Estimation of non-parametric regression models: Production profiles of crude oil

Paul H. C. Eilers[a], Vlasios Voudouris[b,c,d,1,*], Robert Rigby[c], Dimitrios Stasinopoulos[c]

[a]*Department of Biostatistics, Erasmus University, Dr. Molewaterplein 50, 3015 GE Rotterdam, The Netherlands*
[b]*Centre for International Business and Sustainability, London Metropolitan Business School, 84 Moorgate, London EC2M 6SQ, UK*
[c]*Statistics, Operational Research and Mathematics (STORM) centre, London Metropolitan University, 84 Moorgate, London EC2M 6SQ, UK*
[d]*ABM Analytics Ltd, Suite 17 125 145-157 St John Street, EC1V 4PW, London, UK*

Abstract

In many practical statistical situations, it is desirable to restrict the flexibility of nonparametric regression models to accommodate prior information. We propose an estimator for regression models with a smoothness penalty and constraints imposed by the nature of the problem. Our estimator is easy to implement and has an explicit algebraic structure. Alternative or additional constraints can be readily applied. We present production profiles of crude oil to demonstrate possible uses of the proposed estimator.

Keywords: non-parametric regression, smoothing penalties, constraints-based estimator

*Corresponding author
Email address: v.voudouris@londonmet.ac.uk (Vlasios Voudouris)
[1]telephone:+44(0)20 7320 1409; fax:+44(0)20 7320 1585

1. Introduction

The specification and estimation of non-parametric regression models is well established (e.g., Hall and Yatchew, 2007; Yatchew,1998; Eilers and Marx, 1996). In this letter, we propose a generalisation of the estimation of non-parametric regression models given constraints and smoothing penalties in order to incorporate additional prior information, which are exogenously specified, about the regression function.

We wish to estimate a regression function s (e.g., production function) given the data points (y_i,x_i) for $i = 1, 2, \ldots, n$

$$y_i = s(x_i) + \epsilon_i \tag{1}$$

where the $x_i's$ are equally spaced (e.g. in time) and where the ϵ_i are assumed to be independently normally distributed variables with mean of 0 and variance σ^2 and where the smoothing function $s(x)$ is subject to a smoothness penalty on the log of function $s(x)$, and where $s(x)$ is subject to the constraint that the integral of $s(x)$ adds up to a total $\mathcal{T}$ and where the *decline rate* (slope) of $\log s(x)$ is constrained to be below a maximum allowable value within a range of x. In many practical applications, the maximum allowable value of the slope of $\log s(x)$ is based upon forward-looking investment plans and/or historical observations (e.g., WEO, 2008; Hook *et.al.*, 2009). It is important to note that the decline rate springs into action only within the exogenously specified range of x while if the maximum allowable value of the slope of $\log s(x)$ is large enough (e.g., -30%), the effect of the slope is negligible when the regression function $s(x)$ is estimated.

The above conditions are particularly applicable to projections of production profiles of finite commodities (e.g., oil), where past observations are available as well as estimates of the integral (e.g., estimated volumes of oil originally present before any extraction $\mathcal{T}$) and maximum decline rates while the shape of the production function is unknown.

Production profiles of finite commodities is one of the key components in valuing commodities. Our proposed estimator has the advantage that the maximum allowable value of the slope of $\log s(x)$ can be optimised by an 'outer' optimisation algorithm which assumes that a profit-maximising owner of a finite commodity will select a production profile that maximises the net present value of the remaining reserves ($\mathcal{T}$ - minus cumulative production) of the commodity (e.g., Davis and Moore, 1998). Clearly, the optimised maximum allowable value of the slope of $\log s(x)$ depends, for example, on the

projection of the price of the commodity and the total marginal production costs.

In the next section we describe our model and the estimation procedure. Section 3 contains some examples of oil production profiles based on real data. Conclusions are given in section 4.

2. Nonparametric regression estimation under constraints

We consider nonparametric regression, which concerns estimation of regression functions without the straitjacket of a specific functional form, but with specific prior information about constraints on the functional form.

Given past data $(y_i, x_i), i = 1, .., n$ and prior information on the smoothing function $s(x)$, that it is subject to a smoothness penalty on the log of function $s(x)$, and constrained so that the integrand of $s(x)$ adds up to a total $\mathcal{T}$ and that the slope of $\log s(x)$ is below a maximum allowable value within a range of x, we wish to estimate the function of $\mu = s(x)$ in (1).

We extend the equally spaced $x's$ up to a value x_{n+m} from which $s(x)$ is assumed to be negligible. We extend the $y's$ with m zeros and let $\mathbf{y} = (y_1, .., y_n, 0, .., 0)$.

Then $\boldsymbol{\mu} = s(\mathbf{x})$ is estimated by minimising a function T given by:

$$\begin{aligned} T &= (\mathbf{y} - \boldsymbol{\mu})^\top W(\mathbf{y} - \boldsymbol{\mu}) + k(\mathcal{T} - \boldsymbol{\mu}^\top \mathbf{1})^2 + \\ &\quad \lambda \boldsymbol{\eta}^\top \mathbf{D}_2^\top \mathbf{D}_2 \boldsymbol{\eta} + l\left(\mathbf{D}_1 \boldsymbol{\eta} - \mathbf{t}\right)^\top \mathbf{V}\left(\mathbf{D}_1 \boldsymbol{\eta} - \mathbf{t}\right). \end{aligned} \tag{2}$$

where $\boldsymbol{\mu} = (\mu_1, .., \mu_{n+m})^\top$, $\boldsymbol{\eta} = \log(\boldsymbol{\mu})$, $W = diag(1, .., 1, 0, .., 0)$ is a weight matrix with n 1's and m 0's and $\mathbf{D}_2$ is a second order difference penalty matrix for lack of smoothness in $\boldsymbol{\eta}$ and $\mathbf{V} = diag(v_s)$ has $(n+m-1)$ elements where $v_s = 1$ if $(\tilde{\boldsymbol{\eta}}_{s+1} - \tilde{\boldsymbol{\eta}}_s) > t$ and x is within a restricted range (e.g. if x is time, a specific time range), where t is a selected maximum slope in the restricted range of x, and $v_s = 0$ otherwise. $\mathbf{D}_1$ is a first order difference penalty matrix.

Since $\boldsymbol{\eta} = \log(\boldsymbol{\mu})$, $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$ and $\exp(\boldsymbol{\eta})$ is expanded about the current estimate $\tilde{\boldsymbol{\eta}}$ giving:

$$\boldsymbol{\mu} = \exp(\boldsymbol{\eta}) \approx \exp(\tilde{\boldsymbol{\eta}}) + diag\left[\exp(\tilde{\boldsymbol{\eta}})\right] \Delta = \tilde{\boldsymbol{\mu}} + \tilde{\mathbf{G}} \boldsymbol{\Delta} \tag{3}$$

where $\Delta = \boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}$ and $\tilde{\mathbf{G}} = diag(\tilde{\boldsymbol{\mu}})$.

Then T is given by:

$$\begin{aligned} T &= (\mathbf{y} - \tilde{\boldsymbol{\mu}} - \tilde{\mathbf{G}}\Delta)^\top \mathbf{W}(\mathbf{y} - \tilde{\boldsymbol{\mu}} - \tilde{\mathbf{G}}\Delta) + \\ &\quad k(\mathcal{T} - \mathbf{1}^\top \tilde{\boldsymbol{\mu}} - 1^\top \tilde{\mathbf{G}}\Delta)^\top (\mathcal{T} - \mathbf{1}^\top \tilde{\boldsymbol{\mu}} - 1^\top \tilde{\mathbf{G}}\Delta) + \\ &\quad \lambda \boldsymbol{\eta}^\top \mathbf{D}_2^\top \mathbf{D}_2 \boldsymbol{\eta} + l\left(\mathbf{D}_1 \boldsymbol{\eta} - \mathbf{t}\right)^\top \mathbf{V}\left(\mathbf{D}_1 \boldsymbol{\eta} - \mathbf{t}\right). \end{aligned} \tag{4}$$

Differentiate T with respect to $\boldsymbol{\eta}$ to give:

$$\begin{aligned}\frac{dT}{d\boldsymbol{\eta}} &= -2\tilde{\mathbf{G}}^\top \mathbf{W}\left(\mathbf{y}-\tilde{\boldsymbol{\mu}}\right)+2\mathbf{R}\Delta+ \\ &\quad 2\mathbf{u}+2\mathbf{Q}\Delta+2\mathbf{P}_1\boldsymbol{\eta}-2l\mathbf{D}_1^\top\mathbf{V}\mathbf{t}+2\mathbf{P}_2\boldsymbol{\eta}. \end{aligned} \tag{5}$$

Set $\frac{dT}{d\boldsymbol{\eta}}=0$ to give:

$$\boldsymbol{\eta}=(\mathbf{R}+\mathbf{Q}+\mathbf{P}_1+\mathbf{P}_2)^{-1}\mathbf{r} \tag{6}$$

where $\mathbf{R}=\tilde{\mathbf{G}}^\top\mathbf{W}\tilde{\mathbf{G}}$, $\mathbf{Q}=k\tilde{\mathbf{G}}^\top\mathbf{1}^\top\mathbf{1}\tilde{\mathbf{G}}$, $\mathbf{P}_1=l\mathbf{D}_1^\top\mathbf{V}\mathbf{D}_1$, $\mathbf{P}_2=\lambda\mathbf{D}_2^\top\mathbf{D}_2$ and where $\mathbf{r}=\tilde{\mathbf{G}}^\top\mathbf{W}(\mathbf{y}-\tilde{\boldsymbol{\mu}})+\mathbf{R}\tilde{\boldsymbol{\eta}}+\mathbf{u}+\mathbf{Q}\tilde{\boldsymbol{\eta}}+l\mathbf{D}_1^\top\mathbf{V}\mathbf{t}$ and $\mathbf{u}=k(\mathcal{T}-\mathbf{1}^\top\tilde{\boldsymbol{\mu}})\tilde{\mathbf{G}}^\top\mathbf{1}$. Initialize $\tilde{\boldsymbol{\eta}}=\log(\bar{y})$ and then iteratively calculate $\boldsymbol{\eta}$ using (6) and update $\tilde{\boldsymbol{\eta}}$ until convergence.

Note that we do not have a data-driven way to select λ. In most practical applications λ is usually set to a small number so that $s(\mathbf{x})$ interpolates the observations while visual inspection is used to assess the adequacy of the fit.

3. Example: Crude oil production

Concern about the availability of oil emerged since the 19th century. Economist who attempt to model oil production face two key questions, namely i) how much recoverable oil exists? and ii) what path will oil production take over time?

Quantitative understanding of the latter involves curve-fitting techniques, which have been used since the 1950s. The approach of the curve-fitting approaches involves a limited set of parametric mathematical functions to statistically fit to historical production data and the use of $\mathcal{T}$ to improve the quality of model fit (Brandt, 2010).

The proposed non-parametric regression estimator contributes to the second question by converting remaining oil reserves (where $\mathcal{T}$ = cumulative production + remaining reserves) into an estimate of future rate of oil production $s(x)$ without the straitjacket of a parametric functional form for $s(x)$, where x is time. Note that x might be crude oil demand over time if the interest is to explore the optimal production profile given crude oil demand scenarios - based upon the philosophy of 'predict (demand) and provide (production)'. This 'predict and provide' philosophy is not uncommon in the oil industry (e.g., Mitchell, *et. al.*, 2012; Voudouris, *et. al.*, 2011).

Figure 1A shows global production (in million barrels per year) of crude oil over time estimated using the proposed model. The different estimated curves of crude oil production are based upon different geological assumptions for $\mathcal{T}$ and an industry accepted maximum allowable slope of -6.5%. Figure 1B shows the sensitivity of crude oil output to different maximum allowable values of the slope of the $\log s(x)$ while $\mathcal{T}$ is fixed at $3,000$ billion barrels of oil. Note that EUR means estimated ultimate recovery, which is represented by $\mathcal{T}$ in the model, while λ is set to a small number (0.1) so that $s(x)$ 'interpolates' the historical observations. In the examples shown in Figure 1, the slope springs into action from 2100 to 2150.

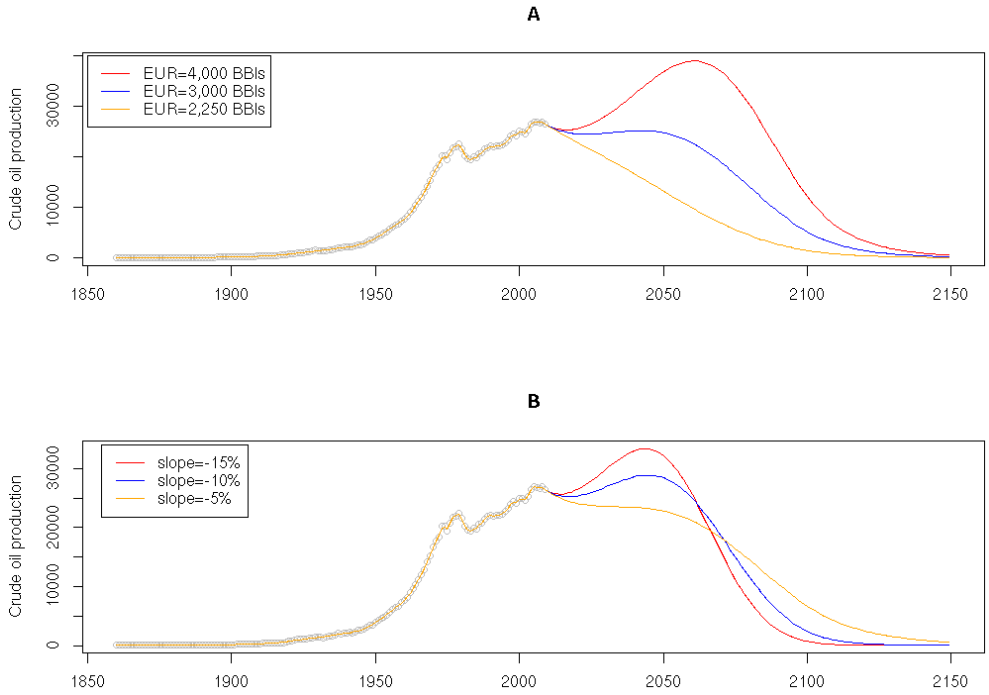

Figure 1: (A) Crude oil production for different $\mathcal{T}$. (B) Sensitivity of crude oil production for different slopes.

The production decline rate that will prevail is determined by forward-looking investments in existing fields and new fields. In many practical applications, the maximum allowable slope and the range of x within which the slope springs into action for the estimation of $s(x)$ are determined by "exam-

ining the situation in each country separately on a field-by-field basis, and drawing on extensive consultation with the oil industry, as well as assumptions about production and investment policies" (WEO, 2008, p. 276). In an optimising model of crude oil production for individual oil fields, however, the maximum allowable value of the slope of $\log s(x)$ can be optimised by an outer algorithm valuing producing crude oil reserves.

4. Conclusion

We propose an estimator of a nonparametric regression model $s(x)$ with smoothness penalties on its log function and constraints by incorporating prior information about the integrating total of $s(x)$ and on the maximum allowable slope of $\log s(x)$ within a range of x.

The estimator is easy to implement and has an explicit algebraic structure. The prior information is exogenously specified by the empirical researcher based upon forward-looking investment plans and/or historical observations. An 'outer' optimisation algorithm, such as maximisation of net present value of a finite commodity, can also be used to optimise the production profile by adjusting the maximum allowable slope of $\log s(x)$ within a range of x. Alternative or additional constraints can be applied in a similar manner.

References

Brandt, A. 2010. Review of mathematical models of future oil supply: Historical overview and synthesizing critique. Energy 35, 3958-3974.

Davis, G.A. and Moore, D.J., 1998. Valuing mineral reserves when capacity constrains production. Economics Letters 60, 121-125.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with b-splines and penalties (with comments and rejoinder). Statistal Science 11, 89-121.

Hall, P., Yatchew, A., 2007. Nonparametric estimation when data on derivatives are available. Annals of Statistics 35, 300-23.

Hook, M., Hirsch, R and Aleklett, K., 2009, Giant oil field decline rates and their influence on world oil production. Energy Policy 37, 2262-2272.

Mitchell, J., Marcel, V and Mitchell, B., 2012, What next for the oil and gas industry? Chatham House, London, UK.

Voudouris, V., Stasinopoulos, D., Rigby, R., Di Maio, C., 2011. The ACEGES laboratory for energy policy: Exploring the production of crude oil. Energy Policy 39, 54809.

WEO (World Energy Outlook) 2008. International Energy Agency, OECD/IEA, Paris. Available from: http://www.iea.org/w/bookshop/add.aspx?id=353

Yatchew, A., 1998. Nonparametric regression techniques in economics. Journal of Economic Literature 62, 669-721.