Austrian Journal of Statistics
*July 2019, Volume 48, 90–100.*
*http://www.ajs.or.at/*
doi:10.17713/ajs.v48i4.851

# A Bayesian Approach to Estimate a Linear Regression Model with Aggregate Data

**Aymen Rawashdeh**
Yarmouk University

**Mohammed Obeidat**
Yarmouk University

**Abstract**

The main purpose of this paper is to perform linear regression analysis on a continuous aggregate outcome from a Bayesian perspective using a Markov chain Monte Carlo algorithm (Gibbs sampling). In many situations, data are partially available due to privacy and confidentiality of the subjects in the sample. So, in this study, the vector of outcomes, $\mathbf{Y}$, is realistically assumed to be missing and is partially available through summary statistics, sum $(\mathbf{Y})$, aggregated over groups of subjects, while the covariate values, $\mathbf{X}$, are available for all subjects in the sample. The results of the simulation study highlight both the efficiency of the regression parameter estimates and the predictive power of the proposed model compared with classical methods. The proposed approach is fully implemented in an example regarding systolic blood pressure for illustrative purposes.

*Keywords*: aggregate information, Bayesian analysis, linear regression, MCMC.

## 1. Introduction

The relation among two or more observable quantities is the main concern of many scientific studies. Particularly, how a response outcome $Y$ varies as function of a p−vector of quantities $\mathbf{X}$. Typically, both $\mathbf{X}$ and $Y$ are available for each individual in the sample, let $D_C$ be the complete data set, where $D_C = \{(\mathbf{X}_i, Y_i), 1 \leq i \leq n\}$. The model in matrix-form is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, such that $\mathbf{X}$ is an $n \times p$ design matrix consisting of $p-1$ explanatory variables with the first column being fixed at 1 for the intercept, therfore, the regression parameters are $\boldsymbol{\beta}' = [\beta_0, \beta_1, \ldots, \beta_{p-1}]$. The error term $\mathbf{e}$ follows the multivariate normal distribution, MVN $\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$, where $\mathbf{I}_n$ is the identity $n \times n$ matrix, and $\mathbf{Y}$ is a vector of $n$ outcomes. Based on this classical scenario, the standard statistical regression models can be used to draw meaningful statistical inferences about the regression parameters $\boldsymbol{\beta}$ and $\sigma^2$.

However, in many situations, data are partially available due to privacy and confidentiality of the subjects in the study, see Moineddin and Urquia (2014). In this article, we realistically assume that the $Y$ values are missing for subjects, but only aggregate information, i.e., sums, minimum, and maximum of $Y$ values of groups of subjects are available, while the $\mathbf{X}$ values are available for each individual in the study. In notation, suppose there are $J$ groups of data. The $j^{th}$ group has $n_j$ subjects, where $j = 1, \ldots, J$ and $n_1 + n_2 + \cdots + n_J = n$. The outcome of the $i^{th}$ subject in the $j^{th}$ group is $Y_{ij}$ with covariates $\mathbf{X}_{ij} = [1, X_{ij\,1}, X_{ij\,2}, \ldots, X_{ij\,p-1}]$,

which is a $(1 \times p)$ matrix. Define the following J-dimensional vectors. For $j = 1, \ldots, J$ let

- $\mathbf{Y}^*$: The vector composed of the group-sum values, $\mathbf{Y}^* = \left[\mathrm{Y}_j^*\right]$, where $\mathrm{Y}_j^* = \sum_{i=1}^{n_j} Y_{ij}$.
- $\mathbf{Y}_{min}$: The vector composed of the group-minimum values, $\mathbf{Y}_{min} = \left[\mathrm{Y}_{j(1)}\right]$, where $\mathrm{Y}_{j(1)} = \min\left\{Y_{1\,j},\, Y_{2\,j}, \ldots,\, Y_{n_j\,j}\right\}$.
- $\mathbf{Y}_{max}$: The vector composed of the group-maximum values, $\mathbf{Y}_{max} = \left[\mathrm{Y}_{j(n_j)}\right]$, where $\mathrm{Y}_{j(n_j)} = \max\left\{Y_{1\,j},\, Y_{2\,j}, \ldots,\, Y_{n_j\,j}\right\}$.

Set $D_{PA}$ to be the partially aggregate data where the design matrix $\mathbf{X}$, $\mathbf{Y}^*$, $\mathbf{Y}_{min}$, and $\mathbf{Y}_{max}$ are available, while $\mathbf{Y}_j{}' = \left[Y_{1\,j},\, Y_{2\,j}, \ldots,\, Y_{n_j\,j}\right]$ are missing for all $j = 1, \ldots, J$. Notice that $D_{PA}$ can not be utilized by the standard statistical methods to estimate $\left(\boldsymbol{\beta},\, \sigma^2\right)$.

The most common approach to deal with such type of missing data is to restrict the analysis to the group level data $D_G = \left\{\left(\mathbf{X}_j^*,\, \mathrm{Y}_j^*\right),\, 1 \leq j \leq J\right\}$, where $\mathbf{X}_j^* = \sum_{i=1}^{n_j} \mathbf{X}_{ij}$, which is a completely-aggregated data. In this case, $D_G$ serves as a bivariate random sample of size $J$, so it can be fitted by the standard statistical regression theory with a design matrix $\mathbf{X}^* = \begin{bmatrix} \mathbf{X}_1^* \\ \vdots \\ \mathbf{X}_J^* \end{bmatrix}$, of dimension $J \times p$, and $\mathbf{Y}^*$ as a vector of $J$ outcomes. Analysis based on completely-aggregated data, $D_G$, is popular especially in meta-analysis, for more information see Moineddin and Urquia (2014) and Riley, Lambert, Staessen, Wang, Gueyffier, Thijs, and Boutitie (2008).

However, such analysis suffers from several disadvantages; (1) Aggregating both $X$ and $Y$ values smooths out variation at the individual level and therefore, synthetically, inflates the correlation coefficient values. (2) Most importantly, assuming that relationships observed for groups necessarily hold for individuals is known as ecological fallacy, therefore models fit group level data can not be applied to individual-level data. For further details about ecological fallacy see Dias, Sutton, Welton, and Ades (2013) and Freedman (1999). (3) When $J$ is small and group sizes are generally large, then $D_G$ has very few amount of information that can be relied on. Retrospectively, it is promising to conduct inference based on partially aggregated data, which is the main idea of this article.

The proposed method is motivated by the work done by Choi, Schervish, Schmitt, and Small (2008), who developed a Bayesian approach to a logistic regression model applied to partially aggregate data. The objective of the proposed method is to, simultaneously, predict the missing values of $Y$ and to estimate the regression model parameters, $\boldsymbol{\beta}$ and $\sigma^2$, based on only the partially aggregate data $D_{PA}$. Predicting the missing values, is tempted by the idea of missing data imputation. For a thorough discussion of missing data imputation see Gelman, Carlin, Stern, and Rubin (2014) and Rubin (2004). It is expected that the inference based on $D_{PA}$ should be more accurate when compared with the one based on $D_G$ especially when $J$ is small and group sizes are, generally, large.

The paper is organized as follows. The main elements for Bayesian inference are presented in section 2. The behavior of both the Bayesian estimates and the predictive power are exploited via a simulation study in Section 3. A discussion is led about the conditions under which the proposed method outperforms traditional method. Section 4 illustrates the approach using a real data. Finally, we make concluding remarks in section 5.

## 2. Bayesian analysis

This section is devoted to propose two Bayesian models; (1) the partially aggregate model, $\text{Model}_{PA}$, and (2) the group level model, $\text{Model}_G$, which are applied to fit both $D_{PA}$ and $D_G$ , respectively. Two important components need to be specified in order to conduct proper

Bayesian analysis; (1) The loss function, and (2) The prior knowledge about the parameters of interest expressed as a probability distribution. The Bayesian estimators of both models are developed using the squared error loss function. Non-informative and weakly-informative priors are used so that posterior densities would be dominated by the sample data, Gelman *et al.* (2014).

Once the prior distributions of the unknown parameters are determined, they are combined with the likelihood function to obtain a joint posterior distribution for the unknown parameters. Typically, the joint posterior distribution of the parameters of interest is analytically intractable. So Markov chain Monte Carlo algorithm (MCMC) together with Gibbs sampling are employed to make inference about the parameters of interest.

### 2.1. The group level model ($Model_G$)

The group level data $D_G$ is modeled by $\mathrm{Y}_j^* = \mathbf{X}_j^*\boldsymbol{\beta} + e_j^*$, where the errors $e_j^*$ are independent normally distributed with mean 0 and variance $n_j\sigma^2$, for $j = 1, \ldots, J$.

*Prior:*

A weakly-informative prior is assigned to parameter space of $(\boldsymbol{\beta}, \sigma^2)$ as follows. Set the prior $\pi(\sigma^2) \propto \left(\sigma^2\right)^{-(a+1)} e^{-\frac{b}{\sigma^2}}$, which is the kernel of the Inverse-Gamma distribution with hyper-parameters shape $= a$ and rate $= b$. The variance of the Inverse-Gamma distribution is $\frac{b^2}{(a-1)(a-2)}$. So, when $a = 2.0005$ and $b = 1$ the variance of $\pi(\sigma^2)$ becomes very large which produces a weakly-informative prior for $\sigma^2$. The parameter vector $\boldsymbol{\beta}$ are assumed to be priori indepndent with flat prior on their domain, that is $\pi(\boldsymbol{\beta})$=1. The parameters $\boldsymbol{\beta}$ and $\sigma$ are assumed to be priori independent, therefore the joint prior distribution is

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \left(\sigma^2\right)^{-(a+1)} e^{-\frac{b}{\sigma^2}}.$$

*Posterior:*

The joint posterior distribution of $\left(\boldsymbol{\beta}, \sigma^2\right)$ given the groupe level data is

$$\pi\left(\boldsymbol{\beta}, \sigma^2 | \, \mathbf{Y}^*, \mathbf{X}^*\right) \propto \pi\left(\boldsymbol{\beta}, \sigma^2\right) P\left(\mathbf{Y}^* | \, \sigma^2, \boldsymbol{\beta}, \mathbf{X}^*\right),$$

where $P\left(\mathbf{Y}^* | \, \sigma^2, \boldsymbol{\beta}, \mathbf{X}^*\right)$ is the likelihood function of the grouped data. After simple algebras, the joint posterior distribution can be written as

$$\pi\left(\boldsymbol{\beta}, \sigma^2 | \, \mathbf{Y}^*, \mathbf{X}^*\right) \propto \left(\sigma^2\right)^{-\left(\frac{J}{2}+a+1\right)} \exp - \left\{ \frac{\frac{1}{2}\left(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}\right)' \mathbf{W}^{-1}\left(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}\right) + b}{\sigma^2} \right\}, \quad (1)$$

where $\mathbf{W} = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & n_J \end{bmatrix}$.

The joint posterior distribution of $\left(\boldsymbol{\beta}, \sigma^2\right)$ can be factored as

$$\pi(\boldsymbol{\beta}, \sigma^2 | \, \mathbf{Y}^*, \mathbf{X}^*) = \pi\left(\boldsymbol{\beta} | \, \sigma^2, \mathbf{Y}^*, \mathbf{X}^*\right) \pi\left(\sigma^2 | \, \mathbf{Y}^*, \mathbf{X}^*\right)$$

Notice that $\pi\left(\boldsymbol{\beta} | \, \sigma^2, \mathbf{Y}^*, \mathbf{X}^*\right)$ is the conditional posterior distribution for $\boldsymbol{\beta}$ given $\sigma^2$ and $\pi\left(\sigma^2 | \, \mathbf{Y}^*, \mathbf{X}^*\right)$ is the marginal posterior distribution for $\sigma^2$, see Gelman *et al.* (2014), which are as follows.

The marginal posterior distribution for $\sigma^2$ can be shown to be the Inv-Gamma distribution with shape A=$a + \frac{J-p}{2}$ and rate $B = b + \frac{1}{2}\left(\mathbf{Y}^*\right)'\left(\mathbf{W}^{-1} - \mathbf{H}\right)\mathbf{Y}^*$, where

$$\mathbf{H} = \mathbf{W}^{-1}\mathbf{X}^*\left[\left(\mathbf{X}^*\right)'\mathbf{W}^{-1}\mathbf{X}^*\right]^{-1}\left(\mathbf{W}^{-1}\mathbf{X}^*\right)'.$$

Notice that the matrix, $\mathbf{H}$, is the same as the Hat matrix for weighted least square regression. Moreover, $\left(\mathbf{Y}^*\right)'\left(\mathbf{W}^{-1} - \mathbf{H}\right)\mathbf{Y}^*$ represents weighted-residual sum of squares. That is

$$\pi\left(\sigma^2|\,\mathbf{Y}^*,\,\mathbf{X}^*\right) \propto \pi\left(\sigma^2,\,\boldsymbol{\beta}\right)\left(\sigma^2\right)^{-\frac{J-p}{2}}\exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}\right)'\mathbf{W}^{-1}\left(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}\right)\right]$$

The posterior distribution of $\boldsymbol{\beta}$ given $\sigma^2$ can be shown to be multivariate normal with mean

$$\hat{\boldsymbol{\beta}} = \left[\left(\mathbf{X}^*\right)'\mathbf{W}^{-1}\mathbf{X}^*\right]^{-1}\left(\mathbf{X}^*\right)'\mathbf{W}^{-1}\mathbf{Y}^*$$

and Variance-Covariance matrix $\sigma^2\left[\left(\mathbf{X}^*\right)'\mathbf{W}^{-1}\mathbf{X}^*\right]^{-1}$. That is,

$$\pi\left(\boldsymbol{\beta}|\,\sigma^2,\,\mathbf{Y}^*,\,\mathbf{X}^*\right) =$$
$$\left(2\pi\sigma^2\right)^{-\frac{p}{2}}\left|\left[\left(\mathbf{X}^*\right)'\mathbf{W}^{-1}\mathbf{X}^*\right]^{-1}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{X}^*\right)'\mathbf{W}^{-1}\mathbf{X}^*\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]$$

The Markov chain Monte Carlo algorithm is implemented by the following steps:

- for $j = 1, \ldots, N$, where $N$ is the number of MCMC iterations,
- sample $\sigma^{2^{(j)}}$ from $\pi\left(\sigma^2|\,\mathbf{Y}^*,\,\mathbf{X}^*\right)$,
- sample $\boldsymbol{\beta}^{(j)}$ from $\pi\left(\boldsymbol{\beta}|\,\sigma^{2^{(j)}},\,\mathbf{Y}^*,\,\mathbf{X}^*\right)$.

## 2.2. The partially aggregate model ($Model_{PA}$)

The major idea of the proposed partially aggregate model is the treatment of the missing responses $\mathbf{Y}$ as additional parameters to be estimated. So a Bayesian inference about the parameter vector $\left(\mathbf{Y},\,\boldsymbol{\beta},\,\sigma^2\right)$, based on $D_{PA}$, is desired. According to Bayes' Theorem the posterior distribution $\pi(\mathbf{Y},\,\boldsymbol{\beta},\,\sigma^2|\,\mathbf{Y}^*,\,\mathbf{Y}_{min},\,\mathbf{Y}_{max},\,\mathbf{X})$ equals to

$$\pi(\boldsymbol{\beta},\sigma^2)\,\pi(\mathbf{Y}\,|\,\boldsymbol{\beta},\,\sigma^2,\,\mathbf{Y}^*,\,\mathbf{Y}_{\min},\,\mathbf{Y}_{\max},\,\mathbf{X}).$$

Set the prior $\pi(\boldsymbol{\beta},\sigma^2) \propto \left(\sigma^2\right)^{-(a+1)}e^{-\frac{b}{\sigma^2}}$, which is the weakly-informative density obtained in sub-section 2.1.1. Markov chain Monte Carlo algorithm with Gibbs sampling is implemented by specifying both $\pi(\boldsymbol{\beta},\,\sigma^2|\,\mathbf{Y},\,\mathbf{X})$ and $\pi(\mathbf{Y}\,|\,\boldsymbol{\beta},\,\sigma^2,\,\mathbf{Y}^*,\,\mathbf{Y}_{\min},\,\mathbf{Y}_{\max},\,\mathbf{X})$ which are provided below:

1. $\pi(\boldsymbol{\beta},\,\sigma^2|\,\mathbf{Y},\,\mathbf{X})$: Given $\mathbf{Y}$, the joint posterior distribution $\pi(\boldsymbol{\beta},\,\sigma^2|\,\mathbf{Y},\,\mathbf{X})$ is factored as $\pi(\boldsymbol{\beta}|\,\sigma^2,\,\mathbf{Y},\,\mathbf{X})\pi(\sigma^2|\,\mathbf{Y},\,\mathbf{X})$. Notice that the conditional posterior distribution $\pi(\boldsymbol{\beta}|\,\sigma^2,\,\mathbf{Y},\,\mathbf{X})$ is a multivariate normal distribution with mean $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$ and variance-covariance matrix $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$. The marginal posterior distribution $\pi(\sigma^2|\,\mathbf{Y},\,\mathbf{X})$ is the Inv-Gamma distribution with shape A=$a + \frac{n-p}{2}$ and rate $B = b + \frac{\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{2}$.
2. $\pi(\mathbf{Y}\,|\,\boldsymbol{\beta},\,\sigma^2,\,\mathbf{Y}^*,\,\mathbf{Y}_{\min},\,\mathbf{Y}_{\max},\,\mathbf{X})$: The next sub-section is devoted to explain how to generate the missing values $\mathbf{Y}$ given $\boldsymbol{\beta}$, $\sigma^2$, $\mathbf{Y}^*$, $\mathbf{Y}_{min}$, and $\mathbf{Y}_{max}$.

*Generating the missing outcomes*

This section aims to discuss how to generate the missing outcomes, $\mathbf{Y}$, Given $\beta$, $\sigma^2, \mathbf{Y}^*, \mathbf{Y}_{min}$, and $\mathbf{Y}_{max}$

Given $\left(\sigma^2, \boldsymbol{\beta}\right)$, $\mathbf{Y}_1, \ldots, \mathbf{Y}_J$ are conditionally independent random vectors, i.e.

$$P(\mathbf{Y} \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{Y}^*, \mathbf{Y}_{\min}, \mathbf{Y}_{\max}, \mathbf{X}) = \prod_{j=1}^{J} P(\mathbf{Y}_{\mathrm{j}} \,|\, \boldsymbol{\beta}, \sigma^2, \mathrm{Y}_{\mathrm{j}}^*, \mathrm{Y}_{\mathrm{j}(1)}, \mathrm{Y}_{\mathrm{j}(\mathrm{n_j})}, \mathbf{X}_{\mathrm{j}}).$$

So, without loss of generality, only the distribution of $P(\mathbf{Y}_{\mathrm{j}} \,|\, \boldsymbol{\beta}, \sigma^2, \mathrm{Y}_{\mathrm{j}}^*, \mathrm{Y}_{\mathrm{j}(1)}, \mathrm{Y}_{\mathrm{j}(\mathrm{n_j})}, \mathbf{X}_{\mathrm{j}})$ is considered. Notice that, $P(\mathbf{Y}_{\mathrm{j}} \,|\, \boldsymbol{\beta}, \sigma^2, \mathrm{Y}_{\mathrm{j}}^*, \mathrm{Y}_{\mathrm{j}(1)}, \mathrm{Y}_{\mathrm{j}(\mathrm{n_j})}, \mathbf{X}_{\mathrm{j}})$ is proportional to

$$P(\mathbf{Y}_{\mathrm{j}} \,|\, \boldsymbol{\beta}, \sigma^2, \mathrm{Y}_{\mathrm{j}}^*, \mathbf{X}_{\mathrm{j}})$$

with a restricted support, where each component $y_{i\,j} \in \left[\mathrm{Y}_{j(1)}, \mathrm{Y}_{j(n_j)}\right]$, for all $1 \leq i \leq n_j$. So $P(\mathbf{Y}_{\mathrm{j}} \,|\, \boldsymbol{\beta}, \sigma^2, \mathrm{Y}_{\mathrm{j}}^*, \mathbf{X}_{\mathrm{j}})$ is firstly obtained and then its truncation, based on the restricted support, is followed. It can be shown that $P(\mathbf{Y}_{\mathrm{j}} \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{X}_{\mathrm{j}})$ is MVN $\left(\boldsymbol{\mu}_{\mathbf{j}}, \sigma^2 \mathbf{I}_{n_j}\right)$, where $\boldsymbol{\mu}_j^{'} = \left[\mathbf{X}_{1j}\boldsymbol{\beta}, \ldots, \mathbf{X}_{n_j j}\boldsymbol{\beta}\right]$. Since $\mathrm{Y}_j^*$ is given, then the number of missing responses in each $\mathbf{Y}_{\mathrm{j}}$ is $(n_j - 1)$. Set $\boldsymbol{\theta}_j^{'} = \left[y_{1\,j}, y_{2\,j}, \ldots, y_{(n_j-1)\,j}\right]$ to be the missing responses that are considered as additional parameters to be estimated. So we need to find $P(\boldsymbol{\theta}_{\mathrm{j}} \,|\, \boldsymbol{\beta}, \sigma^2, \mathrm{Y}_{\mathrm{j}}^*, \mathbf{X}_{\mathrm{j}})$. For this purpose, consider the linear transformation $\mathbf{W} = \mathbf{T}\mathbf{Y}_j$ , where $\mathbf{T}$ is square matrix of size $n_j$ given by $\mathbf{T} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$. As a block matrix, $\mathbf{T}$ can be written as $\begin{bmatrix} \mathbf{I}_{(n_j-1)} & \mathbf{0} \\ \mathbf{1}^{'} & 1 \end{bmatrix}$, where $\mathbf{0}$ is a vector of zeros of dimension $(n_j - 1)$, and $\mathbf{1}$ is a vector of ones of dimension $(n_j - 1)$.

Notice that

$$\mathbf{W} = \begin{bmatrix} \boldsymbol{\theta}_j \\ \mathrm{Y}_j^* \end{bmatrix} \text{ given } \left(\boldsymbol{\beta}, \sigma^2, \mathbf{X}\right) \text{ follows}$$

$$\text{MVN}\left(\begin{bmatrix} \mathbf{X}_{1j}\boldsymbol{\beta} \\ \vdots \\ \mathbf{X}_{(n_j-1)j}\boldsymbol{\beta} \\ \left(\sum_{i=1}^{n_j} \mathbf{X}_{ij}\right)\boldsymbol{\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{I}_{(n_j-1)} & \mathbf{1} \\ \mathbf{1}^{'} & n_j \end{bmatrix}\right)$$

By using the conditional distribution theory of the multivariate normal distribution, derived by Anderson (1957), $P(\boldsymbol{\theta}_{\mathrm{j}} \,|\, \boldsymbol{\beta}, \sigma^2, \mathrm{Y}_j^*, \mathbf{X})$ is a multivariate normal distribution with mean $\boldsymbol{\mu}_j$ and variance-covariance matrix $\sum_j$ such that,

$$\begin{aligned} \boldsymbol{\mu}_j &= \begin{bmatrix} \mathbf{X}_{1j}\boldsymbol{\beta} \\ \vdots \\ \mathbf{X}_{(n_j-1)j}\boldsymbol{\beta} \end{bmatrix} + \frac{1}{n_j} \begin{bmatrix} \mathrm{Y}_j^* - \left(\sum_{i=1}^{n_j} \mathbf{X}_{ij}\right)\boldsymbol{\beta} \\ \vdots \\ \mathrm{Y}_j^* - \left(\sum_{i=1}^{n_j} \mathbf{X}_{ij}\right)\boldsymbol{\beta} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{X}_{1j}\boldsymbol{\beta} \\ \vdots \\ \mathbf{X}_{(n_j-1)j}\boldsymbol{\beta} \end{bmatrix} + \mathbf{1}\left(\bar{\mathrm{Y}}_{\mathbf{j}} - \bar{\mathbf{X}}_{\mathbf{j}}\boldsymbol{\beta}\right), \end{aligned}$$

and

$$\sum\nolimits_j = \sigma^2 \left(\mathbf{I}_{(n_j-1)} - \frac{1}{n_j}\mathbf{1}\mathbf{1}^{'}\right).$$

Notice that, for $i_1 \neq i_2$, the Correlation $(Y_{i_1 j},\, Y_{i_2 j}) = \frac{Cov\left(Y_{i_1 j}, Y_{i_2 j}\right)}{\sqrt{Var\left(Y_{i_1 j}\right) Var\left(Y_{i_2 j}\right)}} = -\frac{1}{(n_j - 1)}$.

Finally a truncated form of the conditional density $P(\boldsymbol{\theta}_{\mathrm{j}} \,|\, \boldsymbol{\beta},\, \sigma^2,\, \mathrm{Y}_j^*,\, \mathbf{X}_j)$ is considered by restricting the support to be $Y_{ij} \in \left[\mathrm{Y}_{j(1)},\, \mathrm{Y}_{j(n_j)}\right]$, for all $1 \leq i \leq (n_j - 1)$. In other words, given $\boldsymbol{\beta}$, $\sigma^2$, $\mathrm{Y}_j^*$, $\mathrm{Y}_{j(1)}$, and $\mathrm{Y}_{j(n_j)}$, $\mathbf{Y}_j$ is generated from truncated multivariate normal distribution. See Horrace (2005) and Wilhelm (2015) for the properties and the generation from truncated multivariate normal distribution, respectively. This completes the Bayesian model specification and it sets the stage for Gibbs sampler to be implemented as follows

- Step One: Given $\mathbf{Y}$, $\left(\boldsymbol{\beta},\, \sigma^2\right)$ is obtained as: Firstly, $\sigma^2$ is generated from $\pi(\sigma^2 |\, \mathbf{Y},\, \mathbf{X})$. Then, given $\sigma^2$, $\boldsymbol{\beta}$ is generated from $\pi(\boldsymbol{\beta} |\, \sigma^2,\, \mathbf{Y},\, \mathbf{X})$.
- Step Two: Given $\left(\boldsymbol{\beta},\, \sigma^2\right)$, $\boldsymbol{\theta}_j = \left[Y_{1 j},\, Y_{2 j}, \ldots,\, Y_{(n_j - 1) j}\right]'$ is generated from the truncated multivariate normal density,

$$P(\boldsymbol{\theta}_{\mathrm{j}} \,|\, \boldsymbol{\beta},\, \sigma^2,\, \mathrm{Y}_j^*,\, \mathrm{Y}_{j(1)},\, \mathrm{Y}_{j(n_j)},\, \mathbf{X}_j).$$

  Then set $\mathbf{Y}_j = \begin{bmatrix} \boldsymbol{\theta}_j \\ \mathrm{Y}_j^* - \sum_{i=1}^{(n_j - 1)} Y_{ij} \end{bmatrix}$. Repeat this step for each $j = 1, \ldots, J$. See Wilhelm and Manjunath (2010) for generating truncated multivariate normal vectors using R-package (tmvtnorm).
- Step Three: Repeat the above two steps for $N$ MCMC iterations

The missing observations $\boldsymbol{\theta}_j$ in each group can be predicted from their posterior predictive distribution given by

$$\int\int P(\boldsymbol{\theta}_{\mathrm{j}} \,|\, \boldsymbol{\beta},\, \sigma^2,\, \mathbf{Y}^*,\, \mathbf{Y}_{\min},\, \mathbf{Y}_{\max},\, \mathbf{X}) \pi(\boldsymbol{\beta},\, \sigma^2 \,|\, \mathbf{Y}^*,\, \mathbf{Y}_{min},\, \mathbf{Y}_{max},\, \mathbf{X}) d\sigma^2 d\boldsymbol{\beta}. \tag{2}$$

## 3. Simulation study

A simulation study is conducted to investigate, based on empirical evidence, the performance of the parameter estimators and the predictive power of both models; Model$_G$ and Model$_{PA}$ in terms of the mean squared errors by considering different values of the parameters. The sample size is varied to observe the effect of small and large samples on the performance. At each scenario, 2000 data sets are simulated from the true model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. Both $D_{PA}$ and $D_G$ are constructed from each simulated dataset. The Bayesian models Model$_G$ and Model$_{PA}$ are applied to each corresponding simulated data set, with an MCMC of length 10,000 iterations such that 6000 burn-in and the last 4000 draws as the posterior sample from the target joint posterior distribution.

### 3.1. Simulation

- Set the regression parameters $\boldsymbol{\beta} = (\beta_0,\, \beta_1) = (0.5,\, 1.1)$.
- Decide the scenario by setting: (1) Number of groups; $J$ to be one of the values from $\{4,\, 6,\, 10,\, 15\}$. (2) Group size; Although we tried two scenarios for group sizes, one with equal group sizes and the other with unequal group sizes, we show here only the results for equal group sizes because the results were similar in both scenarios. We set the group size to be one of the values from $\{n_1 = n_2 = \cdots = n_J = I = 6,\, 8,\, 12,\, or\, 20\}$. (3) $\sigma^2$ to be one of the values from $\{10,\, 20,\, or\, 30\}$.

- For each scenario:
  - Generate $n = IJ$ values from uniform (1, 20) to be the covariate values that constitute the second column of the design matrix $\mathbf{X}$, with the first column being fixed at 1.
  - Decide the three location $\left\{X_{(2)}, \text{Median}, X_{(n-1)}\right\}$ at which prediction is desired.
  - Divide the list $1, \ldots, n$ into $J$ groups, with each group having $I$ observations such that $n = IJ$.
  - Repeat the following steps (1 through 6) 2000 times.
    1. Generate $\mathbf{Y}$ values from MVN $\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$.
    2. For each $1 \leq j \leq J$, obtain $\mathrm{Y}_j^* = \sum_{i=1}^{n_j} Y_{ij}$, $\mathrm{Y}_{j(1)} = \max\left\{Y_{1\,j}, Y_{2\,j}, \ldots, Y_{n_j\,j}\right\}$, $\mathrm{Y}_{j(n_j)} = \max\left\{Y_{1\,j}, Y_{2\,j}, \ldots, Y_{n_j\,j}\right\}$, and $\mathbf{X}_j^* = \sum_{i=1}^{n_j} \mathbf{X}_{ij}$.
    3. The partially aggregate data $D_{PA}$ is composed of $\mathbf{Y}^*$, $\mathbf{Y}_{min}$, $\mathbf{Y}_{max}$, and the design matrix $\mathbf{X}$, while $D_G = \left\{\left(\mathbf{X}_j^*, \mathrm{Y}_j^*\right), 1 \leq j \leq J\right\}$.
    4. Both $\text{Model}_{PA}$ and $\text{Model}_G$ are applied to $D_{PA}$ and $D_G$ to produce $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}\right)_{PA}$ and $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}\right)_G$, respectively.
    5. $\text{Model}_{PA}$ predicts the outcome $Y$, at the three predetermined locations, by using the formula in (2).
    6. $\text{Model}_G$ predicts the outcome $Y$, at the three predetermined locations, by using the formula $\hat{Y} = \left(\frac{1}{I}\right)\left(\hat{\beta}_0\right)_G + \left(\hat{\beta}_1\right)_G X$. Notice that ,$\left(\hat{\beta}_0\right)_G$, the intercept parameter estimate resulted from $\text{Model}_G$, needs to be multiplied by $\left(\frac{1}{I}\right)$. The reason for this is because the prediction is desired for a single $Y$ value, while the estimators are obtained based on $\mathrm{Y}_j^*$, which is a sum of $Y$ values. In other words $(\beta_0)_G$ and $Y$ are not on the same scale because of the aggregation, while the other regression parameters, $\beta_1$, are not affected by aggregation.

### 3.2. Simulation results

The simulation results are listed in this section. The mean square error of the estimates of $\beta_1$ are provided in Table 1. Table 2 shows the coverage probability and half length of a 95% highest posterior credible interval for $\beta_1$. The square root of the mean square error of the estimates of $\sigma^2$ are provided in Table 3. Table 4 provides the prediction error of the missing outcomes using both models $\text{Model}_{PA}$ and $\text{Model}_G$. The prediction power of the proposed model is investigated at 3 distant locations in the range of $X$ values. Two locations on the boundaries; which are $X_{(2)}$ (second minimum) and $X_{(n-1)}$(second maximum) . While, the third location is at the median, which is considered to examine predicting $Y$ values when $X$ is close to the center of the data. Conclusions are drawn regarding the behavior of the estimators and the predictive power, which are summarized in the next subsection. Notice that:

1. The first three scenarios consider the case of increasing $\sigma^2$ while keeping the sample size, $n = IJ = (6)\,(8) = 48$, fixed.
2. The last three scenarios consider the case of increasing the number of groups $J$ while keeping, $\sigma^2 = 30$, fixed.
3. The two middle-scenarios, when$(J,\, I) = (4,\, 12)$ or $(4,\, 20)$, are considered to assess the realistic case when there are few number of groups with a lot of missing data in each group. In this regard, it is expected that $D_{PA}$ is efficiently utilized by$\text{Model}_{PA}$ which takes advantage over $\text{Model}_G$.

Table 1: MSE for $\hat{\beta}_1$

| $\sigma^2 =$ | 10 | 20 | 30 | 20 | 30 | 30 | 30 |
|---|---|---|---|---|---|---|---|
| $(J, I) =$ | (6, 8) | (6, 8) | (6, 8) | (4, 12) | (4, 20) | (10, 6) | (15, 6) |
| $\text{Model}_{PA}$ | 0.669 | 1.09 | 0.654 | 0.805 | 1.429 | 0.422 | 0.413 |
| $\text{Model}_G$ | 0.924 | 7.78 | 1.215 | 1.896 | 9.495 | 0.645 | 0.488 |

Table 2: Coverage prob. and half length (in parenthesis) of a 95% credible interval for $\beta_1$

| $\sigma^2 =$ | 10 | 20 | 30 | 20 | 30 | 30 | 30 |
|---|---|---|---|---|---|---|---|
| $(J, I) =$ | (6, 8) | (6, 8) | (6, 8) | (4, 12) | (4, 20) | (10, 6) | (15, 6) |
| $\text{Model}_{PA}$ | 0.85(0.61) | 0.82(0.82) | 0.84(0.90) | 0.80(0.92) | 0.78(1.00) | 0.85(0.82) | 0.74(0.68) |
| $\text{Model}_G$ | 0.89(1.47) | 0.85(2.13) | 0.86(2.20) | 0.76(2.72) | 0.79(3.27) | 0.91(1.98) | 0.95(1.45) |

### 3.3. Discussion

1. In terms of estimating the slope of the regression line, based on Table 1, as generally expected, the mean square error decreases when sample size increases or the variability exhibited by the data decreases. It is obvious that $\text{Model}_{PA}$ outperforms $\text{Model}_G$ across all scenarios, especially, when $J$ is small and $I$ is large. The improvement of the proposed model in reducing the $MSE\left(\hat{\beta}_1\right)$ is attributed to the practical importance of the truncation of the underlying multivariate normal distribution.

2. Table 2 shows the coverage probabilities and half length of a 95% credible interval for $\beta_1$. It can be seen that $\text{Model}_{PA}$ have a comparable results with $\text{Model}_G$ in terms of coverage probability, while it has smaller half length. That means the interval estimates of $\text{Model}_{PA}$ is more accurate than $\text{Model}_G$.

3. Based on Table 3, as the sample size increases or $\sigma^2$ decreases the MSE of the variance estimates provided by both methods decreases. It seems that the proposed model performs as well, if not slightly worst than the group level model. Both models suffer from under-estimating $\sigma^2$ for completely different reasons. The under-estimation of $\sigma^2$ occurs when using $\text{Model}_{PA}$ is a result of the truncation of the underlying multivariate normal distribution. While, the under-estimation of $\sigma^2$ occurs when using $\text{Model}_G$ is explained by Freedman (1999) who mentioned that aggregating smooths out the variation.

4. A cursory examination of Table 4, the prediction error decreases as the data exhibits less variability or the sample size increases. The smallest prediction error occurs at the median, which is expected because there are more data points in the centre than on the boundaries. Apparently, across all scenarios, $\text{Model}_{PA}$ outperforms $\text{Model}_G$ in predicting $Y$ values especially at the boundaries of the $X$ range, see when$(J, I) =$

Table 3: Square root MSE for $\hat{\sigma^2}$

| $\sigma^2 =$ | 10 | 20 | 30 | 20 | 30 | 30 | 30 |
|---|---|---|---|---|---|---|---|
| $(J, I) =$ | (6, 8) | (6, 8) | (6, 8) | (4, 12) | (4, 20) | (10, 6) | (15, 6) |
| $\text{Model}_{PA}$ | 6.189 | 13.35 | 20.442 | 14.760 | 22.076 | 18.58 | 17.788 |
| $\text{Model}_G$ | 5.607 | 11.58 | 16.917 | 14.071 | 21.179 | 13.73 | 10.880 |

Table 4: Prediction error at three $X$ locations

| | | $X_{(2)}$ | Median | $X_{(n-1)}$ |
|---|---|---|---|---|
| $\sigma^2 = 10,\ J = 6,\ I = 8$ | $\text{Model}_{PA}$ | 17.521 | 9.097 | 18.547 |
| | $\text{Model}_G$ | 21.291 | 9.975 | 23.543 |
| $\sigma^2 = 20,\ J = 6,\ I = 8$ | $\text{Model}_{PA}$ | 29.98 | 17.73 | 26.83 |
| | $\text{Model}_G$ | 95.52 | 19.93 | 82.54 |
| $\sigma^2 = 30,\ J = 6,\ I = 8$ | $\text{Model}_{PA}$ | 29.335 | 26.729 | 33.164 |
| | $\text{Model}_G$ | 35.394 | 29.821 | 44.904 |
| $\sigma^2 = 20,\ J = 4,\ I = 12$ | $\text{Model}_{PA}$ | 22.828 | 17.779 | 28.540 |
| | $\text{Model}_G$ | 31.309 | 18.982 | 45.764 |
| $\sigma^2 = 30,\ J = 4,\ I = 20$ | $\text{Model}_{PA}$ | 48.221 | 28.825 | 38.781 |
| | $\text{Model}_G$ | 162.39 | 29.572 | 114.272 |
| $\sigma^2 = 30,\ J = 10,\ I = 6$ | $\text{Model}_{PA}$ | 28.04 | 25.024 | 27.76 |
| | $\text{Model}_G$ | 34.11 | 28.35 | 32.21 |
| $\sigma^2 = 30,\ J = 15,\ I = 6$ | $\text{Model}_{PA}$ | 26.120 | 25.457 | 26.043 |
| | $\text{Model}_G$ | 30.322 | 29.503 | 29.498 |

(4, 12) or (4, 20). The superiority of the prediction power of the proposed model could be a result of implementing $P(\mathbf{Y} \,|\, \boldsymbol{\beta},\, \sigma^2,\, \mathbf{Y}^*,\, \mathbf{X})$ in $\text{Model}_{PA}$.

5. The overall picture that these results show is the superiority of the proposed Bayesian method in estimating $\boldsymbol{\beta}$ and predicting $\mathbf{Y}$, especially, when $J$ is small and $I$ is large.

# 4. Real dataset analysis

## 4.1. Data description

To illustrate our method on real data, we consider a data set concerning the relationship between systolic blood pressure (BP) for 28 people and their ages. The complete data is available in Kleinbaum, Kupper, Nizam, and Rosenberg (2013). Let $Y$ be Systolic blood pressure and $X$ be the person age, so $D_C = \{(X_i,\, Y_i)\,,\, 1 \leq i \leq 28\}$. The systolic blood pressure value is predicted given the age of the person using both $\text{Model}_{PA}$ and $\text{Model}_G$ when applied to $D_{PA}$ and $D_G$, respectively, which are shown in Table 5 and Table 6.

Table 5: The group level data $D_G$

| $D_G$ | $\mathbf{X}^*$(Year) | $\mathbf{Y}^*$(mm Hg) |
|---|---|---|
| Group1 | 329 | 1005 |
| Group2 | 318 | 1010 |
| Group3 | 263 | 887 |
| Group4 | 328 | 979 |

## 4.2. Data results

This subsection is devoted to investigate the predictive power of the proposed model when applied to a real data. Let $\widehat{Y}_{PA,\,i}$, $\widehat{Y}_{G,\,i}$ be the predicted values of $Y_i$ (for $i = 1, \ldots, 28$) using $\text{Model}_{PA}$ and $\text{Model}_G$, respectively. Since the true value, $Y$, is provided in $D_C$, then we are

Table 6: The partially aggregate data $D_{PA}$

| $D_{PA}$ | $X$ (Year) | | | | | | | $\mathbf{Y}^*$(mm Hg) | $\mathbf{Y}_{\min}$ | $\mathbf{Y}_{\max}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Group1 | 67 | 25 | 63 | 36 | 65 | 56 | 17 | 1005 | 114 | 170 |
| Group2 | 46 | 19 | 64 | 39 | 39 | 44 | 67 | 1010 | 120 | 162 |
| Group3 | 34 | 42 | 29 | 42 | 47 | 21 | 48 | 887 | 110 | 145 |
| Group4 | 56 | 59 | 53 | 45 | 20 | 50 | 45 | 979 | 116 | 158 |

Table 7: Prediction Error for Model$_{PA}$ and Model$_G$

| Model | Prediction Error |
|---|---|
| Model$_{PA}$ | $\sum_{i=1}^{28}\left(\widehat{Y}_{PA,i}-Y_i\right)^2=2073.3$ |
| Model$_G$ | $\sum_{i=1}^{28}\left(\widehat{Y}_{G,i}-Y_i\right)^2=6299.6$ |

able to calculate the prediction error for each model. Apparently, from Table 6, the proposed model, Model$_{PA}$, produces less prediction error compared with Model$_G$.

## 5. Conclusions

The present study proposed a linear regression analysis with a continuous outcome missing for subjects but where only aggregate information is available. Two Bayesian approaches are discussed, namely the group level and the partially aggregate models. For both models, Bayesian estimators cannot be obtained in closed forms and hence Gibbs sampling was used to sample from the conditional posteriors of the parameters. The results of the simulation study and the real data analysis demonstrate the superiority of the proposed Bayesian partially aggregate model over the existed group level model in estimating $\boldsymbol{\beta}$ and predicting the missing observations $\mathbf{Y}$. Therefore, the authors recommend applying partialy aggregate Bayesian method for parameter estimation and prediction in linear regression with missing data. Further research can be done to study the performance of the partially aggregate models in the presence of missing outcomes when the errors are dependent with different structure of the covariance matrix and/or when they follow a t-distribution rather than the normal distribution.

## References

Anderson TW (1957). "An Introduction to Multivariate Statistical Analysis." *Inc, London, Chapman and Hall, Limited*. doi:10.2307/2531310.

Choi T, Schervish MJ, Schmitt KA, Small MJ (2008). "A Bayesian Approach to a Logistic Regression Model with Incomplete Information." *Biometrics*, **64**(2), 424–430. doi:10.1111/j.1541-0420.2007.00887.x.

Dias S, Sutton AJ, Welton NJ, Ades AE (2013). "Evidence Synthesis for Decision Making 3 Heterogeneity-Subgroups, Meta-Regression, Bias, and Bias-Adjustment." *Medical Decision Making*, **33**(5), 618–640. doi:10.1177/0272989X13485157.

Freedman DA (1999). "Ecological Inference and the Ecological Fallacy." *International Encyclopedia of the social & Behavioral sciences*, **6**, 4027–4030.

Gelman A, Carlin JB, Stern HS, Rubin DB (2014). *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA. doi:10.2307/2988417.

Horrace WC (2005). "Some Results on the Multivariate Truncated Normal Distribution." *Journal of Multivariate Analysis*, **94**(1), 209–221. doi:10.1016/j.jmva.2004.10.007.

Kleinbaum DG, Kupper LL, Nizam A, Rosenberg ES (2013). *Applied Regression Analysis and Other Multivariable Methods*. Nelson Education. doi:10.1080/00401706.1989.10488486.

Moineddin R, Urquia ML (2014). "Regression Analysis of Aggregate Continuous Data." *Epidemiology*, **25**(6), 929–930. doi:10.1097/EDE.0000000000000172.

Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F (2008). "Meta-Analysis of Continuous Outcomes Combining Individual Patient Data and Aggregate Data." *Statistics in medicine*, **27**(11), 1870–1893. doi:0.1002/sim.3165.

Rubin DB (2004). *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons. doi:10.1002/9780470316696.

Wilhelm S (2015). "Gibbs Sampler for the Truncated Multivariate Normal Distribution." *Note*.

Wilhelm S, Manjunath BG (2010). "tmvtnorm: A Package for the Truncated Multivariate Normal Distribution." *Sigma*, **2**(2). doi:10.32614/RJ-2010-005.

**Affiliation:**

Aymen Rawashdeh
Department of Statistics
Yarmouk University
Irbid, Jordan
E-mail: ayman.r@yu.edu.jo

Mohammed Obeidat
Department of Statistics
Yarmouk University
Irbid, Jordan
E-mail: mohammad.obidat@yu.edu.jo