UNIVERSITY of York

HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

WP 20/03

Modelling healthcare costs: a semiparametric extension of generalised linear models

Jia Chen: Yuanyuan Gu; Andrew M. Jones and Bin Peng

February 2020

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

Modelling healthcare costs: a semiparametric extension of generalised linear models

Jia Chen[1], Yuanyuan Gu*[1,2], Andrew M. Jones[1,3], and Bin Peng[4]

[1]Department of Economics and Related Studies, University of York

[2]Centre for the Health Economy, Macquarie University

[3]Centre for Health Economics, Monash University

[4]Department of Economics, Deakin University

January 31, 2020

*Correspondence: yuanyuan.gu@mq.edu.au. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.740654. We thank Tingting Cheng, Nigel Rice, James Lomas, Manuel Gomes, Colin Cameron, Debdeep Chattopadhyay, Eugenio Zucchelli and participants at Health Economists' Study Group Winter 2019 Meeting and 28th European Workshop on Econometrics and Health Economics for their assistance, insights and comments.

Abstract

The empirical and methodological efforts in using the generalised linear model to model healthcare costs have been mostly concentrated on selecting the correct link and variance functions. Another type of misspecification - misspecification of functional form of the key covariates - has been largely neglected. In many cases, continuous variables enter the model in linear form. This means that the relationship between the covariates and the response variable is entirely determined by the link function chosen which can lead to biased results when the true relationship is more complicated. To address this problem, we propose a hybrid model incorporating the extended estimating equations (EEE) model and partially linear additive functions. More specifically, we partition the index function in the EEE model into a number of additive components including a linear combination of some covariates and unknown functions of the remaining covariates which are believed to enter the index non-linearly. The estimator for the new model is developed within the EEE framework and based on the method of sieves. Essentially, the unknown functions are approximated using basis functions which enter the model just like the other predictors. This minimises the need for programming as the estimation itself can be completed using existing EEE software programs. The new model and its estimation procedure are illustrated through an empirical example focused on how children's Body Mass Index (BMI) z-score measured at 4-5 years old relates to their accumulated healthcare costs over a 5-year period. Results suggest our new model can reveal complex relationships between covariates and the response variable.

Keywords: Body Mass Index; Extended estimating equations; Generalised linear model; Healthcare cost; Sieve estimation

JEL Classification: C14, I10, P46

1 Introduction

Econometric modelling of healthcare costs serves many purposes: to obtain key parameters in cost-effectiveness analyses (Hoch et al. 2002); to implement risk adjustment in insurance systems (van de Ven & Ellis 2000); and to estimate health care costs attributable to risk factors such as smoking and obesity (Johnson et al. 2003, Cawley & Meyerhoefer 2012). Modelling healthcare costs is challenging because the cost data are typically non-negative, heavy tailed and highly skewed. Early efforts centred on linear regressions with a transformed cost dependent variable. However, with transformed costs, it is necessary to undertake retransformation to obtain predictions and marginal effects of predictors on the original cost scale. The retransformation can be a cumbersome process, especially when heteroscedasticity remains after costs are transformed (Manning 1998, Manning & Mullahy 2001).

Researchers have therefore sought alternative methods to avoid the need to transform costs, and identified the generalised linear model (GLM) as a preferred approach (e.g., Blough et al. 1999). The GLM is built around a link function that specifies the relationship between the conditional mean and a linear function of the covariates (i.e., the index) and a distributional family that specifies the form of the conditional variance as a function of the conditional mean. Apart from being able to model cost on its original scale, the GLM approach also has two advantages: it gains in efficiency (precision) if the estimator matches the data generating process and it provides consistent estimates even if the distribution family is incorrectly specified (i.e., the choice of family only influences efficiency as long as the link function and covariates are correctly specified). However, it is also known that GLM can suffer substantial efficiency losses if data are heavily tailed or the variance function, represented by the distribution family, is misspecified (Jones 2011).

These features of GLMs have motivated efforts to avoid the misspecification of the link and variance functions. For example, a nonparametric GLM was proposed by Chiou & Müller (1998) where both link and variance functions are unknown but smooth functions. This has been rarely used in the healthcare costs modelling literature though, most likely because there is difficulty in implementing the approach when there are a large number of dummy or discrete regressors which is often the case in health economics applications.

Another example is the extended estimating equations (EEE) model of Basu & Rathouz (2005) which is a method to estimate a semiparametric GLM where the index is specified as a Box-Cox transformation of the conditional mean, and the variance function can be either a power function or quadratic function. Whilst not fully nonparametric, the EEE model offers a great deal of flexibility in the choice of the link and variance functions. This has made it increasingly popular in health economics research.

A less explored topic in this area is the potential misspecification of the functional form of variables included in the index. In many cases, continuous variables enter the index in linear form. This means the relationship between the covariate and the dependent variable will be entirely determined by the link function chosen which can lead to biased results when the true relationship is more complicated than the one indicated by the chosen link function. Often, researchers try to use polynomials to ameliorate this problem but this method is limited in the type of functional forms it can accommodate largely due to its undesirable "nonlocal" behaviour (Magee 1998). Another popular approach is to discretise the regressors. However, it may be difficult to identify the relevant cut-off points and the approach may lose substantial information.

An alternative and possibly superior approach is to enter these variables into the index nonparametrically. This leads to the partially linear additive model when it is applied to a linear regression model (Engle et al. 1986) and the generalised (partially linear) additive model (GAM) when it is applied to a GLM (Hastie & Tibshirani 1986). Partially linear models have been applied in health economics (Jones 2000), but not to healthcare costs modelling and GLMs in particular.

In this paper, we extend the EEE model by considering a partially linear additive index function with selected continuous variables entering the index nonparametrically. This represents a marriage of the two lines of research described above resulting in a highly flexible model that can potentially avoid three types of misspecification. The proposed model is a semiparametric extension of GLM with GLM, EEE and GAM as its special cases.

The estimator is developed within the EEE framework by Basu & Rathouz (2005) and based on the method of sieves (Chen 2007). Essentially, the unknown functions in the index are approximated using sets of basis functions which enter the model just like the other predictors. This minimises the need for programming as the estimation itself can be completed using existing EEE software modules (e.g., Basu 2005). The confidence intervals for the unknown function and marginal effects can be obtained using the bootstrap method after the model is estimated.

The method is illustrated through an empirical example that analyses how children's Body Mass Index (BMI) z-score measured at 4-5 years old is related to their medical services costs over a 5-year period. This is a modified replication of Au (2012) in which all the predictors were dummy coded. We considered two continuous variables, the BMI z-score and mother's age at birth, whose functional forms are unknown. These two variables are also of different types, with the BMI z-score distributed on the whole real line while the mother's age at birth is distributed on the positive real line. This provides an opportunity to demonstrate how to use the method of sieves under different situations.

Results suggest a complex relationship between children's BMI z-score measured at 4-5 years and their accumulated medical services costs. The score within the normal weight range is not statistically significantly associated with the costs. The score within the underweight range is negatively associated with the costs but the association is statistically insignificant due to the small number of observations in this group. The score within the overweight range is positively associated with the costs and the association is statistically significant. The function of the score is overall of a "bucket" shape, with decreasing slopes at both ends. By contrast, using polynomials results in a function with an inverted U shape with increasing slopes at both ends. The large deviation in the estimated marginal effects from the two approaches suggests that not modelling the functional form of predictors appropriately may lead to biased results and misleading policy implications.

2 Motivating example

Our motivating example is a modified replication of the Au (2012) study which examined the association between children's BMI at age 4-5 and medical care costs over a 5-year period. The data came from the Longitudinal Study of Australian Children (LSAC) which is a representative panel survey of Australian children. It began in 2004 when the children were aged 4-5. The study was based on the data from the first three waves of the 4-5 years old cohort, collected in 2004, 2006 and 2008, which were linked to each child's Medicare record, covering a 5-year period. Medicare records include items related to the Medicare Benefits Scheme (MBS) costs and the Pharmaceutical Benefits Scheme (PBS) costs, representing medical service costs and pharmaceutical costs respectively. Accumulated MBS costs contain a negligible portion of zeros while accumulated PBS costs contain a relatively large portion of zeros. As our model is focused on the positive cost variable, we only consider MBS costs (adjusted to 2015-2016 price level) in this study. They contain around 1% zeros and these observations were dropped from our analysis.

In Au (2012) the key variable was not the BMI z-score but a discrete variable with three categories: underweight, normal weight, or overweight. Using a GLM model with log link and gamma distribution, Au (2012) found that both the overweight and underweight status would increase the medicare costs but the latter effect is not statistically significant even at 10% level. By contrast we used the original BMI z-score as the key variable and in the meantime selected the mother's age at birth as an additional control variable. The other control variables are largely the same as those used in Au (2012) except that the mother's smoking status was excluded from our analysis since it has a large number of missing values. See the Appendix for the other control variables and their definitions.

The BMI z-score has sparse data points on both ends of the distribution and the mother's age at birth has few data points on the right tail. We dropped these outliers (below the 0.5th percentile and above the 99.5th percentile for the BMI z-score and above the 99.5th percentile for the mother's age at birth) as nonparametric estimation is known to be unreliable in sparse regions. In total 64 observations were dropped and the final sample size is 4,255. The mean and median costs are 841 AUD and 619 AUD. The mean of the BMI z-score is 0.56 (ranging from -2.80 and 3.34). The mean of the mother's age at birth is 30 (ranging from 15 to 44).

3 EEE with a partially linear additive index

3.1 The Model

Consider a sample of observations $(Y_i, \mathbf{X}_i, Z_{1i}, Z_{2i})$, $i = 1, \ldots, n$[1], where Y_i is a dependent variable, Z_{1i} stands for the BMI z-score of individual i, Z_{2i} stands for "mother's age at birth/100"[2] of individual i, and $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})'$ is a $p \times 1$ vector of the rest of the explanatory variables which does not contain a constant term[3]. One of our main goals is to characterise the nonlinear impacts of Z_1 and Z_2 on Y.

Denote $\mu(\mathbf{X}_i, Z_{1i}, Z_{2i}) = E(Y_i|\mathbf{X}_i, Z_{1i}, Z_{2i})$. Suppose that there is a strictly monotone and differentiable link function $f(\cdot)$ and a strictly positive function $g(\cdot)$ such that:

$$f(\mu(\mathbf{X}_i, Z_{1i}, Z_{2i})) = \mathbf{X}_i'\boldsymbol{\beta} + m_1(Z_{1i}) + m_2(Z_{2i}), \tag{1}$$

$$\text{Var}(Y_i|\mathbf{X}_i, Z_{1i}, Z_{2i}) = g(\mu(\mathbf{X}_i, Z_{1i}, Z_{2i})), \tag{2}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, and $m_1(\cdot)$ and $m_2(\cdot)$ are two smooth functions defined on $\mathbb{R}$ and $[0, 1]$, respectively, given the nature of the two variables in our data set.

If the link function $f(\cdot)$ and the variance function $g(\cdot)$ are known, then the model is GAM. If $m_1(\cdot)$ and $m_2(\cdot)$ are also known, then the model is a GLM. In this paper,

[1] We focus on the case where there are two regressors of particular interest, Z_1 and Z_2, to illustrate different families of basis functions. The method could be readily generalised to cases with more or less regressors.

[2] The rescaling of mother's age at birth to within $[0, 1]$ is to facilitate the use of the Fourier basis functions in Section 3.3.

[3] This is simply for the convenience of model identification. The constant is expected to be absorbed into one of the unknown functions.

we adopt the EEE framework by assuming that the link function $f(\cdot)$ and the variance function $g(\cdot)$ have the following parametric forms:

$$f(\mu) = \begin{cases} \frac{\mu^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(\mu), & \text{if } \lambda = 0 \end{cases}, \tag{3}$$

$$g(\mu) = \theta_1 \mu^{\theta_2} \quad \text{or} \quad \theta_1 \mu + \theta_2 \mu^2, \tag{4}$$

where the parameters λ, θ_1 and θ_2 are to be estimated from the data.

The nonparametric additive functions $m_1(\cdot)$ and $m_2(\cdot)$ in (1) can be estimated using different approaches such as kernel estimation and sieve estimation. In the literature of nonparametric estimation, it is well known that these unknown functions are identified only up to a location shift. To uniquely identify them, some location normalisation conditions need to be imposed. A commonly used identification condition in local estimation methods (such as kernel estimation) is $E(m_1(Z_{1i})) = 0$ or $E(m_2(Z_{2i})) = 0$ (note that only one of these two zero mean conditions is required as $\mathbf{X}_i$ does not contain a constant term). Another identification condition, which is often used in series estimation, is $m_1(0) = 0$ or $m_2(0) = 0$ (again only one of these is required) (Li 2000).

In this paper, we will use the method of sieves to estimate $m_1(\cdot)$ and $m_2(\cdot)$ and assume that $m_1(\cdot)$ and $m_2(\cdot)$ are square integrable over their support. As Z_{1i} is defined on $\mathbb{R}$, this implies that the integral of the square of $m_1(z_1)$ on $\mathbb{R}$ is finite, which further implies that $m_1(z_1)$ will approach zero as z_1 approaches infinity. Visually, on a Cartesian graph, this means both ends of the function $m_1(z_1)$ will come near to the horizontal axis when z_1 becomes sufficiently small or large, suggesting the location of the function is identified. In this case the square integrability condition on $m_1(\cdot)$ implicitly places a location normalisation on the function and thus no further identification conditions are required (see related discussion under Assumption C.2 of Dong & Linton (2018)).

3.2 The method of sieves

The method of sieves was first introduced by Grenander (1981). The main difficulty encountered in nonparametric estimation is the need to search for a function on an infinite function space. The core idea of sieve estimation is to convert it to a problem of searching on a finite function space (through "parameterising" the nonparametric problem) which is much easier to handle.

To illustrate, consider a regression model with only one predictor:

$$Y = q(X) + \epsilon$$

where ϵ is the error term and $E(\epsilon|X) = 0$. $q(\cdot)$ is a smooth unknown function belonging to a specific type of function space and defined on $\mathbb{V}$ which could be $[0, 1]$, $\mathbb{R}$, $[0, \infty)$, etc. Estimation of $q(\cdot)$ is essentially a search for a function on the function space which minimises the sum of squared residuals. However, this is not feasible since such a space is infinite. What the method of sieves does is to replace the infinite space with a finite space and then search for the optimal function on that.

For any $q(x)$ on the defined function space, we can write it as a series expansion:

$$q(x) = \sum_{j=0}^{\infty} \pi_j t_j(x), \tag{5}$$

where $t_j(x)$ are the basis functions and π_j are their weights. This sum of an infinite number of basis functions can be approximated by a sum of a finite number of the first few basis functions:

$$q(x) \approx \sum_{j=0}^{k-1} \pi_j t_j(x),$$

where k is a positive integer called the "truncation parameter". This approximation is valid because under some regularity conditions (e.g., Newey 1997) the truncation residual $\sum_{j=k}^{\infty} \pi_j t_j(x)$ is of a negligible order $O(k^{-\nu})$ when k is large enough, where ν is a positive constant whose value is determined by the smoothness of $q(x)$. Thus, for sufficiently smooth $q(x)$ and large enough k, the truncation bias caused by dropping the remainder term $\sum_{j=k}^{\infty} \pi_j t_j(x)$ can be ignored.

The estimation problem has now been reduced to a finite-dimensional optimisation problem. What is left to do is to choose the appropriate basis functions and truncation parameter along with the estimation of the weights. There are many different types of basis functions, suitable for different function spaces and supports (Chen 2007). The choice of basis functions is often based on the support, the smoothness, the shape restrictions (from economic theory) as well as the ease of computation (see Chen (2007) for related discussions). The truncation parameter is typically obtained by using a criterion or method such as the cross-validation.

3.3 Nonparametric estimation

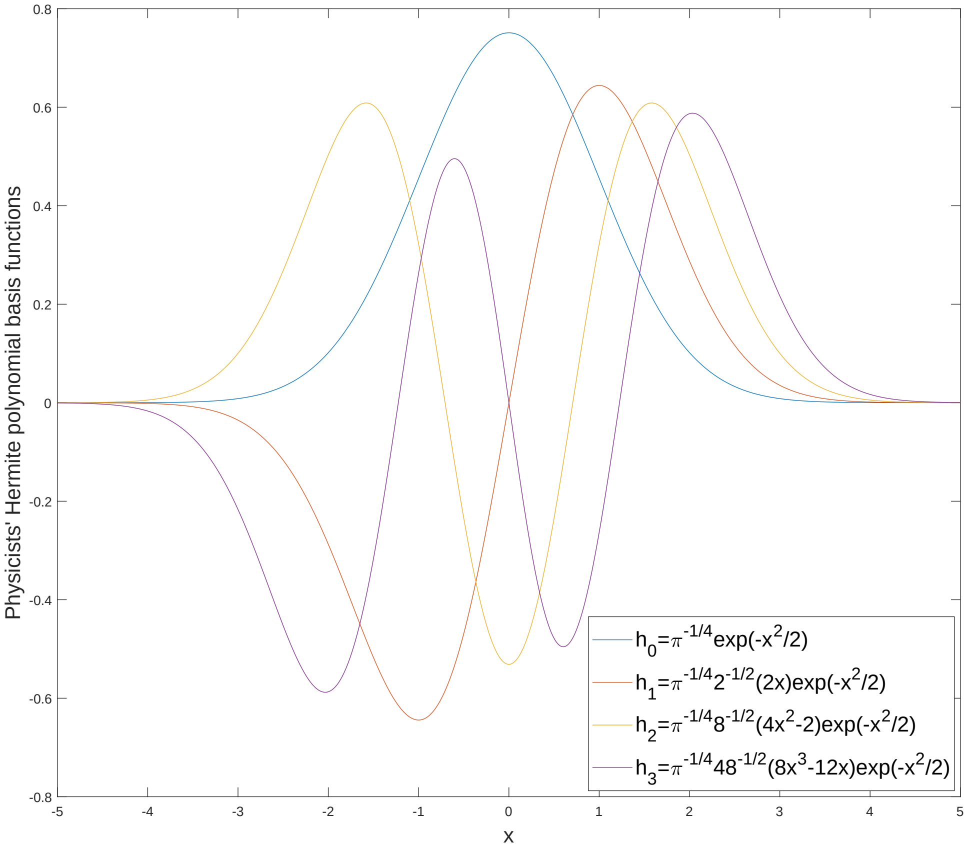
The first step for the sieve estimation of $m_1(Z_{1i})$ and $m_2(Z_{2i})$ is to choose the basis functions. We follow Chen (2007) and make the choice based on the supports (BMI z-score is defined on $\mathbb{R}$ while mother's age at birth/100 is defined on $[0, 1]$) and economic constraints (health expenditures cannot increase unlimitedly). Both basis function systems chosen in our estimation belong to the Hilbert space, ensuring that $m_1(Z_{1i})$ and $m_2(Z_{2i})$ are square integrable on their support.

For the BMI z-score, we use a basis function system defined as below:

$$h_j(w) = \frac{1}{\sqrt[4]{\pi}\sqrt{2^j j!}} H_j(w) \exp(-\frac{w^2}{2}), \quad j = 0, 1, 2, \ldots$$

where $H_j(w)$ represent the physicists' Hermite polynomials (Nevai 1986). We therefore call this system the "physicists' Hermite polynomial system" whose first four basis functions are illustrated in Figure 1. Almost all basis function systems start from the constant function of one (like polynomials). But in the physicists' Hermite polynomial system the first function is not one due to the square integrability condition over $\mathbb{R}$, which in effect enables location identification of $m_1(Z_{1i})$ as discussed in Section 3.1.

Figure 1: Physicists' Hermite polynomial basis functions (first four functions)



For any function $\phi(w)$ belonging to this function space, it can be written as:[4]

$$\phi(w) = \sum_{j=0}^{\infty} \pi_{1j} h_j(w), \tag{6}$$

where $\pi_{1j} = \int \phi(w) h_j(w) dw$. As explained in the last section, $\phi(w)$ can be approximated using the first few basis functions as:

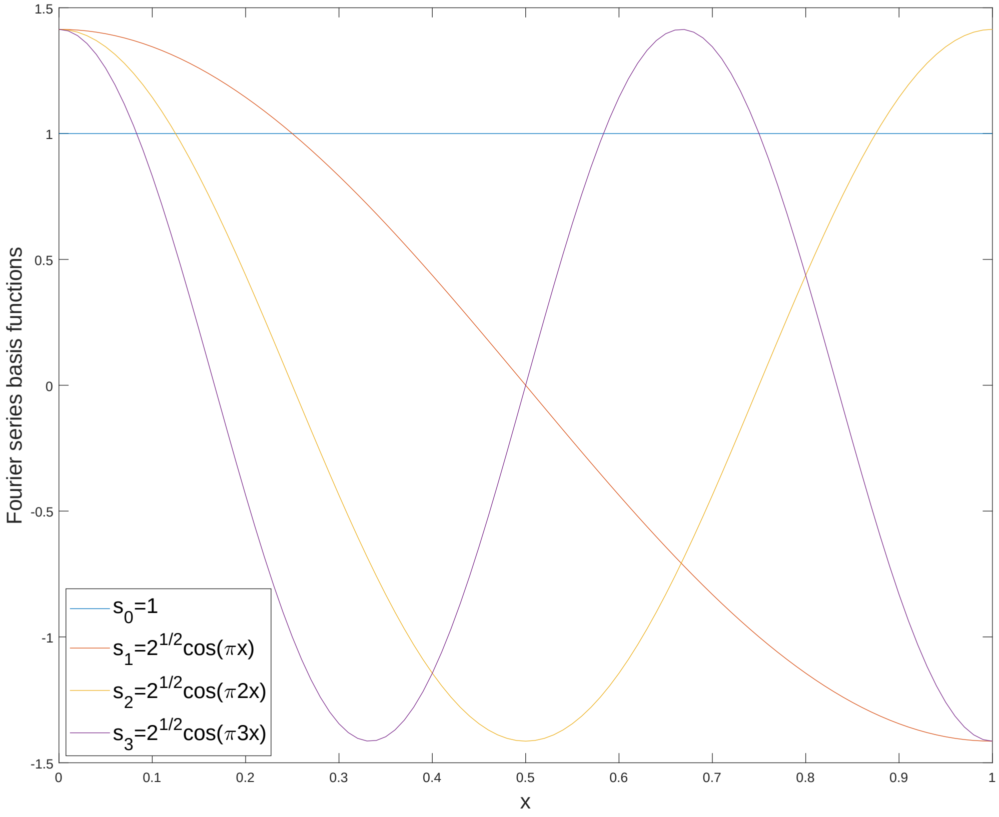
$$\phi(w) \approx \sum_{j=0}^{k_1 - 1} \pi_{1j} h_j(w),$$

where k_1 is a truncation parameter.

Since the values of "mother's age at birth/100" are all between 0 and 1, we use the Fourier series $\{s_0(u) = 1,\ s_j(u) = \sqrt{2}\cos(\pi j u) \text{ for } j \geq 1\}$, which is an orthonormal basis on the Hilbert space with the support $[0, 1]$. The first four basis functions of this system are illustrated in Figure 2. In this case the first function $s_0(\cdot)$ is the constant function of one. As the other control variables $\mathbf{X}$ do not contain a constant and $m_1(z_1)$ approaches zero as z_1 approaches infinity, any unaccounted constant component on the right hand side of (1) is absorbed by $m_2(z_2)$ through $s_0(z_2) = 1$.

[4] As the basis functions are orthonormal this is an orthogonal series expansion.

Figure 2: Fourier series basis functions (first four functions)



Likewise, for any function ψ belonging to this function space, it can be written as

$$\psi(u) = \sum_{j=0}^{\infty} \pi_{2j} s_j(u), \tag{7}$$

where $\pi_{2j} = \int_0^1 \psi(u) s_j(u) du$. This again can be approximated by

$$\psi(u) \approx \sum_{j=0}^{k_2-1} \pi_{2j} s_j(u)$$

where k_2 is a truncation parameter.

Given these choices, we can approximate $m_1(Z_{1i})$ and $m_2(Z_{2i})$ by

$$m_1(Z_{1i}) \approx \sum_{j=0}^{k_1-1} \pi_{1j} h_j(Z_{1i}) = \mathbf{H}(Z_{1i})'\mathbf{\Pi}_1, \tag{8}$$

$$m_2(Z_{2i}) \approx \sum_{j=0}^{k_2-1} \pi_{2j} s_j(Z_{2i}) = \mathbf{S}(Z_{2i})'\mathbf{\Pi}_2, \tag{9}$$

where $\mathbf{H}(Z_{1i}) = (h_0(Z_{1i}), \ldots, h_{k_1-1}(Z_{1i}))'$, $\mathbf{S}(Z_{2i}) = (s_0(Z_{2i}), \ldots, s_{k_2-1}(Z_{2i}))'$, $\mathbf{\Pi}_1 = (\pi_{10}, \ldots, \pi_{1,k_1-1})'$ and $\mathbf{\Pi}_2 = (\pi_{20}, \ldots, \pi_{2,k_2-1})'$.

In view of the series approximations (8) and (9), we can re-write (1) as

$$f(\mu(\mathbf{X}_i, Z_{1i}, Z_{2i})) \approx \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{H}(Z_{1i})'\mathbf{\Pi}_1 + \mathbf{S}(Z_{2i})'\mathbf{\Pi}_2 = \mathbb{X}_i'\mathbf{B}, \tag{10}$$

where $\mathbb{X}_i = (\mathbf{X}_i', \mathbf{H}(Z_{1i})', \mathbf{S}(Z_{2i})')'$ and $\mathbf{B} = (\boldsymbol{\beta}', \mathbf{\Pi}_1', \mathbf{\Pi}_2')'$. Equation (10) now has a linear parametric form, and its parameters $\mathbf{B}$, together with the parameters λ, θ_1 and θ_2 in the link and variance functions, can be estimated using existing EEE procedure (e.g., the -pglm- Stata module was used in our analysis).

The performance of sieve estimation of $m_1(\cdot)$ and $m_2(\cdot)$ depends on the choice of the truncation parameters k_1 and k_2. There are two types of potential errors. One is the approximation error which occurs when the truncation parameter is not sufficiently large. The other is the estimation error when the truncation parameter is too large relative to the given sample size (so that there are too many parameters to estimate). Hence, the optimal choice of k_1 and k_2 involves balancing between these two types of errors. We used 10-fold cross validation (based on the averaged mean squared error (MSE) and mean absolute error (MAE)) to select k_1 and k_2 in the empirical analysis.

Once k_1 and k_2 are chosen, we can use the EEE estimation procedure to obtain an estimate $\widehat{\mathbf{B}} = \left(\widehat{\boldsymbol{\beta}}', \widehat{\mathbf{\Pi}}_1', \widehat{\mathbf{\Pi}}_2'\right)'$ and subsequently estimates of $m_1(u)$ and $m_2(u)$ as $\widehat{m}_1(u) = \mathbf{H}(u)'\widehat{\mathbf{\Pi}}_1$ and $\widehat{m}_2(u) = \mathbf{S}(u)'\widehat{\mathbf{\Pi}}_2$.

3.4 Marginal effects and standard errors

One of the main reasons for undertaking cost regressions is to examine a key variable's impact on costs. It is therefore necessary to derive the marginal effects, which can be challenging in the case of nonparametric estimation (e.g., when using kernel estimation). Often, an additional step of parameterisation is needed to facilitate the derivation of the first order derivative. Sieve estimation has a clear advantage in this case as the core of this approach is to parameterise the unknown function.

Here we demonstrate how to derive the marginal effects of Z_1 and Z_2 given the two different basis function systems. Let $l(\cdot)$ be the inverse function of $f(\cdot)$. Simple algebra gives that

$$\begin{aligned} l(w) &= \begin{cases} (\lambda w + 1)^{\frac{1}{\lambda}} & \text{if } \lambda \neq 0 \\ \exp(w) & \text{if } \lambda = 0 \end{cases}, \\ l^{(1)}(w) &= \begin{cases} (\lambda w + 1)^{\frac{1}{\lambda}-1} & \text{if } \lambda \neq 0 \\ \exp(w) & \text{if } \lambda = 0 \end{cases}. \end{aligned} \tag{11}$$

where $l^{(1)}(w)$ represents the first order derivative of $l(w)$. The marginal effects of Z_1

and Z_2 are then calculated as

$$\begin{aligned}\frac{\partial\mu(\mathbf{X},Z_1,Z_2)}{\partial Z_1} &= l^{(1)}\left(\mathbf{X}'\boldsymbol{\beta}+m_1(Z_1)+m_2(Z_2)\right)\cdot m_1^{(1)}(Z_1), &(12)\\ \frac{\partial\mu(\mathbf{X},Z_1,Z_2)}{\partial Z_2} &= l^{(1)}\left(\mathbf{X}'\boldsymbol{\beta}+m_1(Z_1)+m_2(Z_2)\right)\cdot m_2^{(1)}(Z_2), &(13)\end{aligned}$$

where $m_1^{(1)}(\cdot)$ and $m_2^{(1)}(\cdot)$ are the first order derivatives of $m_1(\cdot)$ and $m_2(\cdot)$, respectively.

Given the approximation (8), we can replace $m_1^{(1)}(Z_{1i})$ by $\mathbf{\Pi}_1'\mathbf{H}^{(1)}(Z_{1i})$ where $\mathbf{H}^{(1)}(Z_{1i})$ is the first order derivative of $\mathbf{H}(Z_{1i})$, i.e., $\mathbf{H}^{(1)}(Z_{1i}) = (h_0^{(1)}(Z_{1i}),\ldots,h_{k_1-1}^{(1)}(Z_{1i}))'$. The first order derivatives for the Physicists' Hermite polynomial basis functions are given below

$$h_j^{(1)}(Z_{1i}) = \begin{cases} -\frac{1}{\sqrt{2}}h_1(Z_{1i}), & j=0,\\ \frac{1}{\sqrt{2}}\left(\sqrt{j}h_{j-1}(Z_{1i})-\sqrt{j+1}h_{j+1}(Z_{1i})\right), & j\geq 1.\end{cases}$$

Similarly, given the approximation (9), $m_2^{(1)}(Z_{2i})$ can be replaced by $\mathbf{\Pi}_2'\mathbf{S}^{(1)}(Z_{2i})$ where $\mathbf{S}^{(1)}(Z_{2i}) = (s_0^{(1)}(Z_{2i}),\ldots,s_{k_2-1}^{(1)}(Z_{2i}))'$. The first order derivatives for the Fourier series basis functions are given below

$$s_j^{(1)}(Z_{2i}) = \begin{cases} 0, & j=0,\\ -\sqrt{2}\pi j\cdot\sin(\pi j Z_{2i}), & j\geq 1.\end{cases}$$

By substituting the estimates of $\boldsymbol{\beta}$, $\mathbf{\Pi}_1$, $\mathbf{\Pi}_2$ and λ into (12) and (13), we can obtain estimates of the marginal effects of Z_1 and Z_2. The standard errors and confidence intervals for estimates of $m_1(Z_1)$ and $m_2(Z_2)$ and the marginal effects of Z_1 and Z_2 can be obtained through a bootstrap method. For example, Stata has an in-built bootstrap command that can be used in conjunction with -pglm-.

One may wonder why we do not directly calculate the standard errors considering the linear combination of basis functions. However, this would work only if we can establish an asymptotic distribution for the estimator of $\mathbf{B}$ in equation(10). The vectors $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ have diverging dimensions as the sample size increases (since the truncation parameters k_1 and k_2 are diverging asymptotically in order for the remainder terms to be negligible as sample size increases). On the other hand, in equation(10), the estimator of $\boldsymbol{\beta}$ and those of $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ have different convergence rates (the sieve estimation part has a slower convergence rate). It is therefore difficult to derive the asymptotic variance of the estimator of $\mathbf{B}$ in equation(10) and hence we recommend using the bootstrap method.

4 Results

Given the nature of the two continuous variables we have chosen suitable basis functions, as described in Section 2.3. The new model can be estimated using the

Stata module -pglm- once the truncation parameters k_1 and k_2 are determined. Based on the formula $k_1 = k_2 = \lfloor n^{1/6} \rfloor$ adapted from Dong & Linton (2018) the initial truncation parameter for both variables was chosen as 4. We then allowed each to vary from 3 to 5 which leads to 9 different models which were then evaluated using a 10-fold cross-validation based on the averaged MSE and MAE. Results are in Table 1 which suggests the model with $k_1 = 3$ and $k_2 = 4$ produced both the smallest MSE (1.8272) and the smallest MAE (1.0149)[5]. The 10-fold cross-validation was also used to select the optimal model based on polynomials resulting in a quadratic function for BMI z-score and a cubic function for mother's age at birth. The average MSE and MAE for this model are 1.8304 and 1.0159, inferior to the model estimated using sieves.[6]

Table 1: Truncation parameters selection: model comparison based on averaged mean squared error (MSE) and mean absolute error (MAE) from 10-fold cross-validation

Based on MSE			
k1\k2	3	4	5
3	1.8369	**1.8272**	Convergence problems
4	1.8380	1.8283	Convergence problems
5	1.8382	1.8285	Convergence problems
Based on MAE			
k1\k2	3	4	5
3	1.0191	**1.0149**	Convergence problems
4	1.0195	1.0153	Convergence problems
5	1.0192	1.0150	Convergence problems

We compare the functions estimated using sieves with the ones estimated using discretisation. For the BMI z-score, we divided it into 11 categories and used one of them (-0.5 0) as the reference level; See Table 2. Using the midpoints in each category and with the coefficient for the reference category fixed (at -0.13 for a location close to the other estimated functions), we were able to plot the implied functional form (the green line) in Figure 3, along with the ones estimated using sieves (the blue lines) and polynomials (the red line). The solid blue line represents the optimal model using sieves when $k_1 = 3$ and the dashed blue line represents the model using sieves when

[5] The MSE and MAE numbers look small because the dependent variable is normalised using its sample mean prior to the estimation.

[6] On the original cost scale, this means an increment of 2263 in MSE and an increment of 0.843 in MAE from the optimal model estimated using sieves.

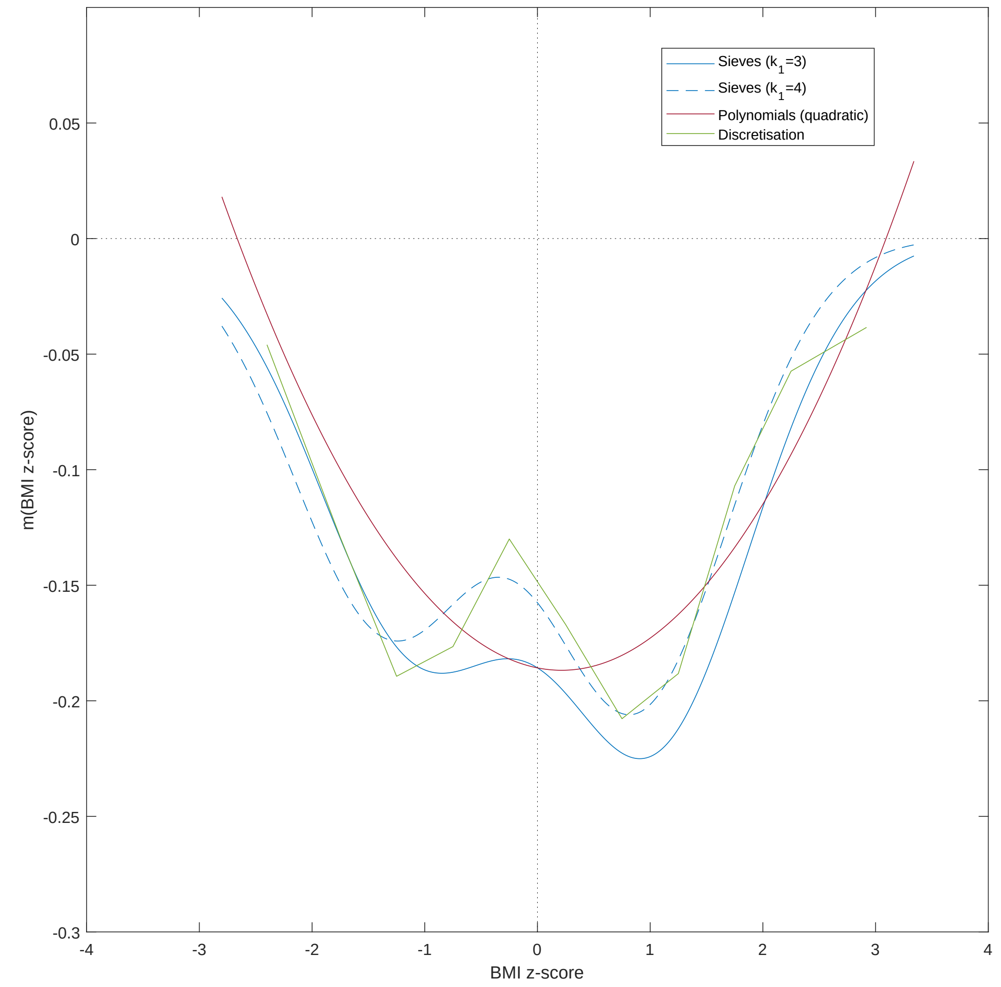
$k_1 = 4$. The figure suggests the functions estimated using sieves are very close to the implied functional form from discretisation. In fact, the one based on $k_1 = 4$ matches it very well, even capturing the hump in the middle. However this is a sign of over-fitting, proven by the cross-validation results.

Using polynomials does not identify the location of the target function. To facilitate the comparison (i.e., to compare functions around the same location) we applied an identification condition that $m_1(0) = -0.19$ representing the point where the optimal model produced by sieve estimation passes the y axis. The figure suggests a clear discrepancy between the functional form estimated by polynomials and the other two. It presents an inverted U shape with increasing slopes at both ends. By contrast, the functional form of the optimal model is of a "bucket" shape overall, with decreasing slopes at both ends.

Table 2: Estimation of m_1(BMI z-score) by using sieves, polynomials, and discretisation

	Est	Std	P-value
Sieves			
x_0	-0.3491	0.1383	0.0120
x_1	-0.0291	0.0343	0.3960
x_2	-0.1439	0.0736	0.0500
Polynomials			
x	-0.0097	0.0183	0.5980
x^2	0.0225	0.0110	0.0400
Discretisation			
-2.8 ∼ -2.0	0.0840	0.2292	0.7140
-2.0 ∼ -1.5	0.0008	0.1259	0.9950
-1.5 ∼ -1.0	-0.0594	0.0687	0.3870
-1.0 ∼ -0.5	-0.0466	0.0635	0.4640
-0.5 ∼ 0	reference		
0 ∼ 0.5	-0.0370	0.0514	0.4720
0.5 ∼ 1.0	-0.0778	0.0487	0.1100
1.0 ∼ 1.5	-0.0582	0.0507	0.2510
1.5 ∼ 2.0	0.0230	0.0591	0.6980
2.0 ∼ 2.5	0.0726	0.0876	0.4070
2.5 ∼ 3.34	0.0916	0.1059	0.3870

Figure 3: Plotting the estimated m_1(BMI z-score) by using sieve estimation ($k_1 = 3$ and $k_1 = 4$), polynomials (a quadratic function), and discretisation



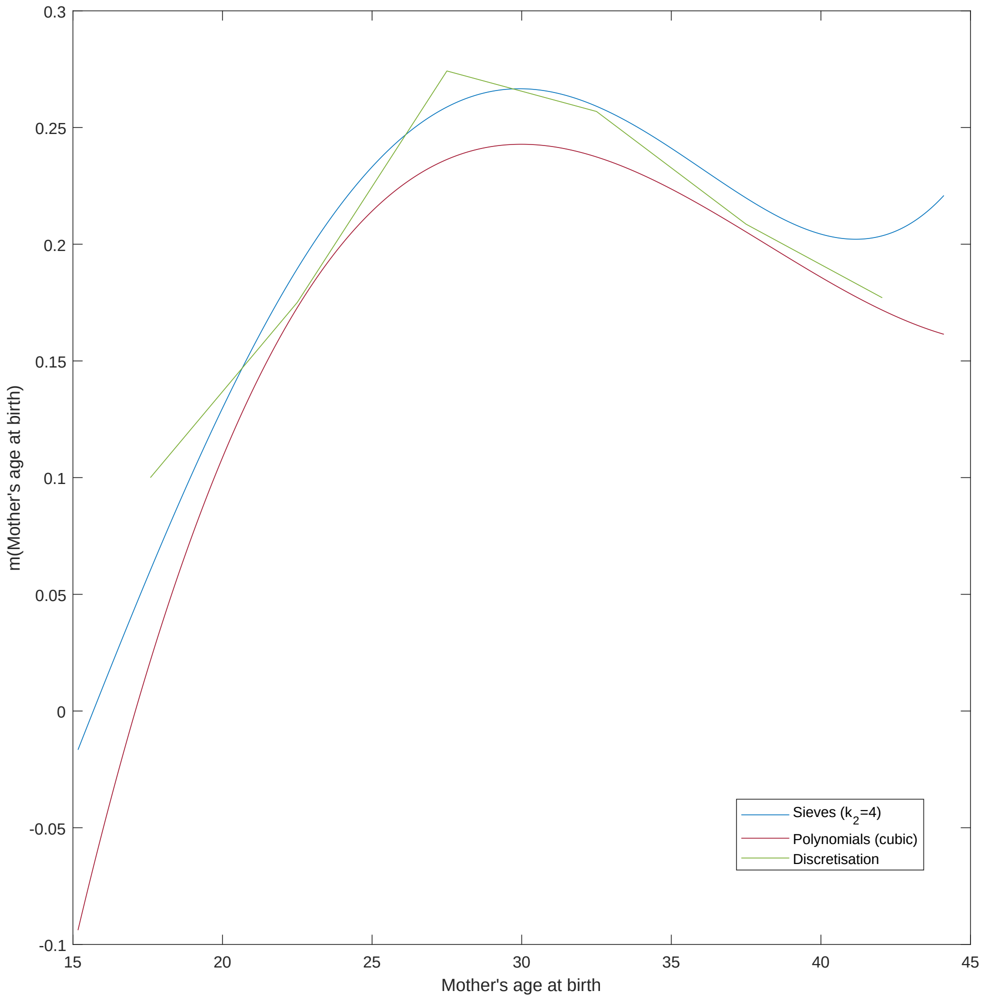
For the mother's age at birth, we divided it into six categories and used the first category ($<= 20$) as the reference level; See Table 3. Similarly, using the midpoints in each category and with the coefficient for the reference category fixed at 0.1 (again, for the convenience of visual comparison), we plotted the implied functional form in Figure 4, along with the ones estimated using the sieves and the polynomials. The figure suggests both the Fourier series with $k_2 = 4$ and polynomials fit well with small differences on both ends.

Table 3: Estimation of m_2(mother's age at birth) by using sieves, polynomials, and discretisation

	Est	Std	P-value
Sieves			
z_0	3.0902	3.4857	0.3750
z_1	-3.5825	4.0819	0.3800
z_2	1.8794	2.0865	0.3680
z_3	-0.7254	0.6204	0.2420
Polynomials			
x	0.1624	0.1376	0.2380
x^2	-0.0045	0.0047	0.3470
x^3	0.00004	0.0001	0.4640
Discretisation			
$15.18 \sim 20$	reference		
$20 \sim 25$	0.0751	0.0741	0.3110
$25 \sim 30$	0.1742	0.0740	0.0190
$30 \sim 35$	0.1569	0.0771	0.0420
$35 \sim 40$	0.1086	0.0781	0.1640
$40 \sim 44.11$	0.0771	0.0969	0.4260

Note: When using the method of sieves, the mother's age at birth variable was scaled into $[0, 1]$. Also, the estimated z_0 combines the intercept in the linear index and the true z_0.

Figure 4: Plotting the estimated m_2(mother's age at birth) by using sieve estimation ($k_2 = 4$), polynomials (a cubic function), and discretisation

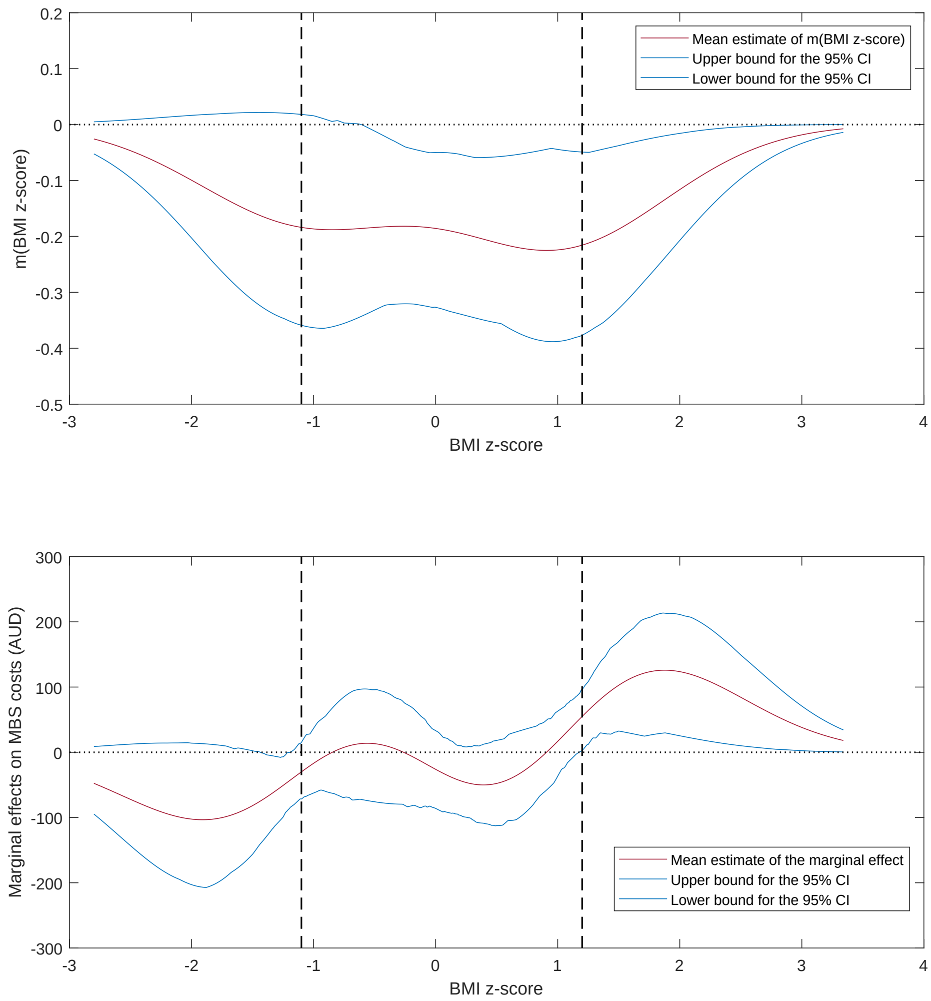


For the preferred model, the estimated λ is 0.93 with its standard error 0.34. The power variance function was used with its parameters θ_1 estimated as 0.80 (standard error 0.06) and θ_2 as 1.41 (standard error 0.18). These suggest the link function is very close to identity and the variance function is close to a Poisson.[7] The former is expected since all the $\mathbf{X}$ variables in our example are dummies so if the relationships between the two continuous variables and the cost are sufficiently recovered then there should be little nonlinearity left for the link function to capture.

We now return to the key research question of the empirical example - how children's BMI z-score measured at 4-5 years old relates to their accumulated MBS costs over a 5-year period. To answer this question, we plot the estimated m_1(BMI z-score) along with its confidence intervals; See Figure 5. The vertical dashed lines indicate the approximate cut-points for the BMI categories using age and gender specific cut-offs from Cole et al. (2000): the one on the left represents the cut-off between underweight and normal weight while the one on the right represents the cut-off between normal and over weight.

[7] Using the quadratic variance function would suggest the variance function is closer to Inverse Gaussian. But as we expected, the choice of variance function structure had little impact on the estimation of the conditional mean.

Figure 5: Estimated m_1(BMI z-score) using the method of sieves and its marginal effects on MBS costs



Note: Top panel: m_1(BMI z-score) estimated using the method of sieves with confidence intervals; Bottom panel: the marginal effects of BMI z-score at age 4-5 on a 5-year accumulated MBS costs (mother's age at birth is fixed at its mean 30 and all the dummy variables are set at their reference level) with confidence intervals; The vertical dashed lines indicate the approximate cut-points for the BMI categories using age and gender specific cut-offs from Cole et al. (2000): the one on the left represents the cut-off between underweight and normal weight while the one on the right represents the cut-off between normal and over weight.

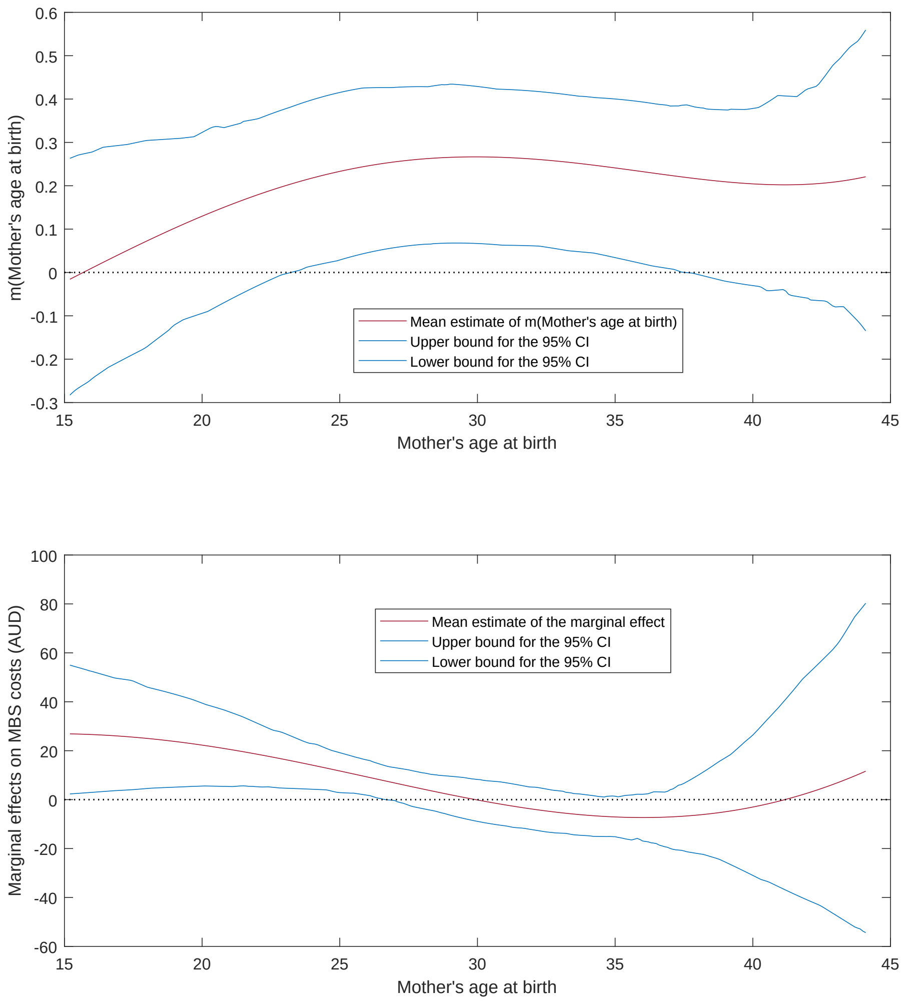
From Figure 5, we can see that when the BMI z-score is around 1, m_1(BMI z-score) reaches its smallest value and as the z-score increases from there m_1(BMI z-score) increases, suggesting the accumulated MBS costs increase as well. On the other hand, when the z-score decreases from there, m_1(BMI z-score) first increases slightly and then becomes almost flat until the z-score reaches around -1, and then increases.

Whether these relationships are statistically significant can be examined through marginal effects analysis. For example, we can fix the mother's age at birth at its mean 30 and all the other control variables at their reference levels, and then use the method outlined in Section 2 to calculate the marginal effects of different BMI z-scores. For illustration, we plot the estimated marginal effects and associated confidence intervals in Figure 5. It suggests that for this specific scenario, all the marginal effects are statistically significant for the overweight range but statistically insignificant for the normal weight range (small size and large standard error). For the underweight range, most are statistically insignificant (but we need to be aware that this group of children is of a small sample size).

Whilst mother's age at birth is only a control variable, out of interest we also plot the estimated m_2(mother's age at birth) and its marginal effects (when the BMI z-score is set to the mean value of 0.56 and all the other control variables are set at their reference levels) in Figure 6. These suggest that the accumulated MBS costs increase as mother's age at birth increases until around 30 years old then decrease after that but for the selected child the marginal effects of mother's age at birth on costs are not statistically significant from zero after age reaches around 27 years old.

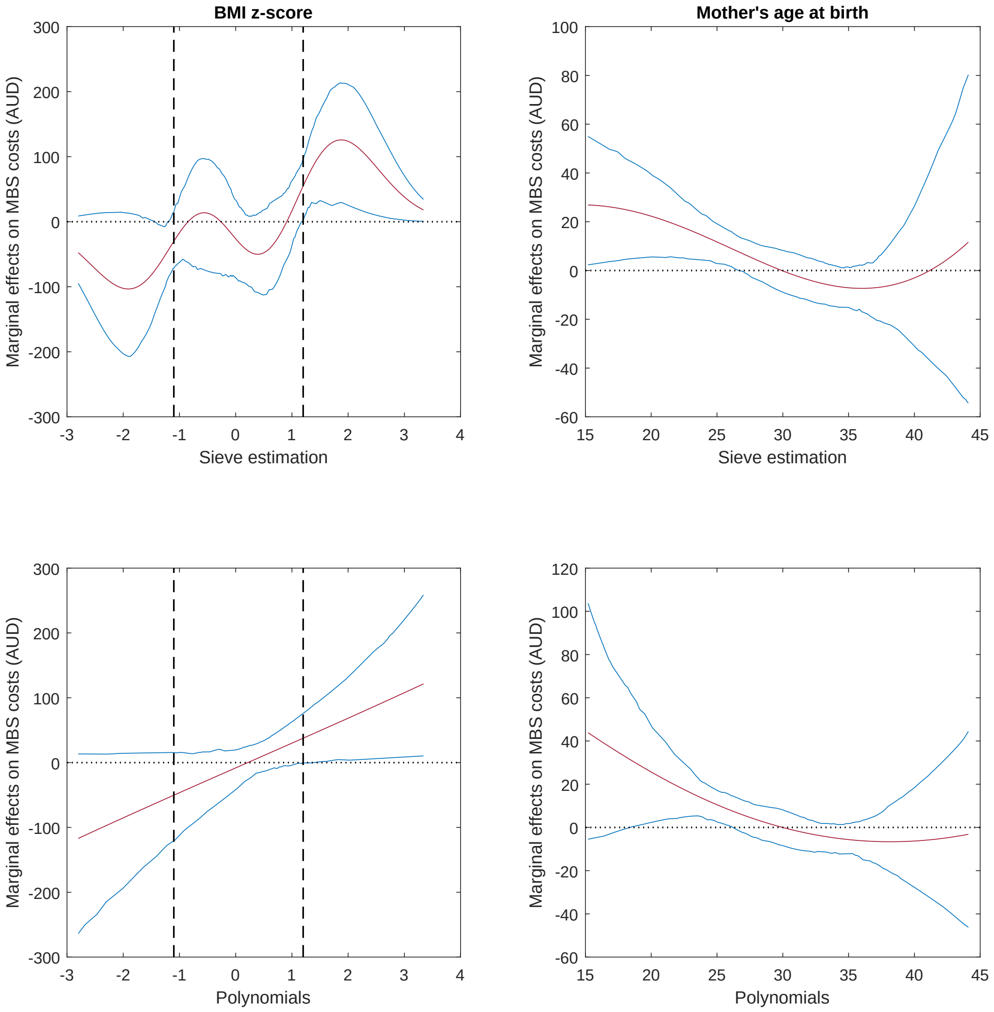
Finally a head-to-head comparison of marginal effects from polynomials and sieves is undertaken; See Figure 7. For the BMI z-score, the difference is clear. One of the key questions policy makers would ask is how healthcare costs are related to being overweight, to which these two methods would give very different answers. For mother's age at birth, the difference is small except that using polynomials generated larger marginal effects for younger ages and it changes at a faster speed.

Figure 6: Estimated m_2(mother's age at birth) using the method of sieves and its marginal effects on MBS costs



Note: Top panel: m_2(mother's age at birth) estimated using the method of sieves with confidence intervals; Bottom panel: the marginal effects of mother's age at birth on a 5-year accumulated MBS costs (the child's BMI z-score at age 4-5 is fixed at its mean of 0.56 and all the dummy variables are set at their reference levels) with confidence intervals.

Figure 7: Estimated marginal effects on MBS costs: comparison between the method of sieves and polynomials



Note: Left panel: the marginal effects of BMI z-score at age 4-5 on a 5-year accumulated MBS costs (mother's age at birth is fixed at its mean 30 and all the dummy variables are set at their reference levels) with 95% confidence intervals; Right panel: the marginal effects of mother's age at birth on a 5-year accumulated MBS costs (the child's BMI z-score at age 4-5 is fixed at its mean 0.56 and all the dummy variables are set at their reference levels) with 95% confidence intervals; Top panel: estimated using the method of sieves; Bottom panel: estimated using polynomials (quadratic for BMI z-score and cubic for mother's age at birth). The dashed lines on the left panel indicate the approximate cut-points for the BMI categories using age and gender specific cut-offs from Cole et al. (2000): the one on the left represents the cut-off between underweight and normal weight while the one on the right represents the cut-off between normal and over weight.

5 Discussion and conclusion

The empirical and methodological efforts in using the GLM to model healthcare costs have been mostly concentrated on selecting the correct link and variance function. The misspecification of functional form of the key covariates has been largely neglected. Using polynomials cannot guarantee accurate approximation of any functional form. To address this problem, we propose a hybrid model incorporating the EEE framework by Basu & Rathouz (2005) and the partially linear additive index. More specifically, we partition the index function in the EEE model into a number of additive components including a linear combination of covariates (either dummy coded or continuous which are believed to enter the index linearly) and unknown functions of continuous variables which are believed to enter the index non-linearly.

These unknown functions can be estimated using various methods such as the kernel method, but the method of sieves is adopted in this paper. With this method, we can approximate the unknown functions using sets of basis functions (similar to using polynomials). The resulting estimation problem can therefore be solved using the existing user-written EEE software programs (e.g., -pglm- in Stata). The standard errors and confidence intervals for these variables and their marginal effects can be obtained using the bootstrap method.

The key to estimating such a model, however, lies in choosing suitable types of basis functions and appropriate truncation parameters. How to choose basis functions in other empirical situations is not in the scope of this study but has been extensively discussed elsewhere (see e.g., Chen 2007). Nevertheless, the two variables considered in our empirical example represent two of the most commonly used types and thus the basis functions adopted in our example may be useful in many other empirical applications.

The physicists' Hermite polynomial system was chosen as the basis function for BMI z-score. By theory it can recover any square integrable function defined on $\mathbb{R}$. As explained before, this type of functions must approach zero as the variable approaches infinity. This implies that there is an upper bound (or lower bound) for the function at the two extreme ends, which is sensible in the case of healthcare costs as costs are typically constrained. However, in practice we do not observe at infinity and only approximate the function wherever we have observations. So on a finite sample space, this basis function system in fact can recover any functional form including those not square integrable on $\mathbb{R}$ (e.g., linear) but on a finite interval.[8]

Unlike kernel estimation, where the selection of bandwidth has been extensively studied, research on how to choose the truncation parameter in sieve estimation has been relatively rare. Following the literature, we adopted out-of-sample forecast ability as the criterion in the selection of truncation parameters through the 10-fold cross validation. We also used a formula proposed by Dong & Linton (2018) which needs to satisfy specific conditions such as the degree of smoothness. It gave a good approximation to the final result. However, we

[8] A square integrable function defined on $\mathbb{R}$ cannot be linear on $\mathbb{R}$ but can be linear on a finite space.

do not recommend it as a rule of thumb for sieve estimation in general like Silverman's rule of thumb (Silverman 2018) for kernel estimation, as it is not a general result. However, it certainly can be used as a starting point.

Our model is a semiparametric extension of GLM with GLM, EEE and GAM as special cases. It inherits the benefits of EEE and has a clear advantage over the GAM which requires a sequential search for the link and variance function as well as all the unknown functions. This can be a laborious process and the optimal model may not even be identified given the strong interplay between the link function and the unknown functions (one changes as the others change). This may also explain why GAM has not been adopted in the healthcare cost regression literature in spite of the popularity of GLM in the area.

In conclusion, this paper identifies an often neglected but important area for modelling healthcare costs and proposes a new model to tackle the potential misspecifications. The estimation of the model can be undertaken using existing software packages with minimal programming needed and thus should be a viable new tool for health economists who are working in modelling healthcare costs and other similar areas.

References

Au, N. (2012), 'The health care cost implications of overweight and obesity during childhood', *Health Services Research* **47**(2), 655–676.

Basu, A. (2005), 'Extended generalized linear models: Simultaneous estimation of flexible link and variance functions', *Stata Journal* **5**(4), 501–516.

Basu, A. & Rathouz, P. J. (2005), 'Estimating marginal and incremental effects on health outcomes using flexible link and variance function models', *Biostatistics* **6**(1), 93–109.

Blough, D. K., Madden, C. W. & Hornbrook, M. C. (1999), 'Modeling risk using generalized linear models', *Journal of Health Economics* **18**(2), 153–171.

Cawley, J. & Meyerhoefer, C. (2012), 'The medical care costs of obesity: An instrumental variables approach', *Journal of Health Economics* **31**(1), 219–230.

Chen, X. (2007), Large sample sieve estimation of semi-nonparametric models, *in* J. J. Heckman & E. E. Leamer, eds, 'Handbook of Econometrics', Vol. 6, North Holland, pp. 5549–5632.

Chiou, J.-M. & Müller, H.-G. (1998), 'Quasi-likelihood regression with unknown link and variance functions', *Journal of the American Statistical Association* **93**(444), 1376–1387.

Cole, T. J., Bellizzi, M. C., Flegal, K. M. & Dietz, W. H. (2000), 'Establishing a standard definition for child overweight and obesity worldwide: international survey', *BMJ* **320**(7244), 1240–1243.

Dong, C. & Linton, O. (2018), 'Additive nonparametric models with time variable and both stationary and nonstationary regressors', *Journal of Econometrics* **207**(1), 212–236.

Engle, R. F., Granger, C. W., Rice, J. & Weiss, A. (1986), 'Semiparametric estimates of the relation between weather and electricity sales', *Journal of the American Statistical Association* **81**(394), 310–320.

Grenander, U. (1981), *Abstract inference*, Wiley, New York.

Hastie, T. & Tibshirani, R. (1986), 'Generalized additive models', *Statistical Science* **1**(3), 297–318.

Hoch, J. S., Briggs, A. H. & Willan, A. R. (2002), 'Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis', *Health Economics* **11**(5), 415–430.

Johnson, E., Dominici, F., Griswold, M. & Zeger, S. L. (2003), 'Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey', *Journal of Econometrics* **112**(1), 135–151.

Jones, A. M. (2000), Health econometrics, *in* A. J. Culyer & J. Newhouse, eds, 'Handbook of Health Economics', Vol. 1, Elsevier, pp. 265–344.

Jones, A. M. (2011), Models for health care, *in* M. P. Clements & D. F. Hendry, eds, 'The Oxford Handbook of Economic Forecasting', Oxford University Press, pp. 1–37.

Li, Q. (2000), 'Efficient estimation of additive partially linear models', *International Economic Review* **41**(4), 1073–1092.

Magee, L. (1998), 'Nonlocal behavior in polynomial regressions', *The American Statistician* **52**(1), 20–22.

Manning, W. G. (1998), 'The logged dependent variable, heteroscedasticity, and the retransformation problem', *Journal of Health Economics* **17**(3), 283–295.

Manning, W. G. & Mullahy, J. (2001), 'Estimating log models: to transform or not to transform?', *Journal of Health Economics* **20**(4), 461–494.

Nevai, P. (1986), 'Géza fredu, orthogonal polynomials and christoffel functions. A case study', *Journal of Approximation Theory* **48**, 3–167.

Newey, W. K. (1997), 'Convergence rates and asymptotic normality for series estimators', *Journal of Econometrics* **79**(1), 147–168.

Silverman, B. W. (2018), *Density estimation for statistics and data analysis*, Routledge.

van de Ven, W. P. M. M. & Ellis, R. P. (2000), Risk adjustment in competitive health plan markets, *in* A. J. Culyer & J. Newhouse, eds, 'Handbook of Health Economics', Vol. 1, Elsevier, pp. 755–845.

Appendix Other control variables

Variable	**Levels**
Child's characteristics	
Gender	female (reference)
	male
Age	age in months at time of survey <60 (reference)
	age in months at time of survey >= 60
Birth weight	>= 2500 grams (reference)
	<2500 grams
Breastfed at 6 months	no (reference)
	yes
Language at home	only English (reference)
	English and other languages
Siblings	none (reference)
	at least one sibling
Schooling	pre-school (reference)
	pre-year one and year one
	day care centre
	other
Residential location	city (reference)
	inner-regional area
	rural area
	remote area
Attention Deficit Disorder	no (reference)
	yes
Hearing problems	no (reference)
	yes
Vision problems	no (reference)
	yes
Eczema	no (reference)
	yes
Ear infections	no (reference)
	yes
Mother's characteristics	
Education level	university degree (reference)
	diploma
	high school and below
Full time employment	no (reference)
	yes
Healthcare card holder	no (reference)
	yes