2016

Association between Size at Birth and Maternal Factors in Rural Bangladesh: A Multivariate Approach

Kabir, A.Y.M. Alamgir

University of Rajshahi

Association between Size at Birth and Maternal Factors in Rural Bangladesh: A Multivariate Approach

This thesis submitted in fulfillment of the requirements for the Degree of Master of Philosophy in Statistics

By

A.Y.M. Alamgir Kabir

Roll no. 09328

Session: 2009-2010

Department of Statistics

University of Rajshahi

June, 2016

Abstract

In this research we explored the potential of the multivariate methods, CCA and PLS regression in studying the relationship between two sets of variables in public health research. This research will help the public health researchers in choosing the appropriate statistical methods if they want to study multiple outcomes and multiple exposures simultaneously. Additionally, we intended to identify a surrogative measure of low birth weight (LBW) using ROC curve and 4 machine learning algorithms, decision tree, random forest, support vector machine and neural network. Both the canonical correlation analysis and the PLS regression analysis has several advantages over the univariate methods. However, CCA is just an exploratory method very similar to Pearson's correlation. Although one variable set is often considered as predictor and the other as criterion, it does not imply causal relationship between the set of exposures and the set of outcomes. On the other hand, PLS regression can help in establishing causal relationship between exposures and the outcomes. So, the choice of CCA or PLS regression depends on the objective of the study. In addition, this study will provide a realistic predictive model of LBW for Rural Bangladesh. The LBW can be predicted with 4 other simpler-to-measure anthropometries, length and head, chest and arm circumferences without measuring their weights, at a greater accuracy with the advent of the information technology.

Acknowledgement

First of all I would like to express my sincere gratitude to my deceased advisor Professor Mohammed Nasser for the continuous support until his death. He was the greatest inspiration for me to be a good researcher and to be a good human being as well. I can't think of what I am now without him. I am also grateful to my current supervisor Dr. Jahanur Rahman who provided me an enormous support after Professor Nasser's death to accomplish the thesis. Furthermore, I am grateful to my professional mentors from Johns Hopkins University, Prof. Keith P West, Prof. Parul Christian, Dr. Alain Labrique and Dr. Rolf Klemm and Abu Ahmed Shamim from JiVitA who also inspired me a lot to accomplish the research by providing their valuable comments, suggestions and edits. I am also thankful to JiVitA and my co-workers at JiVitA from where I got a very big support of real life data and to the Department of Statistics, University of Rajshahi where I grew up as a researcher and all of my respected teachers who put their input in doing this research. Last but not least I would like to express my gratitude to my current professional supervisor, Dr. Firdausi Qadri, Senior Scientist, icddr,b for her inspiration to complete the thesis successfully.

Table of Contents

List of Tables

List of Figures

List of publications from this thesis

From peer reviewed journal

Alamgir Kabir, Rebecca D Merrill, Abu Ahmed Shamim, Rolf D W Clemm, Alain B Labrique, Parul Christian, Keith P. West Jr., **Mohammed Nasser**. Canonical Correlation Analysis of Infant's Size at Birth and Maternal Factors: A Study in Rural Northwest Bangladesh. *PLOS ONE*, 9(4): e94243.

Abstract

This analysis was conducted to explore the association between 5 birth size measurements (weight, length and head, chest and mid-upper arm [MUAC] circumferences) as dependent variables and 10 maternal factors as independent variables using canonical correlation analysis (CCA). CCA considers simultaneously sets of dependent and independent variables and, thus, generates a substantially reduced type 1 error. Data were from women delivering a singleton live birth (n = 14506) while participating in a double-masked, cluster-randomized, placebo-controlled maternal vitamin A or b-carotene supplementation trial in rural Bangladesh. The first canonical correlation was 0.42 (P,0.001), demonstrating a moderate positive correlation mainly between the 5 birth size measurements and 5 maternal factors (preterm delivery, early pregnancy MUAC, infant sex, age and parity). A significant interaction between infant sex and preterm delivery on birth size was also revealed from the score plot. Thirteen percent of birth size variability was explained by the composite score of the maternal factors (Redundancy, R Y/X = 0.131). Given an ability to accommodate numerous relationships and reduce complexities of multiple comparisons, CCA identified the 5 maternal variables able to predict birth size in this rural Bangladesh setting. CCA may offer an efficient, practical and inclusive approach to assessing the association between two sets of variables, addressing the innate complexity of interactions.

From conference proceedings

Alamgir Kabir, A A Shamimm, R Klemm, A B Labrique, P Christian, Keith P West Jr., **Mohammed Nasser**. Classification of low birth weight and very low birth weight in Rural Bangladesh. *International Conference on Applied Statistics Bangladesh.* December 27-29, 2014. Nabab Nawab Ali Chowdhury Senate Bhaban, University of Dhaka, Bangladesh.

Abstract

The consequences of low birth weight, <2.5kg (LBW) or very low birth weight,<1.5kg (VLBW) are universally recognized. The early detection of LBW or VLBW neonate and immediate direct interventions can help to catching up with the normal babies. In resource poor settings like Bangladesh most of the deliveries take place at home and the birth weight is not often measured due the paucity of measurement scale. We aim to propose surrogative measure of LBW and VLBW for rural Bangladesh using univariate and multivariate classification methods and find the best classifier. We studied n=15683 newborn singleton live born neonates whose anthropometric measurements were made at home within 72 hrs of birth. This was a cluster-randomized newborn vitamin A supplementation trial in 19 unions of rural Bangladesh. The anthropometry (weight, length and head, chest and arm circumference) was measured by the trained anthropometrist. Receiver Operating Characteristic (ROC) curve, decision tree (DT), support vector machine (SVM), random forest (RF) and neural network (NNW) were used for classification. Their performance was compared by the misclassification rate, sensitivity and specificity. Chest circumference (CC) had the highest correlation with birth weight. The ROC curve suggested that CC is the best predictor of LBW and VLBW. The thresholds of CC to predict LBW and VLBW are 30.5 cm and 27.2 cm with sensitivity 83% and 82% and specificity 84% and 95% respectively. DT classified LBW and VLBW with 13% and 1% misclassification rate respectively. DT model used only CC and length. Both RF, SVM and NNW classified LBW and VLBW with 12% and 1% misclassification rate respectively. To classify LBW DT gave the maximum sensitivity (89%) and NNW gave the maximum specificity (89%). However, to classify VLBW, NNW gave the maximum sensitivity (65%) and all DT, RF, SVM and NNW gave 100% specificity. In conclusion, we can say that CC is the best predictor of both LBW and VLBW. DT is the best classifier of LBW and ROC curve is the best classifier for VLBW.

Alamgir Kabir, Rebecca D Merrill, Abu Ahmed Shamim, Rolf D W Clemm, Alain B Labrique, Parul Christian, Keith P. West Jr., **Mohammed Nasser**. Partial Least Squares (PLS) Regression in Predicting Infant's Size at Birth in Rural North Western Bangladesh. *International Conference on Statistical Data Mining for Bioinformatics Health Agriculture and Environment.* December 21-24, 2012. Department of Statistics, University of Rajshahi, Bangladesh.

Abstract

Partial least square (PLS) regression is gaining popularity in different branches of research due to its flexibility in use. However, it is relatively new in public health research. It combines features from principal component analysis and multiple linear regression. It is able to solve the problems related to high collinearity among predictors. In this study we attempted to predict 5 birth size measurements from a set of 10 maternal factors using PLS regression and compared its performance with principal component (PC) regression. The data was taken from women with singleton live births (n=14506) participating in a large community-based, double-masked, cluster-randomized, placebo-controlled maternal vitamin A or β-carotene supplementation trial in rural Bangladesh. All the maternal factors except maternal vitamin A and β-carotene supplementation had a significant ($p<0.001$) effect in predicting infant size at birth. Among them, preterm delivery had the largest negative influence on infant's size ($\beta = -0.27$; $p<0.001$). PLS regression required only 2 components to predict infant's size where as PC regression required 5 components. In summary, PLS regression is a useful, efficient method to predict infant size at birth from maternal factors.

DEDICATED

To

Late Dr. Mohammed Nasser

My all time inspiration

Chapter 1: Introduction

Public health research

Public health is "the science and art of preventing disease, prolonging life and promoting health through the organized efforts and informed choices of society, organizations, public and private, communities and individuals" (Winslow & Charles-Edward 1920). Public health incorporates the interdisciplinary approaches of epidemiology, biostatistics and health services. Modern public health practice requires multidisciplinary teams of public health workers and professionals including physicians specializing in public health/community medicine/infectious disease, psychologists, epidemiologists, biostatisticians, medical assistants or Assistant Medical Officers, public health nurses, midwives, medical microbiologists, environmental health officers / public health inspectors, pharmacists, dental hygienists, dietitians and nutritionists, veterinarians, public health engineers, public health lawyers, sociologists, community development workers, communications experts, bioethicists, and others. Public health workers monitor the health of a community by collecting and analyzing data related to health. Statistics is vital part of public health's assessment function, used to identify special risk groups, detect new health threats, plan public health programs and evaluate their success, and prepare government budgets.

Maternal and child health research

Maternal and child health is one of the key areas of public health research specifically for the developing world. Because, maternal and child health is playing a vital role to have better economic growth of a developing country. Children who experience better physical health and fewer negative health shocks during their lifetimes reach a higher, more productive potential and effectively reap the benefits from investments in their health and education. Given that better educated children are expected to be more productive in the future, parents of healthier children are motivated to further invest in their child's schooling. Developments in maternal and child health also contribute to longer life expectancy, thereby creating a stronger rationale for women to invest in their children's education as well as their own. On the other hand children born to malnourished mothers or mothers who experienced a negative health shock (such as suffering from malnutrition or contracting an infectious disease) during pregnancy are more likely to have different health hazards. Children represent the future, and ensuring their healthy growth and development ought to be a prime concern of all societies. So, we should take care of our children from the very beginning of their life, starting from the gestational period.

Birth size and maternal factors

Newborns are particularly vulnerable and children are vulnerable to malnutrition and infectious diseases, many of which can be effectively prevented or treated. Children's healthy growth and development largely depend on the growth of the fetus during the entire gestational period. The effect of intrauterine growth retardation or premature delivery of a newborn will carry all through the life and which constitutes a higher economic burden to the family and eventually to the whole nation. They no longer only have increased perinatal mortality and morbidity, and an increased incidence of sudden infant death syndrome (Øyen et al. 1995). In childhood, they are hampered by deficits in cognitive and neurological development (Taylor & Howie 1989). In adulthood, they are at higher risk of high blood pressure (Williams et al. 1992), elevated cholesterol concentrations (Barker et al. 1993), cardiovascular disease (Osmond et al. 1993), obstructive lung disease (Barker et al. 1993), diabetes (Hales et al. 1991), and renal impairment (Hinchliffe et al. 1992). In contrast, the optimal growth of fetus can ensure the healthier life with a greater likelihood. Size of the infant at birth is the indicator of intrauterine growth or prematurity of the fetus. Fetal growth is largely, but not solely, determined by the availability of nutrients from the mother before and during gestation, as well as placental capacity to supply these nutrients in sufficient quantities to the fetus (Thame et al. 1997), and birth size can reflect the intrauterine environment. Maternal nutritional status largely depends on socio-economic factors. Women from a higher socioeconomic status have increased

access to and consumption of nutritious foods during or prior to gestation and more antenatal care (ANC) visits and nutrition supplementation during gestation. Small birth size is more common in resource poor settings or among more disadvantaged populations (Khatun & Rahman 2008; Leal et al. 2006; Silva et al. 2012). So, there is a growing interest among maternal and child health researchers in studying the relationship between birth size and maternal socio-demographic and health factors. In this study we have five anthropometric measurement of infant at birth to describe size at birth. Birth weight is often the exclusive birth size measure used to evaluate fetal growth. However, other measurements like length and head, chest, and arm circumferences may be important in predicting long-term health and development outcomes (Neggers et al. 1995).

Low birth weight

Many studies used only the birth weight as the birth size measurements and define low birth weight (LBW) (birth weight <2500gm) as the smaller size. They also reported that the early detection of LBW babies and immediate direct interventions can help to catching up with the normal babies. In resource poor settings like Bangladesh most of the deliveries take place at home which is attended by relatives or traditional birth attendants. Weighing scale to measure birth weight is not available at household level. Even if a birth takes place at health facility, babies are not

weighed routinely due to paucity of a suitable weighing scale at the center. Despite having a weighing scale, it is necessary to have regular calibration to avoid any systematic error during measurement. So, a surrogative measure of low birth weight is essential for the early detection of low birth weight in resource poor settings which can be measured without the weighing scale.

Statistical methods in public health research

Statistics is a methodology in science and industry as well as in medicine and in many other fields with broad areas of application. Any phenomenon in biology, psychology, or medicine is usually based on probabilistic model. Because, phenomena are influenced by many factors that in themselves are variable and by other factors that are unidentifiable. That is, various states of a phenomenon occur with certain probabilities. Therefore, the statistical techniques are needed to increase scientific knowledge. The presence of variation requires the use of statistical analysis. A major part of statistics involves the drawing of inferences from samples to a population in regard to some characteristic of interest. The reliability of such inferences or conclusions may be evaluated in terms of probability statements. There are many statistical methods are available in literature. But, some of them are frequently used in public health research which are easy to use and easy to interpret.

There are some other complex methods which are more appropriate to describe a phenomenon in public health research but rarely used.

In public health research most of the time researcher wants of establishes a relationship (causal) between exposure (single or multiple) and a single outcome and most of the times the outcome measurements are binary. So, they need to compare mean or proportion of the outcome at different level of exposure. So, they usually use multiple linear or generalized linear models. The choice of regression model is basically depend on the epidemiological study designs and the type of outcome variable. For example, if it is a population based (cross sectional) study with a binary outcome then logistic or loglineare model is chosen, or if it is a cohort study with an outcome of interest is time-to-vent then cox-proportional hazard model is chosen and if it is a longitudinal follow-up study then multilevel modeling is used. However, in real world there may have multiple outcome measures and multiple exposures. So, the usual statistical methods used in public health research are not applicable to handle multiple outcomes and multiple exposures simultaneously. We often found some public health literature where there are multiple outcomes and multiple exposures and they used the multiple linear models for each individual outcome which leads to increased risk of Type I error (Sherry & Henson 2005; Thompson 1991). However, there are many multivariate statistical methods which are applicable in such situation with reduced Type-I error. For example, multivariate regression analysis, canonical correlation analysis, partial least square regression, structural equation modeling etc.

There may have initially two possible reasons of not using the advanced multivariate statistical methods, one could be lake of knowledge about the advanced statistical methods and the other could be the interpretation complexity of the advanced methods. Additionally, the advanced multivariate statistics are computationally intensive. However, the multivariate statistical methods are gaining popularity with the advent of statistical software.

Additionally, finding a surrogative measure of low birth weight is a classification problem. In literature, there are many classification techniques which can be applied in this problem. Some of them are ROC curve, decision tree, neural networks, random forest, support vector machines. ROC curve is a university classification method and the rests are multivariate methods. Each of method has some advantages and disadvantages over others.

Objectives

As we describe earlier that public health researchers are not using the appropriate multivariate methods for analysis their data when they have multiple outcomes and multiple exposures. We want to explore the potential of multivariate statistical methods which are able to handle multiple outcomes and multiple exposures simultaneously. We also want to explore different classification methods and choose the best one in classifying low birth weight. So, our primary objective is to explore the

association between the birth size and maternal factors and also identify the influential variables in this association using canonical correlation analysis (CCA) and Partial Least squares (PLS) regression and our secondary objective is to find a surrogative measure of low birth weight using an appropriate classification method.

Outline of the thesis paper

This thesis paper consists of six chapters. The introduction chapter discusses the rational and objectives of this research. Sources of data, data collection methods, variables used and the descriptive statistics of each of the variables discussed in chapter 2. Canonical correlation analysis (CCA) of infant size at birth and maternal factors discussed in Chapter 3. Chapter 4 discussed about the methods and applications of Partial Least Squares (PLS) regression to study relationship between size at birth and maternal factors. We also try to identify the surrogative measure of low birth weight in this population using machine learning methods which is discussed in Chapter 5. In Chapter 6, we discussed the summary of this thesis and scope of further research.

References

Barker, D. J., C. N. Martyn, C. Osmond, C. N. Hales & C. H. Fall, 1993. Growth in utero and serum cholesterol concentrations in adult life. *Bmj,* 307(6918), 1524-7.

Hales, C. N., D. J. Barker, P. M. Clark, L. J. Cox, C. Fall, C. Osmond & P. D. Winter, 1991. Fetal and infant growth and impaired glucose tolerance at age 64. *Bmj,* 303(6809), 1019-22.
Hinchliffe, S. A., M. R. J. Lynch, P. H. Sargent, C. V. Howard & D. v. Velzen, 1992. The effect of intrauterine growth retardation on the development of renal nephrons. *BJOG: An International Journal of Obstetrics & Gynaecology,* 99(4), 296-301.

Khatun, S. & M. Rahman, 2008. Quality of antenatal care and its doseâ€“response relationship with birth weight in a maternal and child health training institute in Bangladesh. *Journal of biosocial science,* 40(03), 321-37.

Leal, M. d. C., S. G. N. d. Gama & C. B. d. Cunha, 2006. Consequences of sociodemographic inequalities on birth weight. *Revista de Saúde Pública,* 40(3), 466-73.

Neggers, Y., R. L. Goldenberg, S. P. Cliver, H. J. Hoffman & G. R. Cutter, 1995. The relationship between maternal and neonatal anthropometric measurements in term newborns. *Obstetrics & Gynecology,* 85(2), 192-6.

Osmond, C., D. J. Barker, P. D. Winter, C. H. Fall & S. J. Simmonds, 1993. Early growth and death from cardiovascular disease in women. *Bmj,* 307(6918), 1519-24.

Øyen, N., R. Skjærven, R. E. Little & A. J. Wilcox, 1995. Fetal growth retardation in sudden infant death syndrome (SIDS) babies and their siblings. *American journal of epidemiology,* 142(1), 84-90.

Sherry, A. & R. K. Henson, 2005. Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of personality assessment,* 84(1), 37-48.

Silva, L. M., L. van Rossem, P. W. Jansen, A. C. S. Hokken-Koelega, H. A. Moll, A. Hofman, J. P.

Mackenbach, V. W. V. Jaddoe & H. Raat, 2012. Children of low socioeconomic status show accelerated linear growth in early childhood; results from the generation R study. *PloS one,* 7(5), e37356.

Taylor, D. J. & P. W. Howie, 1989. Fetal growth achievement and neurodevelopmental disability. *BJOG: An International Journal of Obstetrics & Gynaecology,* 96(7), 789-94.

Thame, M., R. J. Wilks, N. McFarlane-Anderson, F. I. Bennett & T. E. Forrester, 1997. Relationship between maternal nutritional status and infant's weight and body proportions at birth. *European journal of clinical nutrition,* 51(3), 134-8.

Thompson, B., 1991. A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development,* 24(2), 80-95.

Williams, S., I. M. St George & P. A. Silva, 1992. Intrauterine growth retardation and blood pressure at age seven and eighteen. *Journal of clinical epidemiology,* 45(11), 1257-63.

Winslow & A. Charles-Edward, 1920. The Untiltled Fields of Public Health. *Science,* 51(1306), 23-33.

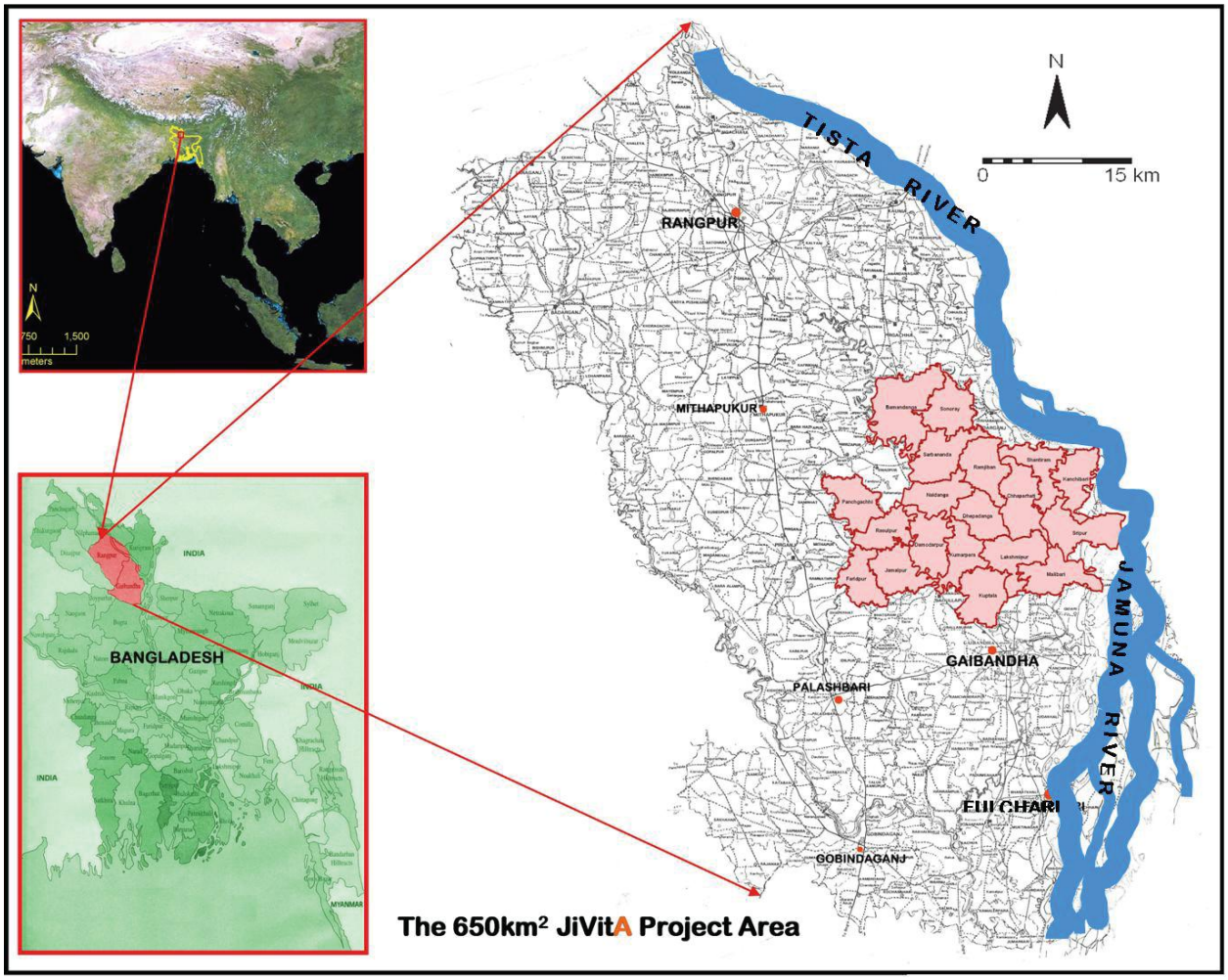
Chapter 2: Data Source and Variables of Interest

Data Source

The data used in this thesis were collected during a field based double-masked, cluster randomized, placebo-controlled trial called JiVitA-1 assessing the efficacy of maternal vitamin A or β-carotene supplementation on maternal and infant mortality through 6 months of age from January 2002 to July 2007. This was a research project of the Department of International Health, Johns Hopkins University, USA. Aim of this study was to conduct the trial in a reasonably accessible, typical rural setting of the country that could be identified from reviews of available reports on population, health, nutrition, agriculture and infrastructure, site visits, and discussions with officials, community groups, and health and nutrition experts. The study area was chosen in the northwest Districts of Gaibandha and Rangpur comprising 19 rural unions (Figure 2.1). This area lies in the 35th percentile of economic and quality of life in rural Bangladesh, based on comparative reports of food production, flood risks and other environmental hazards, maternal and child diet and nutritional and health status and health care utilization reflecting resonance with the national rural context of Bangladesh (Labrique et al. 2011). During March to December 2000, over 120,000 households with ~650000 population in this study area were located and placed on

digitized land-allocation maps as part of a JiVitA Geographic Information System (GIS), described elsewhere (Sugimoto et al. 2007).

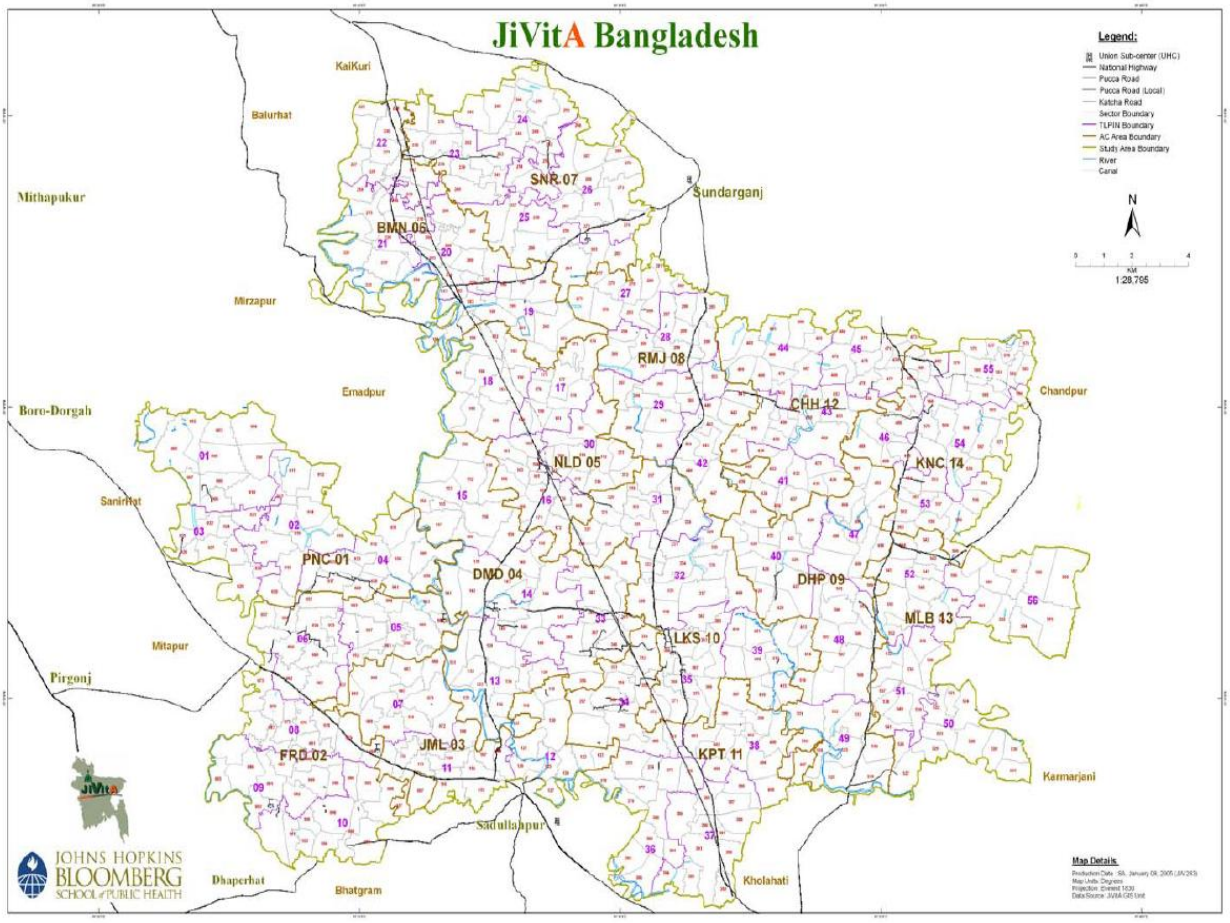
Figure 2-1: JiVitA study area



The area was segregated in to 596 units for both work management and randomization, called “sectors”, each consists of ~250 households and surveillance covered by a single resident field worker (Figure 2.2). A census enumerated 102771 women of reproductive age, married and living with their husbands were eligible to

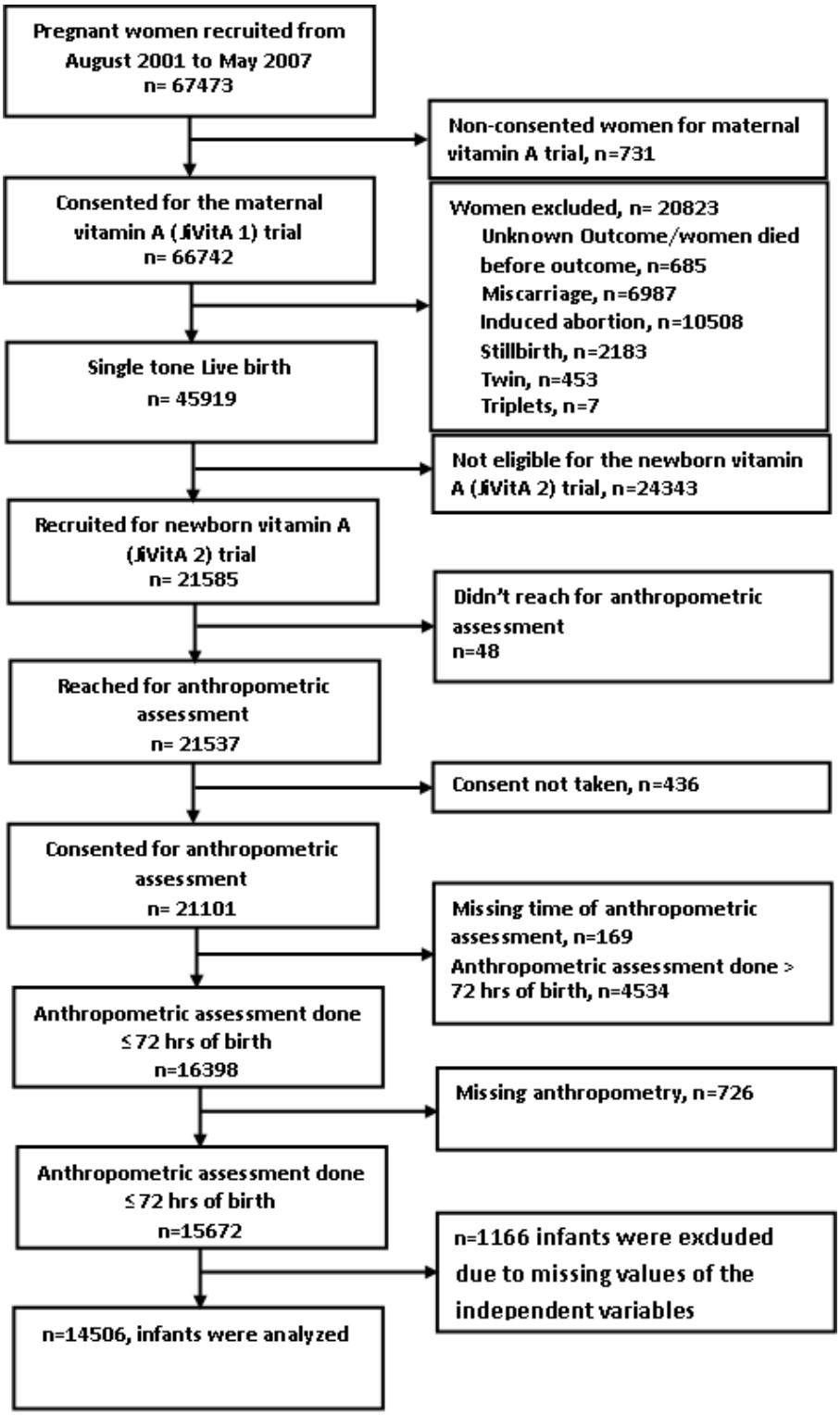
participate in the JiVitA-1 trial. Every 5 weeks thereafter, households were revisited and addresses updated, vital status of enlisted women was recorded, and newly married women were registered. Registrants were delisted if they had permanently moved, become menopausal or had been sterilized, died, or if their husbands had died. In every visit women were asked about their menses in the past 30 days; if amenstrual, they were offered a urine-based hCG pregnancy test (Orchid Biomedical Systems, Goa, India).

Figure 2-2: Sector map of JiVitA area



Following a positive result, consenting pregnant women were administered a coded supplement each week, through 3 months postpartum. As the intervention was cluster randomized, all enrolled pregnant women in a given sector received a weekly capsule containing one of the following masked ingredients: (1) Placebo (consisting of soybean oil with a small amount of vitamin E as an antioxidant), (2) VA (consisting of 7000 µg retinol equivalents, or 23,300 IU, of VA palmitate in soybean oil with a small amount of vitamin E as an antioxidant), or (3) Beta-carotene (consisting of 42 mg of all-trans β-carotene, equivalent to 7000 µg of retinol equivalents, assuming a 6:1 conversion ratio). During the entire enrollment period 59721 pregnant women were recruited. Among them 46379 women had livebirth; singleton 45919, twins 453, triplets 7. The newborn vitamin A trial was initiated 4 years after the maternal vitamin A trial started, 21585 singleton livebirth infant was recruited for dosing. Among these singleton-dosed newborn, 16398 had any anthropometric measurement measured within 72 hrs of birth and 15672 had all 5 anthropometric measurements. Finally, we came up with 14506 infants (Figure 2.3) who had no missing values for other maternal variables of interest and they were analyzed.

Figure 2-3: Assembling the study population



Assessment of outcome

Maternal and infant mortality

Maternal vital status was assessed weekly as women were visited by female supplement distributors through 3 months post partum. Family members of deceased women were visited by a trained research physician, who conducted a “Maternal Verbal Autopsy” interview. Fetal loss due to miscarriage or stillbirth was defined as a pregnancy that ended without a live birth before 28 weeks or ≥ 28 weeks gestation, respectively. Induced abortions, or ‘menstrual regulations’, were also registered as outcomes. Pregnancy losses were assessed weekly during supplementation visits to the home and confirmed with a pregnancy test following the reported loss. Women who experienced a pregnancy loss were visited within a week or two of the reported loss by a study interviewer who conducted a “Miscarriage/Stillbirth” interview. Infant vital status was assessed weekly through 12 weeks of age by a study interviewer and monthly thereafter. Parents of deceased infants were visited by a trained interviewer who conducted an “Infant Verbal Autopsy” interview.

Gestational duration and birth size assessment

Gestational age was assessed by using the reported first date of the last menstrual period (LMP). The recall of this date was aided by the 5-weekly assessment of menstrual histories that were used to administer pregnancy tests. The LMP date was ascertained at the time of the pregnancy enrollment interview, which was conducted

soon after detection of pregnancy, thereby shortening the recall period. Almost 85% of the pregnant women were enrolled within 12 wk of gestation. Local events calendars were used to facilitate recall. In addition, we also cross-checked the reported LMP with the date of the positive urine test. Date of birth is also important in the assessment of the gestational length of a pregnancy. Our birth notification system allowed our workers to know precisely when an infant was born and to record the date and time of birth.

In 2004, after 3 year of enrollment in the study, a second trial of newborn vitamin A supplementation, called JiVitA-2, was nested within the larger pregnancy trial to test the effect of a 50,000-IU single dose on 6-mo infant mortality (Klemm et al. 2008). As part of this nested study, a new birth notification system was set up to reach newborns, of whom 90% were delivered at home. The newborns were visited for vitamin A supplementation with median (inter-quartile range, IQR) time: 7 (2,18) hors of birth. In JiVitA-2 we also initiated birth assessments conducted by our trained team of 56 female anthropometrists, who were equipped with measuring scales and length boards and visited the homes of the newborns to conduct an interview and measure infant size, weight, length, and head, chest, and arm circumferences. Neither procedure was in place for infants born before February 2004. Birth weight was measured by using a Tanita BD-585 digital scale (Tanita Corporation) (Figure 3), which measured weight to

Figure 2-0-4: Birth size measurement

the nearest 10 g. Length was measured to the nearest 0.1 cm by using an infant length board modified after the Infant Shorr board (Shorr Productions) but with the use of local materials. Circumferential measurements were made by using a Ross insertion tape (Abbott Laboratories). Each measurement other than weight was made 3 times, and the median of each was used as the final measurement.

Maternal and household socio-demographic characteristics

On enrollment, women were visited by trained female interviewers to conduct a baseline interview related to their previous pregnancy history and to obtain a 7- and 30-d history of morbidities experienced, 7-d food-frequency intakes, work patterns, and consumption of cigarettes, alcohol, chewing tobacco, and betel nut. In addition, interviews were conducted to elicit data on household sociodemographic characteristics, asset ownership, and house construction. Principal component analysis (PCA) based Living standard index (LSI) was constructed using the asset ownership and house construction variables and the LSI was used in the analysis as a proxy indicator of socio-economic status. The detailed of LSI construction described elsewhere (Gunnsteinsson et al. 2010). Midupper arm circumference (MUAC) of pregnant women was measured in triplicate by using an insertion tape to derive a median of 3 measurements. A three month postpartum interview was conducted to obtain data on maternal diet and morbidity, ANC, events and care during labor and delivery, and conditions of the infant.

Variables to be used and their distribution

Predefined household clusters consisting approximately 250 households called sector (n= 596) were randomized to receive study supplements (Figure 3.2). Married women of reproductive age were enumerated through a baseline census and a subsequent 5 weekly surveillance was carried out to include newly married women. A

5-weekly visit was conducted to assess menstrual history. When a woman reported having missed her menstrual period in the past 30 days, pregnancy was confirmed using human chorionic gonadotropin based on the spot urine test. Once a woman was ascertained her pregnancy, she was asked for consent to receive study supplementation and providing data. Throughout the enrollment period 59721 pregnant women consented and enrolled into the trial.

On enrollment into the trial, mothers were interviewed about household socioeconomic conditions, education, demographic characteristics, previous pregnancy history, frequencies of dietary intake and morbidity in the previous 7 days and measured for mid-upper arm circumference (MUAC) (West et al. 2011). A Living Standard Index (LSI) was constructed using principal component analysis from household socio-economic variables and was used as the main socio-economic variable (Gunnsteinsson et al. 2010). Mothers were visited, provided allocated supplements (vitamin A, b-carotene or placebo) and checked for pregnancy and vital status throughout pregnancy to 3 months post-partum, at which time another interview was completed to obtain further data on maternal diet and morbidity, ANC, events and care during labor and delivery, and conditions of the infant.

Birth anthropometry was collected on infants of consenting mothers who took part in a placebo-controlled newborn vitamin A supplementation trial that was nested into the latter half of the above maternal trial (Klemm et al. 2008). Live-born infants (n= 21,585) were visited for dosing by field staff as soon as possible after birth (median

(Inter Quartile Range, IQR) hrs: 7 (2, 18)). Of this number, 16,290 infants (75%) were singletons who were subsequently visited and measured by trained one of 56 anthropometrists within 72 hours of birth (median (IQR) hrs: 18 (9, 36) and included in the present analysis. Birth size measurements included weight, length, MUAC and head and chest circumferences. Birth weight was measured to the nearest 10 g using a Tanita BD-585 digital pediatric scale (Tanita Corporation, Tokyo, Japan). Length was measured to the nearest 0.1 cm using an affixed headboard and movable footplate that had been fashioned for use with the Tanita scale. Circumferential measurements were made to the nearest 0.1 cm with a Ross insertion tape (Abbott Laboratories, Columbus, OH). All measurements, except for weight, were measured in triplicate with the median taken as the accepted value, as previously described (Christian et al. 2013). The cut-offs used to define a small infant are, weight (<2.5 kg), MUAC (<10 cm), head circumference (<33 cm) and chest circumference (<30.5 cm) (Dhar et al. 2002). Among the 16,290 infants on whom birth anthropometry was collected, 14,506 (89%) had complete data and were included in the CCA which does not allow missing values. The maternal characteristics included in the present analysis are: age at enrollment, parity, early pregnancy mid upper arm circumference (MUAC, cm), education (yrs), LSI, number of ANC visits, and maternal trial supplementation (Vitamin A or bcarotene). Additional infant characteristics included preterm (,37 week of gestation) delivery status and sex. Descriptive statistics of all the variables included in this study were presented in Table 2.1. We observed that more than 50% of the infants were born small. That is they were born with weight

Table 2-1: Descriptive statistics of the indicators of infant's size at birth measured ≤ 72 hrs and maternal socio-demographic factors from rural north west Bangladesh in 2002-2007

Variables	Mean (SD)	Median (IQR)
Indicators of Infant's Size		
Weight, kg	2.44 (0.42)	2.44 (2.18, 2.71)
Length, cm	46.43 (2.41)	46.50 (45.10, 48.00)
MUAC, cm	9.31 (0.84)	9.30 (8.80, 9.90)
HC, cm	32.36 (1.63)	32.50 (31.40, 33.40)
CC, cm	30.40 (2.09)	30.50 (29.20, 31.70)
Maternal Socio-Demographic Factors		
Parity	1.18 (1.41)	1.00 (0.00, 2.00)
Age at enrollment, year	21.96 (5.88)	21.00 (17.00, 26.00)
Early pregnancy MUAC, cm	22.99 (1.97)	22.90 (21.60, 24.10)
Living Standard Index (LSI)	0.08 (0.96)	−0.11 (−0.65, 0.67)
Years of education	3.84 (3.86)	3.00 (0.00, 7.00)
No. of ANC visit	0.52 (1.15)	0.00 (0.00, 1.00)
Vitamin A supplementation	0.34 (0.47)	0.00 (0.00, 1.00)
β-carotene supp supplementation	0.33 (0.47)	0.00 (0.00, 1.00)
Preterm delivery[1]	0.27 (0.44)	0.00 (0.00, 1.00)
Infant's gender (M=1, F=0)	0.51 (0.50)	1.00 (0.00, 1.00)

Abbreviations: ANC, Antenatal Care; CC, Chest Circumference; HC, Head Circumference; MUAC, Mid-Upper Arm Circumference

[1]Any delivery occurred before 37 weeks of gestation

<2.5kg, MUAC<10 cm, head circumference< 33 cm and chest circumference< 30.5 cm. Twenty seven percent of infants were preterm. Half of the infants were male. Mean (SD) maternal age was 22.0 (5.9) years and MUAC was 23.0 (2.0) cm. Most of the women (74%) had not reported an ANC visit. Nearly half of the women (~43%)

were nulliparous and their mean (SD) parity was 1.2 (1.4). Half of the women were literate (52%) and their mean (SD) years of schooling was 3.8 (3.9).

Reference

Christian, P., R. Klemm, A. A. Shamim, H. Ali, M. Rashid, S. Shaikh, L. Wu, S. Mehra, A. Labrique & J. Katz, 2013. Effects of vitamin A and β-carotene supplementation on birth size and length of gestation in rural Bangladesh: a cluster-randomized trial. *The American journal of clinical nutrition,* 97(1), 188-94.

Dhar, B., G. Mowlah, S. Nahar & N. Islam, 2002. Birth-weight Status of Newborns and Its Relationship with Other Anthropometric Parameters in a Public Maternity Hospital in Dhaka, Bangladesh. *Journal of Health, Population and Nutrition (JHPN),* 20(1), 36-41.

Gunnsteinsson, S., A. B. Labrique, K. P. West Jr, P. Christian, S. Mehra, A. A. Shamim, M. Rashid, J. Katz & R. D. W. Klemm, 2010. Constructing indices of rural living standards in Northwestern Bangladesh. *Journal of health, population, and nutrition,* 28(5), 509-19.

Klemm, R. D. W., A. B. Labrique, P. Christian, M. Rashid, A. A. Shamim, J. Katz, A. Sommer & K. P. West, 2008. Newborn vitamin A supplementation reduced infant mortality in rural Bangladesh. *Pediatrics,* 122(1), e242-e50.

Labrique, A. B., P. Christian, R. D. W. Klemm, M. Rashid, A. A. Shamim, A. Massie, K. Schulze, A. Hackman & K. P. West, 2011. A cluster-randomized, placebo-controlled, maternal vitamin A or beta-carotene supplementation trial in Bangladesh: design and methods. *Trials,* 12(1), 102.

Sugimoto, J. D., A. B. Labrique, A. Salahuddin, M. Rashid, R. D. W. Klemm, P. Christian & K. P. West Jr, 2007. Development and management of a geographic information system for health research in a developing-country setting: a case study from Bangladesh. *Journal of health, population, and nutrition,* 25(4), 436.

West, K. P., P. Christian, A. B. Labrique, M. Rashid, A. A. Shamim, R. D. W. Klemm, A. B. Massie, S. Mehra, K. J. Schulze & H. Ali, 2011. Effects of vitamin A or beta carotene supplementation on pregnancy-related mortality and infant mortality in rural Bangladesh. *JAMA: the journal of the American Medical Association,* 305(19), 1986-95.

Chapter 3: Canonical Correlation Analysis of Infant Size at Birth and Maternal Factors

Abstract

This analysis was conducted to explore the association between 5 birth size measurements (weight, length and head, chest and mid-upper arm [MUAC] circumferences) as dependent variables and 10 maternal factors as independent variables using canonical correlation analysis (CCA). CCA considers simultaneously sets of dependent and independent variables and, thus, generates a substantially reduced type 1 error. Data were from women delivering a singleton live birth (n=14506) while participating in a double-masked, cluster-randomized, placebo-controlled maternal vitamin A or β-carotene supplementation trial in rural Bangladesh. The first canonical correlation was 0.42 (P<0.001), demonstrating a moderate positive correlation mainly between the 5 birth size measurements and 5 maternal factors (preterm delivery, early pregnancy MUAC, infant sex, age and parity). A significant interaction between infant sex and preterm delivery on birth size was also revealed from the score plot. Thirteen percent of birth size variability was explained by the composite score of the maternal factors (Redundancy, $R_{Y/X}$=0.131). Given an ability to accommodate numerous relationships and reduce complexities of multiple comparisons, CCA identified the 5 maternal variables able to predict birth size in this rural Bangladesh setting. CCA may offer an efficient, practical and inclusive approach to assessing the association between two sets of variables, addressing the innate complexity of interactions.

Introduction

When exploring the health effects of different exposures, observational epidemiologic studies often deal with data that include both a set of exposure variables and a set of

outcome variables. Routine statistical approaches such as multiple linear regression used to analyze the relationship between exposures and outcomes such as birth size are usually challenged by the potential issues of multiple testing and multicollinearity (Sherry & Henson 2005; Thompson 1991). In some literatures, authors made an effort of analyzing birth size and other maternal, social or environmental variables (Neggers et al. 1995; Rahman et al. 2009; Bhargava 2000; Elshibly & Schmalisch 2009; Ogbonna et al. 2007) used multiple linear regression for analysis despite its limitations. Since CCA assesses the correlation between two composite variables called canonical variate , one representing a set of the exposure variables and the other a set of outcome variables (Sherry & Henson 2005; Thompson 1991), it may be a useful method to evaluate the effect of maternal factors on infant's size at birth.CCA is the most general case of general linear model (Thompson 1991; Fornell 1978; Baggaley 1981; Thompson 1998) and thus it can be used to conduct the univariate and multivariate analyses that CCA subsumes, including multiple regression as a special case (Henson 2000). CCA has several advantages for researchers which were described elsewhere (Fish 1988; Maxwell 1992). Thus CCA is technically able to analyze data involving multiple sets of variables and is theoretically consistent with that purpose (Thompson 1991). Although CCA is used currently in many branches of research: social and behavioral research (Sherry & Henson 2005), bioinformatics (Tripathi et al. 2008), genetics (Naylor et al. 2010), neural network (Bruguier et al. 2008), environmental research (Liu et al. 2009) etc, it is relatively uncommon in public health research and to our knowledge, CCA has not

been applied to analyze the relationship between maternal factors and birth size. The aim of this chapter is to explore the relationship between birth size and maternal factors using CCA. We also want to identify the influential variables in the relationship and the significant interactions between variables.

Canonical Correlation Analysis (CCA)

Canonical correlation analysis is a multivariate statistical model that facilitates the study of linear interrelationships between two sets of variables: one set of variables is referred to as independent and the other as dependent; a canonical variate (i.e. a linear combination) is formed for each set. CCA develops a canonical function that maximizes the correlation between the two canonical variates. Additionally, CCA develops multiple canonical functions; each function is independent (orthogonal) from the others so that they represent different relationships among the sets of dependent and independent variables. CCA can be used for both metric and nonmetric data of either dependent and independent variables (Bartlett 1941). The loadings of the individual variables differ in each canonical function and represent variables' contributions to the specific relationship being investigated.

Suppose we are given a set of p independent variables, **X,** and a set of q dependent variables, **Y**. CCA seeks a number of linear combinations of the two sets of variables

such that they assume maximum correlation across the two data sets, while the transformation within each data set are uncorrelated (Stewart & Love 1968). Suppose

$U = a^T X = a_1 x_1 + a_2 x_2 + ... + a_p x_p$ and $V = b^T Y = b_1 y_1 + b_2 y_2 + ... + b_q y_q$

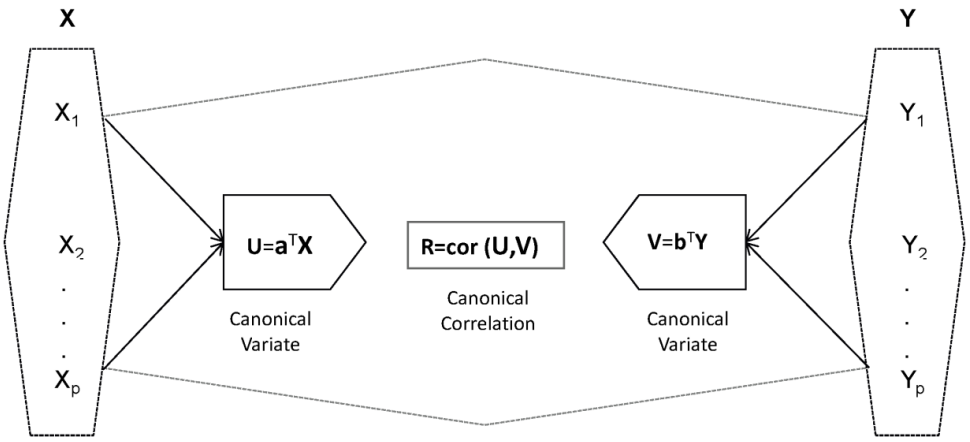
are found to be the sets of transformed variates which maximize the correlation between them. Therefore, the canonical function can be defined as

$$\rho = Corr(U,V) = \max_{a,b} corr(a^T X, b^T Y)$$

Where **U** and **V** are called canonical variates and **a** and **b** are called canonical weights similar to regression weights for **X** and **Y,** respectively. The weights **a** and **b** are chosen to maximize the correlation between the canonical variates **U** and **V** (Figure 1). The first function creates $\mathbf{U_1}$ and $\mathbf{V_1}$ with the highest possible correlation and the unexplainable residual variance left over in the two variable sets. The second function creates $\mathbf{U_2}$ and The challenge is to choose how many functions should be interpreted, however, in most cases the first function will be the most legitimate. Hair et al. (1998) suggested 3 criteria of choosing the important functions as he believed that the use of a single criterion such as the level of significance is too superficial. Because the canonical variates are chosen to maximize the correlation between them, they don't care how much variability they take in to account of each set. The 3 criteria are: (i) level of significance (ii) magnitude of the canonical correlation, and (iii)

redundancy measure for the percentage of variance accounted for from the two data sets like multiple regression's R^2 statistic.

Figure 3-1: Canonical Correlation Analysis



For this analysis all variables were standardized as z-score before applying CCA to have standardized weights. We interpreted the most widely used test, the F statistic (Lambert & Durand 1975), with a level of significance set at 0.05. No generally accepted guidelines have been established regarding suitable sizes for canonical correlations. The decision is usually based on the contribution of the findings to better understand the research problem studied. Because canonical correlations do not give the variance shared between the two sets of variables, a redundancy index is useful for interpretation. The redundancy index is analogous perfectly to the R^2 statistic in multiple regressions. The redundancy index (González et al.

2008)calculates the amount of variance in one set of variables that can be explained by the variables in the other set. According to Sherry and Hensen (2005), any function using this approach that explains <10% of the remaining variance after that explained by a certain number of functions, even if it has significant correlation, the effect sizes of the other functions are considered less impressive. In this paper we applied the criterion of a correlation significance level of 5% and redundancy coefficient of >0.10 to choose the interpretable canonical functions.

If the canonical relationship is statistically significant and the magnitudes of the canonical correlation and the redundancy index are acceptable, the researcher still needs to make substantive interpretation of the results. Making these interpretations involves examining the canonical functions to determine the relative importance of each of the original variables in the canonical relationships. Three methods have been proposed (i) canonical weights (standardized coefficients), (ii) canonical loadings (structural correlations) and (iii) canonical cross-loadings to determine the relative importance of individual variables in to the canonical function. As the canonical weights, like regression weights, are vulnerable to multicollinearity, most of the literatures suggest using canonical loadings or crossing loadings to investigate the relative importance of individual variables in to the canonical relationship. We used both loadings and cross loadings to interpret the canonical variates, however, there is no established cut off for choosing the important contributing variables in to the canonical functions. However, there is a rule of thumb if any variable loading is

>|0.30| then it can be considered to be an important contributing variable in to the function (Lindley et al. 1999). The score plot of canonical variate also helped to find natural variable groupings in to the data set (Stoch & Smythe 1963). We first plotted two canonical variates for the maternal factors, the 1st variate on the horizontal axis and the 2nd variate on the vertical axis.

Multiple linear regression was used to examine the relationship between birth size and maternal factors and to compare the results with canonical correlation analysis. Five models, one for each infant's size variable, were fitted with (i) 10 maternal factors and (ii) with only that factors which had significant loadings (≥30) in the canonical correlation analysis.

To support our CCA findings we stratified our samples by prematurity status and infant sex and investigated their interaction on birth size. We used mean and 95% CI of the 5 anthropometric measurements for 4 strata (Term-Female, Term-Male, Preterm-Female and Preterm-Male). MANOVA was used to investigate interaction effects on infant's size at birth. All analyses were performed using statistical software R version 2.14.1. We used the CCA and yacca R packages.

Results

Table 3-1: Pair wise Pearson's correlation coefficient, r (p-value), between the indicators of birth size, measured ≤ 72 hrs of birth, and maternal socio-demographic factors

	Birth size				
Maternal factors	**Weight, kg**	**Length, cm**	**MUAC, cm**	**HC, cm**	**CC, cm**
Parity	0.14*	0.12*	0.13*	0.10*	0.14*
Age at enrollment, year	0.15*	0.13*	0.13*	0.10*	0.15*
Early pregnancy MUAC, cm	0.17*	0.13*	0.16*	0.13*	0.15*
Living standard index	0.10*	0.09*	0.09*	0.09*	0.09*
Years education	0.04*	0.04*	0.05*	0.05*	0.04*
No. of ANC visit	0.10*	0.09*	0.09*	0.08*	0.08*
Vitamin A supplementation	0.01	0.01	0.00	0.01	0.00
β-carotene supplementation	−0.02	−0.01	−0.01	−0.01	−0.01
Preterm delivery	−0.27*	−0.28*	−0.23*	−0.27*	−0.28*
Infant sex (M=1, F=0)	0.10*	0.12*	0.02**	0.18*	0.06*

ANC: Antenatal Care; CC: Chest Circumference; HC: Head Circumference; MUAC: Mid-Upper Arm Circumference. * P<0.001, **P<0.01

Table 3.1 represents the Pearson's correlation coefficient between the maternal factors and infant's size at birth. All maternal variables except preterm delivery and vitamin A or β-carotene supplementation were positively correlated with infant size at birth. All the infant's anthropometric measurements were negatively correlated with

preterm delivery (P<0.05 for all), however, there was no correlation with maternal vitamin A or β-carotene supplementation.

Table 3-2: Canonical correlation analysis of birth size and maternal socio-demographic factors

Canonical variates	Canonical Correlation	F-statistic	P-value	Redundancy Index, $R_{Y/X}$
Variate-1	0.422	71.977	<0.0001	0.131
Variate-2	0.192	18.483	<0.0001	0.004
Variate-3	0.079	4.939	<0.0001	0.000
Variate-4	0.036	1.933	0.019	0.000
Variate-5	0.024	1.417	0.204	0.000

The canonical correlation coefficients and the redundancy indices are presented in Table 3.2. The canonical correlation analysis is restricted to deriving 5 functions because the dependent set contained the minimum number of 5 variables. The correlations for each successive function were 0.42, 0.19, 0.08, 0.04 and 0.02. All correlations except for the last were statistically significant (P<0.05, F-test). However, the redundancy index for all functions except the first one was zero. Therefore, only the first function is noteworthy in the context of this study.

Table 3-3: Canonical weights, loadings and cross-loadings for the 1st canonical variates of the indicators of birth sizeand maternal factors

Variables	Loadings	Cross loadings
Independent variables		
Parity	0.32	0.13
Age, year	0.34	0.14
Early pregnancy MUAC, cm	0.37	0.15
Living standard index	0.23	0.10
Years of education	0.12	0.05
No. of ANC visit	0.23	0.10
Preterm delivery	-0.74	-0.31
Vitamin A supplementation	0.03	0.01
β-carotene supplementation	-0.03	-0.01
Infant sex (M=1, F=0)	0.35	0.15
Dependent variables		
Weight, kg	0.91	0.38
Length, cm	0.88	0.37
MUAC, cm	0.72	0.31
Head circumference (HC), cm	0.89	0.37
Chest circumference (CC), cm	0.87	0.37

ANC: Antenatal Care; CC: Chest Circumference; HC: Head Circumference; MUAC: Mid-Upper Arm Circumference

Table 3-4: Regression analysis of influence of maternal factors on birth size using canonical correlation analysis

Predictors	Weight, kg		Length, cm		MUAC, cm		CC, cm		HC, cm	
	Model 1* β (p-value)	Model 2* β (p-value)	Model 1 β (p-value)	Model 2 β (p-value)	Model 1 β (p-value)	Model 2 β (p-value)	Model 1 β (p-value)	Model 2 β (p-value)	Model 1 β (p-value)	Model 2 β (p--value)
Parity	0.14 (<0.001)	0.10 (<0.001)	0.11 (<0.001)	0.07 (<0.001)	0.14 (<0.001)	0.09 (<0.001)	0.11 (<0.001)	0.07 (<0.001)	0.14 (<0.001)	0.10 (<0.001)
Age	0.07 (<0.001)	0.08 (<0.001)	0.08 (<0.001)	0.09 (<0.001)	0.05 (<0.001)	0.07 (<0.001)	0.05 (<0.001)	0.06 (<0.001)	0.07 (<0.001)	0.08 (<0.001)
Early pregnancy MUAC	0.12 (<0.001)	0.15 (<0.001)	0.08 (<0.001)	0.10 (<0.001)	0.12 (<0.001)	0.14 (<0.001)	0.09 (<0.001)	0.11 (<0.001)	0.11 (<0.001)	0.13 (<0.001)
LSI	0.05 (<0.001)	-	0.05 (<0.001)	-	0.05 (<0.001)	-	0.04 (<0.001)	-	0.05 (<0.001)	-
Education	0.04 (<0.001)	-	0.04 (<0.001)	-	0.04 (<0.001)	-	0.04 (<0.001)	-	0.04 (<0.001)	-
No. of ANC visit	0.06 (<0.001)	-	0.05 (<0.001)	-	0.05 (<0.001)	-	0.05 (<0.001)	-	0.05 (<0.001)	-
Male	0.11 (<0.001)	0.11 (<0.001)	0.13 (<0.001)	0.13 (<0.001)	0.02 (<0.001)	0.02 (0.003)	0.19 (<0.001)	0.18 (<0.001)	0.07 (<0.001)	0.07 (<0.001)
Preterm	-0.27 (<0.001)	-0.28 (<0.001)	-0.28 (<0.001)	-0.29 (<0.001)	-0.22 (<0.001)	-0.23 (<0.001)	-0.27 (<0.001)	-0.28 (<0.001)	-0.28 (<0.001)	-0.29 (<0.001)
Vitamin A supp.	0.00 (0.722)	-	0.00 (0.747)	-	-0.01 (0.308)	-	0.01 (0.297)	-	-0.01 (0.288)	-
β-carotene supp.	-0.02 (0.043)	-	-0.01 (0.401)	-	-0.01 (0.103)	-	0.00 (0.707)	-	-0.02 (0.025)	-
R^2	**0.15**	**0.14**	**0.14**	**0.13**	**0.11**	**0.09**	**0.14**	**0.13**	**0.14**	**0.13**

MUAC: Mid-Upper Arm Circumference; CC: Chest Circumference; HC: Head Circumference.
*Model 1 consists of all variables and model 2 consists of variables for whose canonical loadings ≥0.30.

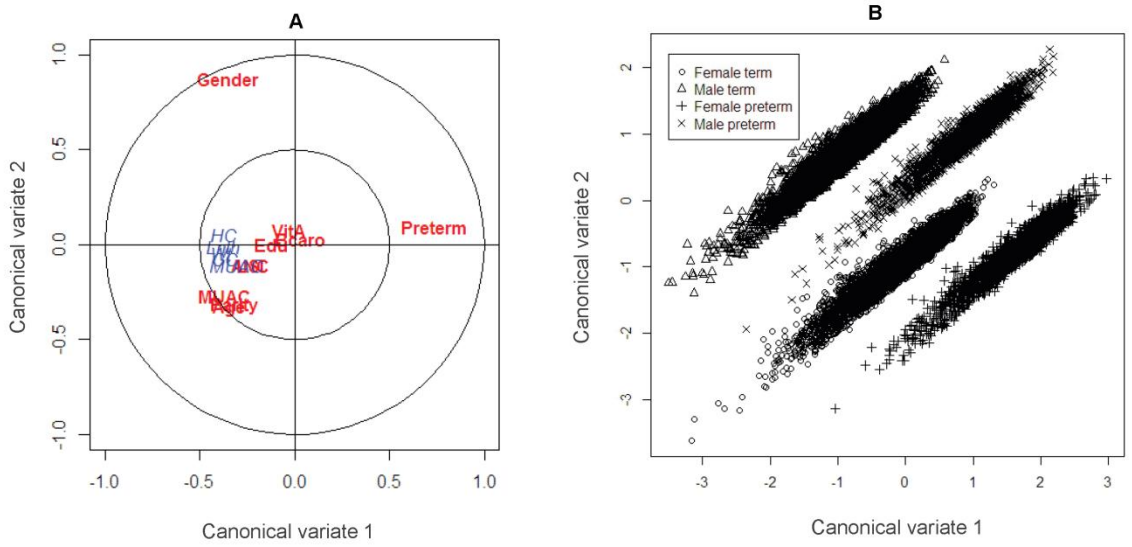
The loadings and cross loadings of the variables for the 1st canonical function are presented in Table 3.3. Looking at the loadings of the variables for function 1 the most important predictor of birth size was preterm delivery (loading: -0.74) followed by maternal early pregnancy MUAC (loading: 0.37), infant's sex (loading: 0.35), maternal age (loading: 0.34) and parity (loading: 0.32). Loadings of the birth size indicators demonstrated that all the anthropometric measurements similarly contributed to the first canonical function. So, all the infant's anthropometric measurements were most strongly negatively correlated with preterm delivery, and positively associated with maternal early pregnancy MUAC, infant sex, age and parity, in that order.

Regression coefficients are presented in Table 3.4. In the models with all 10 maternal factors, except vitamin A and β-carotene supplementation all other factors were significant predictors of infant size at birth. However, in all the models with 5 maternal factors selected through CCA, all 5 factors were significant predictors of infant size at birth. The differences between the coefficients of determination, R2 of the full models and the models with 5 variables varied from 0.01 to 0.02.

Figure 3.2 shows the biplot of the standardized weights for the first two canonical functions for both the maternal factors and infant's anthropometric variables and score plot for the first two canonical variates of the maternal factors. Panel A of Figure 2 illustrates that among the maternal factors preterm delivery had the greatest influence on first canonical function and infant's sex had greatest influence on the

second canonical function but maternal early pregnancy MUAC, age and parity had similar influence on both functions and maternal vitamin A and β-carotene supplementation and maternal education had no influence on either function. The infant size variables had no influence on the second function which implies that

Figure 3-2: Biplot and score plot of the first two canonical functions. Panel A displays the biplot for the indicators of birth size and maternal factors with standardized weights indicated with blue and red, respectively. Panel B displays score plots for the mate



most of the variability in infant size was accounted for by the first canonical variate. Panel B of the Figure 2 shows the score plot of the first and second canonical variates of maternal factors. Four different groups among the infants are revealed.

The grouping results from the interaction effect of preterm delivery and infant sex as they dominate the relationship.

Table 3.5 presents stratum wise mean and 95% confidence interval of birth size. Birth size was significantly different across stratum. Multivariate analysis of variance (MANOVA) showed a significant interaction effect of preterm delivery and infant's sex on birth size; F=161.83, p<0.001.

Table 3-5: Interaction effects of infant sex and preterm delivery on birth size

	Term (Gestational age ≥37wk), Mean (95% CI)		**Preterm (Gestational age <37wk), Mean (95% CI)**	
Birth size	**Female n=5300**	**Male n=5302**	**Female n=1822**	**Male n=2082**
Weight, kg	2.46 (2.45, 2.47)	2.56 (2.55, 2.57)	2.21 (2.19, 2.23)	2.28 (2.27, 2.30)
Length, cm	46.53 (46.47, 46.58)	47.17 (47.11, 47.23)	45.02 (44.89, 45.14)	45.59 (45.47, 45.70)
MUAC, cm	9.41 (9.39, 9.43)	9.45 (9.43, 9.47)	8.97 (8.92, 9.01)	9.02 (8.98, 9.06)
CC, cm	30.61 (30.57, 30.66)	30.91 (30.86, 30.96)	29.27 (29.17, 29.38)	29.56 (29.46, 29.66)
HC, cm	32.31(32.27, 32.35)	32.95 (32.91, 32.99)	31.35 (31.27, 31.43)	31.88 (31.80, 31.96)
MANOVA: F= 161.83, P<0.001				

CC: Chest Circumference; HC: Head Circumference; MUAC: Mid-Upper Arm Circumference

Discussion

We studied the association between birth size and maternal factors using canonical CCA. CCA was used instead of separate linear regression models for each birth size measurement because it simultaneously models effects of multiple independent variables on multiple dependent variables. As CCA uses information from all the variables in both the exposure and outcome variable sets and maximizes the estimation of the relationship between the two sets, CCA may offer a more efficient approach for assessing the effects of the maternal factors on infant size at birth than methods routinely used, such as multiple linear regression. CCA starts with simultaneous consideration of both exposure and outcome variables, limiting the inefficiencies that may accompany conventional multiple testing, and, thus, reducing type-1 error. Furthermore, in CCA the latent variable approach, as used, helped to avoid multicollinearity (Liu et al. 2009). The resulting procedure gives a global view of association between indicators of infant size at birth and maternal factors. We found that infant size at birth in rural Bangladesh had significant but moderate association with maternal nutritional and socioeconomic factors. In addition to providing an assessment of the association between two sets of variables, the application of CCA helped in narrowing down fewer exposure (maternal factors) and outcome variables (birth size) that might contribute to the relationship based on the variable loadings to the composite scores. Thus, CCA could be used as a comprehensive approach to extracting information from data to simultaneously identify both key exposure and

outcome variables so that the assessment of the relationship between an individual exposure and an outcome can be further preceded. Additionally, CCA revealed a significant interaction between preterm delivery and infant's sex on birth size through the score plot of composite scores.

Because the birth size measurements are highly correlated, the combination of the indicators captures more information and, thus, as a composite variable may better predict future health outcomes more efficiently than use of a single birth size measure. For example, head circumference, as an indicator of brain volume (Lindley et al. 1999), may provide important diagnostic and prognostic information, for example related to neurocognitive function (Stoch & Smythe 1963), beyond that provided by birth weight alone. So too, might it be expected that, along with birth weight, other indicators of birth size like length and head, chest and arm circumferences can provide additional information about a wider range of health outcomes related to future child growth, health and development.

WHO suggests that a population with a prevalence of low birth weight of 15% or more or a prevalence of chest circumference at birth < 30 cm experiences a disproportionately elevated risk of infant mortality and morbidity and long-term adverse effects on childhood growth and performance (World Health Organization 1995). We found that approximately half of the infants in this typical rural, Bangladeshi population (Labrique et al. 2011), were born both low birth weight (Christian et al. 2013) and small in chest circumference (<30 cm), revealing a major

public health concern and a subset of infants whose health risks may extend beyond those associated with either criterion alone .

Pearson's correlation coefficients showed that maternal factors, age, parity, MUAC in early pregnancy, LSI of socioeconomic well being, maternal education, number of ANC visits and infant sex were significantly positively associated with birth size whereas, expectedly, preterm delivery was strongly negatively associated with newborn size measures. The individual multiple linear regression analyses also depicted virtually identical results, i.e. in all 5 models, except vitamin A and β-carotene supplementation, all other predictors had significant β-coefficients ($P<0.05$) (data not shown). Christian and colleagues (Christian et al. 2013) also found no significant effect of maternal vitamin A or β-carotene supplementation on newborn's anthropometry in the same population. CCA reduced the number of factors necessary to predict birth size to age, parity, early pregnancy MUAC, infant sex, and preterm delivery (loadings >0.30). If CCA was performed with these 5 predictors instead of 10 then canonical correlation would remain almost the same, $\rho=0.41$ (data not shown). Thus, if CCA was not used before fitting the regression model we would have 3 redundant variables as significant predictors of infant's size. So in addition to evaluating the association between two sets of variables, CCA can also be used as a data mining tool in that it was able to narrow down fewer exposure and outcome variables which might contribute to the relationship.

The score plot of the composite scores can also identify the effect of interaction between factors on outcome of interest (González et al. 2008). The composite scores are the projection of original multidimensional variables to a lower dimension subject to constraint that the correlation between the composite scores of dependent and independent variable sets is maximized. That is, the composite score for the maternal factors was constructed to mirror multiple dimensions of infant size at birth. The effect of interaction between independent variables on the dependent variables was depicted in the score plot of 1st and 2nd composite score of the independent variables. In this study, following the canonical correlation analysis, the multivariate analysis of variance indicated that infant sex and preterm delivery displayed a significant interaction effect on birth size. Infant size was bigger for the male term followed by female term, male preterm and female preterm. Many literature also found this kind of interaction effect on birth size (Storms & Van Howe 2004).

In conclusion, CCA was used to explore the significant association between infant's size at birth and maternal factors. The maternal factors affecting or not affecting infant size at birth, isolated through canonical correlation analysis, were consistent with evidence of these kinds of associations in the literature (Bhargava 2000; Elshibly & Schmalisch 2009; Ogbonna et al. 2007; Hosain et al. 2006; Feleke & Enquoselassie 1999; Yunis et al. 2007; Karim & Mascie-Taylor 1997; Matin et al. 2008; Raum et al. 2001). CCA may offer an efficient, practical and more biologically comprehensive approach to assessing the association between two sets of variables,

by taking into account the innate complexity of interactions and biological pathways that between variables.

References

Baggaley, A. R., 1981. Multivariate Analysis An Introduction for Consumers of Behavioral Research. *Evaluation Review,* 5(1), 123-31.

Bartlett, M. S., 1941. The statistical significance of canonical correlations. *Biometrika,* 32(1), 29-37.

Bhargava, A., 2000. Modeling the effects of maternal nutritional status and socioeconomic variables on the anthropometric and psychological indicators of Kenyan infants from age 0-6 months. *American journal of physical anthropology,* 111(1), 89-104.

Bruguier, A., K. Preuschoff, S. Quartz & P. Bossaerts, 2008. Investigating signal integration with canonical correlation analysis of fMRI brain activation data. *Neuroimage,* 41(1), 35-44.

Christian, P., R. Klemm, A. A. Shamim, H. Ali, M. Rashid, S. Shaikh, L. Wu, S. Mehra, A. Labrique & J. Katz, 2013. Effects of vitamin A and β-carotene supplementation on birth size and length of gestation in rural Bangladesh: a cluster-randomized trial. *The American journal of clinical nutrition,* 97(1), 188-94.

Elshibly, E. M. & G. Schmalisch, 2009. Relationship between maternal and newborn anthropometric measurements in Sudan. *Pediatrics International,* 51(3), 326-31.

Feleke, Y. & F. Enquoselassie, 1999. Maternal age, parity and gestational age on the size of the newborn in Addis Ababa. *East African medical journal,* 76(8), 468-71.

Fish, L. J., 1988. Why multivariate methods are usually vital. *Measurement and Evaluation in Counseling and Development,* 21(2), 130-7.

Fornell, C., 1978. Three approaches to canonical analysis. *Journal of the Market Research Society,* 20(3), 166-81.

González, I., S. Déjean, P. G. P. Martin & A. Baccini, 2008. CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software,* 23(12).

Henson, R. K., 2000. Demystifying parametric analyses: Illustrating canonical correlation analysis as the multivariate general linear model. *Multiple Linear Regression Viewpoints,* 26(1), 11-9.

Hosain, G. M. M., N. Chatterjee, A. Begum & S. C. Saha, 2006. Factors associated with low birthweight in rural Bangladesh. *Journal of tropical pediatrics,* 52(2), 87-91.

Karim, E. & C. G. N. Mascie-Taylor, 1997. The association between birthweight, sociodemographic variables and maternal anthropometry in an urban sample from Dhaka, Bangladesh. *Annals of human biology,* 24(5), 387-401.

Labrique, A. B., P. Christian, R. D. W. Klemm, M. Rashid, A. A. Shamim, A. Massie, K. Schulze, A. Hackman & K. P. West, 2011. A cluster-randomized, placebo-controlled, maternal vitamin A or beta-carotene supplementation trial in Bangladesh: design and methods. *Trials,* 12(1), 102.

Lambert, Z. V. & R. M. Durand, 1975. Some precautions in using canonical analysis. *Journal of Marketing Research,* 12(4), 468-75.

Lindley, A. A., J. E. Benson, C. Grimes, T. M. Cole Iii & A. A. Herman, 1999. The relationship in neonates between clinically measured head circumference and brain volume estimated from head CT-scans. *Early human development,* 56(1), 17-29.

Liu, J., W. Drane, X. Liu & T. Wu, 2009. Examination of the relationships between environmental exposures to volatile organic compounds and biochemical liver tests: application of canonical correlation analysis. *Environmental research,* 109(2), 193-9.

Matin, A., S. K. Azimul, A. K. M. Matiur, S. Shamianaz, J. H. Shabnam & T. Islam, 2008. Maternal Socioeconomic and Nutritional Determinants of Low Birth Weight in Urban area of Bangladesh. *Journal of Dhaka Medical College,* 17(2), 83-7.

Maxwell, S., 1992. Recent developments in MANOVA applications. *Advances in social science methodology,* 2, 137-68.

Naylor, M. G., X. Lin, S. T. Weiss, B. A. Raby & C. Lange, 2010. Using canonical correlation analysis to discover genetic regulatory variants. *PloS one,* 5(5), e10395.

Neggers, Y., R. L. Goldenberg, S. P. Cliver, H. J. Hoffman & G. R. Cutter, 1995. The relationship between maternal and neonatal anthropometric measurements in term newborns. *Obstetrics & Gynecology,* 85(2), 192-6.

Ogbonna, C., G. B. Woelk, Y. Ning, S. Mudzamiri, K. Mahomed & M. A. Williams, 2007. Maternal mid-arm circumference and other anthropometric measures of adiposity in relation to infant birth size among Zimbabwean women. *Acta obstetricia et gynecologica Scandinavica,* 86(1), 26-32.

Rahman, A., M. Vahter, A. H. Smith, B. Nermell, M. Yunus, S. El Arifeen, L.-Å. Persson & E.-C. Ekström, 2009. Arsenic exposure during pregnancy and size at birth: a prospective cohort study in Bangladesh. *American journal of epidemiology,* 169(3), 304-12.

Raum, E., B. Arabin, M. Schlaud, U. Walter & F. W. Schwartz, 2001. The impact of maternal education on intrauterine growth: a comparison of former West and East Germany. *International Journal of Epidemiology,* 30(1), 81-7.

Sherry, A. & R. K. Henson, 2005. Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment,* 84(1), 37-48.

Stewart, D. & W. Love, 1968. A general canonical correlation index. *Psychological Bulletin,* 70(3), 160-3.

Stoch, M. B. & P. M. Smythe, 1963. Does undernutrition during infancy inhibit brain growth and subsequent intellectual development? *Archives of Disease in Childhood,* 38(202), 546-52.

Storms, M. R. & R. S. Van Howe, 2004. Birthweight by gestational age and sex at a rural referral center. *Journal of perinatology,* 24(4), 236-40.

Thompson, B., 1991. A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development,* 24(2), 80-95.

Thompson, B., (1998). Five Methodology Errors in Educational Research: The Pantheon of Statistical Significance and Other Faux Pas, in *Annual Meeting of the American Educational Research Association* San Diego, CA, 102.

Tripathi, A., A. Klami & S. Kaski, 2008. Simple integrative preprocessing preserves what is shared in data sources. *BMC bioinformatics,* 9(1), 111.

World Health Organization, 1995. Physical Status: The Use and Interpretation of Anthropometry. Report of a WHO Expert Committee. . *WHO technical report series,* 854, 1-452.

Yunis, K. A., M. Khawaja, H. Beydoun, Y. Nassif, M. Khogali & H. Tamim, 2007. Intrauterine growth standards in a developing country: a study of singleton livebirths at 28-42 weeks' gestation. *Paediatric and perinatal epidemiology,* 21(5), 387-96.

Chapter 4: Partial Least Squares Regression Analysis of Infant Size at Birth and Maternal Factors

Abstract

Size at birth is an important indicator of fetal and neonatal health on an individual and population level. In this rural setting of Bangladesh more than 50% infant born smaller in size. We conducted this study to assess the effect of 10 maternal socio-demographic factors on 5 birth size measurements using Partial Least Squares (PLS) regression. PLS regression is a rarely used multivariate technique in public health research with a promising ability of handling multicollinearity and handling multiple dependent variables simultaneously. It combines features from principal component analysis and multiple linear regression. The data was taken from women with singleton live births (n=14,506) participating in a large community-based, double-masked, cluster-randomized, placebo-controlled maternal vitamin A or β-carotene supplementation trial in rural Bangladesh. All the maternal factors (parity, age, early pregnancy MUAC, living standard index, years of education, no of ANC visit, preterm delivery and infant gender) except maternal vitamin A and β-carotene supplementation had significant ($p<0.001$) effect on infant size at birth. Among them, preterm delivery had the largest negative influence on infant's size ($\beta = -0.27$; $p<0.001$) as expected. Preterm delivery and infant gender had a significant interaction effect on infant size at birth. PLS regression is computationally more powerful than the Principal Component regression. We recommended to use PLS regression in public health research if multicollinearity exist in the data set or to handle multiple dependent variables simultaneously or for both.

Introduction

In exploring the health effects of different exposures, observational epidemiologic studies often deal with data that include both a set of exposure variables and a set of

outcome variables. More often the variables are interrelated to some extent and consequently multicollinearity often exists. Routine statistical approaches such as multiple linear regression or principal component regression are usually challenged with multiple testing. Specifically, using separate statistical significance tests for each regression equation when there are multiple outcomes substantially increases the risk of Type 1 error (Sherry & Henson 2005). Additionally, the multiple linear regression or principal component regression is not able to address the correlation structure among the dependent variables. There is another multivariate regression approach called partial least squares (PLS) regression developed by Harman Wold in the 1960's (Wold 1985; Abdi 2010) which can simultaneously consider more than one dependent variables, thus can address the correlation structure among the dependent variables, and can efficiently handle the severe multicollineary. Thus, PLS regression has the potential to be a useful method to predict infant size at birth from maternal factors. We aim to estimate the effect of maternal factors on infant size at birth using PLS regression. We also compare the performance of PLS and PC regressions predictive ability.

Partial Least Squares Regression

The classical regression methods usually meet four main challenges with: (i) a large number of variables, (ii) correlated predictors, (iii) smaller sample size with a large number of variables and (iv) having more than one response variables simultaneously (Carrascal et al. 2009). To overcome these problems, the researchers

usually take some measures; they may remove some variables (Draper & Smith 1998) or may use multivariate reduction techniques like principal component analysis to reduce the multidimensionality in the predictor or response variables (Jolliffe 1982). However, removing variables may often incur selection of redundant variables which have no significant effect on the response variable. On the other hand, despite the fact that the dimensionality reduction techniques reduces the number of predictors by using latent variables instead, the latent variables are usually derived by maximizing the covariation among the predictors instead of maximizing the covaraiation among the response variables. Consequently, this may produce patterns or syndromes within the predictor variables making little or no biological sense (Carrascal et al. 2009). The appropriate solution of these challenges is using PLS regression (Carrascal et al. 2009).

PLS regression is comparatively new and its use in research is increasing over time. Initially it was using in analytic chemistry (Geladi & Kowalski 1986; Martens et al. 1986; Mevik & Wehrens 2007) but, now it is gaining popularity in many branches of research including public health (Wimberly et al. 2014; Piccolo et al. 2015), bioinformatics (Yan et al. 2015), ecology (Pérez-Rodríguez et al. 2013; Carrascal et al. 2009) and agriculture (Kwak et al. 2015). As it is computationally much more intensive, the advent of computer packages facilitates the researchers to apply it in this era of information technology. The PLS regression has incorporated in many statistical packages such as, R, SAS, STATA, MatLab and STATISTICA.

PLS regression analysis, like principal component regression (PCR), extract a set of orthogonal factors called latent variables which are used as predictor in the regression model and have the best predictive power (Abdi 2010). The major difference with principal component regression (PCR) is that principal components are determined solely by the ***X*** variables, whereas with PLS, both the ***X*** and ***Y*** variables influence the construction of latent variables. The intention of PLS is to form components (latent variables) that capture most of the information in the ***X*** variables that is useful for predicting ***Y*** variables, while reducing the dimensionality of the regression problem by using fewer components than the number of ***X*** variables. PLS is considered especially useful for constructing prediction equations when there are many explanatory variables and comparatively little sample data (Höskuldsson 1988).

The PLS regression identifies the latent variables stored in matrix ***T*** and they model ***X*** and predict ***Y*** simultaneously. Then the following expression can be written as,

$\boldsymbol{X} = \boldsymbol{TP}^{\boldsymbol{T}}$ and $\widehat{\boldsymbol{Y}} = \boldsymbol{TBC}^{\boldsymbol{T}}$ ……. (1)

Where, ***P*** and ***C*** are loadings and ***B*** is diagonal matrix. These latent variables are ordered according to the variance of $\widehat{\boldsymbol{Y}}$ they explain. $\widehat{\boldsymbol{Y}}$ can also be written as

$\widehat{\boldsymbol{Y}} = \boldsymbol{TBC}^{\boldsymbol{T}} = \boldsymbol{XB}_{\boldsymbol{PLS}}$ where, $\boldsymbol{B}_{\boldsymbol{PLS}} = \boldsymbol{P}^{\boldsymbol{T}+}\boldsymbol{BC}^{\boldsymbol{T}}$

$\boldsymbol{P}^{\boldsymbol{T}+}$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{P}^{\boldsymbol{T}}$. The matrix $\boldsymbol{B}_{\boldsymbol{PLS}}$ has J rows and K columns and is equivalent to the regression weights of multiple regression.

The latent variables are computed iteratively using Singular Value Decomposition (SVD). In each iteration, SVD constructs orthogonal latent variables for ***X*** and ***Y*** and corresponding regression weights (Abdi 2010). The algorithm for PLS regression is as follows:

Step 1: Transform ***X*** and ***Y*** in to Z-score and store in matrices X_0 and Y_0

Step 2: Compute the correlation matrix between X_0 and Y_0, $R_1=X_0^T Y_0$

Step 3: Perform singular value decomposition (SVD) on R_1 and produce two sets of orthogonal singular vectors w_1 and c_1 corresponding to the largest singular value, λ_1.

Step 4: The first latent variable for ***X*** is given by $T_1= X_0^T w_1$.

Step 5: Normalize T_1 such that $T_1^T T_1=1$

Step 6: The loadings of X_0 on T_1 is computed as $P_1=X_0^T T_1$ and $\widehat{X}_1 = T_1^T P_1$.

Step 7: Compute $U_1=Y_0 c_1$ and $\widehat{Y}_1 = U_1 c_1^T = T_1 b_1 c_1^T$ ... ... (3),

Where, $b_1=T_1^T U_1$

The scalar b_1 is the slope of the regression of $\widehat{Y}_1$ on T_1. Equation (3) shows that $\widehat{Y}_1$ is obtained as linear regression from the latent variable extracted from X_0. Matrices $\widehat{X}_1$ and $\widehat{Y}_1$ are then subtracted from the original X_0 and Y_0 respectively to give deflated X_1 and Y_1.

Step 8: Compute the input matrices for the next iteration, $X_1=X_0-\widehat{X}_1$ and $Y_1=Y_0-\widehat{X}_1$

Step 9: The first set of latent variables has now been extracted. Now perform SVD on $\boldsymbol{R_2=X_1^TY_1}$, we get $\boldsymbol{w_2}$, $\boldsymbol{c_2}$, $\boldsymbol{T_2}$ and $\boldsymbol{b_2}$ and the new deflated matrices $\boldsymbol{X_2}$ and $\boldsymbol{Y_2}$.

Step 10: The iterative process continues until ***X*** is completely decomposed in to ***L*** components (where ***L*** is the rank of ***X***). When this is done, the weights (i.e., all the w's) for x are stored in the ***J*** by ***L*** matrix ***W*** (whose ***l***-th column is $\boldsymbol{w_l}$).

The latent variables of ***X*** are stored in matrix ***T***, the weights for ***Y*** are stored in ***C***, the latent variables of **Y** are stored in matrix ***U***, the loadings for ***X*** are stored in matrix ***P*** and the regression weights are stored in a diagonal matrix ***B***. The regression weights are used to predict ***Y*** from ***X***.

Now the question is how many components, **t**'s will have to be retained in the final model. The answer can be obtained by comparing the cross validation Root-Mean Squared Error of Prediction (RMSEP) for different number of components. At which component the cross validation RMSEP does have meaningful change will be used in the final model.

To choose the optimum number of components for both PLS and PC regression, root mean squared error of prediction (RMSEP) were calculated using different number of components. We used variables important for projection (VIP) value to identify the predictors which have the significant effect on the responses **Y**. The predictors with a VIP>1 was considered as significant (Shi et al. 2013). We also performed

approximate t tests of regression coefficients based on jackknife variance estimates (Martens & Martens 2000).

We constructed correlation plot of the variables to observe how variables are correlated with each other and also between the birth size variables and maternal variables. The closer a variable appears on the perimeter of the circle, the better it is represented and if two variables are highly correlated they will appear near each other. If two variables are negatively correlated they will tend to appear in opposite extremes. If two variables are uncorrelated, they will be orthogonal to each other. We plotted the scores of first two components, $\mathbf{t}_1$ vs $\mathbf{t}_2$ which helped us to assess if there is any natural grouping or interactions among variables.

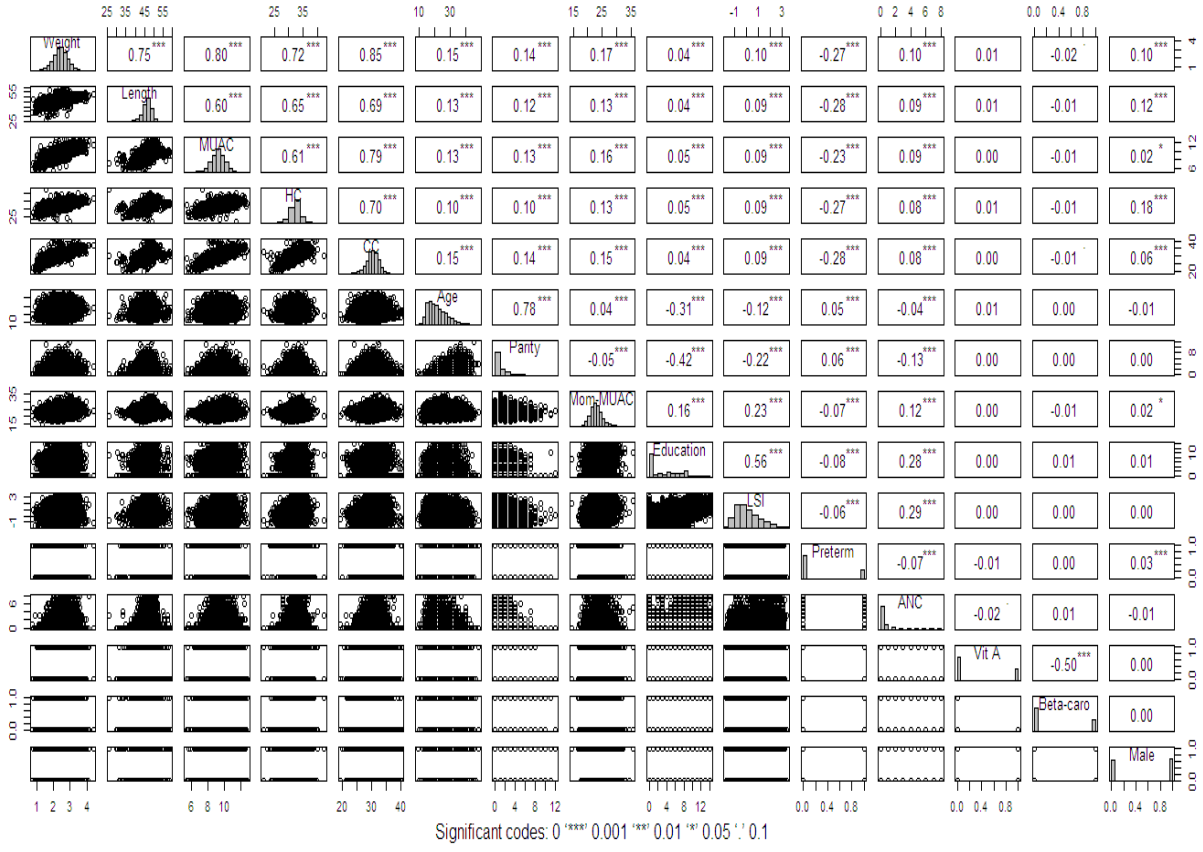
To examine the advantage of PLS regression over PC regression, we calculated Pearson's correlation coefficients between the predicated values (by PLS and PC regression with 1 to 5 components respectively) and the observed values of infant's size variables.

Results

Figure 4.1 simultaneously displays the correlation between variables (superdiagonal), two-way scatter plot (subdiagnal) and the histogram of each variable (diagonal). All the birth size variables are significantly ($p<0.001$) highly correlated with each other. All the predictors except vitamin A and β-carotene supplementation are significantly

(p<0.05) correlated with all the birth size variables. Among them, preterm delivery had highest negative correlation with all the birth size variables (|r|>0.20, p<0.001) and maternal age, parity and MUAC had higher correlation compared to other variables (r≥0.10, p<0.001). Predictors are also correlated with each other, maternal age was highly correlated with parity (r=0.78, p<0.001), maternal education was moderately positively correlated with LSI (r=0.56, p<0.001) and negatively correlated with parity (r= -0.42, p<0.001) and age (r=-0.31, p<0.001).

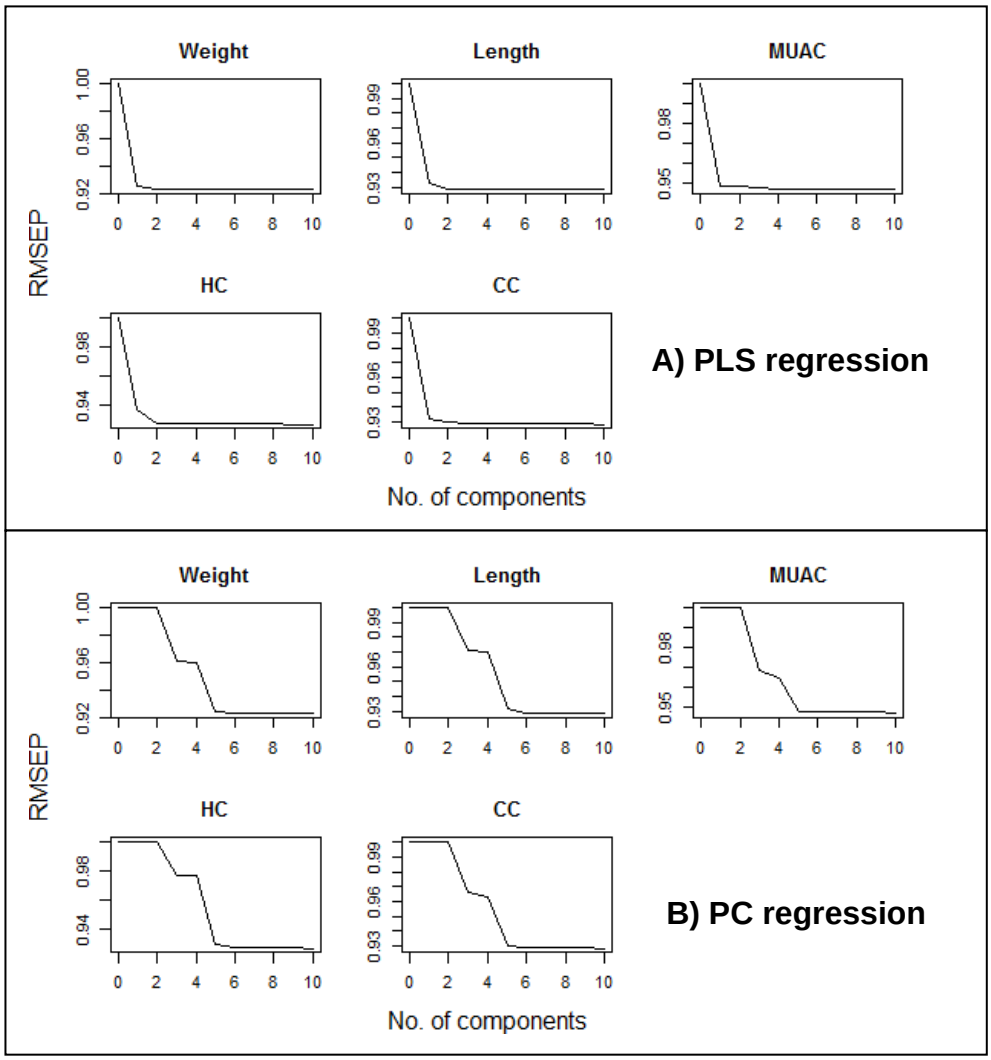
Figure 4-1: Pair wise correlation and scatter plot matrix of the study variables



The root mean squared error of prediction was plotted against number of components used in the PLS and PC regression models in Figure 4.2. This figure suggested to choose 2 components to be included in the PLS regression model and 5 components for the PC regression. Table 2 presents the standardized PLS regression coefficient with 2 components. Except the prenatal supplementation of vitamin A and β-carotene all the variables have significant ($p < 0.001$) effect on infant's size at birth. Preterm delivery was the most influential variable which was negatively associated with infant's size at birth (standardized β= -0.27, -0.27, -0.19, -0.29, and -0.25 for weight, length, MUAC, HC and CC respectively) followed by infant's gender and maternal parity, MUAC, and age.

The correlation plot of the variables for the first two components (Figure 4.3) depicted that maternal education, ANC visit, LSI, age, MUAC and parity are correlated and fall in the 4th quadrant. Among them, age, MUAC and parity are close to each other and they are also closer to the birth size variables, and education, ANC visit, and LSI are very close to each other; however they are less closer to the birth size variables. On the other hand, preterm delivery alone fall in the 3rd quadrant just opposite to the birth size variables and infant's gender alone fall in the 1st quadrant with the birth size variables. Maternal Vitamin A and β-carotene supplementation fall very close to each other. Therefore, this figure also demonstrated that preterm delivery was the most

Figure 4-2: Root mean squared error of prediction (RMSEP) for different number of components of PLS and PC regression



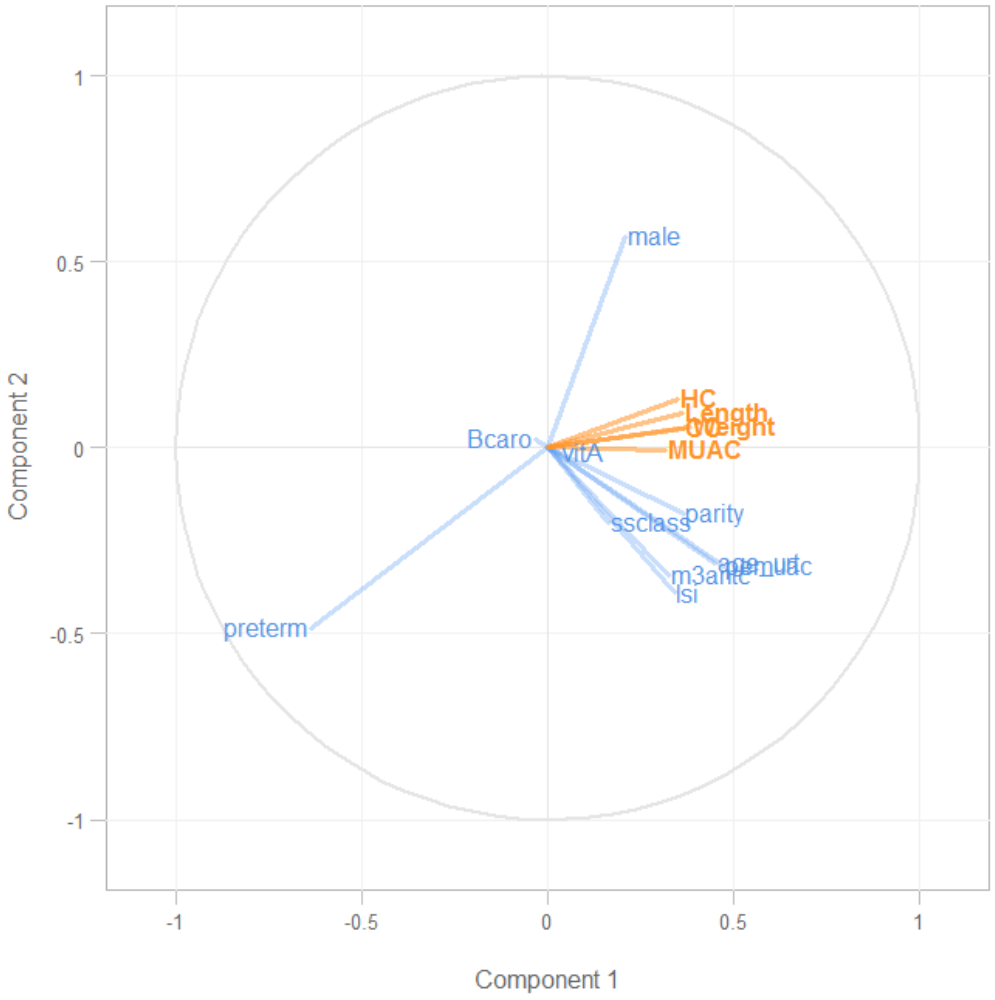
important predictor in opposite direction, followed by infant's gender, parity, age and MUAC.

Table 4-1: Standardized PLS regression coefficients using 2 components with Jackknife SE and p-value to predict infant's size at birth from maternal factors

Maternal factors	Weight		Length		MUAC		HC		CC	
	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value	β (SE)	p-value
Age	0.10 (0.01)	<0.001	0.09 (0.00)	<0.001	0.10 (0.01)	<0.001	0.07 (0.01)	<0.001	0.10 (0.01)	<0.001
Parity	0.11 (0.00)	<0.001	0.10 (0.01)	<0.001	0.09 (0.01)	<0.001	0.09 (0.01)	<0.001	0.10 (0.00)	<0.001
Early pregnancy MUAC	0.11 (0.01)	<0.001	0.10 (0.01)	<0.001	0.11 (0.01)	<0.001	0.09 (0.01)	<0.001	0.11 (0.01)	<0.001
Education	0.03 (0.00)	<0.001	0.03 (0.01)	<0.001	0.03 (0.01)	0.001	0.02 (0.01)	0.009	0.03 (0.00)	<0.001
LSI	0.06 (0.00)	<0.001	0.05 (0.01)	<0.001	0.07 (0.01)	<0.001	0.03 (0.01)	<0.001	0.06 (0.00)	<0.001
Preterm	-0.27 (0.01)	<0.001	-0.27 (0.01)	<0.001	-0.19 (0.01)	<0.001	-0.29 (0.01)	<0.001	-0.25 (0.01)	<0.001
No of ANC visit	0.06 (0.01)	<0.001	0.05 (0.00)	<0.001	0.07 (0.01)	<0.001	0.04 (0.01)	0.001	0.06 (0.01)	<0.001
Vitamin A supp	0.01 (0.01)	0.467	0.00 (0.01)	0.592	0.01 (0.01)	0.371	0.00 (0.01)	0.733	0.00 (0.01)	0.468
β-carotene supp	-0.01 (0.01)	0.178	-0.01 (0.01)	0.224	-0.01 (0.01)	0.142	-0.01 (0.01)	0.300	-0.01 (0.01)	0.176
Male infant	0.12 (0.01)	<0.001	0.12 (0.01)	<0.001	0.07 (0.01)	<0.001	0.16 (0.01)	0.001	0.11 (0.01)	<0.001
R^2	**0.15**	**--**	**0.14**	**--**	**0.10**	**--**	**0.14**	**--**	**0.14**	**--**

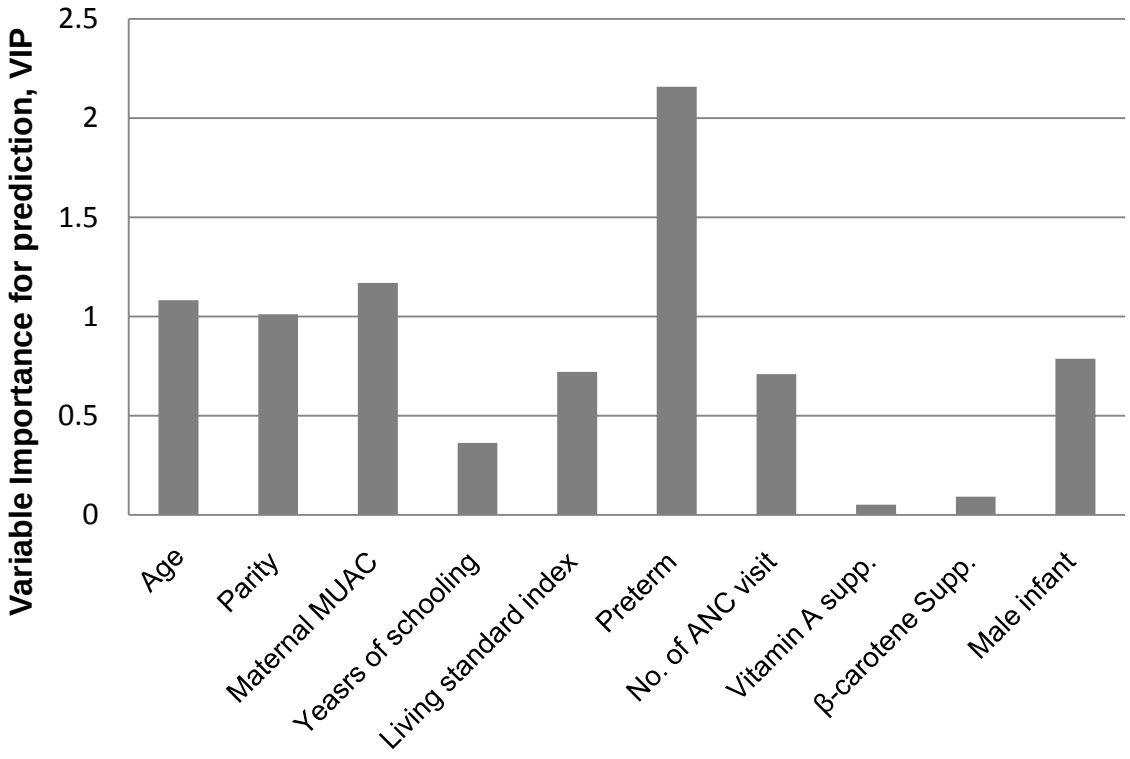
Abbreviations: MUAC,Mid-Upper Arm Circumference; LSI, Living Standard Index; Mom MUAC, Maternal Mid-Upper Arm Circumference.

Figure 4-3: Variables inside the circle of correlation for the 1st and 2nd components (orange color indicates the dependent variables and the blue color indicats the predictors)



Variance important for projection (VIP) value indicated that among the maternal factors, preterm delivery, maternal MUAC, age and parity are the important (VIP>1) ones to predict birth size (Figure 4.4).

Figure 4-4: Variable importance in the projection (VIP) of 10 predictors of birth size by PLS regression

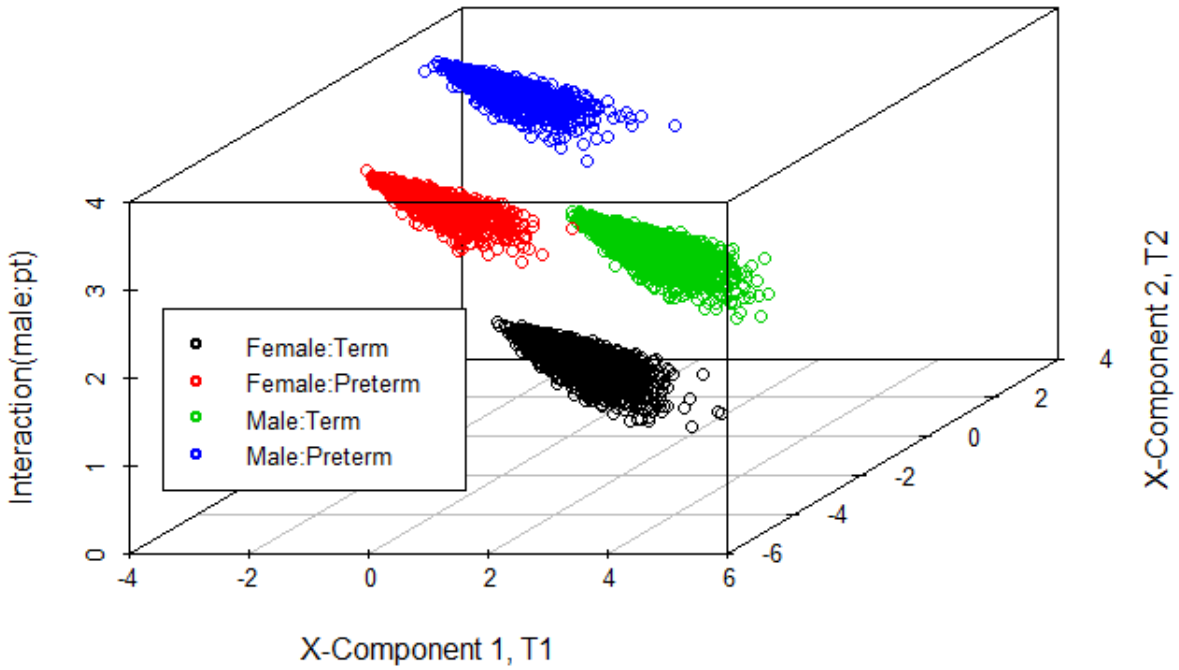


The three dimensional plot with scores of 1st and 2nd components of the predictors and the interactions between male and preterm delivery (male:term, male:preterm, female:term and female:preterm) displayed four different clear groupings among the study participants due to the interaction effect of preterm delivery and infant's gender (Figure 4.5).

We observed that the correlation between the predicted values of birth size variables with PC regression and observed values were increasing with increasing the number

of components, however, the correlation was very poor with the first few components (Table 4.2). On the other hand, prediction with PLS regression did not meaningfully

Figure 4-5: Three dimensional plot of scores of the first two components of predictors and interactions between male and preterm (pt) delivery



change the correlation with increasing the number of components. However, the explained variance of the predictors by the components of PLS regression was always lower than that of the PC regression. Notably, the first 2 PLS components that were derived from maternal factors explained only ~26% of total variation had a predictive ability that was comparable to the PC regression with 5 components that explained ~73% of total variation.

Table 4-2: Comparison between partial least squares and principal component regression: Pearson's correlation coefficient between the predicted and observed value with different number of components

Infant's size	PLS regression					Principal component regression				
	Comp-1	Comp-2	Comp-3	Comp-4	Comp-5	Comp-1	Comp-2	Comp-3	Comp-4	Comp-5
Weight	0.381	0.385	0.386	0.386	0.386	0.003	0.016	0.276	0.281	0.383
Length	0.360	0.372	0.372	0.372	0.372	0.007	0.014	0.240	0.243	0.365
MUAC	0.318	0.318	0.322	0.323	0.324	0.008	0.011	0.251	0.267	0.321
Head circumference	0.351	0.375	0.378	0.378	0.378	0.026	0.030	0.217	0.218	0.371
Chest circumference	0.364	0.368	0.372	0.373	0.373	0.002	0.012	0.257	0.272	0.368
Variance explained (R_x^2)	12.59%	23.64%	37.89%	57.08%	62.84%	23.79%	38.82%	52.92%	63.14%	72.82%

Discussion

We conducted this analysis to assess the effect of maternal socio-demographic factors on infant size at birth using PLS regression. The promising advantage of using PLS regression is that it can mitigate the problem of multicolinearity and it has ability to handle more than one outcome variables simultaneously. The PLS regression revealed that all the maternal variables examined, except vitamin A and β-carotene supplement, had significant effect on birth size. However, maternal education, LSI and no. of ANC visit had practically very minor effects. Preterm delivery had the greatest effect on birth size followed by infant gender, maternal parity, education, and age. We also identified preterm delivery and infant's gender had significant interaction effect on birth size.

PLS regression was used in this study because it is already proved that PLS regression performs equally in ideal situation and it is better than other regression methods which are usually used to handle highly collinear data, such as stepwise multiple regression, PC regression or model fitting techniques that apply Maximum Likelihood or Bayesian theory (Helland 1990; Carrascal et al. 2009). The PLS method simultaneously predicts a set of dependent variables from a set of independent variables, instead of using separate regression models for each dependent variable, eliminating the need for multiple testing and thus reducing the type-I error.

Typically in regression analysis, we always look for a parsimonious model, ie. the model that can predict the response variable at a desired level with as few predictors as possible. Because, however R^2 increases with increasing the number of predictors, variance of regression coefficients also increases. Therefore, it is statistically advantageous to reduce the number of explanatory variables in a given model. PLS can be viewed as a good method to do regression analysis, because, the components are selected so that they describe the dependent variables as much as possible and thus it reduces the number of components to be used in the model for prediction. As a result, the PLS method provides more stable estimates than other regression methods. Criteria that give penalties on the number of variables, like Akaike Information Criterion (AIC), or those where model performance is evaluated, like the Mallows Cp criteria, also give rise to the need to include more variables than the PLS method (Höskuldsson 1988).

In the PLS method, like PC the challenge of choosing an optimum number of latent variables remains in order to obtain the best generalization for the prediction of dependent variables. Based on RMSEP, this also provided the evidence that PLS regression with only 2 components gave the optimum prediction and in contrast PC regression provided similar prediction with 5 components. This is because, in PLS regression, a pair of components is chosen from the dependent and independent variables so that they are closest to each other, however, in PC regression a component is chosen from the independent variables which captures most of the

variability and it does not take into account how close it is to the dependent variables. This is why a smaller number of components in PLS regression is needed to enable better prediction. In PLS regression two components extracted to predict size at birth explained only ~24% variation of the maternal factors; however, for PC regression, 5 components were chosen that explained ~73% variation of the maternal factors. This aspect of efficiency provides a basis for obtaining better prediction without necessarily needing to explain a large amount of variability of the independent variables by the extracted components in PLS regression.

Studying the relationships between infant size at birth and maternal factors have important implications. The combination of all birth size measurements captures more information than a single measurement in isolation; however, they are highly correlated to each other. Head circumference for instance, indicates the brain volume (Lindley et al. 1999) and it may also provide important diagnostic and prognostic information like neurocognitive function (Stoch & Smythe 1963), beyond that provided by birth weight alone. Therefore, it is expected that along with birth weight, other birth size measurements like length and head, chest and arm circumferences can provide more information associated with broader range of health outcomes like future growth, health and development.

It is reported that a population would have a disproportionately high risk of infant mortality and morbidity and long term adverse effect on child growth and

performance if it has low birth weight prevalence ≥15% WHO 1995). The low birth weight prevalence was more than 50% in this rural Bangladeshi population (Kabir et al. 2014; Christian et al. 2013) and thus studying the factors associated with birth size is a major public health concern.

The PLS regression analysis suggested that maternal factors age, parity, early pregnancy MUAC, LSI, maternal education, number of ANC visit and infant sex had significant positive effect on infant's size at birth; however, preterm delivery had the greatest negative effect on birth size as desired. We did not find any effect of maternal vitamin A and β-carotene supplementation on infant's size at birth and which is consistent with the result found by Christian et al. (2013) and Kabir et al. (2014) from the same study. This study also identified that preterm delivery and infant's gender had significant interaction effect on infant's size at birth. Our findings from this study was commensurate with a previous study conducted to find the strength of association between birth size and maternal factors using canonical correlation analysis with the same data set (Kabir et al. 2014).

The correlation plot of first two components of the maternal factors and infant's size at birth indicated that maternal education, no. of ANC visit and LSI are correlated to each other and maternal age and MUAC are also correlated to each other which give the evidence of having colinearity among the predictors. Thus, the PLS regression is justifiable to use in this study.

As the PLS regression is mostly competitive with the PC regression, we compared their performance in terms of their predictive ability with different number of components. We observed that PLS regression with only two components had as similar predictive power as PC regression with 5 components. However, two components in PLS regression explained only about 24% variability of the maternal factors and in contrast, five PC explained about 73% variability. So, our study also depicted that to have a better prediction with the latent variables it is not necessary to have the latent variables should explain most of the variability of the predictors. Thus, PLS regression's coefficients have certainly more stable than the PC regression.

In conclusion, this study conducted to assess the effect of maternal factors on infant's size at birth using an advanced multivariate regression technique, PLS regression. The maternal factors identified as the significant predictors of infant size at birth were commensurate with other studies (Bhargava 2000; Elshibly & Schmalisch 2009; Ogbonna et al. 2007; Hosain et al. 2006; Feleke & Enquoselassie 1999; Yunis et al. 2007; Karim & Mascie-Taylor 1997; Matin et al. 2008; Raum et al. 2001). PLS regression has the promising potential as a multivariate regression method in public health research to address the innate complexity of interactions and biological pathways between variables.

Reference

Abdi, H., 2010. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics,* 2(1), 97-106.

Bhargava, A., 2000. Modeling the effects of maternal nutritional status and socioeconomic variables on the anthropometric and psychological indicators of Kenyan infants from age 0â€“6 months. *American journal of physical anthropology,* 111(1), 89.

Carrascal, L. M., I. GalvÃ¡n & O. Gordo, 2009. Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos,* 118(5), 681-90.

Christian, P., R. Klemm, A. A. Shamim, H. Ali, M. Rashid, S. Shaikh, L. Wu, S. Mehra, A. Labrique & J. Katz, 2013. Effects of vitamin A and β-carotene supplementation on birth size and length of gestation in rural Bangladesh: a cluster-randomized trial. *The American journal of clinical nutrition,* 97(1), 188-94.

Dhar, B., G. Mowlah, S. Nahar & N. Islam, 2002. Birth-weight status of newborns and its relationship with other anthropometric parameters in a public maternity hospital in Dhaka, Bangladesh. *Journal of Health, Population and Nutrition*, 36-41.

Draper, N. R. & H. Smith, (1998). Applied regression analysis 3rd edition, New York: Wiley.

Elshibly, E. M. & G. Schmalisch, 2009. Relationship between maternal and newborn anthropometric measurements in Sudan. *Pediatrics International,* 51(3), 326-31.

Feleke, Y. & F. Enquoselassie, 1999. Maternal age, parity and gestational age on the size of the newborn in Addis Ababa. *East African medical journal,* 76(8), 468-71.

Geladi, P. & B. R. Kowalski, 1986. Partial least-squares regression: a tutorial. *Analytica chimica acta,* 185, 1-17.

Gibson, R. S., 2005. *Principles of nutritional assessment*: Oxford university press.

Gunnsteinsson, S., A. B. Labrique, K. P. West Jr, P. Christian, S. Mehra, A. A. Shamim, M. Rashid, J. Katz & R. D. W. Klemm, 2010. Constructing indices of rural living standards in Northwestern Bangladesh. *Journal of health, population, and nutrition,* 28(5), 509.

Helland, I. S., 1990. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 97-114.

Hosain, G. M. M., N. Chatterjee, A. Begum & S. C. Saha, 2006. Factors associated with low birthweight in rural Bangladesh. *Journal of tropical pediatrics,* 52(2), 87-91.

Höskuldsson, A., 1988. PLS regression methods. *Journal of chemometrics,* (2), 211-28.

Institute of Medicine . Subcommittee on Nutritional, S., P. Weight Gain during, I. Institute of Medicine . Subcommittee on Dietary & P. Nutrient Supplements during, 1990. *Nutrition during pregnancy: part I, weight gain: part II, nutrient supplements*: Natl Academy Pr.

Jolliffe, I. T., 1982. A note on the use of principal components in regression. *Applied Statistics*, 300-3.

Kabir, A., R. D. Merrill, A. A. Shamim, R. D. W. Klemn, A. B. Labrique, P. Christian, K. P. West Jr & M. Nasser, 2014. Canonical Correlation Analysis of Infant's Size at Birth and Maternal Factors: A Study in Rural Northwest Bangladesh. *PloS one,* 9(4), e94243.

Karim, E. & C. G. N. Mascie-Taylor, 1997. The association between birthweight, sociodemographic variables and maternal anthropometry in an urban sample from Dhaka, Bangladesh. *Annals of human biology,* 24(5), 387-401.

Klemm, R. D. W., A. B. Labrique, P. Christian, M. Rashid, A. A. Shamim, J. Katz, A. Sommer & K. P. West, 2008. Newborn vitamin A supplementation reduced infant mortality in rural Bangladesh. *Pediatrics,* 122(1), e242-e50.

Kwak, H. S., B.-H. Ahn, H.-R. Kim & S.-Y. Lee, 2015. Identification of Senory Attributes That Drive the Likeability of Korean Rice Wines by American Panelists. *Journal of food science,* 80(1), S161-S70.

Labrique, A. B., P. Christian, R. D. W. Klemm, M. Rashid, A. A. Shamim, A. Massie, K. Schulze, A. Hackman & K. P. West, 2011. A cluster-randomized, placebo-controlled, maternal vitamin A or beta-carotene supplementation trial in Bangladesh: design and methods. *Trials,* 12(1), 102.

Lindley, A. A., J. E. Benson, C. Grimes, T. M. Cole & A. A. Herman, 1999. The relationship in neonates between clinically measured head circumference and brain volume estimated from head CT-scans. *Early human development,* 56(1), 17-29.

Martens, H., L. Izquierdo, M. Thomassen & M. Martens, 1986. Partial least-squares regression on design variables as an alternative to analysis of variance. *Analytica chimica acta,* 191, 133-48.

Martens, H. & M. Martens, 2000. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food quality and preference,* 11(1), 5-16.

Matin, A., S. K. Azimul, A. K. M. Matiur, S. Shamianaz, J. H. Shabnam & T. Islam, 2008. Maternal socioeconomic and nutritional determinants of low birth weight in urban area of Bangladesh. *Journal of Dhaka Medical College,* 17(2), 83-7.

McCormick, M. C., 1985. The contribution of low birth weight to infant mortality and childhood morbidity. *New England journal of medicine,* 312(2), 82-90.

Mevik, B. r.-H. & R. Wehrens, 2007. The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software,* 18(2), 1-24.

Neggers, Y., R. L. Goldenberg, S. P. Cliver, H. J. Hoffman & G. R. Cutter, 1995. The relationship between maternal and neonatal anthropometric measurements in term newborns. *Obstetrics & Gynecology,* 85(2), 192-6.

Ogbonna, C., G. B. Woelk, Y. Ning, S. Mudzamiri, K. Mahomed & M. A. Williams, 2007. Maternal midâ€arm circumference and other anthropometric measures of adiposity in relation to infant birth size among Zimbabwean women. *Acta obstetricia et gynecologica Scandinavica,* 86(1), 26-32.

Pérez-Rodríguez, A., S. Fernández-González, I. Hera & J. Pérez-Tris, 2013. Finding the appropriate variables to model the distribution of vectorâ€borne parasites with different environmental preferences: climate is not enough. *Global change biology,* 19(11), 3245-53.

Piccolo, B. D., N. L. Keim, O. Fiehn, S. H. Adams, M. D. Van Loan & J. W. Newman, 2015. Habitual Physical Activity and Plasma Metabolomic Patterns Distinguish Individuals with Low vs. High Weight Loss during Controlled Energy Restriction. *The Journal of nutrition,* 145(4), 681-90.

Raum, E., B. Arabin, M. Schlaud, U. Walter & F. W. Schwartz, 2001. The impact of maternal education on intrauterine growth: a comparison of former West and East Germany. *International Journal of Epidemiology,* 30(1), 81-7.

Sherry, A. & R. K. Henson, 2005. Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of personality assessment,* 84(1), 37-48.

Shi, Z. H., L. Ai, X. Li, X. D. Huang, G. L. Wu & W. Liao, 2013. Partial least-squares regression for linking land-cover patterns to soil erosion and sediment yield in watersheds. *Journal of Hydrology,* 498, 165-76.

Stoch, M. B. & P. M. Smythe, 1963. Does undernutrition during infancy inhibit brain growth and subsequent intellectual development? *Archives of Disease in Childhood,* 38(202), 546.

Thame, M., R. J. Wilks, N. McFarlane-Anderson, F. I. Bennett & T. E. Forrester, 1997. Relationship between maternal nutritional status and infant's weight and body proportions at birth. *European journal of clinical nutrition,* 51(3), 134-8.

West, K. P., P. Christian, A. B. Labrique, M. Rashid, A. A. Shamim, R. D. W. Klemm, A. B. Massie, S. Mehra, K. J. Schulze & H. Ali, 2011. Effects of vitamin A or beta carotene

supplementation on pregnancy-related mortality and infant mortality in rural Bangladesh: a cluster randomized trial. *Jama,* 305(19), 1986-95.

Wimberly, M. C., A. Lamsal, P. Giacomo & T.-W. Chuang, 2014. Regional variation of climatic influences on West Nile virus outbreaks in the United States. *The American journal of tropical medicine and hygiene,* 91(4), 677-84.

Wold, H., 1985. Partial least squares. *Encyclopedia of statistical sciences,* 6, 581-91.
World, H. O., 1995. Physical status: The use of and interpretation of anthropometry, Report of a WHO Expert Committee.

Yan, X., J. A. Kruger, P. M. F. Nielsen & M. P. Nash, 2015. Effects of fetal head shape variation on the second stage of labour. *Journal of biomechanics*.

Yunis, K. A., M. Khawaja, H. Beydoun, Y. Nassif, M. Khogali & H. Tamim, 2007. Intrauterine growth standards in a developing country: a study of singleton livebirths at 28-42 weeks' gestation. *Paediatric and perinatal epidemiology,* 21(5), 387-96.

Chapter 5: Prediction of low birth using machine learning techniques

Abstract

Early detection of LBW is very much crucial for adequate care needed for infant survival as well as reducing its magnitude of subsequent adverse effects or catching up normal baby. Unfortunately, many LBW infants are remained undetected due to the paucity of baby weight scale. Our objective was to propose a predictive model for LBW with other easy-to-measure anthropometric measurements using machine learning technique. There is number of machine learning algorithms in literature and finding the optimum one is a major task, as there is no algorithm which is the best for all problems. In this paper we investigated four machine learning algorithms: Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM). We found SVM is the best predictive model of low birth weight. On the other hand, DT is the simplest predictive model amongst the four which can be applied without the help of a computer with highest sensitivity. DT model is more applicable in that study where the researcher or programmer wants to identify as many as LBW as possible keeping false positive or false negative rate <20%. The uniqueness of this study is, we used a very large community based data and used advanced machine learning techniques to predict low birth weight. This study provided the opportunity to the researchers or programmers to choose a optimum predictive model of LBW based on the resources they have without losing much overall accuracy.

Introduction

Low birth weight (LBW) is defined as weight at birth less than 2500 g (WHO 2011). Around 26% infants are born with LBW globally each year. Among them around 97% are from the developing countries (Wardlaw 2004). The rate of LBW significantly varies across the United Nations regions. South-central Asia has the highest

incidence of LBW (27%) and the lowest in Europe (6.4%) (WHO 2011). In Bangladesh around 55% infants have LBW (Klemm et al. 2013). However, the national LBW survey reported to have around 36% LBW (Salam et al. 2004). The consequences of LBW are universally recognized. It contributes to a greater extent to child mortality (Mathews & MacDorman 2012), it also causes long term disability and impaired development (Reichman 2005), delayed motor and social development (Hediger et al. 2002), having a lower IQ (Hack et al. 2002) and many more. Consequently, LBW incurs enormous economic costs, higher medical expenditures, special education and social service expenses and decreased productivity in adulthood. However, the appropriate care including feeding, temperature maintenance, hygienic cord and skin care, and early detection and treatment of complications, can substantially reduce mortality in this highly vulnerable group (WHO 2011). So, it is essential to detect the LBW babies as soon as it borns. Early detection of LBW is very much crucial for adequate care needed for infant survival as well as reducing its magnitude of subsequent adverse effects or catching up normal baby.

Unfortunately, many LBW infants are remained undetected. Because, in resource poor settings like Bangladesh most of the deliveries take place outside the health facility and most of them are attended by neighbor or relatives or traditional birth attendants (Klemm et al. 2008) who are not aware of the importance of recoding birth

weight. Even if a birth takes place at health facility, babies are not weighed routinely due to paucity of a newborn weighing scale at the center. A number of surrogative measures of LBW have been proposed by many literatures (Taksande et al. 2007; Achebe et al. 2013; Sreeramareddy et al. 2008; Mullany et al. 2007), however, most of them are hospital based study and thus the study subjects may not be a representative subset of the target population. Moreover, almost all the surrogative measures are univariate, i.e., authors proposed a single measure with a threshold value as an alternative to the birth weight. So, threshold for other surrogative measures of LBW have been determined by univariate analysis like Receiver Operating Characteristics (ROC) curve. However, multiple measures can be used to increase predictive ability of the model.

There are many machine learning methods like Decision Tree (DT) (Rokach & Maimon 2008), Support Vector Machine (SVM) (Byvatov et al. 2003), Artificial Neural Network (ANN) (Byvatov et al. 2003) and Random Forest (RF) (Friedman et al. 2008) available in literature can be used to develop a predictive model for low birth weight. Machine leaning is a field in computer science which deals with the method of data analysis that automates analytical model building. Machine learning uses algorithms that iteratively learn from data and allows computers to find hidden insights without being explicitly programmed. The entire approach is to build a model by learning from previous experiences and use that model to make predictions for the future observations (Mohri et al. 2012). There is number of machine learning algorithms in

literature and finding the optimum one is a major task. Because, there is no algorithm which is the best for all problems (Müller et al. 2005). In this paper we investigated four machine learning algorithms: Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM). Our objective is to propose a predictive model for LBW with other anthropometric measurements.

The ROC curve

ROC curve (Fawcett 2006) is a graph showing true positive rate on the vertical axis and false positive rate on the horizontal axis, as the classification threshold varies. It is a single curve summarizing the information in the cumulative distribution functions of the scores of the two classes. ROC curve is primarily used in the medical decision making. However, its use in machine learning and data mining research is increasing (Fawcett 2006). It is just a complete representation of classifier performance, as the choice of the classification threshold varies. In this study we used to ROC curve to find an optimum cutoff from each of 4 other anthropometric measurements to predict low birth weight with highest sensitivity (true positive rate) and highest specificity (1-false positive rate).

Suppose that **t** is the value of the threshold **T** in a particular classification, so that an individual is allocated to population classification score **s** exceeds **t** and otherwise to population **N**. In order to assess the efficacy of this classifier we need to calculate the probability of making an incorrect allocation. Such a probability tells us the rate at

which future individuals requiring classification will be misallocated. Given probability densities **p(s| P)**, **p(s| N)**, and the value t, numerical values lying between 0 and 1 can be obtained readily for these four rates and this gives a full description of the performance of the classifier. Clearly, for good performance, we require high “true” and low “false” rates. However, this is for a particular choice of threshold t, and the best choice of t is not generally known in advance but must be determined as part of the classifier construction. Varying **t** and evaluating all the four quantities above will clearly give full information on which to base this decision and hence to assess the performance of the classifier, but since tp + fn = 1 and fp + tn = 1 we do not need so much information. The ROC curve provides a much more easily digestible summary. It is the curve obtained on varying **t**, but using just the true and false positive rates and plotting (fp, tp) as points against orthogonal axes. Here fp is the value on the horizontal axis (abscissa) and tp is the value on the vertical axis (ordinate).

The purpose of the ROC, as indicated above, is to provide an assessment of the classifier over the whole range of potential t values rather than at just a single chosen one. Clearly, the worth of a classifier can be judged by the extent to which the two distributions of its scores p(s| P) and p(s| N) differ. The more they differ, the less in general will there be any overlap between them so the less likely are incorrect allocations to be made, and hence the more successful is the classifier in making correct decisions. Conversely, the more alike are the two distributions, the more overlap there is between them and so the more likely are incorrect allocations to be made.

Decision Tree

Decision tree (DT) (Safavian & Landgrebe 1990) is one of the oldest classification methods. However, it is still a very popular method for classification due to its robustness and simplicity. The goal is to create a model that predicts the value of an outcome variable based on several predictors or features. It is basically used for supervised classification or supervised learning. A DT is like a flowchart. It consists of nodes that form a rooted tree. The node from where the tree starts growing is called root that has no incoming edge. A node with outgoing edges is called an internal or test node. Each internal node represents a “test” on an attribute or a feature, each branch represents the outcome of the test and each leaf node represents a class label which also known as terminal or decision node. Each path from root to leaf represents a decision rule.

Algorithmic Framework for DTs

DT has algorithmic framework that automatically construct a DT from a given dataset. The goal is to find the optimal DT by minimizing the generalization error. There are many algorithms to construct an optimum DT. The method of algorithms is divided in to two groups; top-down and bottom-up. Here we will stick to the top-down algorithms. There are various types of top-down algorithms such as ID3 (Quinlan 1986), C4.5 (Quinlan 1993), CART (Breiman et al. 1984). Some of them consist of both two conceptual phases: growing and pruning and some other perform only the

growing phase. We will use the CART algorithm using the R package: tree. Because, it is better in terms of classifier accuracy (Lakshmi et al. 2013).

The CART algorithm

The algorithm is based on Classification and Regression Trees by Breiman et al (Breiman et al. 1984). A CART tree is a binary DT that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample.

Tree growing process

The basic idea of tree growing is to choose a split among all the possible splits at each node so that the resulting child nodes are the “purest”. In this algorithm, only univariate splits are considered. That is, each split depends on the value of only one predictor variable. All possible splits consist of possible splits of each predictor. If X is a nominal categorical variable of i categories, there are $(2^{i-1} -1)$ possible splits for this predictor. If X is an ordinal categorical or continuous variable with K different values, there are K - 1 different splits on X. Tree growing is started from the root node by repeatedly using the following steps on each node.

1. Find each predictor’s best split: For each continuous and ordinal predictor, sort its values from the smallest to the largest. For the sorted predictor, go through

each value from top to examine each candidate split point (call it v, if x ≤ v, the case goes to the left child node, otherwise, goes to the right.) to determine the best. The best split point is the one that minimize the Gini impurity index the most when the node is split according to it. For each nominal predictor, examine each possible subset of categories (call it A, if xϵA, the case goes to the left child node, otherwise, goes to the right.) to find the best split.

2. Find the node's best split: Among the best splits found in step 1, choose the one that minimizes the Gini Index.
3. Split the node using its best split found in step 2 if the stopping rules are not satisfied.

Pruning

Tight stopping criteria can produce smaller and under-fitted DTs and on the other hand, loose stopping criteria can generate larger DTs and can over-fit the training set. Pruning methods (Breiman et al. 1984) help to get the optimum DT. Pruning determines a nested sequence of subtrees of the supplied tree by recursively "snipping" off the least important splits, based upon the cost-complexity measure.

Random Forest

Random forest (Breiman 2001) is an ensemble learning method for classification. It is also a tree-based classification method. It's a little modification of the concept of the other popular classification method Bagging (Bauer & Kohavi 1999). Bagging

constructs a large number of trees with bootstrap samples, say B. These trees are grown deep and are not pruned. Then, the overall prediction is the most commonly occurring class among the B predictions from the B trees. So, bagging shown to give impressive improvements by combining a large number of trees in a single procedure over using a single tree. The main limitation of bagging is that it uses all the predicators in the training data for all B trees and consequently it produces a large number of correlated trees. Thus, bagging does not lead to as large of a variance reduction as making prediction using many uncorrelated trees. RF has overcome this problem by taking a random subset of the predictors for each tree. Hence, RF is only differs from the bagging only at taking a subset of the predictors. When a large number of trees are bagged, the resulting statistical learning procedure no longer can be presented using a single tree.

Algorithm: RF for Classification

1. For b= 1 to B:
    a. Draw a bootstrap sample **Z*** of size N from the training data.
    b. Grow a random-forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
        i. Select m variables at random from the p variables.
        ii. Pick the best variable and split point among the m.
        iii. Split the node into two child nodes.

2. Output the ensemble of trees $\{\mathrm{T_b}\}_1^B$.

Now make a prediction at a new point x. Let $\hat{\mathrm{C}}_\mathrm{b}(\mathrm{x})$ be the class prediction for the bth random-forest tree. Then $\hat{\mathrm{C}}_{\mathrm{rf}}^{\mathrm{B}}(\mathrm{x}) = Majority\ vote\ \{\hat{C}_b(x)\}_1^B$

We used "randomForest" R-package for this model with tree size B=500 and number of variables randomly chosen for each split was mtry=1. The number of variable used for each split was chosen by a grid search using "Caret" package.

Artificial Neural Network

Artificial neural network (ANN) is a classification method which was designed as a computational model based on the brain to solve certain kind of problems is a mathematical model (Friedman et al. 2008) usually used in pattern recognition and machine learning. The network is constituted by connecting input layer, hidden layer and output layer. There could be only a single hidden layer or could be more. Choice of the number of hidden layers is guided by background knowledge and experimentation. The hidden layers are the weighted sum of the input layers. Each layer consists of some nodes which are called neuron. The neurons for input layer are the input features or predictors and neurons for hidden layers are the weighted sum of the neurons of the previous layer and finally there will be only one neuron for the output layer. The number of neurons for the hidden layers should be reasonably large. It is better to have too many hidden neuron than too few. Each connection from layer to layer is passed on through a weight. The weights are chosen during the

learning phase so as to minimize the prediction cost. Cross-entropy function is used as cost function for classification in ANN and it is minimized by gradient decent method. In the hidden neurons, the ANN model is exactly a linear logistic regression model and the parameters are estimated by maximum likelihood (Friedman et al. 2008). In this study we used “nnet” R-package for ANN model. Through a grid-search using the R-package “Caret” we chose the size of the hidden layer as 6 and the parameter value for “decay” as 0.10.

Support vector machine (SVM)

Support vector machine is typically a kernel-based supervised learning method (Vapnik 2013). SVM can efficiently perform a non-linear classification in addition to linear classification. It is a method of constructing an optimal separating hyperplane or a set of hyperplanes in a high or infinite dimensional space between two perfectly separated classes (Byvatov et al. 2003). It is more focused on the cases where the classes are not separable by a linear boundary. In SVM, the original finite dimensional space is mapped in to a much higher dimensional space so that the separation becomes much easier in that space.

The support vectors are the vectors or points in the training data which are closest to the hyperplane. They are the data points most difficult to classify. Then the decision function is fully specified by a subset (usually very small) of training samples, the support vectors. The separating function can be expressed as a linear combination of kernels associated with the Support Vectors as

$$f(x) = \sum_{x_j \in S} \alpha_j y_j K(x_j, x) + b$$

Where, x_j denotes the p-dimensional vector of training data, S denotes the set of support vectors, α is the corresponding coefficients, K is the kernel function and b is the offset. There are several kernel functions and appropriate kernel has to be chosen based on their classification accuracy. In this analysis we chose the radial basis kernel as it gives more classification accuracy. SVM with radial basis kernel has two hyper parameters, gamma and cost which can be changed by hand. There is a good method for selecting proper values for the parameters which is called "grid-search". The optimum value of the hyper parameters was chosen as gamma=0.8 and cost, C=50 by using grid-search. We used tow R packages for SVM: "e1071" for the SVM model and "Caret" for grid search.

Model's performance assessment

After constructing the classification rule using the training data, the question comes first how effective it will be in assigning future objects to classes. In principle, one could simply apply the rule to the training set, and examine the accuracy with which it classifies those objects. This, however, would be unwise. Since the classification rule has been constructed using the training set it is, in some sense, optimized for those data. After all, since the training data are a sample from the population one is seeking to classify, it would be perverse to have deliberately chosen a classification rule which performed badly on those data. But if the rule has been chosen because it

does well on the training data there is a real possibility that the performance on those data will be better than on another sample drawn from the same distribution. We thus require more subtle assessment methods.

Various approaches have been developed, all hinging around the notion of separating the data set used for constructing the rule from the data set used to evaluate it. The simplest approach is to divide the available data into two sets, a training and test set, the former for choosing the rule and the latter for assessing its performance. This division can be repeated, splitting the data in multiple ways, and averaging the results to avoid misleading estimates arising from the chance way the data are split. Other methods split the data in unbalanced ways, at an extreme putting only one data point in the test set and the others all in the training set, and then repeating this for each available data point, again averaging the results. This is the leave-one-out method. Yet other approaches draw random sets with replacement from the available data, and use each of these as training sets, allowing estimates of the optimism of reclassifying the overall training set to be obtained. These are bootstrap methods. Because of its importance, considerable attention has been devoted to this problem.

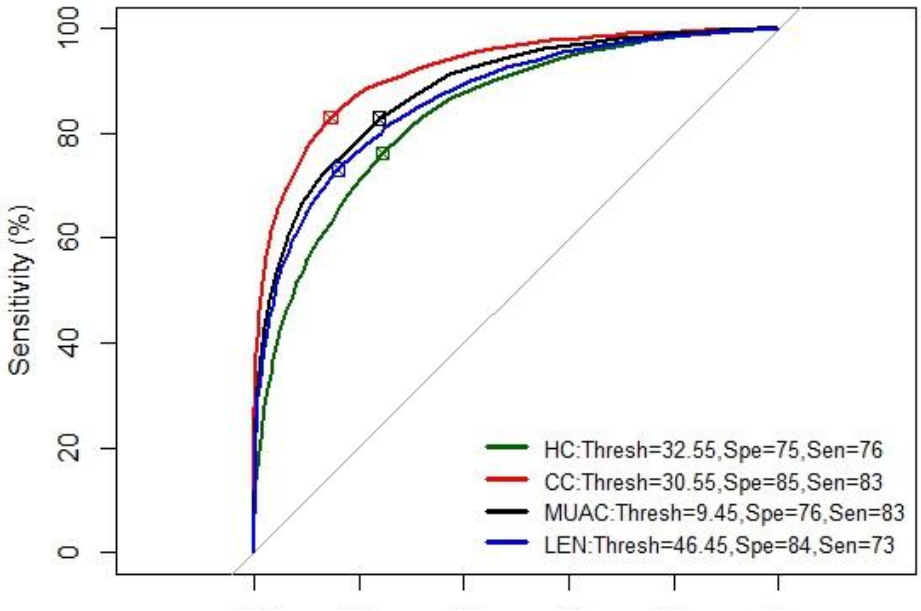
In each supervised learning algorithm, we try find a classification rule so that all the members of positive class and negative class are correctly identified. However, in reality the perfect classification is almost impossible due to the overlapping

characteristics for both positive and negative classes. So, performance is measured by the extent to which a learning algorithm can correctly identify both the positive and negative classes. Many such methods are based on the two-by-two classification table which results from crossclassifying the true class of each object by its predicted class. So in this study we will evaluate machine learning methods by four standard performance measures based on the two-by-two classification table which are accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Accuracy measures the proportion of positive and negative classes that are correctly identified as such, sensitivity measures only the positive class which is correctly identified as such and specificity measures only the negative class which is correctly identified as such. On the other hand, positive predictive value is the proportion of positive class predicted by a machine learning method which is truly positive and the negative predictive value is the proportion of negative class predicted by a machine learning method which is truly positive. A higher result of performance measure indicating higher reliability of a prediction model. Performance of all prediction models will be compared for both training and test data sets.

Results

Figure 5.1 presents the ROC curve for birth length, MUAC, head circumference and chest circumference as a predictor of low birth weight. Chest circumference was

Figure 5-1: Threshold for other anthropometric measurement to predict low birth weight using ROC curve

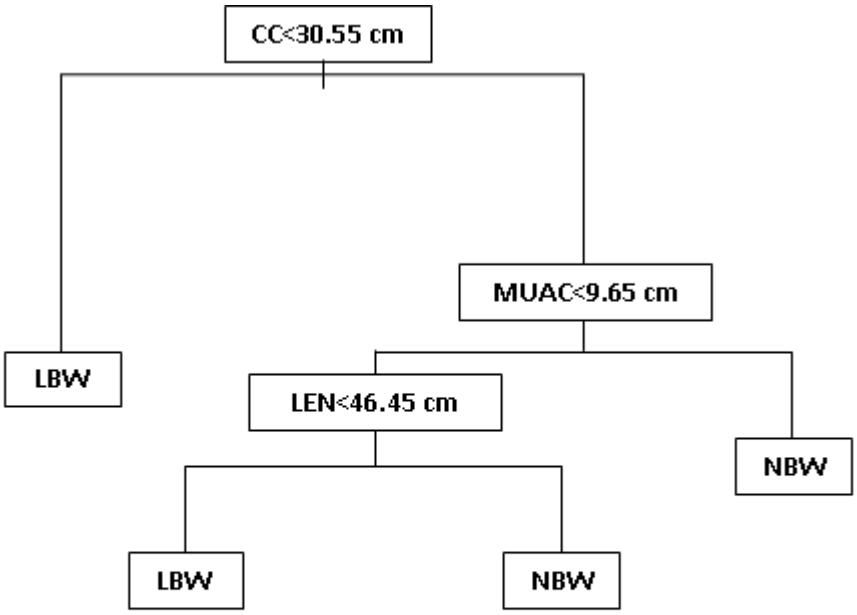


Abbreviations: HC, Head Circumference; CC, Chest Circumference; MUAC, Mid-Upper-Arm Circumference; LEN, Length; Thresh, Threshold; Spe, Specificity; Sen, Sensitivity

depicted to be the best predictor of LBW followed by MUAC, length and head circumference. The threshold for chest circumference to predict LBW was 30.55 cm with 83% sensitivity and 85% specificity. Threshold for MUAC was 9.45 cm with 83% sensitivity and 76% specificity.

DT model to predict LBW is presented in Figure 2. After pruning the DT there were three variables remained in the model which are chest circumference, MUAC and length. With these three variables DT model can predict LBW with 85.88% training and 85% test overall accuracy.

Figure 5-2: Decision tree model to predict low birth weight (<2500 gm)



Abbreviations: CC, Chest Circumference; MUAC, Mid-Upper-Arm Circumference; LEN, Length; LBW, low birth weight; NBW, normal birth weight

Predictive models of LBW are compared in Table 2. Only the chest circumference <30.55 had 84% accuracy for training data and 83% for test data. The DT used chest circumference, MUAC and length to predict LBW with 85.88% and 85.10% training and test accuracy. SVM provided the best accuracy (88.71 for training and 88.57 for test data) followed by ANN (88.42 for training and 88.24 for test data) and RF (87.91 for training and 87.32 for test data). Sensitivity was highest for DT model (89.53 for training and 90.30 for test data) followed by ANN and SVM. Specificity was highest for SVM followed by ANN and RF. SVM was also provided the highest positive

predictive value followed by ANN and RF and ANN provided highest negative predictive value followed by DT and SVM.

Table 5-1: Performance of different machine learning methods to predict low birth weight on both training (n=11763) and test (n= 3920) data set

	Accuracy	Sensitivity	Specificity	PP value	NP value
ROC: CC<30.55					
Training	84.06	82.93	85.45	87.53	80.26
Test	83.14	83.02	83.28	85.70	80.25
Decision tree					
Training	85.88	**89.53**	81.39	85.56	86.32
Test	85.10	**90.30**	78.83	83.74	87.06
Random Forest					
Training	87.91	87.99	87.82	89.90	85.58
Test	87.32	88.48	85.92	88.36	86.07
Neural Network					
Training	88.42	89.17	87.50	89.78	**86.77**
Test	88.24	89.74	86.43	88.87	**87.46**
Support Vector Machine					
Training	**88.71**	88.09	**89.47**	**91.15**	85.92
Test	**88.57**	88.95	**88.12**	**90.04**	86.85

CC: Chest Circumference, PP value: Positive Predictive value, NP value: Negative Predictive value

Discussion

We conducted this study to propose a predictive model of LBW with four other simpler-to-measure anthropometric measurements, length and head, chest, and arm circumferences in a typical rural setting of Bangladesh. We also investigated the use of 4 machine learning methods, DT, RF, ANN and SVM to predict low birth weight. We found SVM is the best predictive model of low birth weight. On the other hand,

DT is the simplest predictive model amongst the four which can be applied without the help of a computer with highest sensitivity and a lower overall accuracy. In a resource poor setting where computer is not available the DT model is the best option. DT model is more applicable in that study where the researcher or programmer wants to identify as many as LBW as possible keeping false positive or false negative rate <20%. The uniqueness of this study is, we used a very large community based data and used advanced machine learning techniques to predict low birth weight. This study provided the opportunity to the researchers or programmers to choose an optimum predictive model of LBW based on the resources they have without losing much overall accuracy.

The present study reported that more than 95% delivery took place at home and this is the common scenario of developing countries. So, birth weight cannot be recorded for all births in developing countries which incur many infant deaths due to low birth weight. Even if we have a sophisticated baby weighing scale, measuring accurate weight at the community level is certainly a very difficult task. Because, the accuracy depends on many things like the machine's precision, placement at ground during measurement, baby placement on the scale, regular calibration, proper recording and so on. On the other hand, baby weighing scale is different from that of the adults and it is very hard to carry from house to house or from village to village. To overcome these problems many studies were performed across the developing countries to provide surrogative measure of low birth weight which are simpler to measure

(Elizabeth et al. 2013; Bhargava et al. 1985; Kapoor et al. 2001; Mullany et al. 2007; Sreeramareddy et al. 2008; Taksande et al. 2007; Diamond et al. 1993). To measure all the 4 anthropometry other than weight only what needed is a Ross insertion tape which can be folded and carried in a pocket. All most all the studies provided a single anthropometric measure with a certain threshold value as the predictor of LBW. Elizabeth and colleagues (Elizabeth et al. 2013) reported that foot length <7.9 cm is the best predictor of LBW in Uganda. Two studies from Nepal suggested two different anthropometric measures as the best predictor of LBW; one suggested head circumference <33.5 cm (Sreeramareddy et al. 2008) and the other one suggested chest circumference <30.3 cm (Mullany et al. 2007). Two studies from India also reported inconsistent results; one showed chest circumference <32.5 cm (Kapoor et al. 2001) and the other one showed thigh circumference <14.5 cm (Taksande et al. 2007) were the best predictors of LBW. All the studies discussed above used ROC curve to find the optimal threshold value for a specific anthropometric measurement to predict LBW and they did not evaluate their performance on any future observations i.e. on the test data. The main limitation of those studies is that almost all of them used a hospital based data which may not represent the community. So this is the most realistic study so far to predict LBW to our knowledge as we used a real community based data. Majority of the previous studies also didn't show whether the combination of more than one anthropometric measurement can predict more preciously than a single anthropometric measurement.

In this study we also identified chest circumference <30.55 cm amongst the 4 is the most important predictor of the LBW using ROC curve. We also defined different predictive models simultaneously using 4 anthropometric measurements to increase the predictive ability of the models using four different machine learning methods. We also assessed their performance using both on the training and test data sets in terms of five indicators: accuracy, sensitivity, specificity, positive predictive value and negative predictive value. This study provided evidence that using multiple anthropometric measurements can improve the predictive accuracy of the model. Among the 4 machine learning methods SVM was the most efficient to predict LBW with highest training and test accuracy (>88.5%), training and test specificity and training and test positive predictive value. On the other hand, the DT model which used three anthropometric measurements chest circumference, MUAC and length had relatively lower training and test accuracy (85.9% vs 85.1%) but had highest training and test sensitivity (89.5% vs 90.3%). The DT model is easy to interpret, can easily be visualized and can apply it by doing hand calculation. In contrast, to apply SVM based prediction we must have to use a computer because it is much more computationally intensive. In some intervention studies or for referral, a decision has to be made instantly at the community where the computer is not available whether the baby is LBW or not, then there are two options; use DT model or use chest circumference only. So the model choice is more or less context specific. Thus, this study facilitates the researchers or programmers to choose a suitable model to predict LBW in the community based studies.

The machine learning techniques have been explored as a more powerful alternative to statistical methods for classification and predictions which use techniques such as unconventional optimization strategies, conditional probabilities or absolute conditionality (McCarthy et al. 2004). Machine learning methods can handle large, noise and complex datasets which is making it popular in various field of research. Nonetheless, it was initially a field in computer science, it is being used in public health (Pineda et al. 2013), bioinformatics (Larrañaga et al. 2006), medicine (Weston & Hood 2004), cancer diagnosis and detection (McCarthy et al. 2004; Wang et al. 2015), study of prevention and treatment policy (Cruz & Wishart 2006) and so on. As machine learning algorithms are completely data driven, they are often much more accurate over traditional methods. In all supervised machine learning methods, a model is initially learned from a training data set and then validate its parameter on validation set and then apply it on a test data set (can be treated as future observation) which is independent of training and validation set but comes from the same population to evaluate its performance. There are many machine learning methods have been developed as there is no unique model which performs best universally. We often have to search for the best one for any given problem. In this study we used four machine learning methods, DT, RF, ANN and SVM.

The main disadvantage of machine learning algorithm is that they are much more computationally intensive than the traditional methods; however, they are

conceptually simple. To apply the machine learning algorithms in a real world setting for classification we must have a computer. However, it was not feasible to apply machine learning algorithm in real world setting due to the paucity of computer at the community level in few years back. But, the primary health care framework of Bangladesh Government is now very much sophisticated (Director General 2014) which may allow us to use machine learning method to predict LBW at the community level. In Bangladesh, Community Clinic, the lowest-level static health facility, has been established at the ward level, sub unit of Union Council (the smallest rural administrative and local government unit of Bangladesh), to provide essentially primary health care services including maternal and child care to the community (Director General 2014). All Community Clinics received Internet connection through a laptop and wireless modem to help collection of local health-related data, provide telemedicine service, community health education, and certain other ICT-based health solutions. Then the service provider can collect anthropometric data other than the weight with a modest effort from the new born baby at the community and with the help of the computer at the clinic they can isolate the LBW baby and provide essential services or referral within a very short time. So use of machine learning algorithm at the community level with a greater accuracy is very much feasible in countries like Bangladesh. Despite the fact that Bangladesh has already made remarkable progress in reducing child mortality, but it has a little improvement on neonatal mortality (Sayem et al. 2011). One of the main causes of neonatal mortality is LBW (Sayem et al. 2011). This study will help the Bangladesh

Government in policymaking to take initiative in reducing neonatal mortality due to LBW.

In conclusion, this study will provide a realistic predictive model of LBW in Rural Bangladesh. The LBW can be identified with measuring 4 simpler-to-measure anthropometries, length and head, chest and arm circumferences without measuring their weights, at a greater accuracy using the SVM-based model at the community level.

Reference

Achebe, C., E. F. Ugochukwu, P. O. U. Adogu & C. Ubajaka, 2013. Prediction of low birth weight from other anthropometric parameters in Nnewi, south eastern Nigeria. *Nigerian Journal of Paediatrics,* 41(1), 59-63.

Bauer, E. & R. Kohavi, 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning,* 36(1-2), 105-39.

Bhargava, S. K., S. Ramji, A. Kumar, M. A. N. Mohan, J. Marwah & H. P. Sachdev, 1985. Mid-arm and chest circumferences at birth as predictors of low birth weight and neonatal mortality in the community. *BMJ,* 291(6509), 1617-9.

Breiman, L., 2001. Random forests. *Machine learning,* 45(1), 5-32.

Breiman, L., J. Friedman, C. J. Stone & R. A. Olshen, 1984. *Classification and regression trees*: CRC press.

Byvatov, E., U. Fechner, J. Sadowski & G. Schneider, 2003. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences,* 43(6), 1882-9.

Christian, P., R. Klemm, A. A. Shamim, H. Ali, M. Rashid, S. Shaikh, L. Wu, S. Mehra, A. Labrique & J. Katz, 2013. Effects of vitamin A and β-carotene supplementation on birth size and length of gestation in rural Bangladesh: a cluster-randomized trial. *The American journal of clinical nutrition,* 97(1), 188-94.

Cruz, J. A. & D. S. Wishart, 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics,* 2, 59.

Diamond, I., J. McDonald & R. Guidotti, 1993. Use of a simple anthropometric measurement to predict birth weight. WHO Collaborative Study of Birth Weight Surrogates. *Bulletin of the World Health Organization,* 71(2), 157-63.

Director General, H. S., (2014). Management Information System, in *Health Bulletin 2014*Government of the People's Republic of Bangladesh.

Elizabeth, N. L., O. G. Christopher & K. Patrick, 2013. Determining an anthropometric surrogate measure for identifying low birth weight babies in Uganda: a hospital-based cross sectional study. *BMC pediatrics,* 13(1), 54.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-74.

Friedman, J., T. Hastie & R. Tibshirani, 2008. *The elements of statistical learning: Data Mining, Inference and Prediction*: Springer series in statistics Springer, Berlin.

Gunnsteinsson, S., A. B. Labrique, K. P. West Jr, P. Christian, S. Mehra, A. A. Shamim, M. Rashid, J. Katz & R. D. W. Klemm, 2010. Constructing indices of rural living standards in Northwestern Bangladesh. *Journal of health, population, and nutrition,* 28(5), 509.

Hack, M., D. J. Flannery, M. Schluchter, L. Cartar, E. Borawski & N. Klein, 2002. Outcomes in young adulthood for very-low-birth-weight infants. *New England Journal of Medicine,* 346(3), 149-57.

Hediger, M. L., M. D. Overpeck, W. Ruan & J. F. Troendle, 2002. Birthweight and gestational age effects on motor and social development. *Paediatric and perinatal epidemiology,* 16(1), 33-46.

Kapoor, S. K., G. Kumar, C. S. Pandav & K. Anand, 2001. Performance of surrogate markers of low birth weight at community level in rural India. *Journal of epidemiology and community health,* 55(5), 366-7.

Klemm, R. D. W., A. B. Labrique, P. Christian, M. Rashid, A. A. Shamim, J. Katz, A. Sommer & K. P. West, 2008. Newborn vitamin A supplementation reduced infant mortality in rural Bangladesh. *Pediatrics,* 122(1), e242-e50.

Klemm, R. D. W., R. D. Merrill, L. Wu, A. A. Shamim, H. Ali, A. Labrique, P. Christian & K. P. West, 2013. Low-birthweight rates higher among Bangladeshi neonates measured during active birth surveillance compared to national survey data. *Maternal & child nutrition,* 2013.

Labrique, A. B., P. Christian, R. D. W. Klemm, M. Rashid, A. A. Shamim, A. Massie, K. Schulze, A. Hackman & K. P. West, 2011. A cluster-randomized, placebo-controlled, maternal vitamin A or beta-carotene supplementation trial in Bangladesh: design and methods. *Trials,* 12(1), 102.

Lakshmi, T. M., A. Martin, R. M. Begum & V. P. Venkatesan, 2013. An analysis on performance of decision tree algorithms using student's qualitative data. *International Journal of Modern Education and Computer Science (IJMECS),* 5(5), 18.

Larrañaga, P., B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé & A. Pérez, 2006. Machine learning in bioinformatics. *Briefings in bioinformatics,* 7(1), 86-112.

Mathews, T. J. & M. F. MacDorman, 2012. Infant mortality statistics from the 2008 period linked birth/infant death data set. *National vital statistics reports,* 60(5).

McCarthy, J. F., K. A. Marx, P. E. Hoffman, A. G. Gee, P. O'Neil, M. L. Ujwal & J. Hotchkiss, 2004. Applications of Machine Learning and High-Dimensional Visualization in Cancer

Detection, Diagnosis, and Management. *Annals of the New York Academy of Sciences,* 1020(1), 239-62.

Mohri, M., A. Rostamizadeh & A. Talwalkar, 2012. *Foundations of machine learning*: MIT press.

Mullany, L. C., G. L. Darmstadt, S. K. Khatry, S. C. Leclerq & J. M. Tielsch, 2007. Relationship between the surrogate anthropometric measures, foot length and chest circumference and birth weight among newborns of Sarlahi, Nepal. *European journal of clinical nutrition,* 61(1), 40-6.

Müller, K.-R., G. Rätsch, S. Sonnenburg, S. Mika, M. Grimm & N. Heinrich, 2005. Classifying'drug-likeness' with kernel-based learning methods. *Journal of chemical information and modeling,* 45(2), 249-53.

Pineda, A. L., F.-C. Tsui, S. Visweswaran & G. F. Cooper, 2013. Detection of Patients with Influenza Syndrome Using Machine-Learning Models Learned from Emergency Department Reports. *Online journal of public health informatics,* 5(1).

Quinlan, J. R., 1986. Induction of decision trees. *Machine learning,* 1(1), 81-106.

Quinlan, R. C., 1993. 4.5: Programs for machine learning Morgan Kaufmann Publishers Inc. *San Francisco, USA*.

Reichman, N. E., 2005. Low birth weight and school readiness. *The Future of Children,* 15(1), 91-116.
Rokach, L. & O. Maimon, 2008. *Data mining with decision trees: theory and applications*: World Scientific Publication Co Inc.

Safavian, S. R. & D. Landgrebe, 1990. A survey of decision tree classifier methodology.

Salam, A., F. Haseen, H. K. M. Yusuf & H. Torlesse, 2004. National low birth-weight survey of Bangladesh, 2003-2004. *Prevention,* 76, 155.

Sayem, A. M., A. T. M. S. Nury & M. D. Hossain, 2011. Achieving the millennium development goal for under-five mortality in Bangladesh: current status and lessons for issues and challenges for further improvements. *Journal of health, population, and nutrition,* 29(2), 92.

Sreeramareddy, C. T., N. Chuni, R. Patil, D. Singh & B. Shakya, 2008. Anthropometric surrogates to identify low birth weight Nepalese newborns: a hospital-based study. *BMC pediatrics,* 8(1), 16.

Taksande, A., K. Y. Vilhekar, P. Chaturvedi, S. Gupta & P. Deshmukh, 2007. Predictor of low birth weight babies by anthropometry. *Journal of tropical pediatrics,* 53(6), 420-3.

Vapnik, V., 2013. *The nature of statistical learning theory*: Springer Science & Business Media.

Wang, G., K.-M. Lam, Z. Deng & K.-S. Choi, 2015. Prediction of mortality after radical cystectomy for bladder Cancer by machine learning Techniques. *Computers in Biology and Medicine*.

Wardlaw, T. M., 2004. *Low Birthweight: Country, regional and global estimates*: UNICEF.

West, K. P., P. Christian, A. B. Labrique, M. Rashid, A. A. Shamim, R. D. W. Klemm, A. B. Massie, S. Mehra, K. J. Schulze & H. Ali, 2011. Effects of vitamin A or beta carotene supplementation on pregnancy-related mortality and infant mortality in rural Bangladesh: a cluster randomized trial. *Jama,* 305(19), 1986-95.

Weston, A. D. & L. Hood, 2004. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of proteome research,* 3(2), 179-96.

World Health Organization (WHO), 2011. Optimal feeding of low birth weight infants in low and middle income countries. *Geneva: WHO*, 16-45.

Chapter 6: Conclusions and Further Scope

Conclusions

In this research we tried to explore the potential of the multivariate methods, CCA and PLS regression in studying the relationship between two sets of variables in public health research. Because, in public health research, investigators often deal with multiple exposure and multiple outcome measures in a single study. They are usually used to apply univariate statistical methods like multiple regression instead of multivariate methods without knowing the consequences of multiple testing. This research will help the public health researchers in choosing the appropriate statistical methods if they want to study multiple outcomes and multiple exposures simultaneously. Additionally, we intended to identify the alternative measure of low birth weight using ROC curve and 4 machine learning algorithms, decision tree, random forest, support vector machine and neural network. This will help researchers or programmers to identify low birth weight infant without measuring their weight with an acceptable accuracy.

CCA and PLS regression analysis are well known techniques for studying the relationship between multiple exposures and multiple outcomes simultaneously. The

basic differences between CCA and PLS regression is that CCA maximizes the correlation while PLS maximizes the covariance. Both the canonical correlation analysis and the PLS regression analysis has several advantages over the univariate methods. However, CCA is just an exploratory method very similar to Pearson's correlation. Although one variable set is often considered as predictor and the other as criterion, it does not imply causal relationship between the set of exposures and the set of outcomes. On the other hand, PLS regression can help in establishing causal relationship between exposures and the outcomes. So, the choice of CCA or PLS regression depends on the objective of the study if we want to study simultaneously multiple exposures and multiple outcomes. In addition, this study will provide a realistic predictive model of LBW for Rural Bangladesh. The LBW can be predicted with measuring 4 simpler-to-measure anthropometries, length and head, chest and arm circumferences without measuring their weights, at a greater accuracy with the advent of the information technology.

Further Scope

The CCA can be used as method of variable selection for PLS regression. So there is a further scope of testing and validating CCA as a method of variable selection for PLS regression. Both the methods are vulnerable to the outlying observations. There are some robust methods of CCA and PLS regression available in literature which

should have been considered. So, there is also some scope of comparing the available robust methods of CCA and PLS regression and also there is always some scope of improving the robust methods as well as developing new robust methods of CCA and PLS regression. Additionally, there are some other multivariate methods available in literature which were not possible to study in this research due to time constraint. For instance, structural equation modeling is now a promising method for studying multiple outcome and multiple exposures simultaneously. We will have further scope to explore all other multivariate methods to study multiple exposure and multiple outcome variables simultaneously.

In this research, we used only 4 machine learning methods for predicting low birth weight; however, there are many more machine learning methods available in literature. For example, “deep learning” is doing very well as a supervised leaning method. So there is also some scope to improve the accuracy of the model to predict low birth weight using some other machine learning methods.