Cross Validation

Žiga Trojer 63200440

I. Introduction

The goal of this assignment was to create at least three models that estimate how different features influence happiness score. We were given a happiness dataset with data from the World Happiness from year 2017 to 2019. We attempted to model happiness score and then compare the models to determine which one is the best.

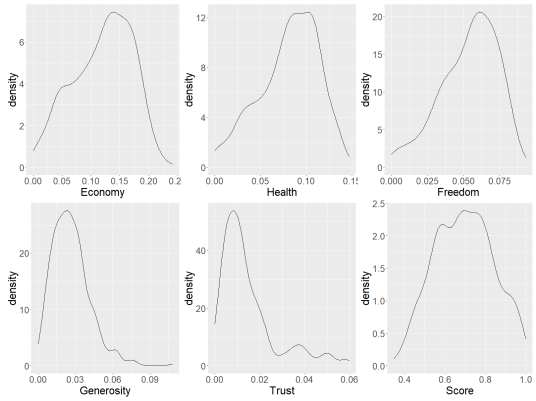
II. Methods

For each year between 2015 and 2019 we collected data from Kaggle Dataset. We first processed the data because the data for the years did not match or had a different number of features. The columns in the merged dataset are as follows:

- *year* - year of the survey,
- *economy* - the extent to which GDP contributes to the calculation of the *score*,
- *health* - the extent to which life expectancy contributes to the calculation of the *score*,
- *freedom* - the extent to which freedom contributed to the calculation of the *score*,
- *generosity* - the extent to which generosity of people contributes to *score*,
- *trust* - the extent to which perception of corruption contributes to *score*,
- *region* - region the country belongs to,
- *score* - a metric measured by asking the sampled people the question: "How would you rate your happiness?".

We kept only the data from 2017 onwards because the data for 2015 and 2016 did not include column *generosity*. The combined dataset contains 7 independent variables and 1 dependent variable (*score*). Figure 1 depicts the distribution of the previously described features. During the modeling, we created new datasets with only a subset of the features or adjusted the dataset as needed - when we used the *region*, we used one-hot-encoding (and removed column for one region due to potential multicollinearity). We also used min-max scaling to ensure numerical feature stability. We made certain that the dataset did not contain any empty values at all times.

Figure 1: Distribution of *economy*, *health*, *freedom*, *generosity*, *trust* and *score* in the merged dataset.



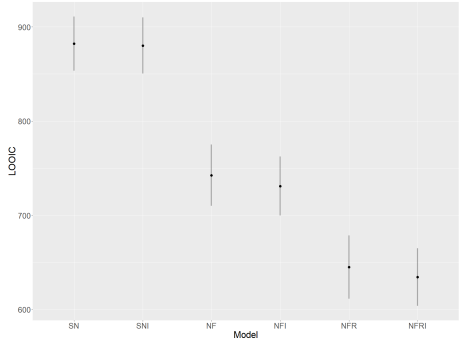
We then plotted the correlation of independent variables and discovered that it was sufficiently low. We used six different models, which are listed in Table I. We then fitted models using four Markov Chains and 500 iterations for warm-up and sampling. For each model, we ran a standard diagnostic procedure to ensure that all chains had appropriate trajectories, that $\hat{R}$ was close to 1, and that the sample size was large enough. A few chains did not converge in the last two models, but because this occurred during the warm-up phase, we ignored the errors.

III. Results

We can now look at the LOOIC graph on Figure 2. It clearly shows how adding more features affects the model's quality. When we compare the first two models, which use only two independent features (*economy* and *trust*), to the models that use all of the features, we see that the model improves when all of the features are used. This is not surprising given that additional features contain more information. We're also curious about how mixed terms, or the interaction of two features, affect model quality. When we compare the models in pairs (SM with SNI, NF with NFI, and NFR with NFRI), we see that when we model feature interactions, the LOOIC decreases.

Then we took a look at the Akaike weights for each model. They suggest that if we had to pick just one model or combine several models, we would go with the NFRI model. This makes sense because the NFR and NFRI models are similar, with NFRI increasing model's complexity by introducing new interactions.

Figure 2: LOOIC for each model (mean and SE). Description of the models could be found in Table I.



IV. Discussion

We attempted to find the best model for predicting happiness score based on the features provided in this homework. In our case, we discovered that the model's complexity improves its performance. A model with quadratic terms would be interesting to try, but we believe it would be unnecessary. We also discovered that all of the available features include additional information about the happiness score. This is not surprising given that, as stated in the data description, the happiness score is calculated from all of the given features. Examining the posterior distribution would be pointless because our task was not to discuss or explain what influences the happiness score, but only to choose the best model. We also learned how to easily compare models in a Bayesian manner, which will be useful in our studies and at work.

Table I: Models description.

Model	**Short name**	**Features Used**	**Interaction**	**Link**
Simple normal	SN	*economy* (e), *trust* (t)	without	$\beta_0 + \beta_e x_e + \beta_t x_t$
Simple normal with interaction	SNI	*economy* (e), *trust* (t)	*economy*, *trust* (et)	$\beta_0 + \beta_e x_e + \beta_t x_t + \beta_{et} x_{et}$
Normal with all features	NF	*economy* (e), *trust* (t), *health* (h) *freedom* (f), *generosity* (g)	without	$\beta_0 + \sum_{i \in [e,t,h,f,g]} \beta_i x_i$
Normal with all features and interactions	NFI	*economy* (e), *trust* (t), *health* (h) *freedom* (f), *generosity* (g)	between all of them	$\beta_0 + \sum_{i \in [e,t,h,f,g]} \beta_i x_i + \sum_{j \in \text{mix}} \beta_j x_j$
Normal with all features and regions	NFR	*economy* (e), *trust* (t), *health* (h) *freedom* (f), *generosity* (g) *regions (r)*	without	$\beta_0 + \sum_{i \in [e,t,h,f,g,r]} \beta_i x_i$
Normal with all features, regions and interactions	NFRI	*economy* (e), *trust* (t), *health* (h) *freedom* (f), *generosity* (g) *regions (r)*	between all of them (except regions)	$\beta_0 + \sum_{i \in [e,t,h,f,g,r]} \beta_i x_i + \sum_{j \in \text{mix}} \beta_j x_j$