Chapter 7

Phoebe Framework and Experimental Results for Estimating Fetal Age and Weight

Loc Nguyen, Truong-Duyet Phan and Thu-Hang T. Ho

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74883

Abstract

Fetal age and weight estimation plays an important role in pregnant treatments. There are many estimation formulas created by the combination of statistics and obstetrics. However, such formulas give optimal estimation if and only if they are applied into specified community. This research proposes a so-called Phoebe framework that supports physicians and scientists to find out most accurate formulas with regard to the community where scientists do their research. The built-in algorithm of Phoebe framework uses statistical regression technique for fetal age and weight estimation based on fetal ultrasound measures such as bi-parietal diameter, head circumference, abdominal circumference, fetal length, arm volume, and thigh volume. This algorithm is based on heuristic assumptions, which aim to produce good estimation formulas as fast as possible. From experimental results, the framework produces optimal formulas with high adequacy and accuracy. Moreover, the framework gives facilities to physicians and scientists for exploiting useful statistical information under pregnant data. Phoebe framework is a computer software available at http://phoebe.locnguyen.net.

Keywords: fetal age estimation, fetal weight estimation, ultrasound measures, regression model, estimation formula

1. Introduction

Fetal age and weight estimation is to predict the birth weight or birth age before delivery. It is very important for doctors to diagnose abnormal or diseased cases so that she/he can decide treatments on such cases. Because this research mentions both age estimation and weight

estimation, for convenience, the term "birth estimation" implicates both of them. There are two methods for birth estimation:

- Determining volume of fetal inside mother womb and then calculating fetal weight based on such volume and mass density of flesh and bone. By the other way, fetal age and weight can be estimated according to size of mother womb.
- Applying statistical regression model: Fetal ultrasound measures such as bi-parietal diameter (*bpd*), head circumference (*hc*), abdominal circumference (*ac*), fetal length (*fl*), arm volume (*arm_vol*), and thigh volume (*thigh_vol*) are recorded and considered as input sample for regression analysis which results in a *regression function*. This function is formula for estimating fetal age and weight according to ultrasound measures such as *bpd*, *hc*, *ac*, *fl*, *arm_vol*, and *thigh_vol*. Data that are composed of these ultrasound measures are called gestational sample or pregnant sample. Terms: "*sample*" and "*data*" have the same meaning in this research. Sample is representation of population where research takes place.

Because the second method reflects features of population from statistical data, the regression model is chosen for birth estimation in this research. Note, some terminologies such as *function*, *regression function*, *estimation function*, *regression model*, *estimation model*, *formula*, *regression formula*, and *estimation formula* have the same meaning.

There are many estimation formulas resulted from gestational researches such as [1–9]. Some of them gain high accuracy, but they are only appropriate to population, community or ethnic group, where such researches are done. If we apply these formulas into other community such as Vietnam, they are no longer accurate. Moreover, it is difficult to find out a new and effective estimation formula or the cost of time and (computer) resources of formula discovery is expensive. Therefore, the first goal of this research is to propose an effective built-in algorithm, which produces highly accurate formulas that are easy to tune with specified population. The process of producing formulas by such algorithm is as fast as possible. In addition, physicians and researchers always want to discover useful statistical information from measure sample and regression model. Thus, the second goal of this research is to give facilities to physicians and researchers by introducing them a framework that is called *Phoebe framework* or *Phoebe system*. Phoebe framework implements such built-in algorithm in the first goal and provides a tool allowing physicians and researchers to exploit and take advantage of useful information under gestational sample. This tool is programmed as computer software. Moreover, Phoebe framework allows software developers to modify its modules. For example, developers can improve the built-in algorithm by adding heuristic constraints.

This chapter is the improved collection of our two articles "A framework of fetal age and weight estimation" [10] and "Experimental Results of Phoebe Framework: Optimal Formulas for Estimating Fetus Weight and Age" [11]. Section 2 gives an overview of the architecture of Phoebe framework. Section 3 is a description of the built-in algorithm to produce optimal formulas which are appropriated to a concrete population like Vietnam. Such algorithm is the core of Phoebe framework. Section 4 discusses main use cases of the framework with respect to gestational sample. As experimental results, some interesting estimation formulas produced by the framework are described in Section 5. A proposal of early weight estimation is proposed in Section 6. Conclusion is given in Section 7. Note that Phoebe framework used statistic software

http://dx.doi.org/10.5772/intechopen.74883

package "Java Scientific Library" of Michael Thomas Flanagan [12] and parsing package "A Java expression parser" of Jos de Jong [13]. The package "Java Scientific Library" is the most important one in the framework. The framework is implemented by Java language [14].

2. General architecture of Phoebe framework

Based on clinical data input which includes fetal ultrasound measures such as *bpd*, *hc*, *ac*, and *fl*, the framework produces optimal formulas for estimating fetal weight and fetal age with the highest precision. Moreover, statistical information about fetus and gestation is also described in detail with two forms: numerical format and graph format. Therefore, the framework consists of four components as follows:

- *Dataset* component is responsible for managing information about fetal ultrasound measures such as *bpd*, *hc*, *ac*, *fl* and extra gestational information in reasonable and intelligent manner. This component allows other components to retrieve such information. Gestational information is organized into some abstract structure, for example, a matrix, where each row represents a sample of *bpd*, *hc*, *ac*, *fl* measures. **Table 1** is an example of this abstract structure.
- *Regression* component represents estimation formula or regression function. This component reads ultrasound information from *Dataset* component and builds up optimal estimation formula from such information. The built-in algorithm, which is used to discover and construct estimation formula, is discussed in Section 3. This component is the most important one because it implements such discovery algorithm.
- *Statistical Manifest* component describes statistical information of both ultrasound measures and regression function, for example, mean and standard deviation of *bpd* samples, sum of residuals, correlation coefficient of regression function, and percentile graph of

bpd	hc	fl	ac	Fetal age (week)	Fetal weight (gram)
74	262	51	255	28	900
72	260	51	232	28	900
68	260	50	229	28	900
72	275	52	240	28	900
72	274	52	240	28	950
74	253	50	235	28	950
71	257	52	239	28	950
71	255	53	236	28	950
70	264	52	246	28	950

Table 1. An example of gestational sample matrix.

fetal weight. Statistical manifest is organized into two forms such as numerical format and graph format.

- *User Interface* (*UI*) component is responsible for providing interaction between system and users such as physicians and researchers. A popular use case is that users enter ultrasound measures and require system to print out both optimal estimation formula and statistical information about such ultrasound measures; moreover, users can retrieve other information in *Dataset* component. *UI* component links to all of other components so as to give users as many facilities as possible.

Three components: *Dataset*, *Regression* and *Statistical Manifest* are basic components. The fourth component *User Interface* is the bridge among them. **Figure 1** shows a general architecture of Phoebe framework.

3. Built-in algorithm of Phoebe framework

Phoebe framework uses a regression model for estimating fetal weight and age. Suppose a linear regression function $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ where Y is fetal weight or age, whereas X_i (s) are gestational ultrasound measures such as *bpd*, *hc*, *ac*, and *fl*. Variable Y is called response variable or dependent variable. Each X_i is called *regression variable*, *regressor*, *regression variable*, or *independent variable*. Each α_i is called *regression coefficient*. Given a set of measure values of X_i (s), the value of Y called *Y-estimated* calculated from this regression function is estimated fetal weight (or age) which is compared with real value of Y measured from ultrasonic machine. The real value of Y called *Y-real* is fetal weight (or age) available in sample. In this research, the notation Y refers implicitly to *Y-estimated* if there is no explanation. The deviation between *Y-estimated* and *Y-real* is a criterion used to assess the quality or the precision of regression function. This deviation is also called *estimation error*. The less the deviation is, the better the regression function is. The goal of this research is to find out the optimal regression function or estimation formula whose precision is highest.

A regression function will be good if it meets two conditions as follows:

- The correlation between *Y-estimated* and *Y-real* is large.

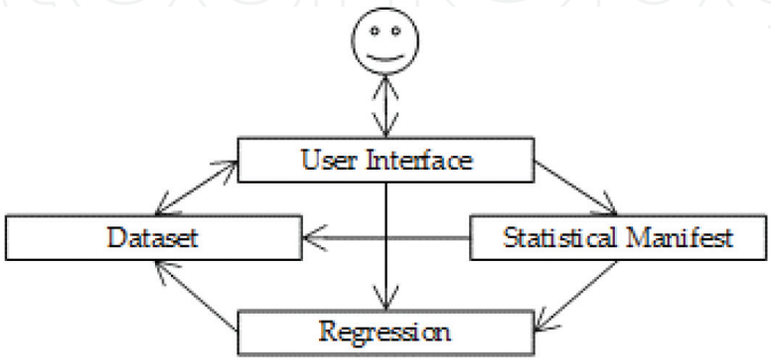

Figure 1. General architecture of Phoebe framework.

- The sum of residuals is small. Note that residual is defined as the square of deviation between *Y-estimated* and *Y-real*. We have:

$$residual = (Y_{estimated} - Y_{real})^{2.}$$

These two conditions are called the *pair of optimal conditions*. A regression function is optimal or best if it satisfies the pair of optimal conditions at most, where correlation between *Y-estimated* and *Y-real* is largest, and the sum of residuals is smallest. Given a set of regression variables X_i (where $i = 1, 2,\ldots, n$), we recognize that a regression function is a combination of k variables X_i (s) where $k \leq n$ so that such combination achieves the pair of optimal conditions. Given a set of possible regression variables $VAR = \{X_1, X_2,\ldots, X_n\}$ being ultrasound measures, brute-force algorithm can be used to find out optimal function, which includes three following steps:

1. Let indicator number k be initialized 1, which responds to k-combination having k regression variables.
2. All combinations of n variables taken k are created. For each k-combination, the function built up by k variables in this k-combination is evaluated on the pair of optimal conditions; if such function satisfies these conditions at most then, it is optimal function.
3. Indicator k is increased by 1. If $k = n$ then algorithm stops, otherwise go back step 2.

The number of combinations which brute-force algorithm browses is:

$$\sum_{k=1}^{n} \frac{n!}{k!(n-k)!}$$

where n is the number of regression variables and notation, and "k!" denotes factorial of k. If n is large enough, there are a huge number of combinations, which causes that the brute-force algorithm never terminates and it is impossible to find out the best function. Moreover, there are many kinds of regression function such as linear, quadric, cube, logarithm, exponent, and product. Therefore, we propose an algorithm which overcomes this drawback and always finds out the optimal function. In other words, the termination of the proposed algorithm is determined, and the time cost is decreased significantly because the searching space is reduced as small as possible. The proposed algorithm is called *seed germination* (***SG***) algorithm. SG is built-in algorithm of Phoebe framework, which is the core of Phoebe framework. It is heuristic algorithm, which is based on the *pair of heuristic assumptions* as follows:

- First assumption: regression variables X_i (s) trends to be mutually independent. It means that any pair of X_i and X_j with $i \neq j$ in an optimal function are mutually independent. The independence is reduced into the looser condition "*the correlation coefficient of any pair of X_i and X_j is less than a threshold δ.*" This is *minimum* assumption.
- Second assumption: each variable X_i contributes to quality of optimal function. The contribution rate of a variable X_i is defined as the correlation coefficient between such variable and *Y-real*. The higher the contribution rate is, the more important the respective variable is. Variables with high contribution rate are called *contributive* variables. Therefore, optimal

function includes only contributive regression variables. The second assumption is stated that "*the correlation coefficient of any regression variable X_i and real response value Y-real is greater than a threshold ε*." This is the *maximum* assumption.

SG algorithm tries to find out a combination of regression variables X_i (s) so that such combination satisfies such pair of heuristic assumptions. In other words, it is expected that this combination can constitute an optimal regression function that satisfies the *pair of heuristic conditions*, as follows ([10] p. 22):

- The correlation coefficient of any pair of X_i and X_j is less than the minimum threshold $\delta > 0$. This condition is corresponding to the minimum assumption, which is called *minimum condition* or *independence condition*.
- The correlation coefficient of any X_i and *Y-real* is greater than the maximum threshold $\varepsilon > 0$. This condition is corresponding to the maximum assumption, which is called *maximum condition* or *contribution condition*.

Given a set of possible regression variables $VAR = \{X_1, X_2, \ldots, X_n\}$ being ultrasound measures, let $f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k$ ($k \leq n$) be the estimation function and let $Re(f) = \{X_1, X_2, \ldots, X_k\}$ be its regression variables. Note that the value of f is fetal age or fetal weight. $Re(f)$ is considered as the representation of f. Let *OPTIMAL* be the output of SG algorithm, which is a set of optimal functions returned. *OPTIMAL* is initialized as empty set. Let $Re(OPTIMAL)$ be a set of regression variables contained in all optimal functions $f \in OPTIMAL$. SG algorithm has four following steps ([10] p. 22):

1. Let *C* be the complement set of *VAR* with regard to *OPTIMAL*, we have $C = \text{VAR} \backslash Re(OPTIMAL)$ where the backslash "\" denotes complement operator in set theory. It means that *C* is in *VAR* but not in $Re(OPTIMAL)$.

2. Let $G \subset C$ be a list of regression variables satisfying the pair of heuristic conditions. Note, *G* is subset of *C*. If *G* is empty, the algorithm terminates; otherwise going to step 3.

3. We iterate over *G* in order to find out the candidate list of good functions. For each regression variable $X \in G$, let *L* be the union set of optimal regression variables and *X*. We have $L = Re(f) \cup \{X\}$ where $f \in OPTIMAL$. Suppose *CANDIDATE* is a candidate list of good functions, which is initialized as empty set. Let *g* be the new function created from *L*; in other words, regression variables of *g* belong to *L*, $Re(g) = L$. If function *g* meets the pair of heuristic conditions, it is added into *CANDIDATE*, $CANDIDATE = CANDIDATE \cup \{g\}$.

4. Let *BEST* be a set of best functions taken from *CANDIDATE*. In other words, these functions belong to *CANDIDATE* and satisfy the pair of heuristic conditions at most, where correlation is the largest and the sum of residuals is the smallest. If *BEST* equals *OPTIMAL*, then the algorithm stops; otherwise assigning *BEST* to *OPTIMAL* and going back step 1. Note that two sets are equal if their elements are the same.

Figure 2 shows the flow chart of SG algorithm.

http://dx.doi.org/10.5772/intechopen.74883

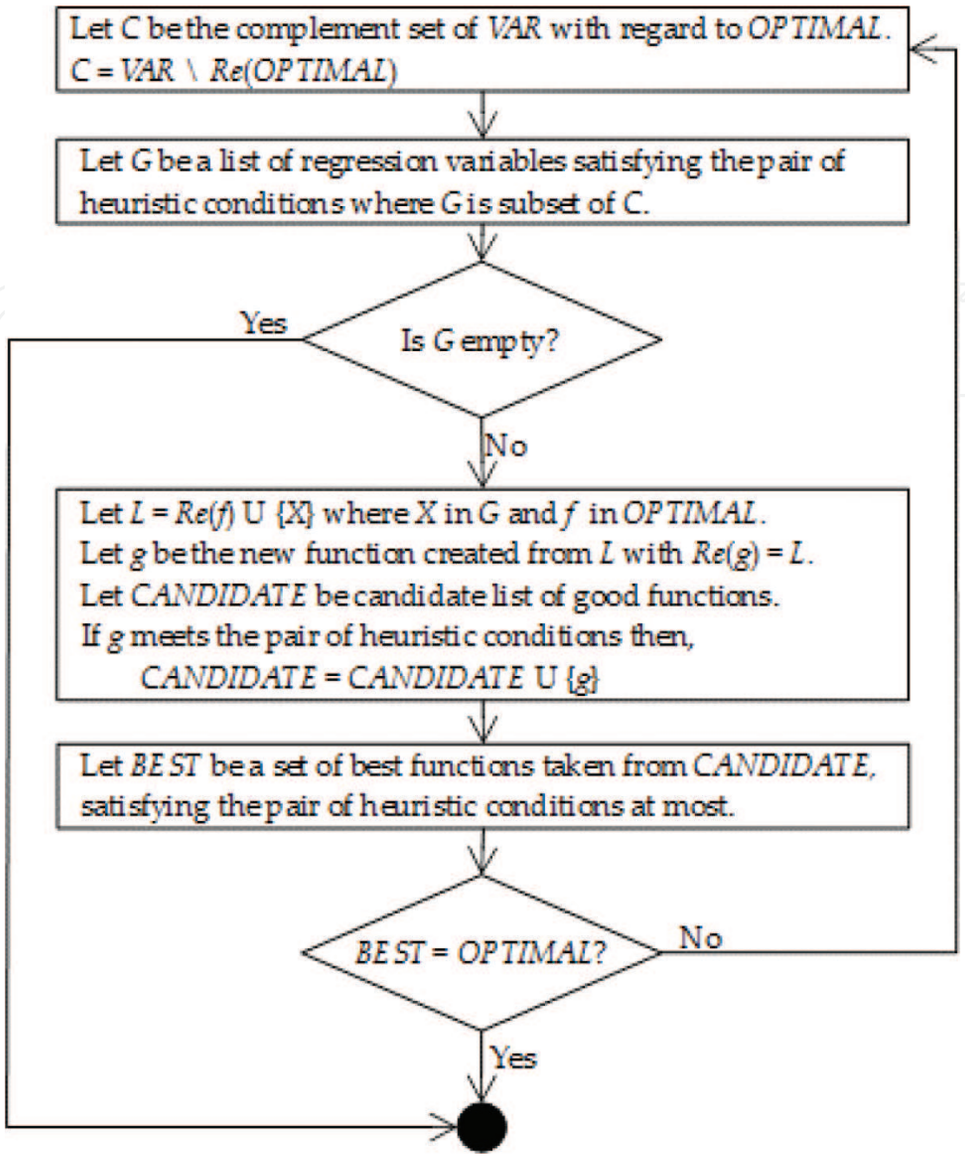

Figure 2. Flow chart of SG algorithm.

SG algorithm was described in article "A framework of fetal age and weight estimation" ([10] pp. 21–23). It is easy to recognize that the essence of SG algorithm is to reduce search space by choosing regression variables satisfying heuristic assumption as "seeds." Optimal functions are composed of these seeds. The algorithm always delivers best functions but can lose other good functions. The length of function is defined as the number of its regression variables. Terminated condition is that no more optimal functions can be found out or possible variables are browsed exhaustedly. Therefore, the result function is the longest and best one, but some other shorter functions may be significantly good.

The current implementation of SG algorithm establishes that the minimum threshold δ is arbitrary. It also supports nonlinear regression models shown in **Table 2** as follows:

Polynomial	$Y = \alpha_0 + \alpha_1(X_1 + X_2 + \ldots + X_n)^k$
Logarithm	$Y = \alpha_0 + \alpha_1 \log(X_1) + \alpha_2 \log(X_2) + \ldots + \alpha_n \log(X_n)$
	$Y = \alpha_0 + \alpha_1 \log(X_1 + X_2 + \ldots + X_n)$
Exponent	$Y = \exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n)$
	$Y = \exp(\alpha_0 + \alpha_1(X_1 + X_2 + \ldots + X_n))$
Product	$Y = \alpha_0 {X_1}^{\alpha_1} {X_2}^{\alpha_2} \ldots {X_n}^{\alpha_n}$

Table 2. Nonlinear regression models.

The notations "exp" and "log" denote exponent function and natural logarithm function, respectively. Most of nonlinear regression models can be transformed into linear regression models. For example, given the product model, the following is an example of linear transformation.

$$\log(Y) = \log(\alpha_0) + \alpha_1 \log(X_1) + \alpha_2 \log(X_2) + \ldots + \alpha_n \log(X_n)$$

Let,

$$U = \log(Y), Z_i = \log(X_i), \beta_0 = \log(\alpha_0), \beta_{i \geq 1} = \alpha_i$$

The product model becomes the linear model with regard to variables U, Z_i and coefficients β_i as follows:

Polynomial transformation	$Y = \alpha_0 + \alpha_1(X_1 + X_2 + \ldots + X_n)^k$
	$Y = \alpha_0 + \alpha_1 Z_1$ where $Z_1 = (X_1 + X_2 + \ldots + X_n)^k$
Logarithm transformation	$Y = \alpha_0 + \alpha_1 \log(X_1) + \alpha_2 \log(X_2) + \ldots + \alpha_n \log(X_n)$
	$Y = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \ldots + \alpha_n Z_n$ where $Z_i = \log(X_i)$
Logarithm transformation	$Y = \alpha_0 + \alpha_1 \log(X_1 + X_2 + \ldots + X_n)$
	$Y = \alpha_0 + \alpha_1 Z_1$ where $Z_1 = \log(X_1 + X_2 + \ldots + X_n)$
Exponent transformation	$Y = \exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n)$
	$U = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ where $U = \log(Y)$
Exponent transformation	$Y = \exp(\alpha_0 + \alpha_1(X_1 + X_2 + \ldots + X_n))$
	$U = \alpha_0 + \alpha_1 Z_1$ where $U = \log(Y)$ and $Z_1 = X_1 + X_2 + \ldots + X_n$
Product transformation	$Y = \alpha_0 {X_1}^{\alpha_1} {X_2}^{\alpha_2} \ldots {X_n}^{\alpha_n}$
	$U = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_n Z_n$ where $U = \log(Y), Z_i = \log(X_i), \beta_0 = \log(\alpha_0), \beta_{i \geq 1} = \alpha_i$

Table 3. Transformation of nonlinear models into linear models.

http://dx.doi.org/10.5772/intechopen.74883

$$U = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_n Z_n$$

Table 3 shows how to transform nonlinear models into linear models.

With the built-in SG algorithm, Phoebe framework can be totally used for any regression application beyond birth estimation.

4. Use cases of Phoebe framework

Phoebe framework has three basic use cases realized by three components: dataset, regression model and statistical manifest as discussed in Section 2. Three basic use cases include:

1. Discovering optimal formulas with high accuracy. Optimal formulas are results of SG algorithm described in Section 3.
2. Providing statistical information under gestational sample. Statistical information is in numeric format and graph format.
3. Comparison among different formulas.

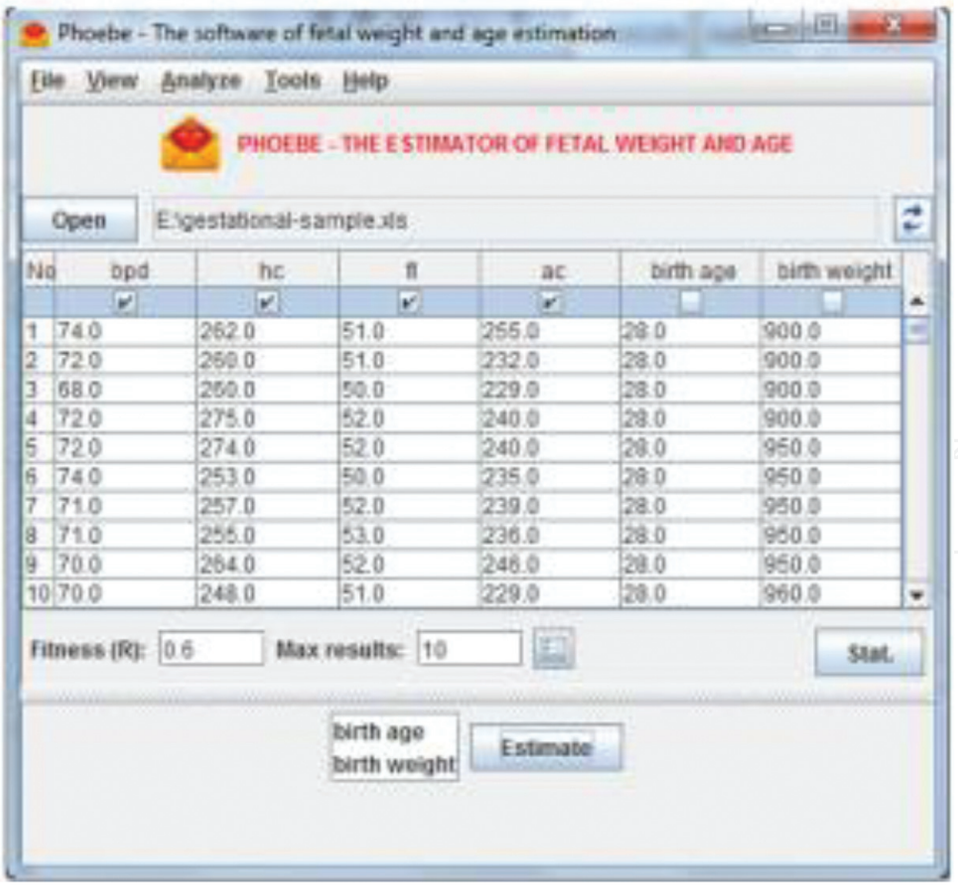

Figure 3. Gestational sample.

Use case 1: Discovering optimal formulas

Given gestational data [15] are composed of two-dimensional ultrasound measures of pregnant women. These measures are taken at Vinh Long General Hospital – Vietnam, which include bi-parietal diameter (*bpd*), head circumference (*hc*), abdominal circumference (*ac*) and fetal length (*fl*). Fetal age is from 28 to 42 weeks. Fetal weight is measured by gram. Gestational sample is shown in **Figure 3**.

After specifying the maximum threshold ε (fitness value) and which measures are regression variables and response variable, user presses button "Estimate" to retrieve optimal formulas as results of SG algorithm. Such optimal formulas are shown in **Figure 4**. Note, in **Figure 4**, regression variables are *bpd*, *hc*, *ac*, and *fl*, whereas response variable is fetal weight. The threshold ε is 0.6.

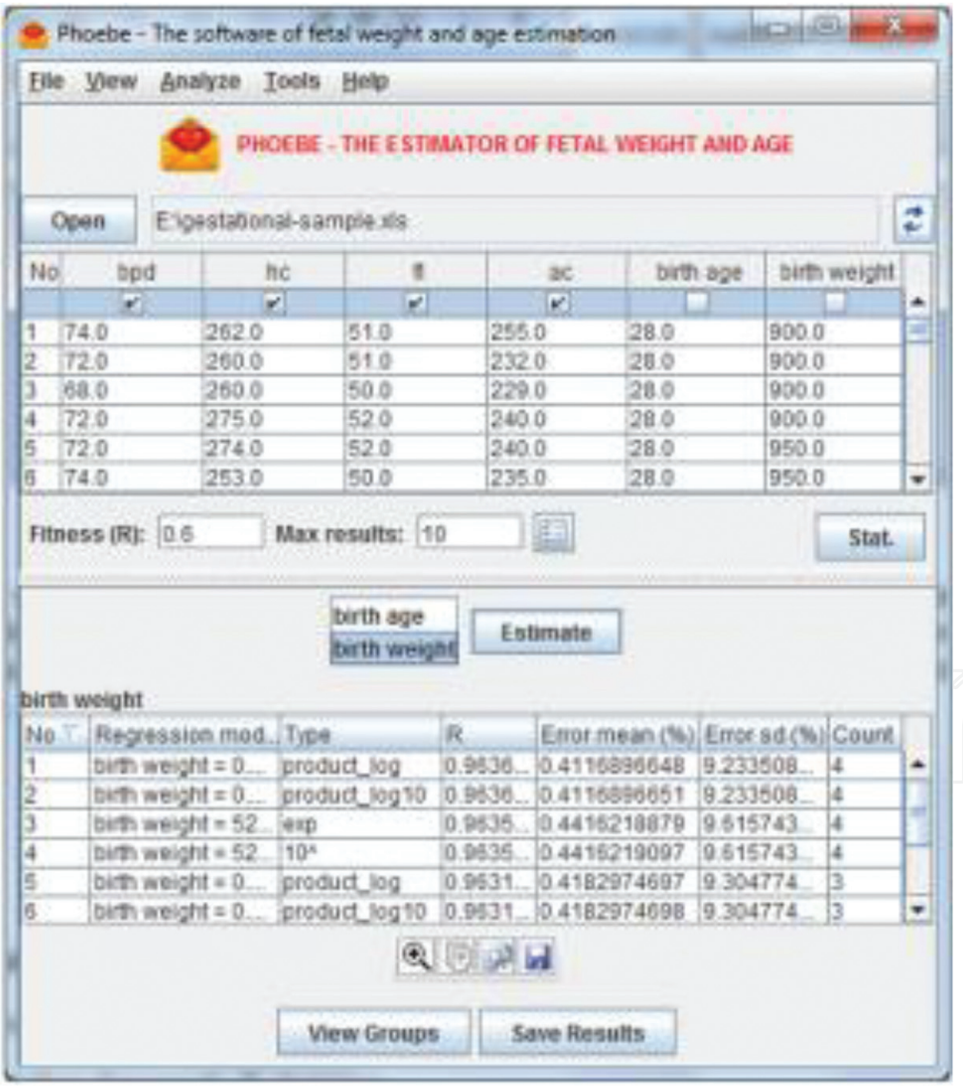

Figure 4. Optimal weight estimation formulas.

An estimation formula with one or two regressors (ultrasound measures) can be represented as a graph. In the illustrative **Figure 5**, the horizontal axis indicates the measure *bpd* in millimeter, and the right vertical axis indicates the measure *ac* in millimeter. The left vertical axis shows the estimated weight.

The graph in **Figure 5** has 11 estimation lines represented as internal (red) lines. Each estimation line corresponds to a small interval of *ac*. Fetal weight on each estimation line ranges from 900 to 4800 g. This is a way to show a three-dimensional function as a two-dimensional graph. For example, given *bpd* = 90 and *ac* = 300, we need to estimate fetal weight. Because *ac* is 300 mm, we look at the sixth estimation line from bottom to up. The intersection point between *bpd* = 90 and the sixth estimation line is projected on the left vertical axis, which results out a fetal weight that approximates to (4800–900)/2 + 900 ≈ 2850 g because such intersection point is near to midpoint of the weight range on the sixth estimation line.

Use case 2: Providing statistical information

Statistical information is classified into two groups: gestational information and estimation information.

- Gestational information contains statistical attributes about fetal ultrasound measures, for example, mean, median and standard deviation of *bpd*.
- Estimation information contains attributes about estimation model, for example, correlation coefficient, sum of residuals and estimation error of estimation formula.

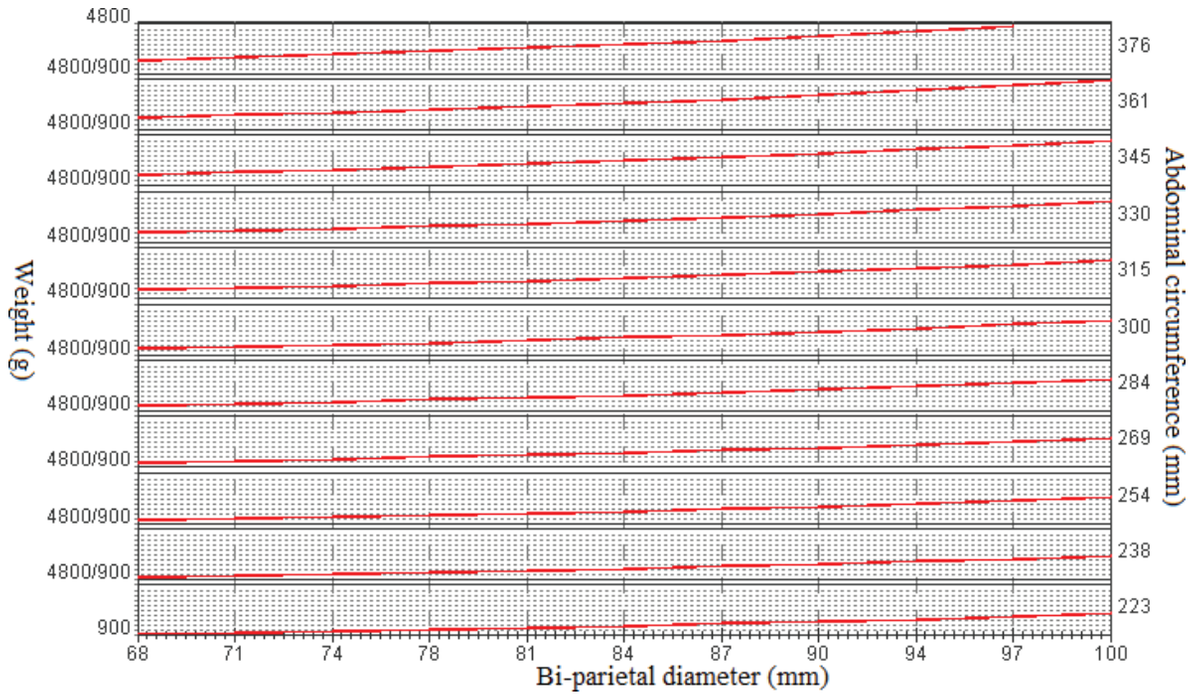

Figure 5. Estimation graph for estimating fetal weight.

In representation, statistical information is described in two forms: numeric format and graph format. **Figure 6** shows statistical attributes (mean, median, standard deviation, histogram, etc.) of fetal age and ultrasound measures *bpd*, *hc*, *ac*, *fl*.

Figure 7 shows a full description of a weight estimation formula: *weight* = 0.000043 * (*bpd*^1.948640) * (*hc*^0.263745) * (*fl*^0.601972) * (*ac*^0.905524). For instance, sum of residuals (SS) is 46412446.0047 and estimation error is -7.4655 ± 212.5571. Note, the sign "^" denotes exponent function, for example, 2^3 = 8.

Use case 3: Comparison among different formulas

There are many criteria to evaluate efficiency and accuracy of estimation formulas. These criteria are called evaluation criteria, for example, correlation coefficient, sum of residuals, estimation error. Each formula has individual strong points and drawbacks. A formula is better than another one in terms of some criteria but may be worse than this other one in terms of different criteria. An optimal formula is the one that has more strong points than drawbacks

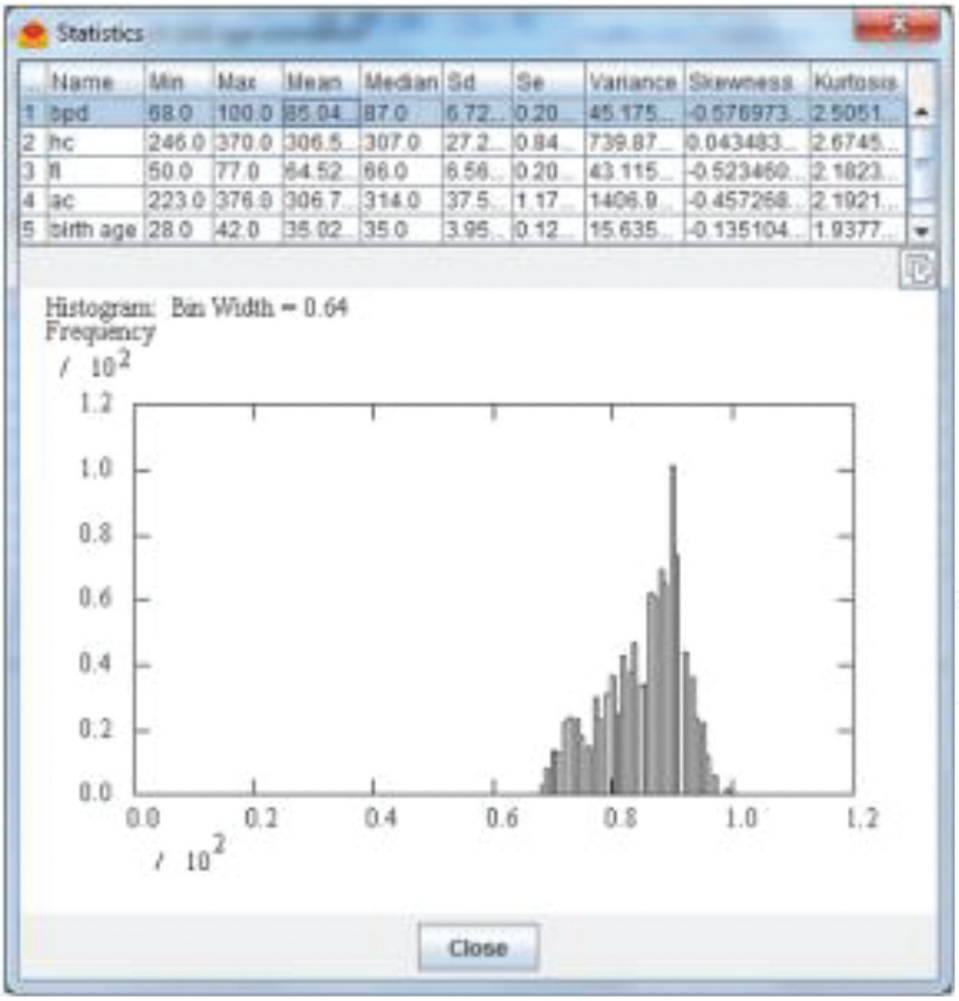

Figure 6. Gestational statistical information.

http://dx.doi.org/10.5772/intechopen.74883

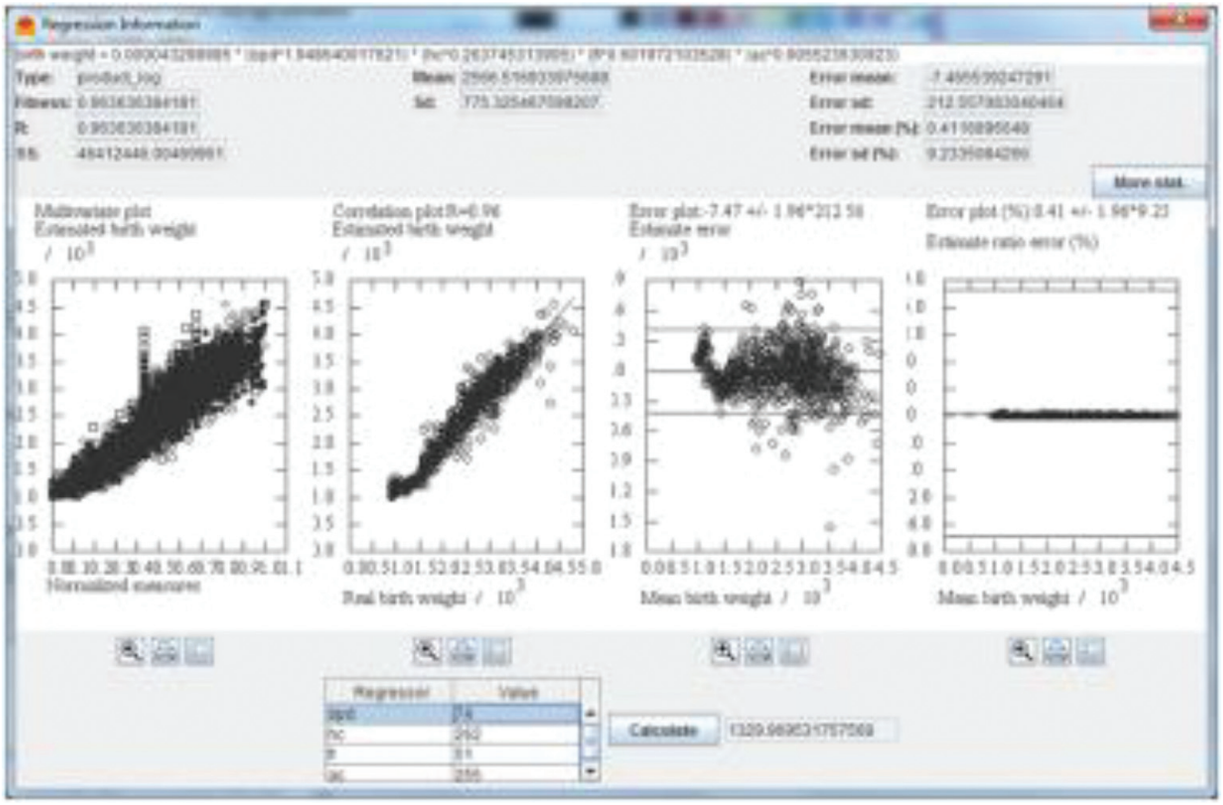

Figure 7. Statistical estimation information.

in most criteria. Hence, Phoebe framework supports the comparison among different formulas via *evaluation matrix* represented in **Figure 8**. Each row in evaluation matrix represents a formula, whereas each column indicates a criterion. For example, first row, second row and third row represent three formulas in form of logarithm function, exponent function and linear function, respectively. Four criteria such as multivariate correlation, estimation correlation, error range and ratio error range are arranged in three respective columns.

Tables 4–8 in the section "experimental results" are numeric interpretations of evaluation matrix in **Figure 8**.

5. Experimental results

We make experiments based on Phoebe framework in order to find out optimal formulas for estimating fetus weight and age with note that such formulas are most appropriate to our gestational samples. We use two samples in which the first sample includes two-dimensional (2D) ultrasound measures of 1027 cases and the second sample includes three-dimensional (3D) ultrasound measures of 506 cases. Ho and Phan [15, 16] collected these samples 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,

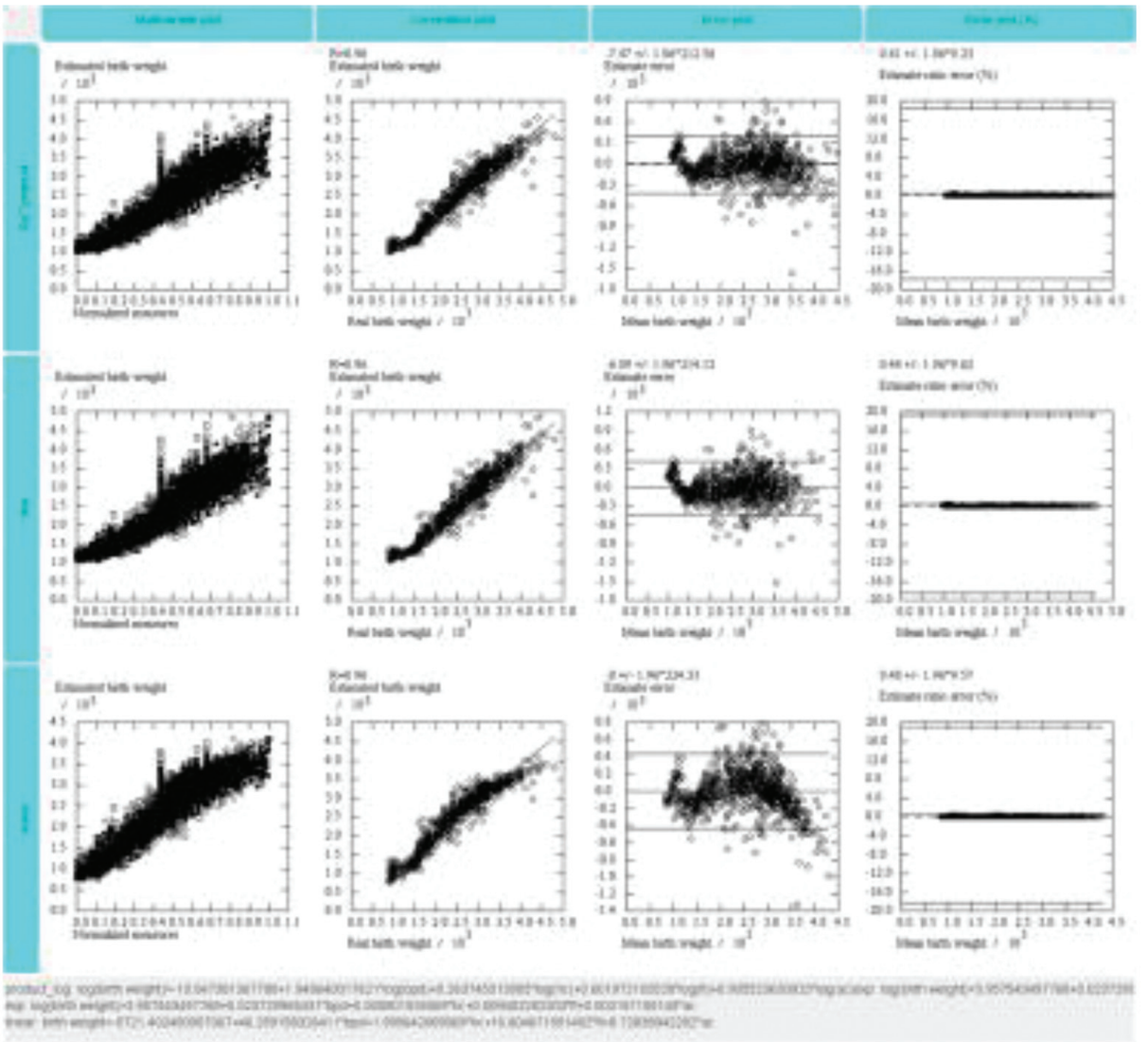

Figure 8. Comparison among different formulas.

and their last periods are determined. Each of them has only one alive fetus. Fetal age is from 28 to 42 weeks. Delivery time is not over 48 h since ultrasound scan. Measures in 2D sample are *bpd*, *hc*, *ac*, and *fl*. Measures in 3D sample are *bpd*, *hc*, *ac*, *fl*, *thigh_vol*, *arm_vol*. The unit of *bpd*, *hc*, *ac*, *fl* is millimeter. The unit of *thigh_vol* and *arm_vol* is cm^3. The units of fetal age and fetal weight are week and gram, respectively. Experimental results mentioned in this section were also published in our article "Experimental Results of Phoebe Framework: Optimal Formulas for Estimating Fetus Weight and Age" [11].

The proposed framework can produce amazing formulas. We compare our optimal formulas with the others according to metrics such as estimation correlation and estimation error range, given such two gestational samples. Let $Y = \{y_1, y_2, \ldots, y_n\}$ and $Z = \{z_1, z_2, \ldots, z_n\}$ be fetal sample age/weight and fetal estimated age/weight, respectively. The estimation correlation denoted R is correlation coefficient of sample response value and estimated response value, according to Eq. (1). The correlation R reflects adequacy of a given formula. The larger the R is, the better the formula is:

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(z_i - \bar{z})^2}} \tag{1}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

An estimation error denoted d_i is deviation between z_i and y_i. The estimation error mean denoted μ is mean of errors. The error mean μ reflects accuracy of a given formula. The smaller the absolute value of μ is, the more accurate the formula is. If μ is positive, the respective formula leans to overestimation. If μ is negative, the respective formula leans to low estimation. The standard deviation σ of estimation errors reflects the stability of a given formula. The smaller the standard deviation σ is, the more stable the formula is. The combination of error mean μ and standard deviation σ results out a so-called *error range*. Eq. (2) explains how to calculate μ, σ, and error range.

$$\begin{aligned} d_i &= z_i - y_i \\ \mu &= \frac{1}{n}\sum_{i=1}^{n} d_i \\ \sigma &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \mu)^2} \\ error\_range &= [\mu - \sigma, \mu + \sigma] = \mu \pm \sigma \end{aligned} \tag{2}$$

For example, if $\mu = -0.0292$ and $\sigma = 1.45$ then, the error range is -0.0292 ± 1.45, which means that the total average error ranges from $-1.4792 = -0.0292\text{-}1.45$ to $1.4208 = -0.0292 + 1.45$. The error range reflects both adequacy and accuracy of a given formula.

Formula	Expression	R	Error range
NH 1	log(*age*) = 2.419638 + 0.002012 * *bpd* + 0.000934 * *hc* + 0.00547 * *fl* + 0.001042 * *ac*	0.9303	−0.0292 ± 1.4500
NH 2	*age* = −3.364759 + 0.056285 * *bpd* + 0.034697 * *hc* + 0.188156 * *fl* + 0.035304 * *ac*	0.9285	0 ± 1.4682
Ho 1	*age* = 331.022308–1.611774 * (*hc* + *ac*) + 0.00278 * ((*hc* + *ac*)^2) - 0.000002 * ((*hc* + *ac*)^3)	0.9212	0 ± 1.5384
Varol 6	*age* = 11.769 + 1.275 * *fl*/10 + 0.449 * ((*fl*/10)^2) - 0.02 * ((*fl*/10)^3)	0.8949	−1.6807 ± 1.8525
Varol 1	*age* = 5.596 + 0.941 * *ac*/10	0.8941	−0.5683 ± 1.7711
Varol 5	*age* = 1.863 + 6.280 * *fl*/10–0.211 * ((*fl*/10)^2)	0.8934	−1.5182 ± 2.1150

The sign "^" denotes exponent operator. The template of formulas aims to flexibility, which can be input of any computational tool. **Table 5** shows a comparison between our best weight formula and the others with 2D sample. As seen in **Table 5**, our formula is the best with $R = 0.9636$ and error range − 7.4656 ± 212.5573 g.

Table 4. Comparison of age estimation with 2D sample.

Table 4 shows a comparison between our best age formula and the others with 2D sample. As a convention, the name of each formula is the name of respective author listed in references section. For example, formula "Ho 1" is the first formula of the author Ho [4]. As seen in **Table 4**, our formula is the best with $R = 0.9303$ and error range -0.0292 ± 1.4500 week (s). As a convention, our formulas have names with prefix "NH"

Table 6 shows comparison between our best age formula and the others with 3D sample. As seen in **Table 6**, our formula is the best with $R = 0.9970$ and error range ± 0.2696 week

Table 7 shows a comparison between our best weight formula and the others with 3D sample. As seen in **Table 7**, our formula is the best with $R = 0.9708$ and error range -0.0001 ± 180.9803 g

Within the context of this research, from section of 3D ultrasound in PhD dissertation of Ho [4], I recognize that fetus weight and fetus age are mutually dependent. For instance, when fetus age increases, fetus weight increases too. As a result, weight estimation is improved significantly if fetus age was known before. If fetus age is added into the regression model of fetus weight as a regression variable (regressor), the resulted weight estimation formula, called *dual formula*, is even better than the most optimal ones shown in **Tables 5** and **6**. Such dual formula is not only precise but also practical because many pregnant women knew their gestational age before taking an ultrasound examination. Given 2D sample and 3D sample, **Table 8** shows dual formulas in comparison with the most optimal ones shown in **Tables 5** and **7** with regard to *R* and error range. As a convention, our dual formulas have names with prefix "NHD". Notation "log10" denotes logarithm function with base 10.

Formula	Expression	R	Error range
NH 3	log(*weight*) = −10.047381 + 1.94864 * log(*bpd*) + 0.263745 * log (*hc*) + 0.601972 * log(*fl*) + 0.905524 * log(*ac*)	0.9636	−7.4656 ± 212.5573
NH 4	log(*weight*) = 3.957543 + 0.02373 * *bpd* + 0.000802 * *hc* + 0.009403 * *fl* + 0.003157 * *ac*	0.9635	−6.0901 ± 214.1153
Sherpard	*weight* = 10^(1.2508 + 0.166 * *bpd*/10 + 0.046 * *ac*/10–0.002646 * *ac* * *bpd*/100)	0.9619	−65.8121 ± 219.0392
Ho 2	*weight* = 10^(1.746 + 0.0124 * *bpd* + 0.001906 * *ac*)	0.9602	−11.5576 ± 223.5124
Hadlock	*weight* = 10^(1.304 + 0.05281 * *ac*/10 + 0.1938 * *fl*/10–0.004 * *ac* * *fl*/100)	0.9395	−76.4960 ± 272.9474
Campbell and Wilkin	*weight* = 1000 * exp.(−4.564 + 0.282 * *ac*/10–0.00331 * *ac* * *ac*/100)	0.9215	68.1261 ± 308.5728

Table 5. Comparison of weight estimation with 2D sample.

Formula	Expression	R	Error range
NH 5	*age* = 20.759763 + 0.170859 * (*thigh_vol* + *arm_vol*) - 0.000545 * ((*thigh_vol* + *arm_vol*) ^2) + 0.000001 * ((*thigh_vol* + *arm_vol*)^3)	0.9970	0 ± 0.2696
NH 6	*age* = 21.816252 + 0.137531 * (*thigh_vol* + *arm_vol*) - 0.000228 * ((*thigh_vol* + *arm_vol*) ^2)	0.9969	0 ± 0.2752
Ho 3	*age* = 21.1148 + 0.2381 * *thigh_vol* - 0.001 * (*thigh_vol*^2) + 0.000002 * (*thigh_vol*^3)	0.9960	−0.0150 ± 0.3173
Ho 4	*age* = 167.079079–1.553705 * *ac* + 0.005559 * (*ac*^2) - 0.000006 * (*ac*^3)	0.8482	0.3723 ± 1.8985

Table 6. Comparison of age estimation with 3D sample.

In **Table 8**, all dual formulas NHD * are better than normal formulas NH * with regard to *R* and error range. Moreover, NHD * do not need too much regressors. Given 2D sample, NHD 1 and NHD 2 use 4 and 3 regressors including age regressor, respectively whereas both NH 3 and NH 4 uses 4 regressors. Given 3D sample, NHD 3 and NHD 4 use 6 and 5 regressors including age regressor, respectively, whereas NH 7 and NH 8 use 5 and 3 regressors, respectively.

Formula	Expression	R	Error range
NH 7	*weight* = −3617.936175 + 0.513171 * *hc* + 1.960176 * *ac* + 39.804645 * *bpd* + 17.016936 * *fl* + 8.366404 * *thigh_vol* + 5.828808 * *arm_vol*	0.9708	−0.0001 ± 180.9803
NH 8	*weight* = −3626.314419 + 43.426744 * *bpd* + 23.645338 * *fl* + 11.414273 * *thigh_vol*	0.9698	0 ± 184.0439
Ho 5	*weight* = −3306 + 55.477 * *bpd* + 13.483 * *thigh_vol*	0.9663	−0.0072 ± 194.0956
Lee 3	*weight* = exp.(0.5046 + 1.9665 * log(*bpd*/10) - 0.3040 * (log(*bpd*/10)^2) + 0.9675 * log(*ac*/10) + 0.3557 * log(*arm_vol*))	0.9620	247.8761 ± 206.1607
Lee 5	*weight* = exp.(2.1264 + 1.1461 * log(*ac*/10) + 0.4314 * log(*thigh_vol*))	0.9514	289.2660 ± 234.0763
Lee 2	*weight* = exp.(−3.6138 + 4.6761 * log(*ac*/10) - 0.4959 * (log(*ac*/10)^2) + 0.3795 * log(*arm_vol*))	0.9472	316.4974 ± 242.7964
Ho 6	*weight* = −882.7049 + 73.9955 * *thigh_vol* - 0.497 * (*thigh_vol*^2) + 0.0014 * (*thigh_vol*^3)	0.9385	−7.5001 ± 260.4596
Lee 4	*weight* = exp.(4.7806 + 0.7596 * log(*thigh_vol*))	0.9298	737.4932 ± 344.1904
Lee 1	*weight* = exp.(4.9588 + 1.0721 * log(*arm_vol*) - 0.0526 * (log(*arm_vol*)^2))	0.9281	867.0836 ± 309.5779
Chang	*weight* = 1080.8735 + 22.44701 * *thigh_vol*	0.9229	456.5168 ± 298.2517

Table 7. Comparison of weight estimation with 3D sample.

Formula	Expression	R	Error range
NHD 1 (2D sample)	log10(*weight*) = −3.715073 + 1.873457 * log10(*bpd*) + 0.363783 * log10(*fl*) + 0.691683 * log10(*ac*) + 0.722245 * log10(*age*)	0.9674	−5.6422 ± 202.0395
NHD 2 (2D sample)	log10(*weight*) = −3.761798 + 2.001731 * log10(*bpd*) + 0.811078 * log10(*ac*) + 0.826279 * log10(*age*)	0.9667	−5.6111 ± 204.1477
NHD 3 (3D sample)	*weight* = −4988.000528 + 66.374156 * *age* + 0.370084 * *hc* + 1.943247 * *ac* + 39.464816 * *bpd* + 13.215505 * *fl* + 3.658463 * *thigh_vol*	0.9715	0 ± 178.8091
NHD 4 (3D sample)	*weight* = −4982.099978 + 68.089354 * *age* + 2.001675 * *ac* + 39.85375 * *bpd* + 13.229377 * *fl* + 3.619405 * *thigh_vol*	0.9714	0 ± 178.9114
NH 3 (2D sample)	log(*weight*) = −10.047381 + 1.94864 * log(*bpd*) + 0.263745 * log(*hc*) + 0.601972 * log(*fl*) + 0.905524 * log(*ac*)	0.9636	−7.4656 ± 212.5573
NH 4 (2D sample)	log(*weight*) = 3.957543 + 0.02373 * *bpd* + 0.000802 * *hc* + 0.009403 * *fl* + 0.003157 * *ac*	0.9635	−6.0901 ± 214.1153
NH 7 (3D sample)	*weight* = −3617.936175 + 0.513171 * *hc* + 1.960176 * *ac* + 39.804645 * *bpd* + 17.016936 * *fl* + 8.366404 * *thigh_vol* + 5.828808 * *arm_vol*	0.9708	−0.0001 ± 180.9803
NH 8 (3D sample)	*weight* = −3626.314419 + 43.426744 * *bpd* + 23.645338 * *fl* + 11.414273 * *thigh_vol*	0.9698	0 ± 184.0439

Table 8. Weight estimation dual formulas.

Although our formulas are better than all remaining ones with high adequacy (large R) and high accuracy (small error range), other researches are always significant because their formulas are very simple and practical. Moreover, our formulas are not global. If they are applied into other samples collected in other communities, their accuracy may be decreased and they may not be still better than traditional formulas such as Sherpard and Hadlock. However, it is easy to draw from our experimental results that if Phoebe framework is used for the same samples with other researches, it will always produce preeminent formulas. In order to achieve global optimality with Phoebe framework, the following are two essential suggestions:

- Experimenting on Phoebe framework with many samples.
- Adding more knowledge of pregnancy study, ultrasound technique, and obstetrics into Phoebe framework. In other words, the additional knowledge will be modeled as constraints of SG algorithm.

These suggestions go beyond this research. In my opinion, we cannot reach absolutely the global optimality because Phoebe framework focuses on local optimality with specific communities. Essentially, the suggestions only alleviate the weak point of the built-in SG algorithm in global optimality.

6. A proposal of early weight estimation

The used ultrasound samples are collected in fetal age from 28 to 42 weeks because delivery time is not over 48 h since last ultrasound scan. Hence, accuracy of weight estimation is only ensured when ultrasound examinations are performed after 28-week old fetal age. This section proposes an early weight estimation, in which ultrasound measures can be taken before 28-week old fetal age. We do not ensure improvement of estimation accuracy yet because we do not make experiments on the proposal yet, but the gestational sample can be totally collected at any appropriate time points in gestational period. In other words, the sample can lack fetal weights. This is a convenience for practitioners because they do not need to concern fetal weights when taking ultrasound examinations. Consequently, early weight estimation is achieved. As a convention, vectors are column vectors if there is no additional information.

Without loss of generality, regression models are linear such as $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ and $Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$ where Y is fetal age and Z is fetal weight, whereas X_i (s) are gestational ultrasound measures such as *bpd*, *hc*, *ac*, and *fl*. Suppose both Y and Z conform normal distribution, according to Eq. (3) ([17] pp. 8–9).

$$\begin{aligned} P\langle Y|X,\alpha\rangle &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(\frac{-\left(Y-\alpha^T X\right)^2}{2\sigma_1^2}\right) \\ P\langle Z|X,\beta\rangle &= \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(\frac{-\left(Z-\beta^T X\right)^2}{2\sigma_2^2}\right) \end{aligned} \tag{3}$$

where $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_n)^T$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_n)^T$ are parameter vectors where $X = (1, X_1, X_2, \ldots, X_n)^T$ is data vector. The means of Y and Z are $\alpha^T X$ and $\beta^T X$, respectively, whereas the variances

of Y and Z are σ_1^2 and σ_2^2, respectively. Note that the superscript "^{T}" denotes transposition operator in vector and matrix. Let $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{z})$ be collected sample in which $\boldsymbol{X}$ is a set of sample measures, $\boldsymbol{y}$ is a set of sample fetal ages, and $\boldsymbol{z}$ is a set of fetal weights with note that $\boldsymbol{z}$ is missed (empty) or incomplete. If $\boldsymbol{z}$ is empty, there is no z_i in $\boldsymbol{z}$. If $\boldsymbol{z}$ is incomplete, $\boldsymbol{z}$ has some values but there are also some missing values in $\boldsymbol{z}$. However, the constraint is that $\boldsymbol{y}$ must be complete, which means that all pregnant women within the research knew their gestational age. Now we focus on estimate α and β based on $\boldsymbol{D}$. As a convention, let α^* and β^* be estimates of α and β, respectively ([17] p. 8).

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

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}$$

Given X, joint probability of Y and Z is product of the probability of Y given X and the probability of Z given X because Y and Z are conditionally independent given X, according to Eq. (4).

$$P\langle Y, Z|X, \alpha, \beta\rangle = P\langle Y|X, \alpha\rangle P\langle Z|X, \beta\rangle = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(\frac{-\left(Y - \alpha^T X\right)^2}{2\sigma_1^2} - \frac{\left(Z - \beta^T X\right)^2}{2\sigma_2^2}\right) \quad (4)$$

Conditional expectation of sufficient statistic Z given X with regard to $P(Z \mid X, \beta)$ is specified by Eq. (5).

$$E\langle Z|X\rangle = \beta^T X \quad (5)$$

When Z is hidden variable, there is a latent dependent relationship between Y and Z, which is specified by joint probability of Y and Z.

$$P(Y, Z) = P(Y)P\langle Z|Y\rangle$$

Variables Y and Z have different measures. For instance, the unit of Y is week, whereas the unit of Z is gram. Suppose Y is considered as discrete variable whose values from 1 to K where K can be up to 42, for example. The $P(Y)$ becomes parameter θ_Y, which is the probability of Y where Y is from 1 to K.

$$P(Y, Z) = \theta_Y P\langle Z|Y\rangle$$

For each Z, suppose the condition probability $P(Z \mid Y)$ is distributed normally with mean μ_Y and variance ${\sigma_Y}^2$. Eq. (6) specifies the joint probability $P(Y, Z)$.

$$P\langle Y, Z|\theta_Y, \mu_Y, \sigma_Y^2\rangle = \frac{\theta_Y}{\sqrt{2\pi\sigma_Y^2}}\exp\left(\frac{-(Z-\mu_Y)^2}{2\sigma_Y^2}\right) \tag{6}$$

Conditional expectation of sufficient statistic Z given Y with regard to $P(Z \mid Y, \mu_Y, {\sigma_Y}^2)$ is specified by Eq. (7).

$$E\langle Z|Y\rangle = \mu_Y \tag{7}$$

Please pay attention to Eq. (7) because Z will be estimated by such expectation later. Eq. (8) specifies expectation of sufficient statistic Z with regard to $P(Y, Z \mid \theta_Y, \mu_Y, {\sigma_Y}^2)$.

$$E(Z) = \sum_{Y=1}^{K}\theta_Y\mu_Y \tag{8}$$

Due to:

$$E\langle Z|\theta_Y, \mu_Y, \sigma_Y^2\rangle = \sum_{Y=1}^{K}\int_Z ZP\langle Y, Z|\theta_Y, \mu_Y, \sigma_Y^2\rangle dZ = \sum_{Y=1}^{K}\theta_Y E\langle Z|\theta_Y, \mu_Y, \sigma_Y^2\rangle = \sum_{Y=1}^{K}\theta_Y\mu_Y$$

The full joint probability of Y and Z given X and parameters α, β, θ_Y, μ_Y, and ${\sigma_Y}^2$ is the product specified by Eq. (9).

$$\begin{aligned} &P\langle Y, Z|X, \alpha, \beta, \theta_Y, \mu_Y, \sigma_Y^2\rangle = P\langle Y, Z|\theta_Y, \mu_Y, \sigma_Y^2\rangle P\langle Y, Z|X, \alpha, \beta\rangle \\ &= P\langle Y, Z|\theta_Y, \mu_Y, \sigma_Y^2\rangle P\langle Y|X, \alpha\rangle P\langle Z|X, \beta\rangle \end{aligned} \tag{9}$$

where $P(Y, Z \mid X, \alpha, \beta)$ and $P(Y, Z \mid \theta_Y, \mu_Y, {\sigma_Y}^2)$ are specified by Eqs. (4) and (6), respectively. Eq. (9) indicates that both explicit dependence via $P(Y, Z \mid X, \alpha, \beta)$ and implicit dependence via $P(Y, Z \mid \theta_Y, \mu_Y, {\sigma_Y}^2)$ between Y and Z. Explicit dependence and implicit dependence share equal influence on Z if $E(Z \mid X)$ specified by Eq. (5) is equal to $E(Z)$ specified by Eq. (8), according to Eq. (10).

$$\sum_{Y=1}^{K}\theta_Y\mu_Y = \beta^T X \tag{10}$$

Given sample $\boldsymbol{D}$, all θ_Y become constants and determined by Eq. (11).

$$\theta_Y = \frac{\text{The number of } y_i = Y}{N} \tag{11}$$

For convenience, let $\Theta = (\alpha, \beta, \mu_Y)^T$ be the compound parameter. The full joint probability specified by Eq. (9) is rewritten as follows:

$$P\langle y, z|X, \Theta\rangle = P\langle y, z|\mu_Y, \sigma_Y^2\rangle P\langle y|X, \alpha\rangle P\langle z|X, \beta\rangle$$

$$= \prod_{i=1}^{N} P\langle y_i, z_i|\mu_Y, \sigma_Y^2\rangle P\langle y_i|\boldsymbol{x_i}, \alpha\rangle P\langle z_i|\boldsymbol{x_i}, \beta\rangle$$

(Due to all observations are independently and identically distributed)

$$= \left(\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}}\right)^N * \exp\left(\frac{-1}{2}\left(\sum_{i=1}^{N}\frac{\left(y_i - \alpha^T\boldsymbol{x_i}\right)^2}{\sigma_1^2} + \sum_{i=1}^{N}\frac{\left(z_i - \beta^T\boldsymbol{x_i}\right)^2}{\sigma_2^2}\right)\right) * \prod_{i=1}^{N}\prod_{Y=1}^{K}\frac{\delta(y_i, Y)\theta_Y}{\sqrt{2\pi\sigma_Y^2}}\exp\left(\frac{-\left(z_i - \mu_Y\right)^2}{2\sigma_Y^2}\right)$$

where

$$\delta(y_i, Y) = \left\{\begin{array}{l} 1 \text{ if } y_i = Y \\ 0 \text{ if } y_i = Y \end{array}\right\}$$

It is conventional that if $\delta(y_i, Y) = 0$ then, the respective probability $P(y_i, z_i \mid \mu_Y, {\sigma_Y}^2)$ is removed from the product. The log-likelihood function is logarithm of the full joint probability as follows:

$$L(\Theta) = \log(P\langle y, z|X, \Theta\rangle) = -N\log(2\pi) - \frac{N\log\left(\sigma_1^2\right)}{2} - \frac{N\log\left(\sigma_2^2\right)}{2}$$

$$-\frac{1}{2\sigma_1^2}\sum_{i=1}^{N}\left(y_i - \alpha^T\boldsymbol{x_i}\right)^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{N}\left(z_i - \beta^T\boldsymbol{x_i}\right)^2$$

$$+\sum_{i=1}^{N}\sum_{Y=1}^{K}\delta(y_i, Y)\left(\log(\theta_Y) - \frac{\log(2\pi)}{2} - \frac{\log\left(\sigma_Y^2\right)}{2} - \frac{\left(z_i - \mu_Y\right)^2}{2\sigma_Y^2}\right)$$

When $\log(2\pi)$ and θ_Y are constants, the reduced log-likelihood function is derived from the log-likelihood as seen in Eq. (12).

$$l(\Theta) = -\frac{N}{2}\log\left(\sigma_1^2\right) - \frac{N}{2}\log\left(\sigma_2^2\right) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{N}\left(y_i - \alpha^T\boldsymbol{x_i}\right)^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{N}\left(z_i - \beta^T\boldsymbol{x_i}\right)^2$$

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{Y=1}^{K}\delta(y_i, Y)\left(\log\left(\sigma_Y^2\right) + \frac{\left(z_i - \mu_Y\right)^2}{\sigma_Y^2}\right) \tag{12}$$

The optimal estimate Θ^* is a maximizer of $l(\Theta)$, according to Eq. (13) ([17] p. 9).

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}}\, L(\Theta) = \underset{\Theta}{\operatorname{argmax}}\, l(\Theta) \tag{13}$$

By taking first-order partial derivatives of $l(\Theta)$ with regard to Θ ([18] p. 34), we obtain:

$$\frac{\partial l(\Theta)}{\partial \alpha} = \frac{1}{\sigma_1^2}\sum_{i=1}^{N}\left(y_i - \alpha^T \boldsymbol{x_i}\right)(\boldsymbol{x_i})^T$$

$$\frac{\partial l(\Theta)}{\partial \beta} = \frac{1}{\sigma_2^2}\sum_{i=1}^{N}\left(z_i - \beta^T \boldsymbol{x_i}\right)(\boldsymbol{x_i})^T$$

$$\frac{\partial l(\Theta)}{\partial \mu_Y} = \sigma_Y^2 \sum_{i=1}^{N}\delta\left(y_i, Y\right)\left(z_i - \mu_Y\right)$$

When first-order partial derivatives of $l(\Theta)$ are equal to zero, it gets locally maximal. In other words, Θ^* is solution of the equation system 14 resulted from setting such derivatives to be zero and setting $E(Z \mid X) = E(Z)$.

$$\left\{\begin{array}{l} \sum_{i=1}^{N}\left(y_i - \alpha^T \boldsymbol{x_i}\right)(\boldsymbol{x_i})^T = \boldsymbol{0}^T \\ \sum_{i=1}^{N}\left(z_i - \beta^T \boldsymbol{x_i}\right)(\boldsymbol{x_i})^T = \boldsymbol{0}^T \\ \sum_{i=1}^{N}\delta\left(y_i, Y\right)\left(z_i - \mu_Y\right) = 0 \\ \sum_{j=1}^{K}\theta_j \mu_j = \beta^T \boldsymbol{x_i} \text{ for some } i \end{array}\right\} \tag{14}$$

where

$$\delta\left(y_i, Y\right) = \left\{\begin{array}{l} 1 \text{ if } y_i = Y \\ 0 \text{ if } y_i \neq Y \end{array}\right\}$$

The notation $\boldsymbol{0} = (0, 0, \ldots, 0)^T$ denotes zero vector. All equations in system 14 are linear, whose unknowns are $\Theta = (\alpha, \beta, \mu_Y)^T$. The last equation in system 14 is Eq. (10) with the heuristic assumption that explicit dependence and implicit dependence share equal influence on Z. Such last equation is only used to adjust μ_Y (s) if the heuristic assumption is concerned; otherwise it is ignored.

We apply expectation maximization (EM) algorithm into estimating $\Theta = (\alpha, \beta, \mu_Y)^T$ with lack of fetal weights. Note that the full joint probability $P(Y, Z \mid X, \alpha, \beta, \mu_Y)$ specified by Eq. (9) is product of regular exponential distributions. EM algorithm has many iterations, and each iteration has expectation step (E-step) and maximization step (M-step) for estimating parameters. Given current parameter $\Theta^t = (\alpha^t, \beta^t, {\mu_Y}^t)^T$ at the t^{th} iteration, the two steps are shown in **Table 9** ([19] p. 4).

The equation system 14 is solvable because missing values z_i (s) were estimated in E-step. The EM algorithm stops if at some t^{th} iteration, we have $\Theta^t = \Theta^{t+1} = \Theta^*$. At that time, $\Theta^* = (\alpha^*, \beta^*, {\mu_Y}^*)^T$ is the optimal estimate of EM algorithm, and hence, linear regression functions of Y and Z are determined with α^*, β^*.

1. E-step: Estimating only missing values z_i (s) as the expectation of themselves based on the current mean $\mu_{y_i}^t$, according to Eq. (7). Note, each missing value z_i is always associated with an observation y_i.
$$z_i = E\langle z_i | y_i \rangle = \mu_{y_i}^t$$
2. M-step: The next parameter Θ^{t+1} is a maximizer of $l(\Theta)$, which is the solution of equation system 14. Note, Θ^{t+1} becomes current parameter for the next iteration.

Table 9. E-step and M-step of EM algorithm.

As usual, all parameters are changed after every iteration of EM algorithm, but fortunately, α^* is determined as a partial solution of equation system 14 at the first iteration of EM process because both $\boldsymbol{X}$ and $\boldsymbol{y}$ are complete. In other words, α^* is fixed, whereas β and μ_Y are changed in EM process. Eq. (15) ([20] p. 417) specifies α^*.

$$\alpha^* = \alpha^1 = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{15}$$

where the superscript "$^{-1}$" denotes the inversion of matrix.

At the first iteration, as usual Θ^1 is initialized arbitrarily but we can improve convergence of EM algorithm by initializing μ_Y^1 as sample mean. Without loss of generality, suppose practitioners obtained $n < N$ fetal weights $z_1, z_2, \ldots, z_n$ from n ultrasound scans. Moreover, the fetal age of all pregnant women over such n scans is the same, which is Y. Thus, μ_Y^1 is initialized by Eq. (16).

$$\mu_Y^1 = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{16}$$

The parameter β^1 at the first iteration is initialized according to previous studies in the literature.

7. Conclusions

According to experimental results, there is no doubt that Phoebe framework produces optimal formulas with high adequacy and accuracy; please see **Tables 4–8** for more details. However, we also recognize the weak point of our research is that the built-in SG algorithm can lose some good formulas due to the heuristic conditions. The suggestive solution is to add more constraints in such conditions; please read the article "A framework of fetal age and weight estimation" ([10] pp. 24–25) for more details. The proposal of early weight estimation uses actually an additional constraint which is the latent relationship between fetal age and fetal weight. Such latent relationship represented by the joint probability of fetal age and weight is a knowledge aspect of pregnancy study. For further research, we will make experiment on the proposal and try our best to discover other knowledge aspects.

Another weak point of our research is difficult to apply our complex formulas for fast mental calculation because we must pay the price for their high accuracy. In the future, we will embed

these formulas into software or hardware of medical ultrasound machine so that users are easy to read estimated values resulted from machine.

Acknowledgements

We express our deep gratitude to the author Michael Thomas Flanagan – University College London and the author Jos de Jong for giving us helpful software packages that help us to implement the framework.

Author details

Loc Nguyen[1]*, Truong-Duyet Phan[2] and Thu-Hang T. Ho[3]

*Address all correspondence to: ng_phloc@yahoo.com

1 Sunflower Soft Company, Ho Chi Minh, Vietnam

2 Hanoi Medical University, Hanoi, Vietnam

3 Vinh Long General Hospital, Vinh Long, Vietnam

References

[1] Hadlock FP, Harrist RB, Sharman RS, Deter RL, Park SK. Estimation of fetal weight with use of head, body and femur measurements: A prospective study. American Journal of Obstetrics and Gynecology. 1 February 1985;**151**(3):pp. 333-337

[2] Phan DT. Ứng dụng siêu âm để chẩn đoán tuổi thai và cân nặng thai trong tử cung. Hanoi: Hanoi University of Medicine; 1985

[3] Phạm TNT. Ước lượng cân nặng thai nhi qua các số đo của thai trên siêu âm. Ho Chi Minh: Ho Chi Minh University of Medicine and Pharmacy; 2000

[4] Ho THT. Nghiên Cứu Phương Pháp Ước Lượng Trọng Lượng Thai, Tuổi Thai Bằng Siêu Âm Hai và Ba Chiều. Hanoi: Hanoi Univerisy of Medicine; 2011

[5] Shepard JM, Richards AV, Berkowitz LR, Warsof LS, Hobbins CJ. An evaluation of two equations for predicting fetal weight by ultrasound. American Journal of Obstetrics and Gynecology. 1 January 1982;**142**(1):47-54

[6] Campbell S, Wilkin D. Ultrasonic measurement of fetal abdomen circumference in the estimation of fetal weight. BJOG: An International Journal of Obstetrics & Gynecology. September 1975;**82**(9):689-697

[7] Lee W, Balasubramaniam M, Deter RL, Yeo L, Hassan SS, Gotsch F, Kusanovic JP, Gonçalves LF, Romero R. New fetal weight estimation models using fractional limb volume. Ultrasound in Obstetrics & Gynecology. 1 November 2009;**34**(5):556-565

[8] Chang F-M, Liang R-I, Ko H-C, Yao B-L, Chang C-H, Yu C-H. Three-dimensional ultrasound-assessed fetal thigh volumetry in predicting birth weight. Obstetrics & Gynecology. September 1997;**90**(3):331-339

[9] Varol F, Saltik A, Kaplan PB, Kilic T, Yardim T. Evaluation of gestational age based on ultrasound fetal growth measurements. Yonsei Medical Journal. June 2001;**42**(3):299-303

[10] Flanagan MT. In: Flanagan MT, editor. Java Scientific Library. London, England: University College London; 2004

[11] Jong Jd, A Java Expression Parser, Rotterdam: SpeQ Mathematics; 2010

[12] Oracle, "Java language," Oracle Corporation, [Online]. Available: https://www.oracle.com/java. [Accessed 25 December 2014]

[13] Nguyen L, Ho H. A framework of fetal age and weight estimation. Journal of Gynecology and Obstetrics (JGO). 30 March 2014;**2**(2):pp. 20-25

[14] Ho THT, Phan DT. Ước lượng cân nặng của thai từ 37–42 tuần bằng siêu âm 2 chiều. Journal of Practical Medicine. December 2011;**12**(797):8-9

[15] Ho T-HT, Phan DT. Ước lượng tuổi thai qua các số đo thể tích cánh tay bằng siêu âm 3 chiều và các số đo bằng siêu âm 2 chiều. Journal of Practical Medicine. December 2011; **12**(798):12-15

[16] Nguyen L, Ho T-HT. Experimental results of phoebe framework: optimal formulas for estimating fetus weight and age. Journal of Community & Public Health Nursing13 March 2017;**3**(2):1-5

[17] Lindsten F, Schön TB, Svensson A, Wahlström N. Probabilistic Modeling – Linear Regression & Gaussian processes. Uppsala: Uppsala University; 2017

[18] Nguyen L. In: Evans C, editor. Matrix Analysis and Calculus. 1st ed. Hanoi: Lambert Academic Publishing; 2015. p. 72

[19] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological). 1977; **39**(1):1-38

[20] Montgomery DC, Runger GC. Applied Statistics and Probability for Engineers. 3rd ed. New York, NY: John Wiley & Sons, Inc.; 2003. p. 706

IntechOpen