Introduction

In this homework I will be investigating happiness on behalf of the mayor of our favourite town. She has promised to increase the happiness of her citizens if she is to take office. Being genuine as she is she has hired me to investigate what factors influence happiness and how she can apply this to her town.

Methods

Data Description

The data provided is a subset of the annual Happiness Survey. The data provides the happiness score across different countries across the years 2017, 2018, and 2019. Happiness score is computed by asking approximately 1000 people from each country at what rung of 10 rung ladder they are on. The 10th run represents the happiest possible life and 0 the worst possible life. This is then weighted and used to construct the population-representative national average we have in our data. Happiness is a difficult metric to measure but this measure is reasonable enough for our analysis and is widely used. Also provided are explanatory metrics about the country namely *perceived_corruption*, a measure between 0 and 1 with higher scores representing higher perceived corruption and *economy* is a measure of how much GDP per capita contributes to a country's happiness score.

Data Wrangling

The *country* feature on its own is not very useful for our analysis. However, I created a new feature *continent* from *country* and for modelling this is one-hot-encoded as it has no natural ordering. The feature year was dropped as it is not needed for our analysis. Finally, the features *perceived_corruption* and *economy* were standardised so we can later compare their relative importance.

Exploratory Data Analysis

There are no missing values in our data but there are some missing observations where the data for a country in a given year is not available. This is not an issue for our analysis as we will not be modelling each country individually. There are some outliers as can be seen in Figure 2, but these should not cause any issues. In figure 1, we can see that the *continent* feature has quite a strong relationship with *score* with the average happiness score varying considerably between continents.

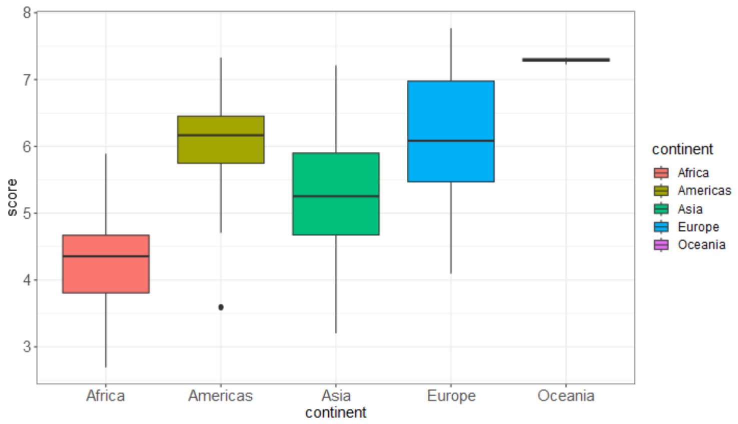

Figure 1. Boxplots of happiness score by continent

Further, in figure 2, we can see continent also has a bearing on how the happiness score is related to our other features. We can see that overall the relationship between *score* and *economy* is positive but the magnitude varies considerably between continents. In Europe increasing *economy* has a greater impact on *score* compared to in Africa for example. For *perceived_corruption* the overall relationship between *score* and *economy* is negative but there are intra-continent differences with the relationship appearing slightly positive for Africa and very negative for Europe. All trends show a linear relationship between economy and happiness score is reasonable and for perceived corruption and happiness score. This suggests that interaction terms will likely prove necessary later but polynomial terms in *economy* and *percieved_corruption* will not.

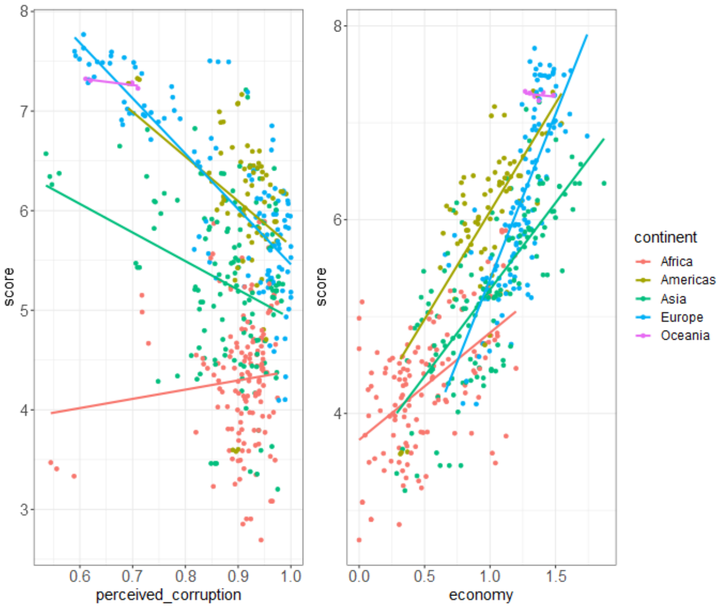

Figure 2. Relationship of corruption and economy to happiness score grouped by continent

Evaluation of the relationships between explanatory variables shows there is a very low correlation between *economy* and *perceived_corruption* and there is some relationship between *continent* and our two continuous features. Therefore, we can safely keep all of these features and use them for modelling.

Models

A variety of models were fitted using Stan however they were all linear models with two main forms. A Normal model where happiness score Y is given by $Y \sim \mathcal{N}(XB, \sigma)$. and a Cauchy model where happiness score Y is given by $Y \sim Cauchy(XB, \sigma)$. In both cases, X is a vector of varying length containing the explanatory features sometimes appended with a 1 for the intercept term. B is a vector of the model parameters representing the contribution of the explanatory variables to the mean happiness score. This flexible formulation allows us to add features to our model by appending to our design matrix. I considered simple models, models with multiple intercepts, and models with interactions across both a Normal and a Cauchy formulation. Finally, to aid convergence and provide regularisation weakly informative priors were used on all beta terms, $B \sim Cauchy(0, 2.5)$.

Diagnostics

Diagnostics statistics of all models were fine (ESS > 400, $\hat{R} \approx 1$). All trace plots indicated no issues with convergence as well. Note for the more complex models, the tree depth had to be increased to 15.

A Note on Probabilities

To distinguish reported Bayesian probabilities from frequentist p-values we denote them with a capital P. Unlike p-values, the reported probabilities directly quantify the degree by which we can claim that our hypotheses are true or not. The probability that the opposite of our claim is true can be calculated as 1 - P. In all cases the uncertainty is reported as the Monte Carlo standard error (MCSE) measure.

Model Comparison

After fitting all of the models I used Leave one out information criterion (”LOOIC”) to assess and compare their fit. A lower LOOIC score indicates a better-fitting model.

In Figure, 3 we can see that including economy in a model substantially decreases the LOOIC as does including the continent. The best model is the most complex model which includes interactions for both economy and perceived corruption with continent. This indicates that there is useful information in perceived corruption for predicting the happiness score too.

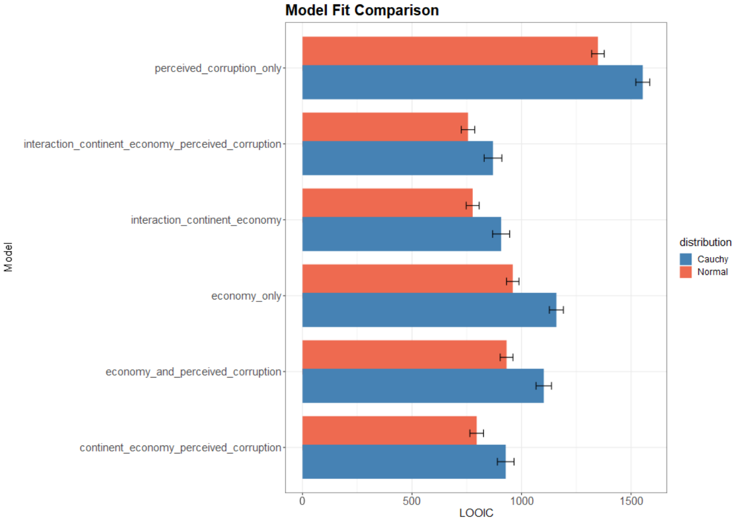

Figure 3. LOOIC of fitted models

Model Choice and Ensemble

I selected the best three models to see if an ensemble approach with LOOIC weight would be suitable. The model with continents and both interaction on the economy and perceived corruption received a weight ≈ 1 and all other models weights ≈ 0. This can be interpreted as with probability P ≈ 1 this model will produce the best estimates given all the models considered. Therefore, we select this model and do not opt for an ensemble.

Final Model

The final model selected by LOOIC has the following explicit form. $y \sim \mathcal{N}(\nu, \sigma)$ where ν the linear predictor is given below.

$$\begin{aligned}
\nu = {} & \mathbb{1}_{Americas}\beta_1 + \ldots + \mathbb{1}_{Europe}\beta_5 \\
& + \mathbb{1}_{Americas}Economy\beta_6 + \ldots + \mathbb{1}_{Europe}Economy\beta_{10} \\
& + \mathbb{1}_{Americas}Percived\_Corruption\beta_{11} + \ldots \\
& + \mathbb{1}_{Europe}Percived\_Corruption\beta_{15}
\end{aligned}$$

Each level of the continent is represented in the intercept terms and the interaction terms with economy and perceived

Results

The following discussion is based off the final model selected. Since the final model is able to distinguish between continents I will focus my analysis on the Americas continent since we are primarily focused on happiness in an American town.

Model Coefficients

In Figure 4, we can see that the theoretical baseline happiness for a country in the Americas. This coefficient is not the focus of our analysis but allows us to better model the data and get better estimates for our slope parameters. More importantly, we can see the impact of the economy and the perceived level of corruption on happiness. The model suggests that the economy has a positive impact and the perceived level of corruption a negative one. Further, the probability P, that increasing economy has a positive impact on happiness is 1 (±0), whereas the probability P, of increasing an increase in perceived corruption having a negative impact is 0.963 (±0.003). Further, we can look at the relative importance of each feature by computing the probability that the economy has a larger absolute impact than perceived corruption, we concur that this is the case with probability P = 1 (±0.005)

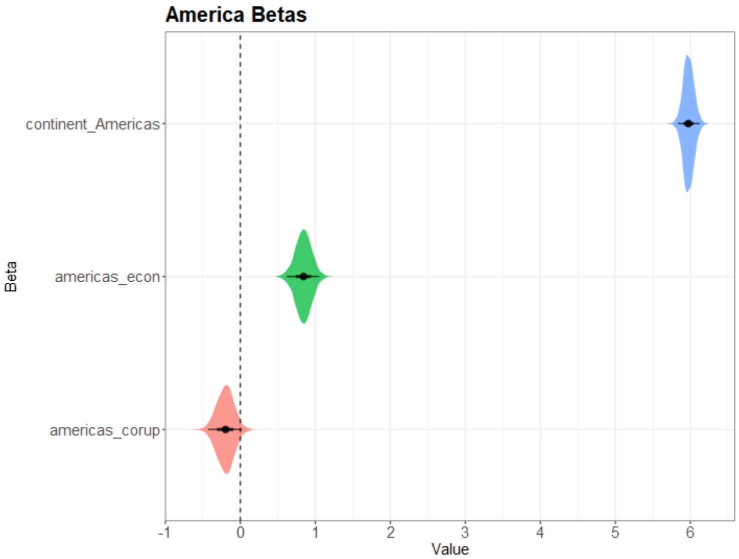

Figure 4. Beta coefficients related to estimates for a country in Americas

We can interpret the coefficients for other continents similarly. For example in Europe, there is the same relationship with happiness for both the economy and the perceived level of corruption. However, we can also compare the coefficients between continents, our model suggests that the economy has a larger impact on happiness in the Americas compared to Europe with probability P = 0.748 (±0.007). Conversely, for the importance of the perceived level of corruption, we observed that with probability P = 0.880 (±0.006) it has a larger impact on happiness in Europe.

Conclusion

My findings suggest that the best way to improve happiness is to focus on growing the economy while appearing not to do this in a corrupt manner. Lowering the sense of corruption in the people is the second way our model suggests to increase happiness in the populace. However, our model suggests that the economy should be the primary area to focus on to improve happiness.

Our conclusion suggests that a better and brighter economy is key to a happy population. Our data is arguably missing key social and cultural factors such as life expectancy, family and freedom. An extension would be to perform this analysis with a wider variety of factors to test whether wealth really is the key to happiness.

An extension of this analysis might be to fit a hierarchical regression model and have regressions for each continent of happiness on the perceived level of corruption and the economy and see how this compared to our best model.