RESEARCH ARTICLE

Identifying maternal and infant factors associated with newborn size in rural Bangladesh by partial least squares (PLS) regression analysis

Alamgir Kabir[1,2]*, **Md. Jahanur Rahman**[1], **Abu Ahmed Shamim**[2,3], **Rolf D. W. Klemm**[4,5], **Alain B. Labrique**[4], **Mahbubur Rashid**[2], **Parul Christian**[4,6], **Keith P. West, Jr.**[4]

1 Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh, **2** The JiVitA Maternal and Child Health and Nutrition Research Project of Johns Hopkins University, Gaibandha, Bangladesh, **3** Helen Keller International, Dhaka, Bangladesh, **4** Center for Human Nutrition, Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America, **5** Helen Keller International, New York, New York, United States of America, **6** The Bill & Melinda Gates Foundation, Seattle, Washington, United States of America

* aym.alamgir@gmail.com

Abstract

Birth weight, length and circumferences of the head, chest and arm are key measures of newborn size and health in developing countries. We assessed maternal socio-demographic factors associated with multiple measures of newborn size in a large rural population in Bangladesh using partial least squares (PLS) regression method. PLS regression, combining features from principal component analysis and multiple linear regression, is a multivariate technique with an ability to handle multicollinearity while simultaneously handling multiple dependent variables. We analyzed maternal and infant data from singletons (n = 14,506) born during a double-masked, cluster-randomized, placebo-controlled maternal vitamin A or β-carotene supplementation trial in rural northwest Bangladesh. PLS regression results identified numerous maternal factors (parity, age, early pregnancy MUAC, living standard index, years of education, number of antenatal care visits, preterm delivery and infant sex) significantly ($p<0.001$) associated with newborn size. Among them, preterm delivery had the largest negative influence on newborn size (Standardized β = -0.29 − -0.19; $p<0.001$). Scatter plots of the scores of first two PLS components also revealed an interaction between newborn sex and preterm delivery on birth size. PLS regression was found to be more parsimonious than both ordinary least squares regression and principal component regression. It also provided more stable estimates than the ordinary least squares regression and provided the effect measure of the covariates with greater accuracy as it accounts for the correlation among the covariates and outcomes. Therefore, PLS regression is recommended when either there are multiple outcome measurements in the same study, or the covariates are correlated, or both situations exist in a dataset.

OPEN ACCESS

Citation: Kabir A, Rahman M.J, Shamim AA, Klemm RDW, Labrique AB, Rashid M, et al. (2017) Identifying maternal and infant factors associated with newborn size in rural Bangladesh by partial least squares (PLS) regression analysis. PLoS ONE 12(12): e0189677. https://doi.org/10.1371/journal.pone.0189677

Editor: Mahfuzar Rahman, BRAC, BANGLADESH

Received: February 18, 2017

Accepted: November 30, 2017

Published: December 20, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Supported by a grant from the Bill and Melinda Gates Foundation, Seattle, WA Global Control of Micronutrient Deficiency (Grant GH614); Micronutrients for Health Cooperative Agreement (HRN-A-00-97-00015); Global Research Activity Cooperative Agreement (GHS-A-00-03-00019-00) between Johns Hopkins University and the Office of Health, Infectious Diseases and Nutrition; and the US Agency for International Development, Washington, DC. Additional director in kind support was provided by the Sight and Life Research Institute (Baltimore, MD); Nutrilite Health Institute (Nutrilite Division, Access Business Group, LLC, Buena Park, CA); the Canadian International Development Agency (CIDA), Ottawa, Canada; and the National Integrated Population and Health Program of the Ministry of Health and Family Welfare of the Government of the People's Republic of Bangladesh. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

In developing countries, small size at birth is common, reflecting combined effects of inadequate intrauterine growth and preterm birth. Both birth conditions are associated with increased risks of poor postnatal growth [1] and infant mortality [2], and diverse maternal nutritional, health and socioeconomic factors [3–5]. Although weight is most commonly measured at birth, other measurements, including length and circumferences of the head, chest and arm provide additional information on newborn size and proportionality, as well as confirmatory and novel insights about fetal growth [2]. While maternal health and socioeconomic status have been shown to be associated with risk of small birth size in low resource settings in South Asia [3,5], there remains a need to adapt analytical techniques to more efficiently and simultaneously explore risk factors of birth size to identify pregnant women who may benefit from antenatal, obstetric and postnatal care. In this paper, we illustrate the use of partial least squares (PLS) regression analysis to quantify associations between materno-infant and household socioeconomic characteristics and multiple measures of newborn size, drawing on a large population cohort of infants whose mothers participated in a maternal vitamin A and beta-carotene supplementation in rural, northern Bangladesh [6].

In exploring associations between health outcomes and multiple risk factors, observational epidemiologic studies often deal with data that include both a set of exposure variables and a set of outcome variables. Frequently, the variables are interrelated to some extent and consequently, multicollinearity often exists. Routine statistical approaches such as multiple linear regression or principal component regression (PCR) 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 [7]. Additionally, the multiple linear or PCR analysis is not able to account for the correlation structure among the dependent variables. A multivariate regression approach called partial least squares (PLS) regression developed by Harman Wold in the 1960's [8,9], which can simultaneously consider more than one dependent variables, can address the correlation structure among the dependent variables, and also efficiently handle multicollineary. Thus, we propose through this study that PLS regression has the potential to be a useful method to predict multiple facets of a health outcome—in this instance, infant size at birth from maternal, infant and other household factors in a rural South Asian population [6]. We also compare the performance of PLS regression with PCR and the individual predictive ability of each these two methods.

Materials and methods

Ethics statement

The overall Jivita study protocol was reviewed and approved by both the Bangladesh Medical Research Council (BMRC) and the Institutional Review Board (IRB) of Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. Documented consent was given by all participating women.

In this study, PLS regression was used to assess the effect of maternal and infant factors on newborn size at birth. The data reported in this analysis were collected during a field based study assessing the efficacy of weekly antenatal vitamin A or β-carotene versus placebo supplementation on maternal and infant mortality through 6 months of age from July 2001 to January 2007. Details of this study conducted in a contiguous ~435 sq km area in rural northwestern Gaibandha and Rangpur Districts of Bangladesh with a population of ~650,000 are available elsewhere [6,10].

On enrollment into the intervention trial, consenting mothers were interviewed in the 1st trimester about household socioeconomic condition, education, demographic characteristics, previous pregnancy history, recent morbidity history and 7-day food intake and mid-upper arm circumference. A Living Standard Index (LSI) was constructed using principal component analysis from household socio-economic variables and was used to describe socio-economic status [11]. At 3 months postpartum another interview was completed to collect dietary, morbidity, antenatal care (ANC) received, and events and care received during labor and delivery, among other postpartum factors. Birth size measurements were added as part of a newborn, placebo-controlled, single-dose vitamin A supplementation trial that was initiated among infants born during the 2nd half of the maternal trial [12]. Shortly after birth, newborns were dosed with allocated supplements and, shortly thereafter [median age (IQR): 18 (9, 36) hours of age], one of 56 standardized, female anthropometrists measured newborn weight, length, and mid-upper arm, 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 following standard methods [13]. Small infant size was defined by a birth weight <2.5 kg, MUAC <10 cm, head circumference <33 cm or chest circumference < 30.5 cm [14]. Singleton live born infants measured within 72 hours of birth (n = 16,290, 75% of 21,585 singleton live born infants) were eligible, out of which 14,506 (89% of 16,290) had complete data and were included in this analysis.

The maternal characteristics included in this study were age at enrollment, parity, mid upper arm circumference (cm), education (yrs), living standard index, number of antenatal care visits, and maternal supplementation group. Additional infant characteristics included preterm (<37 week of gestation) delivery and infant sex (female/male).

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 [15]. To overcome these problems, the researchers usually take some measures; they may remove some variables [16] or may use multivariate reduction techniques like principal component analysis to reduce the multidimensionality in the predictor or response variables [17]. 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 reduce 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 covariation among the response variables. Consequently, this may produce patterns or syndromes within the predictor variables making little or no biological sense [15]. The appropriate solution of these challenges is using PLS regression [15].

Although PLS regression is comparatively new, its use in research is gradually increasing. The great strength of PLS regression is parsimony [18]. Initially, used in analytic chemistry [19–21], PLS now it is gaining popularity in public health [22–24], bioinformatics [25], ecology [15,26] and agriculture [27]. As it is computationally much more intensive, the advent of statistical packages such as, R, SAS, STATA, MatLab and STATISTICA also facilitates its wider application.

Similar to principal component regression (PCR), PLS regression analysis is a data-dimension reduction method that extracts a set of orthogonal factors called latent variables which are used as predictors in the regression model [9]. The major difference with 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 [28].

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,

$$\boldsymbol{X} = \boldsymbol{TP}^{T} \ \ \boldsymbol{and} \ \ \widehat{\boldsymbol{Y}} = \boldsymbol{TBC}^{T} \tag{1}$$

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

$$\widehat{\boldsymbol{Y}} = \boldsymbol{TBC}^{T} = \boldsymbol{XB}_{\boldsymbol{PLS},} \boldsymbol{where} \ \ \boldsymbol{B}_{\boldsymbol{PLS}} = \boldsymbol{P}^{T+}\boldsymbol{BC}^{T} \tag{2}$$

$\mathbf{P}^{\mathbf{T+}}$ is the Moore-Penrose pseudo-inverse of P^{T}. The matrix B_{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 [9]. The algorithm for PLS regression is as follows:

Step 1: Transform X and Y into Z-scores 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

$$\boldsymbol{P}_1 {=} \boldsymbol{X}_0^{T}\boldsymbol{T}_1 \ \ \boldsymbol{and} \ \ \widehat{\boldsymbol{X}}_1 {=} \boldsymbol{T}_1^{T}\boldsymbol{P}_1.$$

Step 7: Compute $U_1 = Y_0c_1$ *and*

$$\widehat{\boldsymbol{Y}}_1 = \boldsymbol{U}_1\boldsymbol{c}_1^{T} = \boldsymbol{T}_1\boldsymbol{b}_1\boldsymbol{c}_1^{T}, \ \ \boldsymbol{where}, \ \ \boldsymbol{b}_1 = \boldsymbol{T}_1^{T}\boldsymbol{U}_1 \tag{3}$$

The scalar b_1 is the slope of the regression of $\hat{Y}$ on T_1. Eq (3) shows that $\hat{Y}$ is obtained as linear regression from the latent variable extracted from X_0. Matrices $\widehat{X}_1$ and $\hat{Y}_1$ 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,

$$\boldsymbol{X}_1 = \boldsymbol{X}_0 - \widehat{\boldsymbol{X}}_1 \boldsymbol{and}\ \ \boldsymbol{Y}_1 = \boldsymbol{Y}_0 - \widehat{\boldsymbol{X}}_1$$

Step 9: The first set of latent variables has now been extracted. Now perform SVD on $R_2 = X_1^T Y_1$ we get w_2, c_2, T_2 and b_2 and the new deflated matrices X_2 and 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 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, or 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. The component at which the cross validation RMSEP has a meaningful change is used in the final model. To choose the optimum number of components for both PLS and principal component regression, root mean squared error of prediction (RMSEP) were calculated using different number of components. We performed approximate t-tests of regression coefficients based on jackknife variance estimates [29].

We constructed a 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 to the perimeter of the circle, the better it is represented, and if two variables are highly correlated they 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, t_1 vs 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 principal component regression, we calculated Pearson's correlation coefficients between the predicated values (by PLS and principal component regression with 1 to 5 components respectively) and the observed values of infant's size variables. This correlation coefficient indicate the predicative power of the model: if the model has perfect predictive ability then the correlation coefficient will be 1. So, the more the correlation coefficient, the higher the predictive power of a given model is. For this analysis, we used the R packages: "plsdepot", "pls" and "mixOmics".

Results

More than half of the infants were small at birth. More than 50% mother had parity of at least 1, mean (SD) maternal age was 21.96 (5.88) yrs and maternal early pregnancy MUAC was 22.99 (1.97) cm. About 75% mothers reported not having had any ANC visits throughout the pregnancy. Preterm delivery was 27% and 51% infants were male (Table 1).

Fig 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 were significantly ($p<0.001$) highly correlated with each other. All the predictors except vitamin A and β-carotene supplementation were 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

Table 1. Descriptive statistics for birth size and maternal socio-demographic factors from rural North West Bangladesh in 2002–2007, n = 14506.

Variables	Mean (SD)/ n (%)	Median (IQR)
Birth size, within 72 hours of birth		
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 and infant 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†	4897 (33.76)	-
β-carotene supplementation†	4803 (33.11)	-
Preterm delivery†	3904 (26.91)	-
Male infant†	7384 (50.90)	-

† Dichotomous variables are presented as n(%)

https://doi.org/10.1371/journal.pone.0189677.t001

correlation compared to other variables (r≥0.10, p<0.001). Predictors were 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).

The RMSEP was plotted against number of components used in the PLS regression model and principal component regression models in Figs 2 and 3 respectively. These figures suggested that 2 components would be included in the PLS regression model while 5 components would be included in PCR.

Table 2 presents the standardized PLS regression coefficient with 2 components. Except the prenatal supplementation of vitamin A and β-carotene, all the variables had significant (p <0.001) association with 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 sex and maternal parity, MUAC, and age.

Some of the coefficients estimated using the ordinary linear regression (LR) were different compared to the coefficients estimated using PLS regression (S1 Table). Coefficients of maternal age estimated using PLS regression were continually higher than the coefficients estimated using LR. Conversely, coefficients of parity as well as maternal education estimated using PLS regression were lower than their respective coefficients estimated using LR. However, in case of LSI, the coefficients were arbitrary (and no observable pattern was present between values estimated using either of the regression models).

The correlation plot of the variables for the first two components (Fig 4) depicted that maternal education, ANC visit, LSI, age, MUAC and parity were correlated and fell in the 4th quadrant. Among them, age, MUAC and parity were close to each other and they were also closer to the birth size variables in the 1st quadrant. Education, ANC visit, and LSI were very close to each other but were further from to the birth size variables compared to the prior

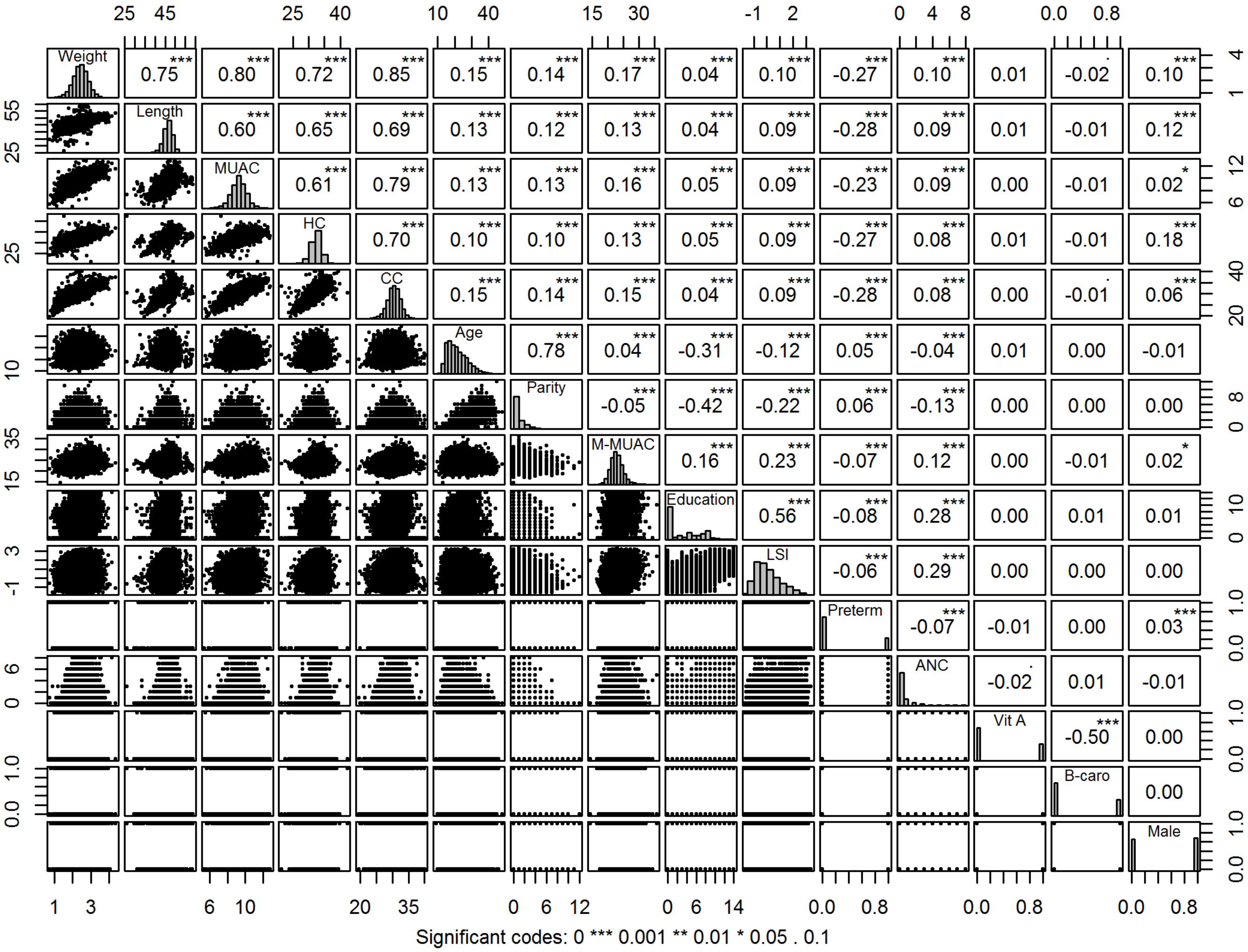

Fig 1. Pair wise correlation and scatter plot matrix of the study variables. (MUAC, Mid-Upper Arm Circumference; HC, Head circumference; CC, Chest Circumference; M-MUAC, Maternal MUAC; LSI, Living Standard Index; ANC, Antenatal care visit; Vit A, Maternal vitamin A supplementation; B-care, Maternal β-carotene supplementation; Male, Infant's sex as male).

https://doi.org/10.1371/journal.pone.0189677.g001

cluster of maternal variables. On the other hand, preterm delivery alone fell in the 3rd quadrant just opposite to the birth size variables and infant's sex alone fell in the 1st quadrant with the birth size variables. Maternal Vitamin A and β-carotene supplementation fell very close to the center indicating no effect on birth size. All the 5 birth size variables in the 1st quadrant were clustered close to each other. Therefore, this figure demonstrated that preterm delivery was the most important predictor in the opposite direction, followed by infant's gender, parity, age and MUAC.

The score plot of the 1st and 2nd PLS components of the predictors displayed four distinct groupings among the study participants due to the interaction effect of preterm delivery and infant's sex (male:term, male:preterm, female:term and female:preterm) (Fig 5).

We observed that the magnitude of correlation between the predicted values using PCR and observed values of birth size was substantially increasing with increasing the number of components, however, the correlation was very poor with the first few components (Table 3). On the other hand, the magnitude of correlation between predicted values using PLS

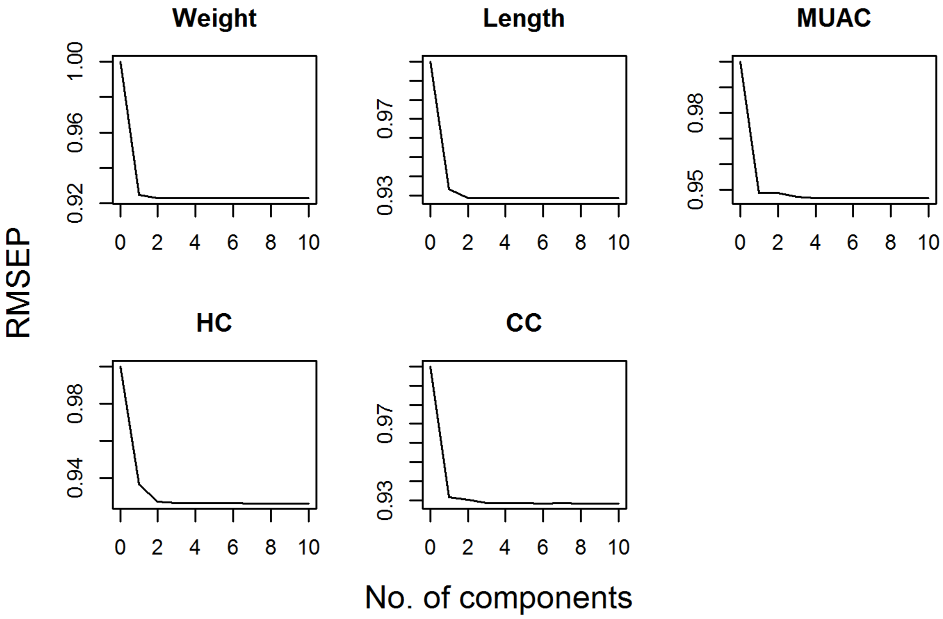

Fig 2. Root mean squared error of prediction (RMSEP) for different number of components of partial least squares (PLS) repression.

https://doi.org/10.1371/journal.pone.0189677.g002

regression and the observed values of birth size did not meaningfully change 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 PCR. Notably, the first 2 PLS components that were derived from maternal factors explained only ~26% of total variation of maternal factors had a predictive ability that was comparable to the PCR with 5 components that explained ~73% of total variation.

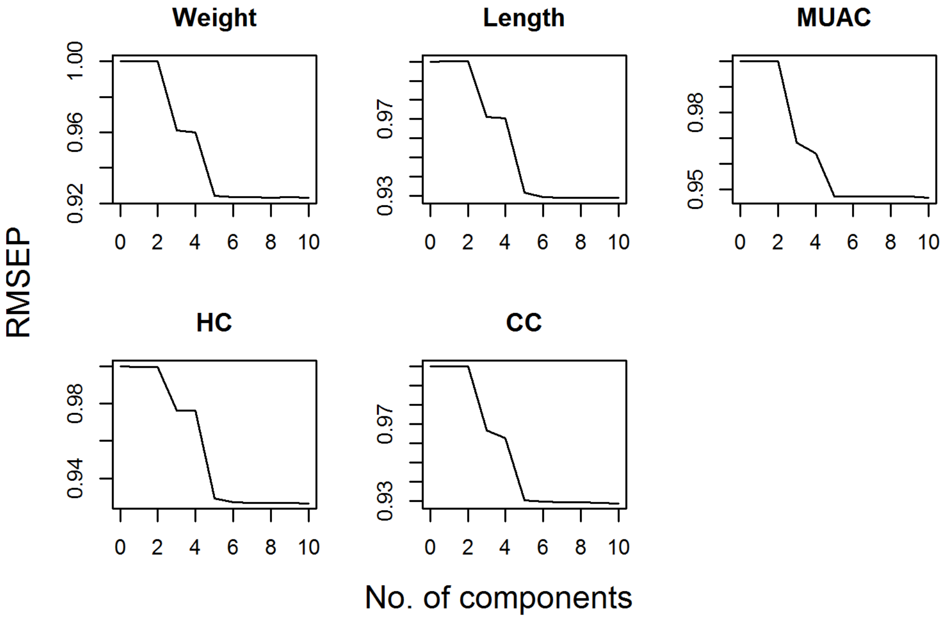

Fig 3. Root mean squared error of prediction (RMSEP) for different number of components of principal component regression (PCR).

https://doi.org/10.1371/journal.pone.0189677.g003

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

Maternal and infant 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 sup	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 sup	-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: SE, Standard Error; MUAC, Mid-Upper Arm Circumference; LSI, Living Standard Index; Mom MUAC, Maternal Mid-Upper Arm Circumference.

https://doi.org/10.1371/journal.pone.0189677.t002

Discussion

We conducted this study to assess the effect of maternal socio-demographic factors on newborn size at birth using PLS regression. Our study revealed that all the maternal variables examined, except vitamin A and β- carotene supplement receipt during pregnancy, were significantly associated with birth size. Preterm delivery had the greatest effect on birth size followed by infant sex, maternal parity, education, and age. In addition, PLS regression facilitated the finding of an interaction effect of preterm delivery and newborn sex on birth size, as revealed by a scatter plot of the first two components (Fig 5). PLS regression was more parsimonious than PCR and the ordinary least squares regression, as it required only two components compared to the five required in PCR regression, while ordinary least squares would have required the inclusion of all covariates. The PLS regression provided effect measures of the covariates with more stability and greater accuracy compared to the ordinary least squares regression.

In addition to birth weight, the combination of all birth size measurements captures more information than a single measurement in isolation; however, they are highly correlated. Head circumference, for instance, indicates the brain volume [30] and it may also provide important diagnostic and prognostic information like neurocognitive function [31], 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. Hence, identification of maternal factors associated with newborn size using PLS regression has an important implication because PLS regression can simultaneously deal with multiple outcomes and collinearity among the covariates. A simulation study reported that in multiple regression settings, the correlation between covariates exceeding 0.35 has a greater impact on both the coefficients and their standard errors [32]. In case of collinearity, the effect of one variable may be confounded by other which cannot be fully extracted by general linear model [33]. Therefore, we need to be more conservative to account for collinearity in investigating the effect of individual risk factors on outcome. Despite the unbiasedness of the ordinary least squares regression coefficients, their standard error (SE) becomes larger

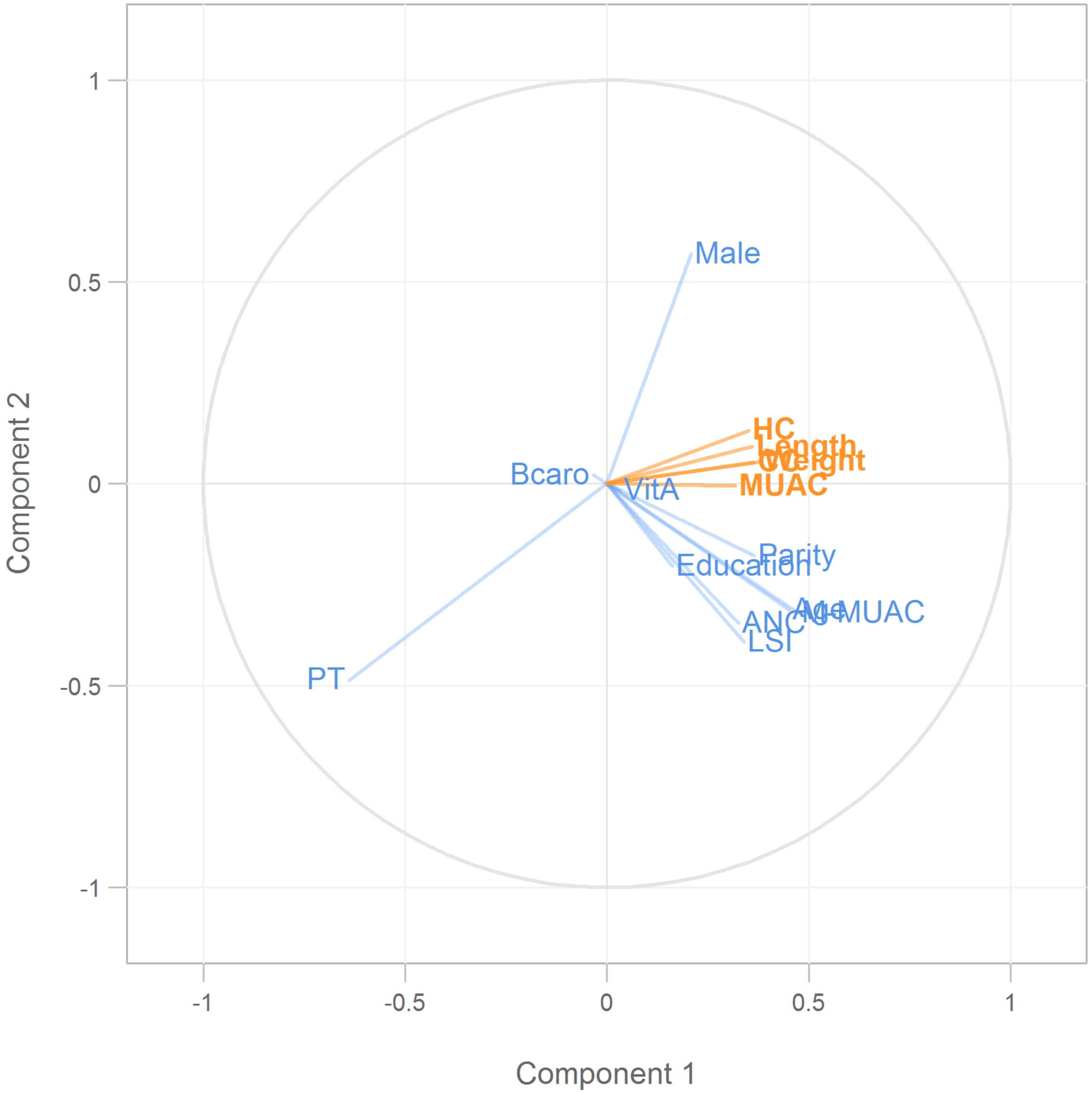

Fig 4. Correlation plot for the first two components. Orange color indicating the weights of dependent variables (birth size variables: weight, length, head (HC), chest (CC) and arm (MUAC) circumferences) and the blue color indicating the weights of the predictors (maternal and infant factors: maternal age, parity, education, MUAC, vitamin A sup. (VitA) and β-carotene sup. (Bcaro), living standard index (LSI), antenatal care visit (ANC), preterm delivery (PT), and infant's sex (Male)).

https://doi.org/10.1371/journal.pone.0189677.g004

with the degree of collinearity among covariates. This analysis also provided the evidence of having larger SE of the coefficients if the ordinary least squares regression was applied instead of PLS regression. However, PLS regression provides biased estimates, it provides a trade-off between bias and precision [34–36]. Moreover, the traditional ordinary least squares regression analysis with the adjustment of multiple confounders inadequately allocates the effect of covariates on outcome in presence of collinearity [33]. Therefore, PLS regression distributes the overall contribution of each of the maternal factors on birth size according to the correlations among covariates and outcomes. In this analysis it is obvious that the effect of the maternal variables with a higher level of collinearity on birth size obtained from PLS regression was substantially different from that of the ordinary least squares regression.

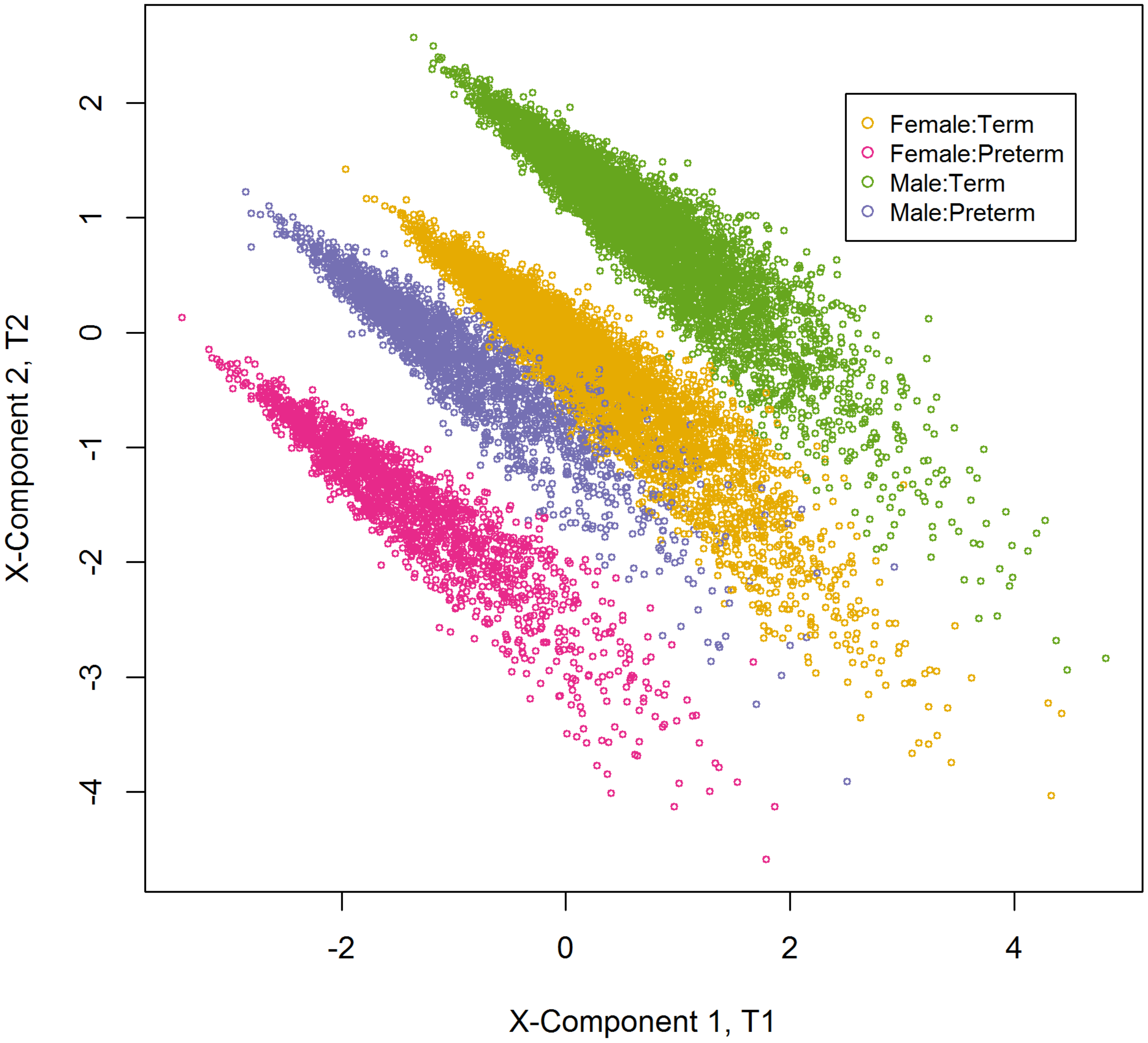

Fig 5. Scatter plot of the scores of first two PLS components of the predictors. Demonstrating 4 groups among the individuals participated in the study due the interaction effect of infant's sex and preterm delivery.

https://doi.org/10.1371/journal.pone.0189677.g005

The PLS regression analysis suggested that maternal and infant factors age, parity, early pregnancy MUAC, LSI, maternal education, number of ANC visits and infant sex were significantly and positively associated with infant's size at birth; however, preterm delivery had the greatest negative effect on birth size as anticipated. 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. [37] and Kabir et al. [38] from the same population. This study also identified that preterm delivery and newborn sex had a significant interaction on size at birth which was also commensurate with the findings from a previous study where canonical correlation analysis was applied to the same dataset [38] and which was also consistent with other study [39]. The correlation plot of first two components of the maternal factors and infant's size at birth indicated that maternal education, the number of ANC visits and the living standards index are correlated and maternal age and MUAC are also correlated which

Table 3. Coefficients of Pearson's correlation between the predicted values of birth size (by PLS and principal component regression) and observed values and the variance of the maternal and infant characteristics explained by the components of PLS and principal component regression analysis.

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
Correlation coefficients										
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)[2]	12.59%	23.64%	37.89%	57.08%	62.84%	23.79%	38.82%	52.92%	63.14%	72.82%

[2] Vaciance explained by the components of maternal characteristics which were generated to model birth size and they were presented as cumulative from comp-1 to comp-5.

https://doi.org/10.1371/journal.pone.0189677.t003

also provided evidence of collinearity existing among the maternal factors. Thus, we think the PLS regression was the appropriate method to use in this study.

PLS regression has the ability of handling multicollinearity and multiple dependent variables simultaneously. It has a proven ability to perform better than other regression methods, which are usually used to handle highly collinear data, such as stepwise multiple regression, PCR or model fitting techniques that apply Maximum Likelihood or Bayesian theory and in addition it equally performs in ideal situations [15,33,40]. 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. Typically in regression analysis, we look for a parsimonious model, i.e. the model that can predict the response variable at the desired level with as few predictors as possible. Because, however, R^2 increases with an increasing number of predictors in the model, the variance of regression coefficients also increases. Thus, it is statistically advantageous to have a reduced number of explanatory variables in a given model. Compared to PCR, PLS regression requires a smaller number of components resulting in more stable estimates of regression coefficients. In this analysis, PLS regression required only two components but PCR required five. Despite the fact that PLS components explained a much lower variance (~24%) of the maternal and infant factors, it had more predictive power over the PCR where components explained ~73% variance. 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 PCR, a component is chosen from the independent variables which capture most of the variability without accounting for how close it is to the dependent variables. 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 encourage the need to include more variables than the PLS method [28]. 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.

Conclusions

PLS regression was a better choice to analyze this data. It needed only two components, while PCR, the most commonly used method, required 5 components to have similar predictive power. The coefficients obtained from PLS regression were more stable and accurate than that of the general linear model. Using PLS regression, the maternal factors identified as the significant predictors of newborn size at birth were commensurate with other studies [41–49]. So, 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.

Supporting information

S1 Dataset. Analytic dataset.
(XLSX)

S1 File. Code book of the analytic data set.
(DOCX)

S1 Table. Comparison between coefficients estimated from PLS and ordinary linear regression.
(DOCX)

Acknowledgments

First of all we would like to express our sincere gratitude to deceased Professor Mohammed Nasser (Department of Statistics, University of Rajshahi) for his continuous support until his death to accomplsh this study. He guided us in every aspects of this research. We gratefully acknowledge the contribution of the JiVitA study team comprised more than 800 Bangladeshi staff who carried out or supervised data collection, entry, and management and provided logistical, mapping, laboratory, and administrative support. We also acknowledge Angela KC who supported us in English editing and the Department of Statistics, University of Rajshahi whose faculties provided expert comments during the development of manuscript.

Author Contributions

Conceptualization: Alamgir Kabir.

Data curation: Alamgir Kabir, Rolf D. W. Klemm, Alain B. Labrique, Parul Christian, Keith P. West, Jr.

Formal analysis: Alamgir Kabir, Md. Jahanur Rahman.

Funding acquisition: Rolf D. W. Klemm, Parul Christian, Keith P. West, Jr.

Investigation: Rolf D. W. Klemm, Alain B. Labrique, Parul Christian, Keith P. West, Jr.

Methodology: Rolf D. W. Klemm, Parul Christian, Keith P. West, Jr.

Project administration: Abu Ahmed Shamim, Rolf D. W. Klemm, Alain B. Labrique, Mahbubur Rashid, Parul Christian, Keith P. West, Jr.

Resources: Rolf D. W. Klemm, Alain B. Labrique, Parul Christian, Keith P. West, Jr.

Supervision: Rolf D. W. Klemm, Parul Christian, Keith P. West, Jr.

Visualization: Alamgir Kabir, Md. Jahanur Rahman.

Writing – original draft: Alamgir Kabir.

Writing – review & editing: Alamgir Kabir, Md. Jahanur Rahman, Parul Christian, Keith P. West, Jr.

References

1. Lawn JE, Blencowe H, Oza S, You D, Lee ACC, Waiswa P, et al. (2014) Every Newborn: progress, priorities, and potential beyond survival. The Lancet 384: 189–205.
2. Katz J, Lee ACC, Kozuki N, Lawn JE, Cousens S, Blencowe H, et al. (2013) Mortality risk in preterm and small-for-gestational-age infants in low-income and middle-income countries: a pooled country analysis. The Lancet 382: 417–425.
3. Kozuki N, Katz J, LeClerq SC, Khatry SK, West KP Jr, Christian P (2015) Risk factors and neonatal/infant mortality risk of small-for-gestational-age and preterm birth in rural Nepal. The Journal of Maternal-Fetal & Neonatal Medicine 28: 1019–1025.
4. Kozuki N, Lee ACC, Silveira MF, Sania A, Vogel JP, Adair L, et al. (2013) The associations of parity and maternal age with small-for-gestational-age, preterm, and neonatal and infant mortality: a meta-analysis. BMC Public Health 13: S2.
5. Christian P, Mullany LC, Hurley KM, Katz J, Black RE. Nutrition and maternal, neonatal, and child health; 2015. Elsevier. pp. 361–372.
6. West KP, Christian P, Labrique AB, Rashid M, Shamim AA, Klemm RDW et al. (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: 1986–1995. https://doi.org/10.1001/jama.2011.656 PMID: 21586714
7. Sherry A, Henson RK (2005) Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. Journal of personality assessment 84: 37–48. https://doi.org/10.1207/s15327752jpa8401_09 PMID: 15639766
8. Wold H (1985) Partial least squares. Encyclopedia of statistical sciences 6: 581–591.
9. Abdi H (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdisciplinary Reviews: Computational Statistics 2: 97–106.
10. Labrique AB, Christian P, Klemm RDW, Rashid M, Shamim AA, Massie A, et al. (2011) A cluster-randomized, placebo-controlled, maternal vitamin A or beta-carotene supplementation trial in Bangladesh: design and methods. Trials 12: 102. https://doi.org/10.1186/1745-6215-12-102 PMID: 21510905
11. Gunnsteinsson S, Labrique AB, West KP Jr, Christian P, Mehra S, Shamim AA, et al. (2010) Constructing indices of rural living standards in Northwestern Bangladesh. Journal of health, population, and nutrition 28: 509. PMID: 20941903
12. Klemm RDW, Labrique AB, Christian P, Rashid M, Shamim AA, Katz J, et al. (2008) Newborn vitamin A supplementation reduced infant mortality in rural Bangladesh. Pediatrics 122: e242–e250. https://doi.org/10.1542/peds.2007-3448 PMID: 18595969
13. Gibson RS (2005) Principles of nutritional assessment: Oxford university press.
14. Dhar B, Mowlah G, Nahar S, Islam N (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.
15. Carrascal LM, Galván I, Gordo O (2009) Partial least squares regression as an alternative to current regression methods used in ecology. Oikos 118: 681–690.
16. Draper NR, Smith H (1998) Applied regression analysis 3rd edition. New York: Wiley.
17. Jolliffe IT (1982) A note on the use of principal components in regression. Applied Statistics: 300–303.
18. Cole TJ (2010) Can partial least squares regression separate the effects of body size and growth on later blood pressure?: Partial least squares regression. Epidemiology 21: 449–451. PMID: 20539106
19. Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Analytica chimica acta 185: 1–17.
20. Martens H, Izquierdo L, Thomassen M, Martens M (1986) Partial least-squares regression on design variables as an alternative to analysis of variance. Analytica chimica acta 191: 133–148.
21. Mevik Br-H, Wehrens R (2007) The pls package: principal component and partial least squares regression in R. Journal of Statistical Software 18: 1–24.
22. Wimberly MC, Lamsal A, Giacomo P, Chuang T-W (2014) Regional variation of climatic influences on West Nile virus outbreaks in the United States. The American journal of tropical medicine and hygiene 91: 677–684. https://doi.org/10.4269/ajtmh.14-0239 PMID: 25092814
23. Piccolo BD, Keim NL, Fiehn O, Adams SH, Van Loan MD, Newman JW (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: 681–690. https://doi.org/10.3945/jn.114.201574 PMID: 25833772

24. Wang H, Leung GM, Schooling CM (2015) Life course body mass index and adolescent self-esteem: Evidence from Hong Kong's "Children of 1997" Birth Cohort. Obesity 23: 429–435. https://doi.org/10.1002/oby.20984 PMID: 25557978

25. Yan X, Kruger JA, Nielsen PMF, Nash MP (2015) Effects of fetal head shape variation on the second stage of labour. Journal of biomechanics.

26. Pérez-Rodríguez A, Fernández-González S, Hera I, Pérez-Tris J (2013) Finding the appropriate variables to model the distribution of vector- borne parasites with different environmental preferences: climate is not enough. Global change biology 19: 3245–3253. https://doi.org/10.1111/gcb.12226 PMID: 23606561

27. Kwak HS, Ahn B-H, Kim H-R, Lee S-Y (2015) Identification of Senory Attributes That Drive the Likeability of Korean Rice Wines by American Panelists. Journal of food science 80: S161–S170.

28. Höskuldsson A (1988) PLS regression methods. Journal of chemometrics: 211–228.

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

30. Lindley AA, Benson JE, Grimes C, Cole TM, Herman AA (1999) The relationship in neonates between clinically measured head circumference and brain volume estimated from head CT-scans. Early human development 56: 17–29. PMID: 10530903

31. Stoch MB, Smythe PM (1963) Does undernutrition during infancy inhibit brain growth and subsequent intellectual development? Archives of Disease in Childhood 38: 546. PMID: 21032415

32. Yoo W, Mayberry R, Bae S, Singh K, He QP, Lillard JW Jr, et al. (2014) A study of effects of multicollinearity in the multivariable analysis. International journal of applied science and technology 4: 9. PMID: 25664257

33. Tu Y-K, Woolston A, Baxter PD, Gilthorpe MS (2010) Assessing the impact of body size in childhood and adolescence on blood pressure: an application of partial least squares regression. Epidemiology 21: 440–448. PMID: 20234316

34. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems 58: 109–130.

35. Boulesteix A-L, Strimmer K (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in bioinformatics 8: 32–44. https://doi.org/10.1093/bib/bbl016 PMID: 16772269

36. Hastie T, Tibshirani R, Friedman J (2009) Overview of supervised learning. The elements of statistical learning: Springer. pp. 9–41.

37. Christian P, Klemm R, Shamim AA, Ali H, Rashid M, Shaikh S, et al. (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: 188–194. https://doi.org/10.3945/ajcn.112.042275 PMID: 23151532

38. Kabir A, Merrill RD, Shamim AA, Klemn RDW, Labrique AB, Christian P, et al. (2014) Canonical Correlation Analysis of Infant's Size at Birth and Maternal Factors: A Study in Rural Northwest Bangladesh. PloS one 9: e94243. https://doi.org/10.1371/journal.pone.0094243 PMID: 24710082

39. Storms MR, Van Howe RS (2004) Birthweight by gestational age and sex at a rural referral center. Journal of perinatology 24: 236–240. https://doi.org/10.1038/sj.jp.7211065 PMID: 15014534

40. Helland IS (1990) Partial least squares regression and statistical models. Scandinavian Journal of Statistics: 97–114.

41. 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: 89. https://doi.org/10.1002/(SICI)1096-8644(200001)111:1<89::AID-AJPA6>3.0.CO;2-X PMID: 10618590

42. Elshibly EM, Schmalisch G (2009) Relationship between maternal and newborn anthropometric measurements in Sudan. Pediatrics International 51: 326–331. https://doi.org/10.1111/j.1442-200X.2008.02756.x PMID: 19400821

43. Ogbonna C, Woelk GB, Ning Y, Mudzamiri S, Mahomed K, Williams MA (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: 26–32. https://doi.org/10.1080/00016340600935664 PMID: 17230285

44. Hosain GMM, Chatterjee N, Begum A, Saha SC (2006) Factors associated with low birthweight in rural Bangladesh. Journal of tropical pediatrics 52: 87–91. https://doi.org/10.1093/tropej/fmi066 PMID: 16014761

45. Feleke Y, Enquoselassie F (1999) Maternal age, parity and gestational age on the size of the newborn in Addis Ababa. East African medical journal 76: 468–471. PMID: 10520356

46. Yunis KA, Khawaja M, Beydoun H, Nassif Y, Khogali M, Tamim H et al. (2007) Intrauterine growth standards in a developing country: a study of singleton livebirths at 28–42 weeks' gestation. Paediatric and perinatal epidemiology 21: 387–396. https://doi.org/10.1111/j.1365-3016.2007.00827.x PMID: 17697069

47. Karim E, Mascie-Taylor CGN (1997) The association between birthweight, sociodemographic variables and maternal anthropometry in an urban sample from Dhaka, Bangladesh. Annals of human biology 24: 387–401. PMID: 9300116

48. Matin A, Azimul SK, Matiur AKM, Shamianaz S, Shabnam JH, Islam T (2008) Maternal socioeconomic and nutritional determinants of low birth weight in urban area of Bangladesh. Journal of Dhaka Medical College 17: 83–87.

49. Raum E, Arabin B, Schlaud M, Walter U, Schwartz FW (2001) The impact of maternal education on intrauterine growth: a comparison of former West and East Germany. International Journal of Epidemiology 30: 81–87. PMID: 11171862