NONPARAMETRIC ANALYSIS OF RETURNS TO SCALE IN THE US HOSPITAL INDUSTRY

PAUL W. WILSONa* AND KATHLEEN CAREYb

a Department of Economics, University of Texas, Austin, USA
b Management Science Group, US Department of Veterans Affairs, Bedford, USA, and School of Public Health, Boston University, USA

SUMMARY
This paper presents new estimates of scale economies for US hospitals. We show that the common translog specification of hospital costs is a misspecification, and employ nonparametric, local linear estimation with both continuous and discrete covariates. A bootstrap method is used to provide inferences regarding ray scale economies and expansion path scale economies for a large sample of hospitals covering 1984–1996. We find evidence of changes in the structure of hospital costs over the sample period, as well as evidence of locally optimal hospital sizes. Copyright © 2004 John Wiley & Sons, Ltd.

1. INTRODUCTION

The last two decades have been a tumultuous period of restructuring within the US hospital industry. Hospitals have been pressured by government and private payers to absorb an increasing share of the financial risk involved with their treatment decisions. Many smaller hospitals have had serious difficulty maintaining viability in this changing environment, leading to a number of hospital closures, particularly during the 1980s. More recently, the US hospital industry has experienced a wave of mergers, acquisitions and other forms of financial consolidation due to continuing financial pressure. This activity peaked in 1996 when it affected 768 facilities, and while on the decline more recently, involved 530 facilities during 1999.1

Merger endeavours among hospitals present particular challenges for antitrust policy. The rise of antitrust enforcement in the hospital industry that began in the late 1970s has led to concern that regulatory action may preclude potential efficiency gains that might be realized through consolidation. In 1992, the Department of Justice and the Federal Trade Commission jointly issued guidelines concerning horizontal mergers aimed at reducing the uncertainty surrounding the likelihood of merger challenges.2 Subsequently, these agencies released extensions of the guidelines directed specifically at the health care industry. Responding to questions over obstructions of efficiency gains, the modified guidelines defined a safe harbour for smaller hospitals3 and identified specific

* Correspondence to: Paul W. Wilson, Department of Economics, University of Texas at Austin, 1 University Station, C3100, Austin, TX 78712, USA. E-mail: wilson@eco.utexas.edu

Modern Healthcare, January 10, 2000, p. 36.

1 Department of Justice and Federal Trade Commission, 1992, Horizontal Merger Guidelines.

considerations by which hospitals can demonstrate that proposed mergers will ultimately result in savings for consumers. Evolution of antitrust standards for the hospital industry continues.

Given the volume of hospital mergers and acquisitions, there is remarkably little consensus on the extent of scale economies in hospitals. This fact may account in part for the unsettled nature of regulation regarding hospital mergers. Early empirical work provides some support for economies of scale among smaller hospitals, but overall results are largely inconclusive, and fail to account for the multidimensional nature of hospital services. More recent studies, while allowing for multiple outputs and offering better controls over case mix, still do not provide a consensus on the nature of scale efficiencies, particularly among larger hospitals.

The recent literature on returns to scale among hospitals has often been criticized for choices of functional form. Vitaliano (1987) demonstrated the sensitivity of the shape of estimated average cost functions to the choice of functional form. He produced conflicting results regarding scale economy estimates using quadratic versus logarithmic forms. Most recent hospital cost function studies employ translog specifications. Although the translog has some desirable properties, its deficiencies for hospital cost function estimation were well demonstrated by Vita (1990). The translog form derives from a second-order Taylor expansion and may approximate costs well in a sample of similar-sized firms. Its global properties are unsatisfactory, however, when used with firms of widely varying size, limiting its usefulness in assessing policies that involve large, incremental changes in output such as mergers and consolidations. An exception to this line of inquiry is the work by Dranove (1998), who used semiparametric techniques to determine the scale effects for non-revenue-producing cost centres. Substantial scale economies were found in small hospitals, but exhausted at a level of 10,000 annual discharges.

This paper employs nonparametric estimation methods to analyse scale economies among a large sample of US hospitals, and how returns to scale may have changed over time. Simple specification tests are shown to reject the conventional parametric, translog specification. Bootstrap methods are used to make inferences about scale economies. The next section discusses measurement of returns to scale and introduces measures that are well-suited to our nonparametric estimation approach. Section 3 describes our data and specifies variables. The formal statistical model is developed in Section 4. In addition, the nonparametric estimation methods, and how the measures defined in Section 2 can be estimated, are discussed in Section 4. Empirical results are discussed in Section 5, and summary and conclusions are given in Section 6.

2. MEASURING RETURNS TO SCALE

Consider a multiple-output cost function \( C(y) \), where \( y \) denotes a vector of output quantities. Returns to scale are frequently measured in terms of elasticities. Product mix remains constant along an arbitrary ray \( \theta y \) emanating from the origin; the elasticity of cost at a point \( y \) along the

---

4 Department of Justice and Federal Trade Commission, 1997, Horizontal Merger Guidelines (with April 8, 1997, Revisions to Section 4 on Efficiencies).

5 For a review of the earlier literature, see Cowing et al. (1983).

6 Among recent studies where hospital cost functions have been estimated, specific measures of returns to scale can be found in Grannemann et al. (1986), Vita (1990), Fournier and Mitchell (1992), Gaynor and Anderson (1995), and Carey (1997).
where \( j \) indexes the different outputs. The elasticity in equation (1) is the multiproduct analogue of marginal cost divided by average cost on a ray from the origin, with \( \eta(y) <, =, > 1 \) implying (increasing, constant, decreasing) returns to scale as output is expanded along the ray from the origin. Hospitals for which \( \eta(y) \neq 1 \) are not competitively viable; if hospitals were subject to the normal rules of competitive behaviour, either a smaller or a larger firm could drive a hospital with \( \eta(y) \neq 1 \) from a competitive market.

The measure defined in equation (1) requires estimation of derivatives of the cost function. We employ fully nonparametric estimation methods in this paper for reasons discussed in Section 4. Since nonparametric estimates of derivatives are typically noisier than similar estimates of the original function,7 we define the ratio

\[
S(y) = \frac{C(\theta y)}{\theta C(y)}
\]

It is straightforward to show that

\[
\frac{dS(y)}{d\theta} \begin{cases} < 0 & \text{if } \theta < 1, \\ > 0 & \text{if } \theta > 1 \end{cases}
\]

i.e., \( S(y) \) is decreasing (constant, increasing) in \( \theta \) if returns to scale are increasing (constant, decreasing) at \( \theta y \) along the ray from the origin passing through \( y \). In addition, \( S(\theta = 1 | y) = 1 \) by definition. Thus, ray scale economies (RSE) along a ray \( \theta y \) can be examined by estimating \( C(y) \) and \( C(\theta y) \) for various values of \( \theta \), and using confidence bands to determine whether \( S(\theta | y) \) is downward or upward sloping.

The RSE measure captures the effect of proportionate increases in outputs, but does not account for variations in output mix among hospitals of different sizes. Output mix may vary among hospitals of different sizes; hospitals may alter their product mix as they expand in terms of size. Hence, we also divide hospitals into groups according to size, and examine returns to scale along paths between typical hospitals in contiguous groups. Consider a pair of output vectors \( (y^a, y^b) \), where \( y^b \approx y^a \). We measure expansion path scale economies (EPSE) along the linear path from \( y^a \) to \( y^b \) in the output space by defining

\[
E(\theta | y^a, y^b) = \frac{C(y^a + \theta(y^b - y^a)) - C(y^a)}{\theta[C(y^b) - C(y^a)]}
\]

for \( \theta \in (0, 1] \).8 Clearly, if \( \theta = 1 \) then \( E(\theta | y^a, y^b) = 1 \). But if \( \theta < 1 \), then the first term in the numerator of equation (4) gives the cost of a hypothetical firm producing at an intermediate point along the linear path from \( y^a \) to \( y^b \). If \( \theta < 1 \) and \( E(\theta | y^a, y^b) > 1 \), then the cost of this hypothetical firm is greater than the weighted costs of two firms producing output vectors \( y^a \) and \( y^b \), given

---

7 This is particularly true for the present case where we would require derivatives in several dimensions; more discussion of this point is provided in Section 4.

8 Both the measures defined by equations (2) and (4) have been used by Wheelock and Wilson (2001) to examine returns to scale among US banks.

by the numerator in (4). This implies that the cost surface is concave from below along the path from \( y^a \) to \( y^b \), which in turn implies that total cost is increasing at a decreasing rate as we move along this path away from \( y^a \) towards \( y^b \). Hence if \( E(\theta | y^a, y^b) > 1 \) for values \( \theta < 1 \), returns to scale are increasing along the expansion path from \( y^a \) to \( y^b \). Similar reasoning demonstrates that if \( E(\theta | y^a, y^b) < 1 \) for values \( \theta < 1 \), decreasing returns to scale prevail along the expansion path from \( y^a \) to \( y^b \). Since \( E(\theta | y^a, y^b) = 1 \) for \( \theta = 1 \) by definition in (4), we need only estimate \( E(\theta | y^a, y^b) \) for small values of \( \theta \) and determine whether these estimates are significantly greater or less than unity.\(^9\) By allowing the product mix to vary as hospitals expand their output, the measure \( E(\theta | y^a, y^b) > 1 \) mixes economies of scale and economies of scope, but may provide a more realistic view of how hospitals expand than measures of RSE.

To implement the analysis, we need merely estimate the measures \( S(\theta | y) \) and \( E(\theta | y^a, y^b) \) by replacing \( C(\cdot) \) on the right-hand sides of (2) and (4) with estimates of cost at the appropriate output levels. Bootstrap methods may be used to make inferences. Details on the variables and data used for estimation are given in the next section, while the estimation methods are discussed in Section 4.

3. A MODEL OF HOSPITAL COST

To estimate the measures of scale economies described in the previous section, a model of hospital costs must be specified. Hospitals use a number of inputs to produce a wide range of services, and in studies of hospital technology researchers are forced to employ simplified models of production due to the lack of reliable data on factor prices.

While the lack of data on factor prices is problematic, for estimation of cost functions, this is true for all large hospital studies. Part of the problem arises from the fact that physicians’ associations with most hospitals are largely unobservable. While physicians’ input is of course important, most physicians are not employees of a hospital, and allocate only part of their time to providing services at one or perhaps several hospitals. Given that there have been a number of hospital cost studies, that these studies have used restrictive parametric functional forms, and that hospital expenditures are a major component of health-care expenditures in the USA, it seems important to make best use of the imperfect data that are available. The issues are simply too important to ignore.

We define three outputs: number of admissions (ADMISS), inpatient days (INDAYS) and outpatient days (OUTDAYS). These are defined on an annual basis, and are typical of output specifications found in studies of hospital costs. Including admissions helps differentiate hospitals with large numbers of inpatient days; some of these might treat the same patients for a long period, while others might treat similar patients for a shorter period, presumably providing more efficient care. In addition, we control for differences across hospitals in the severity of illnesses or injuries treated by including a case-mix index (CASEMIX). Since our data span a 13-year period, we include time (TIME) as an exogenous variable in the cost function.

To allow for possible temporal changes in technology, we include time (TIME) measured as the number of days between 1 January 1960 and the beginning data of a hospital’s fiscal year.

\(^9\) Conceivably, \( E(\theta | y^a, y^b) \) could oscillate around unity for values \( \theta < 1 \), which would suggest both increasing and decreasing returns along different parts of the expansion path. We found no evidence of statistically significant variation about unity in our empirical analysis.
as reported in each of the annual AHA surveys. In the data we use for estimation, there are 244 unique values for TIME, corresponding to dates ranging from 1 December 1982 to 1 June 1996. We control for various environmental differences among hospitals by defining the following binary indicators:

\[
\text{TEACH} = \begin{cases} 
1 & \text{for teaching hospitals} \\
0 & \text{otherwise}
\end{cases}
\]

\[
\text{URBAN} = \begin{cases} 
1 & \text{if hospital is in a metropolitan statistical area (MSA)} \\
0 & \text{otherwise}
\end{cases}
\]

\[
\text{PROFIT} = \begin{cases} 
1 & \text{for for-profit hospitals} \\
0 & \text{otherwise}
\end{cases}
\]

\[
\text{GOVT} = \begin{cases} 
1 & \text{for government-owned hospitals} \\
0 & \text{otherwise}
\end{cases}
\]

Teaching status is determined by whether a hospital is a member of the Council of Teaching Hospitals. The value of URBAN is determined by whether the hospital is located in a metropolitan statistical area as defined by the US Census Bureau.

As with all hospital cost studies that we have seen, we lack extensive information on factor prices. Our data do, however, include the index of local area wage rates used by Medicare for reimbursing hospitals under the Prospective Payment System (PPS). While this measures the price of only a single factor, variation in costs to hospitals for energy and food may be partially reflected in wage rates which must compensate workers for higher costs of living. Moreover, inclusion of the binary variable URBAN may account for some differences in land rents and other factor prices between large urban areas and other locations.

We divide hospitals’ total annual operating expenditures by the wage index to construct our cost variable (COST). Normalizing costs by the wage index ensures homogeneity of the cost function with respect to labour prices.

Most of our data are taken from the American Hospital Association annual survey of hospitals for each year 1984–1996. The wage and the case-mix variables are obtained from the Health Care Financing Administration. Both of the latter measures are used for hospital reimbursement under PPS. After deleting observations with missing or implausible values, we retain 36,963 observations for the roughly 13-year period we examine.

Table I gives numbers of observations (over all years in our study) for each of the 12 categories defined by the discrete variables TEACH, URBAN, PROFIT and GOVT. Hospitals are distributed unevenly over these 12 categories, which has implications for estimation as well as interpretation of results.

4. ESTIMATION METHOD

Having specified variables which affect hospital costs, we must estimate the relation between these variables and hospital costs in order to estimate the RSE and EPSE measures defined in Section 2. Our data have both a cross-sectional as well as a time dimension. We use fully nonparametric estimation methods, and include TIME to control for differences in cost over time. Using subscript \(i\) to index the \(n\) observations in our sample, for each observation let \(y_i\) denote a row vector containing
Table I. Number of observations by category

<table>
<thead>
<tr>
<th>TEACH</th>
<th>URBAN</th>
<th>PROFIT</th>
<th>GOVT</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(b)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(c)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(d)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(e)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(f)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(g)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(h)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(i)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(j)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(k)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(l)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: The last column gives the number of hospitals in each category, determined by teaching status (TEACH), location in an urban area (URBAN), for-profit ownership (PROFIT), or government ownership (GOVT). For example, the 1s and 0s in row (a) indicate teaching hospitals in urban areas that are neither for-profit nor government-owned (i.e., non-profit hospitals).

output quantities ADMISS, INDAYS and OUTDAYS; let $x_i$ denote a row vector containing values of the continuous variables CASEMIX and TIME; and let $w_i$ denote a row vector of length $k(k \in \{0, 1, 2, 3, 4\})$ containing some of the binary dummy variables TEACH, URBAN, PROFIT and GOVT. In addition, let $C_i$ denote $\log($COST$)$ for observation $i$. For notational convenience, let $z_i = [y_i \ x_i]$.

In order to infer scale efficiencies, the mapping $C \leftarrow [z \ w]$ must be estimated. We specify the data-generating process

$$C_i = m(z_i, w_i) + \epsilon_i$$

where $\epsilon_i$ is a stochastic error term with $E(\epsilon_i | z_i, w_i) = 0 \ \forall i = 1, \ldots, n$. The problem then is to estimate the conditional mean function $m(z_i, w_i) = E(C_i | z_i, w_i)$.

Typical studies of hospital costs have used parametric specifications for the conditional mean function; by far the most common choice of functional forms has been the translog specification. For $\ell$ continuous variables and no discrete variables ($k = 0$), the translog specification is

$$m(z_i) = \beta_0 + (\log z_i)\beta + (\log z_i)B(\log z_i)$$

where $\beta_0$ is a scalar parameter, $\beta$ is an $\ell \times 1$ vector of parameters, and $B$ is an $\ell \times \ell$ upper triangular matrix of parameters. Given that this is a quadratic specification, the translog specification is rather restrictive in terms of the variety of shapes for the cost surface admitted by this specification. The translog results from a Taylor expansion of the cost function around the mean of the data; consequently, it seems to make little sense to attempt inference about returns to scale over units of widely varying size using the translog specification. Moreover, the specification in equation (6) is easily rejected by our data.\textsuperscript{10}

\textsuperscript{10}To test the translog specification, we considered three groups of hospitals defined by: (i) TEACH = 0, URBAN = 1, PROFIT = 0, GOVT = 0; (ii) TEACH = 0, URBAN = 0, PROFIT = 0, GOVT = 0; and (iii) TEACH = 0, URBAN = 0, PROFIT = 0, GOVT = 1. We further restricted the analysis to observations for 1996 only, yielding 469, 1150 and
Rejection of the translog functional form is hardly surprising. Several studies have noted the problem: see, for example, Guilkey et al. (1983) and Chalfant and Gallant (1985) for Monte Carlo evidence; Cooper and McLaren (1996) and Banks et al. (1997) for empirical evidence involving consumer demand. Still others have found a similar problem while estimating cost functions for US commercial banks, which, like hospitals, also vary wildly in terms of size; see, for example, McAllister and McManus (1993), Mitchell and Onvural (1996) and Wheelock and Wilson (2001). The problem points to the use of nonparametric estimation methods. Although nonparametric methods are less efficient in a statistical sense than parametric methods when the true functional form is known, nonparametric estimation (asymptotically) avoids any risk of specification error. Moreover, to our knowledge the true functional form remains unknown in the present case.

Fan and Gijbels (1996, ch. 1) and Härdle and Linton (1999) give nice descriptions of nonparametric regression and the surrounding issues. Nonparametric regression models may be viewed as infinitely parameterized; as such, any parametric regression model (such as the translog cost function) is nested within a nonparametric regression model. Clearly, adding more parameters to a parametric model affords greater flexibility. Nonparametric regression models represent the limiting outcome of adding additional parameters.

Several possibilities for nonparametric regression exist. Orthogonal series estimators based on the ideas of Szegő (1959) and Gallant (1981, 1982) involve representing the conditional mean function by an infinite Fourier series and using orthogonal polynomials (e.g., Laguerre or Legendre polynomials) or other functions (e.g., transcendental functions or Muntz–Satz expansions) to represent the basis functions, and have been used in studies of bank costs and elsewhere. One must choose a truncation point for the Fourier series; cross-validation and other methods (e.g., Eastwood, 1991) may be used. As a practical matter, in the present multivariate setting with a large number of observations, this method would incur the numerically challenging problem of inverting very large moment matrices.

Kernel regression offers another possibility for nonparametric estimation of the conditional mean function. The Nadarya–Watson estimator (Nadarya, 1964; Watson, 1964) evaluated at an arbitrary point \( z_0 \) is given by

\[
m(z_0) = \frac{\sum_{i=1}^{n} C_i K(|H|^{-1}(z_i - z_0))}{\sum_{i=1}^{n} K(|H|^{-1}(z_i - z_0))}
\]

with \( n = 641 \) observations, respectively. Thus, each of the three subsamples represent hospitals of a specific type in a single year, removing the need for dummy variables in the translog specification (6) we wish to test. For each subsample, we estimated (6) and recorded the error sum-of-squares, \( ESS_0 \). Next, we divided each subsample into four size groups, depending on number of beds falling into one of the ranges 0–54, 55–115, 116–224 or greater than 224. For each of the original three subsamples, we then estimated (6) using each of the four size groups separately, producing error sums-of-squares \( ESS_j, j = 1, 2, 3, 4 \). Assuming normality in the error terms, the statistic

\[
\hat{F} = \frac{(ESS_0 - ESS_1 - ESS_2 - ESS_3 - ESS_4)/3 K}{(ESS_1 + ESS_2 + ESS_3 + ESS_4)/(N - 4 K)}
\]

is \( F \)-distributed with 45, \( N - 60 \) degrees of freedom, where \( N \) is the number of observations in a particular subsample among (i)–(iii). Estimation with the four subsamples (i)–(iii) described above yields values for \( \hat{F} \) equal to 2.2022, 2.9243 and 1.4693, respectively, with corresponding \( p \)-values of \( 3.1759 \times 10^{-3}, 1.0703 \times 10^{-5} \) and 0.02756. Thus, for each of our ‘homogeneous’ subsamples, we easily reject the null hypothesis of the single translog specification in (6).
for the case of \( \ell \) continuous variables and no discrete variables, where \( K(\cdot) \) is a piecewise continuous multivariate kernel function satisfying \( \int_R K(u)du = 1 \) and \( K(u) = K(-u), u \in R^\ell, \) and \( H \) is an \( \ell \times 1 \) matrix of bandwidths. This estimator is well-known; under mild assumptions, it is weakly consistent and asymptotically normal (see Schuster, 1972). Moreover, several methods (e.g., cross-validation or plug-in methods) exist for choosing the bandwidth or smoothing parameter.

Most consistent nonparametric conditional mean function estimators are asymptotically biased, and the Nadarya–Watson estimator provides no exception. Local linear estimators represent a refinement of the ideas underlying kernel regression, and yield smaller asymptotic bias, but with the same asymptotic variance as the Nadarya–Watson kernel estimator. Local linear estimators are also weakly consistent and asymptotically normal under similar mild assumptions.11

Consider, as before, the case of \( \ell \) continuous variables and no discrete variables. The local linear estimator follows from a first-order Taylor expansion of \( m(z) \) in a neighbourhood of an arbitrary point \( z_0: \)

\[
m(z) \approx m(z_0) + \frac{\partial m(z_0)}{\partial z}(z - z_0)
\]

This suggests estimating the conditional mean function at \( z_0 \) by fitting the locally weighted least-squares regression problem

\[
\min_{a_0,a} \sum_{i=1}^{n} [C_i - a_0 - (z_i - z_0)\alpha]^2 K((H)^{-1}(z_i - z_0))
\]

where \( K(\cdot) \) and \( H \) are as defined earlier, \( a_0 \) is a scalar, and \( \alpha \) is an \( \ell \)-vector. The solution to the least-squares problem is

\[
[\hat{a}_0 \ \hat{a}'] = (Z'\Phi Z)^{-1}Z'\Phi C
\]

where \( C = [C_1 \ldots C_n]' \), \( \Phi = \text{diag}[K((H)^{-1}(z_i - z_0))] \), and \( Z \) is an \( n \times (\ell + 1) \) matrix with \( i \)th row given by \( [1 \ (z_i - z_0)] \). The fitted value \( \hat{a}_0 \) provides an estimate \( \hat{m}(z_0) \) of the conditional mean function at an arbitrary point \( z_0 \).12

Introduction of \( k \) discrete dummy variables into the analysis requires some modification. One possibility is to split the sample into \( 2^k \) subsamples depending on the values of the discrete variables, and then analyse each group separately. Unfortunately, some of the resulting subsamples may be very small—as is the case in our application, as shown by Table I. Moreover, to the extent that each subsample may contain some information that would be useful in estimation on the other subsamples, this approach does not make efficient use of the data.

With the local linear estimator, introduction of dummy variables can be accommodated by use of an augmenting, discrete kernel function. The idea involves smoothing across the \( 2^k \) categories defined by \( k \) dummy variables, and to let the data determine how much smoothing is appropriate. Aitchison and Aitken (1976) discussed use of a discrete kernel for discrimination analysis. Bierens (1987) and Delgado and Mora (1995) suggested augmenting the Nadarya–Watson estimator in

---

11 See Fan and Gijbels (1996) for an extensive discussion and discussion of the properties of local linear estimators.
12 The fitted values in \( \hat{a} \) provide estimates of elements of the vector \( \partial m(z_0)/\partial z \). However, if the object is to estimate first derivatives, the mean square error of the estimates can be reduced by locally fitting a quadratic rather than a linear expression (see Fan and Gijbels, 1996 for a discussion); this increases computational costs, which are already substantial for the local linear fit. Moreover, determining the optimal bandwidths becomes more difficult and computationally more burdensome for estimation of derivatives. See Härdle (1990, pp. 160–162) for a discussion of some of the issues involved with bandwidth selection for derivative estimation.
(7) with a discrete kernel, and proved that the estimator remains consistent and asymptotically normal. Racine and Li (2000) established convergence rates for the Nadarya–Watson estimator with mixed continuous—discrete data; the rate with continuous and discrete covariates is the same as the rate with the same number of continuous variables, but no discrete variables. Thus, introduction of discrete covariates does not exacerbate the curse of dimensionality, at least in the limit. These ideas are easily extended to the local linear estimator; see the Appendix for details.

In order to use a single, scalar bandwidth rather than a matrix of bandwidths \( H \), we transform the continuous covariates to principal-components space, and define \( K(\cdot) \) as a spherically symmetric, \( \ell \)-variante Epanechnikov kernel. The spherically symmetric Epanechnikov kernel is optimal in terms of asymptotic minimax risk; see Fan et al. (1997) for details and a proof. We choose bandwidths using least-squares cross-validation. Again, see the Appendix for technical details.

Once appropriate values of bandwidth parameters have been selected, the conditional mean function can be estimated at any point \( (z_0, w_0) \in \mathbb{R}^\ell \times [0, 1]^k \). We then estimate the RSE and EPSE measures defined in equations (2) and (4) by replacing \( \text{COST} \) with \( \exp(\hat{m}(\cdot)) \) evaluated at the appropriate point. To make inferences about RSE and EPSE, we use the wild bootstrap proposed by Härdle (1990) and Härdle and Mammen (1993). Ordinary bootstrap methods are inconsistent in our context due to the asymptotic bias of the estimator; see Mammen (1992) for additional discussion. The wild bootstrap is used to obtain bootstrap estimates \( \hat{m}_b(\cdot) \), which are substituted into (2) and (4) to obtain bootstrap values \( \hat{S}_b \) and \( \hat{\ell}_b \), respectively, for particular values of \( z \) and \( w \), with \( b = 1, \ldots, B \). For the case of RSE, we sort the values in \( \{(\hat{S}_b - \hat{S})\}_{b=1}^B \) by algebraic value, delete \((\xi \times 100)\% \) of the elements at either end of this sorted array, and denote the lower and upper endpoints of the remaining, sorted array as \( -b^*_a \) and \( -a^*_a \), respectively. Then a bootstrap estimate of a \((1 - \alpha)\% \) confidence interval for \( \hat{S} \) is

\[
\hat{S} + a^*_a \leq \hat{S} \leq \hat{S} + b^*_a \tag{11}
\]

The idea underlying (11) is that the empirical distribution of the bootstrap values \( \hat{S}_b - \hat{S} \) mimics the unknown distribution of \( \hat{S} - \hat{S} \), with the approximation improving as \( n \to \infty \). As \( B \to \infty \), the choices of \( -b^*_a \) and \( -a^*_a \) become increasingly accurate estimates of the percentiles of the distribution of \( \hat{S}_b - \hat{S} \) (we set \( B = 1000 \)). Any bias in \( \hat{S} \) relative to \( \hat{S} \), is reflected in the bias of \( \hat{S} \) relative to \( \hat{S} \). In the case of large bias, it is conceivable that the estimated confidence interval may not include the original estimate \( \hat{S} \), since the estimated confidence interval corrects for the bias in \( \hat{S} \).

5. EMPIRICAL ANALYSIS

5.1. Ray Scale Economies

Once bandwidth(s) have been chosen using least-squares cross-validation as discussed in Section 4, it is straightforward to compute estimates of the RSE and EPSE measures defined in (2) and (4). We first consider the entire sample of 36 963 observations, and ignore the discrete dummy variables defined in Section 3. We include the three output variables ADMISS, INDAYS and OUTDAYS, as well as CASEMIX and TIME. We then compute estimates \( \hat{S}(\theta | y) \) of the RSE measure defined in (2), where \( y \) is a vector containing the sample medians of the output variables. The cost function

\[\text{See Simar and Wilson (2000) for additional discussion of the bootstrap procedure used here.}\]
estimates used to compute $\hat{S}(\theta | y)$ are evaluated at the sample median of CASEMIX in every case, and values of TIME corresponding to June 30 in the first, middle and last complete years covered by our sample (i.e., 1983, 1989 and 1995). Estimates $\hat{S}(\theta | y)$ are computed for various values of $\theta$, i.e., $\theta \in \{0.05, 0.01, 0.15, \ldots, 1.0, 2.0, \ldots, 15.0\}$; $\theta$ scales the median output vector $y$, but not CASEMIX or TIME. Results for 1983, 1989 and 1995 are plotted in Figure 1.14 The vertical axes on the left of each panel in Figure 1 give values of the estimates $\hat{S}(\theta | y)$.

In addition, we used the bootstrap procedure described in the previous section to estimate individual, 95% confidence intervals for $\hat{S}(\theta | y)$ at each value of $\theta$; these interval estimates are represented by shaded bands in Figure 1. Because they are based on individual confidence interval estimates, the shaded bands shown in Figure 1 are narrower than simultaneous 95% confidence bands. However, comparison with Bonferroni confidence interval estimates reveals that the difference is small enough to warrant the conclusions we draw below.

Superimposed on the plots of $\hat{S}(\theta | y)$ in Figure 1 are kernel density estimates of the probability density of number of beds in each hospital divided by the sample median of number of beds; in our sample, number of beds ranges from 8 to 1805, with a median of 120. Dividing number of beds by median beds permits plotting the density estimates and $\hat{S}(\theta | y)$ on the same horizontal axis (e.g., $\theta = 1$ corresponds to 120 beds; $\theta = 0.5$ corresponds to 60 beds; etc.). The density estimates indicate how the data are dispersed, aiding in interpretation of the RSE estimates as seen below. The values of the density estimates are given on the vertical axes on the right in each panel of Figure 1. Number of beds is representative of the size of a given hospital. Finally, note that the horizontal axes in Figure 1 use a log scale; consequently, the density estimates appear not to integrate to unity. This is an illusion, however, due to the compression of horizontal distance on the right.

At $\theta = 1$, we have $\hat{S}(\theta | y) = S(\theta | y) = 1$ by definition, and both the numerator and denominator of $\hat{S}(\theta | y)$ are evaluated at median outputs.15 Figure 1 reveals that in each year, the median hospital size (corresponding to $\theta = 1$) exceeds the modal size; the distribution of hospital sizes is heavily skewed towards smaller hospitals, with a long right tail (even with the log scale on the horizontal axis). In other words, the data become increasingly sparse as we consider larger-sized hospitals.

Differences across time represented by the three panels in Figure 1 are small. In each year, the results suggest increasing returns to scale for hospitals ranging in size from about 50–60% of the median size up to roughly twice the median size. The results also indicate decreasing returns to scale for hospitals beyond about three times median size in each year. The results do not allow us to reject constant returns to scale among hospitals smaller than about 50% of median size in any year (with the possible exception of the anomaly around $\theta = 0.2$ for 1983).

Teaching hospitals may differ from non-teaching hospitals in terms of cost because in addition to inpatient and outpatient treatment, teaching hospitals must also provide medical education to interns and residents. As a group, teaching hospitals are larger and more urban than their non-teaching counterparts, and typically have higher case-mix indices since they must offer opportunities for interns to treat unusual, low-probability cases. We checked for differences between teaching and non-teaching hospitals by repeating the previous, original analysis while including the dummy

---

14 We follow this convention throughout this section, although we computed estimates corresponding to midpoints of the 13 complete years in our sample. Changes from year to year are less dramatic than over six-year periods; we omit the intervening years to save space.

15 Since $\hat{S}(\theta | y) = S(\theta | y) = 1$ by definition at $\theta = 1$, confidence intervals and their estimates necessarily have zero width at this point.
variable TEACH defined in Section 3. Results for the RSE estimates, together with kernel density estimates for the density of size (measured as before), are shown in Figure 2 for teaching hospitals and in Figure 3 for non-teaching hospitals.

Our sample contains 2156 observations on teaching hospitals (5.8% of the full sample). The density plots in Figure 2 reveal that teaching hospitals are typically large hospitals; in our sample, number of beds for teaching hospitals ranges from 121 to 1805, while that for non-teaching hospitals ranges from 8 to 1790. Unsurprisingly, the left-hand portions of the panels in Figure 2
are identical to the corresponding parts of Figure 1—there are no teaching hospitals observed in this region, and our use of a local estimator ensures that these parts of the plots for \( \hat{S}(\theta | y) \) are identified only by (the same) non-teaching hospitals in both Figures 1 and 2. Nonetheless, Figure 2 suggests that in 1983, teaching hospitals faced constant returns to scale for \( \theta = 1 \) to about \( \theta = 3 \) (i.e., 120–360 beds), and decreasing returns for larger sizes. For 1989 and 1995, however, there is evidence of increasing returns to scale for teaching hospitals in the range \( \theta \in [1, 3] \), with decreasing returns again for larger sizes.
Conclusions for non-teaching hospitals based on the results plotted in Figure 3 are similar to those we discussed for the full sample. In addition, over the range $\theta \in [1, 3]$, returns to scale for non-teaching hospitals are qualitatively similar to those for teaching hospitals. Thus, while the two subsamples are apparently quite different in terms of the structure of costs as indicated by our value of $\lambda$ near unity, returns to scale are qualitatively similar across the two groups.
5.2. Expansion Path Scale Economies

As noted in Section 2, RSE considers changes in cost along a single ray, implicitly assuming that output mix does not change as size increases. Because hospitals of different sizes have varying output mixes, RSE may be misleading. Thus, we also estimated EPSE by dividing our sample into deciles based on size (i.e., number of beds), and used estimates of $E(\theta | y^a, y^b)$ defined in equation (4) to examine returns to scale along a path between median output vectors in adjacent deciles.\(^{16}\) As before, we set CASEMIX equal to the sample median in this analysis, and let TIME correspond to the midpoints of the first, middle and last complete years represented in our data. Interest lies in whether $E(\theta | y^a, y^b)$ is greater than or less than one for small values of $\theta$; we thus set $\theta = 0.1$ for estimation purposes.

Results for the original case with no discrete dummies are shown in the column labelled ‘All Hospitals’ in Table II. The last two columns in Table II show EPSE estimates based on inclusion of the teaching dummy, TEACH. The column labelled ‘Teaching Hospitals’ gives estimates conditioned on TEACH = 1, while estimates in the column labelled ‘Non-teaching Hospitals’ are conditioned on TEACH = 0. Note that estimates in these two columns are identical for all but the last three pairs of size deciles, due to the fact that all teaching hospitals are large hospitals. As expected from the preceding discussion of RSE estimates, the results in Table II are qualitatively similar across the three columns; patterns of significance in the various estimates are also similar across the columns. For each pair of deciles where a significant estimate of EPSE is obtained, a downward arrow ($\downarrow$) is shown for estimates greater than one (indicating that hospitals in the smaller decile should expand towards the larger decile), and an upward arrow ($\uparrow$) is shown for estimates less than one (indicating that hospitals in the larger decile should contract towards the smaller decile).

Our estimates of EPSE shown in Table II indicate increasing returns to scale for hospitals beginning in the fourth decile in terms of size, continuing to the largest hospitals. These results contrast with our results for RSE; recall from the earlier discussion that our analysis of RSE indicated that decreasing returns begin at about twice the median size. The analysis of EPSE is a refinement over RSE in the sense that with EPSE we follow the variation in output mix from one decile to the next, where RSE holds output mix at a constant level determined by the median hospital, which may be rather different from the median hospital in any particular size decile. Given that we are using a local estimator, there may be simply too few hospitals along the path where we examined RSE to give a reliable picture of scale economies.

We next examined expansion path scale economies while including each of the four dummy variables TEACH, URBAN, PROFIT and GOVT to control not only for teaching/non-teaching, but urban location and ownership structure as well. Because they are of less interest, and because of space constraints, we omit plots showing estimates of RSE when we include the four dummy variables, and focus instead on estimates of EPSE. Table III displays EPSE estimates for the eight cases labelled (a), (b) and (g)–(l) in Table I, where sufficient observations are available to give meaningful estimates. The estimates differ quantitatively both across time and across groups, but are qualitatively similar. The results here are broadly similar to those in Table II; in particular, we again find increasing returns to scale for the largest half of all hospitals in each year.

---

\(^{16}\) In terms of the notation in (4), we let $y^a$ represent the median output vector of a decile, and let $y^b$ represent the median output vector of the next decile. We repeat this for each of nine pairs of adjacent deciles.
### Table II. Expansion path scale economy estimates

<table>
<thead>
<tr>
<th>Bed size group</th>
<th>All hospitals</th>
<th>Teaching hospitals</th>
<th>Non-teaching hospitals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–36/37–50</td>
<td>1.0414</td>
<td>1.0546</td>
<td>1.0456</td>
</tr>
<tr>
<td>37–50/51–69</td>
<td>1.0279</td>
<td>1.0534</td>
<td>1.0534</td>
</tr>
<tr>
<td>51–69/70–93</td>
<td>1.0918</td>
<td>1.1948</td>
<td>1.1948</td>
</tr>
<tr>
<td>70–93/94–120</td>
<td>3.0675** ↓</td>
<td>3.0341** ↓</td>
<td>3.0341** ↓</td>
</tr>
<tr>
<td>94–120/121–160</td>
<td>3.4013** ↓</td>
<td>3.3822** ↓</td>
<td>3.3822** ↓</td>
</tr>
<tr>
<td>121–160/161–213</td>
<td>1.5799</td>
<td>1.5665</td>
<td>1.5665</td>
</tr>
<tr>
<td>161–213/214–280</td>
<td>1.8556</td>
<td>1.8016** ↓</td>
<td>1.8593* ↓</td>
</tr>
<tr>
<td>214–280/281–400</td>
<td>1.7018</td>
<td>2.1886* ↓</td>
<td>1.7445</td>
</tr>
<tr>
<td>281–400/ ≥ 401</td>
<td>2.2844** ↓</td>
<td>2.1162** ↓</td>
<td>2.4718** ↓</td>
</tr>
<tr>
<td>1989</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–36/37–50</td>
<td>1.2298</td>
<td>1.2299</td>
<td>1.2299</td>
</tr>
<tr>
<td>37–50/51–69</td>
<td>0.9076** ↑</td>
<td>0.9273** ↑</td>
<td>0.9273** ↑</td>
</tr>
<tr>
<td>51–69/70–93</td>
<td>1.3125</td>
<td>1.3560</td>
<td>1.3560</td>
</tr>
<tr>
<td>70–93/94–120</td>
<td>3.4262** ↓</td>
<td>3.4528** ↓</td>
<td>3.4528** ↓</td>
</tr>
<tr>
<td>94–120/121–160</td>
<td>2.9536** ↓</td>
<td>2.9220** ↓</td>
<td>2.9220** ↓</td>
</tr>
<tr>
<td>121–160/161–213</td>
<td>1.6403</td>
<td>1.6831</td>
<td>1.6831</td>
</tr>
<tr>
<td>161–213/214–280</td>
<td>2.2693** ↓</td>
<td>2.2489** ↓</td>
<td>2.2668** ↓</td>
</tr>
<tr>
<td>214–280/281–400</td>
<td>1.7952** ↓</td>
<td>2.2512** ↓</td>
<td>1.8085* ↓</td>
</tr>
<tr>
<td>281–400/ ≥ 401</td>
<td>2.7952** ↓</td>
<td>2.6427** ↓</td>
<td>2.9439** ↓</td>
</tr>
<tr>
<td>1983</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–36/37–50</td>
<td>1.0343</td>
<td>1.0410</td>
<td>1.0410</td>
</tr>
<tr>
<td>37–50/51–69</td>
<td>1.3891</td>
<td>1.4150</td>
<td>1.4150</td>
</tr>
<tr>
<td>51–69/70–93</td>
<td>0.8846* ↑</td>
<td>1.0087</td>
<td>1.0087</td>
</tr>
<tr>
<td>70–93/94–120</td>
<td>2.7284** ↓</td>
<td>2.6911** ↓</td>
<td>2.6911** ↓</td>
</tr>
<tr>
<td>94–120/121–160</td>
<td>3.5949** ↓</td>
<td>3.4809** ↓</td>
<td>3.4809** ↓</td>
</tr>
<tr>
<td>121–160/161–213</td>
<td>1.9724** ↓</td>
<td>1.9813** ↓</td>
<td>1.9813** ↓</td>
</tr>
<tr>
<td>161–213/214–280</td>
<td>2.0829** ↓</td>
<td>1.9410** ↓</td>
<td>2.0436** ↓</td>
</tr>
<tr>
<td>214–280/281–400</td>
<td>2.1507** ↓</td>
<td>1.3535* ↓</td>
<td>2.2846* ↓</td>
</tr>
<tr>
<td>281–400/ ≥ 401</td>
<td>2.9295** ↓</td>
<td>2.6957** ↓</td>
<td>2.9481** ↓</td>
</tr>
</tbody>
</table>

Asterisk (*) indicates significant difference from 1.0 at 90%; double asterisk (**) indicates significant difference from 1.0 at 95%. Where estimates are significant at greater than 90%, upward arrows (↑) indicate decreasing returns to scale, while downward arrows (↓) indicate increasing returns to scale.

### 6. SUMMARY AND CONCLUSIONS

The consensus derived from previous hospital cost functions holds that scale economies are exhausted at a size of 200 to 300 beds. This belief, however, is based on parametrically specified cost functions, often estimated using cross-sectional data drawn from a single state. These results generally have relied on the translog functional form.

Results from parametric analyses depend crucially on the assumption of functional form for the conditional mean function. Although the translog specification is typically described as ‘flexible’, it is anything but that—it is, in the end, simply a quadratic in logs. Moreover, our simple specification test easily rejects the translog form. Nonparametric estimation permits the data to speak for themselves; we have imposed minimal a priori assumptions on the form of the conditional mean function. Similarly, we have not made distributional assumptions on the error term, other than that of zero expectation. Consequently, we have (asymptotically) avoided virtually any risk of specification error, although we have incurred the cost of decreased statistical efficiency relative...
Table III. Expansion path scale economy estimates

<table>
<thead>
<tr>
<th>Bed size group</th>
<th>(a)</th>
<th>(b)</th>
<th>(g)</th>
<th>(h)</th>
<th>(i)</th>
<th>(j)</th>
<th>(k)</th>
<th>(l)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–36/37–50</td>
<td>0.8027</td>
<td>0.8804</td>
<td>0.8027</td>
<td>0.8804</td>
<td>1.1191</td>
<td>1.2811</td>
<td>1.0293</td>
<td>1.5448</td>
</tr>
<tr>
<td>37–50/51–69</td>
<td>2.6834</td>
<td>0.9530</td>
<td>2.6834</td>
<td>0.9530</td>
<td>2.3965</td>
<td>1.2782</td>
<td>1.1716</td>
<td>2.1444</td>
</tr>
<tr>
<td>51–69/70–93</td>
<td>1.9007*</td>
<td>0.8804</td>
<td>1.9007*</td>
<td>0.8804</td>
<td>2.0433</td>
<td>1.0817</td>
<td>1.2282</td>
<td>2.5387</td>
</tr>
<tr>
<td>70–93/94–120</td>
<td>3.8903**</td>
<td>3.0350**</td>
<td>3.8903**</td>
<td>3.0350**</td>
<td>3.2549**</td>
<td>2.5247**</td>
<td>1.1019</td>
<td>3.7098</td>
</tr>
<tr>
<td>94–120/121–160</td>
<td>4.4488**</td>
<td>3.6347**</td>
<td>4.4488**</td>
<td>3.6347**</td>
<td>1.9321*</td>
<td>2.5705**</td>
<td>1.6340**</td>
<td>3.2552**</td>
</tr>
<tr>
<td>121–160/161–213</td>
<td>2.4158**</td>
<td>1.5807**</td>
<td>2.4168**</td>
<td>1.5813**</td>
<td>2.0962*</td>
<td>1.3393**</td>
<td>1.9802**</td>
<td>0.8441*</td>
</tr>
<tr>
<td>161–213/214–280</td>
<td>1.7811**</td>
<td>2.0466**</td>
<td>2.0631**</td>
<td>2.1149**</td>
<td>2.6462**</td>
<td>1.2915**</td>
<td>3.0243**</td>
<td>1.1238**</td>
</tr>
<tr>
<td>214–280/281–400</td>
<td>1.3663</td>
<td>1.6462</td>
<td>2.1135**</td>
<td>1.7810</td>
<td>1.9038</td>
<td>2.3664**</td>
<td>3.8780**</td>
<td>2.2452**</td>
</tr>
<tr>
<td>281–400/401</td>
<td>2.5798**</td>
<td>3.1726**</td>
<td>2.9512**</td>
<td>3.3614**</td>
<td>2.0536**</td>
<td>2.2567**</td>
<td>3.0321**</td>
<td>1.5677**</td>
</tr>
<tr>
<td>1989</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–36/37–50</td>
<td>0.8210*</td>
<td>0.7819*</td>
<td>0.8210*</td>
<td>0.7819*</td>
<td>1.1666</td>
<td>1.3731</td>
<td>0.9957*</td>
<td>1.9484</td>
</tr>
<tr>
<td>37–50/51–69</td>
<td>1.8255</td>
<td>1.1258</td>
<td>1.8255</td>
<td>1.1258</td>
<td>1.4689</td>
<td>1.2120</td>
<td>1.1827</td>
<td>1.8642</td>
</tr>
<tr>
<td>51–69/70–93</td>
<td>1.6427</td>
<td>0.8499</td>
<td>1.6427</td>
<td>0.8499</td>
<td>1.2841</td>
<td>0.8840</td>
<td>1.7622*</td>
<td>2.8755</td>
</tr>
<tr>
<td>70–93/94–120</td>
<td>3.7765**</td>
<td>2.1692**</td>
<td>3.7765**</td>
<td>2.1692**</td>
<td>5.4978**</td>
<td>3.0581**</td>
<td>1.9199</td>
<td>4.9792**</td>
</tr>
<tr>
<td>94–120/121–160</td>
<td>3.2433**</td>
<td>2.3428**</td>
<td>3.2433**</td>
<td>2.3428**</td>
<td>3.1117**</td>
<td>2.3988**</td>
<td>1.3727**</td>
<td>3.3629**</td>
</tr>
<tr>
<td>121–160/161–213</td>
<td>2.1041*</td>
<td>1.8885</td>
<td>2.1061*</td>
<td>1.9202</td>
<td>1.8269</td>
<td>1.5276</td>
<td>1.5818</td>
<td>1.5849</td>
</tr>
<tr>
<td>161–213/214–280</td>
<td>2.1344**</td>
<td>2.3136**</td>
<td>2.2626**</td>
<td>2.4715**</td>
<td>2.6256**</td>
<td>1.7359**</td>
<td>2.4541**</td>
<td>1.5466**</td>
</tr>
<tr>
<td>214–280/281–400</td>
<td>1.9388**</td>
<td>2.1334**</td>
<td>1.9986**</td>
<td>1.8138</td>
<td>1.7995</td>
<td>2.0613**</td>
<td>2.2621**</td>
<td>2.1578**</td>
</tr>
<tr>
<td>281–400/401</td>
<td>2.5223**</td>
<td>2.7766**</td>
<td>2.9260**</td>
<td>3.1498**</td>
<td>1.9046**</td>
<td>2.1269**</td>
<td>2.6043**</td>
<td>1.0071**</td>
</tr>
<tr>
<td>1983</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–36/37–50</td>
<td>0.9493</td>
<td>0.9258</td>
<td>0.9493</td>
<td>0.9258</td>
<td>0.9479</td>
<td>1.1544</td>
<td>1.0457</td>
<td>1.7546**</td>
</tr>
<tr>
<td>37–50/51–69</td>
<td>1.5136</td>
<td>2.5449**</td>
<td>1.5136</td>
<td>2.5449**</td>
<td>2.3006**</td>
<td>1.1240**</td>
<td>1.2387**</td>
<td>2.0807**</td>
</tr>
<tr>
<td>51–69/70–93</td>
<td>1.5514</td>
<td>1.5356</td>
<td>1.5514</td>
<td>1.5356</td>
<td>1.7386</td>
<td>1.2626</td>
<td>1.6150</td>
<td>2.0559**</td>
</tr>
<tr>
<td>70–93/94–120</td>
<td>2.7708</td>
<td>1.6968</td>
<td>2.7708*</td>
<td>1.6968</td>
<td>5.5039**</td>
<td>2.2668</td>
<td>1.8339</td>
<td>5.0209**</td>
</tr>
<tr>
<td>94–120/121–160</td>
<td>2.9533**</td>
<td>2.7841**</td>
<td>2.9533**</td>
<td>2.7841**</td>
<td>3.3535**</td>
<td>2.9512**</td>
<td>1.7710**</td>
<td>3.4796**</td>
</tr>
<tr>
<td>121–160/161–213</td>
<td>1.7788</td>
<td>1.3935</td>
<td>1.7832**</td>
<td>1.4224</td>
<td>1.1661</td>
<td>1.3675</td>
<td>0.8452</td>
<td>0.6719</td>
</tr>
<tr>
<td>161–213/214–280</td>
<td>1.7485**</td>
<td>1.8387</td>
<td>1.8061**</td>
<td>2.0397</td>
<td>1.3024</td>
<td>1.4210</td>
<td>0.8171</td>
<td>1.2588</td>
</tr>
<tr>
<td>214–280/281–400</td>
<td>2.3114</td>
<td>1.9699</td>
<td>2.2231</td>
<td>1.6045</td>
<td>0.6962*</td>
<td>2.0150</td>
<td>1.3430</td>
<td>1.3784</td>
</tr>
<tr>
<td>281–400/401</td>
<td>2.3465**</td>
<td>2.5506**</td>
<td>2.5283**</td>
<td>3.0560**</td>
<td>1.3329**</td>
<td>1.7594**</td>
<td>2.2603**</td>
<td>0.4819</td>
</tr>
</tbody>
</table>

Asterisk (*) indicates significant difference from 1.0 at 90%; double asterisk (**) indicates significant difference from 1.0 at 95%. Where estimates are significant at greater than 90%, upward arrows (↑) indicate decreasing returns to scale, while downward arrows (↓) indicate increasing returns to scale.

Our results differ from those in the existing literature on hospital costs. We find evidence of increasing returns to scale among hospitals above the median size, extending to the largest decile in terms of size, in our analysis of EPSE. The past decade has seen a large number of hospital mergers, including mergers between large hospitals. Our results suggest that these hospitals are exploiting economies of scale. The precise nature in which these economies are exploited may involve both the technology of hospital-based health care, which increasingly relies on expensive capital equipment, as well as increased bargaining power with third-party payers and pharmaceutical companies. Our results also suggest that providing hospital care in small, community-based hospitals remains unknown.

hospitals is expensive and inefficient, and provide justification for consolidating health-care services in larger, regional hospitals.

APPENDIX

For \( k \) dummy variables, let \( w_0, w_i \in \{0, 1\}^k \) be row vectors, and let \( \delta(w_0, w_i) = (w_0 - bw_i)(w_0 - bw_i)' \). Next, define the discrete kernel

\[
G(w_0|w_i, \lambda) \equiv \lambda^{k-\delta(w_0,w_i)}(1 - \lambda)^{\delta(w_0,w_i)}
\]

for \( \lambda \in \left[ \frac{1}{2}, 1 \right] \).

Given \( \ell \) continuous variables in \( z \) and \( k \) discrete dummy variables in \( w \), the local linear estimator of the conditional mean function evaluated at an arbitrary point \( (z_0, w_0) \) is obtained by multiplying the minimand in (9) by \( G(w_0|w_i, \lambda) \), which is equivalent to redefining \( \Phi \) in (10) as

\[
\Phi = \text{diag}[K(H^{-1}(z_i - z_0))G(w_0|w_i, \lambda)]
\]

The matrix \( Z \) defined in Section 4 remains unchanged.

Note that \( \lim G(w_0|w_0, \lambda) \) equals either 1 or 0, depending on whether \( w_0 = w_i \) or \( w_0 \neq w_i \), respectively. In this case, estimation yields the same results as would be the case if the sample were split into \( 2^k \) groups suggested by the dummy variables, with estimation performed independently on each of the \( 2^k \) subsamples. Alternatively, if \( \lambda = \frac{1}{2} \), then \( G(w_0|w_1, \lambda) = 1 \) regardless of whether \( w_0 = w_i \) or \( w_0 \neq w_i \); in this case, there is complete smoothing over the \( 2^k \) categories, and including the dummy variables has no effect relative to the case where they are ignored.

For \( \lambda \in \left( \frac{1}{2}, 1 \right) \), some, but not all, information in each group is used in estimating the conditional mean function for other groups. Cross-validation methods may be used to let the data determine an appropriate value for the smoothing parameter \( \lambda \) on the interval \( \left[ \frac{1}{2}, 1 \right] \).

Given \( 2^k \) groups determined by \( k \) dummy variables, one could employ a different smoothing parameter for each of the \( 2^k(2^k - 1)/2\) \( k \) pairs of groups, but this would result in prohibitive computational costs. The matrix of bandwidths \( H \) presents similar difficulties for computation. We avoid this by transforming the continuous data before performing estimation by pre-whitening. This is accomplished by first computing the \( \ell \times \ell \) matrix \( E \) whose columns are the eigenvectors of the sample correlation matrix for the continuous data. Then, partitioning \( Z \) so that \( Z = [i_n \ Z_1] \), where \( i_n \) is an \( n \times 1 \) vector of 1s, we compute

\[
T = [i_n \ Z_1] \begin{bmatrix} 1 & 0' \\ 0 & E \end{bmatrix}
\]

\[
= [1 \ Z_1 E]
\]

The columns of \( Z_1 E \) contain the principal components of \( Z_1 \); these are orthogonal and have unit variance. Orthogonality eliminates the need for nonzero bandwidths in the off-diagonal elements of the bandwidth matrix \( H \), and identical scale in each dimension eliminates the need for different bandwidths along the diagonal of \( H \). We now have only two bandwidth parameters: \( h \) and \( \lambda \).

We replace (10) with

\[
[\widehat{\alpha}_0 \ \widehat{\alpha}'] = (T'E'T)^{-1}T'E'C
\]

\[
\]
where

\[ \Psi = \text{diag}[K(|H|^{-1}(z_i - z_0)E)G(w_0|w_i, \lambda)] \]  \hspace{1cm} (A.5)

As before, \( \hat{m}_0 \) provides an estimate of the conditional mean function evaluated at the arbitrary point \((z_0, w_0)\). The transformations we use here are similar to those suggested by Silverman (1986), Scott (1992), and others.

After the continuous data have been transformed to principal components, we are able to employ the \( \ell \)-variate spherically symmetric Epanechnikov kernel

\[ K(u) = \frac{\ell(\ell + 2)}{2S_\ell}(1 - uu')(I(uu' \leq 1)) \]  \hspace{1cm} (A.6)

where \( S_\ell = 2\pi^{\ell/2}/\Gamma(\ell/2) \), \( \Gamma(\cdot) \) denotes the gamma function, \( u = h^{-1}(z_i - z_0)E \), and \( I(\cdot) \) denotes the indicator function. The spherically symmetric Epanechnikov kernel is optimal in terms of asymptotic minimax risk; see Fan et al. (1997) for details and a proof.

Although we select only a single, constant (global) bandwidth \( \lambda \) for the discrete data, we use an adaptive, nearest-neighbour bandwidth for the continuous data. We define \( h \) for any particular point \( z_0 \in \mathbb{R}^\ell \) as the maximum Euclidean distance between \( z_0 \) and the \( \kappa \) nearest points in the observed sample \( \{z_i\}_{i=1}^n \), \( \kappa \in \{2, 3, 4, \ldots\} \). Thus, the bandwidth \( h \), determined by \( \kappa \), varies depending on the density of the continuous explanatory variables locally around the point at which the conditional mean function is estimated. This results in a relatively large value for \( h \) where the data are sparse (and where more smoothing is required), and smaller values of \( h \) in regions where the data are relatively dense (where less smoothing is needed).

Note that we are not using a nearest-neighbour estimator, but rather a nearest-neighbour bandwidth. Our bandwidth is used inside a kernel function, and the kernel function integrates to unity. This approach was used by Loftsgaarden and Quesenberry (1965) in the density estimation context to avoid nearest-neighbour density estimates (as opposed to bandwidths) that do not integrate to unity (see Pagan and Ullah, 1999, pp. 11–12 for additional discussion). Fan and Gijbels (1994; 1996, pp. 151–152) discuss nearest-neighbour bandwidths in the regression context.

We choose \( \kappa \) and \( \lambda \) by minimizing the least-squares cross-validation function

\[ \text{CV}(\kappa, \lambda) = \sum_{i=1}^{n} [C_i - \hat{m}_{-i}(z_i)]^2 \]  \hspace{1cm} (A.7)

with respect to \( \kappa \) and \( \lambda \), where \( \hat{m}_{-i}(z_i) \) is computed the same way as \( \hat{m}(z_i) \), except with the \( i \)th diagonal element of \( \Psi \) replaced with zero. The least-squares cross-validation function approximates the part of the mean integrated square error that depends on the bandwidths.\(^{17}\)

**ACKNOWLEDGEMENTS**

This project was supported by a grant of computational time from the Texas Advanced Computing Center at the University of Texas at Austin. We are grateful to Kent Milfeld for technical assistance on the Cray T3E system used for estimation, and to Willard Manning and two anonymous referees for comments on a previous version of the manuscript; as usual, any remaining errors/omissions are solely our responsibility.

\(^{17}\) Choice of \( \kappa \) by cross-validation has been proposed by Fan and Gijbels (1996).
REFERENCES


