SCIENTIFIC SOCIETY
MULTIDISCIPLINAR MAGAZINE

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

MIXTURE REGRESSION MODEL FOR INCOMPLETE DATA

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

[1]Advisory Board, Loc Nguyen's Academic Network, An Giang, Vietnam
ng_phloc@yahoo.com
[2]Department of Mathematics and Statistics, Preston University Islamabad, Islamabad, Pakistan
anumshafiq@ymail.com

ABSTRACT

The Regression Expectation Maximization (REM) algorithm, which is a variant of Expectation Maximization (EM) algorithm, uses parallelly a long regression model and many short regression models to solve the problem of incomplete data. Experimental results proved resistance of REM to incomplete data, in which accuracy of REM decreases insignificantly when data sample is made sparse with loss ratios up to 80%. However, as traditional regression analysis methods, the accuracy of REM can be decreased if data varies complicatedly with many trends. In this research, we propose a so-called Mixture Regression Expectation Maximization (MREM) algorithm. MREM is the full combination of REM and mixture model in which we use two EM processes in the same loop. MREM uses the first EM process for exponential family of probability distributions to estimate missing values as REM does. Consequently, MREM uses the second EM process to estimate parameters as mixture model method does. The purpose of MREM is to take advantages of both REM and mixture model. Unfortunately, experimental result shows that MREM is less accurate than REM. However, MREM is essential because a different approach for mixture model can be referred by fusing linear equations of MREM into a unique curve equation.

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

1. INTRODUCTION

1.1. Main work

As a convention, regression model is a linear regression function $Z = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + \ldots + \alpha_nX_n$ in which variable Z is called response variable or dependent variable whereas each X_i is called regression variable, regressor, predictor, regression variable, or independent variable. Each α_i is called regression coefficient. The essence of regression analysis is to calculate regression coefficients from data sample. When sample is complete, these coefficients are determined by least squares method [1, pp. 452-458]. When sample is incomplete, there are some approximation approaches to estimate regression coefficients such as complete case method, ad-hoc method, multiple imputation, maximum likelihood, weighting method, and Bayesian method [2]. We focus on applying expectation maximization (EM) algorithm into constructing regression model in case of missing data with note that EM algorithm belongs to maximum likelihood approach. In previous research [3], we proposed a so-called Regression Expectation Maximization (REM) algorithm to learn linear regression function from incomplete data in which some values of Z and X_i are missing. REM is a variant of EM algorithm, which is used to estimate regression coefficients. Experimental results in previous research [3] proved that accuracy of REM decreases insignificantly whereas loss ratios increase significantly. We hope that REM will be accepted as a new standard method for regression analysis in case of missing data when there are currently 6 standard approaches such as complete case method, ad-hoc method, multiple imputation, maximum likelihood, weighting method, and Bayesian method [2]. Here we combine REM and mixture model with expectation that the accuracy is improved, especially in case that data is incomplete and has many trends. Our proposed algorithm is called Mixture Regression Expectation Maximization (MREM) algorithm. The purpose of MREM is to take advantages of both REM and mixture model. Unfortunately, experimental result shows that MREM is less accurate than REM. However, MREM is essential because a different approach for mixture model can be referred by fusing linear equations of MREM into a unique curve equation [4], as discussed later. Because this research is the successive one after our previous research [3], they share some common contents related to research survey and experimental design, but we confirm that their methods are not coincide although MREM is derived from REM.

Because MREM is the combination of REM and mixture model whereas REM is a variant of EM algorithm, we need to survey some works related to application of EM algorithm to regression analysis. Kokic [5] 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 [6] calculated the expectation of function of mixture random variable in expectation step (E-step) of EM algorithm and then used such expectation for estimating parameters of multivariate mixed Poisson regression model in the maximization step (M-step). Anderson and Hardin [7] used reject inference technique to estimate coefficients of logistic regression model when response variable Z is missing but characteristic variables (regressors X_i) are fully observed. Anderson and Hardin replaced missing Z by its conditional expectation on regressors X_i where such expectation is logistic function. Zhang, Deng, and Su [8] used EM algorithm to build up linear regression model for studying glycosylated hemoglobin from partial missing data. In other words, Zhang, Deng, and Su [8] 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 [9] 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 [10] proposed a class of inverse probability of censoring weighted estimators for estimating coefficients of regression model. Their approach is based on the dependency of mean vector of response variable Z on vector of regressors X_i when Z has missing values. Robins, Rotnitzki, and Zhao [10] assumed that the probability $\lambda_{it}(\alpha)$ of existence of Z at time point t is dependent on existence of Z at previous time point t–1 but independent from Z. Even though Z is missing, the probability $\lambda_{it}(\alpha)$ is also determined and so regression coefficients are calculated based on the inverse of $\lambda_{it}(\alpha)$ and X_i. The inverse of $\lambda_{it}(\alpha)$ is considered as weight for complete case. Robins, Rotnitzki, and Zhao used additional time-dependent covariates V_{it} to determine $\lambda_{it}(\alpha)$.

In the article “Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models”, Horton and Kleinman [2] classified 6 methods of regression analysis in case of missing data such as complete case method, ad-hoc method, multiple imputation, maximum likelihood, weighting method, and Bayesian method. EM algorithm belongs to maximum likelihood method. According to complete case method, regression model is learned from only non-missing values of incomplete data [2, p. 3]. The ad-hoc method refers missing values to some common value, creates an indicator of missingness as new variable, and finally builds regression model from both existent variables and such new variable [2, p. 3]. Multiple imputation method has three steps. Firstly, missing values are replaced by possible values. The replacement is repeated until getting an enough number of complete datasets. Secondly, some regression models are learned from these complete datasets as usual [2, p. 4]. Finally, these regression models are aggregated together. The maximum likelihood method aims to construct regression model by maximizing likelihood function. EM algorithm is a variant of maximum likelihood method, which has two steps such as expectation step (E-step) and maximization step (M-step). In E-step, multiple entries are created in an augmented dataset for each observation of missing values and then probability of the observation is estimated based on current parameter [2, p. 6]. In M-step, regression model is built from augmented dataset. The REM algorithm proposed in this research is different from the traditional EM for regression analysis because we replace missing values in E-step by expectation of sufficient statistics via mutual balance process instead of estimating the probability of observation. The weighting method determines the probability of missingness and then uses such probability as weight for the complete case. The aforementioned research of Robins, Rotnitzki, and Zhao [10] belongs to the weighting approach. Instead of replacing missing values by possible values like imputation method does, the Bayesian method imputes missing values by the estimation with a prior distribution on the covariates and the close relationship between the Bayesian approach and maximum likelihood method [2, p. 7].

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

1.2. Related Studies

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

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

According to the first approach [11], the mixture probability distribution is formulated as follows:

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

Where $\Theta = (\alpha_k, \sigma_k^2)^T$ is compound parameter whereas α_k and σ_k^2 are regression coefficient and variance of the partial (component) probability distribution $P_k(Z|\alpha_k^T X, \sigma_k^2)$. Note, mean of $P_k(Z|\alpha_k^T X, \sigma_k^2)$ is $\alpha_k^T X$ and mixture coefficient is c_k. In the first approach, regression coefficients α_k are estimated by least squares method whereas mixture coefficients c_k are estimated by logistic function as follows [11, p. 4]:

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

The mixture regression model is:

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

According to the second approach, the joint distribution is defined as follows [12, p. 4]:

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

Where α_k are regression coefficients and σ_k^2 is variance of the conditional probability distribution $P_k(Z|\alpha_k^T X, \sigma_k^2)$ whereas μ_k and Σ_k are mean vector and covariance matrix of the prior probability distribution $P_k(X|\mu_k, \Sigma_k)$, respectively. The mixture regression model is [12, p. 6]:

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

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

Where,

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

The joint probability can be defined by different way as follows [13, p. 21], [14, p. 24], [15, p. 4]:

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

Where $m_k(X)$ and σ_k^2 are mean and variance of Z given the conditional probability distribution $P_k(Z|m_k(X), \sigma_k^2)$ whereas μ_{kX} and Σ_{kX} are mean vector and covariance matrix of X given the prior probability distribution $P_k(X|\mu_k, \Sigma_k)$. When μ_{kX} and Σ_{kX} are calculated from data, other parameters $m_k(X)$ and σ_k^2 are estimated for each k^{th} component as follows [13, p. 23], [14, p. 25], [15, p. 5]:

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

For each k^{th} component, μ_{kZ} is sample mean of Z, Σ_{kZX} is vector of covariances of Z and X, and Σ_{kZZ} is sample variance of Z. The mixture regression model becomes [14, p. 25]:

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

Where,

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

Grün & Leisch [16] mentioned the full application of mixture model into regression model in which regression coefficients are determined by inverse function of mean of conditional probability distribution as follows:

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

In general, the two approaches in literature do not implement regression mixture model according to EM process in full. They aim to simplify the estimation process in which mixture coefficients c_k and regression coefficients α_k are estimated one time.

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

Note that EM process is an iterative process in which parameters are improved gradually until convergence. The EM process is slow, but it can balance many factors to reach most optimal parameters. Here we proposed a so-called Mixture Regression Expectation Maximization (MREM) which is the full combination of REM [3] and mixture model in which we use two EM processes in the same loop. Firstly, we use the first EM process for exponential family of probability distributions to estimate missing values as REM does. Secondly, we use the second EM process to estimate parameters as the full mixture model method does. Anyway, MREM supports fully EM mixture model.

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

- MREM does not use the joint probability distribution. In other words, MREM does not concern the probability distribution of independent variables X_i. MREM does not either use logistic function to estimate mixture coefficients as the first approach does.
- MREM is the full combination of REM [3] and mixture model in which we use two EM processes in the same loop for estimating missing values and parameters.
- Variance σ_k^2 and regression coefficient α_k of the probability $P_k(Z|\alpha_k^TX, \sigma_k^2)$ in MREM are estimated and balanced by both full mixture model and maximum likelihood estimation (MLE). The most similar research to MREM is the weighed least squares algorithm used by Faicel Chamroukhi, Allou Samé, Gérard Govaert, and Patrice Aknin [4]. They firstly split the conditional expectation into two parts at the E-step of EM algorithm and then applied weighed least squares algorithm into the second part for estimate parameters at the M-step [4, pp. 1220-1221].
- Mixture regression models in literature are learned from complete data whereas MREM supports incomplete data.

The methodology of MREM is described in section 2. Section 3 includes experimental results and discussions. Section 4 is the conclusion.

2. METHODOLOGY

The probabilistic Mixture Regression Model (MRM) is a combination of normal mixture model and linear regression model. In MRM, the probabilistic Entire Regression Model (ERM) is sum of K weighted probabilistic Partial Regression Models (PRMs). Equation (12) specifies MRM [17, p. 3].

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

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018
ISSN: 2595-8402
DOI: 10.5281/zenodo.2528978

Where,

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

Note, Θ is called entire parameter,

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

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

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

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

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

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

For each kth PRM, suppose each $x_{ij} \in X_i$ has an inverse regression model (IRM) $x_{ij} = \beta_{kj0} + \beta_{kj1}z_i$. In other words, x_{ij} now is considered as the random variable conforming to normal distribution according to equation (15) [18, p. 8].

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

Where $\beta_{kj} = (\beta_{kj0}, \beta_{kj1})^T$ is an inverse regression coefficient (IRC) and $(1, z_i)^T$ becomes an inverse data vector. The mean and variance of each x_{ij} with regard to the inverse distribution $P_{kj}(x_{ij}|z_i, \beta_{kj})$ are $\beta_{kj}^T(1, z_i)^T$ and τ_{kj}^2, respectively. Of course, for each kth PRM, there are n IRMs $P_{kj}(x_{ij}|z_i, \beta_{kj})$ and n associated IRCs β_{kj}. Totally, there are n*K IRMs associated with n*K IRCs. Suppose IRMs with fixed j have the same mixture model as MRM does. Equation (16) specifies the mixture model of IRMs.

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

$$P_j(x_{ij}|z_i, \beta_j) = \sum_{k=1}^{K} c_k P_{kj}(x_{ij}|z_i, \beta_{kj}) \tag{16}$$

In this research, we focus on estimating the entire parameter $\Theta = (c_k, \alpha_k, \sigma_k^2, \beta_{kj})^T$ where k is from 1 to K. In other words, we aim to estimate c_k, α_k, σ_k^2, and β_{kj} for determining the ERM in case of missing data. As a convention, let $\Theta^* = (c_k^*, \alpha_k^*, (\sigma_k^2)^*, \beta_{kj}^*)^T$ be the estimate of $\Theta = (c_k, \alpha_k, \sigma_k^2, \beta_{kj})^T$, respectively. Let $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{Z})$ be collected sample in which $\boldsymbol{X}$ is a set of regressors and $\boldsymbol{Z}$ is a set of outcome variables plus values 1, respectively [18, p. 8] with note that both $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incomplete. In other words, $\boldsymbol{X}$ and $\boldsymbol{Z}$ have missing values. As a convention, let z_i^- and x_{ij}^- denote missing values of $\boldsymbol{Z}$ and $\boldsymbol{X}$, respectively.

$$\boldsymbol{X} = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix}$$
$$X_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix}, X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{pmatrix} \tag{17}$$
$$Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}, \boldsymbol{Z} = (\boldsymbol{1}, Z) = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_N \end{pmatrix}$$

The expectation of sufficient statistic z_i regard to the k^{th} PRM $P_k(z_i|X_i, \alpha_k, \sigma_k^2)$ is specified by equation (18) [3].

$$E_k(z_i|X_i) = \alpha_k^T X_i = \sum_{j=0}^{n} \alpha_{kj} x_{ij} \tag{18}$$

Where $x_{i0}=1$ for all i. The expectation of the sufficient statistic x_{ij} with regard to each IRM $P_{kj}(x_{ij}|z_i, \beta_j)$ of the k^{th} PRM $P_k(z_i|X_i, \alpha_k, \sigma_k^2)$ is specified by equation (19) [3].

$$E_k(x_{ij}|z_i) = \beta_{kj}^T (1, z_i)^T = \beta_{kj0} + \beta_{kj1} z_i \tag{19}$$

Please pay attention to equations (18) and (19) because missing values of data $\boldsymbol{X}$ and data $\boldsymbol{Z}$ will be estimated by these expectations later.

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

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

that ***X*** and ***Z*** become complete. In M-step, the next parameter $\Theta^{(t+1)} = (c_k^{(t+1)}, \alpha_k^{(t+1)}, (\sigma_k^2)^{(t+1)}, \beta_{kj}^{(t+1)})^T$ is determined based on the complete data ***X*** and ***Z*** fulfilled in E-step. Here we proposed a so-called Mixture Regression Expectation Maximization (MREM) which is the full combination of Regression Expectation Maximization (REM) algorithm [3] and mixture model in which we use two EM processes in the same loop. Firstly, we use the first EM process for exponential family of probability distributions to estimate missing values in E-step. The technique is the same to the technique of REM in previous research [3]. Secondly, we use the second EM process to estimate Θ^* for full mixture model in M-step.

Firstly, we focus on fulfilling missing values in E-step. The most important problem in our research is how to estimate missing values z_i^- and x_{ij}^-. Recall that, for each k^{th} PRM, every missing value z_i^- is estimated as the expectation based on the current parameter $\alpha_k^{(t)}$, according to equation (18) [3].

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

Note, $x_{i0} = 1$. Let M_i be a set of indices of missing values x_{ij}^- with fixed i for each k^{th} PRM. In other words, if $j \in M_i$ then, x_{ij} is missing. The set M_i can be empty. The equation (18) is re-written for each k^{th} PRM as follows [3]:

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

According to equation (19), missing value x_{ij}^- is estimated by [3]:

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

Combining equation (18) and equation (19), we have [3]:

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

It implies [3]:

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

As a result, equation (20) is used to estimate or fulfill missing values for each k^{th} PRM [3].

SCIENTIFIC SOCIETY

MULTIDISCIPLINAR MAGAZINE

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

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

$$x_{ij}^- = \begin{cases} \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)} z_i \text{if} z_i \text{is not missing} \\ \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)} z_i^- \text{if} z_i \text{is missing} \end{cases} \quad (20)$$

Now in M-step we use EM algorithm again to estimate the next parameter $\Theta^{(t+1)} = (c_k^{(t+1)}, \alpha_k^{(t+1)}, (\sigma_k^2)^{(t+1)}, \beta_{kj}^{(t+1)})^T$ with current known parameter $\Theta^{(t)} = (c_k^{(t)}, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}, \beta_{kj}^{(t+1)})^T$ given data ***X*** and data ***Z*** fulfilled in E-step. The conditional expectation $Q(\Theta|\Theta^{(t)})$ with unknown Θ is determined as follows [17, p. 4]:

$$Q\left(\Theta\middle|\Theta^{(t)}\right) = \sum_{k=1}^{K}\sum_{i=1}^{N} P\left(Y = k\middle|X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)\log(c_k) + \sum_{k=1}^{K}\sum_{i=1}^{N} P\left(Y = k\middle|X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)\log\left(P_k(z_i|X_i, \alpha_k, \sigma_k^2)\right)$$

The next parameter $\Theta^{(t+1)}$ is a constrained optimizer of $Q(\Theta|\Theta^{(t)})$. This is the optimization problem.

$$\begin{cases} \Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}}\, Q\left(\Theta\middle|\Theta^{(t)}\right) \\ \text{subject to} \sum_{k=1}^{K} c_k = 1 \end{cases}$$

By applying Lagrange method, each next mixture coefficient $c_k^{(t+1)}$ is specified by equation (21) [17, p. 7].

$$c_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} P\left(Y = k\middle|X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right) \quad (21)$$

Where $P(Y=k \mid X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)})$ is specified by equation (22) [17, p. 3]. It is the conditional probability of the kth PRM given X_i and z_i. Please pay attention to this important probability. The proof of equation (22) is found in [17, p. 3], according to Bayes' rule.

$$P\left(Y = k\middle|X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right) = \frac{c_k^{(t)} P_k\left(z_i\middle|X_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)}{\sum_l^K c_l^{(t)} P_l\left(z_i\middle|X_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)} \quad (22)$$

Note, $P_k(z_i|X_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)})$ is determined by equation (13).

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

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

By applying Lagrange method, each next regression coefficient $\alpha_k^{(t+1)}$ is solution of equation (23) [17, p. 7].

$$\sum_{i=1}^{N}P\left(Y=k\middle|X_i,z_i,\alpha_k^{(t)},(\sigma_k^2)^{(t)}\right)\left(z_i-\left(\alpha_k^{(t)}\right)^TX_i\right)X_i^T=\mathbf{0}^T \qquad (23)$$

Where $\mathbf{0} = (0, 0,\ldots, 0)^T$ is zero vector and $P(Y=k \mid X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)})$ is specified by equation (22). Equation (23) is equivalent to equation (24):

$$\begin{aligned}&\sum_{i=1}^{N}P\left(Y=k\middle|X_i,z_i,\alpha_k^{(t)},(\sigma_k^2)^{(t)}\right)\alpha_k^TX_iX_i^T\\&=\sum_{i=1}^{N}P\left(Y=k\middle|X_i,z_i,\alpha_k^{(t)},(\sigma_k^2)^{(t)}\right)z_iX_i^T\end{aligned} \qquad (24)$$

Let,

$$U_i^{(t)}=\begin{pmatrix}u_{i0}^{(t)}\\u_{i1}^{(t)}\\\vdots\\u_{in}^{(t)}\end{pmatrix}=P\left(Y=k\middle|X_i,z_i,\alpha_k^{(t)},(\sigma_k^2)^{(t)}\right)X_i$$

$$=\begin{pmatrix}P\left(Y=k\middle|X_i,z_i,\alpha_k^{(t)},(\sigma_k^2)^{(t)}\right)\\x_{i1}P\left(Y=k\middle|X_i,z_i,\alpha_k^{(t)},(\sigma_k^2)^{(t)}\right)\\\vdots\\x_{in}P\left(Y=k\middle|X_i,z_i,\alpha_k^{(t)},(\sigma_k^2)^{(t)}\right)\end{pmatrix}$$

Note,

$$u_{ij}^{(t)}=x_{ij}P\left(Y=k\middle|X_i,z_i,\alpha_k^{(t)},(\sigma_k^2)^{(t)}\right)$$

The left-hand side of equation (24) becomes:

$$\sum_{i=1}^{N} \alpha_k^T U_i^{(t)} X_i^T = \alpha_k^T \sum_{i=1}^{N} U_i^{(t)} X_i^T = \alpha_k^T \sum_{i=1}^{N} \begin{pmatrix} u_{i0}^{(t)} \\ u_{i1}^{(t)} \\ \vdots \\ u_{in}^{(t)} \end{pmatrix} (x_{i0}, x_{i1}, \dots, x_{in})$$

$$= \alpha_k^T \sum_{i=1}^{N} \begin{pmatrix} u_{i0}^{(t)} x_{i0} & u_{i0}^{(t)} x_{i1} & \cdots & u_{i0}^{(t)} x_{in} \\ u_{i1}^{(t)} x_{i0} & u_{i1}^{(t)} x_{i1} & \cdots & u_{i1}^{(t)} x_{in} \\ \vdots & \vdots & \ddots & \vdots \\ u_{in}^{(t)} x_{i0} & u_{in}^{(t)} x_{i1} & \cdots & u_{in}^{(t)} x_{in} \end{pmatrix} = \alpha_k^T \left(\boldsymbol{U}^{(t)}\right)^T \boldsymbol{X}$$

Where $\boldsymbol{U}^{(t)}$ is specified by equation (25).

$$\boldsymbol{U}^{(t)} = \begin{pmatrix} \left(U_1^{(t)}\right)^T \\ \left(U_2^{(t)}\right)^T \\ \vdots \\ \left(U_N^{(t)}\right)^T \end{pmatrix} = \begin{pmatrix} u_{10}^{(t)} & u_{11}^{(t)} & \cdots & u_{1n}^{(t)} \\ u_{20}^{(t)} & u_{21}^{(t)} & \cdots & u_{2n}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N0}^{(t)} & u_{N1}^{(t)} & \cdots & u_{Nn}^{(t)} \end{pmatrix} \qquad (25)$$

$$u_{ij}^{(t)} = x_{ij} P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)$$

Let,

$$V_i^{(t)} = \begin{pmatrix} v_{i0}^{(t)} \\ v_{i1}^{(t)} \\ \vdots \\ v_{in}^{(t)} \end{pmatrix} = P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right) Z$$

$$= \begin{pmatrix} z_1 P\left(Y = k \middle| X_1, z_1, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right) \\ z_2 P\left(Y = k \middle| X_2, z_2, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right) \\ \vdots \\ z_N P\left(Y = k \middle| X_N, z_N, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right) \end{pmatrix}$$

Note,

$$v_{ij}^{(t)} = z_i P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)$$

The right-hand side of equation (24) becomes:

$$\sum_{i=1}^{N} V_i^{(t)} X_i^T = \sum_{i=1}^{N} \left(v_{i0}^{(t)} x_{i0}, v_{i1}^{(t)} x_{i1}, \dots, v_{in}^{(t)} x_{in}\right) = \left(V_i^{(t)}\right)^T \boldsymbol{X}$$

Where $V_i^{(t)}$ is specified by equation (26).

$$V_i^{(t)} = \begin{pmatrix} v_{i0}^{(t)} \\ v_{i1}^{(t)} \\ \vdots \\ v_{in}^{(t)} \end{pmatrix}$$
$$v_{ij}^{(t)} = z_i P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right) \tag{26}$$

Equation (24) becomes:

$$\alpha_k^T \left(\boldsymbol{U}^{(t)}\right)^T \boldsymbol{X} = \left(V_i^{(t)}\right)^T \boldsymbol{X}$$

Which is equivalent to the following equation:

$$\boldsymbol{X}^T \boldsymbol{U}^{(t)} \alpha_k = \boldsymbol{X}^T V_i^{(t)}$$

As a result, the next regression coefficient $\alpha_k^{(t+1)}$, which is solution of equation (23), is specified by equation (27).

$$\alpha_k^{(t+1)} = \left(\boldsymbol{X}^T \boldsymbol{U}^{(t)}\right)^{-1} \boldsymbol{X}^T V_i^{(t)} \tag{27}$$

Where $\boldsymbol{X}$, $\boldsymbol{U}^{(t)}$, and $V_i^{(t)}$ are specified by equations (17), (25), and (26), respectively. The proposed equation (27) is most important in this research because it is the integration of least squares method and mixture model. If we think deeply, it is the key to combine REM and mixture model. In other words, it is the key to combine two EM processes in the same loop.

By applying Lagrange method, each next partial variance $(\sigma_k^2)^{(t+1)}$ is specified by equation (28) [17, p. 7].

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^N \left(z_i - \left(\alpha_k^{(t+1)}\right)^T X_i\right)^2 P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)}{\sum_{i=1}^N P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)} \tag{28}$$

Where $P(Y=k \mid z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)})$ is specified by equation (22) and $\alpha_k^{(t+1)}$ is specified by equation (27).The proof of equations (21), (23), and (28) is found in [17, pp. 5-6].

By using maximum likelihood estimation (MLE) method [18, pp. 8-9], we retrieve equation (29) to estimate each next IRC $\beta_{kj}^{(t+1)}$ [1, p. 457].

$$\beta_{kj}^{(t+1)} = (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T X_j \tag{29}$$

Where $\boldsymbol{Z}$ and X_j are specified in equation (17). Not $\boldsymbol{Z}$ and X_j are fulfilled in E-step. In general, MREM is the full combination of REM and mixture model in which two EM processes are applied into the same loop of E-step and M-step. These steps are described in Table 1.

Table1. *Mixture Regression Expectation Maximization (MREM) Algorithm.*

1. E-step: This is the first EM process. Missing values $(z_i^-)_k$ and $(x_{ij}^-)_k$ for each k^{th} PRM are fulfilled by equation (20) given current parameter $\Theta^{(t)}$. Please pay attention that each k^{th} PRM owns a partial complete data $(\boldsymbol{X}_k, \boldsymbol{Z}_k)$. In other words, the whole sample $(\boldsymbol{X}, \boldsymbol{Z})$ has K versions $(\boldsymbol{X}_k, \boldsymbol{Z}_k)$ for K PRMs. Note, such K versions are changed over each iteration.

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

$$\left(x_{ij}^-\right)_k = \begin{cases} \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)}(z_i)_k \text{if} (z_i)_k \text{is not missing} \\ \beta_{kj0}^{(t)} + \beta_{kj1}^{(t)}(z_i^-)_k \text{if} (z_i)_k \text{is missing} \end{cases}$$

The whole sample $(\boldsymbol{X}, \boldsymbol{Z})$ is fulfilled to become complete data when its missing values z_i^- and x_{ij}^- are aggregated from $(z_i^-)_k$ and $(x_{ij}^-)_k$ of K versions $(\boldsymbol{X}_k, \boldsymbol{Z}_k)$, by equations (31) and (16).

$$z_i^- = \sum_{k=1}^{K} c_k^{(t)}(z_i^-)_k$$

$$x_{ij}^- = \sum_{k=1}^{K} c_k^{(t)}\left(x_{ij}^-\right)_k$$

2. M-step: This is the second EM process. The next parameter $\Theta^{(t+1)}$ is determined by equations (21), (27), (28), and (29) and the complete data $(\boldsymbol{X}, \boldsymbol{Z})$ fulfilled in E-step.

$$c_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)$$

$$\alpha_k^{(t+1)} = \left(\boldsymbol{X}^T\boldsymbol{U}^{(t)}\right)^{-1}\boldsymbol{X}^T V_i^{(t)}$$

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^{N}\left(z_i - \left(\alpha_k^{(t+1)}\right)^T X_i\right)^2 P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)}{\sum_{i=1}^{N} P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)}$$

$$\beta_{kj}^{(t+1)} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T X_j$$

Where $\boldsymbol{U}^{(t)}$ and $\boldsymbol{V}^{(t)}$ are specified by equations (25) and (26) and,

$$P\left(Y = k \middle| X_i, z_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right) = \frac{c_k^{(t)} P_k\left(z_i \middle| X_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)}{\sum_l^K c_l^{(t)} P_l\left(z_i \middle| X_i, \alpha_k^{(t)}, (\sigma_k^2)^{(t)}\right)}$$

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

The next parameter $\Theta^{(t+1)}$ becomes current parameter in the next iteration.

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

MREM is also a clustering method whose each resulted cluster is represented by a pair (α_k^*, $(\sigma_k^2)^*$). In other words, each cluster is represented by a PRM. As a convention, these clusters are called conditional clusters or regressive clusters because the mean of each cluster is $\mu_k^* = (\alpha_k^*)^T X_i$ given a data point X_i. This is an unexpecting but interesting result of REM. Given an observation $(X_i, z_i)^T = (x_{i0}, x_{i1},.., x_{in}, z_i)^T$, if the k^{th} PRM gives out the largest condition probability, it is most likely that X_i belongs to the k^{th} cluster represented by such k^{th} PRM. Let $cl(X_i, z_i, k)$ denote the probability of the event that a data point $(X_i, z_i)^T$ belongs to k^{th} cluster (k^{th} PRM). From equation (22), we have:

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

We use the complete case method mentioned in [2, p. 3] to improve the convergence of MREM. The parameters $(\alpha_k^{(1)}, \beta_{kj}^{(1)})^T$ at the first iteration of EM process are initialized in proper way instead that they are initialized in arbitrary way [20]. Let $\boldsymbol{X}_k$' be the complete matrix, which is created by removing all rows whose values are missing from $\boldsymbol{X}_k$. Similarly, let $\boldsymbol{Z}_k$' be the complete matrix, which is created by removing rows whose weights are missing from $\boldsymbol{Z}_k$. The advanced parameters $(\alpha_k^{(1)}, \beta_{kj}^{(1)})^T$ are initialized by equation (30) [1, p. 457].

$$\begin{aligned} \alpha_k^{(1)} &= ((\boldsymbol{X}'_k)^T \boldsymbol{X}'_k)^{-1} (\boldsymbol{X}'_k)^T Z'_k \\ \beta_{kj}^{(1)} &= ((\boldsymbol{Z}'_k)^T \boldsymbol{Z}'_k)^{-1} (\boldsymbol{Z}'_k)^T X'_{kj} \end{aligned} \quad (30)$$

Where Z_k' is the complete vector of non-missing outcome values for each k^{th} PRM and X_{kj}' is the complete column vector of non-missing regressor values for each k^{th} PRM.

The evaluation of MREM follows fully mixture model. For example, given input data vector $X_0 = (x_{01}, x_{02},\ldots, x_{0n})$, let $z_1, z_2,\ldots, z_K$ be the values evaluated from K PRMs with optimal PRCs α_k^* resulted from MREM shown in Table 1.

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

Where $x_{00} = 1$. The final evaluation z is calculated based on mixture coefficients, given data vector $X_0 = (x_{01}, x_{02},\ldots, x_{0n})$, as follows:

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

$$z=\sum_{k=1}^{K}c_k^*z_k=\sum_{k=1}^{K}c_k^*(\alpha_k^*)^TX_0=\sum_{k=1}^{K}c_k^*\sum_{j=0}^{n}\alpha_{kj}^*x_{0j} \tag{31}$$

In general, equation (31) is the final regression model of MREM. Following is the proof of equation (31). From equation (12), let $\hat{z}$ be the estimate of response variable z, we have:

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

We have assumed until now that the number K of PRMs is pre-defined and thus, another problem of MREM is how to determine K. Here we propose a so-called increasing algorithm without pre-defining K. In other words, REM associated with increasing algorithm can automatically determine K. Let k be initilized by 1, Followings are two steps of increasing algorithm:

1. Executing MREM with k PRMs and then, calculating the fitness $f(k)$ of the resulted mixture model with k PRMs. The fitness $f(k)$ measures adequacy of given mixture model.
2. Let $l = k + 1$, trying to execute MREM with k PRMs and then, to calculate the fitness $f(l)$ of the resulted mixture model with l PRMs. If $f(l) > f(k)$ then, setting $k = l$ and going back step 1; otherwise, the increasing algorithm stops with k PRMs.

The essence of increasing algorithm is how to calculate the fitness $f(k)$ because the final mixture model is the one whose fitness is largest. We define $f(k)$ as the sum of optimal partial probabilities $P_c(z_c \mid X_i, \alpha_c^*, (\sigma_c^2)^*)$ over all X_i. Equation (32) is the definition of $f(k)$.

$$f(k)=\sum_{X_i}\max_{c=1,2,\dots,k}P_c(z_c|X_i,\alpha_c^*,(\sigma_c^2)^*) \tag{32}$$

Where,

$$z_c=(\alpha_c^*)^TX_i=\sum_{j=0}^{n}\alpha_{cj}^*x_{ij}$$

For explanation, according to equation (32), for each data point X_i, we determine the largest partial probability $P_c(z_c \mid X_i, \alpha_c^*, (\sigma_c^2)^*)$ over $c = 1, 2,\dots, k$ as the optimal partial probability. Later one, the fitness $f(k)$ is the sum of all optimal partial probabilities over all X_i. We make experiment on MREM associated with increasing algorithm. I feel that increasing algorithm is not optimal because it seems to be a work-around solution for determining the number K of PRMs but I currently cannot think out better algorithm. In furture, we can research hiearchical clustering or BIC criterion [12] as alternative solution.

3. RESULTS AND DISCUSSIONS

The purpose of the experiment here is to compare MREM and REM. We use *xclara* sample of R statistical environment for testing MREM and REM. The xclara dataset was edited and published by Vincent Arel-Bundock [21]. It has 3000 points with 3 clusters. There are two numerical variables *V*1 and *V*2 as *x* and *y* coordinates of points in the xclara dataset. We consider *V*1 as regressor and *V*2 as response variable. The xclara dataset was originally used for clustering by Anja Struyf, Mia Hubert, and Peter Rousseeuw [22] but here it is used for regression analysis.

The dataset is split separately into one training dataset (50% sample) and one testing dataset (50% sample). Later on, the training dataset is made sparse with loss ratios 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%, which is similar to our previous research [20]. Missing values are made randomly regardless of regressors or response variable. For example, the xclara training dataset (50% xclara sample) has 50%*3000=1500 rows and each row has 2 columns (*V*1 and *V*2) and so the training dataset has 1500*5 = 7500 cells. If loss ratio is 10%, there are only 10%*7500=750 missing values which are made randomly among such 7500 cells. In other words, the incomplete training dataset with loss ratio 10% has 7500 – 750 = 6750 non-missing values. Of course, the testing dataset (50% sample) is not made sparse. Each pair of incomplete training dataset and testing dataset is called testing pair. There are ten testing pairs for each sample. As a convention, the origin testing pair which has no missing value in training dataset is the 0th pair. The 0th pair is called complete pair whereas the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, and 9th pairs are called incomplete pairs.

Firstly, we test MREM and REM with xclara sample. Table 2 [20] shows ten testing pairs of xclara sample.

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

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

Table 3 shows ten resulted regression models of REM corresponding to ten testing pairs of xclara sample.

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

Pair	Regression model
0	*V*2 = 34.0445 - 0.2790*(*V*1)
1	*V*2 = 36.8255 - 0.3384*(*V*1)
2	*V*2 = 36.5624 - 0.3393*(*V*1)
3	*V*2 = 37.4537 - 0.4022*(*V*1)
4	*V*2 = 45.8814 - 0.5830*(*V*1)
5	*V*2 = 48.8888 - 0.6477*(*V*1)
6	*V*2 = 55.9764 - 0.8593*(*V*1)
7	*V*2 = 48.8888 - 0.6477*(*V*1)
8	*V*2 = 69.1886 - 1.0823*(*V*1)
9	*V*2 = 62.2939 - 1.1417*(*V*1)

Table 4 shows ten resulted mixture regression models of MREM corresponding to ten testing pairs of xclara sample.

Table 4. *Ten resulted mixture regression models of MREM given xclara sample.*

Pair	Mixture regression model
0	{*V*2 = 34.0445 - 0.2790*(*V*1): coeff=1.0000, var=962.0000}
1	{*V*2 = 16.6425 - 0.3065*(*V*1): coeff=0.6654, var=188.4319}, {*V*2 = 62.3919 - 0.0429*(*V*1): coeff=0.3346, var=86.8709}
2	{*V*2 = 13.2805 - 0.3332*(*V*1): coeff=0.4909, var=124.9130}, {*V*2 = 64.0639 - 0.0980*(*V*1): coeff=0.3031, var=102.6651}, {*V*2 = 32.5172 - 0.2432*(*V*1): coeff=0.2060, var=0.0573}
3	{*V*2 = 13.2047 - 0.3220*(*V*1): coeff=0.4410, var=138.2844}, {*V*2 = 66.5083 - 0.1668*(*V*1): coeff=0.2424, var=91.6464}, {*V*2 = 31.9337 - 0.2667*(*V*1): coeff=0.3166, var=0.0323}
4	{*V*2 = 14.5836 - 0.3404*(*V*1): coeff=0.3772, var=132.5547}, {*V*2 = 65.9884 - 0.1683*(*V*1): coeff=0.2224, var=99.6319}, {*V*2 = 33.3280 - 0.2766*(*V*1): coeff=0.4004, var=0.0547}
5	{*V*2 = 33.5698 - 0.2666*(*V*1): coeff=0.5096, var=0.0346}, {*V*2 = 65.8616 - 0.1536*(*V*1): coeff=0.1906, var=83.6705}, {*V*2 = 13.6946 - 0.3393*(*V*1): coeff=0.2998, var=152.4883}
6	{*V*2 = 73.1729 - 1.1518*(*V*1): coeff=0.8835, var=49.5093}, {*V*2 = 7.3758 + 1.0052*(*V*1): coeff=0.1165, var=296.8865}
7	{*V*2 = 33.5698 - 0.2666*(*V*1): coeff=0.5096, var=0.0346}, {*V*2 = 65.8616 - 0.1536*(*V*1): coeff=0.1906, var=83.6705}, {*V*2 = 13.6946 - 0.3393*(*V*1): coeff=0.2998, var=152.4883}
8	{*V*2 = 69.1886 - 1.0823*(*V*1): coeff=1.0000, var=58.6193}
9	{*V*2 = 62.2939 - 1.1417*(*V*1): coeff=1.0000, var=21.6927}

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

In Table 4, each PRM is wrapped in two brackets "{.}". Notation "coeff" denotes mixture coefficient and notation "var" denotes the variance of a PRM. Note, MREM is also a clustering method where each regressive cluster is represented by a PRM. In other words, each PRM is considered as a regressive mean or regressive representative of a regressive cluster. However, regressive clustering with MREM is different from usual clustering. When data is visualized, we will see that the good number of regressive clusters is 2 whereas the best number of usual clusters in xclara sample is 3 [21]. Figure 1 shows the unique regressive cluster of the training dataset of the 0th testing pair.

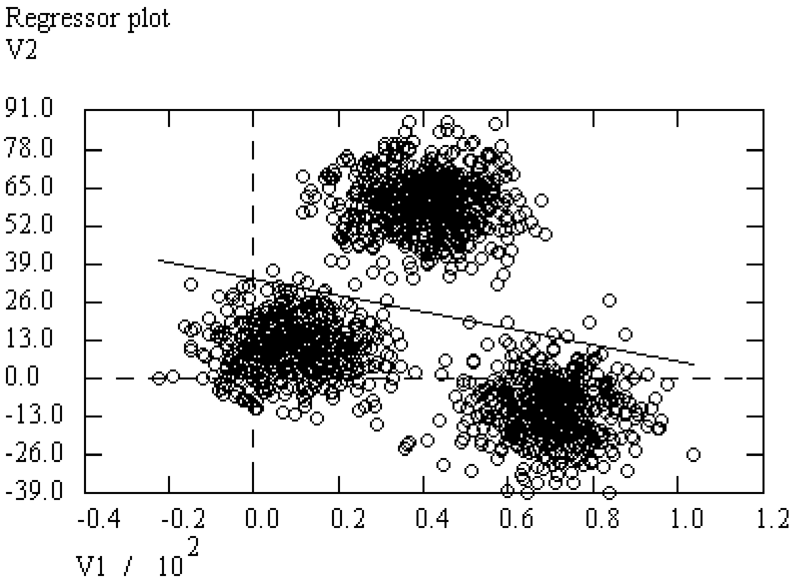

Figure 1. *Unique cluster of the training dataset of the 0th pair.*

The PRM is drawn as a thin and solid line going through the unique regressive cluster. Of course, such solid line shows the line equation of the PRM, $V2 = 34.0445 - 0.2790*(V1)$.

Figure 2 shows two regressive clusters of the training dataset of the 1st testing pair. Note, missing values in the 1st training dataset are fulfilled after MREM finished.

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

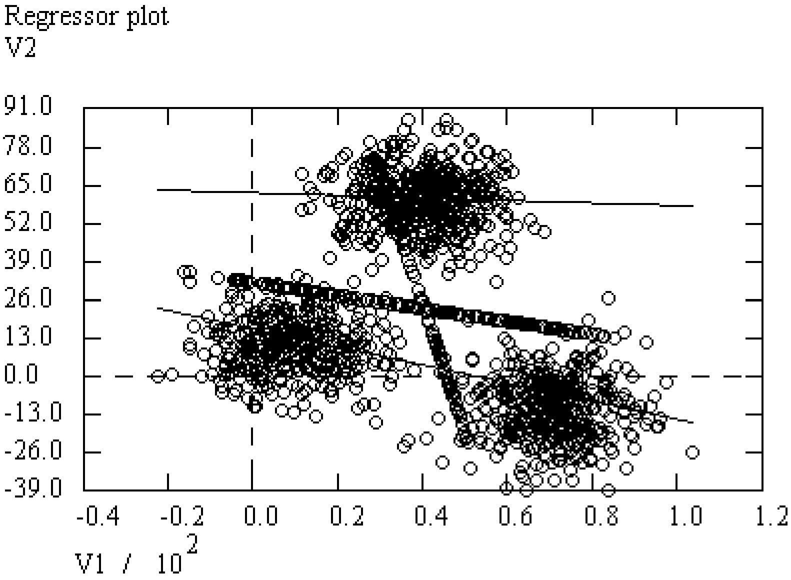

Figure 2. *Two regressive clusters of the training dataset of the 1st pair.*

As seen in figure 2, there are two solid lines which represents two PRMs. The upper solid line represents the PRM *V*2 = 64.0639 - 0.0980*(*V*1) whereas the lower solid line represents the PRM *V*2 = 62.3919 - 0.0429*(*V*1).

Given xclara sample, we compare MREM with REM with regard to the ratio mean absolute error (*RMAE*). Let $W = \{w_1, w_2, \ldots, w_K\}$ and $V = \{v_1, v_2, \ldots, v_K\}$ be sets of actual weights and estimated weights, respectively. Equation (33) specifies the *RMAE* metric [23, p. 814].

$$RMAE = \frac{1}{K}\sum_{i=1}^{K}\left|\frac{v_i - w_i}{w_i}\right| \quad (33)$$

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

Table 5. *Comparison of REM and MREM regarding RMAE, given xclara sample*

Pair	*RMAE* (REM)	*RMAE* (MREM)
0	5.4722	5.4722
1	5.3672	5.7804
2	5.2846	5.6044
3	4.6337	5.1166

SCIENTIFIC SOCIETY
MULTIDISCIPLINAR MAGAZINE

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

4	4.3681	5.3686
5	4.3025	5.5701
8	4.912	5.2689
7	4.3025	5.5701
8	6.1709	6.1709
9	7.2932	7.2932
Average	5.2107	5.7215

From Table 5, given xclara sample, MREM is less accurate than REM according *RMAE* metric. When I test MREM with other samples, it is also not better than REM in accuracy. This is an unexpected result which is easy to lead a conclusion that MREM is not useful. The reason of this unexpected result is that we cannot choose a right regressive cluster for given regressors *X* to estimate response value *z*. The equation (31) is the average formula for evaluating mixture model orver *K* PRMs and so it will produce unexpected bias. For example, I generate a sample in which there is only one regressor *x* and only one response variable *z*. There are 1000 points (*x*, *z*) in the generated sample. The variable *x* is randomized from 0 to 1. From 0 to 0.5, *x* and *z* satisfy the linear equation $z = x$ with variance 0.001. From 0.5 to 1, *x* and *z* satisfy the linear equation $z = 1 – x$ with variance 0.001. The probability of the equation $z = x$ is equal to the probability of the equation $z = 1 – x$, which is 0.5. MREM with pre-defined $K = 2$ produces the mixture model {{z = 0.0036 + 0.9765*(*x*), coeff=0.4634, var=0.0011}, {z = 0.9767 - 0.9713*(*x*), coeff=0.5366, var=0.0009}} which is an approximation of such two linear equations. Without loss of generity, training dataset is also used as testing dataset. Figure 3 shows the mixture model (K = 2) with regressive clusters.

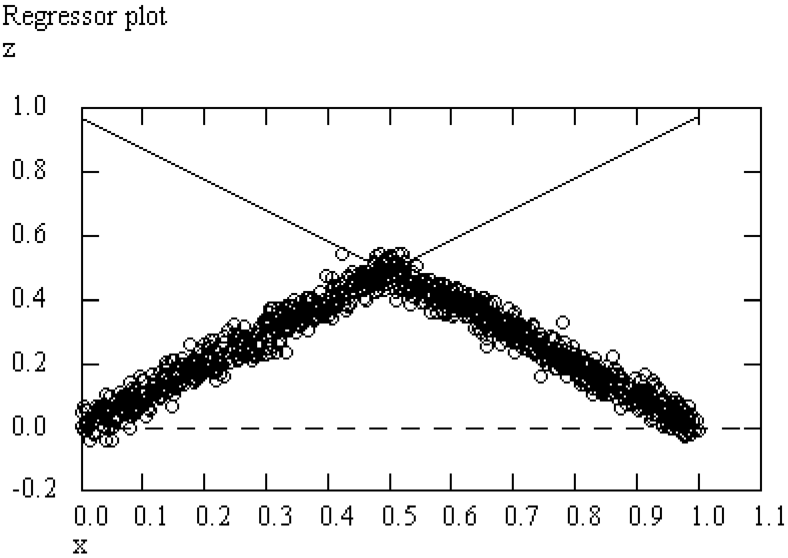

Figure 3. *Regressive clusters of the generated sample.*

In figure 3, the approximation of the equation $z = x$ is $z = 0.0036 + 0.9765*(x)$ with mixture coefficient $c_1 = 0.4634$ whereas the approximation of the equation $z = 1 – x$ is $z = 0.9767 - 0.9713*(x)$ with mixture coefficient $c_2 = 0.5366$. Given generated sample, the *RMAE* of MREM is 5.3957 which is worse than the *RMAE* of REM (2.5790). Obviously, although MREM produces a good approximation of the linear equations $z = x$ and $z = 1 – x$ such as $z = 0.0036 + 0.9765*(x)$ and $z = 0.9767 - 0.9713*(x)$, respectively but it cannot select the right one for estimating response values. MREM instead produces average values according equation (31). As a result, MREM gives out worse accuracy. If MREM can select the equation $z = 0.0036 + 0.9765*(x)$ and the equation $z = 0.9767 - 0.9713*(x)$ for estimating response values for $0 \leq x < 0.5$ and $0.5 \leq x \leq 1$, respectively then, the *RMAE* of MREM becomes 0.4 which is better the *RMAE* of REM (2.5790). In general, MREM is still useful because a different approach for mixture model can be referred by fusing linear equations of MREM into a unique curve equation. The curve modeling with regression analysis was proposed by Faicel Chamroukhi, Allou Samé, Gérard Govaert, and Patrice Aknin [4].

4. CONCLUSIONS

In general, the essence of MREM is to integrate two EM processes (one for exponential estimation of missing values and one for mixture model estimation of parameters) into the same loop with expectation that MREM will take advantages of both REM in fulfilling incomplete data and mixture model in processing complicatedly varied data. The proposed equation (27) is the key to combine REM and mixture model. Unfortunately, experimental result shows that MREM is less accurate than REM because MREM causes biases in estimating response values by

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018
ISSN: 2595-8402
DOI: 10.5281/zenodo.2528978

average formula specified by equation (31). However, MREM is essential because for further research, we will research some approximation techniques to fuse linear equations of mixture model into a unique curve equation. The curve modeling with regression analysis was poposed by Faicel Chamroukhi, Allou Samé, Gérard Govaert, and Patrice Aknin [4].

5. CONFLICTS OF INTEREST

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

6. ACKNOWLEDGMENTS

We express our deep gratitude to Prof. Dr. Thu-Hang Thi Ho (Vinh Long General Hospital – Vietnam) and Prof. Bich-Ngoc Tran who gave us helpful works, helpful comments, and testing sample to develop and test MREM with note that MREM is derived from REM.

7. REFERENCES

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

[2] Horton, N. J.; Kleinman, K. P. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, February 2007, vol. 61, no. 1, pp. 79-90. DOI:10.1198/000313007X172556.

[3] Nguyen, L.; Ho, T.-H. T. Fetal Weight Estimation in Case of Missing Data. *Experimental Medicine (EM)* - Special Issue “Medicine and Healthy Food”, December 17th 2018, vol. 1, no. 2, pp. 45-65. DOI:10.31058/j.em.2018.12004.

[4] Chamroukhi, F.; Samé, A.; Govaert, G.; Aknin, P. (2010, March). A hidden process regression model for functional data description: Application to curve discrimination. (Wang, Z.; Hoi, S.; Eds.) *Neurocomputing*, March 2010, 73(7-9), 1210-1221, DOI:10.1016/j.neucom.2009.12.023. Available online:https://www.sciencedirect.com/science/article/pii/S0925231210000287 (accessed on 12ndOctober 2018).

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

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

[6] Ghitany, M. E.; Karlis, D.; Al-Mutairi, D. K.; Al-Awadhi, F. An EM Algorithm for Multivariate Mixed Poisson Regression Models and its Application. *Applied Mathematical Sciences*, 2012, vol.6, no.137, pp.6843-6856. Available online: http://www.m-hikari.com/ams/ams-2012/ams-137-140-2012/ghitanyAMS137-140-2012.pdf (accessed on 3rd July2018).

[7] Anderson, B.; Hardin, M. J. Modified logistic regression using the EM algorithm for reject inference. *International Journal of Data Analysis Techniques and Strategies*, 1st January 2013, vol. 5, no. 4, pp.359-373. DOI:10.1504/IJDATS.2013.058582.

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

[9] Haitovsky, Y. Missing Data in Regression Analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1st January 1968, vol. 30, no. 1, pp. 67-82. Available online: https://www.jstor.org/stable/2984459 (accessed on 3rdJuly2018).

[10] Robins, J. M.; Rotnitzki, A.; Zhao, L. P. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, March 1995, vol. 90, no. 429, pp. 106-121. DOI:10.2307/2291134.

[11] Lamont, A. E.; Vermunt, J. K.; Lee, V. H. M. Regression mixture models: Does modeling the covariance between independent variables and latent classes improve the results? *Multivariate Behavioral Research*, January 2016, vol. 51, no. 1, pp. 35-52. DOI:10.1080/00273171.2015.1095063.

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

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

[14] Sung, H. G. *Gaussian Mixture Regression and Classification*. Rice University, Houston, 2004. Available online: https://scholarship.rice.edu/handle/1911/18710 (accessed on 4th September2018).

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

[16] Grün, B.; Leisch, F. *Finite Mixtures of Generalized Linear Regression Models*. University of Munich, Munich, 2007. Available online:

VOLUME 1, NÚMERO 3, DEZEMBRO DE 2018

ISSN: 2595-8402

DOI: 10.5281/zenodo.2528978

https://pdfs.semanticscholar.org/e0d5/6ac54b80a1a4e274f11b1d86840461cc542c.pdf (accessed on 4th September2018).

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

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

[19] Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1977, vol. 39, no. 1, pp. 1-38.

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

[21] Arel-Bundock, V. (2018, June 28). R datasets - An archive of datasets distributed with R. *GitHub*, 28th June 2018. Available online: http://vincentarelbundock.github.io/Rdatasets/csv/cluster/xclara.csv (accessed on 11st September 2018).

[22] Struyf, A.; Hubert, M.; Rousseeuw, P. J. (1996). Clustering in an Object-Oriented Environment. *Journal of Statistical Software*, 1996, 1(4), 1-30, DOI:10.18637/jss.v001.i04. Available online: http://www.jstatsoft.org/v01/i04 (accessed on 10th October 2018).

[23] Pinette, M. G.; Pan, Y.; Pinette, S. G.; Blackstone, J.; Garrett, J.; Cartin, A. Estimation of Fetal Weight: Mean Value from Multiple Formulas. *Journal of Ultrasound in Medicine*, 1st December 1999, vol. 18, no. 12, pp. 813-817. Available online: https://www.ncbi.nlm.nih.gov/pubmed/10591444 (accessed on 9th October 2016).