How to achieve happiness

Elena Lahoz Moreno, Claudia Lucio Sarsa, Antonio Martínez Payá

December 31th 2017

1 Abstract

Reaching happiness is something that human beings have always dreamed of, with this project we want to know if we can measure and score it. We have chosen a dataset with 155 rows -which are the countries- and 12 variables. We have applied linear modeling combined with different dimension reduction techniques such as: Principal Component Analysis, Sparse PCA and Partial Least Squares, resulting PLS with 2 components as the appropriate model due to its easier interpretability. We should take into account that we have heteroskedasticity in our data, so we can not apply inference. To know the validity of the model we have used cross-validation. Our interpretation of the results is that happiness is affected positively by good socioeconomic factors while suicides and absence of peace affect negatively.

2 Introduction

The question about happiness is essential in the emergence of ethics in ancient Greece. Philosophers found very different answers, which shows that, as Aristotle said, we all agree that we want to be happy, but as soon as we try to clarify how we can be happy, the discrepancies begin. In the Nicomachean Ethics, written in 350 BCE, Aristotle stated that happiness (also being well and doing well) is the only thing that humans desire for its own sake, unlike wealth, honour, health or friendship. He observed that human beings sought wealth, or honour, or health not only for their own sake but also in order to be happy.

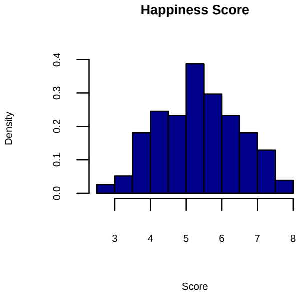
That's why we have asked ourselves: How can people achieve happiness? What are the main social factors influding in our happiness? These are some questions that we want to discover in this project.

Our purpose was inspired by a dataset we found in Kaggle. This dataset was created to help to develop the document World Happiness Record 2017, a survey of the state of global happiness. There are happiness scores and rankings that use data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. Nevertheless, it seemed appealing to us to enrich the data. In the end, each of the 155 rows of the dataset is a country measured with certain variables, where we have added the Human Development Index, the Global Peace Index and the Suicide Rates for each of the countries.

At the end we have 12 columns -which are the variables- and 155 rows -which are the countries-. We want to predict the happines score for countries based on some meassures. Now we are going to describe the variables that we have considered for the study.

2.1 Response variable

Our response variable is the Happiness Score for each country. In the original dataset, it was defined as the sum of the numerical variables: GDP, Family, Life Expectancy, Freedom, Generosity, Government Corruption and Dystopia Residual (those variables will be explained later). Nevertheless, it should be remarked that we added some new variables so that we know it is not a direct linear combination. The Happiness Score is a continuos variable also quite synmmetric, as we can see in the histogram below.



2.2 Predictors

GDP: It is a quantitative continuous scaled variable. It represents the GDP per capita, which meassures the market value of all final goods and services produced in a year per capita.

Family: It is a continuous variable. It represents the social support, as measured by having someone to count on in times of trouble.

Life expectancy: It is a continuous variable and it represents the life expectancy of each country.

Freedom: It is a continuous variable. It measures the freedom to be able to make life decisions.

Generosity: It is a continuous variable. It measures recent donations made by the individuals of the country.

Government corruption: It is a continuous variable. It measures the perceived absence of corruption in government and business.

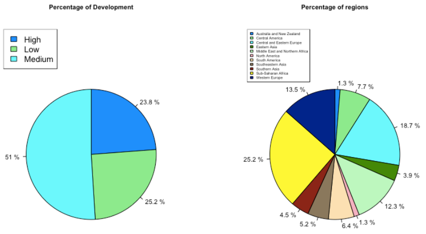
Dystopia residual: It is a continuos variable. We could say that this variable is reserved for other variables that could have influence in happiness beyond the ones we have defined. Somehow, the creators of the original dataset decided to weight more some countries that had low marks in the other variables because they thought in these countries people are also influenced by other factors non present in the data.

Peace: It is a continuous variable, that measures the peace for each country. It takes into account several factors such as country's militarization, violent crimes and political instability. The higher this variable is, the less peace score a country has. For instance, the country Syria has a peace value of 3.814, while Iceland has 1.111.

Suicide: It is a continuous variable that represents the mean suicide daily rates (counted in year 2015) of each country.

Development: It is a qualitative categorical variable that represents the level of development for a country. It can take the values high, medium or low. As we can see in the pie chart, we have a higher number of countries with medium value.

Region: It is a qualitative categorical variable that represents regions of the world. As we can see in the pie chart, we have more number of countries that belongs to the Sub-Saharan Africa region.

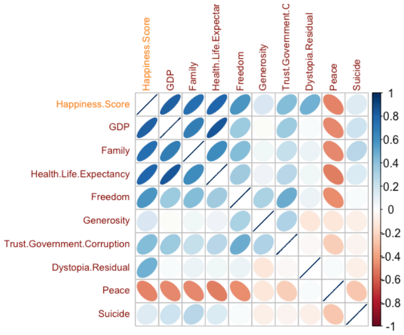


Pie chart for development and region

Variables	GDP	Family	Life Expectancy	Freedom	Generosity	Corruption	Dystopia residual	Peace	Suicide
Mean	0.98	1.19	0.98	0.4	0.25	0.12	1.85	2.1	10.45
Variance	0.18	0.18	0.08	0.02	0.02	0.01	0.25	0.28	43.98
Min	0	0	0	0	0	0	0	1.11	1.4
Median	1.06	1.25	1.06	0.44	0.23	0.09	1.83	2.02	9.2
Max	1.87	1.61	1.87	0.66	0.84	0.46	3.18	3.81	36

2.3 Correlation between variables

If we want to define a linear model, it is very important to analyse the linear correlation between the variables. The following corrplot -which represents the correlations with ellipses- gives us an idea of which are the linear relations between the predictors and the response variable.



With the graphic we can see that some variables are highly correlated. This could gives us problems in terms of multicollinearity, due to those variables that share information. A dimension reduction technique could be an effective way to manage our dataset.

3 Methods

In this section will be explained most of the theory used. First of all, building a probabilistic model consists basically in finding a function f such that

$$Y = f(X_1, ..., X_p) + \varepsilon$$

being $X_1, ..., X_p$ random variables called predictors, Y a random variable called response and ϵ a random variable with mean zero. That is why probabilistic models have a deterministic component $E(Y)$, plus a random component ε. In this course the aim is to study probabistic models in which f is a linear function. Reformulating the problem:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Now our objective is to try to realise if exists a linear relation between the response varible and the predictors, and if it is true, estimate the coeficients in the best possible way. The varibles $X_1, ..., X_p$ get values in n individuals. Let's call X the matrix with n rows and $p + 1$ columns, the first one with only ones and the following ones with the values of the n individuals in each of the p variables, β a vector with the $p + 1$ coefficients. Let's also consider ε the vector with the n errors. We can write the matrix expression:

$$Y = X\beta + \varepsilon$$

So the question now is which is the best criteria to calculate the deterministic part of the model, i.e., β vector. What happens if we try to minimize the square of the errors?

$$RSS(\beta) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2 = (Y - \beta X)'(Y - \beta X)$$

We easily can obtain the minimum of this function:

$$\frac{\delta RSS}{\delta \beta} = -2X'(Y - X\beta) = 0 \Rightarrow X'X\beta = X'XY \Rightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

The following hypothesis are key if we want know if this $\hat{\beta}$ is "the best" to stimate the random variable β.

1. **Linearity**: $E(\varepsilon \mid X) = 0 \Rightarrow E(Y \mid X) = X\beta$.
2. $Var(\varepsilon \mid X) = Var(Y \mid X) = \sigma^2 I$, implying $Var(\varepsilon_i) = \sigma^2$ for each i (**homocedasticity**) and $Cov(\varepsilon_i, \varepsilon_j) = 0$ for each $i \neq j$.
3. X has rank iqual to $p + 1$, meaning that there is **no multicolinearity** between the predictors. This fact implies that we can calculate $(X'X)^{-1}$.

With this three assumptionswe can prove that $\hat{\beta}$ is unbiased (for this particular one we only need linearity), $Var[\hat{\beta} \mid X] = \sigma^2(X'X)^{-1}$ and also that $\hat{\beta}$ is the unbiased estimator with minimum variance, so that we know that this particular method of minimizing RSS gives us "the best" estimator. The unbias of the estimator implies an interesting consecuence:

$$MSE[\hat{\beta}_j] = E[(\hat{\beta}_j - \beta_j)^2] = (E[\hat{\beta}_j] - \beta_j)^2 + Var[\hat{\beta}_j] = Var[\hat{\beta}_j]$$

There is an extra assumption, called **normality in the residuals**: $\varepsilon \mid X \sim N(0, \sigma^2 I) \Rightarrow Y \mid X \sim N(X\beta, \sigma^2 I)$. This assumption is key for the inference of the model. Two interesting things that we could do could be inference on the coefficients or inference in the predictions. For the inference on the coefficients we can test the null hypotesis $H_0 : \beta_j = 0$ either or create confidence intervals with the formula $\hat{\beta}_j \pm t_{n-p,\alpha/2} S \sqrt{a_{jj}}$, being a_{jj} the j element of the diagonal of $(X'X)^{-1}$ matrix. For the inference on the predictions we have the confidence intervals $\hat{y}_i \pm t_{n-2/\alpha/2}\sqrt{x_i'(X'X)^{-1}x_i}$ (for the conditional mean) and $\hat{y}_i \pm t_{n-2/\alpha/2}\sqrt{1 + x_i'(X'X)^{-1}x_i}$ (for the conditional response). Inference is also important for more things, for example for determining if the model obtained is globally significative, i.e., if $\beta_1 = ... = \beta_p$ versus any $\beta_j = 0$. In this case we should use what is called ANOVA. Another example is the coefficient of determination R^2, which is the proportion of variance explained by the model, and gives us a lot of information of how good is a model.

While building a model is also very important to take into account the complexity of the model. For an especific problem it might be better a linear model with a R^2 lower but with less predictors, because having less predictors makes our model easier to interpret. We may be focused on understanding a particular problem and how the predictors affect the response variable rather than predict with a lot of accuracy the response variable. There are some criterias that help us to understand if adding more varibles to our model worths or not. Two of them are:

Bayesian information criterion (BIC). This criterion consideres the number of individuals n, the number of parameters estimated by the model, k (in the case of linear models always $k = p + 2$), and $\hat{L}$, the maximum value of the the likelihood function of the model. So if we take a look at the formula, when the BIC takes low values it means that the complexity is low and the fitness is high, so the model is better.

$$BIC = \underbrace{ln(n)k}_{Complexity} - \underbrace{2ln(\hat{L})}_{Fitness}$$

Akaike information criterion (AIC). This criterion is very similiar to the previous one. We could say that the only difference here is that we are not considering the number of individuals as part of the complexity of the model. When the AIC takes low values it means that the complexity is low and the fitness is high, so the model is better.

$$AIC = \underbrace{2k}_{Complexity} - \underbrace{2ln(\hat{L})}_{Fitness}$$

These two criterias could help us if we want to decide the number of predictors we want for our model, but, are there more ways to decide which variables are the best for predicting the response variable? One answer could be dimention reduction. Apart from the known technique PCA -which maximizes the explained variance in each component- there is a dimention reduction technique called **Partial Least Squares (PLS)**. The main difference is that PCA is unsupervised and PLS is supervised. PLS uses the response variable in order to identify new features that not only approximate the old features of the dataset well, but also that are related to the response.

Now the objective is to describe how we can calculate these components. PLS components are also linear combinations of the original predictors, so the explanation can be reduced to how to calculate the coeffcients of the linear combinations. First, what we need to do is to standardize the predictors, then calculate the first component, PLS_1 as follows: each coefficient φ_{j1} is equal to the coefficient from the simple linear regression of the response Y onto the predictor X_j. We are placing the highest weight on the variables that are most strongly related to the response.

To calculate the second component PLS_2 what is done first is taking the residuals of each predictor X_j regressed onto the previous component PLS_1. These residuals can be interpreted as the remaining information that has not been explained by the first PLS direction. Now we are able to compute the second component being the coefficients φ_{j2} the coefficients from the simple linear regression of the response Y onto the residuals ε_j. And continuing this process we are able to calculate each of the p PLS components. In a practical approach, we should always keep in mind the percentage of variance explained by each of the components and also how complicated are them, meaning that maybe is not useful for us working with linear combination of the predictors instead of the predictors themself. We can work with cross-validation and chose the number of partial least squares directions using that as a tuning parameter of the model.

Talking about cross-validation, this technnique is very useful in order to check weather a model is good or not. But, we have explained how to check it, have not we? The question is, what happens if one of the four key assumptions is not verified? In this case we can not use the inference as a tool for verifying if the model predicts well or not. We have to move from this theoretical approach to a practical one, cross-validation.

A more generalized algorithm is k-fold **cross-validation**. Basically, what is done is a random division of the dataset in k folds, and for each of the k folds we build the desired model with the remaining $k-1$ folds, using the chosen fold for testing the model. In the case of linear regression we can use this technique for testing k MSE, compare them and decide how good is our model.

4 Statistical analysis

In this part, we are going to build a proper linear model for our dataset using the theory explained in the previous section.

A priori, we used the whole dataset, excluding qualitative variables, to build a **linear model**. This model included the variables: GDP, family, life expectancy, freedom, generosity, government corruption, dystopia residual, peace and suicide. As it is shown in the correlation plot on the introduction section, these variables are highly correlated. Moreover, we decided to study the multicolinearity with the Variance Inflation Factor (VIF). These were the results:

Variables	GDP	Life Expectancy	Family	Freedom	Peace	Corruption	Generosity	Suicide	Dystopia residual
VIF	4.633	3.778	2.190	1.820	1.653	1.510	1.246	1.171	1.057

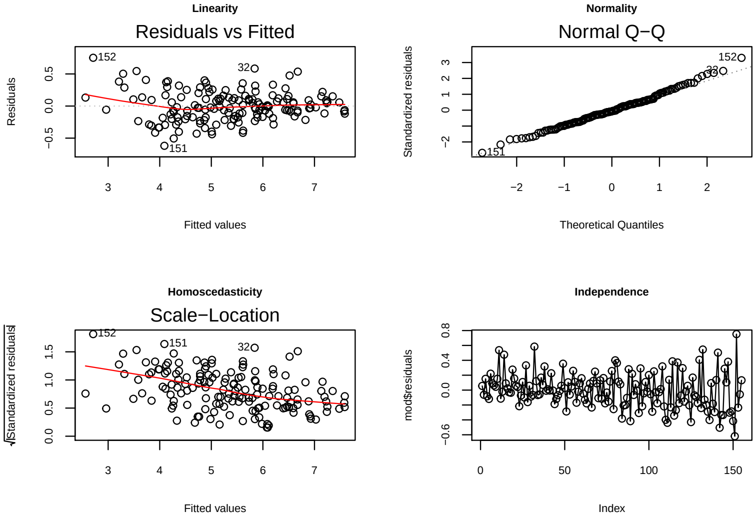
We consider that the variables GDP and life expectancy have a VIF higher enough to assume multicolinearity. Due to these two facts, we made the decision of working with a transformed dataset in which there is no multicolinearity. For this purpose, we used Principal Component Analysis (PCA), Sparse PCA and Partial Least Squares (PLS). Furthermore, these known techniques could be quite useful in order to simplify the model.

Firstly, we have applied **PCA** to our dataset without the response and qualitative variables. Looking at the cumulative proportion, we need at least four principal components to explain the 75% of the variance of our dataset. Nevertheless, we are going to use from one to four components to create the respective models with their Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) parameters.

PCA	Comp.1	Comp.1+Comp.2	Comp.1+Comp.2+Comp.3	Comp.1+Comp.2+Comp.3+Comp.4
AIC	257.5	257.2	26.6	3.6
BIC	266.6	269.4	41.7	21.9

In this table we can see how the value for AIC and BIC decrease when we add more and more components. If we only take into account these values, we would choose a linear model using PCA with four components. We have also checked that we obtained the same estimation and standar error for AIC and BIC, so it's indifferent to choose one of those.

Before continuing with other methods, we should ask ourselves: what is happening with the assumptions that the model has to verify? With the graphs and also with the formal tests, we can check if the model constructed by PCA with four components satisfies the assumptions. As we can see, the model fulfills almost all of them, except homoscedasticity. If we do Breusch-Pagan test we reject the null hypothesis, so we have statistical evidences that our residuals are not homoscedastic. On the other hand, with the Shapiro test and the Durbin-Watson test we do not reject the null hypothesis, so we can assume that we have normality on the residuals and independence between the variables.



In addition, we have analysed the VIF and the outliers for this model and we have seen that we don't have outliers and the values for the VIF are always 1 as we are using PCA.

We have also tried to transform our dataset with the **Sparse PCA** method but we realised that the new coefficients of the sparse PCA components did not simplify the previous PCA components. For example, $PCA_3 = 0.228 Generosity - 0.809 Dystopia.residual - 0.159 Peace + 0.489 Suicide$ and $SPCA_3 = -0.056 GDP - 0.011 Family - 0.084 Life.expectancy - 0.088 Freedom + 0.230 Generosity - 0.057 Government.corruption - 0.803 Dystopia.residual - 0.172 Peace + 0.500 Suicide$.

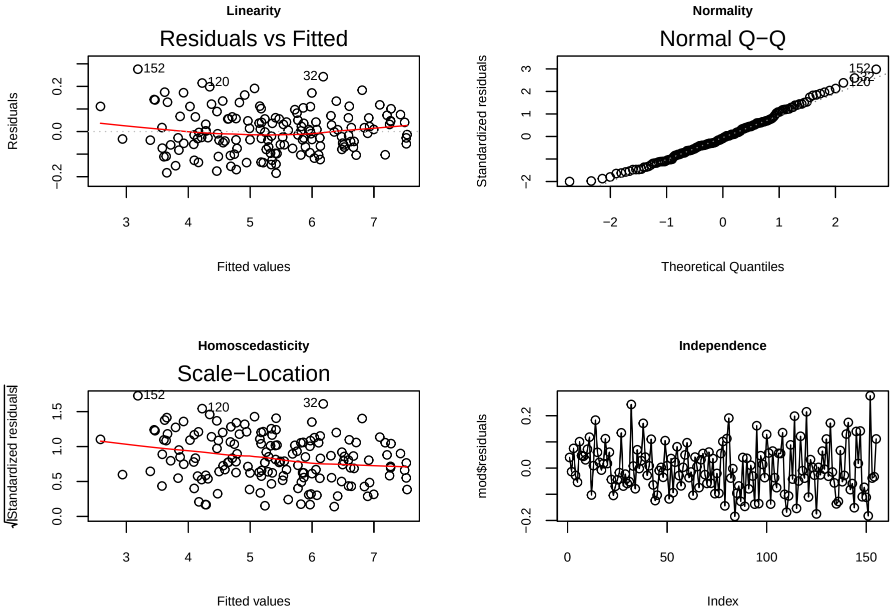
Secondly, we have applied **PLS** to our dataset without the qualitative variables. In this case, we can explain 72% of the variance with four components. However, we are going to use from one to four components to create the respective models with their BIC and AIC parameters.

PLS	Comp.1	Comp.1+Comp.2	Comp.1+Comp.2+Comp.3	Comp.1+Comp.2+Comp.3+Comp.4
AIC	147.3	-287.8	-554.9	-627.2
BIC	156.5	-275.6	-539.7	-608.9

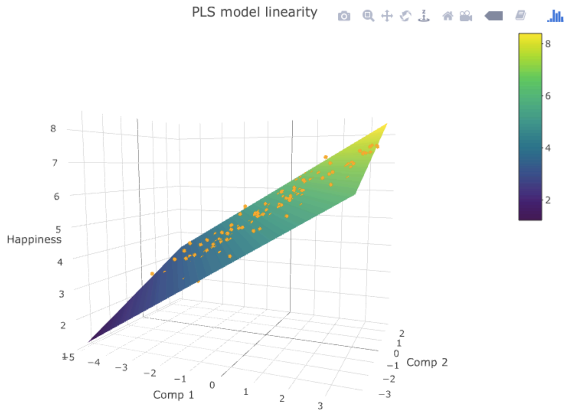
Analogously to PCA, we can see how the value for AIC and BIC decrease when we add more and more components. If we only take into account these values, we would choose a linear model using PLS with four components. We have also checked that we obtained the same estimation and standar error for AIC and BIC, so it is indifferent to choose one of those.

We asked ourselves the same question as before about the model diagnostic. After analysing the hypothesis with four and three components, we have made an study with two components. Our conclusion of this study is that the accepted model is a PLS with two components, because with three and four components independence, normality and homoscedasticity assumptions have not been validated. Notice that now we explain the 49.74% of the variance of our dataset.

Now we are going to evaluate the hypotheses of the model with two components, with the graphs and the formal tests used previously.

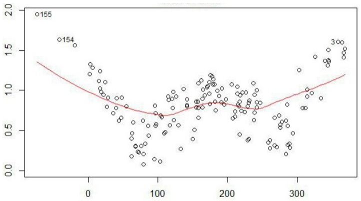


Linearity: As we can see in the first graph, there is barely any trend in the residuals with respect to the estimates of the happiness score. In the following graph, we have represented the two components with the response variable to show the linearity of the model.



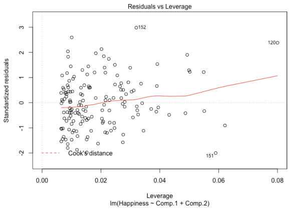
Normality: Observing the qq-plot, the quantiles of the standardized residuals follow a normal distribution. We observed that there are some departures from the diagonal in the extremes. There are formal tests to check the null hypothesis of normality in our residuals. After examining the p-values for Shapiro test -0.15- and Lilliefors (Kolmogorov-Smirnov) test -0.42-, we do not reject the null hypothesis for both tests, so there are statistical evidences that our residuals are normal.

Homoscedasticity: In this hypothesis we have some problems. Studying the graph we can appreciate heteroscedasticity in the residuals. That could make that the confidence intervals for prediction could be wider or shorter than the adequate ones. We have evaluated as well the p-value for the Breusch-Pagan test and we reject the null hypothesis, which means that we do not have homoscedasticity. To solve this problem we have tried the transformations $\log Y$ and $\sqrt{Y}$ in the response variable without success, because we also rejected homocedasticity -with a lower p-value in Breusch-Pagan test-. So we tried the other way, to enlarge the scale of Y with the transformation Y^3, obtaining a large p-value in Breusch-Pagan test and cosequently do not rejecting homocedasticity, but seeing later in the graph that our model has a clear pattern of heteroskedasticity (the test is blind in this situation). **We will deal with heteroskedasticity in the model**.



Independence: The independence is something very important to analyze, because it makes maximal the amount of information obtained in the model, without duplicated information. Observing the graph and the result of the Durbin Watson test, which has a p-value of 0.25, we can say that the model has independency.

In addition, we have also analysed the VIF and the outliers for this mode. With the plot we can see that we don't have outliers and the values for the VIF are always 1 as we are using PLS.



Now, we are in a complicated situation in which we have to choose between two models that verify the same hypotheses. In PLS we only have two components rather than in PCA where we have four. Moreover, let's take a look at both linear transformations:

$PLS_1 = 0.481GDP + 0.448Family + 0.467Life.expectancy + 0.366Freedom + 0.282Government.corruption + 0.108Dystopia.residual - 0.373Peace + 0.128Suicide$

$PLS_2 = -0.114Freedom - 0.173Generosity - 0.139Government.corruption + 0.873Dystopia.residual + 0.340Peace - 0.303Suicide$

$PCA_1 = -0.464GDP - 0.432Family - 0.448Life.expectancy - 0.362Freedom - 0.114Generosity - 0.284Government.corruption + 0.387Peace - 0.151Suicide$

$PCA_2 = 0.208GDP + 0.193Family + 0.173Life.expectancy - 0.3700Freedom - 0.608Generosity - 0.464Government.corruption + 0.110Dystopia.residual + 0.392Suicide$

$PCA_3 = 0.228Generosity - 0.809Dystopia.residual - 0.159Peace + 0.489Suicide$

$PCA_4 = -0.318GDP - 0.317Life.expectancy + 0.265Freedom + 0.212Generosity + 0.510Dystopia.residual - 0.236 * Peace + 0.599Suicide$

As we can see, PLS linear combinations are simpler than PCA linear combinations due to the number of zeros. This fact could make the linear model with the PLS easier to interpretate. Since the aim of this project is to give a social interpretation of the data, we decided to choose interpretability versus complexity, so we chose PLS with two components. It should be remarked that we have tried to include our qualitative variables as dummy variables in order to split the predictions into different groups, but we have not obtained more useful information from the new models.

Due to the fact of heteroskedasticity, inference is not valid in the model. Even if we try to analyse if the coefficients are significative -lm function of R says they are- we could make important errors. Inference is very useful in linear models (see Methods section), but in our case does not make sense to perform it. So what we need to do if we want to know how good the two components of PLS are at predicting countries' happiness score? **Cross validation** is the answer. We performed that manually as it follows.

We split the dataset in five test sets, one set contains the rows whose indexes modulus five equals zero, another set where the indexes modulus five equals one, another where the modulus equals two, and so for the five sets. This way we have our test sets evenly builded as the dataset is ordered from highest happiness score to lowest one. Following this step, we also created a training set for each test set with the indexes that where not used in the corresponding test set. Finally, we kept in a separated variable the real scores for each fold so we can use them later to calculate the MSE. Below are shown the results of the predictions of crossvalidation:

$$MSE: \ 0.0199, \ 0.0113, \ 0.0045, \ 0.0116, \ 0.0067$$

As we can see, the MSE values are low so we can confirm that PLS with two components offers an accurate prediction.

Furthermore, we would like to present a **social explanation** of the model obtained in this report. Previously we saw that the linear combinations of PLS were:
$PLS_1 = 0.481GDP + 0.448Family + 0.467Life.expectancy + 0.366Freedom + 0.282Government.corruption + 0.108Dystopia.residual - 0.373Peace + 0.128Suicide$
$PLS_2 = -0.114Freedom - 0.173Generosity - 0.139Government.corruption + 0.873Dystopia.residual + 0.340Peace - 0.303Suicide$

Also, we can express the final model like this:

$$Happiness.Score = 5.35 + 0.6Component.1 + 0.35Component.2$$

Analysing the previous operations, we can tell that component 1 differentiates between socioeconomic factors such as GDP, family, life expectancy and freedom scores; and the country's state of peace, taking into account that the higher the country's peace index, the less peaceful the country is.

The social interpretation that we can present is that countries with a strong economy where people has a healthy and free life and has social support coming from their family and friends are more likely to achieve a happiness status, while the country's state of peace plays an important role, since the higher it is the more it harms the happiness state of people from that country. On the other hand, violent countries with a really high state of peace and low economic and social factors will have its happiness score affected negatively.

Additionally, component 2 differentiates dystopia residual and peace factors from the freedom, generosity, absence of government corruption and suicide factors. Freedom, generosity and corruption scores take low values (less than 1) therefore, independent of their values, they are not going to have a really high impact on the linear combination of component 2 so we are going to omit these values in the social interpretation of this component.

Suicide is the attribute that takes the highest values so maybe component 2 expresses that suicide rate is a really important factor that can affect the happiness score unfavourably. For example, there might be countries where the state of peace is really peaceful and have a high score of unknown happiness factors evaluated by experts that are not explained by the model that built the dataset (dystopia residual), but because the country has a high rate of suicides, its happiness score is affected adversely.

Lastly, it is clear that component 1 has more weight than component 2 in the calculation of the happiness score, so the state of peace and the economic and social factors are more important elements than the suicide rate of the country.

To conclude this section, we wanted to verify that our fitted model can **predict** new countries' happiness score. To do so we created three imaginary countries where their atributes were generated with random variables, so we could see how well it performed. The first country has good attributes, the second has mediocre attributes and the last one has poor attributes. Here are the attributes and the predicted happiness score for each country:

Nation	GDP	Family	Life	Freedom	Gen.	Corrupt.	Dystopia	Peace	Suicide	Happiness
Good	1.616	1.375	0.799	0.608	0.346	0.277	2.616	1.125	2.9	7.298
Normal	0.092	0.898	0.290	0.379	0.249	0.093	1.482	2.052	7.4	5.881
Violent	0.0316	0.000	0.152	0.121	0.194	0.0566	0.683	3.413	22.3	4.661

We can see that for the first country, the one with the best attributes, it obtains a high happiness score as expected. The second country obtains also a high score but probably it is due to the fact that it has pretty good values for the socioeconomic attributes and not really high peace and suicide rates so the model still considers it a happy country. Finally, it predicted a 4.7 score for the last country, the worst one. This can be explained by the fact that the dataset contains only a few countries with scores below 3.5 so it is possible that it needs more countries of this type to be able to predict them correctly.

5 Conclusions

- We decided to choose PLS with two components to gain interpretability in order to make a social explanation of the model.This model fulfills all the hypotheses, except homoscedasticity so inference is not valid in our model.
- MSE values obtained with crossvalidation are really low, so we justify our model empirically.
- The response variable, happiness score, is affected positively by good socioeconomic factors such as GDP, family and friends support, healthy life expectancy, freedom and absence of government and business corruption, while the suicide rate and the state of peace affect unfavourably.

6 References

- Sustainable Development Solutions Network. World Happiness Report. Happiness scored according to economic production, social support, etc. Recovered from: https://www.kaggle.com/unsdsn/world-happiness/data
- John Helliwell, Richard Layard and Jeffrey Sachs. World happiness report 2017. Recovered from: http://worldhappiness.report/wp-content/uploads/sites/2/2017/03/HR17.pdf
- Gareth James, Daniela Witten, Trevor Hastie y Robert Tibshirani.(2015). An Introduction to Statistical Learning with Applications in R.
- Wikipedia. Developing country. Recovered from: https://es.wikipedia.org/wiki/Pa%C3%ADs_en_v%C3%ADas_de_desarrollo
- Institute for Economics and Peace. Vision of Humanity. Global Peace Index. Recovered from: http://visionofhumanity.org/indexes/global-peace-index/
- World Health Organization. Suicide rates, crude data by country. Recovered from: http://apps.who.int/gho/data/node.main.MHSUICIDE?lang=es