STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. xx–xxx, https://doi.org/10.59139/stattrans-2025-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

An expectation-maximization algorithm for logistic regression based on individual-level predictors and aggregate-level response

Zheng Xu[1]

Abstract

Logistic regression is widely used in complex data analysis. When predictors are at individual level and the response at aggregate level, logistic regression can be estimated using the Maximum Likelihood Estimation (MLE) method with the joint likelihood function formed by Poisson binomial distributions. When directly maximizing the complicated likelihood function, the performance of MLE will worsen as the number of predictors increases. In this article, we propose an expectation-maximization (EM) algorithm to avoid the direct maximization of the complicated likelihood function. Simulation studies have been conducted to evaluate the performance of our EM estimator compared to different estimators proposed in the literature. Two real data-based studies have been conducted to illustrate the use of the different estimators. Our EM estimator proves efficient for the logistic regression problem with an aggregate-level response and individual-level predictors.

Key words: expectation-maximization algorithm, missing values, Poisson binomial distribution, logistic regression, data aggregation, numerical optimization.

1. Introduction

With the fast development in technology, massive complex data have been collected from multiple sources. New methods have been proposed for complex data situations such as (1) how to deal with semi-structured data and structured data in the web (Zhai and Liu, 2006; Getdoor and Mihalkova, 2011), (2) analysis of graph-structured data (Geamsakul et al., 27 2005; Henaff et al., 2015) and (3) multi-level and mixed-level data analysis (Primo et al., 2007; Saramago et al., 2012).

Data can be collected, reported, and are available at different levels due to a range of reasons such as confidentiality, data collection difficulty, and cost saving. For example, the United State Department of Agriculture (USDA) National Agricultural Statistical Services (NASS) (`https://www.nass.usda.gov/`) reports agricultural crop yields at the county level instead of at the farm level, where county-level average or total is aggregated or estimated based on farm-level data in each county and farm-level data are confidential and unavailable to the public. Business data may only publish aggregated commodity purchase data at the store level and the month level to the public and keep individual-level data and

[1]Correspondence Author. Department of Mathematics and Statistics, Wright State University, Dayton, OH, USA. E-mail: zheng.xu@wright.edu. ORCID: https://orcid.org/0000-0003-0311-7004.

daily data confidential. Biological data, social-economic data, survey data, business data are often collected and reported at different levels.

Data can be aggregated in different ways. For example, a sequential two-stage testing method is used to study infectious diseases in epidemiology and bio-statistics. In the first stage, group testing is applied to the combined sample. In the second stage, individuals showing positive in the first stage are called for testing at the individual level. This group-testing strategy has been widely used in coronavirus disease 2019 (COVID-19) testing to increase efficiency and reduce cost (Mercer and Salit, 2021). Group-level Y in Group i can be calculated via the formula $Y_i = 1(\sum_{j=1}^{n_i} Y_{ij} > 0)$, where Y_{ij} is the response for the j-th person in group i, n_i is the number of individuals in group i, and $1(.)$ is the indicator function. The US Census Bureau reports household income as aggregate-level Y and individual income as individual-level Y, the aggregation method is summation, i.e. $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, where Y_{ij} is the income of the j-th person in the i-th household.

When individual-level X and individual-level Y are modeled by logistic regression, individual-level Y follows a Binomial Distribution with success probability as a function of individual-level X, denoted as $\pi(X) = \exp(X^T\beta)/(1+\exp(X^T\beta))$. Then aggregate-level Y, as the sum of individual-level Y, follows a Poisson-Binomial distribution (Hong, 2013; Xu, 2023). A complicated likelihood function $L(\beta)$ is derived and we previously proposed MLE estimator $\hat{\beta}_{MLE} = \text{argmax}_\beta L(\beta)$ with satisfactory statistical performance (Xu, 2023).

Because the maximization of the complicated likelihood function $L(\beta)$ is with respect to $\beta \in \mathscr{R}^p$, the performance of $\hat{\beta}_{MLE}$ will decrease when the dimension p increases (Xu, 2023). Different optimization methods to maximize the likelihood function have been considered and compared in our previous study (Xu, 2023).

We noticed that the limitation of $\hat{\beta}_{MLE}$ is mainly due to the direct optimization of the complicated likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$, formed by Poisson binomial distributions. This optimization can be avoided when an expectation-maximization (EM) method is adopted. As stated in Hastie et al (2009) and Givens and Hoeting (2012), the EM algorithm is a popular tool for simplifying difficult maximum likelihood problems for which maximization of the likelihood function is difficult, but made easier by enlarging the sample with latent data, i.e. a data-argumentation process. For our logistic regression problem with individual-level X and aggregate-level Y, we can enlarge the sample with the latent individual-level Y. One reason for using latent individual-level Y is that the usual logistic regression can be easily conducted when both X and Y are at the same level. Under mild conditions, this usual logistic regression has a unique MLE solution as a convex optimization problem with a convex objective function (Agresti, 2013; Hilbe, 2009). The unique solution can be obtained numerically via Newton's method, which uses the observed second derivative or the Fisher scoring method, which uses the expectation of this second derivative, and the Fisher scoring method is an application of the method of iteratively reweighted least squares (IRLS) (Agresti, 2013; Hilbe, 2009). Our EM algorithm conducts the usual logistic regression using IRLS method with stable performance and avoids the difficult optimization of the complicated likelihood function. Another reason to propose our estimator as an EM algorithm is that the EM algorithm view our problem in the perspective of missing values and data augmentation. This different perspective, compared with our previously proposed MLE estimator, makes our problem easily adapted and extensible to more complicated but

similar problems including (1) the problem that predictors X themselves are at mixed levels, (2) the problem that both X and Y contain missing values and (3) the problem that X and Y are modeled via a generalized linear model (GLM). Both the EM algorithm and our previously proposed MLE estimator in Xu (2023) have their own advantage in model extension to solve more complicated data situations, and choosing which is better depends on specific data situations. Therefore, both EM estimator and MLE estimator are necessary in methodological development of logistic regression.

The aim of this article is to study the performance of EM estimator in logistic regression based on aggregate-level Y and individual-level X. We proposed our EM estimator in Section 2. We conducted simulation studies to evaluate the performance and compare our EM estimator with literature estimators in Section 3. We illustrated the use of different estimators in real data-based studies in Section 4. We provided discussion in Section 5 and made conclusions in Section 6.

2. Materials and Methods

2.1. Data and Model Specification

Suppose there are N independent individuals aggregated into M groups, with group i having n_i individuals, i.e. $N = \sum_{i=1}^{M} n_i$. Denote (X_{ij}, Y_{ij}), $X_{ij} \in \mathscr{R}^p$, $Y_{ij} \in \mathscr{R}$, $i = 1,2,\ldots,M$, $j = 1,2,\ldots,n_i$, as the predictor vector and the response for the j-th individual in the i-th group. Thus, X_{ij} and Y_{ij} are individual-level predictor vector (X) and individual-level response (Y). Aggregate-level Y is obtained by summation within a group, i.e. $Y_i = \sum_{j=1}^{n_i} Y_{ij}$. Suppose there is a logistic regression model at the individual level, i.e.

$$\ln\left(\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right) = X_{ij}^T\beta,\ i = 1,2,...,M,\ j = 1,2,...,n_i. \tag{1}$$

Then $Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$, where $\pi_{ij} = P(Y_{ij}=1) = \frac{\exp(X_{ij}^T\beta)}{1+\exp(X_{ij}^T\beta)}$. When individual-level X and individual-level Y are both available, the logistic model as a generalized linear model (GLM) can be estimated using a range of methods including the Newton-Raphson method and the Fisher scoring method and the Fisher scoring method is an application of the method of iteratively reweighted least squares (IRLS) (Agresti, 2013; Givens and Hoeting, 2012). We name the logistic regression based on X and Y at the same level as the "usual" logistic regression (Agresti, 2013; Givens and Hoeting, 2012), to be compared with our problem of conducting logistic regression based on individual-level X and aggregate-level Y, which was considered in Xu (2023) and this article.

2.2. Joint Likelihood and MLE Method

Then the distribution of aggregate-level response, $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, as the sum of n_i independent Bernoulli random variables $Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$, $j = 1,2,\ldots,n_i$, is a Poisson

binomial distribution, i.e.

$$Y_i \sim \text{PoissonBinomial}(n_i, (\pi_{i1}, \pi_{i2}, \cdots, \pi_{in_i})), \tag{2}$$

where $\pi_{ij} = P(Y_{ij} = 1) = \frac{\exp(X_{ij}^T \beta)}{1+\exp(X_{ij}^T \beta)}$ (Wang, 1993; Hong, 2013; Xu, 2023).

The joint likelihood function is

$$L(\beta) = \prod_{i=1}^{M} P(Y_i | X_{i1}, ..., X_{in_i}; \beta), \tag{3}$$

where $P(Y_i | X_{i1}, ..., X_{in_i}; \beta)$ is the probability of Y_i, as specified in Equation 2.

The calculation of probability for a Poisson binomial distribution is complicated. In general, for a variable $Y \sim \text{PoissonBinomial}(n, (\pi_1, \pi_2, \ldots, \pi_n))$, the probability mass function is $P(Y = y) = \sum_{A \in F_y} \prod_{i \in A} \pi_i \prod_{j \in A^c} (1 - \pi_j)$, where F_y is the set of all subsets of y integers that can be selected from $\{1, 2, 3, ..., n\}$ and A^c is the complement of A (Wang, 1993). The set F_k contains $\binom{n}{k}$ elements so the sum over it is computationally intensive and even infeasible for large n. Instead, more efficient ways were proposed, including the use of a recursive formula to calculate $P(Y = y)$ based on $P(Y = k)$, $k = 0, ..., y-1$, which is numerically unstable for large n (Chen et al., 1994), and the inverse Fourier transform method (Fernandez and Williams, 2010). Hong (2013) further developed it by proposing an algorithm that efficiently implements the exact formula with a closed expression for the Poisson binomial distribution (Hong, 2013). We adopted Hong's algorithm and exact formula in calculating the likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$ in Equation 3 since they are more precise and numerically stable (Xu, 2023). Three optimization methods (Nelder and Mead's simplex method (NM) (Nelder and Mead, 1965), the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (Fletcher, 1970), and the conjugate gradient (CG) method (Fletcher and Reeves, 1964)) to maximize the joint likelihood function $L(\beta)$ were compared in Xu (2023) and the three methods show similar performance with NM method slightly better as our recommended method, and NM method is derivative free (Xu, 2023; Givens and Hoeting, 2012). We note that along the category of methods of directly optimizing the likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$, there can be a range of potential methods including evolutional algorithm and simulated annealing which may have similar or even better performance compared with our recommended directly optimization method (Givens and Hoeting, 2012; Xu, 2023). The search of optimization methods which directly maximizes $L(\beta)$ is not the objective of this article. Instead, we intend to develop methods not in this category of methods, i.e. methods not directly maximizing $L(\beta)$.

2.3. Expectation Maximization Algorithm

As an optimization problem $\max_\beta L(\beta)$, $\beta \in \mathscr{R}^p$, its performance will become worse when the number of predictors p increases. The objective function $L(\beta)$ is a complicated likelihood function so that we consider whether it is possible to circumvent or avoid the direct optimization of $L(\beta)$.

We noticed that the "usual" logistic regression, i.e. logistic regression when X and Y are

at the "same" level, is numerically stable and relatively easy to calculate. However, for our data situation, the usual logistic regression is infeasible because individual-level X is not available. To address this issue, we view our problem as a missing value problem where the latent variable is individual-level Y and we adopt an EM algorithm to substitute it. In each iteration of the EM algorithm, the usual logistic regression is conducted with individual-level Y, i.e. Y_{ij}, substituted with its expectation given current-iteration parameter estimate, i.e. $E(Y_{ij}|Y_i, \beta^{(k)})$, where $\beta^{(k)}$ is the estimated value of parameter β in iteration k.

Illuminated by the materials of EM algorithm in Hastie et al. (2009) and Givens and Hoeting (2012), we developed the EM algorithm for our problem. The EM algorithm is described as Algorithm 1 in the following page. The estimator obtained via the EM algorithm is named as the EM estimator.

One advantage of EM estimator is that it can avoid the direct optimization of the complicated nonlinear likelihood function $L(\beta)$. EM algorithm conducts the usual weighted logistic regression in each iteration. EM estimator is expected to have similar performance or even potentially slightly better performance compared with our MLE estimator in Xu (2023), which directly maximizes $L(\beta)$.

Another advantage of EM estimator is that it views our problem in a different perspective, i.e. missing values and data augmentation. This makes our method easily adapted and extensible for some applications. Potential applications which our EM algorithm may solve after modifications include (1) the situation where X are at mixed levels, i.e. different predictors are at levels, (2) the situation where there are missing values in X and Y, (3) the situation where individual-level X and Y is described by a generalized linear model (GLM), and (4) the situation where the objective is to use a variational Bayes to find a posteriori estimation (MAP) and make use of prior information (Bernardo et al., 2003).

3. Simulation Studies

3.1. Simulation Setups

We conducted simulation studies to evaluate the performance of the following four estimators. Estimator 1, named as "individual-LR", is the logistic regression estimator based on individual-level X and individual-level Y. This estimator is infeasible in our data situation where only aggregate-level Y is available. Because aggregate-level Y contains less information compared to individual-level Y, we expect that this infeasible estimator can provide an upper bound for the performance of feasible estimators based on aggregate-level Y. Estimator 2, named as "naive–LR", is the logistic regression estimator based on the aggregate-level X, which is the mean of X in each group, and the aggregate-level Y, i.e. $Y_i \sim \text{Binomial}(n_i, \sum_{j=1}^{n_i} X_{ij}/n_i)$, $i = 1, 2, \ldots, M$. This estimator can provide a rough approximate estimation. Estimators 3 is our previously recommended MLE estimator via Nelder-Mead optimization (Xu, 2023). Estimator 4 is our proposed EM estimator as described above. The performances of these estimators were compared under three scenarios. In each scenario, simulations were conducted with the number of groups ($M = 300, 500, 1000$), and equal group sizes ($n_i = 5, 10$, $i = 1, 2, \ldots, M$). The setup of data generation is specified as follows:

Algorithm 1 EM Algorithm for Logistic Regression Based on Individual-Level X and Aggregate-Level Y

1. Start with initial value for the parameter β, i.e. $\hat{\beta}^{(0)}$, where the initial value is obtained from the following values: (1) estimated value by the usual logistic regression of aggregate-level Y on aggregate-level X, (2) MLE estimate in Xu (2023), (3) the zero vector $(0,0,\ldots,0) \in \mathscr{R}^p$, (4) the unit vector $(1,1,\ldots,1) \in \mathscr{R}^p$, and (5) the vector $(-1,-1,\ldots,-1) \in \mathscr{R}^p$.

2. Expectation Step: at the j-th step, compute

$$\begin{aligned} Q(\beta', \hat{\beta}^{(j)}) &= E(l(\beta'; \{Y_{ij}\}) | \{Y_i\}, \hat{\beta}^{(j)}) \qquad (4) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{n_i} \{E(Y_{ij}|Y_i, \hat{\beta}^{(j)}) \ln(\pi_{ij}) + (1 - E(Y_{ij}|Y_i, \hat{\beta}^{(j)})) \ln(1-\pi_{ij})\} \end{aligned}$$

   as a function of the dummy argument β'. The expected value of latent value Y_{ij} is computed via the formula

$$\begin{aligned} & E(Y_{ij}|Y_i = y, \hat{\beta}^{(j)}) \qquad (5) \\ = \; & P(Y_{ij} = 1 | Y_i = y, \hat{\beta}^{(j)}) = \frac{P(Y_{ij}=1) P(Y_i - Y_{ij} = y-1)}{P(Y_i = y)} \\ = \; & \frac{\pi_{ij} \times \text{PoissonBinomial}(y-1; n_i - 1, \pi_{i1}, \ldots, \pi_{i,j-1}, \pi_{i,j+1}, \ldots, \pi_{in_i})}{\text{PoissonBinomial}(y-1; n_i, \pi_{i1}, \pi_{i2}, \ldots, \pi_{in_i})}, \end{aligned}$$

   where PoissonBinomial(.) is the probability mass function of a Poisson binomial distribution, and $\pi_{ij} = \exp(X_{ij}^T \beta')/(1+\exp(X_{ij}^T \beta'))$. As the convention in regression analysis, we can treat X as fixed. For random X, we can use the argument of conditioning Y on X and this conditioning is equivalent to treating X as fixed (Hastie et al., 2009; Givens and Hoeting, 2012).

3. Maximization Step: determine the new estimate $\hat{\beta}^{(j+1)}$ as the maximizer of $Q(\beta', \hat{\beta}^{(j)})$ over β'. This step is obtained by conducting weighted logistic regression with the likelihood function specified in Equation 4. To be more specific, our dataset has N observations of individual-level X, i.e. X_{ij}, $i = 1,2,\ldots,M$, $j = 1,2,\ldots,n_i$, $N = \sum_{i=1}^{M} n_i$. A pseudo-dataset of $2N$ pseudo-observations is created with the pseudo-observation represented as $(\tilde{X}_{ijk}, \tilde{Y}_{ijk}, \tilde{W}_{ijk})$, $i = 1,2,\ldots,M$, $j = 1,2,\ldots,n_i$, $k = 0,1$, where $\tilde{X}$, $\tilde{Y}$ and $\tilde{W}$ are respectively the predictor vector, response and weight in the pseudo-dataset. For each observation X_{ij}, two pseudo-observations, i.e. $(\tilde{X}_{ij0}, \tilde{Y}_{ij0}, \tilde{W}_{ij0})$ and $(\tilde{X}_{ij1}, \tilde{Y}_{ij1}, \tilde{W}_{ij1})$, are created as follows:

$$\begin{aligned} &\tilde{X}_{ij0} = X_{ij},\ \tilde{Y}_{ij0} = 0,\ \tilde{W}_{ij0} = 1 - E(Y_{ij}|Y_i, \hat{\beta}^{(j)}) \\ &\tilde{X}_{ij1} = X_{ij},\ \tilde{Y}_{ij1} = 1,\ \tilde{W}_{ij1} = E(Y_{ij}|Y_i, \hat{\beta}^{(j)}). \end{aligned}$$

   Weighted logistic regression is conducted based on the pseudo-dataset.

4. Iterate steps 2 and 3 until convergence.

- In Scenario 1, $p = 5$, $(X_{i1}, X_{i2}) \sim \text{multinormal}_2(0_{2\times1}, \Sigma_{2\times2})$, $0_{2\times1} = (0,0)^T$, $\Sigma_{2\times2} = (\sigma_{ij})$ where $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = 0.6$ for $i \neq j$. $X_{i3} \sim \text{t(df} = 2)$, $X_{i4} \sim \text{Bernoulli}(0.5)$, $X_i = (1, X_{i1}, X_{i2}, X_{i3}, X_{i4})^T \in \mathscr{R}^p$, $Y_i \sim \text{Bernoulli}(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-0.5, 1, -0.5, 2, -1.6)^T$.
- In Scenario 2, $p = 10$, $(X_{i1}, X_{i2}, X_{i3}, X_{i4}) \sim \text{multinormal}_4(0_{4\times1}, \Sigma_{4\times4})$, $0_{4\times1} = (0,0,0,0)^T$, $\Sigma_{4\times4} = (\sigma_{ij})$ where $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = 0.6$ for $i \neq j$. $X_{i5} \sim \text{t(df} = 2)$, $X_{i6} \sim \text{t(df} = 4)$, $X_{i7} \sim \text{chi-square(df} = 2)$, $X_{i8} \sim \text{chi-square(df} = 3)$, $X_{i9} \sim \text{Bernoulli}(0.5)$, $X_i = (1, X_{i1}, X_{i2}, \ldots, X_{i9})^T \in \mathscr{R}^p$, $Y_i \sim \text{Bernoulli}(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-0.5, 1, -2.5, 2, -1.6, 0.7, 0.9, -2.4, 0.5, -1.3)^T$.
- In Scenario 3, $p = 20$, $(X_{i1}, X_{i2}, \ldots, X_{i10}) \sim \text{multinormal}_{10}(0_{10\times1}, \Sigma_{10\times10})$, $0_{10\times1} = (0,0,\ldots,0)^T$, $\Sigma_{10\times10} = (\sigma_{ij})$ where $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = 0.6$ for $i \neq j$. $X_{i11} \sim \text{t(df} = 2)$, $X_{i12} \sim \text{t(df} = 4)$, $X_{i13} \sim \text{t(df} = 6)$, $X_{i14} \sim \text{chi-square(df} = 2)$, $X_{i15} \sim \text{chi-square(df} = 3)$, $X_{i16} \sim \text{chi-square(df} = 4)$, $X_{i17} \sim \text{Bernoulli}(0.3)$, $X_{i18} \sim \text{Bernoulli}(0.5)$, $X_{i19} \sim \text{Bernoulli}(0.7)$, $X_i = (1, X_{i1}, X_{i2}, \ldots, X_{i19})^T \in \mathscr{R}^p$, $Y_i \sim \text{Bernoulli}(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-0.5, -0.8969, 0.1848, 1.5878, -1.1304, -0.0803, 0.1324, 0.7080, -0.2397, 1.9845, -0.1388, 0.4177, 0.9818, -0.3927, -1.0397, 1.7822, -2.311, 0.8786, 0.036, 1.013)^T$. Note that the values of the 19 slope coefficients, i.e. $\beta_1, \ldots, \beta_{19}$, were generated as standard normal random variables, and the generation of the values of the 19 slope coefficients was implemented in R language using the command: "set.seed(2); rnorm(19)". The value of the intercept parameter, i.e. β_0, was fixed at -0.5.

3.2. Performance Evaluation Metrics

The squared bias, variance, mean square error (MSE), and mean absolute deviation (MAD) of each of the four estimators' (E1 to E4) model parameters $(\beta_0, ..., \beta_{p-1}) \in \mathscr{R}^p$ were calculated. Denote the bias, variance, MSE, and MAD of the q-th estimator of β_j as $\text{Bias}(\hat{\beta}_{j,E_q})$,$\text{Var}(\hat{\beta}_{j,E_q})$, $\text{MSE}(\hat{\beta}_{j,E_q})$, and $\text{MAD}(\hat{\beta}_{j,E_q})$. The average squared bias, variance, MSE, and MAD of the q-th estimator were calculated as $\overline{\text{Bias}^2}(E_q) = (1/p)\sum_{j=0}^{p-1}\text{Bias}^2(\hat{\beta}_{j,E_q})$, $\overline{\text{Var}}(E_q) = (1/p)\sum_{j=0}^{p-1}\text{Var}(\hat{\beta}_{j,E_q})$, $\overline{\text{MSE}}(E_q) = (1/p)\sum_{j=0}^{p-1}\text{MSE}(\hat{\beta}_{j,E_q})$, and $\overline{\text{MAD}}(E_q) = (1/p)\sum_{j=0}^{p-1}\text{MAD}(\hat{\beta}_{j,E_q})$.

3.3. Simulation Results

In Table 1, we report the average squared biases and variances for the four estimators (Individual-LR, Naive-LR, MLE and EM) under different scenarios, sample sizes, and group sizes. Regarding bias, the infeasible individual-LR shows smallest bias and the naive-LR shows biggest bias. The reason for individual-LR to have smallest bias is that individual-LR conducts the usual logistic regression based on individual-level X and individual-level Y which makes use of more information than available in our data situation where individual-level Y is not available. Naive-LR is found to have the biggest bias, which was explained by the fact that logistic regression model uses a "non-linear" logit link function and Naive-LR conducts a naive rough approximate using the mean of X, which ignores the nonlinearity in

the link function, so that Naive-LR can induce a big bias. The biases of MLE estimator and EM estimator are found to be between the two extremes, i.e. individual-LR and naive-LR.

Regarding variance, individual-LR and naive-LR have relatively smaller variance, compared with MLE estimator and EM estimator. We explained that the smaller variance in individual-LR is because it uses more information than available in our data situation where individual-level Y is not available. The smaller variance in naive-LR is also reasonable. As a poor rough approximate estimator, naive-LR can have big bias and small variance. For example, suppose that a toy estimator always reports a constant value as its estimate. This toy estimator will have zero variance and a big bias. Thus, we put more focus on mean square error (MSE) and mean absolute deviation (MAD) instead of bias and variance in evaluating estimators.

Next, we check MSE and MAD of the four estimators. In Table 2, we report average MSE and average MAD. The infeasible individual-LR estimator shows the best performance in terms of both MSE and MAD. This is because individual-LR estimators makes use of more information than available in our data situation where individual Y is latent. The naive-LR estimator shows the worst performance in terms of both MSE and MAD. This is because naive-LR is a naive rough approximate estimator which can lead to a big bias due to non-linearity in link function. In terms of MSE and MAD, we found our MLE estimator and EM estimator are between the two extremes (individual-LR and naive-LR). Our MLE and EM estimator show similar performance with EM estimator having potentially slightly better performance.

We add a cautionary note that simulation studies cannot substitute theoretical verification. Simulation studies cannot fully assess theoretical properties of estimators. Theoretical properties of MLE estimators and EM estimators have to be inferred based on theoretical literature on MLE and EM.

4. Real Data-Based Studies

We used real data to illustrate the use of our EM estimator and compare it with different estimators in the literature. Two real data-based studies are shown. One study is wine quality modeling based on physico-chemical tests. The other study is maternal health risk modeling. Both studies used the datasets from UC Irvine machine learning repository (`https://archive.ics.uci.edu/`).

4.1. Wine Quality Modeling

We obtained two datasets of wine quality from UC Irvine machine learning repository (Cortez and Reis, 2009). The two datasets are related to red and white verde wine samples, from the north of Portugal. Due to privacy and logistic issues, only physicochemical (inputs, i.e. X) and sensory (the output, i.e. Y) variables are available. The output variable sensory wine quality score is a score between 1 and 10. This wine quality score was dichotomized into a binary variable, which takes the value of 1 (high-quality) or 0 (low-quality) depending on whether the score is between 6 and 10, or between 1 and 5. Thus, as specified in UC Irvine machine learning repository, the wine quality datasets can be used for both clas-

Table 1: Average Squared Bias and Variance of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM). In the columns for average squared bias and average variance, the unit is 0.001.

			Average Squared Bias				**Average Variance**			
Scenario	**M**	n_i	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1	300	5	0.08	718	0.68	0.37	15.8	21.1	35.3	34.0
1	300	10	0.11	788	3.16	0.74	7.4	16.5	48.3	34.5
1	500	5	0.20	717	0.36	0.90	9.0	11.3	30.7	27.6
1	500	10	0.03	788	0.96	1.12	4.8	13.3	36.4	25.7
1	1000	5	0.03	729	0.29	0.31	4.8	8.3	19.3	12.8
1	1000	10	0.01	799	3.57	0.05	2.3	5.9	23.0	11.2
2	300	5	0.72	1007	6.48	4.39	34.3	24.3	91.2	57.5
2	300	10	0.43	1064	25.23	5.08	13.6	23.0	134.9	46.2
2	500	5	0.43	1021	19.14	0.25	14.6	14.3	63.2	26.4
2	500	10	0.18	1063	49.26	1.54	7.8	11.6	96.0	27.0
2	1000	5	0.18	1018	17.67	0.57	7.8	7.7	57.2	15.1
2	1000	10	0.06	1078	49.59	0.37	4.0	6.8	81.5	11.7
3	300	5	6.25	658	178.2	14.39	48.0	27.6	86.3	78.6
3	300	10	2.15	683	282.9	10.05	22.3	25.3	63.8	65.3
3	500	5	1.08	667	200.2	3.23	28.7	15.2	70.0	43.6
3	500	10	0.40	693	300.0	2.46	13.1	14.4	47.5	33.5
3	1000	5	0.40	668	192.5	0.96	13.1	7.5	59.3	20.3
3	1000	10	0.13	689	306.2	1.03	6.4	6.5	35.1	16.0

sification problem (Y is the binary wine quality variable) and regression problem (Y is the wine quality score which is between 1 and 10). There are 11 continuous features/predictors in X. They are (1) fixed acidity, (2) volatile acidity, (3) citric acid, (4) residual sugar, (5) chlorides, (6) free sulfur dioxide, (7) total sulfur dioxide, (8) density, (9) pH, (10) sulphates and (11) alcohol. For more details in the wine quality datasets, please refer to Cortez and Reis (2009).

We used the wine quality datasets to illustrate the use of logistic regression under the data situation of aggregate-level Y and individual-level X. In practice, there are multiple reasons which can contribute to the situation that Y is available at aggregate level instead of individual level. One reason is confidentiality. For example, suppose the objective is to predict or model wine quality provided by some specific wine association or agency. However, the wine association or agency wants to keep its evaluation in confidentiality and do not want its evaluation to be easily modeled or predicted. In addition, the wine

Table 2: Average Mean Squared Error (MSE) and Mean Absolute Deviation (MAD) of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM). In the columns for average MSE and average MAD, the unit is 0.001.

			Average MSE				**Average MAD**			
Scenario	**M**	n_i	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1	300	5	15.9	739	36.0	34.4	95.5	680.7	146.5	140.1
1	300	10	7.5	804	51.5	35.2	63.7	714.5	169.4	146.4
1	500	5	9.2	729	31.0	28.5	74.6	676.7	137.1	130.7
1	500	10	4.9	801	37.4	26.8	53.6	708.0	140.4	125.3
1	1000	5	4.9	738	19.6	13.1	53.6	685.0	100.2	89.0
1	1000	10	2.3	805	26.6	11.3	37.4	715.0	113.7	82.4
2	300	5	35.0	1032	97.7	61.9	136.9	865.6	228.7	188.2
2	300	10	14.0	1087	160.2	51.3	87.5	885.9	290.2	166.8
2	500	5	15.1	1035	82.3	26.7	90.5	869.1	205.7	122.5
2	500	10	8.0	1075	145.2	28.5	65.2	886.7	264.9	126.0
2	1000	5	8.0	1026	74.9	15.6	65.2	868.5	186.5	91.5
2	1000	10	4.1	1084	131.1	12.0	46.2	897.6	248.6	84.0
3	300	5	54.2	685	264.5	93.0	176.7	645.6	401.8	233.7
3	300	10	24.5	709	346.7	75.3	119.1	658.1	465.2	207.6
3	500	5	29.8	682	270.2	46.8	127.9	643.7	413.7	163.8
3	500	10	13.5	708	347.6	36.0	89.6	658.2	472.1	143.3
3	1000	5	13.5	675	251.8	21.3	89.6	639.1	401.4	112.1
3	1000	10	6.6	695	341.3	17.0	61.2	647.9	466.3	98.3

association is interested in ranking wineries or wine firms based on multiple wine samples submitted by each winery or firm. The rule is that each winery or firm is allowed to submit samples from multiple brands the winery or firm owns. The wine association will only specify how many samples are in high-quality in their submission and does not disclose wine quality of each individual wine sample. In this way, the firms will compete with aggregate-level Y available instead of individual-level Y, and the wine association or agency keep its evaluation result of individual samples to be confidential.

We illustrated the use of our EM estimator and other estimators (infeasible individual-LR, naive aggregate-LR, and MLE estimator in Xu (2023)) in the literature for wine quality modeling. There are 4898 observations in white wine data, and 1599 observations in red wine data. We conducted random aggregation with equal group size $n_i = 5$ and 10. For white wine data, there are $979 = 4895/5$ groups of size $n_i = 5$ formed, and $489 = 4890/10$ groups of size $n_i = 10$ formed. Thus, there are 4895 and 4890 observations used in our data

situation $n_i = 5$ and $n_i = 10$ for white wine data. For red wine data, there are $319 = 1595/5$ groups of size $n_i = 5$ formed, and $159 = 1590/10$ groups of size $n_i = 10$ formed. Thus, there are 1595 and 1590 observations used in our data situation $n_i = 5$ and $n_i = 10$ for red wine data.

We showed the estimated values of the estimators based on our data sets with random aggregation. The estimators are: (1) individual-LR, which conducts logistic regression based on individual-level X and individual-level Y. Individual-LR is considered to be the best estimator since it uses more information (individual-level Y) than the information available in our data situation where aggregate-level Y instead of individual-level Y is available. Thus, individual-LR is infeasible. (2) naive-LR, which conducts logistic regression based on aggregate-level X and aggregate-level Y. (3) our previously proposed MLE in Xu (2023). (4) our EM estimator proposed in this article. We illustrated the use of each estimator based on wine quality data and report the estimated values of parameters for white wine data in Table 3 and the estimated values of parameters for red wine data in Table 4. As shown in Table 3 and 4, these estimators reported numerically different values. We recommend the use of EM estimator and MLE estimator, since individual-LR is infeasible and naive-LR can induce a big bias. Because there is no ground truth (true values) of logistic model parameters known in the real data, no statistical performances (such as bias and variance) were evaluated based on the real data.

Table 3: Estimated Values of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On White Wine Quality Data.

	$n_i = 5$				$n_i = 10$			
Variable	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
β_0	0.920	0.719	0.910	0.923	0.920	0.703	0.913	0.917
β_1	0.032	0.048	0.075	0.018	0.033	0.026	0.065	0.042
β_2	-0.651	-0.456	-0.627	-0.636	-0.650	-0.459	-0.645	-0.658
β_3	0.015	0.066	0.075	0.090	0.015	-0.028	0.010	-0.001
β_4	0.865	0.395	0.599	0.451	0.866	0.984	1.435	1.359
β_5	0.020	-0.066	-0.058	-0.075	0.019	-0.131	-0.078	-0.082
β_6	0.163	0.170	0.170	0.223	0.164	0.214	0.264	0.285
β_7	-0.056	-0.066	-0.019	-0.045	-0.057	-0.141	-0.101	-0.112
β_8	-0.812	-0.267	-0.473	-0.212	-0.814	-0.789	-1.202	-1.086
β_9	0.166	0.131	0.172	0.139	0.167	0.217	0.347	0.323
β_{10}	0.205	0.159	0.209	0.191	0.206	0.182	0.397	0.387
β_{11}	0.915	0.840	1.056	1.210	0.911	0.536	0.681	0.749

Table 4: Estimated Values of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Red Wine Quality Data.

	$n_i = 5$				$n_i = 10$			
Variable	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
β_0	0.239	0.143	0.239	0.226	0.237	0.141	0.291	0.295
β_1	0.247	0.367	0.382	0.693	0.251	0.853	0.880	1.000
β_2	-0.585	-0.337	-0.500	-0.571	-0.588	-0.454	-0.702	-0.771
β_3	-0.248	-0.379	-0.371	-0.532	-0.252	-0.282	-0.356	-0.413
β_4	0.079	-0.043	-0.003	0.108	0.080	0.028	-0.137	-0.296
β_5	-0.183	0.083	-0.023	0.052	-0.180	-0.114	-0.144	-0.161
β_6	0.233	0.337	0.260	0.309	0.230	0.273	0.352	0.083
β_7	-0.539	-0.574	-0.684	-0.721	-0.535	-0.288	-0.317	-0.110
β_8	-0.104	0.051	-0.020	-0.197	-0.104	-0.491	-0.478	-0.544
β_9	-0.052	-0.101	-0.217	-0.117	-0.053	0.159	0.079	0.190
β_{10}	0.475	0.224	0.296	0.326	0.469	0.550	0.643	0.642
β_{11}	0.917	0.688	0.993	0.973	0.917	0.451	0.805	0.876

4.2. Maternal Health Risk Modeling

We obtained the dataset of maternal health risk from UC Irvine machine learning repository (Ahmed, 2023; Ahmed et al., 2020). The data were collected through the IoT-based risk monitoring system from a range of hospitals, community clinics, maternal health care in the rural areas of Bangladesh (Ahmed, 2023). The response variable is the binary maternal health risk level (low risk or high risk). The predictors are (1) age, (2) systolic blood pressure, (3) diastolic blood pressure, (4) blood glucose, (5) body temperature, and (6) heart rate. All these predictors are the responsible and significant risk factors for maternal mortality (Ahmed et al., 2020). UC Irvine machine learning repository specify it as a classification problem since the response variable is binary. There are 1013 individual observations in the dataset. For more details in the maternal health risk data, please refer to Ahmed et al. (2020) and Ahmed (2023).

We conducted random aggregation on the data. There are 202=1010/5 groups of size $n_i = 5$ and 101=1010/10 groups of size $n_i = 10$ formed. Thus, there are 1010 observations in our study of maternal health risk modeling.

Based on the maternal health risk data with random aggregation, we conducted individual-LR, naive-LR, MLE in Xu (2023) and EM estimator proposed in the article. The estimated values of these estimators are shown in Table 5. There is numerical difference in the estimated values of different estimators. We recommend the use of our proposed EM estimator and our previously proposed MLE estimator in the study.

Table 5: Estimated Values of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Maternal Health Risk Data.

	$n_i = 5$				$n_i = 10$			
Variable	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
β_0	0.913	0.544	0.785	0.784	0.913	0.546	0.592	0.669
β_1	-0.079	-0.062	-0.075	-0.079	-0.079	-0.075	-0.019	-0.113
β_2	1.116	2.014	1.060	1.059	1.116	3.413	1.138	1.102
β_3	-0.365	-0.551	-0.346	-0.345	-0.365	-1.205	-0.357	-0.433
β_4	1.631	0.717	1.333	1.329	1.631	0.460	0.640	1.032
β_5	0.668	0.998	0.652	0.650	0.668	1.388	0.746	0.594
β_6	0.272	0.177	0.213	0.214	0.272	-0.104	0.433	0.228

5. Discussion

There are at least two categories of methods to solve the problem of logistic regression based on individual-level X and aggregate-level Y. The first category is directly maximizing the complicated likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$ to find MLE as we previously proposed in Xu (2023). The second category is to avoid the direct optimization of the likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$ using the EM algorithm as we propose in this article. In theory, both categories of methods are valid. Similar but slightly different performances are expected theoretically. We note that the two categories of methods are generic so that there are a range of ways in each category. Along the first category, i.e. obtaining MLE by directly maximizing $L(\beta)$, $\beta \in \mathscr{R}^p$, there can be a range of optimization methods with slightly better or worse performance. A non-exhaustive list of these methods includes: (1) Nelder and Mead's simplex method (NM) (Nelder and Mead, 1965), (2) the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (Fletcher, 1970), (3) the conjugate gradient (CG) method (Fletcher and Reeves, 1964), (4) simulated annealing (Brooks and Morgan, 2018), and (5) evolutional algorithm (Lambora et al., 2019), and their combinations or variants such as Generalized simulated annealing (GSA) and variable step size generalized simulated annealing (VGSA) (Kalivas, 1992). Along the second category avoiding directly maximization of likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$, there can be a range of methods including (1) the standard EM (McLachlan and Krishnan, 2007), (2) Monte Carlo EM (Wei and Tanner, 1990), and (3) variational Bayes EM (Bernardo et al., 2003).

Both categories of methods have their own advantages and are necessary for the logistic regression based on individual-level X and aggregate-level Y. Which category of methods to use in practice depends on the specific problem. The advantages of the second category of methods, including EM algorithms, Bayes methods and their variants, are the convenience in solving a range of data situations including missing values. Along this category of methods, methods can be potentially adapted to solve data situations such as (1) the sit-

uation where X and Y are at mixed-levels, (2) the situation where X and Y contain missing values, (3) the situation where prior information is preferred to use or consider, (4) the situation where individual-level X and individual-level Y is described by a generalized linear model, which can be a linear model, logistic model, Poisson model or other generalized linear models. In comparison, the advantages of the first category of methods include (1) there are a range of optimization methods to try, (2) the potential further extension of methods to penalized likelihood functions which will add a penalty term such as L_p norm of model parameter β with $1 \leq p \leq 2$ to the current complex likelihood function (Hastie et al., 2009), and (3) likelihood-based statistical inferences such as likelihood ratio test, score test, standard errors, and confidence intervals. Studies of these extensions are beyond the scope of this article and are under development as future work.

The dimension p, i.e. the number of predictors, influences the performance of EM estimator and MLE estimator. Given the sample size n, both EM performance and MLE performance are expected to decrease when p increases. The deterioration of both performances with the increase in p is as expected since the optimization problem $\max L(\beta)$, $\beta \in \mathscr{R}^p$ in theory will decrease when p increases, given a fixed sample size n. Both EM and MLE will maximize $L(\beta)$, either indirectly or directly.

However, we need to note although EM algorithms always have likelihood non-decreasing in each step, EM may converge to a local maximum of the observed likelihood function for some starting values instead of a global maximum so that EM estimators may not converge to MLE (Givens and Hoeting, 2012). Our EM estimator is a standard EM estimator, suffering from the (common) limitations of EM estimators while enjoying the (common) benefits and advantages of EM estimators.

In logistic regression, both continuous predictors and categorical predictors can be included. Our simulation studies used both types of predictors. However, for categorical predictors, we only used binary predictors. A categorical predictor with K levels can lead to or amount to $K-1$ binary predictors, which will increase the number of predictors, i.e. p. As the number of levels K increases, the number of predictors, i.e. p, increases, which will make estimation performance become worse. Thus, a categorical predictor with multiple levels may decrease estimation performance of our estimators. Future studies can be on the influence of categorical predictors with more than two levels.

There are some assumptions in our model setup. Firstly, we only consider independent individual-level data, i.e. (X_i, Y_i), $i = 1, 2, \cdots, n$, in this article. In practice, individual-level observations can be correlated or dependent. Secondly, we only consider the situation of "grouping completely at random", which means that the grouping mechanism is completely random, and is not influenced by the values of X and Y. In practice, grouping may not be completely random such as the situation where individuals with similar values in X or Y are more likely to be grouped together. Further studies can be conducted for grouping not completely at random. Thirdly, only summation aggregation $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ is studied. Other aggregations, such as $Y_i = 1(\sum_{j=1}^{n_i} Y_{ij} > 0)$ used in group testing of infectious disease, are not studied in this article, since the group-testing problem with $Y_i = 1(\sum_{j=1}^{n_i} Y_{ij} > 0)$ for logistic regression has been well studied in bio-statistics and epidemiology.

Although the method is proposed for a logistic regression (logistic link function) to deal with binary response variable Y, other link functions can also be used to handle the binary

response. For example, the tobit regression which uses a probit link function can also be used to analyze individual-level X and aggregate-level Y. In addition, the current article is based on the binary response variable Y. A follow-up study to extend our method to handle responses with more than two levels are under development.

6. Conclusions

We proposed an EM estimator for logistic regression based on individual-level predictors (X) and aggregate-level response (Y). We conducted simulation studies to evaluate the performance of the EM estimator and compare it with estimators in the literature (individual-LR, naive-LR and MLE). We then conducted two real data-based studies, i.e. wine quality modeling and maternal health risk modeling, to illustrate the use of different estimators. Both the simulation studies and real data-based studies have shown the use of our EM estimator in conducting logistic regression based on individual-level X and aggregate-level Y. We think both categories of methods (MLE category of methods or EM category of methods) work and are necessary for the problem of logistic regression based on individual-level X and aggregate-level Y. Similar and slightly different performances are expected for estimators along the two categories of methods.

Declarations

Availability of code and data: R functions implementing EM algorithm as described in the manuscript are available in Github repository via the link
`https://github.com/zhengxu0459/EM-Algorithm-Logistics-Regression`.
All data used in the study are publicly available.
Funding: This study receives no funding.

References

Agresti, A., (2013). Categorical Data Analysis, Wiley Series in Probability and Statistics. *Wiley*, Hoboken, NJ, USA.

Ahmed, M., (2023). Maternal Health Risk. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DP5D.

Ahmed, M., Kashem, M. A., Rahman, M. and Khatun, S., (2020). Review and analysis of risk factor of maternal health in remote area using the internet of things (iot). URL https://api.semanticscholar.org/CorpusID:214577407.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M., et al., (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7, (pp. 453–464):210.

Brooks, S. P., Morgan, B. J. T., (2018). Optimization Using Simulated Annealing. *Journal of the Royal Statistical Society Series D: The Statistician*, 44(2), pp. 241–257, 12. ISSN: 2515–7884.

Chen, X., Dempster, A. and Liu, J., (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, pp. 457–469.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J., (2009). Wine Quality. *UCI Machine Learning Repository*. DOI: https://doi.org/10.24432/C56S3T.

Fernandez, M., Williams, S., (2010). Closed-formexpression for the Poisson-binomial probability density function. *IEEE Trans. Aerosp. Electron. Syst.*, 46, pp. 803–817.

Fletcher, R., (1970). A new approach to variable metric algorithms. *Comput. J.*, 13, pp. 317–322.

Fletcher, R., Reeves, C., (1964). Function minimization by conjugate gradient. *Comput. J.*, 7, pp. 149–154.

Geamsakul W., Yoshida T., Ohara K., Motoda H., Yokoi H., and Takabayashi K., (2005). Constructing a decision tree for graph-structured data and its applications. *Fundamenta Informaticae*, 66(1–2), pp. 131–160.

Getoor, L., Mihalkova, L., (2011). Learning statistical models from relational data. *In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 1195–1198.

Givens, G., Hoeting, J., (2012). Computational Statistics, Wiley Series in Probability and Statistics. *Wiley*, Hoboken, NJ, USA.

Hastie, T., Tibshirani, R. and Friedman, J., (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics. *Springer*, Berlin, Germany. ISBN 9780387848846.

Henaff, M., Bruna, J. and LeCun, Y., (2015). Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163.

Hilbe, J., (2009). Logistic Regression Models. Chapman & Hall/CRC Texts in Statistical Science. *CRC Press*, Boca Ration, Florida, USA. ISBN 9781420075779.

Hong, Y., (2013). On computing the distribution function for the Poisson binomial distribution. *Comput. Stat. Data Anal.*, 59, pp. 41–51.

Kalivas, J. H., (1992). Optimization using variations of simulated annealing. *Chemometrics and Intelligent Laboratory Systems*, 15(1), pp. 1–12. ISSN 0169-7439.

Lambora, A., Gupta, K. and Chopra, K., (2019). Genetic algorithm-a literature review. In 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon), pp. 380–384. *IEEE*.

McLachlan, G. J. and Krishnan, T., (2007). The EM algorithm and extensions. John Wiley & Sons, New York City, USA.

Mercer, T. R., Salit, M.. (2021). Testing at scale during the covid-19 pandemic. *Nature Reviews Genetics*, 22(7), pp. 415–426.

Nelder, J., Mead, R., (1965). A simplex method for function minimization. *Comput. J.*, 7, pp. 308–313.

Primo, D. M., Jacobsmeier, M. L. and Milyo, J., (2007). Estimating the impact of state policies and institutions with mixed-level data. *State Politics & Policy Quarterly*, 7(4), pp. 446–459.

Saramago, P., Sutton, A. J., Cooper, N. J. and Manca, A., (2012). Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in medicine*, 31(28), pp. 3516–3536.

Wang, Y., (1993). On the number of successes in independent trials. *Stat. Sin.*, 3, pp. 295–312.

Wei, G. C., Tanner, M. A., (1990). A Monte Carlo implementation of the em algorithmand the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411), pp. 699–704,

Xu, Z., (2023). Logistic regression based on individual-level predictors and aggregate-level responses. *Mathematics*, 11(3), p.746.

Zhai, Y., Liu, B., (2006). Structured data extraction from the web based on partial tree alignment. *IEEE Transactions on Knowledge and Data Engineering*, 18(12), pp. 1614–1628.

Appendix

A. Additional Real Data Study Using Data in Xu (2023)

We conducted additional real data study based on the same data as used in Xu (2023). The dataset is "Social-Network-Ads" data in Kaggle Machine Learning Forum (`https://www.kaggle.com`). The dataset is a categorical dataset to determine whether a user purchases a particular product. It contains 400 observations. Two predictors are age and salary, after data standardization. The same as in Xu (2023), we impose data aggregation on this dataset with the group size equal to 3, 5 and 7. We conducted (1) infeasible individual-level logistic regression, (2) naive logistic regression, (3) MLE estimator in Xu (2023), and (4) our proposed EM estimator in this manuscript. Because true parameter values are unknown, we illustrate the use of different estimators and report estimated values using different estimators in Table 6.

Table 6: Estimated Values of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Social Network Ads Data.

	$n_i = 3$				$n_i = 5$				$n_i = 7$			
Var	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
β_0	-1.14	-0.71	-1.17	-1.17	-1.14	-0.68	-1.25	-1.24	-1.13	-0.64	-1.27	-1.27
β_1	2.45	1.67	2.53	2.53	2.45	1.61	2.79	2.79	2.44	1.59	2.97	2.97
β_2	1.22	0.79	1.47	1.47	1.22	0.64	1.54	1.54	1.22	0.53	1.26	1.26

B. Additional Simulation Study Using Xu (2023)'s Setup

We conducted additional simulation study using Xu (203)'s simulation setup as follows. In each scenario, simulations were conducted with sample sizes ($K = 300, 500, 100$), equal group sizes ($n_g = 7, 30$), and different parameter values. Data were generated as follows:

- In Scenario 1, $X_{i1} \sim N(0,1)$, $X_i = (1, X_{i1})^T$, $Y_i \sim$ Bernoulli$(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (1,-2)^T$ (Scenario 1A) or $(1,3)$ (Scenario 1B).
- In Scenario 2, $X_{i1} \sim N(0,1)$, $X_{i2} \sim t(df = 5)$, $X_i = (1, X_{i1}, X_{i2})^T$, $Y_i \sim$ Bernoulli$(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-1,1,2)^T$ (Scenario 2A) or $(0,-2,1)$ (Scenario 2B).
- In Scenario 3, $(X_{i1}, X_{i2}) \sim$ BivariateNormal$(0,2,1,4,\rho = 0.5)$, $X_{i3} \sim$ Cauchy$(0,1)$, $X_i = (1, X_{i1}, X_{i2}, X_{i3})^T$, $Y_i \sim$ Bernoulli$(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-1,1,0,-1)^T$ (Scenario 3A) or $(0,-2,1,1)$ (Scenario 3B).

We reported squared bias and variance of four estimators (E1: Individual-LR, E2: Naive-LR, E3: MLE and E4: EM) in Table 7. We reported MSE and MAD of the four estimators in Table 8. Results obtained from additional simulation studies confirm our findings based on simulation studies. The same findings were obtained.

Table 7: Average Squared Bias and Variance of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Simulation Setup in Xu (2023). In the columns for average squared bias and average variance, the unit is 0.001.

			Average Squared Bias				**Average Variance**			
Scenario	**M**	n_i	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1A	300	7	0.28	310.34	0.76	0.86	7.69	6.84	23.74	22.58
1A	300	30	0.09	343.26	0.53	0.82	1.43	6.81	30.52	24.37
1A	500	7	0.02	317.16	0.00	0.00	3.56	3.49	11.45	10.60
1A	500	30	0.01	356.36	0.01	0.07	0.91	2.70	11.68	11.33
1A	1000	7	0.08	302.01	0.20	0.26	2.14	1.83	6.90	6.99
1A	1000	30	0.01	352.87	0.24	0.22	0.42	1.76	6.29	6.21
1B	300	7	0.01	1256.06	0.36	0.54	10.90	6.91	37.38	35.02
1B	300	30	0.17	1376.61	1.90	3.24	2.96	5.12	38.64	30.50
1B	500	7	0.16	1264.86	0.35	0.51	6.86	3.30	16.90	16.43
1B	500	30	0.00	1401.36	0.00	0.02	1.58	2.32	23.51	19.81
1B	1000	7	0.02	1267.42	0.05	0.07	3.00	1.74	10.07	10.04
1B	1000	30	0.00	1400.68	0.37	0.07	0.69	1.23	13.19	6.68
2A	300	7	0.00	485.55	0.12	0.14	6.09	6.33	17.56	17.33
2A	300	30	0.02	547.58	0.32	0.23	1.37	4.72	27.61	26.45
2A	500	7	0.09	487.78	0.05	0.08	4.19	3.95	12.23	12.35
2A	500	30	0.01	540.78	0.07	0.13	0.89	3.48	13.58	11.83
2A	1000	7	0.04	484.78	0.04	0.04	1.86	1.70	6.49	6.33
2A	1000	30	0.00	540.90	0.02	0.00	0.47	1.65	7.32	6.88
2B	300	7	0.04	304.21	0.35	0.40	5.37	5.82	17.33	17.08
2B	300	30	0.03	339.23	0.35	0.39	1.25	4.09	18.77	18.82
2B	500	7	0.04	304.27	0.11	0.18	3.21	3.18	10.89	10.67
2B	500	30	0.00	334.51	0.16	0.23	0.80	2.48	12.16	11.68
2B	1000	7	0.06	304.24	0.06	0.09	1.62	1.44	5.31	4.99
2B	1000	30	0.00	333.05	0.12	0.16	0.37	1.38	6.51	6.09
3A	300	7	0.06	336.15	0.55	0.58	4.46	6.19	13.40	12.67
3A	300	30	0.02	336.61	0.60	0.91	0.83	6.34	14.10	13.88
3A	500	7	0.03	342.89	0.12	0.08	2.04	3.22	7.70	7.67
3A	500	30	0.00	345.45	0.74	0.55	0.64	3.39	8.91	7.21
3A	1000	7	0.02	343.02	0.27	0.21	1.24	2.40	4.98	4.38
3A	1000	30	0.00	350.84	0.18	0.01	0.27	1.75	4.83	3.26
3B	300	7	0.27	587.43	0.48	0.62	6.67	4.84	17.43	16.45
3B	300	30	0.04	605.00	0.27	1.11	1.24	3.65	17.57	13.91
3B	500	7	0.10	588.20	0.15	0.18	3.12	2.85	11.54	8.89
3B	500	30	0.01	611.21	0.12	0.36	0.67	2.20	17.26	13.13
3B	1000	7	0.01	592.02	0.03	0.14	1.67	1.46	6.01	4.82
3B	1000	30	0.01	615.74	0.28	0.06	0.33	1.22	6.56	4.26

Table 8: Average Mean Squared Error (MSE) and Average Mean Absolute Deviation (MAD) of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Simulation Setup in Xu (2023). In the columns for average MSE and average MAD, the unit is 0.001.

			Average MSE				Average MAD			
Scenario	**M**	n_i	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1A	300	7	7.98	317.18	24.50	23.44	67.78	529.08	116.88	113.94
1A	300	30	1.53	350.07	31.05	25.19	30.39	558.73	118.68	107.82
1A	500	7	3.58	320.65	11.45	10.60	48.31	534.84	80.78	77.99
1A	500	30	0.92	359.06	11.69	11.40	24.30	567.55	76.76	76.65
1A	1000	7	2.22	303.84	7.10	7.25	35.76	523.73	60.91	61.97
1A	1000	30	0.43	354.63	6.53	6.42	16.33	565.64	56.66	55.27
1B	300	7	10.91	1262.97	37.74	35.56	79.25	1003.77	138.60	135.89
1B	300	30	3.13	1381.73	40.54	33.74	42.20	1051.67	139.06	129.85
1B	500	7	7.03	1268.16	17.25	16.95	66.24	1006.76	97.60	96.12
1B	500	30	1.58	1403.67	23.52	19.84	30.15	1059.46	106.15	99.14
1B	1000	7	3.02	1269.16	10.12	10.11	43.39	1007.77	74.67	73.13
1B	1000	30	0.69	1401.91	13.55	6.75	19.30	1059.49	73.50	59.47
2A	300	7	6.09	491.88	17.68	17.47	62.82	634.87	102.26	101.07
2A	300	30	1.38	552.30	27.93	26.68	29.40	675.59	121.24	117.04
2A	500	7	4.28	491.72	12.27	12.43	51.44	635.26	83.97	83.86
2A	500	30	0.90	544.26	13.65	11.96	23.51	673.15	87.50	82.35
2A	1000	7	1.90	486.48	6.53	6.37	34.90	636.25	62.12	61.98
2A	1000	30	0.47	542.56	7.35	6.89	17.05	671.85	64.59	62.78
2B	300	7	5.40	310.03	17.68	17.48	58.39	444.51	101.07	100.09
2B	300	30	1.28	343.31	19.11	19.21	27.81	460.20	93.87	93.62
2B	500	7	3.26	307.45	11.00	10.85	45.17	441.17	75.70	75.21
2B	500	30	0.80	336.99	12.32	11.91	22.86	458.86	78.42	77.51
2B	1000	7	1.68	305.68	5.37	5.07	31.61	437.72	55.62	53.87
2B	1000	30	0.37	334.43	6.63	6.25	14.96	455.67	56.98	55.07
3A	300	7	4.52	342.34	13.95	13.26	51.25	475.14	89.68	87.77
3A	300	30	0.85	342.95	14.70	14.79	22.81	471.88	90.96	91.01
3A	500	7	2.06	346.11	7.81	7.75	36.19	478.33	66.02	66.15
3A	500	30	0.64	348.84	9.65	7.76	19.24	479.12	72.57	66.37
3A	1000	7	1.27	345.42	5.25	4.59	26.67	474.11	53.40	51.13
3A	1000	30	0.28	352.59	5.00	3.27	12.73	482.34	50.64	43.76
3B	300	7	6.94	592.27	17.91	17.07	62.77	648.17	98.15	96.75
3B	300	30	1.28	608.65	17.84	15.02	27.56	653.18	93.76	88.13
3B	500	7	3.23	591.06	11.70	9.07	41.68	646.40	78.30	70.34
3B	500	30	0.68	613.41	17.38	13.48	20.27	656.09	86.77	79.63
3B	1000	7	1.68	593.48	6.04	4.97	31.43	646.77	55.45	51.70
3B	1000	30	0.34	616.96	6.84	4.32	13.78	657.31	54.93	45.06