Semi-Parametric Regression Models and Connecticut's Hospital Costs

Jeffrey P. Cohen, Ph.D. (corresponding author)
Associate Professor of Economics
University of Hartford
West Hartford, Connecticut 06117 USA
860-768-4834 (Voice)
860-768-4911 (FAX)
professorjeffrey@gmail.com

Jeffrey P. Osleeb, Ph.D.
Professor and Head
Department of Geography
University of Connecticut
Storrs, Connecticut USA
Jeffrey.osleeb@uconn.edu

Ke Yang, Ph.D.
Assistant Professor of Economics
University of Hartford
West Hartford, CT 06117 USA
Kyang@hartford.edu

February 14, 2013

Abstract: The application of spatial analysis in assessing hospital costs has been largely ignored but is deserving of attention. Proximity to other hospitals can lead to spatial spillovers, and recognizing spatial effects can impact hospital economies of scale estimates. In this paper we estimate a variety of cost function models, using annual data for each of Connecticut's 30 hospitals over a 10 year time period, and allow for spatial effects. We consider a variety of semi-parametric regression models as in McMillen and Redfearn (2010). One innovation is that we address both the space and time dimensions in the kernel weights of our panel data semi-parametric regression models. This approach also allows for a general functional form. We find that including a life expectancy measure for years above average lifespan has a negative and significant effect on hospital costs. Finally, we also address potential endogeneity of the life expectancy variable through an instrumental variables estimation approach for panel data semi-parametric models, as first suggested more generally by Baltagi and Li (2002). Monte Carlo simulations indicate our estimator performs well. When addressing the endogeneity with this instrumental variables semi-parametric regression model, the elasticities of scale estimates are smaller but still significant. We also find the hospital cost savings for each year of patients' years above average life expectancy is approximately $4,700 to $7,300 on average, depending on the choice of bandwidth. This life expectancy cost reduction ranges from as low as approximately $480 to as high as $35,000, varying by individual hospitals and by year.

Introduction

Hospitals in the U.S. are situated in both urban and rural areas. Often, there are clusters of hospitals in urban areas but the rural hospitals are spread apart. The spatial nature of hospital locations can affect economies of scale estimates for hospitals. An understanding of economies of scale for hospitals in the U.S. is important because with federal health care reform, greater numbers of Medicaid and uninsured individuals are expected to seek medical treatment. So any information for policy makers on which hospitals are operating most efficiently can be helpful in efforts to decide which hospitals to direct federal and state-level funding. Since focusing exclusively on client counts rather than on the success rates of hospital treatments can be misguided, addressing how improving the “outcomes” of hospital treatment impacts hospital costs is also crucial. Our approach to addressing all of these issues in a hospital cost function model is estimation of semi-parametric regression models, including an instrumental variables specification for an endogenous outcomes variable. We also add to the semi-parametric regression literature with our focus on a weighting kernel that includes a space and a time component.

Cohen et al (2010), Cohen and Morrison Paul (2008), and Li and Rosenman (2001) discuss the broader literature on hospital cost functions. Early studies, such as Carey and Stefos (1992), without relying on economic theory in their cost function estimations focus on a linear functional form that appends quadratic and cubic terms but no interaction terms. Some of the early studies find evidence of internal economies of scale, but follow-up papers, such as Li and

Rosenman (2001), show the importance of allowing for non-linearities that are precluded by using the Cobb-Douglas functional form.

Cohen and Morrison Paul (2008) model hospitals spatial effects through a shift variable in the cost function. They argue that hospitals in urban areas that are clustered together have better access to labor markets, which implies that it may be easier to recruit skilled workers (such as nurses and/or physicians) who are already located nearby. Such an effect, described by O'Sullivan (2010) as labor market pooling, was found to be a significant determinant of costs among the 92 hospitals in the State of Washington that Cohen and Morrison Paul examined. Bates and Santerre (2005) focus on a production function for metropolitan statistical areas in the U.S., and find significant evidence of agglomeration among the hospitals in these areas. Both of these studies modeled the spatial phenomenon directly in the cost or production functions. But imposing such structure on the cost or production function may be considered arbitrary, and for this reason we explore how allowing more general consideration of spatial effects may impact costs.

Little attention has been given to the estimation of spatial econometric models or semi-parametric approaches for hospitals, despite the close link between agglomeration and the spatial locations of hospitals. Exceptions include Mobley et al (2009), although they examine competition among hospitals and also allow for an exogenous spatial shift variable; and Moscone and Knapp (2005), who use a spatial econometrics model to examine mental health expenditures in the UK, although their analysis is at the metropolitan statistical

area rather than the provider level. An alternative approach is to use a non-parametric or semi-parametric technique, which allows greater flexibility without imposing a restrictive functional form as is done with parametric techniques. Wilson and Carey (2005) follow a non-parametric approach. Their rationale is that a priori functional form assumptions made by other researchers (such as Cobb-Douglas, Generalized Leontief, or translog) are arbitrary and unnecessarily impose restrictions on the estimation.

Approach

The analysis we perform is on 30 hospitals over a 10 year period from 1999-2008 in the state of Connecticut, a small prosperous state in the northeastern U.S. Figure 1 shows the locations of the 30 hospitals as well as the population densities. It can be seen that in the larger cities such as Hartford and New Haven, there are clusters of several hospitals, while in the more rural areas of the state the concentration of hospitals is relatively sparse. This variation in hospital locations motivates our analysis of spatial aspects of hospital costs in Connecticut.

We focus on analyzing a hospital cost function, using several approaches to control for spatial heterogeneity with the semi-parametric model, so a more nonlinear specification would be redundant.

In a short-run cost model, the capital stock is assumed fixed, and thus variable costs depend on a vector of input prices, **p** (here, wages and the price of other non-capital inputs); and a vector of shift variables (**R**). Here, these shift

factors include the fixed factor, capital (K);[1] a Case-Mix variable, CMI (or "CASE_MIX"); percent of total patient days that are Medicare patient days, MEDICARE_DAYS; percent of total patient days that are Medicaid patient days, MEDICAID_DAYS; number of total inpatients, Y_{INPAT} ; and number of total outpatients, Y_{OUTPAT}. Other shift variables we include in some of the models (in the vector R) are a time trend (YEAR), and one or more life expectancy variables, such as the years above average lifespan overall (LIFE), or several different specific causes of death. Since we are estimating a short-run model (where the capital stock is a fixed factor), C represents variable costs. The Cobb-Douglas variable cost function (without the life expectancies) takes the form of:

$$C=c(\mathbf{p}, \mathbf{R})=(P_1)^{\beta_1}(P_2)^{\beta_2}(K)^{\beta_3}(CMI)^{\beta_4}(MCARE)^{\beta_5}(MCAID)^{\beta_6}(Y_{INPAT})^{\beta_7}(Y_{OUTPAT})^{\beta}\exp^{u} \quad (1)$$

At this point, we assume that u is iid with mean zero and constant variance. Such a model can be transformed by taking the natural log of both sides, yielding an estimation equation of:

$$\begin{aligned} \text{Log}C_{i,t} = {} & \beta_1 \log(P_{1,i,t}) + \beta_2 \log(P_{2,i,t}) + \beta_3 \log(K_{i,t}) + \beta_4 \log(CMI_{i,t}) + \beta_5 \log(MCARE_{i,t}) \\ & + \beta_6 \log(MCAID_{i,t}) + \beta_7 \log(Y_{INPAT\,i,t}) + \beta_8 \log(Y_{OUTPAT\,i,t}) + u_{i,t}\,. \quad (2) \end{aligned}$$

We begin by estimating this variable cost function model in (2) by OLS, assuming $u_{i,t}$ is iid with mean zero and constant variance, and zero covariances among and across i and t observations.

Since we have data on 30 hospitals (i=1,2,…,30) for each of 10 years (t=1,2,…,10), an alternative, more general approach is a semi-parametric regression model, as in McMillen and Redfearn (2010). We control for spatial

[1] Using a capital stock measure, opposed to a capital price, is appropriate for a short-run cost function model, as has been done in the literature, including by Cohen and Morrison Paul (2011).

effects using this approach, and we also address the issue of internal economies of scale. An additional contribution is that we allow for life expectancy variables to enter the cost function, some of which may be endogenous. Also, the semi-parametric approach is a way for us to control for non-linearities in the functional form. While other hospital studies, such as Cohen and Morrison Paul (2008) and Li and Rosenman (2001) have used a Generalized Leontief functional form, those researchers did not estimate a spatial panel data model and such a non-linear functional form would be relatively difficult to implement in a panel data context using parametric panel data spatial econometrics techniques. While there are few known semi-parametric or nonparametric studies of hospital costs (Wilson and Carey is an exception), there are no known panel data studies on this application that address both the spatial and time dimensions in the kernel weights.

Specifically, McMillen and Redfearn (2010) estimate a semi-parametric regression model to control for spatial effects. Their model is of the form:

$$Y_i = f(Z_i) + \gamma X_i + u_i, \quad (3)$$

where $f(Z_i)$ represents the non-parametric variables, X is the single parametric variable, and u is an iid error term. The focus of a semi-parametric regression model is the estimation of γ by controlling for the other variables in a nonparametric manner. To estimate the coefficients in $f(Z_i)$, we use the Geographically Weighted Regressions (GWR) non-parametric estimator, which can be estimated using weighted least squares.

The advantage of using a semi-parametric model over a fully nonparametric one is for convenience in interpretation and the faster converging rate, the latter being particularly important given our sample size. The estimate of γ provides an estimate of the conditional expectation of Y_{it} given X_{it} after controlling in a general, nonparametric way for the effects of all other variables.

We first consider the case where both X and Z are exogenous. Following Robinson (1988), by taking expectation of (3) conditional on variables in the nonparametric component, z_{it}, then subtracting it from (3) we have

$$Y_{it} - E(Y_{it} \mid Z_{it}) = [X_{it} - E(X_{it} \mid Z_{it})]' \gamma + u_{it} \quad (4)$$

If we use the following notations:

$$\nabla_{it} = Y_{it} - E(Y_{it} \mid Z_{it}), \; V_{it} = X_{it} - E(X_{it} \mid Z_{it}), \quad (5)$$

then we can write the above equation (4) as

$$\nabla_{it} = V_{it} \gamma + u_{it} \quad (6)$$

Then a simple OLS regression of ∇ on V will give a consistent estimator for γ, assuming $E(Y_{it} \mid Z_{it})$ and $E(X_{it} \mid Z_{it})$ are known. In practice, these conditional expectations can be approximated using locally weighted regression[2] (LWR) following McMillen and Redfearn (2010). By rotating each independent variable in the parametric part of the model, X, and leaving the rest of the independent variables in the nonparametric component of the model, f(Z), we can get an

[2] McMillen and Redfearn (2010) note that LWR is equivalent to Geographically Weighted Regressions (GWR). Accordingly, we refer to LWR and GWR interchangeably in our discussion.

estimate of the marginal impact of each individual factor on hospital costs after controlling for the effects of all other variables in a nonparametric way.[3]

We use Geographically weighted regression (GWR) in approximating the conditional expectation in the above discussion. More specifically, the variable $E(X_{it} \mid Z_{i_0t_0})$ is calculated by minimizing the following objective function with respect to a and b,

$$\Sigma_i\Sigma_t \ (y_{it} - a - b'X_{it})^2 K(d_{it}/h_1)\, K(\tau_{it}/h_2) \qquad i=1,2,\ldots,N,\ t=1,2,\ldots,T. \qquad (7),$$

where K(•) is a kernel function that determines the weight that observation (i, t) receives in the regression; d_{it} and τ_{it} are the distance between observation (i; t) and (i_0; t_0) in space and in time, respectively,[4] and h_1, h_2 are the bandwidth on space and time, respectively. This approach is appealing because it leverages the panel nature of the data in a manner that implies hospital i in year t is a different observation than hospital i in year t-r (where r=1,2,…,9, t=1,2,…,10), since we add the additional dimension of time. This distance between observations in the time dimension is reflected in the kernel that depends on τ_{it}.

The Gaussian kernel function is used to calculate the weight assigned to each observation, based on its distance from the target point, both in geographic location and time/year.[5] As McMillen and Redfearn (2010) note, it is well known

---

[3] The semi-parametric regression approach follows McMillen and Redfearn (2010). For each parametric variable z, first we use GWR to regress y on X, and z on X, calculate the fitted residuals u_y and u_z , then regress the fitted u_y on u_z using OLS. The parameter estimate for γ is the coefficient on u_z in this second stage. We repeat this process allowing each of our explanatory variables to be the sole parametric variable, z, in order to obtain semi-parametric regression estimators for the coefficient and standard errors of z, after controlling for any non-linearities in the model.

[4] The distances d_{it} and τ_{it} are normalized with the standard deviation of d_{it} and τ_{it} over all i and t.

[5] The kernel function on time assigns positive weight only for $\tau_{it} \leq 0$ and assigns 0 weight for $\tau_{it} > 0$, i.e., only those observations that precede observation (i0,t0) in time are given positive weights.

that the choices of kernel functions tend to have little effect on the results. The performance of kernel estimation is much more sensitive to the choice of bandwidth, h. Given that in the dataset the hospitals are located densely in some areas and sparsely in other areas, a fixed bandwidth would lead to over-smoothing in areas where many observations are present and under-smoothing in areas with sparse data. Following McMillen and Redfearn (2010) we use a "K^{th} nearest neighbor (K-nn)" approach in calculating the bandwidth. For a target point we chose a bandwidth to include a fixed percentage of the sample into the local averaging.[6]

In addition to the variables included in the cost function models above, we explore the impacts of including an additional set of "output" variables in the R vector in (1), based on life expectancies. Carey and Burgess (1999) included outcomes in the cost function for hospitals, and in our context we include life expectancy as an outcome. The life expectancy data are described below in the data section. In other words, it is possible for hospitals that spend more to end up with better outcomes, while hospitals that treat patients who tend to be healthier may have different operating costs. One way in which patients may be healthier is by participating in wellness programs. While the concept of wellness programs promoting lower medical costs is appealing, there are few studies that report on

[6] McMillen and Redfearn (2010) use two "window" sizes (25% and 100%) and a tri-cube kernel function, and we follow their approach of using these two bandwidths, since there is a lack of guidance in the literature on bandwidth selection in the particular semi-parametric models that we analyze (Ichimura and Todd, 2006). With a Gaussian kernel function (Standard normal density function), which we use in our analysis, the bandwidth includes a specified percentage of the sample points within two standard deviations away from the target point. Sample points outside of the window (two standard deviations) are in the "tails" and essentially get near-zero weights and are ignored in the averaging. The two standard deviations in the Gaussian kernel is analogous to the support of [-1, 1] for the tri-cube kernel used by McMillen and Redfearn (2010).

the actual benefits that accrue from such programs. Ahmed and Rak (2010) report that participation in a large wellness program provided by a healthcare company for diagnostic categories of musculoskeletal and digestive resulted in readmission rates that were lower among individuals who were engaged in a wellness program as compared with those who were not engaged. Individuals not engaged in the wellness program were almost four times more likely to have a hospital readmission than those who participated in the program. Shephard (1999) concluded that studies of work site wellness programs suggest a number of important results including a reduction in healthcare costs with yearly benefits estimated to be between $500 and $700 per worker per year. Naydeck, Pearson, and Ozminkowski (2008) found by using a multivariate model for data for the firm Highmark Inc. estimated that yearly overall health care expenses were on average $176 lower for participants in the program. Inpatient expenses for participants were lowered by $182. They conclude that their study suggests that a comprehensive health promotion program can lower the rate of health care cost increases.

To address this potential endogeneity of the life expectancy variable, we follow an approach outlined by Baltagi and Li (2002) for instrumental variable estimation in semi-parametric panel data models.

Specifically, Baltagi and Li (2002) describe the following type of semi-parametric model:

$$Y_{it} = f(Z_{it}) + \beta X_{it} + u_{it}, \qquad (7)$$

where X_{it} is of dimension one, β is a unknown parameter that is of our main interest, Z_{it} is of dimension dx1, and $f(z_{it})$ is a smooth but otherwise unknown function. They also introduce a kernel function, $K_{it,js} = K((Z_{it} - Z_{js})/b)$, where b is the smoothing parameter.[7] Also, in contrast to our semi-parametric regression model, the Baltagi and Li error term structure is assumed to be a one-way error component, which is the same as that used by Kapoor et al in their spatial econometrics panel data model. In contrast, we address the time dimension in the kernel weights. Baltagi and Li outline a feasible instrumental variables, generalized least squares (IVGLS) estimator for the endogenous variable X_{it}, which depends on the kernel, among other variables. This IVGLS estimator is subsequently used in obtaining a nonparametric estimator for f(Z). We follow the Baltagi and Li (2002) approach, but in the context of a GWR framework, and we also use the Gaussian kernel function.

At this point we only allow the endogenous variable(s) to be X_{it}, i.e. the endogenous variable is in the parametric part of the model.[8] Note that when X_{it} is endogenous, u_{it} and V_{it} are correlated in (6) because of the dependence between X_{it} and u_{it}. We need an instrumental variable that is correlated with V_{it} but independent from u_{it} to get a consistent estimate of β. In this particular case, the life expectancy of a hospital patient in year t likely depends on the hospital's expenses in the previous year, (t-1). A hospital's expenses in (t-1) in turn

[7] It is noteworthy that in the Baltagi and Li (2002) semi-parametric estimation approach, this kernel function can include some of the explanatory variables (which is a regular kernel regression), while in the locally weighted regressions approach the distance between two observations is generally used as the kernel argument.

[8] In the case that the nonparametric components are endogeneous, the asymptotic analysis is more complex and we are not aware of any kernel method that addresses the issue.

depends on factors in year (t - 1), such as wages, fixed costs, and others, i.e. all the exogenous variables included in Z_{t-1}. Therefore, we use $m \equiv E(V_{it} \mid Z_{i,t-1})$ as the instrumental variable.

Finally the estimator for β can be obtained with IV-OLS as follows:[9]

$$\hat{\beta} = (V'mm'V)^{-1}V'mm'(Y - E(Y_{it} \mid Z_{it})) = \beta + (V'mm'V)^{-1}V'mm'u \qquad (8)$$

After obtaining an estimator for β, the nonparametric component $f(z_{it})$ can be estimated by a nonparametric regression of

$(y_{it} - x_{it}\hat{\beta})$ on z_{it} , i=1,…,N, t=1,…,T. A locally weighted estimator of $f(z_{it})$ can be calculated at each observation point.[10] Since

$\hat{\beta}$ converges at rate of $n^{1/2}$, which is faster than the usual nonparametric convergence rate, replacing β with its estimator has minimal impact in the estimation of $f(z_{it})$.

Monte Carlo Simulation

In order to provide some confidence in the performance of the two-stage estimator for a panel-data model with both spatial effects and endogeneity, we developed a simple Monte Carlo experiment based on a stylized model from our

[9] In the situation where we know Var(u) = Σ , Baltagi and Li (2002) show that a potentially more efficient estimator of β is the IV-GLS estimator. However, in their Monte Carlo simulation the IV-GLS estimator performs worse than the IV-OLS estimator, so we focus on the IV-OLS estimator in our study.

[10] It is possible to conduct F-tests on the significance of the nonparametric estimates, as in McMillen and Redfearn (2010), which we have done and reported in Tables 7a and 7b. As an alternative, we could report standard deviations of all N times T coefficient estimates, although these cannot be used for the purpose of statistical inference.

data set. In this Monte Carlo simulation, we envision a hospital cost model where there are N hospitals and each hospital is observed over T time-periods. The hospital costs, Y_{it}, i = 1, …,N; t = 1,…,T, depends on its past, Y_{it-1}, and an exogenous variable, Z_{it}, as in the following data generating process (DGP):

$$Y_{it} = \beta Y_{it-1} + \gamma_1 Z_{it} + \gamma_2 Z_{it}^2 + \mu_{it}\ , \qquad (9)$$

where we set $\beta = 0.5$, $\gamma_1 = \gamma_2 = 1$. We also assume $Y_{i0} = 0$. In the DGP in (9), the variable Z_{it} is i.i.d. from a uniform distribution on [-0.5, 0.5]. The error process, $\mu_{NT} = [\mu_{11},\ldots, \mu_{NT}]$ ', can be written in matrix form as the following:

$$\mu_{NT} = (\rho W_N \alpha_N) \otimes e_T + \alpha_N \otimes e_T + \varepsilon_{NT} \qquad (10)$$

where ρ is a scalar parameter in the first order spatial autoregressive process; W_N is a (N x N) arbitrary (known) weight matrix based on geographic distances (a separately generated variable)[11] between observations; α_N is a (N x 1) vector of random variables following a Normal distribution with mean 0 and standard deviation of 1/3;[12] e_T is a (T x 1) vector of ones; and ε_{NT} is a (NT x 1) vector of random variables. In order to examine the performance of the estimator in different scenarios, we allow the parameters in the model to take the following values: N = 30, 60; T = 10, 20; ρ = 0.25, 0.75; and also the correlation level among observations taken on the same subject (i.e., hospital), *Corr* = 0.4, 0.7.

---

[11] W is constructed using the product of Gaussian kernels on distance in space and time.

[12] This standard deviation for α_N is chosen to have the same scale with the other random variable in the model, Z, which is uniform [-0.5, 0.5].

Note that in our actual data set, N=30 and T=10. For each of the above specifications we performed 500 repetitions (M=500).

The focus in the simulation study is on the estimation of β, which is estimated using equation (6). The conditional expected values used in (6), namely $E(Y_{it} \mid Z_{it})$, $E(X_{it} \mid Z_{it})$ and $E(V_{it} \mid Z_{i;t-1})$, are calculated using GWR. We report in Table 2 the estimated bias, standard deviation (Std), and root mean square error (RMSE) for all model specifications. These quantities are calculated as:

$$\text{Bias}(\hat{\beta}) = M^{-1}\Sigma_i (\hat{\beta}_i - \beta), \quad (11)$$

$$\text{Std}(\hat{\beta}) = \{ M^{-1}\Sigma_i (\hat{\beta}_i - \text{mean}(\hat{\beta}))^2\}^{1/2} \quad (12)$$

$$\text{and Rmse}(\hat{\beta}) = \{ M^{-1}\Sigma_i (\hat{\beta}_i - \beta)^2\}^{1/2}, \quad (13)$$

where i=(1,…,M).

Summarizing the simulation results, the estimator performs reasonably well in all DGP specifications. With the true value $\beta = 0.5$, the estimator gives an average bias of 0.013, average Std of 0.071 and average RMSE of 0.082 across all specifications, indicating that the estimator under study performs reasonably well in all DGP specifications.

As we increase the number of units (N) all three performance measures decrease, which indicates that the estimator under study is likely to be consistent in large samples, as suggested Baltagi and Li (2002). In addition, the Std and RMSE also decreases with the number of time periods (T) on which each unit is observed.

When the correlation level among observations on the same unit taken at different times increases from 0.4 to 0.7, the estimator tends to produce slightly larger Std and RMSE. As expected, the higher correlation has no significant impact on the estimator's bias. Similarly, when ρ (the first order spatial autoregressive spatial parameter) increases from 0.25 to 0.75, the standard deviation of the estimator increases, as expected. However, the bias and RMSE of the estimator shows a mixed pattern with higher spatial dependence.

Figures 2a, 2b through 5a, 5b show the distributions of the b parameter in each of the simulations, for various combinations of assumptions on the time series and spatial autocorrelation parameters. In each set of diagrams, the panel (a) represents the case with N=30, T=10, while the second, panel (b) represents the case with N=60, T=20. By comparing each pair of plots in panels (a) and (b), it is evident that when the sample size in the simulations increases (in both the time and spatial dimensions), the distributions for b become more concentrated around 0.5, which is the true value to be estimated. This implies that our semi-parametric estimator for the endogenous variable is likely to be consistent in large samples.

Data

The annual data covering the years 1999-2008 on the 30 individual Connecticut hospitals which were also used in Cohen, Gerrish, and Galvin (2010) was obtained from the State of Connecticut Department of Public Health. Descriptive statistics are presented in Table 1a. Labor price (consisting of total wages and benefits) and the price of other expenses (excluding labor and depreciation

expenses) are each normalized to a base year (1999). The average hospital's property, plant, and equipment value was about $82,000,000 (in 1999 dollars), with a range of $71,000,000 to $256,000,000. The average number of inpatient days was 66,000 and outpatient visits averaged 211,000. The mean of variable costs (expenses) was $157,000,000 (in 1999 dollars).

The life expectancy data also were obtained from the State of Connecticut Department of Public Health. The descriptive statistics for years above average life expectancies at birth for all causes of death (LIFE), as well as for several specific causes (cardio, cancer, diabetes, stroke, trauma), were calculated and are listed in Table 1b. These annual life expectancy data for each of the 30 Connecticut hospitals also cover the years 1999 to 2008. Trauma patients had the worst life expectancy (below average), while cardio patients had the greatest life expectancy in years above average.

Results

With the pooled OLS regression for the Cobb-Douglas functional form, we also include an intercept term. The OLS regression results are presented in Tables 3a, 3b, and 3d. Both input price variables have positive and significant coefficients. Medicaid has a positive and significant (P=0.03) coefficient, but Medicare has a positive and insignificant coefficient. Capital, case mix index, outpatient, and inpatient all have positive and significant coefficients, implying that more capital increases costs. Also, the inpatient and outpatient coefficients are significantly less than 1, implying economies of scale. In other words, one would expect that serving more patients will lead to lower costs per patient. Also,

changes in the case mix have a significant impact on costs. The YEAR variable is negative and significant, implying the presence of exogenous technical change. Finally, we also include various specifications that include life expectancy variables, including two regressions that have the years above average age of death (LIFE), as well as several individual causes of death. These life expectancy results are discussed in more detail below.

The OLS estimates are based on the Cobb-Douglas functional form, which can be rather inflexible because it does not allow for non-linearity in the data or model. A semi-parametric regression approach, as in the McMillen and Redfearn (2010) approach described above, can help control for nonlinearity with a more general functional form, while at the same time generating parameter estimates and standard errors that can be used for statistical inference. Also, with this semi-parametric regression approach, we omit the constant term from the model, because to obtain parameter estimates for the constant term we would need to run a model with the dependent variable always equal to 1, which introduces additional complications. Although we directly address the time dimension in one of the weighting kernels, we also include a time trend (YEAR) and find that it is negative and highly significant in most specifications. This result for the coefficient on the YEAR variable may be due to the presence of exogenous technical change. These semi-parametric results are shown in Table 4a and 4b.

An important issue to consider in the estimation of the semi-parametric regression models is the appropriate bandwidth. As Ichimura and Todd (2006) discuss, the literature on semi-parametric regression models provides relatively

little guidance on bandwidth choice, and the papers that address the issue focus primarily on special cases (i.e., binary choice models, censored regression models, single index models), none of which are directly applicable in our semi-parametric regression models. A common approach for bandwidth selection in GWR models - cross-validation – is not as straightforward in a semi-parametric regression framework. This is because unlike with nonparametric analysis, the semi-parametric regression models report only one parameter estimate for all observations on each explanatory variable, so the cross-validation approach is not directly applicable. McMillen and Redfearn (2010) report results for each of two different bandwidths - 25% and 100% - so we follow their approach and present results for each of these two bandwidths in each of our semi-parametric regression specifications.

While there is evidence in the GWR literature that the estimation results can be sensitive to the bandwidth choice, we find little difference in the signs and significance of our parameter estimates in the semi-parametric regressions with the two bandwidths (25% and 100%). However, there are several important exceptions to this finding. Specifically, the significance of the Medicaid variable is opposite in the two bandwidths in the regression model where we allow for the endogenous “life” variable; the “YEAR” variable is insignificant in both bandwidth choices;

In the semi-parametric regression models, for both the 25% and 100% bandwidth results the signs of all parameter estimates are positive and all highly significant (Tables 4a, 4b). Furthermore, the “inpatient days” parameter is also

significantly less than 1 for the specifications without the LIFE variable, as well as those including the LIFE variable and including the other life expectancy variables, implying economies of scale for inpatient services. The same is true for the “outpatients” variable in the models where “life” is endogenous. In all of the semi-parametric regression specifications, the “outpatient visits” variable is statistically significantly less than 1, and it is also significantly greater than zero, implying economies of scale for outpatients but significant marginal costs.

Another variable we included was a time trend (“year”), to allow for the possibility of exogenous technical change over the period of our sample. The parameter estimates on the time trend terms are highly significant in all semi-parametric regression models except for the model where “LIFE” is endogenous. This may imply that after controlling for all cost determinants, (except for the case where we include life expectancy), costs fell over time, implying exogenous technical change.

As we mention above, we also have life expectancy data for all of the years (1999-2008) for all hospitals in the data sample. With the models including the LIFE variable, we estimate an OLS specification (Table 3b), and a 2SLS specification (using one-period lagged values for all of the explanatory variables as the instruments for LIFE, with results in Table 3c). For the most part, the signs and significance of the parameter estimates are similar across the OLS and 2SLS models, with a negative and significant parameter estimate for the LIFE variable in the OLS model but insignificant parameter estimate in the 2SLS model. This insignificance leads one to wonder how the results might differ with a

semi-parametric approach, and we address this below. We also re-run the semi-parametric models for a few variations that include life expectancy measures as "outputs" – one including overall lifespan years above average for each hospital, and a separate set of models where individual causes of death are included (cardio, cancer, diabetes, trauma, and stroke). In Tables 5a and 5b, the parameter on LIFE (modeled as exogenous) is negative and highly significant for both bandwidths in the semi-parametric models, implying hospitals with patients who live longer than average also experience significantly lower costs on average.

In Tables 6a and 6b, we include years above average lifespan for individual causes of death as outcomes in the cost function, assuming each of these cause of death variables are exogenous. For the 25% bandwidth, the coefficients on cardio, cancer, trauma and diabetes are all negative and significant, implying hospitals with cardio and stroke cause of death patients who live longer on average also experience significantly lower costs. Stroke patients, on the other hand, tend to increase costs for hospitals when these patients live longer. In contrast, all of the separate cause of death variables are positive and significant in the 100% bandwidth, which implies higher life expectancies for patients who die from each of these diseases also leads to higher hospital costs. Also, the "year" variable is negative and significant in the 25% bandwidth, while it is positive and significant in the 100% bandwidth. An important contrast to these results, however, is when we control for potential endogeneity of these specific categories of life expectancy in a parametric 2SLS estimation procedure, with

results shown in Table 3e. In that context, the parameter estimates for all of the individual life expectancy variables are insignificant. We are unable to estimate the 2SLS model in a semi-parametric context because the Baltagi and Li approach is intended for a situation with only one endogenous variable. Perhaps using the parametric approach when controlling for endogeneity of the life expectancy variables leads to biased parameter estimates and/or insignificance of these variables.

When we estimate the model with the overall life variable as endogenous (Tables 7a, 7b), we obtain an estimate of the parameter on the life variable as ranging from -0.0000308 (for the 25% window) to -0.0000474 (for the 100% window). For the average hospital in the average year, which has costs of $157,000,000, this implies a mean cost increase of approximately $4,700 (for the 25% window) to $7,400 (for the 100% window) for each additional year of life expectancy above the average, after controlling for the endogeneity of the life expectancy variable and allowing for spatial heterogeneity over space and time. We also calculate this cost of greater life expectancy for each hospital in our sample, and this implies a range of $484 to $22,620 per additional year of life for the 25% window, and $758 to $35,438 for the 100% window. Clearly, these findings imply some hospitals benefit much more than others from treating patients with longer life expectancies.

In the semi-parametric regression models where LIFE is endogenous, the first step entails estimating the parameter for the LIFE variable using two-stage least squares semi-parametric estimator; then, the remaining coefficients are

estimated with a nonparametric approach. For these nonparametric estimates, there are 300 separate coefficient estimates, so we list standard deviations of these coefficient estimates, and inference is not straightforward. Instead, we perform F-tests to test for the explanatory power of including each nonparametric variable in the model. The version of the F-test here is slightly different from the one described in McMillen and Redfearn (2010); here, we replace Y with $(Y - X\beta)$, then re-estimate the model using locally weighted regressions for the nonparametric variables, z, while omitting the variable for which we desire the F-statistic. Tables 7a and 7b present the F-statistics for the exogenous variables, as well as the t-statistic for the (endogenous) LIFE variable. While the LIFE variable is negative and highly significant (P-value=0.00000), the F-tests also imply the marginal costs of treating inpatients and outpatients are significantly greater than zero. These inpatient and outpatient treatment cost parameters are significantly less than 1, implying economies of scale. Perhaps this is due to under-utilized capacity at most hospitals in many years, leading to highly significant economies of scale in both inpatient and outpatient services. With the exception of the YEAR variable, the case-mix and Medicaid days variables, all other variables in the semi-parametric regressions with endogenous LIFE are jointly significant for both bandwidths based on the F-tests.

Conclusions

We estimate a cost function model for all hospitals in the state of Connecticut, USA, for each year in the period 1999-2008. Our approaches include least squares, semi-parametric regressions, and semi-parametric

regressions in the presence of an endogenous variable. The semi-parametric regression approach allows for a general functional form that is less restrictive than Cobb-Douglas, and we also introduce an estimator for the situation where life expectancy is considered endogenous. Our Monte Carlo simulations support the notion that this endogenous variable semi-parametric regression model performs well. We incorporate a time trend in the regressions, and the sign of this variable is negative and significant in many of our models, which is indicative of exogenous technical change. To fully leverage the space-time panel nature of our data set, we also introduce a sophisticated kernel structure that incorporates both the time and space dimensions of our data, which is a contribution to the literature on semi-parametric regressions (particularly in the hospitals context).

The “inpatient days” and “outpatient visits” variables are significant in most variations of the model, and we also find evidence of economies of scale in all models. Adding the life expectancy variables in the semi-parametric regression models leads to smaller economies of scale estimates, compared with the models without the life expectancy variables.

These findings imply that allowing for a general functional form with semi-parametric regression models, in a panel (space-time) data framework that also allows for endogeneity of life expectancy, can be crucial in adequately assessing the determinants of hospital costs. Ignoring the spatial elements of the data that are inherently controlled for by the GWR model, and failing to address the potential endogeneity of the LIFE variable can lead to incorrect inferences for the relationship between LIFE and total hospital operating costs.

References

Ahmed, O.I. and Rak, D.J. 2010. "Hospital readmission among participants in a transitional case management program." *American Journal of Managed Care*, 16 (10): 778-783.

Baltagi, B. H. and Q. Li, 2002, On instrumental variable estimation of semiparametric dynamic panel data models. Economics letters, 76, 1-9.

Bates, L. and R. Santerre. 2005. “Do agglomeration economies exist in the hospital services industry?” *Eastern Economic Journal*, 31, 617–28.

Blundell, R., and J.L. Powell. 2003. “Endogeneity in Nonparametric and Semiparametric Regression Models,” in M. Dewatripont, L. Hansen, and S. Turnovsky (eds.), Advances in Economics and Econometrics, Ch. 8, 312-357.

Carey, C. and J.F. Burgess. 1999. “On Measuring the Hospital Cost/Quality Tradeoff.” *Health Economics* 8: 509-520.

Carey, C. and T. Stefos. 1992. “Measuring inpatient and outpatient costs: a cost function approach.” *Health Care Financing Review* 14, 115–124.

Cohen, JP, W. Gerrish and J.R. Galvin. 2010. “Health Care Reform and Connecticut’s Hospitals” Journal of Health Care Finance 37(2):1-7.

Cohen, J. P. and Paul, C. J. M. 2008. “Agglomeration and cost economies for Washington State hospital services,” *Regional Science and Urban Economics*, 38, 553–64.

Cohen, Jeffrey P. and Catherine Morrison Paul. 2011. “Scale and Scope Economies for Drug Abuse Treatment Costs at Hospitals in Washington State”, *Applied Economics* 43(30): 4827-4834.

Ichimura, H. and P. Todd. 2006. “Implementing Nonparametric and Semiparametric Estimators,” Handbook of Econometrics, Volume 6.

Kapoor, M., Kelejian, H.H., Prucha, I.R., 2007. “Panel data models with spatially correlated error components” *Journal of Econometrics* 140, 97_130.

Li, and R. Rosenman. Li, T., Rosenman, R., 2001. “Estimating hospital costs with a generalized Leontief function.” *Health Economics* 10, 523–538.

Daniel P. McMillen & Christian L. Redfearn, 2010. "Estimation And Hypothesis Testing For Nonparametric Hedonic House Price Functions," *Journal of Regional Science*, vol. 50(3), pages 712-733.

Mobley, LR., H. Frech, and L. Anselin. 2009. “Spatial Interaction, Hospital Pricing and Hospital Antitrust”, *International Journal of the Economics of Business*, 16 (1): pp 1-17.

Moscone F, Knapp M. 2005. Exploring the spatial pattern of mental health expenditure, *Journal of Mental Health Policy and Economics*, 8, 205-217.

Naydeck, B.L., Pearson, J.A. and Ozminkowski, R.J. (2008) "The impact of the Highmark employee wellness programs on 4-year healthcare costs." *Journal Of Occupational And Environmental Medicine* 50 (2): 146-156.

O'Sullivan, A., 2010. Urban Economics, 7th edition. McGraw-Hill.

Shephard, R.J. 1999. "Do work-site exercise and health programs work?" *Physician and Sports Medicine* 27(2): 48-72.

Wilson, P. and K. Carey. 2004. Nonparametric Analysis of Returns to Scale in the U.S. Hospital Industry, *Journal of Applied Econometrics* 19: 505-524.

Table 1a – Descriptive Statistics, hospital cost data (N=30, T=10)

	CASE_MIX	INPAT_DAYS	LOGTCOST	OUTPATIENT	MEDICAID	MEDICARE	WAGES	OTH_EXP	PROPERTY_ PLANT_EQUIP	TOTAL_COST
Mean	1.194787	66293.67	8.084383	211019.0	0.123848	0.505252	1.242050	1.521380	82243502	1.57E+08
Median	1.114500	50275.50	8.118361	186844.5	0.109245	0.537152	1.228000	1.341000	71754704	1.31E+08
Maximum	2.223000	272757.0	8.877100	1032759.	0.495339	0.784632	1.861000	3.633000	2.76E+08	7.54E+08
Minimum	0.780000	5544.000	7.207867	8908.000	0.003100	0.000350	0.544000	0.689000	10565488	16138647
Std. Dev.	0.209467	55761.34	0.312007	170646.9	0.072832	0.120243	0.205208	0.559282	61373154	1.23E+08
Skewness	1.102810	1.730478	0.084179	2.104335	2.363721	-2.385731	0.439969	1.388155	1.327233	1.925032
Kurtosis	4.730966	5.759040	2.472964	8.811376	11.39983	10.84071	3.012215	4.754808	4.019425	7.491707
Jarque-Bera	98.26255	244.8815	3.826394	643.5624	1161.323	1053.045	9.680492	134.8406	101.0677	437.4804
Probability	0.000000	0.000000	0.147608	0.000000	0.000000	0.000000	0.007905	0.000000	0.000000	0.000000
Observations	300	300	300	300	300	300	300	300	300	300

Table 1b – Descriptive Statistics, life expectancy data –
average years above expected lifespan overall (LIFE) and for various causes of death (N=30, T=10)

	CANCER	DIABETES	LIFE	STROKE	TRAUMA	CARDIO
Mean	822.65	159.10	5282.22	339.43	-59.91	1481.76
Median	735.95	109.75	4791.86	305.75	-8.50	1169.94
Maximum	2505.02	975.83	15179.30	1080.28	702.25	5667.70
Minimum	-803.90	-60.22	-3176.92	-23.60	-1395.13	-66.80
Std. Dev.	675.58	171.20	4094.11	248.97	289.08	1102.90
Skewness	0.34	1.61	0.28	0.59	-1.43	1.01
Kurtosis	2.43	5.97	2.43	2.63	6.89	3.71
Jarque-Bera	9.90	240.18	8.03	19.28	291.48	57.31
Probability	0.01	0.00	0.02	0.00	0.00	0.00
Observations	300	300	300	300	300	300

Table 2 – Monte Carlo Simulation Results: bias, standard deviation, and Rmse

N=30

	T=10				T=20			
	corr=0.4		corr=0.7		corr=0.4		corr=0.7	
	ρ=0.25	ρ=0.75	ρ=0.25	ρ=0.75	ρ=0.25	ρ=0.75	ρ=0.25	ρ=0.75
Bias	-0.0013	-0.0326	0.0506	-0.0001	0.0366	0.0083	0.0905	0.0576
Std	0.0790	0.0934	0.0868	0.0979	0.0602	0.0710	0.0642	0.0775
Rmse	0.0790	0.0988	0.1004	0.0978	0.0704	0.0714	0.1109	0.0965

N=60

	T=10				T=20			
	corr=0.4		corr=0.7		corr=0.4		corr=0.7	
	ρ=0.25	ρ=0.75	ρ=0.25	ρ=0.75	ρ=0.25	ρ=0.75	ρ=0.25	ρ=0.75
Bias	-0.0295	-0.0601	0.0054	-0.0428	0.0225	0.0095	0.0620	0.0360
Std	0.0618	0.0688	0.0661	0.0779	0.0459	0.0599	0.0508	0.0688
Rmse	0.0684	0.0913	0.0663	0.0888	0.0510	0.0606	0.0801	0.0776

Table 3a – OLS Regression Results

(Note: all variables are in natural logs)

Dependent Variable: LOGTCOST
Method: Least Squares
Date: 02/07/13 Time: 00:10
Sample: 1 300
Included observations: 300

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.881488	0.079541	36.22654	0.0000
LOG(WAGE)	0.122133	0.027179	4.493700	0.0000
LOG(OTH_EXP)	0.125878	0.012573	10.01152	0.0000
LOG(INPAT_DAYS)	0.275012	0.008673	31.71022	0.0000
LOG(OUTPATIENTS)	0.018923	0.004024	4.702901	0.0000
LOG(MEDICAID_DAYS)	0.014619	0.006705	2.180427	0.0300
LOG(MEDICARE_DAYS)	-0.000102	0.003475	-0.029413	0.9766
LOG(PROPERTY_PLANT_EQUIP)	0.111841	0.008508	13.14563	0.0000
LOG(YEAR)	-0.026582	0.007324	-3.629541	0.0003

R-squared	0.980710	Mean dependent var	8.084383
Adjusted R-squared	0.980180	S.D. dependent var	0.312007
S.E. of regression	0.043926	Akaike info criterion	-3.383091
Sum squared resid	0.561476	Schwarz criterion	-3.271978
Log likelihood	516.4637	F-statistic	1849.321
Durbin-Watson stat	1.892277	Prob(F-statistic)	0.000000

Table 3b–Ordinary Least Squares, with years above expected lifespan (LIFE)

(Note: all variables are in natural logs, except for “LIFE” since there are non-positive observations)

Dependent Variable: LOGTCOST
Method: Least Squares
Date: 02/07/13 Time: 00:17
Sample: 1 300
Included observations: 300

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.754142	0.082728	33.29159	0.0000
LOG(WAGE)	0.104108	0.026727	3.895291	0.0001
LOG(OTH_EXP)	0.129709	0.012244	10.59384	0.0000
LOG(INPAT_DAYS)	0.287118	0.008881	32.33083	0.0000
LOG(OUTPATIENTS)	0.022427	0.003992	5.618216	0.0000
LOG(MEDICAID_DAYS)	0.011810	0.006544	1.804632	0.0722
LOG(MEDICARE_DAYS)	0.002536	0.003430	0.739268	0.4603
LOG(PROPERTY_PLANT_EQUIP)	0.110217	0.008272	13.32494	0.0000
LOG(YEAR)	-0.025862	0.007115	-3.634955	0.0003
LIFE	-3.60E-06	8.37E-07	-4.302122	0.0000

R-squared	0.981867	Mean dependent var	8.084383
Adjusted R-squared	0.981305	S.D. dependent var	0.312007
S.E. of regression	0.042661	Akaike info criterion	-3.438292
Sum squared resid	0.527792	Schwarz criterion	-3.314833
Log likelihood	525.7439	F-statistic	1744.800
Durbin-Watson stat	2.042040	Prob(F-statistic)	0.000000

Table 3c – Two-Stage Least Squares, with years above expected lifespan (LIFE)

(Note: all variables are in natural logs, except for “LIFE” since there are non-positive observations)

Instruments for LIFE variable: lagged exogenous variables

Dependent Variable: LOGTCOST
Method: Two-Stage Least Squares
Date: 02/12/13 Time: 23:33
Sample (adjusted): 1 270
Included observations: 270 after adjustments
LOGTCOST = C(1) + C(2)*LOG(WAGE) + C(3)*LOG(OTH_EXP) + C(4)*LOG(INPAT_DAYS) + C(5)*LOG(OUTPATIENTS) + C(6)*LOG(MEDICAID_DAYS) + C(7)*LOG(MEDICARE_DAYS) + C(8)*LOG(PROPERTY_PLANT_EQUIP) + C(9)*LOG(YEAR) + C(10)*LIFE
Instrument list: CONST LOGWAGE LOGOTH_EXP LOGINPAT LOGOUTPAT LOGMEDICAID LOGMEDICARE LOGPROP LOGYEAR LOGWAGE(30) LOGOTH_EXP(30) LOGINPAT(30) LOGOUTPAT(30) LOGMEDICAID(30) LOGMEDICARE(30) LOGPROP(30) LOGYEAR(30)

	Coefficient	Std. Error	t-Statistic	Prob.
C	2.551955	0.216515	11.78648	0.0000
LOG(WAGE)	0.064192	0.040326	1.591800	0.1126
LOG(OTH_EXP)	0.121464	0.015118	8.034514	0.0000
LOG(INPAT_DAYS)	0.302959	0.022015	13.76131	0.0000
LOG(OUTPATIENTS)	0.029717	0.007671	3.873682	0.0001
LOG(MEDICAID_DAYS)	0.005857	0.008963	0.653429	0.5141
LOG(MEDICARE_DAYS)	0.005633	0.005171	1.089483	0.2769
LOG(PROPERTY_PLANT_EQUIP)	0.108001	0.009899	10.90987	0.0000
LOG(YEAR)	-0.019634	0.007839	-2.504738	0.0129
LIFE	-8.67E-06	5.68E-06	-1.526894	0.1280

R-squared	0.980248	Mean dependent var	8.074969
Adjusted R-squared	0.979564	S.D. dependent var	0.309266
S.E. of regression	0.044211	Sum squared resid	0.508188
Durbin-Watson stat	2.187138	Second-stage SSR	0.478407

Table 3d – Ordinary Least Squares, including life expectancy (each cause separately)

(Note: all variables are in natural logs, except for “CARDIO”, “CANCER”, “DIABETES”, “STROKE”, and “TRAUMA” since there are non-positive observations for those variables)

Dependent Variable: LOGTCOST
Method: Least Squares
Date: 02/07/13 Time: 00:23
Sample: 1 300
Included observations: 300

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.732764	0.088275	30.95728	0.0000
LOG(WAGE)	0.109258	0.026786	4.078931	0.0001
LOG(OTH_EXP)	0.135644	0.013008	10.42773	0.0000
LOG(INPAT_DAYS)	0.289377	0.009742	29.70290	0.0000
LOG(OUTPATIENTS)	0.021375	0.004025	5.310570	0.0000
LOG(MEDICAID_DAYS)	0.013162	0.006561	2.006029	0.0458
LOG(MEDICARE_DAYS)	0.002968	0.003444	0.861772	0.3895
LOG(PROPERTY_PLANT_EQUIP)	0.111139	0.008461	13.13535	0.0000
LOG(YEAR)	-0.027685	0.007279	-3.803607	0.0002
CANCER	-1.24E-05	7.32E-06	-1.692662	0.0916
CARDIO	-5.97E-06	4.81E-06	-1.242338	0.2151
DIABETES	-2.53E-05	2.16E-05	-1.168452	0.2436
STROKE	-4.88E-07	2.00E-05	-0.024430	0.9805
TRAUMA	-2.96E-06	1.01E-05	-0.292094	0.7704

R-squared	0.982172	Mean dependent var	8.084383
Adjusted R-squared	0.981361	S.D. dependent var	0.312007
S.E. of regression	0.042596	Akaike info criterion	-3.428548
Sum squared resid	0.518935	Schwarz criterion	-3.255705
Log likelihood	528.2822	F-statistic	1211.984
Durbin-Watson stat	2.038320	Prob(F-statistic)	0.000000

Table 3e – Two-Stage Least Squares, including life expectancy (each cause separately)

(Note: all variables are in natural logs, except for “CARDIO”, “CANCER”, “DIABETES”, “STROKE”, and “TRAUMA” since there are non-positive observations for those variables; instruments are lagged exogenous variables)

Dependent Variable: LOGTCOST
Method: Two-Stage Least Squares
Date: 02/12/13 Time: 23:58
Sample (adjusted): 1 270
Included observations: 270 after adjustments
LOGTCOST = C(1) + C(2)*LOG(WAGE) + C(3)*LOG(OTH_EXP) + C(4)*LOG(INPAT_DAYS) + C(5)*LOG(OUTPATIENTS) + C(6)*LOG(MEDICAID_DAYS) + C(7)*LOG(MEDICARE_DAYS) + C(8)*LOG(PROPERTY_PLANT_EQUIP) + C(9)*LOG(YEAR) + C(10)*CANCER + C(11)*CARDIO + C(12)*DIABETES + C(13)*STROKE + C(14)*TRAUMA
Instrument list: CONST LOGWAGE LOGOTH_EXP LOGINPAT LOGOUTPAT LOGMEDICAID LOGMEDICARE LOGPROP LOGYEAR LOGWAGE(30) LOGOTH_EXP(30) LOGINPAT(30) LOGOUTPAT(30) LOGMEDICAID(30) LOGMEDICARE(30) LOGPROP(30) LOGYEAR(30)

	Coefficient	Std. Error	t-Statistic	Prob.
C	2.880863	0.622008	4.631556	0.0000
LOG(WAGE)	0.069659	0.072489	0.960966	0.3375
LOG(OTH_EXP)	0.098973	0.061302	1.614504	0.1076
LOG(INPAT_DAYS)	0.266198	0.095521	2.786791	0.0057
LOG(OUTPATIENTS)	0.023311	0.012674	1.839201	0.0670
LOG(MEDICAID_DAYS)	0.005820	0.014433	0.403235	0.6871
LOG(MEDICARE_DAYS)	-0.001186	0.008328	-0.142378	0.8869
LOG(PROPERTY_PLANT_EQUIP)	0.112476	0.028518	3.944094	0.0001
LOG(YEAR)	-0.013565	0.020084	-0.675426	0.5000
CANCER	-9.15E-05	6.82E-05	-1.341647	0.1809
CARDIO	2.67E-06	6.86E-05	0.038844	0.9690
DIABETES	7.37E-05	0.000250	0.294902	0.7683
STROKE	0.000217	0.000166	1.306629	0.1925
TRAUMA	4.55E-05	0.000199	0.228530	0.8194

R-squared	0.969539	Mean dependent var	8.074969
Adjusted R-squared	0.967992	S.D. dependent var	0.309266
S.E. of regression	0.055330	Sum squared resid	0.783727
Durbin-Watson stat	2.118389	Second-stage SSR	0.472393

Table 4a – Semi-Parametric Estimation Results (Dependent Variable: LOG(COST)), b=25%

(Note: all variables are in natural logs)

Variable	Estimate	Std Error	P-Value
Wage	0.295214	0.005411	0.000000
Otherexp	0.378311	0.001916	0.000000
medicaid_days	0.021865	0.000291	0.000000
medicare_days	0.000909	9.52E-05	0.000000
Prop_plant_equip	0.233971	0.000422	0.000000
Case_mix	0.091963	0.003856	0.000000
inpat_days	0.652131	0.000487	0.000000
Outpatients	0.040803	8.32E-05	0.000000
Year	-0.0862	0.001503	0.000000

Number of observations = 300

Table 4b - Semi-Parametric Estimation Results (Dependent Variable: LOG(COST)), b=100%

(Note: all variables are in natural logs)

Variable	Estimate	Std Error	P-Value
Wage	0.260904	0.00469	0.000000
Otherexp	0.266857	0.001051	0.000000
medicaid_days	0.026551	0.000262	0.000000
medicare_days	0.000487	6.94E-05	0.000000
Prop_plant_equip	0.252665	0.000395	0.000000
Case_mix	0.104282	0.002662	0.000000
inpat_days	0.631191	0.000437	0.000000
Outpatients	0.047983	8.71E-05	0.000000
Year	-0.08513	0.000396	0.000000

Number of observations = 300

Table 5a – Semi-Parametric Estimation Results, with life expectancy (Dependent Variable: LOG(COST)), b=25%

(Note: all variables are in natural logs, except for “life” since there are non-positive observations)

Variable	Estimate	Std Error	P-Value
Wage	0.220063	0.005508	0.000000
Otherexp	0.385515	0.001913	0.000000
medicaid_days	0.01096	0.000282	0.000000
medicare_days	0.009247	9.33E-05	0.000000
Prop_plant_equip	0.219716	0.000416	0.000000
Case_mix	0.16291	0.003859	0.000000
inpat_days	0.700151	0.000537	0.000000
Outpatients	0.04521	8.16E-05	0.000000
Year	-0.09061	0.001417	0.000000
Life	-9.44E-06	4.42E-12	0.000000

Number of observations = 300

Table 5b – Semi-Parametric Estimation Results, with life expectancy (Dependent Variable: LOG(COST)), b=100%

(Note: all variables are in natural logs, except for “life” since there are non-positive observations)

Variable	Estimate	Std Error	P-Value
Wage	0.206493	0.004568	0.000000
Otherexp	0.276698	0.000989	0.000000
medicaid_days	0.018046	0.000249	0.000000
medicare_days	0.006688	6.75E-05	0.000000
Prop_plant_equip	0.247024	0.000375	0.000000
Case_mix	0.124345	0.002518	0.000000
inpat_days	0.65969	0.000455	0.000000
Outpatients	0.056368	8.61E-05	0.000000
Year	-0.08301	0.000372	0.000000
Life	-8.58E-06	3.75E-12	0.000000

Number of observations = 300

Table 6a – Semi-Parametric Regression Results, with individual causes of death (Dependent Variable: LOG(COST)), b=25%
(Note: all variables are in natural logs, except for “CARDIO”, “CANCER”, “DIABETES”, “STROKE”, and “TRAUMA” since there are non-positive observations for those variables)

Variable	Estimate	Std Error	P-Value
Wage	0.27256	0.004998	0.000000
Otherexp	0.396504	0.001692	0.000000
medicaid_days	0.034832	0.000251	0.000000
medicare_days	0.013531	9.00E-05	0.000000
prop_plant_equip	0.201074	0.000456	0.000000
case_mix	0.254614	0.004519	0.000000
inpat_days	0.716596	0.000606	0.000000
Outpatients	0.050443	7.78E-05	0.000000
Year	-0.13369	0.001334	0.000000
CARDIO	-6.32E-05	1.82E-10	0.000000
CANCER	-5.49E-06	3.30E-10	0.000000
DIABETES	-0.00013	2.66E-09	0.000000
STROKE	7.13E-05	1.97E-09	0.000000
TRAUMA	-2.07E-06	5.34E-10	0.000000

Number of observations = 300

Table 6b – Semi-Parametric Regression Results, with individual causes of death (Dependent Variable: LOG(COST)), b=100%
(Note: all variables are in natural logs, except for “CARDIO”, “CANCER”, “DIABETES”, “STROKE”, and “TRAUMA” since there are non-positive observations for those variables)

Variable	Estimate	Std Error	P-Value
Wage	0.226228	0.294644	0.000000
Otherexp	0.004378	0.001067	0.000000
Medicaid_days	0.226228	0.294644	0.000000
Medicare_days	0.004378	0.001067	0.000000
prop_plant_equip	0.226228	0.294644	0.000000
case_mix	0.004378	0.001067	0.000000
inpat_days	0.226228	0.294644	0.000000
Outpatients	0.004378	0.001067	0.000000
Year	0.226228	0.294644	0.000000
CARDIO	0.004378	0.001067	0.000000
CANCER	0.226228	0.294644	0.000000
DIABETES	0.004378	0.001067	0.000000
STROKE	0.226228	0.294644	0.000000
TRAUMA	0.004378	0.001067	0.000000

Number of observations = 300

Table 7a – Semi-Parametric Regression Results, with life expectancy (LIFE) estimated as endogenous (Dependent Variable: LOG(COST)), b=25%

(Note: all variables are in natural logs, except for “life” since there are non-positive observations)

Instruments for LIFE variable: lagged exogenous variables

Variable	Estimate	Std Dev	F-Stat	P-Value
constant	5.796137	0.78817	118.1396	0.00000
wage	0.182147	0.28947	5868.383	0.00000
otherexp	0.272274	0.240229	5875.4	0.00000
Medicaid_days	-0.03892	0.056794	3.537644	0.06099
medicare_days	0.019896	0.01851	6.874418	0.0092
Prop_plant_equip	0.23904	0.075884	22.48427	0.00000
case_mix	0.220997	0.123942	4.30486	0.03888
inpat_days	0.722866	0.07055	360.9578	0.00000
outpat_visits	0.075609	0.029732	63.42109	0.00000
year	-0.15454	0.211491	0.904629	0.34234

Endogenous Variable	Estimate	Std Error	t-stat	P-Value
life	-3.08E-05	9.38E-19	-3.28E+13	0.00000

Number of observations = 300

Table 7b – Semi-Parametric Regression Results, with life expectancy (LIFE) estimated as endogenous (Dependent Variable: LOG(COST)), b=100%

(Note: all variables are in natural logs, except for “life” since there are non-positive observations)

Instruments for LIFE variable: lagged exogenous variables

Variable	Estimate	Std Dev	F-Stat	P-Value
constant	5.118832	0.200173	24.08115	0.00000
wage	0.09408	0.234105	3829.965	0.00000
otherexp	0.175179	0.109939	3831.318	0.00000
Medicaid_days	-0.02742	0.010274	5.362194	0.02127
medicare_days	0.030073	0.003707	6.114349	0.01398
Prop_plant_equip	0.211017	0.021672	7.097394	0.00815
case_mix	0.188846	0.035332	3.10461	0.07912
inpat_days	0.784437	0.013058	207.2612	0.00000
outpat_visits	0.114883	0.012378	50.7231	0.00000
year	-0.03455	0.019506	0.520402	0.47125

Endogenous Variable	Estimate	Std Error	t-stat	P-Value
life	-4.74E-05	2.85E-19	-1.67E+14	0.00000

Number of observations = 300

Figure 1: Hospitals in the State of Connecticut, USA (n = 1, 2, …, 30), (t = 1999, 2000, …, 2008)

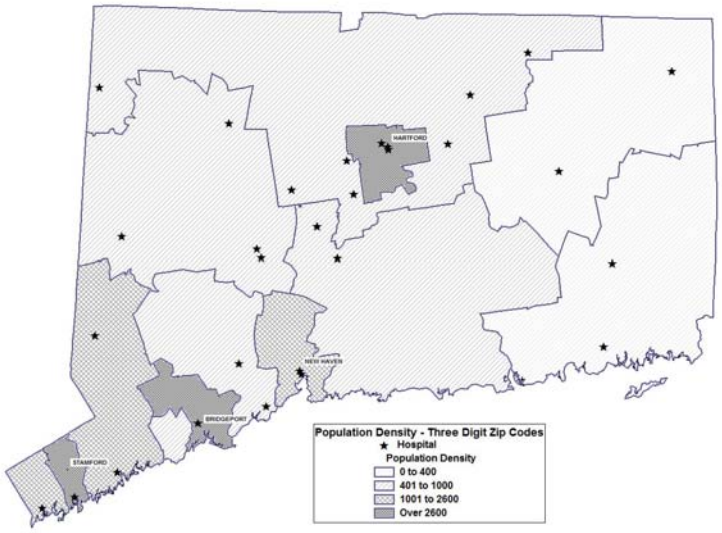

Figure 2a: N=30, T=10

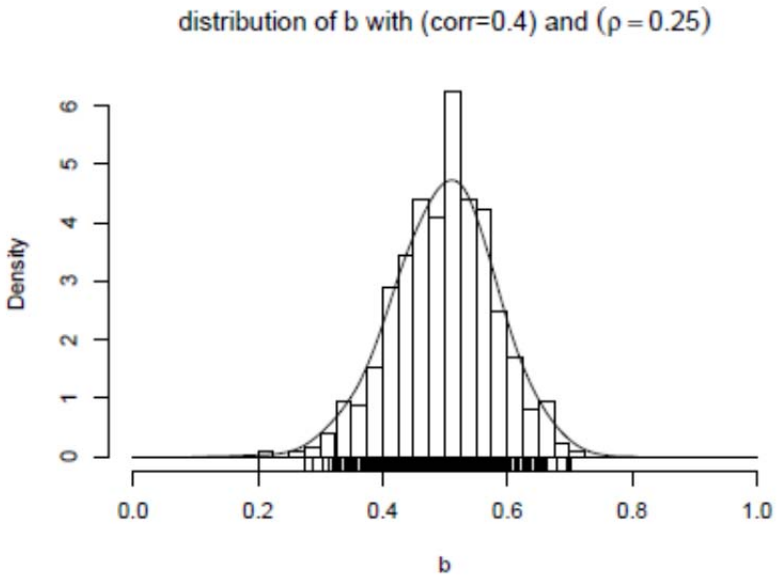

Figure 2b: N=60, T=20

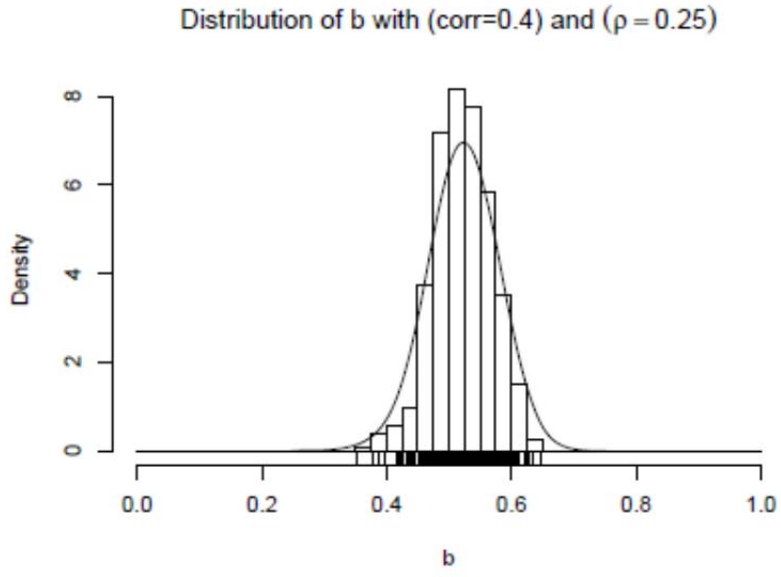

Figure 3a: N=30, T=10

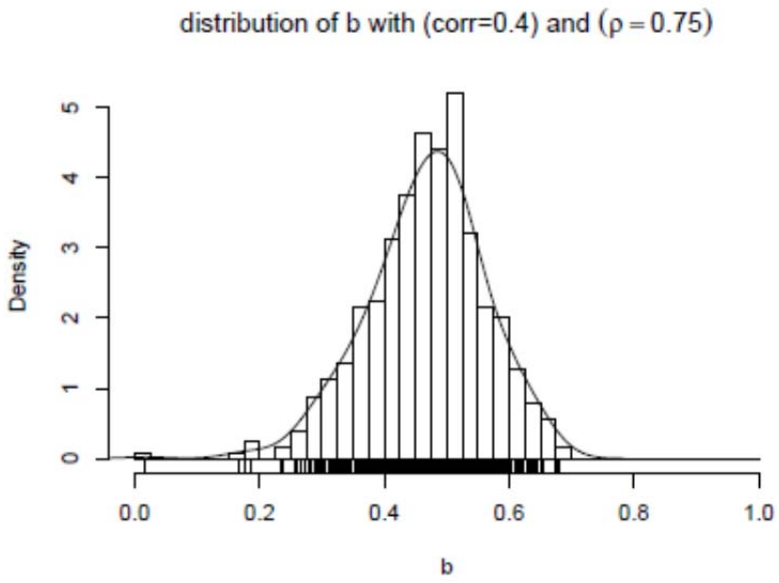

Figure 3b: N=60, T=20

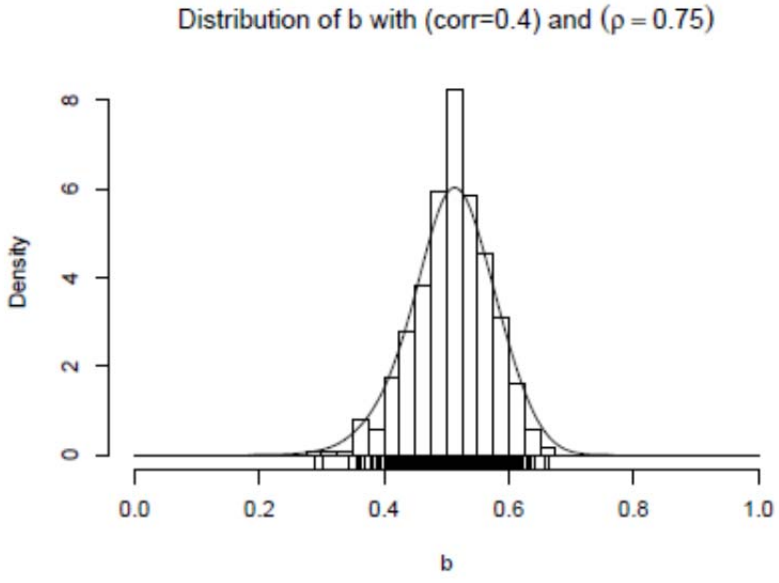

Figure 4a: N=30, T=10

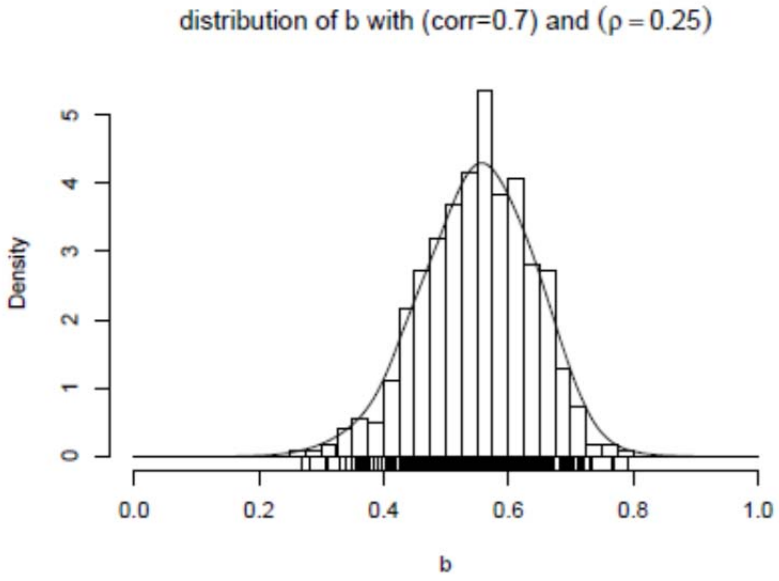

Figure 4b: N=60, T=20

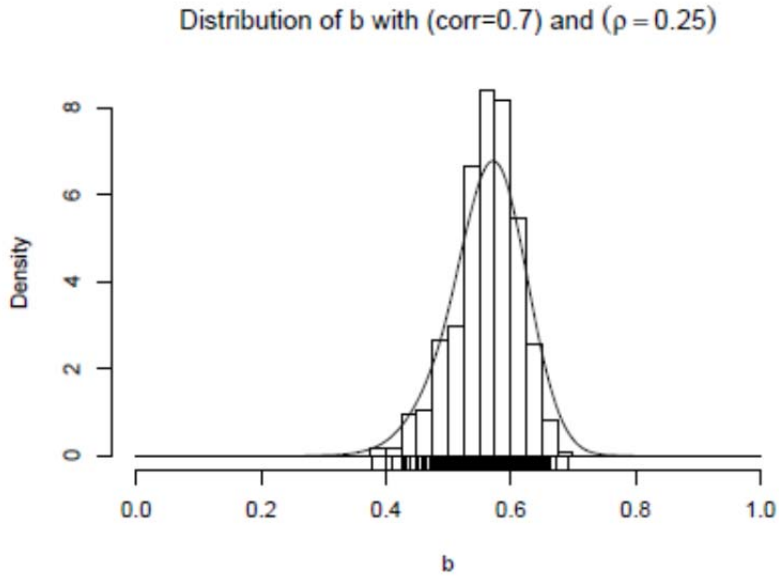

Figure 5a: N=30, T=10

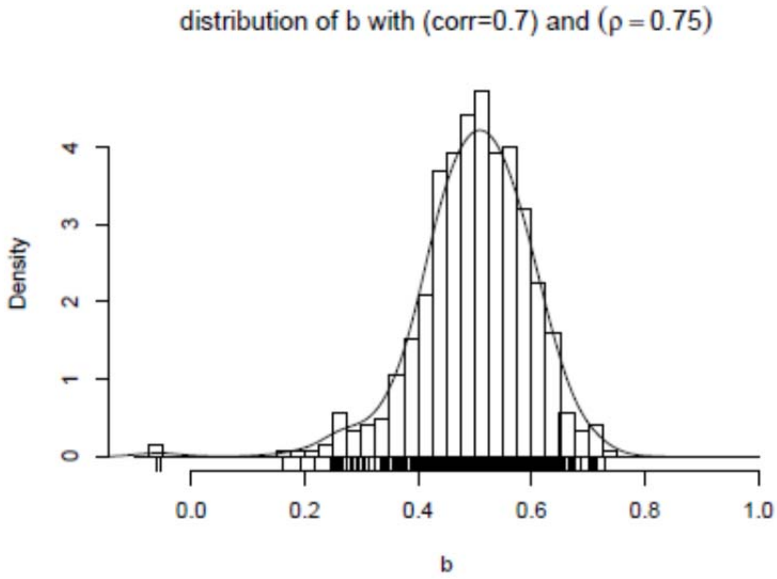

Figure 5b: N=60, T=20