J Prod Anal (2013) 39:101–110
DOI 10.1007/s11123-012-0277-z

Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes

Ryan L. Mutter · William H. Greene · William Spector · Michael D. Rosko · Dana B. Mukamel

Published online: 20 March 2012

Abstract This paper examines the impact of an endogenous cost function variable on the inefficiency estimates generated by stochastic frontier analysis (SFA). The specific variable of interest in this application is endogenous quality in nursing homes. We simulate a dataset based on the characteristics of for-profit nursing homes in California, which we use to assess the impact on SFA-generated inefficiency estimates of an endogenous regressor under a variety of scenarios, including variations in the strength and direction of the endogeneity and whether the correlation is with the random noise or the inefficiency residual component of the error term. We compare each of these cases when quality is included and excluded from the cost equation. We provide evidence of the impact of endogeneity on inefficiency estimates yielded by SFA under these various scenarios and when the endogenous regressor is included and excluded from the model.

Keywords Stochastic frontier analysis · Endogeneity · Efficiency · Quality · Nursing homes

JEL Classification C13 · C15 · I12

R. L. Mutter (✉) · W. Spector
Center for Delivery, Organization and Markets, Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, USA
e-mail: Ryan.Mutter@ahrq.hhs.gov

W. Spector
e-mail: William.Spector@ahrq.hhs.gov

W. H. Greene
Stern School of Business, New York University, 44 West 4th St., New York, NY 10012, USA
e-mail: wgreene@stern.nyu.edu

M. D. Rosko
School of Business Administration, Widener University, One University Place, Chester, PA 19013, USA
e-mail: mdrosko@widener.edu

D. B. Mukamel
Department of Medicine, Health Policy Research Institute, University of California, Irvine, Irvine, CA 92697, USA
e-mail: dmukamel@uci.edu

1 Introduction

Rising expenditures for health care across the continuum of providers has increased interest in identifying ways to reduce inefficiency. At the same time, improving quality of care continues to be a major concern of researchers, providers, and policy makers. Although researchers have provided estimates of inefficiency for nursing homes (and hospitals), there is concern that the analytical techniques used do not necessarily provide unbiased estimates. One concern that is raised, in particular, is that quality is not always explicitly incorporated into health care efficiency measurement approaches (Hussey et al. 2009).

A growing number of studies have employed stochastic frontier analysis (SFA) to analyze cost-inefficiency of health care organizations (Hollingsworth 2008). This technique estimates a best-practice cost frontier and decomposes departures from the frontier into statistical noise (i.e., assumed to be distributed as N[0, σ^2]), v, and positive departures that represent cost-inefficiency, u (Jondrow et al. 1982), where cost-inefficiency includes both technical and allocative inefficiency. Estimated firm-level cost inefficiency is the percentage by which observed costs exceed minimum costs predicted for the best-practice cost frontier (Lovell 1994).

With SFA, it is assumed that u and v are uncorrelated with cost equation variables (Greene 2011). This assumption is reasonable for competitive markets, where input prices are exogenous, and excess supply markets, where output quantities are demand driven, say, where occupancy rates are sufficiently low that providers take all comers. This assumption, however, may be less realistic if quality is included in the cost equation and there are producer-specific differences in the quality of the output produced. Indeed, in competitive markets, the amount of quality a provider produces is simultaneously determined with output and depends on costs. Consequently, it is likely to be endogenous; therefore, including it in the model entails a violation of the assumptions of SFA.

Since SFA decomposes the residual into two parts, endogenous quality may be correlated with u and possibly v as well. Researchers might try to avoid this problem by omitting quality from the cost equation. But if quality is cost enhancing and not included in the cost equation, a producer that provides more quality may be incorrectly measured as being more inefficient compared with a provider that provides less quality. This bias would be a reflection of a missing variable in the cost equation and could result in cost variables being correlated with u or v. Yet if endogenous quality were included in the cost equation it certainly would be correlated with u or v. Both cases violate the assumption that u and v are distributed independently of the cost variables and would result in biased parameter estimates in the cost function. However, the effects of this bias on the inefficiency estimates generated by SFA are unknown. Since analysts generally use SFA for the purpose of estimating inefficiency, this is a particularly important gap in our knowledge.

The problems associated with the use of endogenous variables in econometric modeling have been the subject of extensive consideration for some methods, such as ordinary least squares (OLS) where instrumental variable approaches are commonly used to estimate functions with endogenous variables (although the importance of identifying strong instruments often makes this difficult in practice) (Greene 2011; Stock et al. 2002. However, Greene (2011) pointed out that accounting for endogeneity in non-linear models, such as SFA, is difficult. Indeed, no accepted approach for estimating unbiased efficiency estimates with endogenous variables, such as quality, is currently available for SFA.

Consequently, in SFA studies of health care organizations, the impact of endogenous quality on the bias of inefficiency estimates has not been addressed. Studies of cost inefficiency using SFA in nursing homes (and hospitals—where the majority of papers studying the health care sector have been published) have dealt with quality in a number of ways ranging from leaving it out of the cost equation (Anderson et al. 1999; Chirikos 1998; Li and Rosenman 2001), to only measuring some particular aspect of quality (e.g., a measure of nurse staffing) (Farsi and Filippini 2004) to using a broad measure of quality (e.g., regulator documented deficiencies, liability settlements) (Vitaliano and Toren 1991; Vitaliano and Toren 1996). Rosko and Mutter (2008) and Mutter et al. (2008) examined the potential bias from omitting quality from a hospital cost equation by comparing SFA results with and without different proxy variables for quality. They concluded that mean estimated cost-inefficiency is relatively insensitive to the addition of hospital mortality rate variables in models that already include teaching status and controls for patient burden of illness. However, Rosko and Mutter (2008) and Mutter et al. (2008) did not address the impact of endogeneity on inefficiency estimates.

This paper focuses on the potential impact of an endogenous variable on cost inefficiency estimates generated by SFA. We used simulation techniques with an endogenous variable intended to approximate the characteristics of endogenous quality in the nursing home market. Our analysis simulates the effects of a number of aspects of endogeneity: strength of the endogeneity, whether the correlation is with u or v, and if the correlation is positive or negative. We also compare each of these cases with quality included and excluded from the cost equation. We anchor the simulation by approximating the cost and quality characteristics of nursing homes in California. By conducting our analysis using a cost equation that reflects actual practice, we hope to increase the realism of our findings.

2 Data and methods

2.1 Creating the simulated dataset

We based our simulated dataset on data from three sources. The Medicaid cost reports from the California Office of Statewide Health Planning and Development (OSHPD) are annual financial reports submitted by all nursing homes to the State. They include information about expenditures, revenues, and staffing and are used by the State to determine Medicaid payment rates. The Minimum Data Set (MDS) is a patient-level dataset mandated by the Centers for Medicare & Medicaid Services (CMS) that includes information about the socio-demographics and health status of all nursing home residents. It is used by CMS to calculate the Medicare payment rates for nursing homes and the quality measures included in the web-based Nursing Home Compare report card. The third source, the Online Survey, Certification and Reporting (OSCAR) data, is also maintained by CMS. It is a facility-level dataset that includes information on deficiencies issued to nursing homes not meeting quality standards. Data were obtained for 779 for-profit nursing homes in California in 2005. We used the following variables: total annual costs, case-mix adjusted patient days (i.e., case mix index [CMI]*days), case-mix adjusted admissions (i.e., CMI*admissions), price of capital (i.e., average of lease, depreciation, and interest per bed at the county level), and price of labor (i.e., price of nurse aides at the county level).[1] We employed the standard assumption of linear homogeneity in input prices by normalizing the financial variables in the cost equation by the wage rate, which causes the wage rate to drop out of the cost equation.

We considered four functional forms for the cost equation: Cobb-Douglas, translog, homothetic, and a hybrid model taken from the literature with squared and cubed variables (Nyman 1988). We selected a functional form based on statistical tests. Since the hybrid and translog cost equations are not nested, we used the Akaike information criterion (AIC) and Bayesian information criterion (BIC) to determine which functional form best fit the data. The translog outperformed the Cobb Douglas and hybrid models, but based on a likelihood ratio test, we selected the homothetic cost function as an acceptable simplification of the translog model.

The starting point of our simulated nursing home quality variable, QS, was the actual number of deficiency citations received by a nursing home minus the average number of deficiencies for the licensing and certification region in which the facility is located, QA. This measure adjusts for the fact that nursing home regulatory teams vary in the propensity to cite facilities for care problems and in many States there is more than one regional regulatory team. Therefore, this approach is likely to provide a better measure of relative quality across regions, and it is common in nursing home studies (Li et al. 2010; Mukamel et al. 2011).

Our first step in creating QS was to regress QA on the independent variables in the cost function. The coefficients from that regression were used as parameter values in the quality function used to construct QS. (Note that this is essentially what Gertler and Waldman (1992) did in their quality-adjusted cost function paper.) Our second step was to include a variable, INTER, that reflects the extent to which a hypothetical, cost-neutral, quality-improving intervention has been adopted by nursing homes. The variable, INTER, was created by randomly sampling from a continuous, (0.1) uniform distribution. We included it to impart unexplained variation to QS. We assumed that the coefficient on INTER was 0.5. The interpretation of that coefficient is that a 100 percent adoption of an assumed quality improvement intervention would improve quality by 0.5 units. For ease of interpretation, we assumed that higher QS implies higher facility-level quality. Thus, QS was created as follows:

$$\begin{aligned} \mathrm{QS} &= -23.80 + 0.22 * \mathrm{LN\_PK} + 2.22 * \mathrm{LNCMIDAY} \\ &\quad + 0.09 * \mathrm{LNCMIADM} + 0.50 * \mathrm{INTER} \end{aligned} \tag{1}$$

where LN_PK is the natural log of the price of capital, LNCMIDAY is the natural log of case-mix adjusted patient days, and LNCMIADM is the natural log of case-mix adjusted admissions.

To simulate the natural log of total cost, we first computed the parameters of a stochastic frontier cost function based on the actual cost function variables to generate the variance parameters, σ_u and σ_v, which we describe below. We assumed that the inefficiency residual followed a half-normal distribution. We made this assumption because the half-normal distribution is frequently assumed in the literature and because of its tractability in this simulation. The stochastic frontier cost function was specified as follows:

$$\begin{aligned} \mathrm{LN\_TC} = \alpha_0 &+ \alpha_1 \mathrm{LN\_PK} + \alpha_2 \mathrm{LNCMIDAY} \\ &+ \alpha_3 \mathrm{LNCMIADM} + \alpha_4 \mathrm{CMIDAYSQ} \\ &+ \alpha_5 \mathrm{CMIADMSQ} + \alpha_6 \mathrm{PK\_SQ} + \alpha_7 \mathrm{ADMDAY} \\ &+ \alpha_8 \mathrm{QS} + \mathrm{v_i} + \mathrm{u_i} \end{aligned} \tag{2}$$

where LN_TC is the natural log of total cost, CMIDAYSQ is 0.5*LNCMIDAY*LNCMIDAY, CMIADMSQ is 0.5*LNCMIADM*LNCMIADM, PK_SQ is 0.5*LN_PK*LN_PK, AD MDAY is LNCMIDAY*LNCMIADM, v_i is statistical noise distributed as $N(0, \sigma^2)$, u_i consists of positive departures from the best-practice cost frontier and represents inefficiency, and α_i are parameters to be estimated.

SFA yields estimates of the standard deviations of u, σ_u, and v, σ_v, which we used to create simulated inefficiency, u_γ, and random noise, v_κ, variables. (Those variables were used in the creation of simulated total cost, TC_S, as we describe below.) We created u_γ as follows:

$$\mathrm{u}_\gamma = \sigma_u e^{(\gamma * \mathrm{QS})} |\mathrm{X}| \tag{3}$$

where |X| is the absolute value of a random variable that is N(0, 1), and γ determines the strength of the endogeneity between QS and u (i.e., if $\gamma = 0$ quality is exogenous and higher values of γ denote a stronger association between QS and u). The device in (3) makes the expected value of u_γ dependent on QS when γ is nonzero.

We created v_κ as follows:

$$\mathrm{v}_\kappa = \sigma_v \left(\kappa \mathrm{QS\_STD} + \left(1 - \kappa^2\right)^{.5} \mathrm{Y} \right) \tag{4}$$

where Y is a random variable that is N(0, 1), and QS_STD is QS in standardized form, which gives it a mean of 0 and a standard deviation of 1. Endogeneity between QS and v is

[1] See Mukamel et al. 2011 for more information about this dataset and variables definitions.

determined by κ (i.e., $\kappa = 0$ is exogenous quality and higher values of κ denote a stronger association between QS and v). In (4), v_κ is projected on QS to induce the correlation (endogeneity) when κ is nonzero; if κ equals zero, then v_k is random noise, uncorrelated with QS.

We used a range of values of γ in (3) and κ in (4) to assess the impact of different levels of endogeneity that may be present in empirical data that are used to estimate inefficiency. In conducting this analysis, we did not want to simply select a particular strength of association that could drive our results and influence our conclusions.

Finally, we regressed LN_TC on the actual cost function variables to get parameter values for use in simulating TC_S. We ran the following regression:

$$\begin{aligned} \mathrm{LN\_TC} = {} & \beta_0 + \beta_1 \mathrm{LN\_PK} + \beta_2 \mathrm{LNCMIDAY} \\ & + \beta_3 \mathrm{LNCMIADM} + \beta_4 \mathrm{CMIDAYSQ} \\ & + \beta_5 \mathrm{CMIADMSQ} + \beta_6 \mathrm{PK\_SQ} + \beta_7 \mathrm{ADMDAY} \\ & + e_i \end{aligned} \tag{5}$$

where β_i are parameters to be estimated, and e_i is the error term.

Using these parameter estimates, the variables already created, and assuming that the coefficient on QS is 0.2 (i.e., improvements in quality are associated with cost *increases*), we generated TC_S as follows[2]:

$$\begin{aligned} \mathrm{TC\_S} = {} & 20.01 - 3.28 * \mathrm{LN\_PK} - 1.07 * \mathrm{LNCMIDAY} \\ & + 1.13 * \mathrm{LNCMIADM} + 0.27 * \mathrm{CMIDAYSQ} \\ & + 0.11 * \mathrm{CMIADMSQ} + 0.59 * \mathrm{PK\_SQ} \\ & - 0.16 * \mathrm{ADMDAY} + 0.2 * \mathrm{QS} + u_\gamma + v_\kappa \end{aligned} \tag{6}$$

Our results are based on SFA with TC_S as the dependent variable in the cost equation and LN_PK, LNCMIDAY, LNCMIADM, CMIDAYSQ, CMIADMSQ, ADMDAY, and QS as independent variables. By varying the values of γ and κ, we are able to explore the impact of increasing the degree of endogeneity of QS on the variable of interest, inefficiency.

2.2 Analytic approach

We first generated inefficiency estimates for the case where QS and u are positively related (i.e., the endogeneity is in u) and γ ranged between 0.1 and 1.0. (The relationship between QS and u would be positive, for example, if in some nursing homes, staff are quality leaders, who innovate and take risks. Some of their innovations will improve quality. Others that they may choose to adopt in order to improve quality may not prove themselves, and there will, therefore, be no quality returns on those investments. The implementation of the unsuccessful innovations will raise costs, however, and the facilities will have higher inefficiency because of the extra inputs used to produce the same amount of outputs.) We compared mean estimated inefficiency and the correlation of inefficiency scores between models where $\gamma = 0$ and $\gamma > 0$ with QS included and omitted from the cost function to examine the impact of endogeneity and the decision to include or exclude the endogenous variable on estimates of inefficiency generated by SFA.

Second, we generated inefficiency estimates for the case where endogeneity is in u, γ ranges between 0.1 and 1.0, and the relationship between QS and u is negative. (The relationship between QS and u would be negative, for example, if the innovative staff nursing homes hire to improve quality also find cost savings in other aspects of the facilities' operations [e.g., outsourcing laundry services]. High quality and high efficiency can result from following the principles of Total Quality Management and related approaches [Deming (1982)]).We compared mean estimated inefficiency and the correlation of inefficiency scores between models where $\gamma = 0$ and $\gamma > 0$ with QS included and omitted from the cost function.

Third, we generated inefficiency estimates for the case where endogeneity is in v (and not u) and κ ranges between 0.1 and 1.0. (The relationship between QS and v would be positive, for example, if some nursing homes hired highly skilled staff who generated higher quality care but also commanded higher facility-specific wages. Since the model only accounts for area-level wages, the higher nursing home wages would be swept into the error term, and there would be a positive relationship between QS and v.) We compared mean estimated inefficiency and the correlation of inefficiency scores between models where $\kappa = 0$ and $\kappa > 0$ with QS included and omitted from the cost function.

Finally, as a robustness check, we added random noise into the construction of QS to weaken the association between QS and TC, and we again compared the impact on mean estimated inefficiency and the correlation of inefficiency scores with various values of γ.

3 Results

3.1 Positive relationship between QS and u

We ran SFA on the simulated dataset for the cases where QS and u are unrelated (i.e., $\gamma = 0$) and related (i.e., $\gamma > 0$) and the relationship between QS and u is positive. For the models where there was endogeneity, γ ranged from 0.1 to 1.0 and increased in increments of 0.1. (SFA did not converge when

[2] The negative coefficients on the price and output variables in the above regression are probably due to multicolinearity. We used a Cobb-Douglas model in a robustness check and the coefficients on the price and output variables were all positive.

$\gamma = 0.5$.) Table 1 provides descriptive statistics for estimated inefficiency for each of those levels of endogeneity with QS included and excluded from the cost function.

Model 1—exogenous quality with QS included in the model—serves as our benchmark. The difference between model 1 and model 2 gives a sense of the bias from omitting exogenous quality from the model.

Table 1 demonstrates that as the strength of endogeneity increases, mean estimated inefficiency departs further and further from the estimate generated by model 1. This finding indicates that uncorrected endogeneity can have a substantial impact on inefficiency estimates generated by SFA. Excluding the endogenous variable from the cost function only increases the magnitude of these departures minimally. With the exception of $\gamma = 0$, and $\gamma = 0.1$, mean estimated inefficiency is lower when the quality variable is retained in the cost function, but, for a given value of γ, the differences between mean estimated inefficiency when QS is included and when it is excluded are small. The relationship is illustrated in Fig. 1.

The pattern of results for the correlations with the benchmark model, which are also reported in Table 1, are similar to what we observed for mean estimated inefficiency: inefficiency estimates from models with greater endogeneity are further from the exogenous case, and models that include endogenous QS are closer to model 1 than those that exclude the endogenous variable.

3.2 Negative relationship between QS and u

We next ran SFA on a simulated dataset where we changed Eq. (3) so that the relationship between u and QS was negative:

$$u_\gamma = \sigma_u e^{(\gamma * QS * -1)} |X| \tag{3B}$$

Table 2 provides descriptive statistics for estimated inefficiency for the cases where QS and u are negatively related (i.e., $[\gamma * -1] \le 0$) and when QS is included or excluded from the cost function.

In Table 2, mean estimated inefficiency is below the benchmark model estimate of 0.16020 for low values of γ and above it for $\gamma > 0.4$. The positive difference increases for higher values of γ. The difference between mean estimated inefficiency when QS is exogenous and when it is strongly endogenous is not as large as it was when the relationship between u and QS was positive.

For most values of γ ($\gamma = 0.1$, 0.2, 0.3, 0.8, 0.9, and 1.0), mean estimated inefficiency is closer to the estimates generated by model 1 when QS is included in the model. However, for some values of γ ($\gamma = 0.4$, 0.5, 0.6, and 0.7), mean estimated inefficiency is closer when QS is excluded from the model as shown in Fig. 2. For all values of γ, however, the difference in mean estimated inefficiency when QS is included and when it is excluded is small.

Table 1 Inefficiency estimates by level of endogeneity when relationship between QS and u is positive

Model[a]	Mean estimated inefficiency	SD	Minimum	Maximum	Correlation with model 1
1. $\gamma = 0.0$, QS included	0.16028	0.09737	0.02786	0.50838	1.00000
2. $\gamma = 0.0$, QS excluded	0.15751	0.09349	0.02856	0.53727	0.98015
3. $\gamma = 0.1$, QS included	0.16969	0.10484	0.02621	0.51403	0.99301
4. $\gamma = 0.1$, QS excluded	0.16848	0.10269	0.02901	0.54427	0.97130
5. $\gamma = 0.2$, QS included	0.18307	0.11425	0.02457	0.52030	0.97571
6. $\gamma = 0.2$, QS excluded	0.18365	0.11425	0.02836	0.55137	0.95166
7. $\gamma = 0.3$, QS included	0.19996	0.12491	0.02314	0.52877	0.94932
8. $\gamma = 0.3$, QS excluded	0.20227	0.12663	0.02672	0.55987	0.92396
9. $\gamma = 0.4$, QS included	0.21999	0.13540	0.02209	0.54090	0.92073
10. $\gamma = 0.4$, QS excluded	0.22418	0.13914	0.02541	0.57098	0.89435
11. $\gamma = 0.6$, QS included	0.27258	0.15737	0.02191	0.58930	0.85762
12. $\gamma = 0.6$, QS excluded	0.27895	0.16200	0.02466	0.61185	0.83115
13. $\gamma = 0.7$, QS included	0.30834	0.17033	0.02309	0.62865	0.82014
14. $\gamma = 0.7$, QS excluded	0.31491	0.17449	0.02558	0.64564	0.79378
15. $\gamma = 0.8$, QS included	0.35135	0.18280	0.02514	0.67384	0.78584
16. $\gamma = 0.8$, QS excluded	0.35811	0.18706	0.02757	0.68834	0.75835
17. $\gamma = 0.9$, QS included	0.40138	0.19380	0.02806	0.71731	0.74920
18. $\gamma = 0.9$, QS excluded	0.40811	0.19751	0.03065	0.73162	0.72631
19. $\gamma = 1.0$, QS included	0.45213	0.19781	0.03150	0.73943	0.70711
20. $\gamma = 1.0$, QS excluded	0.46064	0.20199	0.03466	0.75872	0.68991

[a] γ is strength of endogeneity between QS and u. Higher values indicate a stronger relationship

Table 2 also reports the correlation of inefficiency estimates between model 1 and models 2–22. The correlation between the inefficiency estimates of the model with the exogenous variable and a model with an endogenous variable is always higher when the endogeneity is weaker.

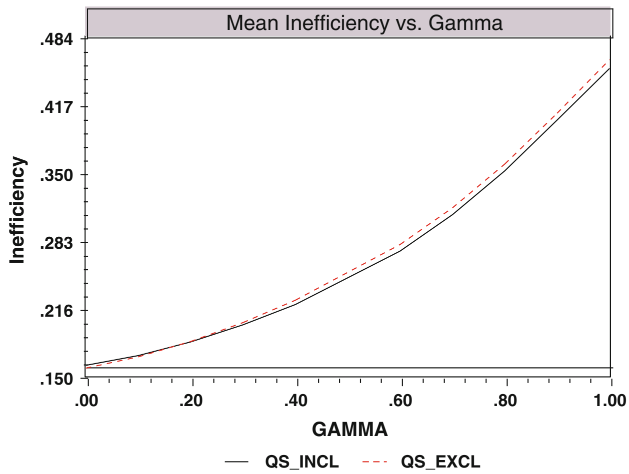

Fig. 1 Impact on mean inefficiency estimates of positive relationship between QS and u

The correlation is also higher when the endogenous variable is not dropped from the model.

3.3 Positive relationship between QS and v

We also examined the effects of endogeneity on inefficiency estimates when there is a positive relationship between QS and v (and no relationship between QS and u). We ran SFA on the simulated dataset for the cases where QS and v are unrelated (i.e., $\kappa = 0$) and related (i.e., $\kappa > 0$). For the models where there was endogeneity, κ ranged between 0.1 and 0.9 and increased in increments of 0.1. (We did not run SFA when $\kappa = 1.0$ since there is no random noise in that case.) Table 3 provides descriptive statistics for estimated inefficiency for each of those levels of endogeneity with QS included and excluded from the cost function.

When the endogenous relationship is between QS and v, mean estimated inefficiency is much less affected than when the relationship is between QS and u. (This result was expected since the v simulates random noise, not inefficiency.) For most values of κ, the difference between mean estimated inefficiency in model 1 and models with $\kappa > 0$ is small. (Indeed, models 1 and 11 have the same mean

Table 2 Inefficiency estimates by level of endogeneity when relationship between QS and u is negative

Model[a]	Mean estimated inefficiency	SD	Minimum	Maximum	Correlation with model 1
1. $\gamma = 0.0$, QS included	0.16020	0.09738	0.02786	0.50838	1.00000
2. $\gamma = 0.0$, QS excluded	0.15751	0.09349	0.02856	0.53727	0.98042
3. $\gamma = 0.1$, QS included	0.15502	0.09328	0.02829	0.50265	0.99432
4. $\gamma = 0.1$, QS excluded	0.15157	0.08867	0.02812	0.53023	0.97754
5. $\gamma = 0.2$, QS included	0.15423	0.09269	0.02774	0.49687	0.97853
6. $\gamma = 0.2$, QS excluded	0.15055	0.08811	0.02767	0.52372	0.96459
7. $\gamma = 0.3$, QS included	0.15725	0.09528	0.02731	0.49148	0.95426
8. $\gamma = 0.3$, QS excluded	0.15365	0.09077	0.02729	0.51823	0.94421
9. $\gamma = 0.4$, QS included	0.16325	0.09980	0.02708	0.48697	0.92518
10. $\gamma = 0.4$, QS excluded	0.16009	0.09604	0.02708	0.51403	0.91625
11. $\gamma = 0.5$, QS included	0.17143	0.10468	0.02713	0.48362	0.89797
12. $\gamma = 0.5$, QS excluded	0.16897	0.10211	0.02716	0.51131	0.88912
13. $\gamma = 0.6$, QS included	0.18155	0.10951	0.02630	0.48164	0.87030
14. $\gamma = 0.6$, QS excluded	0.18007	0.10858	0.02588	0.51007	0.86231
15. $\gamma = 0.7$, QS included	0.19343	0.11361	0.02523	0.48064	0.84473
16. $\gamma = 0.7$, QS excluded	0.19295	0.11404	0.02492	0.51019	0.83882
17. $\gamma = 0.8$, QS included	0.20719	0.11721	0.02453	0.47892	0.81918
18. $\gamma = 0.8$, QS excluded	0.20794	0.11901	0.02426	0.50962	0.81314
19. $\gamma = 0.9$, QS included	0.22270	0.12067	0.02428	0.47744	0.78921
20. $\gamma = 0.9$, QS excluded	0.22479	0.12340	0.02392	0.50734	0.78562
21. $\gamma = 1.0$, QS included	0.23949	0.12332	0.02470	0.48048	0.76292
22. $\gamma = 1.0$, QS excluded	0.24276	0.12664	0.02414	0.50702	0.76044

[a] γ is strength of endogeneity between QS and u. Higher values indicate a stronger relationship

Table 3 Inefficiency estimates by level of endogeneity when relationship between QS and v is positive

Model[a]	Mean estimated inefficiency	SD	Minimum	Maximum	Correlation with model 1
1. $\kappa = 0.0$, QS included	0.16014	0.09743	0.02786	0.50838	1.00000
2. $\kappa = 0.0$, QS excluded	0.15751	0.09349	0.02856	0.53727	0.98041
3. $\kappa = 0.1$, QS included	0.16018	0.09756	0.02773	0.50612	0.99997
4. $\kappa = 0.1$, QS excluded	0.15751	0.09353	0.02849	0.53689	0.97893
5. $\kappa = 0.2$, QS included	0.16025	0.09800	0.02730	0.49922	0.99988
6. $\kappa = 0.2$, QS excluded	0.15760	0.09397	0.02818	0.53247	0.97725
7. $\kappa = 0.3$, QS included	0.16031	0.09863	0.02658	0.48730	0.99945
8. $\kappa = 0.3$, QS excluded	0.15777	0.09478	0.02761	0.52376	0.97523
9. $\kappa = 0.4$, QS included	0.16034	0.09945	0.02552	0.46969	0.99815
10. $\kappa = 0.4$, QS excluded	0.15801	0.09596	0.02676	0.51028	0.97256
11. $\kappa = 0.5$, QS included	0.16014	0.10007	0.02409	0.44536	0.99481
12. $\kappa = 0.5$, QS excluded	0.15819	0.09718	0.02560	0.49123	0.96859
13. $\kappa = 0.6$, QS included	0.15945	0.10027	0.02223	0.41286	0.98715
14. $\kappa = 0.6$, QS excluded	0.15824	0.09845	0.02409	0.46542	0.96170
15. $\kappa = 0.7$, QS included	0.15711	0.09809	0.01986	0.37010	0.97056
16. $\kappa = 0.7$, QS excluded	0.15798	0.09970	0.02218	0.43101	0.94941
17. $\kappa = 0.8$, QS included	0.15152	0.09227	0.01692	0.31403	0.93338
18. $\kappa = 0.8$, QS excluded	0.15633	0.09895	0.01980	0.38497	0.92529
19. $\kappa = 0.9$, QS included	0.13624	0.07629	0.01344	0.24002	0.86105
20. $\kappa = 0.9$, QS excluded	0.15099	0.09354	0.01697	0.32187	0.87220

[a] κ is strength of endogeneity between QS and v. Higher values indicate a stronger relationship

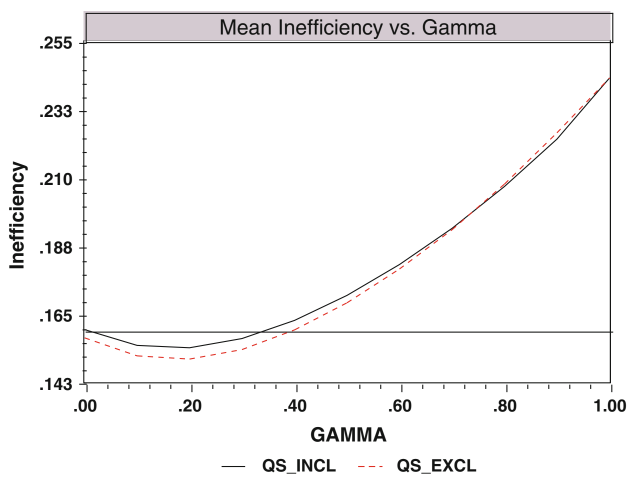

Fig. 2 Impact on mean inefficiency estimates of negative relationship between QS and u

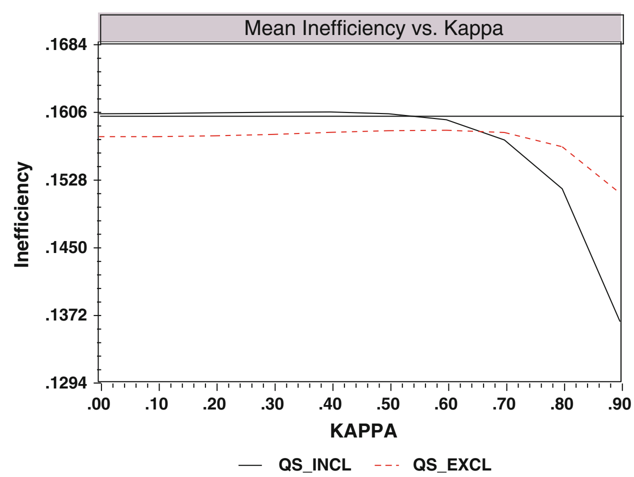

Fig. 3 Impact on mean inefficiency estimates of positive relationship between QS and v

estimated inefficiency, although the standard deviations, minimums, and maximums are different.) For values of $\kappa < 0.7$, mean estimated inefficiency is closer to the model 1 estimate when QS is included in the model. Mean estimated inefficiency is closer to the model 1 estimate when QS is excluded for the highest values of κ as shown in Fig. 3.

The correlation between the estimates generated by model 1 and those generated when there is a positive relationship between QS and the error term are higher when the association is between QS and v than it is when the relationship is between QS and u. Table 3 demonstrates that the correlation between the estimates generated by model 1 and those generated by models when endogeneity is characterized by a positive relationship between QS and v are closer when QS is included than when it is excluded, except when $\kappa = 0.9$.

3.4 Positive relationship between QS and u with noise added to the construction of QS

We wanted to investigate whether our findings were due to the particular relationships that we had modeled. So, we added random noise to the construction of QS as a robustness check. The standard deviation of QS is 1.12. So, to make 25 percent of the variation in QS pure noise, we added a noise variable, NOISE, to Eq. (1) where NOISE was created by randomly sampling from a normal distribution with mean 0 and variance 0.10:

$$\begin{aligned} QS = & -23.80 + 0.22 * LN\_PK + 2.22 * LNCMIDAY \\ & + 0.09 * LNCMIADM + 0.50 * INTER + NOISE \end{aligned} \tag{1B}$$

Table 4 provides mean estimated inefficiency estimates by level of endogeneity with QS included and excluded from the cost function when QS has added noise and QS and u are positively related. (SFA did not converge when $\gamma = 0.5$ or 0.6.)

The pattern of the results in Table 4 is similar, but not identical, to that of Table 1. Mean estimated inefficiency departs more from the exogenous case as γ increases, and with the exception of $\gamma = 0.1$ and when $\gamma > 0.8$, mean estimated inefficiency is closer to the model 1 estimate when QS is included in the cost function. Figure 4 illustrates the relationship.

The correlation results when noise is added to the construction of QS in Table 4 follow a pattern similar to that presented in Table 1: inefficiency estimates from models with greater endogeneity are further from the exogenous case, and models that include endogenous QS are closer to model 1 than those that exclude the endogenous variable, but compared with Table 1 the differences between correlations when including and not including QS for the same

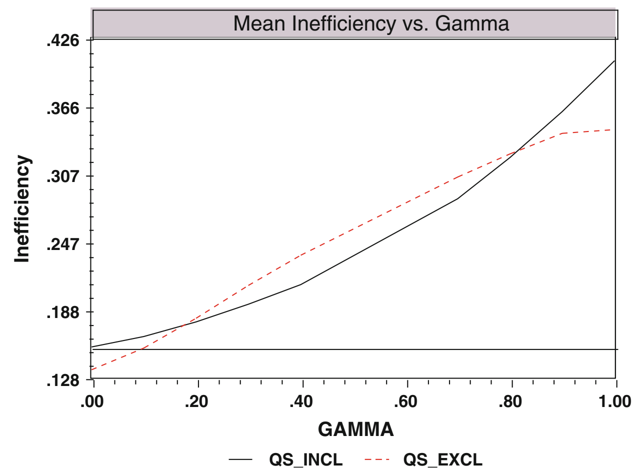

Fig. 4 Impact on mean inefficiency estimates of positive relationship between QS with noise and u

Table 4 Inefficiency estimates by level of endogeneity when relationship between QS and u is positive and noise is added to construction of QS

Model[a]	Mean estimated inefficiency	SD	Minimum	Maximum	Correlation with model 1
1. $\gamma = 0$, QS included	0.15460	0.09549	0.02661	0.43991	1.00000
2. $\gamma = 0$, QS excluded	0.13458	0.06944	0.03766	0.48223	0.91082
3. $\gamma = .1$, QS included	0.16376	0.10174	0.02574	0.44056	0.99343
4. $\gamma = .1$, QS excluded	0.15366	0.08539	0.03901	0.57504	0.89525
5. $\gamma = .2$, QS included	0.17668	0.10955	0.02494	0.43932	0.97344
6. $\gamma = .2$, QS excluded	0.17978	0.10717	0.03941	0.57091	0.87074
7. $\gamma = .3$, QS included	0.19210	0.11779	0.02455	0.44026	0.93947
8. $\gamma = .3$, QS excluded	0.20856	0.12891	0.03681	0.55417	0.83228
9. $\gamma = .4$, QS included	0.20927	0.12577	0.02403	0.45011	0.90439
10. $\gamma = .4$, QS excluded	0.23511	0.14365	0.03316	0.53815	0.79116
11. $\gamma = .7$, QS included	0.28472	0.15578	0.02073	0.54900	0.80373
12. $\gamma = .7$, QS excluded	0.30375	0.15581	0.02459	0.51164	0.69665
13. $\gamma = .8$, QS included	0.32081	0.16683	0.02134	0.59252	0.77281
14. $\gamma = .8$, QS excluded	0.32402	0.15308	0.02367	0.50610	0.66790
15. $\gamma = .9$, QS included	0.36121	0.17596	0.02243	0.63044	0.74358
16. $\gamma = .9$, QS excluded	0.34219	0.14762	0.02381	0.49949	0.63744
17. $\gamma = 1.0$, QS included	0.40585	0.18394	0.02419	0.66705	0.71398
18. $\gamma = 1.0$, QS excluded	0.34534	0.12956	0.02466	0.46382	0.59765

[a] γ is strength of endogeneity between QS and u. Higher values indicate a stronger relationship

level of endogeneity are larger, making the inclusion of QS more important.

4 Discussion

The impact of an endogenous regressor on inefficiency estimates generated by SFA is an important topic for both theory and applied research that has not been addressed in the literature. This study is a first step in trying to better understand endogeneity's effects on SFA inefficiency estimates.

In our analyses, we use simulated data based on the real-world example of the nursing home industry. As health economists and health services researchers continue to study the inefficiency of health care providers and to incorporate recently developed and available measures of provider quality into their models, it will be important to understand the impact of endogeneity on inefficiency measurement since quality is often endogenous in such models. Applied researchers, in particular, need guidance on whether an endogenous variable should be retained in an SFA cost function.

"Endogeneity" is often discussed, tested for, and corrected for in the applied literature; however, researchers are frequently not clear on what exactly it is that they mean by "endogeneity." This paper specifies specific relationships among variables.

When the endogeneity is due to an association between the departures from the best-practice frontier that represent cost-inefficiency, u, and endogenous quality, QS, we find that models with an endogenous variable yield inefficiency estimates that are further from the estimates yielded by the model with an exogenous variable in terms of mean and correlation. The extent of the bias caused by the endogenous variable depends on the strength of the endogeneity (as measured by γ) and its direction (positive or negative relationship). Mean estimated inefficiency is higher when there is a positive relationship between QS and u than when there is a negative relationship between QS and u. This may be because simulated u is larger when the relationship between QS and u is positive than when it is negative.

In general, larger values of γ are associated with further departures from the estimates generated by the model with the exogenous variable, although this is not always the case. The correlation of the inefficiency estimates generated by a model with an exogenous variable and a model with an endogenous variable are almost always higher when the endogenous variable is included in the model. It is frequently true that mean estimated inefficiency is closer to the benchmark, exogenous quality estimate when the endogenous variable is retained. Overall, however, this "omitted variable" effect does not seem to be as large as the influence of the endogeneity itself.

We find that the impact of the association between statistical noise, v, and QS on mean estimated inefficiency and correlation of inefficiency estimates is small for most values of κ. Our results suggest that researchers will get more accurate inefficiency estimates by including the endogenous variable than excluding it if the endogeneity is due to a relationship with v.

As a robustness check, we examined the impact on inefficiency estimates of a positive association between u and QS with noise included in the construction of QS. The pattern of results was similar to what we found in the baseline analysis; however, the impact on the correlation of the inefficiency estimates with the benchmark estimates was greater when the endogenous variable was excluded when there was noise in the construction of QS.

Thus, the impact of an endogenous regressor on inefficiency estimates generated by SFA is driven by the nature and strength of the endogeneity, which suggests that the estimation of inefficiency using SFA would be improved if methods for estimating γ and κ could be identified. An additional topic for future research is identifying a means for correcting for endogeneity in SFA. Without these tools, researchers making inefficiency estimates need to rely on their understanding of the nature of the particular market they are studying to assess the potential size of bias. Our intuition is that endogeneity and the bias in inefficiency estimation that it gives rise to will be greater the larger is the variation in the delivery of care at the facility level because of differences in processes and resources that are unobserved to the analyst (e.g., difference in managerial skill, implementation of interventions designed to improve performance). Also, bias is likely to be larger the more narrowly quality is measured (i.e., the more quality is fully accounted for, the less specification error and endogeneity there will be).

Acknowledgments This paper does not represent the policy of either the Agency for Healthcare Research and Quality (AHRQ) or the U.S. Department of Health and Human Services (DHHS). The views expressed herein are those of the authors and no official endorsement by AHRQ or DHHS is intended or should be inferred. This research was partially funded by contracts from AHRQ: AHRQ Contract Number HHSP233200900285P (Michael Rosko, Dana Mukamel) and HHSN263200500063293 (William Greene).

References

Anderson RI, Lewis D, Webb JR (1999) The efficiency of nursing home chains and the implications of non-profit status. J Real Estate Portfolio Manage 5(3):235–245

Chirikos T (1998) Identifying efficiently and economically operated hospitals: the prospects and pitfalls of applying frontier regression techniques. J Health Polit Policy Law 23(December): 879–904

Deming W (1982) Out of the Crisis. Massachusetts Institute of Technology, Cambridge

Farsi M, Filippini M (2004) An empirical analysis of cost efficiency in non-profit and public nursing homes. Ann Public Cooper Econ 75:339–365

Gertler P, Waldman D (1992) Quality-adjusted cost functions and policy evaluation in the nursing home industry. J Polit Econ 100(6):1232–1256

Greene W (2011) Econometric Analysis, 7th edn. Prentice-Hall, Upper Saddle River

Hollingsworth B (2008) The measurement of efficiency and productivity of health care delivery. Health Econ 17:1107–1128

Hussey P, de Vries H, Romley J, Wang M, Chen S, Shekell P, McGlynn E (2009) A systematic review of health care efficiency measures. Health Serv Res 44(3):784–805

Jondrow J, Lovell CK, Materov I, Schmidt P (1982) On the estimation of technical efficiency in the stochastic frontier production function model. J Econ 19:233–238

Li T, Rosenman R (2001) Cost efficiency in Washington hospitals: a stochastic frontier approach using panel data. Health Care Man Sci 4(2):73–81

Li Y, Harrington C, Spector W, Mukamel D (2010) State regulatory enforcement and nursing home termination from the Medicare and Medicaid programs. Health Serv Res 45(6):1796–1814

Lovell CK (1994) Production frontiers and productive efficiency. In: Fried H, Lovell C, Schmidt S (eds) The Measurement of Productive Efficiency. Oxford University Press, New York, pp 1–67

Mukamel D, Li Y, Harrington C, Spector W, Weimer D, Bailey L (2011) Does state regulation of quality impose costs on nursing homes? Med Care 49(6):529–534

Mutter R, Rosko M, Wong H (2008) Measuring hospital inefficiency: the effects of controlling for quality and patient burden of illness. Health Serv Res 43(6):1992–2013

Nyman J (1988) The marginal cost of nursing home care: New York, 1983. J Health Econ 7:393–412

Rosko M, Mutter R (2008) Assessing the robustness of hospital inefficiency estimates: an assessment of stochastic frontier regression. Med Care Res Rev 65:131–166

Stock JH, Wright JH, Yogo M (2002) A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat 20(4): 518–529

Vitaliano D, Toren M (1991) Cost and efficiency in nursing homes: a stochastic frontier approach. J Health Econ 13:281–300

Vitaliano D, Toren M (1996) Hospital cost and efficiency in a regime of stringent regulation. East Econ J 22:161–173