Non-Parametric Regression for Hypothesis Testing in Hospitality and Tourism Research

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.

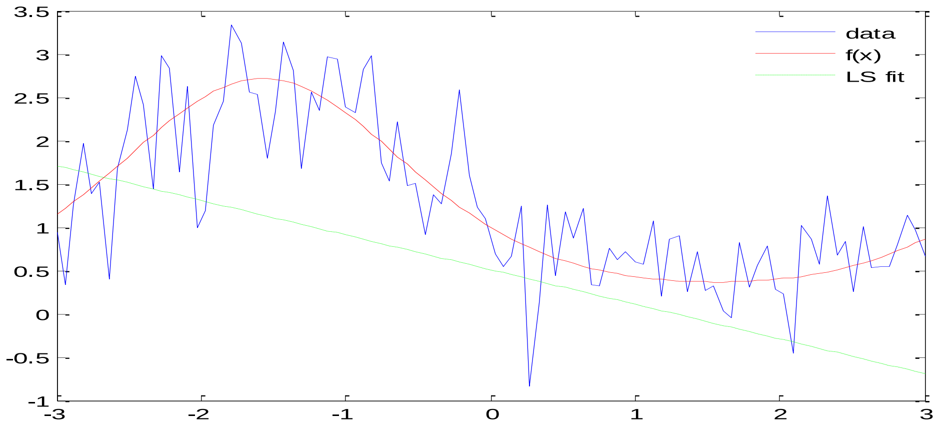
Researchers in the field often model nonlinearities using extensions of the linear model, for example, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$. It is clear, however, that this model accounts only for limited types of nonlinearity of U or inverted U shape, and cannot capture more complicated patterns 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. Moreover, the effect of x on y does not depend on x, a questionable assumption.

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_i s are standard normal random variables. The x_i s are generated as a sequence in the interval [-3, 3] with step 6/99. The results (Figure 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).

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.

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



2. Bayesian nonparametric regression through Gaussian process prior

Let us assume we have a dataset $\{y_i, x_i; i = 1,\ldots,n\}$ 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. The linear regression model is

$$y_i = x_i'\beta + u_i, i = 1,\ldots,n, \tag{1}$$

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

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

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

We use here a Gaussian Process Prior (GPP) to approximate the true but unknown functional form. Suppose $y = [y_1,\ldots,y_n]'$ and $f = [f_1,\ldots,f_n]'$ represent, respectively, the vector of observations for the dependent variable and the vector of unknown function values at the observed regressors. Denote also $X = [x_1',\ldots,x_n']'$ be the $n \times d$ matrix of observations on the regressors. The model can be written in the form:

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

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

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

where $\mathbb{K}$ is an $n \times n$ matrix whose elements are defined by:

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

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

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

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 Figure 2a and in Figure 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.

Figure 2a. Samples from a Gaussian Process Prior (τ=2, η=1)

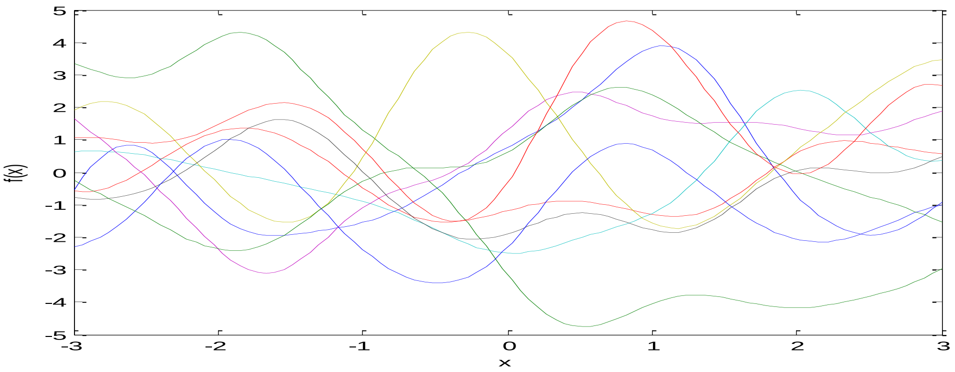
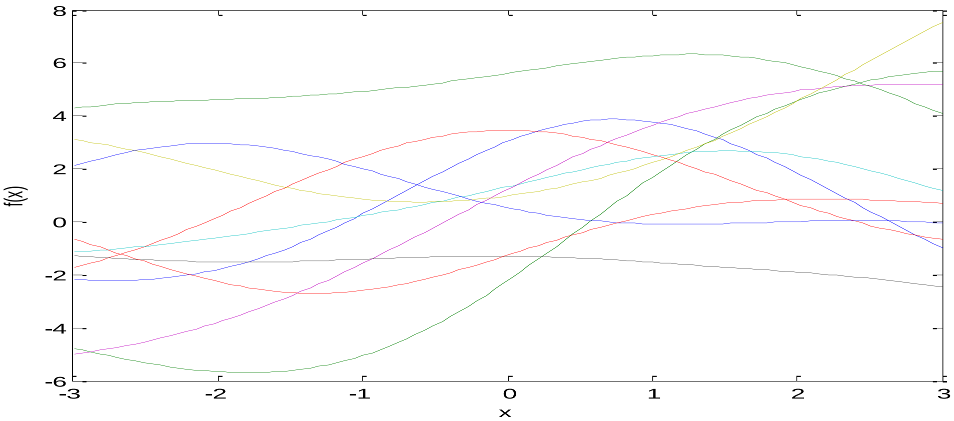


Figure 2b. Samples from a Gaussian Process Prior (τ=3, η=3)



Typically, we are interested in evaluating (and presenting graphically) the unknown functional form at a different set of points, say $X^* = \left[x_1'^*, \ldots, x_m'^*\right]$ where $x_i^* \in \mathbb{R}^d, i = 1, \ldots, m.$ Let $f^* = \left[f_1^*, \ldots, f_m^*\right]$ denote the function values at these points. Therefore, we are interested in the posterior distribution $p(f^* \mid y)$. Our 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 $\left(\mathbb{K}_{xx}\right)_{ij} = \kappa\left(x_i, x_j\right)$, $\left(\mathbb{K}_{xx^*}\right)_{ij} = \kappa\left(x_i, x_j^*\right)$, $\left(\mathbb{K}_{x^*x^*}\right)_{ij} = \kappa\left(x_i^*, x_j^*\right)$. It is simple to show that we have:

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

where

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

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

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

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

where $\theta = [\eta, \tau, \sigma]'$ and $\mathbb{K}_\theta$ shows explicitly the dependence of matrix $\mathbb{K}$ 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.

4. Application

We illustrate the Bayesian non-parametric regression using an interesting application on the relationship between the degree of internationalization and hotel performance. We use data on 45 international hotels companies over a 5-year period (2008-2012).

In line with the literature (Assaf et al. 2017), we measure the degree of internationalization for each company as the percentage of hotel brand properties operating in foreign countries divided by the total number of properties. For hotel performance, we use the Revenue per available room (RevPAR)[1] (Canina, Enz, & Harrison, 2005; Higgins, 2006; Ismail, Dalbor, & Mills, 2002), a global and standard performance measure adopted in the hotel industry. In our model we also controlled for firm size (total number of rooms) and the purchasing power partity (PPP) of the destination where the hotel is located.

The data for this study were mainly collected from three databases: (1) the Euromonitor, (2) the Smith Travel Research (STR), and (3) the World Bank's DataBank. The data for the dependent (revenue per available room) and independent variables (degree of internationalization) were taken from the Euromonitor database.

5. Results

Table 1 provides some descriptive statistics of the model variables. We estimate the Bayesian NPR using the Gauss software. For comparison, we also estimate a simple linear regression. Table 2 presents the marginal effects results we obtained from both models. In Figures 3 to 5 we also present the kernel densities for each variable obtained from the non-parametric regression. As mentioned, the key advantage of the non-parametric regression is that it provides much richer information than OLS. For instance, Figures 3-5 reflect the entire relationship and not just the average effect.

Table 2 clearly highlights the significant differences between the linear regression and non-parametric results. For instance, while linear regression shows a negative effect of internationalization and size on firm performance, the non-parametric results indicate the opposite. The size of the coefficients is also different between the two methods.

[1] We use the log of RevPAR due to the highly skewed nature of this variable.

Why such differences'? Because, it is clear from Figures 3 and 5 that the effect of internationalization and size is not linear. Performance and internationalization, for instance, are related through an inverted-U relationship; performance first rises and then declines (but not rapidly). The relationship between performance and size is more complicated. There is an inverted-U shape followed by a rapid increase of performance after a certain point signifying that large size is beneficial for better performance. The relationship between performance and PPP is monotonic and only slightly nonlinear.

In fact, this should help explain why the linear regression yields very different results as it is assuming a linear relationship and is only considering the average effect. The non-parametric regression, on the other hand, does not impose in advance a functional form and reflects the overall and true relationship between two variables. In fact, our non-parametric results seem also to be more in line with the literature (Lu and Beamish, 2004) which, for the most part, clearly indicate a non-linear relationship between internationalization and firm performance.

6. Concluding Remarks

The goal of this note was to promote the use of NPR in hospitality and tourism research. NPR models are not subject to misspecification error of the functional form and "provide a means of assessing a broad range of hypotheses such as whether the sign of the slope of a relationship changes or whether the relationship is additive, concave, or homothetic" (Yatchew, 1998, p.715). We believe there are many applications in the field where researchers are misspecifying the true functional forms between the hypothesized relationships. We clearly showed through this application the risk of assuming a linear relationship on a relationship that is not linear. As the literature indicates that internationalization and firm performance are related in a nonlinear way, the use of linear regression would be inappropriate and this is something we know in advance. Indeed, our application shows that the relationship is nonlinear in interesting ways and the use of linear regression would have been misleading. We presented a particularly well suited approach to nonparametric analysis in hospitality and tourism research, known as Gaussian Process Priors (GPP). GPPs can deal effectively with the problem of multivariate regression (i.e. in the presence of many covariates) and Bayesian computations are straightforward, requiring only matrix manipulations that can be performed in standard and widely available software.

One can also estimate NPR in a non-Bayesian framework. For instance, STATA can provide such estimates through the "npregress" command. Unlike Bayesian estimation, however, the non-Bayesian framework requires a larger sample size than linear regression to lead consistent estimates. For instance, on the model used in our application, around 500 observations should be at least needed to provide consistent estimates.

Table 1. Descriptive Statistics of Model Variables

Variable	Mean	St.Dev	Median
RevPAR	4.355	0.477	4.323
Internationalization	0.511	0.404	0.417
PPP	2195.56	3626.89	134.908
Size	51.569	67.880	23.600

Table 2. Marginal Effects: Non-Parametric Regression vs. Linear Regression

Variable	NPR	Linear Regression
Internationalization	0.055 (0.101)	-0.085(0.077)
PPP	**0.463** (0.078)	**5.01E-05** (9.23E-06)
Firm Size	**0.860** (0.130)	**-0.003** (0.000)

Numbers in parentheses represent the standard deviations.

Figure 3. Functional Form between Internationalization and Performance

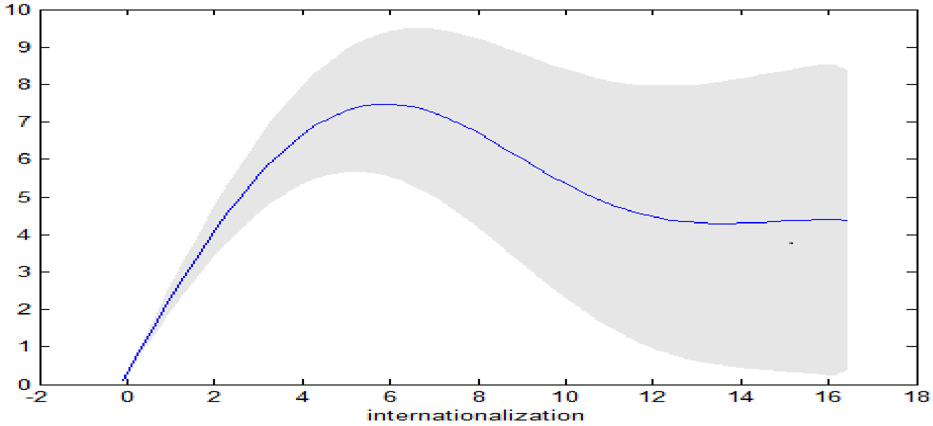


Figure 4. Functional Form between PPP and Performance

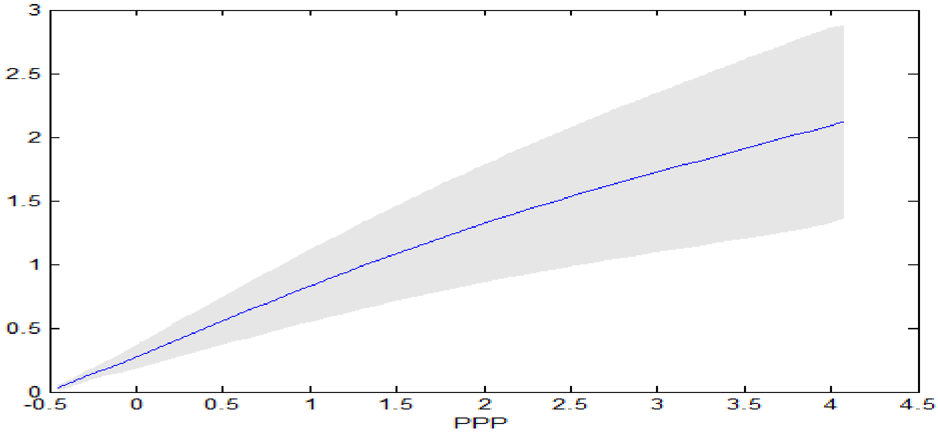
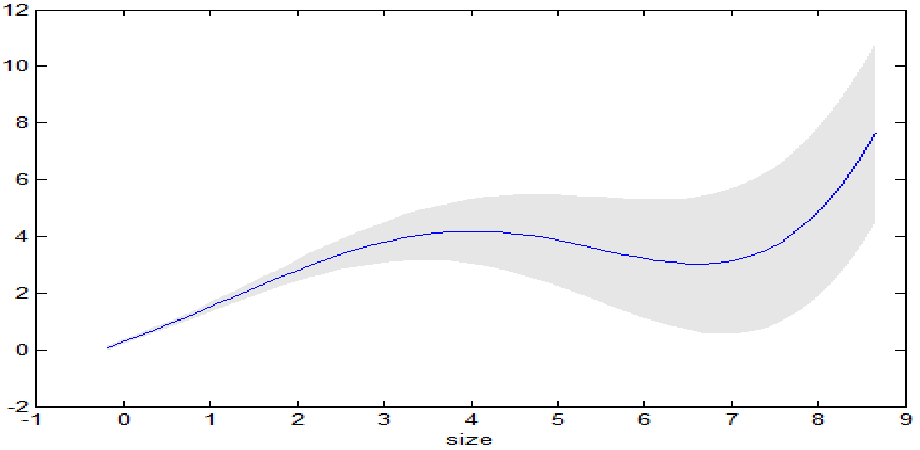


Figure 5. Functional Form between Firm Size and Performance



References

Canina, L., Enz, C. a, & Harrison, J. S. (2005). Agglomeration Effects and Strategic Orientations: Evidence from the U.S. Lodging Industry. *The Academy of Management Journal*, *48*(4), 565–581.

Higgins, S. M. (2006). Higgins, S. M. (2006). RevPAR still king, but GOPPAR on the rise. *Hotel & Motel Mangement*, *1*, 26–30.

Ismail, J. A., Dalbor, M. C., & Mills, J. E. (2002). Using RevPAR to analyze lodging-segment variability. *Cornell Hotel and Restaurant Administration Quarterly*, *43*(6), 73–80.

Lu, J. W., & Beamish, P. W. (2004). International diversification and firm performance: The S-curve hypothesis. *Academy of management journal*, *47*(4), 598-609.

Khanna, T., Palepu, K. G., & Sinha, J. (2005). Strategies That Fit Emerging Markets. *Harvard Business Review*, (June), 1–17

MacKay, D. J. (1998). Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, *168*, 133-166.

Mammen, E., Rothe, C., & Schienle, M. (2012). Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, *40*(2), 1132-1170.

Williams, C. K., & Rasmussen, C. E. (1996). Gaussian processes for regression. In *Advances in neural information processing systems* (pp. 514-520).

Vivarelli, F., & Williams, C. K. (1999). Discovering hidden features with Gaussian processes regression. In *Advances in Neural Information Processing Systems* (pp. 613-619).

Williams, C. K. (1998). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models* (pp. 599-621). Springer, Dordrecht. Lu,

Mammen, E., Rothe, C., & Schienle, M. (2012). Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, *40*(2), 1132-1170.

Müller, H. G. (2012). *Nonparametric regression analysis of longitudinal data* (Vol. 46). Springer Science & Business Media.