VOLUME 1, 2018
DOI: 10.31058/j.em.2018.12004

Fetal Weight Estimation in Case of Missing Data

Loc Nguyen[1*], Thu-Hang T. Ho[2]

[1] Independent Scholar, Loc Nguyen's Academic Network, An Giang, Vietnam
[2] Board of Directors, Vinh Long General Hospital, Vinh Long, Vietnam

Email Address

ng_phloc@yahoo.com (L. Nguyen), bshangvl2000@yahoo.com (H. Ho)
*Correspondence: Timothy.Schmutte@yale.edu

Received: 5 July 2018**; Accepted:** 30 September 2018**; Published:** 17 December 2018

Abstract:

Fetal weight estimation before delivery is important in obstetrics, which assists doctors diagnose abnormal or diseased cases. Linear regression based on ultrasound measures such as bi-parietal diameter (*bpd*), head circumference (*hc*), abdominal circumference (*ac*), and fetal length (*fl*) is common statistical method for weight estimation. There is a demand to retrieve regression model in case of incomplete data because taking ultrasound examinations is a hard task and early weight estimation is necessary in some cases. In this research, we proposed so-called regression expectation maximization (REM) algorithm which is a combination of linear regression method and expectation maximization (EM) method to construct the regression model when both ultrasound measures and fetal weight are missing. The special technique in REM is to build parallelly an entire regression function and many partial inverse regression functions for solving the problem of highly sparse data, in which missing values are fulfilled by expectations relevant to both entire regression function and inverse regression functions. 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%.

Keywords:

Fetal Weight Estimation, Regression Model, Ultrasound Measures, Expectation Maximization Algorithm, Missing Data

1. Introduction

According to the regression approach of fetal weight estimation, without loss of generality, an estimation formula is a linear regression function $Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ where Z is estimated fetal weight whereas X_i (s) are gestational ultrasound measures such as bi-parietal diameter (*bpd*), head circumference (*hc*), abdominal circumference (*ac*), fetal length (*fl*). Variable Z is called response variable or dependent variable. Each X_i is called regression variable, regressor, predictor, regression variable, or independent variable. Each α_i is called regression coefficient. Here we focus on applying expectation maximization (EM) algorithm into constructing regression model. 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. Because this research is the successive one after our previous research [1], they share some common contents, but we confirm that their methods are different. The algorithm in the previous research is dual regression expectation maximization (DREM) algorithm. DREM only accepts incomplete Z but REM accepts both incomplete Z and incomplete X_i. We need to repeat here the survey of fetal weight and age estimation based on regression analysis because this survey was made in our previous research [1].

As pioneers, Hadlock et al. [2] proposed regression models for weight estimation based on head size, abdominal size, and femur length, which is better than those based on measurements of head and body. Error means in percentage of their models are 1.3%, 1.5%, 0.4%, 1.4%, 2.3%, and –0.7% whereas error standard deviations are 10.1%, 9.8%, 7.7%, 7.3%, 7.4%, 7.3%.

Phan [3] proposed some excellent regression formulas for estimating fetal age and weight based on *bpd*, *hc*, *ac*, abdominal area (*aa*), abdominal diameter (*ad*), average abdominal diameter (*aad*). Pham [4] proposed some excellent regression formulas for estimating fetal weight based on *bpd*, ad, arm length (*al*), abdominal diameter (*ad*), average abdominal diameter (*aad*). Ho [5] produced some excellent regression formulas for estimating fetal age and weight based on *bpd*, *ac*, *hc*, and thigh volume in her PhD dissertation. Some of Ho's formulas [5, pp. 155-157] are log(*weight*) = 1.746 + 0,0124**bpd* + 0,001906**ac* with R = 0.962, *weight* = –13099.1862 + 125.662**ac* – 0.3818**ac***ac* + 0.00045**ac***ac***ac* with R = 0.9247, *weight* = –3306 + 55.477**bpd* + 13.483**thigh_volume* with R = 0.9663, *age* = 167.0791 – 1,5537**ac* + 0.00556**ac***ac* – 0.00000618**ac***ac***ac* with R = 0.8980, *age* = 331.0223 – 1.6118 * (*hc* + *ac*) + 0.0028 * (*hc* + *ac*) * (*hc* + *ac*) – 0.0000015 * (*hc* + *ac*) * (*hc* + *ac*) * (*hc* + *ac*) with R = 0.9212, *age* = 21.1148 + 0.2381 * *thigh_volume* – 0.001 * *thigh_volume* * *thigh_volume* + 0.000002 * *thigh_volume* * *thigh_volume* with R= 0,9959. Note that log(.) denotes logarithm function and R is correlation coefficient. The larger the R is, the better the formula is.

Deter, Rossavik, and Harrist [6] reassessed the weight estimation procedure of Rossavik (regression analysis) with particular emphasis on parameter estimation and performance over a wide weight range. As results, Deter, Rossavik, and Harrist assured that "there is no systematic errors over a 250 gram to 4750 gram weight range and random errors (±l standard deviation) of 10% to 13% below 200 gram and 6% to 8% above 2000 gram. The weights of small-and large-for-gestational age fetuses were systematically overestimated (4.1%) and underestimated (–3.0%), respectively, but systematic errors were not found in average-for-gestational age fetuses".

Chien, Owen, and Khan [7] did an evaluation research on formulas of Aoki, Campbell, Shepard, and Hadlock. Chien, Owen, and Khan [7, p. 856] concluded that: "The smallest mean difference was obtained with the Shepard and Aoki formulas (51.4 gram and 60.5 gram, respectively), whereas the Campbell and Hadlock formulas produced larger mean differences (141.8 gram and 190.7 gram, respectively). The Aoki formula generated the smallest range between the limits of agreement (–324.2 to 445.2 gram) whereas the Campbell formula produced the largest range (–286.5 to 570.1 gram). The range between the limits of agreement generated with the Shepard and Hadlock formulas were intermediate between those produced by the Aoki and Campbell formulas. The intraclass correlation coefficients generated with the Aoki and Shepard formulas were identical (0.90). The intraclass correlation coefficients obtained with the Hadlock (0.84) and Campbell formulas (0.85) were lower".

Varol et al. [8] evaluated the growth curve of well-functioned regression models (Hadlock formulas, for example). Their purpose is to contribute to develop national standard growth curve of gestational age and birth weight. Percentile values and correlation coefficients were calculated and well-functioned regression models were produced for growth curve. As a result, the regression model for gestational age *age* = 4.945 + 0.606**ac* + 0.105**bpd* + 0.286**fl* with adjusted R^2 = 0.937 is optimal.

Dudley [9] made a full review of different methods of fetal weight estimation including works of Deter, Hadlock, Dudley, Ott, Rose, McCallum, Miller, Warsof, Simon, Sabbagha, Smulian, Shepard, Blann, Prien, Eden, Jouannic, Medchill, Townsend, Kaaij, Robson, Weinberger, and Weiner [9, pp. 83-85]. The research of Dudley is cohort study with evaluation criteria such as mean of percentage error and standard deviation of percentage error [9, pp. 80-81]. As results, Dudley [9, p. 80] stated that "no consistently superior method has emerged and volumetric methods provide some theoretical advantages". Moreover Dudley [9, p. 80] stated that "random errors are large and must be reduced if clinical errors are to be avoided". Dudley [9, p. 80] also concluded that "the accuracy of weight estimation is compromised by large intra- and interobserver variability and efforts must be made to minimize this variability if weight estimation is to be clinically useful". According to Dudley [9, p. 80], the improvement in weight estimation may be achieved through averaging of multiple measurements, improvements in image quality, uniform calibration of equipment, careful design and refinement of measurement methods, acknowledgment that there is a long learning curve, and regular audit of measurement quality.

Salomon, Bernard, and Ville [10] used polynomial regression approach to compute a new reference chart for weight estimation. Their resulted birth-weight chart showed that the weight estimation was noticeably larger at 25 – 36 weeks. At 28 – 32 weeks, the 50^{th} centile of actual birth weight is approximated to the 50^{th} centile of estimated weight.

A. R. Akinola, I. O. Akinola, and O. O. Oyekan [11] evaluated many regression estimation models. Their results showed that models with *hc* and *ac* are not as good as those with *ac* and *bpd*. The combination of *fl* and *ac* did not improve accuracy. The use of multiple measures gives most accurate estimation.

Lee et al. [12] used multiple linear regression model with standard measures (*bpd*, *fl*, *ac*) and their proposed biometrics such as fractional arm volume (*fav*) and fractional thigh volume (*ftv*). They produced six weight estimation models such as model 1, model 2, model 3, model 4, model 5, and model 6. The model 3 which is log(*weight*) = 0.5046 + 1.9665*log(*bpd*) – 0.3040*log(*bpd*)*log(*bpd*) + 0.9675*log(*ac*) + 0.3557*log(*fav*) and model 6 which is log(*weight*) = –0.8297 + 4.0344*log(*bpd*) – 0.7820*log(*bpd*)*log(*bpd*) + 0.7853*log(*ac*) + 0.0528*log(*ftv*)*log(*ftv*) gain highest accuracy. Model 5 classified an additional 9.1% and 8.3% of fetuses within 5% and 10% of birth weight. Model 6 classified an additional 7.3% and 4.1% of infants within 5% and 10% of birth weight.

Bennini et al. [13] created a total of 210 pregnant women in their research into a formula-generating group (150 women) and prospective validation group (60 women). Polynomial regression is used to generate one formula based on two-dimension measures, one formula based on fetal thigh volume by multi-planar technique, and one formula based on fetal thigh volume by Virtual Organ Computer-aided Analysis. The experimental results showed that their models are significantly good and there is

no significant difference between two-dimension model and three-dimension models. Note that their two-dimension model is *weight* = –562.824 + 11.962**ac***fl* + 0.009**bpd***bpd***ac***ac*. Their three-dimension models are *weight* = 1033.286 + 12.733**thigh_volume* and *weight* = 1025.383 + 12.775* *thigh_volume*.

Cohen et al. [14] used linear regression model to compare estimated weights for births after 6 days after last ultrasound scan and actual weights. Their results indicate that the mean ± standard deviation percentage among deliveries within 1 day of last ultrasound scan is 0.2 ± 9%.

Siggelkow et al. [15] proposed a new algorithm of isotonic regression to construct a birth weight prediction function that increases monotonically with each of input variables (ultrasound measures) and minimizes empirical quadratic loss. As a result, their isotonic regression function gains a small mean absolute error (312 gram).

Mei Wu et al. [16] used measures *bpd*, *hc*, *ac*, and *ft* to estimate fetal weight. Their results [16, p. 540] indicate that there were no significant differences in the fetal AC or body weight evaluated before delivery and recorded after delivery. Mei Wu et al. concluded [16, p. 540] that "their new technique is more convenient and applicable for the evaluation of *ac* as compared to standard method and seems to be reliable and accurate for the assessment of fetal weight". Their technique focuses on how to take and process ultrasound measures from ultrasonic machine [16, pp. 541-542]. The evaluation criteria are absolute error and relative error [16, p. 543].

Pinette et al. [17] used mean weight value from multiple formulas in order to improve the estimation. For instance, Pinette et al. calculated four estimated weight values w_1, w_2, w_3, and w_4 from formulas of Shepard, Hadlock, and Combs and then, they computed the mean $w = (w_1 + w_2 + w_3 + w_4) / 4$ as the optimal estimated value of birth weight.

When fetal weight is estimated based on gestational age, the weight-for-gestational chart is used. In such chart, if gestational age falls below 10^{th} percentile then, it is impossible to estimate respective weight and so such problem is called small-for-gestational-age which often occurs because of missing data. Hutcheon and Platt [18] applied standard epidemiologic approaches to correct the missing data problem. However such approaches does not use regression model. When gestational age is incompletely recorded, Eberg, Platt, and Filion [19] proposed four approaches to estimating missing gestational age: (1) generalized estimating equations for longitudinal data; (2) multiple imputation; (3) estimation based on fetal birth weight and sex; and (4) conventional approaches that assigned a fixed value (39 weeks for all or 39 weeks for full term and 35 weeks for preterm).

There is a demand to construct regression model in case of missing data because taking ultrasound examinations is a hard task and early weight estimation is necessary in some cases [1]. EM algorithm is an approach to solve the problem of incomplete data in regression analysis. Here we browse some researches relevant to EM algorithm and regression model. Kokic [20] 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 [21] calculated the expectation of function of mixture random variable in the 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 [22] 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 [23] used EM algorithm to build up linear regression model for studying glycosylated hemoglobin from partial missing data. In other words, Zhang, Deng, and Su [23] 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 [24] 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 [25] 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 [25] 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 [26] 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 [26, 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 [26, 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 [26, 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 [26, p. 6]. In M-step, regression model is built from the 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 [25] 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 [26, p. 7].

In general, the ideology of applying EM algorithm into regression model is not new but our proposed REM algorithm can build up regression models in case that both response variable Z and regressors X_i have missing values. In other words, REM accepts highly sparse data. From experimental results, the accuracy of REM decreases insignificantly when data sample is made sparse with loss ratios up to 80%. The special technique in REM is to build parallelly an entire regression function $Z = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + ... + \alpha_nX_n$ and many partial inverse regression functions $X_j = \beta_{j0} + \beta_{j1}Z$ for solving the problem of highly sparse data, in which missing values are fulfilled by expectations relevant to both entire regression function and inverse regression functions. Such expectations are re-estimated by a so-called balance process until their bias is small enough.

2. Methodology

Suppose we estimate the linear regression model $Z = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + ... + \alpha_nX_n$ where Z is fetal weight and Y is fetal age whereas X_i (s) are gestational ultrasound measures such as *bpd*, *hc*, *ac*, and *fl*. Suppose the random variable Z conforms normal distribution, according to equation (1) [27, pp. 8-9]. Note, Z is random variable whereas X is data in equation (1).

$$P(Z|X,\alpha) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(Z-\alpha^TX\right)^2}{2\sigma^2}\right) \quad (1)$$

Where $\alpha = (\alpha_0, \alpha_1,\ldots, \alpha_n)^T$ is parameter vector and $X = (1, X_1, X_2,\ldots, X_n)^T$ is data vector. The mean and variance of Z with regard to $P(Z \mid X, \alpha)$ are α^TX and σ^2, respectively. The superscript "^{T}" denotes transposition operator in vector and matrix. Suppose each $X_j \in X$ has an inverse linear regression model $X_j = \beta_{j0} + \beta_{j1}Z$. In other words, X_j now is considered as the random variable conforming normal distribution according to equation (2).

$$P_j\left(X_j\middle|Z,\beta_j\right) = \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{\left(X_j-\beta_j^T(1,Z)^T\right)^2}{2\sigma_j^2}\right) \quad (2)$$

Where $\beta_j = (\beta_{j0}, \beta_{j1})^T$ is a partial parameter vector and $(1, Z)^T$ is a partial data vector. The mean and variance of each X_j with regard to the inverse distribution $P_j(X_j \mid Z, \beta_j)$ are $\beta_j^T(1, Z)^T$ and σ_j^2, respectively. Of course, there are n inverse linear regression models.

Let $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{z})$ be collected sample in which $\boldsymbol{X}$ is a set of sample measures and $\boldsymbol{z}$ is a set of fetal weights with note that both $\boldsymbol{X}$ and $\boldsymbol{z}$ are incomplete. In other words, $\boldsymbol{X}$ and $\boldsymbol{z}$ have missing values. Now we focus on estimating α and β_j based on $\boldsymbol{D}$. As a convention, let α^* and β_j^* be estimates of α and β_j, respectively [27, p. 8].

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix}$$

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix}, \boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{pmatrix} \quad (3)$$

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_N \end{pmatrix}$$

The expectation of sufficient statistic Z regard to the entire linear model $P(Z \mid X, \alpha)$ is specified by equation (4).

$$E(Z|X) = \alpha^T X \quad (4)$$

The expectation of each sufficient statistic X_j with regard to each inverse linear model $P_j(X_j \mid Z, \beta_j)$ is specified by equation (5).

$$E\left(X_j \middle| Z\right) = \beta_j^T (1, Z)^T = \beta_{j0} + \beta_{j1} Z \quad (5)$$

Please pay attention to equations (4) and (5) because Z and X_j will be estimated by these expectations later.

By applying sample D into equations (1) and (2) and using maximum likelihood estimation (MLE) method, we retrieve equation (6) to estimate α^* and β_j^* [27, pp. 8-9].

$$\begin{cases} \alpha^* = (X^T X)^{-1} X^T z \\ \beta_j^* = (Z^T Z)^{-1} Z^T x_j \end{cases} \quad (6)$$

Where $\boldsymbol{X}$, $\boldsymbol{z}$, $\boldsymbol{Z}$, and $\boldsymbol{x}_j$ are specified in equation (3). Because $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incomplete, we apply expectation maximization (EM) algorithm into estimating $(\alpha^*, \beta_j^*)^T$. EM algorithm has many iterations and each iteration has expectation step (E-step) and maximization step (M-step) for estimating parameters. Given current parameter $\Theta^{(t)} = (\alpha^{(t)}, \beta_j^{(t)})^T$ 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 $\Theta^{(t+1)} = (\alpha^{(t+1)}, \beta_j^{(t+1)})^T$ is determined by equation (6) 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 every missing value z_i is estimated as the expectation based on the current parameter $\alpha^{(t)}$, according to equation (4).

$$z_i = \left(\alpha^{(t)}\right)^T \boldsymbol{x}_i = \sum_{j=1}^{n} \alpha_j^{(t)} x_{ij}$$

Note, $x_{i0} = 1$. Let U_i be a set of indices of missing values x_{ij} with fixed i. In other words, if $j \in U_i$ then, x_{ij} is missing. The set U_i can be empty. The equation (4) is re-written:

$$z_i = \sum_{j \in U_i} \alpha_j^{(t)} x_{ij} + \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik}$$

According to equation (5), missing value x_{ij} is estimated by:

$$x_{ij} = \beta_{j0}^{(t)} + \beta_{j1}^{(t)} z_i$$

Combining equation (4) and equation (5), we have:

$$z_i = \sum_{j \in U_i} \alpha_j \left(\beta_{j0}^{(t)} + \beta_{j1}^{(t)} z_i \right) + \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik} = z_i \sum_{j \in U_i} \alpha_j^{(t)} \beta_{j1}^{(t)} + \sum_{j \in U_i} \alpha_j^{(t)} \beta_{j0}^{(t)} + \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik}$$

It implies:

$$z_i = \frac{\sum_{j \in U_i} \alpha_j^{(t)} \beta_{j0}^{(t)} + \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik}}{1 - \sum_{j \in U_i} \alpha_j^{(t)} \beta_{j1}^{(t)}} \quad (7)$$

Missing values z_i and x_{ij} are estimated by the balance process shown in Table 1.

Table 1. *Balance process for estimating missing values.*

Step 1: Missing values z_i are estimated by equation (7), based on the current parameter $\Theta^{(t)} = (\alpha^{(t)}, \beta_j^{(t)})^T$.

$$z_i = \frac{\sum_{j \in U_i} \alpha_j^{(t)} \beta_{j0}^{(t)} + \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik}}{1 - \sum_{j \in U_i} \alpha_j^{(t)} \beta_{j1}^{(t)}}$$

Missing values x_{ij} where $j \in U_i$ are estimated by equation (5) and the estimated values z_i above, based on the current parameter $\Theta^{(t)} = (\alpha^{(t)}, \beta_j^{(t)})^T$.

$$x_{ij} = \beta_{j0}^{(t)} + \beta_{j1}^{(t)} z_i$$

Step 2: For balancing both $P(Z \mid X, \alpha)$ and $P_j(X_j \mid Z, \beta_j)$ in estimation, values z_i and x_{ij} are re-estimated by equations (4) and (5) as new z_i' and x_{ij}', based on the current parameter $\Theta^{(t)} = (\alpha^{(t)}, \beta_j^{(t)})^T$.

$$z_i' = \sum_{j=1}^{n} \alpha_j^{(t)} x_{ij}$$

$$x_{ij}' = \beta_{j0}^{(t)} + \beta_{j1}^{(t)} z_i'$$

Step 3: If the ratio deviation between (z_i', x_{ij}') and (z_i, x_{ij}) is smaller than a small enough threshold or the process reaches a large enough number of iterations, the process stops; at that time z_i' and x_{ij}' are final estimated values. Otherwise, going back step 2 with assignment $x_{ij} = x_{ij}$'.

In fact, the balance process is an iterative process which is the combination of equations (4), (5), and (7). The process starts to estimate missing values z_i without use of x_{ij}. Conversely, the process can start to estimate missing values x_{ij} without use of z_i, which is called inverse balance process.

Recall that U_i is the set of indices of missing values x_{ij} with fixed i. Every missing value x_{il} is estimated as the expectation based on the current parameter $\beta_j^{(t)}$, according to equation (5).

$$x_{il} = \beta_{l0}^{(t)} + \beta_{l1}^{(t)} z_i$$

According to equation (4), missing value z_i is estimated by:

$$z_i = \left(\alpha^{(t)}\right)^T \boldsymbol{x}_i = \sum_{j=1}^{n} \alpha_j^{(t)} x_{ij}$$

Combining equation (5) and equation (4), we have:

$$x_{il} = \beta_{l0}^{(t)} + \beta_{l1}^{(t)} \left(\sum_{j \in U_i} \alpha_j^{(t)} x_{ij} + \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik} \right)$$
$$= \beta_{l1}^{(t)} \alpha_l^{(t)} x_{il} + \beta_{l1}^{(t)} \sum_{j \in U_i, j \neq l} \alpha_j^{(t)} x_{ij} + \beta_{l0}^{(t)} + \beta_{l1}^{(t)} \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik}$$

In other words, we have:

$$\left(\beta_{l1}^{(t)} \alpha_l^{(t)} - 1 \right) x_{il} + \beta_{l1}^{(t)} \sum_{j \in U_i, j \neq l} \alpha_j^{(t)} x_{ij} = -\beta_{l0}^{(t)} - \beta_{l1}^{(t)} C$$

Where,

$$C = \sum_{k \notin U_i} \alpha_k^{(t)} x_{ik}$$

Suppose the cardinality of U_i is k, which means that there are k missing values x_{ij} where $j \in U_i$. Derived from the combination above, missing values $\left(x_{ij}\right)_{j \in U_i}$ are solution of the following system of k equations.

$$\begin{cases} \left(\beta_{11}^{(t)} \alpha_1^{(t)} - 1 \right) x_{i1} + \left(\beta_{11}^{(t)} \alpha_2^{(t)} \right) x_{i2} + \cdots + \left(\beta_{11}^{(t)} \alpha_k^{(t)} \right) x_{ik} = -\beta_{10}^{(t)} - \beta_{11}^{(t)} C \\ \left(\beta_{21}^{(t)} \alpha_1^{(t)} \right) x_{i1} + \left(\beta_{21}^{(t)} \alpha_2^{(t)} - 1 \right) x_{i2} + \cdots + \left(\beta_{21}^{(t)} \alpha_k^{(t)} \right) x_{ik} = -\beta_{20}^{(t)} - \beta_{21}^{(t)} C \\ \vdots \\ \left(\beta_{k1}^{(t)} \alpha_1^{(t)} \right) x_{i1} + \left(\beta_{k1}^{(t)} \alpha_2^{(t)} \right) x_{i2} + \cdots + \left(\beta_{k1}^{(t)} \alpha_k^{(t)} - 1 \right) x_{ik} = -\beta_{k0}^{(t)} - \beta_{k1}^{(t)} C \end{cases}$$

Therefore, missing values x_{ij} are calculated by equation (8) according to Cramer method.

$$\left(x_{ij}\right)_{j \in U_i} = A^{-1} b \tag{8}$$

Where,

$$A = \begin{pmatrix} \beta_{11}^{(t)} \alpha_1^{(t)} - 1 & \beta_{11}^{(t)} \alpha_2^{(t)} & \cdots & \beta_{11}^{(t)} \alpha_k^{(t)} \\ \beta_{21}^{(t)} \alpha_1^{(t)} & \beta_{21}^{(t)} \alpha_2^{(t)} - 1 & \cdots & \beta_{21}^{(t)} \alpha_k^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1}^{(t)} \alpha_1^{(t)} & \beta_{k1}^{(t)} \alpha_2^{(t)} & \cdots & \beta_{k1}^{(t)} \alpha_k^{(t)} - 1 \end{pmatrix}$$

$$b = \begin{pmatrix} -\beta_{10}^{(t)} - \beta_{11}^{(t)} C \\ -\beta_{20}^{(t)} - \beta_{21}^{(t)} C \\ \vdots \\ -\beta_{k0}^{(t)} - \beta_{k1}^{(t)} C \end{pmatrix}$$

Table 2 shows the inverse balance process.

Table 2. *Inverse balance process of missing values.*

Step 1: Missing values x_{ij} where $j \in U_i$ are estimated by equation (8), based on the current parameter $\Theta^{(t)} = (\alpha^{(t)}, \beta_j^{(t)})^T$. Missing values z_i are estimated by equation (7), based on the current parameter $\Theta^{(t)} = (\alpha^{(t)}, \beta_j^{(t)})^T$.

$$\left(x_{ij}\right)_{j \in U_i} = A^{-1} b$$

Missing values z_i are estimated by equation (4) and the estimated values x_{ij} above, based on the current parameter $\Theta^{(t)} = (\alpha^{(t)}, \beta_j^{(t)})^T$.

$$z_i = \sum_{j=1}^{n} \alpha_j^{(t)} x_{ij}$$

Step 2: For balancing both $P(Z \mid X, \alpha)$ and $P_j(X_j \mid Z, \beta_j)$ in estimation, values x_{ij} and z_i are re-estimated by equations (5) and (4) as new x_{ij}' and z_i', based on the current parameter $\Theta^{(t)} = (\alpha^{(t)}, \beta_j^{(t)})^T$.

$$x'_{ij} = \beta_{j0}^{(t)} + \beta_{j1}^{(t)} z_i$$

$$z'_i = \sum_{j=1}^{n} \alpha_j^{(t)} x'_{ij}$$

Step 3: If the ratio deviation between (z_i', x_{ij}') and (z_i, x_{ij}) is smaller than a small enough threshold or the process reaches a large enough number of iterations then, the process stops; at that time z_i' and x_{ij}' are final estimated values. Otherwise, going back step 2 with assignment $z_i = z_i$'.

In fact, the inverse balance process is an iterative process which is the combination of equations (4), (5), and (8). Equation (7) used to estimate missing z_i is based on assumption of appropriate existence of missing x_{ij} and then, equation (7) leans to enhance the inverse models $X_j = \beta_{j0} + \beta_{j1}Z$. Therefore, the balance process aims to adjust equation (7). Similarly, equation (8) used to estimate missing x_{ij} is based on assumption of appropriate existence of missing z_i and then, equation (8) leans to enhance the entire model $Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$. Therefore, the inverse balance process aims to adjust equation (8).

EM algorithm [28, p. 4] associated with the (inverse) balance process for regression model is shown in Table 3. This is our so-called Regression Expectation Maximization (REM) algorithm.

Table 3. *Regression Expectation Maximization (REM) Algorithm.*

1. E-step: Missing values z_i and x_{ij} are estimated by the (inverse) balance process shown in Table 1 (Table 2). The (inverse) balance process is the core of REM.
2. M-step: The next parameter $\Theta^{(t+1)} = (\alpha^{(t+1)}, \beta_j^{(t+1)})^T$ is determined by equation (6) and the complete data $\boldsymbol{X}$ and $\boldsymbol{Z}$ fulfilled in E-step.

EM algorithm stops if at some t^{th} iteration, we have $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$. At that time, $\Theta^* = (\alpha^*, \beta^*)^T$ is the optimal estimate of EM algorithm. Here REM stops if ratio deviation between $\Theta^{(t)}$ and $\Theta^{(t+1)}$ is smaller than a small enough terminated threshold $\varepsilon > 0$ or REM reaches a large enough number of iterations *maximum_iteration*. Followings are two terminated conditions of REM:

$$\begin{cases} \left| \dfrac{\Theta^{(t+1)} - \Theta^{(t)}}{\Theta^{(t)}} \right| \leq \varepsilon \\ \text{the number of iterations} \geq maximum\_iteration \end{cases}$$

The smaller the terminated threshold is, the more accurate REM is. REM uses both the terminated threshold $\varepsilon = 0.1\% = 0.001$ and the maximum number of iterations *maximum_iteration* = 10000. The maximum number of iterations prevents REM from running for a long time.

An technique to improve the convergence of REM [1] is to initialize the parameter $\Theta^{(1)} = (\alpha^{(1)}, \beta^{(1)})^T$ at the first iteration of EM process in proper way instead of initializing $\Theta^{(1)}$ in arbitrary way [1]. Let $\boldsymbol{X}$' be the complete matrix of ultrasound

measures, which is created by removing all rows whose values are missing from $\boldsymbol{X}$. Similarly, let $\boldsymbol{Z}$' be the complete matrix of fetal weights, which is created by removing rows whose weights are missing from $\boldsymbol{Z}$. The advanced $\Theta^{(1)} = (\alpha^{(1)}, \beta^{(1)})^T$ is initialized by equation (9).

$$\begin{cases} \alpha^{(1)} = \left(X'^T X'\right)^{-1} X'^T z' \\ \beta_j^{(1)} = \left(Z'^T Z'\right)^{-1} Z'^T x_j' \end{cases} \quad (9)$$

Where $\boldsymbol{z}$' is the complete vector of non-missing weights and $\boldsymbol{x}_j$' is the complete vector of non-missing measures. Equation (9) is variant of equation (6) where $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{x}_j$, and $\boldsymbol{z}$ are replaced by $\boldsymbol{X}$', $\boldsymbol{Z}$', $\boldsymbol{x}_j$', and $\boldsymbol{z}$'. This improvement technique is the complete case method mentioned in [26, p. 3].

3. Results and Discussion

We use a gestational sample of 1027 cases in which each case includes ultrasound measures, fetus age, and fetus weight. 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 [29], [30] 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 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 [1]. Missing values are made randomly regardless of regressors (*bpd*, *hc*, *ac*, *fl*) or response variable (*weight*). For example, the training dataset (50% 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 according to Table 4 [1]. 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. Experimental results from the incomplete pairs are compared together and are aligned with experimental results from the complete pair in order to evaluate resistance of REM to missing values. The essence of the (inverse) balance process is to improve estimates of missing values at E-step. When making experiments on REM, I recognize that the (inverse) balance process with many iterations shown in Table 1 (Table 2) always results out possible estimates but it does not always result out best estimates because it makes trade-off between the entire model $Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ and many inverse models $X_j = \beta_{j0} + \beta_{j1} Z$. So, we firstly focus on experimental results of REM with one-iteration inverse balance process shown in Table 2 in which only step 1 and step 2 are performed exactly one time in every E-step of REM. In other words, the inverse balance process is degraded

as an *estimation process*. The full (inverse) balance process with many iterations is mentioned later.

Table 4. *Ten testing pair.*

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

Table 5 shows ten regression models corresponding to ten testing pairs with the estimation process.

Table 5. *Ten resulted regression models.*

Pair	Regression model	Iterations
0	*weight* = -5686.8907 + 46.2369**bpd* + 1.7148**hc* + 14.3173**fl* + 9.3881**ac*	1
1	*weight* = -5685.7854 + 43.1103**bpd* + 1.4912**hc* + 17.0387**fl* + 9.8929**ac*	4
2	*weight* = -5853.1375 + 39.5620**bpd* + 2.4174**hc* + 21.7262**fl* + 9.5004**ac*	6
3	*weight* = -6198.2135 + 44.6905**bpd* + 5.2471**hc* + 20.4518**fl* + 6.6326**ac*	7
4	*weight* = -5941.9911 + 39.9082**bpd* + 2.6244**hc* + 23.3244**fl* + 9.2312**ac*	11
5	*weight* = -6496.4041 + 44.6181**bpd* + 3.9971**hc* + 25.8895**fl* + 7.7752**ac*	18
6	*weight* = -5945.7599 + 31.7033**bpd* + 2.8255**hc* + 34.1700**fl* + 9.0212**ac*	20
7	*weight* = -6299.4105 + 66.9913**bpd* + 2.7079**hc* + 16.8104**fl* + 4.0521**ac*	36
8	*weight* = -8991.6524 + 116.5457**bpd* - 0.7010**hc* + 33.5400**fl* - 1.1436**ac*	229
9	*weight* = 20982.7191 - 27.9779**bpd* - 22.6780**hc* - 62.4584**fl* - 17.1056**ac*	269

The third column in Table 5 lists the numbers of iterations that REM converges. The larger the loss ratio is, the more the iterations are required. This implies that the complete case method [26, p. 3] which is the improvement technique mentioned in equation (9) is effective with slightly sparse sample.

Now we assess such ten regression models with subject to two typical metrics such as mean absolute error (MAE) and sample correlation coefficient (R). 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 (10) specifies the MAE metric [31, p. 20].

$$MAE = \frac{1}{K}\sum_{i=1}^{K}|v_i - w_i| \tag{10}$$

The smaller the MAE is, the more accurate the REM is. Table 6 shows MAE metric which evaluates the ten models with the estimation process.

Table 6. *MAE of ten models.*

Pair	MAE
0	162.7412
1	164.2515
2	167.6166
3	168.6956
4	169.4407
5	175.3171
6	176.9861

7	169.4873
8	267.0266
9	2121.2628
Average	374.2825

Let $rMAE_i$ be the bias ratio of MAE between the i^{th} pair and the 0^{th} pair. For example, we have [1]:

$$rMAE_i = \frac{MAE_i - MAE_0}{MAE_0} \quad (11)$$

Where MAE_i is MAE value of the i^{th} pair and MAE_0 is MAE value of the complete pair 0^{th}. For example,

$$rMAE_1 = \frac{164.2515 - 162.7412}{162.7412} \approx 0.0093$$

From equation (11), these bias ratios indicate the resistance of REM to incomplete data. For instance, the value $rMAE_1 = 0.0093$ implies that the accuracy of dual REM decreases 0.93% when the completion of training dataset of the 1st pair decreases 10%. The bias ratios of the pairs 1st (10% missing values), 2nd (20% missing values), 3rd (30% missing values), 4th (40% missing values), 5th (50% missing values), 6th (60% missing values), 7th (70% missing values), 8th (80% missing values) are 0.93%, 3.05%, 0.53%, 0.53%, 0.53%, 0.53%, and 7.43%. Like our previous research [1], we make a one-way paired t-test of X = {10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%} and Y = {0.93%, 3.00%, 3.66%, 4.12%, 7.73%, 8.75%, 4.15%, 64.08%}. Given significant level 95%, the statistic t_0 is calculated by equation (12) [32, p. 376].

$$t_0 = \frac{\overline{D}}{s_D/\sqrt{|X|}} \quad (12)$$

Where $|X| = |Y| = 8$ here and $D = X - Y$ whereas $\overline{D}$ and s_D are sample mean and sample standard deviation of D, respectively. For instance, we have:

$$D = \{0.0907, 0.1700, 0.2634, 0.3588, 0.4227, 0.5125, 0.6585, 0.1592\}$$

$$\overline{D} = 0.3295\,, s_D = 0.1953$$

$$t_0 = \frac{0.3295}{0.1953/\sqrt{8}} \approx 4.7708$$

Because the $t_0 = 4.7708$ is larger than the percentage point $t_{0.05,\ 3} = 2.353$ of t distribution [32, p. 711], difference between the percentage of missing values and the percentage of decrease in accuracy of REM is significant with pairs 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th. We assert that the resistance of REM to missing values given MAE metric is significant because the bias ratios are much smaller than percentages of missing values in case that loss ratios are equal to or smaller than 80%. When the loss ratio is too high (≥ 90%), REM produces unpredictably worse estimates. For instance, the MAE in Table 6 for loss ratio 90% is 2121.2628 which is an unacceptable value in fetal weight estimation.

We continue to assess such ten regression models with subject to R metric. Equation (13) specifies R metric [32, p. 432].

$$R = \frac{\sum_{i=1}^{K}(w_i - \overline{w})(v_i - \overline{v})}{\sqrt{\sum_{i=1}^{K}(w_i - \overline{w})^2}\sqrt{\sum_{i=1}^{K}(v_i - \overline{v})^2}} \quad (13)$$
$$\overline{w} = \sum_{i=1}^{K} w_i\,, \overline{v} = \sum_{i=1}^{K} v_i$$

The *R* reflects adequacy of a given formula. The larger the *R* is, the better the formula is. Table 7 shows R metric which evaluates our models with the estimation process.

Table 7. *R metric of ten models.*

Pair	R
0	0.9615
1	0.9612
2	0.9611
3	0.9602
4	0.9612
5	0.9612
6	0.9594
7	0.9568
8	0.9358
9	-0.9468
Average	0.7672

We make a one-way paired t-test of X = {10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%} and Y = {-0.03%, -0.04%, -0.14%, -0.03%, -0.03%, -0.22%, -0.49%, -2.67%} for R metric. Similarly, because the statistic $t_0 = 5.1191$ is larger than the percentage point $t_{0.05,\,3} = 2.353$, we asserted that the resistance of REM to missing values given R metric is significant in case that loss ratios are equal to or smaller than 80%. When the loss ratio is too high (≥ 90%), REM produces unpredictably worse estimates. For instance, the *R* in Table 7 for loss ratio 90% is –0.9468 which is unacceptable value due to reverse correlation.

As aforementioned we focus experimental results of REM with one-iteration inverse balance process. Here, Table 8 shows experimental MAE values from REM with one-iteration balance process (REM1), REM with balance process (REM2), REM with one-iteration inverse balance process (REM3), REM with inverse balance process (REM4). Note, Table 6 and Table 7 show MAE metric and R metric with regard to REM3. As aforementioned in section 2, REM1 leans to enhance the inverse models $X_j = \beta_{j0} + \beta_{j1}Z$ whereas REM3 leans to enhance the entire model $Z = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + \ldots + \alpha_nX_n$.

Table 8. *MAE metric of REM1, REM2, REM3, and REM4.*

Pair	REM1	REM2	REM3	REM4
0	162.7412	162.7412	162.7412	162.7412
1	167.2224	164.2526	164.2515	164.2515
2	196.4039	167.6164	167.6166	167.6166
3	228.2790	168.6874	168.6956	168.6959
4	233.7819	169.3606	169.4407	169.4411
5	248.7890	175.2555	175.3171	175.3160
6	414.2615	414.2615	176.9861	414.2615
7	358.4372	169.8922	169.4873	169.4820
8	236.2435	165.6578	267.0266	267.6966
9	389.4869	2107.9578	2121.2628	2122.5030
Average	263.5647	386.5683	374.2825	398.2005

MAE values in Table 8 are used to make comparison among REM1, REM2, REM3, and REM4. Except the 9th pair, REM3 gives out best result (least MAE) and REM1 gives out worst result (greatest MAE). The result from REM2 which is REM1 with full balance process is approximate to the result from REM3. Similarly, the result

from REM4 which is REM3 with full inverse balance process is near to the result from REM3. Hence, REM3 and REM4 give out similar results. This implies that the (inverse) balance process does really make trade-off between the entire model $Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ and many inverse models $X_j = \beta_{j0} + \beta_{j1} Z$ so as to reach possible result. For the 9th pair, conversely REM1 gives out best result whereas REM3 gives our worst result, which implies that too sparse sample whose loss ratio is equal to or larger than 90% can produces unpredictable result in regression analysis. When I make the sample randomly sparse many times, REM3 and REM1 can exchange experimental results; concretely REM3 can give out worst result (best result) and REM1 can give out best result (worst result). Anyway, REM2 and REM4 always give out average (trade-off) results. In some cases, REM2 or REM4 can give out best result. In general, the (inverse) balance process for REM is recommended to researchers.

For REM1, REM2, and REM4, the difference between the percentage of missing values and the percentage of decrease in accuracy is insignificant with pairs 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th given paired t-test. In other words, only best model (derived from REM3 here) surely brings out the resistance of REM to missing values given MAE metric with loss ratios up to 80%. Table 9 shows statistic t_0 of REM1, REM2, REM3, and REM4 with pairs 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th.

Table 9. *Statistic t_0 of REM1, REM2, REM3, and REM4 given MAE metric.*

REM1	REM2	REM3	REM4
-1.0695	1.2065	4.7708	0.8698

Figure 1 shows comparison among REM1, REM2, REM3, and REM4 derived from Table 8.

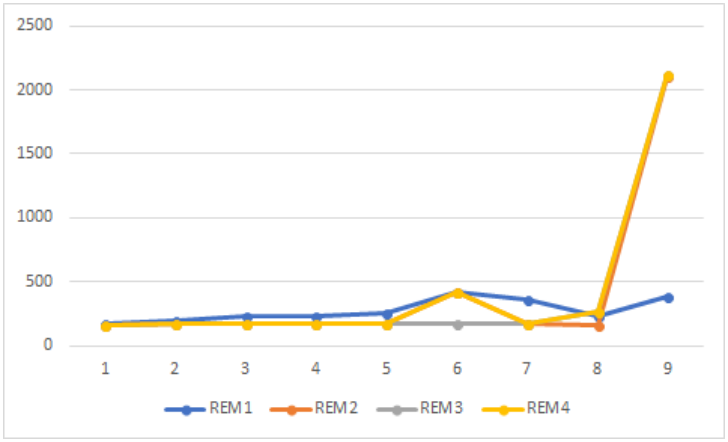

Figure 1. *Comparison among REM1, REM2, REM3, and REM4 given MAE metric.*

As seen in Figure 1, the line of REM2 approximates to the line of REM4.

Here, Table 10 shows experimental R values from REM1, REM2, REM3, and REM4.

Table 10. *R metric of REM1, REM2, REM3, and REM4.*

Pair	REM1	REM2	REM3	REM4
0	0.9615	0.9615	0.9615	0.9615
1	0.9599	0.9612	0.9612	0.9612
2	0.9494	0.9611	0.9611	0.9611
3	0.9314	0.9602	0.9602	0.9602
4	0.9300	0.9612	0.9612	0.9612
5	0.9277	0.9612	0.9612	0.9612

6	0.8575	0.8575	0.9594	0.8575
7	0.8414	0.9566	0.9568	0.9568
8	0.9254	0.9601	0.9358	0.9355
9	0.8372	-0.9468	-0.9468	-0.9469
Average	0.9121	0.7594	0.7672	0.7569

Given R metric, REM2 and REM4 always give out similar a result, which implies again that the (inverse) balance process, makes trade-off between the entire model and many inverse models so as to reach possible result. Given paired t-test, the difference between the percentage of missing values and the percentage of decrease in accuracy is significant with pairs 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th for all REM1, REM2, REM3, and REM4. So, the resistance of REM to missing values given R metric is asserted with loss ratios up to 80%. Unacceptable R values such as -0.9468 and -0.9469 with the 9th pair indicate that too sparse sample whose loss ratio is equal to or larger than 90% can produces unpredictable result in regression analysis. Table 11 shows statistic t_0 of REM1, REM2, REM3, and REM4 with pairs 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th given R metric.

Table 11. *Statistic t_0 of REM1, REM2, REM3, and REM4 given R metric.*

REM1	REM2	REM3	REM4
5.0795	5.0984	5.1191	5.0382

Figure 2 shows comparison among REM1, REM2, REM3, and REM4 derived from Table 10 given R metric.

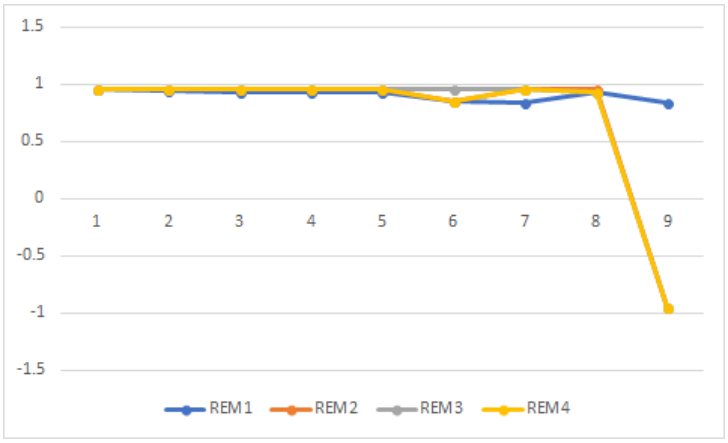

Figure 2. *Comparison among REM1, REM2, REM3, and REM4 given R metric.*

As seen in Figure 2, the line of REM2 approximates to the line of REM4.

The REM3 here produces the best model with lowest MAE up to 80% loss ratio. However, recall that the full (inverse) balance process for REM is recommended to researchers. For instance, when I re-split the gestational sample of Ho and Phan [29] into larger training dataset (70% sample) and smaller testing dataset (30% sample), REM2 with full balance process now gives out the best result (least MAE) except the 9th pair as seen in Table 12.

Table 12. *MAE metric of REM1, REM2, REM3, and REM4 with larger training dataset.*

Pair	REM1	REM2	REM3	REM4
0	169.5586	169.5586	169.5586	169.5586
1	170.4311	170.0891	170.0899	170.0899
2	195.6338	170.6894	170.7017	170.7019
3	234.4279	169.9932	170.0369	170.0370

4	242.2376	172.8648	172.9491	172.9500
5	269.5383	183.0067	183.1272	183.1268
6	261.8110	254.8248	182.1901	254.8248
7	248.2748	176.6183	214.4385	214.4239
8	270.7730	184.1931	363.5646	364.2376
9	1767.8524	2065.6439	2245.0253	2243.3655
Average	383.0539	371.7482	404.1682	411.3316

Table 13 shows the statistic t_0 of REM1, REM2, REM3, and REM4 with pairs 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th given MAE metric and larger training dataset (70% sample).

Table 13. *Statistic t_0 of REM1, REM2, REM3, and REM4 given MAE metric and larger training dataset.*

REM1	REM2	REM3	REM4
1.2671	4.2909	2.5556	2.1507

Now two statistics t_0 of REM2 and REM3 are larger than the percentage point $t_{0.05, 3} = 2.353$ of t distribution. Hence, the larger the training dataset is, the better the resistance of REM to missing values is with loss ratios up to 80%. Figure 3 shows again that the line of REM2 approximates to the line of REM4.

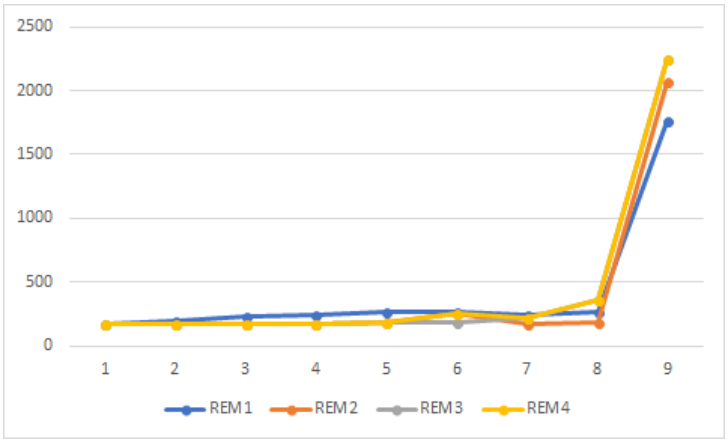

Figure 3. *Comparison among REM1, REM2, REM3, and REM4 given MAE metric and larger training dataset.*

From Figure 3, we concluded again REM2 and REM4 are stable. They produce good enough models or best models. REM3 often gives out best result because it leans to improve the entire model $Z = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + ... + \alpha_nX_n$. Therefore, *REM2 and REM3* are good choices in practice when REM2 leans to improve REM1 and REM4 makes trade-off between REM2 and REM3.

For high loss ratio (≥ 90%), REM1 often results out best models, which is not explained exactly yet. For instance, as seen in Table 10, REM1 gives out good correlation $R = 0.8372$ for the 9th pair whereas other ones give out unacceptable reverse correlation. Similarly, R values for the 9th pair of REM1, REM2, REM3, and REM4 given the larger training dataset (70% sample) are 0.8586, -0.9303, -0.929, and -0.9291, respectively. When training dataset is made sparse with high loss ratio (≥ 90%), the long entire model $Z = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + ... + \alpha_nX_n$ is harmed more than the shorter inverse models $X_j = \beta_{j0} + \beta_{j1}Z$. Moreover, damage caused by high loss ratio is stretched out across many short inverse models and so such damage is alleviated with the inverse models. Hence, because REM1 leans to improve the inverse models $X_j = \beta_{j0} + \beta_{j1}Z$, REM1 often results out best models with high loss ratio (≥ 90%). However, in this research, we do not evaluate the inverse models $X_j = \beta_{j0} + \beta_{j1}Z$ yet

and so the research is still open. In general, REM1 and REM3 are opposite to each other and the (inverse) balance process, which is the core of REM, links them together to produce the trade-off REM2 and REM4.

4. Conclusions

In general, from experimental results on two typical evaluation metrics such as MAE and R, we conclude that REM solves totally the problem in which fetal weight, fetal ages, and ultrasound measures can be missing when the loss ratio is up to 80%. This problem was raised in our previous research [1]. As a result, practitioners will have a lot of benefits when they will not be stressful in taking ultrasound examinations. In other words, it is acceptable for practitioners to make unintentional mistakes when taking ultrasound examinations. Of course, early weight estimation is achieved because ultrasound examination can be taken at any time of gestational period because it is not mandatory to know fetal weights. When the resistance of REM to missing values is proved, we will improve REM with prior distribution of coefficients (α, β_j) and compare REM with other algorithms for further research. When the loss ratio is too high ($\geq$ 90%), I think that we should not construct regression model from too sparse sample because such sample will produce unpredictable biases. The website of REM 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 show our deep gratitude to Prof. Bich-Ngoc Tran who gave us comments to evaluate the resistance of DREM algorithm to missing values. Note that DREM is proposed in our previous research "Early Fetal Weight Estimation with Expectation Maximization Algorithm" published in Experimental Medicine (EM) Journal of International Technology and Science Publications (ITS) on 7th May 2018 but REM and DREM share the same testing way that is one-way paired t-test.

References

[1] 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.

[2] Hadlock, F. P.; Harrist, R. B.; Sharman, R. S.; Deter, R. L.; Park, S. K. Estimation of fetal weight with use of head, body and femur measurements: A prospective study. *American Journal of Obstetrics and Gynecology*, 1st February 1985, 151(3), 333-337, DOI: 10.1016/0002-9378(85)90298-4.

[3] Phan, D. T. *Application of Ultrasonography to Diagnose Fetal Age and Weight in Mother Womb*. Hanoi Medical University: Hanoi, 1985.

[4] Pham, T.-N. T. *Fetal Weight Estimation by Ultrasound Measures*. Ho Chi Minh University of Medicine and Pharmacy: Ho Chi Minh, 2000.

[5] Ho, T. H. T. *Research on Fetal Age and Weight Estimation by Two-Dimensional and Three-Dimensional Ultrasound Measures*. Hanoi Medical University: Hanoi, 2011, DOI: 10.13140/RG.2.2.33184.48645.

[6] Deter, R. L.; Rossavik, I. K.; Harrist, R. B. Development of individual growth curve standards for estimated fetal weight: I. Weight estimation procedure. *Journal of Clinical Ultrasound*, 1988, 16(4), 215-225. Available online: https://www.ncbi.nlm.nih.gov/pubmed/3152508 (accessed on Day Month Year).

[7] Chien, P. F. W.; Owen, P.; Khan, K. S. Validity of Ultrasound Estimation of Fetal Weight. *Obstetrics & Gynecology*, 2000, 95(6), 856-860, DOI: 10.1016/S0029-7844(00)00828-0.

[8] Varol, F.; Saltik, A.; Kaplan, P. B.; Kilic, T.; Yardim, T. Evaluation of Gestational Age Based on Ultrasound Fetal Growth Measurements. *Yonsei Medical Journal*, 2001, 42(3), 299-303, DOI:10.3349/ymj.2001.42.3.299.

[9] Dudley, N. J. A systematic review of the ultrasound estimation of fetal weight. Ultrasound in Obstetrics and Gynecology: *The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 2004, 25(1), 80-89, DOI:10.1002/uog.1751.

[10] Salomon, L. J.; Bernard, J. P.; Ville, Y. Estimation of fetal weight: reference range at 20–36 weeks' gestation and comparison with actual birth-weight reference range. *Ultrasound in obstetrics & gynecology*, 2007, 29(5), 550-555, DOI: 10.1002/uog.4019.

[11] Akinola, R. A.; Akinola, O. I.; Oyekan, O. O. *Sonography in fetal birth weight estimation. Educational Research and Review*, 2009, 4(1), 16-20.

[12] Lee, W.; Balasubramaniam, M.; Deter, R. L.; Yeo, L.; Hassan, S. S.; Gotsch, F.; Kusanovic, J. P.; Gonçalves, L. F.; Romero, R. New fetal weight estimation models using fractional limb volume. *Ultrasound in Obstetrics & Gynecology*, 2009, 34(5), 556-565, DOI: 10.1002/uog.7327.

[13] Bennini, J. R.; Marussi, E. F.; Barini, R.; Faro, C.; Peralta, C. A. F. Birth-weight prediction by two- and three-dimensional ultrasound imaging. *Ultrasound in Obstetrics & Gynecology*, 2009, 35(4), 426-433, DOI: 10.1002/uog.7518.

[14] Cohen, J. M.; Hutcheon, J. A.; Kramer, M. S.; Joseph, K. S.; Abenhaim, H.; Platt, R. W. Influence of ultrasound-to-delivery interval and maternal–fetal characteristics on validity of estimated fetal weight. *Ultrasound in Obstetrics & Gynecology*, 2010, 35(4), 434-441, DOI: 10.1002/uog.7506.

[15] Siggelkow, W.; Schmidt, M.; Skala, C.; Boehm, D.; Forstner, S. v.; Koelb, H.; Tresch, A. A new algorithm for improving fetal weight estimation from ultrasound data at term. *Archives of gynecology and obstetrics*, 2010, 283(3), 469-474, DOI: 10.1007/s00404-010-1390-8.

[16] Wu, M.; Shao, G.; Zhang, F.; Ruan, Z.; Xu, P.; Ding, H. Estimation of fetal weight by ultrasonic examination. *International Journal of Clinical and Experimental Medicine*, 2015, 8(1), 540-545. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4358483 (accessed on 11 July 2018).

[17] 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 11 July 2018).

[18] Hutcheon, J. A.; Platt, R. W. The Missing Data Problem in Birth Weight Percentiles and Thresholds for "Small-for-Gestational-Age. *American Journal of Epidemiology*, 2008, 167(7), 786-792, DOI: 10.1093/aje/kwm327.

[19] Eberg, M.; Platt, R. W.; Filion, K. B. The Estimation of Gestational Age at Birth in Database Studies. *Epidemiology*, 2017, 28(6), 854-862, DOI: 10.1097/EDE.0000000000000713.

[20] 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 11 July 2018).

[21] 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 11 July 2018).

[22] 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.

[23] 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.

[24] 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 11 July 2018).

[25] 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.

[26] 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.

[27] 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).

[28] 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.

[29] 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*, December, 2011, 12(797), 8-9.

[30] 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.

[31] Herlocker, J. L.; Konstan, J. A.; Terveen, L. G.; Riedl, J. T. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems (TOIS)*, 2004, 22(1), 5-53, DOI: 10.1145/963770.963772.

[32] Montgomery, D. C.; Runger, G. C. *Applied Statistics and Probability for Engineers*, 5th ed.; John Wiley & Sons: Hoboken, New Jersey, USA, 2010; p. 792. Available online: https://books.google.com.vn/books?id=_f4KrEcNAfEC (accessed on 11 July 2018).