A Demonstration of Model Selection using Stepwise Regression and K-Fold Cross Validation

Samir Fidai 04/24/2020

samirSfidai@gmail.com

2020

**Abstract:**

The goal of this project is to demonstrate model selection on Happiness scores of all countries. First, forward, backwards, stepwise, and subset regressions will be run. Second, a ten-fold cross validation will be run in order to further aid the process of model selection.

**Introduction:**

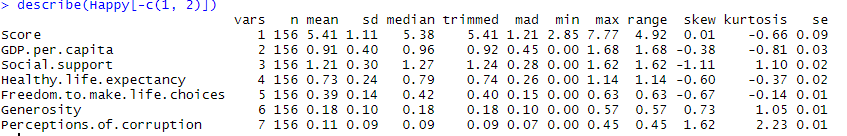
Happiness in a country can be measured. Every year, the Sustainable Development Solutions Network publishes the World Happiness Report using data from The Gallup Poll. The Happiness Report aims to assign a “Happiness Score” to each country that ranges from 0-10. The following 2019 World Happiness Report has been downloaded from Kaggle.com. The 2019 Report uses the following predictor variables in its dataset; Gross Domestic Product, Social Support, Life Expectancy, Freedom, Absence of Corruption, and Generosity.

**Methods:**

Prior to starting the procedure for model selection, descriptive statistics were taken of the 7 variables using the describe() function from the psych package, as well as a density plot of the happiness scores and also a scatterplot matrix to identify potential correlations.

A multiple linear regression model will be first generated using the Happiness scores as the response variable and Generosity, GDP per capita, Social Support, Healthy Life expectancy, Perceptions of corruption, and Freedom to make life choices as the six predictor variables. The oslrr package will be loaded, and a forward, subset, backwards, and stepwise regression of the Happiness dataset will be run in order to take note of the R-square, Mallow’s C(p), and AIC of different models. Finally, a 10-fold Cross Validation will be taken of the models concluded from the stepwise regression, as well as the initial multiple linear regression model from. The models will then be compared, and the model selected will be given in the “Discussions” section.

**Results:**

**Descriptive Statistics:**

**A picture containing different, man, boat, water

Description automatically generated**

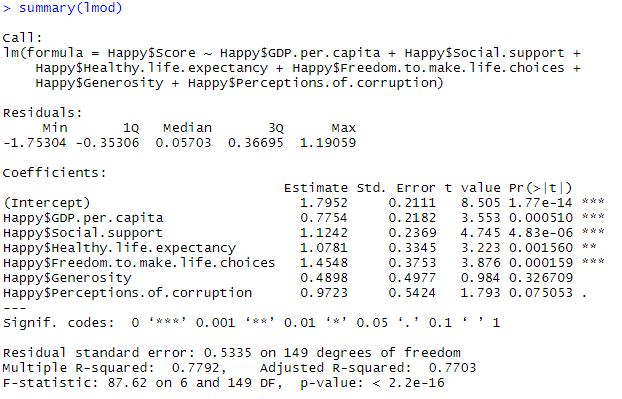
It can be concluded from the density plot of the scores almost follow a normal distribution. Furthermore, there is also a second concentration of happiness scores between 4 and 5.

**A screenshot of a cell phone

Description automatically generated**

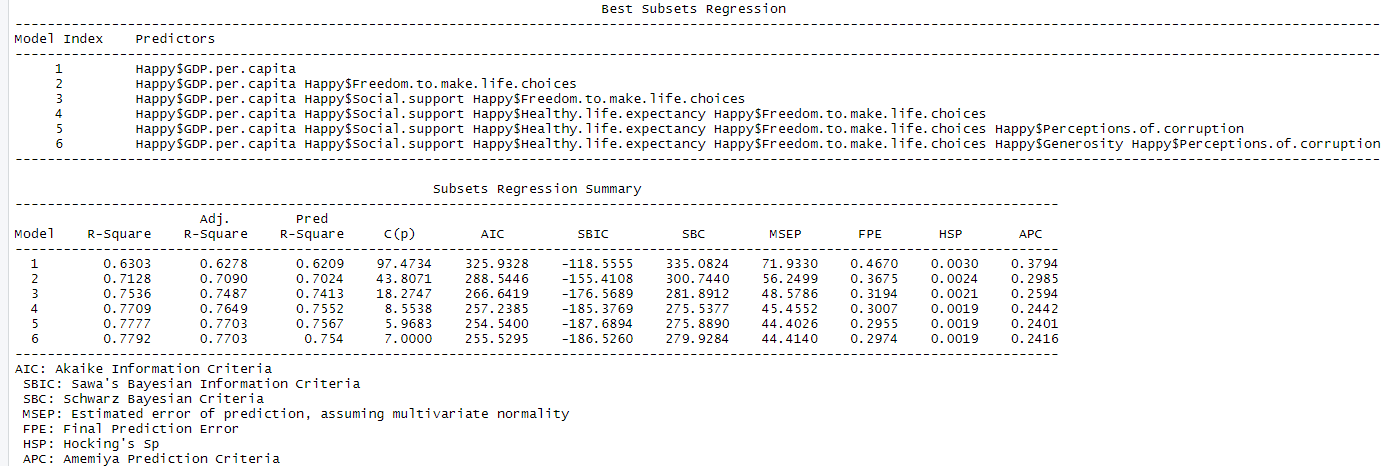
A strong, positive correlation can be seen between score and GDP, score and social report, and score and Life expectation. A weak, positive correlation can be seen between score and life choices. Little to no correlation can be seen between score and generosity and score and perceptions of corruption.

**Multiple Linear Regression:**

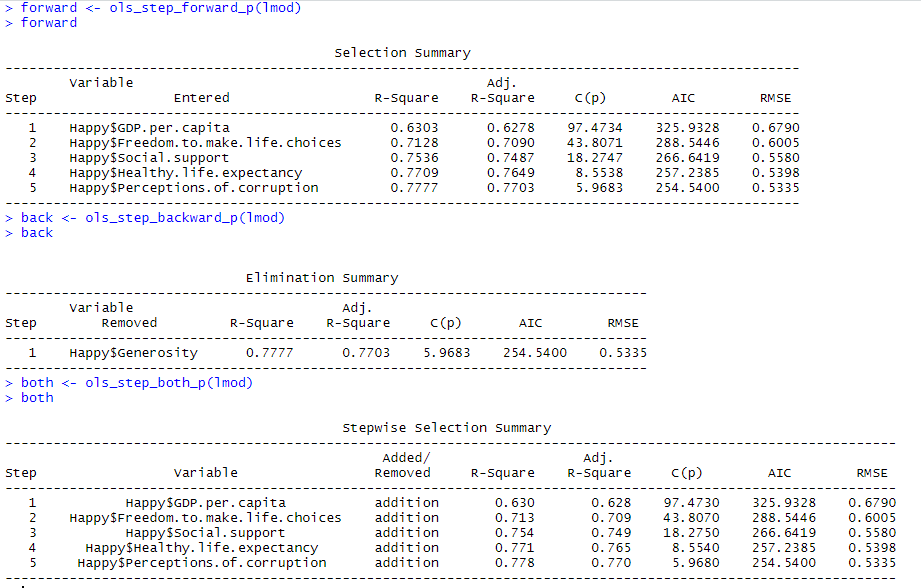
****

There are a couple takeaways from the initial regression model. First, the R-squared is 0.7792, meaning that 77.92% of the data can be explained through the regression model. However, as discussed in class, a model that is overfitted might high R-squared but will perform poor in predicting more cases. Another observation made was the high P-value for Generosity.

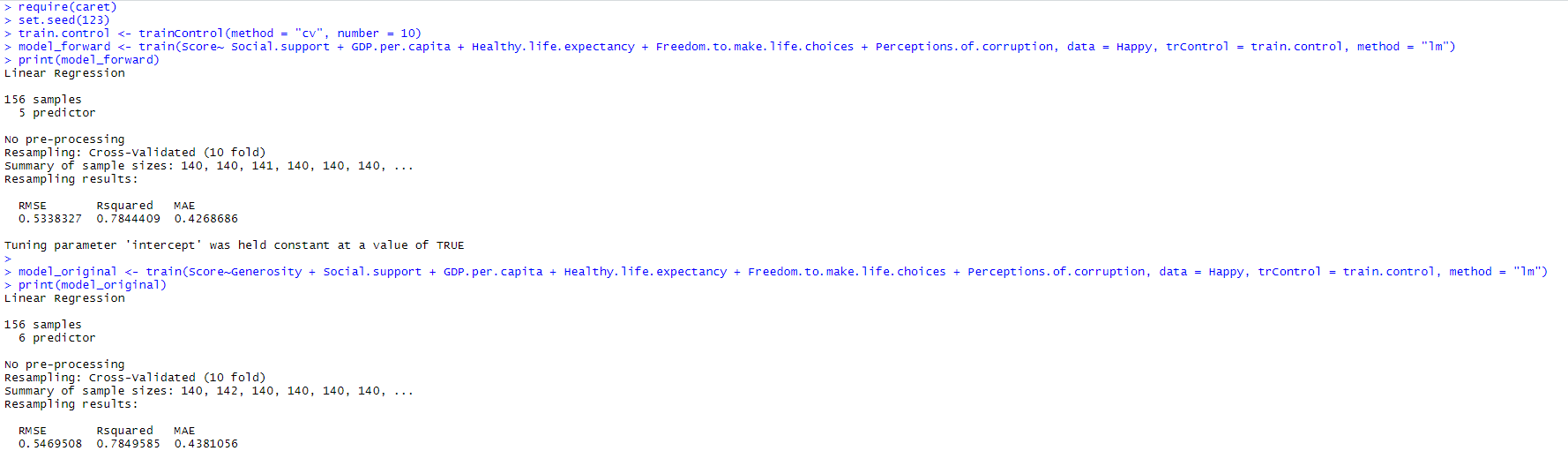
**Subset Regression, Forward Regression, Backward Regression, and Stepwise regression**

****

**Forward Regression:**

****

The stepwise regression confirms my initial observation in the multiple linear regression model. Generosity is insignificant as a predictor variable for a country’s happiness score. Furthermore, the model with 5 predictor variables had the smallest R-Squared, Mallows C(p), and AIC. In order to fully asses the validity, a 10-Fold Cross Validation will now be run twice; first with the five variables not including Generosity, and the second Cross Validation with the original regression model.

**10-Fold Cross Validation:**

**Discussion:**

The model selection process involved comparing many metrics, as well as using different methods of regression. The stepwise regression confirmed my fears from the initial multiple linear regression that Generosity was not significant to a country’s happiness score. The final 10-fold Cross Validation was run between the model with 5 variables and the model with all variables in order to determine which model was more accurate. **The model selected as a result had GDP, Social Support, Healthy Life Expectancy, Freedom to make life choices, and perceptions of corruption.** I was surprised to find my 10-Fold Cross Validation to be very close between five variables and 6 variables. However, The Mean Aboslute Error and Root Mean Squared Error for the 5 variable model was slightly smaller than the full model. The R-Squared was slightly higher for the full model, but the decision to select the model was not solely based on the R-Squared.

Overall, the model selection of country happiness scores was a very hands-on experience. The selection of the dataset involved the installation of many packages, and the process of model selection exposed me to new packages and new commands. The next steps to take are using a dataset that includes many more variables. This project further enhanced my knowledge in R, and I look forward to working more with it in the long run.

**Literature Cited**

Sustainable Development Solutions Network. ([2019; December [Name of the dataset], Version 2. April 24, 2020, <https://www.kaggle.com/unsdsn/world-happiness>.