Random Effects Logistic Regression Model for Ranking Efficiency in Data Envelopment Analysis

So Young Sohn[*]

Department of Information & Industrial Systems Engineering
Yonsei University, 134 Shinchon-dong, Seoul 120-749, Republic of Korea

Abstract *Ranking efficiency based on DEA results can be used for grouping DMUs. The resulting group membership can be partly related to environmental characteristics of DMU, which are not used either as input or output. Utilizing the expert knowledge on super efficiency DEA results, we propose a multinomial Dirichlet regression model which can be used for the purpose of selection of new projects. A case study is presented in the context of ranking analysis of new information technology commercialization projects. It is expected that our proposed approach can complement the DEA ranking results with environmental factors and at the same time it facilitates the prediction of efficiency of new DMUs with only given environmental characteristics.*

Keywords DEA, Ranking analysis, Multinomial Dirichlet regression model, IT project selection

1 Introduction

Data Envelopment Analysis (DEA), initially studied by Charnes et al. (1978), is a methodology used to measure and evaluate the relative efficiency of a set of homogeneous decision-making-units (DMUs) with multiple inputs and outputs. However, several limitations of DEA have been indicated: the risk of evaluating a DMU only with inputs and outputs and inability of efficiency prediction without inputs and accomplished outputs at the planning stage.

In an effort to resolve these limitations, Sohn and Choi (2005) suggested a random effects logistic regression model based on the DEA results. The authors incorporated two different outcomes of the DEA results (efficient DMU or

[*] Corresponding author. Tel.: +82-2-2123-4014; fax: +82-2-364-7807; *e-mail*: sohns@yonsei.ac.kr (S. Y. Sohn).

inefficient DMU) with a random effects logistic model that can accommodate not only the environmental characteristics which were left out from the DEA but also the uncertainty that cannot be explained by such environmental factors. The proposed random effects model can be used for the prediction of efficiency of a new DMU only with given environmental characteristic s.

What their model did not consider was the multi category of DEA outcomes. Several ranking approaches based on DEA have been proposed (Sexton (1986), Andersen & Petersen (1993) and Friedman & Sinuany-Stern (1997)). One of the widely used ranking methods is the super-efficiency technique, suggested by Andersen and Petersen (1993), which ranks DMUs through the exclusion of the unit being scored from the original DEA model. Upon availability of the ranking information, DMUs can be categorized into several groups. For instance, finished R&D projects can be rated as accepted, undetermined, or failed groups.

In this paper, we incorporate the group membership results with a random effects model that can accommodate not only the environmental characteristics which were left out from the super efficiency ranking analysis but also the uncertainty that cannot be explained by such environmental factors. Random effects model has been frequently used to accommodate both 'between cluster variation' as well as 'within cluster variation' (Sohn, 1993, 1996, 1997, 1999, 2002). In our case, the between cluster variation corresponds to the variation due to environmental factors while the within cluster variation reflects the random variation due to uncertainty that cannot be explained by such environmental factors.

Our proposed approach is illustrated in the context of ranking analysis on various technology commercialization projects clustered with respect to the type of information technology (IT), related R&D developer and its receiver (Sohn and Moon, 2003). It is expected that our approach can complement the efficiency-based ranking results with environmental factors and at the same time it facilitates the selection of new technology development scenarios at the planning stage.

The organization of this paper is as follows. In section 2, random effects model for DEA with multi categorical results is proposed. Section 3 presents the case study. Section 4 contains conclusions and future research issues.

2 Random effects model for DEA with multi category outcome

The general DEA model such as CCR (Charnes et al., 1978) and BCC (Banker et al., 1984) cannot generally be used for ranking DMUs, because the efficiency scores of DMUs are compared only with those related to the reference units. Andersen and Petersen (1993) developed the super-efficiency ranking method for only efficient units. The methodology enables an extremely efficient unit *o* to achieve an efficiency score greater than one by removing the o^{th} constraint in the primal formulation. The dual formulation of the super-efficiency model computes the distance between the Pareto frontiers evaluated without DMU$_{o.}$

Upon availability of the ranking results of super-efficiency method, all DMUs can be grouped into C exclusive categories ($c=1,\dots,C$). Additionally, a set of n DMUs can be classified into K homogeneous groups, each of which has n_k DMUs with the same environmental characteristics ($k=1,\dots,K$). Each group k generates responses $I_{k1},\dots,I_{k,n_k}$ and we assume $I_{kd}(d=1,\dots,n_k)$, can take one of C categories. Let n_{kc} be the total number of class c DMUs of the k^{th} group with n_k DMUs. Then one can assume that $n_{k1},\dots,n_{k,C}$ follow a multinomial distribution for given probability $p_{k1}, p_{k2},\dots, p_{k,C}$. The p_{kc} represents the probability that a randomly selected DMU within the kth group is categorized as level c. That is

$$n_{k1},n_{k2},\dots,n_{k,C}\big|p_{k1},p_{k2},\dots,p_{k,C} \sim MND\left(p_{k1},p_{k2},\dots,p_{k,C}:n_k\right) \tag{1}$$

or

$$g\left(n_{k1},n_{k2},\dots,n_{k,C}\big|p_{k1},p_{k2},\dots,p_{k,C}\right)=\frac{n_k!}{\prod_{c=1}^{C} n_{kc}!}\prod_{c=1}^{C} p_{kc}^{n_{kc}}, \tag{2}$$

where $n_k = n_{k1}+n_{k2}+,\dots,+n_{k,C}$ and $\sum_{c=1}^{C} p_{kc}=1$. The marginal mean and variance are $E(n_{kc})=n_k p_{kc}$ and $V(n_{kc})=n_k p_{kc}(1-p_{kc})$ respectively. Often these marginal mean and variance would vary over group mainly due to the environmental factors $z_{1k},\dots,z_{Qk}$ $(q=1,\dots,Q)$ associated with a particular group k. We call this 'between cluster variation'. As both the mean and variance are the functions of p_{kc}, we use a cumulative logit model for p_{kc} against the linear model of z_{qk}. That is

$$P_{kc}=\frac{\exp\left(g_{0c}+g_1 z_{1k}+\cdots+g_Q z_{Qk}\right)}{1+\exp\left(g_{0c}+g_1 z_{1k}+\cdots+g_Q z_{Qk}\right)}, \tag{3}$$

where P_{kc} is the cumulative probability that the result of group k turns out to be less than or equal to level c. g_{0c} and g_q denote intercept and regression coefficients of z_{qk}, respectively. When the p_{kc} is of our interest, it can be obtained as follows: $p_{kc}=P_{kc}-P_{k,c-1}$.

In model (3), we implicitly assume that p_{kc} is completely determined for a given z_{qk}. But it may not be necessarily true always. There could be the remaining part of variation in p_{kc} due to random error even with same environmental factors. We call this 'within cluster variation'. We introduce the following random effects model which can accommodate such variations. That is

$$p_{k1},\ldots,p_{k,C} \sim Dir\left(a_1,\ldots,a_C\right) \tag{4}$$

where Dir represents a Dirichlet distribution:

$$f\left(p_{k1},\ldots,p_{k,C}\right)=\frac{\Gamma\left(a_1+,\ldots,+a_C\right)}{\prod_{c=1}^{C}\Gamma\left(a_c\right)}\prod_{c=1}^{C}p_{kc}^{a_c-1} \tag{5}$$

Here a_c is assumed to be

$$\frac{\exp\left(g_{0c}+g_1z_{1k}+\cdots+g_Qz_{Qk}\right)}{1+\exp\left(g_{0c}+g_1z_{1k}+\cdots+g_Qz_{Qk}\right)}-\frac{\exp\left(g_{0,c-1}+g_1z_{1k}+\cdots+g_Qz_{Qk}\right)}{1+\exp\left(g_{0,c-1}+g_1z_{1k}+\cdots+g_Qz_{Qk}\right)}. \tag{6}$$

This is to reflect the covariate effects on the distribution of p_{kc} based on (3). Choice of Dirichlet distribution is due to the fact that it describes well the distribution of the probability and its conjugate relationship to multinomial distribution. Subsequently, the expected value and variance of p_{kc} can be obtained as follows:

$$E\left(p_{kc}\right)=\frac{a_c}{\sum_{c=1}^{C}a_c}, \tag{7}$$

$$V\left(p_{kc}\right)=\frac{a_c\left(\sum_{c=1}^{C}a_c-a_c\right)}{\left(\sum_{c=1}^{C}a_c\right)^2\left(\sum_{c=1}^{C}a_c+1\right)} \tag{8}$$

As the actual performance data $(n_{k1},\ldots,n_{k,C})$ are observed, the distribution for $p_{k1},p_{k2},\ldots,p_{k,C}$ can be updated:

$$p_{k1},\ldots,p_{k,C}\left|n_{k1},\ldots,n_{k,C}\right. \sim Dir\left(a_1+n_{k1},\ldots,a_C+n_{k,C}\right), \tag{9}$$

with

$$E\left(p_{kc}\middle|n_{k1},\ldots,n_{k,C}\right)=\frac{\left(a_c+n_{kc}\right)}{\left(\sum_{c=1}^{C}a_c+n_k\right)} \tag{10}$$

and

$$V\left(p_{kc}\middle|n_{k1},\ldots,n_{k,C}\right)=\frac{\left(a_c+n_{kc}\right)\left(\sum_{l\neq c}a_l+\sum_{l\neq c}n_{kl}\right)}{\left(\sum_{c=1}^{C}a_c+n_k\right)^2\left(\sum_{c=1}^{C}a_c+n_k+1\right)} \quad . \tag{11}$$

Marginal density of n_{kc} then can be derived as follows:

$$g\left(n_{kc}\right)=\frac{n_k!}{\prod_{c=1}^{C}n_{kc}!}\frac{\Gamma\left(\sum_{c=1}^{C}a_c\right)}{\Gamma\left(\sum_{c=1}^{C}a_c+n_k\right)}\prod_{c=1}^{C}\frac{\Gamma\left(a_c+n_{kc}\right)}{\Gamma\left(a_c\right)}, \tag{12}$$

With

$$E\left(n_{kc}\right)=E\left(E\left(n_{kc}\middle|p_{kc}\right)\right)=n_k\times\frac{a_c}{\sum_{c=1}^{C}a_c}, \tag{13}$$

and

$$V\left(n_{kc}\right)=V\left(E\left(n_{kc}\middle|p_{kc}\right)\right)+E\left(V\left(n_{kc}\middle|p_{kc}\right)\right)=n_k\times\frac{a_c}{\sum_{c=1}^{C}a_c}\times\left(1-\frac{a_c}{\sum_{c=1}^{C}a_c}\right)+f_{kc}$$

where

$$f_{kc}=\left(n_k^2-n_k\right)\times\frac{a_c\left(\sum_{c=1}^{C}a_c-a_c\right)}{\left(\sum_{c=1}^{C}a_c\right)^2\left(\sum_{c=1}^{C}a_{c^2}+1\right)}. \tag{14}$$

From (14), the variance of marginal distribution can accommodate the extra variability at an amount of f_{kc} that could not be captured in (1). Note that, if $n_k=1$, then the variance of marginal distribution is equal to that of conditional distribution.

In order to estimate unknown parameters, g_{0c} and g_q, we obtain the joint distribution of n_{kc} and unknown parameters, $\left(g_{0c}, g_1, \ldots, g_Q\right)$, are obtained by maximizing the following likelihood function of parameters:

$$
\begin{aligned}
&L_r\left(g_{01}, \ldots, g_{0C}, g_1, \ldots, g_Q : n_{11}, \ldots, n_{KC}\right)\\
&= \int \cdots \int \prod_{k=1}^{K} g\left(n_{k1}, \ldots, n_{k,C}\right) \qquad (15)\\
&= \int \cdots \int \prod_{k=1}^{K} f\left(p_{k1}, \ldots, p_{k,C}\right) g\left(n_{k1}, \ldots, n_{k,C} \middle| p_{k1}, \ldots, p_{k,C}\right) dp_1 \ldots dp_{C-1}.
\end{aligned}
$$

MLEs (maximum likelihood estimator) of unknown parameters cannot be found in a closed form. Algorithms to find MLEs require guesses about the initial values of those parameters. We suggest the use of the following fixed effect model to provide initial guesses about $\left(g_{0c}, g_1, \ldots, g_Q\right)$:

$$
L_f\left(g_{01}, \ldots, g_{0C}, g_1, \ldots, g_Q : n_{11}, \ldots, n_{KC}\right) = \prod_{k=1}^{K} g\left(n_{k1}, \ldots, n_{k,C} \middle| p_{k1}, \ldots, p_{k,C}\right), \qquad (16)
$$

where p_{kc} is defined as in (3). The resulting $\left(\tilde{g}_0, \tilde{g}_1, \ldots, \tilde{g}_Q\right)$ can be used as the initial values for MLEs.

After obtaining MLEs $\left(\hat{g}_{0c}, \hat{g}_1, \ldots, \hat{g}_Q\right)$, inferences on unknown parameters can be made evaluating the Fisher information matrix. This is used to find the standard error of each estimator from the inverse of negative Hessian matrix consisting of the second degree of partial derivatives of the corresponding log likelihood function with respect to a set of $\left(\hat{g}_{0c}, \hat{g}_1, \ldots, \hat{g}_Q\right)$. Then $\left(\dfrac{\hat{g}_q - g_q}{s.e\left(\hat{g}_q\right)}\right)^2$ is known to follow a c^2 distribution.

When the resulting MLEs $\left(\hat{g}_{0c}, \hat{g}_1, \ldots, \hat{g}_Q\right)$ replace $\left(g_{0c}, g_1, \ldots, g_Q\right)$ in (13) and (14), one can obtain $\hat{E}\left(n_{new,c}\right)$ and $\hat{V}\left(n_{new,c}\right)$ and they can be used to predict the performance of a new group with n_{new} DMUs for given environmental characteristics ($z_{1,new}, \ldots, z_{Q,new}$). Additionally, a $(1-a)100\%$ confidence interval for the expected performance of a new group, $E\left(n_{new,c}\right)$, can be approximately obtained as follows:

$$L_c = n_{new}\left(\frac{\exp\left(\hat{g}_{0c}+\sum_{q=1}^{Q}\hat{r}_q z_{q,new} - Z_{2/a}\,\hat{s}_{\hat{g}_{0c}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}}\right)}{1+\exp\left(\hat{g}_{0c}+\sum_{q=1}^{Q}\hat{r}_q z_{qnew} - Z_{2/a}\,\hat{s}_{\hat{g}_{0c}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}}\right)} - \frac{\exp\left(\hat{g}_{0,c-1}+\sum_{q=1}^{Q}\hat{r}_q z_{q,new} - Z_{2/a}\,\hat{s}_{\hat{g}_{0,c-1}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}}\right)}{1+\exp\left(\hat{g}_{0,c-1}+\sum_{q=1}^{Q}\hat{r}_q z_{q,new} - Z_{2/a}\,\hat{s}_{\hat{g}_{0,c-1}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}}\right)}\right) \qquad \text{(17-a)}$$

$$U_c = n_{new}\left(\frac{\exp\left(\hat{g}_{0c}+\sum_{q=1}^{Q}\hat{r}_q z_{qnew} + Z_{2/a}\,\hat{s}_{\hat{g}_{0c}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}}\right)}{1+\exp\left(\hat{g}_{0c}+\sum_{q=1}^{Q}\hat{r}_q z_{q,new} + Z_{2/a}\,\hat{s}_{\hat{g}_{0c}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}}\right)} - \frac{\exp\left(\hat{g}_{0,c-1}+\sum_{q=1}^{Q}\hat{r}_q z_{q,new} + Z_{2/a}\,\hat{s}_{\hat{g}_{0,c-1}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}}\right)}{1+\exp\left(\hat{g}_{0,c-1}+\sum_{q=1}^{Q}\hat{r}_q z_{qnew} + Z_{2/a}\,\hat{s}_{\hat{g}_{0,c-1}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}}\right)}\right) \qquad \text{(17-b)}$$

where

$$\hat{s}_{\hat{g}_{0c}+\Sigma_{q=1}^{Q}\hat{r}_q z_{q,new}} = \sqrt{\sum_{q=0}^{Q} z_{qnew}^{2} V\left(\hat{r}_q\right) + 2\sum_{q=0}^{Q-1}\left(z_{qnew} z_{q+1,new} Cov\left(\hat{r}_q, \hat{r}_{q+1}\right) + \cdots + z_{qnew} z_{Qnew} Cov\left(\hat{r}_q, \hat{r}_Q\right)\right)},$$

$\hat{g}_0 = \hat{g}_{0c}$ and $z_{0,new} = 1$. Here, L and U are, respectively, the lower and upper confidence limits for $E\left(n_{new,c}\right)$. This kind of interval estimation can help comparison of the expected performance among several different groups.

3 A Case Study

In this section, we apply the proposed approach to the empirical case. In order to evaluate the relative efficiencies of technology commercialization projects in the area of information technology (IT), we utilize the data obtained from Korean Information Technology Transfer Center in 1998. This covers 489 commercialization projects completed during 1993-1997. The questionnaires regarding thirty one variables were sent to the representatives of technology transferee companies, and 131 out of the received questionnaires were considered as DMUs after eliminating missing values and illogical responses. All thirty one variables were measured in 7 point Likert scale and had potential to be used as inputs or outputs for DEA. We used factor analysis to reduce the dimension of these variables and come up with a total of nine factors.

Each technology commercialization project was then evaluated in terms of six input and three output factors. Input factors used were the R&D ability of a technology provider, the technology receiver's management ability, the technology receiver's application ability of new technology, technology transfer center factor, market condition, and regulation factor. Output factors were technological commercialization success, spreading expect effect, and technology improvement in the company. All these factors were taken from SEM (Structural Equation Model) of Sohn and Moon (2003) where the observed variables in the questionnaires were used as components for latent variables of the confirmatory factor analysis.

Based on these input and output factors, we first used super-efficiency ranking analysis (Andersen and Petersen (1993) to rank 131 commercialization projects (DMUs) and categorized them into the three clusters: highly recommended group, in-between group and failed group. Considering the proportion of DMUs in each category as similarly as possible, DMUs with efficiency score higher than 1.0 are categorized as highly recommended, DMUs with efficiency scores less than 0.8 as failed ones, otherwise undecided ones. Support for the DMUs of undecided cluster is contingent and depends on the resources available.

We then relate the group membership of each DMU with related group characteristics representing the type of technology, technology provider and its receiver using the proposed model (3). The following dummy variables are used for each grouping of environmental characteristics:

i. Technology
  - ▪ Characteristic of technology field
    - Telecommunication: Communication net, interchange, facsimile (z_{11})
    - Information : Computer, S/W, Interface (z_{12})
    - Electric & broadcasting (z_{13})
    - Semiconductor/(machine) parts (Reference group)
  - ▪ Characteristic of product
    - Information and communication service (z_{21})
    - System and finished product (z_{22})
    - (machine) Parts (z_{23})
    - S/W (z_{24})
    - Etc. (Reference group)
  - ▪ Project Type
    - Government-run project (z_3)
    - Other project (Reference group)
  - ▪ Application
    - Existing business (z_4)
    - New business (Reference group)
  - ▪ Technology level
    - Copying level (z_{51})
    - Absorption level (z_{52})

- Improvement level (z_{53})
- Innovation level (Reference group)

ii. Technology provider

- Consortium
  - Joint research (z_6)
  - Independent research (Reference group)
- Institution
  - Corporation (z_7)
  - Research institute or university (Reference group)

iii. Technology receiver

- Company Size
  - 100 or more employees (z_8)
  - Less than 100 employees (Reference group)
- R&D expenditure ratio
  - 2.5% or more R&D expenditure ratio (z_9)
  - Less than 2.5% (Reference group)

These criteria can be considered as the environmental factors of DMU. Details of the levels of each environmental characteristic are given in Figure 1. Note that, the underlined level of each grouping criterion represents the reference group for the linear model in (3).

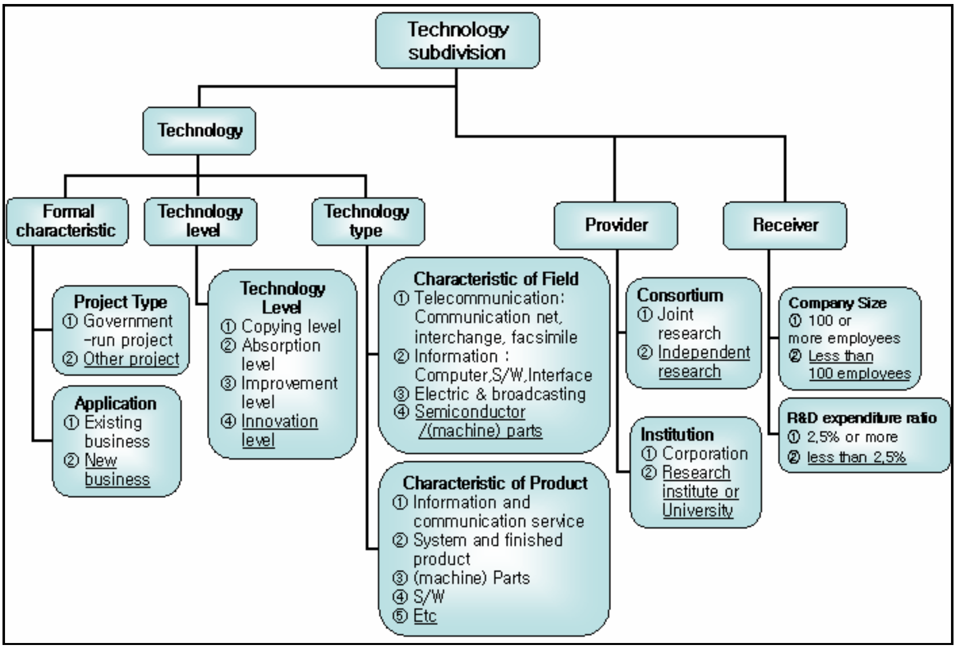

Figure 1. Grouping criteria of technology commercialization projects

All possible combinations of the levels of theses environmental characteristics are apparently 109, where we have a total of 131 DMUs. According to this combination, four groups have three members, 14 groups have two members and, the rest of them, 91 groups consist of a single member.

We apply this information to (15) where the nine kinds of categorical variables $(z_1,\ldots,z_9)$ are used to represent the nine combinations of environmental characteristics. In order to estimate MLE, we first find the initial values, $\tilde{g}_{0c}(c=1,2,3)$ and $\tilde{g}_q(q=1,\ldots,9)$, from the fixed effects model (16) based on the information for n_{kc}, n_k, and z_{qk} $(\mathrm{k}=1,\ldots 109)$. The results of the fixed effects model are displayed in Table 1.

Table 1. Results of the fixed effects model

Parameter		Estimates	Standard Error	Chi-Square	Pr > ChiSq p-value
$\tilde{g}_{01}$ (Intercept1)		-0.0607	1.1933	0.0026	0.9594
$\tilde{g}_{02}$ (Intercept2)		1.9312	1.2053	2.5672	0.1091
Characteristics of product	$\tilde{g}_{11}$ (Information & communication service)	-1.8666	0.7154	6.8081	0.0091
	$\tilde{g}_{12}$ (System & finished product)	-0.9634	0.6627	2.1131	0.1460
	$\tilde{g}_{13}$ (Machine parts)	-1.9446	0.6366	9.3311	0.0023
	$\tilde{g}_{14}$ (S/W)	-0.6035	0.6466	0.8711	0.3507
Characteristics of field	$\tilde{g}_{21}$ (Telecommunication)	0.2703	0.5555	0.2368	0.6265
	$\tilde{g}_{22}$ (Information)	-0.4726	0.6370	0.5505	0.4581
	$\tilde{g}_{23}$ (Electronic & broadcasting)	17.0989	1152.1	0.0002	0.9882
$\tilde{g}_3$ (Project type)		-0.7245	0.3988	3.2995	0.0693
$\tilde{g}_4$ (Application type)		0.4959	0.3776	1.7250	0.1890
Technology level	$\tilde{g}_{51}$ (Copying level)	-0.5514	0.8927	0.3816	0.5368
	$\tilde{g}_{52}$ (Absorption level)	-0.3469	0.9434	0.1352	0.7131
	$\tilde{g}_{53}$ (Improvement level)	0.0176	1.3389	0.0002	0.9895
$\tilde{g}_6$ (Consortium)		1.0043	0.4107	5.9794	0.0145
$\tilde{g}_7$ (Institution)		0.4312	0.4197	1.0554	0.3043

$\tilde{g}_8$ (Numbers of employees)	0.4122	0.4085	1.0186	0.3128
$\tilde{g}_9$ (R&D expenditure ratio)	-0.4143	0.5803	0.5098	0.4752

Next, we applied the Newton-Raphson method available in SAS PROC NLP (SAS Institute, 1998) to find the MLEs based on $(\tilde{g}_{0c}, \tilde{g}_1, \ldots, \tilde{g}_9)$. However, it failed to converge. This might be due to a highly nonlinear structure of our likelihood function and in an effort to reduce the dimension of parameter space, we only included in the linear model those which turn out to be significant in the fixed effects logistic regression.

For this purpose, Chi-Square test is performed at 10% level of significance using the *p*-values given in Table 1. As a result, three significant covariates are selected: characteristics of product, project type, and consortium type of project. According to these three criteria, 131 DMUs can be clustered into twenty groups as displayed in Table 2.

Table 2. Regrouping the 131 DMUs in terms of three environmental characteristics

Characteristics of Technology Combination			Group	Number of highly recommended DMUs in a group	Number of undecided DMUs in a group	Number of rejected DMUs in a group
Information & communication service	Government-run project	Joint research	GR_1	0	1	1
	Government-run project	Independent research	GR_2	2	1	8
	Other project	Joint research	GR_3	1	1	0
	Other project	Independent research	GR_4	1	2	2
System & finished product	Government-run project	Joint research	GR_5	2	2	1
	Government-run project	Independent research	GR_6	0	6	2
	Other project	Joint research	GR_7	2	4	1
	Other project	Independent research	GR_8	3	3	3
Machine parts	Government-run project	Joint research	GR_9	1	2	1
	Government-run project	Independent research	GR_{10}	0	1	7
	Other project	Joint research	GR_{11}	3	3	2
	Other project	Independent research	GR_{12}	2	6	6
S/W	Government-run project	Joint research	GR_{13}	4	2	2
	Government-run project	Independent research	GR_{14}	1	5	3

	Other project	Joint research	GR_{15}	3	4	2
	Other project	Independent research	GR_{16}	1	2	1
Etc	Government-run project	Joint research	GR_{17}	3	3	2
	Government-run project	Independent research	GR_{18}	1	1	1
	Other project	Joint research	GR_{19}	3	0	1
	Other project	Independent research	GR_{20}	3	0	0

After regrouping, we fit again the fixed effects logistic regression and the estimated parameters, $(\tilde{g}_{0c}, \tilde{g}_1, \tilde{g}_3, \tilde{g}_6)$, are set to be the initial values for (g_{0c}, g_1, g_3, g_6) as in Table 3. We then obtain the maximum likelihood estimators, $(\hat{g}_{0c}, \hat{g}_1, \hat{g}_3, \hat{g}_6)$, using (13) and the resulting parameter estimates as well as their standard errors are also displayed in Table 3.

Table 3. ML estimates for the random effects model

Parameter		Initial Estimates	Model based Estimates	Standard Error	Chi-Square	p-value
$\hat{g}_{01}$ (Intercept 1)		0.0763	-0.051171	0.14990	0.116527269	0.73283
$\hat{g}_{02}$ (Intercept 2)		1.8949	1.828600	0.15154	145.6072049	0.0001
(Characteristics of product)	$\hat{g}_{11}$ (Information & communication service)	-1.6159	-1.536953	0.15005	104.9173902	0.0001
	$\hat{g}_{12}$ (System & finished product)	-0.8944	-0.895568	0.14782	36.70562964	0.0001
	$\hat{g}_{13}$ (Machine parts)	-1.7171	-1.618300	0.15140	114.2526595	0.0001
	$\hat{g}_{14}$ (S/W)	-0.8376	-0.746162	0.15726	22.51268967	0.0001
	Reference group (Etc products)					
(Project type)	$\hat{g}_3$ (Government-run project)	0.8178	0.806884	0.04812	281.1685002	0.0001
	Reference group (Other project)					
(Consortium)	$\hat{g}_6$ (Joint research)	0.7372	0.803000	0.04913	267.1391638	0.0001
	Reference group (Independent research)					

According to the results in Table 3, all parameters in three covariates are significant at 10% level. This would indicate that the characteristic of product, project type, and consortium type are the important environmental factors that can be used to predict the efficiency of commercialization scenarios. In terms of the characteristics of product, most of commercialization projects tend to have lower ranks than the other types of IT product such as multimedia contents, security product, and A/S service. These results may be associated with the fact that these areas recently stand a spotlight in IT industry due to the rapid development of internet service, networking, and product liability, respectively.

As for the project type, the government-run projects tend to have higher ranks than the other types of projects. In an effort to escape from the financial crisis in 1998, Korean government has made a continuous investment on the commercialization of new technology in the filed of IT industry. As a result, government-run projects have had strong momentum which in turn induced the remarkable improvement on the efficiency of commercialization.

Similarly, the result for the consortium type shows that the commercialization projects in a form of the joint research have higher ranks than those of the independent research institution. It is concerned with that many companies perform project as a form of consortium in order to not only achieve higher performance but also maintain lower risk.

Using equations (4), (7), (8), (9), (10), and (11), we can obtain the conditional mean and variance of the number of highly recommended DMUs in each group. All of these results are displayed in Table 4.

The results in Table 4 show that mean varies over different technology groups due to the environmental factors to which they are exposed. As referred to in this paper, it is called 'between cluster variation'. At the same time, the DEA efficiencies in each group vary due to the random error following beta distribution. We call this 'within cluster variation'. From the varying variance, one can see that this is not constant over individual clusters.

The results show also that the posterior means of about half of all twenty technology groups (GR_2, GR_3, GR_4, GR_8, GR_9, GR_{13}, GR_{18}, GR_{19}, and GR_{20}) became higher than prior ones while the rest of them are lower than prior ones. However, all the posterior conditional variances are lower than the prior random effects.

When we compare the results in Table 4 with the sample means, which considers no environmental factors in Table 2, we find that the resultant posterior distribution reflects the degree of conformity of the observed data to the prior distribution for the efficiency. For example, it is shown that all the DMUs in GR_{20} are highly recommdended ones without considering any environmental factors. However, it is expected that only about 87% of DMUs would be highly recommended ones according to the result of our random effects model. On the other hand, GR_1, GR_6, and GR_{10} have no highly recommended DMUs from the super-efficiency method while it turns out from random effects analysis that there would exist about 5%, 1%, and 0.8% highly recommended ones in these three groups, respectively.

When unknown parameters are replaced with MLEs in (10), fitted model can be used to obtain the predictive distribution for the number of highly recommended DMUs of new technology group at the selection stage of several alternatives. At this stage, there would be no observed inputs and outputs except for the technology scenarios in terms of the grouping criteria.

In order to illustrate this, we use the seven test data, each of which contains five DMUs. These test data were originally reported by Sohn and Moon (2003) and were described in terms of the nine grouping criteria. We consider them as our scenarios. However, note that only the three of the nine covariates turned out to be significant through our analysis. Thus, we obtain the predictive distribution by using these three covariates. The predictive mean and variance, and 95% confidence interval for each technology group are obtained using (11), (12), (14) and are reported in Table 5.

In general, the expected number of highly recommended DMUs in all seven technology scenarios exceeds the 1/3 of the samples. Among the seven different kinds of technology scenarios, $\mathrm{GR_E}$ turns out to have the highest potential

On the other hand, $\mathrm{GR_A}$ and $\mathrm{GR_B}$ are expected to be the least effective scenarios ($2.613886/5 \approx 0.52$).

These results can be applied to select the potentially effective technology commercialization map among several alternatives at the planning stage of new technology development.

It is interesting to note that all of the 95% confidence intervals for the efficiency of scenarios are overlapping except for that for $\mathrm{GR_E}$. This is mainly due to the inflated variance of the transformed ML estimates for covariate parameters.

We compare these results with Sohn and Choi (2005) that categorized the technology scenarios into efficient and inefficient ones by adopting beta distribution for the random effects for the CCR based DEA. Major difference between the two results is as follows: there were no efficient groups in Sohn and Choi (2005) whereas all DMUs turn out to be the highly recommended groups by having 95% confidence intervals for $E(n_{new,1})$ exceeding 1/3 of the sample DMUs. Such contradictive results are mainly due to the fact that system and finished product turns out to be significantly meaningful in the ranking analysis while not in the efficiency analysis.

For more effective comparison, Table 6 shows the most frequent level of each environmental factor for the technology scenarios which are categorized into failed and undecided ones in ranking analysis. Note that all of these groups turned out to be inefficient ones in CCR efficiency analysis (Sohna and Choi 2005). As can be seen in Table 6, the commercialization projects for system and finished projects in the form of non-governmental projects transferred to the company with more than 100 employees have higher ranks than those for machine parts in the form of governmental projects transferred to the company with less than 100 employees.

Table 6. Most frequent level of each environmental factor for rejected and undecided DMUs

	Rejected DMUs	Undecided DMUs
Characteristic of product	**Machine parts**	**System and finished product**
Characteristic of technology field	Telecommunication: communication net, interchange, facsimile	Telecommunication: communication net, interchange, facsimile
Project Type	**Government-run project**	**Other project**
Application	New business	New business

Consortium	Independent research	Independent research
Institution	Corporation	Corporation
Company Size	**Less than 100 employees**	**100 or more employees**
R&D expenditure ratio	R&D 2.5% or more	R&D 2.5% or more
Technology level	Copying Technology Level	Copying Technology Level

4 Conclusions

In this paper, we proposed a multinomial Dirichlet regression model which can be used for the purpose of selection of new projects. A case study was presented in the context of ranking analysis of information technology commercialization projects. We showed that our approach can complement the DEA ranking results with environmental factors and at the same time it can facilitate the prediction of efficiency of new DMUs with only given environmental characteristics.

We illustrated both the efficiency-based ranking analysis on 131 IT technology scenarios and the prediction of the number of highly recommended DMUs for the given scenarios with six inputs, three outputs, and nine technology grouping criteria considered as covariates. According to our empirical result, technology, technology provider, and technology receiver have at least one influential environmental factor that plays significant roles in terms of ranking prediction. However, there are some other factors that are potentially important but were not included in our study as the characteristics of technology, technology provider, and technology receiver. They are internal factors such as the culture of company, the structure of compensation for commercialization success, the inclination of the company's CEO, and so on. They can be accommodated in the future survey.

Although the proposed approach can be effectively utilized to incorporate the environmental factors with the DEA results, it has also some limitations to be further considered. One of them is related to the difficulty in considering all possible factors due to the highly nonlinear structure of the likelihood function. Another problem is associated with the effective way to set initial values for MLEs.

Another expansion would be the choice of *K*, the number of groups. As can be seen from our empirical implementation, there could be many different ways to determine *K* (Green and Hensher, 2003). They are left for further areas of research.

In summary, the proposed approach can be widely applicable to a large family of real-world problems. They would be especially beneficial to the decision making process in resource management problems where some auxiliary characteristics of the organizations are available.

References

Anderson, P., and Peterson, N C. 1993. A procedure for ranking efficient units in data envelopment analysis. Management Science 39(10) 1261-1264.

Banker, R. D., Charnes, A., and Cooper, W. W. 1984. Some models for estimating technical and scale efficiency in data envelopment analysis. Management Science 30 1078-1092.

Charnes, A, Cooper, W. W., and Rhodes, H. 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2 429-444.

Charnes, A., Cooper, W. W., Lewin, A. Y., and Seiford, L. M. 1994. Data Envelopment Analysis: Theory, Methodology, and Application. Kluwer Academic Publisher. London.

Cooper, W. W. and Tone, K. 1997. Measures of inefficiency in data envelopment analysis and stochastic frontier estimation. European Journal of Operational Research 99 72-88.

Farrell, M. J. 1951. The measurement of productive efficiency. Journal of Royal Statistical Society Series A 120 253-290.

Fernández, C., Koop, G. and Steel M. 2000. A Bayesian analysis of multiple-output production frontiers. Journal of Econometrics 98 (1) 47-79.

Fernández, C., Osiewalski, J. and Mark F. J. Steel. 1997. On the use of panel data in stochastic frontier models with improper priors. Journal of Econometrics 79 (1) 169-193.

Fridman, L. and Sinuany-Stern, Z. 1998. Combining ranking scales and selecting variables in the data envelopment analysis context: The case of industrial branches. Computers and Operations Reserarch 25(9) 781-791.

Green, W. H., Hensher, D. A. 2003. A latent class model for discrete choice analysis: contrast with mixed logit. Transportation Research Part B 37 681-698.

Kohers, T., Huang, M and Kohers, M. 2000. Market perception of efficiency in bank holding company mergers: the roles of the DEA and SFA models in capturing merger potential. Review of Financial Economics 9 (2) 101-120.

Koop, G., Osiewalski, J. and Mark F. J. Steel. 1997. Bayesian efficiency analysis through individual effects: Hospital cost frontiers. Journal of Econometrics 76 (1-2) 77-105.

Korea National Statistical Office. 2002. Major statistics of Korean economy, Research Report, Seoul, Korea.

Kumar, V. and Jain, P. K. 2003. Commercialization of new technologies in India: an empirical study of perceptions of technology institutions. Technovation 23 113-120.

Lovell, C. A. Knox, Reinhard, S. and Thijssen, G. J. 2000. Environmental efficiency with multiple environmentally detrimental variables; estimated with SFA and DEA. European Journal of Operational Research 121 (2) 287-303.

Seiford, L. M. and Thrall, R. M. 1990. Recent development in DEA: the mathematical programming approach to frontier analysis, Journal of Economics 46 7-38.

SAS Institute. 1998. SAS/STAT User's Guide 6.03: SAS Institute, Cary, NC.

Sengupta, J. K. 1995. Dynamics of data envelopment analysis: theory of systems efficiency. Kluwert Academic Publishers. London.

Sohn, S. Y. 1993. A comparative study of four estimators for analyzing the random event rate of the Poisson process. Journal of the Statistical Computing and Simulation 49 1-10.

Sohn S. Y. 1996. Empirical Bayesian analysis for traffic intensity: M/M/1 queues with covariates. Queueing Systems 22 383-401.

Sohn, S. Y. 1997. Bayesian dynamic forecasting for attribute reliability. Computers and Industrial Engineering 33(3-4) 741-744.

Sohn. S. Y. 1999. Robust parameter design for integrated circuit fabrication procedure with respect to categorical characteristic. Reliability Engineering and System Safety 66 253-260.

Sohn, S. Y. 2002. Robust design of server capability in M/M/1 queues with both partly random arrival and service rates. Computers and Operations Research 29 433-440.

Sohn, S. Y. and Moon, T. H. 2003. Structural equation model for predicting technology commercialization success index. Technological Forecasting & Social Change 5553 1-15.

Sohn, S.Y., Yoon, K. B., Jang, I. S., 2005, Random Effects model for the reliability management of modules of a fighter aircraft. Accepted to Reliability Engineering and System Safety.

Sohn, S.Y., Yoon, K. B., 2005, Dynamic preventive maintenance scheduling of the modules of fighter aircrafts based on random effects regression model. submitted for publication, 2005.

Sohn, S. Y. and Choi, H. 2005. Random Effects Logistic Regression Model for Data Envelopment Analysis with Correlated Decision Making Units. JORS, accepted 2005.

Tsionas, E. G. 2003. Combining DEA and stochastic frontier models: An empirical Bayes approach, European Journal of Operational Research, 147 499-510.

West, M. and Harrison, J. 1989. Bayesian forecasting and dynamic models. Springer-Verlag. NY.

Table 4. Random effects & conditioanl means and variances of efficiency of twenty technology groups

Groups	Highly recommended DMUs	Random effect distribution $Dir(a_1,a_2,a_3)$			Highly recommended DMUs		Conditional distribution $Dir(a'_1,a'_2,a'_3)$			Highly recommended DMUs	
	Sample means	a_1	a_2	a_3	Prior mean	Prior variance	a'_1	a'_2	a'_3	Posterior mean	Posterior variance
GR_1	0.000000	0.169	0.402	0.428	0.169102	0.070253	0.169	1.402	1.428	0.056367	0.013297
GR_2	0.181818	0.083	0.290	0.626	0.083554	0.038286	2.083	1.290	8.626	0.17363	0.011037
GR_3	0.500000	0.169	0.402	0.427	0.169648	0.070434	1.169	1.402	0.427	0.389883	0.059469
GR_4	0.200000	0.169	0.402	0.427	0.169648	0.070434	1.169	2.402	2.427	0.194941	0.02242
GR_5	0.400000	0.487	0.374	0.138	0.48721	0.124918	2.487	2.374	1.138	0.414535	0.034671
GR_6	0.000000	0.147	0.383	0.468	0.147591	0.062904	0.147	6.383	2.468	0.016399	0.001613
GR_7	0.285714	0.487	0.374	0.138	0.48721	0.124918	2.487	4.374	1.138	0.310901	0.023805
GR_8	0.333333	0.279	0.438	0.282	0.279541	0.100699	3.279	3.438	3.282	0.327954	0.020036
GR_9	0.250000	0.487	0.374	0.138	0.48721	0.124918	1.487	2.374	1.138	0.297442	0.034828
GR_{10}	0.000000	0.077	0.277	0.644	0.077532	0.035761	0.077	1.277	7.644	0.008615	0.000854
GR_{11}	0.375000	0.487	0.374	0.138	0.48721	0.124918	3.487	3.374	2.138	0.387468	0.023734
GR_{12}	0.142857	0.158	0.393	0.447	0.158495	0.066687	2.158	6.393	6.447	0.1439	0.0077
GR_{13}	0.500000	0.487	0.374	0.138	0.48721	0.124918	4.487	2.374	2.138	0.498579	0.025
GR_{14}	0.111111	0.167	0.401	0.431	0.167393	0.069686	1.167	5.401	3.431	0.116739	0.009374
GR_{15}	0.333333	0.487	0.374	0.138	0.48721	0.124918	3.487	4.374	2.138	0.348721	0.020647
GR_{16}	0.250000	0.310	0.436	0.253	0.310596	0.107063	1.310	2.436	1.253	0.262119	0.032235
GR_{17}	0.375000	0.487	0.374	0.138	0.48721	0.124918	3.487	3.374	2.138	0.387468	0.023734
GR_{18}	0.333333	0.297	0.437	0.264	0.297746	0.104547	1.297	1.437	1.264	0.324436	0.043835
GR_{19}	0.750000	0.679	0.253	0.067	0.679577	0.108876	3.679	0.253	1.067	0.735915	0.032391
GR_{20}	1.000000	0.487	0.374	0.138	0.48721	0.124918	3.487	0.374	0.138	0.871803	0.022353

Table 5. Predictive mean, variance, and 95% *C.I.* for the performance of the seven technology scenarios

Group	Technology Scenario	Sample size	Highly recommended DMUs		
			$E(n_{new,1})$	$V(n_{new,1})$	95% $C.I.$
GR_A	① Information and Communication service ② Telecommunication: communication net, interchange, facsimile ③ Government-run project ④ Existing business ⑤ Innovation technology level ⑥ Independent research ⑦ Corporation ⑧Less than 100 employees ⑨ R&D 2.5% or more	5	2.613886	1.148593	[2.227029,2.995353]
GR_B	① Information and Communication service ② Telecommunication: communication net, interchange, facsimile ③ Government-run project ④ New business ⑤ Innovation technology level ⑥ Independent research ⑦ Corporation ⑧ Less than 100 employees ⑨ R&D 2.5% or more	5	2.613886	1.148593	[2.227029,2.995353]
GR_C	① System and finished product ②Telecommunication: communication net, interchange, facsimile ③ Government-run project ④ New business ⑤ Improvement technology level ⑥ Independent research ⑦ Corporation ⑧ 100 or more employees ⑨ R&D 2.5% or more	5	2.715893	1.887116	[2.316833,3.104166]
GR_D	① System and finished product ② Telecommunication: communication net, interchange, facsimile ③ Government-run project ④ New business ⑤	5	2.715893	1.887116	[2.316833,3.104166]

	Improvement technology level ⑥ Independent research ⑦ Corporation ⑧ 100 or more employees ⑨ R&D Less then 2.5%				
GR_E	① System and finished product ② Telecommunication: communication net, interchange, facsimile ③ Other project ④ Existing business ⑤ Absorption technology level ⑥ Joint research ⑦ Corporation ⑧ 100 or more employees ⑨ R&D 2.5% or more	5	3.519682	3.730697	[3.184723,3.815815]
GR_F	① System and finished product ② Semiconductor / (machine)parts ③ Government-run project ④ New business ⑤ Absorption technology level ⑥ Independent research ⑦ Research institute or University ⑧ Less than 100 employees ⑨ R&D 2.5% or more	5	2.715893	1.887116	[2.316833,3.104166]
GR_G	① Software ② Information: Computer, S/w, Interface ③ Government-run project ④ Existing business ⑤ Copying Technology Level ⑥ Independent research ⑦ Research institute or University ⑧ Less than 100 employees ⑨ R&D 2.5% or more	5	2.750466	2.090589	[2.551572,2.946206]