HHS Public Access

Author manuscript

J Appl Stat. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

J Appl Stat. ; 50(8): 1790–1811. doi:10.1080/02664763.2022.2043254.

An adjusted partial least squares regression framework to utilize additional exposure information in environmental mixture data analysis

Ruofei Du[1,2,*], **Li Luo**[1,2], **Laurie G. Hudson**[3], **Sara Nozadi**[3,4], **Johnnye Lewis**[3,4]

[1]Biostatistics Shared Resource, University of New Mexico Comprehensive Cancer Center, Albuquerque, NM, USA.

[2]Department of Internal Medicine, University of New Mexico, Albuquerque, NM, USA.

[3]Department of Pharmaceutical Sciences, College of Pharmacy, University of New Mexico, Albuquerque, NM, USA.

[4]Community Environmental Health Program, University of New Mexico, Albuquerque, NM, USA.

Abstract

In a large-scale environmental health population study that is composed of many subprojects, often different fractions of participants out of the total enrolled have measures of specific outcomes. It's conceptually reasonable to assume the association study would benefit from utilizing additional exposure information from those with a specific outcome of interest not measured. Partial least squares regression is one of the practical approaches to determine exposure-outcome associations for mixture data. Like a typical regression approach, however, the partial least squares regression requires that each data observation must have both complete covariate and outcome data for model fitting. In this paper, we propose novel adjustments to the general partial least squares regression to estimate and examine the association effects of individual environmental exposure variables to an outcome within a more complete context of the study population's environmental mixture exposures. The proposed framework essentially takes advantage of the bilinear model structure. It allows information from all participants, with or without the outcome values, to contribute to the model fitting and the statistical assessment of association effects. Using this proposed framework, incorporation of additional information will lead to smaller root mean square errors in the estimation of association effects, and improve the ability to assess the significance of the effects.

Keywords

Adjusted SIMPLS; Metal mixture exposure; Mixture analysis; Navajo; Birth Cohort

[*]Corresponding author: rdu@uams.edu.

Introduction

Through interaction with the environment, people are exposed to a multitude of chemicals that can directly or indirectly impact their health (Stern 1993). One example is exposure to heavy metals that are ubiquitous and persist in the environment (Järup 2003; Jaishankar et al. 2014). The effects of individual heavy metals on human health have been studied, but most often without the context of metal mixture exposures (Gidlow 2004; Jomova et al. 2011; Tollett et al. 2009); yet in reality heavy metal exposures often occur as mixtures (Silins et al. 2011; Wu et al. 2016).

Although studying the association of individual exposure variables comprising the mixture related to a health outcome is appealing for a more realistic assessment of the environmental influence on the outcome, there are many challenges to these statistical analyses. Statistical models are required to incorporate the complexity of mixture exposures into analyses. Data characteristics that impede analyses have often been identified, including multicollinearity and high dimensionality of the covariates in the mixtures. A study may also be obscured by relatively low association effects. Partial least squares regression (PLSR) is one of good options to determine exposure-outcome associations for mixture data, as it can create uncorrelated components by maximizing the correlation between the exposure mixture components and the outcome simultaneously. The PLSR approach has been found practically useful and widely applied in association analyses for mixture data in chemometrics (Kettaneh-Wold 1992).

In a large-scale environmental health population study that is composed of many subprojects, basic and essential characteristics are usually obtained from all participants following the consent of the participation, such as demographics (e.g. age, gender, education level) and measures of environmental exposure (e.g. heavy metal concentrations from a urine and/or blood sample). However, out of total enrolled participants often only a fraction have measures of specific outcomes. This may result from an outcome assessment being conducted in a particular time frame when the related subproject was going, and the outcome values from earlier or later enrolled participants were not sampled. It may also be due to limited funding resources in pilot or preliminary projects under the large-scale study that restricting the sample N for specific outcomes.

The association between an outcome of interest and the exposure is usually examined within the scope of the related subproject, i.e. using the complete-case observations only. However, as above described, the environmental exposure data may have been available from all study participants across different subprojects. For a mixture exposure, using the exposure information from a larger sample allows us to build a more accurate picture of the population exposure profiles, especially in the interrelationships among contaminants comprising the mixture. Thus all exposure data, no matter from participants with or without specific outcomes, remain important in a fuller characterization of the exposure; and analyses would benefit from the inclusion of this additional information.

By solely looking at how this full dataset is compiled, which includes both the complete-case part and the other subset having exposure variables only, one could think of the analytic

challenge as PLSR model fitting with missing outcome values. The missing data treatments for PLSR, which focus on missing value imputations, could then be applied. However there are at least two reasons that made us to view the study question as not for missing imputation need, but in the purpose of utilizing additional exposure information, and we further proposed the adjusted PLSR approach. Firstly, that data values are expected to be present but unavailable is considered missing data or missing values in a general sense. For a large-scale environmental study as outlined above, the absence of specific outcomes was in the original study schema, not out of expectation. Secondly, as we observed, the number of observations from the subset with no outcomes could be much greater than the sample size of the complete-case part. In the later presented two real datasets, we have >400 observations without outcome values but only 132 or 76 complete-case samples in each dataset respectively. Holding a conservative opinion, we do not believe using what learned from a small sample is adequate to impute the missing outcome values for a much larger pool. Nonetheless, a concise review of the existing missing data approaches for PLSR is provided in Discussion section.

Unlike a General Linear Model (GLM) which fits the association using one model equation, a PLSR relates the covariates and the outcome through other latent variables (i.e. the components, see details in Methods). Under the assumption of the PLSR approach, an observation with no outcome still informs the relationships between the observed covariates and the latent variables, and thus could contribute in part to improve the model fitting. However, the general PLSR approach was not designed to utilize the information from the observations without outcomes. We suggest applicable adjustments to the general PLSR algorithm that allows utilization of the full exposure dataset in characterizing exposure in the analysis. Starting from the adjustment on the component extraction, we propose an analytic framework that utilizes this more complete information to characterize the population exposure in model fitting, test statistic formation and hypothesis testing association effects. To clarify, in those studied practical situations, the participants no matter with or without specific outcome measured are deemed randomly from the study population. According to the common classification of missing mechanisms, our proposal assumes the unmeasured outcome follows missing completely at random (MCAR). This is further summarized in the Discussion section.

Motivating question: environmental heavy metal datasets from the Navajo Birth Cohort Study

There are more than 500 abandoned uranium mine (AUM) sites located on the lands of Navajo Nation (Lewis et al. 2015). People living there may be exposed by different pathways to AUM waste containing uranium, arsenic, and other co-occurring metals (Blake et al. 2015; Corlin et al. 2016; Orescanin et al. 2011). Several studies are or have been conducted to understand the association between the heavy metal exposures and specific health outcomes in this population (Markstrom and Charley 2003; Hund et al. 2015; Hoover et al. 2019). One of them, the Navajo Birth Cohort Study (NBCS) (Hunter et al. 2015), is a prospective study to investigate the potential associations between exposure

to environmental contaminants from the legacy mine wastes, and birth outcomes, and development of Navajo children.

Participating pregnant women or new mothers were asked to contribute their blood and urine samples at the time of enrollment, and again at their 36 week pregnancy visit or at the time of child delivery. Metal concentrations were measured in blood and urine samples, and used as the environmental exposure variables for the analyses of the two subproject outcome datasets described below. In the oxidative stress study (dataset 1), 132 enrollment urine samples were randomly selected for testing oxidative stress biomarker outcomes; there are an additional 417 samples from the enrollment biomonitoring pool that were not tested for the oxidative stress outcomes. For the Ages and Stages Questionnaire: Inventory (ASQ:I) (Clifford et al. 2018) developmental screening study (dataset 2), 76 infants completed the ASQ developmental screening at age 2, 6 and 12 months. In addition to those 76 infants' mothers, there are the other 447 mothers for whom biomonitoring samples were collected at 36 weeks or delivery, although their children did not participate in all the three-age ASQ:I assessments.

NBCS sample dataset 1: Oxidative stress dataset

Oxidative stress reflects an imbalance between free radicals and antioxidants and is associated with elevated oxidative damage to macromolecules including lipids, proteins and DNA (Sies 1991; Mateos and Bravo 2007). Oxidative stress is a factor during normal pregnancy, but excess oxidative stress is linked to a number of adverse outcomes (Duhig et al. 2016). Exposure to metals such as arsenic can lead to increased oxidative stress and oxidative damage is one proposed mechanism of metal toxicity (Valko et al. 2016; Gentile et al. 2017; Xu et al. 2017; Rehman et al. 2018). A published study (Dashner-Titus et al. 2018) focused on a randomly selected subset of 132 participants of the NBCS. Women's enrollment urine samples were analyzed for selected metals and the oxidative stress biomarkers of lipid peroxidation 8-iso-prostaglandin F2α (8-iso- PGF2α) and the ratio of 8-iso- PGF2α to prostaglandin F2α (PGF2α). The study investigated the relationships between the concentrations of urinary arsenic and uranium and an increased risk of oxidative stress. A significant positive association between urinary total arsenic and the single oxidative stress biomarker 8-iso- PGF2α was reported ($p = 0.012$); however, the association between total arsenic and elevation of the oxidative stress ratio (8-iso- PGF2α/ PGF2α) was marginally significant ($p = 0.053$). Uranium was not found to increase oxidative stress in the study population. This dataset is used here to evaluate how all measured metals related to oxidative stress, in addition to the original aim that was focused on urinary arsenic and uranium.

NBCS sample dataset 2: ASQ:I dataset

The development of the NBCS participants' children in the first year after the birth was assessed using the ASQ:I developmental screener. Mothers or alternate caregivers were interviewed about the child's progress in different developmental domains. Supplementary Figure 1 shows a spaghetti plot of the ASQ:I scores on the *problem-solving* developmental domain from 76 children whose assessments at age 2, 6 and 12 months had all been completed. In each of the three ages, there were about 300 children who participated in the

screening survey; however, due to loss to follow-up and the practical challenges of repeat scheduling in these remote communities with limited infrastructure, less than one-third of them persistently made the assessments at all of the three ages. The developmental trajectory shows a seemingly linear increasing trend over the 3 time-points. We then fit a general linear model for each child (ASQ:I score vs. age) and used the estimated slope as the surrogate endpoint to reflect the child's developmental rate in the first year. GLM and PLSR were employed to fit the associations between the estimated slope and the concentrations of the full suite of metals measured from urine and blood samples, but both methods failed to identify any association effects at a significance level of 0.05.

We were motivated to ask whether the assessment of relationships between metal mixtures and the oxidative stress and developmental outcomes could be improved through the incorporation of more exposure information that cannot be incorporated in the analyses performed. The associations to be examined here are in the context of the population metal mixture exposures using the full panel of measured metal concentrations from urine and blood samples.

Methods

The diagrams in Figure 1 present the method concept in accordance with the notations defined later. Our primary analytic interest is on how exposure variables ($\boldsymbol{X}$) related to an outcome variable ($\boldsymbol{y}$) with the associations being quantified by vector $\boldsymbol{b}$ (Figure 1, (a)). Considering the combined effect of the mixture exposure on the outcome, $\boldsymbol{y}$ is fitted against $\boldsymbol{T}$, a linear combination of $\boldsymbol{X}$ with assumed additive random errors (Figure 1, (b)). An estimate of the association between $\boldsymbol{y}$ and $\boldsymbol{X}$ (i.e. $\widehat{\boldsymbol{b}}$) can then be calculated from the estimate of the association between $\boldsymbol{y}$ and $\boldsymbol{T}$ (i.e. $\widehat{\boldsymbol{a}}$, see Figure 1, (c)).

For dealing with the correlation between the exposure variables, usually orthogonal transformations are performed to obtain an estimate of $\boldsymbol{T}$, such as the algorithm implemented in a principal component analysis (Wold et al. 1987). The PLSR outpaces the principal component regression with the linear transformation of $\boldsymbol{X}$ directed by maximization of the sample covariance between a linear transformed $\boldsymbol{X}$ and the outcome $\boldsymbol{y}$, in addition to the orthogonal transformations. This property is particularly helpful in handling the analytic challenge due to the relatively low association effect.

In other words, by the PLSR approach, the mixture data is rotated to a subspace in favor of detection of the association effects if existing; the estimates obtained using the rotated data are then converted back to estimate the association effect of each of original $\boldsymbol{X}$ variables on $\boldsymbol{y}$. The two model equation structure (Figure 1, (b)) further triggered us to propose the adjustment to the general PLSR that will be capable of including the fuller exposure dataset in the model fitting, and go beyond to benefit the significance assessment of association effects.

Algorithms of PLSR

One of the popular algorithms used for implementing PLSR is the SIMPLS, originally coined from "a straightforward implementation of a statistically inspired modification of the

PLS method" (De Jong 1993). SIMPLS was derived to solve the specific objective function, i.e. to maximize the covariance between the covariates and the outcome(s), which enables SIMPLS to work with the covariance matrix and conduct the deflation on the covariance (De Jong 1993). It offers several advantages over the other exiting algorithms for PLSR, mainly in the simpler interpretation and faster computation time. Nonetheless, it should be clarified that the interest of this paper is in using the two-step SIMPLS procedure to incorporate the additional exposure information for analysis (Hubert and Branden 2003), instead of promoting SIMPLS for the implementation of PLSR.

Notation and model equations

Throughout the paper, we will print a column vector in a bold, italic, and lowercase letter (e.g. $\boldsymbol{y}$), and a matrix in an uppercase letter (e.g. $\boldsymbol{X}$). The dimension of a matrix will be denoted using a subscript, for example $\boldsymbol{X}_{n\times p}$ stands for a matrix with n rows and p columns. A single number of a subscript shows the number of rows/records of a matrix/vector (e.g. $\boldsymbol{y}_s$ contains s records), while two numbers connected by a colon indicate the beginning and ending indices of the rows/records (e.g. $\boldsymbol{y}_{1:s}$ contains the records from 1st up to the sth).

Let $\boldsymbol{y}$ contains the continuous univariate outcome measures, and $\boldsymbol{X}$ is composed of the p-dimensional covariate row vectors. Here $(\boldsymbol{X}_s, \boldsymbol{y}_s)$ denotes the part of the dataset that has the complete observations with paired covariates and outcome; while the other part of the dataset, $\boldsymbol{X}_{s+1:n}$, has the observations with the covariates only and $\boldsymbol{X}_n = \left(\boldsymbol{X}_s', \boldsymbol{X}_{s+1:n}'\right)'$. We assume $\boldsymbol{X}_s$ is randomly selected for measurement of the outcome out of $\boldsymbol{X}_n$.

The x- and y- variables are assumed to be related through a bilinear model:

$$E(\boldsymbol{X}_{n\times p}) = \boldsymbol{T}_{n\times k}\boldsymbol{P}_{k\times p}, \tag{1}$$

$$E(\boldsymbol{y}_s) = \boldsymbol{T}_{s\times k}\boldsymbol{a}_k, \tag{2}$$

where $\boldsymbol{T}_{n\times k}$ is the component matrix of continuous variables with the mth column vector called the mth component. The slope vector of $E(\boldsymbol{y}_S)$ on $\boldsymbol{T}_{s\times k}$ is denoted by $\boldsymbol{a}_k$. Although not seen directly from the notations and equations given here, the slope vector of $E(\boldsymbol{y}_S)$ on $\boldsymbol{X}_{s\times p}$ is of primary interest to estimate, which we denote by $\boldsymbol{b}_p$. The elements in $\boldsymbol{b}_p$ are considered the association effects in our analysis.

Adjusted SIMPLS

As compared to the model fitting in general SIMPLS which only involves $\boldsymbol{X}_s$ and $\boldsymbol{y}_s$, the proposed adjustment in the model fitting below is in which datasets (i.e. $\boldsymbol{X}_s$, $\boldsymbol{y}_s$ or $\boldsymbol{X}_n$) are to be used for specific computations. The dataset *subscript notations* are therefore critical when going through the proposed adjustment steps to differentiate between the approaches.

The estimated values of $\boldsymbol{T}_{n\times k}$ are usually referred to as the component scores. In line with the bilinear model structure, SIMPLS estimates the component scores of the k latent variables $\boldsymbol{t}_{(1)}\cdots\boldsymbol{t}_{(m)}\cdots\boldsymbol{t}_{(k)}$ one after another, where the subscript (m) denotes the procedure

for mth component extraction; secondly as indicated by equation (2) the outcome will be regressed onto those k variables to acquire the estimates of interest.

With SIMPLS, the optimization objective for PLSR implementation is to maximize the covariance between a linear combination of $\boldsymbol{X}$ and $\boldsymbol{y}$. To obtain the first component scores, we solve for $\boldsymbol{r}_{(1)}$ that maximize the sample covariance between $\boldsymbol{X}_s\boldsymbol{r}_{(1)}$ and $\boldsymbol{y}_S$

$$Cov\left(\boldsymbol{X}_s\boldsymbol{r}_{(1)}, \boldsymbol{y}_s\right) = \boldsymbol{r}_{(1)}^{'}\boldsymbol{s}_{xy}, \tag{3}$$

where $\boldsymbol{s}_{xy}$ is the sample covariance vector between x- variables and y of the s subjects. For univariate outcome $\boldsymbol{y}_S$, this maximization has one straightforward solution that $\boldsymbol{r}_{(1)}$ is $\boldsymbol{s}_{xy}$. The component scores are then obtained by $\boldsymbol{X}_n$ multiplied with the normalized vector $\boldsymbol{r}_{(1)}$,

$$\boldsymbol{t}_{(1)} = \boldsymbol{X}_n\boldsymbol{r}_{(1)}/\sqrt{\boldsymbol{r}_{(1)}^{'}\boldsymbol{r}_{(1)}}. \tag{4}$$

From a geometric perspective, the rows of $\boldsymbol{X}_n$ are rotated without change the length of a row vector by the matrix multiplication of a normalized/unit vector here. Of note, although in equation (3) the covariance is estimated through the paired observations having both the covariates and outcomes, $(\boldsymbol{X}_S, \boldsymbol{y}_S)$, we compute the scores for all the n observations in (4). The x- loading vector that describes the linear relation between x- variables and the 1st component can be calculated as

$$\boldsymbol{p}_{(1)} = (\boldsymbol{t}_{(1)}^{'}\boldsymbol{t}_{(1)})^{-1}\boldsymbol{X}_n^{'}\boldsymbol{t}_{(1)}. \tag{5}$$

For the 2nd and above component extraction, for example in the mth component extraction, we construct an orthonormal base of $[\boldsymbol{p}_{(1)},\ldots,\boldsymbol{p}_{(m)}]$ denoted by $\boldsymbol{V}_{(m)} = [\boldsymbol{v}_{(1)},\ldots,\boldsymbol{v}_{(m)}]$ by the Gram–Schmidt process. Next, $\boldsymbol{s}_{xy}$ is deflated as

$$\boldsymbol{s}_{xy}^{(m)} = \boldsymbol{s}_{xy}^{(m-1)} - \boldsymbol{V}_{(m)}(\boldsymbol{V}_{(m)}^{'}\boldsymbol{s}_{xy}^{(m-1)}),\ m > 1 \text{ and } \boldsymbol{s}_{xy}^{(1)} = \boldsymbol{s}_{xy}. \tag{6}$$

The deflation means after the 1st component the successive component extractions will be oriented by maximizing the residual covariance with $\boldsymbol{s}_{xy}$ replaced by $\boldsymbol{s}_{xy}^{(m)}$ in equation (3). This will also assure that a new component is orthogonal to all previously extracted components.

Up to now, the first part of the adjusted SIMPLS has been conducted. To emphasize, the proposed adjustment is to include the component scores of all the n observations, $\boldsymbol{t}_{(m)}$, in the calculation of the x- loading vector $\boldsymbol{p}_{(m)}$ that will impact all the subsequent steps. We provided a proof of concept in Appendix that additional observations in $\boldsymbol{X}$ is able to produce the loading vector $\boldsymbol{p}_{(1)}$ having a smaller variance. For the 2nd and above component extraction, since $\boldsymbol{X}_S$ and $\boldsymbol{X}_n$ derived procedures lead to different deflated sample covariance vectors, the characteristics of $\boldsymbol{p}_{(m)}$'s from the two procedures have not been compared here; but in a general sense additional observations won't bring disadvantages.

The regression of the outcome on the k component scores can be secondly conducted to obtain an estimate of the slope vector $\boldsymbol{a}_k$,

$$\hat{\boldsymbol{a}}_k = (\hat{\boldsymbol{T}}'_{k\times s}\hat{\boldsymbol{T}}_{s\times k})^{-1}\hat{\boldsymbol{T}}'_{k\times s}\boldsymbol{y}_s\,. \tag{7}$$

where $\hat{\boldsymbol{T}}_{s\times k}$ is the submatrix containing 1:s rows of the matrix made of the columns of $\boldsymbol{t}_{(1)},\ldots,\boldsymbol{t}_{(k)}$.

Test statistic

The slope vector of $\boldsymbol{y}_S$ on $\boldsymbol{X}_{S\times p}$ can then be estimated as

$$\hat{\boldsymbol{b}}_p = \boldsymbol{R}_{p\times k}\hat{\boldsymbol{a}}_k = \boldsymbol{R}_{p\times k}(\hat{\boldsymbol{T}}'_{k\times s}\hat{\boldsymbol{T}}_{s\times k})^{-1}\hat{\boldsymbol{T}}'_{k\times s}\boldsymbol{y}_s, \tag{8}$$

where $\boldsymbol{R}_{p\times k} = [\boldsymbol{r}_{(1)},\ldots,\boldsymbol{r}_{(k)}]$. Note, $\hat{\boldsymbol{b}}_p$ is not a simple linear function of $\boldsymbol{y}_S$ because $\boldsymbol{R}_{p\times k}(\hat{\boldsymbol{T}}'_{k\times s}\hat{\boldsymbol{T}}_{s\times k})^{-1}\hat{\boldsymbol{T}}'_{k\times s}$ is dependent on $\boldsymbol{y}_S$. An estimate of the variance of $\hat{\boldsymbol{b}}_p$ is not easily attainable. In literature, people applied the local linearization as an approximation to compute the estimated variance-covariance matrix of $\hat{\boldsymbol{b}}_p$ (Denham 1997; Romera 2010), or utilized the resampling procedures such as jackknife or bootstrap to conduct an estimation (Martens H and Martens M 2000; Bastien et al. 2005). In order to have all the available information contribute in the test statistic construction, also not subject to any specific distribution assumption, we propose here a bootstrap approach for the variance estimation of $\hat{\boldsymbol{b}}_p$.

In the dataset under study, a bootstrapping sample will likewise also include two parts: the part of paired observations, ($\boldsymbol{X}^{bt}_s$, $\boldsymbol{y}^{bt}_s$); and the other subset containing no outcomes, $\boldsymbol{X}^{bt}_{s+1:n}$. The bootstrap resampling is performed for each of the parts separately,

($\boldsymbol{X}^{bt}_s$, $\boldsymbol{y}^{bt}_s$): random sampling from ($\boldsymbol{X}_S$,$\boldsymbol{y}_S$) with replacement,

$\boldsymbol{X}^{bt}_{s+1:n}$: random sampling from $\boldsymbol{X}_{S+1:n}$ with replacement.

Following the above-elaborated steps, we are able to get the estimated slope matrix, denoted as $\hat{\boldsymbol{b}}^{bt}_p$, for a bootstrapping sample. The bootstrap procedure is repeated a large number of times, e.g. 500 times. The sample variance computed on those $\hat{\boldsymbol{b}}^{bt}_p$'s is then used as the estimated variance for $\hat{\boldsymbol{b}}_p$. Eventually, the test statistic we propose is,

$$l_j = \frac{\hat{b}_j}{\sqrt{Var\left(\hat{b}^{bt}_j\right)}}, \quad j = 1,\ldots,p$$

where $\hat{b}_j$ is the jth component of $\hat{\boldsymbol{b}}_p$, and $Var\left(\hat{b}^{bt}_j\right)$ is the sample variance of the jth component of $\hat{\boldsymbol{b}}^{bt}_p$ by above bootstrapping steps.

Hypothesis testing

If the null hypothesis is true, that is changing the exposure will have no effect on the outcome, we can randomly select s exposure observations out of the total n observations and affiliate them to the s outcomes one to one. We can then compute the sampling distribution

of the test statistic under the null hypothesis by using those shuffled datasets. This testing idea is similar to the concept of a permutation test, allowing us to include all the available exposure information (i.e. $\boldsymbol{X}_n$) in the procedure as detailed below:

$\boldsymbol{X}_s^{null}$: random sampling from $\boldsymbol{X}_n$ without replacement,

$\left(\boldsymbol{X}_s^{null}, \boldsymbol{y}_s^{null}\right) = \left(\boldsymbol{X}_s^{null}, \boldsymbol{y}_s\right)$,

$\boldsymbol{X}_{s+1:n}^{null}$: the remaining observations after $\boldsymbol{X}_s^{null}$ excluded from $\boldsymbol{X}_n$.

By applying the adjusted SIMPLS algorithm and the test statistic computing steps aforementioned, we have l_j^{null} for each shuffle of the dataset. The percentage of l_j^{null}'s equal or more extreme than the l_j obtained from the studied dataset is used as the p-value for testing the association effect of a covariate on the outcome, which will be compared to a specified significance value (e.g. 0.05) to make a decision of rejection or not on the null hypothesis.

Simulation study

The bilinear model equations (1) and (2) serve as the fundamental equations for the simulation. We consider similar settings as applied for simulations in other papers in which the statistical uncertainties were assumed to follow normal distributions (Hubert and Branden 2003; Turkmen 2008). Specifically, the datasets were generated according to the sequential steps below:

$$\boldsymbol{T}_{n\times k} \sim N_k\left(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right),$$

$$\boldsymbol{X}_{n\times p} = \boldsymbol{T}_{n\times k}\boldsymbol{P}_{k\times p} + N_p\left(0, \boldsymbol{\Sigma}^x\right),$$

$$\boldsymbol{y}_s = \boldsymbol{T}_{1:s}\boldsymbol{a}_k + N\left(0, \sigma^2\right),$$

where N_k stands for a k-dimensional multivariate normal distribution. Throughout the simulation study, we set $n = 500$, $s = 100$, and carry out two sets of simulations separately.

Setting 1

In this batch of simulations, we have $p = 5$, $k = 2$. The values for the other parameters are set as $\boldsymbol{\mu}^t = (1\ 1)'$, $\boldsymbol{\Sigma}^t = diag(2\ 2)'$, $\boldsymbol{\Sigma}^x = diag(1\ 1\ 1\ 1\ 1)$, $\boldsymbol{a}_k = (1\ 3)'$ and $\sigma^2 = 1$, where *diag* stands for a diagonal matrix with the elements on the diagonal enclosed in the followed vector. The matrix of x- loadings is set,

$$\boldsymbol{P} = \begin{pmatrix} 5 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

which shows x_1 has greater loading than x_2 on the first component (5 vs. 2), x_4 and x_5 share the same loading on the second component (1 vs. 1), and all other loadings are zero. In the simulations, $\boldsymbol{P}$ is row-normalized and served as a rotation matrix of $\boldsymbol{T}$ to $\boldsymbol{X}$ in addition to the variations generated by $N_k(0,\boldsymbol{\Sigma}^x)$.

The slope vector of $\boldsymbol{y}$ on $\boldsymbol{X}$ can then be calculated as,

$$\boldsymbol{b} = (\boldsymbol{P}/rowsum(\boldsymbol{P}))'\boldsymbol{a}_k = (0.7\ 0.3\ 0\ 1.5\ 1.5)',$$

where $(\boldsymbol{P}/rowsum(\boldsymbol{P}))$ denote each element of $\boldsymbol{P}$ divided by its row sum. Figure 2 uses boxplots to show the distribution of the root mean squared errors (RMSE: $\sqrt{\frac{1}{p}\sum_{j=1}^{p}(b_j-\hat{b}_j)^2}$) in the estimation of the slope vector $\boldsymbol{b}$ by the general PLSR approach, which can only use 100 paired observations in a simulated dataset, and the proposed approach. The smaller the RMSE, the better the estimation performance. Figure 2 shows the adjusted PLSR (adj. PLSR) approach does a better job although the improvement looks minimal for this set of simulations. It is worth noting that PLSR is a biased estimation procedure (Frank and Friedman 1993), and so is the adjusted PLSR. Nonetheless, the ability to include more information into the analysis does lead to better estimation.

Table 1 summarizes the hypothesis testing performance of the proposed frame work compared with the general PLSR. The functions from R package *pls* have been applied to carry out the PLSR method in hypothesis testing with a leave-one-out jackknife method for variance estimation (Mevik et al. 2011). We present the rejection rates (the last two columns) using the significance levels of 0.05 or 0.1 as the cut-offs for the testing decision making. When a true value for testing is 0, which the third element of $\boldsymbol{b}$ (i.e. b_3) is, the associated rejection rates are the observed Type I errors. Both approaches have rejection rates lower than 0.05 and 0.1, respectively (row 5 and 6 in the last two columns), representing correct type I error control. When a true value for testing is nonzero, which is the situation for any other elements of $\boldsymbol{b}$ except b_3, the associated rejection rates are the estimated statistical powers. The adjusted PLSR outperforms the PLSR in testing of the two elements (b_1 and b_2) where the true value of b is small. Especially for b_2, which may represent a low association effect in this setting, the adjusted PLSR method has an increased statistical power of more than 30% relative to the general PLSR. For testing of b_4 or b_5, which is considered a large effect here, the two methods perform similarly.

Setting 2

For the second batch of simulations, we have $p = 25$, $k = 5$ that reflect the data dimensions we encountered when analyzing the NBCS datasets. The other values set for simulations are $\boldsymbol{\mu}^t = (1\ 1\ 1\ 1\ 1)'$, $\boldsymbol{\Sigma}^t = diag(2\ 2\ 2\ 2\ 2)'$, $\boldsymbol{\Sigma}^x = diag(3\cdots3)_{25}$, $\boldsymbol{a}_k = (1\ 2\ 3\ 4\ 5)'$ and $\sigma^2 = 2$. The matrix of x- loadings is set,

$$\boldsymbol{P} = \begin{pmatrix} 53100\ 00000\ 00000\ 00000\ 00000 \\ 00000\ 53100\ 00000\ 00000\ 00000 \\ 00000\ 00000\ 55200\ 00000\ 00000 \\ 00000\ 00000\ 00000\ 55200\ 00000 \\ 00000\ 00000\ 00000\ 00000\ 11100 \end{pmatrix}',$$

which shows for a component there are 3 x- variables having non-zero loadings. The loading weights are given (5, 3, 1), (5, 5, 2) or (1, 1, 1) to demonstrate the skewed or evenly distributed loading weights. $\boldsymbol{P}$ is still row-normalized before being used in the normalizations. The slope vector $\boldsymbol{b}$ is calculated using the same formula given in Setting 1, and shown element by element in Table 2.

For the simulations with Setting 2, the boxplots (Figure 3) show the proposed approach has noticeable lower RMSE values compared to the general PLSR. Table 2 displays the results of the evaluation of hypothesis testing performance. The observed Type I errors with both methods are controlled, i.e. the values in the last two columns for b_4, b_5, b_9, b_{10}, b_{14}, b_{15}, b_{19}, b_{20}, b_{24} and b_{25} are all < 0.05. The adjusted PLSR exhibits a consistently superior rate of rejection of the null hypothesis when there is a true association (nonzero value of b_j) compared with the general PLSR, regardless of larger or smaller effect size.

Application

The development of the adjusted PLSR was motivated by the analysis of the two NBCS datasets previously described. We were asking whether less variable estimates of the exposure based on the utilization of all available exposure data to reduce variance will result in a more robust assessment of the association effects when the proposed adjusted PLSR method is applied.

Oxidative stress dataset

We employed one of the endpoints reported in the previous study (Dashner-Titus et al. 2018), the ratio of 8-iso-$PGF_{2\alpha}$ to $PGF_{2\alpha}$. The ratio of 8-iso-PGF2α to prostaglandinF2α (PGF2α) is used to distinguish between enzymatic versus chemical lipid peroxidation as a biomarker of oxidative stress (van't Erve et al. 2015; van't Erve et al. 2016). We fit the transformed ratio in relation to the mixture of the heavy metal concentrations in blood, serum and urine. Allowing no missing data in the mixture exposure data, we identify 129 records having both the outcome and the metal values, and an additional 354 records with the complete metal exposure values only.

Comparing the quantities of the test p values resulting from the general PLSR and the adjusted PLSR approaches (Table 3), we see for some metals the insignificant effect becomes almost certain (e.g. BCD and SCU); while for some other metals, a marginal statistical significance is clarified by the reduced p value given by the adjusted PLSR (e.g. BMN, UAS3, UTAS). Reduced variance resulting from the incorporation of additional exposure records in the adjusted PLSR approach reduces uncertainty for both significant and nonsignificant tested effects.

In the previous publication, the association between total arsenic in urine and the prostaglandin ratio was marginally significant ($p = 0.053$); while total arsenic showed a significant association with another oxidative stress biomarker (i.e. the biomarker 8-iso-$PGF_{2\alpha}$, $p = 0.012$) (Dashner-Titus et al. 2018). Using the adjusted PLSR, the association of urinary total arsenic with the prostaglandin ratio was statistically significant (p=0.046), which may serve as a piece of confirmative evidence that an increase in total urinary arsenic elevates oxidative stress for the study population. The proposed method additionally found UAS3 and BMN had significant positive associations to oxidative stress (for both the metals p=0.05), but the results may need to be further confirmed. The consistency in non-significance of urinary uranium even after the proposed adjustment would suggest that it is not a significant contributor to the outcome of interest.

ASQ:I dataset

The associations between metal mixture exposures and developmental trajectories through age 1 based on the ASQ:I screening provide a second example of the application of the adjusted PLSR approach. As described above, in addition to the 76 mothers whose children had both delivery or 36-week biomonitoring and all ASQ:I scores on the *problem-solving* domain assessed at ages 2, 6 and 12 months, there are an additional 447 mothers having metal concentration measures at 36 weeks or at the child delivery (considered comparable timepoints). Allowing no missing data in the mixture exposure, we identified an additional 315 exposure records for the application of the proposed method. The analysis results are given in Table 4. Copper in serum shows a negative effect (test statistics: −0.76, p-value: 0.022) on children's developmental problem-solving trajectory, which was not detected by the standard PLSR approach, likely due to the lack of statistical power as demonstrated in the simulation study. Relative to the PLSR, the adjusted procedure has again shifted the p-values in both directions, helping to differentiate between the effects of these metals through reduced variance and increased power in the method.

Discussion

Missing data in PLSR.

Missing data are often seen in research. Common reasons accounting for missing data include nonresponse to particular questions, technical limitations in data recording (e.g. instrument detection limit), loss of some participants in follow-ups, practical challenges in scheduling for repeated measures, etc. The widely used categorization of data missingness are (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR). MCAR indicates the missingness is independent to the values of any variables whether observed or missed. MAR implies the cause of the missing data is independent to the missing values, but may be related to the observed values of the other variables in the study. MNAR applies when neither of the MCAR and MAR cannot be assumed. In the data setting for our study, those participants without specific outcomes were randomly from the study population, not related to any values of the variables in the study; so MCAR applies for the adjusted PLSR.

PLSR model fitting with missing data started drawing attention around mid-1990s (Nelson et al. 1996), and different missing data treatments have been proposed since. Among the popular ones are iterative algorithm (IA), singular value decomposition (SVD) (Walczak and Massart 2001), trimmed score regression (TSR) (Camacho et al. 2008), projection to the model plane (PMP), single-component projection (SCP), conditional mean replacement (CMR) (Nelson et al. 1996; Arteaga and Ferrer 2002) and known data regression (KDR) (Folch-Fortuny et al. 2017). These methods can be seen as different ways to impute values for the missing variables. In the majority of previous studies, only missing data in explanatory variables (i.e. covariates in our methods) are considered, although some of methods can be easily adapted to analyze data with missing outcome values (Walczak and Massart 2001; Folch-Fortuny et al. 2017). Those approaches essentially impute the missing covariate values through the component scores estimated from the iteratively fitted PLSRs starting from complete-case observations only, or with the missing data substituted by some reasonable values to initiate the process. Some other generic missing data methods have been studied and applied in PLSR fitting as well, for example, the multiple imputation by chained equations (MICE) (White et al. 2011) and *k*-nearest neighbor imputation (Malarvizhi and Thanamani 2012).

Methods.

A prime step in PLSR is to rotate the covariate matrix (i.e. the linear transformation of $\boldsymbol{X}_S$) supervised by maximization of the covariance between the rotated data and the outcome (see equation (3)). In the proposed adjustments of the SIMPLS procedure, although the vector used for data rotation for the first component extraction ($\boldsymbol{r}_{(1)}$ in equations (3)) is the same as it in PLSR, the *x*- loading vector ($\boldsymbol{p}_{(1)}$ in equation (5)) is now calculated using all the n observations. This loading vector estimates the linear relation between *x*- variables and the 1st component. Utilizing more observations provides a better estimate that is not affected by the outcome variable when the rotation direction is already determined (i.e. $\boldsymbol{r}_{(1)}$ has being obtained). The improved estimate will thereafter benefit the subsequent steps.

The observations without outcomes further contribute to the estimation of the variance of the computed slope vector ($\hat{\boldsymbol{b}}_p$, see equation (8)), and in establishing the sampling distribution of the test statistic under the null hypothesis. Especially since the PLSR outputs create a biased estimate (Frank and Friedman 1993), the resampling procedures that randomly link *x*- variable values to the outcome values to fit the association is a favorable approach for hypothesis testing. The observations with only *x*- variables are accommodated into this scheme reasonably, to provide more realizations of *x*-variables linked to the resampled outcome values to improve the statistical power in the detection of an association effect.

From our experience, additional to the complexity of the mixture itself, the relatively low association effect between the mixture and a study outcome is also a critical challenge in statistical analysis. A statistical method needs to be sensitive in capturing weak, but valid, signals of association for risk effects. Ensuring the analytic method can utilize all available information, reduce variance and increase the robustness of both identifying true effects, and correctly identifying exposures with no association is critical in informing decisions to protect health. High rates of borderline associations in either direction raise questions as

to whether real and non-associations are muddied by the methodology employed. Ensuring the method is as robust as possible by maximizing the use of available information can increase the confidence of the interpretation. By the PLSR approach, the mixture data is first rotated to a subspace in favor of detection of the association effect; the estimates obtained from the regression modeling with the rotated data are then converted back to the effect of each original x-variable. In our simulation study and the applications with real datasets, the results by the PLSR or adjusted PLSR with statistical significance are oftentimes more logical and interpretable, helping to differentiate between metals contributing and not contributing to the observed effect as demonstrated through the simulation studies. Although we here focused on prompting the capacity of the proposed framework in utilizing the additional information, the proposed adjusted PLSR framework likely provides a good tool to tackle the low signal to noise ratio issue in association analyses in population studies where real-world environments contribute higher degrees of the variance than seen in controlled laboratory settings.

The cross-validation (CV) was commented as a practical and reliable way in determining the optimal number of components a PLSR building (Wold et al. 2001), which we applied in analyzing the two real datasets presented above. The complete-case observations were only used this determination. Specifically the leave-one-out cross-validation with adjusted Mean Squared Error of Prediction, combining the consideration of the percentage of variance explained in the covariate matrix, helped select the number of components in analyses.

Oxidative stress dataset.

The proposed method is in agreement with several other population studies investigating arsenic exposure and oxidative stress where positive associations between arsenic exposure and the urinary oxidative stress biomarkers have been reported (Lu et al. 2016; Kubota et al. 2006; Wang et al. 2015). The method also identified manganese from blood negatively associated with the oxidative stress ratio biomarker with a marginal significance (p=0.05). Although manganese plays an important role in redox homeostasis (Li and Yang 2018; Bresciani et al. 2015), there is limited information from population studies on the relationship between blood manganese levels and biomarkers of oxidative stress (Chen et al. 2016; Nascimento et al. 2016; Andrade et al. 2015). The normal reference range of manganese concentration in blood is 4 to 15 μg/L (Coles et al. 2012); however, the concentration in blood can become much higher during pregnancy (Zota et al. 2009; Chung et al. 2015; Kim 2018). From our dataset with all of the 483 observations, the mean (std.) of blood manganese is 19.7 (6.5) μg/L. Figure 4 shows the scatter plot of the prostaglandin ratio against blood manganese from our dataset. It’s not clear whether the claimed negative association with a marginal statistical significance is due to the outlier points or it’s a true signal. The findings from this approach suggest further studying on the consideration of manganese as a factor related to oxidative stress for the pregnant woman population.

ASQ:I dataset.

The scatter plot of the calculated slope for *problem-solving* against copper in serum is displayed in Figure 5. Studies have suggested the risk of copper toxicity contributing to cognitive decline in older adults (Lam et al. 2008) and the general adult population

(Salustri et al. 2010). The association between higher copper level in serum and poorer working memory in schoolboys (aged 10–14 years old) has also been reported from a study conducted in China (Zhou et al. 2015). While copper is known to increase in pregnancy, the normal reference range of serum copper concentration for the third pregnancy trimester is 130–240 μg/dL (Abbassi-Ghanavati et al. 2009). In mother's included in our *problem-solving* domain dataset, 225 out of 391 women (57%) had serum copper concentrations higher than this normal pregnancy reference range (the range between the two dashed lines on the x-axis in Figure 5), which may be a sign of excessive copper in the study population. Our analysis results provide evidence supporting an inverse association between copper in maternal serum and the child's cognitive development in this cohort, which certainly merits further investigation for the Navajo community.

Limitations.

The limitations in this study should also be noted. For analysis of mixture data, there are factors or covariates that may need to be included in parallel with the mixture for regression fitting. With the current version of the proposed analytic framework, however, all the covariates have to be assembled into one design matrix to rotate and then to fit the regression. The approach at the current stage does not provide flexibility for treating variables independent to the exposure mixture. It is an ongoing research topic for us; more broadly we are studying how to fit PLSR into different regression approaches with varied formulation schemes. In addition, the PLSR method itself does not evaluate the interaction effects between the covariates from the mixture; neither does the adjusted PLSR approach. The adjusted PLSR provides a reduced variance estimate that increases the robustness of the analysis to identify individual components of mixtures as they contribute to outcomes of interest, as validated through the simulation studies. When synergistic or antagonistic effects of some covariates are suspected, other means should be applied to determine if interaction effects need to be included to fit the association; e.g. visual examination of contour plot for two-variable interaction effect, etc. A separate variable can be established or coded, as a GLM does, to include an interaction effect in the design matrix when the adjusted PLSR is applied.

Finally, we want to point out that PLSR and the adjusted PLSR are both capable of analyzing multivariate outcome data variables as well. In the literature, the PLSR for handling univariate outcomes is commonly referred to as PLS1, and PLS2 otherwise (Hubert and Branden 2003; Rosipal et al. 2005). The approach covered here provides adjustment to PLS1, although it would be straightforward to apply the proposed adjustments and the subsequent procedures to PLS2. We have provided the R codes for the implementation of the proposed methods (with both PLS1 and PLS2) at https://github.com/rdu2017/adj-PLSR which is publicly accessible.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Appendix

Proof of an element in the loading vector $p_{(1)}$ derived from X_n has a smaller variance compared to which in $p_{(1)}$ derived from X_s when the expectations of $p_{(1)}$ by the two approaches are the same.

In below expressions, we use a superscript to denote the ith column of a matrix, or the ith element of a vector. Since the loading vector $\boldsymbol{p}_{(1)} = (\boldsymbol{t}_{(1)}'\boldsymbol{t}_{(1)})^{-1}\boldsymbol{X}'\boldsymbol{t}_{(1)}$, the ith element of $\boldsymbol{p}_{(1)}$ is $p_{(1)}^i = (\boldsymbol{t}_{(1)}'\boldsymbol{t}_{(1)})^{-1}\boldsymbol{t}_{(1)}'\boldsymbol{x}^i$. Assume the variance of $\boldsymbol{x}^i$ is σ_i^2, the variance of $p_{(1)}^i$ can be calculated,

$$Var(p_{(1)}^i) = (\boldsymbol{t}_{(1)}'\boldsymbol{t}_{(1)})^{-1}\boldsymbol{t}_{(1)}'Var(\boldsymbol{x}^i)\boldsymbol{t}_{(1)}(\boldsymbol{t}_{(1)}'\boldsymbol{t}_{(1)})^{-1} = \frac{\sigma_i^2}{\boldsymbol{t}_{(1)}'\boldsymbol{t}_{(1)}}.$$

Below we show that $\boldsymbol{t}_{(1)}'\boldsymbol{t}_{(1)}$ obtained using all the n observations in $\boldsymbol{X}$ is greater than that obtained using only s observations in $\boldsymbol{X}$, and consequently have the argument proved.

In the first component extraction, the same $\boldsymbol{r}_{(1)}$ is used in generating the score vector $\boldsymbol{t}_{(1)}$ no matter with $\boldsymbol{X}_s$ by the general PLSR, or with $\boldsymbol{X}_n$ as the proposed adjustment,

$$\boldsymbol{t}_{(1)} = \boldsymbol{X}\boldsymbol{r}_{(1)}/\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}.$$

Having all n observations in $\boldsymbol{X}$,

$$\begin{aligned}\boldsymbol{t}_{(1)}'\boldsymbol{t}_{(1)} &= \frac{\boldsymbol{r}_{(1)}'}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}}\boldsymbol{X}_n'\boldsymbol{X}_n\frac{\boldsymbol{r}_{(1)}}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}} \\ &= \frac{\boldsymbol{r}_{(1)}'}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}}(\boldsymbol{X}_s' \ \boldsymbol{X}_u')\begin{pmatrix}\boldsymbol{X}_s \\ \boldsymbol{X}_u\end{pmatrix}\frac{\boldsymbol{r}_{(1)}}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}} \\ &= \frac{\boldsymbol{r}_{(1)}}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}}\boldsymbol{X}_s'\boldsymbol{X}_s\frac{\boldsymbol{r}_{(1)}}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}} + \frac{\boldsymbol{r}_{(1)}'}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}}\boldsymbol{X}_u'\boldsymbol{X}_u\frac{\boldsymbol{r}_{(1)}}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}},\end{aligned}$$

where $\boldsymbol{X}_u$ is the remaining part of the design matrix additional to the s observations in $\boldsymbol{X}_n$. The last expression is a summation of all positive numbers, since the inner product of a vector and itself is always nonnegative, and not all elements in either $\boldsymbol{r}_{(1)}$, $\boldsymbol{X}_s$, or $\boldsymbol{X}_u$ are assumed 0s. We thus have,

$$\frac{\boldsymbol{r}_{(1)}'}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}}\boldsymbol{X}_s'\boldsymbol{X}_s\frac{\boldsymbol{r}_{(1)}}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}} + \frac{\boldsymbol{r}_{(1)}'}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}}\boldsymbol{X}_u'\boldsymbol{X}_u\frac{\boldsymbol{r}_{(1)}}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}} > \frac{\boldsymbol{r}_{(1)}'}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}}\boldsymbol{X}_s'\boldsymbol{X}_s\frac{\boldsymbol{r}_{(1)}}{\sqrt{\boldsymbol{r}_{(1)}'\boldsymbol{r}_{(1)}}}.$$

That says $\boldsymbol{t}_{(1)}'\boldsymbol{t}_{(1)}$ obtained using $\boldsymbol{X}_n$ is greater than it obtained using only $\boldsymbol{X}_s$.

Another fact is that the expectation of $\boldsymbol{p}_{(1)}$ derived from $\boldsymbol{X}_n$ is the same as it derived from $\boldsymbol{X}_s$, since simply $E(\boldsymbol{X})$ is not related to the number of observations, and the same $\boldsymbol{r}_{(1)}$ being used in both the derivations.

Reference

Abbassi-Ghanavati M, Greer LG, & Cunningham FG (2009). Pregnancy and laboratory studies: a reference table for clinicians. Obstetrics & Gynecology, 114(6), 1326–1331. [PubMed: 19935037]

Andrade VM, Mateus ML, Batoreu MC, Aschner M, & Dos Santos AM (2015). Lead, arsenic, and manganese metal mixture exposures: focus on biomarkers of effect. Biological trace element research, 166(1), 13–23. [PubMed: 25693681]

Arteaga F, & Ferrer A (2002). Dealing with missing data in MSPC: several methods, different interpretations, some examples. Journal of Chemometrics: A Journal of the Chemometrics Society, 16(8–10), 408–418.

Bastien P, Vinzi VE, & Tenenhaus M (2005). PLS generalised linear regression. Computational Statistics & data analysis, 48(1), 17–46.

Blake JM, Avasarala S, Artyushkova K, Ali AMS, Brearley AJ, Shuey C, … & Hirani C (2015). Elevated concentrations of U and co-occurring metals in abandoned mine wastes in a northeastern Arizona Native American community. Environmental science & technology, 49(14), 8506–8514. [PubMed: 26158204]

Bresciani G, da Cruz IBM, & González-Gallego J (2015). Manganese superoxide dismutase and oxidative stress modulation. In Advances in clinical chemistry (Vol. 68, pp. 87–130). Elsevier. [PubMed: 25858870]

Camacho J, Picó J, & Ferrer A (2008). Bilinear modelling of batch processes. Part II: a comparison of PLS soft-sensors. Journal of Chemometrics: A Journal of the Chemometrics Society, 22(10), 533–547.

Clifford J, Chen CI, Xie H, Chen CY, Murphy K, Ascetta K, … & Hansen S (2018). Examining the technical adequacy of the ages & stages questionnaires: INVENTORY. Infants & Young Children, 31(4), 310–325.

Coles C, Crawford J, McClure PR, Roney N, & Todd GD (2012). Toxicological profile for manganese.

Corlin L, Rock T, Cordova J, Woodin M, Durant JL, Gute DM, … & Brugge D (2016). Health effects and environmental justice concerns of exposure to uranium in drinking water. Current environmental health reports, 3(4), 434–442. [PubMed: 27815781]

Chen P, Culbreth M, & Aschner M (2016). Exposure, epidemiology, and mechanism of the environmental toxicant manganese. Environmental Science and Pollution Research, 23(14), 13802–13810. [PubMed: 27102617]

Chung SE, Cheong HK, Ha EH, Kim BN, Ha M, Kim Y, … & Oh SY (2015). Maternal blood manganese and early neurodevelopment: the mothers and children's environmental health (MOCEH) study. Environmental health perspectives, 123(7), 717–722. [PubMed: 25734517]

Dashner-Titus EJ, Hoover J, Li L, Lee JH, Du R, Liu KJ, … & Hudson LG (2018). Metal exposure and oxidative stress markers in pregnant Navajo Birth Cohort Study participants. Free Radical Biology and Medicine, 124, 484–492. [PubMed: 29723666]

De Jong S (1993). SIMPLS: an alternative approach to partial least squares regression. Chemometrics and intelligent laboratory systems, 18(3), 251–263.

Denham MC (1997). Prediction intervals in partial least squares. Journal of Chemometrics: A Journal of the Chemometrics Society, 11(1), 39–52.

Duhig K, Chappell LC, & Shennan AH (2016). Oxidative stress in pregnancy and reproduction. Obstetric medicine, 9(3), 113–116. [PubMed: 27630746]

Frank LE, & Friedman JH (1993). A statistical view of some chemometrics regression tools. Technometrics, 35(2), 109–135.

Folch-Fortuny A, Arteaga F, & Ferrer A (2017). PLS model building with missing data: new algorithms and a comparative study. Journal of Chemometrics, 31(7), e2897.

Gentile F, Arcaro A, Pizzimenti S, Daga M, Cetrangolo GP, Dianzani C, … & Barrera G (2017). DNA damage by lipid peroxidation products: implications in cancer, inflammation and autoimmunity. AIMS genetics, 4(2), 103. [PubMed: 31435505]

Gidlow DA (2004). Lead toxicity. Occupational medicine, 54(2), 76–81. [PubMed: 15020724]

Hoover J, Erdei E, Nash J, & Gonzales M (2019). A Review of Metal Exposure Studies Conducted in the Rural Southwestern and Mountain West Region of the United States. Current epidemiology reports, 6(1), 34–49. [PubMed: 30906686]

Hubert M, & Branden KV (2003). Robust methods for partial least squares regression. Journal of Chemometrics: A Journal of the Chemometrics Society, 17(10), 537–549.

Hund L, Bedrick EJ, Miller C, Huerta G, Nez T, Ramone S, … & Lewis J (2015). A Bayesian framework for estimating disease risk due to exposure to uranium mine and mill waste on the Navajo Nation. Journal of the Royal Statistical Society. Series A (Statistics in Society), 1069–1091.

Hunter CM, Lewis J, Peter D, Begay MG, & Ragin-Wilson A (2015). DIRECT FROM ATSDR: The Navajo Birth Cohort Study. Journal of environmental health, 78(2), 42–45. [PubMed: 26502566]

Jaishankar M, Tseten T, Anbalagan N, Mathew BB, & Beeregowda KN (2014). Toxicity, mechanism and health effects of some heavy metals. Interdisciplinary toxicology, 7(2), 60–72. [PubMed: 26109881]

Jomova K, Jenisova Z, Feszterova M, Baros S, Liska J, Hudecova D, … & Valko M (2011). Arsenic: toxicity, oxidative stress and human disease. Journal of Applied Toxicology, 31(2), 95–107. [PubMed: 21321970]

Järup L (2003). Hazards of heavy metal contamination. British medical bulletin, 68(1), 167–182. [PubMed: 14757716]

Kettaneh-Wold N (1992). Analysis of mixture data with partial least squares. Chemometrics and Intelligent Laboratory Systems, 14(1–3), 57–69.

Kim Y (2018). Sex, pregnancy, and age-specific differences of blood manganese levels in relation to iron status; what does it mean?. Toxicology Reports, 5, 28–30. [PubMed: 29270364]

Kubota R, Kunito T, Agusa T, Fujihara J, Monirith I, Iwata H, … & Tanabe S (2006). Urinary 8-hydroxy-2′-deoxyguanosine in inhabitants chronically exposed to arsenic in groundwater in Cambodia. Journal of Environmental Monitoring, 8(2), 293–299. [PubMed: 16470262]

Lam PK, Kritz-Silverstein D, Barrett-Connor E, Milne D, Nielsen F, Gamst A, … & Wingard D (2008). Plasma trace elements and cognitive function in older men and women: the Rancho Bernardo study. The Journal of Nutrition Health and Aging, 12(1), 22–27.

Lewis J, Gonzales M, Burnette C, Benally M, Seanez P, Shuey C, … & Nez S (2015). Environmental exposures to metals in native communities and implications for child development: basis for the Navajo Birth Cohort Study. Journal of social work in disability & rehabilitation, 14(3–4), 245–269. [PubMed: 26151586]

Li L, & Yang X (2018). The essential element manganese, oxidative stress, and metabolic diseases: links and interactions. Oxidative medicine and cellular longevity, 2018.

Lu S, Ren L, Fang J, Ji J, Liu G, Zhang J, … & Fan R (2016). Trace elements are associated with urinary 8-hydroxy-2′-deoxyguanosine level: a case study of college students in Guangzhou, China. Environmental Science and Pollution Research, 23(9), 8484–8491. [PubMed: 26782679]

Malarvizhi R, & Thanamani AS (2012). K-nearest neighbor in missing data imputation. International Journal of Engineering Research and Development, 5(1), 5–7.

Markstrom CA, & Charley PH (2003). Psychological effects of technological/human-caused environmental disasters: examination of the Navajo and uranium. American Indian and Alaska Native Mental Health Research: The Journal of the National Center, 11(1), 19–45. [PubMed: 12955630]

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(1–2), 5–16.

Mateos R, & Bravo L (2007). Chromatographic and electrophoretic methods for the analysis of biomarkers of oxidative damage to macromolecules (DNA, lipids, and proteins). Journal of separation science, 30(2), 175–191. [PubMed: 17390612]

Mevik BH, Wehrens R, & Liland KH (2011). pls: Partial least squares and principal component regression. R package version, 2(3).

Nascimento S, Baierle M, Göethel G, Barth A, Brucker N, Charão M, … & Jager M (2016). Associations among environmental exposure to manganese, neuropsychological performance,

oxidative damage and kidney biomarkers in children. Environmental research, 147, 32–43. [PubMed: 26844420]

Nelson PR, Taylor PA, & MacGregor JF (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. Chemometrics and intelligent laboratory systems, 35(1), 45–65.

Orescanin V, Kollar R, Nad K, Mikelic IL, & Kollar I (2011). Characterization and treatment of water used for human consumption from six sources located in the Cameron/Tuba city abandoned uranium mining area. Journal of Environmental Science and Health Part A, 46(6), 627–635.

Rehman K, Fatima F, Waheed I, & Akash MSH (2018). Prevalence of exposure of heavy metals and their impact on health consequences. Journal of cellular biochemistry, 119(1), 157–184. [PubMed: 28643849]

Romera R (2010). Prediction intervals in Partial Least Squares regression via a new local linearization approach. Chemometrics and Intelligent Laboratory Systems, 103(2), 122–128.

Rosipal R, & Krämer N (2005, February). Overview and recent advances in partial least squares. In International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection" (pp. 34–51). Springer, Berlin, Heidelberg.

Salustri C, Barbati G, Ghidoni R, Quintiliani L, Ciappina S, Binetti G, & Squitti R (2010). Is cognitive function linked to serum free copper levels? A cohort study in a normal population. Clinical Neurophysiology, 121(4), 502–507. [PubMed: 20097602]

Sies H (1991). Oxidative Stress: Oxidants and Antioxidants Academic Press. New York.

Silins I, & Högberg J (2011). Combined toxic exposures and human health: biomarkers of exposure and effect. International journal of environmental research and public health, 8(3), 629–647. [PubMed: 21556171]

Stern PC (1993). A second environmental science: human-environment interactions. Science, 260(5116), 1897–1899. [PubMed: 17836719]

Tollett VD, Benvenutti EL, Deer LA, & Rice TM (2009). Differential toxicity to Cd, Pb, and Cu in dragonfly larvae (Insecta: Odonata). Archives of environmental contamination and toxicology, 56(1), 77. [PubMed: 18421495]

Turkmen A (2008). Robust partial least squares for regression and classification (Doctoral dissertation).

Valko M, Jomova K, Rhodes CJ, Kuča K, & Musílek K (2016). Redox-and non-redox-metal-induced formation of free radicals and their role in human disease. Archives of toxicology, 90(1), 1–37. [PubMed: 26343967]

van't Erve TJ, Lih FB, Kadiiska MB, Deterding LJ, Eling TE, & Mason RP (2015). Reinterpreting the best biomarker of oxidative stress: The 8-iso-PGF2α/PGF2α ratio distinguishes chemical from enzymatic lipid peroxidation. Free Radical Biology and Medicine, 83, 245–251. [PubMed: 25772010]

van't Erve TJ, Lih FB, Jelsema C, Deterding LJ, Eling TE, Mason RP, & Kadiiska MB (2016). Reinterpreting the best biomarker of oxidative stress: the 8-iso-prostaglandin F2α/prostaglandin F2α ratio shows complex origins of lipid peroxidation biomarkers in animal models. Free Radical Biology and Medicine, 95, 65–73. [PubMed: 26964509]

Walczak B, & Massart DL (2001). Dealing with missing data: Part I. Chemometrics and Intelligent Laboratory Systems, 58(1), 15–27.

Wang T, Feng W, Kuang D, Deng Q, Zhang W, Wang S, … & Guo H (2015). The effects of heavy metals and their interactions with polycyclic aromatic hydrocarbons on the oxidative stress among coke-oven workers. Environmental Research, 140, 405–413. [PubMed: 25956561]

White IR, Royston P, & Wood AM (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine, 30(4), 377–399. [PubMed: 21225900]

Wold S, Esbensen K, & Geladi P (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1–3), 37–52.

Wold S, Sjöström M, & Eriksson L (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 58(2), 109–130.

Wu X, Cobbina SJ, Mao G, Xu H, Zhang Z, & Yang L (2016). A review of toxicity and mechanisms of individual and mixtures of heavy metals in the environment. Environmental Science and Pollution Research, 23(9), 8244–8259. [PubMed: 26965280]

Xu J, Wise JT, Wang L, Schumann K, Zhang Z, & Shi X (2017). Dual roles of oxidative stress in metal carcinogenesis. Journal of Environmental Pathology, Toxicology and Oncology, 36(4).

Zota AR, Ettinger AS, Bouchard M, Amarasiriwardena CJ, Schwartz J, Hu H, & Wright RO (2009). Maternal blood manganese levels and infant birth weight. Epidemiology (Cambridge, Mass.), 20(3), 367. [PubMed: 19289966]

Zhou G, Ji X, Cui N, Cao S, Liu C, & Liu J (2015). Association between serum copper status and working memory in schoolchildren. Nutrients, 7(9), 7185–7196. [PubMed: 26343713]

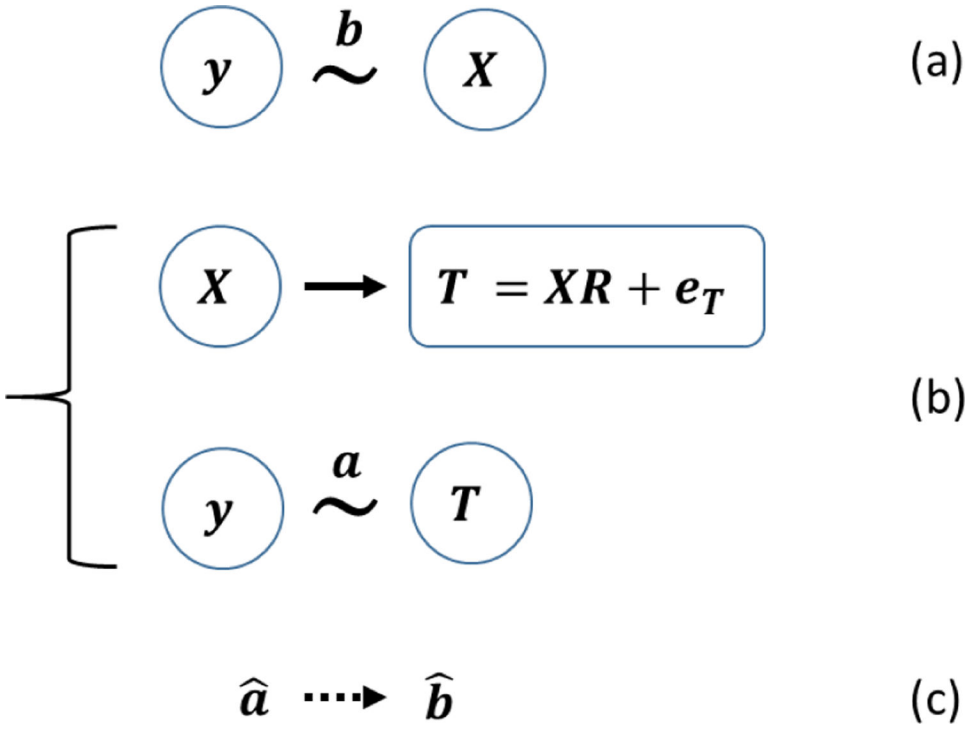

Figure 1:
Diagrams of the concept of the proposed analytic framework.

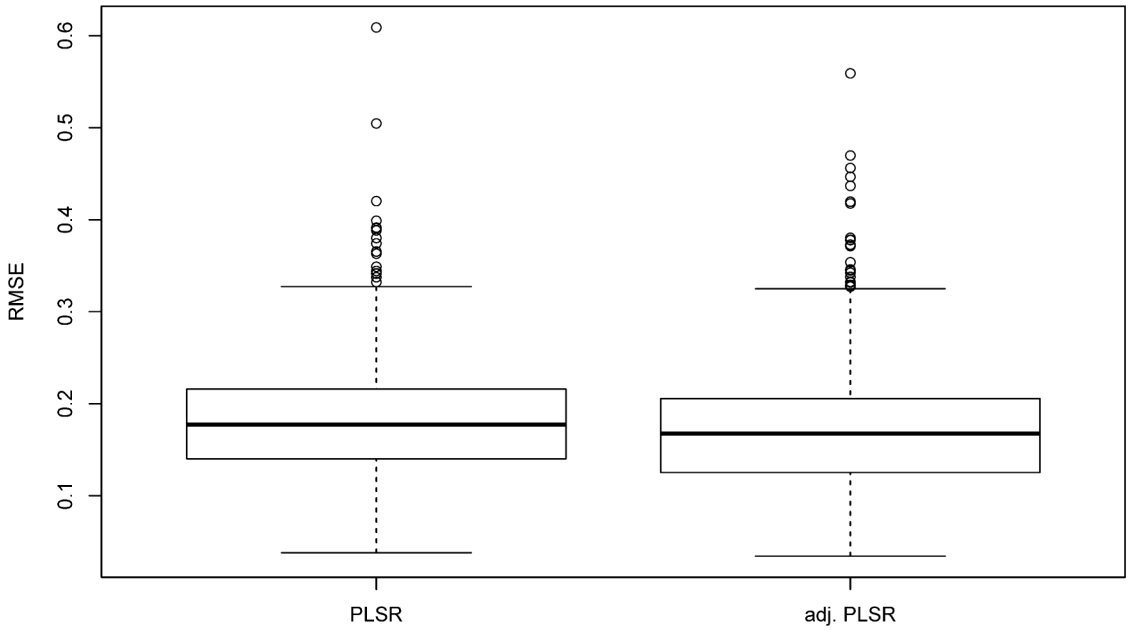

Figure 2.
Boxplot of the RMSE in the estimation of the slope vector ***b*** by PLSR and adjusted PLSR approaches from 1000 simulations for setting 1 ($p = 5$, $k = 2$).

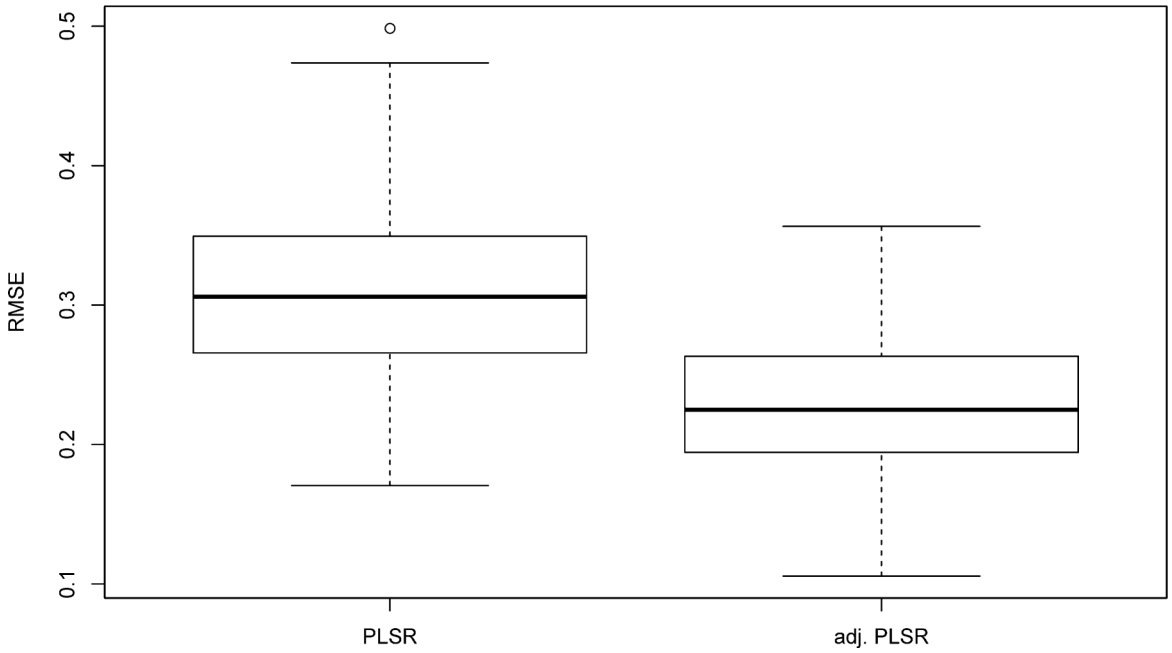

Figure 3.

Boxplot of the RMSE in the estimation of the slope vector ***b*** by PLSR and adjusted PLSR approaches from 1000 simulations for *setting 2* ($p = 25$, $k = 5$).

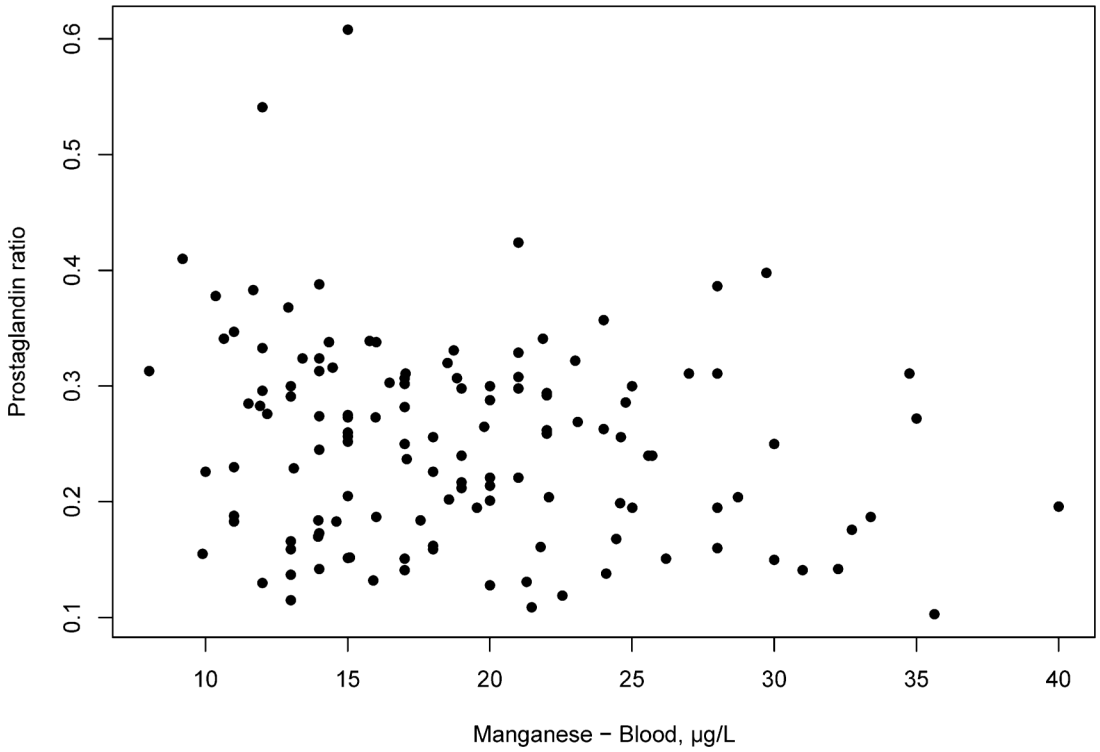

Figure 4:
Scatter plot of the raw prostaglandin ratio versus manganese from blood in oxidative stress dataset.

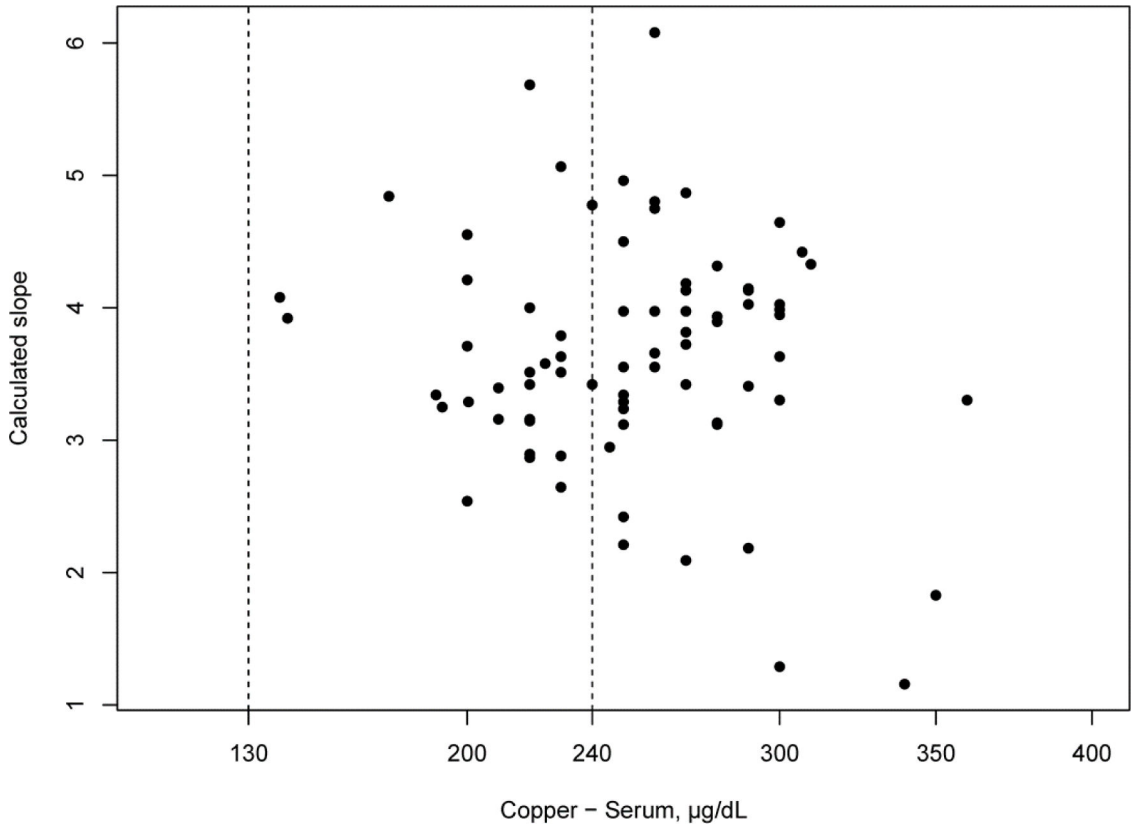

Figure 5:
Scatter plot of the calculated slope with ASQ:I scores of *problem-solving* versus copper in serum in ASQ:I dataset.

Table 1:

Hypothesis testing performance of the adjusted PLSR approach compared with the general PLSR from 1000 simulated datasets for *setting 1* ($p = 5$, $k = 2$).

Setting	Elements in b		Rejection rate	
			PLSR	adj. PLSR
$p = 5, k = 2$	$b_1 = 0.7$	$p \le 0.05$	0.853	0.955
		$p \le 0.1$	0.921	0.974
	$b_2 = 0.3$	$p \le 0.05$	0.172	0.485
		$p \le 0.1$	0.257	0.590
	$b_3 = 0$	$p \le 0.05$	0.028	0.035
		$p \le 0.1$	0.060	0.069
	$b_4 = 1.5$	$p \le 0.05$	0.999	0.999
		$p \le 0.1$	0.999	1
	$b_5 = 1.5$	$p \le 0.05$	0.999	0.998
		$p \le 0.1$	1	0.999

Table 2:

Hypothesis testing performance of the adjusted PLSR approach compared with the general PLSR from 1000 simulations for *setting 2* ($p = 25$, $k = 5$).

		Rejection rate*	
Setting	**Elements in *b***	**PLSR**	**adj. PLSR**
	$b_1 = 0.56$	0.077	0.149
	$b_2 = 0.33$	0.038	0.080
	$b_3 = 0.11$	0.030	0.033
	$b_4 = 0$	0.029	0.020
	$b_5 = 0$	0.028	0.017
	$b_6 = 1.11$	0.190	0.288
	$b_7 = 0.67$	0.083	0.170
	$b_8 = 0.22$	0.030	0.032
	$b_9 = 0$	0.022	0.016
	$b_{10} = 0$	0.025	0.020
	$b_{11} = 1.25$	0.261	0.436
	$b_{12} = 1.25$	0.275	0.410
$p = 25$, $k = 5$	$b_{13} = 0.5$	0.056	0.093
	$b_{14} = 0$	0.034	0.018
	$b_{15} = 0$	0.028	0.023
	$b_{16} = 1.67$	0.492	0.644
	$b_{17} = 1.67$	0.435	0.600
	$b_{18} = 0.67$	0.087	0.140
	$b_{19} = 0$	0.033	0.024
	$b_{20} = 0$	0.021	0.025
	$b_{21} = 1.67$	0.500	0.675
	$b_{22} = 1.67$	0.495	0.674
	$b_{23} = 1.67$	0.508	0.684
	$b_{24} = 0$	0.022	0.020
	$b_{25} = 0$	0.033	0.022

*The cutoff value for testing is set 0.05.

Table 3:

Oxidative stress dataset analysis output by PLSR and adjusted PLSR.

Metal	PLSR		adj. PLSR	
	Test statistic	***P* value**	**Test statistic**	***P* value**
BCD: Cadmium - Blood	0.615	0.540	−0.026	0.910
BMN: Manganese - Blood	−1.708	0.090	−1.834	0.050
BPB: Lead - Blood	1.030	0.305	0.780	0.156
BSE: Selenium - Blood	−0.269	0.788	−2.842	0.650
THG: Mercury Total - Blood	−0.263	0.793	−0.386	0.420
SCU: Copper - Serum	−0.538	0.591	−1.963	0.952
SSE: Selenium - Serum	0.730	0.467	0.245	0.166
UAS3: Arsenous (III) acid - Urine	1.568	0.119	1.497	0.050
UBA: Barium - Urine	−0.811	0.419	−0.570	0.290
UCD: Cadmium - Urine	0.736	0.463	0.734	0.150
UCO: Cobalt - Urine	−0.440	0.661	−0.504	0.374
UCS: Cesium - Urine	1.016	0.311	0.998	0.064
UDMA: Dimethylarsinic Acid - Urine	1.050	0.296	0.755	0.200
UIO: Iodine - Urine	1.195	0.234	0.805	0.316
UMN: Manganese - Urine	−0.730	0.467	−0.630	0.456
UMO: Molybdenum - Urine	0.456	0.649	0.027	0.620
UPB: Lead - Urine	1.022	0.309	0.663	0.208
USB: Antimony - Urine	0.124	0.902	−0.156	0.794
USN: Tin - Urine	−1.241	0.217	−1.011	0.228
USR: Strontium - Urine	−0.829	0.409	−1.350	0.376
UTAS: Arsenic Total - Urine	1.658	0.100	1.685	0.046
UTL: Thallium - Urine	−1.615	0.109	−1.212	0.054
UTU: Tungsten - Urine	1.403	0.163	1.221	0.070
UUR: Uranium – Urine	0.546	0.586	0.597	0.240

Table 4:

ASQ:I dataset analysis of *problem-solving* by PLSR and adjusted PLSR.

Metal*	PLSR		adj. PLSR	
	Test statistic	***P* value**	**Test statistic**	***P* value**
BCD: Cadmium - Blood	−0.269	0.789	0.347	0.518
BMN: Manganese - Blood	−0.008	0.994	0.411	0.680
BPB: Lead - Blood	−0.484	0.630	−1.034	0.128
BSE: Selenium - Blood	0.892	0.375	2.555	0.516
SCU: Copper - Serum	−1.403	0.165	−0.760	0.022
SSE: Selenium - Serum	1.432	0.156	1.785	0.224
SZN: Zinc - Serum	−0.344	0.732	−0.304	0.480
UBA: Barium - Urine	0.261	0.795	0.473	0.850
UCD: Cadmium - Urine	−0.341	0.734	−0.456	0.340
UCO: Cobalt - Urine	0.219	0.828	0.760	0.174
UCS: Cesium - Urine	−0.488	0.627	−0.115	0.824
UDMA: Dimethylarsinic Acid - Urine	−0.346	0.730	0.184	0.686
UIO: Iodine - Urine	0.885	0.379	1.217	0.254
UMN: Manganese - Urine	−0.327	0.744	−0.669	0.234
UMO: Molybdenum - Urine	1.693	0.095	1.988	0.116
UPB: Lead - Urine	−0.494	0.623	−0.953	0.156
USB: Antimony - Urine	−0.441	0.661	−0.475	0.336
USN: Tin - Urine	−0.193	0.847	0.296	0.690
USR: Strontium - Urine	0.010	0.992	−0.115	0.820
UTAS: Arsenic Total - Urine	−0.532	0.597	−0.547	0.216
UTL: Thallium - Urine	−0.383	0.703	−0.145	0.714
UTU: Tungsten - Urine	0.302	0.763	0.040	0.982
UUR: Uranium – Urine	−0.218	0.828	0.365	0.436