***Job Market Paper***

# CHILD MALNUTRITION, NEONATAL AND INFANT MORTALITY: NEW INSIGHTS LEARNT FROM 'MACHINE LEARNING'[§]

- Dweepobotee Brahma[1]
Western Michigan University

## ABSTRACT

This paper investigates neonatal mortality and infant mortality in India using a large household survey data and by employing multiple parametric and non-parametric "Statistical Learning" techniques. The Statistical Learning methods address two common econometric challenges- predicting 'rare-events' and obtaining valid inference in the presence of many correlated covariates. Several Machine Learning (ML) methods along with traditional logistic regression are used to predict the incidences of neonatal and infant mortality, with Learning methods displaying substantially higher prediction accuracy than the conventional logistic model. Using the results this paper tries to identify a 'high-risk mortality group' for babies to help implement more targeted health policies. Additionally, a selection and shrinkage-based machine-learning technique (LASSO) is used to select from a large set of predictors for infant malnutrition measured by weight-for-length Z score. Inference is performed on the selected model using Statistical Learning based post-selection inference method. Based on these exercises we try to assess as well as recommend certain policy measures for combating child mortality and malnutrition.

**Keywords:** Neonatal and infant mortality, Machine Learning, Imbalanced Data, Child malnutrition, Post-Selection inference.

**JEL Classification:** C52, C53, I12, J13, O53

[§] I would like to thank Dr. Debasri Mukherjee, Department of Economics, Western Michigan University, Dr. Eskander Alvi, Department of Economics, Western Michigan University, Dr. Kevin Lee, Department of Statistics, Western Michigan University and Dr. Yuekai Sun, Department of Statistics, University of Michigan for helpful comments and discussions on this work. All remaining errors are mine.

[1] 5438 Friedman Hall, Department of Economics, 1093 W. Michigan Avenue, Kalamazoo MI 49008
Email: dweepobotee.brahma@wmich.edu

## 1. INTRODUCTION

The issues of persistent high rates of infant mortality and child undernutrition in developing countries have worried social scientists, medical professionals and policy-makers alike. These have been the focus of the United Nation's *Millennium Development Goals 2015* and subsequently the *Sustainable Development Goals 2030*. While India has made considerable progress in this regard there is still a long way to go. As of 2015, India accounts for one-sixth of the world's population but alone bears the burden of *one-fourth* of the world's neonatal mortality and *one-third* of the world's malnourished children. This paper investigates the trio of neonatal mortality, infant mortality and child malnutrition in India using cutting-edge 'Statistical Learning' techniques (based on Machine Learning (ML) algorithms) which provide newer perspectives[2] to these issues as well as help overcome two challenges faced by traditional econometric techniques – prediction of 'rare-events' and robust inference in the presence of many correlated covariates.

The first econometric challenge in development economics while investigating child mortality using survey data is that the incidence of mortality is after all a 'rare event'. Rare-events in econometrics are characterized by a binary dependent variable with few ones (events) and many times higher number of zeros (non-events). Conventional techniques like logistic techniques underestimate the conditional probability of such events (neonatal and infant mortality in this paper) in heavily imbalanced data (King and Zeng 2001; Firth 2008). This is similar to the challenges faced in the finance and health literature in predicting fraudulent transactions, loan defaults or mortality from specific diseases. Although such events are rare by nature, their consequences are grave, demanding serious attention. New machine-learning (ML) techniques have been proven better in an empirical sense compared to the traditional statistical/econometric methods in predicting such rare events. This paper uses several parametric and non-parametric ML algorithms (to be explained later) and combines them with commonly used re-sampling techniques (to be explained in detail later) to model and predict neonatal and infant mortality in India and compares their prediction performance to that of logistic regression. To the best of our knowledge, this is the *first attempt* to utilize the wisdom from the machine learning literature in the context of development economics. In the process of obtaining the best predictors, this paper pin-points certain socio-economic-biological characteristics of mothers and babies to identify 'high-mortality risk' groups and suggests that more aggressive policy interventions be taken for them in particular. The identification of this 'high-mortality risk' group can be particularly useful for public health policies such as the India Newborn Action Plan (INAP). Additionally, we use a post-selection-inference (PoSI) technique (to be explained later) as well as standard logistic regression to examine the roles of such characteristics on mortality from the inference angle. This is the *first contribution* of our study.

The second econometric challenge commonly faced in development economics is the presence of many correlated covariates associated with malnutrition necessitating the need for a sparsity

[2] The utility of the fresh perspective that these techniques can provide has been pointed out recently by (Varian 2014) and (Mullainathan and Spiess 2017).

analysis and model selection[3]. However, "*estimates in selected models should tend to generate more type I errors than conventional theory allows because the typical selection procedure favors models with strong, hence highly significant predictors*" (Berk et al. 2013). Modern post-selection inference (PoSI) techniques (to be explained later) can largely guard against such pitfalls. Our study employs LASSO method to select important predictors of child malnutrition (measured by weight-for-length Z score) from a large set of demographic, socio-economic and hereditary indicators, dietary practices, health status, health beliefs, and policy variables, and uses post-selection-inference (PoSI) technique to perform inference on them. It examines whether the regression coefficients of the 'selected' predictors are statistically significant or not. See Lee, Sun, Sun and Taylor (2016) and Taylor and Tibshirani (2018) for details. This work is the *first attempt* to utilize such Statistical Learning technique in development economics, and to consider the most comprehensive set of covariates from the framework suggested by Mosley and Chen (1984) for examining infant malnutrition in India - the *second contribution* of the work.

The rest of the paper is organized as follows. Section 2 introduces the issue of neonatal and infant mortality and child malnutrition in India and discusses the relevant background literatures from development economics, health and machine-learning. In section 3 data and predictors are discussed. In section 4 estimation strategies, machine-learning techniques and evaluation criteria are presented. Section 5 discusses the results. Finally, section 6 concludes with a discussion on future research to follow.

## 2. BACKGROUND OF THE PROBLEM, POLICIES UNDERTAKEN AND ASSOCIATED LITERATURE

### 2.1. Overview of the problem in India and the associated policies

Between 1990 to 2015[4], the Indian economy experienced rapid growth with per capita GDP growth rate peaking at 8.8% in 2010. During the same time, India's infant mortality[5] rate fell by 59% (from roughly 88 deaths to 36 deaths per 1,000 live births) and the child mortality rate[6] fell by 64% (from roughly 123 deaths to 45 deaths per 1,000 live births), but neonatal mortality rate[7] fell by only 53% (from 57 deaths to 26 deaths per 1,000 live births). As a policy response India instituted the Janani Suraksha Yojana (JSY) which provides cash-incentives to mothers and health workers for institutional deliveries and adopted the India Newborn Action Plan (INAP)[8] which aims to bring down India's neonatal mortality rate to single digits by 2030. The leading causes of neonatal mortality and infant mortality in India are prematurity, intra-partum complications, neonatal infections and vaccine-preventable diseases (The Million Death Study Collaborators 2010). As of 2015, neonatal mortality comprises 73% of infant mortalities and 59% of child

[3] Note, that in some of our sample groups the number of potential predictors often comes close to the number of observations approaching a near high-dimension setting.

[4] World Bank estimates.

[5] Infant mortality rate refers to the number of deaths before 1 year (0-11months) per 1,000 live births.

[6] Child mortality rate refers to the number of deaths before 5 years (0-59 months) per 1,000 live births.

[7] Neonatal mortality rate refers to the number of deaths before 28 days from birth per 1,000 live births.

[8] For more details see http://nhm.gov.in/india-newborn-action-plan.html

mortalities in India. *Thus, the issue of neonatal mortality in India warrants a separate investigation along with that of infant mortality.*

In addition to a high burden of child mortality, 43% of children under five years suffer from malnutrition with 20% suffering from wasting due to acute undernutrition in India (UNICEF, India). Malnutrition continues to be the leading cause of child disability, deaths and indirectly leads to deaths from other infectious diseases by reducing immunity and delaying recovery. In 1975 India launched the Integrated Child Development Scheme (ICDS) to combat malnutrition through Anganwadi Centers (AWD) across rural areas. Within India, the incidence of malnutrition is higher in rural areas and in the Empowered Action Group (EAG) states of Bihar, Chhattisgarh, Jharkhand, Rajasthan, Madhya Pradesh, Uttar Pradesh, Uttaranchal, Orissa and Assam (Radhakrishna and Ravi 2004). Figure A.2 in the Appendix illustrates these states.

## 2.2. Background Literature

The background literature for this paper spans the disciplines of development economics, machine-learning and health policy.

### 2.2.1. Relevant Literature on child mortality and child malnutrition

The literature in development economics has conventionally investigated neonatal and infant mortality from a macro perspective (national, state or regional level analysis) focusing on mortality rates, instead of investigating from a micro perspective using mortality incidences or events. For example, Bhalotra (2007) find a negative impact of state health expenditure on infant mortality rates in India using the National Family Health Survey-II. Other studies have investigated neonatal and infant mortality in specific states or geographical areas in India (Singh, Kumar, and Kumar 2013; Arokiasamy and Gautam 2008). Mosley and Chen (1984) has rightly pointed out that literature focusing on the prediction of ‘child-death event’ using survey data is scant in economics and calls for serious attention of the researchers. Although some limited studies such as Bhalotra and Soest (2008) investigate the incidence from a micro perspective using National Family Health Survey-II, their focus is on just one state/region in India. They find that the incidences of older sibling deaths lead to subsequent neonatal deaths by shortening birth-intervals-- evidence of which is also found in this study. Rammohan, Iqbal, and Awofeso (2013) investigate the link between the availability of emergency obstetric care facilities as well as ‘distance to the district hospital’ to the neonatal mortality in rural areas using an older dataset (compared to ours). But, none of these studies have focused on the issue of *prediction accuracy and appropriate predictor search,* which the present paper aims at. This is also the *first* comprehensive study of neonatal and infant mortality in India that *addresses* the issue of ‘rare-events’ using national survey level data in a micro setting.

Although most of the irreversible damages due to malnutrition occur very early in childhood with growth faltering starting as early as 3 months and continues until 12 months (Victora et al. 2010), literature on *infant* malnutrition is sparse and it is even more scant for India. Duflo (2003) found that conditional cash transfers to elder female recipients in South African households improved weight-for-height Z-scores in girls but not for boys. Engebretsen et al. (2008) study mother-infant pairs in Uganda and find that low household wealth, poor infant feeding practices, age, gender and number of children are associated with lower weight-for-length Z scores. Jalan and Ravallion

(2003) find that the incidence and duration of diarrhea causing rapid weight loss (leading to low weight-for-length score) is significantly lower in families with piped water compared to families without it. The only study on early-childhood (age 0-2) malnutrition in India is by Jain (2015) which used data from an earlier survey (compared to ours) and found that under the ICDS scheme receiving supplementary nutrition raised children's height only by around 1 cm. To the best of our abilities we do not find any literature investigating malnutrition in infancy in recent times in India using a comprehensive set of covariates and this paper aims to fill that gap.

### 2.2.2. **Relevant Machine Learning literature**

The problem of predicting events is ubiquitous and machine-learning techniques find applications across health (in predicting specific diseases, epidemics), finance (predicting fraud), political science (presidential vetoes, coups, wars), information technology (software malfunctions). In the field of political science Muchlinski et al. (2016) apply Random Forest, Ordinary Logistic Regression, Firth Logistic Regression, Logistic LASSO Regression to predict civil wars where the incidence of civil wars is 1% (like ours) and find that Random Forest has the highest prediction accuracy. In the field of medicine, Khalilia, Chakraborty, and Popescu (2011) use data on individuals' health status and utilization of healthcare services and compared the performance of multiple machine-learning techniques in predicting disease risk where the incidence of disease ranges from 0.01% to 29.1% and find superiority of Random Forest. In the world of finance, Bhattacharyya et al. (2011) use data on credit card fraud (where the incidence of fraud was between 2% to 15%) and compared Random Forests, Support Vector Machines with Logistic Regression and find that Random Forest has higher overall prediction accuracy than the others. The applied machine-learning literature on prediction is vast with little consensus about the best algorithm in terms of predictability. Each dataset has its own unique challenges and there is no one-size-fits-all solution to predicting such events, thus necessitating the adoption of 'data-mining' techniques. We apply multiple parametric and nonparametric ML methods (to discuss below) and compare their prediction accuracy. Among the algorithms that are used in this paper, we also find the superiority of Random Forest and Boosted Trees (nonparametric ML methods) over standard logistic model or logistic LASSO.

Out of most machine-learning techniques, LASSO performs variable *selection* which is important when the set of covariates is very large as in our study in case of malnutrition. However, the coefficient estimates on the selected predictors generated by LASSO are biased and cannot directly be used to conduct inference. Additionally, as stated earlier, inference using classical statistical theory on selected predictors tend to produce higher than accepted rates of Type I errors since a "*typical selection procedure favors models with strong, hence highly significant predictors*" (Berk et al. 2013). As Berk et al. (2013) point out "*The reason for the invalidation of classical inference guarantees is that a data-driven variable selection process produces a model that is itself stochastic, and this stochastic aspect is not accounted for by classical theory.*" This necessitates the adoption of a post-selection inference technique. There is an emerging body of literature on post-selection inference techniques each with its own advantage and relevance (see Bühlmann 2013; Belloni, Chernozhukov, and Hansen 2013; Lee et al. 2016; Taylor and Tibshirani 2018 for more details). The PoSI technique popular in economics by Belloni, Chernozhukov, and Hansen (2013) focusses on post-selection inference on one treatment effect parameter. However, in this study we are interested in the PoSI on multiple policy parameters and use a PoSI technique following Lee, Sun, Sun and Taylor (2016) and Taylor and Tibshirani (2018) that obtains p-values to test if the regression coefficients in the selected model are statistically significant or not. To the best of our knowledge this is the *first* study to apply a PoSI technique in development economics.

## 3. DATA AND PREDICTORS

The dataset used in this study comes from the second round (2011-2012) of the Indian Human Development Survey (IHDS-II). The survey identified one ever-married women between the ages of 15-49 in each household and asked her a series of questions about the antenatal and postnatal care received for the last live birth since 2005. The survey also recorded the heights and weights for all the children in the household which was used to construct the weight-for-length anthropometric Z score.

### 3.1. Data and Predictors for Neonatal and Infant Mortality

The data on neonatal and infant mortality consisted of more than 7,000 observations on the last live birth since 2005 of which approximately 1% were neonatal and infant deaths. Deaths within 1 month[9] of birth are considered as neonatal mortality and deaths within 1 year of birth are regarded as infant mortality. The choice of the predictors was guided by (i) the previous literature in development economics and medicine and (ii) availability of data. They were compiled and merged from the three different modules of the survey - ‘eligible women’, ‘individual’ and ‘household’ modules. Early medical intervention is accounted for using the predictor ‘newborn care’ which measure whether the newborn received the first medical checkup within seven days of birth (for neonates), and within one month of birth (for infants). The predictor ‘lab-test’ measures whether the mother has received an intensive medical checkup including a blood test and a urine test that requires access to laboratory testing services. The variables ‘Tetanus Toxoid’ and ‘folic acid’ measures whether the mother received Tetanus Toxoid injections and folic acid tablets respectively during pregnancy. Prematurity and poor fetal growth which are leading cause of neonatal and infant mortality are measured using ‘low birth weight’ where ‘1’ refers to babies who were born small or very small and ‘0’ refers to babies who were born average or large. The variables ‘pregnancy-complication, ‘delivery-complication’ and ‘post-complication’ measure whether the mother had any medical complications during pregnancy, during delivery and in the two months following delivery respectively.

While ‘income’ measures real annual household income, ‘urban’ is a binary variable. The indicator variable ‘above poverty line’ measures whether the household is above the World Bank’s $1.90 estimate for the poverty line. ‘Forward group’ is a binary covariate that controls for the composite effect of caste and religion. Here ‘0’ includes the marginalized populations of Dalits, Adivasis,

[9] Strictly, neonatal deaths refer to deaths in the first 28 days of life but deaths within 30 days are considered in this paper to allow for age-heaping.

Muslims and other backward castes (OBC) while ‘1’ includes the non-marginalized populations of Brahmins, Forward caste Hindus, Christians, Jains, Sikhs. Characteristics of the mother’s fertility history is measured using the variables such as ‘total births’, ‘prior deaths’ (which measures if any of the mother’s previous births died) and ‘first-born’ (measuring if this birth is the mother’s first-born). The amount of cash transfer received by the mothers from government schemes (JSY) to promote institutionalized delivery is included as ‘JSY delivery money’. In the analysis for infant mortality, predictors ‘BCG’, ‘DPT’, ‘measles’ and ‘polio’ measures whether the infant received the Bacillus Calmette-Guérin vaccination, the vaccination against diphtheria-pertussis-tetanus, the measles vaccination, and the polio vaccination respectively. Table A.1 in the Appendix presents a tabular description of the predictors and Tables A.3 and A.4 present their descriptive statistics.

### 3.2. Data and Predictors for Malnutrition

For the sample of children under 1 year who remained alive at the time of the survey, anthropometric weight-for-length Z score[10] was constructed using the standards in WHO Growth Reference (2006) and is detailed in (Mercedes 2008). A low weight-for-length score also referred to as ‘wasting’ or ‘thinness’ indicates insufficient weight gain relative to height or recent or continuing severe weight loss. The Z score is widely used as a measure of the nutritional status of the population. Using the Z-score to determine the nutritional status of infants has several advantages. First, a fixed Z-score interval implies a fixed difference in height or weight of children in the same age making it interpretable and comparable across cohorts. Second, Z-scores are independent of gender and thus allow the evaluation of children’s growth status by combining sex and age cohorts.

An infant’s nutritional status is determined by a large number of factors ranging from infant feeding practices, family dietary practices, household socio-economic characteristics, family nutritional status, hereditary characteristics, mother’s health beliefs, incidence of diseases and policy outreach, as outlined in Mosley and Chen (1984). Over 60 predictors were included encompassing these various aspects[11]. To account for the wide range of dietary practices in India, the relative monthly expenditure of the family on various kinds of food are included using the predictors ‘Relative carb expenditure’ ‘Relative meat expenditure’ and ‘Relative milk expenditure’. The predictor ‘three square meals’ captures whether the household members typically eat three or more meals per day.

The socio-economic characteristics of the household are included through ‘income’ (same as before), ‘household consumption’ which measures the per capita monthly household consumption expenditure, primary source of income for the household using binary predictors for ‘agriculture’,

[10] Weight-for-length Z score was constructed the ‘zscorer’ R package following the WHO Growth Reference (2006) standards.

[11] In case of the mortality, the data comprised births over a period of seven years (last birth between 2005 to 2011) and hence many of these covariates measured at the time of survey may not be relevant at the time of mortality and hence were excluded.

‘artisan/business’, ‘rent/pension’ with salaried households as the reference category. ‘Water’ measures whether the household has access to safe drinking water and ‘purify water’ captures whether the household purifies the drinking water either through boiling or filtering. ‘Toilet’ captures whether the household has a flush/semi-flush toilet facility. The range of caste and religion which affects dietary practices are accounted for using dummies for ‘Brahmins’, ‘Dalits’, ‘Adivasis’, ‘Muslims’, ‘Other Backward Castes (OBC)’ with ‘non-brahmin forward Hindus’ being the reference caste. ‘Below Poverty Line’ captures whether the household is located below the World Bank’s $1.90 estimated poverty line. The size of the household is included using the ‘total household members’. Overall family demographic characteristics are included through ‘male household head age’, ‘female household head age’, ‘male household head education’ and ‘female household head education’. The age composition of the household members can have opposing effects on the nutritional status of children. The presence of elderly family members requiring care can divert the mother’s attention away from her children causing the children to be less healthy. On the other hand, elderly family members who receive pension can contribute to household income and raise the overall nutritional status of children in the household—as discovered by Duflo (2003) in the context of the South African pension scheme. To account for both possibilities the predictor ‘old pension’ is included which measures whether there are any pension recipients in the household, and ‘dependency ratio’ is included which captures the dependency ratio in the household.

Transitory shocks to the household are likely to affect the nutritional status of children and are included using the dummy variable ‘household shock’ that takes ‘1’ if the household suffered a major loss of expenditure resulting from illness/accident/ drought/flood/fire/loss of job/ death/crop failure. The location of the household is accounted for with three dummies measuring ‘developed village’, ‘non-metro urban’ and ‘metro urban’ with ‘less-developed rural’ being the reference category.

The mother’s socio-economic and demographic characteristics are including using ‘mother’s age’, ‘mother’s education’, ‘mother’s BMI’, ‘mother’s employment’. Infant feeding practices are included through ‘breast fed within 1 hour’ which measures whether the newborn was first breastfed within an hour of birth, and through ‘Breast fed for six months’ that measures whether the infant was breastfed exclusively for six months. Women’s exposure to mass media potentially improving infant feeding practices as well as various health beliefs is captured in ‘women media’ variable measuring whether the mother regularly watches television, reads newspapers or listens to the radio. The mother’s health beliefs are accounted for using three dummies ‘health belief about first milk’ taking the value ‘1’ if the mother believes that the first milk is good for the baby, ‘health belief about malaria’ taking the value ‘1’ if the mother believes that mosquitos cause malaria, and ‘health belief about diarrhea’ using the value ‘1’ if the mother believes that it is important to drink more water during diarrhea. The mother’s autonomy in intra-household decision is accounted in ‘bargaining power’ which measures if the mother makes the baby’s health decisions. The mother’s fertility related health status is measured by ‘number of children alive’.

Similar to the previous analysis on mortality, vaccinations are included through ‘BCG’, ‘DPT’, ‘measles’ and ‘polio’ and premature birth and low fetal growth is included through ‘low birth weight' variable. Policy outreach under the 'Anganwadi scheme' was included with 'AWD vaccines' (measuring whether the infant received the immunization vaccines at the Anganwadi center), 'AWD food' (whether the infant received supplemental food from the Anganwadi center), 'AWD health' (assessing if the infant received health checkups at the Anganwadi centers), 'AWD growth' (measuring if the infant received growth monitoring services from the Anganwadi center, and 'AWD mother' (whether the mother received supplemental food while pregnant or lactating). Similarly, 'vit A' measures whether the infant received vitamin A supplement--an essential micronutrient for children in preventing childhood blindness and the incidence of diarrhea. Personal illness can cause short term weight loss and contribute to a low weight-for-length Z score. To account for this the number of 'days ill in the last 30 days', 'medical expenditure' and a dummy for 'sought treatment' are included. In addition, dummies for 'fever', 'cough', 'diarrhea' and 'ORS' for whether Oral Rehydration Solution was administered are included. Under the Public Distribution System (PDS) in India state governments issue identification documents called 'ration cards' which entitles households to purchase food grains at a subsidized rate. Three types of ration cards were included as three predictors - 'Above Poverty Line ration card', 'Below Poverty Line ration card' and 'Antyodaya ration card' (issued to the 'poorest of the poor' households enabling them to receive higher quantity of subsidized food grains). Table A.2 in the Appendix presents a tabular description of the predictors and Tables A.5 and A.6 present their descriptive statistics.

## 4. ESTIMATION AND EVALUATION STRATEGIES

For ease of understanding Figure A.1 in the Appendix provides a flowchart for the empirical analysis in this paper.

### 4.1 MORTALITY

First, data on the last live birth (based on the survey) have been used to construct predictive models for neonatal and infant mortality using Logistic LASSO, Random Forest, Boosted Logistic Regression and Boosted Trees. This allows us to classify a "high mortality risk" group for policy makers to use as target.

While building these predictive models, the dataset was first divided randomly into training and test set (James et al. 2000). Each of the algorithms (including logistic regression) were fitted (trained) on the training data and this fitted model was used to predict mortality on the test data. Dividing the data into training and test is important because a good in-sample prediction by an algorithm does not guarantee good out-of-sample prediction. The prediction performance of each of the algorithms on the test set was compared to that of baseline logistic regression.

The results from Machine Learning (ML) techniques are relatively harder to interpret compared to traditional regression techniques. The interpretations are derived from various measures of the influence the predictors yield on the response. The statistical theory and accompanying mathematics behind the machine-learning algorithms are reasonably complex and the interested readers are referred to (Hastie, Tibshirani, and Friedman 2009) for a detailed exposition. A simplistic explanation of the ML techniques used in this paper and their interpretations are provided below.

### Logistic LASSO

LASSO[12] (Least Angle Shrinkage and Selection Operator) is a parametric ML technique in statistics that constrains and regularizes the coefficient estimates (shrinks the coefficient estimates to zero) by applying an $l_1$ norm (sum of the absolute value of the coefficients) on the regression coefficients. Mathematically, LASSO coefficients $\hat{\beta}_\lambda^L$ solve the following optimization problem.

$$\text{Minimize} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where *i=1,…,n* is the number of observations and *j=1,..,p* is the number of covariates.

Here λ is the tuning parameter that controls the degree of regularization and is chosen by cross-validation. In the process of improving prediction, LASSO selects some predictors and de-selects others (by shrinking their coefficients to zero) thereby selecting a sub-model. *The mere selection or de-selection of predictors reveal the extent of their influence on the response and can be used for policy evaluation following the suggestions in Mullainathan and Spiess (2017).* In addition, significance test can be sought on the selected predictors using Post-Selection-Inference (PoSI) techniques. The predictors can also be ranked according to their influence on the response based on the order in which they enter the LASSO path.

### Random Forest

A decision tree is a nonparametric ML technique (as opposed to LASSO which is a parametric ML method) that involves stratifying or segmenting the predictor space into a number of simple regions using a splitting rule which can be summarized into a tree format. A sample decision tree is illustrated in Figure A.3 in the Appendix. However, a single decision tree often tends to suffer from high variance (a phenomenon known as overfitting in ML) by conforming too much to the noise rather than the information in the data.

Bagging (bootstrapped aggregating) is used to overcome overfitting where multiple decision trees are fit on bootstrapped resamples of the training data. The prediction for each data point is then obtained by averaging (or taking a majority vote) over all the fitted decision trees. However, using the same set of predictors repeatedly can cause the trees to be highly correlated with each other especially if there are one or two very strong predictors. Random Forest[13] overcomes this issue by bootstrapping the predictors as well each time a split is considered. Thus, in a Random Forest a number of decision trees are fit on bootstrapped resamples of the training data and by using a bootstrapped resample of the predictors at each split.

[12] LASSO was implemented using the ‘glmnet’ package in R Studio v3.4.5.

[13] Random Forest was implemented using the ‘randomforest’ package in R Studio v3.4.5.

The interpretation from Random Forest is derived from the measure of "purity of the predictor space" resulting from using that predictor to perform the split known as Gini Index (see James et al. (2000) for details). Each time any predictor is used to perform the split, the difference in the Gini index in the resulting child nodes and the mother node is calculated. This difference is then summed for each variable and normalized to get the mean decrease in the Gini index. Put simply, it measures how much each predictor contributes to the purity in the leaves in growing the random forest. Predictors that result in nodes with higher purity have a higher mean decrease in Gini and thus a higher association with the response (James et al. 2000).

**Boosting**

Boosting is a general iterative machine-learning approach that can be applied to many estimation algorithms. It is a process by which an algorithm is applied sequentially to modified versions of the same data using information learnt from the previous iteration. The predictions from each of these iterations are combined using a weighted majority vote to produce a final prediction where higher weights are assigned to more accurate predictions. In the first iteration all the observations receive equal weight and the algorithm is applied to the original data. In each successive iteration the data are re-weighted using a different weight for those observations that were mislabeled in the previous iteration from those observations that were correctly labeled in the previous iteration. Thus, with every iteration, the algorithm is compelled to 'learn' from its mistakes by focusing more intensely on the observations it mislabeled previously. This study uses a non-parametric algorithm - boosted classification trees[14]. The influence of the predictor on the response is measured using "relative influence" (Friedman 2001). A higher number indicates a higher influence of a predictor in explaining the variation in the response and vice-versa.

**Resampling Techniques**

In modeling 'rare-events', Machine-learning algorithms are frequently combined with various re-sampling techniques to improve prediction accuracy further. Resampling techniques involve balancing the distribution of the two classes using observations in the training data[15] to enable the algorithm to *learn* more efficiently from the events. The commonly used resampling techniques involve over-sampling (where the 'events' are replicated in the training data), and under-sampling (where a fraction of the 'non-events' are removed from the data). However, simple oversampling can lead to overfitting and simple under-sampling can lead to information loss. In this study we consider two modified versions of resampling[16] combined with Boosted Classification Trees- SMOTEBoost and RUSBoost. SMOTEBoost introduced by Chawla et al. (2003) employs Synthetic Minority Oversampling Technique (SMOTE) (where additional synthetic observations on 'events' are generated using information on the events already present in the data) in each iteration of boosting to enable the algorithm to have more 'events' to learn from in addition to assigning them higher weights. SMOTE overcomes the issue of overfitting partially by generating synthetic observations that are not exactly equal to the original observations. RUSBoost introduced

[14] Boosted Classification Trees were implemented using the 'gbm' package and 'xgboost' package in RStudio v3.4.5

[15] Note the test data is not balanced so that it reflects the true state of nature when testing the prediction accuracy of the ML algorithms.

[16] SMOTEBoost and RUSBoost were implemented using the 'ebmc' package in RStudio v3.4.5

by Seiffert et al. (2010) employs Random Under Sampling (RUS) (where a fraction of randomly selected 'non-events' are removed) in each boosting iteration. By combining RUS with boosting its main drawback of information loss can be overcome. Any observation dropped in one iteration is likely to be included in subsequent iterations. In addition, RUSBoost is computationally less time-consuming. Initially all the predictors (discussed earlier in section 3.1) were including in building predictive models. However, the initial analysis of the census of the ML algorithms revealed that some of these predictors were more informative that others-a finding that leads us to identify a 'high-mortality risk' group (discussed later in sections 4 and 5). While employing SMOTEBoost and RUSBoost only the important predictors[17] identified by the consensus from ML techniques (listed in Table 4) were used to incorporate the learning from the previous analysis.

**Measuring Prediction Accuracy**

Prediction problems involving a binary categorial response variable (like mortality in our case) are called classification problems. An algorithm predicts the probability of the event for each observation which are used to get predicted classes at different probability cut-off values (for instance predicting all observations with predicted probability above 0.5 as events/deaths). For a typical two-class classification problem, the possible outcomes are illustrated as a matrix popularly called confusion matrix shown below in Table 1.

**Table 1: Confusion matrix for a two-class classification problem**

| | **Positive Prediction** | **Negative Prediction** |
|---|---|---|
| **Actual Positive Class** | True Positive (TP) | False Negative (FN) |
| **Actual Negative Class** | False Positive (FP) | True Negative (TN) |

The columns in Table 1 are the predicted classes at a given cut-off while the rows represent the actual observed classes of the data. Each cell in the confusion matrix contains the count of observations with the true class and the predicted class. For instance, the first element in Table 1 is a count of the number of observations that the algorithm predicts as positive[18] that are truly positive (true positive). For this study 'positives' / 'events' refer to neonatal and infant deaths. The aim of the algorithms is to maximize the count of elements along the diagonal of the confusion matrix (which represent accurate predictions) and minimize the count of the elements along the off-diagonal of the confusion matrix (which represent the misclassifications).

In studying mortality, the consequences of misclassifying the positive class (failing to predict death) is graver than that of misclassification of the negative class. To take this into account we measure the area under precision-recall curve (AUPRC) to compare the prediction performances of the machine learning algorithms to that of logistic regression (See (Davis and Goadrich 2006) for details). Precision (TP/(TP+FP)) measures the probability that a randomly selected positive prediction is correct. Recall (TP/(TP+FN)) measures the probability that a randomly selected

[17] Following economic intuition 'gender' was included along with the list of predictors from Table 4

[18] In keeping with the literature, 'positive' refers to 'events' and 'negative' refers to 'non-events'.

positive observation is predicted as positive by the algorithm. Precision-Recall curve captures the trade-off between successfully predicting the events and being selective in producing positive predictions and enables us to focus on the events.

### Inference for Mortality

In keeping with the standard applied economics literature, we also employ conventional logistic regression and further perform PoSI (as a comparison) on predictors selected through LASSO in order to check if the regression coefficients of the selected predictors on mortality are significant. This exercise substantiates our findings from prediction exercises in identifying high-mortality risk group ---which will be illustrated later in section 5.

## 4.2. MALNUTRITION

We now focus on only the infants that remain alive at the time of the survey to investigate malnutrition. For this, the dataset is divided into five subgroups (national rural sample, national rural below poverty line sample, overall EAG states, rural EAG states sample, rural below poverty line EAG states sample), where the incidence of malnutrition is expected to be more severe. The number of predictors in some of these sub-groups comes close to the sample size, approaching near high-dimensional setting and necessitating the need for predictor selection. Variable selection is performed using LASSO, and inference and hypothesis testing is then performed on the selected predictors using PoSI. We test the statistical significance of the regression coefficients of the *selected* predictors. In other words, we test the null hypothesis that the regression coefficient of the selected predictors is zero in the model selected by LASSO. The PoSI technique described in Taylor and Tibshirani (2018) and in Lee, Sun, Sun and Taylor (2016) provide asymptotically valid p-values for the test. [19]

# 5. RESULTS

## 5.1 RESULTS FOR NEONATAL AND INFANT MORTALITY

### Prediction:

The prediction performance of the machine-learning algorithms (using AUPRC and the percentage increase in AUPRC) compared to that of ordinary logistic baseline regression are presented in Table 2 below. AUPRC for ordinary logistic regression is 0.42-0.43 for both neonatal mortality and infant mortality. All the machine-learning techniques have a higher predictive power than ordinary logistic regression demonstrated by a higher AUPRC. We also note that nonparametric ML methods like Random Forest and Boosted Trees perform better than standard logistic regression or logistic LASSO. Combining resampling techniques with Boosted Trees further improves prediction accuracy and RUSBoost demonstrates a higher prediction accuracy compared to SMOTEBoost. RUSBoost has the highest prediction accuracy with AUPRC values of 0.60-0.62 which is a 42%-44% increase in AUPRC over that of logistic regression.

[19] PoSI was performed using the ‘selectiveInference’ package in R studio v3.4.5

**Table 2: Prediction Performance of the Machine Learning Algorithms**

| **Machine Learning Algorithm** | **Neonatal Mortality** | | **Infant Mortality** | |
|---|---|---|---|---|
| | **AUPRC** | **Percentage increase in AUPRC** | **AUPRC** | **Percentage increase in AUPRC** |
| Standard Logistic Regression | 0.43 | ---- | 0.42 | -------- |
| Logistic LASSO | 0.45 | 5.2% | 0.47 | 10.6% |
| Classification Random Forest | 0.48 | 12.2% | 0.50 | 16.6% |
| Boosted Classification Trees | 0.54 | 25.5% | 0.55 | 27.1% |
| SMOTEBoost Classification Trees | 0.56 | 30.2% | 0.58 | 38.1% |
| RUSBoost Classification Trees | 0.62 | 44.2% | 0.60 | 42.8% |

Next, we present the ranking of predictors from LASSO along with their signs in Table 3 below. Across both neonatal and infant mortality 'First-Born', 'Prior-Death' and 'Total Births' enter the LASSO path early indicating strong association with the response variable. In case of infant mortality, the vaccines 'BCG', 'DPT', 'Measles' and 'Polio' enter the LASSO path early. In addition, JSY cash received for delivery ('JSY delivery money') the incidence of complications during delivery ('Delivery complications'), prematurity ('low birth weight'), early medical intervention received by the infant ('New born care') enter moderately early in the LASSO path for both neonatal and infant mortality.

Note that the dependent variable here is binary (mortality). Thus, a positive sign on the predictor implies higher mortality and vice-versa. Across both neonatal and infant mortality, first-borns have a higher incidence of mortality evidenced by the selection of 'First-Born' with a positive sign and 'Total Births' with a negative sign. The incidence of 'Prior Deaths' and being born premature ('Low Birth Weight') and intra-partum complications (delivery complications) is associated with higher mortality. The JSY cash received for delivery ('JSY delivery money') is associated with a lower incidence of mortality. In case of infant mortality, all the vaccines are associated with lower mortality as evidenced by their selection with a negative sign.

**Table 3: Ranking of Predictors from LASSO based on their order of entering the LASSO path**

| NEONATAL | INFANT |
|---|---|
| Prior deaths (+) | Polio (-) |
| Total births (-) | Prior deaths (+) |
| First born (+) | DPT (-) |
| Delivery complication (+) | BCG (-) |
| JSY Delivery money (-) | Measles (-) |
| New born care (-) | Total births (-) |
| Mother education (-) | Low birth weight (+) |
| Mother employed (-) | First born (+) |
| Low birth weight (+) | Folic acid (+) |
| Tetanus Toxoid (-) | New born care (-) |
| Pregnancy complication (+) | JSY Delivery money (-) |
| Age at birth (+) | Lab test (+) |
| Lab test (-) | Delivery complication (+) |
| Urban (-) | Mother's employment (+) |
| Income (-) | Sonogram (+) |
| Forward group (+) | Income (-) |
| Above poverty line (+) | Forward group (+) |
| Folic acid (+) | Mother's education (+) |
| Sonogram (+) | Gender (+) |
| Gender (+) | Tetanus toxoid (-) |
| Post-delivery complication (+) | Above poverty line (+) |
|  | Urban (+) |
|  | Post-delivery complication (+) |
|  | Pregnancy complication (+) |
|  | Age at birth (+) |

The ranking of the full set of predictors based on the measures of predictor influence from Random[20] Forest and Boosting is illustrated using bar-charts in Figures A.4 through A.7 in the Appendix. For both these algorithms, whether the mother has had a child die before (prior deaths), the total births a mother has had (total births), the mother's age at the time of the baby's birth (age at birth), whether the birth is the mother's first-born, and the household income as well as the cash received for delivery (JSY delivery money) are ranked higher that other predictors for neonatal as well as infant mortality. In addition, the immunizations (BCG, DPT, measles, and polio) are influential predictors in reducing infant mortality and the incidence of delivery complications is an important predictor in case of neonatal mortality. These predictors have also been selected in the model *selected* by LASSO and most of them are statistically significant after PoSI (discussed below). A summary of the most important predictors based on the consensus from the Machine Learning algorithms is listed in Table 4 below.

[20] The reader is advised to interpret the ranking of predictors from Random Forest with caution since how well a predictor predicts the response in one node depends on how well another predictor predicted the response at a node higher up in the tree.

**Table 4: Summary of the most important predictors based on the consensus from the Machine Learning Algorithms**

| NEONATAL MORTALITY | INFANT MORTALITY |
|---|---|
| Prior deaths | Prior deaths |
| Total births | Total births |
| First-born | First-born |
| JSY Delivery money | Low Birth Weight |
| Delivery Complication | JSY Delivery money |
| New Born Care | New Born Care |
| Mother Education | Polio |
| Mother Employed | BCG |
| Low Birth Weight | DPT |
| Income | Measles |
| | Income |
| | Folic Acid |

**Inference:**

After building predictive models and identifying strong predictors for mortality, we now examine the same through the lens of inference and investigate whether the predictors selected by ML algorithms have significant effects. Results from standard logistic regression are reported in Table A.6. Note that for the analysis on mortality the dimensionality does not necessarily call for sparsity analysis.[21] Nevertheless, we apply PoSI after logistic LASSO and compare results with baseline logistic regression. See, Table 5 below. We find that the strong predictors identified by the Machine Learning algorithms- ‘prior deaths’, ‘first-born’ ‘delivery complication’ as well as the ‘immunizations’ also produce statistically significant regression coefficient, substantiating our identification of the ‘high risk’ group. Thus, from both prediction as well as inference perspectives we can suggest that first-born children with prior sibling death(s), mother’s delivery complications and families which act careless about immunizations need to be under stronger surveillance than others. In addition, we find evidence in favor of effectiveness of a key policy variable, “*JSY delivery money”* based on logistic regression.

[21] The number of covariates is around 20 to 25, and therefore standard logistic regressions can work well given the large number of observations. This is also corroborated by high AUROC (around 96%) obtained from logistic regressions.

**Table 5: Statistical Significance of Predictors from Ordinary Logistic Regression and PoSI on Neonatal and Infant Mortality**

| NEONATAL MORTALITY | | INFANT MORTALITY | |
|---|---|---|---|
| Ordinary Logistic Regression | PoSI | Ordinary Logistic Regression | PoSI |
| First-born (+) *** | First-born (+) *** | First-born (+) *** | First-born (+) *** |
| Prior Deaths (+) | Prior Deaths (+) *** | Prior Deaths (+) | Prior Deaths (+) *** |
| Total Births (-) *** | Total Births (-) *** | Total Births (-) ** | Total Births (-) ** |
| Low birth weight (+) | Low birth weight (+) | Low birth weight (+) ** | Low birth weight (+) ** |
| New born care (-) | New born care (-) | New born care (-) | New born care (-) |
| JSY Delivery Money (-) * | JSY Delivery Money (-) | JSY Delivery Money (-) ** | JSY Delivery Money (-) |
| Mother Employed (-) | Mother Employed (-) | Mother Employed (+) | Mother Employed (+) |
| Delivery Complication (+) ** | Delivery Complication (+) ** | Delivery Complication (+) | Delivery Complication (+) |
| Mother Education (-) | Mother Education (-) | Mother Education (-) | Mother Education (-) |
| Gender (+) | ---- | Gender (+) | Gender (+) |
| Folic Acid (+) | --- | Folic Acid (+) | Folic Acid (+) |
| Sonogram (+) | --- | Sonogram (+) | Sonogram (+) |
| Income (-) | --- | Income (-) | Income (-) |
| Forward group (+) | --- | Forward group (+) | Forward group (+) |
| Above Poverty Line (+) | --- | Above Poverty Line (+) | Above Poverty Line (+) |
| Lab test (+) | --- | Lab test (+) | Lab test (+) |
| Pregnancy Complication (+) | --- | Pregnancy Complication (+) | --- |
| Post Complication (+) | --- | Post Complication (+) | --- |
| Mother’s age at birth (+) | --- | Mother’s age at birth (+) | --- |
| Urban (-) | --- | Urban (-) | --- |
| Tetanus Toxoid (-) | --- | Tetanus Toxoid (-) | Tetanus Toxoid (-) |
| | | BCG (-) *** | BCG*** |
| | | DPT (-) | DPT (-) |
| | | Polio (-) *** | Polio (-) *** |
| | | Measles (-) *** | Measles (-) *** |

(*** 1%, **5%)

## 5.2. RESULTS FOR MALNUTRITION

Our next focus is to investigate infant malnutrition using weight-for-length Z score of infants. Note that here we use a continuous response variable and thus a positive sign on the covariate implies higher weight-for height. For this part we consider a large comprehensive set of covariates that are available following the framework proposed by Mosley and Chen (1984) (see, Tables A.5 and A.6) and use LASSO to perform model selection. As stated earlier, we investigate malnutrition separately in the subgroups of rural infants, infants in rural and below poverty line households, infants in EAG states, rural infants in EAG states and infants in rural, below poverty line households in EAG states. Selection of such sub-groups allows us to focus on populations where the malnutrition is especially stark. Recall that we are now dealing with a large set of (possibly correlated) covariates and in some of these subgroups, the number of covariates come close to sample size, thus necessitating sparsity analysis.

Figures A.8 to A.12 in the Appendix plot the 95% confidence intervals from PoSI and the corresponding naive confidence intervals from performing OLS for the selected predictors. The naive confidence intervals are shown in orange and the PoSI confidence intervals are shown in green. Table A.8 in the Appendix reports the point estimates and the corresponding p-values (significance test) from PoSI on the sub-groups considered. Table 6 below summarizes only the statistically significant (at 95%) covariates obtained from the analysis from each sub-group using PoSI.

**TABLE 6: Statistically Significant Covariates from PoSI on Malnutrition**

| **National Rural Sample** | **National Rural and Below Poverty Line** | **EAG states** | **EAG states Rural** | **EAG States Rural and Below Poverty Line** |
|---|---|---|---|---|
| Dependency ratio (-) | Water (+) | Water (+) | Dependency ratio (-) | Water (+) |
| AWD Vaccine (-) | Income (-) | AWD vaccine (-) | AWD vaccine (-) | Income (-) |
| Below Poverty Line ration card (+) | | | | |

Access to safe drinking water is selected as a predictor in most samples and is statistically significant in the EAG states and in the rural below poverty line samples in raising Z-scores. This result corroborates findings from (Jalan and Ravallion 2003). Safe drinking water improves Z-scores by protecting them from various water borne diseases (like diarrhea) which causes rapid substantial weight loss. A higher dependency ratio in the rural samples are associated with lower weight-for height Z scores in infants. This is likely due to a crowding out effect of the family's attention from the infants.

Infants who received vaccinations at the Anganwadi centers have Z scores that are lower than infants who didn't receive their vaccinations at Anganwadi centers. This apparently puzzling result is presumably due to inadequacy in data collection. The survey asked the respondents whether the infant received the vaccine at the Anganwadi center, but not specifically *where* the vaccines were received. Thus, the infants who didn't receive their vaccines at the Anganwadi centers includes both infants in poor households who didn't receive vaccines at all as well as infants in better-off households who received their vaccines in other private medical clinics that are equipped with *better preservation facilities*. This is corroborated by the fact that we don't find this perverse result in the sample below the poverty line. However, the preservation of vaccines, power outages as well as corruption, leakage and lack of monitoring of resources prevalent in Anganwadi centers are well-recognized matters of concerns for policy makers.[22]

Interestingly, in the "rural below poverty line" groups (both for national level and EAG states), weight-for-length Z scores decreases with increase in annual household income. This is not true

[22] Only for the national sample we find that infants whose mothers received nutrition supplement at the Anganwadi centers have a lower weight-for-height Z score. Mothers who didn't receive supplementary nutrition at the Anganwadi centers included mothers who received supplementary nutrition from other sources (including the ones from wealthier families).

for other samples. This can be well explained by the fact that in these households most women are employed and as a result the infants are likely to be left in the care of older children who are not vigilant about proper infant feeding practices.[23]

## 6. CONCLUDING DISCUSSION WITH POTENTIAL FUTURE EXTENSIONS

Tackling India's high rates of neonatal and infant mortality as well as malnutrition are matters of utmost importance. In 2015, more than 600,000 deaths occurred during the neonatal period and more than 900,000 deaths occurred during infancy[24]. "*India Newborn Action Plan*" aims at bringing mortality to single digit per thousand by 2030. This study uses several parametric and non-parametric machine-learning tools to build predictive models for the incidence of neonatal and infant mortality which are 'rare-events' in survey data. Of all the techniques applied, RUSBoost with boosted classification trees stood out as the best performing Learning technique for prediction. Also a baseline logistic regression and PoSI (Taylor and Tibshirani 2018) were employed. Combining the results, we identify a "high mortality risk" group (first-borns or prior sibling deaths, delivery complications, not being vaccinated) for implementing a more aggressive and targeted policy. We also find support in favor of JSY delivery money in bringing down mortality. Despite the existence of the Universal Immunization Program in India (under which vaccinations are provided free of cost) for over 35 years, India is yet to achieve full coverage of immunization across the country. Achieving nation-wide full coverage of these essential vaccines can prevent the deaths from vaccine-preventable diseases. Families that are averse to vaccinating their children are also likely to make other mistakes in health care for children, raising the probability of infant death. Our study strongly indicates the importance of these vaccines and tracking down children that are not immunized in reducing infant mortality.[25]

This study also uses a comprehensive set of covariates and PoSI approach (a separate Learning method) to investigate the multi-faceted problems of *infant malnutrition* in India. Our results strongly suggest a pressing need to provide safe drinking water to all households. As a related issue, some states in India are increasingly suffering from drought during the summer months. While collecting data, it is important to collect a comprehensive measure (quality, quantity and consistency) of access to safe drinking water in households. The various health beliefs possessed by the mother as well as mother's exposure to mass media (women media) are sporadically selected as predictors across sub-samples. More aggressive public health campaigns disseminated

[23] Many women from such households work for long hours with low pay and several of them are single mothers as well.

[24] Author's back of the envelop calculations based on World Bank's estimates of IMR and NMR.

[25] Predictive modeling for mortality involves tacking data that are imbalanced. Providing good causal inference in case of imbalanced data, especially when sample size is small is a challenging task (see, (King and Zeng 2001; Firth 1993) for details). Machine Learning techniques are generally known for producing better prediction of events. However. emerging research in Machine Learning, for example by Fithian and Hastie (2014) tailored towards the issue of data imbalance can potentially improve the performances of models even further. Such techniques can be considered in the future to achieve even higher prediction accuracy in predicting mortality, especially very rare mortality events such as maternal mortality or still birth.

through mass media could have a wider reach and increase the effectiveness of these beliefs. The multitude of services provided by Anganwadi centers (especially the provision of supplemental feeding to pregnant and lactating mothers as well as vaccinating the pregnant mothers) do not seem to have expected outcomes. Problem of corruption and leakage of resources is rampant across many of India’s public health policies especially the Public Distribution Scheme. Similar problems have been observed for India’s school lunch program (Mid-Day Meal Scheme). In addition, air pollution which leads to respiratory diseases in children is an emerging problem in urban India and needs to be measured while collecting data. Finally, there is pressing need to collect more detailed data on neonatal and infant mortality especially *cause-specific* data at the micro-level. This will allow researchers to prescribe even more *targeted* health policies than the existing ones.

## REFERENCES

Arokiasamy, Perianayagam, and Abhishek Gautam. 2008. “Neonatal Mortality in the Empowered Action Group States of India: Trends and Determinants.” *Journal of Biosocial Science* 40 (2): 183–201. https://doi.org/10.1017/S0021932007002623.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2013. “Inference on Treatment Effects after Selection among High-Dimensional Controls.” *Review of Economic Studies* 81 (2): 608–50. https://doi.org/10.1093/restud/rdt044.

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. 2013. “Valid Post-Selection Inference.” *Annals of Statistics* 41 (2): 802–37. https://doi.org/10.1214/12-AOS1077.

Bhalotra, Sonia. 2007. “SPENDING TO SAVE? STATE HEALTH EXPENDITURE AND INFANT MORTALITY IN INDIA.” *Health Economics* 16: 911–28. https://doi.org/10.1002/hec.

Bhalotra, Sonia, and Arthur van Soest. 2008. “Birth-Spacing, Fertility and Neonatal Mortality in India: Dynamics, Frailty, and Fecundity.” *Journal of Econometrics* 143 (2): 274–90. https://doi.org/10.1016/j.jeconom.2007.10.005.

Bhattacharyya, Siddhartha, Sanjeev Jha, Kurian Tharakunnel, and J. Christopher Westland. 2011. “Data Mining for Credit Card Fraud: A Comparative Study.” *Decision Support Systems* 50 (3): 602–13. https://doi.org/10.1016/j.dss.2010.08.008.

Bühlmann, Peter. 2013. “Statistical Significance in High-Dimensional Linear Models.” *Bernoulli* 19 (4): 1212–42. https://doi.org/10.3150/12-BEJSP11.

Chawla, Nitesh V., Aleksandar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. 2003. “SMOTEBoost: Improving Prediction of the Minority Class in Boosting.” In *European Conference on Principles of Data Mining and Knowledge Discovery PKDD 2003*, 107–19. https://doi.org/10.1007/978-3-540-39804-2_12.

Davis, Jesse, and Mark Goadrich. 2006. “The Relationship between Precision-Recall and ROC Curves.” In *Proceedings of the 23rd International Conference on Machine Learning - ICML ’06*, 233–40. https://doi.org/10.1145/1143844.1143874.

Desai, Sonalde, Reeve Vanneman and National Council of Applied Economic Research, New Delhi. India Human Development Survey-II (IHDS-II), 2011-12. ICPSR36151-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-07-31. http://doi.org/10.3886/ICPSR36151.v2

Duflo, Esther. 2003. “Grandmothers and Granddaughters.” *World Bank Economic Review* 17 (1): 1–25. https://doi.org/10.1093/wber/lhg013.

Engebretsen, Ingunn Marie Stadskleiv, Thorkild Tylleskär, Henry Wamani, Charles Karamagi, and James K. Tumwine. 2008. “Determinants of Infant Growth in Eastern Uganda: A Community-Based Cross-Sectional Study.” *BMC Public Health* 8: 1–12. https://doi.org/10.1186/1471-2458-8-418.

Firth, David. 1993. “Bias Reduction of Maximum Likelihood Estimates Author ( s ): Published by : Biometrika Trust Stable URL.” *Biometrika* 80 (1): 27–38.

Fithian, William, and Trevor Hastie. 2014. “Local Case-Control Sampling: Efficient Subsampling in Imbalanced Data Sets.” *Annals of Statistics* 42 (5): 1693–1724. https://doi.org/10.1214/14-AOS1220.

Friedman, Jerome H. 2001. "Greedy Function Approximation : A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–1232. https://www.jstor.org/stable/2699986.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. *Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-84858-7.

Jain, Monica. 2015. "India's Struggle against Malnutrition-Is the ICDS Program the Answer?" *World Development* 67: 72–89. https://doi.org/10.1016/j.worlddev.2014.10.006.

Jalan, J., and M. Ravallion. 2003. "Does Piped Water Reduce Diarrhoea for Children in Rural India?" *Journal of Econometrics* 112 (August): 153–73. https://doi.org/10.1016/S0304-4076(02)00158-6.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2000. *An Introduction to Statistical Learning*. *Springer*. Vol. 7. https://doi.org/10.1007/978-1-4614-7138-7.

Khalilia, Mohammed, Sounak Chakraborty, and Mihail Popescu. 2011. "Predicting Disease Risks from Highly Imbalanced Data Using Random Forest." *BMC Medical Informatics and Decision Making* 11 (1): 51. https://doi.org/10.1186/1472-6947-11-51.

King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9 (02): 137–63. https://doi.org/10.1093/oxfordjournals.pan.a004868.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling [Hardcover]*. https://doi.org/10.1007/978-1-4614-6849-3.

Lee, Jason D., Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. 2016. "Exact Post-Selection Inference, with Application to the Lasso." *Annals of Statistics* 44 (3): 907–27. https://doi.org/10.1214/15-AOS1371.

Mercedes, De Onis. 2008. "The New WHO Child Growth Standards." *Paediatria Croatica Supplement* 52 (SUPP.1): 13–17. https://doi.org/10.4067/S0370-41062009000400012.

Mosley, W. Henry, and Lincoln C Chen. 1984. "An Analytical Framework for the Study of Child Survival in Developing Countries." *Population and Development Review* 10: 25–45.

Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher. 2015. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24 (1): 1–17. https://doi.org/10.1093/pan/mpv024.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. http://www.aeaweb.org/articles?id=10.1257/jep.31.2.87.

Radhakrishna, R., and C. Ravi. 2004. "Malnutrition in India: Trends and Determinants." *Economic and Political Weekly* 39 (7): 671–76. https://doi.org/10.2307/4414642.

Rammohan, Anu, Kazi Iqbal, and Niyi Awofeso. 2013. "Reducing Neonatal Mortality in India: Critical Role of Access to Emergency Obstetric Care." *PLoS ONE* 8 (3). https://doi.org/10.1371/journal.pone.0057244.

Seiffert, Chris, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2010. "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance." *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans* 40 (1): 185–97. https://doi.org/10.1109/TSMCA.2009.2029559.

Singh, Aditya, Abhishek Kumar, and Amit Kumar. 2013. "Determinants of Neonatal Mortality in Rural India, 2007–2008." *PeerJ*, May. https://doi.org/10.7717/peerj.75.

Taylor, Jonathan, and Robert Tibshirani. 2018. “Post-Selection Inference for L1-Penalized Likelihood Models.” *Canadian Journal of Statistics* 46 (1): 41–61. https://doi.org/10.1002/cjs.11228.

The Million Death Study Collaborators. 2010. “Causes of Neonatal and Child Mortality in India: A Nationally Representative Mortality Survey.” *The Lancet* 376 (9755): 1853–60. https://doi.org/10.1016/S0140-6736(10)61461-4.

Varian, Hal R. 2014. “Big Data: New Tricks for Econometrics.” *Journal of Economic Perspectives* 28 (2): 3–28. https://doi.org/10.1257/jep.28.2.3.

Victora, C. G., M. de Onis, P. C. Hallal, M. Blossner, and R. Shrimpton. 2010. “Worldwide Timing of Growth Faltering: Revisiting Implications for Interventions.” *Pediatrics* 125 (3): e473–80. https://doi.org/10.1542/peds.2009-1519.

## APPENDIX

### FIGURE A.1: FLOWCHART ILLUSTRATING THE ANALYSIS IN THIS PAPER

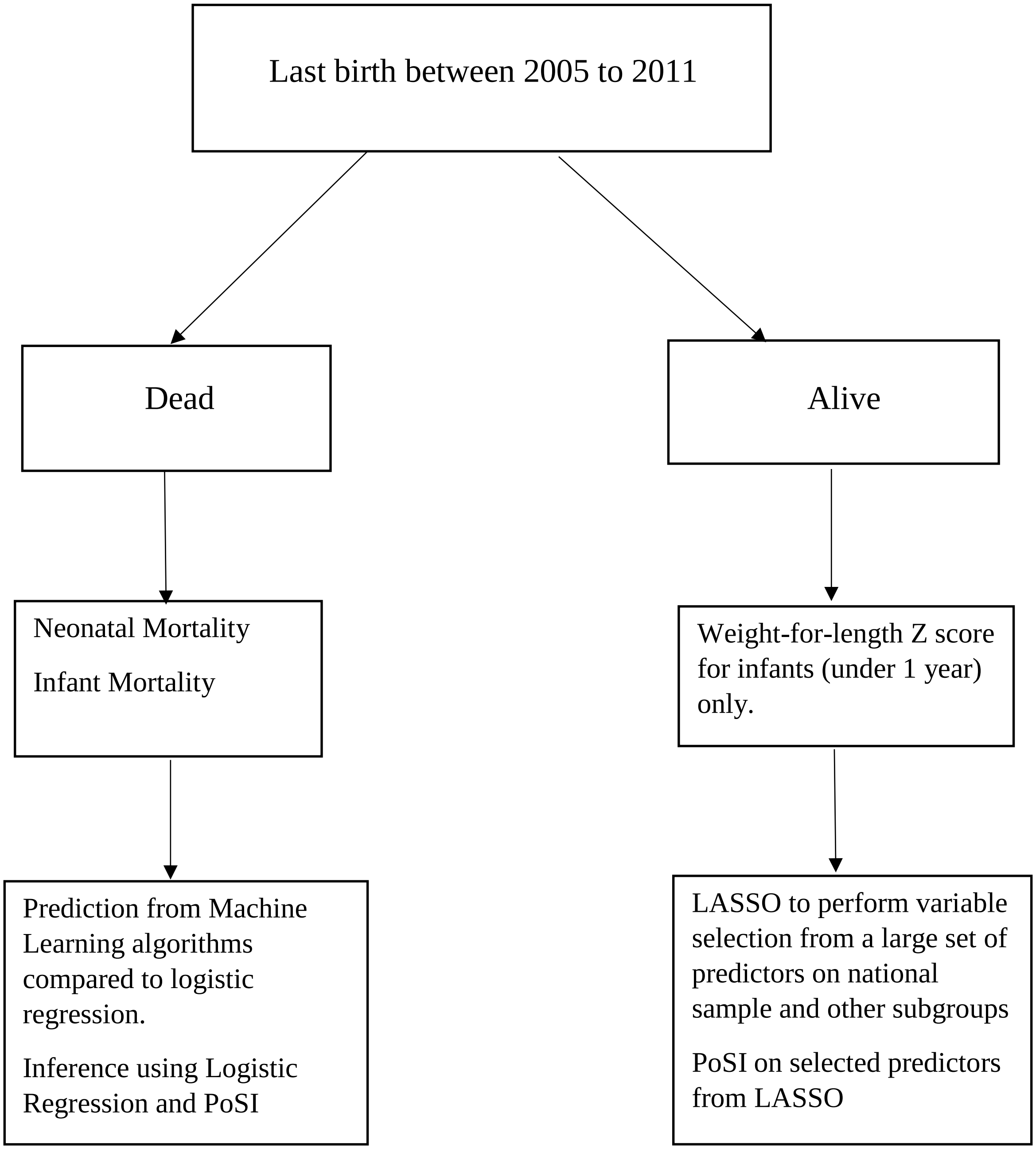
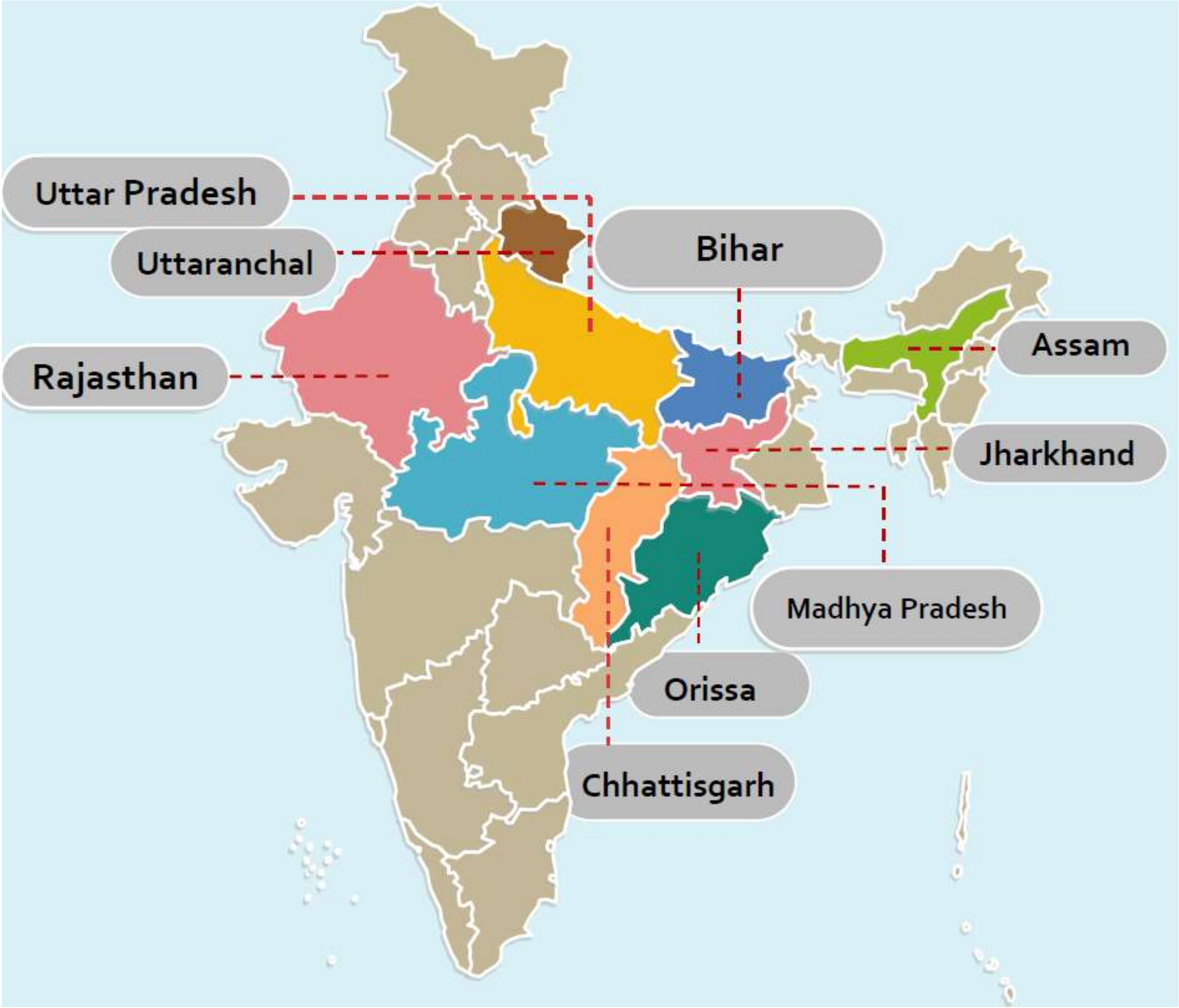

Last birth between 2005 to 2011
Dead
Alive
Neonatal Mortality
Infant Mortality
Weight-for-length Z score for infants (under 1 year) only.
Prediction from Machine Learning algorithms compared to logistic regression.
Inference using Logistic Regression and PoSI
LASSO to perform variable selection from a large set of predictors on national sample and other subgroups
PoSI on selected predictors from LASSO

**FIGURE A.2: MAP OF INDIA SHOWING THE EMPOWERED ACTION GROUP STATES**



## TABLE A.1: DESCRIPTION OF PREDICTORS FOR NEONATAL AND INFANT MORTALITY

| Variable | Type | Description |
|---|---|---|
| Sonogram | Binary | Whether the mother received a sonogram during pregnancy. |
| New-born care | Binary | Whether the baby received the first medical checkup during the first week of birth (neonatal) or first month (infant) |
| Lab test | Binary | Whether the mother received an intensive medical checkup including a blood test and urine test which requires access to a medical laboratory |
| Folic acid | Binary | Whether the mother received folic acid tablets during pregnancy. |
| TT | Binary | Whether the mother received TT injections during pregnancy. |
| Low birth weight | Binary | Whether the newborn had low weight at birth |
| Pre-complication | Binary | Whether the mother experienced any complication during pregnancy |
| Delivery complication | Binary | Whether the mother experienced any complication during delivery |
| Post complication | Binary | Whether the mother experienced complication in the two months after delivery. |
| JSY Delivery money | Continuous | The amount of money received for delivery under the Janani Suraksha Yojana scheme. |
| Gender | Binary | Gender of the baby. =0 if male; =1 if female |
| Total births | Discrete | Total number of births of the mother |
| First-born | Binary | Whether this live birth was the mother’s first-born |
| Prior deaths | Binary | Whether the mother has had any children die before this live birth. |
| Age at birth | Binary | Mother’s age at the time of baby’s birth |
| Mother education | Binary | Whether the mother has studied until primary school |
| Employed | Binary | Whether the mother is employed. |
| Income | Continuous | Real annual household income |
| Above poverty line (APL) | Binary | Whether the household is above the poverty line. The poverty line was constructed using the World Bank’s estimate of $1.90 for the 2011 PPP conversion factor. |
| Urban | Binary | Whether the household is in the urban area. |
| Forward group | Binary | Whether the household belongs to the group with higher social status.<br>=0 for OBC/Dalit/Adivasi/Muslim;<br>=1 for Brahmin/Forward Caste/Christian/Jain/Sikh |
| BCG | Binary | Whether the baby received the BCG vaccination |
| DPT | Binary | Whether the baby received at least one dose of DPT vaccination |
| Measles | Binary | Whether the baby received at least one dose of Measles vaccination |
| Polio | Binary | Whether the baby received at least one dose of the Oral Polio Vaccination |

## TABLE A.2: DESCRIPTION OF PREDICTORS FOR MALNUTRITION

| VARIABLE | TYPE | DESCRIPTION |
|---|---|---|
| Relative carb expenditure | continuous | Ratio of monthly household expenditure on rice, lentils and vegetables to total expenditure on food. |
| Relative meat expenditure | continuous | Ratio of monthly household expenditure on meat and eggs to total expenditure on food. |
| Relative milk expenditure | continuous | Ratio of monthly household expenditure on milk to total expenditure on food. |
| Three square meals | binary | Whether the household typically consumes three or more meals each day |
| Gender | binary | Gender of the baby; Male=0, Female=1 |
| Water | binary | Whether the household has access to safe drinking water |
| Purify water | binary | Whether the household purifies drinking either by boiling of filtering |
| Toilet | binary | Whether the household has a flush or semi-flush toilet |
| Household shock | binary | Whether the household experienced a shock recently due to illness/accident/ drought/flood/fire/loss of job/ death/crop failure |
| Household agriculture | binary | Whether the primary source of income for the household is from agriculture |
| Household art | binary | Whether the primary source of income for the household is from artisanal enterprise or petty business |
| Household pension | binary | Whether the primary source of income for the household is from rent/pension. |
| Brahmin | binary | Whether the household belongs to the Brahmin caste |
| OBC | binary | Whether the household belongs to the Other Backward caste group |
| Dalit | binary | Whether the household belongs to the 'Dalit' caste |
| Adivasi | binary | Whether the household belongs to the 'Adivasi' caste |
| Muslim | binary | Whether the household members are Muslims |
| BPL ration card | binary | Whether the household has a 'Below Poverty line' ration card |
| APL ration card | binary | Whether the household has a 'Above Poverty line' ration card |
| Antyodaya ration card | binary | Whether the household has a 'Antyodaya' ration card |
| Ill last 30 days | continuous | Number of days the infant was ill in the last 30 days |
| Fever | binary | Whether the infant had a fever in the last 30 days |
| Cough | binary | Whether the infant had a cough in the last 30 days |
| Diarrhea | binary | Whether the infant had diarrhea in the last 30 days |
| ORS | binary | Whether the infant was administered Oral Rehydration Salts solution |
| Treatment | binary | Whether treatment was sought for illness in the last 30 days |
| Medical Expenditure | continuous | Expenditure on medical services (including medicines) in the last 30 days |

| | | |
|---|---|---|
| Dependency ratio | continuous | The ratio of the number of children below 14 and adults above 60 to the number of adults between 15 to 60 years of age in the household |
| No. of household members | continuous | The total number of household members |
| No. of children | continuous | The total number of children of the mother who are alive |
| Income | continuous | Real annual household income |
| Household consumption | continuous | Per capita household monthly consumption |
| Below Poverty Line | binary | Whether the household is below the World Bank's poverty line estimate of $1.90 per person per day |
| Male head age | continuous | Age of the male head of the household |
| Female head age | continuous | Age of the female head of the household |
| Male head education | binary | Whether the male head of the household finished primary school |
| Female head education | binary | Whether the female head of the household finished primary school |
| Old pension | binary | Whether there are any elderly pension recipients in the household |
| Women Media | binary | Whether the women regularly watch TV, listen to the radio, read the newspaper. |
| Health Belief first milk | binary | Whether the mother believes that the first milk is good for the baby |
| Health Belief diarrhea | binary | Whether the mother believes that during diarrhea the baby should be given more to drink |
| Health Belief malaria | binary | Whether the mother believes that mosquitos cause malaria |
| Mom Autonomy | binary | Whether the mother makes decisions about the infant's health |
| Mom employed | binary | Whether the mother is employed |
| Mom age | continuous | Mother's age |
| Mom education | binary | Whether the mother finished primary school |
| Mom BMI | continuous | Mother's Body Mass Index |
| BCG | binary | Whether the infant received the BCG vaccination |
| DPT | binary | Whether the infant received the DPT vaccination |
| Polio | binary | Whether the infant received the Oral Polio vaccination |
| Measles | binary | Whether the infant received the measles vaccination |
| Vitamin A | binary | Whether the infant received vitamin A supplement |
| Low birth weight | binary | Whether the baby was born very small or smaller than average in size |
| AWD food | binary | Whether the infant received supplementary food from Anganwadi centers |
| AWD growth | binary | Whether the infant received growth monitoring services from Anganwadi centers |

| | | |
|---|---|---|
| AWD health | binary | Whether the infant received health check-up services from Anganwadi centers |
| AWD mother | binary | Whether the mother received supplemental food or immunization while pregnant or lactating |
| Breast fed 1 hour | binary | Whether the infant was breast fed within 1 hour of birth |
| Breast fed 6 months | binary | Whether the infant was breast fed exclusively for six months |
| Developed village | binary | Whether the household lives in a relatively developed village |
| Non-metro urban | binary | Whether the household lives in an urban area that is not one of the six major metro areas |
| Metro urban | binary | Whether the household lives in one of the six major urban metropolitan areas |

### TABLE A.3 DESCRIPTIVE STATISTICS OF DISCRETE PREDICTORS FOR NEONATAL AND INFANT MORTALITY

| **Predictors** | **Neonatal Mortality** | **Infant Mortality** |
|---|---|---|
| | **Percentages** | **Percentages** |
| Gender (Boy) | 55.42 | 55.57 |
| First-born | 32.33 | 32.73 |
| Folic acid | 85.98 | 86.0 |
| Newborn care | 52.34 | 66.65 |
| Sonogram | 62.08 | 61.97 |
| Tetanus Toxoid | 90.9 | 90.96 |
| Prior deaths | 10.08 | 9.54 |
| Lab test | 83.65 | 83.63 |
| Pregnancy Complication | 64.6 | 64.7 |
| Delivery complication | 35.66 | 35.6 |
| Post complication | 31.64 | 31.56 |
| Low birth weight | 15.3 | 15.3 |
| Mother finished primary school | 64.54 | 64.32 |
| Above Poverty Line | 67.82 | 67.74 |
| Urban | 36.83 | 37.04 |
| Forward group | 23.51 | 23.49 |
| BCG | -- | 79.63 |
| DPT | -- | 90.9 |
| Measles | -- | 73.53 |
| Polio | -- | 95.1 |

## TABLE A.4. DESCRIPTIVE STATISTICS OF CONTINUOUS PREDICTORS FOR NEONATAL AND INFANT MORTALITY

| Predictors | | Neonatal | Infant |
|---|---|---|---|
| Income | Min | -192600 | -192600 |
| | Mean | 121651.2 | 122011.5 |
| | Standard Deviation | 111330.4 | 112019 |
| | Max | 600000 | 600000 |
| JSY Delivery Money | Min | 0 | 0 |
| | Mean | 555.93 | 556.74 |
| | Standard Deviation | 980.97 | 976.35 |
| | Max | 8000 | 8000 |
| Total births | Min | 1 | 1 |
| | Mean | 2.23 | 2.24 |
| | Standard Deviation | 1.39 | 1.39 |
| | Max | 14 | 14 |
| Mother’s age at birth | Min | 14 | 14 |
| | Mean | 25 | 25 |
| | Standard Deviation | 4.72 | 4.72 |
| | Max | 49 | 49 |

## TABLE A.5: DESCRIPTIVE STATISTICS OF DISCRETE PREDICTORS FOR MALNUTRITION

| Predictors | | Rural and Below Poverty Line Sample Percentages |
|---|---|---|
| Gender (Boy) | | 47.37 |
| Women Media | | 42.6 |
| Three square meals | | 61.1 |
| Water | | 84.7 |
| Purify water | | 9.5 |
| Toilet | | 13.7 |
| Household shock | | 58.4 |
| Household main source of income | Agriculture | 85.3 |
| | Art / business | 10.0 |
| | Pension /rent | 1.0 |
| | Salaried (reference) | 3.7 |
| Caste and Religion | Brahmin | 2.1 |
| | OBC | 38.9 |
| | Dalit | 21.6 |
| | Adivasi | 16.3 |
| | Muslim | 12.6 |
| | Non-Brahmin forward caste (reference) | 8.4 |
| Ration card | None (reference) | 22.1 |
| | BPL ration card | 33.7 |
| | APL ration card | 33.7 |
| | Antyodaya ration card | 10.5 |
| Illness | Fever | 44.2 |
| | Cough | 37.3 |
| | Diarrhea | 20.5 |
| | None | --- |
| Received ORS | | 9.5 |
| Received treatment | | 47.9 |
| Old pension | | 9.5 |
| Bargaining power | | 14.8 |
| Breast fed within 1 hour of birth | | 52.1 |
| Breast fed for six months | | 78.5 |
| Health belief about first milk | | 84.2 |
| Health belief about diarrhea | | 63.2 |
| Health belief about malaria | | 87.9 |
| Mother employed | | 10.0 |
| Mother finished primary school | | 52.2 |
| Sector | Less developed village | 71.5 |
| | More developed village | 28.5 |
| | Non-metro urban | NA |
| | Metro urban | NA |
| BCG | | 86.4 |
| DPT | | 75.8 |
| Measles | | 16.8 |
| Polio | | 86.8 |
| Vitamin A supplement | | 12.1 |

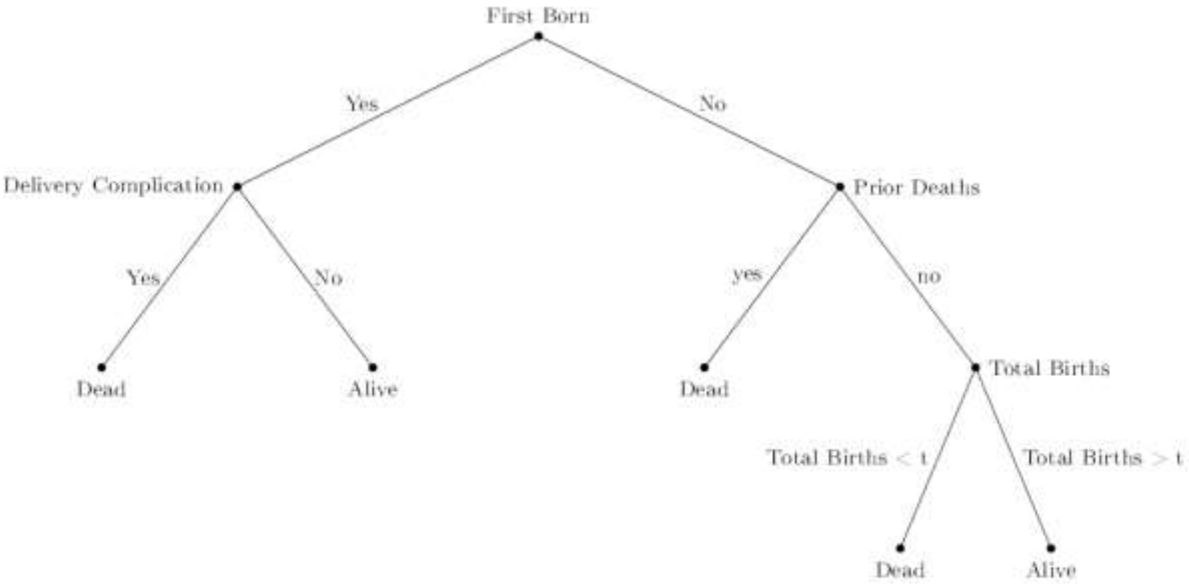
| | |
|---|---|
| Low birth weight | 16.4 |
| AWD health checkup | 32.1 |
| AWD food supplement | 37.9 |
| AWD growth monitoring | 42.6 |
| AWD vaccine | 59.5 |
| AWD mother received supplement | 45. 3 |
| Male household head finished primary school | 72.1 |
| Female household finished primary school | 52.1 |

**TABLE A.6: DESCRIPTIVE STATISTICS OF THE CONTINUOUS PREDICTORS FOR MALNUTRITION**

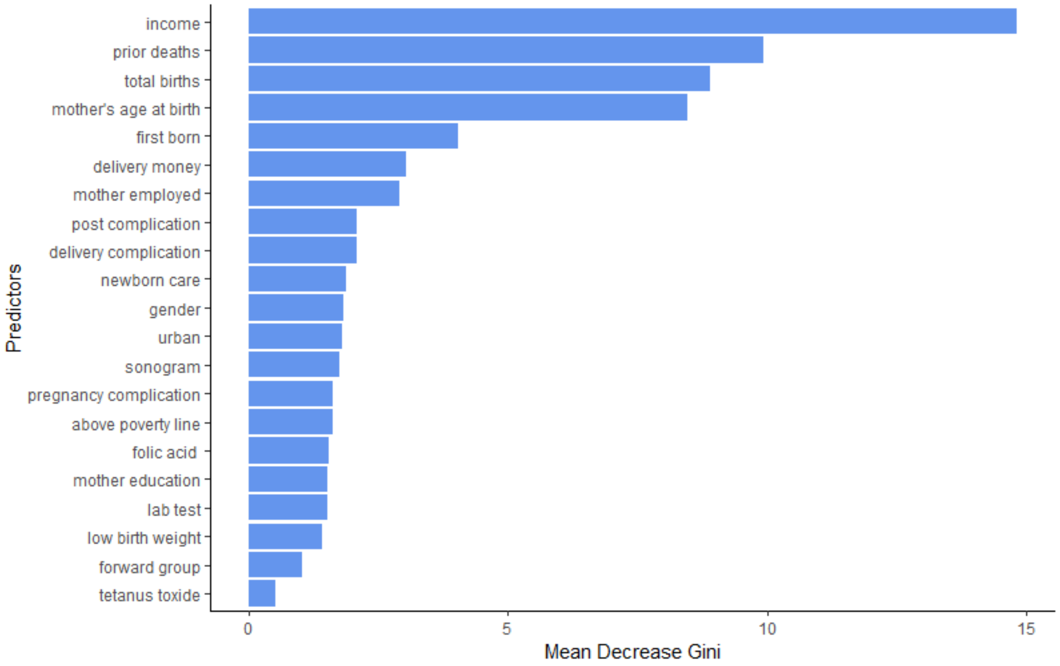
| | | **Rural and Below Poverty Line Sample** |
|---|---|---|
| Relative carb expenditure | Min | 0.23 |
| | Mean | 0.70 |
| | Standard Deviation | 0.16 |
| | Max | 1.00 |
| Relative meat expenditure | Min | 0.00 |
| | Mean | 0.11 |
| | Standard Deviation | 0.09 |
| | Max | 0.46 |
| Relative milk expenditure | Min | 0.00 |
| | Mean | 0.18 |
| | Standard Deviation | 0.17 |
| | Max | 0.71 |
| Number of days ill in the last months | Min | 0 |
| | Mean | 3.0 |
| | Standard Deviation | 4.6 |
| | Max | 30 |
| Medical Expenditure (in INR) | Min | 0 |
| | Mean | 209.96 |
| | Standard Deviation | 588.73 |
| | Max | 5000 |
| Dependency ratio | Min | 0.14 |
| | Mean | 1.00 |
| | Standard Deviation | 0.59 |
| | Max | 3 |
| Total People | Min | 3 |
| | Mean | 6.93 |
| | Standard Deviation | 2.98 |
| | Max | 21 |
| Income | Min | 1875 |
| | Mean | 41955.51 |
| | Standard Deviation | 29009.11 |
| | Max | 185715 |

| | | |
|---|---|---|
| Household consumption | Min | 2614 |
| | Mean | 12529.43 |
| | Standard Deviation | 5748.54 |
| | Max | 32377 |
| Male household head age | Min | 21 |
| | Mean | 43.19 |
| | Standard Deviation | 14.84 |
| | Max | 78 |
| Female Household head age | Min | 19 |
| | Mean | 39.26 |
| | Standard Deviation | 14.22 |
| | Max | 70 |
| Number of children alive | Min | 1 |
| | Mean | 2.4 |
| | Standard Deviation | 1.4 |
| | Max | 8 |
| Mother BMI | Min | 12.80 |
| | Mean | 19.55 |
| | Standard Deviation | 2.99 |
| | Max | 30.37 |

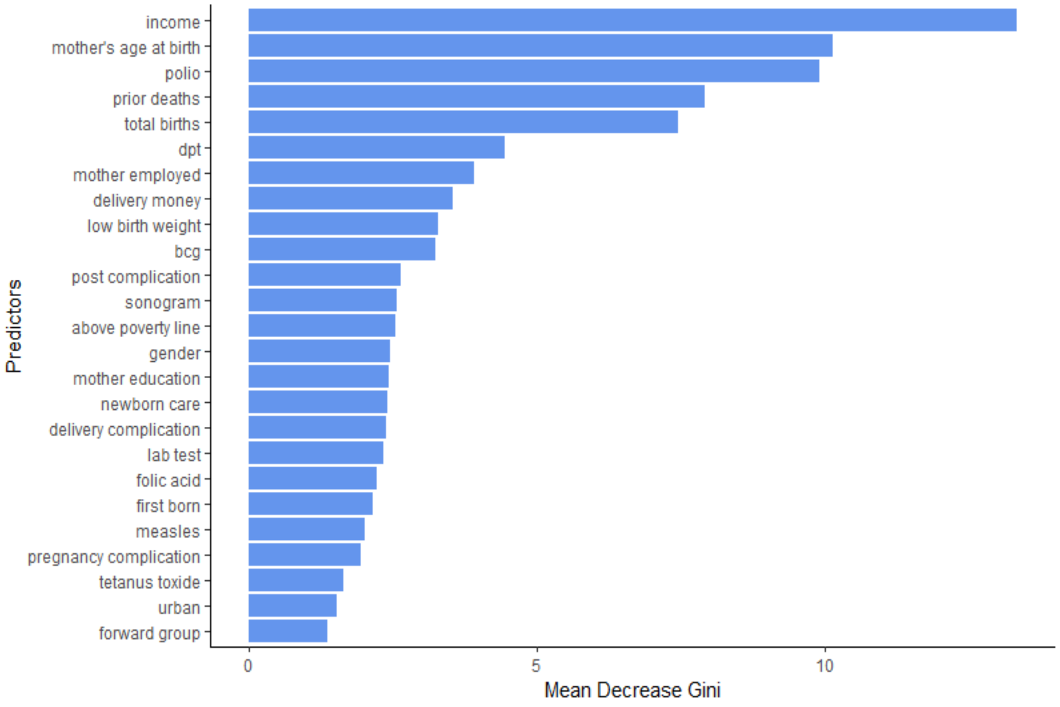
## FIGURE A.3: A SAMPLE DECISION TREE



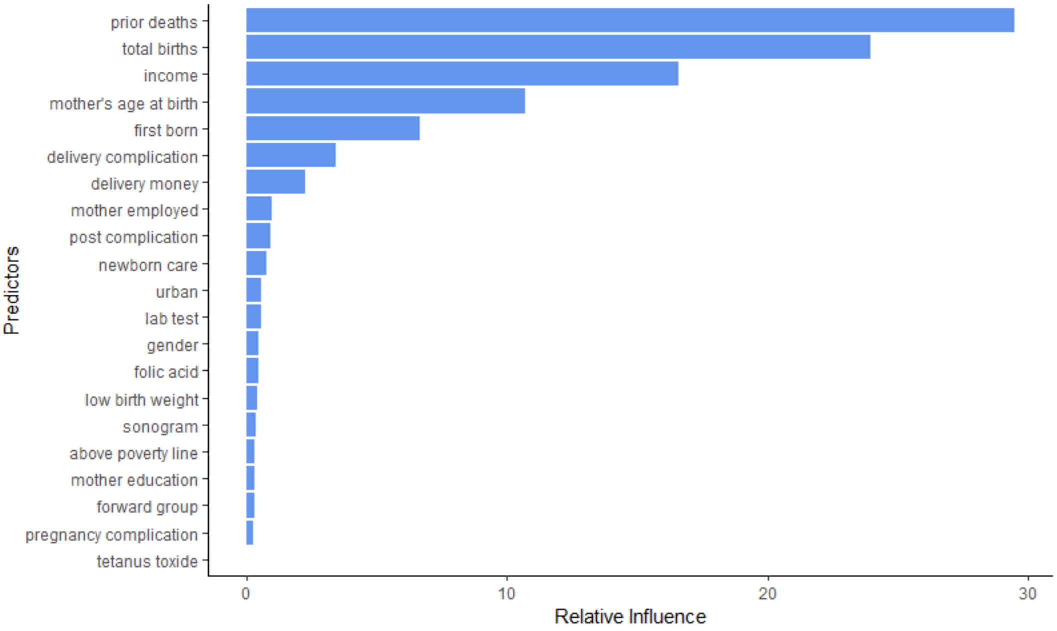
## FIGURE A.4: RANKING OF PREDICTORS FROM RANDOM FOREST FOR NEONATAL MORTALITY



## FIGURE A.5: RANKING OF PREDICTORS FROM RANDOMFOREST FOR INFANT MORTALITY



## FIGURE A.6: RANKING OF PREDICTORS FROM BOOSTING FOR NEONATAL MORTALITY



## FIGURE A.7: RANKING OF PREDICTORS FROM BOOSTING FOR INFANT MORTALITY

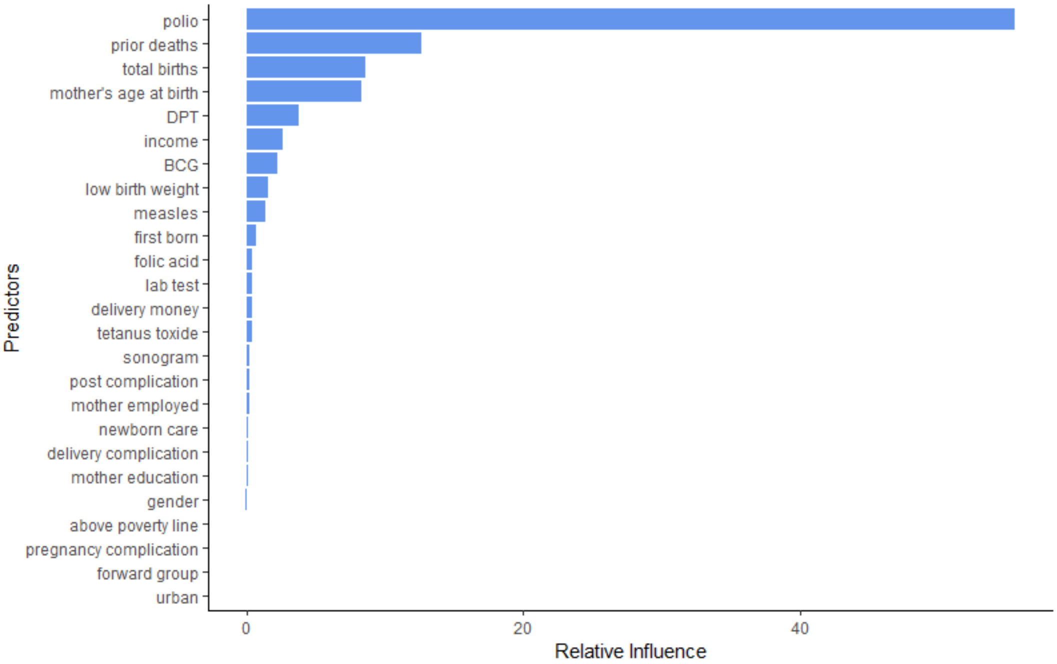


**Table A.7: Results from Logistic Regression for Neonatal and Infant Mortality**

| | NEONATAL MORTALITY | INFANT MORTALITY |
|---|---|---|
| Gender | 0.01 | 0.01 |
| | (0.97) | (0.52) |
| First born | 2.08*** | 17.47 |
| | (0.000) | (0.97) |
| Folic Acid | 0.07 | 0.49 |
| | (0.84) | (0.17) |
| Sonogram | 0.04 | 0.19 |
| | (0.90) | (0.59) |
| Tetanus Toxoid | -0.70 | -0.03 |
| | (0.38) | (0.36) |
| Prior deaths | 2.29 | 1.98 |
| | (0.97) | (0.97) |
| JSY Delivery money | -0.0004* | -0.0005** |
| | (0.08) | (0.03) |
| Post complication | -0.01 | 0.02 |
| | (0.97) | (0.94) |
| Forward group | 0.08 | 0.36 |
| | (0.83) | (0.35) |
| Above Poverty Line | 0.09 | 0.20 |
| | (0.78) | (0.54) |
| Mother Employed | -0.09 | -0.15 |
| | (0.34) | (0.14) |
| Mom’s Age at birth | 0.04 | 0.004 |
| | (0.25) | (0.90) |
| Low birth weight | 0.20 | 0.77 ** |
| | (0.55) | (0.01) |
| Delivery complication | 0.70 ** | 0.15 |
| | (0.02) | (0.60) |
| Pregnancy complication | 0.29 | 0.008 |
| | (0.42) | (0.98) |
| New-born care | -0.24 | -0.36 |
| | (0.40) | (0.20) |
| Lab test | -0.08 | 0.15 |
| | (0.81) | (0.68) |
| Mother education | -0.46 | 0.24 |
| | (0.16) | (0.47) |
| Total number of births | -0.50 *** | -0.34 *** |
| | (0.000) | (0.008) |
| Urban | -0.16 | 0.04 |
| | (0.64) | (0.88) |
| Income | -8.6e-07 | - 3.04e-06 |
| | (0.66) | (0.12) |
| BCG | | -1.43 *** |
| | | (0.000) |
| DPT | | -0.69 |
| | | (0.13) |
| Measles | | -2.42*** |
| | | (0.000) |
| Polio | | -1.82*** |
| | | (0.000) |
| n | 7,656 | 7,257 |

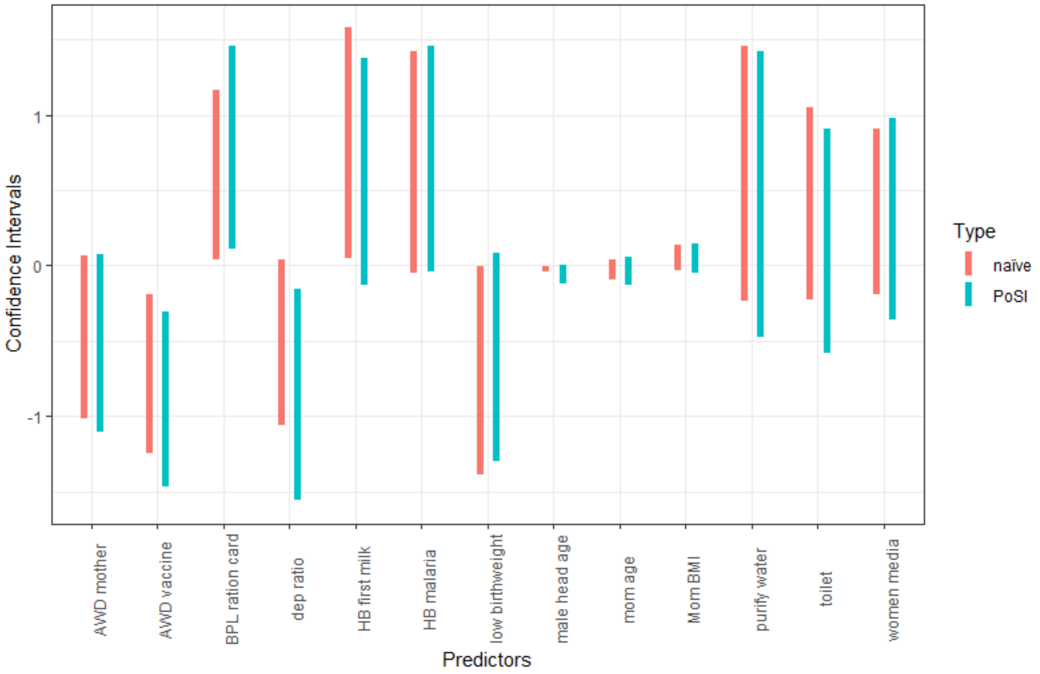
(p-values are reported in parenthesis)
(*** 1%, **5)

**Table A.8: Post-Selection Inference on Weight-for-Length Z Scores in the population subgroups (selected covariates only)**

| COVARIATES | RURAL SAMPLE | RURAL BELOW POVERTY LINE SAMPLE | EAG STATES | EAG RURAL SAMPLE | EAG RURAL BELOW POVERTY LINE |
|---|---|---|---|---|---|
| Water | ____ | 1.574**<br>(0.014) | 1.002**<br>(0.033) | 0.739<br>(0.128) | 1.922**<br>(0.014) |
| Income | _____ | -0.00002**<br>(0.032) | ____ | ____ | -0.00003**<br>(0.030) |
| AWD vaccine | -0.894**<br>(0.007) | _____ | -1.088**<br>(0.006) | -1.191**<br>(0.010) | -0.977<br>(0.108) |
| AWD mother | -0.560<br>(0.071) | _____ | ____ | ____ | ____ |
| Mom age | -0.061<br>(0.193) | -0.119<br>(0.071) | _____ | _____ | ______ |
| Health belief about malaria | 0.777<br>(0.057) | 0.765<br>(0.256) | _____ | _____ | 0.095<br>(0.569) |
| Women media | 0.420<br>(0.177) | _____ | ____ | ____ | _____ |
| Purify water | 0.627<br>(0.164) | ______ | _____ | _____ | ______ |
| Toilet | 0.326<br>(0.273) | ______ | ______ | _____ | _____ |
| Dependency ratio | -0.861**<br>(0.023) | ______ | _____ | -1.039**<br>(0.018) | _____ |
| Male head age | -0.059<br>(0.069) | _____ | _____ | ____ | ____ |
| Health belief about first milk | 0.667<br>(0.080) | _____ | _____ | 0.550<br>(0.226) | 1.064<br>(0.124) |
| Low birth weight | -0.659<br>(0.068) | _____ | _____ | ___ | ____ |
| Mom BMI | 0.065<br>(0.175) | _____ | ____ | ____ | ____ |
| Below Poverty Line ration card | 0.794**<br>(0.029) | 0.601<br>(0.177) | ____ | _____ | 0.910<br>(0.139) |
| Relative Carb Expenditure | ____ | 0.608<br>(0.304) | ____ | _____ | 1.173<br>(0.131) |
| Mother Employed | ____ | ____ | _____ | -1.111<br>(0.078) | ____ |

(p-values are reported in parenthesis)
(*** 1%, **5%)

## Figure A.8: Naïve and PoSI Confidence Intervals in the National Rural Sample of Infants



## Figure A.9: Naïve and PoSI Confidence Intervals in the National Rural Below Poverty Line Sample of Infants

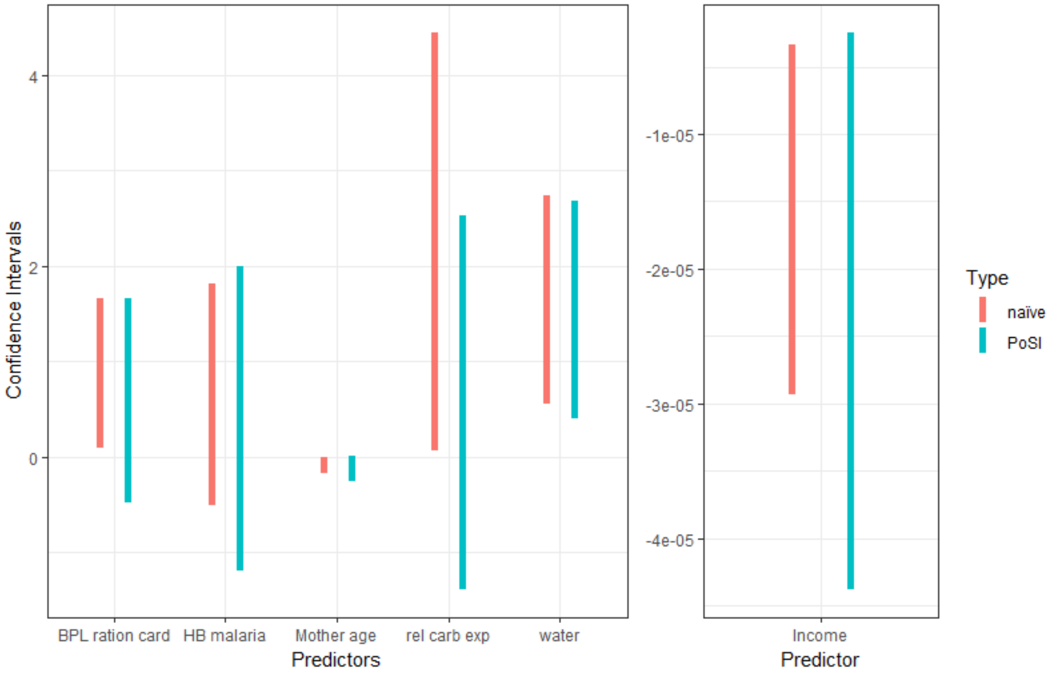


**Figure A.10: Naïve and PoSI Confidence Intervals in the Sample of Infants in EAG States**

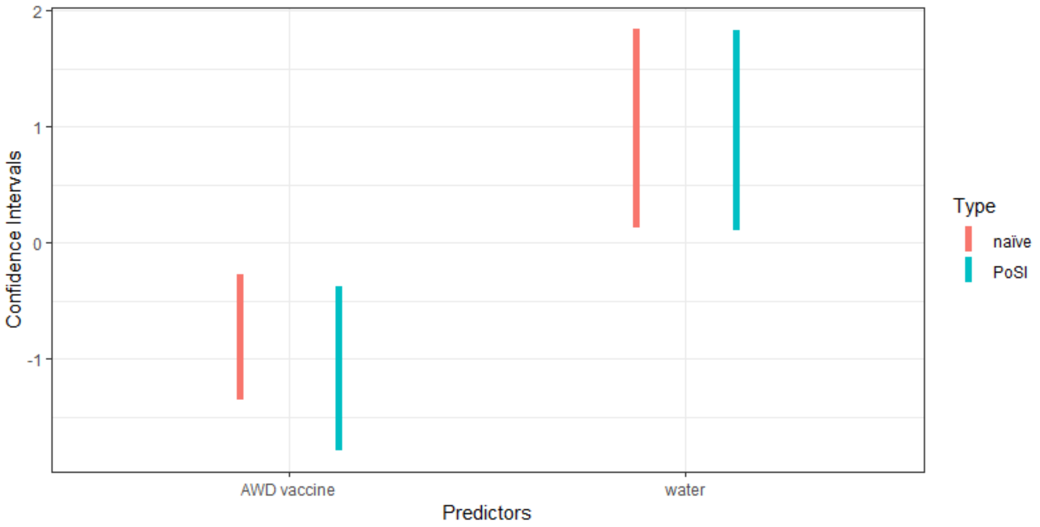


**Figure A.11: Naïve and PoSI Confidence Intervals in the Rural Sample of Infants in EAG States**

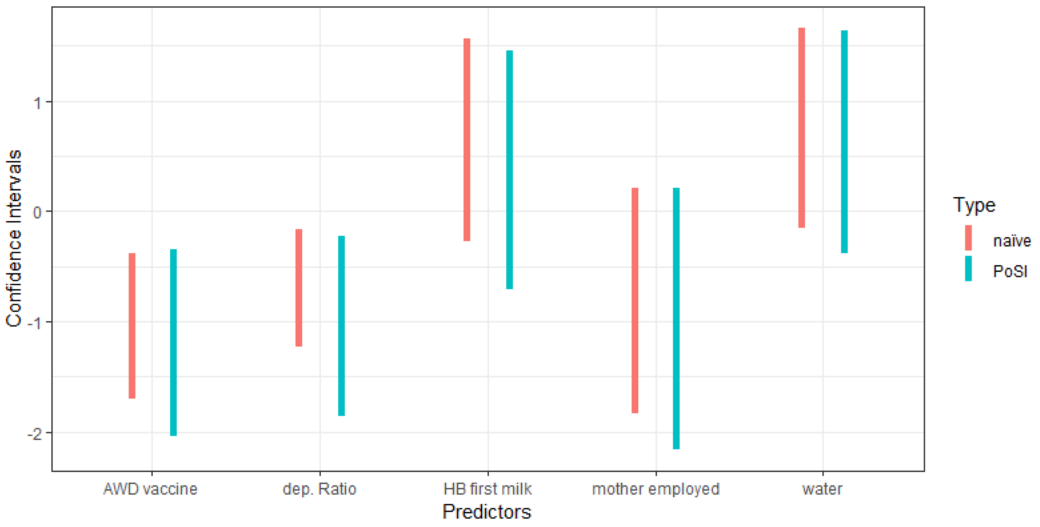
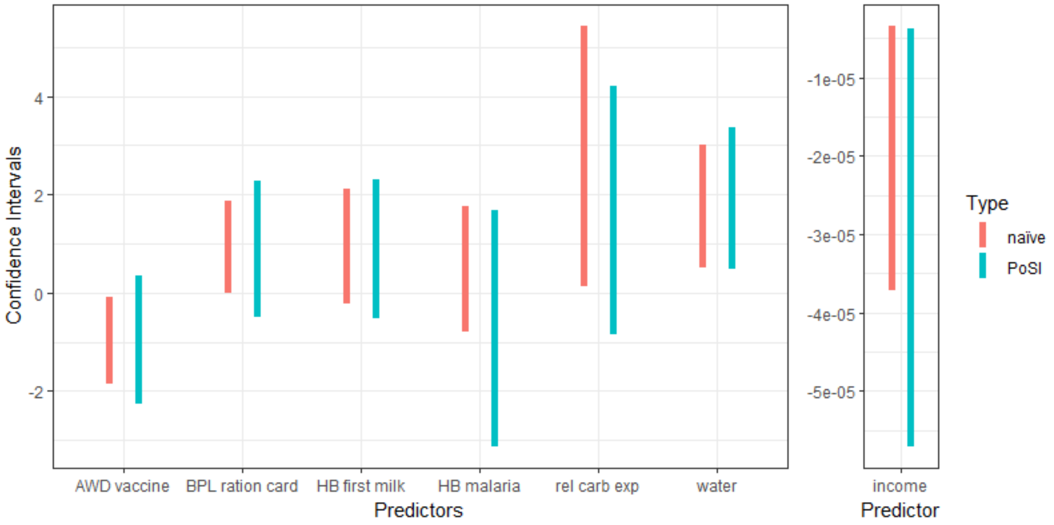


**Figure A.12: Naïve and PoSI Confidence Intervals in the Rural Below Poverty Line Sample of Infants in EAG States**


Confidence Intervals
AWD vaccine
BPL ration card
HB first milk
HB malaria
rel carb exp
water
Predictors
-1e-05
-2e-05
-3e-05
-4e-05
-5e-05
income
Predictor
Type
naïve
PoSI