*mathematics*

MDPI

*Article*

# Logistic Regression Based on Individual-Level Predictors and Aggregate-Level Responses

**Zheng Xu**

Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, USA; zheng.xu@wright.edu; Tel.: +1-937-775-2103

**Abstract:** We propose estimation methods to conduct logistic regression based on individual-level predictors and aggregate-level responses. We derive the likelihood of logistic models in this situation and proposed estimators with different optimization methods. Simulation studies have been conducted to evaluate and compare the performance of the different estimators. A real data-based study has been conducted to illustrate the use of our estimators and compare the different estimators.

**Keywords:** Poisson binomial distribution; logistic regression; data aggregation; likelihood; numerical optimization

**MSC:** 62J12

## 1. Introduction

Data can be reported at different levels due to various considerations including economic, confidentiality, and data collection difficulty. For example, the US Census Bureau reports income at the household level. The aggregate-level data in this example are household income, which is a measure of the combined incomes of all people sharing a particular household or place of residence. The individual-level data in this example are individuals' incomes. The aggregate-level data are defined as data aggregated from individual-level data by groups. Although there are risks in estimating individual-level relationships based on aggregate-level data, such as unequal correlations between variables in aggregate-level data and between the same variables in individual-level data [1,2], researchers continue to use aggregate-level data because in many situations, individual-level data are not available and valid methods for estimating individual-level relationships based on aggregate-level data can be derived [1,3]. The terms "individual" and "aggregate" refer to the different levels and units of analysis [1].

This article intends to solve the problem of estimating models describing an individual-level relationship based on an aggregate-level response variable $Y$ and individual-level predictors $X$. Examples of data situations include survey data, multivariate time series, social data, and biological data, collected and reported at different levels.

Our interest in developing methods to analyze aggregate data was motivated by real-life examples. One example is group testing of infectious diseases in bio-statistics. To reduce the costs, a two-stage sequential testing strategy is applied. In the first stage, group testing is conducted. Individuals showing positive in the first stage are called back for a follow-up individual test. With the first-stage group testing data available, analyses can be conducted. The second example is consumer demand studies in economics. The consumer's characteristics data are available at the individual level, whereas the consumer's purchase data are available only at the aggregate level. The third example is the analysis of multivariate time series. It is likely that different time series are reported at different frequencies. To study the relationships between multiple time series with different frequencies, researchers need to develop statistical methods.

check for updates

**Citation:** Xu, Z. Logistic Regression Based on Individual-Level Predictors and Aggregate-Level Responses. *Mathematics* **2023**, *11*, 746. https://doi.org/10.3390/math11030746

Academic Editors: Niansheng Tang and Shen-Ming Lee

Received: 17 January 2023
Revised: 27 January 2023
Accepted: 30 January 2023
Published: 2 February 2023

Suppose there are $n$ observations in the sample, $(X_i, Y_i), i = 1, 2, \ldots, n, X \in \mathcal{R}^p, Y \in \mathcal{R}$, aggregated into $K$ groups, $G_1, G_2, \ldots, G_K$, with group sizes, respectively, of $n_1, n_2, \cdots, n_K$, $\sum n_g = n$. Denote the set of observations in Group $g$ as $G_g = \{g1, g2, \ldots, gn_g\}$. Aggregate-level $X$ and $Y$, i.e., $(X_g^*, Y_g^*), g = 1, 2, \ldots, K$ are

$$X_g^* = \sum_{i \in G_g} X_i = \sum_{i=1}^{n_g} X_{gi} \text{ and } Y_g^* = \sum_{i \in G_g} Y_i = \sum_{i=1}^{n_g} Y_{gi}. \quad (1)$$

Note that $Y_g^*$ can be any summary statistic calculated from individual-level $Y$ in Group $g$, and we study summation aggregation in this paper.

Researchers have solved this problem for linear models [4–6]. Suppose the linear model describing individual-level data $(X_i, Y_i)$ is

$$Y_i = X_i^T \beta + \epsilon_i, i = 1, 2, \ldots, n.$$

Then, the corresponding model describing the aggregated data $(X_g^*, Y_g^*)$ is

$$Y_g^* = (X_g^*)^T \beta + \epsilon_g^*, g = 1, 2, \ldots, K,$$

where $\epsilon_g^* = \sum_{i \in G_g} \epsilon_i$ is the aggregate-level error so that weighted least squares (WLS) can be applied when $\epsilon_i \sim i.i.d.\ N(0, \sigma^2)$ [4]. More estimators have been proposed for linear regression based on aggregate data or partially aggregate data including Palm and Nijman's MLE estimator [5] and Rawashdeh and Obeidat's Bayesian estimator [6].

Although the estimations of linear regression models in the above data situation have been well studied, more studies are needed for the estimations of other regression models. The aim of this article is to study the estimations of logistic models in the data situation of aggregate-level $Y$ and individual-level $X$. We derive the likelihoods and our estimators with different optimization methods in Section 2, conduct simulation studies to evaluate and compare the performances of different estimators in Section 3, illustrate the use of different estimators in real data-based studies in Section 4, provide discussions in Section 5, and draw conclusions in Section 6.

## 2. Methods

Suppose $n$ independent observations $(X_i, Y_i)$ are modeled by a logistic model

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = X_i^T \beta, i = 1, 2, \ldots, n. \quad (2)$$

Then, $Y_i \sim$ Bernoulli $(\pi_i)$, where $\pi_i = P(Y_i = 1) = \frac{exp(X_i^T \beta)}{1 + exp(X_i^T \beta)}$. When individual-level $X$ and $Y$ are both available, the logistic model as a general linear model can be estimated using a range of methods including the Newton–Raphson method and Fisher's scoring method [7,8].

### *2.1. Likelihood of Aggregate-Level Y and Individual-Level X*

When individual-level $Y$ is not available, we can derive estimators based on aggregate-level $Y$ and individual-level $X$. Suppose the $n$ observations of $(X_i, Y_i)$ are aggregated into $K$ groups, as described in the introduction section, with the aggregated data $(X_g^*, Y_g^*)$, $g = 1, \ldots, K$, defined in Equation (1).

Aggregate-level $Y$ is obtained by summing all $Y$ within each group. Thus, the distribution of the sum of multiple independent random variables is helpful for studying data aggregation. In our logistic regression scenario, we need to calculate the sum of multiple Bernoulli random variables. In statistics, the Poisson binomial distribution is the distribution of a sum of independent Bernoulli random variables, which do not necessarily have different success probabilities [9,10]. The term PoissonBinomial$(n, (\pi_1, \pi_2, \cdots, \pi_n))$

is used to refer to the distribution of the sum of $n$ independent Bernoulli random variables with success probabilities $\pi_1, \pi_2, \cdots, \pi_n$ [9].

Because $Y_g^*$ is the sum of $n_g$ independent Bernoulli random variables,

$$Y_g^* \sim \text{PoissonBinomial}(n_g, (\pi_{g1}, \pi_{g2}, \cdots, \pi_{gn_g})), \tag{3}$$

where the success probability for the $i$th individual in Group $g$ is

$$\pi_{gi} = P(Y_{gi} = 1) = \frac{exp(X_{gi}^T \beta)}{1 + exp(X_{gi}^T \beta)}. \tag{4}$$

Denote the individual likelihood for $Y_g^*$ as $L_g(\beta) = P(Y_g^*; X_{g1}, \ldots, X_{gn_g}, \beta)$. Then, the aggregate likelihood $L(\beta) = \prod_{g=1}^{K} L_g(\beta)$.

### 2.2. Calculation and Maximization of Likelihood

Computing the likelihood function needs to calculate the probability mass function for $Y_g^* \sim \text{PoissonBinomial}(n_g, (\pi_{g1}, \pi_{g2}, \ldots, \pi_{gn_g}))$. The variable $Y_g^*$ will reduce to $\text{Binomial}(n_g, \pi)$ when $\pi_{g1} = \pi_{g2} = \cdots = \pi_{gn_g}$. This case can happen when aggregation is based on the values of $X$ and the individual-level predictors $X_i$ are the same within each group. This specific aggregation has been well studied in the topic of logistic regression based on aggregate data [7,11]. We consider aggregation not based on $X$, i.e., allowing different values of $X$ in a group, in this paper.

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 $\{1, 2, 3, \ldots, n\}$ and $A^c$ is the complement of $A$ [9]. 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 $Pr(Y = k)$, $k = 0, \ldots, y - 1$, which is numerically unstable for large $n$ [12], and the inverse Fourier transform method [13]. Hong [10] further developed it by proposing an algorithm that efficiently implements the exact formula with a closed expression for the Poisson binomial distribution. We adopted Hong's algorithm [10] and exact formula in calculating the likelihood function $L(\beta)$ since they are more precise and numerically stable.

Commonly used optimization methods were adopted to maximize the likelihood $L(\theta)$, including (1) Nelder and Mead's simplex method (NM) [14], (2) the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method [15], and (3) the conjugate gradient (CG) method [16].

### 2.3. Large-Sample Properties of Estimators

As mentioned above, our proposed estimators are obtained by maximizing the aggregate likelihood $L(\beta)$ using different optimization methods (NM, BFGS, and CG). The MLE $\hat{\beta}_{MLE}$ is an estimator that maximizes the aggregate likelihood function, i.e., $\hat{\beta}_{MLE} = argmax_\beta L(\beta)$. If our three optimization methods can always obtain the maximizer of $L(\beta)$, the three estimators will be equal and exactly the same as the MLEs.

In practice, the three optimization methods may not obtain the same value as the MLE. We observed that as the sample size increases, the values obtained using the three optimizations become closer and nearly the same for a large sample size. In discussing large-sample properties, we refer to the scenario of an infinite number of observations and assume that the three optimization methods can always obtain MLEs under the scenario of large samples, i.e., the scenario of an infinite number of observations. Then, our three estimators are identical to the MLE and have the same large-sample properties as the MLE. We add a cautious note that if our estimators are still quite different from the MLE under the large-sample scenario, we cannot state that our estimators have the same large-sample properties as the MLE.

The large-sample properties of the MLE $\hat{\beta}_{MLE}$ [17] include (i) consistency, i.e., $\hat{\beta}_{MLE} \rightarrow \beta$ in probability, and (ii) asymptotic normality, i.e., $\hat{\beta}_{MLE} \sim N(\beta, I(\beta)^{-1})$, where $I(\beta)$ is the expected information matrix, defined as the negative expectation of the second derivative of the log-likelihood. The expected information matrix can be approximated using the observed information matrix, which is the negative of the second derivative (the Hessian matrix) of the log-likelihood function [17].

### *2.4. Software Implementation*

All analyses in this paper were conducted using R software (version 4.2.0). Multiple R packages were used as follows:

- The *PoissonBiomial* package. This package implements multiple exact and approximate methods to calculate Poisson binomial distributions [10]. We used this package to calculate the Poisson binomial distributions and aggregate likelihood $L(\beta)$.
- The *stats* package. This package contains the $optim()$ function, which can conduct general-purpose optimization based on multiple optimization methods, including the Nelder–Mead, BFGS, and CG methods. We used this function to obtain our three estimators using three optimization methods.
- The *glm* package. This package can be used to fit generalized linear models including logistic regression. We used this package to conduct logistic regression.

### *2.5. Computational Burden*

The computational burden of our method relies on three factors: (1) $p$, (2) aggregate-level data sample size $K$, and (3) group size $n_g$.

Our estimator for $\beta$ is obtained by maximizing the aggregate likelihood $L(\beta) = \prod_{g=1}^{K} L_g(\beta)$, $\beta \in \mathcal{R}^p$ using three optimization methods (NM, BFGS, and CG). The number of evaluations of the optimization function $L(\beta)$ and the derivatives will increase with respect to an increase in $p$. Large $p$ will decrease the performance. Given a small fixed number $p$, the number of function evaluations is $O(1)$. Because $L(\beta) = \prod_{g=1}^{K} L_g(\beta)$, the computational amount for $L(\beta)$ is $K$ times the computational amount for $L_g(\beta)$.

The computation of $L_g(\beta)$ includes two steps. In Step 1, the success probabilities are calculated using Equation (4). The computational burden of Step 1 is $O(n_g)$. In Step 2, the probability mass for a Poisson binomial random variable described in Equation (3) is calculated. This step adopts Hong's Algorithm A, which is an efficient implementation of the discrete Fourier transform of the characteristic function (DFT-CF) of the Poisson binomial distribution [10]. The computational burden of Step 2 is $O(n_g^2)$. In total, the computational burden of our estimation method is $O(1) \times K \times O(n_g^2) = O(Kn_g^2)$, given a small constant $p$.

## 3. Simulation Studies

We conducted simulation studies to evaluate the performance of the five estimators. The first estimator, named individual-LR, is the logistic regression estimator based on individual-level $X$ and $Y$. This estimator is infeasible when only aggregate $Y$ 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 $Y$. The second estimator, named naive LR, is the logistic regression estimator based on the mean $X$ in each group and the aggregate $Y$, i.e., $Y_g^* \sim Bin(n_g, X_g^*/n_g)$, $g = 1, 2, \ldots, K$. This estimator can provide a rough approximate estimation.

Estimators 3 to 5 are our estimators that maximize the aggregate likelihood $L(\beta)$ using the Nelder–Mead optimization, CG optimization, and BFGS optimization, named aggregate LR with NM, aggregate LR with CG, and aggregate LR based on BFGS, respectively.

The performances of the estimators were compared in three scenarios. In each scenario, simulations were conducted with sample sizes ($K = 300, 500, 1000$), 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 $(1,3)^T$ (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 $(0,-2,1)^T$ (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 $(0,-2,1,1)^T$ (Scenario 3B).

The bias, variance, mean square error (MSE), and mean absolute deviation (MAD) of each of the five estimators' (E1 to E5) model parameters $(\beta_0, \ldots, \beta_p)$ were calculated. Denote the bias, variance, MSE, and MAD of the $q$-th estimator of $\beta_j$ as $Bias(\hat{\beta}_{j,E_q})$,$Var(\hat{\beta}_{j,E_q})$, $MSE(\hat{\beta}_{j,E_q})$, and$MAD(\hat{\beta}_{j,E_q})$. The average squared bias, variance, MSE, and MAD of the $q$th estimator were calculated as

$$\begin{aligned}
\overline{Bias^2}(E_q) &= [(Bias^2(\hat{\beta}_{0,E_q}) + \cdots + (Bias^2(\hat{\beta}_{p,E_q})]/(p+1),\\
\overline{Var}(E_q) &= [Var(\hat{\beta}_{0,E_q}) + \cdots + Var(\hat{\beta}_{p,E_q})]/(p+1),\\
\overline{MSE}(E_q) &= [MSE(\hat{\beta}_{0,E_q}) + \cdots + (MSE(\hat{\beta}_{p,E_q})]/(p+1),\\
\overline{MAD}(E_q) &= [MAD(\hat{\beta}_{0,E_q}) + \cdots + (MAD(\hat{\beta}_{p,E_q})]/(p+1).
\end{aligned}$$

Please note that we averaged over the squared bias instead of the bias because the positive bias and negative bias can cancel out when averaging the bias. The average across the parameters allows us to obtain the average performance in terms of the squared bias, variance, MSE, and MAD and still maintain the equality of the bias, variance, and MSE, i.e.,

$$\overline{MSE}(E_q) = \overline{Bias^2}(E_q) + \overline{Var}(E_q).$$

In Table 1, we report the average squared biases and variances for the five estimators (E1 to E5) under the different scenarios, sample sizes $K$, and aggregation sizes $n_g$. As we expected, there was a relatively large bias for the naive estimator E2, which used an approximate likelihood by conducting logistic regressions using the average $X$. Our estimators (E3 to E5) had relatively small biases because these estimators were working on the correct and exact likelihood functions. The first estimator E1 had the smallest bias by working on individual-level $X$ and individual-level $Y$. This estimator is widely used when individual-level $Y$ is available. However, under the scenario we intended to solve, only aggregate-level $Y$ was available. Thus, the E1 estimator is infeasible. We still report the performance of E1 to provide some measurements of the possible upper bound of the performance. Because data aggregation will discard information, we expect that estimator E1 will generally perform better than the estimators based on aggregate $Y$.

Next, we check the variances of all five estimators. The variances of all five estimators were similar in the same magnitude level. There was no estimator that performed uniformly better or even generally better than the other estimators. The naive estimator E2 had similar performance or even slightly better performance in the average variance compared with the other estimators (E1, E3–E5). Our estimators (E3 to E5) were slightly worse in terms of variance. We think the slightly worse performance of our estimators (E3–E5) was likely due to the nonlinear optimization to find the MLE in our estimators. In comparison, the logistic regression estimators (E1 and E2) were calculated using iteratively re-weighted least squares (IRLS) (logistic regression ensures global concavity so that it is simpler to find the MLE), which was numerically more stable compared to the nonlinear optimization of a general likelihood function using (1) Nelder and Mead's simplex method [14], (2) the BFGS method [15], and (3) the conjugate gradient (CG) method [16].

**Table 1.** Average Squared Bias and Variance of Estimators E1 to E5 in Scenarios 1A to 3B. $K$ is the sample size of the aggregate data. $n_g$ is the group size in the aggregation.

| | | | **Average Squared Bias** | | | | | **Average Variance** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Scen.** | **K** | $n_g$ | **E1** | **E2** | **E3** | **E4** | **E5** | **E1** | **E2** | **E3** | **E4** | **E5** |
| 1A | 300 | 7 | 0.001 | 0.281 | 0.001 | 0.001 | 0.001 | 0.027 | 0.025 | 0.077 | 0.077 | 0.077 |
| 1A | 300 | 30 | 0.000 | 0.344 | 0.001 | 0.001 | 0.001 | 0.006 | 0.018 | 0.071 | 0.071 | 0.071 |
| 1A | 500 | 7 | 0.000 | 0.293 | 0.000 | 0.000 | 0.000 | 0.009 | 0.008 | 0.020 | 0.020 | 0.020 |
| 1A | 500 | 30 | 0.000 | 0.358 | 0.000 | 0.000 | 0.000 | 0.002 | 0.005 | 0.017 | 0.017 | 0.017 |
| 1A | 1000 | 7 | 0.000 | 0.288 | 0.000 | 0.000 | 0.000 | 0.005 | 0.005 | 0.011 | 0.011 | 0.011 |
| 1A | 1000 | 30 | 0.000 | 0.351 | 0.000 | 0.000 | 0.000 | 0.001 | 0.003 | 0.012 | 0.012 | 0.012 |
| 1B | 300 | 7 | 0.001 | 1.176 | 0.002 | 0.002 | 0.002 | 0.050 | 0.025 | 0.108 | 0.108 | 0.108 |
| 1B | 300 | 30 | 0.000 | 1.367 | 0.000 | 0.000 | 0.000 | 0.012 | 0.014 | 0.099 | 0.098 | 0.099 |
| 1B | 500 | 7 | 0.000 | 1.193 | 0.000 | 0.000 | 0.000 | 0.017 | 0.006 | 0.032 | 0.032 | 0.032 |
| 1B | 500 | 30 | 0.000 | 1.369 | 0.000 | 0.000 | 0.000 | 0.004 | 0.004 | 0.032 | 0.030 | 0.033 |
| 1B | 1000 | 7 | 0.000 | 1.181 | 0.000 | 0.000 | 0.000 | 0.009 | 0.004 | 0.018 | 0.018 | 0.018 |
| 1B | 1000 | 30 | 0.000 | 1.388 | 0.000 | 0.000 | 0.000 | 0.002 | 0.003 | 0.019 | 0.019 | 0.019 |
| 2A | 300 | 7 | 0.000 | 0.471 | 0.002 | 0.002 | 0.002 | 0.031 | 0.023 | 0.073 | 0.073 | 0.073 |
| 2A | 300 | 30 | 0.000 | 0.523 | 0.004 | 0.004 | 0.004 | 0.007 | 0.016 | 0.071 | 0.071 | 0.071 |
| 2A | 500 | 7 | 0.000 | 0.462 | 0.000 | 0.000 | 0.000 | 0.008 | 0.007 | 0.020 | 0.020 | 0.020 |
| 2A | 500 | 30 | 0.000 | 0.538 | 0.000 | 0.000 | 0.000 | 0.002 | 0.006 | 0.019 | 0.019 | 0.019 |
| 2A | 1000 | 7 | 0.000 | 0.464 | 0.000 | 0.000 | 0.000 | 0.005 | 0.004 | 0.012 | 0.012 | 0.012 |
| 2A | 1000 | 30 | 0.000 | 0.532 | 0.000 | 0.000 | 0.000 | 0.001 | 0.003 | 0.013 | 0.013 | 0.013 |
| 2B | 300 | 7 | 0.000 | 0.291 | 0.000 | 0.000 | 0.000 | 0.025 | 0.018 | 0.059 | 0.059 | 0.059 |
| 2B | 300 | 30 | 0.000 | 0.336 | 0.003 | 0.003 | 0.003 | 0.006 | 0.016 | 0.066 | 0.066 | 0.066 |
| 2B | 500 | 7 | 0.000 | 0.277 | 0.000 | 0.000 | 0.000 | 0.007 | 0.007 | 0.017 | 0.017 | 0.017 |
| 2B | 500 | 30 | 0.000 | 0.340 | 0.000 | 0.000 | 0.000 | 0.002 | 0.005 | 0.017 | 0.017 | 0.017 |
| 2B | 1000 | 7 | 0.000 | 0.282 | 0.000 | 0.000 | 0.000 | 0.005 | 0.004 | 0.012 | 0.012 | 0.012 |
| 2B | 1000 | 30 | 0.000 | 0.340 | 0.000 | 0.000 | 0.000 | 0.001 | 0.003 | 0.012 | 0.012 | 0.012 |
| 3A | 300 | 7 | 0.000 | 0.332 | 0.001 | 0.000 | 0.000 | 0.018 | 0.020 | 0.045 | 0.052 | 0.055 |
| 3A | 300 | 30 | 0.000 | 0.345 | 0.003 | 0.002 | 0.001 | 0.004 | 0.019 | 0.045 | 0.049 | 0.055 |
| 3A | 500 | 7 | 0.000 | 0.336 | 0.000 | 0.000 | 0.000 | 0.006 | 0.006 | 0.014 | 0.015 | 0.015 |
| 3A | 500 | 30 | 0.000 | 0.344 | 0.000 | 0.000 | 0.000 | 0.001 | 0.006 | 0.013 | 0.016 | 0.017 |
| 3A | 1000 | 7 | 0.000 | 0.340 | 0.000 | 0.000 | 0.000 | 0.003 | 0.004 | 0.008 | 0.009 | 0.009 |
| 3A | 1000 | 30 | 0.000 | 0.346 | 0.000 | 0.000 | 0.000 | 0.001 | 0.004 | 0.008 | 0.008 | 0.010 |
| 3B | 300 | 7 | 0.000 | 0.567 | 0.002 | 0.001 | 0.001 | 0.025 | 0.020 | 0.056 | 0.064 | 0.068 |
| 3B | 300 | 30 | 0.000 | 0.603 | 0.005 | 0.004 | 0.003 | 0.006 | 0.014 | 0.063 | 0.069 | 0.077 |
| 3B | 500 | 7 | 0.000 | 0.578 | 0.001 | 0.000 | 0.000 | 0.007 | 0.005 | 0.015 | 0.019 | 0.018 |
| 3B | 500 | 30 | 0.000 | 0.614 | 0.000 | 0.000 | 0.000 | 0.002 | 0.005 | 0.018 | 0.025 | 0.022 |
| 3B | 1000 | 7 | 0.000 | 0.587 | 0.000 | 0.000 | 0.000 | 0.005 | 0.003 | 0.010 | 0.010 | 0.010 |
| 3B | 1000 | 30 | 0.000 | 0.608 | 0.000 | 0.000 | 0.000 | 0.001 | 0.003 | 0.010 | 0.012 | 0.012 |

We point out that although the naive estimator E2 worked on an incorrect (or approximate) likelihood function, which can lead to a large bias due to the incorrect likelihood

function, the performance of the variance of E2 did not necessarily become worse. A similar phenomenon was the under-fitting in the data analysis. Suppose the true relationship is a quadratic function. If a linear function is used in model fitting, the estimator will have a large bias due to model mis-specification, whereas the variance may not increase. We note that the main disadvantage of estimator E2 was the use of an incorrect or approximate likelihood function, which can lead to a large bias. Using the correct exact likelihood, i.e., our estimators (E3 to E5), can solve the issue of bias due to the slight increase in variance from the switch in finding the MLE from iteratively reweighted least squares (IRLS) to nonlinear optimization using the Nelder and Mead's simplex, BFGS, and CG methods. We compared the decrease in bias and increase in variance and think the bias reduction will dominate the variance increase in our estimators. We calculated the overall performance in terms of the MSE and MAD to confirm it.

Our simulation results showed that the naive estimator had a large bias due to the use of an incorrect or approximate likelihood function, which can hurt the MSE. Thus, in Table 2, we report the average performance of the five estimators (E1 to E5) in terms of the MSE and MAD. Our simulation results indicated that our proposed estimators (E3 to E5) were better than the naive LR estimator (E2). As expected, the infeasible estimator (E1) based on individual-level $Y$ performed better than the other four feasible estimators (E2 to E5) based on aggregate-level $Y$ due to the loss of information in the data aggregation. Our estimator based on Nelder and Mead's simplex optimization (E3) performed slightly better than our estimator based on BFGS optimization (E4) and CG optimization (E5).

**Table 2.** Average MSE and MAD of Estimators E1 to E5 in Scenarios 1A to 3B. $K$ is the sample size of the aggregate data. $n_g$ is the group size in the aggregation.

| | | | **Average MSE** | | | | | **Average MAD** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Scen.** | **K** | $n_g$ | **E1** | **E2** | **E3** | **E4** | **E5** | **E1** | **E2** | **E3** | **E4** | **E5** |
| 1A | 300 | 7 | 0.027 | 0.307 | 0.078 | 0.078 | 0.078 | 0.129 | 0.504 | 0.198 | 0.198 | 0.198 |
| 1A | 300 | 30 | 0.006 | 0.362 | 0.072 | 0.072 | 0.072 | 0.062 | 0.558 | 0.192 | 0.192 | 0.192 |
| 1A | 500 | 7 | 0.009 | 0.302 | 0.020 | 0.020 | 0.020 | 0.075 | 0.515 | 0.109 | 0.109 | 0.109 |
| 1A | 500 | 30 | 0.002 | 0.363 | 0.017 | 0.017 | 0.017 | 0.036 | 0.568 | 0.093 | 0.093 | 0.093 |
| 1A | 1000 | 7 | 0.005 | 0.293 | 0.011 | 0.011 | 0.011 | 0.057 | 0.509 | 0.080 | 0.080 | 0.080 |
| 1A | 1000 | 30 | 0.001 | 0.354 | 0.012 | 0.012 | 0.012 | 0.028 | 0.563 | 0.078 | 0.078 | 0.078 |
| 1B | 300 | 7 | 0.051 | 1.200 | 0.109 | 0.109 | 0.109 | 0.173 | 0.970 | 0.235 | 0.235 | 0.235 |
| 1B | 300 | 30 | 0.012 | 1.380 | 0.099 | 0.098 | 0.099 | 0.084 | 1.046 | 0.222 | 0.221 | 0.222 |
| 1B | 500 | 7 | 0.017 | 1.200 | 0.032 | 0.032 | 0.032 | 0.098 | 0.977 | 0.130 | 0.130 | 0.130 |
| 1B | 500 | 30 | 0.004 | 1.373 | 0.033 | 0.031 | 0.033 | 0.048 | 1.048 | 0.129 | 0.125 | 0.129 |
| 1B | 1000 | 7 | 0.009 | 1.185 | 0.018 | 0.018 | 0.018 | 0.072 | 0.973 | 0.100 | 0.100 | 0.100 |
| 1B | 1000 | 30 | 0.002 | 1.390 | 0.019 | 0.019 | 0.019 | 0.038 | 1.054 | 0.098 | 0.098 | 0.098 |
| 2A | 300 | 7 | 0.031 | 0.494 | 0.075 | 0.075 | 0.075 | 0.138 | 0.627 | 0.204 | 0.204 | 0.204 |
| 2A | 300 | 30 | 0.007 | 0.539 | 0.075 | 0.075 | 0.075 | 0.065 | 0.661 | 0.200 | 0.200 | 0.200 |
| 2A | 500 | 7 | 0.008 | 0.469 | 0.021 | 0.021 | 0.021 | 0.070 | 0.622 | 0.111 | 0.111 | 0.111 |
| 2A | 500 | 30 | 0.002 | 0.543 | 0.020 | 0.020 | 0.020 | 0.036 | 0.669 | 0.106 | 0.106 | 0.106 |
| 2A | 1000 | 7 | 0.005 | 0.468 | 0.012 | 0.012 | 0.012 | 0.057 | 0.620 | 0.084 | 0.084 | 0.084 |
| 2A | 1000 | 30 | 0.001 | 0.535 | 0.013 | 0.013 | 0.013 | 0.030 | 0.667 | 0.085 | 0.085 | 0.085 |
| 2B | 300 | 7 | 0.025 | 0.309 | 0.059 | 0.059 | 0.059 | 0.124 | 0.445 | 0.182 | 0.182 | 0.182 |
| 2B | 300 | 30 | 0.006 | 0.352 | 0.068 | 0.068 | 0.068 | 0.060 | 0.464 | 0.180 | 0.180 | 0.180 |
| 2B | 500 | 7 | 0.007 | 0.284 | 0.017 | 0.017 | 0.017 | 0.065 | 0.424 | 0.099 | 0.099 | 0.099 |

**Table 2.** *Cont.*

| | | | Average MSE | | | | | Average MAD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Scen.** | **K** | $n_g$ | **E1** | **E2** | **E3** | **E4** | **E5** | **E1** | **E2** | **E3** | **E4** | **E5** |
| 2B | 500 | 30 | 0.002 | 0.344 | 0.018 | 0.018 | 0.018 | 0.034 | 0.463 | 0.093 | 0.093 | 0.093 |
| 2B | 1000 | 7 | 0.005 | 0.286 | 0.012 | 0.012 | 0.012 | 0.054 | 0.425 | 0.081 | 0.081 | 0.081 |
| 2B | 1000 | 30 | 0.001 | 0.343 | 0.012 | 0.012 | 0.012 | 0.024 | 0.461 | 0.075 | 0.075 | 0.075 |
| 3A | 300 | 7 | 0.018 | 0.352 | 0.046 | 0.052 | 0.055 | 0.104 | 0.486 | 0.162 | 0.168 | 0.170 |
| 3A | 300 | 30 | 0.004 | 0.364 | 0.047 | 0.051 | 0.056 | 0.049 | 0.495 | 0.161 | 0.165 | 0.170 |
| 3A | 500 | 7 | 0.006 | 0.342 | 0.014 | 0.015 | 0.015 | 0.058 | 0.474 | 0.090 | 0.091 | 0.091 |
| 3A | 500 | 30 | 0.001 | 0.350 | 0.014 | 0.016 | 0.017 | 0.028 | 0.481 | 0.088 | 0.090 | 0.091 |
| 3A | 1000 | 7 | 0.003 | 0.344 | 0.008 | 0.009 | 0.009 | 0.043 | 0.475 | 0.066 | 0.067 | 0.067 |
| 3A | 1000 | 30 | 0.001 | 0.350 | 0.008 | 0.008 | 0.009 | 0.022 | 0.480 | 0.069 | 0.069 | 0.070 |
| 3B | 300 | 7 | 0.025 | 0.587 | 0.058 | 0.065 | 0.069 | 0.119 | 0.645 | 0.178 | 0.184 | 0.188 |
| 3B | 300 | 30 | 0.006 | 0.617 | 0.068 | 0.073 | 0.079 | 0.057 | 0.656 | 0.180 | 0.184 | 0.190 |
| 3B | 500 | 7 | 0.007 | 0.584 | 0.015 | 0.019 | 0.018 | 0.066 | 0.644 | 0.092 | 0.095 | 0.095 |
| 3B | 500 | 30 | 0.002 | 0.619 | 0.019 | 0.025 | 0.022 | 0.033 | 0.659 | 0.096 | 0.102 | 0.099 |
| 3B | 1000 | 7 | 0.005 | 0.590 | 0.010 | 0.010 | 0.010 | 0.055 | 0.647 | 0.075 | 0.075 | 0.075 |
| 3B | 1000 | 30 | 0.001 | 0.611 | 0.011 | 0.012 | 0.012 | 0.026 | 0.655 | 0.072 | 0.074 | 0.074 |

We found the performances of our estimators (E3, E4, E5) were slightly worse when the group size $n_g = 30$ compared with the performances of our estimators when the group size $n_g = 7$. We expect that the performance of our estimators may decrease for a large group size $n_g$ due to rounding errors in computation.

## 4. Real Data-Based Studies

We used real data to illustrate the use of our estimators and compare the different estimators. The dataset used was the "Social-Network-Ads" dataset from the Kaggle Machine Learning Forum (https://www.kaggle.com, accessed on 12 January 2023).

The dataset has been used by statisticians and data scientists to illustrate the use of logistic regression in categorical data analysis. We used the dataset to illustrate the use of our method to conduct logistic regression in the presence of data aggregation.

The Social-Network-Ads dataset in Kaggle is a categorical dataset for determining whether a user purchased a particular product. The dataset (https://www.kaggle.com/datasets, accessed on 12 January 2023) contains 400 people/observations. The information about the person's purchase action (purchased with a binary variable of 1 denotes purchased and 0 denotes not purchased), as well as the person's age and estimated salary, is provided. Logistic regression has been recommended in Kaggle to model the person's purchase action based on the person's age and estimated salary. We intend to apply our method to this dataset in the presence of data aggregation.

The original dataset is at the individual level, which allows us to conduct logistic regression based on individual-level $Y$ and $X$. We standardized $X$ by $X^* = (X - mean(X))/sd(X)$ in data pre-processing. Standardization of $X$ allows for better estimation and interpretation. Standardized coefficients $\beta^*$ are obtained by logistics regression of $Y$ on standardized data $X^*$. The original slope coefficients in $\beta$ can be calculated by the formula $\hat{\beta} = \hat{\beta}^* \times sd(X)$ and then the intercept coefficient can be calculated.

We imposed data aggregation on this dataset with an aggregation size $n_g = 3, 5, 7$. We randomly divided the persons into groups of size $n_g$ and calculated the group aggregate of the purchase actions $Y$. Due to confidentiality and the cost of collecting individual-level data, businesses and organizations can choose to post data information at an aggregate

level. We mimicked the data aggregation process by random grouping and calculated the aggregate-level $Y$ based on the individual-level $Y$. We repeated the data aggregation 300 times. In this way, we generated 300 datasets, with the individual-level $X$ and aggregate-level $Y$ calculated.

For each dataset, we conducted logistic regression based on individual-level $X$ and $Y$ and obtained our estimator E1. Since data aggregation discards information, we evaluated the other estimators by checking whether they were close to estimator E1. Because the true values of the coefficients in individual-level logistic regression models are not known in real data-based studies, we used estimator E1 as a gold-standard estimator. We compared the other estimators based on aggregate-level $Y$ to determine which estimator was closer to our gold-standard estimator E1. Note that E1 is an infeasible estimator when individual-level $X$ is not available.

The estimator E1 was calculated based on individual-level $X$ and individual-level $Y$. The estimated value of estimator E1 remained the same in our 300 generated datasets and E1 was treated as the gold-standard estimator; thus, we denote it as $(\beta_0, \beta_1, \beta_2)$.

Denote the estimated value of $\beta_i$ for the $j$-th estimator in the $k$-th dataset by $\hat{\beta}_{i,Ej}(D_k)$. The bias, variance, MSE, and MAD of estimators E2 to E5 for $\beta_0$,$\beta_1$, and $\beta_2$ were calculated by the formulae

$$\begin{aligned}
\overline{\hat{\beta}_{i,Ej}} &= \sum_{k=1}^{300} \hat{\beta}_{i,Ej}(D_k)/300 \\
Bias(\hat{\beta}_{i,Ej}) &= \overline{\hat{\beta}_{i,Ej}} - \beta_i \\
Var(\hat{\beta}_{i,Ej}) &= \sum_{k=1}^{300} \{\hat{\beta}_{i,Ej}(D_k) - \overline{\hat{\beta}_{i,Ej}}\}^2/300 \\
MSE(\hat{\beta}_{i,Ej}) &= \sum_{k=1}^{300} \{\hat{\beta}_{i,Ej}(D_k) - \beta_i\}^2/300 \\
MAD(\hat{\beta}_{i,Ej}) &= \sum_{k=1}^{300} |\hat{\beta}_{i,Ej}(D_k) - \beta_i|/300
\end{aligned}$$

For the four estimators based on aggregate-level Y and individual-level X, i.e., E2 to E5, we report the biases and variances in Table 3. We can see that in most cases, there are large biases in estimating $\beta_0$ and $\beta_2$ and relatively smaller biases in estimating $\beta_1$ using the naive estimator E2. Our proposed estimators (E3 to E5) always achieved smaller biases compared to the naive estimator E2. This is because the naive estimator E2 used an approximate likelihood instead of an exact likelihood, which our proposed estimators are based on. In terms of variance, the naive estimator had a relatively smaller variance compared with our estimators E3 to E5. We point out that the calculation algorithm used in E2, i.e., iteratively reweighted least squares (IRLS), was more numerically stable compared with the nonlinear optimization algorithms adopted by our estimators, i.e., Nelder and Mead's simplex method, the BFGS method, and the conjugate gradient method.

We then checked the overall performance of the different estimators and report the MSE and MAD in Table 4. We found that our estimators (E3 to E5) had better performance than the naive estimator (E2) in terms of the MSE and MAD in all situations based on the Social-Network-Ads dataset.

**Table 3.** Biases and Variances of Estimators E2 to E5 based on Aggregate-Level $Y$ and Individual-Level $X$. $n_g$ is the group size in the aggregation.

| | | **Bias** | | | | **Variance** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Coef.** | $n_g$ | **E2** | **E3** | **E4** | **E5** | **E2** | **E3** | **E4** | **E5** |
| $\beta_0$ | 3 | 1.580 | −0.018 | 0.620 | −0.017 | 0.094 | 0.160 | 0.125 | 0.160 |
| $\beta_0$ | 5 | 1.721 | −0.054 | 0.877 | −0.054 | 0.115 | 0.383 | 0.218 | 0.383 |
| $\beta_0$ | 7 | 1.769 | −0.156 | 1.019 | −0.156 | 0.194 | 0.721 | 0.319 | 0.721 |
| $\beta_1$ | 3 | −0.163 | 0.001 | −0.073 | 0.001 | 0.002 | 0.003 | 0.002 | 0.003 |
| $\beta_1$ | 5 | −0.176 | 0.005 | −0.108 | 0.005 | 0.002 | 0.006 | 0.004 | 0.006 |
| $\beta_1$ | 7 | −0.180 | 0.016 | −0.127 | 0.016 | 0.004 | 0.012 | 0.006 | 0.012 |
| $\beta_2$ | 3 | −1.007 | 0.039 | −0.195 | 0.039 | 0.025 | 0.062 | 0.026 | 0.062 |
| $\beta_2$ | 5 | −1.123 | 0.075 | −0.258 | 0.075 | 0.043 | 0.174 | 0.053 | 0.174 |
| $\beta_2$ | 7 | −1.141 | 0.150 | −0.266 | 0.150 | 0.053 | 0.260 | 0.087 | 0.260 |

**Table 4.** MSE and MAD of Estimators E2 to E5 based on Aggregate-Level $Y$ and Individual-Level $X$. $n_g$ is the group size in the aggregation.

| | | **MSE** | | | | **MAD** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Coef.** | $n_g$ | **E2** | **E3** | **E4** | **E5** | **E2** | **E3** | **E4** | **E5** |
| $\beta_0$ | 3 | 2.591 | 0.160 | 0.509 | 0.160 | 1.580 | 0.317 | 0.644 | 0.317 |
| $\beta_0$ | 5 | 3.076 | 0.385 | 0.986 | 0.385 | 1.721 | 0.480 | 0.909 | 0.480 |
| $\beta_0$ | 7 | 3.322 | 0.743 | 1.356 | 0.742 | 1.769 | 0.651 | 1.036 | 0.650 |
| $\beta_1$ | 3 | 0.028 | 0.003 | 0.008 | 0.003 | 0.163 | 0.041 | 0.077 | 0.041 |
| $\beta_1$ | 5 | 0.033 | 0.006 | 0.016 | 0.006 | 0.176 | 0.060 | 0.112 | 0.060 |
| $\beta_1$ | 7 | 0.036 | 0.012 | 0.023 | 0.012 | 0.181 | 0.085 | 0.130 | 0.085 |
| $\beta_2$ | 3 | 1.039 | 0.063 | 0.064 | 0.063 | 1.007 | 0.186 | 0.222 | 0.186 |
| $\beta_2$ | 5 | 1.304 | 0.179 | 0.120 | 0.179 | 1.123 | 0.325 | 0.297 | 0.325 |
| $\beta_2$ | 7 | 1.356 | 0.282 | 0.157 | 0.282 | 1.141 | 0.409 | 0.331 | 0.409 |

## 5. Discussion

Our estimators are obtained by maximizing the nonlinear likelihood function $L(\beta)$, $\beta \in \mathcal{R}^p$. Different optimization methods can influence the performance of our estimators. Further studies can be conducted on other optimization methods such as the genetic algorithm or using multiple starting values. The performance of optimization is expected to decrease when $p$ increases.

We only consider independent individual-level data, i.e., $(X_i, Y_i)$, $i = 1, 2, \cdots, n$. The $n$ observations are randomly divided into groups of size $n_g$ and the aggregate-level $Y$ is calculated after grouping. In this paper, we only consider the situation of "grouping completely at random", which means that the grouping mechanism is completely random. The values of $X$ and $Y$ do not influence the grouping. Further studies can be conducted beyond this type of grouping mechanism.

Our aggregation scheme is based on independent individual-level data. There are more aggregations schemes. For example, temporal aggregation can aggregate dependent data, which can generate aggregated low-frequency time series based on high-frequency time series by summing every $m$ consecutive time points. For example, we can aggregate daily time series into weekly time series by summing every $m = 7$ consecutive daily observations. Temporal aggregation is often based on a time series model such as an integer-valued generalized autoregressive conditional heteroskedasticity (INGARCH) [18].

We note that the proposed methods also allow for other link functions in addition to the logit link. For example, when a probit link function is used, we can estimate individual-level probit models based on aggregate-level $Y$ and individual-level $X$. In addition, we only consider binary responses in this paper. A follow-up study to extend our methods to handle responses with more than two levels are under development.

## 6. Conclusions

We proposed methods to estimate logistic models based on individual-level predictors and aggregate-level responses. We conducted simulation studies to evaluate the performance of the estimators and show the advantage of our estimators. We then used the Social-Network-Ads dataset to illustrate the use of our estimators in the presence of data aggregation and compared the different estimators. Both the simulation studies and real data-based studies have shown the advantage of our estimators in estimating logistics models describing individual-level behaviors based on aggregate-level Y and individual-level X, i.e., when there is data aggregation in the response variable.

**Funding:** This research received no external funding.

**Data Availability Statement:** All data used in the study are publicly available.

**Conflicts of Interest:** The author declares no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| BFGS | Broyden–Fletcher–Goldfarb–Shanno method |
| CF | Characteristic function |
| CG | Conjugate gradient |
| DFT | Discrete Fourier transform |
| IRLS | Iteratively re-weighted least squares |
| LR | Logistics regression |
| MAD | Mean Absolute Deviation |
| MSE | Mean Square Error |
| NM | Nelder-Mead method |

## References

1. Firebaugh, G. A rule for inferring individual-level relationships from aggregate data. *Am. Sociol. Rev.* **1978**, *43*, 557–572. [CrossRef]
2. Robinson, W.S. Ecological correlations and the behavior of individuals. *Int. J. Epidemiol.* **2009**, *38*, 337–341. [CrossRef]
3. Hammond, J.L. Two sources of error in ecological correlations. *Am. Sociol. Rev.* **1973**, *38*, 764–777. [CrossRef]
4. Hsiao, C. Linear regression using both temporally aggregated and temporally disaggregated data. *J. Econom.* **1979**, *10*, 243–252. [CrossRef]
5. Palm, F.C.; Nijman, T.E. Linear regression using both temporally aggregated and temporally disaggregated data. *J. Econom.* **1982**, *19*, 333–343. [CrossRef]
6. Rawashdeh, A.; Obeidat, M. A Bayesian Approach to Estimate a Linear Regression Model with Aggregate Data. *Austrian J. Stat.* **2019**, *48*, 90–100. [CrossRef]
7. Agresti, A. *Categorical Data Analysis*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2013.
8. Givens, G.; Hoeting, J. *Computational Statistics*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2012.
9. Wang, Y.H. On the number of successes in independent trials. *Stat. Sin.* **1993**, *3*, 295–312.
10. Hong, Y. On computing the distribution function for the Poisson binomial distribution. *Comput. Stat. Data Anal.* **2013**, *59*, 41–51. [CrossRef]
11. Bilder, C.; Loughin, T. *Analysis of Categorical Data with R*; Chapman & Hall/CRC Texts in Statistical Science; CRC Press: Boca Raton, FL, USA, 2014.
12. Chen, X.H.; Dempster, A.P.; Liu, J.S. Weighted finite population sampling to maximize entropy. *Biometrika* **1994**, *81*, 457–469. [CrossRef]
13. Fernández, M.; Williams, S. Closed-form expression for the poisson-binomial probability density function. *IEEE Trans. Aerosp. Electron. Syst.* **2010**, *46*, 803–817. [CrossRef]
14. Nelder, J.A.; Mead, R. A simplex method for function minimization. *Comput. J.* **1965**, *7*, 308–313. [CrossRef]

15. Fletcher, R. A new approach to variable metric algorithms. *Comput. J.* **1970**, *13*, 317–322. [CrossRef]
16. Flecher, R.; Reeves, C. Function minimization by conjugate gradient. *Comput. J.* **1964**, *7*, 149–154. [CrossRef]
17. Shao, J. *Mathematical statistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2003.
18. Su, B.; Zhu, F. Temporal aggregation and systematic sampling for INGARCH processes. *J. Stat. Plan. Inference* **2022**, *219*, 120–133. [CrossRef]