Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 166 (2005) 520–527
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/dsw

Interfaces with Other Disciplines

A distribution-free approach to estimating best response values with application to mutual fund performance modeling

Marvin D. Troutt [a,*], Michael Y. Hu [b], Murali S. Shanker [a]

[a] *Department of M&IS, College of Business, Kent State University, Rm. 426A, Business Adm. Bldg., P.O. Box 5190, Kent, OH 44242-0001, USA*
[b] *Department of Marketing, College of Business, Kent State University, Kent, OH 44242, USA*

Received 9 October 2003; accepted 3 February 2004
Available online 25 March 2004

Abstract

Frontier regression models seek to model and estimate best rather than average values of a response variable. Our proposed frontier model has similar intent, but also allows for an additional error term. The composed error approach uses the sum of two error terms, one an inefficiency error and the other as white noise. Previous research proposed assumptions on the distributions of the error components so that the distribution of this total error can be specified. Here we propose a distribution free approach to specifying these errors. In addition, our approach is completely data driven, rendering model specification an unnecessary step. We also outline, step-by-step, an approach to implementing this procedure. Our entire approach is illustrated with a mutual fund data set from the Morning Star database.

Keywords: Finance; Financial DSS; Securities; Statistics; Parameter estimation; Frontier estimation

1. Introduction

Frontier regression models seek to model and estimate the best rather than average values of response variables. Here we develop a model with similar intent, but also allow for an additional error term. In our model, the composed error term is separated into two components. The first, a non-negative error that represents inefficiency, or shortfall from ideal performance, and second, a white noise error term. Such approaches are desirable from the perspective of modeling observed performances. For example, with mutual funds, we can now distinguish between performance due to luck, the white noise error, and that due to performance ability, represented by the inefficiency error term.

Models to estimate the best response values have been proposed in the literature. For example, Aigner and Chu (1968) proposed a Cobb–Douglas function as a frontier model to fit firms' level data on production levels and factors. The parameter values, estimated using mathematical programming, were later shown to be maximum likelihood estimates (MLE) under some restrictive

* Corresponding author. Tel.: +1-330-672-1145; fax: +1-330-672-2953.
E-mail addresses: mtroutt@bsa3.kent. edu, mtroutt@kent. edu (M.D. Troutt), mhu@bsa3.kent.edu (M.Y. Hu), mshanker@kent.edu (M.S. Shanker).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2004.02.006

assumptions, specifically, that the errors were independent and identically distributed from the half-normal distribution. A modification of this model was considered by Meeusen and Van Den Broeck (1977) and Aigner et al. (1977). Here, actual performance is modeled as the frontier model plus an error term composed of two parts. The first error term is assumed to be normally distributed with mean zero, often described as accounting for uncertainty in the frontier model. The second error term is a nonnegative one representing a measure of inefficiency error or deviation from the efficient frontier. This term is also called the inefficiency effect in Coelli et al. (1998). The Aigner et al. (1977) method assumes that such nonnegative inefficiencies are distributed as half-normal. Stevenson (1980) extended the method to permit assumption of truncated normal and gamma distributions. Another approach was proposed by Green (1990). It assumes a gamma distribution for the inefficiency error terms. However, Ritter and Léopold (1997) have found that such models are difficult to estimate accurately. Recently, Van Den Broeck et al. (1994) have considered Bayesian models. These models require a large number of assumptions such as the form of the likelihood function and a choice of prior distribution for the estimated parameters.

The model proposed in this paper does not require any distributional assumptions on the errors. Our approach, based on mathematical programming, is distribution free and totally data driven. We show that meaningful parameter estimates can be derived without having to make restrictive distributional assumptions.

The rest of the paper is organized as follows. The next section explores the MLE principle for estimation of these two errors. We show that the likelihood function is U-shaped and is unbounded at solutions corresponding to a pure frontier solution, and to an ordinary least squares (OLS) regression solution. Thus, the MLE principle does not yield a meaningful solution to our problem, thereby providing a rationale to consider different approaches to parameter estimation. We then present our model in Section 3, followed by our data driven procedure to estimate parameter values in Section 4. Experimental results are in Section 5, followed by conclusions in Section 6.

2. Maximum likelihood estimation

The general composed error frontier estimation model can be written as

$$y_j = f(x_j, \theta) + \varepsilon_j - \omega_j, \tag{1}$$

where for $j = 1, \ldots, n$, y_j is a measurement on a dependent variable, x_j is a vector of measurements on independent variables in $\Re^m$, θ is a vector of model parameters in $\Re^p$, ε_j is a white noise error term with variance σ^2, and ω_j is a nonnegative inefficiency error for observation j, independent of ε_j. $f(x_j, \theta)$ is a "ceiling" type frontier model—that is, observations without other errors will fall beneath the level given by the ceiling model. A "floor" model is the opposite, and model specification becomes

$$y_j = f(x_j, \theta) + \varepsilon_j + \omega_j. \tag{2}$$

If the ε_j's are set to zero then the resulting model is called the *pure frontier* model. If the ω_j are set to zero then the resulting model is called the *pure OLS* model.

We claim first that no MLE estimates of the parameters of this model are possible. More precisely, the likelihood function is unbounded at two distinct sets of parameter values, namely, those corresponding to the pure frontier and pure OLS solutions, respectively. Consider, for example, the case for which the ω_j's are considered to be exponentially distributed with density

$$f_\omega(\omega) = \mu \exp(-\mu\omega). \tag{3}$$

The maximum likelihood estimation problem for the ceiling model may be written as

PL:

$$\text{Max}\left\{ \Pi_j \left(\frac{1}{\sigma\sqrt{2\pi}} \right) \exp\left(\frac{-\varepsilon_j^2}{2\sigma^2} \right) \right\} \{ \Pi_j \mu \exp(-\mu\omega_j) \} \tag{4}$$

s.t.

$$y_j = f(x_j, \theta) + \varepsilon_j - \omega_j \quad \text{for all } j, \tag{5}$$

$$\omega_j \geqslant 0 \quad \text{for all } j,$$

$$\sum_j \varepsilon_j = 0,$$

$$\sigma^2 > 0,$$

$$\mu > 0.$$

We note that for any feasible ε_j and ω_j, the conditionally optimal values of σ^2 and μ are known from the elementary MLE results for the two densities in question. Namely, we have

$$\sigma^{*2} = \frac{1}{N-1}\sum_j \varepsilon_j^2 \quad \text{and} \quad \mu^* = \frac{N}{\sum_j \omega_j}. \tag{6}$$

Then the objective function in (4) can be rewritten as

$$\sigma^{-N}(2\pi)^{-N/2}\exp\left(-\sum_j \varepsilon_j^2/2\sigma^2\right)\mu^N \times \exp\left(-\mu\sum_j \omega_j\right). \tag{7}$$

Substitution of (6) into (7) yields the simplified result

$$(2\pi)^{-N/2}\left[\frac{\sum_j \varepsilon_j^2}{N-1}\right]^{-N/2} \times \exp\left(\frac{-(N-1)}{2}\right)\left[\frac{N}{\sum_j \omega_j}\right]^N \exp(-N). \tag{8}$$

Neglecting constant terms in the simplified objective function, the revised estimation problem becomes

PLS:

$$\text{Max}\left[\sum_j \varepsilon_j^2\right]^{-N/2}\left[\sum_j \omega_j\right]^{-N} \tag{9}$$

s.t.

$$y_j = f(x_j,\theta) + \varepsilon_j - \omega_j \quad \text{for all } j,$$

$$\omega_j \geqslant 0 \quad \text{for all } j,$$

$$\sum_j \varepsilon_j = 0. \tag{10}$$

Assume for simplicity that $f(x_j,\theta)$ has range all of $\Re$ as θ is varied. This is the case, for example, with linear regression models. Let $\sum_j \varepsilon_j^2 = \delta$ and consider the case of $\delta \to 0$. The first term of (9) has as infinite limit in that case due to the negative exponent. Furthermore, the corresponding ω_j must tend to those of the pure frontier model and give a finite limiting value to the second term of (9). It follows that as $\delta \to 0$, the value of (9) tends to infinity. Similarly, considering $\sum_j \omega_j = \delta \to 0$, shows that the corresponding ε_j tend to the pure OLS solution and that the value of (9) tends to infinity, as in the first case.

Thus, the likelihood function in (9) is U-shaped, being unbounded at each extreme pure solution case. Hence, a unique bounded MLE solution does not exist for this problem. It is easily seen that the assumption of the exponential density for the ω_j is not critical to the argument so that the same result is expected for other density assumptions.

One interpretation of the foregoing discussion is that such composed error models, while intuitively attractive, are not estimable based on the maximum likelihood criterion. That is, given any sets of feasible error terms of the above types, their likelihood will be dominated by solutions for which one of the sets of errors is identically zero. This suggests that with respect to likelihood, both a pure frontier model and a pure OLS model will dominate other feasible composed error models. The result would appear to suggest that the model builder should decide to use precisely one or the other of the pure frontier or pure OLS models. However, since both their likelihoods are unbounded, a choice cannot be made on that basis.

Ritter and Léopold (1997) have studied the estimation of normal-gamma composed error models and noted several difficulties with their accurate estimation. In view of the foregoing discussion, such difficulties might be expected. More generally, previous likelihood based studies in this area may possibly have obtained only relative maxima solutions of some kind.

Thus, it appears that composed error models are not estimable in the maximum likelihood sense. Nevertheless, they are intuitively compelling. In most cases the dependent variable values are likely to have at least some measurement error. Moreover, when the dependent variable values are performance results, common experience suggests that performance often involves both chance or luck variations as modeled by ε_j, and skill or performance variations as modeled by ω_j. As noted earlier, the white noise error components are often described as representing uncertainty in the

model form. Here we are employing a different interpretation. Namely, we assume these represent a chance component of performance; while the nonnegative errors represent the variation due to purposeful attempts at ideal performance. Most importantly, the predictive ability of such models may vary depending on the level of the white noise component being assumed. Hence, it is desirable to propose a reasonable remedial approach for this class of models. In the next section, we develop an application to model mutual funds performances. An estimation procedure is proposed that permits separate estimates of the two kinds of error components rather than their sum.

3. Application to mutual fund data

Indro et al. (1999) provided theoretical argument and empirical evidence for the relationship between mutual fund size and fund returns. A sample of 683 nonindexed, actively managed US equity funds was obtained from the Morningstar's Mutual Funds OnDisc in the 1993–1995 period. Results showed a curvilinear relationship between size and returns. Funds have to exceed a certain size in order to efficiently cover the cost of information. Yet, the marginal returns diminished and became negative as the fund exceeds an optimal fund size.

The same data set is used in this study to illustrate how the frontier model approach can be used to estimate the relationship between returns and fund size. Fund size is typically measured as the net assets amount. The three-year average for the sample is \$843.6 million with a standard deviation of \$1,869.28 million. The natural logarithm of month-end net assets under management is used in this study. The mutual funds in our sample achieved a 13.40% average annual return for the three years (standard deviation = 5.36%) as compared to average return on the S&P 500 of 15.28%. The optimal fund size according to Indro et al. (1999) was between \$946 million and \$1.1 billion of net assets.

We now construct a frontier model for the foregoing mutual fund data. Our model is based on the following:

Observed performance
= Predicted optimal performance (frontier)
+ model and other random error
− performance shortfall for the observed period.

Let

d^* ideal fund size for managing funds in the chosen category,
k model parameter, $k \geqslant 0$,
y^* expected optimal relative performance measure for funds in the chosen category if operated at ideal size, d^*,
ε_j normally distributed mean zero, error term reflecting model error and errors from all other sources except managerial efficiency/inefficiency relative to the estimated frontier model,
ω_j nonnegative underachievement of ideal performance for mutual fund (MF) j-distribution assumption discussed below,
d_j actual fund size in net assets for MF j.

Thus, the frontier model can be written as:

$$y_j = f(x_j, \theta) + \varepsilon_j - \omega_j = y^* - k(d^* - d_j)^2 + \varepsilon_j - \omega_j, \quad (11)$$

where the non-linear model $f(x_j, \theta)$ represents the predicted optimal performance. In our mutual fund example, the x_j's are the d_j's, and θ has three non-negative components: $\theta_1 = y^*$, $\theta_2 = k$, and $\theta_3 = d^*$. We now consider how to estimate the parameters and residuals for this model without additional distributional assumptions on the latter.

3.1. Frontier estimation given d^*

We will solve model (11) by providing estimates for d^*. We assume that $m = \min(d_j) \leqslant d^* \leqslant \max(d_j) = M$. That is, we believe that ideal fund size is in the interval of actually observed fund sizes (net assets). So we break up the interval $[m, M]$ into, say, n steps and solve each problem with a trial value of d^* given by these steps. For a given such d^* value, say, d, then the mutual fund model of (11) becomes:

$$y_j = y^* - k(d - d_j)^2 + \varepsilon_j - \omega_j. \quad (12)$$

Given d^*, this model is linear in the parameters y^* and k, except that we must enforce their nonnegativities.

4. Procedure

Let d^* denote the ideal fund size. To generate potential solutions to Model (12), we consider the revised model below:

$$\min \quad \lambda \sum_j \varepsilon_j^2 + (1-\lambda) \sum_j \omega_j^2 \tag{13}$$

s.t.

$$y_j = y^* - k(d^* - d_j)^2 + \varepsilon_j - \omega_j \quad \text{for all } j,$$
$$\omega_j \geqslant 0 \quad \text{for all } j,$$
$$\sum_j \varepsilon_j = 0,$$
$$y^*, k \geqslant 0.$$

Model (13) uses λ as a parameter to generate potential solutions for a fixed fund size. When λ is near zero, then all effort is placed on minimizing the sum of the ω^2 errors. When λ is near unity, all effort is focused on minimizing the sum of the squared white noise error terms. Thus, a range of intermediate possible solutions is generated as λ ranges from zero to unity. It should be observed that this model requires λ to be at least positive. That is, $\lambda = 0$ cannot be permitted in this model. If $\lambda = 0$, then the second term in Model (13) might be made smaller than that of the pure frontier model by an appropriate adjustment of the ε terms, which would be essentially unrestricted in that case. For similar reasons, we restrict λ to be strictly less than 1. As such, we consider $\lambda \in (0, 1)$. As we average three years of performance data to produce the observation for each fund, the dependence is only on subscript j.

As the values of d^* and λ need to be specified before solving Model (13), we proceed as follows:

1. We first randomly divide our dataset into two equal parts A and B. Initially, let A be the training set, and B be the test set.
2. Let n be the predetermined number of d^* values that we would like to consider. Initially, set $d^* = \min(d_j)$, where d_j is a fund size observed in the training set. Subsequent values of d^* are then determined as $d^* \leftarrow d^* + \frac{d_{\text{range}}}{n}$, where d_{range} represents the range of fund sizes in the training set.
3. For each of the values of d^*, solve Model (13) by varying the value of λ from 0 to 1.
4. For each trial combination of d^* and λ, apply the results of the model solution on the test set and calculate the average residual. Specifically, For a particular trial l, let $\bar{\varepsilon}_l$ and $\bar{\omega}_l$ represent the average values of the ε and ω from the model solution. Then, for each fund i in the test set, calculate the residual as follows:

$$R_{il} = \left(y_i - (y^* - k(d^* - d_i)^2 + \bar{\varepsilon}_l - \overline{\omega}_l)\right)^2,$$
$$\overline{R}_l = \frac{\sum_i R_{il}}{N}, \tag{14}$$

   where it is assumed that the values of y^*, k, and d^* are those specific for iteration l, and N is the test sample size.
5. The ideal value of the fund size is then the solution that produces the smallest average residual $\min(\overline{R}_l)$. The R_{il} squared residuals may be interpreted as follows. From the solution on the training set, the best prediction for a hold-out sample MF would be based on the mean residual. Thus the best prediction would be given by the inside parenthetical expression in (14). Hence, the R_{il} give the squared residuals between that prediction and a given hold-out MF i. Thus the criterion we propose is to choose that solution for which the average squared hold-out sample error is as small as possible.
6. As validation, repeat steps 2–5 by reversing the training and test sets. That is, now B is the training set, and A the test set.

It is important to emphasize that the optimization of the objective function in (13) does not extend over λ. The parameter λ is used to generate different allocations between the white noise errors and the inefficiencies. Thus the optimal value of λ is not determined from the programming problem at (13). Instead, as different sets of solutions of (13)

are generated, their impact on the average squared residuals based on (14) is calculated. From these results, the optimal λ can then be determined.

We illustrate the technique on a mutual fund performance data set in the next section.

5. Results

The above procedure was applied to our dataset. Towards that end, the data of 684 observations were equally divided into two sets A and B. The variables in this data set consisted of the size $d_j(=\ln((d_1+d_2+d_3)/3))$ and the performance y_j of a fund. To determine the value of d^* for a particular optimization of Model (13), the range of fund sizes d_{range} was first determined from the training set, and n, the number of d^* values to consider was chosen to be 22. λ was similarly chosen by setting the initial value of $\lambda = 0.1$, and then incrementing λ by 0.1 for each new iteration of the model, until $\lambda = 0.9$. In principle, the results may be affected by the granularity, i.e., the number of values, of λ and d^*. Preliminary experiments with this particular dataset indicated that the results were robust with respect to the granularity, and as such we only present a representative set of results.

The results of running the above search procedure leads to 198 (22×9) model solutions counting each trial combination of d^* and λ. The results are presented in Table 1, where the minimum values of the average squared residuals are shown for each respective size d^*. The optimal size is at $d^* = 6.9489$ giving an average residual of 22.197 when using A as the training set; and $d^* = 6.7979$ with an average residual of 23.284 when using B as the training set. It should be noted that for these optimal sizes, we get feasible solutions across all ranges of chosen λ values. These feasible solutions are shown in Table 2, and also plotted in Figs. 1 and 2 for the range of λ values. Clearly, as $\lambda \to 1$, the model goes towards a pure frontier model.

Table 1
Optimal sizes, and minimum test residuals

Training set A, test set B		Training set B, test set A	
d^*	$\overline{R}$	d^*	$\overline{R}$
0.51083	28.862	0.03279	26.826
0.94003	28.862	0.45561	26.826
1.3692	28.862	0.87843	26.826
1.7984	28.862	1.3013	26.826
2.2276	28.862	1.7241	26.826
2.6568	28.862	2.1469	26.826
3.086	28.862	2.5697	26.826
3.5152	28.862	2.9925	26.826
3.9444	28.862	3.4154	26.826
4.3736	28.845	3.8382	26.826
4.8028	28.254	4.2610	26.826
5.232	26.604	4.6838	26.830
5.6613	24.119	5.1067	26.484
6.0905	22.594	5.5295	25.496
6.5197	22.201	5.9523	24.210
6.9489	**22.197**	6.3751	23.486
7.3781	22.331	**6.7979**	**23.284**
7.8073	22.504	7.2208	23.333
8.2365	22.680	7.6436	23.462
8.6657	22.844	8.0664	23.602
9.0949	22.992	8.4892	23.732
9.5241	23.122	8.912	23.848

Table 2
Model performance across lambda

λ	Training set A: $d^* = 6.9489$				Training set B: $d^* = 6.7979$			
	y^*	k	$\overline{\omega}$	$\overline{R}$	y^*	k	$\overline{\omega}$	$\overline{R}$
0.1	15.11	0.347	0.182	22.205	15.38	0.359	0.183	23.284
0.2	15.30	0.345	0.386	22.202	15.55	0.353	0.387	23.284
0.3	15.52	0.342	0.619	22.200	15.75	0.347	0.618	23.286
0.4	15.78	0.340	0.891	22.198	15.99	0.340	0.885	23.290
0.5	16.10	0.338	1.215	22.197	16.27	0.334	1.203	23.296
0.6	16.50	0.337	1.622	22.197	16.64	0.329	1.596	23.304
0.7	17.04	0.338	2.160	22.197	17.14	0.325	2.116	23.310
0.8	17.85	0.340	2.952	22.198	17.89	0.324	2.874	23.310
0.9	19.43	0.354	4.476	22.219	19.29	0.337	4.206	23.293

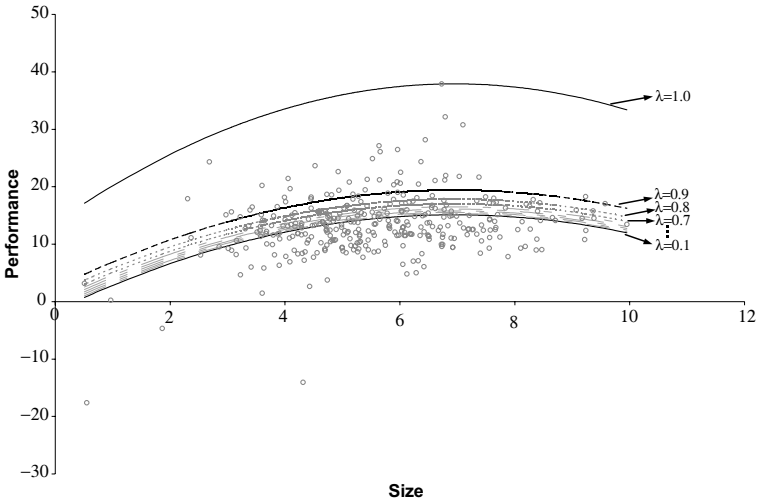

Fig. 1. Training set *A*.

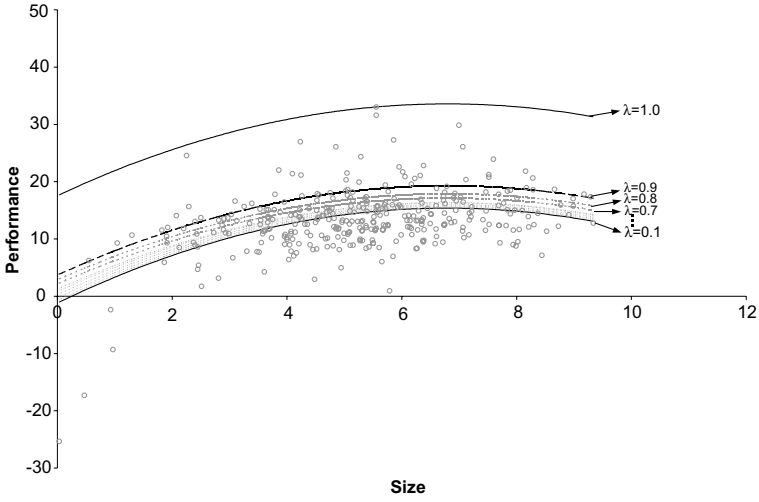

Fig. 2. Training set *B*.

That is, potential solutions can be generated for each $\lambda \in (0,1)$. However, these various solutions have better or worse predictive capabilities as measured by the criterion (14).

The optimal fund size corresponds to either $d^* = 6.9489$ (equivalent to \$1.042 billion), or $d^* = 6.7979$ (equivalent to \$896 million). For these two optimal sizes, a range of λ values worked equally well. These results are also consistent with the results found in the regression approach of Indro et al. (1999).

This work has practical significance for investors wishing to select mutual funds. As is well known, best performers in the previous year attract greater investments in the subsequent year. While attention to good performers is quite natural, investors may also wish to look at the standing of the fund in relation to optimal fund size.

Although a fund may have done well in a previous year, it may have been operating near or over optimal size at that time. If it attracts a great influx of new investments, as will be expected, it may move far away from optimal size so that its performance could suffer in subsequent years. A fund that did moderately well, and which is likely to remain at, or move into optimal fund size range, may be a better choice.

6. Conclusions

This paper discusses frontier estimation. We first demonstrate that maximum likelihood does not yield a plausible solution to models of this kind. However, since the practical modeling value of the composed error approach is very compelling for performance data, we seek a suitable criterion for such estimates.

Maximum likelihood estimation requires restrictive distributional assumptions and the functional form of the model needs to be specified. Here we propose a distribution-free, data-driven approach where the functional form is the result of our estimation procedure. As an example, we use mutual fund data to provide a step-by-step procedure to implement our model. Our procedure involves training and test samples, and cross validation is done by switching samples.

The key parameter in our model, λ, specifies the proportional split of the total error term into white-noise error and inefficiencies. For the mutual fund data in question, our model shows that for various levels of λ, parameters can be reasonably estimated, thus overcoming the major deficiencies of the maximum likelihood approach. Furthermore, the best model as suggested by our procedure shows the optimal fund size may be smaller than that previously obtained by ordinary least squares regression.

This study shows how a distribution-free approach can be used for frontier estimation, and the application is illustrated with mutual fund data. To that end, our sample data set has validated our procedure. But, as is true with illustrations with a single data set, results from this study may not be generalized to other problem settings. Thus, one area of future research is to focus on the generalizability of this technique to other problem domains.

References

Aigner, D.J., Chu, S.F., 1968. On estimating the industry production function. American Economic Review 58, 826–839.

Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6, 21–37.

Coelli, T., Prasada Rao, D.S., Battese, G.E., 1998. An Introduction to Efficiency and Productivity Analysis. Kluwer Academic Publishers, Boston, US.

Green, W.H., 1990. A gamma-distributed stochastic frontier model. Journal of Econometrics 46, 141–163.

Indro, D.C., Jiang, C.X., Hu, M.Y., Lee, W.Y., 1999. Mutual fund performance: Does fund size matter? Financial Analysts Journal (May/June), 74–87.

Meeusen, W., Van Den Broeck, J., 1977. Efficiency estimation from Cobb–Douglas production functions with composed error. International Economic Review 8, 435–444.

Ritter, C., Léopold, S., 1997. Pitfalls of normal-gamma stochastic frontier models. Journal of Productivity Analysis 8, 167–182.

Stevenson, R.E., 1980. Likelihood functions for generalized stochastic frontier estimation. Journal of Econometrics 13, 57–66.

Van Den Broeck, J., Koop, G., Osiewalski, J., Steel, M.F.J., 1994. Stochastic frontier models: A Bayesian perspective. Journal of Econometrics 61, 273–303.