Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

A Methodology to Improve the Predictability of the CER with Insufficient Data in Korean Weapon System R&D Environment

Yong Bok Lee, Dong Kyu Kim

Ph. D student, Korea National Defense University

miliman@naver.com, kdk1216@hanmail.net

Sung Jin Kang

Professor in the department of OR/SA, Korea National Defense University

Sjkang20559@naver.com

ABSTRACT

Parametric cost estimating models have been used widely to obtain appropriate cost estimates in the early phases of weapon system acquisition. Parametric cost estimating models are composed of some CERs(Cost Estimation Relationships) based on regression analysis with historical data. However, there are many restrictions in developing a Korean version CER because of the insufficient number of projects and also abnormal data characteristics such as multicollinearity, existing outliers, heteroscedasticity, etc. As a result, the diverse regression methods have been studied in Korea to improve the predictability and stability of each CER respectively. We propose a CER development process suitable for the Korean weapon system R&D environment and a newly developed combining method of each regression model which is able to provide a better predictive ability of a CER. Real world data from historic weapon system R&D records are used to verify the performance of the developed method. Our linear combination CER method was more accurate than each regression model. Our study will provide an appropriate methodology to develop our CERs and a more accurate method in the Korean weapon system R&D environment.

1. INTRODUCTION

The acquisition cost of weapon system has continued to increase in accordance with the changes in aspects of war that are becoming modernized and precise. Moreover, such environmental changes have demanded the elevation of the efficiency of budget use and decision support for more economical acquisitions by estimating the costs in early phases of the weapon system acquisition for the total life cycle. In order to estimate a weapon system cost, the methods of parametric estimation, analogy and build-up and expert opinion are used developing on available data. Among these methods, the parametric cost estimating method is widely used in early phases of weapon system acquisition due to its promptness and convenience. As the method of estimating future costs by statistically analyzing the historical data of similar projects, the parametric cost estimating method uses the CER(Cost Estimation Relationship) that expresses the relationship between the cost and cost drivers(see ISPA 2007). To this day in Korea, the commercial models of foreign countries have been used for the parametric cost estimating method. However, many questions have been raised regarding the model not being able to reflect upon the Korean defense industry environment by using a CER based on a foreign database. In accordance with this line of thinking, Korea has recently attempted to develop Korea's own CER for the torpedo and tank weapon systems based on an expert survey(see Lee J.Y., et. Al., 2006 and 2008). We tried to develop the CER using data generation and the principal component regression method which has enabled the overcoming of the multicollinearity problem and small number of data points for the field of movement weapon systems(see Eo W.J., 2010). However, such methods have the weaknesses that they are unable to present methods for improving the accuracy of the CER and of the general methodologies that can process the abnormal data of various forms that can occur in Korean situations where the number of weapon system R&D data are insufficient.

Therefore, this study will propose a generalized CER development method enabled to develop a CER appropriate for the various characteristics of the data that can occur in Korean situations and the methods to improve the accuracy through a linear combination of the developed CERs.

2. Background

2.1 Cost Estimation Relationship (CER)

The CER is the expression that explains the relationship between the dependent variable of cost and the independent variables of cost drivers and it expresses how the cost changes according to the changes of the cost driver. It is well known that regression analysis is the most appropriate method to develop the CER(see ISPA 2007). The general regression analysis model for the independent variables of X_1, X_2, ... , X_k and the dependent variable of Y is as shown in Equation 1.

[1] $$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

β_i indicates the regression coefficient as the parameters of the population while ε indicates the residual. The hypothesis for ε that occurs when measuring Y follows the multivariate normal distribution and is hypothesized as $E(\varepsilon)=0$, $Var(\varepsilon)=\sigma^2$. Generally, the regression analysis most approximate to the actual value can be found when the sum of the residual becomes the minimum. The regression line where the squared sum of the residual becomes the minimum by using the Least Squares Method is regarded as the most appropriate regression line to the actual value.

2.2 CER Linear Combination Method

The CER Linear Combination Method proposed in this study is the method of estimating the cost by using the linearly combined CER by administering the weight based on the degree of accuracy upon each of the single CER developed according to the characteristics of the data. In connection with the linear combination model, Bates and Granger verified that the combined model reduces the errors in comparison to the unassociated models through experimental analysis(see Bates J.M. and Granger C.W.J., 1965). Furthermore, 83% of the scholars participating in the verification test related to the combination in 1992's “International Journal of Forecasting” stated that the forecasted error for the combination model had been minimized in comparison with the single models.

By proposing Rule-based forecasting, Armstrong said the method was the most effective application of the combination model. Moreover, as a result of the meta-analysis for 30 study results(1938~2000), the excellence of the combination model was verified by presenting its validity as the empirical method for the combination model reducing the error by an average of 12.5% in comparison to the single models(see Armstrong J. S., 1989 and 2001).

3. Development Method for the CER Linear Combination Method

The CER linear combination model is combined by placing the weight on the 6 developed regression models such as the Principle Component Regression, Ridge Regression, Robust Regression, Weighted Least Square Regression, Linear Regression and Log-Linear Regression according to its accuracy.

3.1 Development Process for the CER Linear Combination Method

The development of the CER linear combination method progresses in the order of data collection and normalization, data analysis and single CER development, the development of a linear combination CER and CER validation as shown in Figure 1.

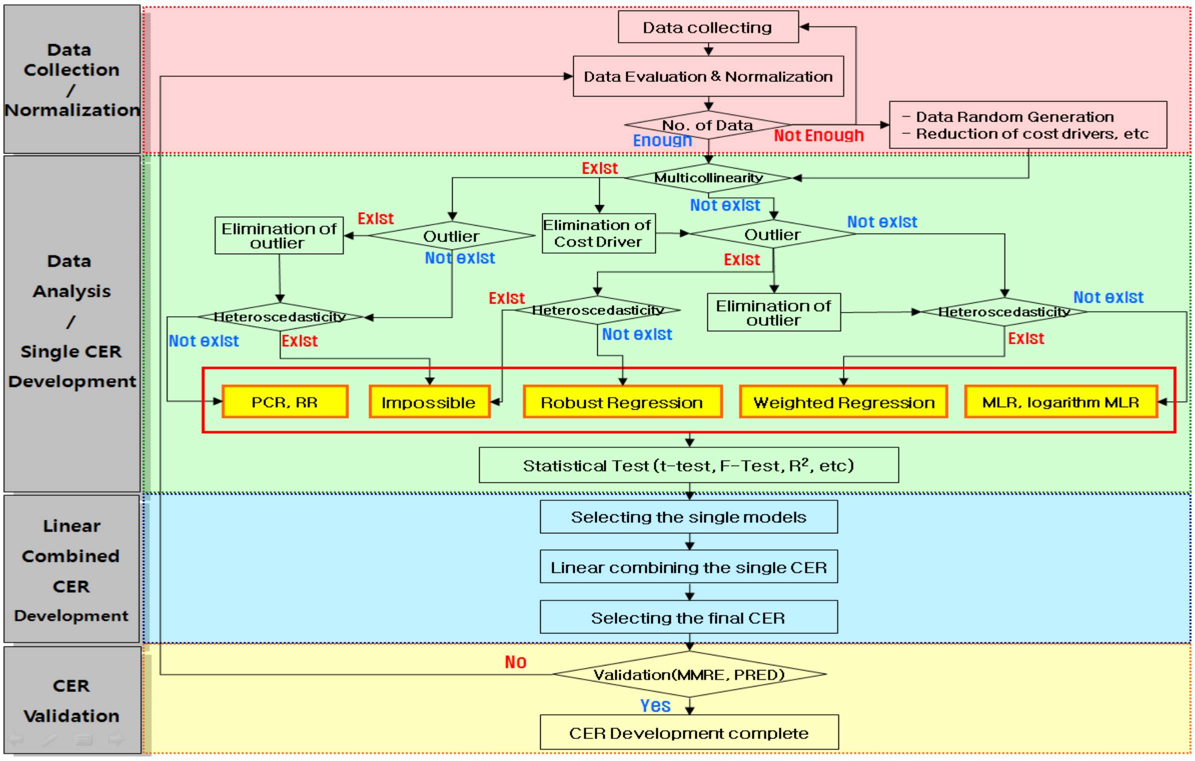


Figure 1. CER linear combining process

3.1.1 Data Collection and Normalization

The data for the analysis are collected from companies and institutions related to cost, technologies and the project.

The collected data are classified under the 6 domains of cost, specification, homogeneity, recurring/non-recurring, quantity and operational environment in order to evaluate and normalize the data(see Kang S.J., 2010). In the domain of cost, the quantity and financial standard year are normalized and in the domain of specification, the scale is normalized. In the domain for the distinguishing of homogeneity, the data are normalized by calibrating the similar data and removing the data that cannot be handled. In the recurring/non-recurring domain, recurring/non-recurring costs are distinguished and normalized by applying the learning rate in the quantity domain. Lastly, the operational environment is equally normalized.

The normalized data are selected as the core cost drivers through basic statistical analysis and the interviews with experts of each field. In this process, the number of data points is very important in order to conduct a statistical analysis. Usually, it's very difficult to obtain many cases in the Korean R&D environment. In many cases, we have only one or two samples available. Direct statistical analysis cannot be performed when the number of projects(n) is compared with the number of cost drivers(k) to result in n – k < 2 during this process. In Korea, there are not many cases where the development numbers for similar weapon system are sufficient enough for statistical analysis. Therefore, additional treatments must be performed to satisfy n - k ≥ 2 such as the reduction of cost factors or random variable generation.

3.1.2 Data Analysis & the Development of a Single CER

As part of the phase of developing the CER according to the characteristics of the selected data, data analysis is performed with the determination and measures for multicollinearity, outliers and heteroscedasticity.

3.1.2.1 Judgment & Measures for Multicollinearity

Often times, there are cases where multicollinearity exists among the independent variables of the cost drivers within the process of CER development. Multicollinearity means that the independent variables are correlated among the cost drivers. This means that the common information included among the cost drivers cannot calculate the regression coefficients or cannot implement them accurately by largely overstating the standard error of regression coefficients even when the regression analysis is made possible. In these cases, another method must be used for the diagnosis since multicollinearity cannot be detected through the residual analysis of regression.

In general, multicollinearity exists when the VIF(Variance Inflation Factor) is over 10 or the CI(Condition Index) is over 30(see Montgomery D.C., Peck E.A., Vining G.G., 2001). Measures are taken by using the following two methods when the multicollinearity exists.

First, the multicollinearity problem can be solved by applying the Ridge Regression or the Principal Component Regression model as the alternatives of the least squares method in the condition where the cost driver has not been removed(see Chatterjee S., 2000 and Montgomery D.C., Peck E.A., Vining G.G., 2001).

As the method of application when there is only the multicollinearity problem without the existence of outliers and heteroscedasticity, ridge regression solves the multicollinearity problem by using the ridge estimator to minimize the variance while acknowledging the partial bias. This method is enabled to effectively solve the multicollinearity problem that can easily occur when the number of cost drivers is small during CER development. Moreover, ridge regression normally has the tendency of reducing the mean squared error that is smaller than the OLS estimates(see Hoerl A.E. and Kennard R.W, 1970).

The principal component regression is able to effectively overcome the multicollinearity problem just like in ridge regression with the original cost drivers correlated to induce the mutual independent principal components that have linearly combined. Although the principal component analysis has a bias, a more stable estimated value for the regression coefficients can be gained.

Second, the selection method for a variable combination based on the removal of cost drivers is applied to select the cost driver combination without multicollinearity. The selection for a variable combination reevaluates the multicollinearity for the cost driver combination that are selected multiple from the 6 different selection methods(R^2, Adjusted R^2, Forward Regression, Backward Regression, Stepwise Regression, C(p) Selection). If the multicollinearity is re-founded in all the combinations that are multiple selected, the multicollinearity is diagnosed gradually from the higher order combination among that R^2 is more than 0.8 and perform the next phase using the first combination which multicollinearity is removed.

3.1.2.2 Judgment & Measures for Outlier

As the few influential observations that greatly affect the results of regression analysis, the outlier uses the studentized residual and the studentized deleted residual to diagnose the outlier for the cost and uses the leverage to diagnose the outlier of the cost driver. The judgment for the outlier must diagnose the existence of the outlier and the influential degree of the outlier.

The outlier judgment by the standardized residual is based on the critical value presented by Lund(see Lund R.E, 1975). By considering the insufficiency in the R&D data for the weapon systems in Korea, the significance level of 1% was judged as the standard. The leverage value is considered as an outlier when the leverage value is larger that 2(p(the number of estimated regression coefficients)+1)/n(see Chatterjee S., 2000). For the value judged as outlier, the Cook's Distance is applied to judge the influence for the overall fitted value and judges the influence of the regression coefficients through DFBETAS(Difference in Betas).

When the outlier exists as a result of the judgment and its influence is high, measures are taken with the following two methods.

First, the outlier is solved through robust regression, that is, the method of not removing the outlier. Robust regression is a method that places great weight on the normal error to reduce the influence of the outlier. In this study, M estimation applied with the Tukey-bisquare for estimating the robust regression coefficients by minimizing the influence of the estimation of regression coefficients and LTS(Least Trimmed Square) estimation to estimate the minimizing regression coefficients for the sum of squared residuals by ordering the squared residuals after excluding the ones with large values are used.

Second, the cost drivers are reselected after removing the outlier in case of the number of data is sufficient relatively than applying robust regression and judge the heteroscedasticity for applying other appropriate regression models. With no relation to the existence of the outlier, these measures are omitted when there is no or very small influence of each cost driver in judging the heteroscedasticity.

3.1.2.3 Judgment & Measures for Heteroscedasticity

The homoscedasticity of error terms is one of the basic hypotheses for the least squares theory. In the case of heteroscedasticity, the standard error and the estimate of the regression coefficients may be inaccurate. Therefore, the theoretical validity for applying the least squares method cannot be guaranteed. The heteroscedasticity is diagnosed according to the dispersion of residuals. When the residuals irregularly disperse, it is judged that the heteroscedasticity does not exist and when the residual disperses in the form of being proportionate according to the cost driver, it is judged that the heteroscedasticity does exist.

When the heteroscedasticity exists, the weighted regression of transforming the residual to the equal variance is used to solve the heteroscedasticity problem. The weighted regression gains better estimation using the Ordinary Least Squares method by solving the heteroscedasticity through the variance-stabilizing transformation that stabilizes the dispersion. In short, this method estimates the cost by minimizing the weight error square sum that gives a weight value that is reciprocally proportionate to the dispersion of the error term.

When the heteroscedasticity does not exist, the appropriate CER according to the characteristics of the data is developed using one of ridge regression, principal component regression, robust regression, linear regression or log linear regression.

In accordance with the above process, the development process for a single CER according to the characteristics of the data is as shown in Table 1. For example, when there is the existence of multicollinearity, the nonexistence of an outlier and the nonexistence of heteroscedasticity, the single CER is developed by applying ridge regression or principal component regression.

TABLE 1. Single CER development process according to data characteristics

CER type	Multicollinearity	Outlier	Heteroscedasticity	Remark
Linear / Log Linear	○ (cost driver elimination)	×	×	• Multicollinearity no exists • Heteroscedasticity no exists • Outlier no exists
	○ (cost driver elimination)	○ (outlier elimination)	×	
	×	×	×	
	×	○ (outlier elimination)	×	
Ridge / Principal Component	○	×	×	• Multicollinearity only exists
	○	○ (outlier elimination)	×	
Robust	○ (cost driver elimination)	○	×	• Outlier only exists
	×	○	×	
Weighted	○ (cost driver elimination)	×	○	• Heteroscedasticity only exists
	○ (cost driver elimination)	○ (outlier elimination)	○	
	×	×	○	
	×	○ (outlier elimination)	○	

3.1.2.4 STATISTICAL TEST OF THE SINGLE CER

The R^2 test, t-test and the F test were used for the statistical tests of the single CERs. In this study, the CER was appropriately good when the value of R^2 is over 0.8, the fitness for the regression coefficient by the t-test was based on the significance level of 5% and the goodness of fit of the entire regression coefficient by the F-test was based on the significance level of 5%.

3.1.3 DEVELOPMENT OF LINEAR COMBINATION CER

Linear combination CER is the linearly combined CER with the all selected single CERs which are based on the data characteristics of Figure 1 and Table 1. The weight was placed on the single CERs based on the 4 different methods for linear combination and the final CER which was selected had the minimal value of the RMSE (Root Mean Squared Error) where R^2 was over 0.8. During the combination of CERs, all weight calculation methods(j) are performed and the number of selected single CERs of type k(m) is one of the number of 2~6 according to selection of the single CERs. For example, if 3 CERs are selected to combine, C_j is one of the ${}_6C_3$, that is, C_j could be $W_{j1}CER_1$+ $W_{j2}CER_3$+ $W_{j6}CER_6$ or $W_{j2}CER_2$+ $W_{j4}CER_4$+ $W_{j6}CER_6$, etc.

$$Min(RMSE_1,\ RMSE_2,\ldots,RMSE_j) \tag{2a}$$

subject to

$$C_j = \sum_{k=1}^{m} W_{jk} CER_k \tag{2b}$$

$$\sum_{k=1}^{m} W_{jk} = 1 \tag{2c}$$

where

[2] j : Weight calculation method

k : Single CER type

m : the number of selected single CERs of type k

RMSE $_j$: RMSE value of C_j within R^2 value $\geq$ 0.8

C_j : Linear Combining CER of method j

CER_k : Selected single CER of type k

W_{jk} : Weight of CER_k with method j

In accordance with the method of placing the weight according to the accuracy, the linear combination was classified as a linear combination by SSE(Sum of Squares due to Residual Errors), linear combination by MMRE(Mean Magnitude of Relative Error), the linear combination by adjusted coefficient of determination(adjusted R^2) and the linear combination by partial regression coefficient.

First, the linear combination by SSE placed weight according to Equations 3a and 3b.

$$W_{jk} = (1/SSE_k)/(\sum_{k=1}^{m} 1/SSE_k) \tag{3a}$$

where

[3] $$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3b}$$

y_i : Actual cost of observation i

$\hat{y}_i$: Estimated cost of observation i

Second, the linear combination by MMRE placed weight according to Equations 4a and 4b.

$$W_{jk} = (1/MMRE_k)/(\sum_{k=1}^{m} 1/MMRE_k) \tag{4a}$$

[4] *where*

$$MMRE = (1/n)\sum_{i=1}^{n} [(\left|y_i - \hat{y}_i\right|)^2 / y_i] \tag{4b}$$

n : Number of observations

Third, the linear combination by adjusted R^2 placed weight according to Equations 5a, 5b and 5c.

$$W_{jk} = R^2_{adj,k} / \sum_{k=1}^{m} R^2_{adj,k} \quad (5a)$$

[5] *where*

$$R^2_{adj,k} = 1 - [SSE/(n-(k+1))]/[SST/(n-1)] \quad (5b)$$

$$SST(Total\ Sum\ of\ Squared\ deviations) : \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5c)$$

Fourth, the CER combination by partial regression coefficient combine the CER linearly as shown in Equation 6 using the regression coefficient values of multiple regressions that placed the estimating cost of the single model as the independent variable and placed the actual cost as the dependent variable.

$$W_{jk} = \beta_k / \sum_{k=1}^{m} \beta_k$$

[6] *where*

β_k : Regression coefficient of selected single CER_k, $\sum_{k=1}^{m} \beta_k = 1$

3.1.4 Model Selection

The final CER is selected with an R^2 value is over 0.8 and with the minimum RMSE value.

3.2 CER Validation

The final CER requires the actual validation and verification of accuracy. This study evaluated the validity through the comparison of the actual cost and the estimates, the verification by the MMRE(Mean Magnitude of Relative Error) and PRED(ℓ) (Prediction at level ℓ) that displayed the validity of the CER due to the rack of the number of weapon system R&D data. When the verification results by the scale is better than the utilized standards, it signifies the success of CER development and that CER can be used in actual projects. However, when the accuracy does not satisfy the utilized standard, it signifies the failure of CER development and therefore, the CER must be re-developed by going back to the past phase.

First, the MMRE is the average for the accuracy of cost estimation and is as shown in Equation 7 and indicates the degree of accuracy according to the utilized standard. In this study, MMRE≤0.25 was utilized as the standard of judgment for the validation.

[7] $$MMRE = (1/n)\sum_{i=1}^{n}[(\left|y_i - \hat{y}_i\right|)^2 / y_i] = (1/n)\sum_{i=1}^{n} MRE_i$$

Second, PRED(ℓ) is as shown in Equation 8 with the range of error(ℓ) and the number(q) included in MRE ≤ ℓ (see Boehm B.W., et al., 2000 and Conte S. D., Dunsmore H. E., Shen V. Y., 1986). In this study, PRED(0.3) ≥ 0.3 was utilized as the standard of judgment for model validity.

$$PRED(l) = q/n$$

[8] *where*

q : Obseervation number of $MRE_i \le l$

n : Total number of observations

4. CER Development Case Study

4.1 Data Collection & Normalization

The worth of the CER depends on the reliability of the collected data. To collect the reliable data for a weapon system, it is important to collect the official data from sources that possess the actual data such as the national research institutes and defense industry firms. In the attempt to perform the present study, the 25 types of R&D data as shown in Table 2 were acquired for the 284 development cases by cooperating with the Agency for Defense Development that leads the R&D of weapon system in Korea. As a result of analyzing the acquired data by dividing them into the 8 major classifications of weapon system(C4I, surveillance-reconnaissance, mobile, shipment, aircraft, firepower, protection and other), it was concluded that it was appropriate to develop the CER within the standards of mid-classifications, where the number of entities had been appropriate and the characteristics amongst the entities had been similar. However, the statistical analysis was restricted due to 9 fields having lesser similarities amongst weapon systems and 16 fields lacking in number of data amongst the total of 27 fields. As a result of the final analysis, it was judged that the statistical analysis was enabled in the field of artillery and this study researched the cases for the 9 artillery weapon systems.

TABLE 2. Data collection

cost data	4	R&D Cost, Production Cost, Import Cost, Inverted Cost
Specification data	17	Combat Weight, No. of passengers, Engine power, Range, Max Velocity, Max Range, Caliber, Weight, Length, Max rapidity, Continue rapidity, etc.
Project data	5	R&D Duration, Quantity, Company, Military Type, Arrangement Year

The collected data normalized the cost data by reflecting on inflation and its results are as shown in Table 3. Moreover, the maximum distance, caliber, weight and length were judged as the cost drivers through the interviews with experts in the field of cost as well as the analysis of a scatter plot between the cost and cost drivers. With regard to the amount of data, there were no restrictions in performing the statistical analysis by regression analysis because of the 4 cost drivers and 9 observations.

TABLE 3. Normalization result of the data

Weapon	Max Range (km)	Caliber (mm)	Weight (kg)	Length (cm)	Max rapidity of fire (R/min)	Continuous rapidity of fire (R/min)	R&D cost (100M$, 2010)
1	3.59	60	18	99	30	20	18.2027
2	1.8	60	21	82	30	18	12.7289
3	6.473	81	41	155	30	11	35.2546
4	4.737	81	81	130	12	5	17.8506
5	11.274	105	2,260	231	3	1	37.6372
6	14.7	105	2,650	392	5	2	27.069
7	18	155	6,890	701	4	2	43.0712
8	18	155	25,000	912	4	1	74.0739
9	41	155	47,000	810	6	2	1,342.85

4.2 DATA ANALYSIS & DEVELOPMENT OF SINGLE CER

4.2.1 DATA ANALYSIS

First of all, the selection for the variable combination(R^2, Adjusted R^2, Forward Regression, Backward Regression, Stepwise Regression and the C(p) Selection) was executed to select main cost drivers. Four cost drivers(Max Range, Weight, Length and Caliber) were selected as main cost drivers

TABLE 4. Cost driver selection

R^2	Adjusted R^2	Forward	Backward	Stepwise	C(p) Selection
Max Range, Weight, Length			Caliber	Max Range, Weight, Length	

After selecting main cost drivers, we calculated VIF and CI to judge whether the multicollinearity exists or not. The VIF values of all cost drivers exceeded 10 as seen in Table 5 and the CI value of the 4th cost driver exceeded 30 and as a result it was judged that multicollinearity existed in the data.

TABLE 5. VIF, CI values of cost driver

Statistics	Max Range	Caliber	Weight	Length
VIF	15.76	30.2	14.54	29.74
C I	2.75	6.7	10.87	42.8

And we analyzed the correlation matrix to judge which cost drivers are highly correlated. As shown as Table 6, two cost drivers(Max Range and Length) were highly correlated with the others(Caliber and Weight) and it means that Max Range and Length may be eliminated to decrease the multicollinearity.

TABLE 6. Correlation Matrix

	Max Range	Caliber	Weight
Caliber	0.827	-	-
Weight	0.925	0.740	-
Length	0.818	0.964	0.804

Secondly, outliers were suspected in the 9th weapon system as shown in Table 7 and it can be said that the data point of the 9th weapon system has greater influence because Cook's Distance and the DFBETAS value are higher than the standard value.

TABLE 7. Outlier and influence power test

Weapon	SR	SDR	Hat Diag	Cook' Distance	DFBETAS				
					Constant	Max Range	Caliber	Weight	Length
1	0.15	0.13	0.41	0	0.08	0.03	-0.07	-0.02	0.05
2	0.68	0.62	0.36	0.05	0.24	-0.09	-0.12	0.09	0.07
3	0.15	0.13	0.18	0	0	-0.01	0.01	0.01	-0.02

4	0.35	0.31	0.34	0.01	-0.09	-0.13	0.14	0.13	-0.15
5	-1.52	-2.04	0.53	0.52	1.54	0.65	-1.68	-0.72	1.74
6	-1.48	-1.91	0.69	0.98	-1.76	-2.44	2	2.59	-1.95
7	1.89	5.02	0.62	1.18	-1.86	1.02	1.04	-2.6	0.34
8	-1.94	-7.06	0.89	6.37	-0.44	10.48	-0.53	-6.62	-4.22
9	1.99	19.08	0.98	45.17	2.73	30.47	-3.84	24.81	-16.4

Third, the heteroscedasticity did not exist since the dispersion for the 8 weapon systems were irregularly dispersed.

4.2.2 Single CER Development

The 7 methods of ridge regression, principal component regression, robust regression, linear regression(I), log linear regression(I), linear regression(II) and log linear regression(II) were applied according to the results of data analysis to develop the single CER and the results of statistical tests are as shown in Table 8.

Table 4. Test statistics of single CER

CER Type	Cost Driver	$R^2(R^2_{adj})$	Test result
Principal component	Max range(A), Caliber(B), Weight(C), Length(D)	0.82 (0.80)	• Model Pr>F : 0.002 • Coefficient Pr>|t| : intercept=1, Prin1=0.002
Ridge		0.93 (0.83)	• Model Pr>F : 0.047 • Coefficient Pr>|t| : intercept=0.8337, A=0.636, B=0.372 C=0.099, D=0.311
Robust (LTS)	Caliber, Weight	0.86 (0.81)	• Coefficient Pr>ChiSq : intercept=0.477, B=0.127, C=0.007
Linear(□)	Caliber, Weight,	0.89 (0.85)	• Model Pr>F : 0.004 • Coefficient Pr>|t| : intercept=0.509, B=0.187, C=0.043
Log-linear(□)		0.78 (0.69)	• Model Pr>F : 0.023 • Coefficient Pr>|t| : intercept=0.431, B=0.218, C=0.832
Linear(□)	Weight	0.84 (0.81)	• Model Pr>F : 0.001 • Coefficient Pr>|t| : intercept=0.0005, C=0.0014
Log-linear(□)		0.69 (0.64)	• Model Pr>F : 0.011 • Coefficient Pr>|t| : intercept=0.0002, C=0.01

First, the CER was developed by principal component regression for the 8 weapon systems that used the 4 cost drivers and R^2 is 0.82 while the results of the F-test and t-test were under the significant level of 5% for the fitness of the regression coefficients and the appropriateness of the model is seen to be good. Therefore, the CER was developed through the principal component regression(Equation 9) can be seen as being appropriate as the single CER.

[9] $$Y_{PCR} = 5.747 + 0.714 Range + 0.127 Caliber + 0.001 Weight + 0.016 Length$$

Second, the CER was developed by linear regression(II) for the 8 weapon systems that had 1 cost driver removed had a value for R^2 which was 0.84 with the significant level of the F-test and t-test being under 5%. The appropriateness of that model and the fitness of the regression coefficients were good. Hence, the CER developed(Equation 10) according to the linear regression(II) can be explained as being appropriate as the single CER.

[10] $Y_{Lin2} = 23.52163 + 0.0021 Weight$

Other than those mentioned, the CER was developed through ridge regression and robust regression. Linear regression(I), log linear regression(I) and log linear regression(II) did not satisfy the fitness of regression coefficients or the appropriateness of models to be determined as being inappropriate as the single CER. Hence, the CER developed by the principal component regression and the linear regression(II) were finally selected as the single CER to perform the CER linear combination.

4.3 Development of Linear Combination CER

4.3.1 CER Linear Combination

Based on the initial data as shown in Table 9 for the selected single CER, the 4 different methods were used for the linear combination.

Table 5. Basic statistics for CER linear combining

weapon	y_i	$\hat{y}_{PCR,i}$	$\hat{y}_{Lin2,i}$	$(\hat{y}_{PCR,i}-y_i)^2$	$(\hat{y}_{Lin2,i}-y_i)^2$	$\lvert\hat{y}_{PCR,i}-y_i\rvert$	$\lvert\hat{y}_{PCR,i}-y_i\rvert$	$R^2_{adj,PCR}$	0.795
1	18.203	17.525	23.559	0.460	28.695	0.678	5.357	$R^2_{adj,Lin2}$	0.812
2	12.729	15.985	23.566	10.600	117.438	3.256	10.837	β_{PCR}	0.469
3	35.255	23.160	23.607	146.280	135.673	12.095	11.648	β_{Lin2}	0.543
4	17.851	21.539	23.692	13.605	34.119	3.689	5.841	$MMRE_{PCR}$	0.214
5	37.637	31.983	28.268	31.967	87.790	5.654	9.370	$MMRE_{Lin2}$	0.284
6	27.069	37.173	29.087	102.089	4.071	10.104	2.018		
7	43.071	52.960	37.991	97.783	25.813	9.889	5.081		
8	74.074	65.565	76.022	72.394	3.794	8.508	1.948		
sum				475.178	437.393	53.873	52.100		

The CER by principal component regression was named as CER$_1$ while the CER by the linear regression(II) was named as CER$_2$ for the linear combination and evaluation. The modeling results of those mentioned above are as follows.

[11]

$$Min(RMSE_1,\ RMSE_2,\ RMSE_3,\ RMSE_4) \quad (11a)$$

subject to

$$C_j = \sum_{k=1}^{2} W_{jk} CER_k \quad (11b)$$

$$\sum_{k=1}^{2} W_{jk} = 1 \quad (11c)$$

where

j : 1(Combining model by SSE), 2(Combining model by $MMRE$),
3(Combining model by R^2), 4(Combining model by *regeression coefficient*)

k : 1(PCR), 2(*Linear regression* 2)

First, the linear combination by SSE used the weight W_{11}=0.479, W_{12}=0.521 and is as shown in Equation12.

[12]
$$\begin{aligned} C_1 &= 0.479CER_1 + 0.520CER_2 \\ W_{11} &= (1/475.178)/(1/475.178 + 1/437.393) = 0.479 \\ W_{12} &= (1/437.393)/[1/475.178 + 1/437.393] = 0.520 \end{aligned}$$

Second, the linear combination by MMR calculated the weight of W_{21}=0.570, W_{22}=0.430 and is as shown in Equation 13.

[13]
$$\begin{aligned} C_2 &= 0.570CER_1 + 0.430CER_2 \\ W_{21} &= (1/0.214)/(1/0.214 + 1/0.284) = 0.570 \\ W_{22} &= (1/0.284)/(1/0.214 + 1/0.284) = 0.430 \end{aligned}$$

Third, the linear combination by adjusted R^2 calculated the weight of W_{31}=0.495, W_{32}=0.505 and is as shown in Equation 14.

[14]
$$\begin{aligned} C_3 &= 0.495CER_1 + 0.505CER_2 \\ W_{31} &= 0.795/(0.795 + 0.812) = 0.495 \\ W_{32} &= 0.812/(0.795 + 0.812) = 0.505 \end{aligned}$$

Fourth, the linear combination by the partial regression coefficient executed the linear regression analysis that placed the value estimated by CER_1 and CER_2 as the independent variables and the R&D cost as the dependent variable, respectively, to gain the estimated regression coefficient as W_{41}=0.469 and W_{42}=0.543 and is as shown in Equation 15.

[15]
$$C_4 = 0.469CER_1 + 0.543CER_2,\ R^2 = 0.97,\ p - value < 0.0001$$

Model Selection

For the model selection, the RMSE and R^2 for the 4 CER linear combination models and 2 single CER models were compared as shown in Table 10.

Table 6. Comparison the RMSE and R2 values of each model

	CER linear combining model				Single CER model	
	by SSE	by MMRE	by R^2_{adj}	by Regression	PCR	Linear
R^2	0.879	0.879	0.879	0.880	0.825	0.839
RMSE	7.3988	7.3974	7.3968	7.3734	12.582	8.5381

As the result, C_4 was selected as the final CER linear combination model(Equation 16) that is the minimum RMSE with the R^2 being over 0.8.

[16]
$$C_4 = 15.46 + 0.3350Range + 0.00598Caliber + 0.0014Weight + 0.0074Length$$

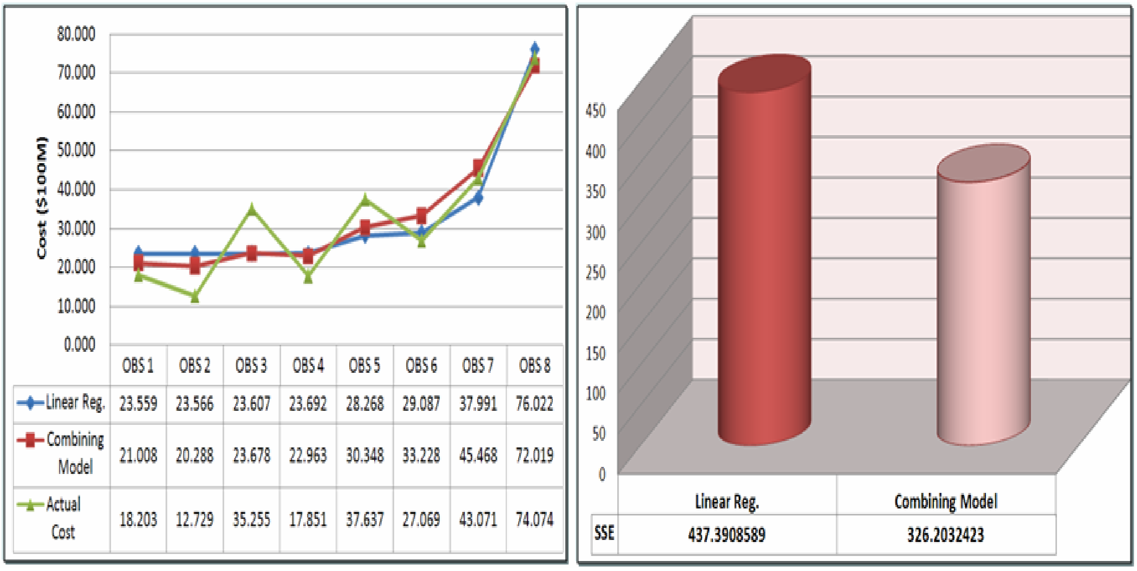
4.4 CER Validation

As a result of calculating the MMRE and PRED(0.3) for C_4 using Table 11, MMRE satisfied the condition MMRE≤0.25 with 0.23 to be a model with goodness(see Conte S.D, Dunsmore V.Y, Shen V.Y., 1986). Moreover, the validity of C_4 could be stated to be high when considering that the PRED(0.3)≥0.3 of the commercial model as PRED(0.3) is 0.75 since q=6 and n=8(see Boehm B.W., et al., 2000).

TABLE 7. MRE, MMRE values of C_4

	1	2	3	4	5	6	7	8	Average(MRE)
MRE	0.154	0.594	0.328	0.286	0.194	0.228	0.056	0.028	0.23

Second, the difference of the actual cost with the estimated cost and the SSE were compared as shown in Figure 2 for C_4 and the general linear regression. As a result of comparing the actual cost and the estimated cost, the cost estimated by C_4 was closer to the actual cost with the exclusion of the 6th and 8th weapon system and the SSE of C_4 indicated a higher accuracy by being smaller than the general linear regression.



	OBS 1	OBS 2	OBS 3	OBS 4	OBS 5	OBS 6	OBS 7	OBS 8
Linear Reg.	23.559	23.566	23.607	23.692	28.268	29.087	37.991	76.022
Combining Model	21.008	20.288	23.678	22.963	30.348	33.228	45.468	72.019
Actual Cost	18.203	12.729	35.255	17.851	37.637	27.069	43.071	74.074

	Linear Reg.	Combining Model
SSE	437.3908589	326.2032423

FIGURE 2. Comparison the estimates and SSE between the C4 with the linear regression

Third, Equation 17 was applied for the linear combination model to evaluate the degree in improvement of accuracy.

[17] $$Accuracy\ improvement\ rate = [1 - \min(SSE_k)/(SSE_j)] \times 100$$

The SSE of C_j is as shown in Table 12 and in comparison to CER_2 with the lowest SSE among the single CERs, the accuracy of C_j was each elevated by 24.91%, 24.94%, 24.95% and 25.42%.

TABLE 8. Accuracy improvement rate of the CER linear combining model

	by SSE	by MMRE	by R^2_{adj}	by Regression
CER	$C = 0.479CER_1 + 0.521CER_2$	$C = 0.570CER_1 + 0.430CER_2$	$C = 0.495CER_1 + 0.505CER_2$	$C = 0.469CER_1 + 0.543CER_2$
SSE	328.449	328.326	328.275	326.203
Accuracy improvement rate(%)	24.91	24.94	24.95	25.42

5. CONCLUSION AND FUTURE STUDY

This study holds the following significance as the study that has primarily presented the development process for CER linear combination in the field of cost estimation for weapon system.

First, the linear regression or the log linear regression is generally used to develop the CER in cases where the historical number of weapon system is sufficient and where it is possible to collect data from various sites. However, it is not easy to develop the CER by considering the various characteristics of data such as multicollinearity, heteroscedasticity and outlier within situations where the number of weapon systems is insufficient. Therefore, the process for CER development and validation proposed in this study can be used as the standard process for general CER development that has widely considered the numerous problems that may occur according to the characteristics of data within the development process of CERs.

Second, the CER linear combination method is enabled to solve the possibilities for omission of the critical factors within the process of cost estimation by forecasting based on more information than the single model since it uses all the cost related information held by each of the single models.

Third, the linear combination method is able to reduce the errors that occur by the single model by inducing the estimations with great errors to become closer to the actual value through placing greater weight in the observed value with smaller error according to the degree of accuracy.

This study has proposed a CER development methodology which has enabled the overcoming of the restrictions of an insufficient amount of weapon system R&D data under the situations in Korea. However, additional studies are required to theoretically establish the validity to place weight for the CER linear combination.

REFERENCES

Armstrong J. S., 1989, *Combining Forecasts : The End of the Beginning or the Beginning of the End*, International Journal of Forecasting.

Armstrong J. S., 2001, *Combining Forecasts : Principles of Forecasting, A Handbook for researchers and Practitioners*, Kluwer Academic Publishers, 417-439.

Bates J. M. and Granger C. W. J., 1965, *The Combination of Forecasts, Operational Research Quarterly*, Vol. 20 : 451-468.

Boehm B. W, et al., 2000, *Software Cost Estimation With COCOMO* , Prentice Hall.

Chatterjee S., Hadi A. S, Price B; 2000, *Regression analysis by example*, John Wiley and Sons.

Conte S. D., Dunsmore H. E., Shen V. Y., 1986, *Software Epreyngineering Metrics and Models*, Benjamin-Cummings.

Eo W. J ., et al., 2010, *Development of R&D CER considering the Korean weapon system environment*, Journal of Society of Korea Industrial and Systems Engineering. Vol.33(3).

Hoerl A. E. and Kennard R. W., 1970, *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, Vol.12 : 69-82.

International Society of Parametric Analysts(ISPA), 2007, *Parametric Estimating Handbook*, ISPA Asian Chapter.

Kang S. J., 2010, *Cost Estimation*, Seoul:Du Nam.

Lee J. Y., et al., 2006, *Development of defense cost estimation model based on expert survey*, Agency for Defense Development.

Lee J. Y., et al., 2008, *Development of defense cost estimation model based on expert survey(   )*, Agency for Defense Development.

Lund R. E., 1975, *Tables for an approximate test for outliers in linear regression*, Technometrics, Vol.17.

Montgomery D. C., Peck E. A., Vining G. G., 2001, *Introduction to linear regression analysis*, John Wiley and Sons.

Rencher A. C. and Schaalje G. B., 2008, *Linear Models in Statistics*, Wiley-Interscience.