ORIE 4741 Midterm Report

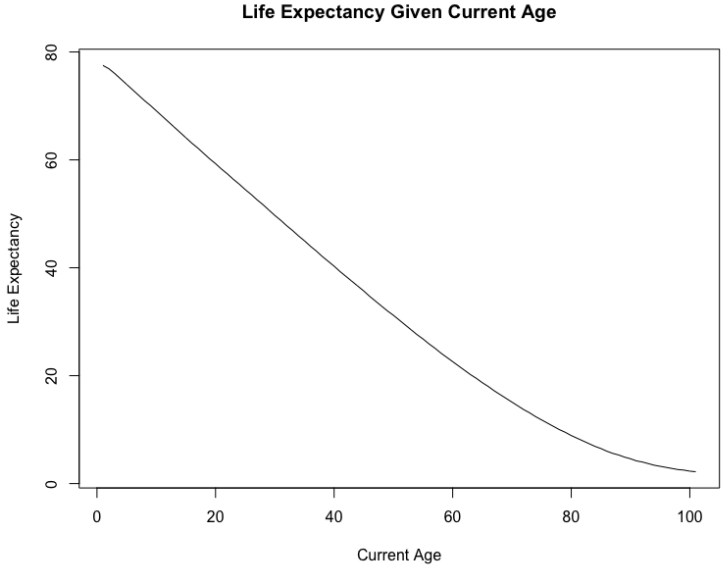
Sean Yang (smy54), Rochelle Kris (rhk79)

October 2021

1 Data Set

We will use a data set that contains features such as the probability of dying between ages x and x + 1, the number of individuals surviving to age x, the number of individuals dying between ages x and x + 1, etc. Furthermore, the data set contains a column for the expectation of life at age x, which will serve as the feature label column. We have a collection of such data sets that are divided by race and year. We have combined these data sets into one large feature set by creating a feature for the year and a feature for the race of the population. For each of the original data sets, we had an entry for every age 0 through 100, so there were 100 feature vectors. This data set was split into our data and label data structures.

Using one representative sample from our data, we can plot the set of points using only current age as a feature with the label to produce the following plot:



From this plot, it is safe to assume that we should have an overly complex model. Further, a linear model could be mostly accurate on this data. However this is an oversimplification of our data that only depends on the current age of an individual. This serves as a confirmation of the reliability of this feature in our dataset because we know that we should be seeing life expectancy decrease for older individuals compared to younger people.

2 Over/Under-Fitting Prevention

Balancing the complexity of the model may be tricky and is necessary to prevent under and over fitting of the training data set. We have learned several methods of creating the training and validation sets from the original data. Here we will focus on using n-fold cross validation to produce a data splitting method that should produce somewhat accurate results at a reasonable time complexity.

3 Model Evaluation

This model will predict values as a regression rather than a classification system, so we need a way of evaluating the model that tells us how close to the true value our predicted value is. Here we will a measure of Mean Squared Error to calculate our accuracy, or lack thereof. The formula we will use for measuring this accuracy is the following:

$$MSE = \frac{1}{n} \sum_{i \in D} (y_i - \hat{y_i})^2 \tag{1}$$

Where n is the number of samples in our dataset D, y_i is the actual label of the i^{th} sample, and $\hat{y_i}$ is the predicted value of the sample.

4 Preliminary Analyses

We investigate the changes in life expectancy across various age cohorts over a short length of time to get a preliminary understanding of where the most significant changes are likely to be. In order to do this, we plot the variances in life expectancy for each age cohort between 2000 and 2005 to identify which cohorts saw the most significant fluctuations in that period.

As we can see, the plot illustrates that Americans of most ages saw changes in their life expectancy over that 6 year period, with the notable exception of seniors, whose life expectancies were relatively stationary. This suggests that the most influential factors for Americans' life expectancy are likely focused on the lives of younger people, but do not have significant impacts on the mortality rates of senior citizens. This insight will help guide our future work as we try to identify which features are likely to be strong predictors of life expectancy changes.

5 Next Steps

We have several hopes for the future of our project. We would like to add features to the set that will allow us to predict not only the life expectancy of the individuals, but also the expected cause of death. This will help to understand the disparities in deaths across different demographics. As mentioned in the previous section, we will focus our investigations on the features most likely to impact life expectancy for younger Americans.

Further, we would like to improve the model by using more local data: currently the data is for the United States as a whole, but we would like to add a feature for location that would be the state within the country or ideally the zip-code or town.