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 the Korean Weapons Systems R&D Environment

June. 2011

Maj. Dong Kyu Kim, Maj. Yong Bok Lee, Dr. Sung Jin Kang

Korea National Defense University

Contents

Introduction

- **Recently Situation**
- **Goals**

Recently Situation

The Cost Analysis In Korean Defense Acquisition Environment

The Increase of Complexity, Art-of-technology of Weapons systems → The Increase of Acquisition Cost → The Efficiency and Transparency in the Acquisition Process → The Importance of Cost Analysis

The Limitations of the Cost Analysis In Korea

- No cost estimation tools suitable for Korean defense industry environment
- The insufficiency of R&D, production experiences of weapons systems

Recently Situation (cont') / Goals

The Recently Changes to Strengthen the Cost Analysis System In Korea

- **The Application of TLCCSM to Weapons Systems Acquisition Process**

  TLCCSM : Total Life-Cycle Cost System Management

- **The Development of the Korean Version Cost Estimation Model**

Goals

- **Propose the Methodology and Process of CER Development Suitable for Korean Defense Industry Environment**
- **Propose the Methodology to Improve the Predictability of the CER with Insufficient Data**

Background

- **CER** (Cost Estimation Relationships)
- **CER Linear Combination Model**

CER (Cost Estimation Relationships)

What is a CER ?

The relationship between the dependent variable of cost and the independent variable of cost drivers

How to Develop a CER ?

- **Regression Analysis** → **The most appropriate method**

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

- **The Assumption for the Linear Regression**

  - $E(\varepsilon_i) = 0$ **for i = 1,2,…, n or, equivalently,** $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k x_{ik}$
  - $\mathrm{var}(\varepsilon_i) = \sigma^2$ **for i = 1,2,…, n or, equivalently,** $\mathrm{var}(Y_i) = \sigma^2$
  - $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ **for all i ≠j, or, equivalently,** $\mathrm{cov}(Y_i, Y_j) = 0$

CER (Cost Estimation Relationships) (cont')

If One or More assumptions do not hold ?

- **The estimators may be poor**
- **If then, how to develop the CER if one or more assumptions do not hold ?**

- Step #1 : Analyze the characteristics of data

- Step #2 : Select appropriate method and Apply it to the development of CER

CER Linear Combination Model

What is a CER Linear Combination Model ?

The linear combined CER by administering the weight based on the degree of accuracy upon each of the single CERs

How to Develop a CER Linear Combination Model ?

- **Develop the single CERs**
- **Select the single CERs to combine them**
- **Calculate the weights**
- **Combine the single CERs based the weights**

$$\textbf{CER} = W_1 CER_1 + W_2 CER_2 + \ldots + W_k CER_k$$

CER Linear Combination Model (cont')

Prior Experimental Analysis and Test for the Linear Combination Model

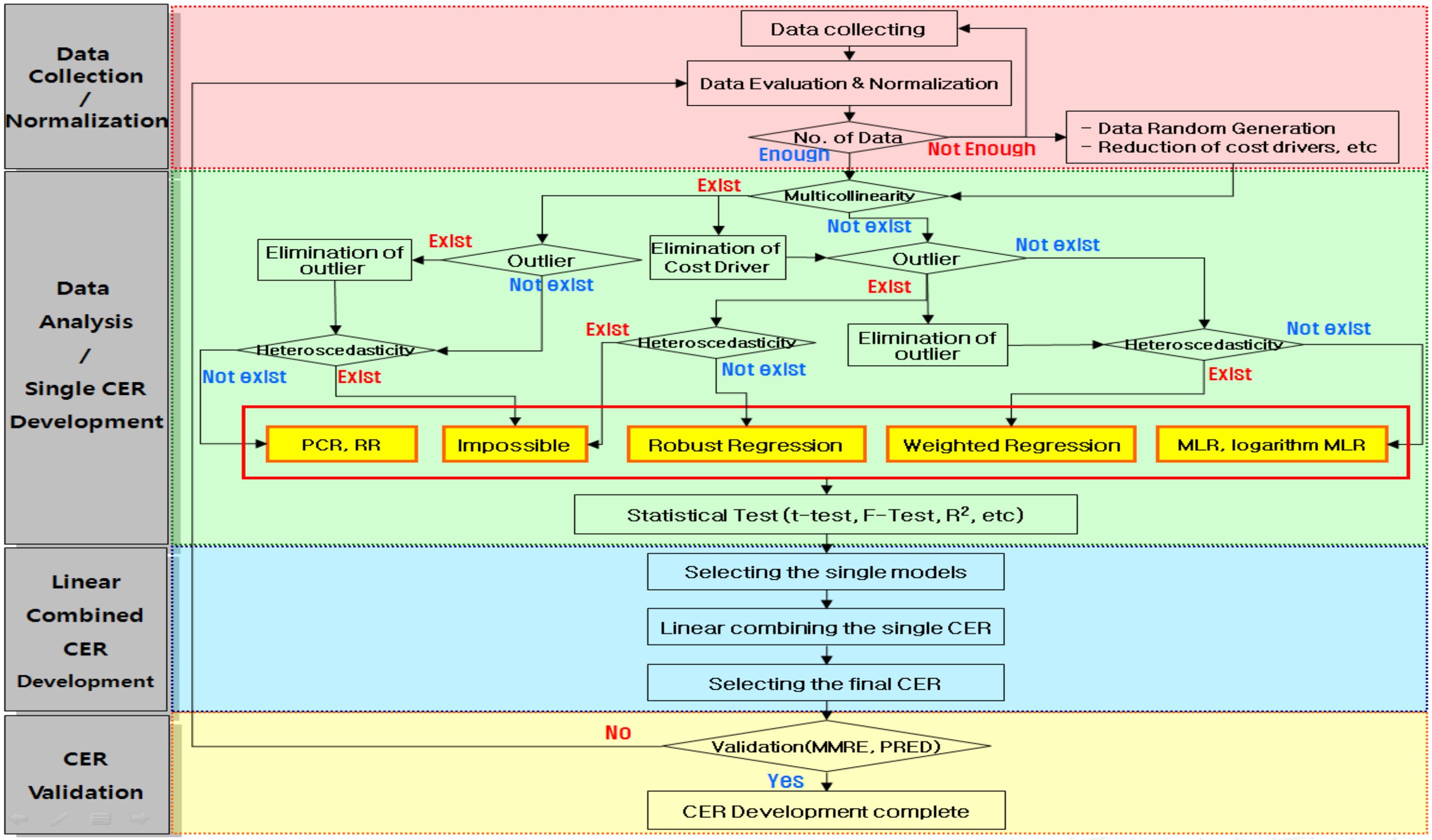
- **Bates and Granger(1965)**
  - **Reduce the errors in comparison to the unassociated models through experimental analysis**
- **International Journal of Forcasting(1992)**
  - **83% of scholars participating in the verification test stated that the forecasted error for the combination model had been minimized**
- **Armstrong(1989, 2001)**
  - **The combination model reduced the error by average of 12.5% in comparison to the single model**

Development Method for the CER Linear Combination Model

- Development Process
- Data Collection and Normalization
- Data Analysis & the Dev. of a Single CER
- Dev. Of CER Linear Combination Model
- CER Validation

Development Process

Development Process



Data Collection and Normalization

Data Collection

- **Definition of the scope of weapon system for estimating cost**
  - **- Homogeneity and Consistency**
  - **- Feasibility of the data collection**
    (Example) Armored vehicle and Tank ? R&D cost / Product cost / O&S cost ?
- **Data collection**
  - **- Cost data as the dependent variable**
  - **- Cost drivers as the independent variable**
    (Example) Weight, Speed, Fire range, etc.
- **The number of data**
  - **- It is important in Korean defense environment**
  - **- If n (the number of projects) – k (the number of cost drivers) ≥ 2, feasible**
  - **- If not feasible, stop/eliminate cost drivers/data generation …**

Data Collection and Normalization (cont')

Data Collection (cont')

- **The most important thing is to collect the reliable data from authorized institutes or companies which produce those cost drivers**

Normalization

Provide consistent data by neutralizing the impacts of external influences

Judgment & Solutions for Multicollinearity

Multicollinearity

The Independent variables and correlated among the cost drivers

- Violate the assumption for the Linear Regression
- Regression Coefficients and Estimated Values may not be confident

Judgment

- Correlation Analysis : Correlation Matrix
- Variance Inflation Factor(VIF) : $VIF_k = 1/(1-R_k^2) \geq 10$ $E(VIF_k) = \frac{1}{k}\sum_{i=1}^{k}(VIF_i) > 1$
- Condition Index(CI) : $\eta_j = \sqrt{\frac{\max\lambda}{\lambda_j}} > 30$

Solution

- Eliminate one or more unimportant cost drivers among high correlated ones
- Apply the Principal Component Regression(PCR), Ridge Regression(RR)

Data Analysis & The Dev. Of a single CER **(cont')**

Judgment & Solutions for Outliers

Outliers

Some residuals are much larger than the others

- **Unusual data may have a big influence on the regression model**

Judgment

- **Studentized Residual, Studentized Deleted Residual**
- **Cook's Distance, Difference in Betas(DFBETAS), etc.**

Solution

- **Eliminate outliers if data are enough**
- **Apply the Robust Regression if data are not enough or important**

Data Analysis & The Dev. Of a single CER **(cont')**

Judgment & Solutions for Heteroscedasticity

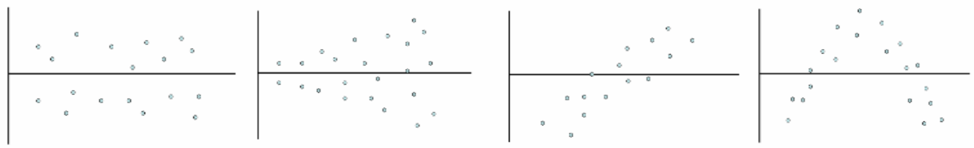
Heteroscedasticity

The residuals disperse regularly

- Violate the basic hypotheses for the least squares theory
- The estimate of the regression coefficients may be inaccurate

Judgment

- Scatter Plot of residuals



(a) (b) (c) (d)

Solution

- Weighted Regression

Data Analysis & The Dev. Of a single CER (cont')

Development of a Single CER and Statistical test

Development of a Single CER

Apply the appropriate method according to the characteristics of data

Statistical test of a Single CER

R^2, $_{adj.}R^2$, T-test, F-test → Commonly Acceptable Criteria

Dev. of CER Linear Combination Model

Dev. of CER Linear Combination Model

Objective

$$Min(RMSE_1,\ RMSE_2, \ldots, RMSE_j)$$

Subject to

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

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

Definition of parameters

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

Dev. of CER Linear Combination Model (cont')

Methods of placing the weight

Error Sum of Squares(SSE)

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

where

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

y_i : Actual cost of observation *i*

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

Mean Magnitude of Relative Error(MMRE)

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

where

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

n : Number of observations

Adjusted R^2

$$W_{jk} = R^2_{adj,k} / \sum_{k=1}^{m} R^2_{adj,k}$$

where

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

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

Partial Regression Coefficient

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

where

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

Dev. of CER Linear Combination Model (cont')

Model Selection

- **The Final CER Selection**
  - **R^2 or adjusted R^2 value over a commonly(generally) acceptable criteria**
  - **The minimum Root Mean Squared Error(RMSE) value**

CER Validation

Validation Methods of CER linear Combination Model

The comparison of the actual cost and the estimates

Mean Magnitude of Relative Error(MMRE)

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

Prediction al level l (PRED(l))

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

where

q : Obseervation number of $MRE_i \le l$

n : Total number of observations

Case Study

- Data Collection & Normalization
- Data Analysis & the Dev. of a Single CER
- Dev. Of CER Linear Combination Model
- CER Validation

Data Collection and Normalization

Data Collection

- **Definition of the scope of weapon system**

  → **Artillery Weapon System / R&D Cost**

- **Data collection**

Cost data	**1**	**R&D Cost**
Specification data	**6**	**Max Range, Caliber, Weight, Length, Max rapidity of Fire, Continue rapidity of Fire**

- **The number of data**

  - The number of projects : 9

  - The number of cost drivers : 6

  → **9 (the number of projects) – 6 (the number of cost drivers) ≥ 2, it is feasible to develop the CER by regression analysis**

Data Collection and Normalization (cont')

Normalization

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

Data Analysis & The Dev. Of a single CER

Data Analysis

Cost Driver Selection

- **The selection for the variable was executed**

R^2, Adj.R^2, Forward/Backward/Stepwise Regression, C(p) Selection

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

Feasible Cost Drivers : Max Range, Weight, Caliber, Length

Data Analysis & The Dev. Of a single CER (cont')

Data Analysis (cont')

Multicollinearity

- **Judgment**

*** Correlation Matrix**

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

*** VIF and CI**

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**

Existence of multicollinearity

- **Solution**

#1 : Eliminate cost drivers(Max. Range and Length)

#2 : PCR, RR

Data Analysis & The Dev. Of a single CER (cont')

Data Analysis (cont')

Outlier

- **Judgment**

* **Studentized Residual, Studentized Deleted Residual, Cook's Distance, DFBETAS, etc.**

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

Outlier is suspected in the 9th data

- **Solution**

#1 : Eliminate the 9th data

#2 : Robust Regression

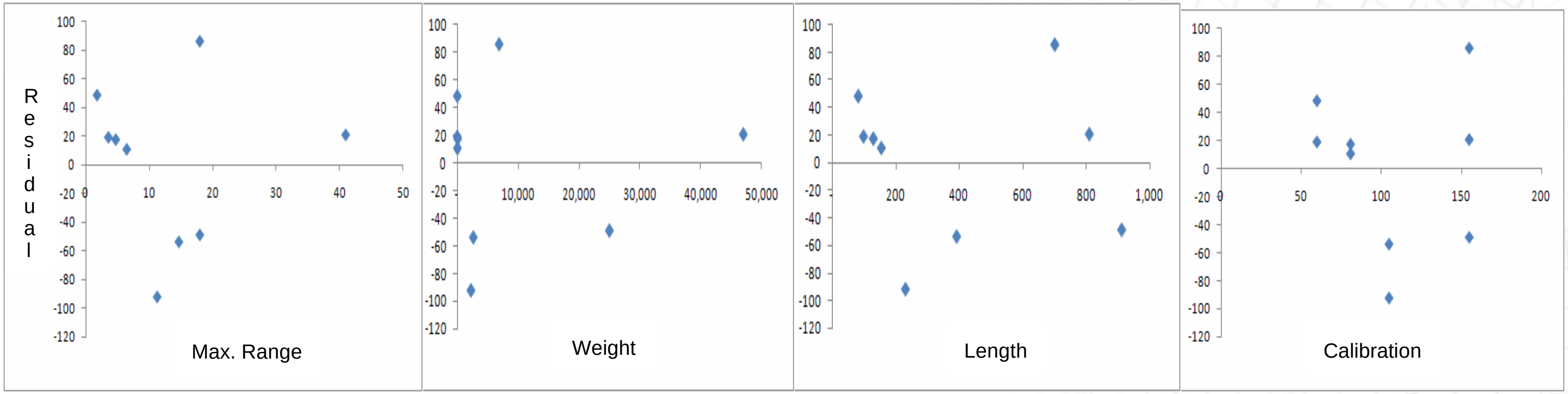
Data Analysis & The Dev. Of a single CER (cont')

Data Analysis (cont')

Heteroscedasticity

- **Judgment**

* **Scatter plot of residual**

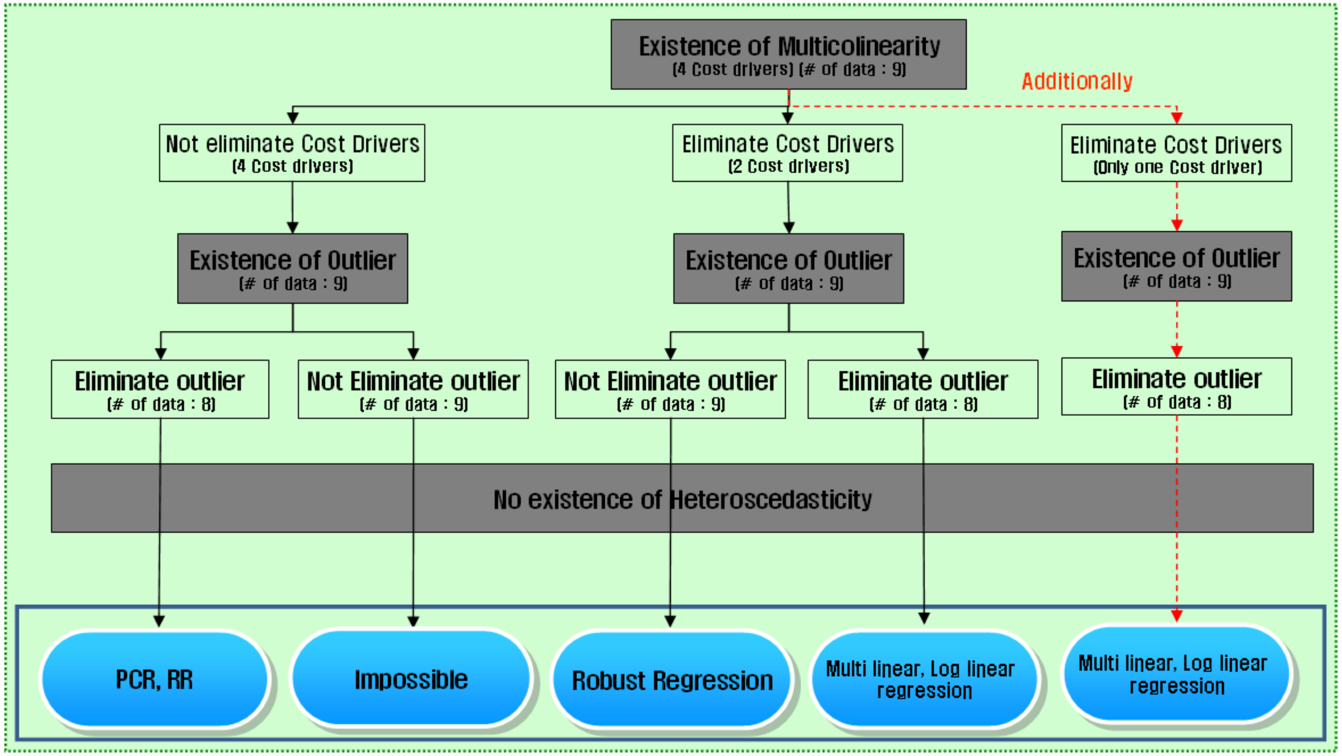


The residuals disperse irregularly. So no heteroscedasticity

Data Analysis & The Dev. Of a single CER (cont')

Development of a single CER

Feasible Single CERs according to the characteristics of data



Data Analysis & The Dev. Of a single CER (cont')

Development of a single CER (cont')

Test statistics of single CERs

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(B) Weight(C)	0.86 (0.81)	• Coefficient Pr>ChiSq : intercept=0.477, B=0.127, C=0.007
Linear(□)	Caliber(B) Weight(C)	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(C)	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

Dev. of CER Linear Combination Model

Selecting the Single CERs

Criteria to select the single CERs

$R^2 \geq 0.8$, Confidence Level : 5%

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(B) Weight(C)	0.86 (0.81)	• Coefficient Pr>ChiSq : intercept=0.477, B=0.127, C=0.007
Linear(I)	Caliber(B) Weight(C)	0.89 (0.85)	• Model Pr>F : 0.004 • Coefficient Pr>|t| : intercept=0.509, B=0.187, C=0.043
Log-linear(I)		0.78 (0.69)	• Model Pr>F : 0.023 • Coefficient Pr>|t| : intercept=0.431, B=0.218, C=0.832
Linear(II)	Weight(C)	0.84 (0.81)	• Model Pr>F : 0.001 • Coefficient Pr>|t| : intercept=0.0005, C=0.0014
Log-linear(II)		0.69 (0.64)	• Model Pr>F : 0.011 • Coefficient Pr>|t| : intercept=0.0002, C=0.01

CER # 1

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

CER # 2

$$Y_{Lin2} = 23.52163 + 0.0021Weight$$

Dev. of CER Linear Combination Model (cont')

Dev. Of CER Linear Combination Model

Formula

$$Min(RMSE_1, RMSE_2, RMSE_3, RMSE_4)$$

subject to

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

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

Definition of parameters

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)

Combination Models according to methods of placing the weight

	By SSE		By MMRE		By adj.R2		By Regression	
	CER_1	CER_2	CER_1	CER_2	CER_1	CER_2	CER_1	CER_2
SSE/MMRE/adjR^2	475.178	437.393	0.214	0.284	0.795	0.812	-	-
Weights	0.479	0.521	0.570	0.430	0.495	0.505	0.469	0.543
Model	$0.479*CER_1+0.521*CER_2$		$0.570*CER_1+0.430*CER_2$		$0.495*CER_1+0.505*CER_2$		$0.469*CER_1+0.543*CER_2$	

Dev. of CER Linear Combination Model (cont')

Model Selection

Comparison the RMSE and R^2 values of each combination model

	CER linear combining model				Single CER model	
	by SSE	by MMRE	By adj.R^2	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

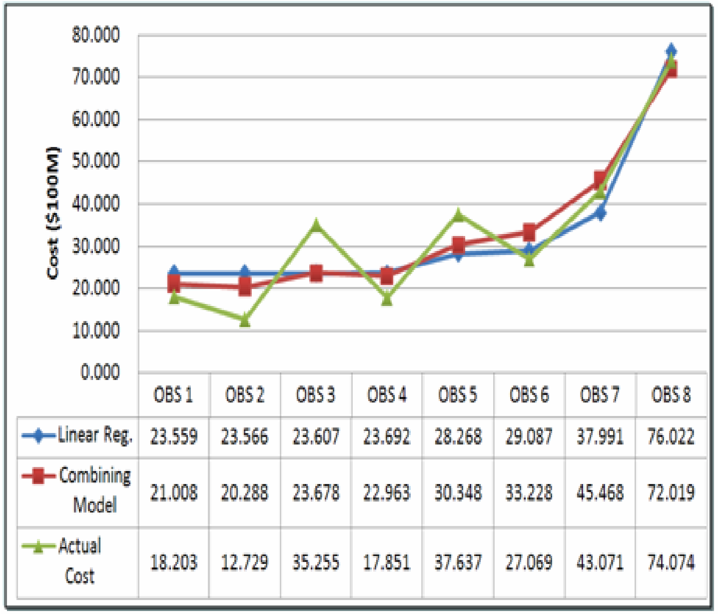
The final CER Linear Combination Model

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

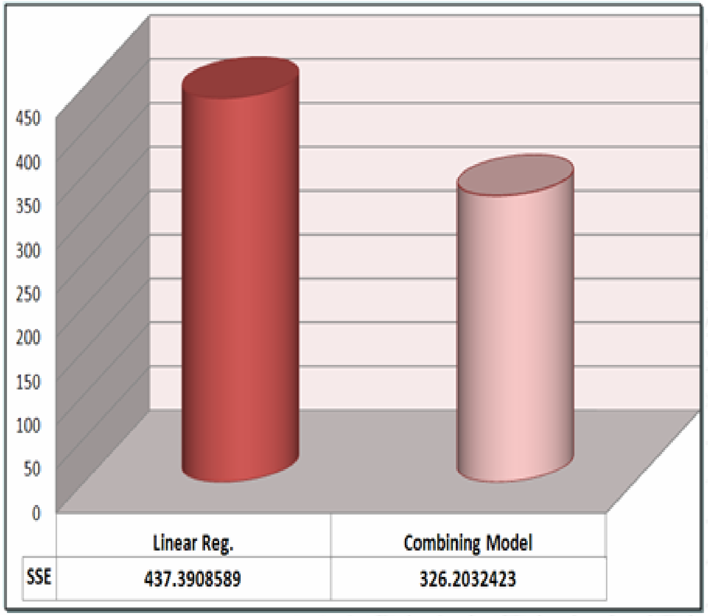
CER Validation

CER Validation

The comparison of the actual cost and the estimates



	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

The CER linear combination model is closer to the actual cost than the linear regression model

CER Validation

CER Validation (cont')

Accuracy Improvement

	By SSE	By MMRE	By adj.R^2	by Regression
SSE	328.449	328.326	328.275	326.203
Accuracy improvement rate(%)	24.91	24.94	24.95	25.42

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

The accuracy of CER linear combination model is better than the other regression model in this case study

CER Validation (cont')

CER Validation (cont')

Mean Magnitude of Relative Error(MMRE)

$$MMRE = (1/n)\sum_{i=1}^{n} MRE_i = 0.23 \leq 0.25$$

	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

Prediction al level l (PRED(l))

$$PRED(0.3) = q/n = 6/8 = 0.75 \geq 0.3$$

The final CER Linear Combination Model is acceptable

Conclusion

Conclusion

- **The process for CER development considering the various characteristics of data proposed in this study can be used as the standard process**

- **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**

- **The CER linear combination method is able to reduce the errors that occur by the single model**

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, production data under the situations in Korea.

Thank you

(Q&A)