---

**Article:**
Gu, Y. and Wei, H.L. orcid.org/0000-0002-4704-7346 (2018) Significant indicators and determinants of happiness: Evidence from a UK survey and revealed by a data-driven systems modelling approach. Social Sciences, 7 (4). 53. ISSN 2076-0760

https://doi.org/10.3390/socsci7040053

---

social sciences

MDPI

Article

# Significant Indicators and Determinants of Happiness: Evidence from a UK Survey and Revealed by a Data-Driven Systems Modelling Approach

**Yuanlin Gu and Hua-Liang Wei ***

Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, UK; ygu11@sheffield.ac.uk
* Correspondence: w.hualiang@sheffield.ac.uk; Tel.: +44-114-2225198

Received: 1 March 2018; Accepted: 27 March 2018; Published: 29 March 2018


**Abstract:** This study aims to establish a quantitative relationship between lifestyle and happiness in the UK based on over 10,000 surveyed samples with 63 lifestyle variables from the UK Understanding Society Data. Transparent parametric models are built and a number of significant explanatory variables (lifestyle indicators) have been identified using a systems engineering modelling approach. Specifically; based on the traditional orthogonal forward regression (OFR) algorithm; the study introduces a new metrics; with which the impacts of lifestyle variables (and/or their interactions) can be quantitatively measured and identified one by one. These identified significant indicators provide a meaningful parsimonious representation of the relationship between happiness and lifestyle; revealing how happiness quantitatively depends on lifestyle; and how the lifestyle variables interactively affect happiness. For example; the quantitative results of a linear model indicate that lifestyle variables such as 'health'; 'income'; and 'retirement'; impacts happiness significantly. Furthermore; the results of a bilinear model show that some interaction variables such as 'retired' together with 'elder'; 'fair health' together with 'low-income' and so on; are significantly related to happiness.

**Keywords:** happiness; lifestyle; life satisfaction; nonlinear system; data-driven modelling; systems engineering

## 1. Introduction

Lifestyle is known to be an effective cause of happiness (life satisfaction) worldwide. As people's lifestyle may include views and habits on fields such as health, religion and politics and so on, it is necessary to find how these aspects of lifestyle independently or interactively affect happiness, in order to obtain transparent representation of the relationship between lifestyle and happiness. Numerous studies have been conducted to find out which aspects of lifestyle have significant impacts on happiness (Easterlin 1995; Fletcher et al. 1990; Hills and Argyle 1998). Firstly, happiness is affected by employment, which can be described by working hours, type of the job and incomes (Booth and Van Ours 2008; Ekici and Koydemir 2016; Lim et al. 2017; Köksa et al. 2017; Ward and King 2016). Secondly, there is accumulating evidence that well-being is associated with health status as well as healthy lifestyle, for example, the consumption of fruits and vegetables (Gschwandtner et al. 2015; Kvintova et al. 2016; Mujcic and Oswald 2016; Puvill et al. 2016). Thirdly, according to some research, higher level of education neither produces higher happiness, nor the wealth, nor health (Veenhoven 1996; Sabatini 2014). However, IQ, which can be improved by education, can independently affect health status to indirectly influence happiness (Hartog and Oosterbeek 1998). There are many other aspects of lifestyle that could be potentially

effective for affecting people's happiness, such as age, religion, marital status, personality and past lives (Elmslie and Tebaldi 2014; Frey and Stutzer 2002; Fujita and Diener 2005; Jewell and Kambhampati 2014; Carr et al. 2014). The happiness of older adults is especially found to be closely related to, for example, health status, personality, retirement and other aspects like psychosocial variables and so forth (Berg et al. 2006, 2008, 2009, 2011; Enkvist et al. 2012; Hansson et al. 2017; Henning et al. 2017; von Humboldt et al. 2014; Fagerström et al. 2012).

Quantitative methods used in previous research on the relationship between lifestyle and happiness have mainly focused on traditional linear model and parameter estimation methods such as ordinary least squares and instrumental variable. However, there is still large room for improvement. For example, ordinary least squares method can only be used for normal linear model parameter estimation. Ordinary least squares method is not applicable for small $n$ and large $p$ regression problem where the number of parameters is larger than the number of observations. For multicollinearity (i.e., collinearity) regression problem, the solution given by ordinary least squares method may not be reliable. In addition, the ordinary least square estimation usually uses all the available explanatory variables for regression, ignoring the fact that only a small number of variables (lifestyles) might be significantly important for explaining the response variables (happiness). It is also time-consuming for researchers to manually choose the variables according to the vast existing literature. In some situations when nonlinear interaction among variables (lifestyle) need to be considered for nonlinear modelling, there often exists a huge number of potential interactions and the initial full model could be very complicated and cannot be used for analysis or prediction. These situations also increase the difficulties using statistical analysis approaches such as t test because the implementations of statistics usually require the information of estimated model. If the model itself is too complicated or inaccurate, the results brought by these tests can be unreliable. Thus, the question raises: how to select the significant variables from a large number of candidate interactive variables and how to determine the number of interactions that should be included in the model?

In the existing literature, it still remains an open question about how the interactions of lifestyle variables have significant impacts on life satisfaction. To fill such a gap, a novel approach is proposed in this study to find out the most significant lifestyle interaction variables and to provide an effective tool that can easily be adapted for quantitative analysis of well-being using big data. There are several unresolved questions in the research filed of life satisfaction and happiness, for example, how to quantitatively measure the significance and impact of political, culture, religion and so forth, on happiness.

Based on the above concerns, this study aims to introduce a systems engineering modelling approach to reveal and characterize the dependent relationship of happiness on lifestyles. More specifically, based on the traditional orthogonal forward regression (OFR) algorithm (Chen et al. 1989), the study introduces a new metrics, with which the impacts of lifestyle variables (and/or their interactions) can be quantitatively measured and identified one by one.

The main contribution of this study is twofold. First, it introduces three new measures, namely, orthogonal error reduction ratio ($oERR$), non-orthogonal error reduction ratio ($nERR$) and contribution factor ($CI$), which enhance the explanatory ability of the OFR algorithm. These measures, together with the OFR algorithm, can effectively, quantitatively measure the significance of lifestyle variables to happiness. The proposed approach in this study has the following two advantages: (i) the algorithm is easy to implement and compute; (ii) the algorithm provides an efficient way to identify the most significant variables (lifestyles) and the associated cross terms (interactions of the lifestyle variables) for representing life satisfaction and this enables the establishment of a parsimonious representation of the relationship between lifestyles and happiness. Second, this study contributes to the literature by providing new evidence regarding the relationship between lifestyle and happiness. Our results clearly indicate that the most effective lifestyle variables, in a national scale of the UK, are 'health,' 'income,' 'retirement,' among others. Most of the significant variables picked out by the proposed algorithm are in line with previous findings (Ball and Chernova 2008; Bender and Jivan 2005; Gorry et al. 2015;

Gerdtham and Johannesson 2001; Puvill et al. 2016). However, the proposed method provide a clearer quantitative indication of the key lifestyle variables (or their interactions). Furthermore, our bilinear model also demonstrates that some interaction variables such as 'retired' together with 'elder,' 'fair health' together with 'low-income' and so on, are significantly related to happiness.

## 2. Data

This study uses the UK Understand Society Data which have been collected since 2009. Following over 40,000 households in the UK with more than five waves of surveys, the data consist of a wide range of investigations, covering comprehensive features of basic daily life information of the society. This study uses the data from wave 2 where general questions about lifestyle were asked and answers were recorded.

The life satisfaction variable, denoted by $y$ in the table, is a categorical variable described by 7 integers (i.e., 1, 2, . . . , 7). Participants are asked to score their own satisfaction of life using a 7-point scale, from 1 (very unsatisfied) to 7 (very satisfied). Lifestyle variables are meanly described by two kinds of categorical variables. First, there are categorical variables such as age, job, education, gender and region. These variables are dummy coded and each species of the group is presented by an individual variable, as shown in Table 1. Dummy coding is a simple and effective method that uses only 0 and 1 to convey all the information from a group (Suits 1957): 1 denotes the confirmation of what the variable describes, while 0 means rejection. For example, variables $u1$ to $u7$ describe the type of jobs involved in the investigation; variables $u30$ to $u34$ describe the groups of age; variables $u42$ to $u53$ describe the types of degree hold by the individuals. Second, there are a few variables which just describe a simple 'yes or no' questions like 'if have children,' 'if married,' 'if cohabited,' 'if married' and 'if smoke.' They can be treated as binary categorical variable, with a value of either 1 (confirmation) or 0 (rejection).

After data pre-processing (dummy coding), a total number of 63 lifestyle variables are considered (as shown in Table 1), from which the most important indicators will be identified using our newly proposed systems engineering modelling approach and the most important interactions among the 42 variables will also be investigated. It can be noticed that the age of the respondents is 49 on average, ranging from 21 to 103. About 52% of the respondents are married and 21% of the respondents have children. Most of the individuals have a full-time or part-time job, while a small percentage of them is sick and disabled. The respondents come from different regions of the UK, for example, about 15% of people are from South East and many of them are from North West, East and London and so forth.

**Table 1.** Statistic Description of UKUS Data.

| Variable | Term | Explanation | Mean | Std.Dev | Min | Max |
|---|---|---|---|---|---|---|
| Job | $u1$ | FT and PT | 0.5434 | 0.2481 | 0 | 1 |
| | $u2$ | unemployed | 0.0610 | 0.0573 | 0 | 1 |
| | $u3$ | retired | 0.1545 | 0.1307 | 0 | 1 |
| | $u4$ | family care | 0.0586 | 0.0552 | 0 | 1 |
| | $u5$ | FT student | 0.0577 | 0.0544 | 0 | 1 |
| | $u6$ | sick and disabled | 0.0186 | 0.0183 | 0 | 1 |
| | $u7$ | other | 0.0064 | 0.0063 | 0 | 1 |
| Children | $u8$ | if has children | 0.2115 | 0.1668 | 0 | 1 |
| Days eat fruit per week | $u9$ | never | 0.0631 | 0.0592 | 0 | 1 |
| | $u10$ | 1–3 days | 0.2938 | 0.2075 | 0 | 1 |
| | $u11$ | 4–6 days | 0.2005 | 0.1603 | 0 | 1 |
| | $u12$ | everyday | 0.4425 | 0.2467 | 0 | 1 |
| Days eat vegetables per week | $u13$ | never | 0.0165 | 0.0162 | 0 | 1 |
| | $u14$ | 1–3 days | 0.1846 | 0.1506 | 0 | 1 |
| | $u15$ | 4–6 days | 0.2786 | 0.2010 | 0 | 1 |
| | $u16$ | everyday | 0.5203 | 0.2496 | 0 | 1 |

**Table 1.** *Cont.*

| Variable | Term | Explanation | Mean | Std.Dev | Min | Max |
|---|---|---|---|---|---|---|
| Smoke | $u17$ | if smoke | 0.3320 | 0.2218 | 0 | 1 |
| Sports frequency | $u18$ | 3+ times a week | 0.2235 | 0.1735 | 0 | 1 |
| | $u19$ | 1–3 times a week | 0.2832 | 0.2030 | 0 | 1 |
| | $u20$ | at least once/month | 0.2104 | 0.1661 | 0 | 1 |
| | $u21$ | at least 3–4 times/year | 0.1775 | 0.1460 | 0 | 1 |
| | $u22$ | twice in past year | 0.0648 | 0.0606 | 0 | 1 |
| | $u23$ | once in past year | 0.0406 | 0.0390 | 0 | 1 |
| Sport Facilities | $u24$ | very difficult | 0.0118 | 0.0116 | 0 | 1 |
| | $u25$ | difficult | 0.0361 | 0.0348 | 0 | 1 |
| | $u26$ | not difficult nor easy | 0.0831 | 0.0762 | 0 | 1 |
| | $u27$ | easy | 0.4023 | 0.2405 | 0 | 1 |
| | $u28$ | very easy | 0.4636 | 0.2487 | 0 | 1 |
| Marriage | $u29$ | if married | 0.5227 | 0.2495 | 0 | 1 |
| Age | $u30$ | 16–24 | 0.0437 | 0.0418 | 0 | 1 |
| | $u31$ | 25–34 | 0.1533 | 0.1298 | 0 | 1 |
| | $u32$ | 35–49 | 0.3325 | 0.2220 | 0 | 1 |
| | $u33$ | 50–64 | 0.2530 | 0.1890 | 0 | 1 |
| | $u34$ | above 65 | 0.2175 | 0.1702 | 0 | 1 |
| Cohabitation | $u35$ | if cohabited | 0.1469 | 0.1254 | 0 | 1 |
| Education | $u36$ | high degree | 0.1005 | 0.0904 | 0 | 1 |
| | $u37$ | other high degree | 0.0807 | 0.0742 | 0 | 1 |
| | $u38$ | A-level | 0.0080 | 0.0079 | 0 | 1 |
| | $u39$ | GCSE | 0.0471 | 0.0449 | 0 | 1 |
| | $u40$ | other qualification | 0.0002 | 0.0002 | 0 | 1 |
| | $u41$ | none | 0.1772 | 0.1458 | 0 | 1 |
| Region | $u42$ | North East | 0.0407 | 0.0390 | 0 | 1 |
| | $u43$ | North West | 0.1090 | 0.0971 | 0 | 1 |
| | $u44$ | Yorkshire and the Humber | 0.0750 | 0.0694 | 0 | 1 |
| | $u45$ | East Midlands | 0.0825 | 0.0757 | 0 | 1 |
| | $u46$ | West Midlands | 0.0710 | 0.0660 | 0 | 1 |
| | $u47$ | East of England | 0.1046 | 0.0936 | 0 | 1 |
| | $u48$ | London | 0.1047 | 0.0937 | 0 | 1 |
| | $u49$ | South East | 0.1594 | 0.1340 | 0 | 1 |
| | $u50$ | South West | 0.0988 | 0.0890 | 0 | 1 |
| | $u51$ | Wales | 0.0462 | 0.0441 | 0 | 1 |
| | $u52$ | Scotland | 0.0755 | 0.0698 | 0 | 1 |
| | $u53$ | Northern Ireland | 0.0327 | 0.0317 | 0 | 1 |
| Health | $u54$ | excellent | 0.1744 | 0.1440 | 0 | 1 |
| | $u55$ | very good | 0.3782 | 0.2352 | 0 | 1 |
| | $u56$ | good | 0.2969 | 0.2088 | 0 | 1 |
| | $u57$ | fair | 0.1205 | 0.1060 | 0 | 1 |
| | $u58$ | poor | 0.0300 | 0.0291 | 0 | 1 |
| Income | $u59$ | living comfortably | 0.2812 | 0.2021 | 0 | 1 |
| | $u60$ | doing alright | 0.3402 | 0.2245 | 0 | 1 |
| | $u61$ | just about getting by | 0.2623 | 0.1935 | 0 | 1 |
| | $u62$ | find it quite difficult | 0.0830 | 0.0761 | 0 | 1 |
| | $u63$ | find it very difficult | 0.0334 | 0.0323 | 0 | 1 |
| Life Satisfaction | $y$ | Satisfaction Level | 5.2430 | 1.9785 | 1 | 7 |

## 3. Methods

We first consider the linear regression model where the response variable $y$ linearly depends on the candidate explanatory variables $u_1, u_2, \ldots, u_m$, as below:

$$y = a_0 + a_1u_1 + a_2u_2 + \ldots + a_mu_m + \varepsilon \quad (1)$$

Note that in many applications, linear models may not provide sufficient representation of the dependent relationship of the response variable and the explanatory variables. In this case, a nonlinear model for example a bilinear model below many give a better representation:

$$y = a_0 + a_1u_1 + a_2u_2 + \ldots + a_mu_m + a_{1,2}u_1u_2 + \ldots + a_{1,m}u_1u_m + \ldots + a_{2,2}u_2u_2 + \ldots + a_{m,m}u_mu_m + \varepsilon \quad (2)$$

where $u_1,\ u_2 \ldots u_m,\ \ldots,\ u_mu_m$ are the explanatory variables, $y$ is the response variable, $\alpha_1,\ \alpha_2 \ldots \alpha_m,\ \ldots \alpha_{m,m}$ are the estimated parameters and $\varepsilon$ is the model residual.

In many previous studies, linear models have been developed to investigate the relationship between lifestyle (represented by explanatory variables $u_1,\ u_2,\ \ldots,\ u_{42}$) and happiness (represented by response variable $y$). For such linear regression problems, the conventional way is to establish a model using all the available variables and investigate the significance of the variables by statistical tests. However, for some data analytic problems where there is a huge number of variables or some interactions of the variables that need to be considered, the initial full model with all the available variables (including linear variables and interactive variables) can be extremely complicated. For example, there is 63 explanatory variables in this study and the initial full model of nonlinear degree 2 of form (2) contains a total number of 2080 linear variables and interactive variables. Such a model is barely useful in analyze the relationship between lifestyle and happiness. Actually, since not all the variables are equally important and essential in explaining the variation of the response variable, some of the variables may be irrelevant and they can be removed from the full model. Thus, the decision on which variables are important and should stay in the model and which variables are trivial and should be removed from the model become crucially important. An orthogonal forward search (OFR) algorithm (Billings and Wei 2008; Chen et al. 1989), initially developed in the field of control and systems engineering, is applied to select explanatory variables/regressors (they are usually called 'candidate model terms' in the selection process) based on their contribution to explaining the response variable. The OFR algorithm is widely applied in many real-world applications including ecological systems (Marshall et al. 2016), environmental systems (Bigg et al. 2014), space weather (Balikhin et al. 2011; Boynton et al. 2011; Solares et al. 2016), medicine (Billings et al. 2013) and neurophysiological sciences (Li et al. 2016), social science (Wei and Bigg 2017) and so forth. In this study, we further develop the OFR algorithm by introducing two new measures which are '*oERR*' and '*CI*,' to analyze the relationship between happiness and lifestyles.

### *3.1. Orthogonal Forward Search Algorithm and Orthogonal Error Reduction Ratio (oERR)*

Note that both the linear model (1) and the bilinear model (2) are a linear-in-the-parameter representation, which can be re-arranged to a compact matrix form (4). Note that we use normal letters (for example '$y$') to represent the variables/model terms and **bold** letters (for example '$\boldsymbol{y}$') to represent the associated vectors:

$$\boldsymbol{y} = \boldsymbol{\Phi\theta} + \boldsymbol{\varepsilon} \quad (3)$$

where

$$\boldsymbol{y} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_M \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(N) \end{bmatrix}, \quad \boldsymbol{\Phi} = [\boldsymbol{\varphi}_1,\ \boldsymbol{\varphi}_2,\ \ldots,\ \boldsymbol{\varphi}_M] \quad (4)$$

and

$$\boldsymbol{\varphi}_i = \begin{bmatrix} \varphi_i(1) \\ \varphi_i(2) \\ \vdots \\ \varphi_i(N) \end{bmatrix} \quad i = 1,\ 2,\ \ldots.,\ M \quad (5)$$

where $N$ is the number of observations, $M$ is the number of candidate model terms, $\{\theta_1, \theta_2, \ldots, \theta_M\}$ are the unknown model parameters, $\{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_M\}$ are the associated candidate basis vectors generated from the candidate model terms $\{u_1, u_2 \ldots u_m\}$. The full dictionary $D$ contains all the candidate model terms $\{\varphi_1, \varphi_2, \ldots, \varphi_M\}$. Assume that all the $M$ candidate model terms are significant and the matrix $\boldsymbol{\Phi}$ is full rank in columns, then $\boldsymbol{\theta}$ can be estimated by solving:

$$\hat{\boldsymbol{\theta}} = \left[\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right]^{-1} \boldsymbol{\Phi}^T \boldsymbol{y} \tag{6}$$

The above procedure is referred as the ordinary least square (OLS) estimation. As mentioned previously, it might be the case that the terms that should be included in the model cannot be known in advance and the initial full model may contain many redundant variables/regressors. The least squares problem can therefore be severely ill-conditioned and the solution may become less or not reliable. Moreover, there is nothing to know which explanatory variables/regressors are important. To overcome the ill-conditioning and clearly know which explanatory variables/regressors are important, the OFR algorithm was developed to find a subset of the most important model terms, based on which a parsimonious representation can be achieved (Chen et al. 1989). The basic idea behind the OFR algorithm is to select significant terms in a stepwise manner. In the first step, the significance of each model term is measured by error reduction ratio ($ERR$) index. A rank is then generated according to the contribution made by each of the model terms to explaining the variation of the response variable (Billings and Wei 2008; Chen et al. 1989; Wei and Billings 2008). The ERR value of each model term is defined as:

$$ERR^{(1)}[i] = \frac{\left(\boldsymbol{y}^T \boldsymbol{\varphi}_i\right)^2}{\left(\boldsymbol{y}^T \boldsymbol{y}\right)\left(\boldsymbol{\varphi}_i^T \boldsymbol{\varphi}_i\right)} \tag{7}$$

The index of the first important model term can be identified as:

$$l_1 = arg \max_{1 \le i \le M} \left\{ ERR^{(1)}{}_i \right\} \tag{8}$$

Thus, the first model term can be selected as $\varphi_{l_1}$ and the first associated orthogonal vector can be defined as $\boldsymbol{q}_1 = \boldsymbol{\varphi}_{l_1}$. Note that once $\varphi_{l_1}$ is selected by the algorithm, the $l_1$-th candidate model term should be removed from the initial dictionary $D$ and the $l_1$-th column of the matrix $\boldsymbol{\Phi}$ (i.e., the $l_1$-th candidate basis vector $\boldsymbol{\varphi}_{l_1}$) should also be removed from the matrix accordingly. After removal $\varphi_{l_1}$ from the full dictionary, the dictionary is then reduced to a sub-dictionary, consisting of $M-1$ unselected candidate model terms.

At step $s$ ($s \ge 2$), the $M-s+1$ basis vectors of unselected candidate model terms are all firstly transformed into a new group of orthogonalised basis vectors. The orthogonlization transformation is defined as:

$$\boldsymbol{q}_j^{(s)} = \boldsymbol{\varphi}_j - \sum_{r}^{s-1} \frac{\boldsymbol{\varphi}_j^T \boldsymbol{q}_r}{\boldsymbol{q}_r^T \boldsymbol{q}_r} \boldsymbol{q}_r \tag{9}$$

where $\boldsymbol{q}_r (r = 1, 2, \ldots, s-1)$ are orthogonal vectors, $\boldsymbol{\varphi}_j (j = 1, 2, \ldots, M-s+1)$ are the basis vectors of unselected model terms and $\boldsymbol{q}_j^{(s)} (j = 1, 2, \ldots, M-s+1)$ are the new orthogonalised basis vectors. At step $s$, the contribution of each model term can be measured by the ERR value of its associated orthogonalised basis vector. It is defined as orthogonal error reduction ratio ($oERR$), as:

$$oERR^{(s)}[j] = \frac{\left(\boldsymbol{y}^T \boldsymbol{q}_j^{(s)}\right)^2}{\left(\boldsymbol{y}^T \boldsymbol{y}\right)\left(\boldsymbol{q}_j^{(s)^T} \boldsymbol{q}_j^{(s)}\right)} \tag{10}$$

So, the $s$th important model term is selected to be $\varphi_{l_s}$ and the $s$th orthogonal vector is $\boldsymbol{\varphi}_{l_s}$, where:

$$l_s = arg \underset{1 \le j \le M-s+1}{max} \left\{ oERR^{(s)}[j] \right\} \tag{11}$$

Thus, the model terms of the subset $\left[\varphi_{l_1}, \varphi_{l_2}, \ldots, \varphi_{l_n}\right]$ can be selected step by step, on at each time. The $n$ selected model terms will be included in the final model, which can be written as:

$$y = \theta_{l_1} \varphi_{l_1} + \theta_{l_2} \varphi_{l_2} + \ldots + \theta_{l_n} \varphi_{l_n} + \varepsilon \tag{12}$$

Normally, the number of selected model terms by OFR algorithm is much less than the total number of the candidate model terms (n $\ll$ M), so that a parsimonious representation can be achieved.

### 3.2. Model Length Detection, Non-Orthogonal Error Reduction Ratio (nERR) and Contribution Factor (CI)

The *oERR* is a simple but effective index in selecting the important model terms in a forward stepwise manner. *oERR* is a measure defined in a space where the basis vectors are orthogonal to each other. In order to measure the contribution made by each of the selected model terms (basis vector) to explaining the response variable in the original non-orthogonal space, a new measure called non-orthogonal error reduction ratio (*nERR*) is proposed, which is defined as:

$$nERR\left[l_i\right] = \frac{\left(\boldsymbol{y}^T \boldsymbol{\varphi}_{l_i}\right)^2}{(\boldsymbol{y}^T \boldsymbol{y})(\boldsymbol{\varphi}_{l_i}^T \boldsymbol{\varphi}_{l_i})} \tag{13}$$

where $\boldsymbol{\varphi}_{l_i}$ $(i = 1, 2, \ldots, n)$ are the associated basis vectors of selected model terms. The nERR value is calculated directly from the non-orthogonalised basis vectors of model terms so that it is not affected by the orthogonlization procedure. Then, we can define the contribution indicator $\Phi$ of each selected model term as:

$$\Phi = \frac{nERR\left[l_i\right]}{\sum_{j=1}^{n} nERR\left[l_j\right]} \times \sum_{j=1}^{n} oERR\left[l_j\right] \times 100\% \tag{14}$$

where $\frac{nERR\left[l_i\right]}{\sum_{j=1}^{n} nERR\left[l_j\right]} (i = 1, 2, \ldots, n)$ describes the amount of the contribution made by each of the model terms selected by the algorithm and $\sum_{j=1}^{n} oERR\left[l_j\right]$ indicates how much of the variation in the response variable can be explained by the selected model terms (variables/regressors).

## 4. Results

Two types of models, namely, linear regression model and bilinear regression model are considered and described as below.

### 4.1. Linear Model

The OFR algorithm was applied to analyze the UKUS data, with all the lifestyle variables from $u_1$ to $u_{63}$ and constant variable as predictors and happiness variable $y$ as response. The APRESS suggested that a model consisting of 11 variables (excluding the constant variable) is the best choice to fit the data (as shown in Table 2). Note that the constant variable (denoted by $u_0$) is included in the model because of the bias or shift of the mean but it is meaningless for explaining the happiness. Thus, we define $z = y - u_0$ as the new response variable and re-estimated the model, to avoid selecting constant variable as significant variable. The estimated parameters of the 11 selected significant variables, along with the associated contribution indicators and t test result, are shown in Table 2. As depicted in the OFR algorithm, the variables in Table 2 are listed in the order of their entrance into the model in a forward stepwise way, step by step and one in each step.

**Table 2.** Linear model.

| No | Term | Description/Feature | Parameter | oERR (100%) | CI (100%) |
|---|---|---|---|---|---|
| 1 | $u59$ | income (living comfortably) | 0.9818 | 18.3247 | 10.1596 |
| 2 | $u60$ | income (doing alright) | 0.7231 | 11.1139 | 6.1618 |
| 3 | $u54$ | health (excellent) | 0.5612 | 1.7470 | 6.2711 |
| 4 | $u55$ | health (every good) | 0.2980 | 2.1867 | 7.7090 |
| 5 | $u58$ | health (poor) | −0.9582 | 0.9018 | 0.2194 |
| 6 | $u3$ | retired | 0.3950 | 0.8436 | 4.4210 |
| 7 | $u63$ | income (find it very difficult) | −0.4853 | 0.5598 | 0.2340 |
| 8 | $u57$ | health (fair) | −0.4121 | 0.3685 | 0.0480 |
| 9 | $u61$ | income (just about getting by) | 0.3191 | 0.5905 | 0.9533 |
| 10 | $u48$ | region (London) | −0.2229 | 0.1610 | 0.5815 |
| 11 | $u6$ | sick and disable | −0.4226 | 0.1273 | 0.1661 |

It can be observed that the following factors appear to have a significantly positive impact on happiness: 'income (living comfortably),' 'income (doing alright),' 'income (just about getting by),' 'retired,' 'health (excellent),' 'health (every good).' On the contrary, 'income (find it very difficult),' 'health (poor),' health (fair),' 'sick and disable' and 'region (London)' impact significantly negatively on happiness.

With only 11 variables chosen from the 63 candidate variables, the model provides a simple representation of the relationship between happiness and lifestyle, revealing how happiness quantitatively depends on lifestyle and how the lifestyle variables individually and collectively affect happiness. The contribution indicator is calculated to measure the contribution of each selected variable for explaining the response variable. For example, the contribution factor of 'retired' means that it explains 4.42% of the relationship between lifestyles and happiness.

Compared with the traditional ordinary least squares and t test methods, the new metrics provides an efficient and effective way to select the important explanatory variables in a stepwise manner. The term selection process can be automatically terminated by using APRESS criterion when all the important variables are selected. Overall, the new metrics not only reduces the time on large dataset processing and model estimation but also makes it possible to distinguish the most important variables (linear and nonlinear model terms or regressors for nonlinear modelling). The model could be potentially applied in many areas, such as healthcare and policy making, to provide a transparent reference of which lifestyle variables have significant impacts on happiness.

### *4.2. Nonlinear Model*

Since two or more lifestyle behaviors of an individual might affect the happiness simultaneously and interactively, nonlinear models, which could capture the effect of the interactions of lifestyle variables, are often more meaningful. For example, the variables 'income (find it quite difficult)' and 'health (fair)' are known to decrease the happiness separately according to the linear model. Then, how could the interaction variable of 'income (find it quite difficult)' and 'health (fair)' affect happiness? Thus, a bilinear model (that is a model with a nonlinear degree of 2) is developed to solve this question and reveal which of the interactions of variables are significantly effective for happiness.

The bilinear model was estimated using the OFR algorithm and the APRESS criterion. The most important 10 model terms selected from all the candidate model terms, are listed in Table 3, where the variables are listed in the order of their entrance into the model in a forward stepwise way, step by step and one in each step. From the nonlinear model, it is clear that 'income,' 'health' and some other linear variables always matter a lot either in linear or nonlinear models. It might be inferred that some interactions of the candidate terms play an important role in explaining the relationship between lifestyle and happiness. For example, the interaction term '$u3 \times u34$' indicates that retired people aged above 65 are more likely to be happy. In this way, the nonlinear model could provide a broader picture of how happiness is affected by different lifestyles.

**Table 3.** Nonlinear model.

| No | Term | Variable | Parameter | *oERR* (100%) | *CI* (100%) |
|---|---|---|---|---|---|
| 1 | $u59$ | income (living comfortably) | 0.7033 | 18.3247 | 10.0211 |
| 2 | $u60$ | income (doing alright) | 0.4526 | 11.1139 | 6.0778 |
| 3 | $u54$ | health (excellent) | 0.8181 | 1.7470 | 6.1856 |
| 4 | $u55$ | health (very good) | 0.5553 | 2.1867 | 7.6038 |
| 5 | $u58$ | health (poor) | −0.7844 | 0.9018 | 0.2164 |
| 6 | $u3 \times u34$ | retired & aged above 65 | 0.4065 | 0.8811 | 4.3129 |
| 7 | $u63$ | income (find it very difficult) | −0.7593 | 0.5577 | 0.2308 |
| 8 | $u57 \times u62$ | health (fair) & income (find it quite difficult) | −0.6460 | 0.5045 | 0.2694 |
| 9 | $u29 \times u56$ | married & health (good) | 0.3261 | 0.4511 | 1.8685 |
| 10 | $u17 \times u62$ | smoke & income (find it quite difficult) | −0.4398 | 0.2560 | 0.1383 |

## 5. Discussions and Conclusions

This work proposed a new method to investigate the relationship between happiness and lifestyle by analyzing the UK Understanding Society (UKUS) Data. Using the OFR algorithm, the most significant variables of lifestyle for happiness were identified. Both linear and bilinear models were created to represent the relationship between happiness and lifestyle using the selected significant variables.

The results from our study show that, on the collective national level of the UK, the following factors appear to have significantly positive impact on happiness: 'income (living comfortably),' 'income (doing alright),' 'income (just about getting by),' 'retired,' 'health (excellent),' 'health (every good).' For the role of income, our finding re-confirms the conclusion of previous studies by Ball and Chernova (2008) that "both absolute and relative income are positively and significantly correlated with happiness." As for the importance of the status of retirement, our finding is perfectly consistent with that reported in Bender and Jivan (2005), where their findings include two aspects. Firstly, based on a nationally representative sample of the U.S. elderly population, a majority of people fully retired in 2000 stated that their well-being in retirement became better. Secondly, older retirees have higher retirement satisfaction than those who are under 62 years old. Our result also re-confirms other previous research outcome for example the finding that given in Gorry et al. (2015), concluding that retirement improves both health and life satisfaction. On the contrary, 'income (find it very difficult),' 'health (poor),' health (fair),' 'sick and disable' and 'region (London)' impact significantly negatively on happiness. Our finding is in line with previous studies (e.g., Gerdtham and Johannesson 2001; Puvill et al. 2016) that good health is related with a higher level of happiness, while bad health could reduce individuals' happiness. With respect to the negative effect of region on happiness, it might be because people in a metropolis like London face more enhanced competition than their counterparts in small cities.

From the bilinear model, it was observed that some interaction terms affect happiness significantly and positively, while some other interaction terms have significantly negative impacts on happiness. For example, the interaction variable of 'retirement' and 'above 65' indicates that the retired people who are over 65 years old are more likely to be happy, which supports the findings of previous study (see for example (Henning et al. 2017)). We also find that marriage could enhance the positive relationship between good health status and happiness, while smoke could enhance the negative effect of low income on happiness.

Rather following the literature and simply assuming that certain variables are 'significant' or 'important' for representing happiness, this study advocates to use a data-driven modelling approach (such as the improved variant of OFR algorithm), together with available qualitative analysis knowledge, to identify the most important variables from a huge number of candidate variables and their interactions. It is known that the size and complexity of the well-being data is increasing rapidly, there is an increasing demand for quantitative methods for automatic identification of important lifestyle variables and detection of interaction variables. In this sense, the proposed method provides

an effective automatic tool which can save data analysis cost and time and meanwhile produces a ranked list of important lifestyle variables.

A limitation of the study is that model do not consider the impacts of politics, religion and other factors, as the data set used does not contain variables that properly describe these aspects. Thus, one of our future research directions would be to fuse information from different resources to include variables closely related to these aspects and investigate their influence using the proposed approach.

**Acknowledgments:** The authors acknowledge that this work was supported in part by the Engineering and Physical Sciences Research Council Platform Grant EP/H00453X/1.

**Author Contributions:** H.W. conceived, designed and conducted the project; Y.G. performed the numerical experiments and analyzed the data; H.W. and Y.G. wrote the paper.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## References

Balikhin, Michael A., Richard J. Boynton, Simon N. Walker, Joe E. Borovsky, Stephen A. Billings, and Hua-Liang Wei. 2011. Using the NARMAX approach to model the evolution of energetic electrons fluxes at geostationary orbit. *Geophysical Research Letters* 38: 1–5. [CrossRef]

Ball, Richard, and Kateryna Chernova. 2008. Absolute income, relative income and happiness. *Social Indicators Research* 88: 497–529. [CrossRef]

Bender, Keith A., and Natalia A. Jivan. 2005. *What Makes Retirees Happy?* An Issue in Brief, 25(5); Boston: Center for Retirement Research at Boston College, Available online: http://www.bc.edu/centers/crr/issues/ib_28.pdf (accessed on 18 December 2016).

Berg, Anne I., Linda B. Hassing, Gerald E. McClearn, and Boo Johansson. 2006. What matters for life satisfaction in the oldest-old? *Aging and Mental Health* 10: 257–64. [CrossRef] [PubMed]

Berg, Anne I., Linda B. Hassing, Sven E. Nilsson, and Boo Johansson. 2008. "As long as I'm in good health." The relationship between medical diagnoses and life satisfaction in the oldest-old. *Aging Clinical and Experimental Research* 21: 307–13. [CrossRef]

Berg, Anne I., Lesa Hoffman, Linda B. Hassing, Gerald E. McClearn, and Boo Johansson. 2009. What matters and what matters most, for change in life satisfaction in the oldest-old? A study over 6 years among individuals 80+. *Aging and Mental Health* 13: 191–201. [CrossRef] [PubMed]

Berg, Anne I., Linda B. Hassing, Valgeir Thorvaldsson, and Boo Johansson. 2011. Personality and personal control make a difference for life satisfaction in the oldest-old: Findings in a longitudinal population-based study of individuals 80 and older. *European Journal of Ageing* 8: 13–20. [CrossRef] [PubMed]

Bigg, Grant R., Hua-Liang Wei, David J. Wilton, Yifan Zhao, Stephen A. Billings, Edward Hanna, and Visakan Kadirkamanathan. 2014. A century of variation in the dependence of Greenland iceberg calving on ice sheet surface mass balance and regional climate change. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 470. [CrossRef] [PubMed]

Billings, Stephen A., and Hua-Liang Wei. 2008. An adaptive orthogonal search algorithm for model subset selection and non-linear system identification. *International Journal of Control* 81: 714–24. [CrossRef]

Billings, Catherine G., Hua-Liang Wei, Patrick Thomas, Seamus J. Linnane, and Ben D. M. Hope-Gill. 2013. The prediction of in-flight hypoxaemia using non-linear equations. *Respiratory Medicine* 107: 841–47. [CrossRef] [PubMed]

Booth, Alison L., and Jan C. Van Ours. 2008. Job satisfaction and family happiness: The part-time work puzzle. *The Economic Journal* 118: F77–99. [CrossRef]

Boynton, Richard J., Michael A. Balikhin, Stephen A. Billings, Hua-Liang Wei, and Natalia Ganushkina. 2011. Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere. *Journal of Geophysical Research: Space Physics* 116: 1–8. [CrossRef]

Carr, Deborah, Vicki A. Freedman, Jennifer C. Cornman, and Norbert Schwarz. 2014. Happy marriage, happy life? Marital quality and subjective well-being in later life. *Journal of Marriage and Family* 76: 930–48. [CrossRef] [PubMed]

Chen, Sheng, Stephen A. Billings, and Wanlong Luo. 1989. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control* 50: 1873–96. [CrossRef]

Easterlin, Richard A. 1995. Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior & Organization* 27: 35–47. [CrossRef]

Ekici, Tufan, and Selda Koydemir. 2016. Income expectations and happiness: Evidence from British panel data. *Applied Research in Quality of Life* 11: 539–52. [CrossRef]

Elmslie, Bruce T., and Edinaldo Tebaldi. 2014. The determinants of marital happiness. *Applied Economics* 46: 3452–62. [CrossRef]

Enkvist, Åsa, Henrik Ekström, and Sölve Elmståhl. 2012. What factors affect life satisfaction (LS) among the oldest-old? *Archives of Gerontology and Geriatrics* 54: 140–45. [CrossRef] [PubMed]

Fagerström, Cecilia, Magnus Lindwall, Anne I. Berg, and Mikael Rennemark. 2012. Factorial validity and invariance of the Life Satisfaction Index in older people across groups and time: Addressing the heterogeneity of age, functional ability and depression. *Archives of Gerontology and Geriatrics* 55: 349–56. [CrossRef] [PubMed]

Fletcher, Garth J. O., Julie Fitness, and Neville M. Blampi. 1990. The link between attributions and happiness in close relationships: The roles of depression and explanatory style. *Journal of Social and Clinical Psychology* 9: 243–55. [CrossRef]

Frey, Bruno S., and Alois Stutzer. 2002. What Can Economists Learn from Happiness Research? *Journal of Economic Literature* 40: 402–35. [CrossRef]

Fujita, Frank, and Ed Diener. 2005. Life satisfaction set point: Stability and change. *Journal of Personality and Social Psychology* 88: 158–64. [CrossRef] [PubMed]

Gerdtham, Ulf G., and Magnus Johannesson. 2001. The relationship between happiness, health and socio-economic factors: Results based on Swedish microdata. *Journal of Socio-Economics* 30: 553–57. [CrossRef]

Gorry, Aspen, Devon Gorry, and Sita Slavov. 2015. *Does Retirement Improve Health and Life Satisfaction?* NBER Working paper, No. 21326; Cambridge: National Bureau of Economic Research. [CrossRef]

Gschwandtner, Adelina, Sarah L. Jewell, and Uma Kambhampati. 2015. On the relationship between lifestyle and happiness in the UK. Paper presented at 89th Annual Conference, Warwick University, Coventry, UK, April 13–15.

Hansson, Isabelle, Sandra Buratti, Valgeir Thorvaldsson, Boo Johansson, and Anne I. Berg. 2017. Changes in life satisfaction in the retirement transition: Interaction effects of transition type and individual resources. *Work, Aging and Retirement*, 1–15. [CrossRef]

Hartog, Joop, and Hessel Oosterbeek. 1998. Health, wealth and happiness: Why pursue a higher education? *Economics of Education Review* 17: 245–56. [CrossRef]

Henning, Georg, Isabelle Hansson, Anne I. Berg, Magnus Lindwall, and Boo Johansson. 2017. The role of personality for subjective well-being in the retirement transition—Comparing variable- and person-oriented models. *Personality and Individual Differences* 116: 385–92. [CrossRef]

Hills, Peter, and Michael Argyle. 1998. Positive moods derived from leisure and their relationship to happiness and personality. *Personality and Individual Differences* 25: 523–35. [CrossRef]

Jewell, Sarah, and Uma S. Kambhampati. 2014. Are happy youth also satisfied adults? An analysis of the impact of childhood factors on adult life satisfaction. *Social Indicators Research* 121: 543–67. [CrossRef]

Köksa, Onur, Harun Uçak, and Faruk Şahin. 2017. Happiness and domain satisfaction in Turkey. *International Journal of Happiness and Development* 3: 323–41. [CrossRef]

Kvintova, Jana, Michal Kudlacek, and Dagmar Sigmundova. 2016. Active lifestyle as a determinant of life satisfaction among university students. *Anthropologist* 24: 179–85. [CrossRef]

Li, Yang, Hua-Liang Wei, Stephen A. Billings, and Ptolemaios G. Sarrigiannis. 2016. Identification of nonlinear time-varying systems using an online sliding-window and common model structure selection (CMSS) approach with applications to EEG. *International Journal of Systems Science* 47: 2671–81. [CrossRef]

Lim, Hock Eam, Daigee Shaw, and Pei-shan Liao. 2017. Revisiting the income-happiness paradox: The case of Taiwan and Malaysia. *Institutions and Economies* 9: 53–69.

Marshall, Abigail M., Grant R. Bigg, Sonja M. van Leeuwen, John K. Pinnegar, Hua-Liang Wei, Thomas J. Webb, and Julia L. Blanchard. 2016. Quantifying heterogeneous responses of fish community size structure using novel combined statistical techniques. *Global Change Biology* 22: 1755–68. [CrossRef] [PubMed]

Mujcic, Redzo, and Andrew J. Oswald. 2016. Evolution of well-being and happiness after increases in consumption of fruit and vegetables. *American Public Health Association* 106: 1504–10. [CrossRef] [PubMed]

Puvill, Thomas, Jolanda Lindenberg, Antonius J. M. de Craen, Joris P. J. Slaets, and Rudi G. J. Westendorp. 2016. Impact of physical and mental health on life satisfaction in old age: a population based observational study. *BMC Geriatrics* 16: 194. [CrossRef] [PubMed]

Sabatini, Fabio. 2014. The relationship between happiness and health: Evidence from Italy. *Social Science & Medicine* 114: 178–87. [CrossRef]

Solares, Jose R. A., Hua-Liang Wei, Richard J. Boynton, Simon N. Walker, and Stephen A. Billings. 2016. Modeling and prediction of global magnetic disturbance in near-Earth space: A case study for Kp index using NARX models. *Space Weather* 14: 899–916.

Suits, Daniel B. 1957. Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association* 52: 548–51. [CrossRef]

Veenhoven, Ruut. 1996. Developments in satisfaction-research. *Social Indicators Research* 37: 1–46. [CrossRef]

von Humboldt, Sofia, Isabel Leal, and Filipa Pimenta. 2014. Living well in later life: The influence of sense of coherence and socio-demographic, lifestyle and health-related factors on older adults' satisfaction with life. *Applied Research in Quality of Life* 9: 631–42. [CrossRef]

Ward, Sarah J., and Laura A. King. 2016. Poor but happy? Income, happiness and experienced and expected meaning in life. *Social Psychological and Personality Science* 7: 463–70. [CrossRef]

Wei, Hua-Liang, and Grant R. Bigg. 2017. The dominance of food supply in changing demographic factors across Africa: A model using a systems identification approach. *Social Science* 6: 122. [CrossRef]

Wei, Hua-Liang, and Stephen A. Billings. 2008. Model structure selection using an integrated forward orthogonal search algorithm assisted by square correlation and mutual information. *International Journal of Modelling, Identification and Control* 3: 341–56. [CrossRef]