Fetal Weight Estimation in Case of Missing Data

**Loc Nguyen1**\***, Thu-Hang T. Ho2**

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](mailto:ng_phloc@yahoo.com) (Loc Nguyen), [bshangvl2000@yahoo.com](mailto:bshangvl2000@yahoo.com) (Thu-Hang T. Ho)

\*Correspondence: [Timothy.Schmutte@yale.edu](file:///E:\文章\文章\ITS精排\ITS%2020180502%20精排文章\EM1003\Timothy.Schmutte@yale.edu)

**Received:; Accepted:; Published:**

**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 a so-called regression expectation maximization (REM) algorithm which is a combination of maximum likelihood estimation (MLE) 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 filled in 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 high loss ratio 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 = α*0 *+ α*1*X*1 *+ α*2*X*2 *+ … + αnXn* where *Z* is estimated fetal weight whereas *Xi* (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 *Xi* is called regression variable, regressor, predictor, regression variable, or independent variable. Each *αi* is called regression coefficient. In previous research [1] we survey many researches related to the regression approach. Hence, 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 *Xi* 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*. We need to repeat the survey of fetal weight and age estimation based on regression analysis. The 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 amazing 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 amazing regression formulas for estimating fetal weight based on *bpd*, ad, arm length (*al*), abdominal diameter (*ad*), average abdominal diameter (*aad*). Ho [5] produced some amazing regression formulas for estimating fetal age and weight based on *bpd*, *ac*, *hc*, thigh volume, and thigh volume in her PhD dissertation. Some of Ho’s formulas [5] 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”.

Varol et al. [7] 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 R2 = 0.937 is optimal.

Salomon, Bernard, and Ville [8] 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 50th centile of actual birth weight is approximated to the 50th centile of estimated weight.

A. R. Akinola, I. O. Akinola, and O. O. Oyekan [9] 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. [10] 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. [11] 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. [12] 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. [13] 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).

Pinette et al. [14] 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 10th 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 [15] 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 [16] 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 [17] 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 [18] 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 [19] used reject inference technique to estimate coefficients of logistic regression model when response variable *Z* is missing but characteristic variables (regressors *Xi*) are fully observed. Anderson and Hardin replaced missing Z by its conditional expectation on regressors *Xi* where such expectation is logistic function. Zhang, Deng, and Su [20] used EM algorithm to build up linear regression model for studying glycosylated hemoglobin from partial missing data. In other words, they 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 [21] 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 [22] 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 *Xi* when *Z* has missing values. Robins, Rotnitzki, and Zhao [22] assumed that the probability *λit*(*α*) 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 *λit*(*α*) is also determined and so the coefficients *β* is calculated based on the inverse of *λit*(*α*) and *Xi*. The inverse of *λit*(*α*) is considered as weight for complete case. Robins, Rotnitzki, and Zhao used additional time-dependent covariates *Vit* to determine *λit*(*α*).

In the article “Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models”, Horton and Kleinman [23] 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 [23, 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 [23, 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 [23, 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 [23, 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 statistic 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 [22] 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 [23, 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 *Xi* 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 high loss ratio 80%. The special technique in REM is to build parallelly an entire regression function *Z* = *α*0 *+ α*1*X*1 *+ α*2*X*2 *+ … + αnXn* and many partial inverse regression functions *Xj* = *βj*0 *+ βj*1*Z* for solving the problem of highly sparse data, in which missing values are filled in 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.

1. **Methodology**

Suppose we estimate the linear regression model *Z* = *α*0 *+ α*1*X*1 *+ α*2*X*2 *+ … + αnXn* where *Z* is fetal weight and *Y* is fetal age whereas *Xi* (s) are gestational ultrasound measures such as *bpd*, *hc*, *ac*, and *fl*. Suppose the random variable *Z* conforms normal distribution, according to equation (1) [24, pp. 8-9]. Note, *Z* is random variable whereas *X* is data in equation (1).

|  |  |
| --- | --- |
|  | (1) |

Where *α* = (*α*0, *α*1,…, *αn*)*T* is parameter vector and *X* = (1, *X*1, *X*2,…, *Xn*)*T* is data vector. The mean and variance of *Z* with regard to *P*(*Z* | *X*, *α*) are *αTX* and *σ*2, respectively. The superscript “*T*” denotes transposition operator in vector and matrix. Suppose each has an inverse linear regression model *Xj* = *βj*0 *+ βj*1*Z.* In other words, *Zj* now is considered as the random variable conforming normal distribution according to equation (2).

|  |  |
| --- | --- |
|  | (2) |

Where *βj* = (*βj*0, *βj*1)*T* is a partial parameter vector and (1, *Z*)*T* is a partial data vector. The mean and variance of each *Xj* with regard to the inverse distribution *Pj*(*Xj* | *Z*, *βj*) are *βjT*(1, *Z*)*T* and *σj*2, respectively. Of course, there are *n* inverse linear regression models.

Let ***D*** = (***X***, ***z***) be collected sample in which ***X*** is a set of sample measures and ***z*** is a set of fetal weights with note that both ***X*** and ***z*** are incomplete. In other words, ***X*** and ***z*** have missing values. Now we focus on estimating *α* and *βj* based on ***D***. As a convention let *α\** and *βj\** be estimates of *α* and *βj*, respectively [24, p. 8].

|  |  |
| --- | --- |
|  | (3) |

The expectation of sufficient statistic *Z* regard to the entire linear model *P*(*Z* | *X*, *α*) is specified by equation (4).

|  |  |
| --- | --- |
|  | (4) |

The expectation of each sufficient statistic *Xj* with regard to each inverse linear model *Pj*(*Xj* | *Z*, *βj*) is specified by equation (5).

|  |  |
| --- | --- |
|  | (5) |

Please pay attention to equations (4) and (5) because *Z* and *Xj* 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 *βj\** [24, pp. 8-9].

|  |  |
| --- | --- |
|  | (6) |

Where ***X***, ***z***, ***Z***, and ***x****j* are specified in equation (3). Because ***X*** and ***Z*** are incomplete, we apply expectation maximization (EM) algorithm into estimating (*α\**, *βj\**)*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 Θ(*t*) = (*α*(*t*), *βj*(*t*))*T* at the *t*th iteration, missing values *zi* and *xij* are calculated in E-step so that ***X*** and ***Z*** become complete. In M-step, the next parameter Θ(*t*+1) = (*α*(*t*+1), *βj*(*t*+1))*T* is determined by equation (6) and the complete data ***X*** and ***Z***.

The most important problem in our research is how to estimate missing values *zi* and *xij*. Recall that every missing value *zi* is estimated as the expectation based on the current parameter *α*(*t*), according to equation (4).

Let *Ui* be a set of indices of missing values *xij*. In other words, if then, *xij* is missing. The set *Ui* can be empty. The equation (4) is written:

Note, *xi*0 = 1. According to equation (5), missing value *xij* is estimated by:

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

It implies:

|  |  |
| --- | --- |
|  | (7) |

Missing values *zi* and *xij* are estimated by the balance process shown in Table 1.

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

|  |
| --- |
| 1. Step 1: Missing values *zi* are estimated by equation (7), based on the current parameter Θ(*t*) = (*α*(*t*), *βj*(*t*))*T*.   Missing values *xij* where are estimated by equation (5) and the estimated values *zi* above, based on the current parameter Θ(*t*) = (*α*(*t*), *βj*(*t*))*T*.   1. Step 2: For balancing both *P*(*Z* | *X*, *α*) and *Pj*(*Xj* | *Z*, *βj*) in estimation, values *zi* and *xij* are re-estimated by equations (4) and (5) as new *zi*’ and *xij*’, based on the current parameter Θ(*t*) = (*α*(*t*), *βj*(*t*))*T*. 2. Step 3: If the deviation between (*zi*’, *xij*’) and (*zi*, *xij*) is smaller than a small enough threshold or the process reaches a large enough number of iterations, the process stops; at that time *zi*’ and *xij*’ are final estimated values. Otherwise, going back step 2 with assignment *xij* = *xij*’. |

In fact, the balance process is an iterative process which is a combination of equations (4), (5), and (7). The process starts to estimate missing values *zi* without use of *xij*. Conversely, the process can start to estimate missing values *xij* without use of *zi*, which is called inverse balance process.

Recall that *Ui* is the set of indices of missing values *xij*. Every missing value *xil* is estimated as the expectation based on the current parameter *βj*(*t*), according to equation (5).

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

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

In other words, we have:

Where,

Suppose the cardinality of *Ui* is *k*, which means that there is *k* missing values *xij*. Derived from the combination above, missing values are solution of the following system of *k* equations.

Therefore, missing values *xij* are calculated by equation (8) according to Cramer method.

|  |  |
| --- | --- |
|  | (8) |

Where,

Table 2 shows the inverse balance process.

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

|  |
| --- |
| 1. Step 1: Missing values *xij* where are estimated by equation (8), based on the current parameter Θ(*t*) = (*α*(*t*), *βj*(*t*))*T*. Missing values *zi* are estimated by equation (7), based on the current parameter Θ(*t*) = (*α*(*t*), *βj*(*t*))*T*.   Missing values *zi* are estimated by equation (4) and the estimated values *xij* above, based on the current parameter Θ(*t*) = (*α*(*t*), *βj*(*t*))*T*.   1. Step 2: For balancing both *P*(*Z* | *X*, *α*) and *Pj*(*Xj* | *Z*, *βj*) in estimation, values *xij* and *zi* are re-estimated by equations (5) and (4) as new *xij*’ and *zi*’, based on the current parameter Θ(*t*) = (*α*(*t*), *βj*(*t*))*T*. 2. Step 3: If the deviation between (*zi*’, *xij*’) and (*zi*, *xij*) is smaller than a small enough threshold or the process reaches a large enough number of iterations then, the process stops; at that time *zi*’ and *xij*’ are final estimated values. Otherwise, going back step 2 with assignment *zi* = *zi*’. |

In fact, the inverse balance process is an iterative process which is a combination of equations (4), (5), and (8). We use the balance process shown in Table 1 for experiments in this research with note that the balance process leans to enhance the inverse models *Xj* = *βj*0 *+ βj*1*Z* and the inverse balance process leans to enhance the entire model *Z* = *α*0 *+ α*1*X*1 *+ α*2*X*2 *+ … + αnXn*. As a result, EM algorithm [25, p. 4] associated with the 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 *zi* and *xij* are estimated by the (inverse) balance process shown in Table 1 (Table 2). The balance process is the core of REM. 2. The next parameter Θ(*t*+1) = (*α*(*t*+1), *βj*(*t*+1))*T* is determined by equation (6) and the complete data ***X*** and ***Z*** fulfilled in E-step. |

The REM algorithm stops if at some *t*th iteration, we have Θ(*t*) = Θ(*t*+1) = Θ*\**. At that time, Θ*\** = (*α\**, *β\**)*T* is the optimal estimate of EM algorithm. In practice, the algorithm can stop if deviation between Θ(*t*) and Θ(*t*+1) is smaller than a small enough threshold or REM reaches a large enough number of iterations. The smaller the terminated threshold is, the more accurate REM is. REM uses both the terminated threshold *ε* = 0.1% = 0.001 and the maximum number of iterations *maximum-iteration* = 10000. The parameter *maximum-iteration* = 10000 prevents REM from running for a long time.

An technique to improve the convergence of REM is to initialize the parameter Θ(1) = (*α*(1), *β*(1))*T* at the first iteration of EM process in proper way instead of initializing Θ(1) in arbitrary way [1]. As usual, Θ(1) is initialized arbitrarily. Let ***X***’ be the complete matrix of ultrasound measures, which is created by removing all rows whose values are missing from ***X***. Similarly, let ***Z***’ be the complete matrix of fetal weights, which is created by removing rows whose weights are missing from ***Z***. The advanced Θ(1) = (*α*(1), *β*(1))*T* is initialized by equation (9).

|  |  |
| --- | --- |
|  | (9) |

Where ***z***’ is the complete vector of non-missing weights and ***x****j*’ is the complete vector of non-missing measures. Equation (9) is variant of equation (6) where ***X***, ***Z***, ***x****j*, and ***z*** are replaced by ***X***’, ***Z***’, ***x****j*’, and ***z***’.

1. **Results and Discussion**

We use a gestational sample of 127 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 [26], [27] collected the sample of pregnant women at Vinh Long General Hospital – Vietnam with obeying strictly all medical ethical criteria. 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 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9, 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 has 2 rows and each row has 5 columns (*bpd*, *hc*, *ac*, *fl*, *weight*) and so the dataset has 10 cells. If loss ratio is 20%, there are only 2 missing values which are made randomly among such 10 cells. 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 0th pair.

**Table 4.** Ten testing pairs

|  |  |  |  |
| --- | --- | --- | --- |
| 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% |

The 0th pair is called completed pair whereas the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, and 9th 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 = α*0 *+ α*1*X*1 *+ α*2*X*2 *+ … + αnXn* and many inverse models *Xj* = *βj*0 *+ βj*1*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 5 shows ten regression models corresponding to ten testing pairs with the estimation process.

**Table 5.** Ten resulted regression models

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

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,…, *wK*} and *V* = {*v*1, *v*2,…, *vK*} be sets of actual weights and estimated weights, respectively. Equation (10) specifies the MAE metric [28, p. 20].

|  |  |
| --- | --- |
|  | (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 *rMAEi* be the bias ratio of MAE between the *i*th pair and the 0th pair. For example, we have [1]:

|  |  |
| --- | --- |
|  | (11) |

Where *MAEi* is MAE value of the *i*th pair and *MAE*0 is MAE value of the complete pair 0th. For example,

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%, 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) [29, p. 376].

|  |  |
| --- | --- |
|  | (12) |

Where *D* = *X* – *Y* whereas and *sD* are sample mean and sample standard deviation of *D*, respectively. For instance, we have:

Because the *t*0 = 3.3735 is larger than the percentage point *t*0.05, 3 = 2.353 of *t* distribution [29, 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 [29, p. 432].

|  |  |
| --- | --- |
|  | (13) |

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 = 3.9173 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 reversed 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, Tables 6 and 7 show MAE metric and R metric with regard to REM3.

**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 |

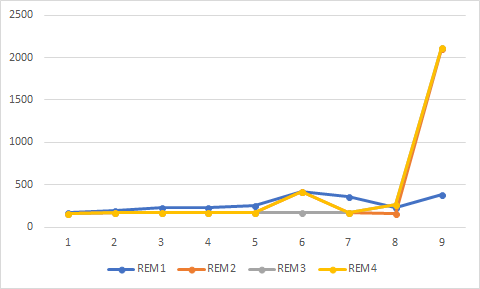
MAE values in Table 8 are used to make comparison among REM1, REM2, REM3, and REM4. Except the 9th pair, REM3 gives out best result (least MAE) and REM1 gives out worst result (greatest MAE). The result from REM2 which is REM1 with full balance process is approximate to the result from REM3. Similarly, 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 = α*0 *+ α*1*X*1 *+ α*2*X*2 *+ … + αnXn* and many inverse models *Xj* = *βj*0 *+ βj*1*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 and REM2 can give out best result. Anyway, REM2 and REM4 always give out average (trade-off) results. In some cases, REM2 or REM4 gives 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 bring out the resistance of REM to missing values given MAE metric. 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 |
| -0.7563 | 0.8531 | 3.3735 | 0.6151 |

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 coincides approximately with 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

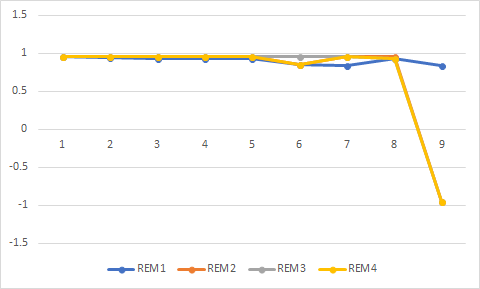
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pair | REM1 | REM2 | REM3 | REM4 |
| 0 | 0.9615 | 0.9615 | 0.9615 | 0.9615 |
| 1 | 0.9599 | 0.9612 | 0.9612 | 0.9612 |
| 2 | 0.9494 | 0.9611 | 0.9611 | 0.9611 |
| 3 | 0.9314 | 0.9602 | 0.9602 | 0.9602 |
| 4 | 0.9300 | 0.9612 | 0.9612 | 0.9612 |
| 5 |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

Given R metric, REM2 and REM4 always give out similar results, 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. Unacceptable R values such as -0.9468 and -0.9469 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 |
| 3.5918 | 3.6051 | 3.9173 | 3.5625 |

Figure 11 shows comparison among REM1, REM2, REM3, and REM4 derived from Table 10 given R metric.



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

As seen in Figure 2, the line of REM2 coincides approximately with the line of REM4.

1. **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*, *βj*) and compare REM with other algorithms for further research. When the loss ratio is too high (≥ 90%), I think that we should not construct regression model from too sparse sample because such sample will produce unpredictable biases.

**Conflicts of Interest**

The authors Loc Nguyen and Thu-Hang T. Ho 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] | L. Nguyen and T.-H. T. Ho, "Early Fetal Weight Estimation with Expectation Maximization Algorithm," *Experimental Medicine (EM),* vol. 1, no. 1, pp. 12-30, 7 May 2018. |
| [2] | F. P. Hadlock, R. B. Harrist, R. S. Sharman, R. L. Deter and S. K. Park, "Estimation of fetal weight with use of head, body and femur measurements: A prospective study," *American Journal of Obstetrics and Gynecology,* vol. 151, no. 3, pp. 333-337, 1 February 1985. |
| [3] | D. T. Phan, "Ứng dụng siêu âm để chẩn đoán tuổi thai và cân nặng thai trong tử cung," Hanoi University of Medicine, Hanoi, 1985. |
| [4] | T. N. T. Phạm, "Ước lượng cân nặng thai nhi qua các số đo của thai trên siêu âm," Ho Chi Minh University of Medicine and Pharmacy, Ho Chi Minh, 2000. |
| [5] | T. H. T. Ho, "Nghiên Cứu Phương Pháp Ước Lượng Trọng Lượng Thai, Tuổi Thai Bằng Siêu Âm Hai và Ba Chiều," Hanoi Univerisy of Medicine, Hanoi, 2011. |
| [6] | R. L. Deter, I. K. Rossavik and R. B. Harrist, "Development of individual growth curve standards for estimated fetal weight: I. Weight estimation procedure," *Journal of Clinical Ultrasound,* vol. 16, no. 4, pp. 215-225, 1 May 1988. |
| [7] | F. Varol, A. Saltik, P. B. Kaplan, T. Kilic and T. Yardim, "Evaluation of Gestational Age Based on Ultrasound Fetal Growth Measurements," *Yonsei Medical Journal,* vol. 42, no. 3, pp. 299-303, June 2001. |
| [8] | L. J. Salomon, J. P. Bernard and Y. Ville, "Estimation of fetal weight: reference range at 20–36 weeks' gestation and comparison with actual birth-weight reference range," *Ultrasound in obstetrics & gynecology,* vol. 29, no. 5, pp. 550-555, 1 May 2007. |
| [9] | R. A. Akinola, O. I. Akinola and O. O. Oyekan, "Sonography in fetal birth weight estimation," *Educational Research and Review,* vol. 4, no. 1, pp. 16-20, January 2009. |
| [10] | W. Lee, M. Balasubramaniam, R. L. Deter, L. Yeo, S. S. Hassan, F. Gotsch, J. P. Kusanovic, L. F. Gonçalves and R. Romero, "New fetal weight estimation models using fractional limb volume," *Ultrasound in Obstetrics & Gynecology,* vol. 34, no. 5, pp. 556-565, 1 November 2009. |
| [11] | J. R. Bennini, E. F. Marussi, R. Barini, C. Faro and C. A. F. Peralta, "Birth-weight prediction by two- and three-dimensional ultrasound imaging," *Ultrasound in Obstetrics & Gynecology,* vol. 35, no. 4, pp. 426-433, 9 December 2009. |
| [12] | J. M. Cohen, J. A. Hutcheon, M. S. Kramer, K. S. Joseph, H. Abenhaim and R. W. Platt, "Influence of ultrasound-to-delivery interval and maternal–fetal characteristics on validity of estimated fetal weight," *Ultrasound in Obstetrics & Gynecology,* vol. 35, no. 4, pp. 434-441, 1 April 2010. |
| [13] | W. Siggelkow, M. Schmidt , C. Skala, D. Boehm, S. v. Forstner, H. Koelb and A. Tresch, "A new algorithm for improving fetal weight estimation from ultrasound data at term," *Archives of gynecology and obstetrics,* vol. 283, no. 3, pp. 469-474, 20 February 2010. |
| [14] | M. G. Pinette, Y. Pan, S. G. Pinette, J. Blackstone, J. Garrett and A. Cartin, "Estimation of Fetal Weight: Mean Value from Multiple Formulas," *Journal of Ultrasound in Medicine,* vol. 18, no. 12, pp. 813-817, 1 December 1999. |
| [15] | J. A. Hutcheon and R. W. Platt, "The Missing Data Problem in Birth Weight Percentiles and Thresholds for "Small-for-Gestational-Age"," *American Journal of Epidemiology,* vol. 167, no. 7, pp. 786-792, 1 April 2008. |
| [16] | M. Eberg, R. W. Platt and K. B. Filion, "The Estimation of Gestational Age at Birth in Database Studies," *Epidemiology,* vol. 28, no. 6, pp. 854-862, 1 November 2017. |
| [17] | P. Kokic, "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. |
| [18] | M. E. Ghitany, D. Karlis, D. K. Al-Mutairi and F. Al-Awadhi, "An EM Algorithm for Multivariate Mixed Poisson Regression Models and its Application," *Applied Mathematical Sciences,* vol. 6, no. 137, pp. 6843-6856, 2012. |
| [19] | B. Anderson and M. J. Hardin, "Modified logistic regression using the EM algorithm for reject inference," *International Journal of Data Analysis Techniques and Strategies,* vol. 5, no. 4, pp. 359-373, 1 January 2013. |
| [20] | X. Zhang, J. Deng and R. Su, "The EM algorithm for a linear regression model with application to a diabetes data," in *The 2016 International Conference on Progress in Informatics and Computing (PIC)*, Shanghai, China, 2016. |
| [21] | Y. Haitovsky, "Missing Data in Regression Analysis," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 30, no. 1, pp. 67-82, 1 January 1968. |
| [22] | J. M. Robins, A. Rotnitzki and L. P. Zhao, "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association,* vol. 90, no. 429, pp. 106-121, March 1995. |
| [23] | N. J. Horton and K. P. Kleinman, "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models," *The American Statistician,* vol. 61, no. 1, pp. 79-90, February 2007. |
| [24] | F. Lindsten, T. B. Schön, A. Svensson and N. Wahlström, "Probabilistic modeling – linear regression & Gaussian processes," Uppsala University, Uppsala, 2017. |
| [25] | A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B (Methodological),* vol. 39, no. 1, pp. 1-38, 1977. |
| [26] | T. H. T. Ho and D. T. Phan, "Fetal Weight Estimation from 37 Weeks to 42 Weeks by Two-Dimensional Ultrasound Measures," *Journal of Practical Medicine,* vol. 12, no. 797, pp. 8-9, December 2011. |
| [27] | T. H. T. Ho and D. T. Phan, "Fetal Age Estimation by Three-Dimensional Ultrasound Measure of Arm Volume and Other Two-Dimensional Ultrasound Measures," *Journal of Practical Medicine,* vol. 12, no. 798, pp. 12-15, December 2011. |
| [28] | J. L. Herlocker, J. A. Konstan, L. G. Terveen and J. T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," *ACM Transactions on Information Systems (TOIS),* vol. 22, no. 1, pp. 5-53, 2004. |
| [29] | D. C. Montgomery and G. C. Runger, Applied Statistics and Probability for Engineers, 5th ed., Hoboken, New Jersey: John Wiley & Sons, 2010, p. 792. |

|  |  |
| --- | --- |
| F:\期刊所需资料\文章template\已准备的资料\完成文档\图片\ccby4.0.jpg | © 2018 by Loc Nguyen and Thu-Hang T. Ho; licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/) |