JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **30**: 705–717 (2015)
Published online 2 September 2014 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2410

SPLINE REGRESSION IN THE PRESENCE OF CATEGORICAL PREDICTORS

SHUJIE MA,[a,b] JEFFREY S. RACINE[c,d]* AND LIJIAN YANG[e]

[a] *Department of Statistics, University of California, Riverside, CA, USA*
[b] *Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou, China*
[c] *Department of Economics and Graduate Program in Statistics, McMaster University, Hamilton, Ontario, Canada*
[d] *School of Economics, La Trobe University, Melbourne, Victoria, Australia*
[e] *Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou, China*

SUMMARY

We consider the problem of estimating a relationship nonparametrically using regression splines when there exist both continuous and categorical predictors. We combine the global properties of regression splines with the local properties of categorical kernel functions to handle the presence of categorical predictors rather than resorting to sample splitting as is typically done to accommodate their presence. The resulting estimator possesses substantially better finite-sample performance than either its frequency-based peer or cross-validated local linear kernel regression or even additive regression splines (when additivity does not hold). Theoretical underpinnings are provided and Monte Carlo simulations are undertaken to assess finite-sample behavior; and two illustrative applications are provided. An implementation in R is available; see the R package 'crs' for details. Copyright © 2014 John Wiley & Sons, Ltd.

Received 26 August 2013; Revised 14 May 2014


Supporting information maybe found in the online version of this article.

1. BACKGROUND

Applied researchers must frequently model relationships involving both continuous and categorical predictors, and a range of nonparametric kernel regression methods have recently been proposed for such settings. These developments have extended the reach of kernel smoothing methods beyond the traditional continuous-only predictor case and have had a marked impact on applied nonparametric research; see Li and Racine (2007) for examples and an overview. Although kernel methods hold much appeal for practitioners, many in the applied community continue to resist their use, often for valid reasons. In particular, nonparametric kernel methods are local in nature, bandwidth selection can be fragile and numerically demanding and interpretation can be challenging, while imposing constraints on the resulting estimate can be difficult.

Regression spline methods, on the other hand, are global in nature and involve straightforward least squares solutions; hence unconstrained and constrained estimation is much easier to handle and faster to compute. Furthermore, their least squares underpinnings render the methods immediately accessible to those who routinely use least squares approaches. As such, regression splines provide an immediately accessible and attractive alternative to kernel methods for the nonparametric estimation of regression models. For excellent overviews of spline modeling we direct the interested reader to Stone (1985, 1994), Huang (2003) and Wang and Yang (2009). For applications of spline approaches, see Huang (1998) for functional ANOVA models, Huang and Yang (2004), Wang and Yang (2007) and Xue (2009) for additive models, Wang and Yang (2009) and Wang (2009) for single-index models,

* Correspondence to: Jeffrey S. Racine, Department of Economics, Kenneth Taylor Hall, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4M4, Canada. E-mail: racinej@mcmaster.ca

Greiner (2009) for the use of penalized splines in economic applications, Liu and Yang (2010) and Xue and Liang (2010) for additive coefficient models, Ma and Yang (2011) for jump detection in regression functions and Haupt *et al.* (2014) for smooth quantile modeling of economic retail scanner data via splines. See also the related work of Andrews (1991) and Newey (1991) for theoretical underpinnings of semi- and nonparametric series estimators and Chen (2007) for semiparametric sieve estimation.

However, just like their traditional kernel-based continuous-only predictor kin, nonparametric regression splines and series estimators are limited by their inability to handle the presence of categorical predictors without resorting to sample-splitting, which can entail a substantial loss in efficiency. In this paper we consider a regression spline alternative motivated by recent developments in the kernel smoothing of relationships involving categorical covariates. The proposed tensor-product spline approach possesses intuitive appeal by producing a piecewise polynomial, computational expedience as discussed before and theoretical reliability according to the mean squared and uniform convergence rates, and the pointwise asymptotic distribution results established in this paper. Ma and Racine (2013) recently proposed an additive regression spline approach with categorical and continuous predictors using discrete-support kernel functions, as is done in this paper. Since Ma and Racine's (2013) approach imposes an additive structure on the conditional mean function, it thereby fails to consider possible interactive effects of covariates unless explicitly provided by the researcher. Such models can be misspecified and perform poorly when the additivity assumption is violated, as illustrated in our simulations. Furthermore, not only is the proposed tensor-product spline approach more flexible, but it can provide practitioners with a useful approach for discovering underlying structure when additivity is violated. While the additive spline estimator achieves a univariate convergence rate, it suffers from the limitation that it has no asymptotic distribution, as discussed in Stone (1985). In this paper we furnish an asymptotic normality result for the tensor-product spline estimator, which allows the construction of pointwise confidence intervals, thereby overcoming yet another limitation inherent to the additive approach.

The remainder of this paper proceeds as follows. Section 2 outlines the framework and presents theorems detailing rates of convergence and the asymptotic distribution of the proposed approach. Section 3 considers cross-validatory selection of the spline knot vector and kernel bandwidth vector. Section 4 examines the finite-sample performance of the proposed method versus the traditional 'frequency' (i.e. 'sample-splitting') estimator, the additive regression spline estimator and the cross-validated local linear kernel regression estimator, and also offers some practical advice for practitioners. Section 5 presents illustrative applications and Section 6 presents some summary remarks. Proofs are to be found as supporting information in the supplementary Web Appendix, which also includes a supplementary illustration, a Monte Carlo simulation designed to assess the effect of 'distance' between the regression functions for different categories, and a Monte Carlo simulation designed to assess how confidence interval coverage based on our asymptotic results is affected by the DGP and cross-validated smoothing parameters.

2. METHODS AND MAIN RESULTS

We consider a nonparametric regression model containing both categorical and continuous predictors. In what follows we presume that the reader is interested in the unknown conditional mean in the following location-scale model:

$$Y = g(\mathbf{X}, \mathbf{Z}) + \sigma(\mathbf{X}, \mathbf{Z})\varepsilon \tag{1}$$

where $g(\cdot)$ is an unknown function, $\mathbf{X} = (X_1, \ldots, X_q)^{\mathrm{T}}$ is a q-dimensional vector of continuous predictors and $\mathbf{Z} = (Z_1, \ldots, Z_r)^{\mathrm{T}}$ is an r-dimensional vector of categorical predictors. Letting $\mathbf{z} = (z_s)_{s=1}^{r}$, we assume that z_s takes c_s different values in $D_s \equiv \{0, 1, \ldots, c_s - 1\}$, $s = 1, \ldots, r$,

DOI: 10.1002/jae

and let c_s be a finite positive constant. Let $\left(Y_i, \mathbf{X}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}}\right)_{i=1}^{n}$ be an i.i.d. copy of $\left(Y, \mathbf{X}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}}\right)$. Assume for $1 \leq l \leq q$, each X_l is distributed on a compact interval $[a_l, b_l]$ and, without loss of generality, we take all intervals $[a_l, b_l] = [0, 1]$. To handle the presence of categorical predictors, we propose to estimate $g(\cdot)$ by tensor-product polynomial splines weighted by categorical kernel functions. Let $\mathcal{B}(\mathbf{x})$ be the tensor-product polynomial splines and $L(\mathbf{Z}, \mathbf{z}, \lambda)$ be a product categorical kernel function, both of which will be defined later in this section. Then the nonparametric function $g(\mathbf{x}, \mathbf{z})$ can be approximated by $\mathcal{B}(\mathbf{x})^{\mathrm{T}}\beta(\mathbf{z})$, where $\beta(\mathbf{z})$ is a $\mathbf{K}_n \times 1$ vector with $\mathbf{K}_n \to \infty$ as $n \to \infty$. We estimate $\beta(\mathbf{z})$ by minimizing the following weighted least squares criterion:

$$\hat{\beta}(\mathbf{z}) = \arg \min_{\beta \in \mathbb{R}^{\mathbf{K}_n}} \sum_{i=1}^{n} \left\{Y_i - \mathcal{B}(\mathbf{X}_i)^{\mathrm{T}}\beta\right\}^2 L(\mathbf{Z}_i, \mathbf{z}, \lambda) \tag{2}$$

Thus $g(\mathbf{x}, \mathbf{z})$ is estimated by $\hat{g}(\mathbf{x}, \mathbf{z}) = \mathcal{B}(\mathbf{x})^{\mathrm{T}}\hat{\beta}(\mathbf{z})$. In this section, we will study the asymptotic properties of the proposed estimator.

We define a variant of Aitchison and Aitken's (1976) univariate categorical kernel function as

$$\begin{aligned} l(Z_s, z_s, \lambda_s) &= \begin{cases} 1, \text{ when } Z_s = z_s \\ \lambda_s, \text{ otherwise} \end{cases} \\ L(\mathbf{Z}, \mathbf{z}, \lambda) &= \prod_{s=1}^{r} l(Z_s, z_s, \lambda_s) = \prod_{s=1}^{r} \lambda_s^{1(Z_s \neq z_s)} \end{aligned} \tag{3}$$

where $L(\cdot)$ is a product categorical kernel function and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)^{\mathrm{T}}$ is the vector of bandwidths for each of the categorical predictors (for ordered categorical predictors we could instead use $l(Z_s, z_s, \lambda_s) = \lambda_s^d$ where $|Z_s - z_s| = d$ $(0 \leq d \leq c_s)$, which results in estimation bias identical to using Equation (3); see Racine and Li, 2004, for details). We point out that this kernel has been used extensively for kernel-based estimation involving the mix of continuous and categorical variables considered herein; see Li and Racine (2007) for the theoretical underpinnings of this kernel for kernel regression, kernel density estimation and kernel quantile estimation, among others. The distinguishing feature of this kernel is its ability to pool across categories (this occurs when a smoothing parameter hits its upper bound, i.e. $\lambda_s = 1$) and otherwise to deliver a smooth model whose coefficients are allowed to vary with respect to the realizations of the categorical variables (unlike, for example, a partially linear specification where only the 'intercept' term is allowed to change with respect to the values assumed by the categorical variables).

Let $G_l = G_l^{(m_l-2)}$ be the space of polynomial splines of order m_l and pre-select an integer $N_l = N_{n,l}$. Divide $[0, 1]$ into $(N_l + 1)$ subintervals $I_{j_l,l} = [t_{j_l,l}, t_{j_l+1,l})$, $j_l = 0, \ldots, N_l - 1$, $I_{N_l,l} = [t_{N_l,l}, 1]$, where $\left\{t_{j_l,l}\right\}_{j_l=1}^{N_l}$ is a sequence of equally spaced points, called interior knots, given as

$$t_{-(m_l-1),l} = \ldots = t_{0,l} = 0 < t_{1,l} < \ldots < t_{N_l,l} < 1 = t_{N_l+1,l} = \ldots = t_{N_l+m_l,l}$$

in which $t_{j_l,l} = j_l h_l$, $j_l = 0, 1 \ldots, N_l + 1$, $h_l = 1/(N_l + 1)$ is the distance between neighboring knots. Then G_l consists of functions ϖ satisfying: (i) ϖ is a polynomial of degree $m_l - 1$ on each of the subintervals $I_{j_l,l}$, $j_l = 0, \ldots, N_l$; (ii) for $m_l \geq 2$, ϖ is $m_l - 2$ times continuously differentiable on $[0, 1]$. Let $K_l = K_{n,l} = N_l + m_l$, where N_l is the number of interior knots and m_l is the spline order, $B_l(x_l) = \left\{B_{j_l,l}(x_l) : 1 - m_l \leq j_l \leq N_l\right\}^{\mathrm{T}}$ be a basis system of the space G_l. We define the space of tensor-product polynomial splines by $\mathcal{G} = \otimes_{l=1}^{q} G_l$. It is clear that $\mathcal{G}$ is a linear space of dimension $\mathbf{K}_n = \prod_{l=1}^{q} K_l$. Then

$$\mathcal{B}(\mathbf{x}) = \left[\left\{\mathcal{B}_{j_1,\ldots,j_q}(\mathbf{x})\right\}_{j_1=1-m_1,\ldots,j_q=1-m_q}^{N_1,\ldots,N_q}\right]_{\mathbf{K}_n \times 1} = B_1(x_1) \otimes \cdots \otimes B_q(x_q)$$

DOI: 10.1002/jae

is a basis system of the space $\mathcal{G}$, where $\mathbf{x} = (x_l)_{l=1}^q$. Let $\mathbf{B} = \left[\{\mathcal{B}(\mathbf{X}_1), \ldots, \mathcal{B}(\mathbf{X}_n)\}^{\mathrm{T}}\right]_{n \times \mathbf{K}_n}$. Let $\mathcal{L}_z = \mathrm{diag}\{L(\mathbf{Z}_1, \mathbf{z}, \lambda), \ldots, L(\mathbf{Z}_n, \mathbf{z}, \lambda)\}$ be a diagonal matrix with $L(\mathbf{Z}_i, \mathbf{z}, \lambda)$, $1 \leq i \leq n$ as the diagonal entries. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Then $\hat{\beta}(\mathbf{z})$ defined in (2) can be written as

$$\hat{\beta}(\mathbf{z}) = \left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{B}\right)^{-1}\left(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{Y}\right) \tag{4}$$

Given any $\mathbf{z} \in \mathcal{D}$, for $\mathcal{D} = D_1 \times \ldots \times D_r$, for any $\mu \in (0, 1]$, we denote by $C^{0,\mu}[0,1]^q$ the space of order μ Hölder continuous functions on $[0,1]^q$, i.e.

$$C^{0,\mu}[0,1]^q = \left\{\phi : |\phi|_{0,\mu,\mathbf{z}} = \sup_{\mathbf{x}\neq\mathbf{x}',\mathbf{x},\mathbf{x}'\in[0,1]^q} \frac{|\phi(\mathbf{x},\mathbf{z}) - \phi(\mathbf{x}',\mathbf{z})|}{|\mathbf{x}-\mathbf{x}'|_2^{\mu}} < +\infty\right\}$$

in which $|\mathbf{x}|_2 = \left(\sum_{l=1}^q x_l^2\right)^{1/2}$ is the Euclidean norm of $\mathbf{x}$, and $|\phi|_{0,\mu,\mathbf{z}}$ is the $C^{0,\mu}$ -norm of ϕ. Let $C[0,1]^q$ be the space of continuous functions on $[0,1]^q$. Given a q-tuple $\alpha = (\alpha_1, \ldots, \alpha_q)$ of non-negative integers, let $[\alpha] = \alpha_1 + \ldots + \alpha_q$ and let D^α denote the differential operator defined by $D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1}\cdots\partial x_q^{\alpha_q}}$. For positive numbers a_n and b_n, $n \geq 1$, let $a_n \asymp b_n$ mean that $\lim_{n\to\infty} a_n/b_n = c$, where c is some nonzero finite constant. The assumptions employed for the asymptotic results are listed below:

A1. For any given $\mathbf{z} \in \mathcal{D}$, the regression function satisfies $D^\alpha g \in C^{0,1}[0,1]^q$, for all α with $[\alpha] = p-1$ and $1 \leq p \leq \min(m_1, \ldots, m_q)$.
A2. The marginal density $f(\mathbf{x})$ of $\mathbf{X}$ satisfies $f(\mathbf{x}) \in C[0,1]^q$ and $f(\mathbf{x}) \in [c_f, C_f]$ for constants $0 < c_f \leq C_f < \infty$. There exists a constant $c_P > 0$, such that $P(\mathbf{Z} = \mathbf{z}\,|\mathbf{X}) \geq c_P$ for all $\mathbf{z} \in \mathcal{D}$.
A3. The noise ε satisfies $E(\varepsilon\,|\mathbf{X}, \mathbf{Z}) = 0$, $E(\varepsilon^2\,|\mathbf{X}, \mathbf{Z}) = 1$. There exists a positive value $\delta > 0$ and finite positive M_δ such that $\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} E\left(|\varepsilon|^{2+\delta}\,|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}\right) < M_\delta$. The standard deviation function $\sigma(\mathbf{x}, \mathbf{z})$ is continuous on $[0,1]^q$ for each given $\mathbf{z} \in \mathcal{D}$, and $0 < c_\sigma \leq \inf_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} \sigma(\mathbf{x}, \mathbf{z}) \leq \sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} \sigma(\mathbf{x}, \mathbf{z}) \leq C_\sigma < \infty$.
A4. The number of interior knots N_l, $1 \leq l \leq q$ satisfy as $n \to \infty$, $N_l \to \infty$ for $1 \leq l \leq q$, and $\prod_{l=1}^q N_l = o\{n(\log n)^{-1}\}$, and the bandwidths λ_s, $1 \leq s \leq r$ satisfy $\sum_{s=1}^r \lambda_s = o(1)$.

Assumption A1 provides a smoothness condition of the regression function $g(\cdot)$, which is also given in Section 5.2 of Huang (2003). Assumption A3 states the moment requirements of the standardized error term as well as the uniform bounds of the standard deviation. A similar assumption is given in Section 4.1 of Huang (2003). Assumption A4 states the order requirements for the smoothing parameters. The theorem below gives the uniform convergence rate of the estimator to the true mean function.

Theorem 1. Under assumptions A1–A4

$$\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} |\hat{g}(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})| = O_{\mathrm{a.s.}}\left[\sum_{l=1}^q N_l^{-p} + \sum_{s=1}^r \lambda_s + \left\{n^{-1}\left(\prod_{l=1}^q N_l\right)\log n\right\}^{1/2}\right]$$

and then for $N_l \asymp n^{1/(2p+q)}$ for $1 \leq l \leq q$ and $\sum_{s=1}^r \lambda_s = o\left[\{n^{-1}\left(\prod_{l=1}^q N_l\right)\log n\}^{1/2}\right]$, we have $\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} |\hat{g}(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})| = O_{\mathrm{a.s.}}\left\{n^{-p/(2p+q)}(\log n)^{1/2}\right\}$.

DOI: 10.1002/jae

Let

$$\begin{aligned}\Sigma_{\mathbf{z},n} &= E\left\{\mathcal{B}_{j_1,\ldots,j_q}(\mathbf{X})\mathcal{B}_{j_1',\ldots,j_q'}(\mathbf{X})L^2(\mathbf{Z},\mathbf{z},\lambda)\sigma^2(\mathbf{X},\mathbf{Z})\right\}\\ \mathbf{V}_{\mathbf{z},n} &= E\left\{\mathcal{B}_{j_1,\ldots,j_q}(\mathbf{X})\mathcal{B}_{j_1',\ldots,j_q'}(\mathbf{X})L(\mathbf{Z},\mathbf{z},\lambda)\right\}\end{aligned} \tag{5}$$

For $(\mathbf{x},\mathbf{z}) \in [0,1]^q \times \mathcal{D}$ define

$$\hat{\sigma}_n^2(\mathbf{x},\mathbf{z}) = n^{-1}\mathcal{B}(\mathbf{x})^{\mathrm{T}}\mathbf{V}_{\mathbf{z},n}^{-1}\Sigma_{\mathbf{z},n}\mathbf{V}_{\mathbf{z},n}^{-1}\mathcal{B}(\mathbf{x}) \tag{6}$$

Theorem 2. Under assumptions A1–A4, $\max_{1\le l\le q}(N_l^{-1}) = o\{n^{-1/(2p+q)}\}$, and $\sum_{s=1}^{r}\lambda_s = o\left\{(n^{-1}\prod_{l=1}^{q}N_l)^{1/2}\right\}$, we have as $n\to\infty$, for $\hat{\sigma}_n^2(\mathbf{x},\mathbf{z})$ in equation (6), $\hat{\sigma}_n^{-1}(\mathbf{x},\mathbf{z})\{\hat{g}(\mathbf{x},\mathbf{z}) - g(\mathbf{x},\mathbf{z})\} \longrightarrow \mathbf{N}(0,1)$. For any given $(\mathbf{x},\mathbf{z}) \in [0,1]^q \times \mathcal{D}$, with probability 1, $c_\sigma^* n^{-1}\left(\prod_{l=1}^{q}N_l\right) \le \hat{\sigma}_n^2(\mathbf{x},\mathbf{z}) \le C_\sigma^* n^{-1}\left(\prod_{l=1}^{q}N_l\right)$, for some constants $0 < c_\sigma^* < C_\sigma^* < \infty$

Comments. Theorem 1 gives the uniform convergence rate of the estimator to the true mean function. Theorem 2 presents the asymptotic normality of the proposed estimator with the bias vanished given that the order assumptions of smoothing parameters in Theorem 2 holds. Based on results in Theorem 2, we obtain the L_2 convergence rate of the estimator which is $O_{\mathrm{a.s.}}\{n^{-1/2}\left(\prod_{l=1}^{q}N_l\right)^{1/2}\}$.

Proofs of these theorems are presented in the supplementary Web Appendix. Having outlined the theoretical underpinnings of the proposed method, we now consider an illustrative simulated example that demonstrates how smoothing the categorical predictors in the manner prescribed above impacts the resulting estimator $\hat{g}(\mathbf{x},\mathbf{z})$.

2.1. Marginal Effects

Marginal effects are routinely required and reported by applied econometricians, and one of the appealing aspects of our approach is that, since the estimator is simply a (weighted) least squares estimator where the continuous predictors have been replaced with their B-spline representations, the computation of marginal effects is a straightforward exercise requiring nothing more than the ability to compute derivatives of the B-spline basis with respect to each of the continuous predictors. The R (R Core Team, 2013) package 'crs' (Racine and Nie, 2014; Nie and Racine, 2012) that implements the proposed approach produces marginal effects for an estimated model, rendering the method directly applicable in econometric applications. The estimation of marginal effects is illustrated in Figures 2 and 3 in Section 5.

3. CROSS-VALIDATED CHOICE OF N AND λ

Cross-validation (Stone, 1977) has a rich pedigree in the regression spline arena and has been used for decades to choose the appropriate number of interior knots and is the basis for Friedman's (1991) multivariate adaptive regression spline (MARS) methodology, among others; see Wahba (1990) for an overview in the spline context. More recently, Ma and Racine (2013) provide theoretical underpinnings for cross-validated selection of bandwidths, number of interior knots, and spline orders for semiparametric additive models that admit categorical predictors in the same manner as that used above, though here we consider selection of bandwidths and number of knots for a fully nonparametric model (the cross-validated selection of the spline order in Ma and Racine (2013) was to address the semiparametric additive case where a predictor may be irrelevant but this is not known a priori). Cross-validation has also been used extensively for bandwidth selection for kernel estimators such as the local linear kernel estimator proposed by Li and Racine (2004) that appears in the simulations in Section 4 (see also Racine and Li (2004) for the local constant counterpart), while Hall and

DOI: 10.1002/jae

Racine (2013) consider selecting both the polynomial order and bandwidths for local polynomial kernel regression via cross-validation. Following in this tradition we choose the number of interior knots (i.e. the vector (N)) and smoothing parameters (i.e. the bandwidth vector λ) by minimizing the cross-validation function defined by

$$CV(N,\lambda) = n^{-1}\sum_{i=1}^{n}\left(Y_i - B_m(X_i)^T\hat{\beta}_{-i}(Z_i)\right)^2 \tag{7}$$

where $\hat{\beta}_{-i}(Z_i)$ denotes the leave-one-out estimate of β. Theoretical properties of this approach follow directly from Ma and Racine (2013, Theorem 1), with the only difference being the replacement of the additive spline used in Ma and Racine (2013) with the tensor-product spline. Hence we do not replicate the proof here; rather we refer the interested reader to Ma and Racine (2013) (note that when $q = 1$ the additive and tensor bases coincide; hence results in Ma and Racine (2013) are directly applicable without modification in the univariate continuous predictor case). We investigate the performance of the cross-validated estimator via Monte Carlo simulation in Section 4 below, which reveals that cross-validation indeed delivers smoothing parameters that are consistent with theoretical underpinnings of the estimator provided in Section 2 above and in Ma and Racine (2013) (with appropriate substitution of basis function when $q > 1$ as indicated above).

To illustrate the finite-sample behavior of the data-driven cross-validated selection of N and λ, we consider four simple data-generating processes (DGPs) and use the popular cubic B-spline basis throughout. For each figure we simulate $n = 500$ observations and regress Y_i on X_i and Z_i using the proposed regression spline method. Results are plotted in Figure 1.

Figure 1 illustrates how the cross-validated choices of N and λ differ depending on the underlying DGP. For instance, the plot at the upper left is one for which $Y_i = \cos(\pi X_i) + \varepsilon_i$ if $Z_i = 0$ and

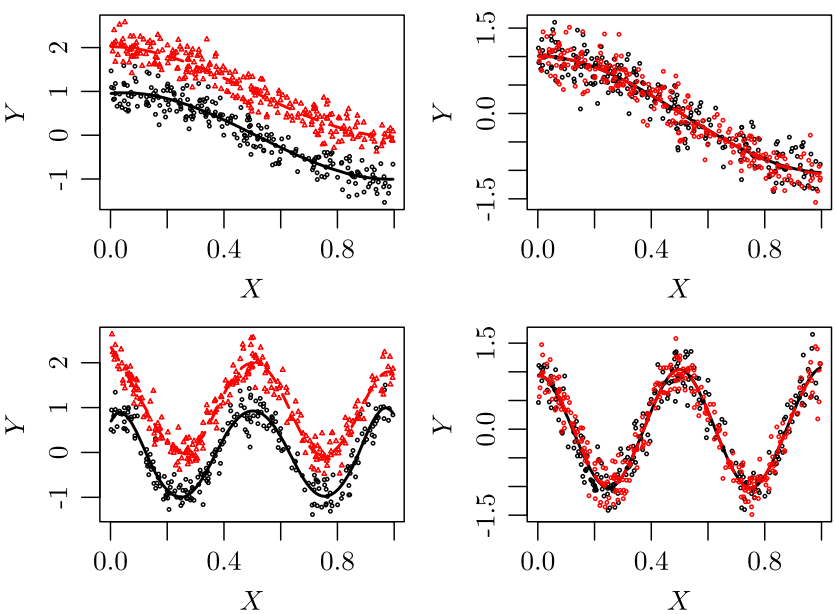

Figure 1. Four illustrative DGPs, $n = 500$. For the plots on the left there is a vertical shift in the DGP when $Z = 0/1$, while for the plots on the right there is no difference in the function when $Z = 0/1$. For the latter, cross-validation ought to select a larger value of λ and for the former a smaller value of λ would be expected ($\lambda \in [0, 1]$).

DOI: 10.1002/jae

$Y_i = 1 + \cos(\pi X_i) + \varepsilon_i$ if $Z_i = 1$. It is evident that the cross-validated choices, $N \approx 0$ and $\lambda \approx 2.2e-16$, are appropriate here. The plot at the upper right is one for which $Y_i = \cos(\pi X_i) + \varepsilon_i$ regardless of the value taken on by Z_i. It is evident that the cross-validated choices, $N \approx 0$ and $\lambda \approx 0.91$, are appropriate here. The plots at the lower left and lower right are similar but use $\cos(4\pi X_i)$ instead, and the cross-validated choices $N \approx 3$ and $\lambda \approx 2.2e-16$ were selected for the DGP at the bottom left, $N \approx 3$ and $\lambda \approx 0.82$ for the bottom right. The relatively large values of λ for the figures at the right are appropriate since Z_i is independent of Y_i. These simple examples serve to illustrate how cross-validation is delivering values of N and λ that are tailored to the DGP at hand.

Before proceeding, a few words on the numerical optimization of Equation (7) are in order. Search takes place over $N_1, \ldots, N_q$ and $\lambda_1, \ldots, \lambda_r$ where the λ are continuous lying in $[0, 1]$ and the N are integers. Clearly this is a mixed-integer combinatorial optimization procedure which would render exhaustive search infeasible when facing a non-trivial number of predictors. However, in settings such as these one could leverage recent advances in mixed-integer search algorithms, which is the avenue we pursue in the Monte Carlo simulations reported below. In particular, we adopt the 'nonsmooth optimization by mesh adaptive direct search' (NOMAD) approach (Abramson *et al.* (Abramson et al., 2011)). Given that the objective function can be trivially computed for large sample sizes as it involves nothing more than computing the hat matrix for weighted least squares, it turns out that the computational burden is in fact nowhere near as costly as, say, cross-validated kernel regression for moderate to large datasets. As such, the proposed approach constitutes a computationally attractive alternative to multivariate cross-validated kernel regression. In addition, in the next section we shall also see that the proposed approach constitutes a statistically attractive alternative as well, at least from the perspective of finite-sample squared error loss.

4. MONTE CARLO SIMULATIONS

In this section we consider a modest Monte Carlo experiment designed to assess the finite-sample performance of the proposed method. We consider two DGPs with $q = 2$ continuous predictors and $r = 2$ categorical predictors given by

$$\text{DGP-M: } \cos(4\pi X_{i1}) \times \sin(4\pi X_{i2}) \times (Z_{i1} + 1) \times (Z_{i2} + 1) \tag{8}$$

$$\text{DGP-A: } \cos(4\pi X_{i1}) + \sin(4\pi X_{i2}) + (Z_{i1} + Z_{i2})/10 \tag{9}$$

then set $Y_i = \text{DGP-M}/\hat{s}_{\text{DGP-M}} + \varepsilon_i$, where $\hat{s}_{\text{DGP-M}}$ is the sample standard deviation of DGP-M (we do the same for DGP-A), where the continuous predictors are drawn from the uniform ($X_j \sim U[0, 1]$, $j = 1, 2$), the categorical predictors (Z_j, $j = 1, 2$) are drawn from the rectangular distribution with equal probability ($z_s \in \{0, 1, \ldots, c_s - 1\}$, where c_s is the number of categorical outcomes, $c_s \geq 2$, $s = 1, 2$) and $\varepsilon \sim N(0, \sigma^2)$ with $\sigma = \{0.25, 0.5, 1.0, 2.0\}$. Dividing the DGPs by their sample standard deviations controls the signal-to-noise ratio across all simulations and corresponds to an expected coefficient of determination of 0.95, 0.8, 0.5 and 0.2, respectively. For what follows we set $c_s = c$, $s = 1, 2$. Observe that DGP-M is multiplicative in all components, while DGP-A is additive in all components.

We draw $M = 1,000$ Monte Carlo replications and for each replication we compute the cross-validated frequency estimator (i.e. that based only on the $(Y, \mathbf{X})$ pairs lying in each 'cell' defined by $\mathbf{Z}$), the proposed cross-validated categorical regression spline estimator, the cross-validated additive categorical regression spline estimator (Ma and Racine, 2013) and the cross-validated local linear kernel estimator that is often used to model continuous and categorical predictors in a manner similar to that undertaken here (note that the local linear kernel estimator is minimax efficient and has the best boundary bias correction properties of the class of kernel regression estimators; see Li and Racine,

DOI: 10.1002/jae

Table I. Relative median MSE of the proposed regression spline estimator versus the frequency regression spline, additive regression spline and local linear kernel regression estimators. Numbers less than one indicate superior performance of the spline estimator. c denotes the number of outcomes for the discrete predictors, n the sample size.

DGP-M: multiplicative specification					DGP-A: additive specification				
n	c	Frequency	Additive	Kernel	n	c	Frequency	Additive	Kernel
$\sigma = 0.25$							$\sigma = 0.25$		
500	2	0.376	0.029	0.642	500	2	0.206	2.313	0.607
500	3	0.089	0.045	0.854	500	3	0.058	1.922	0.664
500	4	0.137	0.054	0.964	500	4	0.087	1.773	0.717
1000	2	0.799	0.016	0.463	1000	2	0.418	2.448	0.544
1000	3	0.191	0.027	0.606	1000	3	0.124	2.070	0.595
1000	4	0.078	0.037	0.749	1000	4	0.060	1.863	0.625
$\sigma = 0.5$							$\sigma = 0.5$		
500	2	0.389	0.081	0.652	500	2	0.215	2.557	0.704
500	3	0.189	0.110	0.743	500	3	0.108	2.065	0.731
500	4	0.255	0.129	0.792	500	4	0.137	1.846	0.755
1000	2	0.741	0.049	0.535	1000	2	0.353	2.311	0.609
1000	3	0.295	0.076	0.623	1000	3	0.132	1.983	0.668
1000	4	0.169	0.096	0.677	1000	4	0.092	1.836	0.704
$\sigma = 1.0$							$\sigma = 1.0$		
500	2	0.407	0.204	0.724	500	2	0.236	2.410	0.819
500	3	0.297	0.256	0.775	500	3	0.158	2.131	0.799
500	4	0.317	0.292	0.804	500	4	0.156	1.978	0.825
1000	2	0.605	0.133	0.645	1000	2	0.292	2.511	0.753
1000	3	0.319	0.180	0.692	1000	3	0.140	2.178	0.763
1000	4	0.253	0.212	0.724	1000	4	0.115	1.930	0.777
$\sigma = 2.0$							$\sigma = 2.0$		
500	2	0.447	0.532	0.948	500	2	0.335	2.542	1.091
500	3	0.326	0.587	0.952	500	3	0.220	2.280	1.081
500	4	0.257	0.623	0.953	500	4	0.166	2.278	1.077
1000	2	0.460	0.321	0.777	1000	2	0.283	2.458	0.968
1000	3	0.299	0.385	0.817	1000	3	0.176	2.383	0.980
1000	4	0.248	0.428	0.839	1000	4	0.136	2.198	0.956

2004, for details). For the regression spline estimators we set the spline degree vector equal to $(3, 3)$ (a popular choice) and cross-validate the number of knots vector (N_1, N_2) and the bandwidth vector (λ_1, λ_2). We then compute the mean squared error (MSE) of each estimator based upon Equation (8) for each replication and report the relative median MSE over all M replications in Table I (results for the mean MSE over all M Monte Carlo replications are comparable to those for the median MSE over the M replications and are not reported here for space considerations). MSE is computed via $n^{-1}\sum_{i=1}^{n}(\hat{g}(\mathbf{x}_i, \mathbf{z}_i) - g(\mathbf{x}_i, \mathbf{z}_i))^2$, where $\hat{g}(\mathbf{x}_i, \mathbf{z}_i)$ is defined immediately following (2) and $g(\mathbf{x}_i, \mathbf{z}_i)$ is a Monte Carlo DGP defined in Equation (8). Table II reports a summary of the smoothing parameters chosen by cross-validation.

Table I illustrates how, for a given sample size, the relative performance of the proposed approach that smooths the categorical predictors versus the frequency approach that breaks the data into $c_1 \times c_2 = (4, 9, 16)$ subsets improves as c increases, as expected (each categorical predictor takes on $c_1 = c_2 = c = (2, 3, 4)$ values). Table II reveals how the cross-validated bandwidths tend to zero as n increases. These findings are consistent with the theoretical properties detailed in the supplementary Web Appendix and in Ma and Racine (2013), as previously noted. Furthermore, for a given c, as n increases the proposed estimator approaches the frequency estimator since $\lambda \to 0$ as $n \to \infty$. Table I further illustrates how the proposed cross-validated method dominates the popular local linear kernel estimator for (almost) all sample sizes and values of c and σ considered.

DOI: 10.1002/jae

Table II. Median values for the number of interior knots and bandwidths for the proposed regression spline estimator. $c = c_1 = c_2$ denotes the number of outcomes for the discrete predictors, n the sample size, N_j the number of interior knots for continuous predictor X_j, $j = 1, 2$ and λ_j the bandwidth for continuous predictor Z_j, $j = 1, 2$.

DGP-M: multiplicative specification						DGP-A: additive specification					
n	c	N_1	N_2	λ_1	λ_2	n	c	N_1	N_2	λ_1	λ_2
$\sigma = 0.25$								$\sigma = 0.25$			
500	2	4	5	0.08	0.08	500	2	4	5	0.64	0.66
500	3	4	3	0.09	0.09	500	3	4	5	0.49	0.50
500	4	4	3	0.10	0.10	500	4	4	5	0.38	0.38
1000	2	4	5	0.05	0.04	1000	2	4	6	0.46	0.48
1000	3	4	5	0.05	0.05	1000	3	4	5	0.33	0.33
1000	4	4	5	0.05	0.05	1000	4	4	5	0.25	0.25
$\sigma = 0.5$								$\sigma = 0.5$			
500	2	4	3	0.17	0.17	500	2	4	3	0.85	0.86
500	3	4	3	0.19	0.18	500	3	4	3	0.76	0.74
500	4	4	3	0.18	0.18	500	4	4	3	0.63	0.64
1000	2	4	5	0.11	0.11	1000	2	4	5	0.77	0.74
1000	3	4	5	0.11	0.11	1000	3	4	5	0.61	0.61
1000	4	4	3	0.11	0.11	1000	4	4	5	0.48	0.49
$\sigma = 1.0$								$\sigma = 1.0$			
500	2	4	3	0.38	0.39	500	2	4	3	0.97	0.98
500	3	4	3	0.36	0.37	500	3	4	3	0.90	0.91
500	4	4	3	0.38	0.37	500	4	4	3	0.89	0.87
1000	2	4	3	0.24	0.25	1000	2	4	3	0.93	0.93
1000	3	4	3	0.25	0.25	1000	3	4	3	0.84	0.81
1000	4	4	3	0.24	0.24	1000	4	4	3	0.76	0.76
$\sigma = 2.0$								$\sigma = 2.0$			
500	2	4	3	0.69	0.67	500	2	2	3	1.00	1.00
500	3	2	3	0.65	0.65	500	3	2	3	1.00	1.00
500	4	2	3	0.66	0.66	500	4	2	3	0.92	0.98
1000	2	4	3	0.51	0.50	1000	2	4	3	1.00	1.00
1000	3	4	3	0.50	0.50	1000	3	4	3	0.93	0.95
1000	4	4	3	0.50	0.51	1000	4	4	3	0.90	0.89

Often additive spline models are recommended in applied settings due to the curse of dimensionality (the property that the multiplicative tensor regression spline method has a rate of convergence that deteriorates with the number of continuous predictors, q). Of course, this curse is not unique to the proposed method and is a function of many fully nonparametric estimators, including the local linear kernel estimator considered in the simulations below, by way of illustration. Observe, however, that even in small-sample settings such as $n = 500$, if the additive model is used when additivity is not present the squared error properties of the additive regression spline model can be much worse than the multiplicative tensor regression spline model that we propose. Naturally, if additivity is appropriate, the additive model that incorporates this information will have better finite-sample properties (the tensor model has roughly 2 times the MSE of the additive model for DGP-A, the additive DGP). Simulations not reported here for space considerations indicate that the finite-sample mean squared error improvement over the kernel regression estimator holds (a) whether or not there exist categorical predictors and (b) in higher-dimension settings than those reported here.

In the above simulations the tensor-based multivariate regression spline approach outperforms the popular local linear kernel regression approach. However, we caution the reader that this is not guaranteed always to be the case (see, for example, DGP-A, $\sigma = 2.0, n = 500$). The dimension of the tensor spline basis grows exponentially with the number of continuous predictors for a given order/knot combination for each predictor. Thus, for a fixed sample size n, as the number of continuous predictors q

DOI: 10.1002/jae

increases, degrees of freedom will quickly fall and the squared error properties of the resulting estimator will naturally deteriorate. Thus, in settings with small n, large q and low degrees of freedom, one could readily construct instances in which the local linear kernel regression approach will display better finite-sample behavior than the regression spline approach. We therefore offer the following advice for the sound practical application of the methods proposed herein:

1. The proposed methods are best suited to settings in which q is not overly large and n not overly small.

   Experience indicates that for a range of DGPs the regression spline outperforms kernel regression when $n \geq 500$ and $q \leq 5$.

   One practical advantage is the reduced computational burden of cross-validation for regression splines versus their kernel counterpart, and in large sample settings (say, $n \geq 10,000$) one can push the dimension of q much higher than that considered here.
2. Of course, when the dimension of the multivariate tensor spline becomes a practical barrier to their sound application, one can always resort to additive spline models; see Ma and Racine (2013) for details. The drawback of the additive spline approach is that if the DGP is non-additive the inefficiency of the additive spline approach can be much worse than the multivariate kernel approach, as clearly demonstrated above. Of course, it is a simple matter to compare the value of the cross-validation function for each of the tensor-based, additive-based and kernel-based cross-validated approaches and it is perfectly sensible to use this as a guide in applied settings. But our experience is that the tensor-based multivariate regression spline will indeed be competitive and ought to be part of every practitioner's toolkit.

In summary, the simulations outlined above indicate that the proposed method is capable of outperforming the frequency estimator that breaks the sample into subsets, while it provides a compelling alternative to kernel methods when faced with a mix of categorical and continuous predictors and to additive regression spline models for general nonlinear DGPs for which additivity is not fully present.

5. APPLICATIONS

In this section we consider two illustrative applications chosen to highlight the flexibility of the proposed approach in applied settings.

5.1. GDP Growth and OECD Status

By way of illustration, we model growth rates of per capita GDP where the predictors are OECD status ('oecd'), the initial level of GDP ('initgdp') and investment ('inv'). We treat OECD status as categorical and initial GDP and investment as continuous. Initial income estimates are from Heston *et al.* (2009), as are the estimates of the the average investment/GDP ratio for 5-year periods. The average growth rate of the per capita GDP is from the World Bank. This data represent a cross country GDP growth panel covering the period 1960–1995. There were $n = 616$ observations in total.

We first compute parametric specifications which are linear in predictors, then quadratic in initial DGP (a popular parametric specification found in the literature) and then compute the proposed categorical spline approach. For each model the coefficient of determination was 0.176, 0.19 and 0.342, respectively. For the categorical spline model the cross-validated values of N_{initgdp}, N_{inv} and λ_{OECD} were 1, 4 and 0.041, respectively. Figure 2 presents the partial regression surfaces and marginal effects (first derivatives) along with their 95% pointwise confidence intervals using results presented in Theorem 2 (a 'partial regression plot' is simply a 2D plot of the estimated model versus one predictor when all other predictors are held constant at their respective medians/modes).

DOI: 10.1002/jae

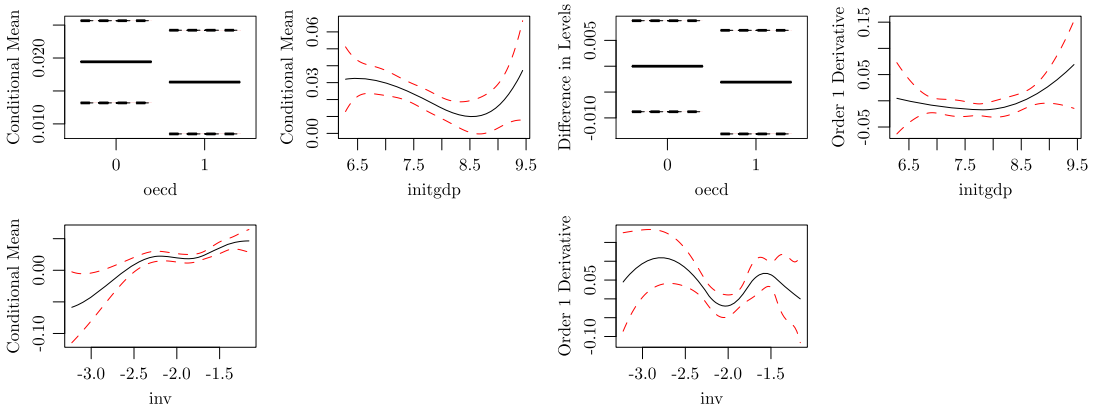

Figure 2. Partial regression plots (left) and marginal effects plots (right) for the OECD data with asymptotic 95% pointwise confidence bounds.

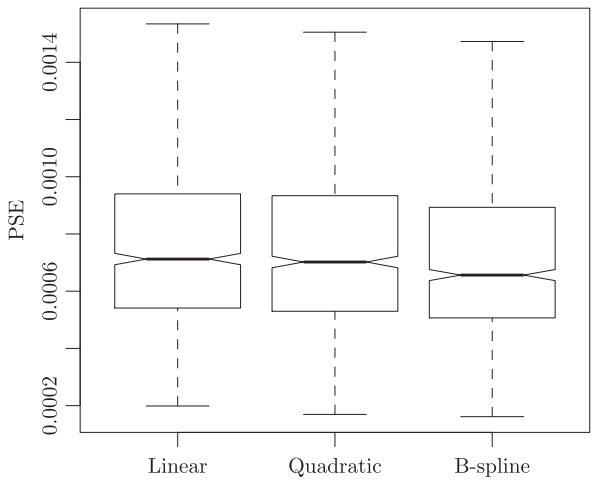

Figure 3. Out-of-sample predicted squared error (PSE) box plots for the linear, quadratic and B-spline models for the OECD data. The median PSEs of the categorical spline model relative to the linear and quadratic parametric models are 0.9215 and 0.9350, respectively.

One might naturally wonder whether the improvement in fit is illusory and merely reflects 'overfitting' by the spline model. In order to investigate whether or not the categorical spline model is merely overfitting we conduct a pseudo Monte Carlo experiment in which we randomly shuffle the 616 observations into a training sample of size $n_1 = 600$ and an evaluation sample of size $n_2 = 16$ observations. For each random shuffle of the data, we fit the categorical spline model and the linear and quadratic parametric regression models on the $n_1 = 600$ training observations and compute the out-of-sample predicted squared error (PSE) for the n_2 hold-out observations. We repeat this $M = 1000$ times and compute the median PSE of the categorical spline model relative to the linear and quadratic parametric models, which were 0.9215 and 0.9350, respectively. We also present a box plot of the PSEs in Figure 3.

DOI: 10.1002/jae

The second illustration involving hourly earnings can be found in the supplementary Web Appendix. For both of the illustrative applications considered we can see that the proposed approach produces models that appear to be more faithful to the underlying DGP than two simple parametric specifications that might be considered in applied settings.

6. CONCLUDING REMARKS

Applied researchers frequently must model relationships containing categorical predictors, yet may require nonparametric estimators of, say, regression functions. The traditional kernel and spline estimators break the data into subsets defined by the categorical predictors and then model the resulting relationship involving continuous predictors only. Though consistent, these approaches are acknowledged to be inefficient. In this paper we provide an approach combining regression splines with categorical kernel functions that is capable of overcoming the efficiency losses present in the traditional sample-splitting approach. Furthermore, the proposed approach constitutes an attractive alternative to cross-validated kernel estimators that admit categorical predictors. Theoretical underpinnings are provided and simulations are undertaken to assess the finite-sample performance of the proposed method. We hope that these methods are of interest to those modeling regression functions nonparametrically when faced with both continuous and categorical predictors. An implementation in R (R Core Team, 2013) is available from the Comprehensive R Archive Network (http://cran.r-project.org); see the R package 'crs' for details (Racine and Nie, 2014; Nie and Racine, 2012).

ACKNOWLEDGEMENTS

Racine would like to gratefully acknowledge support from the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca). Ma's research was partly supported by NSF grant DMS 1306972.

REFERENCES

Abramson M, Audet C, Couture G, Dennis J Jr, Le Digabel S. 2011. The NOMAD project. *Technical report*, GERAD (Groupe d'études et de recherche en analyse des décisions), Montreal. Software available: http://www.gerad.ca/nomad [Accessed on 24 August 2014].

Aitchison J, Aitken CGG. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* **63**(3): 413–420.

Andrews D. 1991. Asymptotic normality of series estimators for non-parametric and semiparametric regression models. *Econometrica* **59**: 307–345.

Chen X. 2007. Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics*, Vol. 6. Elsevier: Amsterdam; 5549–5632.

Friedman JH. 1991. Multivariate adaptive regression splines. *Annals of Statistics* **19**(1): 1–67.

Greiner A. 2009. Estimating penalized spline regressions: theory and application to economics. *Applied Economics Letters* **16**: 1831–1835.

Hall P, Racine JS. 2013. (forthcoming) Infinite Order Cross-Validated Local Polynomial Regression. Journal of Econometrics.

Haupt H, Kagerer K, Steiner W. 2014. Smooth quantile based modeling of brand sales, price and promotional effects from retail scanner panels. *Journal of Applied Econometrics*. DOI 10.1016/j.jeconom.2014.06.003, (in press) Available at: http://www.sciencedirect.com/science/article/pii/S0304407614001432.

Heston A, Summers R, Aten B. 2009. Penn world table version 6.3, *Technical report*, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, Philadelphia.

Huang JZ. 1998. Projection estimation in multiple regression with application to functional ANOVA models. *Annals of Statistics* **26**: 242–272.

Huang JZ. 2003. Local asymptotics for polynomial spline regression. *Annals of Statistics* **31**: 1600–1635.

DOI: 10.1002/jae

Huang JZ, Yang L. 2004. Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society Series B* **66**: 463–477.
Li Q, Racine JS. 2004. Cross-validated local linear nonparametric regression. *Statistica Sinica* **14**(2): 485–512.
Li Q, Racine J. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press: Princeton, NJ.
Liu R, Yang L. 2010. Spline-backfitted kernel smoothing of additive coefficient model. *Econometric Theory* **26**: 29–59.
Ma S, Racine JS. 2013. Additive regression splines with irrelevant regressors. *Statistica Sinica* **23**: 515–541.
Ma S, Yang L. 2011. A jump-detecting procedure based on spline estimation. *Journal of Nonparametric Statistics* **23**: 67–81.
Newey W. 1991. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79**: 147–168.
Nie Z, Racine JS. 2012. The crs package: nonparametric regression splines for continuous and categorical predictors. *The R Journal* **4**(2): 48–56.
Racine JS, Li Q. 2004. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **119**(1): 99–130.
Racine J S, Nie Z. 2014. crs: categorical regression splines. R package version 0.15-22. Available from: https://github.com/JeffreyRacine/R-Package-crs/ [Accessed on 24 August 2014].
Core Team R. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna.
Stone CJ. 1985. Additive regression and other nonparametric models. *Annals of Statistics* **13**: 689–705.
Stone CJ. 1994. The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics* **22**: 118–184.
Stone M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* **39**(1): 44–47.
Wahba G. 1990. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics: Philadelphia, PA.
Wang J, Yang L. 2009. Polynomial spline confidence bands for regression curves. *Statistica Sinica* **19**: 325–342.
Wang L. 2009. Single-index model-assisted estimation in survey sampling. *Journal of Nonparametric Statistics* **21**: 487–504.
Wang L, Yang L. 2007. Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Annals of Statistics* **35**: 2474–2503.
Wang L, Yang L. 2009. Spline estimation of single index model. *Statistica Sinica* **19**: 765–783.
Xue L. 2009. Variable selection in additive models. *Statistica Sinica* **19**: 1281–1296.
Xue L, Liang H. 2010. Polynomial spline estimation for the generalized additive coefficient model. *Scandinavian Journal of Statistics* **37**: 26–46.

DOI: 10.1002/jae