Three-stage-DEA model selections and managerial decision

Long-Fei Chen¹, Chih-Hui Hsiao²* and Chin-Fa Tsai¹

¹Department of Business Administration, National Chiayi University, 580, Hsin-Ming road, Chiayi City, 500, Taiwan.
²Institute of Marketing and Logistics Management, National Chiayi University, 600, Taiwan.

Accepted 16 August, 2010

Previous DEA studies have provided useful managerial information on improving the productivity of managers’ organizations. This study tries to clarify the implicit assumptions of Ruggiero’s (1998) three-stage DEA model and the data fitness that was overlooked in the three-stage-DEA. However, the model is extended to avoid possible bias or faulty results. As such, the distinct extended features are presented and discussed. With the presented features and a designated procedure including normality test, practitioners and researchers can apply this work to other related managerial applications as well as cited areas.

Key words: Model, assumption, stochastic frontier analysis (SFA), data envelopment analysis (DEA), three-stage DEA, environmental effects, normality test, linear regression.

INTRODUCTION

Data envelopment analysis (DEA) approach, first created by Charnes et al. (1978) and Banker et al. (1984), is an excellent empirical model that compares a decision unit with an efficient frontier using performance indicators. DEA is a popular tool used to analyze efficiency in many fields (Barros and Leach, 2006; Lin et al., 2010). For its importance and wide applications, Joe (2004a: 4) said “Managers are often under great pressure to improve the performance of their organizations…. Performance evaluation and benchmarking are a widely used method to identify and adopt the best practices as means to improve performance and increase productivity”. He also said “DEA uses mathematical programming techniques and models to evaluate the performance of peer units (for example, bank branches, hospitals and schools) in terms of multiple inputs used and multiple outputs produced”; “Since DEA was first introduced in 1978, over 2000 DEA-related articles have been published”; “DEA applications involve a wide range of contexts, such as education, health care, banking, armed forces, auditing, market research, retail outlets, organization effectiveness, transportation, public housing and manufacturing” and “Such previous DEA studies provide useful managerial information on improving the productivity of service business”. As such, its model form will be introduced in this paper. Since three-stage DEA is an extension based on DEA to consider the effects of environmental variables (or uncontrollable variables), its development of new models can also have important implications to DEA practitioners and researchers.

Stochastic frontier analysis (SFA) and data envelopment analysis (DEA) are popular models that are applied to many diversified fields to measure efficiency (Kilic et al., 2009; Hu et al., 2010). SFA is a well-developed statistical test to identify the effectiveness of the model description and has the ability to decompose the deviations from efficiency levels into noise and pure inefficiency (Barros, 2005). Its production frontier model is proposed by Aigner et al. (1977) and Meesusen and van den Broeck (1977). The model is of the form:

\[ y_i = \alpha_0 + x_i'\beta + \xi_i + \epsilon_i, \quad \xi_i = v_i - u_i, \quad i = 1, 2, ..., N \]

(where \( y_i \) is log output, \( x_i \) is KX1 vector of log inputs and \( \beta \) is the vector of regression coefficients)

The error term, \( \epsilon_i = v_i - u_i \), is made up of both a statistical noise term (vi) and the technical inefficiency (ui ≥ 0). It is often assumed that v is normal and u is half-normal. Also, v, u and x are assumed to be independent.
Table 1. A sample of adjusted $R^2$ (Model summary (b)).

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Std. error of the estimate</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.784(a)</td>
<td>0.614</td>
<td>0.607</td>
<td>0.090322</td>
<td>2.141</td>
</tr>
</tbody>
</table>

(a) Predictors: (Constant), ZB; (b) Dependent variable: DEA score.

Fried et al.’s (2002) three-stage DEA is a combination of the two models to purge the influence of environmental effects and statistical noise. This new DEA model involves a three-stage analysis. In the first stage, DEA is applied to outputs and inputs only, to obtain initial measures of producer efficiency. In the second stage, SFA is used to regress efficiency measures from the first stage against a set of exogenous environment variables. In the final stage, inputs are adjusted to account for the influence of environmental effects and statistical noise (good and bad luck) uncovered in the second stage and DEA is used to re-evaluate producer efficiency.

However, its identical individual distributions (i.i.d) are strong assumptions to apply (Amsler et al., 2009). When the data of applications do not fit normal distribution assumptions for SFA, the bias or adjusted $R^2$ needs to be explored. However, it is overlooked from an example presented in this paper. Several survey papers also ignore the fitness of SFA in their studies. Ruggiero’s (1998, 2010) three-stage DEA replace SFA in stage two with ordinary least squares (OLS) linear regression and was thought as a model without a specified functional form relative to the restrictions of Fried et al.’s (2002) three-stage DEA. However, after careful examination, there are implicit assumptions postulated on OLS linear regression (Poole and O’Farrell, 1971) than normal distribution restrictions. To apply the aforementioned models, one can not, but consider the goodness of fit for real world data. However, Reimann and Filzmoser (1999) have said “real world data are rarely well-behaved as classical statistical tests assume”. These give challenge to applications and motivate this study to have a little clarification and extension of it.

The following content begins with a literature review, as well as, a little clarification of Riggiero’s (1998, 2010) work and an extension to the situation when the data given are not normally distributed. Finally, a brief discussion and conclusion is presented.

MATERIALS AND METHODS

DEA has a wide spread application. Liang et al. (2006) extend DEA in the supply chain efficiency evaluation, while Seiford and Zhu (1998) have a study on decision support. Zhu (2004b) has an application in business negotiation, whereas Liang et al. (2008) have an extension in game theory with DEA study. Zhu (1998) applies DEA in evaluating the economic performance of Chinese cities, whereas Morita et al. (2005) studied the context-dependent data envelopment analysis. Chen and Zhu (2004) and Chen et al. (2006) studied the indirect impact the measuring information technology has on firm performance, while Cook et al. (2004) evaluated the effect of e-business activities. However, Seiford and Zhu (2002) and Yang and Pollitt (2009) studied the undesirable factors with application of carbon emissions reduction, which in recent times, is a high profile topic in energy consumption. Recently, Kao and Hwang (2010), Guan and Chen (2010) and Cooke et al. (2010) studied DEA’s extension to network DEA. Besides, DEA has been studied by several authors of African Journal of Business Management and closely related journal in diversified fields recently (Kareem et al., 2008; Sreekumar and Mahapatra, 2009; Sufian and Habibullah, 2009; Liu et al., 2010; Tung et al., 2010).

Fried et al. (2002) provided a three-stage DEA model for incorporating environmental factors and statistical noise into producer performance evaluation based on data envelopment analysis (DEA). According to Fried et al.’s proposal (Fried et al., 2002), DEA is applied to inputs and outputs to get initial measure of DMUs performance in the first stage. In the second stage, stochastic frontier analysis (SFA) is used to regress the measures from the first stage against a set of environmental variables. This gives a kind of decomposition in the variation of performance into a part attributable to environmental effects, a part attributable to managerial inefficiency and a part attributable to statistical noise. In the last stage (that is, the third stage), either outputs or inputs (depending on the input-oriented or output-oriented mode of the first stage DEA) are adjusted to account for the impact of the environmental effects and the statistical noise explored in the second stage, and as such, DEA was used to re-evaluate producer performance. Shang et al. (2008) use the cross sectional data of 2004 for 57 hotels in Taiwan to evaluate their relative performance and environmental effects from age, size and market condition with Fried et al.’s (2002) three-stage DEA. As mentioned before, there is an i.i.d restriction on the applied data. After screening the data for regression in stage two, neither the log technical efficiency nor the log environmental effects fit the truncated and normal distributions as needed. Besides, no bias and adjusted R square figures were reported. Other three-stage-DEA literatures are listed in Table 2. No adjusted R$^2$ and normality test are reported for their works.

In the survey papers, the industries include insurance and bank (Eling and Luhnen, 2010; Fethi and Pasiouras, 2010). In Eling and Luhnen’s (2010) work for efficiency measures in insurance industry, DEA approach has been most frequently used in empirical studies. Out of the 95 surveyed studies, 55 used DEA and 22 applied SFA. Sixty-one out of the 95 studies use at least labor and capital as inputs and most of them also added a third category (miscellaneous, mostly business services). There are three main insurance inputs: labor, business services and materials, and capital. However, neither normality test for SFA nor adjusted R square figures were reported. In Fethi and Pasiouras’s (2010) survey on DEA and artificial techniques for bank efficiency, four cases were reported to adjust the bank efficiency scores for risk and/or external environmental factors using a multi-stage DEA. As such, no normality test and adjusted R square figures were reported (Table 3).

It is assumed that the S output is a dependent variable for n DMUs in stage one, while the R exogenous variables are assumed to be in-discretionary independent variables in stage two for SFA analysis. There are R corresponding SFA regression equations to be estimated, and as such, the functional form (for output oriented mode) can be:

$$TOS_{ij} = f^j(Z_{ij}; \beta^j) + \nu_{ij} + \mu_{ij}, i = 1...S; j = 1...n$$

(1)
Table 2. Recapitulation of studies on the 3S-DEA frontier efficiency.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Method</th>
<th>Units</th>
<th>Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee (2008)</td>
<td>3S-DEA</td>
<td>89 global forest and paper companies</td>
<td>Expenses and interest expenses</td>
<td>Total sales and operating costs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Environmental variables: Exchange rate, GDP and population</td>
<td></td>
</tr>
<tr>
<td>Gorman and Ruggiero (2008)</td>
<td>3S-DEA</td>
<td>49 U.S. States</td>
<td>Sworn officers, employees and vehicles</td>
<td>Murders, violent crimes and total property crimes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Environmental factors: Percent of single mothers, the poverty rate, percent of individuals in the labor force, population and population per square mile.</td>
<td></td>
</tr>
<tr>
<td>Shang et al. (2008)</td>
<td>3S-DEA</td>
<td>57 hotels in Taiwan</td>
<td>Guest rooms, food and beverage capacity, employees and operating expense</td>
<td>Room revenue expense</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Environmental variables: Market condition, hotel size and age</td>
<td></td>
</tr>
<tr>
<td>Yang and Pollitt (2009)</td>
<td>3S-DEA</td>
<td>221 Chinese coal-fired power plants</td>
<td>Installed capacity, labour and fuel</td>
<td>Annual generation and undesirable output: SO₂ emissions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uncontrollable variables: Vintage, vintage squared, calorific value of coal, Scale 1, Scale 2, Scale 3 and CHP.</td>
<td></td>
</tr>
</tbody>
</table>

Where S is the number of output variables, n is the number of DMUs and \( f (Z) \) is the r-th deterministic feasible surplus frontier of deterministic output. \( \beta r \) is r-th estimated coefficient for \( Z_r \). The \( v_r \) is a statistical noise and is assumed to be \( v_r \sim N \left( 0, \sigma_r^2 \right) \). \( u_r \) is assumed to be the inefficiency of j-th output due to r factor, and has a truncated distribution \( u_r \sim N \left( 0, \sigma_r^2 \right) \) which is independent from \( v_r \). Thus, TOS (total output slack for output-oriented mode) can be decomposed into the effects of environmental factors and statistical noise, are filtered out like:

\[
y^A_{ij} = y_{ij} - \min \{ Z^j - Z^j \} - \min \{ v_r \} - v_r, \quad r = 1 \ldots S; j = 1 \ldots n \tag{2}
\]

Thus, some effects due to environmental effects and statistical noise can be filtered out by SFA in stage two, where the assumptions for the independence of \( x_r \) and \( v_r \), and \( u_r \) and \( v_r \) and \( x_r \) are too strong and the dependent variables can not be 0 (Amsler et al., 2009).

For cross-sectional stochastic frontier models, they rely on two kinds of strong assumptions. The first is the specific distribution assumptions that need to be made for noise and technical inefficiency, while the other is the errors which must be independent of the inputs.

Even with these strong assumptions, the estimates of the technical inefficiency are not consistent. Panel data allow us to relax some or all of these strong assumptions. However, the advantage comes at a cost dependent on the assumption that technical inefficiency is time invariant (Amsler et al., 2009, p8). Ruggiero (1998) provided a three-stage model to discuss non-discretionary inputs when \( S>1 \) (number of output variables) uses ordinary least squares (OLS) regression and gives a relaxation to the strong assumptions in stage two. In the first stage, in the analysis of DMU “0”, the standard DEA model is applied using only outputs and discretionary variables:

\[
F_0 = \min \Phi \quad \text{s.t.}
\]

\[
\sum_{i=1}^{n} \lambda_i y_{ij} \geq y_{0j}, \quad j=1, \ldots, S,
\]

\[
\sum_{i=1}^{n} \lambda_i x_{ik} \leq \sigma x_{0k}, \quad k=1, \ldots, M,
\]

\[
\sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0.
\]

Where S is the number of output variables, M is the number of input variables and n is the number of DMUs. The resulting index is composed of inefficiency and the environmental effect that non-discretionary variables have on the production. In the second stage, the following regression is applied:

\[
F_i = \alpha + \beta z_i + \epsilon
\]

Where \( \alpha \) is assumed to be a constant, \( z \) is the environmental effect on DMU i, \( \beta \) is the vector of regression coefficients and \( \epsilon \) is the estimated error.

Ruggiero (1998) showed that an overall index of environmental harshness can be obtained from \( z' = F^Z(y, x_i) \), which is the predicted first- stage index. This index is then used as a control variable in a third-stage model:

\[
\sum_{i=1}^{n} \lambda_i y_{ij} \geq y_{0j}, \quad j=1, \ldots, S,
\]

\[
\sum_{i=1}^{n} \lambda_i x_{ik} \leq \sigma x_{0k}, \quad k=1, \ldots, M,
\]

\[
\sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0, \lambda_i = 0 \text{ if } z' > z^0.
\]

This model only uses the parameter weights \( Z = \sum_{i=1}^{R} \beta_i z_i \) (where R is the number of non-discretionary variables) to construct the environmental harshness index, and there are no distributional...
assumptions to be made like those in SFA (Ruggiero, 1998, p 465). From the aforementioned review, it is seen that DEA has more application researches than SFA. With the limit of time and the fact that this paper is focused on three-stage-DEA model, other multiple-stage-DEA models and efficiency models are not in the scope of this study. Driven by the problem solving of more complicated applications, networked DEA and multiple-stage-DEA are two major theoretical research directions in trends that may be integrated together. This study meets this trend to do a little contribution to the theoretical research. Although Ruggiero’s (1998) three-stage-DEA and efficiency models are not in the scope of this study. Driven by the problem solving of more complicated needs less computation, the aforementioned simple review finds together. This study meets this trend to do a little contribution to the theoretical research directions in trends that may be integrated applications, networked DEA and multiple-stage-DEA are two major

<table>
<thead>
<tr>
<th>Model type</th>
<th>Fried et al.</th>
<th>Ruggiero</th>
<th>Proposed revision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage-I model</td>
<td>SFA</td>
<td>OLS regression</td>
<td>OLS regression</td>
</tr>
<tr>
<td>Data restriction</td>
<td>i.i.d for u and v</td>
<td>Normal distribution</td>
<td>Neo-normal distribution</td>
</tr>
<tr>
<td>Operation sequence and special care</td>
<td>Stage I, II and III</td>
<td>Stage I, II and III</td>
<td>Stage I, II* and III**</td>
</tr>
</tbody>
</table>

Note: *Normality test (Note 1) and adjust R² (Note 2) for regression model validation reported; **extension from the proposed model.

### New revision of three-stage DEA

However, it would be better if intensive consideration for hidden assumptions (Poole and O’Farrell, 1971; Ray and Das, 2010, pp304) and linear regression, through OLS in the second stage, is considered. When the samples are not normally distributed, the hidden assumptions for linear regression in Ruggiero’s (1998) three-stage-DEA second stage would be violated (Poole and O’Farrell, 1971). Then the error e in regression (4) can not be omitted after regression (that is, mean of e, is not zero).

With this consideration, a new proposed extended revision is derived as follows:

\[ z_i^2 = z_i^0 \] (where \( F_1 = \alpha + \beta z_0 + e_i \) and \( F_0 = \alpha + \beta z_0 + e_0 = 1 \))

When \( z_i^2 = z_i^0 \), the equations in form (5) needs to be revised to avoid multiple DMUs with \( \lambda = 1 \). A criterion is to filter out those DMUs with smaller noise. Regarding the previous equations, an improved model for the third-stage model (5) is as follows:

\[ \sum_{i=1}^{n} \lambda_i y_{ij} \geq y_{ij}, \text{ if } j=1, \ldots, S, \]

\[ \sum_{i=1}^{n} \lambda_i x_{ik} \leq e x_{ik}, \text{ if } k=1, \ldots, M, \]

\[ \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0, \lambda_0 = 0 \text{ if } \{ \lambda_i > z_i^2 \text{ or if } z_i^2 < z_i^0 \text{ and } e_i > e_0 \} \]

### DISCUSSION

In applying Fried et al.’s three-stage-DEA, the log-adjusted TE efficiency neither fit a normal distribution nor a well truncated-normal distribution from the P-P and histogram plots for ln (TE) score, like Figures 1 and 2 in the appendix from Shang et al. (2008) data. Furthermore, log-environmental effect also does not fit a well normal distribution as in Figure 3 of the appendix. So the log-linear regression model of SFA in Fried et al.’s three-stage-DEA is not well satisfied. In applying Ruggiero’s three-stage-DEA, the DEA efficiency in stage-1 did not fit a well normal distribution as in Figure 4. Likewise, the adjusted TE did not fit a well truncated normal distribution, nor did the environmental effect (ZB) fit a well normal distribution as illustrated in Figures 5 and 6 in the appendix. To apply the three-stage-model, the data can do a common transformation (like logarithm, square root and reciprocal) if the normality test in it failed. The practitioners can apply Jarque-Bera test (Wikipedia) to check which of the common transformation can bring the best goodness of fit to the applied data and then select the best one for the transformation.

Besides, the wikipedia (Wikipedia) also gives some advanced transformation (like Box-Cox transformation) for reference, as well as, a list of common transformations. Researchers and practitioners can do a further study to find an approximate parameter \( \alpha \) with least bias for all variables’ data. From this case, it can be seen that the data do not guarantee a normal distribution even after a common transformation (like logarithm transformation). Thus, a tolerance for the non-well normal distribution data is needed and the proposed three-stage-DEA revision model is to fit the necessity. However, the adjusted \( R^2 \) for the transformed data is a future work for further study. A sample of reported adjusted \( R^2 \) is shown in Table 1.

In daily business environments, one can face many situations that need managerial decision (Akinwale and Abiola, 2007; Chen, 2009; Duygulu et al., 2009; Ahangar, 2010; Soh, 2010; Yilmaz and Flouris, 2010; Yukcu et al., 2010). An alternative way is to find their normal inputs, outputs and environmental input variables, and apply the proposed model following the procedure or mix the proposed model with the current methods to solve the
current problems. Note that normal one-stage DEA model can be viewed as a special case of the three-stage-DEA without environmental effects. Here, the study recommends Zhu's (2004) book as a good reference which gets the highest five-star top rating in Amazon bookstore than other similar books and includes abundant spreadsheet DEA models together with over a dozen of his publications in application.

Zhu's (2004a) spreadsheet DEA modeling is a suitable tool without the number limit of DMUs to facilitate managers in conducting performance evaluation and analyzing decision alternatives without the help of sophisticated modeling programs. New DEA models for performance evaluation and benchmarking are collected there to facilitate evaluating business operations and processes in a variety of contexts. Particularly, the impact of IT to incorporate performance and evaluate the efficiency of a supply chain is both an important and practical problem for DEA applications. With a further consideration of uncontrollable (environmental) variables, problems can be solved in a three-stage-DEA way. In the past, managers have relied upon narrowly focused decision support systems (DSS) to facilitate the conventional decision process. More contemporary theory advances the integration of expert systems (ES) with DSS to overcome the deficiencies found in the conventional decision support systems. Yaiverbaum and Reynolds (1991) present a managerial problem identifier (MPI) that integrates decision support and expert systems. It can be a reference for practitioners to integrate three-stage-DEA with ES for the benefits of managerial problem identifier of ES.

Effective computer-based support for the use of analytic models in management decision making requires model management systems (MMS) that facilitate model selections in all phases (Banerjee and Basu, 1993). Existing approaches to the design of MMS commonly assume that the type of model needed to solve each problem is predetermined by the decision maker. Banerjee and Basu (1993) propose the design of the model selection subsystem in an integrated decision support system (DSS) to facilitate managers in their difficulty of selecting the model (type) process. This can facilitate practitioners to select models in the first two stages of the three-stage-DEA.

Conclusions

This study has done some clarification of assumptions and extension of Ruggiero's (1998) three-stage DEA. By exploring the assumptions of OLS regression, Ruggiero's (1998) claim for no assumption of efficiency overlooks the assumptions in the OLS regression, as well as, the goodness of fit in the real world data. Since real world data are rarely well-behaved as normality tests assume, one can not but consider the goodness of fit for real world data. Therefore neglecting the fitness of models will lead to biased or faulty results. This study has an extension to supplement Ruggiero's (1998) work by data checking and tolerance, in order for it to have a wider application.

The distinct extended features are presented and discussed to enable practitioners apply the proposed model to their applications, and to integrate it with decision support systems for complex decision problems. Since real data do not guarantee a normal distribution even after a common transformation (like logarithm transformation), this problem has always been ignored. Thus, the tolerance for non-well normal distribution data is needed and the proposed three-stage-DEA revision model is important to fit the necessity. With the presented features and a designated procedure including normality test, practitioners and researchers can apply this work to other related managerial applications, as well as, cited areas.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referees for their helpful and valuable comments.

REFERENCES


Appendix

Figure 1. The P-P plot for ln (TE) score is not well normally distributed.

Figure 2. The Histogram plot for ln (TE) score is non-well-truncated normally distributed.
Figure 3. The Histogram plot for ln (environmental effects) is not well normally distributed.

Figure 4. The Histogram plot for DEA scores in stage-I is not well normally distributed.
Figure 5. The Histogram plot for TE is non-well-truncated normally distributed.

Figure 6. The Histogram plot for environmental effects is not well normally distributed.
Normality test: P-P plot, Q-Q plot and histogram plot

For the visualization of normality test for data, P-P plot, Q-Q plot and histogram plot in SPSS can be applied in convenience (SPSS, 2006). Normality plots with tests can display normal probability and de-trended normal probability plots. If there are specified non-integer weights, the Shapiro-Wilk statistic is calculated when the weighted sample size lies between 3 and 50. For no weights or integer weights, the statistic will be calculated when the weighted sample size lies between 3 and 5000.

Adjusted $R^2$

Since $R^2$ will not decline with more added possible variables and approximate 1 in longer regression, adjusted $R^2$ can increase or decrease depending on if a new variable fit the regression in test with a penalty score and can supplement $R^2$ (Greene, 2008).