Application of Regression for Modeling Medical Cost Projections

Final Project Proposal

Carroll, Kenneth J.

2021

Medical costs are a constant threat to economic stability for a number of individuals in the United States. While there are various schools of thoughts and political ideas towards how medical costs should be, the fact remains that the costs exist and will remain to for the foreseeable future. The problem is, medical costs can vary for a variety of factors ranging from age and gender to region and genetic risk factors, which makes planning personal finances to account for these potential additional costs much more difficult. The task at hand will be to use data gathered for the purposes of determining insurance premiums for insurance companies, and instead estimating future medical costs for use by the average consumer. The data set was already geared towards predicting cost, since that cost is the input for most insurance companies to calculate premiums for their customer. The tasks at hand are determining the most influential attributes that determine overall medical expenses, identify features that lack any significant correlation, and attempt to account for factors not reflected in existing datasets or additional gathered data. The goal will be to create a regression model that will allow users to modify their personal information, so they can plan what to expect in medical costs under various conditions and different points of life.

The selected data set for this project is sourced from Kaggle, found [here](https://www.kaggle.com/noordeen/insurance-premium-prediction). The dataset includes 1338 unique entries, with the attributes of age, sex, BMI, number of children, smoker status, and regional categorization. It also includes the medical expenses for the individual and is the value under observation for this project. Within the system proposed for this project, the attributes regarding a person will be used to estimate the medical expenses they can expect. The system will allow them to enter these attributes for the models predicted price; they can use this info to see how changes in their life can impact their future medical expenses. The cost of an additional child, the savings of reducing bodyweight or quitting smoking, the way their age will increase their expenses as a whole – these are the questions that the system will hopefully answer for an individual allowing them to better prepare for the future. The initial approach for the dataset is going to be a polynomial regression, with comparisons to approaches with other algorithms as further research continues. As with most regression, R-Squared value will be used as an initial verification, but the tuning of the model will come more from the adjusted R-Squared value since it penalizes unnecessary features and with any dataset there may be features that do not contribute significantly to the overall accuracy of a given model. Since is the goal is to accurately model the cost for an individual, R-Square works by minimizing overestimates and underestimates; since either could drastically impact an individual’s quality of life, with overestimates impacting quality of life before an incident and with underestimates impacting quality of life after an incident, trying to focus on one or the other could negatively impact the other.

While goal of the project is believed to be reasonable, there are multiple challenges that could impact the validity of the model. The current data set does not specify regions in a sense that can be used in a literal sense, since the description of the region is just a cardinal direction. However, it can be useful for establishing a pattern in this current dataset and help determine to what degree it may impact the accuracy; if it is determined to have little correlation, it may not be worth finding data including that feature. Additionally, the data does not include any information for how long someone has been in a particular region, so someone who moved there a day ago will be considered the same as someone who has lived a lifetime in that region. Lack of time data for the region could lead someone who fits better with “northeast” to be classified as “southwest” or any other possible interactions with region. Alongside that, the data set uses BMI as a metric which is not a direct indicator of health. Bodyfat percentages and activity levels can be more accurate measures in their own regards which may impact the model for not including these features. The dataset was also sourced from Kaggle and indicates it was used as part of a separate course – this makes verifying the validity of the data difficult and could indicate the need to source multiple datasets for accuracy and additional testing. I do believe the current dataset is enough for initial verification but sourcing additional data could be a required task but also a challenge where health information is usually private unless modified to a degree of anonymity.

Related Works:

Predicting Health Care Costs Using Evidence Regression: This article uses different attributes but attempts to solve the same fundamental problem. It and the other related works serve as similar comparisons to the type of problem I wish to solve, while also being different enough that my work should hopefully be a more unique contribution rather than a rehash of their existing project.

<https://www.mdpi.com/2504-3900/31/1/74?type=check_update&version=2>

Development of a predictive model for integrated medical and long-term care resource consumption based on health behavior: application of healthcare big data of patients with circulatory diseases: This resource is an example where information is used to predict resource consumption for individuals afflicted by cardiovascular disease. While only focusing on a subset of medical patients, it does form a relationship basis between individuals and their incurred resource consumptions – the goal is to take a similar approach but specifically towards that individual’s costs and with a broader scope in types of people observed; there will be no isolation of a specific condition or lack thereof.

<https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01874-6>