International Journal of Statistics and Probability; Vol. 10, No. 4; July 2021
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Use of Shape Restricted Regression Methods for Fitting Model of Per Capita GDP: A Global Economic Scenario of 2018

Sanjida Tasnim

Correspondence: Lecturer, Department of Statistics, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh

Received: April 15, 2021 Accepted: May 19, 2021 Online Published: June 1, 2021

doi:10.5539/ijsp.v10n4p52 URL: https://doi.org/10.5539/ijsp.v10n4p52

Abstract

The aim of the study is to analyze the pattern of gross domestic product (GDP) according to human development index (HDI) for 184 countries of the world. GDP per capita indicates only economic prosperity but not the overall development of the citizens of a country. This research tries to find out the beneath relationship of the financial state and human development of countries using the data of 2018. For demonstrating this analysis several parametric and non-parametric regression methods subject to shape restriction have been used. The study targets to shed light on comparative performance of shape constrained regression with cone projection, polynomial regression, LOESS, isotonic regression with pooled adjacent violators algorithm, kernel regression, smoothing spline and generalized additive model in convex situation.

Keywords: non-linear regression, polynomial regression, shape constrained regression, GAM

1. Introduction

Gross domestic product (GDP) is the market value of all final goods and services produced within a country in a specific time period. It is a globally used indicator of the economic size and growth of a country. GDP per capita is another indicator of economic prosperity per person which is calculated mainly by overall GDP by the country's total population. GDP per capita does not only determine the national wealth of the country but it is also closely related with the living standard of the citizen of that country. Economically developed countries tend to have a higher GDP per capita with smaller population and better civic facilities including mass education accessibility, medical facilities, social security etc. than the less developed countries. Human development index (HDI) is a tool annunciated yearly by United Nations Development Program (UNDP), which evaluates the actual development of the citizen of a country not only their monetary wellbeing. HDI actually assesses three aspects such as living healthy life, education attainability and an admissible standard of livelihood for capturing a crude picture of genuine development of subjects living in a particular region. Statistically, HDI is the geometric mean of normalized indices of three (health, social and economic) attributes, of which the first attribute is characterized by life expectancy, the second one is arithmetic mean of expected years of schooling for children of school entering age and average years of schooling for at least 25 years aged people and the last one is determined by a logarithmic transformation of Gross national income (GNI) per capita.

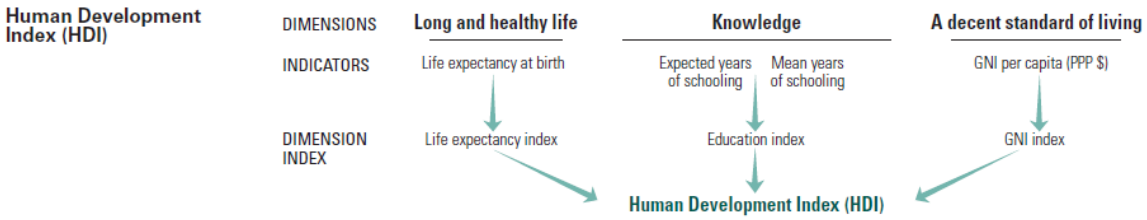


Figure 1. Graphical representation of calculating HDI (UNDP, Human Development Index)

According to the cutoff points referred in Human Development Report 2019, 32.6 percent of 184 countries show very high human development (HDI>= 0.8), 28.8 percent show high human development (0.7-0.8), 19.6 percent show medium development (0.55 -0.7) and 19.6 percent of those countries show low human development (HDI<0.55).

In this study this is tried to observe the relationship between the country's economic growth with the overall prosperity of its citizen's life. For serving this purpose, different statistical regression models (both parametric and non-parametric) have been studied. In section 2 theoretical background of the methods are reviewed along with their practical performance in section 3. Comparison among the models has been presented in section 4.

2. Materials and Methods

2.1 Data Source

To study the impact of human development index in gross domestic product per capita over the time period 2018, the country specific data values of these variables are employed in this research. The secondary data of nominal GDP per capita (in us dollar) of 184 countries have been taken from International monetary fund (IMF). Luxemburg shows highest per capita in 2018 and Burundi possesses the minimum GDP per capita, only 306.9 US dollar per year. Another key variable is human development index (HDI). This values of HDI is taken from United Nations Development Programme (UNDP)'s official database for the corresponding countries for specific year 2018. Statistical software R (version 4.0.2) is used for computational purpose.

2.2 Parametric and Nonparametric Approaches

2.2.1 Polynomial Regression

In presence of one predictor, holding both normality and linearity assumption, the relationship between the response and predictor can be explained by simple linear regression model such as,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\,, \qquad E(\varepsilon_i|X_i) = 0, \qquad i = 1,2,3,\dots,n$$

But when the linearity assumption violates i.e., the response and predictor does not have a linear relationship, then a nonlinear regression model can be imposed for better fit. The non-linear regression model can be written as,

$$Y_i = f(X_i) + \varepsilon_i\,, \qquad i = 1,2,3,\dots,n$$

Here, the function $f(X_i)$ can take any form and shape. If the function is replaced by a polynomial function of order p, then it is called polynomial regression model and the regression parameters can be obtained by least square method (Gergonne, 1815).

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \cdots + \beta_p X_i^p + \varepsilon_i$$

2.2.2 Local Regression and Loess Approach

Local regression is a different approach for fitting flexible non-linear functions involving the fit of a specific point only by its neighborhood observations. The basic concept of local regression comes from the graphical approach of Lowess (Loess). It fits weighted least square by minimizing $\sum_{i=1}^{n} w(y_i - \beta_0 - \beta_1 x_i)^2$ at $X = x_0$ using weight $w(x)$ and k nearest neighbor algorithm with an appropriate choice of smoothing span f. Greater value of f tends to result in more global and smoother fit. Usually quadratic function is used to fit locally with the tri-cube weight function w(x) as following,

$$w(x) = (1 - |d|^3)^3$$

where d is the numeric distance from fitted value and given sample value and $0 \le d \le 1$.

2.2.3 Shape Restricted Regression

This is a semi parametric model estimation criteria where a constrained regression function and a vector of parameters are estimated using a single cone projection. The shape restricted regression model is used by Liao and Meyer (2014) is as follows,

$$Y_i = f(X_i) + \alpha Z_i + \varepsilon, \qquad i = 1,2,\dots\dots,n$$

In a matrix form, it can be written as, $Y = \theta + \alpha Z + \varepsilon$, where θ is the vector of parameter and Z is the covariate matrix. For k covariate, Z is full column rank $n \times k$ matrix and for no covariate case, it contains only a column of ones. This model provides a good picture of predictor X non-parametrically using a flexible assumption of the shape of $f(X_i)$ with the presence of parametrically modelled covariate. The model accedes eight shapes of the function i.e., increasing, decreasing, convex, concave, increasing convex and increasing concave. Estimation of θ and α is executed from a single polyhedral cone rather than using back fitting algorithm introduced by Cheng (2009).

2.2.4 Isotonic Regression and PAVA

When the predictor holds the assumption of non-decreasing ($x_i \le x_{i+1}$) pattern, hence the function g(x) is a monotonic (isotonic) function. The isotonic regression is really helpful because of its flexibility to any functional form. Isotonic regression uses weighted least square subject to non-decreasing constraint. The most commonly used algorithm to get this regression is pooled adjacent violators algorithm (PAVA). If $x_1 < x_2 < x_3 \dots < x_n$, the PAVA algorithm works with starting with initial value y_1, moving to right until any pair (y_i, y_{i+1}) violates monotonicity constraint $y_i \le y_{i+1}$.

If any $y_i > y_{i+1}$, then they will be replaced by $\bar{y}_i = \frac{y_i + y_{i+1}}{2}$. Then $y_{i-1} > \bar{y}_i$ condition will be checked. If the condition is violated, then the process goes back to previous step and performs for the rest values. Otherwise, the adjacent three values are pooled and replace them with $\bar{y}_i = \frac{y_{i-1}+y_i+y_{i+1}}{3}$. This step is repeated until the monotonicity constraint prevails (Barlow et al., 1972).

2.2.5 Kernel Regression for Single Parameter

Nonparametric regression can be modelled as,

$$Y_i = \ m(X_i) + \ \sigma(X_i).\varepsilon_i \ , i = 1,2, \dots, n \, .$$

While performing monotonic regression, this function $m(X_i)$ is estimated through any unconstrained non-parametric method such as Nadaraya-Watson estimate or local linear estimate. Then inverse of the monotonic function ($\widehat{m}_I$) can be obtained by the following formula,

$$\widehat{m}_I^{-1} = \frac{1}{N.h_d} \sum_{i=1}^{N} \int_{-\infty}^{t} K_d \left(\frac{\frac{\widehat{m}.i}{N} - u}{h_d} \right) du$$

Here, K_d is a function which measures the probability to find specific distant neighbors from certain values. Kernel regression is used here to estimate density. The kernel K_d uses the bandwidth h_d. Many options are available to choose an admissible bandwidth of a kernel density estimator such as Sheather and Jones (1991) approach, unbiased and biased cross validation approach etc.

2.2.6 Regression Spline With Basis Function

Regression spline is a technique which fits lower order polynomial regression models into k distinct regions of predictor. It is a very popular regression technique because of its flexibility than higher order polynomials and step functions. It is often called piecewise cubic polynomial regression as it fits third order polynomial model of the form

$$Y_i = \beta_0 + \beta_1 X_i \ + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$

for every knots. For k knots, the number of associated spline coefficients is (k+p+1) as there are k+1 polynomials with p degree of polynomials having p*k constraints. Let the basis functions can be represented with $b_1, b_2, \dots, b_{k+3}$ and the model becomes

$$Y_i = \beta_0 + \beta_1 b_1(X_i) + \beta_2 b_2(X_i) + \cdots + \beta_{k+3} b_{k+3}(X_i) + \varepsilon_i \, .$$

It is needed to choose the appropriate basis function for a specific application among different spline basis such as, truncated power basis, B-spline, cardinal spline, penalized spline etc. (Knott, 2000).

2.2.7 Smoothing Spline

Smoothing spline aims to find a function which minimizes the loss function incorporating a nonnegative tuning parameter λ that actually accounts for the roughness of smoothing spline ranging 0 to ∞. The loss function used here is

$$\sum_{i=1}^{n} \{Y_i - g(X_i)\}^2 + \lambda \int g''(t)^2 dt \quad , \quad \lambda \geq 0$$

where $g''(t)$ measures the amount by which the slope of a function is changing at t (Green and Silverman, 1994). Unlike cubic splines, the main problem of smoothing spline is to fix an appropriate value of λ to balance between bias and variance of smoothing spline rather than knots. LOOCV (Leave One Out Cross Validation) method can be employed here to find a possible solution using the following formula,

$$RSS_{cv}(\lambda) = \sum_{i=1}^{n} \{Y_i - \widehat{g}_{\lambda}^{(-i)}(X_i)\}^2$$

2.2.8 Generalized Additive Model (GAM)

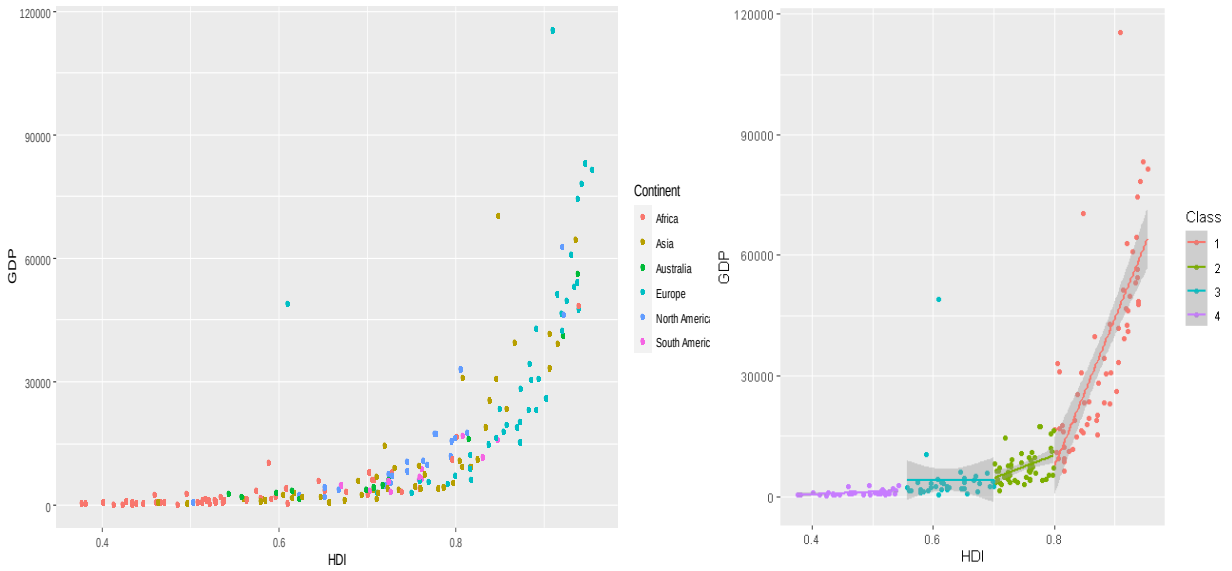
Generalized Additive Model (GAM) is a technique to extend a linear model by imposing linear or non-linear functions for each variable for smoothing purpose prevailing additivity. So if there are multiple predictors $X_1, X_2, \dots, X_p$ and the response is Y, then the GAM model can be written as,

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i \ , \qquad i = 1,2,3 \ldots, n$$

where ε_i is iid random variable with mean 0 and variance σ^2 (Hastie & Tibshirani, 1990). As GAM provides a linear equation, the impact of every single predictor variable in response can be identified individually. GAM with a natural spline is a simpler way to fit if an appropriate set of basis function can be chosen. On the contrary, GAM with smoothing spline includes a tuning parameter to deal with the roughness of the fitted curve which makes the computation little bit more complex. As least square estimates cannot be obtained here, backfitting method can obtain an admissible result.

3. Practical Performance

We use a dataset of GDP and HDI of 184 countries to demonstrate the methodology for several types of constrained regression models. In this study, it is aimed to fit a nonlinear regression model of GDP with shape constraint, assuming GDP as response variable and HDI as predictor. A scatterplot (fig 1) is drawn to present the worldwide scenario of GDP along with HDI according to continents.



(a) (b)

Figure 2. (a) Scatterplot of GDP with respect to HDI of 184 countries according to continents, (b) Illustration of simple linear regression model individually for four classes of human development index

Figure 2(a) clearly shows an upward convex relationship between GDP and HDI. As log transformation does not result in a linear pattern, it is wise to fit a nonlinear regression model with a shape restriction of convexity rather than a linear one. Countries are classified into four classes such as very highly developed (0.8 and above), highly developed (0.7 to 0.8), medium developed (0.55 to 0.7) and poorly developed (0.55 or less) according to their human development index which was introduced in Human Development Report 2014. Demonstration of linear regression models for each of the classes (figure 2(b)) result in poor fitting ($R^2 = 0.58, 0.18, 0.000008$ and 0.18) for each class and also for overall data ($R^2 = 0.2$). Such situation can be handled in various parametric and nonparametric way where linearity assumption is disregarded.

3.1 Polynomial Regression

Polynomial regression is parametric way to deal with intrinsically nonlinear data where least square estimation is used to predict parameters. As larger degree overfits the data, it is crucial to find the optimal choice of order of polynomial function. Polynomials of several orders have been fitted for this data where the dependent variable is GDP and the predictor is HDI. F test statistic $(F = 7.92,\ p\ value = 0.0054)$ suggests polynomial of 4^{th} order fits better than any other order. So the non-linear effect of the GDP data can be modelled through a polynomial of degree 4 given by,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \varepsilon_i, \qquad i = 1,2, \ldots, n$$

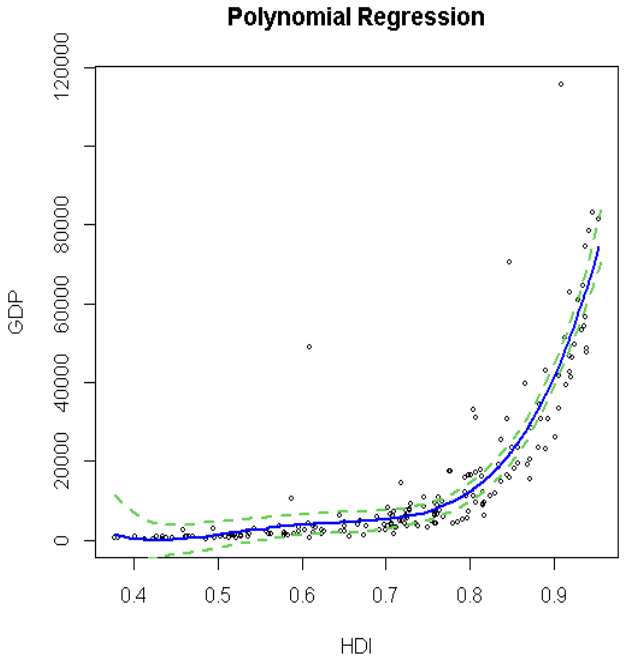


Figure 3. Fitting of polynomial regression of order 4 with 95% confidence band

This model leads to Multiple R^2= 0.8033 (adjusted R^2=0.7987) and shows the significance of $\beta_1, \beta_2, \beta_3$and β_4 coefficients. Though polynomial regression model is often inflexible to interpret, it has several enviable properties to figure out curvilinear relationship of the response and predictor parametrically.

3.2 Local Regression

For the GDP data, Loess is applied for two tuning parameter f. From figure 4 it is observed that, for f=0.8 the fitted line does not capture the total intrinsic pattern as higher value of f lead to use more of the data as training observations.

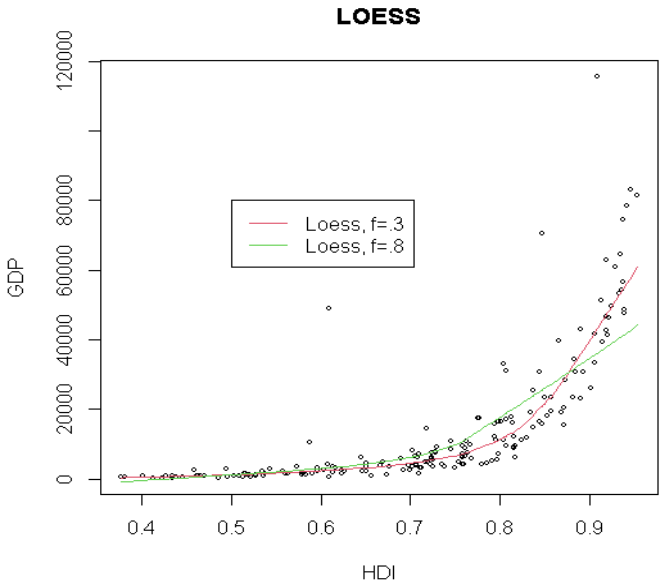


Figure 4. Local piecewise polynomial regression illustrated for two smoothing parameters (red line represents f=0.3 and green line represents f=0.8)

For this study, the admissible choice of smoothing span is 0.6 as the residual standard error is lower for this bandwidth.

This is a non-parametric approach employing moving polynomial although parametric weighted least square is used to fit the estimate in every point of x. Figure 5 shows the Loess fit of GDP data for f= 0.6 with 95% confidence band.

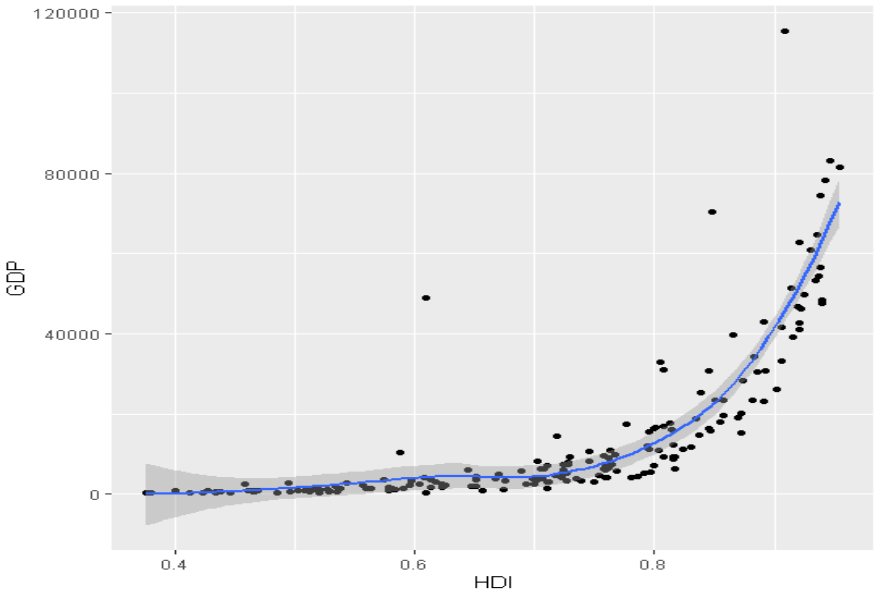


Figure 5. Fitting of local regression for f=0.6 with 95% confidence band

3.3 Shape Restricted Regression

Shape restricted regression is a semi parametric technique where the response is modelled non-parametrically with predictor with a clear assumption of its shape as well as modelled parametrically with one or more covariates. As our GDP data exhibits an increasing convexity, using this assumption we get the following (figure 6) estimates.

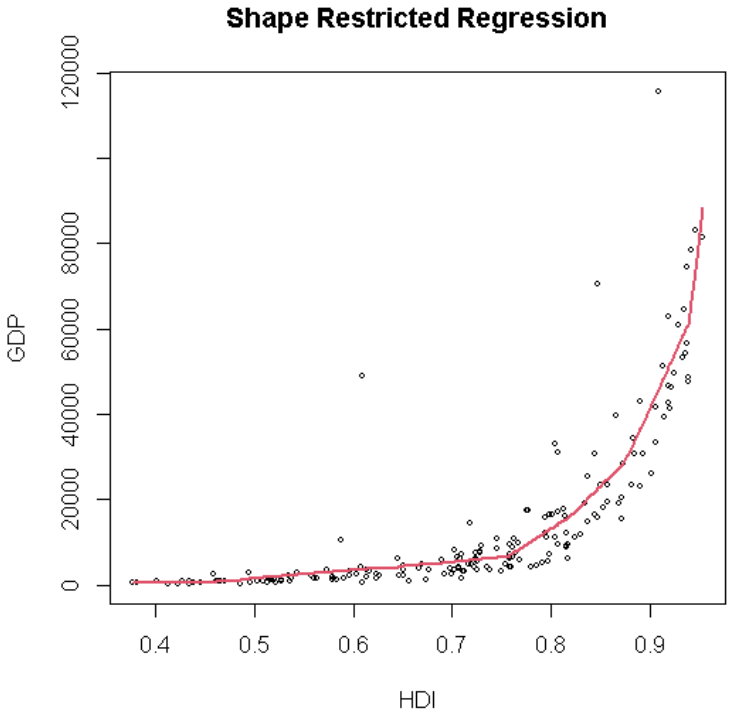


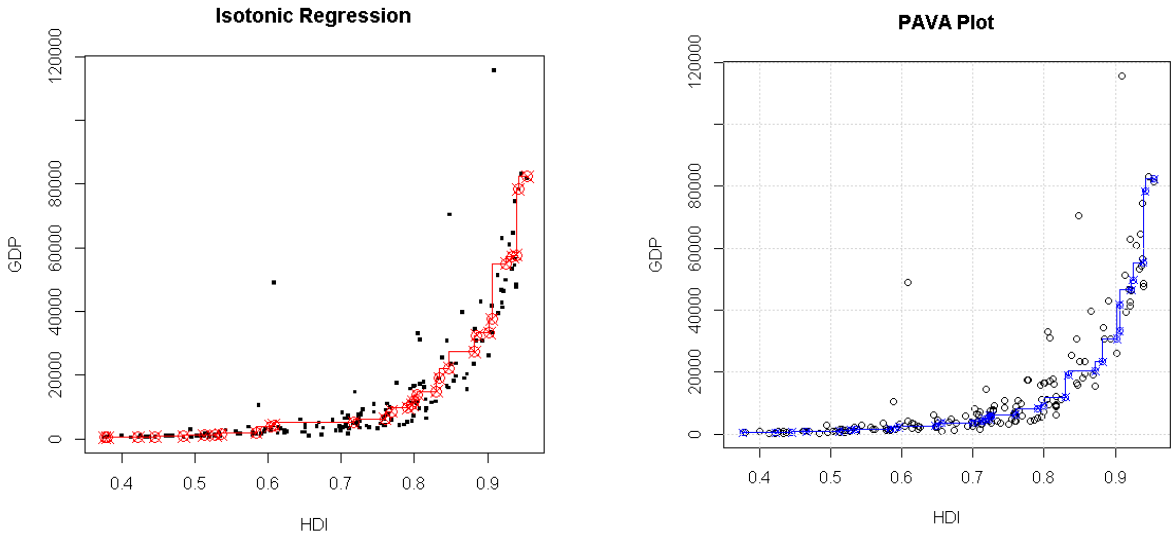
Figure 6. Fitting of convex regression with presence of no covariate (x=HDI, y=GDP)

As the regression is demonstrated for only one predictor with no categorical or scaled covariate, the parametric part of

this regression is absent. The hypothesis testing of $H_0: Full\ constrained\ model$ vs $H_1: Linear\ model$ is an exact one sided test which is illustrated with $E_{01} = \frac{SSE_0 - SSE_1}{SSE_0}$. We get the value sum of squared residuals for the linear part, $SSE_0 = 7.177$ and sum of squared residuals for the shape constrained model SSE_1=1.366. P value (<.001) of this test justifies the use of shape constraint of the model.

3.4 Isotonic Regression and PAVA

Isotonic regression is applied to fit monotonically increasing model which has actually piecewise linear form. Pooled adjacent violators algorithm is also used as it is a general approach to deal with convex function and ties to solve isotonic problem. Estimation of isotonic regression with active set algorithm is figured out in 7(a) whereas 7(b) shows the popular PAVA algorithm employing weighted median solver.



(a) (b)

Figure 7. (a) Isotonic regression with active set algorithm and (b) PAVA optimization using weighted median solver

3.5 Kernel Regression

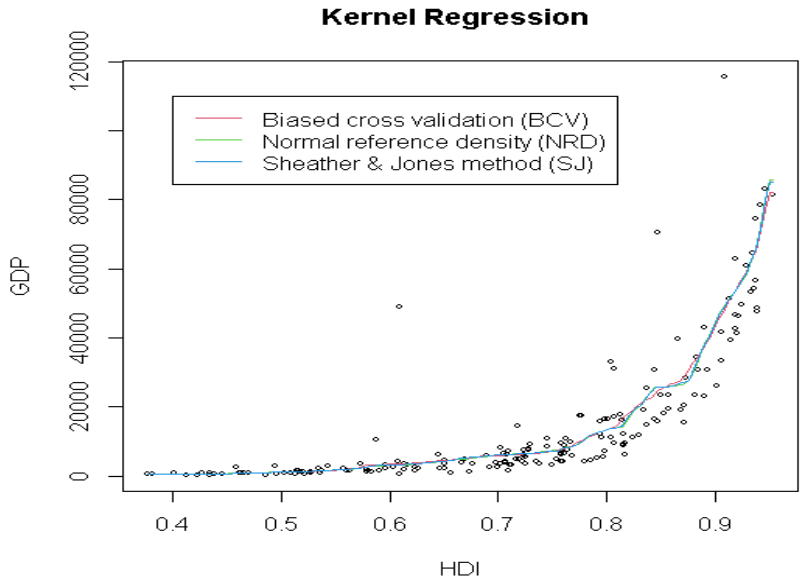


Figure 8. Kernel smoothing with different bandwidth selectors

For modeling the dependent variable GDP with one predictor HDI Kernel regression with three different optimal

bandwidth selection techniques have been employed (figure 8). BCV, SJ and NRD method has given almost similar result as they all use pairwise binned distance.

3.6 B-Spline and Smoothing Spline

Countries are categorized into four distinct groups according to their HDI for classification. So imposing these quartile cutoff values (0.55, 0.7, 0.8) as knots we get the following model with six predictors and intercept.

$$Y_i = \beta_0 + \beta_1 b_1(X_i) + \beta_2 b_2(X_i) + \cdots + \beta_6 b_6(X_i)$$

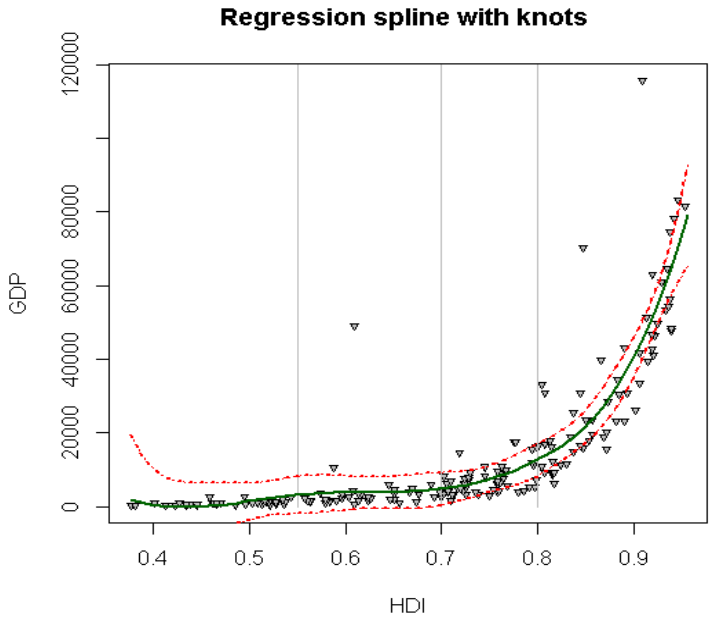


Figure 9. Cubic regression spline using basis function with three knots at 0.55, 0.7, 0.8

Spline fitting to GDP data with three knots is displayed graphically in figure 9. The spline function is colored as green and 95% confidence bands are colored as red. It is evident that confidence bands in the boundary region appears narrower. This model results into adjusted R^2 =0.7962 and multiple R^2 =0.8029, which is a clear indication of better fit.

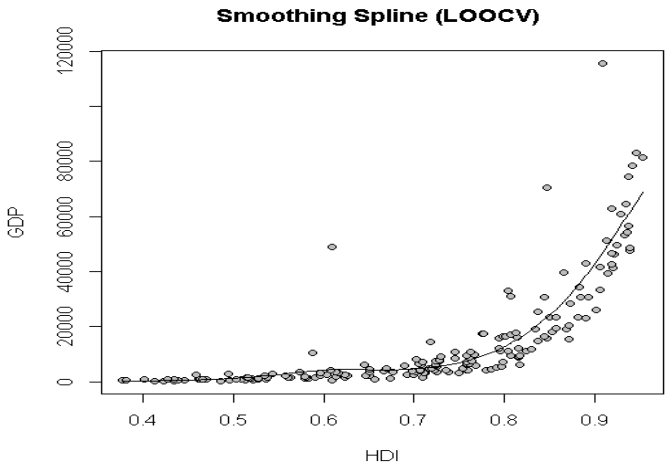


Figure 10. Fitting of Smoothing spline by leave-one-out cross-validation (lOOCV), resulted in 6.56 df

3.7 Generalized Additive Model (GAM)

Generalized additive model (GAM**)** is a linear model in which the linear predictor depends linearly on unknown smooth

functions of some predictor variables. It is called additive model because smooth functions for separate variables are added together. For GDP data, three models are employed. The first model uses smoothing spline of predictor with 6 degrees of freedom. Model 2 uses bs=cr, which implies cubic spline basis defined by a modest sized set of knots spread evenly through the covariate values. They are penalized by the conventional integrated square second derivative cubic spline penalty. Model 3 uses REML of maximum likelihood which may be used for smoothness selection, by viewing the smooth components as random effects.

Formula:

Model 1: GDP ~ s (HDI, 6)

Model 2: GDP ~ s (HDI, bs = "cr")

Model 3: GDP ~ s (HDI), method="REML"

Table 1. Comparative performance of three GAM models

	AIC	Adjusted R^2	Deviance
Model 1	4046.095	0.4879	49.07%
Model 2	3882.856	0.794	80.1%
Model 3	3700	0.795	80.2%

A comparison between these three GAM models has been illustrated in table 1. Model 2 and 3 gives almost similar result with an admissible value of adjusted R^2. Model 2 is displayed in figure 11 with 95% confidence band. Model 3 shows better estimate as it resulted in lower value of Akaike information criterion.

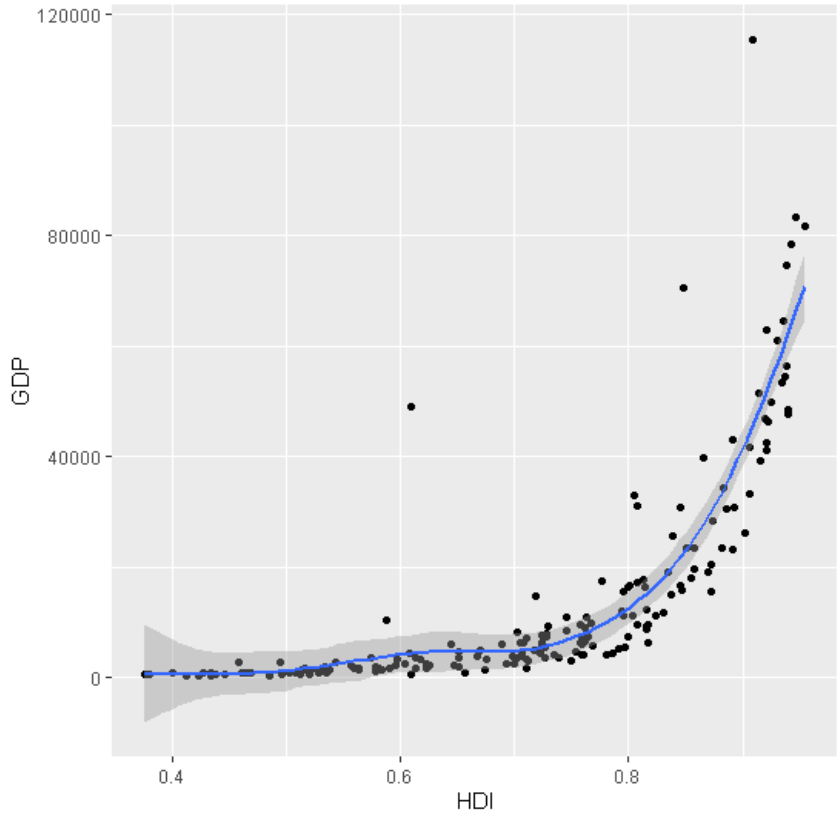


Figure 11. Fitting of generalized additive model (GAM) for model 2

4. Discussion

A comparative performance of the above fitted non-linear models has been presented in table 2. The performance of the models has been evaluated by some quality measures such as root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). It is apparent from the table that among several parametric and non-parametric approaches, the regression model restricted to convex shape and general additive model (GAM) using maximum likelihood estimation for smoothing parameter gives better estimation as their error measurements are lower

than others.

Table 2. Comparative performance of fitted models

Model	RMSE	MAE	MAPE
Polynomial regression (4)	5083.5	3351.1	19.48%
Local regression (f=0.6)	5168.2	3418.7	21.07%
Convex regression	4053.7	2481.6	11.41%
Spline regression (BS)	5353.3	3809.1	24.80%
Smoothing Spline	5298.1	3944.7	25.32%
GAM3	4299.3	2911.9	18.75%

5. Conclusion

The study is intended to fit a model between two key variables human development index (HDI) and gross domestic product (GDP) of worldwide countries. It is evident from the data that the countries which has low human development index implying low life expectancy, less year of schooling and lower income, has also less GDP per capita. Most of the African countries and few Asian countries belong to this class. On the contrary, most of the European and North American countries show very high human development along with a good prospect in GDP. As the data exhibits convexity pattern, several non-linear methods have been conducted for modelling purpose. Among parametric techniques, polynomial with degree 4 gives a better fit. Among non-parametric techniques, general additive model (GAM) performed well for smoothing. For semi-parametric situation, shape constrained regression cone projection works well under increasing convexity with the presence of scaled or categorical covariates.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. New York, NY: John Wiley & Sons.

Cheng, G. (2009). Semiparametric Additive Isotonic Regression. *Journal of Statistical Planning and Inference*, *139*(6), 1980-1991. https://doi.org/10.1016/j.jspi.2008.09.009

Green, P. J., Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach.* London, England: Chapman and Hall. https://doi.org/10.1007/978-1-4899-4473-3

International Monetary Fund. (2020). *World Economic Outlook Database* [Dataset]. Retrieved from https://www.imf.org/en/Publications/WEO/weo-database/2020/April

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning with Application in R*. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Knott, G. D. (2000). *Interpolating Cubic Splines.* New York, NY: Springer. https://doi.org/10.1007/978-1-4612-1320-8

Meyer, M. C. (2013). Semi-parametric Additive Constrained Regression. *Journal of Nonparametric Statistics*, *25*(3), 715-730. https://doi.org/10.1080/10485252.2013.797577

Meyer, M. C., & Liao, X. (2014). coneproj: An R Package for the Primal or Dual Cone Projections with Routines for Constrained Regression. *Journal of Statistical Software*, *61*(12), 1-22. https://doi.org/10.18637/jss.v061.i12

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A Review of Spline Function Procedures in R. *BMC Medical Research Methodology, 19*(46). https://doi.org/10.1186/s12874-019-0666-3

Sheather, S. J., & Jones, M. C. (1991). A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation. *Journal of Royal Statistical Society B*, *53*(3), 683-690. https://doi.org/10.1111/j.2517-6161.1991.tb01857.x

United Nations Development Programme. (2019). Human Development Reports Database [Dataset]. Retrieved from http://hdr.undp.org/en/indicators/137506

United Nations Development Programme. (n.d.). *Human Development Index (HDI)* [Technical Note]. Retrieved from http://hdr.undp.org/en/content/human-development-index-hdi

Copyrights