MODELLING SOCIOECONOMIC AND HEALTH DETERMINANTS OF HEALTH-CARE USE: A SEMIPARAMETRIC APPROACH

JÜRGEN MAURER*
Mannheim Research Institute for the Economics of Aging (MEA), University of Mannheim, Mannheim, Germany

SUMMARY
This paper suggests bivariate semiparametric index models as a tool for modelling the interplay of socioeconomic and health characteristics in determining health-care utilisation. These models allow for a fully nonparametric relationship between socioeconomic status, health-care need and care utilisation. The only parametric restriction imposed is that multiple socioeconomic and health indicators can be aggregated into two distinct indices that measure the broader concepts of socioeconomic status and health-care need, respectively. We demonstrate the usefulness of this class of models based on an illustrative empirical example. The estimations highlight complex interactions of socioeconomic status and health-care need in determining care use, which may be difficult to grasp via standard parametric modelling approaches. Copyright © 2007 John Wiley & Sons, Ltd.

Received 9 October 2006; Revised 13 March 2007; Accepted 27 April 2007

KEY WORDS: semiparametric methods; health-care utilisation

INTRODUCTION
Empirical assessments of the effects of socioeconomic status (SES) on health-care use are of great interest in the health policy sphere, as ensuring socioeconomic equity of the health care system is often considered a high priority. Since health care is instrumental for improving and maintaining individual health and functioning, a lack of adequate medical attention among the socially disadvantaged may cause or at least exacerbate socioeconomic gradients in health.1 However, not all individuals require the same intensity of health care. The amount of care delivered to a specific individual should rather depend on her own health characteristics and risk factors as well as on the availability of (cost-effective) treatment technology for improving or managing these. The latter consideration highlights the important distinction between health-care need on the one hand and common notions of ill-health on the other. Particularly, it permits ‘the non-ill to be said to be in need of medical care, in the sense that their health in the future could be better than it would otherwise be if they received (preventive) care now’ (Wagstaff and van Doorslaer, 2000, p. 1813). The allocation of health-care resources should therefore appropriately reflect individual differences in health-care need, but not be dependent on the patients’ SES.

*Correspondence to: Mannheim Research Institute for the Economics of Aging (MEA), University of Mannheim, L13, 17, 68131 Mannheim, Germany. E-mail: maurer@mea.uni-mannheim.de, juergen.maurer@iue.it

The distinction between health-care need and SES as legitimate and illegitimate sources of differences in medical care utilisation is central to almost all empirical studies of equity in health-care utilisation. Although this conceptual dichotomy is omnipresent in equity assessments of health-care delivery, modelling it in a flexible, yet parsimonious way poses several challenges for the applied researcher. Parametric models may often appear too restrictive to incorporate fully flexible SES–need interactions, as they usually impose strong a priori assumptions on permissible functional forms. Without extensive prior knowledge, nonparametric approaches seem better suited for modelling the interplay of SES and health-care need with respect to care use. Yet, actual applications of these methods are frequently impractical. Modelling SES and health-care need typically requires the use of multiple indicators to capture the multi-facetted nature of either concepts, and nonparametric methods have well-known difficulties in handling high-dimensional problems.

This paper suggests the use of bivariate semiparametric index models as a potentially powerful tool for modelling the dichotomy of SES and health-care need in health-care delivery. A blend of parametric and nonparametric approaches, semiparametric index models combine the two in order to mitigate the dimensionality issues of nonparametric modelling, while maintaining its flexibility. Specifically, while this class of models places some parametric structure on the constituents of SES and care need, respectively, it retains a fully nonparametric approach with respect to how these concepts may interact to determine care use. The model is, therefore, especially well-suited for assessing observable heterogeneity in the effects of SES on health-care use across different levels of care need.

The reminder of the paper is organised as follows: the second section reviews parametric and nonparametric strategies for modelling the SES and care need interactions with respect to health-care use, and highlights their respective advantages and disadvantages. The third section then tries to strike a balance between these considerations by introducing the class of bivariate semiparametric index models. The fourth section brings the proposed approach to life by means of a simple empirical illustration. Specifically, we use a bivariate semiparametric model to analyse the effects of education and health-care need on the average number of yearly doctor visits of older Italian men. This section also describes how such a model can be estimated and highlights some of the most important findings revealed by this modelling strategy. The fifth section concludes with a short summary of the paper and potential directions for future research.

MODELLING SES AND HEALTH-CARE NEED INTERACTIONS

This section provides a brief conceptual review of alternative approaches for modelling the SES–need dichotomy in care use. We first consider parametric approaches for modelling the conditional expectations of care use $m_i$ given multiple measures of SES $S_j$ and health-care need $H_i$. We then move to nonparametric generalisations of these approaches, as these do not rely on any a priori restrictions regarding the way $S_j$ and $H_i$ are allowed to affect $m_i$. However, fully nonparametric models are often difficult to employ in practice. Their application to multi-dimensional data requires very large samples, and these may not be available in many applications.

See, e.g. Wagstaff and van Doorslaer (2000) for an overview.

Manski (2005) notes, for example, the importance of assessing the effects of observable heterogeneity for policy targeting. This is due to the fact that policy makers can usually only discriminate between individuals based on observable characteristics in assigning different treatments.

This section gives a review of various well-known modelling strategies and considers some peculiarities of applications in health economics. A more extensive overview of different assumptions implied by parametric, semiparametric and nonparametric modelling approaches can be found in Powell (1994) among others.

Throughout the paper, our discussion will focus on modelling the conditional expectations $E[m_i|S_j, H_i]$. We note, however, that this is only done for expositional ease and because most models concentrate on the conditional mean as their main parameter of interest. In fact, the approaches suggested here can also be generalised to other features of care use – such as access to care – that may be of independent interest (see, e.g. Maurer, 2007 for an example).
Parametric models

In parametric approaches, the researcher assumes a functional form for the relationship between health-care use $m_i$, SES $S_i$, and health-care need $H_i$, which is determined \textit{a priori} up to some finite-dimensional parameter vector $\beta$. This single parameter $\beta$ fully characterises a specific data generating process within the supposed parametric family. A parametric specification for the conditional expectation of $m_i$, given $S_i$ and $H_i$ will therefore have the form

$$E[m_i | S_i, H_i] = m(S_i, H_i, \tilde{\beta})$$

where $m(\cdot, \cdot, \cdot)$ denotes a fixed deterministic family of regression functions assumed by the researcher and $\tilde{\beta}$ a free parameter to be estimated from the data.\footnote{Throughout this paper, we always use tildes to indicate flexible parameters, which are determined in the estimation rather than \textit{a priori}.} Of course, the form of $m(\cdot, \cdot, \cdot)$ may be chosen after prior inspection of the data, but the functional form itself is not allowed to be determined by the data directly. Finding an appropriate parametric specification seems especially difficult for typical health-care data, which feature numerous peculiarities such as ‘excess zeros’ or ‘overdispersion’. As a response, a myriad of different functional forms for $m(\cdot, \cdot, \cdot)$ has been used in practice. These range from simple linear regression models over more sophisticated nonlinear specifications such as the negative binomial model up to even more complex approaches like two-part models, which combine different parametric families such as logit or probit models with say truncated negative binomial models to obtain an expression for $m(\cdot, \cdot, \cdot)$ that depends on a finite-dimensional parameter vector $\beta$ only.\footnote{See, e.g. Jones (2000) for an overview.} Such parametric approaches have some well-known merits and shortcomings. Particularly, if the functional form $m(\cdot, \cdot, \cdot)$ is correctly specified, parametric maximum likelihood estimators for $\beta$ will be root-N-consistent, asymptotically normal and efficient. If the specification of $m(\cdot, \cdot, \cdot)$ is incorrect, however, the resulting estimate $m(S_i, H_i, \tilde{\beta})$ will generally be inconsistent for $E[m_i | S_i, H_i]$ and may therefore be seriously misleading about how average levels of care use vary by SES and care need. It is easy to see that this issue is indeed a cause for concern for the applied researcher. The multitude of different models used in the past already highlights that an \textit{a priori} picking of the correct functional form for $m(\cdot, \cdot, \cdot)$ is far from trivial. In this context, it is also worth noting that choosing $m(\cdot, \cdot, \cdot)$ does not only require deciding on the right distributions for the stochastic error terms, but also taking a stance on how potentially important interactions between $S_i$ and $H_i$ can be accurately reflected in the model. The various different treatments adopted in the previous literature again illustrate the difficulties in resolving these issues using prior reasoning alone. For example, some studies simply impose additivity in $S_i$ and $H_i$, and thereby rule out any SES–need interactions in the use of health care. Yet, other approaches include linear interactions in the model to allow for some linear interdependence in the way $S_i$ and $H_i$ affect $m_i$, whereas a third strand of the literature employs arbitrarily discretised values of $S_i$ and then runs separate regressions for each SES group, corresponding to a model with full SES–need interactions at the SES-group level, however these are defined.\footnote{See, e.g. Jones (2000) and Wagstaff and van Doorslaer (2000) for further references.}

The strong requirement of extensive prior information as well as the arbitrariness of parametric assumptions have long-been identified as key shortcoming of parametric approaches. As McFadden notes in his famous 1985 presidential address to the Econometric Society, parametric modelling ‘interposes an untidy veil between econometric analysis and the propositions of economic theory, which are most abstract without specific dimensional or functional restrictions’. This argument seems to apply with even stronger force to the empirical analysis of equity in health-care utilisation, since there is rarely any formal theory that could be used to justify specific parametric restrictions \textit{ex ante}.

---

\footnotesize


\footnotesize

\textit{DOI: 10.1002/hec}
Nonparametric models

Fully nonparametric models resolve the arbitrariness of parametric approaches, which stems from the need to choose a functional form for \( m(\cdot, \cdot, \cdot) \) \textit{ex ante}. Formally, a model for the conditional expectation of care use \( m_i \), given SES \( S_i \), and care need \( H_i \) can be written as

\[
E[m_i|S_i, H_i] = \hat{m}(S_i, H_i)
\]

Rather than restricting \( m(\cdot, \cdot, \cdot) \) to lie in some \textit{a priori} specified parametric family, nonparametric methods treat \( \hat{m}(\cdot, \cdot) \) as an infinite-dimensional unknown parameter to be estimated from the data. Apart from mild regularity conditions, these methods do not impose any restrictions on the relationship between \( S_i, H_i \) and \( E[m_i|S_i, H_i] \). Thus, ‘the main strength of nonparametric over parametric regression is the fact that it assumes no functional form for the relationship, allowing the data to choose, not only the parameter estimates, but the shape of the curve itself’ (Deaton, 1997, p. 193). A variety of approaches, such as kernel or series methods can be used to estimate \( \hat{m}(\cdot, \cdot) \).\(^9\) While such a general estimation approach provides a desirable safeguard against the potential adverse effects of parametric misspecification, the prices of this added flexibility and robustness are much greater data requirements for actual implementation. Particularly, the precision of fully nonparametric estimators is often poor and their rate of convergence is usually slower than in parametric models. This is especially true in higher-dimensional problems like the one considered here, where both SES and health-care need may only be measurable via multiple indicators. Hence, there is a practical trade-off between the use of parametric or nonparametric methods for estimation based on finite samples. Given this trade-off, it would seem desirable to combine the two approaches to retain some flexibility and robustness in modelling the function \( E[m_i|S_i, H_i] \), but mitigate the ‘curse of dimensionality’ associated with nonparametric regression. Semiparametric models do this by introducing parametric components to the model to attain some dimensionality reduction. Of course, the plausibility of such an approach depends on whether such parametric elements seem justifiable by theory, as indicated by the McFadden quote. We use the nature of the policy discourse to motivate a bivariate semiparametric framework for modelling \( E[m_i|S_i, H_i] \), which ought to strike a reasonable balance between precision, flexibility and coherency with theory.

BIVARIATE SEMIPARAMETRIC INDEX MODELS

Moving from fully nonparametric to semiparametric regression requires the introduction of some theoretically justifiable parametric elements in (2). At first, this appears challenging, as we have not provided a formal model on how SES and health-care need bring about specific levels of care use. Moreover, looking at the previous literature as well as the policy debate, the only apparent theoretical differentiation is their categorisation into legitimate and illegitimate sources of differences in medical care utilisation, what we have labelled SES \( S_i \) on the one hand, and health-care need \( H_i \) on the other. At the conceptual level, the discourse, therefore, appears merely dichotomous, with additional complexities solely arising from the use of multiple measures to capture ‘need’ and ‘non-need’ determinants of health-care use. In other words, while SES and health-care need appear sufficiently well-defined for theory, they seem inherently difficult to actualise when bringing theory to the data. The class of bivariate semiparametric index models suggested here reflects this dichotomous structure of the debate – aggregating illegitimate and legitimate sources of differences in care use.

\(^9\) See, for example Härdle (1990), Härdle and Linton (1994), Deaton (1997) or Yatchew (2003) for accessible introductions to the topic.
into one-dimensional concepts of SES and care need, respectively. The analysis then proceeds fully nonparametrically when estimating how expected care use varies with SES and health-care need. Specifically, it does not impose any additional functional form assumption on $m(\cdot, \cdot)$ beyond the two index restrictions that allow aggregation of the multiple SES and care need indicators.

To understand how this procedure works, assume that the concept of SES can be measured as a linear combination of say education, income or wealth, while health-care need may be measured as a linear combination of multiple need indicators such as age, diseases, functioning measures, respectively. We can then represent SES and health-care need as two indices $I_i^{\text{SES}}$ and $I_i^{\text{HCN}}$ constructed as a linear combination of the multiple SES and need indicators $S_i$ and $H_i$ that are measured in the data. We thus obtain

$$I_i^{\text{SES}} = S_i \delta$$

$$I_i^{\text{HCN}} = H_i \beta$$

where $\delta$ and $\beta$ denote specific aggregation parameters pertaining to the broader concepts of SES and health-care need, which are now measured via $I_i^{\text{SES}}$ and $I_i^{\text{HCN}}$, respectively. Assuming that expected health-care utilisation $E[m_i|S_i, H_i]$ depends on $S_i$ and $H_i$ through the two one-dimensional indices $I_i^{\text{SES}}$ and $I_i^{\text{HCN}}$ only, we obtain a semiparametric model of the form

$$E[m_i|S_i, H_i] = E[m_i|\tilde{I}_i^{\text{SES}}, \tilde{I}_i^{\text{HCN}}] = \tilde{m}(\tilde{I}_i^{\text{SES}}, \tilde{I}_i^{\text{HCN}}) = \tilde{m}(S_i, \delta, H_i, \beta)$$

where $\tilde{\delta}, \tilde{\beta}$ and $\tilde{m}(\cdot, \cdot)$ denote unknown (finite- and infinite-dimensional) parameters that need to be estimated from the data. The only parametric restrictions involved in this model are the index assumptions (3) and (4). Particularly, the approach remains fully nonparametric with respect to $\tilde{m}(\cdot, \cdot)$, and does therefore not incorporate any a priori constraint on how $I_i^{\text{SES}}$ and $I_i^{\text{HCN}}$ affect $E[m_i|S_i, H_i]$.

Being a hybrid of parametric and nonparametric regression, the semiparametric double index model in (5) features obvious similarities with both approaches. Like in parametric models, (5) includes some finite-dimensional unknown parameter vectors $\tilde{\delta}$ and $\tilde{\beta}$. Beyond $\tilde{\delta}$ and $\tilde{\beta}$, however, the model also includes the infinite-dimensional parameter $\tilde{m}(\cdot, \cdot)$ which represents a fully flexible link function for mapping SES $I_i^{\text{SES}}$ and care need $I_i^{\text{HCN}}$ into expected care use $E[m_i|S_i, H_i]$. Like in the nonparametric approach, the functional form of $\tilde{m}(\cdot, \cdot)$ is not at all constrained a priori, but, in fact, is estimated from the data. One important advantage of this modelling strategy is that it flexibly incorporates observable heterogeneity in the effects of SES $I_i^{\text{SES}}$ for different levels of health-care need $I_i^{\text{HCN}}$. Hence, SES gradients are allowed to vary freely across the need distribution, which may reveal some useful information for policy design.

The class of semiparametric index models does, however, not only inherit the advantages of both approaches, but also their disadvantages. On the one hand, semiparametric estimators for $E[m_i|S_i, H_i]$ are generally consistent under a wider range of circumstances than their parametric counterparts and therefore more robust to potential misspecification. At the same time, they are usually more precise than their nonparametric counterparts, due to the built-in dimensionality reduction implied by the index restrictions. Specifically, the parametric components of the model, $\tilde{\delta}$ and $\tilde{\beta}$, can be estimated with the usual parametric rate of convergence, whereas the estimate for the conditional

---

10Ichimura and Lee (1991) have introduced multiple index model into the theoretical econometrics literature, which have been subsequently applied to diverse research questions with bipartite structures such as supply and demand (Stern, 1996; Maurer and Pohl, 2007) or interactions of macro and micro determinants of the income distribution (Farré-Ollalla and Vella, 2006) to name just a few.
expectation $\tilde{m}(\bar{S}, \tilde{H}, \tilde{\beta})$ converges at the (slower) rate of a nonparametric estimate of a conditional mean function with two arguments, and thus much faster than if $S_i$ and $H_i$ would be treated fully nonparametrically. On the other hand, semiparametric estimators may be considerably less efficient than their fully parametric counterpart, at least if the latter can be correctly specified based on prior information. Also, unlike the fully nonparametric approach, semiparametric estimation may still lead to inconsistent estimates, if the parametric part of the model – here the two index restrictions – are inaccurate. We, therefore, consider semiparametric models as complementary to the other two approaches.

EMPIRICAL ILLUSTRATION

This section presents an empirical example to bring the above concepts to life, and shows the applicability of such semiparametric index models to typical survey data on health-care use.

Data

To illustrate the kinds of insights that a semiparametric double index approach may deliver, we estimate a model for health-care utilisation of older Italian men using data from the first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE) collected in 2004. SHARE is a multidisciplinary, cross-national micro database containing information on health and SES of some 22,000 Continental Europeans aged 50+ from 10 European countries. Yet, for the sake of this simple illustration, we restrict our sample to the 1017 male respondents from Italy.

Model specification

We measure health-care utilisation as the total number of doctor visits during the last 12 months. This measure includes both visits to GPs as well as specialists, but does not account for inpatient care as a potential substitute. For SES, we use years of education as our proxy variable. As we are considering an elderly population, alternative SES measures such as income or wealth appear heavily confounded by the respondent’s labour market status as well as typical life-cycle trajectories of asset holdings, respectively. It is, however, important to note that the suggested model could easily incorporate multiple SES indicators if desired. Specifically, these indicators would then form an actual SES index, replacing our one-dimensional SES measure based on education. The corresponding index coefficients would then be estimated in the same way as the index coefficients of the care need index, to which we turn now. Our model uses a large number of health indicators to comprehensively capture individual differences in health-care need. In fact, the abundance of health measures available in SHARE makes it an ideal data source for the kind of exercise considered here.

Apart from age and dynamometer-measured maximum grip strength, we include a set of 15 binary indicators for different doctor-diagnosed health conditions. These ought to capture various aspects of

---

11 This paper uses data from the early release 1 of SHARE 2004. This release is preliminary and may contain errors that will be corrected in later releases. The SHARE data collection has been primarily funded by the European Commission through the fifth framework programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life). Additional funding came from the US National Institute on Ageing (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064). Data collection in Austria (through the Austrian Science Foundation, FWF), Belgium (through the Belgian Science Policy Office) and Switzerland (through BBW/OFES/UFES) was nationally funded. The SHARE data set is introduced in Börsch-Supan et al. (2005); methodological details are contained in Börsch-Supan and Jürges (2005).

12 Results from a more comprehensive semiparametric cross-country comparison of care utilisation using data for both sexes and all 10 initial SHARE countries can be found in Maurer (2007).

13 See Maurer (2007) for a more detailed justification of this choice. Banks et al. (2002) provide a more comprehensive theoretical discussion of the issues involved. An empirical assessment can, for example, be found in de Vos (2005).
individual health-care need, including acute conditions such as heart attacks or stroke as well as chronic
diseases like diabetes or mere risk factor such as hypertension, all of which require a different form and
intensity of disease management. Table I presents the basic descriptive statistics for all variables used in
the analysis. The unconditional mean of the number of doctor visits in the last 12 months is 7.23:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard error</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of doctor visits</td>
<td>7.230</td>
<td>11.280</td>
<td>0</td>
<td>98</td>
</tr>
<tr>
<td>Years of education</td>
<td>7.600</td>
<td>4.340</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Age</td>
<td>64.752</td>
<td>8.580</td>
<td>50</td>
<td>94</td>
</tr>
<tr>
<td>Maximum grip strength (in kg)</td>
<td>39.942</td>
<td>10.680</td>
<td>7</td>
<td>70</td>
</tr>
<tr>
<td>Asthma</td>
<td>0.044</td>
<td>0.206</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cancer</td>
<td>0.025</td>
<td>0.155</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cataracts</td>
<td>0.048</td>
<td>0.214</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>0.176</td>
<td>0.381</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.122</td>
<td>0.327</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Heart attack</td>
<td>0.112</td>
<td>0.316</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hip fracture</td>
<td>0.012</td>
<td>0.108</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.354</td>
<td>0.478</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Lung disease</td>
<td>0.080</td>
<td>0.271</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>0.015</td>
<td>0.121</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Parkinson</td>
<td>0.005</td>
<td>0.070</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Stroke</td>
<td>0.030</td>
<td>0.172</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ulcer</td>
<td>0.078</td>
<td>0.268</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Other condition</td>
<td>0.127</td>
<td>0.333</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Number of observations</td>
<td>1021</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimation

As our empirical illustration employs many health-care need indicators but only one SES measure,\textsuperscript{14} the
general double index model for estimating $E[m|S, H, \beta]$ based on (5) simplifies to

$$E[m|S, H] = \tilde{m}(S, H|\beta)$$

and thus requires estimation of only one index parameter $\beta$ in addition to nonparametric link function
$\tilde{m}(\cdot, \cdot)$.\textsuperscript{15}

We use Ichimura and Lee’s (1991) multiple index extension of Ichimura’s (1993) semiparametric
least-squares estimator for single index models to estimate the index coefficient $\beta$ in (6) along with
the nonparametric link function $\tilde{m}(\cdot, \cdot)$. Specifically, Ichimura and Lee’s estimator for $\beta$ is based on

\textsuperscript{14}From the derivations that follow, it is straightforward to see how the estimation can also handle richer index specifications for $S$. Yet, for reasons outlined in the specification section, as well as in Maurer (2007), we prefer the specification based on
education alone, which seems still sufficient to illustrate the main features of bivariate semiparametric models.

\textsuperscript{15}Note that, as always in semiparametric index models, separate identification of the components in $\tilde{m}(S, H|\beta)$ can only
be attained up to location and scale (see e.g. Horowitz, 1998). This is why the first argument of $\tilde{m}(\cdot, \cdot)$ does not contain any parameter
and there is also no intercept in $\beta$ (to normalise location) while its first element is set equal to one (to normalise scale).
minimising a semiparametric least-squares criterion function of the form

$$SSR(\hat{\beta}) = \sum_{i=1}^{N} (m_i - \hat{m}(S_i, H_i \hat{\beta}))^2$$

(7)

where an estimate of $\hat{m}(S_i, H_i \hat{\beta})$ for any given candidate parameter vector $\hat{\beta}$ is constructed via bivariate nonparametric kernel regression of $m_i$ on $S_i$ and $H_i \hat{\beta}$, respectively.\(^{16}\) The estimate of $\hat{m}(S_i, H_i \hat{\beta})$ is thus computed via

$$\hat{m}(S_i, H_i \hat{\beta}) = \frac{\sum_{j=1}^{n} K_h(X_i - X_j)m_j}{\sum_{j=1}^{n} K_h(X_i - X_j)}$$

(8)

with $X_i = (S_i, H_i \hat{\beta})$, i.e. a two-dimensional vector consisting of SES $S_i$ and health-care need $H_i \hat{\beta}$, respectively, and $K_h(X_i - X_j) = \det(h_i)^{-1} \cdot K(h_i^{-1}(X_i - X_j))$ for some bivariate kernel function $K(\cdot)$ and matrices of local bandwidths $h_i$ regulating the degree of smoothing in the two directions of the $(S_i, H_i \hat{\beta})$-space.

### Selected results

As highlighted in the above discussion, the semiparametric modelling approach features two main estimation parameters – the finite-dimensional parameter $\hat{\beta}$ as well as the infinite-dimensional parameter $\hat{m}(\cdot, \cdot)$. Our presentation of the results reflects this bipartite structure, starting with a brief discussion of what we estimate as health-care need $H_i \hat{\beta}$ before turning to the semiparametric estimate for the conditional expectations $E[m_i | S_i, H_i]$ that also incorporates an estimate $\hat{m}(\cdot, \cdot)$ of the nonparametric link function $\hat{m}(\cdot, \cdot)$.

Table II presents the estimates for the parameter vector $\hat{\beta}$, which aggregates all health indicators into the one-dimensional health-care need index $H_i \hat{\beta}$. Of course, we cannot investigate the effects of these health controls on actual care utilisation without knowledge of the unspecified mapping $\hat{m}(\cdot, \cdot)$.

---

\(^{16}\)The actual estimator requires some additional adjustments such as trimming of the criterion for observations where the data are sparse as well as the use of so-called higher-order kernels or local bandwidth selection (local smoothing). While the actual estimations incorporate trimming and local smoothing, these technicalities are omitted from the discussion here for the sake of brevity and expositional ease.
However, we can check whether their aggregation into a single care need index is consistent with our prior expectations regarding their relative signs. For identification, the index does not include an intercept and we have normalised the coefficient of age to 0.01 to fix its location and scale. Given this normalisation, we would expect that all of our health controls, with the exception of maximum grip strength, enter the model with a positive sign. In fact, a remarkably consistent pattern emerges. Almost all of our health conditions enter the model positively, whereas maximum grip strength has indeed the expected negative coefficient. Specifically, the only health conditions that enter the model with a very small and statistically insignificant negative coefficient are ‘having been diagnosed with high cholesterol’ and ‘having ever been diagnosed with cataracts’, two conditions for which we would not also have expected a large effect on care need. Also, the relative sizes of the estimated coefficients seem largely in line with prior expectations. Given this encouraging first glimpse on the

\[ \text{(a)} \]

\[ \text{(b)} \]

Figure 1. Bivariate density estimates for the controls: (a) surface plot and (b) contour plot

---

17 This presumes, of course, that both higher age and the prevalence of a health condition indicate more care need, whereas higher grip strength indicates less need for medical attention. So far, we do not know how the care need index maps into actual utilisation, but we can assess this relationship once we turn to our estimate of \( \tilde{m}(\cdot, \cdot) \).

18 This is especially true for the latter, which can basically be completely cured by one-time surgery.
constituents of care need, we can now turn to a more comprehensive assessment of the effects that our health controls on actual care use. Also, and arguably more interestingly, we can assess the effect of our SES measure – education – on medical care utilisation, and how its effects vary across the distribution of healthcare need. To do so, we present our estimate of the nonparametric function \( \hat{m}(.,.) \) which links years of education \( S_i \) and the aggregated care need index \( H_i \beta \) with the expected number of doctor visits.

Since semi- and nonparametric methods leave it to the data to choose the shape of the regression function \( \hat{m}(.,.) \), we cannot, of course, come up with a corresponding estimate \( \hat{m}(.,.) \) out of sample. It is, therefore, important to first clarify the relevant support of \( S_i \) and \( H_i \beta \) over which we can estimate the conditional expectation function \( \hat{m}(.,.) \) nonparametrically. Figure 1 presents bivariate density estimates for the joint distribution of \( S_i \) and \( H_i \beta \) to highlight the relevant support of the data. The joint distribution of \( S_i \) and \( H_i \beta \) in the sample is concentrated at five to six years of education and a care need index value of around 0.4, but covers a fairly wide range of educational attainment and health-care need. In the discussion of our estimates, we will make sure to just consider points that lie well inside the support of the data to avoid invalid out-of-sample prediction as well as spurious results due to a lack of sufficient data in the tails.

Figure 2 presents our semiparametric estimate of the conditional expectations function \( E[m_i|S_i, H_i] \), which is the main object of interest in this study. Exploiting the semiparametric structure of (6), the two
plots display $\hat{m}(\cdot, \cdot)$ as a function of its two arguments, education $S_i$ and health-care need $H_i\hat{\beta}$. The figure reveals several interesting patterns regarding the interplay of $S_i$ and $H_i\hat{\beta}$ in the determination of care use. Particularly intriguing is the vast observable heterogeneity in the effects of educational attainment across different points in the care need distribution. For low levels of care need, say $H_i\hat{\beta} = 0.1$, the average number of doctor visits is only slightly increasing with increasing levels of educational attainment. Among these relatively healthy respondents, we estimate only a small positive education gradient, with conditional means of $\hat{m}(2, 0.1) = 3.12$ for two years of education and $\hat{m}(14, 0.1) = 3.88$ for 14 years of education, respectively. Yet, the effects of education change dramatically as we move along the distribution of care need. Specifically, the slightly positive education gradient gradually reverses, being almost zero at $H_i\hat{\beta} = 0.44$ before reemerging as a pronounced negative gradient for those with higher levels of care need. For example, for respondents with a care need level of $H_i\hat{\beta} = 1$, the average number of doctor visits in the past year varies from 13.26 to 9.46 for two and 14 years of education, respectively. We thus find a fairly strong dependence of the effects of SES on care use for different levels of care need, indicating important SES–need interactions that seem to call for explicit consideration when modelling SES gradients in care use.

CONCLUSION

This paper suggests semiparametric double index models as a potentially valuable tool for applied researchers analysing the interplay of SES and health-care need in determining medical care utilisation. The previous literature has mostly employed parametric methods, in which the functional form of the regression function is a priori restricted by the researcher. While these approaches surely have certain advantages in terms of estimation efficiency and inference if correctly specified, they are also known to perform poorly if the assumed parametric structure is inaccurate. Semiparametric approaches on the other hand, combine parametric modelling with nonparametric estimation. This approach seems especially advantageous if there is little a priori knowledge about potentially complex features in the data, but some parametric structure can nonetheless be justified by theory. The semiparametric double index approach suggested here is motivated by these observations. With regard to the micro-determinants of health-care use, almost all studies feature a bipartite conceptual distinction between SES and care need as illegitimate and legitimate sources of differences in medical care utilisation, even if either concept can only be actualised via multiple proxies. The parametric component of our suggested modelling approach therefore assumes that multiple SES and care need indicators can be aggregated into a single SES and health-care need index, respectively. Given these one-dimensional SES and care need measures, the analysis proceeds fully nonparametrically. Particularly, we do not assume any specific functional form with respect to how SES and care need interact in bringing about certain intensities of care use. Our approach thus allows for observable heterogeneity in SES gradients across the care need distribution, which may deliver important insights for targeting health policies.

We demonstrate how the suggested method works based on actual data on health-care use of older Italian men taken from the first wave of SHARE. Particularly, our example considers a typical research question in equity analysis, namely, the conditional effects of education on health-care use, given multiple controls for care need. Our illustration turns out to be an interesting case of partially offsetting education gradients which vary considerably across the distribution of health-care need. Particularly, our estimates indicate that the highly educated healthy respondents consume more care than their less educated counterparts. The education gradient then gradually reverses with the less educated using considerably more care when sick than the well-educated respondents. Similar to Abasolo et al. (2001), we can interpret our regression results in terms of inequity in health-care delivery. Yet, our approach has the additional advantage that we can explicitly consider how SES gradients may vary across the
need distribution. Our example has highlighted that such observable heterogeneity may indeed be important and by itself informative for targeting policies. In addition, it may also pose some challenges for conventional parametric approaches as to how to account for such heterogeneity.

Admittedly, the paper leaves some open issues for future research. Firstly, it would be most useful to also attempt an empirical comparison of the various parametric models with the semiparametric approach suggested here, which may deliver further insights on the practical relevance of some of the theoretical concerns that have served as its motivation. Secondly, investigating the usefulness of the proposed estimator for formal approaches of measuring and testing for inequity in health-care delivery seems another promising route for further work. As it stands, we nonetheless deem semiparametric estimation as a promising tool for flexibly modelling SES–need interactions in the delivery of care, even if additional evidence on its relative performance is surely desirable.

ACKNOWLEDGEMENTS

I would like to thank Roger Klein and Frank Vella for sharing some of their GAUSS routines with me. I would also like to thank Martin Salm, Emily Sinnott, seminar participants at the University of Mannheim, participants of the 15th European Workshop on Econometrics and Health Economics in Thessaloniki (especially my discussant, Eddy van Doorslaer), the editor, Andrew Jones, as well as an anonymous referee for very helpful comments and suggestions. Of course, all remaining errors are my own responsibility.

REFERENCES


Maurer J, Pohl V. 2007. *Who has a Clue to Preventing the Flu?* MEA, Mannheim.


