Final Destination: Predicting Death in India

Nan Chen, Andrew Thappa, and Zihui Zhao

Abstract– Using an individual-level dataset gathered across 9 Indian states and composed of 19 predictors we try to predict the response variable age of death. First, we run four linear models in order to predict age of death which can take integer values from 0 to 99. The resulting models lacked predictive efficacy and as such, we binned the response variable into groups from 0-33, 34-66, and 67-99 and ran two classification models, random forests and boosting. We find binning improves prediction accuracy compared to the continuous setting. By plotting importance of variables, we find that asset possession, location of residency, hygiene conditions and living habits are essential for lifespan.

	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

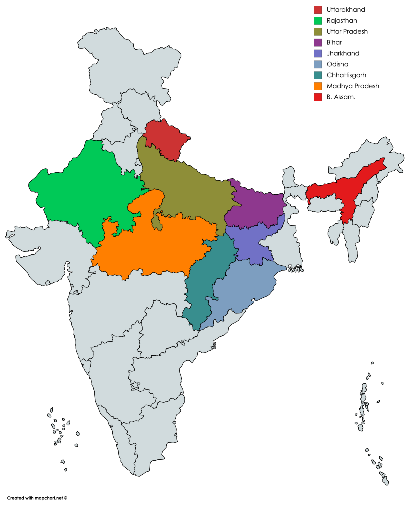
Introduction

Our goal is to create a model in order to predict death age given an individual-level dataset composed of 103 predictors for 770,000 Indian citizens across 9 states in a single year, 2010. Our underlying assumption is that demographic factors, hygiene conditions, lifestyles, current health conditions, community environment will contribute to death at certain age.

Data

For this project we 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 (Predict Mortality/Death Rate, 2017). 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.

While each of the datasets have unique ID's, the data documentation does not make clear whether the two update datasets include previously sampled individuals or whether IDs' match across different years. As such, we will only use the first baseline sample survey data collected in from 2010-11 which has the added benefit of simplifying our analysis because we will not have to account for time.



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.

First, we append nine dataset together, each of which contains observations from one state in India. After merging, the combined dataset includes 125 variables and 769,582 observations.

Second, we only keep variables we deem relevant to predicting the age of death. For example, different codes that identify an individual are redundant. In terms of whether an individual has radio, television, computer etc., we only keep one variable that is enough to capture living and hygiene conditions for an individual.

Third, we drop observations with irregular values. As an example, the data dictionary tells us that the variable rural can only take on the values 1 and 2, yet there were values outside this set, which were dropped.

Finally, despite having data from multiple years we only keep the first year. Doing this ensures that no time-series techniques are necessary which greatly simplifies our analysis. Additionally, we drop observations that contain null and missing values and finally we have 20 variables with 55,267 observations.

Age_of_death_above_one_year is our response variable, that means we exclude samples for people who don't survive their infant periods. The rest variables are predictors of which only asset is a continuous variable. After factorizing predictors, we get a clean dataset which was stored as a CSV on Github.

Methodology and Analysis

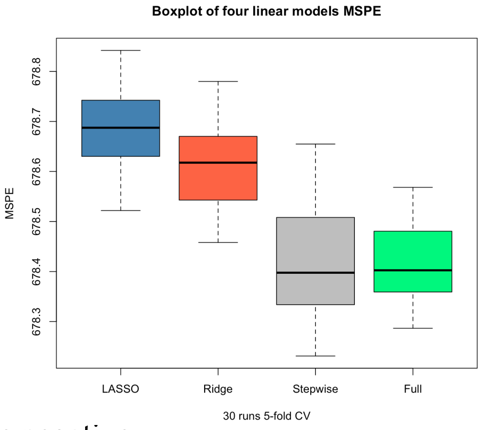
Our initial research question centered around predicting the age of death given an individual-level dataset composed of 121 predictors. After choosing our research topic and dataset, we explored the data and identified 19 features we thought would be most relevant to predict the age of death. As a first pass, we assumed the output was continuous and use a full linear model, stepwise AIC, LASSO, and Ridge Regression. However, this approach yielded disappointing results.

As such, we reframed our question as a classification problem by taking our response variable, age_of_death_above_one_year, and binning it into the categories 0-33, 34-66, and 67-99. The age cutoffs were chosen because This binning results in 19,041, 15,752, and 20,474 in each of the respective categories.

	Variable	Data Definition
1	treatment_source	Source of medical attention before death
2	state	States with different labels
3	factors_contributing_death	Top two factors contributing to death in order of priority 1
4	sex	Male-1/Female-2
5	marital_status	Marital Status
6	occupation_status	Occupation Status during last 1 year
7	smoke	smoke type
8	drinking_water_source	main source of drinking water
9	toilet_used	Type of toilet facility mainly used
10	as	Asset Score
11	death_symptoms	Symptoms leading to death
12	factors_contributing_death_2	Top two factors contributing to death in order of priority 2
13	rural	Rural-1/Urban-2
14	religion	Religion
15	highest_qualification	Highest educational qualification attained (for age 7 years above)
16	chew	Chew Status
17	alcohol	Consume Alcohol
18	is_water_filter	Yes-1/No-2
19	household_have_electricity	Yes-1/No-2
20	age_of_death_above_one_year	1 year and above (in completed years)

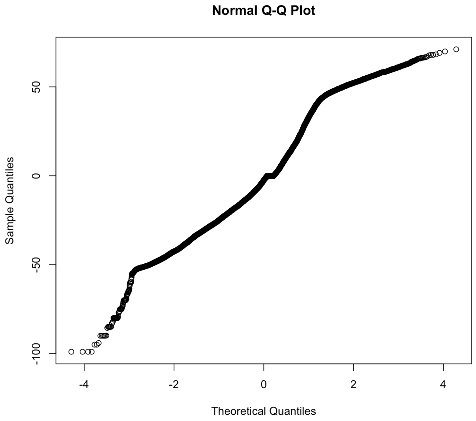
Continuous Response Variable

In our initial analysis, we took the response variable, age_of_death_above_one_year as continuous and tried to predict it with our design matrix. In predicting, we consider four regression models: Full Linear model, Stepwise AIC, LASSO, and Ridge Regression. To compare, we considered the mean-squared prediction error (MSPE) and chose the model that minimized this value using 30 runs of 5-fold cross validation.



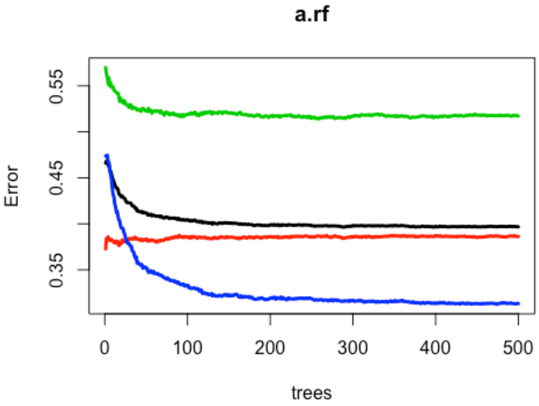
One issue when predicting with regression was that there were negative predicted ages which occurs because our variable as, which represents assets of an individual has a negative mean for the entire population. To mitigate this, we set death_age_above_one_year to zero if they are predicted to be negative.

The Stepwise and Full linear model minimize mean-squared prediction error, but in terms of variation, the full linear model has a smaller variance as shown in Figure 1. Examining the residual diagnostic, we don't see any systematic pattern in the residual plot, and can therefore conclude that the full linear model can be used to predict individual's death age provided he or she survived at least one year. But when we look into the actual value, the MSPE of these four models are all above 670. Of these models, the best one is the full linear model which has a MSPE with mean 678.4 and smallest variance. Although lower bound of MSPE in stepAIC is the lowest among all four linear models, it is still a large value. If we take square root of 670, it is approximately 26 which indicates a large and inaccurate bandwidth of death age prediction.



Categorical response variable

When dealing with our continuous integer response variable, the four regression models do not perform well. To salvage the analysis we recast the problem as one of classification. We divide individuals in our dataset into three age groups based on a criteria that separates them based on general life stages. This is justifiable because, due to the inherent randomness of life, it is hard to predict death age precisely.

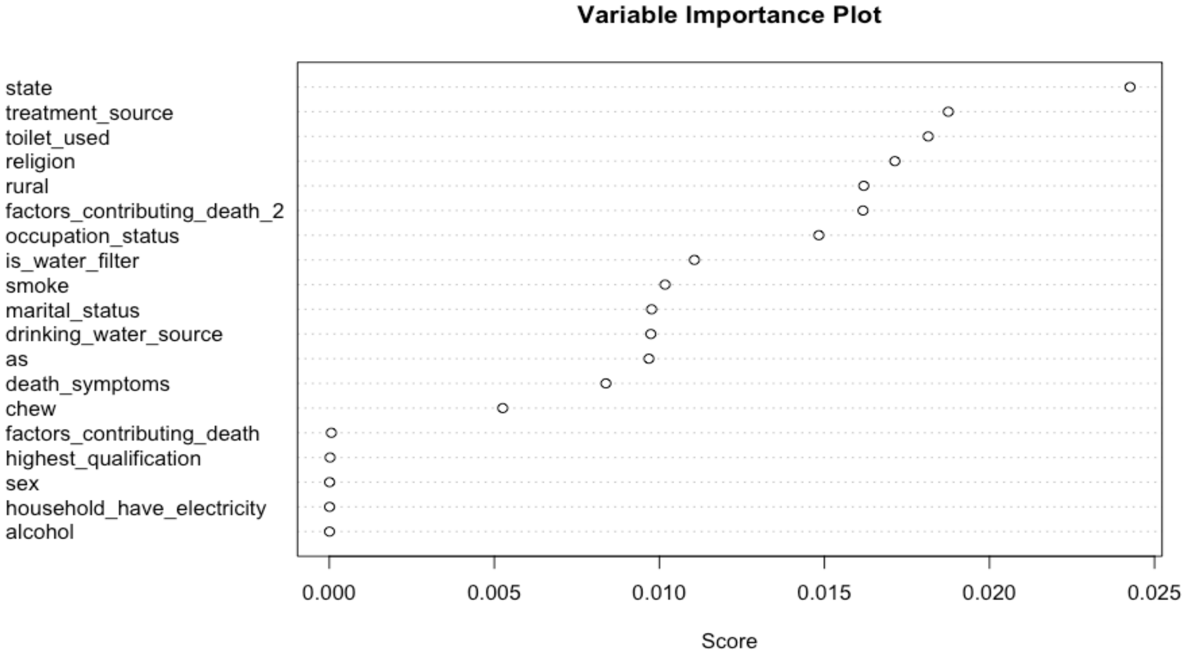


Thus we divide people into three age groups, which we hypothesized would allow us to better predict age of death. Intuitively, this makes sense because actual year of death is hard to predict due to random accidents that might occur at any point. But, whether people die in an early age when they have not entered labor force, or die after they retire will be more predictable given economic resources and demographic characteristics which are variables captured in our design matrix.

To do this we chose 0-33, 34-66, and 67-99 as year cutoffs. These bins are fairly well balanced with 19,041, 15,752, and 20,474 individuals in each respective bin. To classify individuals, we run Random Forest and Boosting.

First, we ran a random forest with a default value of 500 trees. It is reasonable to use 500 because classification error rates don't change much after 200 iterations as Figure 3 shows. The blue line represents age (66,99] with the lowest error rate. The average OOB error rate is 39.3% with 4 variables tried at each split.

Second, we run a boosting model. Boosting grows trees sequentially by incorporating information from previously grown trees. In other words, boosting avoids overfitting by learning slowly. With boosting the key parameters one can tune are the number of iterations as well as the number of splits (d). In *ISLR*, James et al. note that d controls the complexity. We set the depth to 2 and the number of trees to be 500. After fitting the model we observe prediction error of 40.44% and this ensemble error was invariant to the number of trees/iterations.



Challenges/Obstacles

The main challenge we faced in this project was figuring out how to handle relatively novel survey data and frame it in a way that would allow us to develop a model we could use to derive predictions.

- With the benefit of hindsight it is clear that a dataset like the one used for this project is best suited to help answer inferential questions about specific regions rather than being used for predictive purposes.
- Running k-fold cross validation for boosting to find the proper number of iterations & tree depth is time consuming. Instead, This corresponds to a 2-dimentional optimization problem, the first attempts takes more than 6 hours. Given time and resource limits, we only solve it by trying different combinations of these two parameters. We could have mitigated this by utilizing R packages to take advantage of parallelization and/or running all analyses on a remote server. However, due to time constraints, setting up this infrastructure was infeasible.
- When examining the Variable Importance Plot, you can see that the "most important" predictor is state. However, the fifth "most important" predictor is rural. We would expect these to capture much of the same information which is problematic.

Results/Conclusion

In predicting the age of death of individuals, our variable importance plot shows that the state in which one lives is of prime importance. This makes sense given that the states from which the sample was derived include B. Assam and Bihar which are two of the poorest regions in India. Hygiene conditions such as whether an individual uses a toilet, whether water is filtered or not, and the source of drinking water also play a role in explaining the age of ones death.

In short, our model tells us what is already common sense: if one lives in a state that is rural with bad drinking water and no toilet, our model predicts you will die at an earlier age. In this sense, our model is not terribly novel. Perhaps the biggest takeaway is that using survey data to develop a predictive model is inherently challenging since most variables are categoricial.

Future Exploration

- We would ideally want to run the cross validation analysis with different partitions of the response variable. Currently, the response variable is segmented into 3 buckets: 0-32, 33-65, and 66-99. Ideally we would cross validate different partitions including increments of 5 and 10.
- Ideally, we need 2 optimal parameters to perform boosting : iteration & maxdepth. This corresponds to a 2-dimentional optimization problem, given time and resource limits, we only solve it by trying different combinations of these two parameters. Generally you can set your depth parameter to be 1 but we would like to check out different values of the iteration parameter.

Sources

1. "Predict Mortality/Death Rate." 2017. Kaggle. Accessed November 26, 2017. https://www.kaggle.com/rajanand/mortality.
2. James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2013). *An Introduction to Statistical Learning: with Applications in R*.