Jake Schinto

Math 312

Prof. Ping-Shi Wu

**Dataset Introduction - Health Insurance**

In the world’s current state, health has become a major concern for people. The coronavirus has been especially costly on the average person because of the effects it has on people but also the limited testing can be expensive if someone is uninsured. According to the last count in 2018, 27.5 million Americans are uninsured; this makes the cost of testing a large burden on the uninsured. Costs to visit the doctor alone could be as much as $1,151 and adding the testing could make it as much as $3,270 [1]. These costs and public crises are all part of the complex equation that insurance companies need to take into account when deciding costs to cover an individual. In addition to my curiosity about the current pandemic, my family also has a long running background in healthcare. For many generations before me, my family has practiced dentistry and are therefore influenced by people who have insurance coverage. Lastly, I will soon graduate and be in need of health insurance, so knowing the factors that go into deciding costs are important for minimizing expenses. For these reasons, I have decided to take a closer look at the insurance dataset [2].

The dataset is split up into 7 different columns: age, sex, bmi (Body Mass Index, the ratio of weight to height), children, smoker (binary yes/no), region (quadrant of US), and charges (costs billed by insurance). Using these features, I hope to create a model using my knowledge of statistics and data science that will allow me to accurately predict a new given person’s insurance costs based only on the features of that individual.

This ability to predict medical costs is important because it is the deciding factor of how much someone will have to spend on having health insurance. This information is therefore both valuable to the individual who will end up having to pay the bills as well as to the insurance companies themselves so that they can ensure that they are not losing money to a particularly pricy individual to cover.

When looking for others’ work with this kind of data, I found a Medium article by Bayu Galih Prianda who documented his attempt at finding a linear regression model to use with health insurance data. He had attempted to use Python with multiple linear regression (multiple independent variables) in order to find a best fit line in order to accurately predict the medical costs. His results came out to y = -11676.830 + 259.547x1 + 322.615x2 + 23823.684x3, where (y = charges, x1 = age, x2 = bmi, x3 = smoking(0/1)). What strikes me as interesting about these findings is that smoking is equivalent to about 90 years of age, which seems wild and possibly could lead to alarm for any smokers looking to get insurance [3].

In conclusion, I look forward to experimenting myself with this data and finding the true factors behind what can cause a certain person to have a higher bill than others.

**References:**

1. Srikanth, Anagha. “How Much Will Getting Coronavirus Cost You?” *TheHill*, 3 Mar. 2020, thehill.com/changing-america/respect/poverty/485666-how-much-will-getting-coronavirus-cost-you.
2. Stedy. “Stedy/Machine-Learning-with-R-Datasets.” *GitHub*, 28 Mar. 2017, github.com/stedy/Machine-Learning-with-R-datasets.
3. Prianda, Bayu Galih. “Prediction of Health Insurance Costs with Linear Regression.” *Medium*, Medium, 24 Dec. 2018, medium.com/@BAYUGALIH/prediction-of-health-insurance-costs-with-linear-regression-8fd95a905a40.