METHODS AND INITIAL RESULTS

MACS 30200, Dr. Evans
Sushmita V Gopalan

Methods

As I described in my proposal and literature review, I will first use a logistic regression model to identify the key variables that predict under-5 mortality in the Indian context, i.e. I will explore a host of socio-economic variables, self-reported data on maternal health, health-seeking behaviour and some anthropometric measures (for the mother and the child) that are easily obtained, to predict the likelihood of the child making it to their fifth birthday.

I start with a standard logistic regression model, as specified below. My response variable Y takes on the value 1 if the child dies within five years of birth and 0, if they survive.

$$P(Y = 1|\, x) = \frac{e^{x\beta}}{1 + e^{x\beta}}$$

The probability that event Y = 1 will occur, conditional on a covariate vector **x** is determined by a logistic function of the vector **x** and the vector of coefficients β.

As I discussed in my literature review, the factors that have been found to be predictive of infant mortality can be classified into three major categories

- Personal and biological characteristics of the mother and the child - weight with respect to median, malnutrition, age of the mother, time since previous birth, etc.
- Parents's health status and behaviour - smoking, drinking, dietary habits, tetanus, anemia, attitude towards health seeking, awareness levels
- Community - sanitation, public health facilities, access to insurance, communicable diseases, cultural attitudes toward health care, women's empowerment, etc.

At the moment, I have access to the first category of variables from the mother's questionnaire. The second category of variables will become available when I have linked records between the women's questionnaire and the household and men's schedules. The third category of variables will involve the usage of multilevel modeling, which I am currently learning about and hope to apply soon. Note that all of these categories are particularly important for the consideration of under-5 mortality - this includes neonatal mortality, postnatal mortality, infant mortality and child mortality and is thus influenced by a variety of characteristics. Some

Song et al. (2004) suggest that the relationships between health outcomes and their predictors are highly non-linear. After I complete the logistic regression using the rich set of variables proffered by the NFHS, I will proceed to use a decision tree approach, such as that used by Tesfaye et al. (2010) to classify children into high-risk and low-risk categories, given the values they take on for key variables and contrast its predictive ability with the logistic regression. Time permitting, I would like to try to use a neural net to identify a handful of key variables and then

use pruned decision trees to arrive at a simple set of predictive rules, as Chen at al. (2011) did, while building a model to predict preterm birth.

Data Source

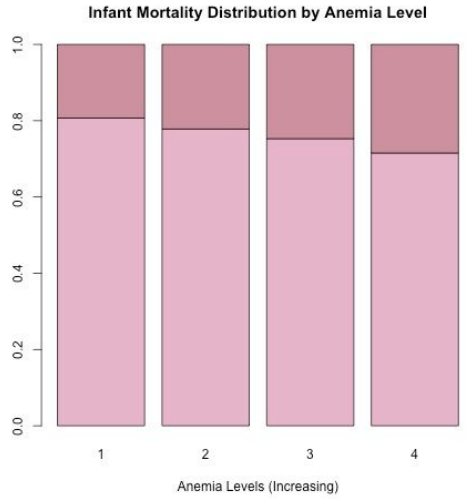
The National Family Health Survey (NFHS) is a nationally representative survey that is conducted in India every ten years. It is carried out by the Ministry of Health and Family Welfare, Government of India, along with the International Institute for Population Sciences, Mumbai. The last two rounds were carried out in 2015-16 and 2005-06, respectively. Funding for the NH-3 was obtained from the United States Agency for International Development (USAID), the Department for International Development (DFID), the Bill and Melinda Gates Foundation, UNICEF, the United Nations Population Fund, and the Government of India. For this paper, I will be using data from NHFS-3 (2005-06). The NFHS is a monumental operation - up to 18 different research organizations conducted interviews with over 230,000 individuals over the course of 18 months.

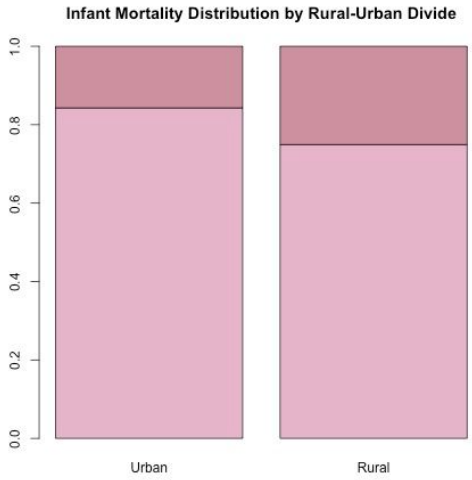
The schedule of primary interest to me at this point is the Women's Questionnaire. It covers a range of questions on deliveries, infant-care, maternal health-care, nutrition, a few questions on the woman's agency and safety in the household and as of the NFHS-3, a host of anthropometric measures such as height, weight, hemoglobin levels for women and children.

As a first pass, I am only considering the Women's Questionnaire at this point. Note, however, that several key variables like family income, access to health insurance, etc. are only obtained from the Men's or the Household Schedules. At the outset, there were 124,385 recorded surveys. From these, I dropped all observations about women who had never given birth to a child. This brought the sample down to 84,609. Further, for each woman interviewed, information was recorded on up to 20 past births she had given. I extracted this data and converted it from wide to long, so that my unit of analysis while running my logistic regression is 'births' and not 'mothers'.

I will first present some key summary statistics that describe the sample of women interviewed.

Statistic	N	Mean	St. Dev.	Min	Max
age	84,609	33.180	8.125	15	49
no_of_children	84,609	3.035	1.791	1	16
no_dead	84,609	0.305	0.712	0	10
education	84,609	1.109	1.038	0	9
anemia	77,364	1.704	0.763	1	4





The histogram on the left shows that a larger proportion of women with more severe anemia have had a children die. Similarly, the histogram on the right shows that a larger proportion of women living in rural areas lose children than women in urban areas.

I ran my logit regression on a random subset of the sample, due to constraints of computational effort. The results show that some variables are indeed significant at the 0.01 level, but as mentioned above, several more variables are required for the prediction to be meaningful.

	Dependent variable:
	u5-death
sex	0.070**
	(0.032)
age_first_birth	0.018***
	(0.006)
year	-0.096***
	(0.002)
anemia_severe	-0.200*
	(0.121)
years_educ	0.014***
	(0.004)
Constant	188.528***
	(4.521)
Observations	50,000
Log Likelihood	-14,019.170
Akaike Inf. Crit.	28,050.340

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