International Journal of Hospitality Management 76 (2019) 43–47

Contents lists available at ScienceDirect

International Journal of Hospitality Management

journal homepage: www.elsevier.com/locate/ijhm

Non-parametric regression for hypothesis testing in hospitality and tourism research

A. George Assaf[a,*,1], Mike Tsionas[b,c,1]

[a] Isenberg School of Management, University of Massachusetts-Amherst, 90 Campus Center Way, 209A Flint Lab, Amherst, MA, 01003, United States
[b] Lancaster University Management School, United Kingdom
[c] Athens University of Economics and Business, Greece

ARTICLE INFO

Keywords:
Non-Parametric Regression
Bayesian
GPP

ABSTRACT

The goal of this paper is to promote the use of Non-Parametric Regression (NPR) for hypothesis testing in hospitality and tourism research. In contrast to linear regression models, NPR frees researchers from the need to impose a priori specification on functional forms, thus allowing more flexibility and less vulnerability to misspecification problems. Importantly, we discuss in this paper a Bayesian approach to NPR using a Gaussian Process Prior (GPP). We illustrate the advantages of this method using an interesting application on internationalization and hotel performance. Specifically, we show how in contrast to linear regression, NPR decreases the risk of making incorrect hypothesis statements by revealing the true and full relationship between the variables of interest.

1. Introduction

Despite the increased popularity of non-parametric regression (NPR), its use in the tourism and hospitality literature remains very limited. We aim in this note to highlight the advantages of NPR, and illustrate how it can be used to provide a more accurate reflection on the true relationship between a set of variables. We show through an example that hospitality researchers might be missing some important input for hypothesis testing when estimating the traditional linear regression model.

NPR, like linear regression, estimates mean outcomes for a given set of covariates. However, unlike linear regression, NPR is not subject to misspecification error arising from potentially wrong functional forms as it does not impose a priori a functional form on the regression model (Müller, 2012; Mammen et al., 2012). The linear model ($y = \beta_0 + \beta x + u$) is generally assumed for convenience, and not because we truly believe that the model is linear in reality.

Researchers in the field often model nonlinearities using extensions of the linear model, for example, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$. It is clear, however, that this model accounts only for limited types of nonlinearity of U or inverted U shape, and cannot capture more complicated patterns in the data. When more than one regressor is available, nonlinearities are often modeled using interactions: $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + u$. The interpretation is that the effect of x on y depends on z: $\frac{\partial E(y)}{\partial x} = \beta_1 + \beta_3 z$. This is, of course, a deviation from the simple linear model where the main assumption is that the effect of x on y is constant across all values of x or other explanatory variables. However, the effect of x on y depends on z in a linear way, an assumption that may or may not hold in practice.

Let us illustrate here the above with a small example: we generate for instance, 100 observations from the model: $y_i = \exp(-\sin(x_i)) + 0.5\varepsilon_i$, where the ε_is are standard normal random variables. The x_is are generated as a sequence in the interval $[-3, 3]$ with step 6/99. The results (Fig. 1) illustrate nicely what happens when a linear model is fitted to data, which have been generated through a nonlinear model. It is a complete miss. As mentioned, the linear model is only an approximation to an unknown regression function of the form: $y = f(x) + u$. The non-parametric regression does not assume that $f()$ is linear; it can in fact be non-linear. NPR does not also assume that $f()$ is linear in the parameters. It could be actually anything. In nonparametric analysis, we seek to estimate directly the unknown function $f(x)$ when observations $\{x_i, y_i, i = 1,..., n\}$ are available. The model for each observation is $y_i = f(x_i) + u_i$ or $y_i = f_i + u_i$, $i = 1, ..., n$ where $f_i = f(x_i)$. The unknown function values $f_1, ..., f_n$ are then treated as parameters. Clearly, the number of parameters in this instance, rises with the sample size. However, it is possible to obtain consistent estimates if we assume that the regression function is sufficiently smooth (i.e. possesses continuous derivatives of a certain order).

* Corresponding author.
E-mail addresses: assaf@isenberg.umass.edu (A.G. Assaf), m.tsionas@lancaster.ac.uk (M. Tsionas).
[1] Both authors have contributed equally to the paper.

https://doi.org/10.1016/j.ijhm.2018.04.002
Received 7 February 2018; Received in revised form 4 April 2018; Accepted 10 April 2018

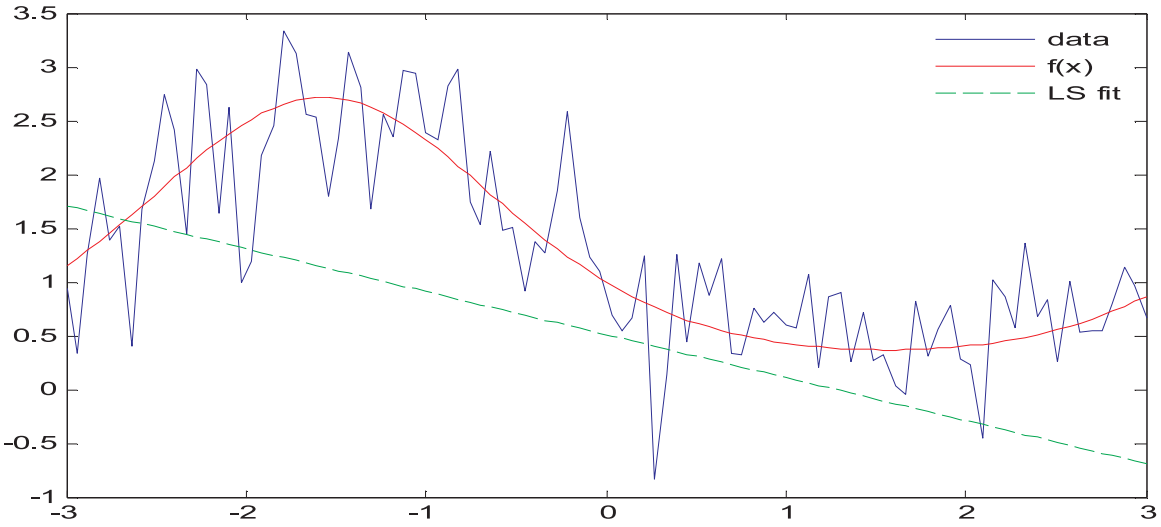

Fig. 1. NPR vs. Linear Regression: Results from Artificial Data.

Some popular non-parametric techniques include the Nadaraya – Watson estimator, kernel smoothing, local linear estimation etc. The situation is more difficult when the underlying model is: $y_i = f(x_i) + u_i$, where $x_i \in \mathbb{R}^k$ is a vector of explanatory variables. This situation is of interest because rarely if ever we have only one explanatory variable. The problem of non-parametric regression with multiple explanatory variables is a difficult problem. One approach is additive non-parametric regression: $y_i = f_1(x_{i1}) + f_2(x_{i2}) + ... + f_k(x_{ik}) + u_i$, where f_1, f_2, ..., f_k are unknown functional forms. In this model, however, the effect of any regressor on the dependent variable does not depend on the values of the other regressors; an assumption that is unlikely to be met in practice.

In this paper, we describe a Bayesian approach to NPR, using a Gaussian Process Prior (GPP), which is a popular and effective way of dealing with the problems of non-parametric multivariate regression (Williams and Rasmussen, 1996; Williams, 1998; MacKay, 1998; Vivarelli and Williams, 1999). We elaborate more on this method in the next section. We also present an application from the hotel literature.

2. Bayesian nonparametric regression through Gaussian process prior

Let us assume we have a dataset $\{y_i, x_i; i = 1,..., n\}$ where $x_i \in \mathbb{R}^d$ is a vector of predictors and y_i is the dependent variable. It is customary to use a linear regression model to perform statistical inferences:

$$y_i = x'_i\beta + u_i,\ i = 1,..., n, \tag{1}$$

where $\beta \in \mathbb{R}^d$ is a vector of fixed coefficients. The linear regression model is, in reality, only an approximation to a true regression model of the form

$$y_i = f(x_i) + u_i,\ i = 1,..., n, \tag{2}$$

where $f(x_i)$ is an unknown functional form. We assume $u_i \sim iidN(0, \sigma^2)$.

We use here a Gaussian Process Prior (GPP) to approximate the true but unknown functional form. Suppose $y = [y_1,...,y_n]'$ and $f = [f_1,...,f_n]'$represent, respectively, the vector of observations for the dependent variable and the vector of unknown function values at the observed regressors. Denote also $X = [x'_1,...,x'_n]'$ be the $n \times d$ matrix of observations on the regressors. The model can be written in the form:

$$y|f \sim N(f, \sigma^2 I). \tag{3}$$

The GPP places a prior upon the class of unknown functional forms:

$$f \sim N(0, \mathbb{K}), \tag{4}$$

where K with double bars is n*n matrix whose elements are defined by:

$$\mathbb{K}_{ij} = \kappa(x_i, x_j),\ i, j = 1,..., n,$$

where $\kappa(x_i, x_j)$ is a certain kernel function that measures the distance between different points. A popular choice is

$$\kappa(x_i, x_j) = \tau^2 e^{-(x_i - x_j)'(x_i - x_j)/\eta^2}, \tag{5}$$

where τ and η are hyperparameters to be selected along with σ.

It is instructive to consider what types of functions can be delivered through a GPP. Samples from a GPP with $\tau = 2$, $\eta = 1$ are shown in Fig. 2a and in Fig. 2b when $\tau = 3$, $\eta = 3$ in which case the resulting functions are closer to what we would expect in typical economic and management studies.

Typically, we are interested in evaluating (and presenting graphically) the unknown functional form at a different set of points, say $X^* = [x'^*_1,...,x'^*_m]$ where $x^*_i \in \mathbb{R}^d$, $i = 1,..., m$. Let $f^* = [f^*_1,...,f^*_m]$ denote the function values at these points. Therefore, we are interested in the posterior distribution $p(f^*|y)$. The model then is as follows:

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbb{K}_{xx} & \mathbb{K}_{xx^*} \\ \mathbb{K}'_{xx^*} & \mathbb{K}_{x^*x^*} \end{bmatrix} \right), \tag{6}$$

where $(K_{xx})_{ij} = \kappa(x_i, x_j)$, $(K_{xx^*})_{ij} = \kappa(x_i, x^*_j)$, $(K_{x^*x^*})_{ij} = \kappa(x^*_i, x^*_j)$. It is simple to show that we have:

$$f^*|y \sim N(\overline{f}^*, V), \tag{7}$$

where

$$\overline{f}^* = \mathbb{K}'_{xx^*}(\mathbb{K}_{xx} + \sigma^2 I)y, \tag{8}$$

$$V = \mathbb{K}_{x^*x^*} - \mathbb{K}'_{xx^*}(\mathbb{K}_{xx} + \sigma^2 I)\mathbb{K}_{xx^*}. \tag{9}$$

Based on (8) we can plot the unknown function at selected points. The log marginal likelihood of the model is:

$$\log M(y) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log\left|\mathbb{K}_\theta\right| - \tfrac{1}{2}y'\mathbb{K}_\theta^{-1}y, \tag{10}$$

where $\theta = [\eta, \tau, \sigma]'$ and K_θ shows explicitly the dependence of matrix K on the hyperparameters in θ. The log marginal likelihood can be maximized numerically with respect to the hyperparameters to provide the best possible choices that can, in turn, be used in (8) to provide the function values at the desired points.

3. Application

We illustrate the Bayesian non-parametric regression using an interesting application on the relationship between the degree of