Case Study: Life Expectancy

Justine Chu and Kailyn Pudleiner

Data: Our dataset was collected by the World Health Organization and United Nations, and it contains information about 193 different countries and their average life expectancy from 2000 to 2015, as well as many other associated factors.

- dependent variable: life expectancy, which measures the average expected age of the population
- independent variables: many variables related to life expectancy having to do with economic, social, and health related factors

Summary: In our analysis, we used backward stepwise regression to eliminate variables that are not useful for predicting life expectancy in our model. The resulting model had 14 variables, so we further reduced the number of variables by looking at the VIF for multicollinearity and p-values for individual terms in the model. We were able to reduce the model down to 8 linear terms and examined the residual plots to ensure that the normality assumptions are met. Then, we decided to look at a model with second order terms and compare it to the model with only first order terms to see if adding squared terms improved the performance. We tested the utility of each of the models and performed a nested F test to determine which of the models is better at predicting life expectancy. We found that the second order model performed better.

First order model: $\hat{y} = 272.5 - 0.110x_1 - 0.0193x_2 - 0.0693x_3 + 0.000553x_4 + 0.0484x_5 + 0.0302x_6 - 0.457x_7 + 1.374x_8$ (adj-$R^2 = 0.8096$)

Second order model: $\hat{y} = 15000 + 1.62x_1 - 14.8x_2 - 0.0133x_3 + 0.342x_4 + 0.0128x_5 + 0.00136x_6 - 0.0674x_7 - 1.04x_8 - 0.0189x_1^2 + 0.00367x_2^2 - 0.0320x_3^2 - 0.0333x_4^2 + 0.000378x_5^2 - 0.0000000639x_6^2 + 0.000839x_7^2 + 0.0154x_8^2$ (adj-$R^2 = 0.8353$)

Residual Plots: