B07

Effect of data aggregation on inefficiency estimates

T. Kim | Seoul National University, Graduate School of International Agricultural Technology | Pyeongchang, Ganwon-do | Korea

B. Wade Brorsen | Oklahoma State University, Department of Agricultural Economics | Stillwater, Oklahoma | Usa

P. Kenkel | Oklahoma State University, Department of Agricultural Economics | Stillwater, Oklahoma | Usa

DOI: 10.1481/icasVII.2016.b07c

ABSTRACT

The objective of this research is to determine the effects of data aggregation on firm-level efficiency measures. Aggregation creates heteroscedasticity since the error variance of average cost decreases as total output increases. Inefficiency indexes from stochastic frontier functions estimated assuming both heteroscedasticity and homoscedasticity are compared with those from data envelopment analysis (DEA) using a Monte Carlo study. Stochastic frontier functions hold up rather well in the presence of data aggregation, but efficiency measurement from DEA diverges from true efficiency measurement. In particular, DEA is biased towards finding more inefficiency in small firms.

Keywords: aggregation, cost function, data envelopment analysis, stochastic frontier.

PAPER

1. Introduction

Since Farrell (1957) developed his efficiency index using a deterministic frontier function, efficiency measurements from a stochastic frontier model and data envelopment analysis (DEA) have been frequently used to estimate firm-level efficiency measures. Stochastic frontier functions were suggested by Aigner et al. (1977), who allowed deviations from the frontier to arise from random factors where the disturbance term was the sum of symmetric normal and half-normal random variables. Jondrow et al. (1982) introduced firm-specific inefficiency measurement. DEA was introduced by Charnes et al. (1978), who assumed constant returns to scale (CRS). This idea was extended to variable returns to scale (VRS) by Banker et al. (1984). DEA is a nonparametric method and therefore does not require distributional assumptions about error terms. Also, DEA can handle multiple outputs and inputs. On the other hand, the fundamental merit of using a stochastic frontier function is to measure inefficiency in the presence of statistical noise, but it is subject to potential bias if an incorrect error structure is assumed.

One concern about the error structure is possible heteroscedasticity. Caudill and Ford (1993) find biases in frontier estimation due to heteroscedasticity of the one-sided error and later Caudill et al. (1995) find that the rankings of firms by efficiency measures are significantly affected by correcting for heteroscedasticity. These followed Schmidt's suggestion (1986) that a one-sided error can be associated with factors controlled by the firm while the random component can be associated with factors outside the firm's control. Hadri (1999) finds heteroscedasticity of both error terms with the same data of Caudill et al. (1995). This past research, however, has not formally derived how aggregation leads to heteroscedasticity.

Greene (2003) argues that the most common occurrence of heteroscedasticity is, in general, when data are aggregated, which is called "groupwise heteroscedasticity". Dickens (1990) shows that using data weighted by the square root of group size is only appropriate if individual error terms are not correlated within groups. In empirical work, disaggregated data are often not available so that economic research is often done using aggregated data. For example, the Macdonald and Michael (2000) study of the hog slaughter industry and the Ollinger et al. (2005) study of the poultry processing industry are aggregated over packing plants owned by the same firm. Adkins and Moomaw (2003) study public schools aggregated across teachers, while the Featherstone et al.

(1997) study of beef cow farms uses data aggregated over cows. These studies typically conclude that small firms are much more inefficient than large firms. When the average cost (or average utput) function with aggregated data is estimated, the error variance of average cost decreases as group size increases, which raises the question as to how much of this finding of small-firm inefficiency is due to not considering heteroscedasticty created from using aggregate data.

This article uses a Monte Carlo study to estimate the biases in inefficiency measurement that are created from aggregate data. We begin with a disaggregate model with random effects, and the aggregate model is obtained by summing the disaggregate observations. The log cost model is obtained by a first order Taylor approximation. The resulting model is heteroscedastic. A Monte Carlo study compares inefficiency measurement in the presence of this heteroscedasticity with parameter estimates that both consider and disregard heteroscedasticity. With the same data, inefficiency indexes from DEA are also computed in order to provide comparisons with the stochastic frontier function.

2. Theory

Consider the following disaggregated cost function with random effects:

$$(1) \qquad C_{ij} = \mathbf{x}'_{ij}\beta + u_j + w_{ij}, \quad i = 1, \dots, n_j, \; j = 1, \dots, J,$$

$$u_j \sim iid\, N(0, \sigma_u^2), \; w_{ij} \sim iid\, N(0, \sigma_w^2), \; \mathrm{cov}(u_j, w_{ij}) = 0,$$

where C_{ij} is the cost of the ith unit in the jth firm, $\mathbf{x}_{ij}$ is a vector of explanatory variables including input prices, $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated, u_j is the random effect of the jth firm, and w_{ij} is the unexplained portion of the cost of the ith unit in the jth firm.

Examples of such a disaggregated cost function include a firm with multiple packing plants, a farmer with many fields, and a school with many teachers. The unit refers to each packing plant, each field, or each teacher. Note that the units within a firm all have the same random effect. The heteroscedasticity comes from the effects of w_{ij} being diversified away in larger firms.

In a stochastic frontier cost function, the inefficiency is represented with a one-sided error term (Aigner et al. 1977). Thus, a stochastic frontier cost function can be defined as

$$(2) \qquad C_{ij} = \mathbf{x}'_{ij}\beta + u_j + w_{ij} + v_j, \; v_j \sim iid\, \left|N(0, \sigma_v^2)\right|, \mathrm{cov}(u_j, v_j) = 0,$$

$$\mathrm{cov}(w_{ij}, v_j) = 0,$$

where v_j is the inefficiency and a one-sided error with $\mathrm{E}(v) = \sigma_v \sqrt{2/\pi}$ and $\mathrm{Var}(v) = \sigma_v^2(1 - 2/\pi)$. Especially, $\sigma_v \sqrt{2/\pi}$ is known as an average inefficiency measurement by Aigner et al. (1977). The term ($\mathbf{x}'_{ij}\beta$) can be interpreted as the minimum expected cost. Note that each unit within a firm is assumed to have the same inefficiency, which is consistent with Schmidt's (1986) view that the inefficiency error represents factors under control of the firm.

The (total) stochastic frontier cost function is the sum over all units of the firm:

$$(3)\qquad \sum_{i=1}^{n_j} C_{ij} = \sum_{i=1}^{n_j} \mathbf{x}'_{ij}\boldsymbol{\beta} + n_j u_j + \sum_{i=1}^{n_j} w_{ij} + n_j v_j,$$

where n_j is the number of units produced by the jth firm.

A dot subscript is the common notation to denote that the variable has been averaged over the corresponding index. A (total) stochastic frontier cost function (3) using the dot notation is

$$(4)\qquad TC_j = n_j \mathbf{x}'_{\bullet j}\beta + n_j(u_j + w_{\bullet j} + v_j),\ \ j = 1,\ldots,J,\ \ w_{\bullet j} \sim N(0, \frac{\sigma_w^2}{n_j}),$$

where TC_j is the total cost for the jth firm, and the dot subscript indicates that the variable has been averaged over units. $\mathbf{x}_{\bullet j}$ is the averaged vector of explanatory variables over units, and $w_{\bullet j}$ is the averaged unexplained error over units.

Here, heteroscedasticity related with units is shown, which is typically called groupwise heteroscedasticity (Greene, 2003). Dickens (1990) showed similar heteroscedasticity in the presence of firm specific error, which is similar to the random effect shown here.

Logarithmic cost functions (translog or double log) are typically used in empirical work (Melton and Huffman, 1995) due to several conveniences such as including multiple outputs, calculating elasticities easily, and adjusting for heteroscedasticity. Taking the natural log of equation (4) gives

$$(5)\qquad \ln TC_j = \ln n_j + \ln\left(\mathbf{x}'_{\bullet j}\beta + u_j + w_{\bullet j} + v_j\right),$$

which is the double log cost function. Then, since error terms are the only random variables, applying a first-order Taylor approximation of $\ln\left(\mathbf{x}'_{\bullet j}\beta + u_j + w_{\bullet j} + v_j\right)$ around the mean of the random and unexplained error, and the frontier of inefficiency error such as $u_j = 0$, $w_{\bullet j} = 0$ and $v_{\bullet j} = 0$ gives the following model[1]:

$$(6) \qquad \ln TC_j \approx \ln n_j + \ln \mathbf{x}'_{\bullet j}\beta + \frac{1}{\mathbf{x}'_{\bullet j}\beta}\left(u_j + w_{\bullet j} + v_j\right).$$

The variance of all error terms is $\left(1/\mathbf{x}'_{\bullet j}\beta\right)^2\left(\sigma_u^2 + \sigma_w^2/n_j + \sigma_v^2\right)$, which shows a combination of dependent variable heteroscedasticity and groupwise heteroscedasticity.

The usual stochastic frontier model is a special case of (3) where $\sigma_u^2 = 0$ and $n_j = 1, \forall j$. To define the usual model, let $e = w + v$ where w is unexplained error and v is the random inefficiency term. The density function for this case developed by Weinstein (1964) is

$$(7) \qquad f(e) = \frac{2}{\sigma} f^*\left(\frac{e}{\sigma}\right) F^*\left(\frac{\lambda e}{\sigma}\right), -\infty < e < +\infty,$$

where $\sigma^2 = \sigma_w^2 + \sigma_v^2$, $\lambda = \sigma_v/\sigma_w$, and f^* and F^* are the standard normal probability density function and the standard normal cumulative density function, respectively. Note that the density function of a stochastic frontier function is a special case of skew normal distributions (Dominguez-Molina et al. 2003, Genton 2004; Gonzalez-Farias et al. 2004).

Here, λ is an indicator of the relative variability of error terms. As Aigner et al. (1977) argues, $\lambda \to 0$ means $\sigma_v \to 0$ and/or $\sigma_w \to \infty$, or that inefficiency error is dominated by random error. With data aggregation, λ increases as firms become larger since the variance of unexplained error does not increase as rapidly with firm size.

[1] A second order Taylor approximation gives the following model: $\ln TC_j \approx \ln n_j + \ln \mathrm{x}'_{\bullet j}\beta + \frac{1}{\mathrm{x}'_{\bullet j}\beta}\left(u_j + w_{\bullet j} + v_j\right) - \frac{1}{2}\left(\frac{1}{\mathrm{x}'_{\bullet j}\beta}\left(u_j + w_{\bullet j} + v_j\right)\right)^2$. This model is not considered here since the primary concern is to investigate heteroscedasticity easily in the stochastic frontier function. Also, first order Taylor approximation for an average cost function can be expressed as $\ln AC_j \approx \ln \mathrm{x}'_{\bullet j}\beta + \frac{1}{\mathrm{x}'_{\bullet j}\beta}\left(u_j + w_{\bullet j} + v_j\right)$, which has the same error structure.

Two measurements for the firm-specific inefficiency are given by Jondrow et al. (1982). Both are based on the conditional distribution of inefficiency error (v) given overall error (e). The first measure is given by

$$\text{(8)} \qquad \mathrm{E}(v|e) = \sigma_* \left[\left(\frac{\lambda e}{\sigma}\right) + f^*\left(\frac{\lambda e}{\sigma}\right) \Big/ F^*\left(\frac{\lambda e}{\sigma}\right)\right],$$

where $\sigma_*^2 = \left(\sigma_v \sigma_w / \sigma\right)^2$. The other variables are the same as in equation (7).

The second measure, which is based on the conditional mode, is given by

$$\text{(9)} \qquad \begin{aligned} \mathrm{M}(v|e) &= e\left(\sigma_v^2/\sigma^2\right) && \text{if } e \geq 0 \\ &= 0 && \text{if } e < 0. \end{aligned}$$

The log-likelihood function is derived from the density function in equation (7) by substituting the appropriate variances for the aggregated model:

$$\text{(10)} \qquad \sum_{j=1}^{J} \ln(f_j(e_j)) = \sum_{j=1}^{J} \ln \frac{2}{\sigma_j} f^*\left(\frac{e_j}{\sigma_j}\right) F^*\left(\frac{\lambda_j e_j}{\sigma_j}\right),$$

where $e_j = \frac{1}{\mathrm{x}'_{\bullet j}\beta}\left(u_j + w_{\bullet j} + v_j\right)$, $\sigma_j = \left(\frac{1}{\mathrm{x}'_{\bullet j}\beta}\right)\sqrt{\sigma_u^2 + \frac{\sigma_w^2}{n_j} + \sigma_v^2(1-2/\pi)}$, and $\lambda_j = \sqrt{\frac{\sigma_v^2(1-2/\pi)}{\sigma_u^2 + \sigma_w^2/n_j}}$.

Maximizing (10) gives maximum likelihood estimates for the stochastic frontier cost function with heteroscedastiecty due to aggregation.

In the presence of data aggregation, two firm-specific inefficiency measurements like equation (8) and (9) are slightly modified. The first measurement can be expressed as

$$\text{(11)} \qquad \mathrm{E}(v_{*j}|e_j) = \sigma_{*j}\left[\left(\frac{\lambda_j e_j}{\sigma_j}\right) + f^*\left(\frac{\lambda_j e_j}{\sigma_j}\right) \Big/ F^*\left(\frac{\lambda_j e_j}{\sigma_j}\right)\right],$$

where $v_{*j} = v_j / \mathrm{x}'_{\bullet j}\beta$, $\sigma_{*j} = \left(\frac{1}{\mathrm{x}'_{\bullet j}\beta}\right)\sqrt{\frac{\sigma_v^2(1-2/\pi)\cdot\sigma_w^2/n_j}{\sigma_u^2 + \sigma_w^2/n_j + \sigma_v^2(1-2/\pi)}}$. The other variables are the same as in equation (10).

The second measure which is based on the conditional mode is given by

$$(12)\qquad \begin{aligned} \mathrm{M}(v_{*j}\,|\,e_j) &= e_j\left(\sigma_{*v}^2/\sigma_j^2\right) && \text{if } e_j \geq 0 \\ &= 0 && \text{if } e_j < 0\,, \end{aligned}$$

where $\sigma_{*v}^2 = \left(1/\mathrm{x}'_{\bullet j}\beta\right)^2 \sigma_v^2 (1-2/\pi)$. The other variables are the same as in equation (10).

3. Data Envelopment Analysis (DEA)

Since input-oriented efficiency indexes with CRS and VRS were proposed by Charnes et al. (1978) and Banker et al. (1984), respectively, these two techniques have both been widely used (Fare et al. 1994) and therefore efficiency measurement with both CRS and VRS is discussed here.

Assuming M different outputs, N different inputs, and J different firms, the input-oriented model with VRS is

$$(13)\qquad \begin{aligned} &F_{\mathrm{j}}(\mathbf{x}_{\mathrm{j}},\mathbf{y}_{\mathrm{j}}) = \min_{\theta_j,\boldsymbol{\lambda}}\ \theta_j \\ &\qquad\qquad \text{s.t.}\ \ \mathbf{y}_{\mathrm{j}} \leq \mathbf{Y}\boldsymbol{\lambda},\ \ \mathbf{X}\boldsymbol{\lambda} \leq \theta_j\mathbf{x}_{\mathrm{j}},\ \ \mathbf{j}'\boldsymbol{\lambda}=1,\ \ \boldsymbol{\lambda}\geq 0,\ \ j=1,\ldots,J \end{aligned},$$

where $F_{\mathrm{j}}(\mathbf{x}_{\mathrm{j}},\mathbf{y}_{\mathrm{j}})$ is the Farrell efficiency estimate (or technical efficiency) given a $N\times 1$ input vector ($\mathbf{x}_{\mathrm{j}}$) and a $M\times 1$ output vector ($\mathbf{y}_{\mathrm{j}}$) for the j^{th} firm, $\mathbf{Y}$ is a $M\times J$ matrix for outputs, $\mathbf{X}$ is a $N\times J$ matrix for inputs, θ_j is a shrinking factor, $\boldsymbol{\lambda}$ is a $J\times 1$ vector of weights for firms, and $\mathbf{j}$ is a vector of ones.

The Farrell efficiency estimate is the reciprocal of the input distance function with the input-oriented model. Also, if there is no restriction of $\mathbf{j}'\boldsymbol{\lambda}=1$, then the model is the case of CRS.

The cost minimization model with VRS can be specified as

$$(14)\qquad \begin{aligned} &\min_{\boldsymbol{\lambda},\mathbf{x}_{\mathrm{j}}}\ \mathbf{r}_{\mathrm{j}}'\mathbf{x}_{\mathrm{j}}^{*} \\ &\text{s.t.}\ \ \mathbf{y}_{\mathrm{j}} \leq \mathbf{Y}\boldsymbol{\lambda}\,,\ \ \mathbf{X}\boldsymbol{\lambda} \leq \mathbf{x}_{\mathrm{j}}^{*},\ \ \mathbf{j}'\boldsymbol{\lambda}=1,\ \ \boldsymbol{\lambda}\geq 0\,,\ \ j=1,\ldots,J, \end{aligned}$$

where $\mathbf{r}_{\mathrm{j}}$ is a $N\times 1$ vector of input prices for the j^{th} firm; $\mathbf{x}_{\mathrm{j}}^{*}$ is the cost-minimizing $N\times 1$ vector of input quantities for the j^{th} firm, which is calculated by the linear programming given a vector of output quantities for the j^{th} firm ($\mathbf{y}_j$) and a vector of input prices for the j^{th} firm ($\mathbf{r}_j$); and the other variables are the same as above.

Then, efficiency measurements for the j^{th} firm can be defined as

$$(15) \qquad CE = \frac{\text{minimized cost}}{\text{actual cost}} = \frac{\mathbf{r}_j'\mathbf{x}_j^*}{\mathbf{r}_j'\mathbf{x}_j}, \quad AE = \frac{CE}{TE}, \quad TE = \theta_j,$$

where *CE* is the cost efficiency, *AE* is the allocative efficiency, and *TE* is the technical efficiency derived from the linear programming problem in equation (13).

4. Data and procedures

A Monte Carlo study is used to examine the effects of heteroscedasticity due to data aggregation. Based on equation (2), our true model is assumed as

$$(16) \quad C_{ij} = r_{ij} + u_j + w_{ij} + v_j,$$

where r_{ij} is the input price of the *i*th unit in the *j*th firm, while the other variables are as previously defined.

Aggregation over all units yields the following model:

$$(17) \quad \sum_{1}^{n_j} C_{ij} = TC_j = n_j r_{\bullet j} + n_j(u_j + w_{\bullet j} + v_j).$$

Taking the natural log and a first-order Taylor series around the mean of random errors (u_j and $w_{\bullet j}$) and the frontier (zero) of inefficiency error (v_j) gives

$$(18) \quad \ln TC_j \approx \ln(r_{\bullet j}) + \ln n_j + \frac{1}{r_{\bullet j}}(u_j + w_{\bullet j} + v_j).$$

So, our stochastic frontier cost function of equation (18) can be rewritten as

$$(19) \quad \ln TC_j = \beta_1' \ln r_{\bullet j} + \beta_2' \ln n_j + (u_j' + w_{\bullet j}' + v_j'),$$

where heteroscedasticity is incorporated into the variances by assuming

$\mathrm{var}(u'_j + w'_{\bullet j}) = \left(\sigma_u^2 + \frac{\sigma_w^2}{n_j}\right)\frac{1}{(r_{\bullet j})^2}$ and $\mathrm{var}(v'_j) = \left(\sigma_v^2(1-2/\pi)\right)\frac{1}{(r_{\bullet j})^2}$. Here, β'_1, β'_2, σ_u^2, σ_w^2, and σ_v^2

are unknown parameters to be estimated.

Input prices are generated as $r_{ij} \sim N(12,4)$. Also, numerous small firms and a few large firms are assumed by truncating a random number drawn from a distribution of 5*exp(N(0,1))+1; the mean units is 8.89 with variance around 112. To see the changes in relative variability of error terms, three scenarios of variances are considered: $[\sigma_u^2, \sigma_w^2, \sigma_v^2(1-2/\pi)] = [1, 4, 1.45]$, $[1, 4, 5.81]$, and $[1, 4, 13.08]$. The relative variability for these are, on average, $\lambda \approx 1$, $\lambda \approx 2$, and $\lambda \approx 3$, respectively. These scenarios show how much the inefficiency indexes change as the variability of inefficiency increases.

Using NLMIXED in SAS with 100 samples[2] of 100 observations, the stochastic frontier cost function with heteroscedasticity and without heteroscedasticity is estimated. Since one output and one input are assumed, cost inefficiency is the same as technical inefficiency from DEA. Inefficiency measurement of DEA using the data envelopment analysis program (DEAP) is also calculated and compared with those from the stochastic frontier cost function. Both constant return to scale (CRS) and variable return to scale (VRS) are used.

5. Results

Table 1 shows mean values of estimated parameters for the stochastic frontier cost function. In no case are the estimates significantly different from the true value. The variability of parameter estimates is slightly larger when homoscedasticity is incorrectly assumed. Certainly, the results

[2] The simulation takes considerable time and the differences between DEA and the stochastic efficiency measures are large enough that using one hundred samples is sufficient.

show that ignoring the effects of data aggregation does not create a serious problem for estimating a stochastic frontier cost function.

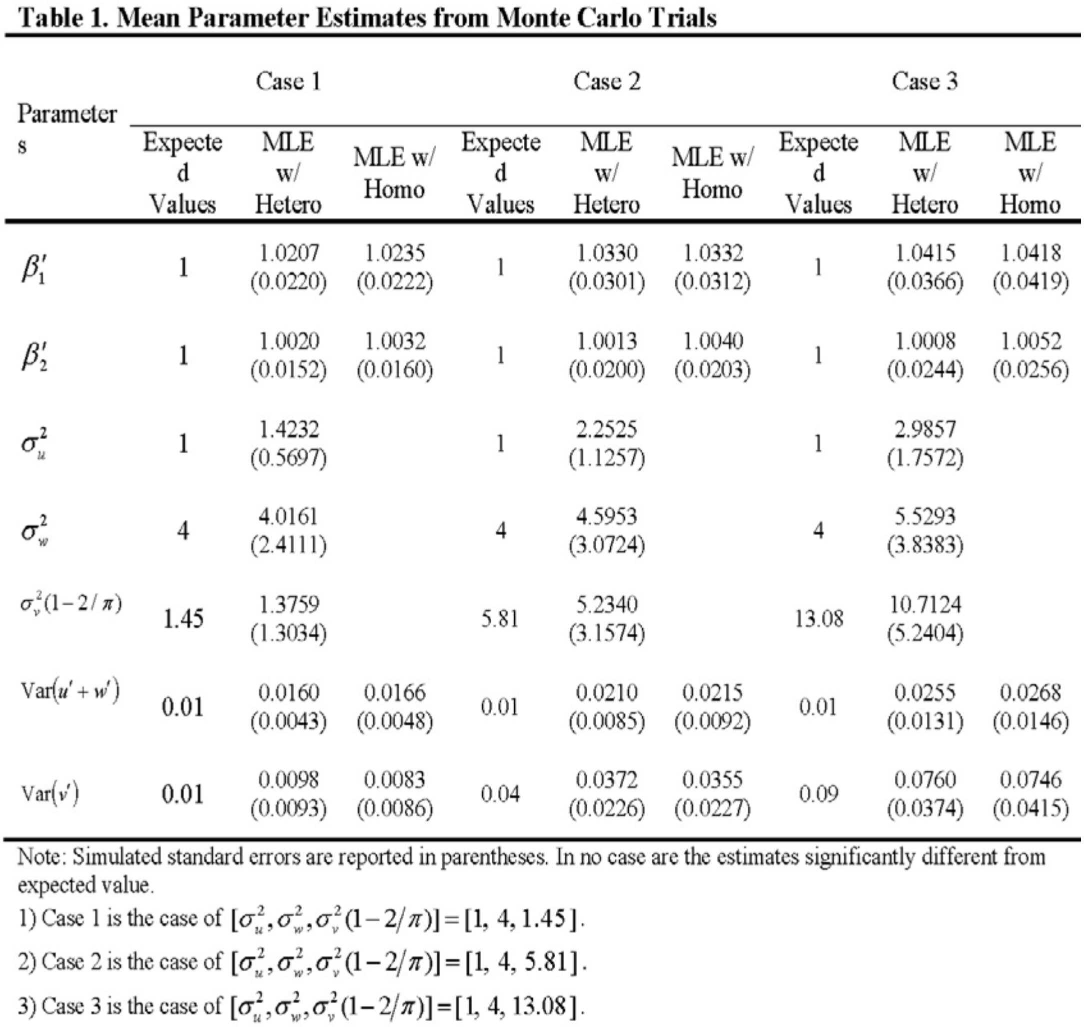


Table 1. Mean Parameter Estimates from Monte Carlo Trials

Parameters	Case 1			Case 2			Case 3		
	Expected Values	MLE w/ Hetero	MLE w/ Homo	Expected Values	MLE w/ Hetero	MLE w/ Homo	Expected Values	MLE w/ Hetero	MLE w/ Homo
β_1'	1	1.0207 (0.0220)	1.0235 (0.0222)	1	1.0330 (0.0301)	1.0332 (0.0312)	1	1.0415 (0.0366)	1.0418 (0.0419)
β_2'	1	1.0020 (0.0152)	1.0032 (0.0160)	1	1.0013 (0.0200)	1.0040 (0.0203)	1	1.0008 (0.0244)	1.0052 (0.0256)
σ_u^2	1	1.4232 (0.5697)		1	2.2525 (1.1257)		1	2.9857 (1.7572)	
σ_w^2	4	4.0161 (2.4111)		4	4.5953 (3.0724)		4	5.5293 (3.8383)	
$\sigma_v^2(1-2/\pi)$	1.45	1.3759 (1.3034)		5.81	5.2340 (3.1574)		13.08	10.7124 (5.2404)	
$\mathrm{Var}(u'+w')$	0.01	0.0160 (0.0043)	0.0166 (0.0048)	0.01	0.0210 (0.0085)	0.0215 (0.0092)	0.01	0.0255 (0.0131)	0.0268 (0.0146)
$\mathrm{Var}(v')$	0.01	0.0098 (0.0093)	0.0083 (0.0086)	0.04	0.0372 (0.0226)	0.0355 (0.0227)	0.09	0.0760 (0.0374)	0.0746 (0.0415)

Note: Simulated standard errors are reported in parentheses. In no case are the estimates significantly different from expected value.

1) Case 1 is the case of $[\sigma_u^2, \sigma_w^2, \sigma_v^2(1-2/\pi)] = [1,\ 4,\ 1.45]$.

2) Case 2 is the case of $[\sigma_u^2, \sigma_w^2, \sigma_v^2(1-2/\pi)] = [1,\ 4,\ 5.81]$.

3) Case 3 is the case of $[\sigma_u^2, \sigma_w^2, \sigma_v^2(1-2/\pi)] = [1,\ 4,\ 13.08]$.

Table 2 shows the mean of average inefficiency. Inefficiency indexes with heteroscedasticity are slightly bigger than those assuming homoscedasticity, which agrees with previous findings by Caudill, Ford, and Gropper (1995), but the differences between inefficiency indexes with and without heteroscedasticity are small. In no case are the estimates significantly different from true values, which agrees with Table 1. So, biases in terms of average inefficiency are also small.

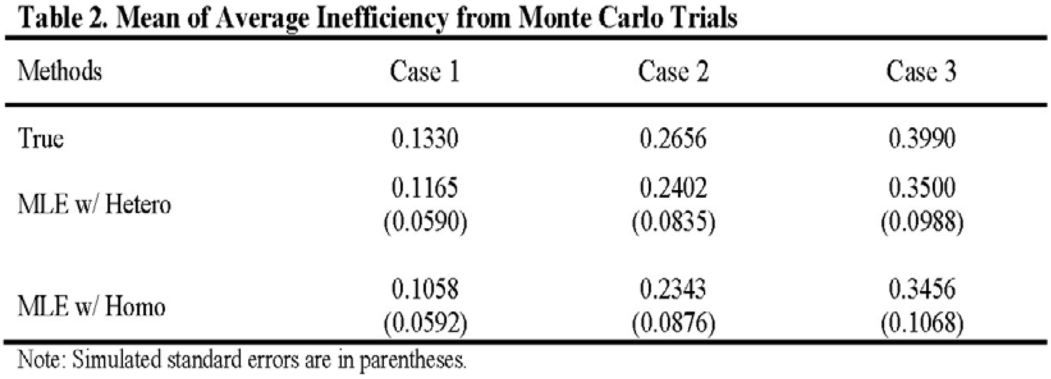


Table 2. Mean of Average Inefficiency from Monte Carlo Trials

Methods	Case 1	Case 2	Case 3
True	0.1330	0.2656	0.3990
MLE w/ Hetero	0.1165 (0.0590)	0.2402 (0.0835)	0.3500 (0.0988)
MLE w/ Homo	0.1058 (0.0592)	0.2343 (0.0876)	0.3456 (0.1068)

Note: Simulated standard errors are in parentheses.

Table 3 shows correlations between true firm-specific inefficiency indexes based on the conditional mean in equation (11) and firm specific inefficiency indexes from each method. Since the units of firm specific inefficiency indexes resulting from parametric and non-parametric methods are different, it is better to look at their correlations to compare the two methods. Also, rank correlations are reported in parentheses because ranks are frequently used after estimating an efficiency index. The stochastic frontier inefficiency measures show high correlations (mostly greater than 0.9) with the true values. DEA, however, has much smaller correlations ranging from 0.4 to 0.6.

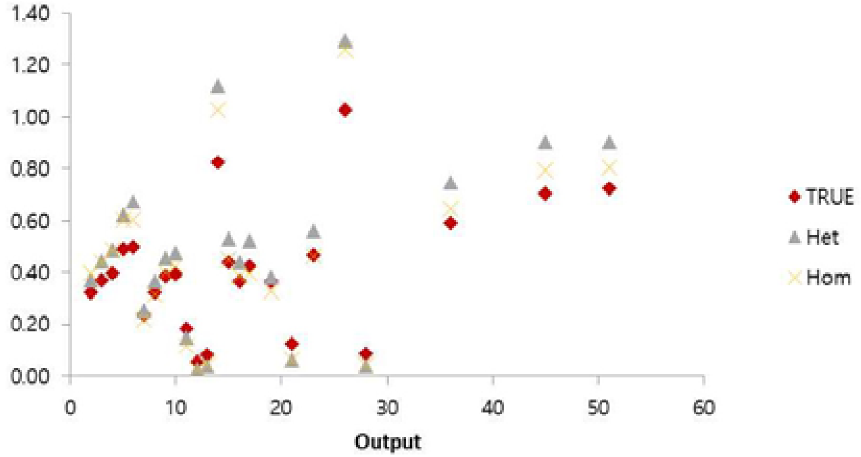
Table 3. Correlations from Monte Carlo Trials of True Inefficiency and Estimated Firm Specific Inefficiency

Methods	Case 1	Case 2	Case 3
MLE w/ Hetero	0.9155 (0.9784)	0.9259 (0.9876)	0.9331 (0.9891)
MLE w/ Homo	0.8959 (0.9717)	0.9217 (0.9880)	0.9265 (0.9914)
DEA-CRS	0.3767 (0.8518)	0.4839 (0.8371)	0.5614 (0.8300)
DEA-VRS	0.4044 (0.5720)	0.5202 (0.6537)	0.5899 (0.6985)

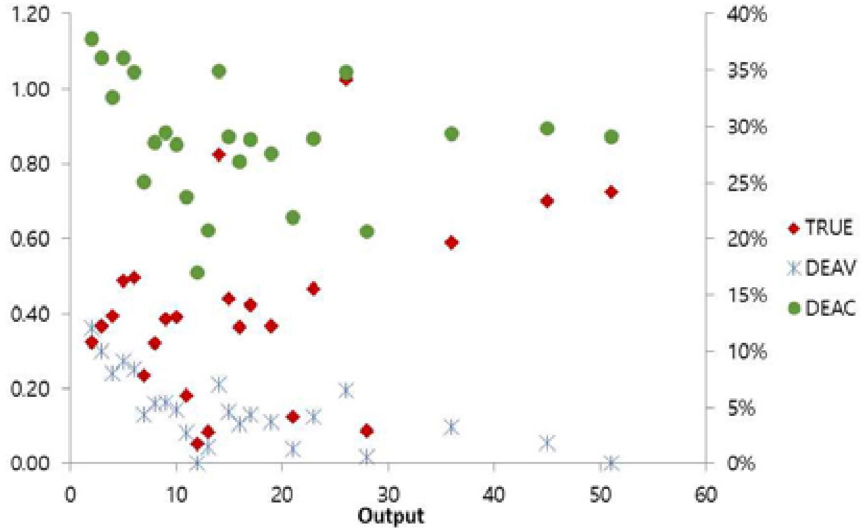
Note: Rank Correlations are reported in parentheses.

Figure 1 shows the relationship between firm-specific inefficiency indexes and output in the first sample in case 3 where variability of inefficiency error is high. The figure shows how data aggregation can affect estimates of inefficiency by size of firm. As shown in Figure 1(a), inefficiency indexes from stochastic frontier cost functions have a similar pattern regardless of whether heteroscedasticity is considered or not.

Figure 1 - Firm specific inefficiency index over averaged output in the 1st sample in case 3



(a) Inefficiency Index from Stochastic Frontier Cost Function



(b) Inefficiency Index from DEA of Constant Return to Scale and Variable Return to Scale

However as shown in Figure 1(b), inefficiency indexes from DEA with VRS have a much different pattern compared to true inefficiency indexes. DEA with CRS does not show greater inefficiency with small farms, but this result is due to imposing constant returns to scale. DEA with VRS, however, tends to have inefficient small firms relative to efficient large firms even though the true inefficiency indexes do not vary by size. Thus, as the example shows, DEA falsely leads to finding small firms having greater inefficiency, but this result is driven both by the heteroscedasticity and the larger number of small firms.

6. Conclusions

This article studies estimation of stochastic frontier (total) cost functions with heteroscedasticity from using aggregated data. Aggregation creates heteroscedasticity in the unexplained error. Each unit within a firm is assumed to have identical inefficiency. Future research may want to consider the effects of aggregation when the inefficiency varies by individual unit. The stochastic frontier functions hold up rather well in the presence of data aggregation, but DEA shows low correlations with actual inefficiency and also DEA with VRS incorrectly finds that small firms have more inefficiency.

References

Adkins, L.C., and R.L. Moomaw. (2003). The impact of local funding on the technical efficiency of Oklahoma schools. Economic Letters 81:31-37.

Aigner, D., C.A.K. Lovell, and P. Schmidt. (1977). Formulation and estimation of stochastic frontier production models. Journal of Econometrics 6: 21-37.

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

Caudill, S.B., and J.M. Ford. (1993). Biases in frontier estimation due to heteroscedasticity. Economics Letters 41: 17-20.

Caudill, S.B., J.M. Ford, and D.M. Gropper. (1995). Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. Jounal of Business & Economic Statistics 13(1): 105-111.

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

Dickens, W.T. 1990. Error components in grouped data: Is it ever worth weighting. The Review of Economics and Statistics 72(2): 328-333.

Dominguez-Molina, J. A., G. Gonzalez-Farias, and R. Ramos-Quiroga. (2003). Skew-normality in stochastic frontier analysis. Comunicacion Tecnica I-03-18: 1-13.

Fare, R., S. Grosskopf, and C. A. K. Lovell. (1994). Production Frontiers. Cambridge University Press.

Farrell, M.J. 1957. The measurement of productive efficiency. Journal of the Royal Statistical Society. Series A(General) 120: 253-90.

Featherstone, A.M., M.R. Langemeier, and M. Ismet. (1997). A nonparametric analysis of efficiency for a sample of kansas beef cow farms. Journal of Agricultural and Applied Economics 29:175-184.

Genton, M. G. 2004. Skew-Elliptical Distributions and Their Applications: A Journey beyond Normality. Florida: Chapman & Hall/CRC.

Gonzalez-Farias, G., A. Dominguez-Molina, and A. K. Gupta. (2004). Additive properties of skew normal random vectors. Journal of Statistical Planning and Inference 126: 512-534.

Greene, W.H. (2003). Econometric Analysis, 5th ed. New Jersey: Pearson Education, Inc. Hadri, K. (1999). Estimation of a doubly heteroscedastic stochastic frontier cost function. Journal of Business & Economic Statistics 17(3): 359-363

Jondrow, J., C.A.K. Lovell, I.S. Materov, and P. Schmidt. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. Journal of Econometrics 19: 233-238.

MacDonald, J.M., and E.O. Michael. (2000). Scale economies and consolidation in hog slaughter. American Journal of Agricultural Economics 82(2): 334-346.

Melton, B. E., and W. E. Huffman. (1995). Beef and pork packing costs and input demands: Effect of unionization and technology. American Journal of Agricultural Economics 77(3): 471-485.

Ollinger, M., J.M. MacDonald, and M. Madison. (2005). Technological change and economies of scale in U.S. poultry processing. American Journal of Agricultural Economics 87(1): 116-129.

Schmidt, P. (1986). Frontier production functions. Econometric Reviews 4: 289-328.

Weinstein, M.A. (1964). Query 2: The sum of values from a normal and a truncated normal distribution. Technometrics 6(1): 104-105.