UNIVERSITY
OF YORK

CENTRE FOR HEALTH ECONOMICS

HEALTH ECONOMICS CONSORTIUM

Measuring Technical Efficiency in the National Health Service: A Stochastic Frontier Analysis

by
ADAM WAGSTAFF

DISCUSSION PAPER 30

UNIVERSITY OF YORK

CENTRE FOR HEALTH ECONOMICS

MEASURING TECHNICAL EFFICIENCY IN THE NATIONAL HEALTH SERVICE : A STOCHASTIC FRONTIER ANALYSIS

by

Adam Wagstaff

Copyright : A. Wagstaff

Abstract

This paper reports the results of an investigation into the technical efficiency (i.e. the ability to convert inputs into outputs) the NHS hospital sector. The method employed is the 'stochastic frontier production function'. In contrast to the approach adopted by Feldstein in his 'Economic Analysis for Health Service Efficiency', the stochastic frontier approach recognizes that a hospital's failure to produce exactly what would be expected of it on the basis of the parameters of its production function may be due not only to technical inefficiency but also to random influences outside its control (e.g. viruses). The model is estimated on data from 193 maternity hospitals for the financial year 1971/72. Various special cases of a transcendental logarithmic frontier production function are estimated, including a Cobb-Douglas function; all are estimated under the assumption that the 'error' term reflecting inefficiency has a half-normal distribution. Surprisingly, in none of the models estimated was there any evidence of technical inefficiency. Interpreted literally, therefore, all the hospitals in the sample were operating at 100% technical efficiency, obtaining maximum 'output' (deliveries) from their bundles of inputs.

The Author

Adam Wagstaff is a Lecturer in Economics in the School of Social Sciences at the University of Sussex and a Visiting Research Fellow in the Centre for Health Economics at the University of York.

Acknowledgements

I am grateful to Yock Chong for computational assistance and to Bob Lavers for making available the data used in the paper. The research for the paper was supported by the Nuffield Provincial Hospitals Trust.

Further Copies

Further copies of this document are available (at price £2.00 to cover the costs of publication, postage and packing) from:

The Secretary
Centre for Health Economics
University of York
Heslington
YORK YO1 5DD

Please make cheques payable to the University of York. Details of other Discussion Papers can be obtained from the same address, or telephone York (0904) 430000, extension 5751/2.

The Centre for Health Economics is a Designated Research Centre of the Economic and Social Research Council and the Department of Health and Social Security.

1. Introduction

Attempts by economists to measure the efficiency of hospitals in the British National Health Service (NHS) have been relatively few. The great majority of the work which has been undertaken to date has been directed at the issue of allocative efficiency. Comparatively little research effort has been directed at the issue of technical efficiency. Indeed, there has been only one study to date which claims to shed light on the issue of technical efficiency in the NHS hospital sector, namely that of Feldstein (1967). Feldstein estimated a variety of production functions on data for large, acute non-teaching hospitals and used the residuals to identify hospitals of below-average and above-average technical efficiency. Feldstein's work represents an important step forward in the measurement of technical efficiency in NHS hospitals. It does, however, suffer from certain shortcomings, which - though difficult to avoid with the techniques available in the 1960s - may be more easily avoided with techniques developed over the course of the last few years. Specifically, the stochastic frontier production function (cf Forsund et al, 1980; Schmidt, 1986) provides a more satisfactory - though not perfect - means of estimating the extent of technical inefficiency than the method used by Feldstein.

This paper employs the stochastic frontier production function to analyse technical efficiency in a sample of NHS maternity hospitals. It begins in Section 2 with a discussion of the limitations of the previous work in the area and shows how these can be avoided using the stochastic frontier approach. Section 3 then outlines the data and model used, and discusses the estimation problems. The next section - Section 4 - presents the empirical results and the final section - Section 5 - contains a discussion.

2. The Previous Literature

The only study to date purporting to shed light on the issue of technical efficiency in the NHS hospital sector is that of Feldstein (1967). Feldstein used a variety of functional forms to estimate a production function for NHS acute hospitals and interpreted the residuals as a measure of technical efficiency. Thus hospitals with residuals equal to zero were said to be of average technical efficiency, whilst hospitals with residuals which were greater (smaller) than zero were said to be of above-average (below-average) technical efficiency (Feldstein, 1967, pp 110-115). The rationale behind this is that the output of a hospital with a zero residual is exactly the output that would be expected of it on the basis of the estimated input coefficients. A hospital with a positive (negative) residual, by contrast, produces more (less) than what it would have been expected to produce on the basis of the estimated parameters of the production function.

There are two main problems with this method. First, it only enables a ranking of hospitals by technical efficiency : it provides no information on how far a hospital is from the frontier. Second, it implicitly assumes that all cross-sample variation in the error term is due to variation in efficiency. In reality the residuals are likely to reflect random influences outside the hospital's control (viruses, for example) as well as "statistical noise".

Both problems can be overcome - at least partially - within the frontier production function approach. The first can be overcome by constraining the error term in the production function to be one-sided: hospitals can therefore produce on or beneath the frontier, but not above. Modifying the production function model in this way gives rise to the

deterministic production frontier model of Aigner and Chu (1968) and Afriat (1972). The second problem can be resolved by assuming the error term to be comprised of two parts: the first is a symmetric term capturing random shocks and noise, and the second is a one-sided term reflecting inefficiency. This gives rise to the stochastic frontier model of Aigner et al. (1977) and Meeusen and van den Broech (1977). The stochastic frontier model therefore allows both shortcomings in Feldstein's approach to be overcome.

3. Model Specification and Estimation Issues

The empirical analysis in the present paper is based on a sample of NHS maternity hospitals. Like other single speciality hospitals, maternity hospitals have the advantage that their output is relatively homogenous (cf Lavers and Whynes, 1978; Steele and Gray, 1982). The considerable problems encountered when trying to measure output in general hospitals can therefore be side-stepped (cf eg Tatchell, 1983).

3.1 The Data

The data used are those used by Lavers and Whynes (1978) - hereafter LW - and relate to 193 NHS maternity hospitals in England for the financial year 1971/72. A description of the data is to be found in LW.

3.2 Model Specifications

As in LW it is assumed the hospital's production frontier is of the translog variety. The model is therefore

$$(1)\quad \ln y_t = \alpha + \Sigma_i \beta_i \ln x_{it} + (1/2)\,\Sigma_i \Sigma_j \gamma_{ij}\, \ln x_{it} \ln x_{jt} + \varepsilon_t$$

where y_t is output for hospital t (t=1,...,T), the x_{it} are inputs, α and the β_i and γ_{ij} are parameters and ε_t an error term. As in LW output is measured by cases treated and four inputs are used: x_m is medical staff, x_b is beds, x_d is drugs and dressings and x_n is nursing staff. All inputs except x_b are measured on a per annum expenditure basis. In contrast to LW it is assumed that ε_t is composed of two parts: thus

$$(2)\quad \varepsilon_t = v_t - u_t$$

where v_t is a symmetric term capturing exogenous shocks and statistical noise and u_t is a one-sided term ($u_t \geq 0$ all t) reflecting technical inefficiency. Thus the core of the hospital's production frontier is given by

$$\alpha + \Sigma_i \beta_i \ln x_{it} + (1/2)\Sigma_i \Sigma_j \gamma_{ij} \ln x_{it} \ln x_{jt}$$

which is common to all hospitals and is non-stochastic. The actual frontier is given by

$$\alpha + \Sigma_i \beta_i \ln x_{it} + (1/2)\Sigma_i \Sigma_j \gamma_{ij}\, \ln x_{it} \ln x_{jt} + v_t$$

which is stochastic and hence varies from one hospital to the next. The extent to which the hospital operates beneath its stochastic frontier is given by the one-sided term, u_t, which provides a measure of technical inefficiency. The primary objective of the exercise is therefore to obtain an estimate of u_t for each hospital in the sample.

In passing it is worth noting that there are two features of the NHS hospital which make it a particularly interesting area of application for the frontier model. The first is that, unlike the firm, the NHS hospital may reasonably be expected to be technically inefficient. This is because there is no obvious reason why the primary decision-maker (the hospital doctor) should choose to be technically efficient. In the theory of firm, technical efficiency is a simple corrolorary of utility maximizing behaviour (cf Stigler, 1976). In the NHS hospital, by contrast, being technically efficient may generate disutility for the doctor, since treating most cases will be associated with a heavier workload (cf Culyer and Cullis, 1975). Technical inefficiency may well be consistent, therefore, with utility maximizing behaviour. The criticism of the frontier production function approach that it "flies in the face of classical microeconomics" (Greene, 1986, p 336) would seem, therefore, to carry less weight in the context of the NHS hospital than in the context of private sector industry.

The other feature of the NHS hospital which is of special interest in the present context is the fact that hospitals of comparable size will tend to facebroadly the same technology. Stigler (1976) has emphasized that since technology is costly to acquire, firms will differ in their investments in new technology and will therefore face different frontiers. The stochastic frontier model allows for cross-sample variations in the frontier, but does so by adding the symmetric error term, v_t, to the non-stochastic core of the production frontier. Each firm's frontier is therefore a "neutrally scaled transform" of every other firm's frontier, with the parameters of the production technology being common to all firms. This is, as Schmidt (1986) acknowledges, probably too simplistic: if technologies differ across firms then the perameters of the production frontier will also differ. Failure to allow for such differences will

almost certainly have consequences for the distribution of the inefficiency error term in the stochastic frontier model (cf Forsund, 1986). In a study of private sector firms, therefore, one may well end up ascribing to technical inefficiency the effects of systematic cross-sample variations in technologies. Because hospitals in the NHS may reasonably be assumed to face the same technology, however, this criticism of the frontier approach would seem to carry less weight in the context of the NHS hospital than in the context of the private sector firm.

3.3 Estimation of the Frontier Production Function

In order to estimate (1) assumptions need to be made about the distributions of v_t and u_t. The present paper follows Aigner et al. (1977) and assumes that the v_t are normally distributed with zero mean and constant variance σ_v^2, and that the u_t are half-normal: thus $u_t = |u_t^*|$ with $u_t^* \sim N(0, \sigma_u^2)$. It is also assumed that u_t and v_t are independent of one another and are independent of the x_{it}.

Since (1) is intrinsically linear, it can be estimated by maximum likelihood using the method proposed by Greene (1982). In the first step (1) is estimated by OLS : the parameter estimates and the second and third moments of the residuals are then used to obtain consistent estimates of α β_i, γ_{ij}, σ_v^2 and σ_u^2. Except in the case where the third moment of the OLS residuals, μ_3, is positive, the OLS estimates are then used as starting values and the maximum likelihood estimates are obtained by iteration. If the OLS residuals are positively skewed ($\hat{\mu}_3 > 0$), the iterative procedure becomes unnecessary. Waldman (1982) has shown that in this case the maximum likelihood estimates of α, β_i, γ_{ij}, σ_v^2 and σ_u^2 are $\hat{\alpha}$, $\hat{\beta}_i$, $\hat{\gamma}_{ij}$, S^2, and 0 respectively, where $\hat{\alpha}$, $\hat{\beta}_i$ and $\hat{\gamma}_{ij}$ denote the OLS estimates of α, β_i, γ_{ij},

and S^2 is the OLS estimate of the variance of ε. Since $\sigma_u^2 = 0$ implies $u_t = 0$ all t, a positive μ_3 would indicate that each hospital in the sample is operating at 100% technical efficiency.

4. Empirical Results

As in LW, (1) is estimated on rescaled data along the lines suggested by Sargan (1971). Specifically all variables were rescaled so that at their means the values of $\ln y$ and $\ln x_i$ (i=1,...,4) were zero. (1) can then be interpreted as a second-order Taylor series approximation to any general production function, with the β_i and γ_{ij} interpreted as first and second derivatives at the sample mean.

The equation was then estimated subject to five sets of restrictions. Model M1 is equation (1) subject to symmetry restrictions - ie $\gamma_{ij} = \gamma_{ji}$ all i,j - and is estimated using the equation

$$(3) \quad \ln y_t = \alpha + \Sigma_i \beta_i \ln x_{it} + (1/2)\Sigma_i \gamma_{ii} (\ln x_{it})^2 + \Sigma_i \Sigma_{j>i} \gamma_{ij} \ln x_{it} \ln x_{jt} + \varepsilon_t$$

Model M2 is (1) subject to symmetry and homogeneity restrictions - ie $\gamma_{ij} = \gamma_{ji}$ all i, j plus $\Sigma_i \gamma_{ij} = 0$ $j = 1,...,k$ - and is estimated from

$$(4) \quad \ln y_t = \alpha + \Sigma_i \beta_i \ln x_{it} + \Sigma_i \Sigma_{j>i} \gamma_{ij} \{ \ln x_{it} \ln x_{jt} - (1/2)[(\ln x_{it})^2 + (\ln x_{jt})^2] \} + \varepsilon_t$$

Model M3 is (1) subject to the restrictions implied by homogeneity and constant returns to scale (CTRS) - ie $\gamma_{ij} = \gamma_{ji}$ all i,j plus $\Sigma_i \gamma_{ij} = 0$ $j = 1,...,k$ and $\Sigma_i \beta_i = 1$: model M3 estimated from

$$(5) \quad \ln y_t - \ln x_{1t} = \alpha + \Sigma_{i=2}^{k} \beta_i (\ln x_{it} - \ln x_{1t}) + \Sigma_i \Sigma_j \gamma_{ij} \{ \ln x_{jt} \ln x_{it} - (1/2)[(\ln x_{it})^2 + (\ln x_{jt})^2] \} + \varepsilon_t$$

Model M4 is the Cobb-Douglas - ie $\gamma_{ij} = 0$ all i,j - and is estimated from

$$(6) \quad \ln y_t = \alpha + \Sigma_i \beta_i \ln x_{it} + \varepsilon_t$$

Finally, model M5 is the Cobb-Douglas with CRTS imposed: it is estimated using

$$(7) \quad \ln y_t - \ln x_{it} = \alpha + \sum_{i=2}^{k} \beta_i (\ln x_{it} - \ln x_{1t}) + \varepsilon_t$$

The maximum likelihood (ML) estimates of α, β_i, γ_{ij}, σ_v^2, and σ_u^2 for models M1 to M5 are given in Table 1. In each model μ_3 was positive indicating that the ML estimate of σ_u^2 is zero - ie $u_t = 0$ all t - and that the OLS estimates of α, β_i, γ_{ij} and σ_v^2 are ML. The extent of skewness in the OLS residuals can be assessed using the test statistic $\sqrt{b_1}$ (cf Schmidt and Lin, 1984), where

$$(8) \quad \sqrt{b_1} = \hat{\mu}_3 / \hat{\mu}_2^{2/3}$$

The 1% critical value of $\sqrt{b_1}$ for a one-tailed test is 0.403 (Biometrika Tables for Statisticians, Vol I), which is greater than any of the values of $\sqrt{b_1}$ in Table 1. Thus, though the OLS residuals are skewed in the "wrong" direction for a production frontier, the skewness is not significant.

Because the estimates in Table 1 are the OLS estimates, they are similar to those reported by LW. The models can be set up in nested sequences, with sequence A as M1, M2, M3 and M5, and sequence B as M1, M2, M4 and M5, where M1 is the most general model. The validity of the extra restrictions implied by each model can be tested using the likelihood

ratio. The likelihood ratio for testing the homogeneity restrictions imposed in moving from model M1 to M2 i , 30.372. The upper 0.01 point of the χ^2_4 distribution is 13.28, indicating that the homogeneity restrictions are rejected. Since each of the models M2 to M5 are special cases of model M1, the most general specification, M1, is to be preferred.

5. Discussion

Interpreted literally, the results reported in the previous section indicate that all hospitals in the sample are operating at 100% technical efficiency. Before accepting this result at face value, however, it is important to consider how far it might be due to model misspecifications.

There are two obvious sources of potential model misspecification. The first is in the specification of the systematic component of (1). It may be the case, for example, that certain relevant inputs have not been included or that those that have been included have not been measured correctly. To some extent both are likely to be true in the present context. Some inputs have been excluded from (1). However, they tend to be relatively unimportant in terms of their share of the hospital's budget and exhibit little cross-sample variation (cf Lavers and Whynes, 1978).

Moreover, it is usually argued that the effect of omitting relevant inputs will be to bias upwards rather than downwards the estimate of technical inefficiency (cf Schmidt, 1986, p321). It is also true that there may be measurement error in the inputs. As Feldstein (1967) noted, measuring the inputs in terms of expenditure rather than physical units may result in biased parameter estimates. Whether or not it will result in a biased estimate of σ^2_u , however, is not clear. On balance, therefore, it is not at all obvious that the results are due to a misspecification in the

systematic component of (1).

The second and more serious source of potential misspecification is in the error structure of (1). The estimates are based on the assumption that the v_t are $N(0, \sigma_v^2)$ and the u_t are half-normal - ie $u_t = |u_t^*|$, where $u_t^* \sim N(0, \sigma_u^2)$. This is by far the most popular assumption in applied work in the field. It appears to be the case, however, that if the u_t^* are not normally distributed, $\hat{\mu}_3 > 0$ may not necessarily imply that the ML estimate of σ_u^2 is equal to zero (Waldman, 1982, fn 5). A $\hat{\mu}_3 > 0$ may therefore be compatable with some inefficiency if the u_t^* are non-normal. Whether $\hat{\mu}_3 > 0$ is compatable with $\sigma_u^2 > 0$ when the u_t^* are normally distributed but do not have a zero mean (cf Stevenson, 1980) is not clear from the literature. If so, this might be another source of potential misspecification.

Since there are no strong a priori reasons for believing the u_t to be half-normal, this is not an entirely satisfactory state of affairs. The problem would be worse, however, if the $\sqrt{b_1}$ statistics had been significantly different from zero. As it is, it is not clear how the results of the $\sqrt{b_1}$ tests could be consistent with $\sigma_u^2 > 0$, unless the v_t were non-normal. Ultimately, therefore, as recently acknowledged by Schmidt (1986), one's conclusions regarding the extent of technical inefficiency depend critically on the correctness of the distributional assumptions about the error structure of the frontier. All than can be concluded in the present context is that providing one is prepared to defend the assumption that the v_t are $N(0,\sigma_v^2)$, there would not appear to be any evidence of technical inefficiency in this particular sample of maternity hospitals.

References

Afriat, S N, 1972, Efficiency estimation of production functions, International Economic Review 13, 568-98.

Aigner, D J and S F Chu, 1968, On estimating the industry production function, American Economic Review 58, 226-39.

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

Culyer, A J and J G Cullis, 1975, Hospital waiting lists and the supply and demand of inpatient care, Social and Economic Administration 9, 13-25.

Feldstein, M S, 1967, Economic analysis for health service efficiency: econometric studies of the British National Health Service (North-Holland, Amsterdam).

Forsund, F R, 1986, Comment on "Frontier production functions" by P Schmidt, Econometric Reviews 4, 329-34.

Forsund, F R, C A K Lovell and P Schmidt, 1980, A survey of frontier production functions and of their relationship to efficiency measurement, Journal of Econometrics 13, 5-25.

Greene, W H, 1982, Maximum likelihood estimation of stochastic frontier production models, Journal of Econometrics 18, 285-9.

Greene, W H, 1986, Comment on "Frontier production functions" by P Schmidt, Econometric Reviews 4, 335-8.

Lavers, R J and D K Whynes, 1978, A production function analysis of English maternity hospitals, Socio-Economic Planning and Sciences 12, 85-93.

Meussan, W and J van den Broek, 1977, Efficiency estimation from Cobb-Douglas production functions with composed error, International Economic Review 18, 435-44.

Sargan, I D, 1971, Production functions, in P R Layard et al,. eds., Qualified manpower and economic performance (The Penguin Press, London).

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

Schmidt, P and T F Lin, 1984, Simple tests of alternative specifications in stochastic frontier models, Journal of Econometrics 24, 349-61.

Steele, R and A M Gray, 1982, Statistical cost analysis: the hospital case, Applied Economics 14, 491-502.

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

Stigler, G J, 1976, the Xistence of X-efficiency, American Economic Review 66, 213-6.

Tatchell, M, 1983, Measuring hospital output: a review of the service-mix and case-mix approaches, Social Science and Medicine 17, 871-83.

Waldman, D M , 1982, A stationary point for the stochastic frontier likelihood, Journal of Econometrics 18, 275-9.

Table 1 - Maximum likelihood estimates of translog production frontier

	M1	M2	M3	M4	M5
α	-0.038 (1.34)	-0.034 (1.48)	-0.045 (1.97)	-0.016 (0.68)	0.014 (0.59)
β_m	-0.052 (2.69)	-0.036 (1.80)	-0.005 (2.90)	0.002 (0.35)	0.000 (0.10)
β_b	0.492 (6.68)	0.383 (5.43)	0.414 (5.94)	0.216 (4.42)	0.225 (4.73)
β_d	0.213 (3.28)	0.246 (3.73)	0.185 (3.14)	0.186 (3.90)	0.161 (4.14)
β_n	0.259 (3.02)	0.343 (3.92)	0.456 (6.72)	0.566 (7.86)	0.613 (12.25)
γ_{mm}	-0.004 (1.36)	-0.006 (8.63)	-0.008 (2.00)		
γ_{bb}	0.151 (1.69)	0.376 (5.68)	0.378 (5.76)		
γ_{dd}	0.129 (0.81)	0.429 (3.71)	0.394 (3.41)		
γ_{nn}	0.124 (0.32)	-0.508 (2.48)	-0.489 (2.34)		
γ_{mb}	0.076 (1.81)	-0.016 (0.83)	-0.013 (0.66)		
γ_{md}	-0.006 (0.30)	-0.001 (0.08)	-0.004 (0.47)		
γ_{mn}	-0.142 (2.59)	0.023 (1.25)	0.025 (1.42)		
γ_{bd}	-0.636 (2.42)	-0.637 (5.90)	-0.609 (5.64)		
γ_{bn}	-0.429 (1.28)	0.277 (2.16)	0.243 (1.90)		
γ_{dn}	0.746 (1.79)	0.208 (1.70)	0.220 (1.78)		
σ_v^2	0.036	0.041	0.042	0.062	0.062
σ_u^2	0.000	0.000	0.000	0.000	0.000
$\sqrt{b_1}$	0.099	0.111	0.109	0.155	0.154
$\bar{R}^2$	0.948	0.941	0.997	0.910	0.996
$\ln\mathcal{L}$	54.910	39.724	37.607	-3.858	-4.278