In Managing Data Mining Technologies in Organizations, 2002

Chapter

Frontier Versus Ordinary Regression Models For Data Mining

Marvin D. Troutt, **Michael Hu, Murali Shanker** and **William Acar**
Graduate School of Management
Kent State University
Kent, OH 44242
(mtroutt@bsa3.kent.edu)
(mhu@bsa3.kent.edu)
(mshanker@bsa3.kent.edu)
(wacar@kent.edu)

October 2001

Frontier regression models seek to explain boundary, frontier or optimal behavior rather than average behavior as in ordinary regression models. Ordinary regression is one of the most important tools for data mining. Frontier models may be desirable alternatives in many circumstances. In this chapter, we discuss frontier regression models and compare their interpretations to ordinary regression models. Occasional contact with stochastic frontier estimation models is also made, but we concentrate primarily on pure ceiling or floor frontier models. We also propose some guidelines for when to choose between them.

1 Introduction

Frontier or benchmark estimation models were first discussed by Aigner and Chu (1968), who fitted a Cobb-Douglas industry production

function to data on production levels and factors. Here the Cobb-Douglas model was proposed as the best possible, frontier or benchmark model for the data. Observed production levels were modeled by subtracting nonnegative errors or inefficiency shortfalls from the frontier. More generally, such models may be called frontier regression models. They seek to explain boundary, frontier or optimal behavior rather than average behavior as in ordinary regression models. Such a model may also be called a ceiling model as it lies above all the observations. (The opposite case is similarly called a floor model). Ordinary regression is one of the most important tools for data mining. Frontier models may be desirable alternatives in some circumstances. In this chapter, we discuss frontier regression models and compare them to ordinary regression models. We also propose guidelines for when to choose between them.

There are a related class of models called Stochastic Frontier Estimation (SFE) models. These are modifications of the pure frontier method first considered separately by Meeusen and van den Broeck (1977) and Aigner, Lovell and Schmidt (1977). Here actual performance is modeled as the frontier model plus an error term composed of two parts. The first error part is normally distributed with mean zero. It is usually justified as accounting for uncertainty in the frontier model. The second error part is a nonnegative one, representing a measure of inefficiency error or deviation from the efficient frontier as in the pure frontier model. This term is also called the *inefficiency effect* in Coelli et al. (1998). The Aigner, Lovell and Schmidt (1977) method assumes that such nonnegative inefficiencies are distributed as half-normal. This permits the distribution of the total error to be specified and its parameters to be estimated by the maximum likelihood method. Stevenson (1980) extended that method to permit assumption of truncated normal and gamma distributions. However, Ritter and Léopold (1997) have found that such models are difficult to accurately estimate. Recently, Troutt et al. (2001) have pointed out theoretical problems in maximizing the likelihood function for such models. In that research, it was found that the likelihood function is U-shaped. One end corresponds to assuming a pure frontier model and the other corresponds to a pure ordinary least squares regression model. Thus, that research suggests that the maximum likelihood principle will choose one of those end point cases and not a mixed or SFE type model. We therefore concentrate on pure frontier models in this chapter.

2 Pure Frontier Models

Consider the general composed error stochastic frontier estimation model given by

$$y_j = f(\mathbf{x}_j , \boldsymbol{\theta}) + \varepsilon_j - \omega_j \qquad (1)$$

where for j = 1, …,n

y_j is the dependent variable

$\mathbf{x}_j$ is a vector of measurements on independent variables in $\Re^m$

$\boldsymbol{\theta}$ is a vector of model parameters in $\Re^p$

$f(\mathbf{x}_j , \boldsymbol{\theta})$ 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 the model specification becomes

$$y_j = f(\mathbf{x}_j , \boldsymbol{\theta}) + \varepsilon_j + \omega_j \qquad (2)$$

ε_j is a white noise error term with variance σ^2

ω_j is a nonnegative inefficiency error for observation j, independent of the ε_j. Thus, a pure frontier (ceiling) model would be given by

$$y_j = f(\mathbf{x}_j , \boldsymbol{\theta}) - \omega_j \qquad (3)$$

and similarly, a pure frontier (floor) model would appear as

$$y_j = f(\mathbf{x}_j , \boldsymbol{\theta}) + \omega_j \qquad (4)$$

The ceiling model would be most appropriate if the behavior of the observations is such that 'more is always better'. That is, the observations represent attempts to maximize the dependent variable. Similarly, the floor model would be appropriate in the opposite case. From this point of view, the OLS model can sometimes be regarded as a 'middle is better' situation. Fig. 1 depicts a scattergram of data pairs with the Ordinary Least Squares (OLS) regression model and a ceiling type frontier model.

Figure 1 About Here

It can be noted that the ceiling model is not necessarily the same as that obtained by raising the OLS model upwards until it just envelopes the data points from above. This is, of course, a heuristic approach to estimating a ceiling model in two dimensions. A more dramatic difference can be seen for regression or frontier model that are specified to have zero-intercepts (so-called regression through the origin). Fig. 2 depicts this contrast between and OLS and a floor frontier model for the same scattergram.

Figure 2 About Here

Figures 1 and 2 also suggest a motivation for the SFE models. Namely, the position and slope of the ceiling model depends heavily on just a few of the upper most data pairs. On the possibility that those data pairs are unrepresentative outliers, then one has less confidence that the correct frontier model has been estimated. From the point of view of optimum seeking or purposeful behavior, such data points would represent unusually good performance in the nature of a lucky event. For the purposes of this research, we assume that any such data have been removed or adjusted appropriately. One might consider SFE models as motivated by a desire to smooth out the upper or lower boundaries with white noise adjustments as a more mechanical approach to this issue, however.

Various approaches to estimating such pure frontier models have been proposed. For example, if the sum of squares of the ω_i is minimized as the model fitting criterion, then that procedure is a

maximum likelihood estimation (MLE) procedure when the ω_j are distributed as half-normal. Similarly, minimizing the sum of the ω_j is MLE when they are distributed with the exponential probability density function. In this work, we propose the latter criterion. Namely, we assume in the rest of this chapter that models of the type (3) and (4) will be fitted or estimated by the criterion of minimum $\Sigma_j\omega_j$, but we use a different, more general rationale than the above kinds of distribution assumptions to be explained below.

There are three reasons for this choice of criterion. First is the rationale of purposeful behavior. If each observation is interpreted as an attempt to reach the target or goal given by the frontier model, then the ω_j may be regarded as distances from the target. As each attempt seeks to minimize such a distance, then over all instances or observations the sum of these would be minimized. That is, the criterion of minimum $\Sigma_j\omega_j$ is taken as modeling purposeful behavior over repetitions of a single unit, or over a set of units. This has been formalized in Troutt et al. (2000) as the Maximum Performance Efficiency (MPE) estimation principle.

The second reason is that computation of a model solution for this criterion is flexible and straightforward. Typically, the model is a simple linear programming model easily solved with a spreadsheet solver. The model used in Troutt, Gribbin, Shanker and Zhang (2000) is an example.

The third reason is that a model aptness test is available for this criterion. Called the normal-like-or-better (NLOB) criterion, it consists of examining the fitted ω_j values. If these have a density sufficiently concentrated on the mode at zero, then the performance can be said to be as good or better than a bivariate normal model for a target such as a bull's-eye in throwing darts. Note that sometimes throwing darts is used as a metaphor for completely random guessing. Here we use it differently, however. If a data scattergram for dart hits is modeled well by a univariate density as steep or steeper than the normal, and with mode coincident to the target, then we regard this as very good performance, allowing for natural efficiency variation. The NLOB criterion can, in principle, be used with any distributional form for the fitted ω_j values. However, it appears most naturally suited to the case

when these are gamma distributed. Complete details on applying the NLOB criterion may be found in Troutt, Gribbin, Shanker and Zhang (2000).

3 Contrasts of Meaning and Purpose

Let us write $f^{OLS}(x)$ as a model for the data based on the OLS criterion, and similarly $f^*(x)$ as a, say, ceiling frontier model. Then we have two representation of the data values, y_j . Namely,

$$y_j = f^{OLS}(x_j) + \varepsilon_j = f^*(x_j) - \omega_j \qquad (5)$$

We note first that the two model values can and will likely be different. Which is valid or perhaps, most valid? We can apply the test of normality to the OLS residuals and the NLOB criterion to the frontier residuals as an obvious first step that may support a choice. But what if both tests are acceptable? Then we suggest relying on context. Namely, we ask whether the value y_j can be regarded as in the nature of an attempt at getting high values. If so, then the ceiling model would appear more appropriate. By contrast, if these data values are thought to be merely random deviations from a mean response then the OLS model should be preferred. A familiar example is that of a set of test scores for a student examination. Either case, or even mixtures of these cases might apply. If the exam were in a required course, for which an average grade is most desirable for the majority of the students, then the OLS model is compelling. On the other hand, if the test is a college entrance or professional qualification exam, then a higher grade is likely to be better for most students. In that case, the frontier model would likely be best.

Such difficult to call cases might be especially expected to occur with large sample sizes. As an illustration, we consider an example in Madansky (1988). There a large data set of 100 observations was simulated according to a gamma distribution anchored at zero. Then various well-known tests of normality were applied to see whether they could correctly reject the normal distribution hypothesis. Surprisingly, several of the tests did not reject the normal hypothesis. The gamma distribution chosen appeared to be well modeled by a normal density according to several tests.

This example leads to several observations in connection with the representation (5). First let us denote by *a* the average of the frontier residuals, ω_j . Next we write $\varepsilon_j^* = \omega_j - a$. Then $\omega_j = \varepsilon_j^* + a$. Substitution into (5) yields

$$y_j = f^{OLS}(x_j) + \varepsilon_j = f^*(x_j) - (\varepsilon_j^* + a) = (f^*(x_j) - a) - \varepsilon_j^*$$

or

$$f^{OLS}(x_j) + a + \varepsilon_j = f^*(x_j) - \varepsilon_j^*$$

One is tempted to suggest in such a case that a suitable frontier model may then be obtained from the OLS model by translating it upwards by the amount *a*. However, this would require the two types of epsilon residuals to be equal and opposite in sign. That this is not generally true is most easily seen from Fig. 2. Namely, if the OLS model were translated upward it would apparently intersect the frontier model. Thus even if the frontier residuals are normally distributed, the frontier model may differ from the OLS model adjusted for the mean frontier residuals. In this case we may even assume that the frontier model for model was fitted by an OLS criterion with the constraint of passing through the origin. As the fitting criterion is changed and more complex constraints are present, it is reasonable to expect differences between the forms of models obtained – despite the possibility of normally distributed residuals for both models.

Moreover, the foregoing points suggest the possibility of obtaining somewhat misleading regression analyses on comparative performance data analysis. Let us consider the context of explaining the performances of firms according to a single independent variable, x. We might think of x as some measure such as size in dollar valuation of assets, say, along with a dependent variable, y, as some acceptable measure of performance. For simplicity, we assume that the performance variable is not directly proportional to size. The analyst may well obtain an acceptable OLS model for such data. Perhaps the fitted model suggests that $y = a_0 + b_0x$ explains the firm performance data very well. However, as an OLS regression model, this result can be

interpreted in an average or typical sense. Namely, on average, firms of size x in this context will be expected to have observed performances given by that fitted model. But what exactly is the role of x here? Does the level of x reflect a higher potential performance for firms, or does the level of x pertain to performance ability? To better see this contrast, suppose that a frontier model is fitted to the same data and yields the different model $y = a_1 + b_1x$. In this frontier model, x affects the estimated upper limit of performance rather than the performance with respect to that goal. One may also compute an average performance based on this model from its residuals. It may happen that performance, as measured by the ω_j values, does not really vary with the value of x. Alternatively; it may vary in some other fashion. It may be proposed that when a goal such as highest possible performance is present, then both the level of that goal and performance with respect to it can be affected by independent variables. An OLS or average oriented model may therefore be confounding two phenomena. Variation of performance can be expected in almost any goal directed behavior. Such performance fluctuation may be a regarded as an always present effect or variable, which itself may be affected by variables proposed as influential in the OLS model.

4 Data Mining Uses and Suggested Guidelines

The ceiling frontier model seeks to explain best performance as a function of one or more independent variables. Such models may be especially attractive for many data mining (DM) applications. For example, interest often centers on best instances such as customers most responsive to mailings, or safest drivers, etc. For mailings of a given type, it would be desirable to predict a ranking of most responsive customers so that efforts can be best directed. Namely, if only, say, 1000 are to be mailed, then those predicted as the top 1000 would be attractive for consideration. Similarly, an insurer may be interested in characterizing its best and worst customers according to a model.

Here we briefly propose several potential DM and related applications:

Supplier Ranking - In Supply Chain Management, firms must often consider and choose among potential suppliers. While cost is an important variable, many others may need to be considered. These

include, for example: lead time performance, quality measures, capacity and flexibility measures, to name only a few. Generally, it will be desirable to select and characterize the best or highest performing suppliers among these.

Technology Choice – For the choice of industrial robots, instrumentation and similar technology, it is often possible to test and collect data on several possible choices. Obviously, the firm will be interested in the attributes of the best performers for the selection decision.

Total Quality Management - Every year in the total quality management area, the Malcolm Baldrige National Quality Award is given to a small group of firms. The guidelines for winning this prestigious award may not be clearly spelled out. One would be interested in the dimensions along which the award winners may differ from the 'average' firm.

Marketing - In marketing, the 20/80 rule, sometimes called Pareto's Law or Principle, is appropriate for modeling the usage rate of the heavy users. Typically, the top twenty percent (heavy users) of the total number of consumers in the marketplace will account for roughly 80% of total revenue. Thus, it will be quite misleading for a firm to base its marketing strategy on the average purchase behavior of its consumers. A firm in most cases will develop a ceiling model for its 'heavy' user group. Likewise, a 'floor' model would be appropriate for the non-users (users that a company may have no hope of getting – or perhaps those they do not wish to have such as high risk drivers for car insurance).

Airline Productivity - The efficiency frontier models have been used to examine productivity in the airline industry. A Cobb-Douglas total cost function is used for the estimation of the efficiency frontier. The dependent variable is the total cost for an airline. The independent variables are (a) passenger output (number of passengers times distance traveled); (b) labor cost for an airline; (c) fuel cost for an airline; and (d) capital costs for an airline. Airlines that lie on the efficient frontier are presumably the most production efficient. The inefficiency measures of airlines can be constructed by taking the distance between any airline with that of the frontier.

Comparison of Stocks and Mutual Fund - Investors are not only interested in the average performance of the stock market. A ceiling model corresponds to the top performing stocks while a floor model will provide insights into the 'poor' performers, firms that may potentially declare bankruptcy. Similarly, investors would like to decide between mutual funds with deeper information than just a simple comparison of recent performances.

Employee Loyalty - Employee turnover refers to the loss of trained employees, especially when such losses are early and costly. Data Mining could be applied to Human Resources data marts with the help of a floor frontier regression models. Which available attributes best explain the earliest termination cases? Such information could be used to score potential new hires on likelihood to terminate early. In this context the opposite case of most loyal employees might similarly be modeled as a ceiling type model.

By considering the general features of the above examples, we may propose the following suggested guidelines for considering frontier models instead of, or in connection with regression data mining applications:

1. There is interest in characterizing and modeling the best and/or worst cases in the data.

2. Behaviors of both customers and the businesses that serve them are of the managed kind. In general, such 'managed data' or data from purposeful or goal directed behavior will be amenable to frontier modeling.

3. Some loss of inferential capability can be tolerated. (See limitations below).

4. High-lier data (for ceiling models) and low-lier data (for floor models) can easily be identified and/or adjusted.

5 Other Models and Applications

The general approach of frontier models might be carried over to other models and contexts. For example, logistic frontier regression might be aimed at modeling the most probable cases. Applications of this kind are attractive for predicting poor and/or excellent credit risks and for income tax filers most likely to be evading taxes.

Policy capturing studies have been around for a long time. Generally, this approach uses regression, classification, or other data models in order to explain and predict dichotomous, categorical or ordinal outcomes. For example, it may be useful for corporate legal planners to predict the likely outcome of legal actions based on a human resources case profile. The prediction of attribute levels or ranges for the best cases might be compared to those for the average case. A planned legal strategy may be judged average or better than average and compared to the appropriate model in that case to improve decision making about out-of-court settlements, for instance.

6 Limitations and Further Research

As noted earlier, a potential limitation of these models arises in connection with outliers. In the present setting, one may have two kinds, which might be called high-liers and low-liers, respectively. High-liers would be problematical for ceiling frontier models. Such observations suggest fortunate high performance unrelated to the predictor model. Similarly, low-liers would be of concern for floor models. As noted above, these concerns can be regarded as a motivation for SFE type models. Unfortunately, SFE models are difficult to estimate at the present time. Additional research on this class of models

would be helpful when low-liers or high-liers are not easy to identify or accommodate.

A generalization of frontier type models would be to what may be called percentile and stratification response type models. One may envision a modeling approach that uses a parameter, z, with range [0,1]. Such a model would seek to associate observed values with a z value so that the model with $z = z_0$ provides the best prediction for the z_0-percentile response variable. A closely related, percentile estimation type of model would seek to best explain the upper z_0 percent of the potential responses. That is, if one starts with a pure frontier model, it would be possible to estimate, say, the 50^{th}-percentile level of the dependent variable for a given value of independent variable values. However, if we know in advance that we wish to model specifically the upper 50^{th}-percentile of cases, then it is conceptually possible to obtain a different model than that based on the pure frontier one.

Of course, a great advantage of OLS regression models lies in the inferential capabilities of normal distribution based theory. The NLOB criterion provides some help in this direction for the frontier models. However, more statistical theory work along those lines would clearly be useful.

7 Conclusions

Frontier regression models seek to explain topmost or bottommost performers in the data. Many data mining applications can be so conceived. Several potential applications of this type were discussed. Such models are also natural when the data arise from purposeful, goal-directed or managed activities. A test of this characteristic, called the Normal-Like-Or-Better (NLOB) performance criterion has recently been developed. Using the fitting criterion called Maximum Performance Efficiency (MPE) estimation, the sum of efficiency residuals is minimized. This criterion often reduces to a linear programming model and is therefore straightforward to perform in spreadsheet models with solver capabilities.

References

Aigner, D.J. and Chu, S.F. (1968). On Estimating the Industry Production Function. *American Economic Review* **58**, 826-839.

Aigner,D.J., Lovell, C.A.K., and Schmidt, P. (1977). Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics* **6**, 21-37.

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

Madansky, A. (1988). *Prescriptions for working statisticians.* Springer-Verlag, New York.

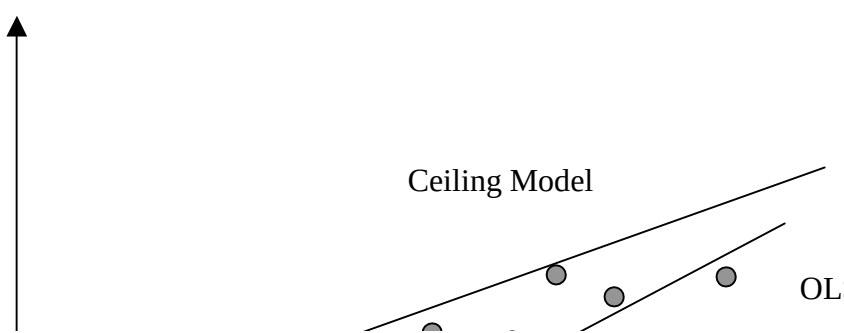
Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* **8**, 435-444.

Ritter, C. and 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.

Troutt, M.D., Gribbin, D.W., Shanker, M. and Zhang, A. (2000). Cost efficiency benchmarking for operational units with multiple cost drivers. *Decision Sciences* **31**(4), Fall, 813-832.

Troutt, M. D., Hu, M. and Shanker, M. (2001). Unbounded Likelihood in Stochastic Frontier Estimation: Signal-To-Noise Ratio-Based Alternatives. *Working paper*, Department of Management & Information Systems, Kent State University, Kent, Ohio.



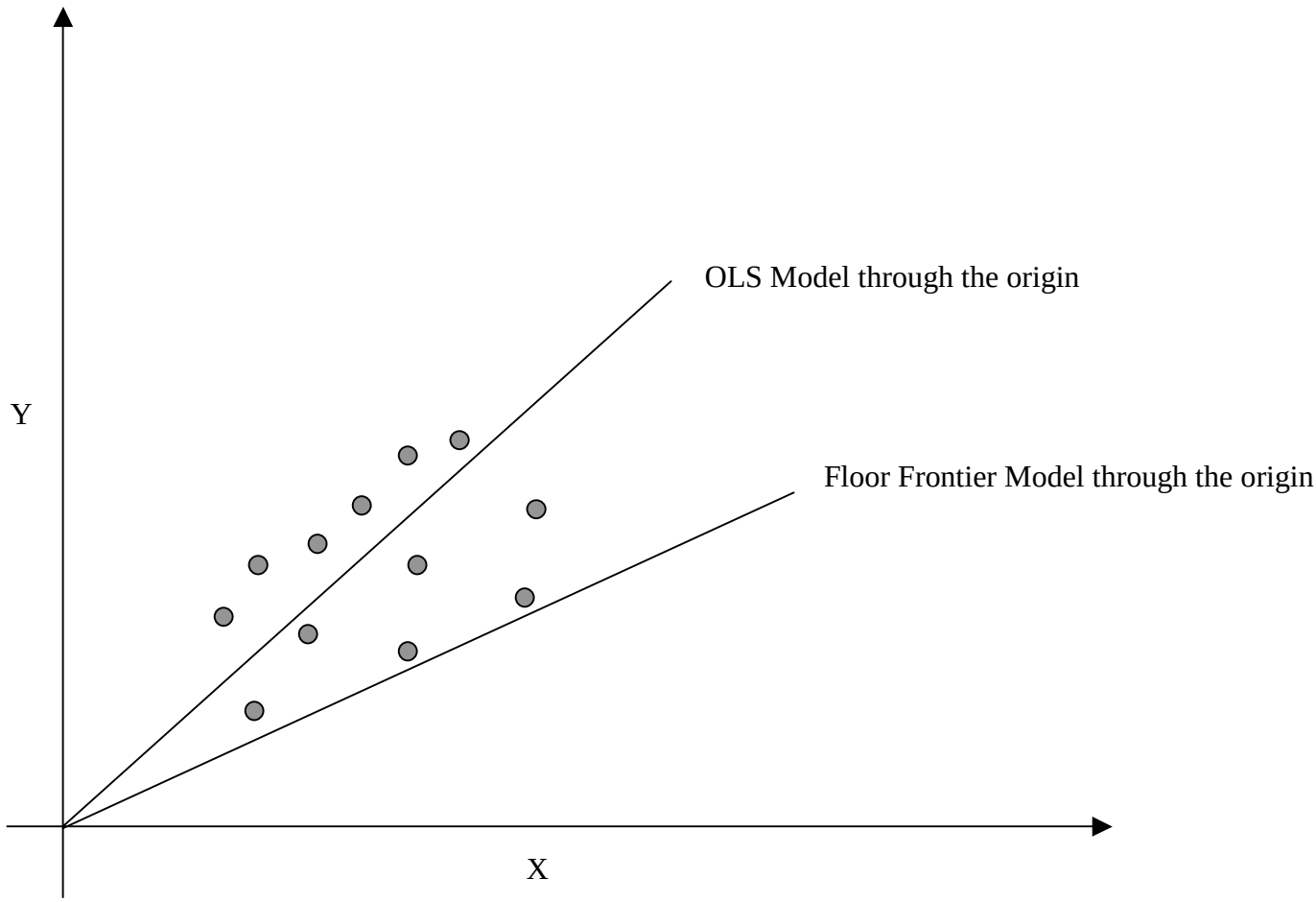


Figure 2: Models forced to pass through the origin