Predicting Under-5 Mortality in India

Sushmita V Gopalan
M.A. Computational Social Science, University of Chicago

Background

•At present, over 35 out of every 1000 children born in India die before their fifth birthday.

•My goal is to develop a model to predict whether or not a child will make it to 5, given a set of easily accessible information about

- •Mother and child's personal biological characteristics
- •Parents's health- related behavior
- •Community-level and socioeconomic variables

Data

•India's National Family Health Survey (NFHS) from 2005-06, part of the Demographic and Health Surveys V.

- •Women's Questionnaire for information on the mother and child. Mother and child's personal biological characteristics
- •Household Schedule for community and socioeconomic indicators..
- •For the logistic regression, I used a subset of 33,642 births for which complete information was available.

Methods

Logistic Regression

-Logit 1:Mother's age, education, religion, wealth index etc.

-Logit 2: Variables in Logit 1 + maternal healthcare details + access to health insurance

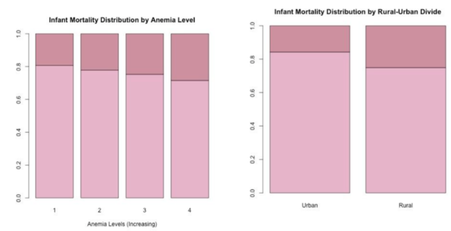
-Logit Stepwise: Stepwise Backward Elimination using Akaiki Information Criterion (AIC) values

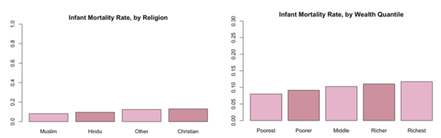
Decision Trees

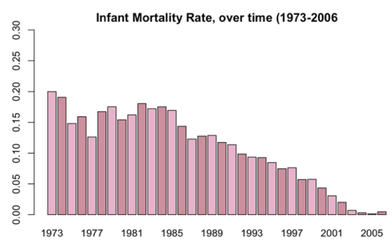
-Default Controls

-Pruned to 6 variables, based on MSE

Some Bivariate Relationships







Results

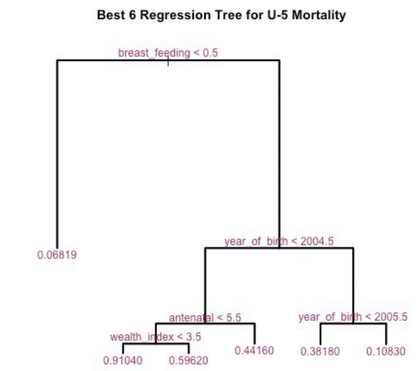
Results of Stepwise Logistic Regression

	Under-5 Death
Multiple Births	0.599*** (0.220)
Scheduled Tribe	0.337** (0.163)
Christian	0.346 (0.263)
Mother's Age	0.018* (0.011)
Wealth Index	-0.176*** (0.050)
Mother's Age at First Birth	-0.046*** (0.018)
Mother's Education	-0.201*** (0.075)
Partner's Education	-0.188*** (0.067)
Birth In Private Institution	-0.385*** (0.140)
Breastfed for More than Six Month	3.262*** (0.132)
Year of Birth	-0.471*** (0.041)
Scheduled Caste* Hindu	0.462 (0.429)
Scheduled Caste	-0.188 (0.406)
Hindu	0.052 (0.148)
Constant	941.594*** (81.331)
Observations	33,642
Log Likelihood	-1,194.868
Akaike Inf. Crit.	2,415.736

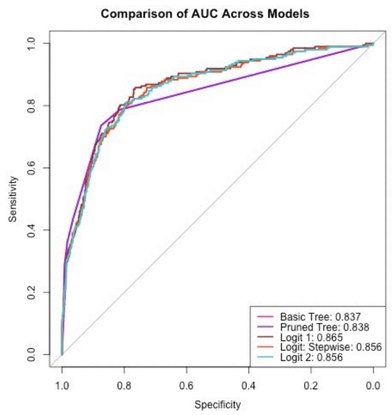
Note: *p<0.1; **p<0.05; ***p<0.01

Stepwise Backward Elimination: In this method, a variable is considered for inclusion to or exclusion from a model in each step and a decision is made based on a predetermined criterion. In this case, the decision was whether or not the change to the model lowered the AIC by a certain amount.

Results



Running a cross-validation check on the decision tree with default controls revealed that a tree of size 6 would have lowest MSE. As the tree shows, the variables it finds to be most important, match, for the most part, with the results of the logistic regression.



All the five models I fit had very similar results in terms of the AUC (Area Under Curve) of their ROC (Receiver Operating Characteristic) Curves. This means that they are almost equally likely to rank a randomly chosen positive instance higher than a randomly chosen negative one.

Conclusion

The variables that are most predictive of under-5 mortality are

- whether or not the child was breast-fed for > 6 months
- family's wealth index
- year of birth
- private, institutional birth
- mother's education

Limitations

-Prudence is warranted while using an algorithm to pick variables to include in a regression because spurious correlations could be flagged as significant.

-Far more information is available regarding children who survived past the age of 5 than those who didn't - this makes training and testing prediction models challenging because if I use the sample as is, even if the model predicted 'Survive' a 100% of the time, it would be right close to 96% of the time. To this end, I used a random, smaller subsample of the data on alive children. Other methods exist however, to analyze such 'rare events' such as Firth's Penalized Likelihood Method.

Call Me, Maybe!

To talk more about public health, please email me at sushmitavgopalan@uchicago.edu!