Medical Cost Analysis Plan and Outline

Sophia Haass, Samson Tessema, Robbie Durham, Aditi Kishore

Motivation:

Through the entire Democratic Primaries, health care has been the center of attention. Rising health care costs and lack of coverage for millions of Americans is turning health care into a luxury item. Bernie Sanders, Elizabeth Warren, and others have constantly mentioned Medicare for All, while their opposition will argue that costs will be way too high for this to be feasible. While we cannot predict what these costs will be, we can predict things that are in our control, like BMI, whether we smoke, and how many children we have. We want to predict how health care costs can change for each person and show how to limit an individual's health care spending.

Hypothesis:

We believe that younger people will have lower healthcare costs due to high metabolism, strong immune systems, and lower risk for diseases. Additionally, men will more likely have lower healthcare costs than women due to increasing breast cancer and cervical cancer cases and numerous pregnancy complications. Non-smokers will have lower risk of lung cancer, loss of eyesight, and cardiovascular disease, so their health care costs would be less than people who smoke. We predict medical costs will be lower for a smaller number of children. Also, people with healthy BMI scores will have lower healthcare costs because BMI is an indicator of obesity and high body fat so having a healthy BMI score will reduce health care costs. We predict that people with low BMI scores will also have increased health care costs relative to individuals with healthy BMI scores, but lower than people with high BMI scores. This is because being underweight is still a health risk, but not as likely to cause diabetes or heart disease, which are expensive.

Research Approach:

In order to gather information on this subject, we visited known websites with data on this topic, such as FiveThirtyEight and Kaggle. We decided on using the data from Kaggle, as this gave us a wide range of data and multiple other methods of analysis to build off of. We will plot the cost for each category of data then we will plot different categories together to find the covariance

and correlation to costs. We will use regression models to predict health care costs and compare the models.

Methods:

We will use Linear Regression in Python to estimate healthcare costs given a number of factors inputted from the user. We will also use other regression models and compare their results. We can also determine the correlation and the covariance of the factors both with each other and with the final cost of the healthcare coverage.

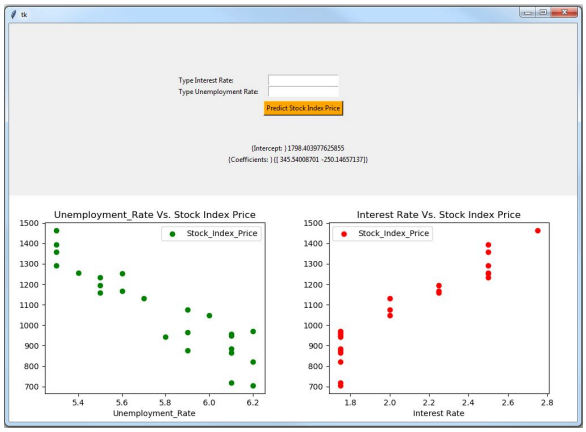

Figure 1. Example of Linear Regression in Python

Data Sources:

Kaggle https://www.kaggle.com/mirichoi0218/insurance

Timeline:

This written plan and outline will be submitted and reviewed on April 2nd, 2020, and any changes that are recommended will be implemented as quickly as possible.

We will present our one slide summarizing our project to the class on April 2nd, 2020.

We plan to have the written code and analysis portion of the project done by April 20th, 2020, one week before the final poster presentation is due.

We aim to complete the presentation by April 25th, 2020. It will give us some time to rehearse prior to presenting.

Task Division:

We will work together on all aspects of the project through live zoom meetings. Samson will be in charge of coding; however, we will all contribute to the code. Aditi will do the data analysis, but we can work on that as a team as well. Sophia will work on the presentation. We will keep the presentation on google slides, so we can all work on different parts. Robbie will work on the final report. We will complete the final report on google docs, so we can all collaborate on it and edit it.