VOLUME 1, 2019
DOI: 10.31058/j.adp.2019.11001

Semi-mixture Regression Model for Incomplete Data

Loc Nguyen[1*], Anum Shafiq[2]

[1] Advisory Board, Loc Nguyen's Academic Network, An Giang, Vietnam
[2] Department of Mathematics and Statistics, Preston University Islamabad, Islamabad, Pakistan

Email Address

ng_phloc@yahoo.com (L. Nguyen), anumshafiq@ymail.com (A. Shafiq)
*Correspondence: ng_phloc@yahoo.com

Received: 6 September 2018**; Accepted:** 16 October 2018**; Published:** 29 January 2019

Abstract:

The regression expectation maximization (REM) algorithm, which is a variant of expectation maximization (EM) algorithm, uses parallelly a long regression model and many short regression models to solve the problem of incomplete data. Experimental results proved resistance of REM to incomplete data, in which accuracy of REM decreases insignificantly when data sample is made sparse with loss ratios up to 80%. However, the convergence speed of REM can be decreased if there are many independent variables. In this research, we use mixture model to decompose REM into many partial regression models. These partial regression models are then unified in the so-called semi-mixture regression model. Our proposed algorithm is called semi-mixture regression expectation maximization (SREM) algorithm because it is combination of mixture model and REM algorithm, but it does not implement totally the mixture model. In other words, only mixture coefficients in SREM are estimated according to mixture model whereas regression coefficients are estimated by REM. The experimental results show that SREM converges faster than REM does although the accuracy of SREM is not better than the accuracy of REM in fair tests.

Keywords:

Regression Model, Mixture Regression Model, Expectation Maximization Algorithm, Incomplete Data

1. Introduction

1.1. Main Work

As a convention, regression model is a linear regression function $Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ in which variable Z is called response variable or dependent variable whereas each X_i is called regression variable, regressor, predictor, regression variable, or independent variable. Each α_i is called regression coefficient. The essence of regression analysis is to calculate regression coefficients from data sample. When sample is complete, these coefficients are determined by least squares method [1, pp. 452-458]. When sample is incomplete, there are some approximation approaches to estimate regression coefficients such as complete case method, ad-hoc method, multiple imputation, maximum likelihood, weighting method, and Bayesian method

[2]. We focus on applying expectation maximization (EM) algorithm into constructing regression model in case of missing data with note that EM algorithm belongs to maximum likelihood approach. In previous research [3], we proposed a so-called Regression Expectation Maximization (REM) algorithm to learn linear regression function from incomplete data in which some values of Z and X_i are missing. REM is a variant of EM algorithm, which is used to estimate regression coefficients. Experimental results in previous research [3] proved that accuracy of REM decreases insignificantly whereas loss ratios increase significantly. We hope that REM will be accepted as a new standard method for regression analysis in case of missing data when there are currently 6 standard approaches such as complete case method, ad-hoc method, multiple imputation, maximum likelihood, weighting method, and Bayesian method [2]. Here we combine REM and mixture model to improve convergence speed of REM. Our proposed algorithm is called Semi-mixture Regression Expectation Maximization (SREM) algorithm. Experimental results mentioned later show that SREM converges faster than REM although it is not as accurate as REM. Because this research is the successive one after our previous research [3], they share some common contents related to research survey and experimental design, but we confirm that their methods are not coincide although SREM is derived from REM.

Because SREM is the combination of REM and mixture model whereas REM is a variant of EM algorithm, we need to survey some works related to application of EM algorithm to regression analysis. Kokic [4] proposed an excellent method to calculate expectation of errors for estimating coefficients of multivariate linear regression model. In Kokic's method, response variable Z has missing values. Ghitany, Karlis, Al-Mutairi, and Al-Awadhi [5] calculated the expectation of function of mixture random variable in expectation step of EM algorithm and then used such expectation for estimating parameters of multivariate mixed Poisson regression model in the maximization step. Anderson and Hardin [6] used reject inference technique to estimate coefficients of logistic regression model when response variable Z is missing but characteristic variables (regressors X_i) are fully observed. Anderson and Hardin replaced missing Z by its conditional expectation on regressors X_i where such expectation is logistic function. Zhang, Deng, and Su [7] used EM algorithm to build up linear regression model for studying glycosylated hemoglobin from partial missing data. In other words, Zhang, Deng, and Su [7] aim to discover relationship between independent variables (predictors) and diabetes.

Besides EM algorithm, there are other approaches to solve the problem of incomplete data in regression analysis. Haitovsky [8] stated that there are two main approaches to solve such problem. The first approach is to ignore missing data and to apply the least squares method into observations. The second approach is to calculate covariance matrix of regressors and then to apply such covariance matrix into constructing the system of normal equations. Robins, Rotnitzki, and Zhao [9] proposed a class of inverse probability of censoring weighted estimators for estimating coefficients of regression model. Their approach is based on the dependency of mean vector of response variable Z on vector of regressors X_i when Z has missing values. Robins, Rotnitzki, and Zhao [9] assumed that the probability $\lambda_{it}(\alpha)$ of existence of Z at time point t is dependent on existence of Z at previous time point t–1 but independent from Z. Even though Z is missing, the probability $\lambda_{it}(\alpha)$ is also determined and so regression coefficients are calculated based on the inverse of $\lambda_{it}(\alpha)$ and X_i. The inverse of $\lambda_{it}(\alpha)$ is considered as weight for complete case. Robins, Rotnitzki, and Zhao used additional time-dependent covariates V_{it} to determine $\lambda_{it}(\alpha)$.

In the article "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models", Horton and Kleinman [2] classified 6 methods of regression analysis in case of missing data such as complete case method, ad-hoc method, multiple imputation, maximum likelihood, weighting method, and Bayesian method. EM algorithm belongs to maximum likelihood method. According to complete case method, regression model is learned from only non-missing values of incomplete data [2, p. 3]. The ad-hoc method refers missing values to some common value, creates an indicator of missingness as new variable, and finally builds regression model from both existent variables and such new variable [2, p. 3]. Multiple imputation method has three steps. Firstly, missing values are replaced by possible values. The replacement is repeated until getting an enough number of complete datasets. Secondly, some regression models are learned from these complete datasets as usual [2, p. 4]. Finally, these regression models are aggregated together. The maximum likelihood method aims to construct regression model by maximizing likelihood function. EM algorithm is a variant of maximum likelihood method, which has two steps such as expectation step (E-step) and maximization step (M-step). In E-step, multiple entries are created in an augmented dataset for each observation of missing values and then probability of the observation is estimated based on current parameter [2, p. 6]. In M-step, regression model is built from augmented dataset. The REM algorithm proposed in this research is different from the traditional EM for regression analysis because we replace missing values in E-step by expectation of sufficient statistics via mutual balance process instead of estimating the probability of observation. The weighting method determines the probability of missingness and then uses such probability as weight for the complete case. The aforementioned research of Robins, Rotnitzki, and Zhao [9] belongs to the weighting approach. Instead of replacing missing values by possible values like imputation method does, the Bayesian method imputes missing values by the estimation with a prior distribution on the covariates and the close relationship between the Bayesian approach and maximum likelihood method [2, p. 7].

1.2. Related Studies

Recall that SREM is the combination of REM and mixture model and so we need to survey other works related to regression model with support of mixture model. As a convention, such regression model is called mixture regression model. In literature, there are two approaches of mixture regression model:

- The first approach is to use logistic function to estimate the mixture coefficients.
- The second approach is to construct a joint probability distribution as product of the probability distribution of response variable Z and the probability distribution of independent variables X_i.

According to the first approach [10], the mixture probability distribution is formulated as follows:

$$P(Z|\Theta) = \sum_{k=1}^{K} c_k P_k(Z|\alpha_k^T X, \sigma_k^2) \quad (1)$$

Where $\Theta = (\alpha_k, \sigma_k^2)^T$ is compound parameter whereas α_k and σ_k^2 are regression coefficients and variance of the partial (component) probability distribution $P_k(Z|\alpha_k{}^T X, \sigma_k{}^2)$. Note, mean of $P_k(Z|\alpha_k{}^T X, \sigma_k{}^2)$ is $\alpha_k{}^T X$ and mixture coefficients are c_k. In the first

approach, regression coefficients α_k are estimated by least squares method whereas mixture coefficients are estimated by support of logistic function as follows [10, p. 4]:

$$c_k = \frac{\exp\left(P_k(Z|\alpha_k^T X, \sigma_k^2)\right)}{\sum_{l=1}^{K} \exp\left(P_l(Z|\alpha_l^T X, \sigma_l^2)\right)} \tag{2}$$

The mixture regression model is:

$$\hat{Z} = \sum_{k=1}^{K} c_k \alpha_k^T X \tag{3}$$

According to the second approach, the joint distribution is defined as follows [11, p. 4]:

$$\begin{aligned} P(Z|\Theta) &= \sum_{k=1}^{K} c_k P_k(Z, X|\alpha_k^T X, \sigma_k^2, \mu_k, \Sigma_k) \\ &= \sum_{k=1}^{K} c_k P_k(Z|\alpha_k^T X, \sigma_k^2) P_k(X|\mu_k, \Sigma_k) \end{aligned} \tag{4}$$

Where α_k are regression coefficients and σ_k^2 is variance of the conditional probability distribution $P_k(Z|\alpha_k{}^T X, \sigma_k{}^2)$ whereas μ_k and Σ_k are mean vector and covariance matrix of the prior probability distribution $P_k(X|\, \mu_k, \Sigma_k)$, respectively. The mixture regression model is [11, p. 6]:

$$\hat{Z} = E(Z|X) = \sum_{k=1}^{K} \pi_k \alpha_k^T X \tag{5}$$

Where,

$$\pi_k = \frac{c_k P_k(X|\mu_k, \Sigma_k)}{\sum_{l=1}^{K} c_l P_l(X|\mu_l, \Sigma_l)} \tag{6}$$

The joint probability can be defined by different way as follows [12, p. 21], [13, p. 24], [14, p. 4]:

$$P(Z|\Theta) = \sum_{k=1}^{K} c_k P_k(Z|m_k(X), \sigma_k^2) P_k(X|\mu_{kX}, \Sigma_{kX}) \tag{7}$$

Where $m_k(X)$ and σ_k^2 are mean and variance of Z given the conditional probability distribution $P_k(Z|m_k(X), \sigma_k{}^2)$ whereas μ_{kX} and Σ_{kX} are mean vector and covariance matrix of X given the prior probability distribution $P_k(X|\, \mu_k, \Sigma_k)$. When μ_{kX} and Σ_{kX} are calculated from data, other parameters $m_k(X)$ and σ_k^2 are estimated for each k^{th} component as follows [12, p. 23], [13, p. 25], [14, p. 5]:

$$\begin{aligned} m_k(X) &= \mu_{kZ} + \Sigma_{kZX}\Sigma_{kX}^{-1}(X - \mu_{kX}) \\ \sigma_k^2 &= \Sigma_{kZZ} - \Sigma_{kZX}\Sigma_{kX}^{-1}\Sigma_{kZX} \end{aligned} \tag{8}$$

For each k^{th} component, μ_{kZ} is sample mean of Z, Σ_{kZX} is vector of covariances of Z and X, and Σ_{kZZ} is sample variance of Z. The mixture regression model becomes [13, p. 25]:

$$\hat{Z} = m(X) = \sum_{k=1}^{K} \pi_k m_k(X) \tag{9}$$

Where,

$$\pi_k = \frac{c_k P_k(X|\mu_k, \Sigma_k)}{\sum_{l=1}^{K} c_l P_l(X|\mu_l, \Sigma_l)} \tag{10}$$

Grün & Leisch [15] mentioned the full application of mixture model into regression model in which regression coefficients are determined by inverse function of mean of conditional probability distribution as follows:

$$\begin{aligned} P(Z|\Theta) &= \sum_{k=1}^{K} c_k P_k(Z|\mu_k, \sigma_k^2) \\ g^{-1}(\mu_k) &= \alpha_k^T X \end{aligned} \tag{11}$$

In general, the ideology of combination of regression analysis and mixture model which produces mixture regression is not new, but our proposed SREM is different from other methods in literature because of followings:

- SREM does not use the joint probability distribution. In other words, SREM does not concern the probability distribution of independent variables X_i.
- Variance and mean of the conditional probability $P_k(Z|{\alpha_k}^T X, {\sigma_k}^2)$ in SREM are not estimated by mixture model. They are instead estimated by one-time balance process of REM. SREM also does not use logistic function to estimate mixture coefficients as the first approach does. However, SREM is similar to the first approach most because both SREM and the first approach use the conditional probability distribution to estimate mixture coefficients except that SREM takes advantages of the mean of component probabilities whereas the first approach takes advantages of logistic function.
- SREM does not re-compute mixture coefficients when evaluating regression function.
- Mixture regression models in literature are learned from complete data whereas SREM supports incomplete data.

In general, SREM does not implement totally mixture model because only mixture coefficients in SREM are estimated by the estimation process of mixture model. In this research, we do not compare SREM with other mixture regression methods because the purpose of SREM is different from the purpose of mixture regression model. SREM aims to speed up the convergence of REM in case of missing data whereas mixture regression model aims to improve accuracy of regression analysis in case that data varies complicatedly with many trends. At the first stage of this research, I aim to decompose REM by SREM with hope that SREM is more accurate than REM in fair testing. Unexpectedly, the accuracy of SREM is not better than the accuracy of REM in fair tests but SREM converges faster than REM. Because speed is a significant aspect of an algorithm when data is large, I write this paper as a contribution of SREM. I guesstimate that SREM can be worse than full mixture regression model when data is complete and varies in many trends. On the other hand, full mixture model combined with REM will be better than SREM when data is incomplete and varies in many trends. However, we need an experimental research to

assert this assumption. The methodology of SREM is described in section 2. Section 3 focuses on experimental results. Section 4 is the conclusion.

2. Methodology

The probabilistic Mixture Regression Model (MRM) is a combination of normal mixture model and linear regression model. In MRM, the probabilistic Entire Regression Model (ERM) is sum of K weighted probabilistic Partial Regression Models (PRMs). Equation (12) specifies MRM [16, p. 3].

$$P(z_i|X_i,\Theta)=\sum_{k=1}^{K}c_kP_k(z_i|X_i,\alpha_k,\sigma_k^2) \tag{12}$$

Where,

$$\sum_{k=1}^{K}c_k=1$$

Note, Θ is called entire parameter,

$$\Theta=\left(c_k,\alpha_k^T,\sigma_k^2,\beta_{kj}\right)^T$$

The superscript "T" denotes transposition operator in vector and matrix. In equation (12), the probabilistic distribution $P(z_i|X_i, \Theta)$ represents the ERM where z_i is the response variable, dependent variable, or outcome variable. The probabilistic distribution $P_k(z_i|X_i, \alpha_k, \sigma_k^2)$ represents the k^{th} PRM $z_i = \alpha_{k0} + \alpha_{k1}x_{i1} + \alpha_{k2}x_{i2} + \ldots + \alpha_{kn}x_{in}$ with suppose that each z_i conforms to normal distribution according to equation (13) with mean $\mu_k = \alpha_k^T X_i$ and variance σ_k^2.

$$P_k(z_i|X_i,\alpha_k,\sigma_k^2)=\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(-\frac{(z_i-\alpha_k^TX_i)^2}{2\sigma_k^2}\right) \tag{13}$$

The parameter $\alpha_k = (\alpha_{k0}, \alpha_{k1},\ldots, \alpha_{kn})^T$ is called the k^{th} Partial Regression Coefficient (PRC) and $X_i = (1, x_{i1}, x_{i2},\ldots, x_{in})^T$ is data vector. Each x_{ij} in every PRM is called a regressor, predictor, or independent variable.

In equation (12), each mixture coefficient c_k is the prior probability that any z_i belongs to the k^{th} PRM. Let Y be random variable representing PRMs, $Y = 1, 2,\ldots, K$. The mixture coefficient c_k is also called the k^{th} weight, which is defined by equation (14). Of course, there are K mixture coefficients, K PRMs, and K PRCs.

$$c_k=P(Y=k) \tag{14}$$

For each k^{th} PRM, suppose each $x_{ij} \in X_i$ has an inverse regression model (IRM) $x_{ij} = \beta_{kj0} + \beta_{kj1}z_i$. In other words, x_{ij} now is considered as the random variable conforming to normal distribution according to equation (15) [17, p. 8].

$$P_{kj}\left(x_{ij}\middle|z_i,\beta_{kj}\right)=\frac{1}{\sqrt{2\pi\tau_{kj}^2}}\exp\left(-\frac{\left(x_{ij}-\beta_{kj}^T(1,z_i)^T\right)^2}{2\tau_{kj}^2}\right) \tag{15}$$

Where $\beta_{kj} = (\beta_{kj0}, \beta_{kj1})^T$ is an inverse regression coefficient (IRC) and $(1, z_i)^T$ becomes an inverse data vector. The mean and variance of each x_{ij} with regard to the inverse distribution $P_{kj}(x_{ij}|z_i, \beta_{kj})$ are $\beta_{kj}^T(1, z_i)^T$ and τ_{kj}^2, respectively. Of course, for

each k^{th} PRM, there are n IRMs $P_{kj}(x_{ij}|z_i, \beta_{kj})$ and n associated IRCs β_{kj}. Totally, there are $n*K$ IRMs associated with $n*K$ IRCs.

In this research, we focus on estimating the entire parameter $\Theta = (c_k, \alpha_k, \sigma_k^2, \beta_{kj})^T$ where k is from 1 to K. In other words, we aim to estimate c_k, α_k, σ_k^2, and β_{kj} for determining the ERM in case of missing data. As a convention, let $\Theta^* = (c_k^*, \alpha_k^*, (\sigma_k^2)^*, \beta_{kj}^*)^T$ be the estimate of $\Theta = (c_k, \alpha_k, \sigma_k^2, \beta_{kj})^T$, respectively. Let $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{Z})$ be collected sample in which $\boldsymbol{X}$ is a set of regressors and $\boldsymbol{Z}$ is a set of outcome variables plus values 1, respectively [17, p. 8] with note that both $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incomplete. In other words, $\boldsymbol{X}$ and $\boldsymbol{Z}$ have missing values. As a convention, let z_i^- and x_{ij}^- denote missing values of $\boldsymbol{Z}$ and $\boldsymbol{X}$, respectively.

$$\boldsymbol{X} = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix}$$
$$X_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix}, X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{pmatrix} \quad (16)$$
$$Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}, \boldsymbol{Z} = (\boldsymbol{1}, Z) = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_N \end{pmatrix}$$

The expectation of sufficient statistic z_i regard to the k^{th} PRM $P_k(z_i|X_i, \alpha_k, \sigma_k^2)$ is specified by equation (17) [3].

$$E_k(z_i|X_i) = \alpha_k^T X_i = \sum_{j=0}^{n} \alpha_{kj} x_{ij} \quad (17)$$

Where x_{i0}=1 for all i. The expectation of the sufficient statistic x_{ij} with regard to each IRM $P_{kj}(x_{ij}|z_i, \beta_j)$ of the k^{th} PRM $P_k(z_i|X_i, \alpha_k, \sigma_k^2)$ is specified by equation (18) [3].

$$E_k(x_{ij}|z_i) = \beta_{kj}^T (1, z_i)^T = \beta_{kj0} + \beta_{kj1} z_i \quad (18)$$

Please pay attention to equations (17) and (18) because missing values of data $\boldsymbol{X}$ and data $\boldsymbol{Z}$ will be estimated by these expectations later. By applying sample $\boldsymbol{D}$ into equations (12) and (13) and using maximum likelihood estimation (MLE) method [17, pp. 8-9], we retrieve equation (19) to estimate α_k^*, β_{kj}^* [1, p. 457], and $(\sigma_k^2)^*$ for each k^{th} PRM where $\boldsymbol{X}$, $\boldsymbol{Z}$, Z, X_i, and X_j are specified in equation (16). Appendix A1 is the proof of equation (19).

$$\begin{aligned} \alpha_k^* &= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Z \\ \beta_{kj}^* &= (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T X_j \\ (\sigma_k^2)^* &= \frac{1}{N} \sum_{i=1}^{N} (z_i - (\alpha_k^*)^T X_i)^2 \end{aligned} \quad (19)$$

From sample $\boldsymbol{D}$, the optimal regression coefficients (α_k^*, $(\sigma_k^2)^*$) and β_{kj}^* estimated by equation (19) whereas the optimal mixture coefficient c_k^* for each k^{th} PRM is estimated by equation (20) as follows [16, p. 7]:

$$c_k^* = \frac{1}{N}\sum_{i=1}^{N} P(Y = k|z_i, X_i, \alpha_k^*, (\sigma_k^2)^*) \quad (20)$$

Where [16, p.3],

$$P(Y = k|z_i, X_i, \alpha_k^*, (\sigma_k^2)^*) = \frac{c_k P_k(z_i|X_i, \alpha_k^*, (\sigma_k^2)^*)}{\sum_l^K c_l P_l(z_i|X_i, \alpha_k^*, (\sigma_k^2)^*)} \quad (21)$$

Note, each optimal PRM $P_k(z_i|X_i, \alpha_k^*, (\sigma_k^2)^*)$ is determined by equation (13).

$$P_k(z_i|X_i, \alpha_k^*, (\sigma_k^2)^*) = \frac{1}{\sqrt{2\pi(\sigma_k^2)^*}} \exp\left(-\frac{(z_i - (\alpha_k^*)^T X_i)^2}{2(\sigma_k^2)^*}\right)$$

Because $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incomplete, we apply expectation maximization (EM) algorithm into estimating $\Theta^* = (c_k^*, \alpha_k^*, (\sigma_k^2)^*, \beta_{kj}^*)^T$. According to [18], EM algorithm has many iterations and each iteration has expectation step (E-step) and maximization step (M-step) for estimating parameters. Given current parameter $\Theta^{(t)} = (c_k^{(t)}, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}, \beta_{kj}^{(t)})^T$ at the t^{th} iteration, missing values z_i^- and x_{ij}^- are calculated in E-step so that $\boldsymbol{X}$ and $\boldsymbol{Z}$ become complete. In M-step, the next parameter $\Theta^{(t+1)} = (c_k^{(t+1)}, \alpha_k^{(t+1)}, (\sigma_k^2)^{(t+1)}, \beta_{kj}^{(t+1)})^T$ is determined by equations (19) and (20) and the complete data $\boldsymbol{X}$ and $\boldsymbol{Z}$.

The most important problem in our research is how to estimate missing values z_i^- and x_{ij}^-. Recall that, for each k^{th} PRM, every missing value z_i^- is estimated as the expectation based on the current parameter $\alpha_k^{(t)}$, according to equation (17) [3].

$$z_i^- = E_k(z_i|X_i) = \left(\alpha_k^{(t)}\right)^T X_i = \sum_{j=0}^{n} \alpha_{kj}^{(t)} x_{ij}$$

Note, $x_{i0} = 1$. Let U_i be a set of indices of missing values x_{ij}^- with fixed i for each k^{th} PRM. In other words, if $j \in U_i$ then, x_{ij} is missing. The set U_i can be empty. The equation (17) is re-written for each k^{th} PRM as follows [3]:

$$z_i^- = \sum_{j \in U_i} \alpha_{kj}^{(t)} x_{ij}^- + \sum_{l \notin U_i} \alpha_{kl}^{(t)} x_{il}$$

According to equation (18), missing value x_{ij}^- is estimated by [3]:

$$x_{ij}^- = E_k(x_{ij}|z_i^-) = \left(\beta_{kj}^{(t)}\right)^T (1, z_i^-)^T = \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)} z_i^-$$

Combining equation (17) and equation (18), we have [3]:

$$\begin{aligned} z_i^- &= \sum_{j \in U_i} \alpha_{kj}^{(t)} \left(\beta_{kj0}^{(t)} + \beta_{kj1}^{(t)} z_i^-\right) + \sum_{l \notin U_i} \alpha_{kl}^{(t)} x_{il} \\ &= z_i^- \sum_{j \in U_i} \alpha_{kj}^{(t)} \beta_{kj1}^{(t)} + \sum_{j \in U_i} \alpha_{kj}^{(t)} \beta_{kj0}^{(t)} + \sum_{l \notin U_i} \alpha_{kl}^{(t)} x_{il} \end{aligned}$$

It implies [3]:

$$z_i^- = \frac{\sum_{j \in U_i} \alpha_{kj}^{(t)} \beta_{kj0}^{(t)} + \sum_{l \notin U_i} \alpha_{kl}^{(t)} x_{il}}{1 - \sum_{j \in U_i} \alpha_{kj}^{(t)} \beta_{kj1}^{(t)}}$$

As a result, equation (22) is used to estimate or fulfill missing values for each k^{th} PRM [3].

$$z_i^- = \frac{\sum_{j\in U_i} \alpha_{kj}^{(t)} \beta_{kj0}^{(t)} + \sum_{l\notin U_i} \alpha_{kl}^{(t)} x_{il}}{1 - \sum_{j\in U_i} \alpha_{kj}^{(t)} \beta_{kj1}^{(t)}}$$
$$x_{ij}^- = \begin{cases} \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)} z_i \text{ if } z_i \text{ is not missing} \\ \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)} z_i^- \text{ if } z_i \text{ is missing} \end{cases} \quad (22)$$

In previous research, we proposed a so-called Regression Expectation Maximization (REM) which is a variant of EM algorithm for estimating α_k^* and β_{kj}^*. Equation (22) is used in the E-step of REM to fulfill missing values. However, REM does not support mixture model. Here we proposed a so-called Semi-mixture Regression Expectation Maximization (SREM) which is a variant of REM, in which M-step is modified to calculate the optimal mixture coefficient c_k^*. SREM is described in Table 1. We will explain later why SREM does not conform fully to mixture model although it supports mixture model.

Table 1. *Semi-mixture Regression Expectation Maximization (SREM) Algorithm.*

1. E-step: Missing values z_i^- and x_{ij}^- for each k^{th} PRM are fulfilled by equation (22) given current parameter $\Theta^{(t)}$.

$$z_i^- = \frac{\sum_{j\in U_i} \alpha_{kj}^{(t)} \beta_{kj0}^{(t)} + \sum_{l\notin U_i} \alpha_{kl}^{(t)} x_{il}}{1 - \sum_{j\in U_i} \alpha_{kj}^{(t)} \beta_{kj1}^{(t)}}$$
$$x_{ij}^- = \begin{cases} \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)} z_i \text{ if } z_i \text{ is not missing} \\ \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)} z_i^- \text{ if } z_i \text{ is missing} \end{cases}$$

2. M-step: The next parameter $\Theta^{(t+1)}$ is determined by equations (19), (20), and (21) and the complete data ($\boldsymbol{X}_k$, $\boldsymbol{Z}_k$) fulfilled in E-step. Please pay attention that each k^{th} PRM owns a particular complete data ($\boldsymbol{X}_k$, $\boldsymbol{Z}_k$). In other words, original sample ($\boldsymbol{X}$, $\boldsymbol{Z}$) has K complete versions ($\boldsymbol{X}_k$, $\boldsymbol{Z}_k$) fulfilled in E-step for K PRMs. Note, such K complete versions are changed over each iteration.

$$\alpha_k^{(t+1)} = (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T Z_k$$
$$\beta_{kj}^{(t+1)} = (\boldsymbol{Z}_k^T \boldsymbol{Z}_k)^{-1} \boldsymbol{Z}_k^T X_{kj}$$
$$(\sigma_k^2)^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \left(z_{ki} - \left(\alpha_k^{(t+1)}\right)^T X_{ki} \right)^2$$
$$c_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} P\left(Y = k \middle| z_{ki}, X_{ki}, \alpha_k^{(t+1)}, (\sigma_k^2)^{(t+1)}\right)$$

Where,

$$P\left(Y = k \middle| z_{ki}, X_{ki}, \alpha_k^{(t+1)}, (\sigma_k^2)^{(t+1)}\right) = \frac{c_k^{(t)} P_k\left(z_{ki} \middle| X_{ki}, \alpha_k^{(t+1)}, (\sigma_k^2)^{(t+1)}\right)}{\sum_l^K c_l^{(t)} P_l\left(z_{ki} \middle| X_{ki}, \alpha_k^{(t+1)}, (\sigma_k^2)^{(t+1)}\right)}$$

$$P_k\left(z_{ki} \middle| X_{ki}, \alpha_k^{(t+1)}, (\sigma_k^2)^{(t+1)}\right) = \frac{1}{\sqrt{2\pi (\sigma_k^2)^{(t+1)}}} \exp\left(- \frac{\left(z_{ki} - \left(\alpha_k^{(t+1)}\right)^T X_{ki} \right)^2}{2(\sigma_k^2)^{(t+1)}} \right)$$

Note that Z_k is Z that belongs to $\boldsymbol{Z}_k$, X_{ki} is X_i that belongs to $\boldsymbol{X}_k$, X_{kj} is X_j that belongs to $\boldsymbol{X}_k$, and z_{ki} is z_i that belongs to Z_k. The next parameter $\Theta^{(t+1)}$ becomes current parameter in the next iteration.

EM algorithm stops if at some t^{th} iteration, we have $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$. At that time, $\Theta^* = (c_k^*, \alpha_k^*, (\sigma_k^2)^*, \beta_{kj}^*)$ is the optimal estimate of EM algorithm. Note, $\Theta^{(1)}$ at the first iteration is initialized arbitrarily. Here SREM stops if ratio deviation between $\Theta^{(t)}$ and $\Theta^{(t+1)}$ is smaller than a small enough terminated threshold $\varepsilon > 0$ or SREM reaches a large enough number of iterations. The smaller the terminated threshold is, the more accurate SREM is. SREM uses both the terminated threshold $\varepsilon = 0.1\% = 0.001$ and the maximum number of iterations (10000). The maximum number of iterations prevents SREM from running for a long time.

In traditional Gaussian mixture model, variances $(\sigma_k^2)^{(t+1)}$ and means $\mu_k^{(t+1)}$ are estimated by different way based on $c_k^{(t)}$ and PRMs. Therefore, our model is called semi-mixture regression model when only $c_k^{(t+1)}$ is estimated by PRMs. The reason is that $(\sigma_k^2)^{(t+1)}$ and $\alpha_k^{(t+1)}$ were optimized by maximum likelihood estimation (MLE) method and it may be overfitting or redundant to re-estimate $(\sigma_k^2)^{(t+1)}$ and $\alpha_k^{(t+1)}$ by Gaussian mixture model. As a result, we save computation cost by estimating $(\sigma_k^2)^*$ and c_k^* after EM process finished. In other words, $(\sigma_k^2)^{(t+1)}$ and $c_k^{(t+1)}$ are not re-computed many times at E-step of every iteration and so $(\sigma_k^2)^*$ and c_k^* are computed only one time after EM process finished, according to equation (23).

$$\begin{aligned}
(\sigma_k^2)^* &= \frac{1}{N}\sum_{i=1}^{N}(z_{ki} - (\alpha_k^*)^T X_{ki})^2 \\
c_k^* &= \frac{1}{N}\sum_{i=1}^{N} P(Y = k|z_{ki}, X_{ki}, \alpha_k^*, (\sigma_k^2)^*) \\
P(Y = k|z_{ki}, X_{ki}, \alpha_k^*, (\sigma_k^2)^*) &= \frac{P_k(z_{ki}|X_{ki}, \alpha_k^*, (\sigma_k^2)^*)}{\sum_l^K P_l(z_{ki}|X_{ki}, \alpha_k^*, (\sigma_k^2)^*)}
\end{aligned} \tag{23}$$

Note that Z_k is Z that belongs to $\boldsymbol{Z}_k$, X_{ki} is X_i that belongs to $\boldsymbol{X}_k$, and z_{ki} is z_i that belongs to Z_k where $(\boldsymbol{X}_k, \boldsymbol{Z}_k)$ is owned by the k^{th} PRM, which the k^{th} version of the original sample $(\boldsymbol{X}, \boldsymbol{Z})$.

We use the complete case method mentioned in [2, p. 3] to improve the convergence of SREM. The parameters $(\alpha_k^{(1)}, \beta_{kj}^{(1)})^T$ at the first iteration of EM process are initialized in proper way instead that they are initialized in arbitrary way [19]. Let $\boldsymbol{X}_k$' be the complete matrix, which is created by removing all rows whose values are missing from $\boldsymbol{X}_k$. Similarly, let $\boldsymbol{Z}_k$' be the complete matrix, which is created by removing rows whose weights are missing from $\boldsymbol{Z}_k$. The advanced parameters $(\alpha_k^{(1)}, \beta_{kj}^{(1)})^T$ are initialized by equation (24).

$$\begin{aligned}
\alpha_k^{(1)} &= ((\boldsymbol{X}_k')^T \boldsymbol{X}_k')^{-1} (\boldsymbol{X}_k')^T Z_k' \\
\beta_{kj}^{(1)} &= ((\boldsymbol{Z}_k')^T \boldsymbol{Z}_k')^{-1} (\boldsymbol{Z}_k')^T X_{kj}'
\end{aligned} \tag{24}$$

Where Z_k' is the complete vector of non-missing outcome values for each k^{th} PRM and X_{kj}' is the complete column vector of non-missing regressor values for each k^{th} PRM. Equation (24) is a variant of equation (19) where $\boldsymbol{X}_k$, $\boldsymbol{Z}_k$, X_{kj}, and Z_k are replaced by $\boldsymbol{X}_k$', $\boldsymbol{Z}_k$', X_{kj}', and Z_k'.

The evaluation of SREM is different from traditional regression model. It follows mixture model. For example, given input data vector $X_0 = (x_{01}, x_{02}, \ldots, x_{0n})$, let $z_1, z_2, \ldots, z_K$ are values evaluated from K PRMs, we have:

$$z_k = (\alpha_k^*)^T X_0 = \sum_{j=0}^{n} \alpha_{kj}^* x_{0j}$$

Where $x_{00} = 1$. The final evaluation z is calculated based on mixture coefficients as seen in equation (25).

$$z = \sum_{k=1}^{K} c_k^* z_k = \sum_{k=1}^{K} c_k^* (\alpha_k^*)^T X_0 = \sum_{k=1}^{K} c_k^* \sum_{j=0}^{n} \alpha_{kj}^* x_{0j} \quad (25)$$

Following is the proof of equation (25). From equation (12), let $\hat{z}$ be the estimate of response variable z, we have:

$$\hat{z} = E\left(z \middle| P(z|X, \Theta^*)\right) = \sum_{k=1}^{K} c_k^* E_k\left(z \middle| P_k(z|X, \alpha_k^*, (\sigma_k^2)^*)\right) = \sum_{k=1}^{K} c_k^* (\alpha_k^*)^T X \blacksquare$$

Equation (25) is the semi-mixture regression model where mixture coefficients α_{kj}^* are resulted from the EM process of SREM shown in Table 1 and c_k^* is calculated by equation (23). Note, semi-mixture regression model does not re-compute mixture coefficients c_k^* when evaluating z from X_0. In other words, after SREM finished, c_k^* are fixed.

3. Results and Discussions

We use two data samples for testing SREM. The first one is the gestational dataset of 1027 cases in which each case includes ultrasound measures (regressors) and fetus weight (response variable). Ultrasound measures are bi-parietal diameter (*bpd*), head circumference (*hc*), abdominal circumference (*ac*), and fetal length (*fl*). The unit of *bpd*, *hc*, *ac*, and *fl* is millimeter whereas the unit of fetal weight is gram. Ho and Phan [20], [21] collected the ultrasound measure sample of pregnant women at Vinh Long General Hospital – Vietnam with obeying strictly all medical ethical criteria. These women and their husbands are Vietnamese. Their periods are regular and their last periods are determined. Each of them has only one alive fetus. Fetal age is from 28 weeks to 42 weeks. Delivery time is not over 48 hours since ultrasound scan.

The second sample is the dataset which contains 9568 data points collected from a Combined Cycle Power Plant (CCPP) [22]. Regressors in CCPP dataset are hourly average Ambient Temperature (*AT*), Ambient Pressure (*AP*), Relative Humidity (*RH*) and Exhaust Vacuum (*V*) to predict the net hourly electrical energy output (PE) as response variable. *AT* is in the range 1.81 ℃ and 37.11 ℃. *AP* is in the range 992.89-1033.30 millibar. *RH* is in the range 25.56% to 100.16%. *V* is in the range 25.36-81.56 cm Hg. *PE* is in the range 420.26-495.76 MW.

In general, we have two samples such as gestational sample and CCPP sample. The dataset is split separately into one training dataset (50% sample) and one testing dataset (50% sample). Later on, the training dataset is made sparse with loss ratios 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%, which is similar to our previous research [19]. Missing values are made randomly regardless of regressors or response variable. For example, the gestational training dataset (50% gestational sample) has 50%*1027 ≈ 513 rows and each row has 5 columns (*bpd*, *hc*, *ac*, *fl*, *weight*) and so the training dataset has 513*5 = 2565 cells. If loss ratio is 10%, there are only 10%*2565 ≈ 256 missing values which are made randomly among such 2565 cells. In other words, the incomplete training dataset with loss ratio 10% has 2565 –

256 = 2309 non-missing values. Of course, the testing dataset (50% sample) is not made sparse. Each pair of incomplete training dataset and testing dataset is called testing pair. There are ten testing pairs for each sample. As a convention, the origin testing pair which has no missing value in training dataset is the 0^{th} pair. The 0^{th} pair is called complete pair whereas the 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, 6^{th}, 7^{th}, 8^{th}, and 9^{th} pairs are called incomplete pairs.

Firstly, we test SREM with gestational sample. Table 2 [19] shows ten testing pairs of gestational sample.

Table 2. *Ten testing pairs of gestational sample.*

Pair	Training dataset	Testing dataset	Loss ratio
0	Ges.sample.base	Ges.sample.test	0%
1	Ges.sample.base.0.1.miss	Ges.sample.test	10%
2	Ges.sample.base.0.2.miss	Ges.sample.test	20%
3	Ges.sample.base.0.3.miss	Ges.sample.test	30%
4	Ges.sample.base.0.4.miss	Ges.sample.test	40%
5	Ges.sample.base.0.5.miss	Ges.sample.test	50%
6	Ges.sample.base.0.6.miss	Ges.sample.test	60%
7	Ges.sample.base.0.7.miss	Ges.sample.test	70%
8	Ges.sample.base.0.8.miss	Ges.sample.test	80%
9	Ges.sample.base.0.9.miss	Ges.sample.test	90%

SREM may be better than REM if SREM has a large enough number of PRMs and each PRM has many enough regressors. Thus, for fair testing, the number of PRMs in SREM is equal to the number of regressors and each PRM has only one regressor. Table 3 shows ten resulted regression models of REM corresponding to ten testing pairs, given gestational sample.

Table 3. *Ten resulted regression models of REM given gestational sample.*

Pair	Regression model
0	*weight* = -5686.8907 + 46.2369*(*bpd*) + 1.7148*(*hc*) + 14.3173*(*fl*) + 9.3881*(*ac*)
1	*weight* = -5685.7848 + 43.1103*(*bpd*) + 1.4912*(*hc*) + 17.0387*(*fl*) + 9.8929*(*ac*)
2	*weight* = -5853.1212 + 39.5619*(*bpd*) + 2.4174*(*hc*) + 21.7261*(*fl*) + 9.5005*(*ac*)
3	*weight* = -6198.2399 + 44.6901*(*bpd*) + 5.2472*(*hc*) + 20.4527*(*fl*) + 6.6325*(*ac*)
4	*weight* = -5941.9821 + 39.9089*(*bpd*) + 2.6238*(*hc*) + 23.3260*(*fl*) + 9.2312*(*ac*)
5	*weight* = -6496.2424 + 44.6131*(*bpd*) + 3.9980*(*hc*) + 25.8861*(*fl*) + 7.7759*(*ac*)
6	*weight* = -5940.9170 + 31.6952*(*bpd*) + 2.8293*(*hc*) + 34.1356*(*fl*) + 9.0107*(*ac*)
7	*weight* = -6296.7603 + 66.8602*(*bpd*) + 2.7111*(*hc*) + 16.8848*(*fl*) + 4.0660*(*ac*)
8	*weight* = -5362.1163 + 35.6642*(*bpd*) + 4.7398*(*hc*) + 14.8123*(*fl*) + 8.2385*(*ac*)
9	*weight* = -5923.3220 + 87.5165*(*bpd*) + 3.4471*(*hc*) - 0.2822*(*fl*) - 0.0753*(*ac*)

Table 4 shows ten resulted semi-mixture regression models of SREM corresponding to ten testing pairs, given gestational sample.

Table 4. *Ten resulted semi-mixture regression models of SREM given gestational sample.*

Pair	Semi-mixture regression model
0	{*weight* = -6651.5534 + 108.5531*(*bpd*): coeff=0.2721, var=113888.6649}, {*weight* = -4986.7292 + 24.6736*(*hc*): coeff=0.2041, var=188973.6069}, {*weight* = -4505.6926 + 109.6790*(*fl*): coeff=0.2450, var=119971.2307}, {*weight* = -3385.5925 + 19.4249*(*ac*): coeff=0.2788, var=97458.0445}
1	{*weight* = -6802.9586 + 110.3231*(*bpd*): coeff=0.2700, var=98865.5195}, {*weight* = -5089.2989 + 25.0105*(*hc*): coeff=0.2012, var=163173.8380}, {*weight* = -4744.7739 + 113.4482*(*fl*): coeff=0.2426, var=103291.1775}, {*weight* = -3515.6183 + 19.8394*(*ac*): coeff=0.2862, var=77538.8650}

2	{*weight* = -6977.8017 + 112.5302*(*bpd*): coeff=0.2628, var=86798.8614}, {*weight* = -5312.2016 + 25.7209*(*hc*): coeff=0.2000, var=138635.9025}, {*weight* = -4866.0218 + 115.1670*(*fl*): coeff=0.2514, var=78089.6011}, {*weight* = -3615.3155 + 20.1265*(*ac*): coeff=0.2858, var=61080.1223}
3	{*weight* = -7044.9992 + 113.0877*(*bpd*): coeff=0.2850, var=49040.4530}, {*weight* = -5765.6434 + 27.0933*(*hc*): coeff=0.2164, var=74811.7832}, {*weight* = -4850.8460 + 114.8885*(*fl*): coeff=0.2321, var=60022.4336}, {*weight* = -3639.0104 + 20.2068*(*ac*): coeff=0.2665, var=47595.5654}
4	{*weight* = -7176.0173 + 115.2133*(*bpd*): coeff=0.2716, var=38495.8850}, {*weight* = -5580.7794 + 26.5639*(*hc*): coeff=0.1939, var=71575.3627}, {*weight* = -5143.9012 + 119.9590*(*fl*): coeff=0.2319, var=46756.6663}, {*weight* = -3824.3390 + 20.8679*(*ac*): coeff=0.3026, var=29724.0337}
5	{*weight* = -7660.3431 + 120.8204*(*bpd*): coeff=0.2693, var=30819.8738}, {*weight* = -6110.0704 + 28.2196*(*hc*): coeff=0.2138, var=48373.8766}, {*weight* = -5331.1994 + 122.4455*(*fl*): coeff=0.2369, var=36807.6503}, {*weight* = -3967.5178 + 21.3295*(*ac*): coeff=0.2800, var=27240.5556}
6	{*weight* = -8097.3745 + 125.7068*(*bpd*): coeff=0.2302, var=22289.0842}, {*weight* = -7015.6149 + 31.3566*(*hc*): coeff=0.2103, var=28635.3775}, {*weight* = -5480.6406 + 125.3284*(*fl*): coeff=0.2674, var=13952.3164}, {*weight* = -3676.3555 + 20.3238*(*ac*): coeff=0.2920, var=11540.1306}
7	{*weight* = -7076.9202 + 112.8536*(*bpd*): coeff=0.3705, var=3375.2380}, {*weight* = -5497.9202 + 26.2185*(*hc*): coeff=0.1612, var=18787.3282}, {*weight* = -4947.5898 + 117.8865*(*fl*): coeff=0.2113, var=9967.0914}, {*weight* = -3653.8140 + 20.3827*(*ac*): coeff=0.2569, var=8618.4241}
8	{*weight* = -7018.2030 + 112.6524*(*bpd*): coeff=0.2678, var=3654.3436}, {*weight* = -5235.5481 + 25.2899*(*hc*): coeff=0.2162, var=5459.5803}, {*weight* = -5647.3688 + 127.7972*(*fl*): coeff=0.2054, var=5974.4689}, {*weight* = -3285.2965 + 19.3967*(*ac*): coeff=0.3106, var=2526.1926}
9	{*weight* = -6350.5284 + 104.5601*(*bpd*): coeff=0.1787, var=204.7618}, {*weight* = -5140.6601 + 24.4881*(*hc*): coeff=0.0745, var=1245.6621}, {*weight* = -6791.1342 + 152.4635*(*fl*): coeff=0.3553, var=68.6443}, {*weight* = -3831.9687 + 21.4992*(*ac*): coeff=0.3915, var=53.0970}

In Table 4, each PRM is wrapped in two brackets "{.}". Notation "coeff" denotes mixture coefficient and notation "var" denotes the variance of a PRM. For explanation, the 1th regression model is interpreted according to equation (25) as follows: *weight* = 0.2700*(-6802.9586 + 110.3231*(*bpd*)) + 0.2012*(-5089.2989 + 25.0105*(*hc*)) + 0.2426*(-4744.7739 + 113.4482*(*fl*)) + 0.2862*(-3515.6183 + 19.8394*(*ac*)) = -5018.02 + 5.6780(*ac*) + 29.7872(*bpd*) + 27.5225(*fl*) + 5.0321(*hc*).

Given gestational sample, we compare SREM with REM given with regard to the ratio mean absolute error (*RMAE*) and the number *t* of iterations. The number *t* reflects speed of an algorithm. The smaller the number *t* is, the faster the algorithm is. Let $W = \{w_1, w_2, \ldots, w_K\}$ and $V = \{v_1, v_2, \ldots, v_K\}$ be sets of actual weights and estimated weights, respectively. Equation (26) specifies the *RMAE* metric [23, p. 814].

$$RMAE = \frac{1}{K}\sum_{i=1}^{K}\left|\frac{v_i - w_i}{w_i}\right| \tag{26}$$

The smaller the *RMAE* is, the more accurate the algorithm is. Table 5 is the comparison of REM and SREM with regard to *RMAE* and *t* given gestational sample.

Table 5. *Comparison of REM and SREM regarding RMAE and t, given gestational sample.*

Pair	*RMAE* (REM)	*RMAE* (SREM)	*t* (REM)	*t* (SREM)
0	0.0711	0.0786	1	2

1	0.0722	0.0759	4	4
2	0.0739	0.0738	6	4
3	0.0724	0.0720	7	4
4	0.0746	0.0727	11	5
5	0.0780	0.0721	18	5
8	0.0777	0.0745	22	4
7	0.0709	0.0706	37	5
8	0.0729	0.0752	112	4
9	0.0853	0.1147	444	4
Average	0.0749	0.0780	66.2	4.1

From Table 5, given gestational sample, SREM is faster than REM according to *t* but the accuracy of REM is better than the accuracy of SREM according to *RMAE*. Note [19], values of paired t-test statistic t_0 [1, p. 376] of *RMAE* for REM and SREM are 5.3294 and 6.4541, respectively. Because all these values are larger than the percentage point $t_{0.05,8}$ = 1.860 [1, p. 711] given significant level 95%, the resistance of REM and SREM to missing values given gestational sample is proved.

We continue to test SREM with CCPP sample. Table 6 shows ten testing pairs of CCPP sample.

Table 6. *Ten testing pairs of CCPP sample.*

Pair	Training dataset	Testing dataset	Loss ratio
0	CCPP.sample.base	CCPP.sample.test	0%
1	CCPP.sample.base.0.1.miss	CCPP.sample.test	10%
2	CCPP.sample.base.0.2.miss	CCPP.sample.test	20%
3	CCPP.sample.base.0.3.miss	CCPP.sample.test	30%
4	CCPP.sample.base.0.4.miss	CCPP.sample.test	40%
5	CCPP.sample.base.0.5.miss	CCPP.sample.test	50%
6	CCPP.sample.base.0.6.miss	CCPP.sample.test	60%
7	CCPP.sample.base.0.7.miss	CCPP.sample.test	70%
8	CCPP.sample.base.0.8.miss	CCPP.sample.test	80%
9	CCPP.sample.base.0.9.miss	CCPP.sample.test	90%

Table 7 shows ten resulted regression models of REM corresponding to ten testing pairs, given CCPP sample.

Table 7. *Ten resulted regression models of REM given CCPP sample.*

Pair	Regression model
0	*PE* = 469.7296 - 1.9885*(*AT*) - 0.2332*(*V*) + 0.0474*(*AP*) - 0.1602*(*RH*)
1	*PE* = 415.9687 - 1.9131*(*AT*) - 0.2579*(*V*) + 0.0979*(*AP*) - 0.1272*(*RH*)
2	*PE* = 416.5671 - 1.8401*(*AT*) - 0.2940*(*V*) + 0.0963*(*AP*) - 0.1047*(*RH*)
3	*PE* = 401.8042 - 1.8324*(*AT*) - 0.2999*(*V*) + 0.1099*(*AP*) - 0.0869*(*RH*)
4	*PE* = 369.4165 - 1.7559*(*AT*) - 0.3281*(*V*) + 0.1410*(*AP*) - 0.0789*(*RH*)
5	*PE* = 346.6202 - 1.7208*(*AT*) - 0.3237*(*V*) + 0.1615*(*AP*) - 0.0633*(*RH*)
6	*PE* = 341.1562 - 1.6900*(*AT*) - 0.3300*(*V*) + 0.1647*(*AP*) - 0.0383*(*RH*)
7	*PE* = 346.4257 - 1.6501*(*AT*) - 0.3776*(*V*) + 0.1618*(*AP*) - 0.0467*(*RH*)
8	*PE* = 302.7665 - 1.5758*(*AT*) - 0.3174*(*V*) + 0.1942*(*AP*) + 0.0391*(*RH*)
9	*PE* = 564.1434 - 2.1327*(*AT*) + 0.0188*(*V*) - 0.0684*(*AP*) + 0.0205*(*RH*)

Table 8 shows ten resulted semi-mixture regression models of SREM corresponding to ten testing pairs, given CCPP sample.

Table 8. *Ten resulted semi-mixture regression models of SREM given CCPP sample.*

Pair	Semi-mixture regression model
0	{*PE* = 497.0645 - 2.1763*(*AT*): coeff=0.4227, var=29.6573},

	{*PE* = 517.8105 - 1.1672*(*V*): coeff=0.2769, var=71.6045}, {*PE* = -1058.5211 + 1.4933*(*AP*): coeff=0.1597, var=211.5011}, {*PE* = 421.6716 + 0.4486*(*RH*): coeff=0.1406, var=248.4670}
1	{*PE* = 497.4977 - 2.1979*(*AT*): coeff=0.4280, var=24.2516}, {*PE* = 519.3656 - 1.1965*(*V*): coeff=0.2768, var=60.5493}, {*PE* = -1214.8271 + 1.6475*(*AP*): coeff=0.1584, var=180.1064}, {*PE* = 417.8420 + 0.5020*(*RH*): coeff=0.1368, var=217.2895}
2	{*PE* = 497.6871 - 2.2081*(*AT*): coeff=0.4291, var=20.0817}, {*PE* = 520.7027 - 1.2180*(*V*): coeff=0.2841, var=48.7344}, {*PE* = -1304.3280 + 1.7359*(*AP*): coeff=0.1541, var=157.6827}, {*PE* = 413.7453 + 0.5563*(*RH*): coeff=0.1327, var=189.0623}
3	{*PE* = 498.5778 - 2.2479*(*AT*): coeff=0.4453, var=13.8541}, {*PE* = 522.2781 - 1.2467*(*V*): coeff=0.2830, var=37.3203}, {*PE* = -1512.8163 + 1.9414*(*AP*): coeff=0.1472, var=128.9238}, {*PE* = 405.7745 + 0.6610*(*RH*): coeff=0.1245, var=156.3326}
4	{*PE* = 498.5320 - 2.2546*(*AT*): coeff=0.4335, var=10.5627}, {*PE* = 523.8185 - 1.2793*(*V*): coeff=0.2961, var=26.0347}, {*PE* = -1714.4568 + 2.1407*(*AP*): coeff=0.1511, var=91.7893}, {*PE* = 401.0777 + 0.7325*(*RH*): coeff=0.1192, var=123.9264}
5	{*PE* = 498.4271 - 2.2470*(*AT*): coeff=0.4353, var=7.9534}, {*PE* = 523.2183 - 1.2717*(*V*): coeff=0.2939, var=19.5630}, {*PE* = -1857.9068 + 2.2820*(*AP*): coeff=0.1528, var=67.4559}, {*PE* = 392.9270 + 0.8393*(*RH*): coeff=0.1181, var=90.1255}
6	{*PE* = 498.0319 - 2.2315*(*AT*): coeff=0.4395, var=5.0596}, {*PE* = 524.2077 - 1.2912*(*V*): coeff=0.2861, var=13.3621}, {*PE* = -1963.7000 + 2.3864*(*AP*): coeff=0.1552, var=42.8344}, {*PE* = 387.3950 + 0.9189*(*RH*): coeff=0.1192, var=60.0808}
7	{*PE* = 498.3792 - 2.2522*(*AT*): coeff=0.4358, var=2.9110}, {*PE* = 525.1901 - 1.3086*(*V*): coeff=0.2879, var=7.2515}, {*PE* = -2134.9587 + 2.5554*(*AP*): coeff=0.1520, var=23.8247}, {*PE* = 381.4177 + 0.9984*(*RH*): coeff=0.1243, var=29.7061}
8	{*PE* = 496.4571 - 2.1705*(*AT*): coeff=0.4590, var=1.1633}, {*PE* = 524.3892 - 1.2790*(*V*): coeff=0.2884, var=3.3270}, {*PE* = -2349.5928 + 2.7669*(*AP*): coeff=0.1334, var=12.5737}, {*PE* = 369.4027 + 1.1507*(*RH*): coeff=0.1192, var=16.3293}
9	{*PE* = 497.3288 - 2.1356*(*AT*): coeff=0.5466, var=0.1691}, {*PE* = 532.0547 - 1.4489*(*V*): coeff=0.2210, var=1.0673}, {*PE* = -2537.2255 + 2.9526*(*AP*): coeff=0.1349, var=2.7906}, {*PE* = 369.2398 + 1.1183*(*RH*): coeff=0.0975, var=4.1247}

In Table 8, each PRM is wrapped in two brackets "{.}". Notation "coeff" denotes mixture coefficient and notation "var" denotes the variance of a PRM.

Table 9 is the comparison of REM and SREM with regard to *RMAE* and *t* given CCPP sample.

Table 9. *Comparison of REM and SREM regarding RMAE and t, given CCPP sample.*

Pair	*RMAE* (REM)	*RMAE* (SREM)	*t* (REM)	*t* (SREM)
0	0.0081	0.0123	1	2
1	0.0081	0.0119	5	5
2	0.0081	0.0116	10	7
3	0.0082	0.0111	27	8
4	0.0082	0.0109	23	10
5	0.0083	0.0109	68	10
8	0.0084	0.0109	994	9
7	0.0084	0.0109	47	8

8	0.0089	0.0110	90	13
9	0.0101	0.0104	1780	23
Average	0.0085	0.0112	304.5	9.5

From Table 9, given CCPP sample, SREM is faster than REM according to *t* but the accuracy of REM is better than the accuracy of SREM according to *RMAE*. Note [19], values of paired t-test statistic t_0 [1, p. 376] of *RMAE* for REM and SREM are 6.1786 and 5.9070, respectively. Because all these values are larger than the percentage point $t_{0.05,8}$ = 1.860 [1, p. 711] given significant level 95%, the resistance of REM and SREM to missing values given CCPP sample is proved.

From experimental results of both gestational sample and CCPP sample, the convergence of SREM is always faster than the convergence of REM because SREM decomposes a long regression model into many shorter regression models. In optimization process of SREM, of course each short model with only one independent variable in two-dimension space will converge faster than the long model because the long model needs much more iterations to reach and balance the optimal point (optimizer) in multi-dimension space with many independent variables.

4. Conclusions

From the number of iterations, we conclude that SREM converges faster than REM does. According to *RMAE* metric, the accuracy of REM is better than the accuracy of REM but their distance in accuracy is not large. Moreover, the number of PRMs in fair tests is equal to the number of regressors and each PRM has only one regressor. If the number of PRMs is large enough and each PRM has many enough regressors with some combination of regressors, SREM may be better than REM. Note, Bayesian Information Criterion (BIC) was proposed to estimate the number of PRMs in [11, p. 5]. This may be true but finding the optimal number of PRMs and regressors for SREM is not a methodological ideology because the essence of SREM is decomposition of REM. Therefore, for further research, we will modify SREM so that it implements fully mixture model in which both mixture coefficients c_k and regression coefficients α_k are estimated by normal mixture model and balance process (estimation of missing values) of REM. We expect that taking advantages of both mixture model and REM via iterative process will result out better estimation at least in the case that incomplete data varies in many trends.

In general, the combination of REM and mixture model like SREM is potential. The website of REM and SREM is http://rem.locnguyen.net.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgments

We express our deep gratitude to Prof. Dr. Thu-Hang Thi Ho (Vinh Long General Hospital – Vietnam) who provided us the gestation sample of ultrasound measures and fetal weights for testing REM and SREM. Prof. Dr. Thu-Hang Thi Ho is also the co-author of REM algorithm in the previous research "Fetal Weight Estimation in Case of Missing Data" [3]. We also express our deep gratitude to Prof. Bich-Ngoc Tran who gave us comments relevant to one-way paired t-test for evaluating the resistance of REM and SREM to missing values.

Appendix

A1. Proof of equation (19)

The joint probability of data $\boldsymbol{X}$ and data Z for each k^{th} PRM is defined as follows:

$$P_k(\boldsymbol{X}, Z|\alpha_k, \sigma_k^2, \beta_{kj}) = P_k(Z|\boldsymbol{X}, \alpha_k, \sigma_k^2) P_k(X_j|Z, \beta_{kj})$$

When $\boldsymbol{X}$, X_j, and Z are specified in equation (16), we have:

$$P_k(\boldsymbol{X}, Z|\alpha_k, \sigma_k^2, \beta_{kj}) = \left(\prod_{i=1}^{N} P_k(z_i|X_i, \alpha_k, \sigma_k^2)\right)\left(\prod_{i=1}^{N} P_k(x_{ij}|z_i, \beta_{kj})\right)$$

(Because all z_i are mutually independent given X_i and all x_{ij} with fixed j are mutually independent given z_i)

$$= \left(\frac{1}{2\pi\sqrt{\sigma_k^2\tau_{kj}^2}}\right)^N * \exp\left(-\sum_{i=1}^{N}\frac{\left(z_i - \alpha_k^T X_i\right)^2}{2\sigma_k^2}\right) * \exp\left(-\sum_{i=1}^{N}\frac{\left(x_{ij} - \beta_{kj}^T(1, z_i)^T\right)^2}{2\tau_{kj}^2}\right)$$

(Due to equations (13) and (15))

The log-likelihood function is natural logarithm of the joint probability $P_k(\boldsymbol{X}, Z|\alpha_k, \sigma_k^2, \beta_{kj})$ as follows:

$$\begin{aligned} L(\alpha_k, \sigma_k^2, \beta_{kj}) &= \log\left(P_k(\boldsymbol{X}, Z|\alpha_k, \sigma_k^2, \beta_{kj})\right) \\ &= -N\log(2\pi) - \frac{N}{2}\left(\log(\sigma_k^2) + \log(\tau_{kj}^2)\right) - \sum_{i=1}^{N}\frac{\left(z_i - \alpha_k^T X_i\right)^2}{2\sigma_k^2} \\ &\quad - \sum_{i=1}^{N}\frac{\left(x_{ij} - \beta_{kj}^T(1, z_i)^T\right)^2}{2\tau_{kj}^2} \end{aligned}$$

The optimal estimate $(\alpha_k^*, (\sigma_k^2)^*, \beta_{kj}^*)^T$ is a maximizer of $L(\alpha_k, \sigma_k^2, \beta_{kj})$ [17, p. 9].

$$\left(\alpha_k^*, (\sigma_k^2)^*, \beta_{kj}^*\right)^T = \underset{\alpha_k, \sigma_k^2, \beta_{kj}}{\operatorname{argmax}} L(\alpha_k, \sigma_k^2, \beta_{kj})$$

By taking first-order partial derivatives of $L(\alpha_k, \sigma_k^2, \beta_{kj})$ with regard to α_k, σ_k^2, and β_{kj}, we obtain [24, p. 34]:

$$\begin{aligned} \frac{\partial L(\alpha_k, \sigma_k^2, \beta_{kj})}{\partial \alpha_k} &= -\frac{1}{\sigma_k^2}\sum_{i=1}^{N}\left(z_i - \alpha_k^T X_i\right)X_i^T \\ \frac{\partial L(\alpha_k, \sigma_k^2, \beta_{kj})}{\partial \beta_{kj}} &= -\frac{1}{\tau_{kj}^2}\sum_{i=1}^{N}\left(x_{ij} - \beta_{kj}^T(1, z_i)^T\right)(1, z_i) \\ \frac{\partial L(\alpha_k, \sigma_k^2, \beta_{kj})}{\partial \sigma_k^2} &= -\frac{N}{2\sigma_k^2} + \frac{1}{2(\sigma_k^2)^2}\sum_{i=1}^{N}\left(z_i - \alpha_k^T X_i\right)^2 \end{aligned}$$

When first-order partial derivatives of $L(\alpha_k, \sigma_k^2, \beta_{kj})$ are equal to zero, it gets local maximal. In other words, $(\alpha_k^*, (\sigma_k^2)^*, \beta_{kj}^*)^T$ is solution of the following system of linear equations:

$$\begin{cases} -\frac{1}{\sigma_k^2}\sum_{i=1}^{N}\left(z_i - \alpha_k^T X_i\right)X_i^T = \mathbf{0}^T \\ -\frac{1}{\tau_{kj}^2}\sum_{i=1}^{N}\left(x_{ij} - \beta_{kj}^T(1, z_i)^T\right)(1, z_i) = \mathbf{0}^T \\ -\frac{N}{2\sigma_k^2} + \frac{1}{2(\sigma_k^2)^2}\sum_{i=1}^{N}\left(z_i - \alpha_k^T X_i\right)^2 = 0 \end{cases}$$

The notation $\mathbf{0} = (0, 0, \ldots, 0)^T$ denotes zero vector. Solution of the system of linear equations above is [1, p. 457]:

$$\begin{cases} \alpha_k^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Z \\ \beta_{kj}^* = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T X_j \\ \left(\sigma_k^2\right)^* = \frac{1}{N}\sum_{i=1}^{N}(z_i - (\alpha_k^*)^T X_i)^2 \end{cases}$$

Where $\boldsymbol{X}$, X_i, X_j, and $\boldsymbol{Z}$ are specified by equation (16). Therefore, the equation (19) is established.

References

[1] Montgomery, D. C.; Runger, G. C. Applied Statistics and Probability for Engineers, 5th ed.; John Wiley & Sons: Hoboken, New Jersey, USA, 2010, 792. Available online: https://books.google.com.vn/books?id=_f4KrEcNAfEC. (accessed on 6 September 2016)

[2] Horton, N. J.; Kleinman, K. P. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 2007, 61(1), 79-90, DOI: 10.1198/000313007X172556.

[3] Nguyen, L.; Ho, T.-H. T. Fetal Weight Estimation in Case of Missing Data. *Experimental Medicine (EM)* - Special Issue "Medicine and Healthy Food", 2018.

[4] Kokic, P. *The EM Algorithm for a Multivariate Regression Model: including its applications to a non-parametric regression model and a multivariate time series model*. Qantaris GmbH, Frankfurt, 2002. Available online: https://www.cs.york.ac.uk/euredit/_temp/The%20Euredit%20Software/NAG%20Prototype%20platform/WorkingPaper4.pdf. (accessed on 30th June 2018)

[5] Ghitany, M. E.; Karlis, D.; Al-Mutairi, D. K.; Al-Awadhi, F. An EM Algorithm for Multivariate Mixed Poisson Regression Models and its Application. *Applied Mathematical Sciences*, 2012, 6(137), 6843-6856. Available online: http://www.m-hikari.com/ams/ams-2012/ams-137-140-2012/ghitanyAMS137-140-2012.pdf (accessed on 3 July 2018).

[6] Anderson, B.; Hardin, M. J. Modified logistic regression using the EM algorithm for reject inference. *International Journal of Data Analysis Techniques and Strategies*, 2013, 5(4), 359-373. DOI: 10.1504/IJDATS.2013.058582.

[7] Zhang, X.; Deng, J.; Su, R. The EM algorithm for a linear regression model with application to a diabetes data. In *Proceedings of the 2016 International Conference on Progress in Informatics and Computing (PIC)*, Shanghai, China, 2016, DOI: 10.1109/PIC.2016.7949477.

[8] Haitovsky, Y. Missing Data in Regression Analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1968, 30(1), 67-82. Available online: https://www.jstor.org/stable/2984459 (accessed on 3 July 2018).

[9] Robins, J. M.; Rotnitzki, A.; Zhao, L. P. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, 1995, 90(429), 106-121, DOI: 10.2307/2291134.

[10] Lamont, A. E.; Vermunt, J. K.; Lee, V. H. M. Regression mixture models: Does modeling the covariance between independent variables and latent classes improve the results? *Multivariate Behavioral Research*, 2016, 51(1), 35-52, DOI: 10.1080/00273171.2015.1095063.

[11] Hoshikawa, T. Mixture regression for observational data, with application to functional regression models. *arXiv preprint*, 30th June 2013. arXiv:1307.0170.

[12] Nguyen, H. D. *Finite Mixture Models for Regression Problems*. The University of Queensland, Brisbane, 2015, DOI: 10.14264/uql.2015.584.

[13] Sung, H. G. *Gaussian Mixture Regression and Classification*. Rice University, Houston, 2004. Available online: https://scholarship.rice.edu/handle/1911/18710 (accessed on 4 September 2018).

[14] Tian, Y.; Sigal, L.; Badino, H.; Torre, F. D. l.; Liu, Y. Latent Gaussian Mixture Regression for Human Pose Estimation. In *Lecture Notes in Computer Science*, vol 6494, Proceedings of The 10th Asian Conference on Computer Vision (ACCV 2010), Queenstown, 2010. DOI: 10.1007/978-3-642-19318-7_53.

[15] Grün, B.; Leisch, F. Finite Mixtures of Generalized Linear Regression Models. University of Munich, Munich, 2007. Available online: https://pdfs.semanticscholar.org/e0d5/6ac54b80a1a4e274f11b1d86840461cc542c.pdf (accessed on 4 September 2018).

[16] Bilmes, J. A. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. University of Washington, Berkeley, 1998. Available online: http://melodi.ee.washington.edu/people/bilmes/mypubs/bilmes1997-em.pdf (accessed on 17 September 2013).

[17] Lindsten, F.; Schön, T. B.; Svensson, A.; Wahlström, N. *Probabilistic modeling – linear regression & Gaussian processes*. Uppsala University, Uppsala, 2017. Available online: http://www.it.uu.se/edu/course/homepage/sml/literature/probabilistic_modeling_compendium.pdf (accessed on 24 January 2018).

[18] Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1977, 39(1), 1-38.

[19] Nguyen, L.; Ho, T.-H. T. Early Fetal Weight Estimation with Expectation Maximization Algorithm. *Experimental Medicine (EM)*, 2018, 1(1), 12-30, DOI: 10.31058/j.em.2018.11002.

[20] Ho, T. H. T.; Phan, D. T. Fetal Weight Estimation from 37 Weeks to 42 Weeks by Two-Dimensional Ultrasound Measures. *Journal of Practical Medicine*, 2011, 12(797), 8-9.

[21] Ho, T. H. T.; Phan, D. T. Fetal Age Estimation by Three-Dimensional Ultrasound Measure of Arm Volume and Other Two-Dimensional Ultrasound Measures. *Journal of Practical Medicine*, 2011, 12(798), 12-15.

[22] Tüfekci, P.; Kaya, H. Combined Cycle Power Plant Data Set, Irvine, California: Center for Machine Learning and Intelligent Systems, 2014.

[23] Pinette, M. G.; Pan, Y.; Pinette, S. G.; Blackstone, J.; Garrett, J.; Cartin, A. Estimation of Fetal Weight: Mean Value from Multiple Formulas. *Journal of Ultrasound in Medicine*, 1999, 18(12), 813-817. Available online: https://www.ncbi.nlm.nih.gov/pubmed/10591444 (accessed on 9 October 2016).

[24] Nguyen, L. *Matrix Analysis and Calculus*. Matrix Analysis and Calculus, 1st ed.; Evans, C., Ed.; Hanoi, Vietnam: Lambert Academic Publishing, 2015, 72. Available online: https://www.shuyuan.sg/store/gb/book/matrix-analysis-and-calculus/isbn/978-3-659-69400-4 (accessed on 3 March 2014).