ORIE 4741 Final Report

Sean Yang (smy54), Rochelle Kris (rhk79)

December 2021

1 The Problem

What factors, such as health conditions or resource accessibility, affect life expectancy?

Our original aim of this project was to determine how factors such as health conditions and resource accessibility may influence life expectancy. During our preliminary analysis of the data, we observed consistently higher variance in the life expectancies of Americans under 50, after which that variance quickly declined, suggesting that the average American's experience during the prime years of their adulthood plays a significant role in their ultimate health outcomes. Thus, we chose to focus on race and sex as features of importance, characteristics which inescapably affect nearly every facet of adult life in America.

Figure 1: Results of preliminary analysis

Ultimately, we posed the following question: do the demographic features of race and sex play a statistically significant role in determining the life expectancies of Americans and, if so, can we predict this influence? Specifically, we focus on the task of predicting disparities in life expectancy across race and sex for Americans of every age cohort.

2 Data

2.1 Obtaining the Data

We developed a combined dataset catered towards answering this question. We found individual data sets for different years and demographics, so these became features in our models and we were able to merge the data tables together into one. This was done by combining data from the CDC's National Vital Statistics Survey (NVSS) life tables, a series of 405 tables capturing life expectancy and other statistics by age, sex, and race across a 17 year long period from 2001 to 2017, with extracts from US Census Bureau's annual American Community Survey (ACS) over the same period of time drawn from the Integrated Public Use Microdata Series (IPUMS), which provided higher resolution data on the size and distribution of each demographic subgroup under consideration.

Our datasets presented with no missing data, simplifying the task of data cleaning. The only major challenge was tailoring our scripts to handle the various file formats that the CDC switched between from 2001 to 2017. After all collation was completed, we were left with 15387 data points of 10 features each.

The collated dataset included features from the NVSS 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, all in addition to the demographic information provided by the ACS. The value of x ranges from 0 to 100. Furthermore, the data set contains a column for the expectation of life at age x, which we treat as the feature label column.

2.2 Feature Engineering

We employed several feature engineering techniques such as one-hot encoding, min/max scaling, etc. One example of feature engineering used in our analysis was converting the categorical sex data to be a binary vector that is 1 for male and 0 for female. In order to produce this feature, we first used one-hot encoding on the sex feature and then dropped the female column, leaving us with this binary classification.

Furthermore, we had a feature representing the year which was significantly larger than most other features in the data set. Since we know that year should not be a tremendous influence on life expectancy (especially since we only have data for sixteen years), this feature was scaled down by subtracting the minimum value across all data points from each sample's year feature. Specifically, the

minimum year seen was 2001, so the "Year" feature was turned into a "Years Since 2001" feature.

2.3 Splitting the Data

Once we concluded the feature engineering steps, we divided the data into three groups: training, validation, and testing. These sets contained 80%, 10%, and 10% of the full dataset, respectively. These sets were used to have "new" data to evaluate the data on and help reduce the risks of over- and under-fitting the data to the training set. The training set was used to produce an initial assessment of the models. Next the models used the validation set to produce a score of how well the model performs on new data. Finally the scores from each model predicting the validation sets were compared to select the best model and this model was used to predict labels on the testing set.

3 Algorithms & Approaches

3.1 Models

1. **Linear Regression**
   The linear regression model develops a model that represents a line in the dimension of the feature spaces. For our dataset in d dimensions, this model produces $\vec{\boldsymbol{w}} = [w_1, ...w_d]$ and a scalar b for the offset. Thus we have a model that looks like:

   $$y_i = w \cdot x_i + b$$

   Further, the function used the develop this model relies on ordinary least squares loss in tuning the parameters to produce the optimal configuration.

2. **Stochastic Gradient Descent Regression**
   The stochastic gradient descent model relies on the SGD learning algorithm with a squared error loss function. While SGD is used for classification more often than for regression, it can also be used as a regressor for larger data sets. After experimenting with different parameters to the SGD Regression function, we found the ideal configuration was to use at most 1000 iterations of the algorithm with termination threshold of $1e-3$, meaning the algorithm terminates when the loss at an iteration is greater than the best loss seen minus this threshold.

3. **Kernel Ridge Regression**
   The third model we developed relies on kernel ridge regression. This is a variation of support vector regression which uses kernelization of the data and squared error loss.

3.2 Measures of Accuracy

We used mean squared error (MSE) as a measure of accuracy to assess our models. MSE is defined as follows:

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

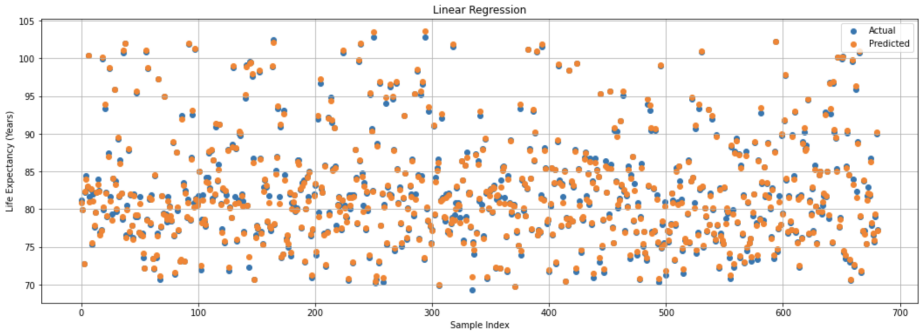
Where y_i is the actual label and $\hat{y}_i$ is the model's prediction on the i^{th} data point. Further, D is the set of all data points and n is the number of samples.

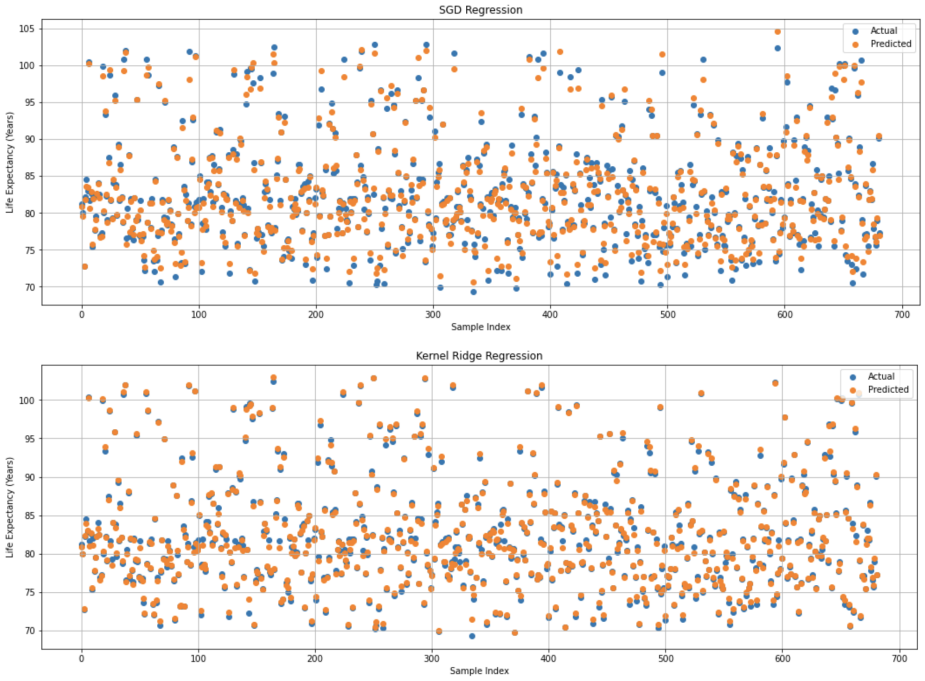
We chose to use MSE over other measures of accuracy like mean absolute deviation because this approach gives greater weight to outliers.

4 Results

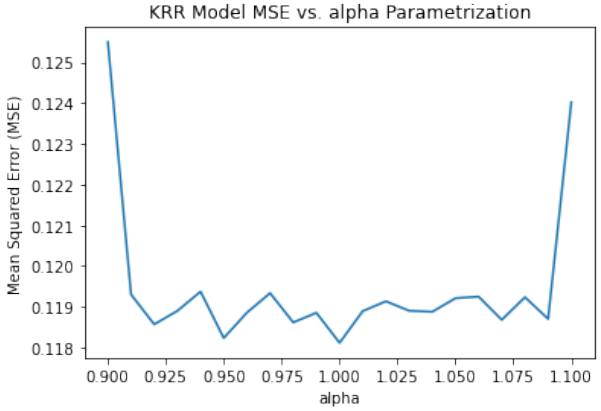
We have developed a machine learning model which produces predictions on the life expectancy of individuals based on features such as demographics, sex, and other features like the number of individuals surviving to the current age of this individual. Below are the graphs and MSE values from each model on the validation set:

Model	MSE
Linear Regression	0.12
SGD Regression	0.61
Kernel Ridge Regression	0.11



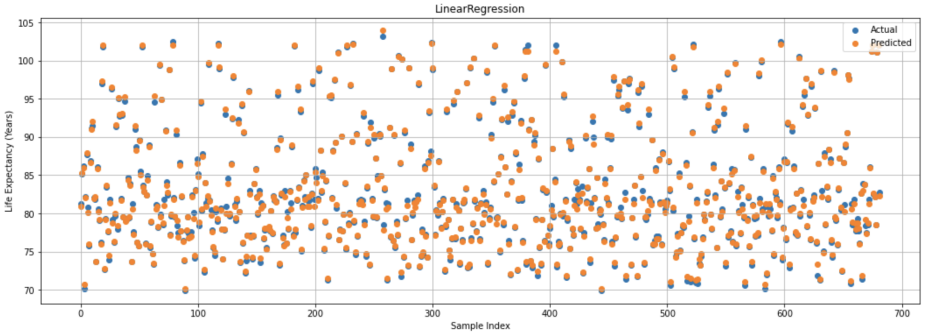


Ultimately we found Kernel Ridge Regression (KRR) to be the most promising model class. From here, we tested a series of values for the α parameter of the KRR model in order to determine its optimal value, the results of which are shown in the plot below.



Our final model has α = 1.0, an MSE of 0.119 against the validation set, and predicts with 88.1% accuracy when evaluated against the test set. This was concluded by comparing the MSE values of the validation set on each of the models and selecting the best MSE score. From this we were able to select

a model and evaluate its performance on the testing set. Below is the graph presenting the performance of the final model on the testing set.



Ultimately, these results are strong enough to illustrate the existence of a correlation between race and sex and disparities in life expectancy. Since the only features we trained our models against were related to race and sex, we know that if there were truly no relationship between these features and individuals' life expectancies and any disparities were instead the results of random fluctuations, then the models' best prediction for each age group would simply be the average life expectancy of the age group, making the model MSE equal to the variance in life expectancy at that age group. However, we find that our validation MSE of 11.9% is significantly higher than even the highest variance of any age group, which sits around 8%, indicating that there is, in fact, an issue of correlation between race, sex, and life expectancy. While not enough to produce a specific recommendation of subsequent actions to take, our results illustrate the existence of a bias and the need for more targeted work.

Could these results be used to produce a weapon of math destruction? Were they used to guide policy change it is possible that they could do just that, as the data is fundamentally coarse-grained in its categorization of the American populace into discrete demographic groups.