Project Proposal for STAT 6214

Junchi Zhang

Life expectancy is the average length of life every person in the population. Different countries have different life expectancies. What cause life expectancies to differ for different country? What contribute to higher life expectancy? In this project, I plan to explore the relationship between life expectancy and more than 20 variables to see if these variables matter. This research will cover 194 countries including both developing countries and developed countries. Data used in this research is from 2000 to 2015 considering social aspect, health factors, economic performance, and education. In each aspect I will choose several variables which are representative. Overall, I plan to use 23 variables to do the analysis. Take social aspect social aspect as an example, clean water usage, mortality rates, Gini coefficient, government expenditure and Intentional homicides. Most of the data can be obtained from Kaggle (Life Expectancy WHO Statistical Analysis on factors influencing Life Expectancy) and The World Band website.

In my opinion, a high dimensional linear regression will be generated in the end of the research. Since the amount of data given is large, according to central limit theorem, normality can be assumed for each variable.

$$Life\ Expectancy_i = \sum X_i \beta_i + \varepsilon_i$$

X_i represent each variable and β_i is the corresponding coefficient. ε_i is the random noise (assumed to be normally distributed). In addition, X_i are the remaining variables after variables selection using AIC, BIC, R square, and cross validation. Transformation will be applied if variables are not seemed to be linearly correlated with life expectancy. β_i is calculated using OLS estimate. However, I will also use Lasso and Ridge regressions as well to compute the β_i. The final beta estimate will be the one with better model performance. Generalized linear model will be used if residuals are non-constant or correlated with the independent variables.

There should be high multicollinearity between variables since usually, economic performance, education, and social aspects are correlated. For example, countries that have better education usually have a higher GDP per capita. Variables with very high multicollinearity (high VIF) should be deleted from the model.

Since there are missing values in data, use a prediction by linear regression for the missing variable is an option or row with missing value will be deleted instead of filling in with artificial values since there are enough data and artificial data might mislead the research. Personally, I do not think using average or a randomly generated value to fill in the missing value is a good idea. For example, population of the country is one of the variables, by taking the average population of all countries and put it in the missing place will obviously create a bad data.

For this project, I plan to use packages that relate to high dimension linear regression. “ggplot” to see the distribution and generate thoughts for model. “MASS” for variable selection and model comparison. “reshape” and “Amelia” will be used for the data cleaning process. “glmnet” will be used to generate LASSO and Ridge regression.