Aix Marseille School of Economics

Master 2: Econometrics, Big Data and statistics

Advanced Econometrics: Homework

Author:
Coraline BEST

Professor: Emmanuel FLACHAIRE

February 12, 2024

Table of contents

1 Presentation of the application

For this assignment, I decided to provide an application based on the world happiness score. In the first section, I will conduct an analysis using standard parametric econometric models, primarily simple linear regressions. Subsequently, I will focus on nonparametric econometrics, with a specific emphasis on Generalized Additive Models. Finally, the assignment concludes with a definition of a modified parametric model and a comparative analysis between these two distinct approaches.

1.1 What is the study about?

The World Happiness Report (2019), serves as a pivotal survey outlining the global landscape of happiness. Acknowledged worldwide, the report has become a significant reference for governments, organizations, and civil society, leveraging happiness metrics to shape their policy-making strategies. Esteemed experts from various disciplines including economics, psychology, national statistics, health, and public policy shed light on the effective use of well-being metrics for evaluating a nation's progress. This report offers an insightful review of the contemporary state of happiness worldwide, delving into the science of happiness and unveiling the intricacies behind personal and national variations in well-being.

1.2 Description of the data

The happiness scores and rankings are derived from data collected through the Gallup World Poll, which assesses individual's satisfaction. Participants are asked to envision a ladder, where 0 represents the worst possible life and 10 represents the best possible life, and then rate their own current lives on this scale. The collected scores are representative of national populations from 2019 and regroups 9 variables and 156 observations. The "happiness score" (or subjective well-being) reflects the national average response to the question: "On which step of the ladder would you say you personally feel you stand at this time?"
Here is a brief description of the other variables:

- **GDP per capita:** a measure of a country's economic output that considers its population size.
- **Healthy life expectancy:** the average number of years a newborn is expected to live in good health, without debilitating illnesses or injuries.
- **Social support:** the national average of binary responses (0 or 1) to the question: "Do you have relatives or friends you can count on in times of trouble?"
- **Freedom to make life choices:** the national average of responses to the question: "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
- **Generosity:** the residual of regressing the national average of the question: "Have you donated money to a charity in the past month?" on GDP per capita.
- **Corruption Perception:** this measure represents the national average of survey responses to the questions: "is corruption widespread throughout the government? and "Is corruption widespread within businesses?"

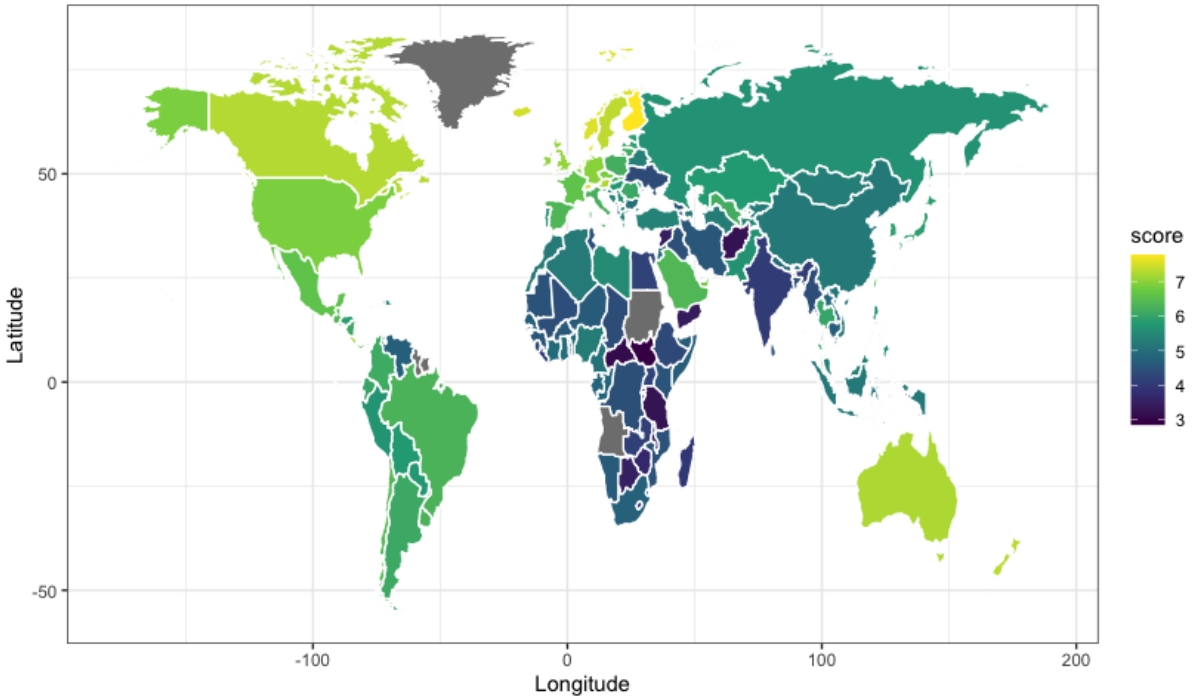
Descriptive statistics

This subpart will display some descriptive statistics about the data in order to have an overview. Table 1 provides some statistics like the min, max, first and third quartile and the mean of the variables.

Table 1: Descriptive Statistics

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Score	2.85	4.54	5.38	5.41	6.18	7.77
GDP per capita	0.00	0.60	0.96	0.91	1.23	1.68
Social support	0.00	1.06	1.27	1.21	1.45	1.62
Healthy life expectancy	0.00	0.55	0.79	0.73	0.88	1.14
Freedom to make life choices	0.00	0.31	0.42	0.39	0.51	0.63
Generosity	0.00	0.11	0.18	0.18	0.25	0.57
Perceptions of corruption	0.00	0.05	0.09	0.11	0.14	0.45

Figure 1: Geographical distribution of Happiness Scores



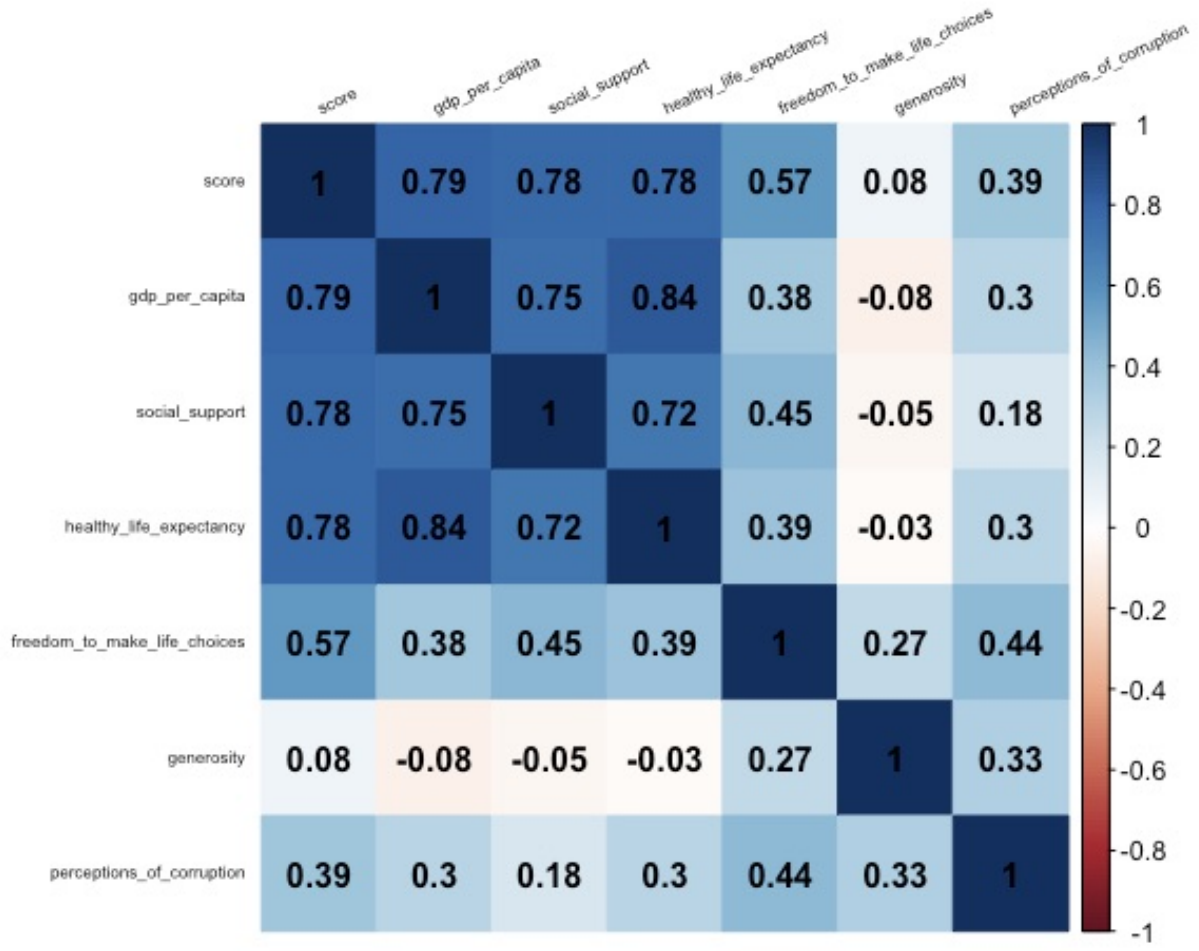
Note: Some countries appear in grey as they are not part of the dataset.

The map above illustrates happiness scores across various countries, highlighting a significant variation in happiness levels among continents. For instance, most African countries exhibit notably low happiness scores, represented by darker colors. Conversely, higher scores, depicted with lighter shades, are more prevalent in regions such as America and Europe.

1.3 Regression model

In this subsection, I will outline the execution of the linear regression. However, before delving into it, it is crucial to identify the relevant variables. To accomplish this, I can use the correlation matrix, which will present the correlation coefficients among the variables. Each cell in the matrix depicts the correlation between two specific variables.

Figure 2: Correlation matrix



The target variable is the happiness score, a continuous variable, for which I use a linear regression model incorporating the following key variables: GDP per capita (GDP), social support, healthy life expectancy (HealthyLife), and freedom to make life choices (Freedom). These variables were selected due to their significant correlations with the outcome variable 'score'. Notably, 'gdp per capita' demonstrates a strong 79% correlation with the happiness score. Therefore, the aim is to estimate the following equation:

$$\text{Happiness Score}_i = \beta_0 + \beta_1 \text{GDP}_i + \beta_2 \text{SocialSupport}_i + \beta_3 \text{HealtyLife}_i + \beta_4 \text{Freedom}_i + \epsilon_i \quad (1)$$

1.4 Estimation results and interpretation

Table 2: OLS results

	Dependent variable:	
	Happiness score	Happiness score
	OLS	*OLS Robust*
	(1)	(2)
GDP per capita	0.811***	0.811***
	(0.216)	(0.214)
Social support	1.017***	1.017***
	(0.235)	(0.257)
Healthy life expectancy	1.141***	1.141***
	(0.337)	(0.364)
Freedom to make life choices	1.846***	1.846***
	(0.340)	(0.333)
Constant	1.892***	1.892***
	(0.199)	(0.220)
Observations	156	
R^2	0.771	
Adjusted R^2	0.765	
Residual Std. Error	0.540 (df = 151)	
F Statistic	127.048*** (df = 4; 151)	
Note:		*p<0.1; **p<0.05; ***p<0.01

To ensure unbiased OLS estimators, it is imperative to verify certain assumptions, notably the Gauss-Markov assumptions. One of these involves assessing heteroscedasticity, represented by $Var(\epsilon_i) = \sigma^2 f(x_i)$. I conducted a Breusch-Pagan test for heteroscedasticity, yielding a p-value of 0.006, which offers statistically significant evidence to reject the null hypothesis of homoscedasticity. Consequently, I have reason to suspect the presence of heteroscedasticity within the residuals. Although OLS remains an unbiased linear estimator, it no longer represents the best estimator, as it is not BLUE. In the context of heteroscedasticity, it is feasible to find an estimator that surpasses OLS in efficiency. Accounting for heteroscedasticity, I inferred using the OLS estimator with the White "sandwich" correction applied to the variance-covariance matrix. The results are displayed in the second column, titled "coefficient test". The first column, "OLS" represents the linear regression without accommodating robust standard errors. Initially, it's notable that all the estimated coefficients exhibit statistical significance at the 1% level. To illustrate, when the GDP per capita increases by one unit, it leads to an estimated increase of 0.811 units in the happiness score.

2 Nonparametric approach

Having estimated the standard linear regression model, this section will present a nonparametric approach to identify socioeconomic key determinants for happiness. To this end, I estimate and evaluate Generalized Additive Models (GAM) based on the spline method. The first GAM will capture the effect of each variable in a fully flexible manner. In a second step, I will linearize the model where applicable. The last part of this section will examine plausible interactions between variables.

2.1 Generalized additive model

Model
In a first step, I estimate and evaluate a GAM model assuming a nonparametric functional form for all regressors that I considered in the linear regression. This way, I allow the model to additively capture nonlinear effects for each explanatory variable separately. More concretely, I estimate the following equation (GAM1):

$$\text{Happiness Score}_i = m_1(\text{GDP}_i) + m_2(\text{SocialSupport}_i) + m_3(\text{HealtyLife}_i) + m_4(\text{Freedom}_i) + \epsilon_i \quad (2)$$

where m_i denotes the i^{th} nonlinear model capturing the effect of each regressor on the happiness score, respectively.

Results
Different from a standard OLS regression, the p-values of a GAM model correspond to test:

$$H_0\text{: linear vs. } H_1\text{: nonlinear.}$$

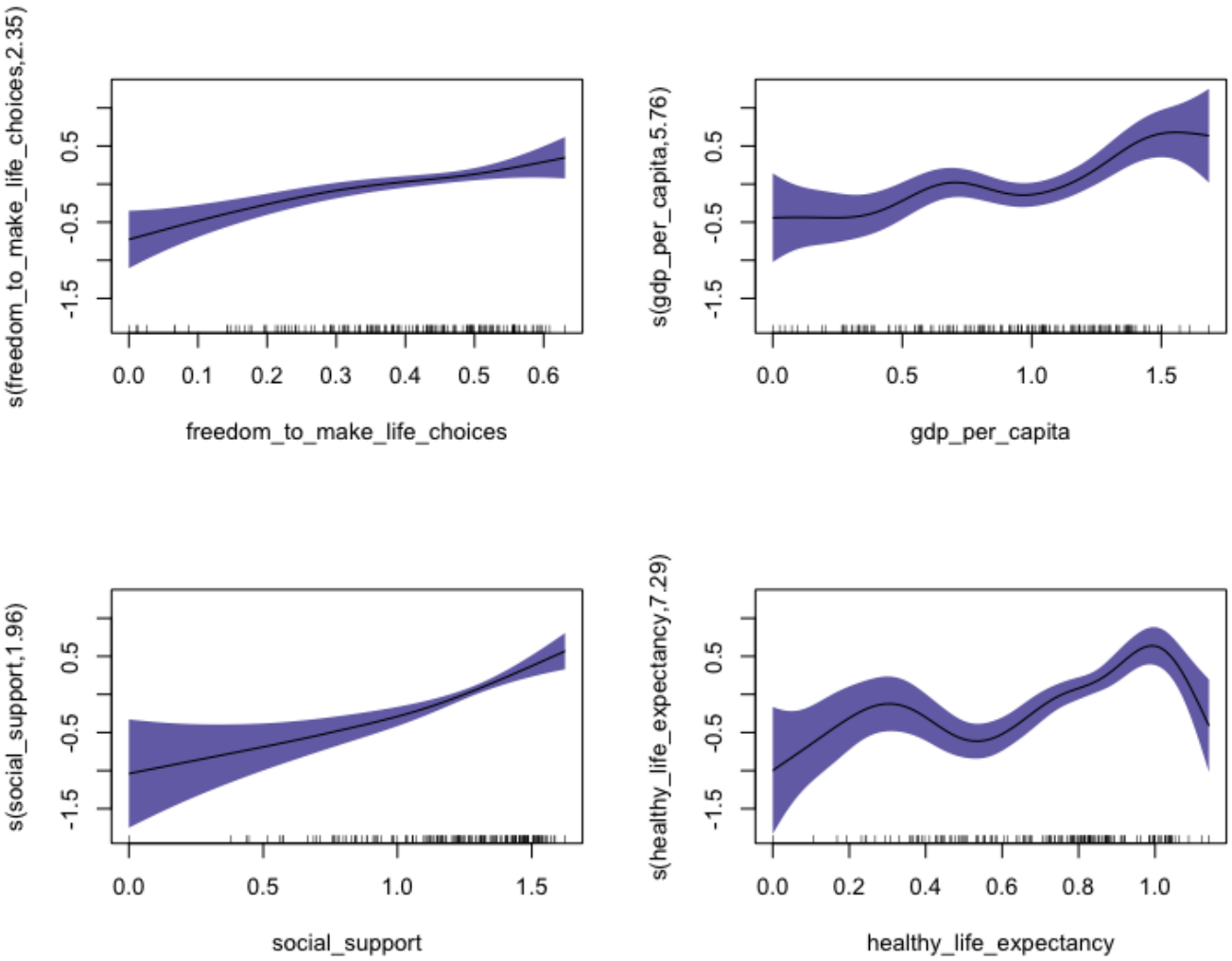
As shown in Table 3, the smooth terms of all variables show $p < 0.001$, so I can reject the null hypothesis and assume nonlinear effects. However the low edf values (effective degrees of freedom) for social support (1.956) and freedom to make life choices (2.35) imply low complexity (functions of degree 2). Complementing these results with the visualization of the smooth functions (see Figure 3), reveals that in both cases I am dealing with quasilinear functions. I can see that these heavily stretched parabolas show weak nonlinearities, and thus suggest exclusively positive effects on a country's happiness score. Put differently, linear estimates are very likely to capture the positive effects of the mentioned variables on the happiness score sufficiently fine.

In contrast, the first GAM estimates more complex effects for GDP per capita and healthy life expectancy, which is confirmed by the corresponding edf values of 5.76 and 7.285. Despite a positive effect of GDP per capita on happiness, I can observe a plateau, i.e. significantly weaker effects, at the lower and upper income range. Thus, the positive effect on happiness is mainly driven by countries that surpass a certain threshold of GDP per capita. After having reached this threshold, the effect is diminishing and, interestingly, becomes even ambiguous at the very upper income range (which is also due to sparse data). Put short, more money does not necessarily make you happier. Similarly, healthy life expectancy shows ambiguous effects on happiness. For the range between 0 and 0.4, the confidence bands are too large to conclude any clear effect. From 0.5 on wards an increase in healthy life expectancy has a positive effect on happiness even though this effect diminishes and becomes ambiguous at the end of the age range.

As rule of thumb, I confirm the significance of a smooth term by assessing whether a horizontal line could be drawn through the 95 percent confidence interval. As Figure 3 shows, this is not possible for any of the smooth functions, and thus the estimates are perceived to be highly significant.

Against this background, in the next sections I aim to simplify the GAM model by linearizing the effects for the mentioned variables. Here, the adjusted R^2 measurement indicates that the model explains 84.1 percent of the variance of the data. I will return to this value as benchmark for the explanatory power when optimizing the GAM.

Figure 3: Plot of the nonparametric components



2.2 Modified Generalized additive model

I now aim to optimize the efficiency of the GAM by predefining a linear relation for the variables of freedom to make life choices and social support. This way, I decrease the model's complexity and thus its explanatory power but in return increase the interpretability. With regards to the results, the later effect is expected to outweight the former, i.e., the gain in interpretability should largely compensate the introduction of bias.

Model
More precisely, I estimate the following equation (GAM2):

$$\text{Happiness Score}_i = \beta_1 \text{Freedom}_i + \beta_2 \text{SocialSupport}_i + m_1(\text{HealthyLife}_i) + m_2(\text{GDP}_i) + +\epsilon_i \quad (3)$$

with m_i denoting the nonlinear and β_i capturing the linear effects of the respective variables.

Results
Compared to the previous results, the modified GAM provides simliar findings. As for the linear regression model, social support and freedom to make life choices have a highly significant and positive effect on happiness (p-value inferior to 0.001). As assumed, partially restricting the model to linear effects did slightly reduce its explanatory power but gained (back) the interpretability of the effects for these variables. This can be observed by a similar adjusted R^2, measuring that 83.7 percent of variance of the data is explained by the model (reduction of 0.4 percentage points).

Table 3: GAM estimation results

	GAM1			GAM2			GAM3		
	coefficient	edf	p-value	coefficient	edf	p-value	coefficient	edf	p-value
Intercept	5.407	-	***	3.5074	-	***	4.2500	-	***
GDP	-	-	-	-	-	-	-	-	-
Social Support	-	-	-	1.0943	-	***	0.9572	-	***
Healthy Life	-	-	-	-	-	-	-	-	-
Freedom	-	-	-	1.4694	-	***	-	-	-
m(GDP)	-	5.760	**	-	5.806	**	-	-	-
m(Social Support)	-	1.956	***	-	-	-	-	-	-
m(HealthyLife)	-	7.285	***	-	7.332	***	-	7.593	***
m(Freedom)	-	2.350	***	-	-	-	-	-	-
m(GDP, Freedom)	-	-	-	-	-	-	-	18.060	***
R^2	0.841			0.837			0.818		

Note: $***p < 0.001$; $**p < 0.01$; $*p < 0.05$

2.3 Variable interactions

One powerful feature of the GAM model is that it allows us to study and plot, in the most flexible manner, the interactions between two variables. The GAM can model the combined effect of a set of multiple variables on the outcome variable. In my case, I choose two variables, as this allows for graphical representation and clearer interpretation.

Based on the results from section 2.1 that after a certain threshold, more money does not necessarily bring more happiness, I want to investigate potential interactions of GDP per capita with other variables. Following concepts of social science, e.g. Maslow's hierachy of needs, I propose that freedom of making life choices will have a stronger effect on happiness with an increase in financial means. This theory proposes that the need for self-fulfilment comes for individuals with secure financial and material situation. This seems plausible, since the income might be perceived as essential precondition for further wellbeing. I would therefore observe a larger impact of freedom to make life choices

on happiness for countries with a higher GDP per capita.

Model

To check whether Maslow's pyramid of needs can be observed in out data, I specify a new GAM model. The model contains an interaction term including GDP per capita and freedom to make life choices, a smooth term for healthy life expectancy - since I have seen its non linear effect, and a linear term for social support which was found to have a linear effect on the happiness score. The model takes the following form (GAM3):

$$\text{Happiness Score}_i = \beta_1 \text{SocialSupport}_i + m_1(\text{HealthyLife}_i) + m_2(\text{GDP}_i, \text{Freedom}_i) + \epsilon_i \quad (4)$$

Results

Figure 4: Plot of the interaction effect on the happiness score

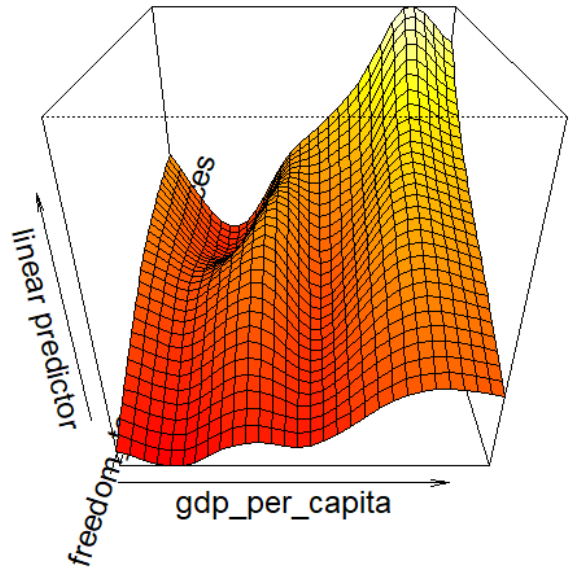


Figure 4 illustrates the effect of the interaction of the two variables on the happiness score. Here, the maximum happiness is reached when both variables, freedom of life choices and GDP per capita are at their highest levels. A high GDP per capita with low freedom to make life choices leads to a relatively low happiness score. Also, I observe the stated hypothesis that moral fulfilment is more important at higher levels of gdp per capita, as the combination of high freedom to make life choices and low GDP per capita leads to a considerably lower level of happiness. Further results of this GAM3 are illustrated in Table 3.

3 Comparison between the two approaches

In the first part I have seen the parametric linear model and in the second part I have implemented a nonparametric and flexible model with a GAM model. The parametric linear model hold a few advantages over the GAM model. Because it is parametric, for a given sample size the variance of the model will be lower than for the nonparametric model. Also because I assume a linear functional form, inference and interpretation of the coefficient is much easier to perform to understand the insights from the model. The GAM model loses some of these aspects but, because it assumes no prior functional form on the individual covariates, it is much more flexible and is less prone to be miss specified compared to the linear model.

One approach here would be to look at the results of the GAM model and try to find a parametric specification which could try to capture some of the non linearities highlighted by the GAM model. We use two types of models, a piecewise linear model or segmented linear model and a polynomial model. The segmented linear model is given by the following equation :

$$\begin{aligned}\text{Happiness Score}_i &= \beta_1 \text{SocialSupport}_i + \beta_2 \text{Freedom}_i + \beta_3 \text{GDP}_i + \beta_4 (\text{GDP}_i - 0.7)_+ \\ &+ \beta_5 (\text{GDP}_i - 1)_+ + \beta_6 \text{HealthyLife}_i + \beta_7 (\text{HealthyLife}_i - 0.32)_+ \\ &+ \beta_8 (\text{HealthyLife}_i - 0.55)_+ + \beta_9 (\text{HealthyLife}_i - 0.9)_+ + \epsilon_i \\ \text{with} \quad (x_i - \kappa)_+ &= \begin{cases} 0, & \text{if} \quad x_i < \kappa \\ \\ x_i - \kappa, & \text{if} \quad x_i \geq \kappa \end{cases}\end{aligned} \tag{1}$$

It segments the covariate space to fit individual linear model in between these segments. This way I create a model with multiple parametric and linear pieces put together. Based on the shape of the plots from figure 3, I assume breaks in the line of regression at $GDP = (0.7; 1)$ and $HealthyLife = (0.32; 0.55; 0.9)$. By being locally linear and parametric the piecewise model produces estimates with a lower variance than the nonparametric GAM model, and provides easier interpretation and with its linearity.
The polynomial model is given by the following equation :

$$\begin{aligned}\text{Happiness Score}_i &= \beta_1 \text{SocialSupport}_i + \beta_2 \text{Freedom}_i + \beta_3 \text{GDP}_i + \beta_4 \text{GDP}_i{}^2 + \beta_5 \text{GDP}_i{}^3 \\ &+ \beta_6 \text{GDP}_i{}^4 + \beta_7 \text{GDP}_i{}^5 + + \beta_8 \text{HealthyLife}_i + \beta_9 \text{HealthyLife}_i{}^2 + \beta_{10} \text{HealthyLife}_i{}^3 \\ &+ \beta_{11} \text{HealthyLife}_i{}^4 + \beta_{12} \text{HealthyLife}_i{}^5 + \beta_{13} \text{HealthyLife}_i{}^6 + \beta_{14} \text{HealthyLife}_i{}^7 + \epsilon_i\end{aligned} \tag{2}$$

By adding 5 polynomial terms for *GDP* and 7 for *HealthyLife* I have a parametric model but much more flexible that can account for the non linearities in these two variables. This parametric form provides estimates with a lower variance than the GAM model. However this gain is decreased by the increased variance coming from the large number of regressors compared to the amount of data, and the correlation between the polynomial terms. Nonetheless, this clear functional form helps with the interpretation and inference. The polynomial orders were chosen based on the rounded lower value of the effective degrees of freedom coming from GAM2, observed in Table 3. This is not a perfect method, but it serves as an indicator of how much flexibility is needed.

After fitting the models, I can make the following observations. First of all, the coefficients of the piecewise model are coherent with the graphs seen in Figure 3. The coefficients have the right signs and are not significant when the regression line is flat or has large confidence intervals, like for values of *GDP* between 0.7 and 1. For the polynomial model many of the polynomial coefficients are non significant but for both the largest one is, leading us to think that the number of degrees is about right. In terms of fit when GAM2 had an adjusted R^2 of 0.837 and the linear model a R^2 of 0.765, the piecewise model has an adjusted R^2 of 0.811 and of 0.830 for the polynomial model. This shows that the new parametric models take the non linearities observed in the variables more into account and provide an improved fit compared to the linear model. Nonetheless this fit is not as good as the the GAM model. This comes with the benefit of increased interpretability, easier inference and, but not as surely for the polynomial model, reduced variance of the estimates. Of the two new parametric approaches the weaker one seems the polynomial model as its large number of polynomial degrees makes it less interpretable and less efficient.

4 Conclusion

In this elaboration I have presented several parametric as well as nonparametric models to identify the key determinants of the happiness score, and the way they interact with this variable. The linear model seemed to perform well, but its assumption of linearity was too restrictive. I then switched to a fully flexible GAM model. In this model, all variables are taken individually by separately estimating a flexible functional form, based on the splines method. This model proved to give a better fit of the data by relaxing the linear assumption and allowing for a flexible representation of variable interaction not possible with a standard linear model. However its nonparametric form comes at the cost of easy interpretability, easy inference and increased variance of the estimates. The last approach was to construct a parametric model but with a form based on the insights coming from the GAM model. I've proposed two models, a piecewise linear model and polynomial model. These two models showed a similar, even if slightly worse, fit of the data compared to the GAM model. However they provide increased interpretability, easier inference and lower variance of the estimates.