Proposal

Group 8

2019-03-11

Group Members

Name	Uni
Noah T. Kreski	ntk2109
Shuwei Liu	sl4471
Jie Yu	jy2944

Project Title

A Predictive Model for Life Expentancy using Community Health Status Indicators (CHSI) Data

Motivation

Average life expectancy is a general health measure stating average lifespan for a given area, and so our goal is to build a model that 'best' predicts life expectancy per county in the United States.

Life expectancy in the United States has been declining recently, and as a key indicator for the Sustainable Development Goal to "Ensure healthy lives and promote well-being for all at all ages", improving lifespan is an urgent and necessary goal.

Reference:

Sustainable Development Goal's utilization of life expectancy

U.S. life expectancy in decline for the second year in a row

Anticipated Data Source

Data: Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer, from Centers for Disease Control and Prevention

Short Description:

Community Health Status Indicators (CHSI) to combat obesity, heart disease, and cancer are major components of the Community Health Data Initiative. This dataset, which contains over 200 measures for each of the 3,141 United States counties, provides key health indicators for local communities and encourages dialogue about actions that can be taken to improve community health (e.g., obesity, heart disease, cancer).

Response and predictors

Response variable:

- `ALE`: Average life expectancy on county level

Identifiers:

- `CHSI_County_Name`: Name of county
- `CHSI_State_Abbr`: Two-character postal abbreviation

Possible Predictors (on county level)

1) Demographic Information:

- `Population_Size`: County population size
- `Population_Density`: People per square mile
- `Poverty`: Percent of people living below the poverty line
- Population by Age (`Age_19_Under`, `Age_19_64`, `Age_65_84`, `Age_85_and_Over`): Age-specific population sizes
- Population by Race/Ethnicity (`White`, `Black`, `Native_American`, `Asian`, `Hispanic`): Race- and ethnicity-specific population sizes

2) Summary Measures of Health:

- `All_Death`: Mortality rate per 100,000 from all causes of death
- `Health_Status`: Percent of population self-rating health as poor or fair
- `Unhealthy_Days`: Average number of unhealthy days a person had in the past month

3) Risk Factors / Lifestyle Choices:

- `No_Exercise`: the percentage of adults reporting of no participation in any leisure-time physical activities or exercises in the past month
- `Few_Fruit_Veg`: the percentage of adults reporting an average fruit and vegetable consumption of less than 5 servings per day
- `Obesity`: Percentage with a BMI >= 30.0
- `High_Blood_Pres`:the percentage of adults who responded yes to the question, "Have you ever been told by a doctor, nurse, or other health professional that you have high blood pressure?")
- `Smoker`: the percentage of adults who responded "yes" to the question, "Do you smoke cigarettes now?"
- `Diabetes`: the percentage of adults who responded "yes" to the question, "Have you ever been told by a doctor that you have diabetes?"

4) Access to Care:

- `Uninsured`: The number of uninsured individuals in a county
- `Elderly_Medicare`: Number of those aged 65+ on Medicare
- `Disabled_Medicare`: Number of those with disabilities on Medicare
- `Prim_Care_Phys_Rate`: Total active, non-federal physicians per 100,000)
- `Dentist_Rate`: Total active dentists per 100,000 people
- `Community_Health_Center_Ind`: Indicates the presence of a health center for low-income and uninsured families with funding from HRSA
- `HPSA_Ind`: Indicates that a county has a shortage of health professionals as determined by the department of HHS

5) Vulnerable Populations:

- `No_HS_Diploma`: The number of individuals aged 25 years and older who have not graduated from high school
- `Unemployed`: The number of persons who had no employment
- `Sev_Work_Disabled`: The number of individuals who are severely work disabled

- `Major_Depression`: The number of individuals aged 18 years and older experiencing a major depressive episode during the past year
- `Recent_Drug_Use`: The number of individuals aged 12 years and older using illicit drugs within the past month

Planned analyses

1.Data exploration: descriptive and visualization

2.Variable reduction: using Principle Component Analysis (PCA)

3.Model building

- Odinary Least Squares model: using stepwise regression / best subsets regression
- Ridge regression
- Lasso
- Principal components regression (PCR)
- Partial least squares (PLS) regression
- GAM
- KNN
- MARS

4.Determine the best predictive model: use k-folds Cross Validation to check the predictability of our models; made density plots to compare the distributions of testing RMSE for all models, and select the best predictive model