**EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS**
Vol. 5, No. 3, 2012, 333-356
ISSN 1307-5543 – www.ejpam.com

# Data Fusion Using Weighted Likelihood

Pengfei Guo, Xiaogang Wang, Yuehua Wu*

*Department of Mathematics and Statistics, York University, Toronto, Canada*

**Abstract.** This article proposes to perform data fusion by using an adaptive weighted likelihood function when data sets are available from related populations. The main objective of data fusion is to integrate information from different sources to improve the quality of inference when the sample size from the target population is small or moderate. The weighted likelihood function is employed simply as an instrument to facilitate the data fusion process. The weighted likelihood method has information-theoretic justification and embraces the widely used classical likelihood method which utilizes only on the data set from the target population. The degree of information integration in the proposed data fusion process is determined by the likelihood weights which should be chosen in a reasonable and adaptive way. The major challenge in the proposed data fusion process is then to choose likelihood weights adaptively and effectively when the deterministic relationships among all related parameters are unknown. We propose adaptive likelihood weights based on the estimated likelihood ratio. We show that the data fusion involving all relevant data sets could significantly improve the mean squared error (MSE) of the classical maximum likelihood estimator which only uses data set from the target population. It also increases the power for hypothesis testing. The proposed estimator is shown to be consistent and asymptotically normally distributed in the framework of generalized linear models. The advantage of the proposed weighted likelihood estimator for linear models is illustrated numerically by a simulation study. A real data example is also provided.

**2010 Mathematics Subject Classifications**: 62F12, 62J05, 62J12

**Key Words and Phrases**: Cross-validation, Nonparametric regression, Relative likelihood ratio, Semi-parametric estimation, Weighted likelihood

## 1. Introduction

In clinical trials for medical research, a common problem that often arises is how to efficiently estimate the parameter of interest when the sample size from the population of inferential interest is small due to the cost or other limitations associated with the experiment. When the sample size from the population of inferential interest is small and observations from related studies are available, one important question is whether there is any merit to

*Corresponding author.

*Email addresses:* pguo@mathstat.yorku.ca (P. Guo), stevenw@mathstat.yorku.ca (S. Wang), wuyh@mathstat.yorku.ca (Y. Wu)

 

combine information from populations that are known to be different than the population of inferential interest.

The classical likelihood method would concentrate solely on the sample from the the target population without incorporating other relevant information. The advantage of doing so is to avoid possible bias or contamination of the data sampled directly from the target population. The maximum likelihood estimator (MLE) is well known to enjoy asymptotic properties such as consistency and asymptotic normality. It is one of the most widely used method in statistical inference. However, the asymptotic properties are of little help when the sample size is small or very moderate. The MLE could provide quite misleading results due to a insufficient sample size. By contrast, the Bayesian method can effectively combine information when the parameters from these populations are assumed to be random variables from a hyper-distribution. It has many advantages since prior information can be formally incorporated and the inferences are conditional on all the data. Bayesian inference might be computationally intensive for non- trivial cases. It is well known, however, the powerful Monte Carlo Markov Chain method could be computationally intensive and challenging for the high dimensional case when a conjugate prior can not be assumed.

For exploratory purposes, we propose to integrate all available information through a data fusion process based on an adaptive weighted likelihood function. The weighted likelihood function that we employ can be very adaptive and easy to implement. Since there are many weighted likelihoods proposed for various purposes, it is necessary to provide an brief overview and avoid any possible confusion later on. [4] introduced the idea of local likelihood to derive local inference. [2] defined her version of local likelihood in the context of non-parametric regression. [17] presented the general form of the local likelihood. [9] and [10] proposed their relevant weighted likelihood function to combine information in which the weights are very general. [12] proposed their adaptive weights when the time trend is present. [11] reviewed most related weighted likelihood methods proposed in the literature. All the aforementioned weighted likelihood methods focus on a setting in which the number of populations goes to infinity. In this article, however,we focus on situations in which the number of populations is actually fixed. For example, one might be interested in the efficacy of one particular treatment when the results of the current and several historical clinical trials are available.

Extending the *relevant weighted likelihood* by [9], [21] proposed their version of the WL when the number of populations is fixed. We now briefly their weighted likelihood function as follows. Suppose that we observe independent random response vectors $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_m$ with probability density functions $f(.;\theta_1),\ldots,f(.;\theta_m)$, where $\boldsymbol{Y}_i=(Y_{i1},\ldots,Y_{in_i})^{\mathrm{T}}, i=1,2,\cdots,m$. Further suppose that only population 1, in particular $\theta_1$, an unknown vector of parameters, is of inferential interest. Given data, $\boldsymbol{Y}_1=\boldsymbol{y}_1$, the classical likelihood would be

$$L_1(\boldsymbol{y}_1,\theta_1)=\prod_{j=1}^{n_1} f(y_{1j};\theta_1).$$

When the parameters $\theta_2,\ldots,\theta_m$ are thought to interconnected with $\theta_1$, for given

$\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_m)$, the weighted likelihood (WL) for $\theta_1$ is defined as

$$\mathrm{WL}(\boldsymbol{y}; \theta_1) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(y_{ij}; \theta_1)^{\lambda_i},$$

where $\lambda = (\lambda_1, \ldots, \lambda_m)$, the "weights vector", must be specified. We emphasize that the parameters from the related populations, $\theta_2, \ldots, \theta_m$, which are unknown, do not appear in the WL, since the inferential interest focuses on $\theta_1$ of the first population.

The WL estimator (WLE) for $\theta_1$, say $\tilde{\theta}_1$, is defined as the maximizer of the objective function $\mathrm{WL}(\boldsymbol{y}; \theta_1)$:

$$\tilde{\theta}_1 = \arg \sup_{\theta_1 \in \Theta} \mathrm{WL}(\boldsymbol{y}; \theta_1).$$

In order for the weighted likelihood to be effective, the weights must be chosen adaptively so that all information obtained from related population must be evaluated to determine their relevance. [7] proposed a cross-validatory procedure for weights selection and showed that the resulting WLE is consistent and asymptotically normal. However, the cross-validatory approach is challenged with the following problems: (a) it is computationally intensive if the WLE does not have an analytical form, and (b) numerical performances could be unstable when the sample sizes are unequal or the sample sizes are very small. Thus, we propose a simple and effective method to choose the likelihood weights adaptively. This method is better than the cross-validated weights proposed by [7] in that its computation is robust and implementation is straightforward. More importantly, it avoids all the numerical problems with the cross-validated weights when the sample sizes are small or unequal. We derive the consistency and asymptotic distribution of the resulting WLE.

The article is organized as follows. In Section 2, we develop a WL procedure. The asymptotic properties of the WLE for generalized linear models using the proposed adaptive weights are presented in Section 3. Section 4 presents the results from simulation experiments that illustrate the numerical performance of our method. Section 5 shows how to obtain the empirical distribution of the WLE by bootstrap. Section 6 analyzes a real data set from by using the proposed approach. Section 7 provides conclusions and discussions.

## 2. The Weighted Likelihood and Adaptive Weights

### 2.1. The Weighted Likelihood

We assume the existence of the $m$ population density functions are unknown and play purely conceptual roles. More specifically, assume $\sigma$-finite probability spaces $(\Omega, \mathscr{F}, \mu_i), i = 1, 2, \ldots, m$, with probability measures $\mu_i$'s that are *absolutely continuous* with respect to one another. The existence of a $\sigma$-finite measure $\nu$ that dominates the $\mu_i$'s then follows. We take the $f_i$ to be the Radon-Nikodym derivatives of $\mu_i$ with respect to $\nu$ for $i = 1, 2, \ldots, m$.

Suppose that we observe independent random response vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ with probability density functions $f(.; \theta_1), \ldots, f(.; \theta_m)$, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}, i = 1, 2, \cdots, m$. Further

suppose that only population 1, in particular $\theta_1$, an unknown vector of parameters, is of inferential interest. Given data, $\boldsymbol{Y}_1 = \boldsymbol{y}_1$, the classical likelihood of $\theta_1$ is defined by

$$L_1(\boldsymbol{y}_1;\theta_1) = \prod_{j=1}^{n_1} f(y_{1j};\theta_1).$$

Similarly, we can define the likelihood for each $\theta_i$ as follows:

$$L_i(\boldsymbol{y}_i,\theta_1) = \prod_{j=1}^{n_i} f(y_{ij};\theta_i).$$

where $i = 2,\cdots,m$.

If all parameters are not related in any way, the correct likelihood for $\theta_1$ should be $L_1(\boldsymbol{y}_1;\theta_1)$. However, $L_1(\boldsymbol{y}_1;\theta_1)$ would not be the correct likelihood if there exist functional relationship among all parameters. To be more specific, assume that
$\theta_2 = g_1(\theta_1),\ldots,\theta_m = g_m(\theta_1)$, where $g_i$ are measurable functions of $\theta_1$. Furthermore, if all observations from different populations are independent of each other, the likelihood of $\theta_1$ involving all samples should be

$$\tilde{L}(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_m;\theta_1) \;=\; L_1(\boldsymbol{y}_1,\theta_i) \;\times\; \prod_{i=2}^{m} L_i(\boldsymbol{y}_i, g_i(\theta_1)).$$

We observe that all parameters, $\theta_2,\ldots,\theta_m$ disappeared from the above likelihood since they are replaced by $g_2(\theta_1),\ldots,g_m(\theta_m)$ respectively. We see that all data sets are integrated into one likelihood function that we call fusion likelihood.

When the functional relationship, $g_i$, are indeed known, we can derive the estimating equation by consider the following:

$$\begin{aligned}\frac{\log\tilde{L}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m;\theta_1)}{\partial\theta_1} &= \frac{\log L_1(\boldsymbol{y}_1;\theta_1)}{\partial\theta_1} + \sum_{i=2}^{m}\frac{\log L_i(\boldsymbol{y}_i;\theta_i)}{\partial\theta_i}\;\frac{g_i(\theta_1)}{\partial\theta_1}\\ &= \frac{\log L_1(\boldsymbol{y}_1;\theta_1)}{\partial\theta_1} + \sum_{i=2}^{m} w_i(\theta_1)\;\frac{\log L_i(\boldsymbol{y}_i;\theta_i)}{\partial\theta_i}.\end{aligned}$$

where $w_i(\theta_1) = \partial g_i(\theta_1)/\partial\theta_1, i = 2,\cdots,m$. By setting the above function to be zero, we obtain an estimating equation for $\theta_1$ which is based on all available data.

In practice, however, the functional form of $g_i$'s are not known. Therefore, the analytical forms of $g_i$'s are simply not available. When facing such a difficulty, one could abandon the data fusion process and proceed with $L_1$ by severing the connections among all parameters. Although this would avoid any possible bias, this approach would clearly lose some information. The loss of information should not be of great concern if the sample size from the target population is large enough to ensure a valid and effective inference. The benefit of integrating information from other sources is clearly negligible. When the sample from the target population is very small or moderate, the related data sets discarded as the result of avoiding bias could contain very valuable information.

The alternative is try to incorporate information from other sources in a meaningful way. This must be done in a reasonable and careful fashion without the knowledge of the link functions $g_i$, $i = 2, \cdots, m$. First, we must adopt a meaningful measure to evaluate the discrepancy among the probability density distributions when the parameters are known. Second, we must tackle the problem without the knowledge of the true values of the parameters. To resolve the first challenge, we propose to use some general measure for information discrepancy to characterize the connections among all parameters. Entropy or relative entropy has been widely used in information theory. For density functions, $g_1(x)$ and $g_2(x)$, with respect to a $\sigma-$finite measure $\nu$, the relative entropy, also called *Kullback-Leibler* divergence, is defined as:

$$KL(g_1, g_2) = E_1\left(\log \frac{g_1(X)}{g_2(X)}\right) = \int \log \frac{g_1(x)}{g_2(x)} g_1(x) d\nu(x). \tag{1}$$

In this expression, $\log(g_1(x)/g_2(x))$ is defined as $+\infty$, if $g_1(x) > 0$ and $g_2(x) = 0$. Therefore the expectation could be $+\infty$. Although $\log(g_1(x)/g_2(x))$ is defined as $-\infty$ when $g_1(x) = 0$ and $g_2(x) > 0$, the integrand, $\log(g_1(x)/g_2(x))g_1(x)$ is defined as zero in this case. It is widely used in information theory. Detailed discussions and application of the entropy in information theory can be found in [19]. Some theoretical properties of the entropy can be found in [14]. In particular, the relative entropy is not symmetric and therefore not a distance. The relative entropy, also known as Kullback-Leibler divergence, is also called the entropy loss. [20] introduce it as a performance criterion in estimating the multinormal variance-covariance matrix. [15] shows the entropy is a loss function in a Bayesian framework. [18] provides detailed discussions on the relative entropy including the connection between Fisher information and relative entropy. [13] has shown that the classical maximum likelihood principle can be considered a method of asymptotic realization of an optimum estimate with respect to the relative entropy.

Since the metric of the manifold that connects all the parameters can not be specified, we now define the neighborhood enveloping by [17]. For any fixed value of $\theta_1$, the parameter of the population of inferential interest, we can define a neighborhood by using the relative entropy as

$$N_{f_1}(\epsilon) = \cup_{\theta\in\Theta}\{f(\theta) : I(f(\theta), f_1) \leq \epsilon\}, \tag{2}$$

where $f_1 = f(x, \theta_1)$ and $\epsilon \geq 0$.

Assume that the density functions, $f_1, \ldots, f_m \in V$, are all assumed to be continuous where $V$ is a reflexive Banach space. Although $V$ can be quite arbitrary, we take $V = L^p = L^p(\Omega, \nu)$. It is known that the $L^p$ spaces ($1 < p < \infty$) are reflexive but that $L^1$ is not [see 16, for example]. For $i = 2, \ldots, m$, we define

$$\mathscr{E}_i = \{g \in L^p : ||g - f_i||_p < C_i, \int f_i(x) \log \frac{g(x)}{f_i(x)} d\nu(x) \leq a_i, \int g(x) d\nu(x) = 1, g(x) > 0\}, \tag{3}$$

where $a_i \geq 0$ and $C_i$, $i = 2, 3, \ldots, m$, are constants. We then define the interception of these individual enveloping neighborhood.

$$\mathscr{E} = \cap_{i=2}^m \mathscr{E}_i. \tag{4}$$

We remark that the set $\mathscr{E}$ will be bounded with respect to the $L^p$ norm and non-empty if the constraints are not too restrictive. The latter is assumed throughout.

Thus, for a given set of density functions, $f_1(x)$ being primary, we seek a probability density function $g \in \mathscr{E}$ which minimizes $I(f_1, g) = \int f_1(x) \log \frac{g(x)}{f_1(x)} d\nu(x)$ over all probability densities, $g$, satisfying

$$I(f_i, g) \leq a_i, \quad i = 2, \ldots, m, \tag{5}$$

where $a_i$, $i = 2, 3, \ldots, m$, are non-negative constants.

[6] showed that the the optimal solution takes the form of a *mixing distribution*:

$$g^* = \sum_{i=1}^{m} \lambda_i \, f_i(x), \tag{6}$$

where $g^*$ is the optimal solution for optimization probelm described above.

Assuming that $f_i$ are members of one parametric family while the difference is due to different value for the parameter, *i.e.* $f_i(x) = f(x; \theta_i)$. Without the knowledge of the exact values for $\theta_i$, [6] show that the weighted likelihood function:

$$\mathrm{WL}(\boldsymbol{y}; \theta_1) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(y_{ij}; \theta_1)^{\lambda_i}, \tag{7}$$

is the correct device to used by following the argument of [13]. The weights are functions of the $\lambda_i$'s which measures the discrepancies among all parameters.

We remark that this classical likelihood is embraced by the proposed likelihood by setting the weights to be zero for relevant samples. In addition, [5] showed that the estimator derived from the weighted likelihood function can be viewed as an approximate Bayes decision by following the same argument by [3]. The coefficients $a_i$'s, however, are not known since the exact functional form of the connections among all parameters are not assumed to be known. Therefore, the weights must be chosen adaptively to avoid introduction of significant bias into the inference for the target population.

## 2.2. Adaptive Weights Based on Likelihood Ratio

Since the weighted likelihood is the chosen instrument for the data fusion process, the likelihood weights play a central role in integrating all relevant information in a meaningful way. As we have seen in previous section, the likelihood weights should be chosen to reflect the true relevance between a relevant sample and the target population. Without any prior knowledge, the evidence expressed in the relevant sample should be the only source for evaluation and determination of a set of appropriate weights. The fundamental question is how to measure the discrepancy between to two populations that could be related.

In a likelihood ratio context, a likelihood-ratio test is a statistical test for making a decision between two hypotheses based on the value of this ratio.

Since the parameters of these two populations are not known, therefore we plug in the MLE estimates. When the ratio measures the resemblance between the two values. similarly

to the hypothesis test, it implies that the two parameter values are too different in the context of likelihood ratio. Following the same line of argument, we propose to use the likelihood ratio as a measure of discrepancy. Since the true values of the parameters are not known, it is natural to use estimated likelihood ratio with the parameter values replaced by corresponding MLE estimates.

Given a parametric density function $f_1$ and a sample $\boldsymbol{Y}_1 = (Y_{11}, Y_{12}, \ldots, Y_{1n_1})$, the classical likelihood ratio for testing a null hypothesis is based on the following

$$LR = \frac{\sup_{\Omega_H} \prod_{j=1}^{n_1} f_1(Y_{1j}; \theta_1)}{\sup_{\Omega} \prod_{j=1}^{n_1} f_1(Y_{1j}; \theta_1)},$$

where $\Omega_H$ is a subspace in the parameter space under the null hypothesis $H$.

Assume that the likelihood function attains an unique maximum in $\Omega_H$ and $\Omega$ respectively. Denote the maximizers as $\widehat{\theta}_1^H$ and $\widehat{\theta}_1$ respectively. The likelihood ratio (LR) can then be rewritten as

$$LR = \frac{\prod_{j=1}^{n_1} f_1(Y_{1j}; \widehat{\theta}_1^H)}{\prod_{j=1}^{n_1} f_1(Y_{1j}; \widehat{\theta}_1)}.$$

Therefore, the validity of the proposed hypothesis is evaluated according to the likelihood ratio.

By using a similar idea, we propose to choose likelihood weights by using the estimated pairwise likelihood ratio. For simplicity, let $m = 2$, *i.e.*, there is only one related population together with the target population. Let $\widehat{\theta}_2$ denote the estimate derived from the second sample. We consider the estimated pairwise likelihood ratio (PLR) as follows:

$$\gamma_i = PLR_i = \frac{\prod_{j=1}^{n_1} f_1(Y_{1j}; \widehat{\theta}_i)}{\prod_{j=1}^{n_1} f_1(Y_{1j}; \widehat{\theta}_1)}, \quad i = 1, 2.$$

It then follows that $\gamma_1$ equals to 1. Furthermore, we have $\gamma_i \leq 1$, $i = 2, \ldots, m$ by the definition of the MLE.

The likelihood weights $\lambda_1$ and $\lambda_2$ will then be chosen as

$$\lambda_1 = \frac{1}{1+\gamma_2}, \quad \lambda_2 = \frac{\gamma_2}{1+\gamma_2}.$$

For the simple linear models presented before, a simplification yields that the WLE of $\beta_1$ takes the following form:

$$\widehat{\beta}_1^{\text{WLE}} = w_1 \widehat{\beta}_1^{\text{MLE}} + w_2 \widehat{\beta}_2^{\text{MLE}},$$

where

$$w_1 = \frac{\sum_{j=1}^{n_1} x_{1j}^2}{\sum_{j=1}^{n_1} x_{1j}^2 + \gamma_2 \sum_{k=1}^{n_2} x_{2k}^2} \text{ and } w_2 = \frac{\gamma_2 \sum_{k=1}^{n_2} x_{2k}^2}{\sum_{j=1}^{n_1} x_{1j}^2 + \gamma_2 \sum_{k=1}^{n_2} x_{2k}^2}.$$

To further illustrate the behavior of the proposed likelihood weights, we revisit the simple linear models given by Equation (1). It then follows that

$$\begin{aligned}\gamma_2 &= \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n_1}(y_{1j}-\widehat{\beta}_2^{\text{MLE}}x_{1j})^2 + \frac{1}{2\sigma_1^2}\sum_{j=1}^{n_1}(y_{1j}-\widehat{\beta}_1^{\text{MLE}}x_{1j})^2\right\} \\ &= \exp\left\{-(\widehat{\beta}_1^{\text{MLE}}-\widehat{\beta}_2^{\text{MLE}})^2\sum_{j=1}^{n_1}x_{1j}^2/(2\sigma_1^2)\right\}.\end{aligned}$$

Since $\lambda_2 = 1/(1+\gamma_2)$, we then have

$$\lambda_2 = \frac{1}{1+\exp\left\{(\widehat{\beta}_1^{\text{MLE}}-\widehat{\beta}_2^{\text{MLE}})^2\sum_{j=1}^{n_1}x_{1j}^2/(2\sigma_1^2)\right\}}.$$

Thus the proposed likelihood weight for the second sample is a function of the estimated difference between the regression coefficients, the variance of the error terms and the "variance" of the covariate. For a fixed design matrix, a large difference between $\widehat{\beta}_1^{\text{MLE}}$ and $\widehat{\beta}_2^{\text{MLE}}$ will reduce the assigned importance to the second sample. On the other hand, a large variance in the first model would result in a bigger value for the weight assigned to the second sample.

## 2.3. Estimation with Fixed Likelihood Weights

In this section we develop the WL procedure for the linear models as an illustration. For simplicity, we assume that $m = 2$ in Sections 2.1 and 2.2. Extension to the case of $m > 2$ is straightforward. Assume that data $\{x_{1j}, y_{1j}; j = 1,\ldots,n_1\}$ and $\{x_{2k}, y_{2k}; k = 1,\ldots,n_2\}$ are generated from the following models:

$$\begin{aligned} y_{1j} &= \beta_1 x_{1j} + \epsilon_j, \quad \epsilon_j \sim N(0,\sigma_1^2), \quad j = 1,\ldots,n_1, \\ y_{2k} &= \beta_2 x_{2j} + e_k; \quad e_k \sim N(0,\sigma_2^2), \quad k = 1,\ldots,n_2, \end{aligned} \tag{8}$$

where the $\{x_{1j}, j = 1,\ldots,n_1; x_{2j}, k = 1,\ldots,n_2\}$ are fixed. Assume that $\{\epsilon_j\}$'s and $e_k$'s are *i.i.d.* respectively. The error terms from these two models are assumed to be independent as well. For the purpose of our demonstration we assume $\sigma_1$ and $\sigma_2$ are known although that would rarely be the case in practice. Denote $\boldsymbol{y}_1 = (y_{11}, y_{12}, \ldots, y_{1n_1})^{\text{T}}$, and $\boldsymbol{y}_2 = (y_{21}, y_{22}, \ldots, y_{2n_2})^{\text{T}}$.

Suppose that the parameter, $\beta_1$, is of primary inferential interest. The parameter $\beta_2$ is thought or expected to be not too different from $\beta_1$ due to the "similarity" of the two experiments. Note that a bivariate distribution is not assumed in the above model.

Assuming marginal normality, the marginal likelihoods for $\beta_1$ and $\beta_2$ are

$$L_1(\boldsymbol{y}_1;\beta_1) \propto \prod_{j=1}^{n_1} \exp\left\{-\frac{(y_{1j}-\beta_1 x_{1j})^2}{2\,\sigma_1^2}\right\},$$

and

$$L_2(\boldsymbol{y}_2;\beta_2) \propto \prod_{k=1}^{n_2} \exp\left\{-\frac{(y_{2k}-\beta_2 x_{2k})^2}{2\,\sigma_2^2}\right\}.$$

A direct calculation deduces the MLE of $\beta_1$ and $\beta_2$ as follows

$$\widehat{\beta}_1^{\text{MLE}} = \left(\sum_{j=1}^{n_1} x_{1j}^2\right)^{-1} \sum_{j=1}^{n_1} x_{1j}y_{1j}, \text{ and } \widehat{\beta}_2^{\text{MLE}} = \left(\sum_{k=1}^{n_2} x_{2k}^2\right)^{-1} \sum_{k=1}^{n_2} x_{2k}y_{2k}.$$

If one knows that $\beta_1$ and $\beta_2$ are similar to each other according to past studies or expert opinions, then it is reasonable to expect that this information might be used to yield a better estimate of the parameter $\beta_1$.

The WL for inference about $\beta_1$ can be represented as:

$$\text{WL}(\beta_1;\boldsymbol{y}_1,\boldsymbol{y}_2) = L_1^{\lambda_1}(\boldsymbol{y}_1;\beta_1)L_2^{\lambda_2}(\boldsymbol{y}_2;\beta_1), \tag{9}$$

where $\lambda_1$ and $\lambda_2$ are weights selected according to the relevance of the likelihood to which they attached. The non-negative requirement for the weights is not assumed in the formulation although the optimum weights should be non-negative according to the assertion by [6].

It is worth noting that $L_2(\boldsymbol{y}_2;\beta_1)$ instead of $L_2(\boldsymbol{y}_2;\beta_2)$ is used to define WL($\beta_1$) since $\beta_1$ is of our primary inferential interest at this stage and the marginal distributions of the elements of the $\boldsymbol{Y}_2$ are thought to resemble the marginal distributions of those of the $\boldsymbol{Y}_1$. Note that the WL for $\theta_1$ depends on the distribution of the $\boldsymbol{Y}_1$. However, it does not depend on the distribution of $\boldsymbol{Y}_2$. Notice that the joint distribution of the $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ does not appear in the formulation of the WL for $\theta_1$ and no assumptions are made about it.

The WLE of $\beta_1$ is obtained by maximizing the weighted likelihood function for given weights $\lambda_1$ and $\lambda_2$. It follows from (9) that

$$\log\{\text{WL}(\beta_1)\} = \lambda_1 \log\{L_1(\boldsymbol{y}_1;\beta_1)\} + \lambda_2 \log\{L_2(\boldsymbol{y}_2;\beta_1)\}.$$

Note that

$$\frac{\partial \log\{\text{WL}(\beta_1)\}}{\partial \beta_1} = \frac{\lambda_1}{2\sigma_1^2}\sum_{j=1}^{n_1} x_{1j}(y_{1j}-\beta_1 x_{1j}) + \frac{\lambda_2}{2\sigma_1^2}\sum_{k=1}^{n_2} x_{2k}(y_{2k}-\beta_1 x_{2k}).$$

A simplification yields the WLE of $\beta_1$ as follows

$$\widehat{\beta}_1^{\text{WLE}} = w_1\widehat{\beta}_1^{\text{MLE}} + w_2\widehat{\beta}_2^{\text{MLE}}. \tag{10}$$

where

$$w_1 = \frac{\lambda_1 \sum_{j=1}^{n_1} x_{1j}^2}{\lambda_1 \sum_{j=1}^{n_1} x_{1j}^2 + \lambda_2 \sum_{k=1}^{n_2} x_{2k}^2} \text{ and } w_2 = \frac{\lambda_2 \sum_{k=1}^{n_2} x_{2k}^2}{\lambda_1 \sum_{j=1}^{n_1} x_{1j}^2 + \lambda_2 \sum_{k=1}^{n_2} x_{2k}^2}.$$

It can be seen that the WLE of $\beta_1$ is a linear combination of $\widehat{\beta}_1^{\text{MLE}}$ and $\widehat{\beta}_2^{\text{MLE}}$, under the over-simplified model. The WLE of $\beta_1$ coincides with the MLE of $\beta_1$ obtained from the first if the weight for the second marginal likelihood function is set to be zero. Therefore, the weights $w_1$ and $w_2$ reflect the importance of $\widehat{\beta}_1^{\text{MLE}}$ and $\widehat{\beta}_2^{\text{MLE}}$.

Let $\beta_1^0$ and $\beta_2^0$ be the true values of the parameters $\beta_1$ and $\beta_2$ respectively. If one knows that $|\beta_1^0 - \beta_2^0| \leq C$, where $C$ is a known constant according to past studies or expert opinions, we then have the following theorem.

**Theorem 1.** *Under the normality assumption with known variances, the WLE $\tilde{\beta}_1^{\text{WLE}}$ takes the form*

$$\tilde{\beta}_1^{\text{WLE}}(w_1, w_2) = w_1\widehat{\beta}_1^{\text{MLE}} + w_2\widehat{\beta}_2^{\text{MLE}},$$

*where $w_1 + w_2 = 1, 0 < w_1 \leq 1, 0 \leq w_2 < 1$. Furthermore, $|\beta_1^0 - \beta_2^0| \leq C$, $C > 0$, then $|E(\tilde{\beta}_1 - \beta_1^0)| \leq w_2\, C$. In addition,*

$$\max_{w_1,w_2} MSE(\tilde{\beta}_1^{\text{WLE}}) < MSE(\widehat{\beta}_1^{\text{MLE}}) \text{ if and only if } \frac{M}{M+1} < w_1 < 1,$$

*where*

$$M = \frac{C^2 + \sigma_2^2\left(\sum_{k=1}^{n_2} x_{2k}^2\right)^{-1} - \sigma_1^2\left(\sum_{j=1}^{n_1} x_{1j}^2\right)^{-1}}{2\sigma_1^2\left(\sum_{j=1}^{n_1} x_{1j}^2\right)^{-1}}.$$

*In addition, $Var(\tilde{\beta}_1^{\text{WLE}}) < Var(\widehat{\beta}_1^{\text{MLE}})$ if $\max_{w_1,w_2} MSE(\tilde{\beta}_1^{\text{WLE}}) < MSE(\widehat{\beta}_1^{\text{MLE}})$.*

*Thus, for small value of C, the WLE could have smaller MSE and variance simultaneously with the cost of small bias. The magnitude of the bias is controlled by the weight assigned to the second population. The lower bound for the weight $w_1$ is connected with the worst MSE of the WLE instead of the optimal one.*

However, this theorem does not provide any guidance on how to choose the optimum weights since the bound $C$ is rarely known in practice. Even though the exact value of the bound is known or can be estimated fairly accurately, this theorem still could not be used to choose the best set of weights. For example, suppose that there are two independent experiments with exactly the same coefficients and design matrices. Therefore, we should simply merge these two experiments. However, the above theorem merely suggests that the lower bound is zero for $w_1$ since $M$ is zero in this case.

## 3. Asymptotic Properties

In this section we investigate the asymptotic properties of the WLE in the framework of generalized linear models (GLM) for simple presentation. The GLMs have the following structure. The sample $Y = (Y_1, \ldots, Y_n) \in \Re^n$ has independent components and $Y_i$ has the p.d.f.

$$\exp\left\{\frac{\eta_i y_i - \zeta(\eta_i)}{\phi_i}\right\} h(y_i, \phi_i), \ \ldots, n,$$

w.r.t. a $\sigma$-finite measure $\nu$, where $\eta_i$ and $\phi_i$ are unknown, $\phi_i > 0$,

$$\eta_i \in \Xi = \{\eta : 0 < \int h(y, \phi) e^{\eta y/\phi} d\nu(y) < \infty\} \subset \Re$$

for all $i$, $\zeta$ and $h$ are known functions, and $\zeta''(\eta) > 0$ is assumed for all $\eta \in \Xi^\circ$, the inferior of $\Xi$. Note that the p.d.f. belongs to an exponential family if $\phi$ is known. As a consequence, for $i = 1, \ldots, n$,

$$\begin{aligned} E(Y_i) &= \zeta'(\eta_i) \stackrel{\triangle}{=} \mu(\eta_i), \\ Var(Y_i) &= \phi \zeta''(\eta_i). \end{aligned}$$

It is assumed that $\eta_i$ is related to $X_i$, the $i$th value of a $p$-vector covariates, through

$$g(\mu(\eta_i)) = \boldsymbol{\beta}^T X_i, \ i = 1, \ldots, n.$$

In GLM, $\boldsymbol{\beta}$ is the parameter of interest and $\phi_i$'s are considered to be nuisance parameters. The range of $\boldsymbol{\beta}$ is assumed to be $\mathscr{B} = \{\boldsymbol{\beta} : (g \circ \mu)^{-1}(\boldsymbol{\beta}^T x) \in \Xi^\circ \forall x \in \mathscr{X}\}$, where $\mathscr{X}$ is the range of $X_i$'s. It is often assumed that

$$\phi_i = \phi / t_i, \ i = 1, \ldots, n,$$

with unknown $\phi > 0$ and known positive $t_i$'s.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$ and $\psi = (g \circ \mu)^{-1}$, then the log-likelihood function is

$$\log l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\log h(y_i, \phi/t_i) + \frac{\psi(\boldsymbol{\beta}^T X_i) y_i - \zeta(\psi(\boldsymbol{\beta}^T X_i))}{\phi/t_i}\right]$$

and the score function for $\boldsymbol{\beta}$ is

$$s_n(\boldsymbol{\beta}) = \frac{\partial \log l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \frac{1}{\phi} \sum_{i=1}^{n} \{[y_i - \mu(\psi(\boldsymbol{\beta}^T X_i))] \psi'(\boldsymbol{\beta}^T X_i) t_i X_i\}.$$

Let

$$M_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} [\psi'(\boldsymbol{\beta}^T X_i)]^2 \zeta''(\psi(\boldsymbol{\beta}^T X_i)) t_i X_i X_i^T.$$

Then the Fisher information evaluated at $\boldsymbol{\beta}$ is

$$I_n(\boldsymbol{\beta}) = Var\left(\frac{\partial \log l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) = M_n(\boldsymbol{\beta})/\phi.$$

We now present the asymptotic properties for the relevant weighted likelihood estimator and the hypothesis testing based on this estimator. We use double index as the subscript of observations and establish the following notations.

- $\beta_0$: The true parameter value of main group.
- $y_{1j}, j = 1, \ldots, n_1$ are the observations from main group.
- $y_{2j}, j = 1, \ldots, n_2$ are the observations from contamination group.
- $wl(\beta) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} w_{ij} \log f(y_{ij} \mid x_i, \beta) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} w_{ij} l_{ij}(\beta)$.
- $l(\beta) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} \log f(y_{ij} \mid x_i, \beta) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} l_{ij}(\beta)$.

Besides the regularity conditions and assumptions given in [1], we need the following additional assumptions for the present problem.

**(A1)** The information matrix of main group $I_{n_1}(\beta_0)$ is positive definite, and the eigenvalues of $n_1^{-r} I_{n_1}(\beta_0)$ are bounded, for some $r > 0$, so are the eigenvalues of $n_1^r I_{n_1}^{-1}(\beta_0)$. That is, there exist constants $c_1$ and $c_2$, such that

$$c_1 \le \lambda_{\min} n_1^r I_{n_1}^{-1}(\beta_0) \le \lambda_{\max} n_1^r I_{n_1}^{-1}(\beta_0) \le c_2.$$

**(A2)** There exists a neighborhood of $\beta_0$, $O$, such that for all $\beta \in O$ and $y_{2j}$, $j = 1, \ldots, n_2$,

$$E\left\{\sup_{\beta \in O, 1 \le j \le n_2} \left|\log \frac{f(y_{2j}, \beta)}{f(y_{2j}, \beta_0)}\right|\right\} \le K,$$

where $K$ is a constant.

**(A3)** The proportion of contamination converges to 0 in the order of $n^{-1/2}$, i.e., $\epsilon_n = o(n^{-1/2})$.

**(A4)** The criteria to choose weights guarantees that the weights are consistent. Assume the first $n_1$ observations are from main group and the rest are from contamination group. Then as $n \to \infty$ and $\epsilon_n \to 0$,

$$\boldsymbol{w} = (w_1, \ldots, w_n)^T \to \boldsymbol{v} = (v_1, \ldots, v_n) = (\underbrace{1, \ldots, 1}_{n_1}, \underbrace{0, \ldots, 0}_{n_2})^T.$$

Furthermore, $\max_i |w_i - v_i| = o(n^{-1/2})$.

We now provide the asymptotic results in Theorems 2-3 below.

**Theorem 2.** *Under the regularity conditions specified in [1] and assumptions (A1)–(A4), there is a sequence of estimators* $\{\hat{\beta}_n\}$ *such that*

$$P(s_n(\hat{\beta}_n) = 0) \to 1$$

*and* $\hat{\beta}_n \to \beta_0$ *in probability.*

*Proof.* In order to present the proof clearly, we use double index as the subscript of observations. Then we need to specify the following notations.

- $\boldsymbol{\beta}_0$: The true parameter of main group.
- $\boldsymbol{\gamma}_0$: The true parameter of second group.
- $y_{1j}, j = 1, \ldots, n_1$ are the observations from main group.
- $y_{2j}, j = 1, \ldots, n_2$ are the observations from second group.
- $wl(\boldsymbol{\beta}) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} w_{ij} \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{\beta}) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} w_{ij} l_{ij}(\boldsymbol{\beta})$.
- $l(\boldsymbol{\beta}) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{\beta}) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} l_{ij}(\boldsymbol{\beta})$.

Since there are two populations, we distinguish the score functions, Fisher information, design matrices in the following way. For main group, we denote them as $s_{n_1}(\boldsymbol{\beta})$, $I_{n_1}(\boldsymbol{\beta})$ and $X$; and for contamination group, we have $s_{n_2}(\boldsymbol{\beta})$, $I_{n_2}(\boldsymbol{\beta})$ and $Z$. For simplicity, We use $M_{n_1}$ to denote $M_{n_1}(\boldsymbol{\beta}_0)$.

Define the neighborhood of $\boldsymbol{\beta}_0$, $N_{n_1}(\delta)$, $\delta > 0$, as

$$N_{n_1}(\delta) = \{\boldsymbol{\beta} : \ \|I_{n_1}(\boldsymbol{\beta}_0)^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\| \le \delta)\},$$

where $I_{n_1}(\boldsymbol{\beta}_0)$ is the Fisher information of the main group data evaluated at true value $\boldsymbol{\beta}_0$. Let $\partial N_{n_1}(\delta)$ be the boundary of $N_{n_1}(\delta)$. To prove $\boldsymbol{\beta} \xrightarrow{p} \boldsymbol{\beta}_0$, it is sufficient to show for any $\eta > 0$, there exist $\delta > 0$ and $N > 0$, such that for $n \ge N$ and all $\boldsymbol{\beta} \in \partial N_{n_1}(\delta)$, we have

$$P(wl(\boldsymbol{\beta}) - wl(\boldsymbol{\beta}_0) < 0 \text{ for all } \boldsymbol{\beta} \in \partial N_{n_1}(\delta)) > 1 - \eta.$$

We start the proof by partitioning $wl(\boldsymbol{\beta}) - wl(\boldsymbol{\beta}_0)$.

$$\begin{aligned} &wl(\boldsymbol{\beta}) - wl(\boldsymbol{\beta}_0) \\ &= \sum_{j=1}^{n_1}\left[l_{1j}(\boldsymbol{\beta}) - l_{1j}(\boldsymbol{\beta}_0)\right] - \sum_{j=1}^{n_1}(1 - w_{1j})\left[l_{1j}(\boldsymbol{\beta}) - l_{1j}(\boldsymbol{\beta}_0)\right] + \sum_{j=1}^{n_2} w_{2j}\left[l_{2j}(\boldsymbol{\beta}) - l_{2j}(\boldsymbol{\beta}_0)\right] \\ &\triangleq A - B + C. \end{aligned}$$

In [1], it is shown that there exist $\delta > 0$ and $N$ such that for $n \ge N$ and all $\boldsymbol{\beta} \in \partial N_{n_1}(\delta)$,

$$P(A < 0) > 1 - \eta.$$

Then we need to show that $B$ and $C$ are dominated by $A$, which will follows from $B$ and $C$ are $o_p(1)$ since it is shown that $A$ is $O_p(1)$ in [1].

In the following, we show that $B$ is dominated by $A$. Define $s_{n_1}^{1-w}(\boldsymbol{\beta})$ as

$$\begin{aligned} s_{n_1}^{1-w}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{j=1}^{n_1} (1-w_{1j}) l_{1j}(\boldsymbol{\beta}) \\ &= \left( \frac{\partial \log f(y_{11}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \ldots, \frac{\partial \log f(y_{1n_1}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \begin{pmatrix} 1-w_{11} \\ \vdots \\ 1-w_{1n_1} \end{pmatrix} \\ &\triangleq \left( \boldsymbol{U}_1, \ldots, \boldsymbol{U}_{n_1} \right) \boldsymbol{W}_1 \\ &\triangleq \boldsymbol{S}_1 \boldsymbol{W}_1. \end{aligned}$$

Let $s_{n_1}(\boldsymbol{\beta})$ be the score function for the main group, which can be similarly written as

$$s_{n_1}(\boldsymbol{\beta}) = \boldsymbol{S}_1 \mathbf{1}_{n_1},$$

where $\mathbf{1}_n$ denotes the length-$n$ vector with all element equal to 1. Then by taking Taylor expansion of $B$, we have

$$B = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0) + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \nabla s_{n_1}^{1-w}(\boldsymbol{\gamma}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

where $\boldsymbol{\gamma}^* \in N_{n_1}(\delta)$. In the following, we will show that both these two terms are $o_p(1)$.

To show $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0) = o_p(1)$, we need to show for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P(\|(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0)\| < \varepsilon) = 1$$

By Markov inequality, we have

$$P(\|(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0)\| < \varepsilon) \geq 1 - (\frac{1}{\varepsilon})^2 E\|(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0)\|^2$$

and so it is sufficient to show $E\|(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0)\|^2 \to 0$.

Let $\lambda = I_{n_1}(\boldsymbol{\beta}_0)^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)/\delta$, then $\|\lambda\| = 1$ and

$$\begin{aligned} &E\|(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0)\|^2 = E\|(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \boldsymbol{S}_1 \boldsymbol{W}_1\|^2 \\ &= E\|\delta \lambda^T I_{n_1}^{-1/2}(\boldsymbol{\beta}_0) \boldsymbol{S}_1 \boldsymbol{W}_1\|^2 \leq \delta^2 \|\boldsymbol{W}_1\|^2 E\|I_{n_1}^{-1/2}(\boldsymbol{\beta}_0) \boldsymbol{S}_1\|^2 \\ &= \delta^2 \|\boldsymbol{W}_1\|^2 E\left\{ tr[\boldsymbol{S}_1^T I_{n_1}^{-1}(\boldsymbol{\beta}_0) \boldsymbol{S}_1] \right\} = \delta^2 \|\boldsymbol{W}_1\|^2 E\left\{ \sum_{j=1}^{n_1} \boldsymbol{U}_j^T I_{n_1}^{-1}(\boldsymbol{\beta}_0) \boldsymbol{U}_j \right\}. \end{aligned}$$

Because the eigenvalues of $I_{n_1}(\boldsymbol{\beta}_0)/n_1^r$ are bounded by assumption (*A*4), so are eigenvalues of $n_1^r I_{n_1}^{-1}(\boldsymbol{\beta}_0)$. That is

$$c_1 \leq \lambda_{\min}(n_1^r I_{n_1}^{-1}(\boldsymbol{\beta}_0)) \leq \lambda_{\max}(n_1^r I_{n_1}^{-1}(\boldsymbol{\beta}_0)) \leq c_2,$$

where $c_1$ and $c_2$ are constants. Because $E\boldsymbol{U}_j^T\boldsymbol{U}_k = 0$, let $I_p$ be the $p \times p$ identity matrix, we have

$$\begin{aligned}
E\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0)\|^2 &\leq \delta^2\|\boldsymbol{W}_1\|^2\frac{1}{n_1}E\left\{\sum_{j=1}^{n_1}\boldsymbol{U}_j^T(n_1^r I_{n_1}^{-1}(\boldsymbol{\beta}_0))\boldsymbol{U}_j\right\}\\
&\leq \delta^2\|\boldsymbol{W}_1\|^2\frac{1}{n_1}E\left\{\sum_{j=1}^{n_1}\boldsymbol{U}_j^T[c_2 I_p]\boldsymbol{U}_j\right\}\\
&\leq \delta^2\|\boldsymbol{W}_1\|^2\frac{c_2}{n_1^r c_1}E\left\{s_{n_1}(\boldsymbol{\beta}_0)^T[n_1^r I_{n_1}^{-1}(\boldsymbol{\beta}_0)]s_{n_1}(\boldsymbol{\beta}_0)\right\}\\
&\leq \delta^2\|\boldsymbol{W}_1\|^2\frac{c_2}{n_1^r c_1}Etr\left\{s_{n_1}(\boldsymbol{\beta}_0)^T[n_1^r I_{n_1}^{-1}(\boldsymbol{\beta}_0)]s_{n_1}(\boldsymbol{\beta}_0)\right\} = \delta^2\|\boldsymbol{W}_1\|^2\frac{c_2 p}{c_1}.
\end{aligned}$$

Because $\max_i(1-w_i) = o(n^{-1/2})$, $\|W_1\|^2 = o(1)$. Then $\delta^2\|\boldsymbol{W}_1\|^2\frac{c_2 p}{c_1} = o(1)$. So $E\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T s_{n_1}^{1-w}(\boldsymbol{\beta}_0)\|^2 \to 0$. This proves that the first term is $o_p(1)$.

In order to show the second term is $o_p(1)$, we need to show $\|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\nabla s_{n_1}^{1-w}(\boldsymbol{\gamma}^*)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|$ is $o_p(1)$. Since

$$(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\nabla s_{n_1}^{1-w}(\boldsymbol{\gamma})(\boldsymbol{\beta}-\boldsymbol{\beta}_0) = \delta^2\lambda^T I_{n_1}^{-1/2}(\boldsymbol{\beta}_0)\nabla s_{n_1}^{1-w}(\boldsymbol{\gamma})I_{n_1}^{-1/2}(\boldsymbol{\beta}_0)\lambda,$$

it is sufficient to show

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta)}\|I_{n_1}^{-1/2}(\boldsymbol{\beta}_0)\nabla s_{n_1}^{1-w}(\boldsymbol{\gamma})I_{n_1}^{-1/2}(\boldsymbol{\beta}_0)\| \to 0.$$

Let

$$\begin{aligned}
R_{n_1}^w(\boldsymbol{\beta}) &= \sum_{j=1}^{n_1}(1-w_{1j})[y_{1j}-\mu(\psi(\boldsymbol{\beta}^T X_j))]\psi''(\boldsymbol{\beta}^T X_j)t_j X_j X_j^T,\\
M_{n_1}^w(\boldsymbol{\beta}) &= \sum_{j=1}^{n_1}(1-w_{1j})[\psi'(\boldsymbol{\beta}^T X_j)]^2\zeta''(\psi(\boldsymbol{\beta}^T X_j))t_i X_j X_j^T.
\end{aligned}$$

Then

$$\nabla s_{n_1}^{1-w}(\boldsymbol{\gamma}) = [R_{n_1}^w(\boldsymbol{\gamma}) - M_{n_1}^w(\boldsymbol{\gamma})]/\phi.$$

So it suffices to show

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta)}\|M_{n_1}^{-1/2}M_{n_1}^w(\boldsymbol{\gamma})M_{n_1}^{-1/2}\| \to 0,$$

and

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta)} \|M_{n_1}^{-1/2} R_{n_1}^w(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\| \to 0.$$

Because

$$\|M_{n_1}^{-1/2} M_{n_1}^w(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\| \le \max_j (1-w_{1j}) \|M_{n_1}^{-1/2} M_{n_1}(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\|.$$

Using similar argument as in [1], $\|M_{n_1}^{-1/2} M_{n_1}(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\|$ is bounded by

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta)} \|M_{n_1}^{-1/2} M_{n_1}(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\| \le \sqrt{p} \max_{\boldsymbol{\gamma}\in N_{n_1}(\delta), j\le n_1} |\varphi(\boldsymbol{\gamma}^T X_j)/\varphi(\boldsymbol{\beta}^T X_j)|,$$

which converges to 0 since $\varphi$ is continuous and, for $\boldsymbol{\gamma} \in N_{n_1}(\delta)$, $|\boldsymbol{\gamma}^T X_j - \boldsymbol{\beta}^T X_j|^2 \to 0$. Because $\max_j (1-w_{1j}) \to 0$, we have

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta)} \|M_{n_1}^{-1/2} M_{n_1}^w(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\| \to 0.$$

Let

$$\begin{aligned}
e_j &= y_{1j} - \mu(\psi(\boldsymbol{\beta}^T X_j)),\\
U_{n_1}^w(\boldsymbol{\gamma}) &= \sum_{j=1}^{n_1} (1-w_{1j})[\mu(\psi(\boldsymbol{\beta}^T X_j)) - \mu(\psi(\boldsymbol{\gamma}^T X_j))]\psi''(\boldsymbol{\beta}^T X_j) t_j X_j X_j^T,\\
V_{n_1}^w(\boldsymbol{\gamma}) &= \sum_{j=1}^{n_1} (1-w_{1j}) e_j [\psi''(\boldsymbol{\gamma}^T X_j) - \psi''(\boldsymbol{\beta}^T X_j)] t_i X_j X_j^T,\\
W_{n_1}^w(\boldsymbol{\beta}) &= \sum_{j=1}^{n_1} (1-w_{1j}) e_j \psi''(\boldsymbol{\beta}^T X_j)] t_i X_j X_j^T.
\end{aligned}$$

Then $R_{n_1}^w(\boldsymbol{\gamma}) = U_{n_1}^w(\boldsymbol{\gamma}) + V_{n_1}^w(\boldsymbol{\gamma}) + W_{n_1}^w(\boldsymbol{\beta})$. Using a similar argument as above, we can show that

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta)} \|M_{n_1}^{-1/2} U_{n_1}^w(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\| \to 0.$$

Note that $\|M_{n_1}^{-1/2} U_{n_1}^w(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\|$ is bounded by the product of

$$\max_j (1-w_{1j}) M_{n_1}^{-1/2} \sum_{j=1}^{n_1} |e_j| t_j X_j X_j^T M_{n_1}^{-1/2} = o_p(1)$$

and

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta), j\le n_1} |\psi''(\boldsymbol{\gamma}^T X_j) - \psi''(\boldsymbol{\beta}^T X_j)|,$$

which can be shown to be $o(1)$ using the same argument. Hence,

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta)} \|M_{n_1}^{-1/2} V_{n_1}^w(\boldsymbol{\gamma}) M_{n_1}^{-1/2}\| \to 0.$$

Finally we need to show

$$\max_{\boldsymbol{\gamma}\in N_{n_1}(\delta)} \|M_{n_1}^{-1/2} W_{n_1}^{w}(\boldsymbol{\beta}) M_{n_1}^{-1/2}\| \to 0.$$

Since $E(e_j)=0$ and $e_j$'s are independent, it suffices to show that

$$\sum_{j=1}^{n_1} E|(1-w_{1j})e_j\psi''(\boldsymbol{\beta}^T X_j)t_i X_j^T M_{n_1}^{-1} X_j|^{1+\tau} \to 0$$

for some $\tau\in(0,1)$. It is trivial since $\max_j(1-w_{1j})\to 0$ and it is shown in [1] that

$$\sum_{j=1}^{n_1} E|e_j\psi''(\boldsymbol{\beta}^T X_j)t_i X_j^T M_{n_1}^{-1} X_j|^{1+\tau} \to 0.$$

Now we have shown that the first term and the remainder of the Taylor expansion of $B$ are both $o_p(1)$. So $B=o_p(1)$.

In the following we show that $C$ is dominated by $A$. By assumption (A5), we have

$$\begin{aligned} E|C| &= E\left|\sum_{j=1}^{n_2} w_{2j}\left[l_{2j}(\boldsymbol{\beta})-l_{2j}(\boldsymbol{\beta}_0)\right]\right| \\ &\le \sum_{j=1}^{n_2} w_{2j} E\left|\log\frac{f(y_{2j},\boldsymbol{\beta})}{f(y_{2j},\boldsymbol{\beta}_0)}\right| \\ &\le K\sum_{j=1}^{n_2} w_{2j} \\ &= K\cdot n_2 o(n^{-1/2}) \\ &\to 0. \end{aligned}$$

This implies $C\to_p 0$.

**Theorem 3.** *Under the assumptions of Theorem 2,*

$$I_{n_1}^{1/2}(\beta_0)(\hat{\beta}-\beta_0)\to N_p(0,I_p)$$

*in distribution.*

*Proof.* Let

$$s_n^w(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}\sum_{i=1}^{2}\sum_{j=1}^{n_i} w_{ij}l_{ij}(\boldsymbol{\beta})$$

and $\nabla s_n^w(\boldsymbol{\beta}) = \dfrac{\partial}{\partial\boldsymbol{\beta}} s_n^w(\boldsymbol{\beta})$. Then

$$\nabla s_n^w(\boldsymbol{\beta}) = \nabla s_{n_1}^w(\boldsymbol{\beta}) + \nabla s_{n_2}^w(\boldsymbol{\beta}),$$

where $\nabla s^w_{n_1}(\boldsymbol{\beta})$ and $\nabla s^w_{n_2}(\boldsymbol{\beta})$ are similarly defined on main group and contamination group,

$$\begin{aligned}\nabla s^w_{n_1}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{j=1}^{n_1} w_{1j} l_{1j}(\boldsymbol{\beta}),\\ \nabla s^w_{n_2}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{j=1}^{n_2} w_{2j} l_{2j}(\boldsymbol{\beta}).\end{aligned}$$

Since $C = o_p(1)$, we know that

$$\nabla s^w_{n_2}(\boldsymbol{\beta}) \to 0.$$

By the proof of $B = o_p(1)$, we have

$$\max_{\boldsymbol{\gamma} \in N_{n_1}(\delta)} \| I^{-1/2}_{n_1}(\boldsymbol{\beta}_0) \nabla s^{1-w}_{n_1}(\boldsymbol{\gamma}) I^{-1/2}_{n_1}(\boldsymbol{\beta}_0) \| \to 0,$$

so, as $\nabla s_{n_1}(\boldsymbol{\gamma}) = \nabla s^{1-w}_{n_1}(\boldsymbol{\gamma}) + \nabla s^w_{n_1}(\boldsymbol{\gamma})$ and $N_{n_1}(\delta)$ converges to $\boldsymbol{\beta}_0$,

$$\begin{aligned} & I^{-1/2}_{n_1}(\boldsymbol{\beta}_0) \nabla s^w_{n_1}(\boldsymbol{\beta}_0) I^{-1/2}_{n_1}(\boldsymbol{\beta}_0) \\ &\quad = I^{-1/2}_{n_1}(\boldsymbol{\beta}_0) \nabla s_{n_1}(\boldsymbol{\beta}) I^{-1/2}_{n_1}(\boldsymbol{\beta}_0) + o(1) \\ &\quad \to -I_p, \end{aligned}$$

and

$$\nabla s^w_{n_1}(\boldsymbol{\beta}_0) \to -I_{n_1}(\boldsymbol{\beta}_0).$$

Then

$$\nabla s^w_n(\boldsymbol{\beta}_0) = \nabla s^w_{n_1}(\boldsymbol{\beta}_0) + \nabla s^w_{n_2}(\boldsymbol{\beta}_0) \to -I_{n_1}(\boldsymbol{\beta}_0).$$

Taking the Taylor expansion of $s^w_n(\hat{\boldsymbol{\beta}})$ at $\boldsymbol{\beta}_0$, where $\hat{\boldsymbol{\beta}}$ is the WL estimate, it follows that

$$s^w_n(\hat{\boldsymbol{\beta}}) = s^w_n(\boldsymbol{\beta}_0) + \nabla s^w_n(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(\nabla s^w_n(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))$$

and so

$$s^w_n(\hat{\boldsymbol{\beta}}) = s^w_n(\boldsymbol{\beta}_0) - I_{n_1}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(I_{n_1}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)) \tag{11}$$

Setting $s^w_n(\hat{\boldsymbol{\beta}}) = 0$, and multiplying the both sides of (11) by $I^{-1/2}_{n_1}(\boldsymbol{\beta}_0)$, we obtain

$$I^{1/2}_{n_1}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = I^{-1/2}_{n_1}(\boldsymbol{\beta}_0) s^w_n(\boldsymbol{\beta}_0) + o_p(1).$$

Because $C = o_p(1)$, we have $var(s^w_{n_2}(\boldsymbol{\beta})) \to 0$, then

$$\begin{aligned} var(s^w_n(\boldsymbol{\beta}_0)) &= var(s^w_{n_1}(\boldsymbol{\beta}_0)) + var(s^w_{n_2}(\boldsymbol{\beta}_0)) \\ &= \frac{1}{\phi} \left[ \sum_{j=1}^{n_1} w^2_{1j} \varphi((\boldsymbol{\beta}^T_0 X_j)) t_{1j} X_j X^T_j \right] + o_p(1) \end{aligned}$$

$$\begin{aligned} &= I_{n_1}(\boldsymbol{\beta}_0) - \frac{1}{\phi}\left[\sum_{j=1}^{n_1}(1-w_{1j}^2)\varphi((\boldsymbol{\beta}_0^T X_j))t_{1j}X_j X_j^T\right] + o_p(1) \\ &= I_{n_1}(\boldsymbol{\beta}_0) - o_p(var(s_n^w(\boldsymbol{\beta}_0))), \end{aligned}$$

and hence

$$I_{n_1}(\boldsymbol{\beta}_0) = var(s_n^w(\boldsymbol{\beta}_0)) + o_p(var(s_n^w(\boldsymbol{\beta}_0))).$$

By the CLT [Corrollary 1.3, 1] and Slutsky's theorem, we have

$$I_{n_1}^{-1/2}(\boldsymbol{\beta}_0)s_n^w(\boldsymbol{\beta}_0) \xrightarrow{d} N_p(0, I_p),$$

and hence

$$I_{n_1}^{1/2}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N_p(0, I_p).$$

## 4. Empirical Distribution by Bootstrap

We have given the asymptotic distribution of WLE, however, the asymptotic distribution is hard to use in real data analysis. At the same time, the sample size of data sometime is not big enough, and therefore it may cause bias to use the asymptotic distribution. In order to assess the significance of a parameter based on WLE, we proposed the bootstrap method to generate the empirical distribution of WLE.

Without loss of generality, we assume that $m = 2$ and number of independent variables to be 1. Suppose the data we have is $(x_i, y_i)$ for $i = 1, \ldots, n$. Below are the main steps to generate the empirical distribution.

**Step 1:** In the main group, let $z_j$ $(j = 1, \ldots, k)$ to be the unique values of $x_i$ $(i = 1, \ldots, n_1)$, where $k \leq n_1$. Calculate the frequency of $z_j$ and let $p_0 = \sum y_i/n_1$.

**Step 2:** Draw $n_1$ sample from $z_j$, with probabilities corresponds to frequencies of $z_i$. Denotes the new sample of the main group as $x_i'$ $(i = 1, \ldots, n_1)$.

**Step 3:** Regenerate $y_i'$ $(i = 1, \ldots, n_1)$ from binomial distribution $bin(n_1, p_0)$.

**Step 4:** Combine $(x_i', y_i')$ $(i = 1, \ldots, n_1)$ and $(x_i, y_i)$ $i = n_1 + 1, \ldots, n$, calculate the WLE of this new sample.

Repeat step 2 to 4 for $N$ times, to get the empirical distribution of WLE with $N$ values.

## 5. Simulation Study

In this section we present the simulation results to illustrate the numerical performance of the proposed methods. We mainly compare the estimated values and the standard deviations based on the MLE and WLE estimators and present the ratios of the MSE of the WLE to the MSE of the MLE under different simulation scenarios.

**Example 1.** *The data are generated from models* (8) *with* $\beta_1 = 1$ *and* $\beta_2 = 0.92$. *Let* $X_1$ *and* $X_2$ *are independent and identically distributed with* $N(0,1)$. *Also, we set that* $\sigma_1 = \sigma_2 = 0.3$. *The simulation study is conducted as follows. We consider three scenarios: (i)* $n_1 < n_2$; *(ii)* $n_1 = n_2$; *and (iii)* $n_1 > n_2$. *The range of* $n_1$ *and* $n_2$ *is between 8 and 90. For each fixed sample size,* 1000 *independent data sets are generated. Table 1 summarizes the results. It can be seen that both the MLE and WLE are close to the true values. The standard deviations of the WLE's are consistently smaller than those of the MLE. Furthermore, the ratios of the MSE of the WLE to the MSE of the MLE are always smaller than 1. This implies that the WLE could reduce the MSE in contrast with the MLE. To assess the sensitivity of the ratio to actual value of the related parameter* $\beta_2$, *we run simulations with different* $\beta_2$ *values ranging from* 0.7 *to* 0.98. *The trend of the ratios against the values of* $\beta_2$ *is provided in Figure 1. In general, we find that the smaller* $n_1$, *the smaller the ratio under the current setup. For each fixed* $n_1$, *the ratios are inversely proportional to the value of* $\beta_2$. *They would also increase with larger value for* $n_2$. *We see that the WLE outperforms the MLE remarkably in small sample size* $n_1$. *This feature could be useful in practice if the first sample size is small and we have abundant related information.*

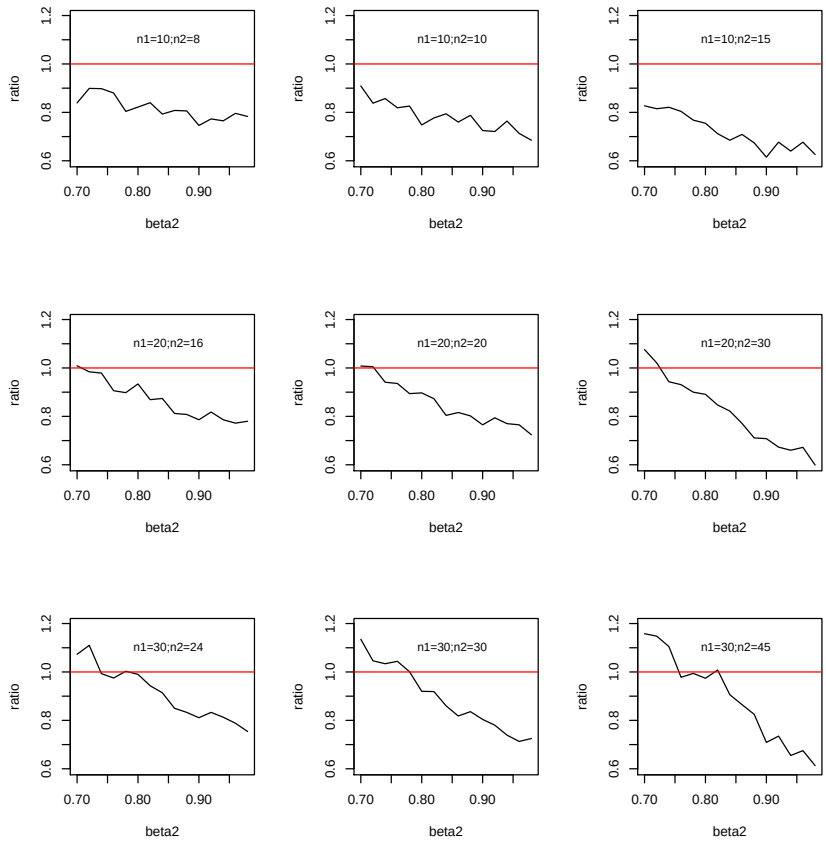


Figure 1: The trend of the ratio values against $\beta_2$ for 9 combinations of $(n_1, n_2)$. The horizontal line indicates ratio= 1.

Table 1: Estimated values based on the MLE and WLE, the corresponding MSE and the ratio of the MSE(WLE) to the MSE(MLE) for Example 1.

| | | MLE | | WLE | | | |
|---|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | $\hat{\beta}$(std) | MSE(std) | $\hat{\beta}$(std) | MSE(std) | $\hat{\lambda}$(std) | Ratio |
| 10 | 8 | 1.009(0.239) | 0.057(0.093) | 0.997(0.21) | 0.044(0.069) | 0.744(0.187) | 0.768 |
| 20 | 16 | 1.008(0.166) | 0.028(0.041) | 0.992(0.149) | 0.022(0.033) | 0.757(0.186) | 0.807 |
| 30 | 24 | 1.006(0.128) | 0.016(0.025) | 0.993(0.113) | 0.013(0.02) | 0.754(0.188) | 0.778 |
| 40 | 32 | 1.001(0.114) | 0.013(0.019) | 0.991(0.102) | 0.01(0.016) | 0.767(0.187) | 0.813 |
| 50 | 40 | 1.005(0.098) | 0.01(0.014) | 0.994(0.09) | 0.008(0.012) | 0.781(0.194) | 0.851 |
| 60 | 48 | 0.999(0.094) | 0.009(0.013) | 0.99(0.088) | 0.008(0.012) | 0.774(0.192) | 0.875 |
| 10 | 10 | 1.004(0.241) | 0.058(0.085) | 0.985(0.206) | 0.042(0.063) | 0.742(0.186) | 0.734 |
| 20 | 20 | 1.011(0.164) | 0.027(0.04) | 0.996(0.143) | 0.02(0.032) | 0.743(0.188) | 0.751 |
| 30 | 30 | 1.003(0.134) | 0.018(0.026) | 0.989(0.117) | 0.014(0.02) | 0.758(0.188) | 0.765 |
| 40 | 40 | 1.001(0.112) | 0.013(0.018) | 0.988(0.1) | 0.01(0.015) | 0.761(0.188) | 0.808 |
| 50 | 50 | 0.994(0.099) | 0.01(0.015) | 0.981(0.088) | 0.008(0.012) | 0.764(0.189) | 0.812 |
| 60 | 60 | 0.999(0.088) | 0.008(0.011) | 0.986(0.081) | 0.007(0.009) | 0.77(0.187) | 0.867 |
| 10 | 15 | 0.986(0.249) | 0.062(0.108) | 0.966(0.196) | 0.04(0.077) | 0.717(0.181) | 0.638 |
| 20 | 30 | 1.006(0.161) | 0.026(0.039) | 0.98(0.131) | 0.017(0.026) | 0.72(0.186) | 0.675 |
| 30 | 45 | 0.995(0.13) | 0.017(0.025) | 0.977(0.106) | 0.012(0.018) | 0.74(0.184) | 0.693 |
| 40 | 60 | 1.002(0.109) | 0.012(0.017) | 0.984(0.095) | 0.009(0.014) | 0.753(0.188) | 0.78 |
| 50 | 75 | 1.002(0.101) | 0.01(0.015) | 0.982(0.084) | 0.007(0.011) | 0.754(0.184) | 0.729 |
| 60 | 90 | 1.002(0.096) | 0.009(0.014) | 0.986(0.086) | 0.007(0.011) | 0.761(0.185) | 0.815 |

## 6. A Real Data Example

The low birth weight data comes from a study of risk factors associated with low infant birth weight.The data were collected at Baystate Medical Center, Springfield, Massachusetts, during 1986 and presented in [8]. The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby who weighs less than 2500 grams. In this study data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. Variables which were thought to be of importance were weight of the subject at her last menstrual period and race. These two variables are denoted as "LWT" and "RACE". The response "LOW", is a binary variable which indicates a low birth weight baby if equals to 1, and normal birth weight if equals to 0. The variable "RACE" has been recoded using the two dummy variables. First of all, we investigate on the model

$$g(\mu) = \beta_0 + \beta_1 LWT,$$

where $\mu$ is the expected value of the binary response and $g$ is the link function for binomial family. We find that the variable LWT is significant. The estimate of the coefficients and the corresponding p-values are given in Table 2. One can also verify that the interaction between LWT and RACE is not significant.

Table 2: Estimated Coefficients for general logistic regression model using the variables LWT.

| | Estimate | $p$-value |
|---|---|---|
| Intercept | 0.9983 | 0.2036 |
| LWT | -0.0141 | 0.0227 |

Then we check the significance of LWT within each RACE individually. The $p$-values are all greater than the $p$-value of LWT for each group is given in Table 3. Despite the fact that the variable LWT is significant for the overall model, it is very surprising to see that the LWT is not significant within any group. By applying the WLE method based on distribution from bootstraping, we find the LWT is significant in "Other" group.

Table 3: Estimated Coefficients for general logistic regression model using the variables LWT in each RACE group.

| | Estimate | $p$-value |
|---|---|---|
| | "White" Group | |
| Intercept | 0.7925 | 0.528 |
| LWT | -0.0151 | 0.123 |
| | "Black" Group | |
| Intercept | 0.6941 | 0.663 |
| LWT | -0.0069 | 0.517 |
| | "Other" Group | |
| Intercept | 2.7135 | 0.1066 |
| LWT | -0.0275 | 0.0559 |

Table 4: Estimated Coefficients for weighted logistic regression model using the variables LWT with each RACE group as the main group. P-values are based on bootstrap distribution.

| | Estimate | $p$-value |
|---|---|---|
| | "White" Group | |
| Intercept | 0.7925 | 0.1296 |
| LWT | -0.0148 | 0.1332 |
| | "Black" Group | |
| Intercept | 0.6941 | 0.5396 |
| LWT | -0.0075 | 0.5788 |
| | "Other" Group | |
| Intercept | 2.7135 | 0.0408 |
| LWT | -0.0276 | 0.0432 |

## 7. Discussion

To obtain a more efficient estimator with a smaller MSE when related information is available, we proposed the weighted likelihood inference for the parameters in linear and partially linear models. We developed a data-driven approach to estimate the weights to address the vexing issued in weighted likelihood inference. The proposed estimators have the same asymptotically normal distribution as the maximum likelihood estimators. The advantage of the WLE over the classical MLE was illustrated by simulation studies. Results from these simulation studies suggest that the proposed estimators have prospective promises. The simplicity and effectiveness of the proposed adaptive weights are also appealing.

We remark that the real data set used in this paper actually came from a longitudinal study. Thus one must recognize the within-subject and between-subject variations. Further investigation of the inference based on the weighted likelihood for mixed-effect models is required.

**ACKNOWLEDGEMENTS** The research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

## References

[1] J Shao. *Mathematical Statistics.* Springer, New York, 2003.

[2] J Staniswalis. The kernel estimate of a regression function in a likelihood-based models. *J. Am. Statist. Assoc.*, 84:276–83, 1989.

[3] M Stone. Strong inconsistencies from uniform priors. *J. Am. Statist. Assoc.*, 71:114–116, 1976.

[4] R Tibshirani and T Hastie. Local likelihood estimation. *J. Am. Statist. Assoc.*, 82:559–67, 1987.

[5] X Wang. Approximating bayesian inference by weighted likelihood. *Can. J. Statist.*, 34:279–298, 2006.

[6] X Wang and J Zidek. Derivation of mixture distribution and weighted likelihood as minimizers of kl-divergence subject to constraints. *Ann. Inst. Statist. Math.*, 57:687–701, 2005.

[7] X Wang and J Zidek. Selecting likelihood weights by cross-validation. *Ann. Statist.*, 33:463–500, 2005.

[8] D. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 1989.

[9] F. Hu. The asymptotic properties of the maximum-relevance weighted likelihood estimators. *Can. J. Statist.*, 25:45–59, 1997.

[10] F. Hu and J. Zidek. The relevance weighted likelihood with applications. In Ahmed, S.E. & Reid, N. , editor, *Empirical Bayes and Likelihood Inference.*, pages 211–235, New York., 2001. Springer Verlag.

[11] F Hu and J Zidek. The weighted likelihood. *Can. J. Statist.*, 25:347–71, 2002.

[12] F. Hu and W. Rosenberger. Analysis of time trends in adaptive designs with applications to a neurophysiology experiments. *Statist. Med.*, 19:2067–75, 2000.

[13] H. Akaike. Information theory and an extension of the maximum likelihood principle. In Petrov, B.N. & Csaki, F. , editor, *Proc. 2nd International Symposium on Information Theory.*, pages 267–281, Budapest., 1973. Akademiai Kiado.

[14] I. Csiszár. *I*-divergence geometry of probability distributions and minimization problems. . *Ann. Statist.*, 3:146–158, 1975.

[15] J. Bernardo. A maximum likelihood methodology for clusterwise linear regression. *Ann. Statist.*, 7:686–690, 1979.

[16] H Royden. *Real Analysis*. Prentice Hall, New York, 3 edition, 1988.

[17] S. Eguchi and J. Copas. A class of local likelihood methods and near-parametric asymptotics. *J. R. Statist. Soc. B*, 60:709–24, 1998.

[18] S Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[19] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[20] W James and C Stein. Estimation with quadratic loss. In *Proc 4th Berkeley Symp. Math Statist Prob.*, volume 1, pages 361–379. Berkely: University of California Press., 1961.

[21] X Wang, C van Emden and J Zidek. Asymptotic properties of maximum weighted likelihood estimator. *J. Statist. Plan. Infer.*, 119:37–54, 2004.