Predicting Life Expectancy using Community Health Status Indicators (CHSI) Data

Group 8: Jie Yu, Shuwei Liu, and Noah Kreski

Introduction

Life expectancy is one of the most important measures of health. It serves as a meaningful health indicator that can be compared between countries or regions within countries (Mathers et al., 2001), and helps to answer the most fundamental question concerning health: how long can the typical person expect to live? Life expectancy may be the outcome of multiple factors, such as lifestyle choices, access to healthcare, socioeconomic status, community and social support (WHO, 2016).

According to the latest annual report from the National Center for Health Statistics (2018), the average American could expect to live 78.6 years, down from 78.7 years in 2016, and as a key indicator for the Sustainable Development Goal to “Ensure healthy lives and promote well-being for all at all ages” (WHO, 2016), prolonging lifespan is an urgent and necessary goal. Public health professionals need to ensure the efficient allocation of resources to combat this decline in life expectancy. Under such circumstances, prediction of life expectancy is necessary, providing vital information for relevant health policy making.

In this project, we aim to build a model that 'best' predicts the average life expectancy per county in the United States, and examine the association between multiple factors, including those mentioned above, and average life expectancy.

Methods

Data Description and Preparation

We used the Community Health Status Indicators (CHSI) dataset which was released in 2012 by CDC. The thirty-three predictor variables we picked provide information at the county level related to residents’ sociodemographic characteristics, lifestyle choices, access to healthcare and other health-related risk factors. When we checked the correlations among predictor variables, we found that there was a group of eight variables measuring the total count of individuals for a county which were all highly correlated with county population size. We felt that proportions of key factors were more pertinent than pure counts, and so by dividing those variables by county population size, we successfully eliminated most of the correlations. Our final predictors include age, race, population size and density, death rate, percent with poor health status, number of monthly unhealthy days, percent not exercising, percent with few fruits or vegetables, percent obese, percent with high blood pressure, percent smoking, percent with diabetes, physician rate, dentist rate, presence of a community health center, health professional shortage (HPSA) designation, percent poverty, percent unemployed, percent uninsured, percent medicare based on elderly status and disability, percent with no high school diploma, percent severely disabled, percent with major depression, and percent with recent drug use. Twelve predictor variables have missing data, and we decided to perform the K-nearest neighbors (KNN) imputation on them in the subsequent model building part.

Our response variable of interest, average life expectancy, is also provided by this data set at the county level. We excluded two observations with missing values for the response, therefore we had a total of 3139 county-level observations.

Exploratory Data Analysis

The descriptive statistics for all continuous variables is shown in Table 1. There were two categorical variables, HPSA designation which was seen in 75.95% of the sample, and presence of a community health center, 55.11% of the sample. Since average life expectancy is approximately normally distributed, we didn't apply any transformations to it.

According to the correlation plot (Figure 1), some of the variables in our data were highly correlated with one another. Percent with age 85+, percent with age 65-84, and percent medicare based on elderly status show strong positive correlations, which makes sense because they are all related to the elderly. Percent uninsured and percent poverty are positively correlated as individuals below poverty are at high risk of being uninsured. Number of monthly unhealthy days, percent with poor health status, percent medicare based on disability, percent with no high school diploma, and percent not exercising are positively correlated with one another, which may provide redundant information. Additionally, it is easy to understand that percent white and percent black show a strong negative correlation.

Models

Multiple models were examined to choose the best one, including ridge regression, LASSO, principal components regression (PCR), partial least squares regression (PLS), generalized additive model (GAM) and multivariate adaptive regression splines (MARS). An ordinary least squares regression (OLS) was also prepared to compare with other models. Ultimately, LASSO was isolated as the ideal model to fit the data.

A design matrix of predictors was prepared, and data was split into training and testing datasets with 75% chosen as the training rows. Scaled, centered and KNN-imputed data was used to build models during cross-validation period since there are a number of missing values in the original dataset.

Except for OLS, all the other models have tuning parameters in their model building. We utilized 10-fold cross validation for 5 times to find the optimal tuning parameters which gave us the smallest cross-validation error, and refit the models on the training data using the optimal tuning parameters. Models were then applied on the testing data to make predictions for the response. Root mean square error (RMSE) were calculated to compare the model performance by summary statistics (Table 3) and violin box plot (Plot 3).

RIDGE: As one of the shrinkage methods, it can perform particularly well when there is a subset of true coefficients that are very small. We tested a sequence of tuning parameter λ from -5 to 2, and the optimal one chosen by cross-validation was 0.428. In terms of mean RMSE (1.2747), ridge performed better than OLS, but not that well compared to others.

LASSO: As another shrinkage method, it can penalize the coefficient estimates of unimportant predictors to zero and offer a clearer interpretation compared with ridge regression. A sequence of tuning parameter λ from -4 to 1 was tested and the optimal one chosen was 0.058 (Figure 2: the optimal log(λ) is at the bottom of the "U" shape).The final model provided 15 non-zero coefficients, which are listed in Table 2.These variables played an important role in predicting the response "average life expectancy". As expected, LASSO removed some highly correlated variables and kept the most essential ones, for example, keeping percent not exercising, and removing the variables highly correlated with it as mentioned above. Mean RMSE (0.9036) of this model turned out to be the smallest among all models (Table 3).

PCR: This method was utilized to examine the predictors as linear combinations that reduce dimension and capture joint variation. Here, the response does not supervise this identification of these principal components. This method led to the identification of 25 principal components, based on cross-validated results that showed the lowest RMSE at this number, reduced from 33 predictors overall. In terms of mean RMSE (0.9340), PCR performed better than OLS or ridge, but worse than LASSO and MARS(Table 3).

PLS: This method operates similar to PCR, but in a supervised way to ensure that new features are related to the response. Here, the optimal number of components was only 4, a severe reduction from the initial sample. While the mean RMSE from this model (1.0197) was higher than PCR, the PLS model offers helpful simplicity (Table 3).

GAM: This model obtained its optimal smoothing parameter through generalized cross-validation. It extends the linear model to allow for certain non-linear relationships, was not only computationally intensive and prohibitively complex, but yielded the lowest mean RMSE (1.5669) among all the models (Table 3). While it may serve some use in isolating non-linearity within the data, it does not serve as a useful model for these data in terms of practicality.

MARS: This method was used to create a piecewise linear model with an automatically chosen cut point. The best combination of tuning parameters were identified by a grid where the result showed that 11 degree of freedom allowing interactions in the model is appropriate (degree of features = 2, maximum number of terms including intercept = 11). The predictors chosen by MARS include percent with no high school diploma, death rate, percent native American, percent black, population size, percent poverty, percent obese, interaction between population size and death rate, interaction between native American and Hispanic, interaction between poverty and obesity, interaction between poverty and recent drug use, and interaction between native American and poverty. RMSE of MARS (0.9275) is slightly larger than lasso regression but smaller than others (Table 3).

LASSO ultimately demonstrated the lowest mean cross-validated RMSE (Figure 3), 0.9036, with the smallest RMSE variance 0.0153, showing the best predictive performance to fit these data and a certain degree of stability. The final model kept 15 predictors with non-zero coefficient estimates (Table 2) which were important in predicting the response. Predictors related to the response ranged from demographics, where the percent of a county whose population is Black was negatively associated with life expectancy, to health behaviors, where lower average life expectancy was predicted by more people smoking or not exercising, and more. The percent age 85 or more, percent Asian or Hispanic, and the dentist rate all positively predicted average life expectancy, while, poverty, death rate, poor health status, high blood pressure, lack of insurance and not having a high school diploma all negatively predicted life expectancy. Curiously, the percent with recent drug use had a positive association with average life expectancy, but this coefficient is the smallest of all non-zero coefficients. There may be underlying differences in counties with higher substance use that confound this association. One limitation of LASSO is that it is relatively less flexible than linear regression. However, it is also more interpretable because in the final model the response variable is only related to a small subset of the predictors, leaving the others with coefficients of zero. This helps simplify results down to the most vital associations.

Conclusions

In conclusion, longer life expectancy should be pursued with a focus on health behaviors like smoking and low exercise, as well as structural efforts, like improved insurance and education. Areas with higher poverty and higher proportions of Black residents should receive specific attention in order to limit disparities in life expectancy. While these findings make sense, many other factors that could have been significant were not, and so having concrete evidence lets us understand which areas of health need the most urgent intervention.

References

Mathers, C.D., Sadana, R., Salomon, J.A., Murray, C.J.L., Lopez, A.D (2001). Healthy life expectancy in 191 countries, 1999. Lancet 2001;357:1685-91.

National Center for Health Statistics (NCHS). Health, United States, 2017: With special feature on mortality. Hyattsville, MD, 2018. Retrieved April 5, 2019, from https://www.cdc.gov/nchs/data/hus/hus17.pdf

World Health Organization (WHO). World Health Statistics 2016: Monitoring the Health Goal - Indicators of Overall Progress, Retrieved March 11, from https://www.who.int/gho/publications/world_health_statistics/2016/EN_WHS2016_Chapter3.pdf