Cross validation

Urša Zrimšek

Introduction

Our goal is to build a couple of models that estimate how various features influence happiness. We will train them on a part of datasets available on Kaggle - data from years 2017, 2018 and 2019. We will first analyse the data, to figure out what should our models be, and then use LOOIC to determine their quality and Akaike weights and LOOIC to gain additional insight into how the models compare against each other and how we would weigh decisions of models if we were to combine them together.

Data preparation and model choice

In this section we will describe how we prepared the data and how we analyzed it to decide on which models to build for further testing.

Data preparation

We build our dataset by combining the dataset prepared directly for this homework with the Kaggle datasets. To the existing ones: **Year**, **Country**, **Score** (happiness score that we are predicting, the higher it is more happy the populants of a country are), **GDP** (GDP per capita) and **Corruption** (how populants of a country think of corruption in it, there 0 denotes no corruption, while 1 denotes maximum corruption); we added additional columns: **Region** (region the country belongs to - 10 unique regions), **Life** (health / life expectancy), **Freedom** (perception of freedom), **Trust** and **Generosity**. The last 4 columns are not some direct measures of of these parameters, but describe the extent to which these factors contribute in evaluating the happiness in each country (all positive). More information about the parameters can be read in the description of Kaggle datasets.

Data analysis

The mentioned description also explains that the data has an addition column, called Dystopia Residual metric, that is actually the Dystopia (a benchmark imaginary country that has the world's least-happy people) Happiness Score(1.85) + the Residual value or the unexplained value for each country. First thing we did, we looked at these residuals, and saw that they are approximately normaly distributed, so we decided that all our models will also use normal distribution.

Our next question was which of the above features we should use. We plotted their densities grouped by years, to confirm there are no discrepancies between them and that the distributions are nice. To decide which affect happiness score the most and how they are connected, we also calculated correlations and scatter plots between them, all can be seen on figure 1. From this figure we can see that the **most correlated with happiness score is GDP and close second is Life**, but they are also quite highly correlated with each other. Since we won't deal with interpretation of our models (it would be really hard in this case, considering different scales and completely different meanings of our features), this shouldn't be a big problem, and we will use both. Also we can immediately **remove Trust** as it is just a negative of Corruption and Generosity because of almost no correlation. We also checked correlation with some of the interactions, and found high correlation with **GDP** $\cdot$ **Life** (0.83) and with $\frac{\textbf{GDP}}{\textbf{Corruption}}$ (0.8). We put Corruption in the denominator because it has a negative impact on happiness.

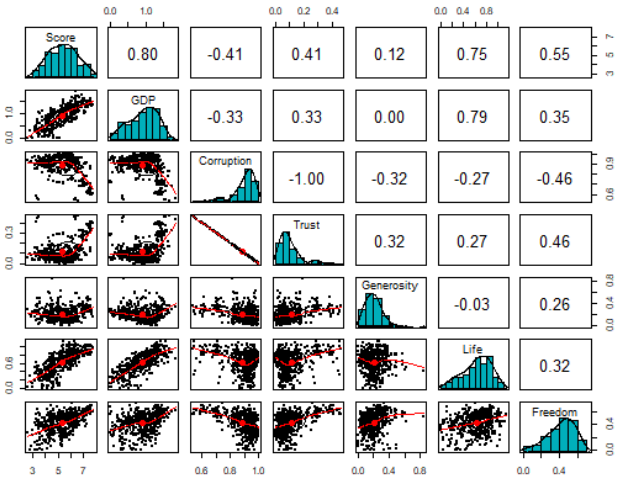

Figure 1. **Correlations, distributions and scatter plots of features.** In the top row we can see all correlations with happiness score, and in the first column scatter plots of all independent parameters with it. Notice that *Corruption* and *Trust* are just negatives of each other.

Models

Let's describe the models we chose based on data analysis and prior knowledge about the domain. To simplify further equations we define x_g, x_c, x_l, x_f as values of features for one datapoint, indices corresponding with first letter of above selected features.

Linear: Our benchmark model is simple **linear regression** model, that uses **GDP** as independent variable:

$$y \sim N(\beta_0 + \beta_g x_g, \sigma^2).$$

We chose GDP because of highest correlation with the target. We can also see from the scatter plot that the linear model should be the right choice.

Multilinear: Next was **multilinear regression** model, that uses **GDP, Life, Corruption** and **Freedom** as independent variables:

$$y \sim N(\beta_0 + \beta_g x_g + \beta_c x_c + \beta_l x_l + \beta_f x_f, \sigma^2).$$

We chose all parameters that we described above.

Regions: This is also a **multilinear regression** model, that besides all above parameters also uses the **Region** column. We didn't analyze data to see how Regions are connected to happiness, but we know that there is similar situation in multiple nearby countries, so we should also build models that use them. We have 10 distinct regions, that's why we one-hot-encoded them into 10 columns. We can describe this model as:

$$y \sim N(\beta_0 + \beta_g x_g + \beta_c x_c + \beta_l x_l + \beta_f x_f + \sum \beta_{ri} x_{ri}, \sigma^2),$$

where x_{ri} are zeros and ones corresponding with region of point x.

Interactions: This is the same model as *Multilinear*, but with above mentioned interactions included:

$$y \sim N(sum_{multilin} + \beta_{gl} x_g x_l + \beta_{gc} \frac{x_g}{x_c}, \sigma^2).$$

where $sum_{multilin}$ is the sum written in Multilinear model.

Inter_regions: This is again the same model as *Regions*, with the same interactions included:

$$y \sim N(sum_{regions} + \beta_{gl} x_g x_l + \beta_{gc} \frac{x_g}{x_c}, \sigma^2).$$

where $sum_{regions}$ is the sum written in Regions model.

log_GDP: For this model we used our prior knowledge about the domain. We know that the happiness rises fast with income if people are on the brink of poverty. But as income is above the line of comfort life, it doesn't bring as much additional happiness anymore. We tried modeling this with logarithmic function:

$$y \sim N(\beta_0 + \beta_g \log(x_g + c) + \beta_c x_c + \beta_l x_l + \beta_f x_f + \sum \beta_{ri} x_{ri}, \sigma^2),$$

where c is a horizontal shift of the logarithmic function. It tells us the starting point of the logarithmic function. Without it we also had negatively infinite values and the model didn't converge, so we needed to set at least some small value to it. We didn't know what it should be, so we let the model infer it. We only limited it between 0 and 2, so that the model did converge.

To infer model parameters we used MCMC with 500 warmup iterations and 1000 sampling iterations. We didn't use any priors, to be sure we don't brake Akaike constraints. We also checked that the posterior distributions are approximately multivariate Gaussian and we know that we have much more datapoints than parameters in all the models above. We also made sure that the diagnostics of all chains are ok - we checked traceplots, $\hat{R}$ and effective sample sizes.

Model selection

To select the best model out of the above, we used the state-of-the-art Bayesian cross validation approximation technique, **Leave one out information criterion (LOOIC)**. It can be seen on figure 2.

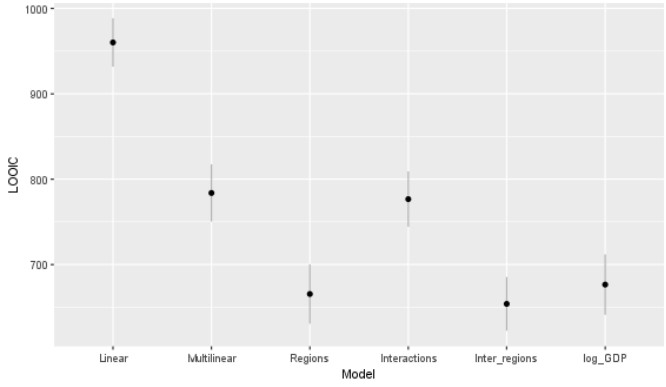

Figure 2. **LOOIC of our models.** Model that uses interactions and regions has the lowest mean, but the standard errors overlap with those of the model with only regions and with the logarithmic one.

We didn't include exact numbers, as the visualization is more informative and quicker to understand, as we only want to compare the models with each other. To further compare them and decide how we should combine them, we calculated **Akaike weights** from LOOIC, and found out that we get **weight 1 for** model **Inter_regions**, so this should be the only model we base our decision on. We also calculated *AIC* (seen on figure 3) and *WAIC* (exactly the same as LOOIC) and the results were the same.

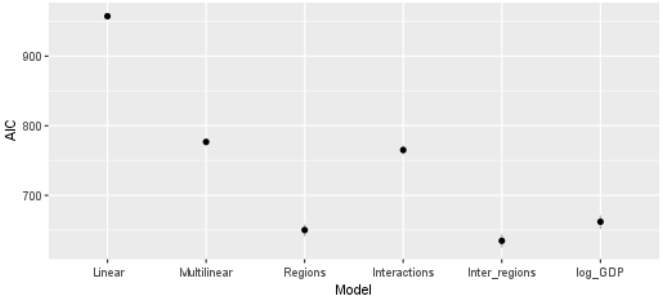

Figure 3. **AIC of our models.** We can see that means are similar to the ones of LOOIC. We showed uncertainty with 90% HDI, that are much smaller than SE in LOOIC. There is no intercept between HDI of best model ([625, 644]) and the log_GDP model ([652, 671]), and very little with Regions model ([641, 658]).

Discussion

Based on LOOIC we found that the **model we should use is Inter_regions**. Based on the models that are the best, we can conclude that regions are important feature to predict the country populants happiness, and that GDP is not the only feature that is important, besides the region. Our hypothesis about logarithmic impact of GDP didn't prove to be useful, but maybe it should be researched further with some other basis.