Project Proposal

	Name	Student ID	Email
1	Andrew Thappa	57683154	c1l0b@ugrad.cs.ubc.ca
2	Nan Chen	46937141	nanchen1@alumi.ubc.ca
3	Zihui Zhao	37480143	lilaczhaozihui@alumni.ubc.ca

Key Question

Our goal is to create a model in order to predict death age given an individual-level dataset composed of 121 predictors for 770,000 Indian citizens across 9 states in a year. Our underlying assumption is that demographic factors, hygiene conditions, lifestyles, current health conditions, community environment will contribute to death at certain age.

Our analysis plan first consists of an exploratory data analysis. From this, we will determine the functional form of our model and will then use feature selection in order to determine the subset of variables that are most impactful for improving the predictive power of our model.

Data

For this project we will use a unit level dataset sourced from Kaggle that is a subsample of a survey sample containing information related to death of residents from 9 Indian states during the reference period. The data were collected in a tripartite manner. In 2010-11, baseline sample survey data was collected with 4.14 million households in the overall sample with 447,223 observations included in our subsample. In 2011-12 the survey data were subsequently updated and in this update, 4.28 million households were in the overall sample, and 174,192 observations included in our subsample. In 2012-13 there was a third update in which data for 4.32 million households was added, with 148,167 observations included in our subsample.

The data include information on sex of deceased, date of death, age at death, registration of death and source of medical attention received before death. In aggregate there are total of 770k observations and 121 variables in this dataset. The 9 states included in the data are B.
Assam., Bihar, Chhattisgarh, Jharkhand, Madhya Pradesh, Odisha, Rajasthan, Uttar Pradesh, Uttarakhand.

Challenges and Difficulties

- We are proposing a project that is trying to predict age of death. One problem is that the variable in our data that involves death is broken out in two variables, `"age_of_death_below_one_month"` and `"age_of_death_above_one_year"`. One proposed way of handling this would be to create a new variable with age of death measured in days. Another option would be to split our sample for outliers and normal part.
- How to measure social demographic variables quantitatively.
- Data quality is another challenge. Because the data are a subsample and are sourced from Kaggle there is an outstanding question about whether individuals included in one of the two data updates form 2011 and 2012 could be repeated. There are no repeated id variables in the dataset, but this does not necessarily mean that all the data are unique. We are searching for a definitive answer to this question.
- How to deal with lots of missing values.

Sources

1. 'Predict Mortality/Death Rate.' URL: https://www.kaggle.com/rajanand/mortality