀모형의 제약사항을 명확히 이해하고 있어야 한다.

3.7 CER 검증

최종 단계로서 선택된 비용추정관계식은 모형의 추정값에 대한 정확도를 평가해야 한다. 모형의 추정값에 대한 타당성을 평가하는 방법은 일반적으로 MMRE(Mean Magnitude of Relative Error), PRED(Prediction)를 가장 많이 사용하며 MMRE와 PRED는 다음 식과 같다.

$$MMRE = \frac{1}{n} \sum_{i=1}^{n} \frac{| y_i - \hat{y_i} |}{y_i} \qquad \text{(식 10)}$$

$$PRED = \frac{k}{n}, \qquad \text{(식 11)}$$

(k : MRE $\leq \ell$ 에 해당하는 데이터의 수)

〈표 1〉 국내 연구개발 포병무기체계에 대한 표준화 데이터 현황

무기 체계	최대 사거리 (km)	구경 (mm)	중량 (kg)	전장 (cm)	최대발사 속도(분)	지속발사 속도(분)	연구 개발비 (억원, 2010년)
A	3.59	60	18	99	30	20	18.2027
B	1.8	60	21	82	30	18	12.7289
C	6.473	81	41	155	30	11	35.2546
D	4.737	81	81	130	12	5	17.8506
E	11.274	105	2,260	231	3	1	37.6372
F	14.7	105	2,650	392	5	2	27.0690
G	18	155	6,890	701	4	2	43.0712
H	18	155	25,000	912	4	1	74.0739
I	41	155	47,000	810	6	2	1,342.847

모형이 타당성 검증에서 적합한 것으로 결정되면 비용추정관계식을 채택한다.

그러나 이러한 다양한 검증을 거친 비용추정관계식 조차도 모형 개발자의 주관이 개입될 수 있는 소지를 안고 있으므로 공식적인 인증(Certification) 절차를 받는 것이 필요하다. 본 연구에서는 비용추정관계식 개발 절차에 한정하여 서술하며 이후의 절차에 대한 내용은 생략한다.

4. 사례분석

본 장에서는 제안된 비용추정관계식 개발방법 및 절차를 기반으로 실제 적용가능성을 판단하기 위해 국내 포병 무기체계 연구개발비용에 대한 사례 연구를 수행한다. 모든 방법 및 절차는 앞장에서 제안한 비용추정관계식 개발방법 및 절차를 그대로 적용하였으며, 통계분석도구로 SAS 9.1, MINTAB 14, Excel 2007 Package를 함께 사용하였다.

4.1 비용추정 대상 및 범위 설정

본 사례연구는 국내 포병 무기체계의 연구개발비에 대한 비용추정관계식 개발을 대상으로 한다. 포병 무기체계의 범위는 박격포, 견인곡사포, 자주곡사포로 한정하며, 한국적 방산환경에 부합하는 비용추정관계식을 개발하기 위해 모든 데이터는 국내 연구개발 실적 데이터를 기반으로 한다. 따라서 구매 및 국외 데이터는 비용수집 대상에서 제외한다. 추정범위는 연구개발 비용중 선행연구비, 사업관리 부서의 사업관리 비용을 제외한 탐색개발 및 체계개발비용으로 한정한다.

4.2 데이터 수집, 평가 및 표준화

설정된 비용추정 대상 및 범위에 따라 실제 연구개발을 수행한 기관을 방문해 관련 데이터 9종을 수집하였다. 연구개발비와 더불어 비용인자로 선택 가능한 무기체계 특성치들은 6가지, 즉 최대사거리, 구경, 중량, 전장, 최대발사속도, 지속발사속도가 수집되었다. 대상 무기체계의 국내 연구개발 실적이 적어 많은 데이터를 수집할 수는 없었으나 비용인자와 데이터의 수는 다양한 회귀기법을 사용하는데 무리가 없어 추가적인 데이터 수집이나 확률적 임의생성, 비용인자 축소 등은 수행하지 않았다.

수집된 데이터는 <표 1>에서처럼 비용인자에 대해 표준 척도로, 연구개발비에 대해서는 2010년을 기준년도로 전환하였다.

4.3 비용인자(Cost Driver)의 선택

〈표 2〉 독립변수 개수별 선택된 비용인자

변수의 개수	R^2	R^2_{adj}	Cp	최대사거리	중량	구경	전장	최대발사속도	지속발사속도
1	0.7886	0.7584	52.1937		○				
	0.7707	0.738	57.0321	○					
	0.2589	0.153	195.5042				○		
2	0.9478	0.9303	11.1332	○		○			
	0.9077	0.877	21.9617		○		○		
	0.9027	0.8703	23.3222	○			○		
3	0.9837	0.974	3.397	○	○		○		
	0.97	0.952	7.1153	○	○	○			
	0.959	0.9344	10.0941	○		○		○	
4	0.9904	0.9807	3.6052	○	○		○		○
	0.9902	0.9804	3.6502	○	○		○	○	
	0.9841	0.9682	5.3077	○	○	○	○		
5	0.9924	0.9797	5.0622	○	○	○	○		○
	0.9909	0.9757	5.4658	○	○	○	○	○	
	0.9908	0.9754	5.4923	○	○		○	○	○
6	0.9926	0.9704	7	○	○	○	○	○	○

종속변수, 즉 연구개발비용에 높은 연계성을 보이는 비용인자를 추출하기 위해 비용인자의 수와 비용인자의 종류에 따른 모든 조합을 회귀식으로 분석하였으며 결과는 <표 2>와 같다.

표에서 보는 바와 같이 변수가 3개 이상에서 95%이상의 높은 R^2, R^2_{adj} 값이 도출되었으며, 변수 3개중 일부와 변수 4, 5개의 경우 Cp값도 6이하로 상대적으로 좋은 값들을 보였다. 또한 비용인자들은 주로 최대사거리, 중량, 구경, 전장의 선택도가 기타 비용인자보다 높았다. 이상과 같은 정량적인 분석결과와 함께 최종적으로 전문가들의 의견수렴을 통해 최대사거리, 중량, 구경, 전장의 4개 비용인자를 선택하였다.

4.4 데이터 특성분석

4개의 비용인자와 9개의 데이터를 기준으로 비용인자간 다중공선성, 이상치 및 이분산성 존재유무를 분석하였다. 먼저 4개의 비용인자는 비용인자간 상관관계 분석<그림 3>을 통해 다중공선성을 의심할 수 있었으며, <표 3>에서 보는 바와 같이 VIF 및 CI로 다중공선성이 존재함을 확인할 수 있었다.

	최대사거리(㎞)	구경(㎜)	중량(㎏)
구경(㎜)	0.827 0.006		
중량(㎏)	0.926 0.000	0.740 0.023	
전장(㎝)	0.818 0.007	0.964 0.000	0.804 0.009

셀 내용: Pearson 상관 계수
P-값

〈그림 3〉 비용인자간 상관관계

다중공선성이 존재하는 경우 변수를 제거하거나 주성분회귀 또는 능형회귀를 수행할 수 있다. 변수 제거는 최대사거리와 전장이 나머지 변수와 모두 높은 상관관계를 가지므로 이들을 제거한 구경과 중량만을 비용인자로 하는 회귀모형을 구현할 수 있다.

다음으로 표준화잔차와 표준화제외잔차 등을 이용하여 이상치를 분석한 결과 <표 4>와 같이 I

〈표 3〉 비용인자별 VIF 및 CI

구분	최대 사거리	구경	중량	전장
VIF	14.41	24.83	9.53	39.98
CI	2.73	6.63	20.76	37.41

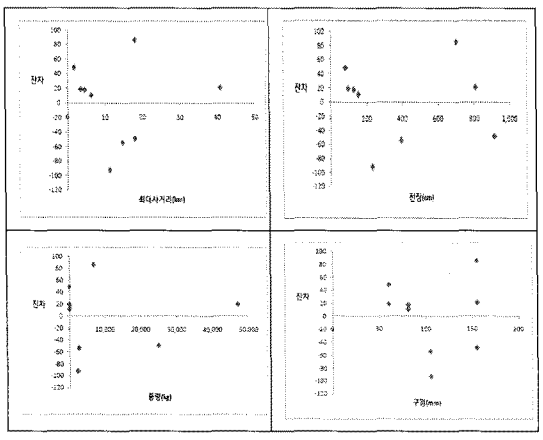
〈표 4〉 데이터의 표준화 잔차, 표준화제외 잔차 등 이상치 판단 자료

무기체계	표준화 잔차	표준화 제외잔차	Hat Diag	Cov Ratio	DFFITS	DFBETAS				
						상수	최대사거리	구경	중량	전장
A	0.145	0.126	0.4064	6.9137	0.1043	0.0837	0.025	-0.065	-0.0233	0.052
B	0.676	0.6218	0.3578	3.5791	0.4642	0.2388	-0.0882	-0.122	0.0879	0.0747
C	0.151	0.1313	0.1811	5.0005	0.0618	0.0006	-0.0066	0.0116	0.0069	-0.018
D	0.353	0.3109	0.337	5.424	0.2216	-0.0928	-0.1293	0.1399	0.1287	-0.1508
E	-1.524	-2.0391	0.5286	0.1156	-2.1591	1.5368	0.6477	-1.6795	-0.716	1.7418
F	-1.48	-1.9048	0.6905	0.2586	-2.8453	-1.7583	-2.4362	1.995	2.5839	-1.9422
G	1.89	5.0151	0.6219	0.0002	6.4319	-1.863	1.0166	1.0368	-2.5922	0.342
H	-1.942	-7.0574	0.894	0	-20.4945	-0.4367	10.4762	-0.528	-6.6161	-4.2224
I	1.992	19.0875	0.9827	0	143.9452	2.7271	30.4635	-3.8353	24.8136	-16.3969

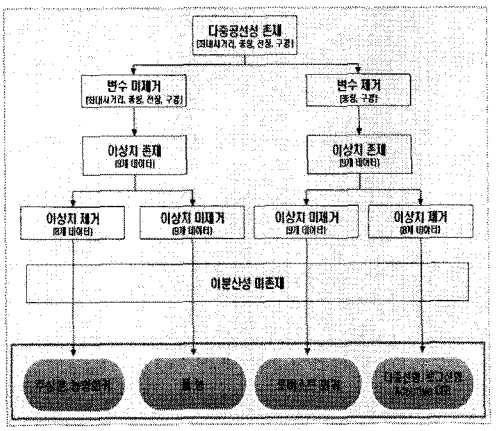
무기체계가 이상치로 식별되었다. 이상치로 식별된 I 무기체계는 이를 제거하고 회귀모형을 구현하거나 로버스트회귀방법을 사용할 수 있다.

마지막으로 데이터에 대한 이분산성을 판단하기 위해 비용인자별 잔차 산점도를 분석하였으며 결과는 <그림 4>과 같다. 데이터 수가 적어 정확한 판단에 어려움이 있으나 대부분 불규칙한 경향을 보이고 있어 데이터들은 이분산성이 존재하지 않는 것으로 판단하였다.

이상과 같이 세 가지 측면에서 데이터를 분석한 결과 다중공선성 및 이상치가 존재하였고 이분산성은 존재하지 않는 것으로 식별되었다. 이런 경우 <그림 5>과 같은 절차를 적용하여 회귀모형을 수행할 수 있다.



〈그림 4〉 비용인자별 잔차 산점도



〈그림 5〉 데이터 특성에 기반한 다양한 회귀방법

4.5 CER 개발

국내 포병무기체계 연구개발 데이터가 가지는 다중공선성, 이상치를 고려하여 회귀식을 구현하는 방법은 <그림 5>과 같이 다양하다. 따라서 가능한 모든 회귀방법을 사용하여 도출된 비용추정 관계식에 대해 t-검정, F-검정, R^2, R^2_{adj} 와 같은 통계적 평가를 수행하고 상호 비교·평가하여 가장 적합한 모형을 선정하여야 한다. 도출된 결과는 <표 5>와 같다.

표에서 보는 바와 같이 ①은 주성분회귀를 사용함으로써 다중공선성의 영향을 감소시켰으며 동시에 이상치를 제거함으로써 이상치의 영향도 제거하였다. 또한 모형 및 회귀계수의 신뢰도 역

시 타 모델에 비해 우수하다. 그러나 R^2, R^2_{adj} 는 타 모델에 비해 낮게 분석되었으나 80%이상으로 모형 자체의 정확도는 높다고 볼 수 있다.

〈표 5〉 도출된 CER 및 통계적 분석 내용

구분	①	②
변수 개수	4개 : 최대사거리(A), 구경(B), 중량(C), 전장(D)	
이상치	제거	
회귀 방법	주성분회귀	능형회귀
회귀식	5.747 + 0.714A + 0.127B + 0.001C + 0.016D	0.312A + 0.815B + 1.142C - 1.202D
R^2	0.82	0.93
R^2_{adj}	0.80	0.83
회귀 계수 검정 값	(모형) Pr>F : 0.002 (계수) Pr>|t| intercept = 1 Prin1(주성분) = 0.002	(모형) Pr>F : 0.048 (계수) Pr>|t| intercept = 1 A=0.636, B=0.372 C=0.099, D=0.311
장점	· 다중공선성 제거 · 이상치 영향 제거 · R^2, R^2_{adj} 회귀계수 및 모형의 신뢰성 좋음	· 다중공선성 제거 · 이상치 영향 제거 · R^2, R^2_{adj}, 값 좋음
단점	-	· 회귀계수의 신뢰성 나쁨

구분	③	④
변수 개수	2개 : 구경(B), 중량(C)	
이상치	미제거	
회귀 방법	로버스트회귀	
	LTS	MM
회귀식	7.702 + 0.188B + 0.002C	7.702 + 0.188B + 0.002C
R^2	0.86	0.86
R^2_{adj}	0.81	0.81
회귀 계수 검정값	(계수) Pr>ChiSq intercept=0.477 B=0.127, C=0.007	(계수) Pr>ChiSq intercept=0.477 B=0.127, C=0.007
장점	· 다중공선성 영향 감소 · 이상치 영향 감소 · R^2 값 좋음	· 이상치 양향 제거 · R^2, R^2_{adj} 값 좋음
단점	· 회귀계수의 신뢰성 나쁨	· 다중공선성 존재 · 이상치 영향 · 회귀계수의 신뢰성 나쁨

구분	⑤	⑥
변수 개수	2개 : 구경(B), 중량(C)	
이상치	제거	
회귀 방법	다중선형회귀	로그선형회귀
회귀식	7.702 + 0.187B + 0.001C	$e^{-3.670}\ B^{1.59}\ C^{-0.033}$
R^2	0.89	0.78
R^2_{adj}	0.85	0.69
회귀 계수 검정값	(모형) Pr>F : 0.004 (계수) Pr>|t| intercept=0.509 B=0.187, C=0.043	(모형) Pr>F : 0.023 (계수) Pr>|t| intercept=0.431 B=0.218, C=0.832
장점	· 다중공선성 영향 감소 · 이상치 영향 제거 · R^2, R^2_{adj} 값 좋음	· 다중공선성 영향 감소 · 이상치 영향 제거
단점	· 회귀계수의 신뢰성 나쁨	· R^2, R^2_{adj} 회귀계수의 신뢰성 나쁨

구분	⑦	⑧
변수 개수	2개 : 구경(B), 중량(C)	
이상치	제거	
회귀 방법	Adaptive CER	
	Piece 1	Piece 2
회귀식	-60.1 + 1.4*B - 0.443*C	21.2 + 0.066*B + 0.0017*C
R^2	0.97	0.948
R^2_{adj}	0.91	0.845
회귀 계수 검정값	(모형) Pr>F : 0.173 (계수) Pr>|t| intercept=0.149 B=0.111, C=0.146	(모형) Pr>F : 0.227 (계수) Pr>|t| intercept=0.569 B=0.822, C=0.223
장점	· 다중공선성 영향 감소 · 이상치 영향 감소 · R^2 값 좋음	
단점	· 회귀계수의 신뢰성 나쁨	

따라서 본 연구사례에서는 최대사거리, 중량, 전장, 구경의 4개 비용인자를 사용하면서 이상치를 제거하고 동시에 주성분회귀를 사용하여 다중공선성을 제거한 ①번 회귀모형을 비용추정관계식으로 선정하였다.

4.6 CER 검증

최종 단계로서 선택된 비용추정관계식은 MMRE가 0.214으로 0.25보다 낮은 값을 보였으며 PRED(0.3) 역시 0.75로 높은 값으로 계산됨에 따라 이 비용추정관계식의 예측력은 매우 높다고 할 수 있다[13].

따라서 최종적으로 선택된 비용추정관계식은 이상치(무기체계 I)를 제거하고 4개의 비용인자로 주성분회귀를 수행하여 얻은 다음의 식을 선택하였다.

연구개발비 = 5.747 + 0.714*(최대사거리)
+ 0.127*(구경) + 0.001*(중량)
+ 0.016*(전장)

5. 결론 및 향후 연구방향

수집된 데이터가 가지는 다양한 특성을 고려하여 적절한 회귀방법을 선택하고, 선택된 다수의 회귀방법으로 유도된 다양한 회귀모형을 통계적으로 분석함으로써 가장 적합한 비용추정관계식을 개발하는 알고리즘을 제안하였다. 이는 앞으로도 지속적으로 연구될 비용추정관계식 개발을 위한 기본 틀을 제공하였다는데 그 의의를 가지며, 특히 현재 진행되고 있는 한국형 비용추정 전산모델의 핵심 논리인 비용추정관계식 개발을 위한 기본 알고리즘을 제안하였다는데 그 의의를 둘 수 있다. 이를 통해 일반 사용자에게는 최소한의 통계적 지식으로 비용을 추정할 수 있는 자동화 절차를 제시할 수 있고, 고급 사용자에게는 회귀분석에 요구되는 각각의 요소들을 직접 선택하게 함으로써 더 신뢰할 수 있는 비용추정 절차를 제공할 수 있다.

그러나 앞에서도 언급한 바와 같이 최종적으로 선택된 비용추정관계식은 개발자의 정성적 요소가 개입될 소지가 있다. 따라서 개발자는 사용한 데이터, 개발 알고리즘을 투명하게 공개하고 권위있는 기관이나 위원회를 통해 공식적인 인증 절차를 받는 것이 필요하다. 또한 가용한 다수의 회귀모형에 대해 장점과 단점을 적절히 조합할 수 있는 방법론에 대한 연구도 추가적으로 이루어져야 할 것이다.

참고문헌

[1] 국방대학교, 한국형 비용분석 전산모델 개발(사전 연구), 방위사업청 정책연구보고서, 2009.

[2] 국방부, 국방전력발전업무규정, 국방부 훈령 제1251호, 2010.

[3] 국방부, 총소유비용 분석체계 확립방안 연구, 국방기술품질원, 2010.

[4] 김진욱, "다중공선성 상태의 주성분회귀와 능형회귀", 한국체육학회지, 45(4), 2006.

[5] 류민규 외 2명, "수리부속비 비용추정식 개발과 활용방안", 한국군사과학기술학회지, 2010.

[6] 박성현, 회귀분석 제3판, 민영사, 2007.

[7] 방위사업청, 방위사업법 제9561호, 2009.

[8] 방위사업청, 분석평가업무 실무지침서, 방위사업청 지침 제2008-19, 2008.

[9] 백종문 외 3명, "국방 소프트웨어 비용추정모델 개발방안 연구", 국방과학연구소, 2009.

[10] 성내경, 회귀분석, 자유아카데미, 2004.

[11] 어원재, 이용복, 강성진, "한국 무기체계 실적을 고려한 연구개발 비용추정관계식 개발", 산업경영시스템학회지, 2010.

[12] 이재용 외 7명, 전문가 사전지식기반 국방연구개발 비용추정 모델 개발, 국방과학연구소, 2006.

[13] Conte, S. D., Dunsmore, H. E., Shen, V. Y., "Software Engineering Metrics and Models", Benjamin-Cummings, 1986.

[14] Hoerl, A., Kennard, R. W., "Ridge regression : application to non-orthogonal problem",

Technometrics, Vol.12, 1970.

[15] Horak, John A.외 3명, "Integrated Performance Cost Model(IPCM) SYstem Level Systems", Technomics, Inc., 2005.

[16] Lund, R. E., "Tables for an approximate test for outliers in linear regression", Technometrics, 17, 1975.

[17] Book, Stephen A., "Statistical Foundations of Adaptive Cost-Estimating Relationships", SCEA-ISPA Joint Annual Conference, 2008.

▌저자소개▐

정 원 일(E-mail: keven0824@naver.com)

2003	육군사관학교 전자공학(학사)
현재	국방대학교 관리대학원 운영분석 석사과정/육군 대위
관심분야	M&S, 비용 대 효과분석, OR/SA, 재고관리

김 동 규(E-mail: kdk1216@hanmail.net)

1999	육군사관학교 화학과(학사)
2005	모스크바국립대 화학공학(석사)
2006~2008	방위사업청 획득기획국
현재	국방대학교 관리대학원 운영분석 박사과정/육군 소령
관심분야	비용분석, 비용 대 효과분석, 사업관리

강 성 진(E-mail: sjkang@kndu.ac.kr)

1983	미해대원(석사, 운영분석(OR/SA))
1988	텍사스 A&M 대학(박사, 산업공학)
1989	국방대학교(OR/SA) 교수
1999~2005	한국국방경영분석학회 편집위원
2005~2007	국방대학교 교수부장
2007	한국국방경영분석학회 회장
현재국	방대학교(OR/SA) 교수/육군 대령
관심분야	체계분석, 비용대효과분석, 자원할당, EVMS, CAIV 등