Journal of Statistical Planning and Inference 231 (2024) 106121

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Inference on regression model with misclassified binary response

Arindam Chatterjee [a,*,1], Tathagata Bandyopadhyay [b], Ayoushman Bhattacharya [c]

[a] Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, New Delhi 110016, India
[b] Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
[c] Department of Statistics and Data Science, Washington University in St. Louis, Missouri, United States of America

ARTICLE INFO

MSC:
primary 62J12
secondary 62F12

Keywords:
Contaminated data
Validation sample
Binary response
M-estimator
Bootstrap

ABSTRACT

Misclassification of binary responses, if ignored, may severely bias the maximum likelihood estimators (MLEs) of regression parameters. For such data, a binary regression model incorporating non-differential classification errors is extensively used by researchers in different application contexts. We strongly caution against indiscriminate use of this model considering the fact that it suffers from a serious estimation problem due to confounding of the unknown misclassification probabilities with the regression parameters, and thus, may lead to a highly biased estimate. To overcome this problem, we propose here the use of an internal validation sample in addition to the main sample. Assuming differential classification errors, we consider MLEs of the regression parameters based on the joint likelihood of the main sample and the internal validation sample. We then develop a rigorous asymptotic theory for the joint MLEs under standard assumptions. To facilitate its easy implementation for inference, we propose a bootstrap approximation to the asymptotic distribution and prove its consistency. The results of the simulation studies suggest that even an extremely small validation sample may lead to a vastly improved inference. Finally, the methodology is illustrated with a real-life survey data.

1. Introduction

Misclassified binary responses are common in sociological, economic, education and health surveys (Hausman et al., 1998, Hausman, 2001, Bollinger and David, 1997, Savoca, 2011, Bound et al., 2001, Bollinger and David, 2001, Poterba and Summers, 1995, Kreider and Pepper, 2008, Black et al., 2003). Examples include participation in welfare programs, medicaid enrolment, workers' union status, employment status, educational attainment, language proficiency, self-reported health conditions, self-reported voting behaviour, physical and mental impairment, and disease status. Misclassification is also common in political science, epidemiological data (Katz and Katz, 2010, Hug, 2010, Lyles et al., 2011, Gilbert et al., 2014, Zawistowski et al., 2017) and also in genome-wide association studies (Smith et al., 2013, Rekaya et al., 2016). Misclassified remotely sensed binary measure of deforestation is considered in Alix-Garcia and Millimet (2023). Researchers (Copas, 1988, Carroll et al., 2006 (Section 15.3), Neuhaus, 1999) observe that, not accounting for misclassification of the dependent variable in binary choice model results in severely biased and inconsistent parameter estimates, inaccurate standard errors, and thus potentially distorting the relative impact of different explanatory variables of interest on the response variable leading to erroneous conclusions.

Suppose, $Y \in \{0,1\}$, denotes the true response, and $\widetilde{Y} \in \{0,1\}$, denotes the misclassified version of Y. Then, Y is said to be misclassified, if we observe $\widetilde{Y} = 1$, when $Y = 0$, or vice-versa. In the absence of misclassification, the binary choice model for the

* Corresponding author.
E-mail addresses: cha@isid.ac.in (A. Chatterjee), tathagata_b@daiict.ac.in (T. Bandyopadhyay), abhattacharya@wustl.edu (A. Bhattacharya).
[1] Research partially supported by SERB, India grant MTR/2017/000224.

https://doi.org/10.1016/j.jspi.2023.106121
Received 28 November 2022; Received in revised form 10 October 2023; Accepted 9 November 2023
Available online 29 November 2023

response Y is assumed to be of the form,

$$\mathbf{P}_0\,(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \psi(\mathbf{x}^T\boldsymbol{\beta}_0), \quad \text{for all } \mathbf{x} \in \mathbb{R}^p, \tag{1.1}$$

where, $\psi(\cdot)$ is a specified cumulative distribution function (cdf), $\mathbf{x} = (x_1, \ldots, x_p)^T$ is the vector of covariates, $\boldsymbol{\beta}_0 = \left(\beta_{1,0}, \ldots, \beta_{p,0}\right)^T$ is the unknown regression parameter, and $\mathbf{P}_0$ denotes the underlying probability distribution. Common examples of $\psi(\cdot)$ are $\{1 + e^{-u}\}^{-1}$ (logistic) and $\Phi(u)$ (probit), where $\Phi(\cdot)$ is the standard normal cdf.

To correct for misclassification, a regression model for $\widetilde{Y} \in \{0, 1\}$ assuming non-differential misclassification is used by many researchers (Copas, 1988, Magder and Hughes, 1997, Bollinger and David, 1997, Hausman et al., 1998, Roy et al., 2005, Gilbert et al., 2014, Katz and Katz, 2010, Hug, 2010, Savoca, 2011, Lyles et al., 2011, Smith et al., 2013). The model assumes,

$$\mathbf{P}_0\big(\widetilde{Y} = 1 \mid Y = 0\big) = \theta_{1,0} \quad \text{and} \quad \mathbf{P}_0\big(\widetilde{Y} = 0 \mid Y = 1\big) = \theta_{2,0}. \tag{1.2}$$

Thus, the resulting regression model for $\widetilde{Y}$ is,

$$\mathbf{P}_0\big(\widetilde{Y} = 1 \mid \mathbf{X} = \mathbf{x}\big) = \theta_{1,0} + (1 - \theta_{1,0} - \theta_{2,0}) \cdot \psi(\mathbf{x}^T\boldsymbol{\beta}_0). \tag{1.3}$$

This kind of misclassification is often referred to as "conditionally random", because given the true value Y, the misclassification probabilities are constants.

However, this model suffers from some serious issues. First, MLE of the regression parameter $\boldsymbol{\beta}_0$ often performs poorly even for very large sample sizes. This is due to the confounding of the regression parameter $\boldsymbol{\beta}_0$ with the misclassification probabilities $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0})^T$. This is a serious issue (Copas, 1988, Cox (p. 256 of Copas, 1988), Ekholm and Palmgren (p. 262 of Copas, 1988) that needs to be addressed properly. Second, in many applications, the assumption of non-differential misclassification seems to be inappropriate (Bollinger and David, 1997, Meyer and Mittag, 2017, Yi, 2017). As pointed out by Meyer and Mittag (2017), "there is little ex ante reason to believe that misclassification is independent of the covariates".

In this paper, we address both these issues. To overcome the confounding problem, we consider a likelihood-based methodology for inference on $\boldsymbol{\beta}_0$ using an internal validation sample in addition to the main sample. We assume a differential misclassification model by allowing $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0})^T$ to depend on the covariates. It needs mentioning that, in various application contexts, researchers (Lyles et al., 2011, Edwards et al., 2013, Katz and Katz, 2010) consider the likelihood-based inference on $\boldsymbol{\beta}_0$ utilizing internal validation. However, first, to the best of our knowledge the asymptotic properties of these estimators are not studied before. Also, the standard asymptotic arguments for MLE's based on independent and identically distributed (i.i.d.) data do not work in the presence of internal validation data. In fact, we show in Section 3, the log-likelihood for the proposed estimator is a function of non-identically distributed summands. Thus, to derive the asymptotic distribution, we invoke empirical process results for proving limit laws of M-estimators. Further, to draw inference on $\boldsymbol{\beta}_0$, the estimation of the asymptotic covariance matrix is essential. However, its computation is tedious, because it depends on the unknown distribution of the covariates, the link function $\psi(\cdot)$ and its derivatives, and also on the unknown parameters. To avoid its direct computation from data, we propose a bootstrap approximation to the asymptotic distribution and prove its consistency. To summarize, we present here a rigorous asymptotic theory for the likelihood estimator of $\boldsymbol{\beta}_0$ using internal validation data assuming both differential and non-differential misclassification.

The paper is organized as follows. In Section 2, we discuss the confounding issue in detail, and investigate its impact on the performance of the estimator of $\boldsymbol{\beta}_0$ by carrying out a small numerical study. We then introduce a generalization of model (1.3) with differential misclassification. In Section 3, under mild assumptions, we prove the asymptotic normality of the proposed estimator assuming differential misclassification. Next in Section 3.1, we present a modification of the asymptotic theory in case one of the misclassification probabilities is known a priori to be zero, a situation often found in practice (cf. Kothari and Bandyopadhyay, 2011 and Section 4.1). In Section 3.2, we propose a bootstrap approximation to the asymptotic distribution and prove its consistency. Numerical studies, presented in Section 4, suggest that both in terms of the coverage and expected length, the bootstrap confidence intervals (CIs) match closely with the corresponding asymptotic CIs. A real-life data study is presented in Section 4.1. Concluding remarks are given in Section 5. Proofs of results and definitions of some matrices used in main results are provided in the Appendix.

2. Confounding and differential misclassification

In this section, first, we discuss in detail the root cause of the confounding, and provide a statistical explanation for it. We then present the results of a small numerical study showing its impact on the accuracy of the following MLE of $\boldsymbol{\beta}_0$: MLE based on (i) the likelihood of the main sample, (ii) the likelihood of the validation sample, (iii) the likelihood of the combined sample, *i.e.*, the likelihood of the main and validation sample combined, and (iv) the pseudo-likelihood, obtained by plugging in an estimate of $\boldsymbol{\theta}_0$ from the validation sample to the likelihood of the combined sample. Finally, we propose a differential misclassification model for developing the likelihood-based methodology for estimation of $\boldsymbol{\beta}_0$.

2.1. Confounding

Hausman et al. (1998) propose the MLE of $\boldsymbol{\beta}_0$ based on the contaminated data $\{(\widetilde{Y}_i, \mathbf{X}_i) : i = 1, \ldots, n\}$. We call this estimator NMLE, or naive MLE. They prove its consistency if the cdf $\psi(\cdot)$ is symmetric about zero, and $\theta_{1,0} + \theta_{2,0} < 1$. Although it is consistent, its performance may be very poor (Copas, 1988, Cox (p. 256 of Copas, 1988), Ekholm and Palmgren (p. 262 of Copas, 1988) even for very large sample sizes. Thus, it is important to raise a caveat against indiscriminate use of NMLE by providing a strong statistical reason for it.

Notice that, if $\psi(\cdot)$ is linear, the model (1.3) is not identifiable. Also, note that, the common choices for $\psi(\cdot)$, the logit and the probit functions, are almost linear except near the boundaries of the interval $(0,1)$. Thus, for such choices, θ_0 and β_0 are virtually confounded (cf. (1.3)) unless enough misclassified responses lie near the boundaries of the interval $(0,1)$ to capture the non-linearity of $\psi(\cdot)$ (Copas, 1988) leading to a meaningful estimate of β_0. Otherwise, the estimate of β_0 is highly unreliable. Cox (p. 256 of Copas, 1988) argues that, misclassifying observations with high chance of success or failure is unlikely in a practical situation. Thus, to get round this difficulty, "independent data for the error probabilities are essential for adjustment". Ekholm and Palmgren (p.262 of Copas, 1988) suggest, "One remedy for the identifiability problem is to restrict the use of these models to cases where a random subsample of the units has been diagnosed with an errorless device".

Thus, to avoid the confounding problem, information about the true error probabilities θ_0 is essential. In this paper, we assume that complete information on $(Y, \widetilde{Y}, \mathbf{X})$ for a randomly chosen sub-sample of the main sample, known as internal validation sample (Lyles et al., 2011, Edwards et al., 2013, Katz and Katz, 2010, Bound et al., 2001), is available. Bound et al. (2001) say, "the usefulness of validation data in telling us about errors in survey measures can be enhanced if validation data is collected for a random portion of major surveys (rather than, as is usually the case, for a separate convenience sample for which validation data could be obtained relatively easily)". In survey literature, the use of two-phase sampling design for collection of internal validation data is as old as Neyman (1938) and Cochran (1977). Recently, for causal inference with unobserved confounders, renewed interest in the use of internal validation data is observed among the researchers (Wang et al., 2009, Yang and Ding, 2020).

One of the peripheral goals of this paper is to show that it is possible to obtain very accurate estimates of β_0 by utilizing the information in the internal validation data, even for relatively small validation sample sizes. Thus, whenever possible, the collection of internal validation data should be planned. It should be considered as a good practice for the data collection process considering its huge impact on the accuracy of the estimator. In the next section (cf. Section 2.2), we present the results of an empirical study to drive this point home.

2.2. An empirical study: Effect of internal validation data on accuracy

Here, we report the results of an empirical study for non-differential misclassification comparing the performances of the MLEs of β_0 discussed in the beginning of this section in a simple logistic regression model set up. We present the results of a similar study for differential misclassification in Section 2.3. Let us write the logistic regression function,

$$\psi_L(\mathbf{x}^T\beta_0) \equiv \frac{1}{1+\exp\{-\beta_{1,0}-\beta_{2,0}x\}}, \quad \text{for all } x \in \mathbb{R}, \tag{2.1}$$

where, $\mathbf{x} = (1,x)^T$, and $\beta_0 = (\beta_{1,0}, \beta_{2,0})^T$. Let $\{(\widetilde{Y}_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ represent the main sample, $\{(Y_i, \widetilde{Y}_i, \mathbf{X}_i) : i = 1, \ldots, n_1\}$, the internal validation sample, and $f(= n_1/n)$, the internal validation sampling fraction. Evidently f is a proper fraction. Write $\beta = (\beta_1, \beta_2)^T$ and $\theta = (\theta_1, \theta_2)^T$ to denote the regression parameters and the misclassification parameters, respectively. The following MLEs of β_0 are considered.

(i) Naive MLE (NMLE): It is obtained by maximizing the conditional log-likelihood of $[\widetilde{Y} \mid \mathbf{X}]$ (cf. (1.3)) of the main sample $\{(\widetilde{Y}_i, \mathbf{X}_i) : 1 \le i \le n\}$. The log-likelihood is given by

$$\log L_N(\beta,\theta) = \sum_{i=1}^{n} \Big[\widetilde{Y}_i \cdot \log\{\theta_1 + (1-\theta_1-\theta_2)\cdot\psi_L(\mathbf{X}_i^T\beta)\} + (1-\widetilde{Y}_i)\cdot\log\{1-\theta_1-(1-\theta_1-\theta_2)\cdot\psi_L(\mathbf{X}_i^T\beta)\}\Big].$$

(ii) Validation-data MLE (VMLE): It is obtained by maximizing the log-likelihood of the validation sample $\{(Y_i, \mathbf{X}_i) : 1 \le i \le n_1\}$. The log-likelihood is given by,

$$\log L_V(\beta) = \sum_{i=1}^{n_1} \left[Y_i \cdot \log \psi_L(\mathbf{X}_i^T\beta) + (1-Y_i)\cdot\log\{1-\psi_L(\mathbf{X}_i^T\beta)\}\right].$$

(iii) Combined data MLE (CMLE): It is obtained by maximizing the log-likelihood based on the combined sample, *i.e.*, combining both the main and the validation samples. The log-likelihood is given by,

$$\begin{aligned}\log L_C(\beta,\theta) &= \log L_V(\beta) \\ &+ \sum_{i=1}^{n_1}\left[Y_i\widetilde{Y}_i\cdot\log(1-\theta_2) + Y_i(1-\widetilde{Y}_i)\cdot\log\theta_2 + (1-Y_i)(1-\widetilde{Y}_i)\cdot\log(1-\theta_1) + (1-Y_i)\widetilde{Y}_i\cdot\log\theta_1\right] \\ &+ \sum_{i=n_1+1}^{n}\left[\widetilde{Y}_i\cdot\log\{\theta_1+(1-\theta_1-\theta_2)\cdot\psi_L(\mathbf{X}_i^T\beta)\} + (1-\widetilde{Y}_i)\cdot\log\{1-\theta_1-(1-\theta_1-\theta_2)\cdot\psi_L(\mathbf{X}_i^T\beta)\}\right].\end{aligned}$$

Table 1
Bias and MSE (in parenthesis) of four different likelihood-based estimators of $\boldsymbol{\beta}_0$ (cf. (2.1)) for **non-differential misclassification**, at various choices of n and f. The results are based on the simulation plan described in Section 2.2.

n	f	Estimators of $\beta_{1,0}$				Estimators of $\beta_{2,0}$			
		CMLE	VMLE	PMLE	NMLE[a]	CMLE	VMLE	PMLE	NMLE[a]
200[b]	0				118				−445
					(9.94×10^5)				(52.2×10^5)
	0.1	0.074	0.285	0.018		−0.174	−0.488	−0.156	
		(0.337)	(1.94)	(0.358)		(0.622)	(3.07)	(0.284)	
	0.15	0.078	0.139	0.036		−0.108	−0.198	−0.111	
		(0.219)	(0.392)	(0.222)		(0.210)	(0.555)	(0.177)	
	0.2	0.103	0.140	−0.070		−0.128	−0.182	−0.126	
		(0.162)	(0.314)	(0.174)		(0.195)	(0.439)	(0.178)	
	0.3	0.027	0.047	0.012		−0.058	−0.077	−0.058	
		(0.090)	(0.123)	(0.103)		(0.119)	(0.193)	(0.112)	
1000	0				0.072				−0.388
					(0.439)				(1.46)
	0.05	0.017	0.076	0.019		−0.024	−0.097	−0.064	
		(0.076)	(0.188)	(0.199)		(0.062)	(0.245)	(0.077)	
	0.1	0.003	0.050	0.012		−0.030	−0.053	−0.038	
		(0.048)	(0.078)	(0.083)		(0.036)	(0.098)	(0.039)	
	0.2	0.022	0.039	−0.002		−0.021	−0.044	−0.017	
		(0.026)	(0.041)	(0.031)		(0.023)	(0.047)	(0.023)	
	0.3	−0.001	−0.005	−0.007		−0.004	−0.007	0.003	
		(0.016)	(0.022)	(0.019)		(0.017)	(0.027)	(0.017)	

[a] NMLE does not use validation data, hence $f = 0$.

[b] We drop $f = 0.05$ for $n = 200$ case, to avoid problems with numerical convergence.

Table 2
Bias and MSE (in parenthesis) of the CMLE, PMLE and NMLE estimators of $\boldsymbol{\theta}_0$ for **non-differential misclassification**, at $n \in \{200, 1000\}$ and $f = 0.2$. The results are based on the simulation plan described in Section 2.2.

Parameter	$n = 200,\ f = 0.2$			$n = 1000,\ f = 0.2$		
	CMLE	PMLE[a]	NMLE[b]	CMLE	PMLE[a]	NMLE[b]
$\theta_{1,0}$	−0.004	0.032	0.094	−0.001	0.007	0.025
	(0.007)	(0.008)	(0.043)	(0.001)	(0.001)	(0.018)
$\theta_{2,0}$	0.003	0.016	−0.010	0.000	−0.001	−0.005
	(0.003)	(0.006)	(0.017)	(0.001)	(0.001)	(0.007)

[a] The PMLE estimator is $\widetilde{\boldsymbol{\theta}}_n$ (cf. (2.2)).

[b] The NMLE does not use validation data, hence $f = 0$ for NMLE.

(iv) Pseudo-MLE (PMLE): It is obtained by maximizing the pseudo log-likelihood $L_P(\boldsymbol{\beta}) = L_C(\boldsymbol{\beta}, \widetilde{\boldsymbol{\theta}}_n)$, where $\widetilde{\boldsymbol{\theta}}_n$ is an estimator of $\boldsymbol{\theta}_0$ based on the internal validation data. The estimator $\widetilde{\boldsymbol{\theta}}_n = \left(\widetilde{\theta}_{1,n}, \widetilde{\theta}_{2,n}\right)^T$ that we consider is a contingency table based estimator (cf. Haldane, 1956 and Gart and Zweifel, 1967) of the misclassification probabilities,

$$\widetilde{\theta}_{1,n} = \frac{\frac{1}{2} + \sum_{i=1}^{n_1} \mathbf{1}\left(\widetilde{Y}_i = 1, Y_i = 0\right)}{1 + \sum_{i=1}^{n_1} \mathbf{1}(Y_i = 0)} \quad \text{and} \quad \widetilde{\theta}_{2,n} = \frac{\frac{1}{2} + \sum_{i=1}^{n_1} \mathbf{1}\left(\widetilde{Y}_i = 0, Y_i = 1\right)}{1 + \sum_{i=1}^{n_1} \mathbf{1}(Y_i = 1)}. \tag{2.2}$$

The PMLE is included in the simulation study for comparison purposes following the suggestion of one of the reviewers.

Remark 2.1. Note that, the VMLE only produces estimates of $\boldsymbol{\beta}_0$. On the other hand, to find the NMLE, CMLE and PMLE of $\boldsymbol{\beta}_0$, $\boldsymbol{\theta}_0$ is to be estimated. Also note that, the above estimators (and log-likelihoods) have been defined in a non-differential misclassification setting (cf. (1.2)), but these definitions can be easily extended in the differential misclassification setting (see (2.3)). Also, the definitions stated above remain valid if the logistic link in (2.1) is replaced by any other known binary link function.

In order to generate samples, we use $\boldsymbol{\beta}_0 = (\beta_{1,0}, \beta_{2,0})^T = (1, -1)^T$ in (2.1), and the covariate $X \sim N(0, 1)$. Here, we study the impact of non-differential misclassification (cf. (1.2)) on the accuracy of the estimators assuming constant misclassification probabilities $\boldsymbol{\theta}_0 = (0.1, 0.2)^T$. We consider two different sample sizes, $n = 200$ and $n = 1000$, and $f \in \{0.05, 0.1, 0.15, 0.2, 0.3\}$. In Table 1, we present the bias and mean squared errors (MSE's) of the estimates of $\beta_{1,0}$ and $\beta_{2,0}$ for NMLE, VMLE, CMLE and PMLE. In Table 2 we present similar results for NMLE and CMLE of $\boldsymbol{\theta}_0$ as well as for $\widetilde{\boldsymbol{\theta}}_n$. The misclassification probabilities are usually treated as nuisance parameters, but their estimation may be of interest in some situations. Hence, following the suggestion of one of the reviewers, we have presented the bias and MSE of its estimators in Table 2. Although several choices of the sampling fraction f are considered, the primary interest is to compare the performance of CMLE with that of NMLE, PMLE and VMLE when f is small. The results in are based on 500 Monte Carlo replications.

Table 3
Bias and MSE (in parenthesis) of four different likelihood-based estimators of β_0 (cf. (2.1)) for **differential misclassification**, at various choices of n and f_n. The results are based on the simulation plan described in Sections 2.2 and 2.3.

n	f_n	Estimators of $\beta_{1,0}$				Estimators of $\beta_{2,0}$			
		CMLE	VMLE	PMLE	NMLE[a]	CMLE	VMLE	PMLE	NMLE[a]
200	0				125				−152.1
					(6.52×10^5)				(12.96×10^5)
	0.1[b]	0.090	0.251	0.108		−0.188	−0.386	−0.126	
		(0.765)	(1.394)	(0.963)		(1.128)	(2.873)	(1.596)	
	0.15	0.099	0.179	0.014		−0.111	−0.207	−0.068	
		(0.373)	(0.527)	(0.229)		(0.610)	(0.710)	(0.263)	
	0.2	0.030	0.107	−0.018		−0.072	−0.160	−0.046	
		(0.143)	(0.249)	(0.140)		(0.260)	(0.506)	(0.193)	
	0.3	0.040	0.065	0.009		−0.052	−0.086	−0.026	
		(0.099)	(0.142)	(0.097)		(0.134)	(0.196)	(0.123)	
1000	0				53.91				−54.99
					(12.86×10^5)				(13.61×10^5)
	0.05	0.030	0.068	0.015		−0.047	−0.129	−0.043	
		(0.084)	(0.156)	(0.177)		(0.095)	(0.261)	(0.178)	
	0.1	0.010	0.022	−0.008		−0.008	−0.015	−0.019	
		(0.038)	(0.068)	(0.063)		(0.037)	(0.080)	(0.063)	
	0.2	0.004	0.015	−0.009		−0.019	−0.031	−0.011	
		(0.021)	(0.033)	(0.029)		(0.027)	(0.044)	(0.035)	
	0.3	−0.008	0.003	−0.022		0.002	−0.006	0.009	
		(0.015)	(0.021)	(0.018)		(0.021)	(0.030)	(0.023)	

[a] NMLE does not use validation data, hence $f_n = 0$.
[b] Numerical convergence issues arise in this case.

As seen from Table 1, the NMLE is highly inaccurate and suffers from serious numerical convergence issues due to reasons discussed above, especially at small sample sizes. Based on our simulations, we find that the NMLE is highly erratic and shows extreme fluctuations due to numerical instability in the optimization algorithms, which is caused by the shape of the NMLE log-likelihood surface. In comparison, the CMLE is highly accurate even for small values of n and f_n. As it turns out, the PMLE has nearly comparable performance as the CMLE for $n = 200$, but the CMLE has lower MSE and bias for $n = 1000$. This implies, the CMLE effectively utilizes the information in the combined sample to improve estimation. Also, as expected the performance of CMLE is substantially better than VMLE. It needs mentioning in this context, for smaller sample sizes, PMLE may be considered as a worthy alternative to CMLE. In Section 3, we develop the asymptotic theory for CMLE assuming a differential misclassification model which we introduce in Section 2.3.

Similarly in Table 2, we show the bias and the MSE for CMLE and NMLE based estimators of θ_0 as well as for $\widetilde{\theta}_n$. We find that CMLE is the most accurate. To save space, we present results for only one choice of validation fraction, *viz.*, $f = 0.2$, but the conclusions remain similar for other choices of f. Table 2 shows that CMLE of θ_0 performs much better than $\widetilde{\theta}_n$. The figures for NMLE in Table 2 suggest that even though estimates for β_0 are highly erratic, due to the constrained nature of the optimization problem, the NMLE based θ_0 estimates are less erratic, but still it remains an unreliable estimator of θ_0. Overall, on the basis of results in , we may conclude that the CMLE is the preferred choice among these four types of estimators.

2.3. Model with differential misclassification

In sociological and economic surveys, it is common to observe that the misclassification probabilities of the binary response (*viz.*, whether beneficiary of a program or not) depend on the covariates, like union membership (a member or not), income (above or below median income), age (above the median age or not) etc. (cf. Bollinger and David, 1997 and Abrevaya and Hausman, 1999). Thus, in the context of real applications, it may be reasonable to assume that the covariates are categorical, and accordingly, we partition the sample space of $\mathbf{X}$ into R non-overlapping subsets, say, $G_1, \ldots, G_R$, each representing the profile of a distinct group, and possibly, with different misclassification probabilities. We assume that R is fixed. The true model connecting Y and $\mathbf{X}$ remains the same as in (1.1), but the misclassification probabilities are now allowed to depend on the covariates in the following manner; for all $r = 1, \ldots, R$

$$\mathbf{P}_0\big(\widetilde{Y} = 1 \mid Y = 0, \mathbf{X} \in G_r\big) = \theta_{1,0}^{(r)}, \quad \text{and} \quad \mathbf{P}_0\big(\widetilde{Y} = 0 \mid Y = 1, \mathbf{X} \in G_r\big) = \theta_{2,0}^{(r)}. \tag{2.3}$$

We write $\boldsymbol{\theta}_0^{(r)} = \big(\theta_{1,0}^{(r)}, \theta_{2,0}^{(r)}\big)^T$ to denote the misclassification probabilities associated with the group G_r. It should be noted that the non-differential case in (1.2) corresponds to $R = 1$ (with G_1 equal to the entire sample space of $\mathbf{X}$).

Following the suggestions of reviewers, we present a simulation study similar to the one shown in comparing the performances of the same set of estimators in case of differential misclassification. The true model (2.1) remains same as above, and we consider two groups ($R = 2$) partitioning the sample space of X, *viz.*, $G_1 = \{x : x < 0\}$ and $G_2 = \{x : x > 0\}$. Since $X \sim N(0, 1)$ (see description of simulation study in Section 2.2), we have $\mathbf{P}(X \in G_1) = \mathbf{P}(X \in G_2) = 1/2$. This is designed to evenly compare the

Table 4
Bias and MSE (in parenthesis) of the CMLE, PMLE and NMLE estimators of $\theta_0^{(r)}$, $r = 1, 2$, for **differential misclassification**, at $n \in \{200, 1000\}$ and $f = 0.2$.

Parameter	$n = 200,\ f = 0.2$			$n = 1000,\ f = 0.2$		
	CMLE	PMLE[a]	NMLE[b]	CMLE	PMLE[a]	NMLE[b]
$\theta_{1,0}^{(1)}$	−0.002	0.123	0.190	−0.002	0.026	0.056
	(0.031)	(0.035)	(0.175)	(0.006)	(0.006)	(0.051)
$\theta_{2,0}^{(1)}$	−0.007	0.019	0.006	0.000	0.004	0.000
	(0.004)	(0.007)	(0.013)	(0.001)	(0.001)	(0.006)
$\theta_{1,0}^{(2)}$	−0.011	0.035	0.092	−0.005	0.005	0.028
	(0.012)	(0.014)	(0.058)	(0.002)	(0.003)	(0.022)
$\theta_{2,0}^{(2)}$	−0.004	0.023	−0.022	−0.002	0.005	−0.052
	(0.010)	(0.014)	(0.048)	(0.002)	(0.003)	(0.021)

[a] The PMLE estimator is the group-wise version of $\widetilde{\theta}_n$ (cf. (2.2)).
[b] The NMLE does not use validation data, hence $f = 0$ for NMLE.

performance for two groups. We assume $\theta_0^{(1)} = (0.1, 0.15)^T$ and $\theta_0^{(2)} = (0.15, 0.20)^T$ for groups G_1 and G_2, respectively. It must be noted that with $R = 2$, the effective validation sample size available for the rth group is proportional to $\mathbf{P}(\mathbf{X} \in G_r) \cdot n_1$. Thus, in our simulation, the effective validation sample sizes for both the groups G_1 and G_2 are equal to $n_1/2$. Thus, in practice, the validation sample size n_1 needs to be large enough to ensure that each groups has enough observations for estimating the misclassification probabilities. An extremely small validation sample size in a group may lead to the convergence problem of optimization algorithms used for obtaining the MLE's. The simulated bias and MSE values are presented in Table 3 (for the estimators of $\boldsymbol{\beta}_0$) and Table 4 (for the estimators of $\boldsymbol{\theta}_0^{(r)}$, $r = 1, 2$). Like Table 2, in Table 4, we present the results for $f = 0.2$ only.

As observed in the case of non-differential misclassification, in this case too, the naive MLE (NMLE) solutions either fail to converge, or converge to erratic values leading to high bias and MSE. This effect is more prominent in the differential misclassification case. Among the three estimators that utilize validation sample data, the CMLE performs the best, although for $n = 200$ the PMLE performs well reinforcing the belief that for small sample sizes PMLE may be considered to be a decent alternative to CMLE. However, for $n = 1000$, it seems that the CMLE utilizes the information in the combined sample more effectively than the PMLE, and thus, leading to much superior performance. Like non-differential misclassification, in this case too, the CMLE comes out as the preferred estimator of $\boldsymbol{\beta}_0$. In the next section, we present the asymptotic theory for the CMLE assuming differential misclassification.

3. Asymptotic properties of the CMLE

Suppose $\mathcal{X}_{n_1} = \{(Y_i, \widetilde{Y}_i, \mathbf{X}_i) : 1 \le i \le n_1\}$, $\mathcal{X}_{n_2} = \{(\widetilde{Y}_i, \mathbf{X}_i) : (n_1 + 1) \le i \le n\}$ and $\mathcal{X}_n = \mathcal{X}_{n_1} \cup \mathcal{X}_{n_2}$ (with $n_2 = n - n_1$) represent the validation, non-validation and the combined sample, respectively. The *non-validation* sample consists of the observations for which the true response Y is not recorded. It is assumed that the validation and non-validation samples are independent of each other, and the observations within each sample $\mathcal{X}_{n_i}$, $i = 1, 2$, are independent. Let $\mathcal{X}$ denote the sample space of $(Y, \widetilde{Y}, \mathbf{X})$. Define the empirical measures $\mathbb{P}_{n_1} \equiv n_1^{-1} \sum_{i=1}^{n_1} \delta_{(Y_i, \widetilde{Y}_i, \mathbf{X}_i)}$ and $\mathbb{P}_{n_2} \equiv n_2^{-1} \sum_{i=n_1+1}^{n_1+n_2} \delta_{(\widetilde{Y}_i, \mathbf{X}_i)}$, for $n_1, n_2 \ge 1$, based on the validation and non-validation samples, respectively, where $\delta_{(y,\widetilde{y},\mathbf{x})}(\cdot)$ and $\delta_{(\widetilde{y},\mathbf{x})}(\cdot)$ denote point masses at $(y, \widetilde{y}, \mathbf{x})$ and $(\widetilde{y}, \mathbf{x})$, respectively. For any measurable function g and a probability measure $\mathbf{P}$, we write $\mathbf{P}g \equiv \int g \, d\mathbf{P}$. Let $\{f_n\}$ denote the sequence of validation sampling fractions defined as

$$f_n = \frac{n_1}{n} = 1 - \frac{n_2}{n}, \quad \text{for all } n \ge 1. \tag{3.1}$$

Following the true model defined in (1.1) along with the R group differential misclassification model defined in (2.3), we define a new $(p + 2R)$ dimensional parameter $\boldsymbol{\eta} = \left(\boldsymbol{\beta}^T, \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_1^{(R)}, \theta_2^{(R)}\right)^T$, and its true value $\boldsymbol{\eta}_0 = \left(\boldsymbol{\beta}_0^T, \theta_{1,0}^{(1)}, \theta_{2,0}^{(1)}, \ldots, \theta_{1,0}^{(R)}, \theta_{2,0}^{(R)}\right)^T$. Let $\mathbf{P}_0$ denote the joint distribution of $(Y, \widetilde{Y}, \mathbf{X})$ under $\boldsymbol{\eta}_0$. Based on the combined sample $\mathcal{X}_n$, we define the CMLE $\widehat{\boldsymbol{\eta}}_n$ as

$$\widehat{\boldsymbol{\eta}}_n = \underset{\boldsymbol{\eta} \in \Delta}{\operatorname{argmax}}\ M_{n,R}(\boldsymbol{\eta}), \quad \text{where,} \quad M_{n,R}(\boldsymbol{\eta}) = f_n \cdot \mathbb{P}_{n_1} g_{1,R,\boldsymbol{\eta}}(Y, \widetilde{Y}, \mathbf{X}) + (1 - f_n) \cdot \mathbb{P}_{n_2} g_{2,R,\boldsymbol{\eta}}(\widetilde{Y}, \mathbf{X}), \tag{3.2}$$

and Δ denotes the underlying parameter space for $\boldsymbol{\eta}$ (cf. assumptions (A1) and (A3)) and

$\widehat{\boldsymbol{\eta}}_n = \left(\widehat{\boldsymbol{\beta}}_n^T, \widehat{\theta}_{1,n}^{(1)}, \widehat{\theta}_{2,n}^{(1)}, \ldots, \widehat{\theta}_{1,n}^{(R)}, \widehat{\theta}_{2,n}^{(R)}\right)^T$. The first p components of $\widehat{\boldsymbol{\eta}}_n$ is the CMLE $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$. The maps $g_{1,R,\boldsymbol{\eta}}$ and $g_{2,R,\boldsymbol{\eta}}$ are defined on the sample space of $(Y, \widetilde{Y}, \mathbf{X})$ and represent the scaled log-likelihoods of the validation and non-validation samples, respectively. They are defined as

$$g_{1,R,\boldsymbol{\eta}}(y, \widetilde{y}, \mathbf{x}) = \sum_{r=1}^{R} \mathbf{1}(\mathbf{x} \in G_r) \cdot g_{1,\boldsymbol{\beta},\theta^{(r)}}(y, \widetilde{y}, \mathbf{x}), \quad \text{and} \quad g_{2,R,\boldsymbol{\eta}}(\widetilde{y}, \mathbf{x}) = \sum_{r=1}^{R} \mathbf{1}(\mathbf{x} \in G_r) \cdot g_{2,\boldsymbol{\beta},\theta^{(r)}}(\widetilde{y}, \mathbf{x}). \tag{3.3}$$

Further, for any regression parameter $\boldsymbol{\beta}$ and a pair of misclassification probabilities $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ (cf. (2.3) or (1.2)), the maps $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}$ and $g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}$ are defined on $\mathcal{X}$ as

$$\left.\begin{aligned} g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x}) &= y\cdot\log\psi(\mathbf{x}^T\boldsymbol{\beta}) + (1-y)\cdot\log\{1-\psi(\mathbf{x}^T\boldsymbol{\beta})\} + y\widetilde{y}\cdot\log(1-\theta_2) + y(1-\widetilde{y})\cdot\log\theta_2 \\ &\quad + (1-y)(1-\widetilde{y})\cdot\log(1-\theta_1) + (1-y)\widetilde{y}\cdot\log\theta_1, \\ g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x}) &= \widetilde{y}\cdot\log h_{\boldsymbol{\beta},\boldsymbol{\theta}}(\mathbf{x}) + (1-\widetilde{y})\cdot\log\{1-h_{\boldsymbol{\beta},\boldsymbol{\theta}}(\mathbf{x})\}, \quad \text{and} \\ h_{\boldsymbol{\beta},\boldsymbol{\theta}}(\mathbf{x}) &= \theta_1 + (1-\theta_1-\theta_2)\cdot\psi(\mathbf{x}^T\boldsymbol{\beta}). \end{aligned}\right\} \tag{3.4}$$

Note that $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}$ and $g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}$ represent the log-likelihoods of the validation and non-validation samples, respectively, assuming non-differential misclassification model (1.2). For the differential misclassification model (2.3), the discrete dependence of the misclassification probabilities on the underlying covariates provides a connection to the non-differential case. As is evident from (3.3), the combined log-likelihood for the differential misclassification is a weighted sum of the log-likelihoods arising from R separate groups of covariates, and within each such group G_r, the misclassification probabilities are fixed (non-differential). The model (2.3) is flexible enough to handle nearly all types of differential misclassification. Even if misclassification rates depend on $\mathbf{X}$ in a continuous manner, one may be able to discretize the dependence to a sufficient degree, and use the groupwise misclassification model (2.3). At the end of this section we make a brief remark (see Remark 3.3) on the difficulty of handling any arbitrary differential misclassification model, from a theoretical standpoint.

Thus, there is a close connection between the non-differential and differential misclassification models. If we use, $R = 1$, $G_1 =$ sample space of $\mathbf{X}$, and write, $\boldsymbol{\theta}_0^{(1)} = \boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0})^T$, then the differential misclassification model (2.3) reduces to the non-differential case (cf. (1.2)). The CMLE $\widehat{\boldsymbol{\eta}}_n = \left(\widehat{\boldsymbol{\beta}}_n^T, \widehat{\boldsymbol{\theta}}_n^T\right)^T$ can be defined similarly as in (3.2) with the above modifications. In spite of its simplicity and the widespread use of non-differential misclassification model (1.2), as discussed in Section 1, the asymptotic properties of the CMLE have not yet been studied in the existing literature, even for the non-differential setting. Thus, first, we prove an asymptotic normality result for the CMLE assuming the non-differential misclassification model (1.2), which we state as Lemma 3.1. Building on Lemma 3.1, we obtain the main result (cf. Theorem 3.2) on the asymptotic normality of the CMLE (cf. (3.2)) in the differential misclassification case (2.3). As we will see, the technical results for proving Lemma 3.1 are indispensable for proving Theorem 3.2 (see Appendix A). In the following, we introduce some notations and technical assumptions before stating the results.

Convergence in probability to zero with respect to underlying true distribution $\mathbf{P}_0$ is denoted by $o_{\mathbf{P}_0}(1)$. Expectation with respect to $\mathbf{P}_0$ is denoted by $\mathbf{E}_0$ (or $\mathbf{P}_0$). Convergence in distribution is denoted by $\xrightarrow{d}$. Unless stated otherwise, $\mathbf{E}$ and $\mathbf{Var}$ will denote usual expectation and variance with respect to the underlying probability law. For any $\mathbf{u} = (u_1, \ldots, u_p)^T \in \mathbb{R}^p$, with $p \geq 1$, we write, $\|\mathbf{u}\| = \left(\sum_{j=1}^p u_j^2\right)^{1/2}$. All limits are taken as $n \to \infty$. We now state some technical assumptions which we will be requiring to prove our results.

(A1) There exists a bounded set $\mathbb{K} \subset \mathbb{R}^p$ with non empty interior such that the underlying regression parameter $\boldsymbol{\beta}_0 = \left(\beta_{1,0}, \ldots, \beta_{p,0}\right)^T \in \text{int}(\mathbb{K})$ (cf. (1.1)).

(A2) Let $f_{\boldsymbol{\beta}}(y, \mathbf{x})$ denote the joint density of $(Y, \mathbf{X})$ w.r.t. a dominating measure ν, when the underlying parameter is $\boldsymbol{\beta}$. We assume,

$$\sup\left\{\mathbf{P}_0 \log f_{\boldsymbol{\beta}}(Y,\mathbf{X}) : \boldsymbol{\beta} \in \mathbb{K},\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \geq \delta\right\} < \mathbf{P}_0 \log f_{\boldsymbol{\beta}_0}(Y,\mathbf{X}), \quad \text{for all } \delta > 0, \tag{3.5}$$

where, $\mathbf{P}_0$ denotes the true distribution of $(Y, \mathbf{X})$ under the parameter $\boldsymbol{\beta}_0$.

(A3) There exists constants, $0 < \delta_1 < \delta_2 < 1$, and a common parameter space Θ, such that the groupwise misclassification parameter $\boldsymbol{\theta}_0^{(r)} \in \Theta$, for all $r = 1, \ldots, R$ (see (2.3)), where

$$\Theta = \left\{(a,b) \in \mathbb{R}^2 : \delta_1 < a, b < \delta_2,\ a+b \neq 1\right\}.$$

The combined parameter space for $\boldsymbol{\eta}_0 = \left(\boldsymbol{\beta}_0^T, \theta_{1,0}^{(1)}, \theta_{2,0}^{(1)}, \ldots, \theta_{1,0}^{(R)}, \theta_{2,0}^{(R)}\right)^T$ will be denoted by Δ, where $\Delta = \mathbb{K} \times \Theta^R$.

(A4) Let $Q(\mathbf{x})$ denote the marginal distribution of the covariate vector $\mathbf{X} = (X_1, \ldots, X_p)^T$. We assume, there exists a constant $K_0 \in (0, \infty)$, such that, (i) $\|\mathbf{X}\| \leq K_0$, with probability 1, and (ii) $\mathbf{Var}(\mathbf{X})$ exists, and is positive definite.
If an intercept term is explicitly added in the true model (1.1), then we will have $X_1 = 1$, and we will require: X_j are bounded for all $j = 2, \ldots, p$, and $\mathbf{Var}\left((X_2, \ldots, X_p)^T\right)$ exists and is positive definite.

(A5) The validation sampling fractions $\{f_n : n \geq 1\}$, defined in (3.1), satisfy the following condition: there exists a constant $f \in (0, 1]$ such that, $|f_n - f| = o(1)$.

(A6) The cdf $\psi(\cdot)$ used in (1.1) is specified and satisfies the following conditions:

(i) $\psi : \mathbb{R} \mapsto (0, 1)$, is strictly increasing and $0 < \psi(u) < 1$, for all $u \in \mathbb{R}$.
(ii) The second derivative $\psi^{(2)}(\cdot)$ exists everywhere and is continuous at all points.

Assumption (A1) ensures that the parameter space $\mathbb{K}$ for the regression parameter $\boldsymbol{\beta}_0$ is bounded. Further, it is assumed that $\mathbb{K}$ has a non-empty interior and $\boldsymbol{\beta}_0$ is an interior point of $\mathbb{K}$. This allows us to obtain an open neighbourhood around $\boldsymbol{\beta}_0$. The boundedness condition on $\mathbb{K}$ and the boundedness assumption on the covariates in assumption (A4) enable us to greatly simplify the technical arguments. They ensure that $\mathbf{P}_0|g_{i,R,\boldsymbol{\eta}}| < \infty$, for $i = 1, 2$, for all $\boldsymbol{\eta} \in \Delta$, and also ensure the existence of the expectations of the first and second order derivatives of $g_{i,R,\boldsymbol{\eta}}$, $i = 1, 2$. If we drop the boundedness assumptions, then it becomes difficult to

control the behaviour of terms like $\{h_{\beta,\theta}(\mathbf{x})\}^{-1}$ and $\{1-h_{\beta,\theta}(\mathbf{x})\}^{-1}$ (cf. (3.4)), which arise while finding expected value of higher order derivatives of $\log g_{2,R,\eta}$ (see Appendix B for details).

Assumption (A2) is a high level 'well-separatedness' condition, that is required to prove consistency of M-estimators (see Chapter 5 of van der Vaart, 1998). It should be noted that assumption (A2) involves the joint density of $(Y,\mathbf{X})$, and we do not require a similar 'well-separatedness' assumption for the joint density of $(\widetilde{Y},\mathbf{X})$ or $(Y,\widetilde{Y})$. If assumption (A2) fails to hold, then the VMLE based on $(Y,\mathbf{X})$ data will even be inconsistent. Assumption (A2) can be directly established if the family of densities $\{f_\beta : \beta \in \mathbb{K}\}$ is identifiable and, (a) $\mathbf{P}_0 \log f_\beta$ is a strictly concave function of β, or (b) $\mathbb{K}$ is compact. Depending on the choice of the link function, we can use either (a) or (b). For example, the logit and probit links satisfy the strict concavity assumption,[2] and for other link functions, one can use the compactness condition. Instead of focusing on separate technical conditions required for proving (A2) in case of various link functions, we assume (A2), which is a bare necessity for consistent estimation of β_0 based on $(Y,\mathbf{X})$ data. For more details, one may refer to Newey and McFadden (1994), who provide an excellent discussion on the conditions required to establish (3.5) in various settings. For arbitrary binary link functions, unbounded parameter spaces, and unbounded covariates, in the absence of compactness and/or strict concavity, typically a modified estimating function is used to define the MLE, van de Geer (2000) provides more details on such cases.

Assumption (A3) is a standard condition that prevents misclassification rates within each group from becoming arbitrarily close to zero or unity. Later in Section 3.1, we discuss the case where one of the misclassification probabilities is known to be equal to zero. The condition $a+b \neq 1$, ensures that the non-validation components have a contribution to the log-likelihood. It should be noted that Θ is not a compact parameter space and we require each $\theta_0^{(r)}$ to be an interior point of Θ. In assumption (A4), other than the boundedness condition on $\mathbf{X}$, we require the positive definite condition on $\mathbf{Var}(\mathbf{X})$ along with monotonicity of ψ (cf. Assumption (A6)(i)) to ensure that the binary model (1.1) for $(Y,\mathbf{X})$ remains identifiable (see Chapter 5 of van der Vaart, 1998). In assumption (A5), we allow the validation sampling fraction f_n to change with n and converge to a limiting value $f \in (0,1]$. If $f=0$ then the non-validation component becomes the dominant term in the log-likelihood, and consequently our results will fail. The strict monotonicity condition along with the assumption (A4) is required for identifiability of β_0, and assumption (A6)(ii) provides the required smoothness condition on $\psi(\cdot)$.

These assumptions are similar to those used in studying asymptotic properties of MLE's in binary regression models (cf. Newey and McFadden, 1994, Amemiya, 1985, Gouriéroux and Monfort, 1981, Fahrmeir and Kaufmann, 1985 and van der Vaart, 1998). Now we state Lemma 3.1 for the non-differential case and Theorem 3.2 for the differential case.

Lemma 3.1. *Suppose, the non-differential misclassification model (1.2) holds with a fixed misclassification parameter θ_0 and assumptions (A1)-(A6) hold with $R=1$. Then, the CMLE $\widehat{\boldsymbol{\eta}}_n = \left(\widehat{\beta}_n^T, \widehat{\theta}_n^T\right)^T$ (cf. (3.2)) satisfies,*

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_n-\beta_0\\ \widehat{\theta}_n-\theta_0\end{pmatrix} \xrightarrow{d} N_{p+2}\left(\mathbf{0},\ \left[f\cdot\boldsymbol{\Sigma}_{1,0}+(1-f)\cdot\boldsymbol{\Sigma}_{2,0}\right]^{-1}\right), \tag{3.6}$$

where, $\boldsymbol{\Sigma}_{1,0}$ and $\boldsymbol{\Sigma}_{2,0}$ are described in (B.8).

Theorem 3.2. *Suppose the differential misclassification model (2.3) holds for $R>1$, and assumptions (A1)-(A6) also hold. Then, the MLE $\widehat{\boldsymbol{\eta}}_n$ defined in (3.2) satisfies,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\eta}}_n-\boldsymbol{\eta}_0\right) \xrightarrow{d} N_{p+2R}\left(\mathbf{0},\ \left[f\cdot\Gamma_{1,0}+(1-f)\cdot\Gamma_{2,0}\right]^{-1}\right),$$

where, $\Gamma_{1,0}$ and $\Gamma_{2,0}$ are defined in (A.15), (B.9) and (B.10).

Remark 3.3. Researchers (cf. Lyles et al., 2011, Bollinger and David, 1997) have considered parametric misclassification models, which depend continuously on the covariate value $\mathbf{x}$, with $\theta_{i,0}=\theta_{i,0}(\mathbf{x}:\boldsymbol{\gamma}_{i,0})$, for each $i=1,2$, and all $\mathbf{x}$, where, $\boldsymbol{\gamma}_{1,0}$ and $\boldsymbol{\gamma}_{2,0}$ are finite dimensional parameters. For example, in case of a logistic misclassification model, $\theta_{i,0}(\mathbf{x}:\boldsymbol{\gamma}_{i,0}) = \left(1+\exp\{-\mathbf{x}^T\boldsymbol{\gamma}_{i,0}\}\right)^{-1}$, for $i=1,2$. For such parametric models, the combined log-likelihood in (3.3) will depend on the parameters $(\beta,\boldsymbol{\gamma}_{1,0},\boldsymbol{\gamma}_{2,0})$, and asymptotic properties of the resulting CMLE's will depend on the structure of the log-likelihood function. Since the technical arguments and the asymptotic results will depend on the *specific* choice of the misclassification model, we do not provide general results in this case. Similar technical problems will arise if we consider nonparametric models for the misclassification probabilities.

3.1. One of the misclassification probabilities equal to zero

The misclassification probabilities are usually asymmetric. One-sided misreporting is common when programme participants self-report not receiving a treatment when they did (Lynch et al., 2007, Meyer et al., 2022). In this context, Nguimkeu et al. (2019) recently considered estimation of treatment effects when misreporting is endogeneous. An extreme case of asymmetry is, one of the misclassification probabilities is known to be zero a priori while the other is non-zero, but unknown. Kothari and Bandyopadhyay (2011) observe that the Indian Census data on literacy collected by the interview method misclassifies a large number of illiterates ($Y=0$) as literates ($\widetilde{Y}=1$), while a negligible number of literates ($Y=1$) as illiterates ($\widetilde{Y}=0$). This leads to a substantial over

[2] See Pratt (1981) for more examples of concave link functions.

estimation of literacy rates. Assuming non-differential misclassification rates, in our notation, it means $\theta_{2,0} = 0$, but $\theta_{1,0}$ is non-zero. The details are deferred to Section 4.1, where this particular data set is analysed. Such extreme asymmetric misclassification may also arise in medical diagnostic studies, Demidenko (2004) (p. 512) provides an example related to cancer detection.

In such a situation, the asymptotic results provided in Theorem 3.2 (and Lemma 3.1) are not directly applicable, because Assumption (A3) fails to hold if one the misclassification probabilities is known to be zero. However, we show that the above mentioned asymptotic results continue to hold with minor modifications.

For simplicity of presentation, we focus on the non-differential case (cf. (1.2)) with a fixed misclassification probability $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0})^T$. If we assume $\theta_{2,0} = 0$ ($\theta_{1,0} = 0$) then the misclassification parameter $\boldsymbol{\theta}_0$ reduces to $\theta_{1,0}$ ($\theta_{2,0}$). Thus, the misclassification parameter reduces to a scalar. The existing definition of the MLE in (3.2) can still be used in this setting, with $R = 1$, and the following modified definition of $g_{1,\boldsymbol{\beta},\theta}$ and $h_{\boldsymbol{\beta},\theta}$ (cf. (3.4)),

$$\left.\begin{aligned} g_{1,\boldsymbol{\beta},\theta}(y,\widetilde{y},\mathbf{x}) &= \begin{cases} \log\left[\left(\psi(\mathbf{x}^T\boldsymbol{\beta})\right)^y \cdot \left(1-\psi(\mathbf{x}^T\boldsymbol{\beta})\right)^{(1-y)}\right] + \log\left[(1-\theta_1)^{(1-y)(1-\widetilde{y})} \cdot {\theta_1}^{(1-y)\widetilde{y}}\right] & \text{if } \theta_{2,0} = 0, \\ \log\left[\left(\psi(\mathbf{x}^T\boldsymbol{\beta})\right)^y \cdot \left(1-\psi(\mathbf{x}^T\boldsymbol{\beta})\right)^{(1-y)}\right] + \log\left[(1-\theta_2)^{y\widetilde{y}} \cdot {\theta_2}^{y(1-\widetilde{y})}\right] & \text{if } \theta_{1,0} = 0, \end{cases} \\ \text{and,} \quad h_{\boldsymbol{\beta},\theta}(\mathbf{x}) &= \begin{cases} \theta_1 + (1-\theta_1)\cdot\psi(\mathbf{x}^T\boldsymbol{\beta}) & \text{if } \theta_{2,0} = 0, \\ (1-\theta_2)\cdot\psi(\mathbf{x}^T\boldsymbol{\beta}) & \text{if } \theta_{1,0} = 0. \end{cases} \end{aligned}\right\} \tag{3.7}$$

The definition of $g_{2,\boldsymbol{\beta},\theta}$ remains the same with the above definition of $h_{\boldsymbol{\beta},\theta}$. The CMLE's $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\theta}_{1,n}$ are defined similarly as in (3.2) (with $R = 1$). In case $\theta_{2,0}$ or $\theta_{1,0}$ is known to be zero, then assumption (A3) needs to be slightly modified.

(A3*) If $\theta_{2,0} = 0$, we assume that there exist constants $\delta_1, \delta_2 \in (0,1)$, such that, $\delta_1 < \theta_{1,0} < \delta_2$. A similar condition applies on $\theta_{2,0}$, in case it is known that $\theta_{1,0} = 0$.

Corollary 3.4. *The result is stated separately for the non-differential and differential cases.*

(i) Consider the non-differential case with $R = 1$ (cf. (1.2)) and assume all conditions for Lemma 3.1 hold, except (A3) is replaced by (A3*). Then the CMLE's satisfies

$$\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \\ \widehat{\theta}_{i,n} - \theta_{i,0} \end{pmatrix} \xrightarrow{d} N_{p+1}\left(\mathbf{0},\ \left[f\cdot\boldsymbol{\Sigma}_{(-i),1,0} + (1-f)\cdot\boldsymbol{\Sigma}_{(-i),2,0}\right]^{-1}\right), \quad \text{for both } i = 1,2,$$

where, $\boldsymbol{\Sigma}_{(-i),1,0}$ and $\boldsymbol{\Sigma}_{(-i),2,0}$ are obtained by removing the $(p+i)$-th row and column from $\boldsymbol{\Sigma}_{1,0}$ and $\boldsymbol{\Sigma}_{2,0}$, respectively, for $i = 1,2$.

(ii) In case $R > 1$, and the differential misclassification model (2.3) holds and all assumptions of Theorem 3.2 hold. Then the distribution convergence result of Theorem 3.2 holds for all active misclassification parameters, and the rows and columns corresponding to inactive (zero) misclassification parameters are dropped from the expression of $\Gamma_{1,0}$ and $\Gamma_{2,0}$ (see Theorem 3.2), in the same manner as shown in part (i) above.

3.2. Bootstrapped CMLE

Here, we define a bootstrapped version of the CMLE and provide an approximation to its asymptotic distribution stated in Theorem 3.2. Note that for drawing inference on $\boldsymbol{\beta}_0$ using the asymptotic distribution of the CMLE the asymptotic variance matrix (cf. Appendix B) needs to be evaluated analytically term-by-term. This is typically a tedious task, and especially difficult for the practitioners to implement because of its complicated form. A bootstrap algorithm to approximate the limiting distribution of the CMLE as given in Theorem 3.2, may facilitate easy implementation of inference procedures for $\boldsymbol{\beta}_0$, like obtaining bootstrap-confidence intervals (CIs) for the underlying model parameters.

There can be several approaches to select a bootstrap sample in this set up. However, we focus on the usual with replacement n-out-of-n bootstrap (Efron, 1979). For bootstrapping, we select independent random samples with replacement of sizes n_1 and n_2 from the validation, and the non-validation samples, respectively. We denote these bootstrap samples as, $\mathcal{X}^\star_{n_1} = \left\{\left(Y_i^\star, \widetilde{Y}_i^\star, \mathbf{X}_i^\star\right) : 1 \le i \le n_1\right\}$ and $\mathcal{X}^\star_{n_2} = \left\{\left(\widetilde{Y}_i^\star, \mathbf{X}_i^\star\right) : (n_1+1) \le i \le n\right\}$. The bootstrap sample based CMLE's are defined similarly as in (3.2) using the observations in $\mathcal{X}^\star_{n_1}$ and $\mathcal{X}^\star_{n_2}$,

$$\left.\begin{aligned} \boldsymbol{\eta}_n^\star &= \underset{\boldsymbol{\eta}\in\Delta}{\operatorname{argmax}}\ M_n^\star(\boldsymbol{\eta}), \quad \text{where,} \\ M_n^\star(\boldsymbol{\eta}) &= \left[f_n\cdot\frac{1}{n_1}\sum_{i\le n_1} g_{1,R,\boldsymbol{\eta}}\left(Y_i^\star, \widetilde{Y}_i^\star, \mathbf{X}_i^\star\right) + (1-f_n)\cdot\frac{1}{n_2}\sum_{i>n_1} g_{2,R,\boldsymbol{\eta}}(\widetilde{Y}_i^\star, \mathbf{X}_i^\star)\right], \end{aligned}\right\} \tag{3.8}$$

where, $g_{1,R,\boldsymbol{\eta}}$ and $g_{2,R,\boldsymbol{\eta}}$, are defined in (3.3). One can show that the scaled and centred bootstrapped CMLE's, $\sqrt{n}(\boldsymbol{\eta}_n^\star - \widehat{\boldsymbol{\eta}}_n)$, will converge in distribution (conditionally in probability $\mathbf{P}_0$) to the limiting distribution stated in Theorem 3.2. The proof follows by using results on bootstrap consistency for M-estimators and by showing that underlying log-likelihood function classes are Donsker (see proof of Lemma 5.1). We skip the details, since the arguments are routine (cf. Cheng and Huang, 2010, Wellner and Zhan, 1996 or Gine, 1997).

Following the confidence intervals (CIs) described in (4.1), the level $(1-\alpha)$ bootstrap CI for the parametric function $\mathbf{c}^T\boldsymbol{\eta}_0$ (for any $\mathbf{c}\in\mathbb{R}^{p+2}$) is,

$$\left[\mathbf{c}^T\widehat{\boldsymbol{\eta}}_n-\frac{\xi^{\star}_{n,1-\alpha/2}}{\sqrt{n}},\ \mathbf{c}^T\widehat{\boldsymbol{\eta}}_n-\frac{\xi^{\star}_{n,\alpha/2}}{\sqrt{n}}\right], \tag{3.9}$$

where, $\xi^{\star}_{n,\alpha}$ denotes the α-quantile of $\mathbf{c}^T\sqrt{n}\left(\boldsymbol{\eta}^{\star}_n-\widehat{\boldsymbol{\eta}}_n\right)$. In the next section, we compare the performance of the bootstrap percentile CIs with the corresponding asymptotic CIs based on Theorem 3.2.

4. Simulation results and real-data analysis

In this section, we conduct a simulation experiment to study the empirical coverage and average length of bootstrapped CMLE based confidence intervals (CIs) (cf. (3.9)) and also, compare them with the CIs based on the CMLE (with asymptotic covariance matrix), and VMLE. The simulation setting is same as the one used in Section 2.2 for non-differential misclassification, and Section 2.3 for differential misclassification with two groups. The non-differential and differential misclassification cases are considered separately. In the following, we give the details of the estimators and the confidence intervals considered in each of these two cases along with the details of the simulation parameter settings.

(S.1) **Non-differential misclassification**: We compare CMLE based asymptotic CIs (cf. (4.1)), CMLE based bootstrap percentile CIs (cf. (3.9)) and VMLE based asymptotic CIs. The sample size is fixed at $n=1000$. We use $f\in\{0.05,0.1,0.2,0.3,0.9\}$ and CIs are constructed for three different model parameters, *viz.*, $\beta_{1,0}$, $\beta_{2,0}$ and $(\beta_{1,0}-\beta_{2,0})$. The results are presented in Table 5, and they are based on 300 Monte Carlo replications.

(S.2) **Differential misclassification**: We consider $R=2$ groups (cf. (2.3)). As stated above, evaluation of CMLE based asymptotic CIs involve cumbersome variance expressions (see Appendix B). Thus, only bootstrap percentile based CIs (cf. (3.9)) are studied. We consider $f\in\{0.1,0.2,0.3\}$ and the choice of $f=0.05$ is dropped to avoid numerical convergence issues. Other than the three parameters described in (S.1) above (for non-differential case), we also consider inference on the unknown misclassification parameters. Two different sample sizes are used: $n=200$ and $n=1000$. The simulation results are presented in Table 6 and they are based on 150 Monte Carlo replications.

If we write $\boldsymbol{\eta}_0=(\boldsymbol{\beta}^T,\boldsymbol{\theta}_0^T)^T$, and $\widehat{\boldsymbol{\eta}}_n=(\widehat{\boldsymbol{\beta}}_n^T,\widehat{\boldsymbol{\theta}}_n^T)^T$, then for any $\mathbf{c}\in\mathbb{R}^{p+2}$, a level $(1-\alpha)$ asymptotic confidence interval (CI) for the parametric function $\mathbf{c}^T\boldsymbol{\eta}_0$ is,

$$\left[\mathbf{c}^T\widehat{\boldsymbol{\eta}}_n-\frac{\tau_{1-\alpha/2}}{\sqrt{n}}\cdot\sqrt{\mathbf{c}^T\widehat{A}_n\mathbf{c}},\ \mathbf{c}^T\widehat{\boldsymbol{\eta}}_n-\frac{\tau_{\alpha/2}}{\sqrt{n}}\cdot\sqrt{\mathbf{c}^T\widehat{A}_n\mathbf{c}}\right], \tag{4.1}$$

where, $\widehat{A}_n$ is a plugin estimate of the asymptotic covariance matrix given in Lemma 3.1, and τ_α denotes the α-quantile of standard normal distribution. The asymptotic covariance matrix depends on $\boldsymbol{\beta}_0,\boldsymbol{\theta}_0$ and the unknown distribution of the covariates. In practice, a plugin estimate $\widehat{A}_n$ can be obtained by plugging in $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\theta}}_n$ in place of the unknown true parameters, and by using the empirical distribution of the covariates, in place of true distribution. The asymptotic CIs described in (4.1) and used in Table 5 are obtained in this manner.

As seen from the results in Table 5, in the non-differential misclassification setting, the CMLE based asymptotic and bootstrap percentile CIs both attain coverages close to the nominal one (95%). The bootstrap percentile CIs have slightly larger length than their asymptotic counterparts. Quite expectedly, the VMLE based asymptotic CIs attain close to nominal coverage but have substantially larger lengths. This indicates that the CMLE utilizes the main sample and the validation sample effectively.

In the differential misclassification setting, the results in Table 6 (for $n=1000$) suggest that CMLE based bootstrap percentile CIs for the primary model parameters (cf. (1.1)) attain close to nominal coverage. However for $n=200$, at $f=0.1$ the performance of bootstrapped CMLE is marginally weaker, as f increases, the performance improves. In case of the misclassification parameters $\boldsymbol{\theta}_0^{(r)}$, $r=1,2$, except for the parameter $\theta_{1,0}^{(1)}$, the performance of the CMLE is satisfactory, at both choices of n. The reason for the poor coverage of CIs for $\theta_{1,0}^{(1)}$ is not clear and needs further investigation.

Remark 4.1. Following the suggestion made by one of the reviewers, in this remark we briefly explain the rationale behind the choices of estimators and sample sizes used in the simulation results displayed in . The primary goal is to focus on the performance of the bootstrapped CMLE based CIs against other competing CIs, in both differential and non-differential settings, at different choices of n and f.

Table 5 focuses on the non-differential case and compares the bootstrapped CMLE based CIs against two other competing CIs, which use asymptotic normality to construct the confidence intervals. Thus, a large sample size ($n=1000$) is used in order to extract optimum performance from these competing CIs. We expect, on the basis of results obtained in Table 6, that for $n=200$, and in the non-differential setting, the bootstrapped CMLE based CIs will continue to have good coverage. The NMLE is not considered as a competing estimator, as it is not expected to perform well in view of the poor bias and MSE figures presented in Section 2.2. The case of $f=0.9$ in Table 5 is considered mainly to see the effect of the quantum of validation data on the performance of CMLE. It may be noticed (in Table 5) that the performance of CMLE between the cases $f=0.3$ and $f=0.9$ is not significantly different, at least in terms of coverage accuracy.

Table 5
Empirical coverages and average lengths (in parenthesis) for various types of CIs for underlying model parameters, in the **non-differential misclassification** setting described in (S.1) in Section 4; at $n = 1000$ and various choices of f. The nominal coverage level is 0.95.

Parameter	Validation sample fraction f				
	0.05	0.1	0.2	0.3	0.9
	CMLE based asymptotic CI				
$\beta_{1,0}$	0.955	0.94	0.945	0.93	0.94
	(0.989)	(0.763)	(0.578)	(0.488)	(0.323)
$\beta_{2,0}$	0.935	0.95	0.945	0.95	0.93
	(0.807)	(0.653)	(0.547)	(0.490)	(0.358)
$(\beta_{1,0} - \beta_{2,0})$	0.955	0.965	0.935	0.94	0.935
	(1.576)	(1.227)	(0.957)	(0.823)	(0.557)
Parameter	CMLE based bootstrap percentile CI				
$\beta_{1,0}$	0.96	0.947	0.940	0.967	0.967
	(1.110)	(0.808)	(0.588)	(0.494)	(0.320)
$\beta_{2,0}$	0.906	0.967	0.973	0.926	0.96
	(1.016)	(0.769)	(0.572)	(0.502)	(0.359)
$(\beta_{1,0} - \beta_{2,0})$	0.934	0.940	0.940	0.953	0.96
	(1.887)	(1.369)	(0.992)	(0.842)	(0.553)
Parameter	VMLE based asymptotic CI				
$\beta_{1,0}$	0.975	0.97	0.935	0.94	0.935
	(1.523)	(1.030)	(0.714)	(0.576)	(0.329)
$\beta_{2,0}$	0.96	0.965	0.96	0.945	0.95
	(1.757)	(1.154)	(0.798)	(0.642)	(0.365)
$(\beta_{1,0} - \beta_{2,0})$	0.985	0.96	0.95	0.96	0.94
	(2.720)	(1.801)	(1.242)	(0.997)	(0.566)

Table 6
Empirical coverages and average lengths (in parenthesis) for **CMLE based bootstrap percentile** CIs for underlying model parameters, in the **differential misclassification** setting described in (S.2) in Section 4; at various choices of f. The nominal coverage level is 0.95.

n	f	Parameter						
		$\beta_{1,0}$	$\beta_{2,0}$	$(\beta_{1,0} - \beta_{2,0})$	$\theta^{(1)}_{1,0}$	$\theta^{(1)}_{2,0}$	$\theta^{(2)}_{1,0}$	$\theta^{(2)}_{2,0}$
200	[a]0.1	0.882	0.882	0.824	0.224	0.706	0.447	0.647
		(2.51)	(3.248)	(5.004)	(0.533)	(0.263)	(0.396)	(0.395)
	0.2	0.933	0.947	0.893	0.260	0.773	0.787	0.853
		(1.57)	(1.86)	(2.86)	(0.427)	(0.225)	(0.341)	(0.352)
	0.3	0.987	0.953	0.933	0.413	0.833	0.813	0.887
		(1.286)	(1.552)	(2.388)	(0.368)	(0.206)	(0.304)	(0.313)
1000	0.1	0.960	0.913	0.933	0.507	0.933	0.847	0.907
		(0.787)	(0.847)	(1.363)	(0.265)	(0.129)	(0.230)	(0.204)
	0.2	0.940	0.953	0.953	0.733	0.887	0.933	0.933
		(0.582)	(0.622)	(0.992)	(0.213)	(0.105)	(0.178)	(0.162)
	0.3	0.960	0.960	0.940	0.693	0.933	0.940	0.940
		(0.489)	(0.538)	(0.844)	(0.192)	(0.091)	(0.149)	(0.139)

[a] In case of $f = 0.1$, 85 Monte-Carlo replications were used.

In Table 6, the differential misclassification case is considered at both $n = 200$ and $n = 1000$. The VMLE is not considered, as we are also interested in CIs for the misclassification parameters. Neither the asymptotic CIs based on CMLE are considered, as they would be difficult to compute in this setting.

The PMLE (cf. Section 2.2) was not included in any table presented in Section 4. It is expected that the PMLE should have a *higher* asymptotic variance than the CMLE (cf. Chatterjee et al., 2016 for details on the PMLE). However, as suggested by , the PMLE may be comparable to the CMLE, especially in small samples. We do not explore this further in this article.

4.1. Real data analysis

In this section, we illustrate our methodology by using a household literacy survey data collected from four Indian states, and draw inference on the parameters of interest. For a detailed description of the survey, we refer to Kothari and Bandyopadhyay (2011).

One of the goals of the survey is to assess the bias in the literacy rate computed from the census data by collecting a sample of gold standard data. In the census method, the literacy status ($\widetilde{Y}$) of an individual in a household is recorded as reported by the head of the household. For an individual $\widetilde{Y} = 1$, if the individual is reported literate, and $\widetilde{Y} = 0$, otherwise. On the other hand, in

Table 7
CMLE estimates and 95% asymptotic and bootstrap CIs for model parameters for the literacy data example, based on a single sample of size $n = 740$. The simulated performance of CMLE based 95% bootstrap percentile CIs are shown in the last two columns, by repeatedly drawing sub-samples of size $n = 740$. We used $f = 0.2$ in all cases.

Parameter	CMLE	Asymp. CI	Bootstrap CI	Performance of bootstrap CIs[a]	
				Emp. Cov.	Avg. Length
$\beta_{1,0}$	1.907	(1.288, 2.526)	(1.278, 2.516)	0.805	1.240
$\beta_{2,0}$	−0.072	(−0.090, −0.050)	(−0.094, −0.050)	0.840	0.045
$\theta_{1,0}$	0.199	(0.128, 0.270)	(0.133, 0.275)	0.940	0.139

[a] Based on 200 Monte-Carlo sub-samples of size $n = 740$ from the population.

the gold standard data, the literacy status (Y) of each individual is collected by giving a proper written/oral test. In this context, it is worth mentioning that the census data on the literacy status of an individual is subject to substantial misclassification.

Data on 7409 individuals within the age group 15–45 are collected in the survey using census method followed by a written/oral test. The chance of misclassifying a literate person ($Y = 1$) as illiterate by the head of the household ($\widetilde{Y} = 0$) is found to be nearly zero, and thus, we assume the chance to be zero. For each individual, age (X) is also recorded. We treat it as a covariate.

Complete information on the triplet, $(Y, \widetilde{Y}, X)$ is available for all 7409 individuals. From this data, we find $\theta_{1,0} = \mathbf{P}(\widetilde{Y} = 1 \mid Y = 0) = 0.177$. We assume a linear logistic regression model as shown below,

$$\psi_L(\mathbf{x}^T \boldsymbol{\beta}_0) = \frac{1}{1 + \exp\{-\beta_{1,0} - \beta_{2,0} \cdot x\}}, \quad \text{for all } x. \tag{4.2}$$

Based on complete information over 7409 individuals, the *true* parameter values are found to be, $\beta_{1,0} = 1.66$, $\beta_{2,0} = -0.0623$. However, in practice such detailed information would be unavailable. For our data-analysis, we select a random sample of size $n = 740$ from the complete set (about 10% of the available number of observations), out of which the first $n_1 = 148$ observations are treated as the validation sample. This leads to a validation sample fraction $f \approx 0.2$. Table 7 shows the CMLE based estimates and the asymptotic and bootstrap percentile CIs, which are obtained from this sample of $n = 740$ individuals. Both types of CIs are performing similarly, and contain the *true* population parameters. Both bootstrap and asymptotic CIs for the slope parameter indicates that, plausibly with increase in age, the chances of a person being literate decreases, which is true in general for the Indian population. Following the suggestion made by one of the reviewers, we have also checked the *actual* the coverage accuracy of the bootstrap percentile CIs. In order to do this, samples of size $n = 740$ are repeatedly selected from the population of 7409 individuals, and the corresponding bootstrap percentile CIs are computed. The empirical coverage and average lengths of these bootstrap based CIs are evaluated. The last two columns of Table 7 show the results. It is seen that there is undercoverage in case $\beta_{1,0}$ and $\beta_{2,0}$, but coverage for $\theta_{1,0}$ attains its nominal value. The under coverage of the CI, contradicting the results obtained in the simulation experiments, may be caused by the lack of goodness of fit of the assumed logistic model to the data.

5. Concluding remarks

The binary regression model (1.3) is used extensively, especially by the researchers from other disciplines, to handle misclassified responses. This article addresses two issues related to this model. First, the model suffers from a serious estimation problem because of confounding of the regression parameters with the misclassification probabilities. We discuss this issue in detail and provide statistical insights into the phenomenon. Through a numerical study we show that the confounding may lead to a hugely biased estimator of the regression parameter. Second, in many practical situations the non-differential misclassification model (1.2) may not be appropriate. We propose a differential misclassification model which subsumes the non-differential misclassification model (1.2). To overcome the confounding problem, it is essential to gather extra information on the misclassification probabilities. To capture this extra information, we propose use of internal validation sample in addition to the main sample, and the maximum likelihood estimator (CMLE) of the regression parameter based on the combined sample. We then develop a rigorous asymptotic theory for CMLE. It is worthwhile to mention that PMLE performs nearly as good as CMLE for smaller sample sizes. For researchers, it seems to be a decent alternative especially when CMLE is infeasible because of convergence problem or validation data may not be available. For its easy implementation, a bootstrap approximation to the asymptotic distribution is proposed, and we prove its consistency. The results of the numerical studies show that, in terms of both coverage and expected length, the bootstrap confidence intervals of the regression parameters match very well with that of the asymptotic intervals. Finally, the proposed methodology is illustrated using a real data set.

We feel that the present work is a significant methodological contribution to the existing literature on binary regression models with misclassified response. To be precise, the present work is the first to develop a likelihood-based methodology backed by a rigorous asymptotic theory for drawing inference on the regression parameter with misclassified response using internal validation data. Our results are quite general. First, our methodology considers differential misclassification and, second, it is valid for any specified and strictly increasing cdf $\psi(\cdot)$ on the real line, provided it satisfies certain mild regularity conditions.

Acknowledgements

The first author's research is partially supported by SERB-MATRICS grant MTR/2017/000224. The authors wish to thank two anonymous reviewers for their constructive comments and suggestions which helped to improve this article. The authors also wish to thank Professor A. W. van der Vaart for a helpful comment.

Appendix A. Proofs of main and auxiliary results

Proofs of main results are provided. Some auxiliary Lemma's which are required to prove the main results, are also stated and proved. We define the scaled and centred empirical processes, $\mathbb{G}_{n_i} \equiv \sqrt{n_i}\left(\mathbb{P}_{n_i} - \mathbf{P}_0\right)$, for $i = 1, 2$. For any two $x, y \in \mathbb{R}$, $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. Also, $\|\mathbf{u}\|$ denotes usual Euclidean norm for any $\mathbf{u} \in \mathbb{R}^k$, for any $k \geq 1$. We will write $\mathcal{X}$ to denote the sample space for $(Y, \widetilde{Y}, \mathbf{X})$.

Initially we consider the non-differential misclassification model (1.2) and provide a proof of Lemma 3.1 for the non-differential case. The proof of Theorem 3.2 (for the differential misclassification model) will follow on similar lines.

Lemma 5.1 (*Consistency of the CMLE for Non-Differential Misclassification*). *Suppose, assumptions (A1)-(A6) hold, the true model is given by (1.1) and also assume that the non-differential misclassification model (1.2) is valid. Then, the CMLE's $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\theta}}_n)$ (with $R = 1$) defined in (3.2) are consistent and satisfy, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_{\mathbf{P}_0}(1)$.*

Proof of Lemma 5.1. The proof consists of verifying three sufficient conditions needed for showing consistency of M-estimators, as stated in Theorem 5.7 of van der Vaart (1998).

Consider the class of functions $\{\mathbf{x} \mapsto \mathbf{x}^T\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{K}\}$. This is a finitely generated class of functions and hence it is a VC class. Now, $u : \mathbb{R} \mapsto \psi(u)$, and $u : \mathbb{R} \mapsto 1 - \psi(u)$, are monotone mappings, and $u : (0, 1) \mapsto \log u$, is also monotone. Also, $(y, \widetilde{y}, \mathbf{x}) \mapsto y$, and $(y, \widetilde{y}, \mathbf{x}) \mapsto (1 - y)$, are fixed maps on the underlying sample space. Thus, using VC class preservation properties we can claim, the classes, $\{(y, \widetilde{y}, \mathbf{x}) \mapsto y \cdot \log \psi(\mathbf{x}^T\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ and $\{(y, \widetilde{y}, \mathbf{x}) \mapsto (1 - y) \cdot \log \psi(\mathbf{x}^T\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$, will be both VC classes. The boundedness assumptions in (A1) and (A3) imply, there exists a $C_0 = \text{diam}(\mathbb{K}) \cdot K_0 \in (0, \infty)$, such that $|\mathbf{x}^T\boldsymbol{\beta}| \leq C_0$, for all $\boldsymbol{\beta} \in \mathbb{K}$ and $\mathbf{x} \in \text{support}(Q)$. As ψ is a strictly increasing cdf on $\mathbb{R}$, there exists some $c_0 \in (0, 1)$, such that, $c_0 < \psi(-C_0) \leq \psi(\mathbf{x}^T\boldsymbol{\beta}) \leq \psi(C_0) < 1 - c_0$. This implies, $\log \psi(\mathbf{x}^T\boldsymbol{\beta}) \in (\log c_0, \log(1 - c_0))$, for each $\mathbf{x}$ in the support of Q and all $\boldsymbol{\beta} \in \mathbb{K}$. Thus, the two aforementioned classes have bounded uniform entropy integral (BUEI) (cf. Chapter 9 of Kosorok, 2008) with respect to the constant envelope function, $C_1 \equiv \max\{|\log c_0|, |\log(1 - c_0)|\}$. The boundedness assumptions play an extremely critical role in obtaining this envelope function and allows us to work with the objective function $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}$. Sums (and subsets) of BUEI classes are BUEI, with respect to the sum of corresponding envelope functions. Hence the class, $\{(y, \widetilde{y}, \mathbf{x}) \mapsto y \cdot \log \psi(\mathbf{x}^T\boldsymbol{\beta}) + (1 - y) \cdot \log \psi(\mathbf{x}^T\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ will be a BUEI class with respect to the envelope function $2C_1$.

Recall that $\mathcal{X}$ denotes the sample space of $(Y, \widetilde{Y}, \mathbf{X})$. Consider the class, $\{u_{(\boldsymbol{\beta},\boldsymbol{\theta})}(y, \widetilde{y}, \mathbf{x}) = \theta_1 : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$, of constant functions. The subgraph of $u_{(\boldsymbol{\beta},\boldsymbol{\theta})}$ is $A(\boldsymbol{\beta}, \boldsymbol{\theta}) = \{(y, \widetilde{y}, \mathbf{x}, t) \in \mathcal{X} \times \mathbb{R} : t < u_{\boldsymbol{\beta},\boldsymbol{\theta}}(y, \widetilde{y}, \mathbf{x})\} = \{(y, \widetilde{y}, \mathbf{x}, t) : t < \theta_1\} = \mathcal{X} \times (-\infty, \theta_1)$, for each $\boldsymbol{\beta} \in \mathbb{K}$ and $\boldsymbol{\theta} \in \Theta$. Consider any two points $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{X} \times \mathbb{R}$. Note, each $\mathbf{v}_i$, $i = 1, 2$, are $(p + 3)$ dimensional vectors of the form, $(y_i, \widetilde{y}_i, \mathbf{x}_i^T, a_i)^T$, where $a_i < \theta_1$. Consider the $(p+3)$-th component for each $\mathbf{v}_i$, for $i = 1, 2$. Without loss of generality, assume $a_1 \leq a_2$. If $a_1 = a_2$, then there does not exist any $A(\boldsymbol{\beta}, \boldsymbol{\theta})$, such that $A(\boldsymbol{\beta}, \boldsymbol{\theta}) \cap \{\mathbf{v}_1, \mathbf{v}_2\} = \{\mathbf{v}_1\}$ (or $\{\mathbf{v}_2\}$). In case, $a_1 < a_2$ and $a_1 < \delta_2$ (cf. assumption (A2)), we can find a set $A(\boldsymbol{\beta}, \boldsymbol{\theta})$ (or equivalently, a choice of θ_1), such that $A(\boldsymbol{\beta}, \boldsymbol{\theta}) \cap \{\mathbf{v}_1, \mathbf{v}_2\} = \{\mathbf{v}_1\}$. But, there does not exist any $A(\boldsymbol{\beta}, \boldsymbol{\theta})$, such that $A(\boldsymbol{\beta}, \boldsymbol{\theta}) \cap \{\mathbf{v}_1, \mathbf{v}_2\} = \{\mathbf{v}_2\}$. And if, $\delta_2 \leq a_1$, then there does not even exist any $A(\boldsymbol{\beta}, \boldsymbol{\theta})$, such that $A(\boldsymbol{\beta}, \boldsymbol{\theta}) \cap \{\mathbf{v}_1, \mathbf{v}_2\} = \{\mathbf{v}_1\}$ (and same conclusion holds for $\{\mathbf{v}_2\}$). As a result, any two point set $\{\mathbf{v}_1, \mathbf{v}_2\} \in \mathcal{X} \times \mathbb{R}$ cannot be shattered by this class of subgraphs. Hence this class of subgraphs is a VC subgraph class, and by definition, $\{u_{(\boldsymbol{\beta},\boldsymbol{\theta})}(y, \widetilde{y}, \mathbf{x}) = \theta_1 : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ becomes a VC class with VC index equal to 2. As log is a fixed monotone map and $(y, \widetilde{y}, \mathbf{x}) \mapsto (1 - y)\widetilde{y}$, is a fixed mapping not depending on $(\boldsymbol{\beta}, \boldsymbol{\theta})$, thus the class of functions $\{(y, \widetilde{y}, \mathbf{x}) \mapsto (1 - y)\widetilde{y} \cdot \log \theta_1 : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ becomes a VC class. Assumption (A2) ensures, $\log \delta_1 < \log \theta_1 < \log \delta_2$, for all $\boldsymbol{\theta} \in \Theta$. Thus the aforementioned class has a constant envelope function $C_2 \equiv \max\{|\log \delta_1|, |\log \delta_2|, |\log(1 - \delta_1)|, |\log(1 - \delta_2)|\}$. This implies it will be a BUEI class with respect to this envelope function.

Similar arguments can be used to claim that the classes, $\{y\widetilde{y} \cdot \log(1 - \theta_2) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$, $\{y(1 - \widetilde{y}) \cdot \log \theta_2 : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ and $\{(1 - y)(1 - \widetilde{y}) \cdot \log(1 - \theta_1) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ are also BUEI classes with respect to similarly defined constant envelope functions. This ensures that the class of maps $\{g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}(y, \widetilde{y}, \mathbf{x}) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ is a BUEI class with respect to the constant envelope function $2C_1 + 4C_2$. This class can be shown as pointwise measurable, by considering the subclass indexed by all rational $\boldsymbol{\beta} \in \mathbb{K}$ and $\boldsymbol{\theta} \in \Theta$. The constant envelope function will be trivially square integrable (with respect to $\mathbf{P}_0$). Thus, $\{g_{1,\boldsymbol{\beta},\boldsymbol{\theta}} : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ will be a $\mathbf{P}_0$-Donsker class. Exactly similar arguments can be used to show that $\{g_{2,\boldsymbol{\beta},\boldsymbol{\theta}} : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\}$ will be a $\mathbf{P}_0$-Donsker class. As Donsker classes are Glivenko–Cantelli (GC), this implies,

$$\sup\left\{\left|\left(\mathbb{P}_{n_i} - \mathbf{P}_0\right) g_{i,\boldsymbol{\beta},\boldsymbol{\theta}}\right| : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\right\} = o_{\mathbf{P}_0}(1), \quad \text{for } i = 1, 2.$$

Now define,

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}) \equiv f \cdot \mathbf{P}_0 g_{1,\boldsymbol{\beta},\boldsymbol{\theta}} + (1 - f) \cdot \mathbf{P}_0 g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}. \tag{A.1}$$

Note that $M(\boldsymbol{\beta}, \boldsymbol{\theta})$ (cf. (A.1)) is well defined, because the boundedness assumptions in (A1) and (A3) ensure that $\mathbf{P}_0|g_{i,\boldsymbol{\beta},\boldsymbol{\theta}}| < \infty$, for $i = 1, 2$, for all $\boldsymbol{\beta} \in \mathbb{K}$ and $\boldsymbol{\theta} \in \Theta$. Also note that,

$$\sup\left\{\left|\mathbf{P}_0 g_{i,\boldsymbol{\beta},\boldsymbol{\theta}}\right| : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta\right\} < \infty, \quad \text{for } i = 1, 2, \tag{A.2}$$

because $\mathbb{K}$ and Θ are bounded sets. Assumption (A3) now ensures that,

$$\sup_{\boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\beta}, \boldsymbol{\theta}) - M(\boldsymbol{\beta}, \boldsymbol{\theta})| = o_{\mathbf{P}_0}(1),$$

which completes the verification of the first sufficient condition. Also note, by definition of $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\theta}}_n)$, we have $M_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\theta}}_n) \geq M_n(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) - o_{\mathbf{P}_0}(1)$, which verifies the second sufficient condition. To verify the third and final sufficient condition for showing consistency, we need to check the 'well-separatedness' condition, *i.e.*, show that,

$$\sup\left\{M(\boldsymbol{\beta}, \boldsymbol{\theta}) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \geq \delta^2\right\} < M(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0), \quad \text{for any } \delta > 0,$$

The supremum above is taken outside an open ball of radius δ around $(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$. In order to establish the 'well-separatedness' condition, let us consider $\mathbf{P}_0 g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}$ and $\mathbf{P}_0 g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}$ separately. Note, we can write,

$$\begin{aligned}
\mathbf{P}_0 g_{1,\boldsymbol{\beta},\boldsymbol{\theta}} &= \mathbf{P}_0\left[Y \cdot \log \psi(\mathbf{X}^T\boldsymbol{\beta}) + (1-Y)\cdot \log(1-\psi(\mathbf{X}^T\boldsymbol{\beta}))\right] \\
&\quad + \mathbf{P}_0\left[Y\widetilde{Y}\cdot \log(1-\theta_2) + Y(1-\widetilde{Y})\cdot \log\theta_2 + (1-Y)(1-\widetilde{Y})\cdot\log(1-\theta_1) + (1-Y)\widetilde{Y}\cdot\log\theta_1\right] \\
&= \mathbf{P}_0 \log f_{\boldsymbol{\beta}}(Y,\mathbf{X}) + \mathbf{P}_0 \log f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}}(Y,\widetilde{Y},\mathbf{X}),
\end{aligned} \tag{A.3}$$

where, $f_{\boldsymbol{\beta}}(y,\mathbf{x})$ denotes the joint density of $(Y,\mathbf{X})$ with respect to a dominating measure ν (see van de Geer, 2000 or Example 1.2 of Newey and McFadden, 1994 for further details), when the underlying parameter is $(\boldsymbol{\beta},\boldsymbol{\theta})$. Similarly let $f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x})$ denote the joint density of $(Y,\widetilde{Y},\mathbf{X})$, when the underlying parameter is $(\boldsymbol{\beta}_0,\boldsymbol{\theta})$. This is feasible because, the conditional law of $[\widetilde{Y} \mid Y,\mathbf{X}]$ does not depend on $\mathbf{X}$ as per the misclassification rule given in (1.2). The conditional density of $[\widetilde{Y} \mid Y]$ does not depend on $\boldsymbol{\beta}$, but the marginal law of Y (which depends on $\mathbf{X}$) depends only on $\boldsymbol{\beta}_0$.

Under assumption (A3), which ensures $\mathbf{Var}(\mathbf{X})$ is positive definite, it can be checked that the family of joint distributions $\{f_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathbb{K}\}$ for $(Y,\mathbf{X})$ is identifiable. Similarly it can be checked that if, $f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x}) = f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}^\dagger}(y,\widetilde{y},\mathbf{x})$, a.e. $(y,\widetilde{y},\mathbf{x})$ (with respect to the dominating measure ν), for some $\boldsymbol{\theta},\boldsymbol{\theta}^\dagger \in \boldsymbol{\Theta}$, then it trivially implies, $\boldsymbol{\theta} = \boldsymbol{\theta}^\dagger$. Thus, the family of joint distributions $\{f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ of $(Y,\widetilde{Y},\mathbf{X})$ will be identifiable. The identifiability of these families of distributions ensures that the corresponding expected log-likelihoods (under $\mathbf{P}_0$) are uniquely maximized at the true parameter values (cf. Lemma 5.35 of van der Vaart, 1998). Hence, for all $\boldsymbol{\beta} \in \mathbb{K}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\mathbf{P}_0 \log f_{\boldsymbol{\beta}}(Y,\mathbf{X}) < \mathbf{P}_0 \log f_{\boldsymbol{\beta}_0}(Y,\mathbf{X}), \quad \text{and} \quad \mathbf{P}_0 \log f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}}(Y,\widetilde{Y},\mathbf{X}) < \mathbf{P}_0 \log f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}_0}(Y,\widetilde{Y},\mathbf{X}). \tag{A.4}$$

Now, assumption (A2) ensures that for any $\delta > 0$, we have,

$$\sup\left\{\mathbf{P}_0 \log f_{\boldsymbol{\beta}}(Y,\mathbf{X}) : \boldsymbol{\beta} \in \mathbb{K},\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \geq \delta\right\} < \mathbf{P}_0 \log f_{\boldsymbol{\beta}_0}(Y,\mathbf{X}), \quad \text{for any } \delta > 0. \tag{A.5}$$

Now, define the maps,

$$h_i(\theta) = \theta_{i,0}\cdot\log\theta + (1-\theta_{i,0})\cdot\log(1-\theta), \quad \text{for any } \theta \in \mathbb{R}, \text{ for } i = 1,2, \tag{A.6}$$

where, $\boldsymbol{\theta}_0 = (\theta_{1,0},\theta_{2,0})$, are the true misclassification parameters. Then, we can write the second term in (A.3) as,

$$\begin{aligned}
\mathbf{P}_0 \log f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}}(Y,\widetilde{Y},\mathbf{X}) &= \mathbf{E}_0(Y)\cdot\log\left\{\theta_2^{\theta_{2,0}}\cdot(1-\theta_2)^{(1-\theta_{2,0})}\right\} + \left(1-\mathbf{E}_0(Y)\right)\cdot\log\left\{\theta_1^{\theta_{1,0}}\cdot(1-\theta_1)^{(1-\theta_{1,0})}\right\} \\
&= \mathbf{E}_0(Y)\cdot h_2(\theta_2) + (1-\mathbf{E}_0(Y))\cdot h_1(\theta_1), \quad \text{for any } \boldsymbol{\theta} = (\theta_1,\theta_2) \in \boldsymbol{\Theta}.
\end{aligned}$$

Since the map, $h_i(\theta)$ has maxima at $\theta = \theta_{i,0}$, for $i = 1,2$, it implies, for any small enough $\delta > 0$,

$$\begin{aligned}
&\sup\left\{\left(1-\mathbf{E}_0(Y)\right)\cdot h_1(\theta_1) : \theta_1 \in \boldsymbol{\Theta},\ |\theta_1 - \theta_{1,0}| \geq \delta\right\} \\
&= \left(1-\mathbf{E}_0(Y)\right)\cdot\max\left\{h_1\left(\max\{\delta_1, \theta_{1,0}-\delta\}\right), h_1\left(\min\{\delta_2,\theta_{1,0}+\delta\}\right)\right\} < (1-\mathbf{E}_0(Y))\cdot h_1(\theta_{1,0}),
\end{aligned} \tag{A.7}$$

and a similar conclusion holds if, h_1, θ_1 and $\theta_{1,0}$, are replaced by, h_2, θ_2 and $\theta_{2,0}$. Recall that δ_1,δ_2 are the lower and upper end points in the parameter space for θ_i's (see assumption (A3)). Now, fix a small enough $\delta > 0$, and define the regions, $A_\delta = \left\{(\theta_1,\theta_2) \in \boldsymbol{\Theta} :\ \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \delta\right\}$, and $A_{i,\delta} = \left\{(\theta_1,\theta_2) \in \boldsymbol{\Theta} : |\theta_i - \theta_{i,0}| \geq \delta/\sqrt{2}\right\}$, for $i = 1,2$. Then, $A_\delta \subseteq A_{\delta,1} \cup A_{\delta,2}$. As a result,

$$\begin{aligned}
&\sup\left\{\mathbf{P}_0 \log f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}} : \boldsymbol{\theta} \in A_\delta\right\} = \sup\left\{\mathbf{E}_0(Y)\cdot h_2(\theta_2) + (1-\mathbf{E}_0(Y))\cdot h_1(\theta_1) : \boldsymbol{\theta} \in A_\delta\right\} \\
&\leq \mathbf{E}_0(Y)\cdot\sup\left\{h_2(\theta_2) : \boldsymbol{\theta} \in A_\delta\right\} + \left(1-\mathbf{E}_0(Y)\right)\cdot\sup\left\{h_1(\theta_1) : \boldsymbol{\theta} \in A_\delta\right\} \\
&\leq \mathbf{E}_0(Y)\cdot\sup\left\{h_2(\theta_2) : \boldsymbol{\theta} \in A_{\delta,1}\cup A_{\delta,2}\right\} + \left(1-\mathbf{E}_0(Y)\right)\cdot\sup\left\{h_1(\theta_1) : \boldsymbol{\theta} \in A_{\delta,1}\cup A_{\delta,2}\right\} \\
&= \mathbf{E}_0(Y)\cdot\sup\left\{h_2(\theta_2) : \left[|\theta_1-\theta_{1,0}| \geq \delta/\sqrt{2}\right] \cup \left[|\theta_2-\theta_{2,0}| \geq \delta/\sqrt{2}\right]\right\} \\
&\quad + \left(1-\mathbf{E}_0(Y)\right)\cdot\sup\left\{h_1(\theta_1) : \left[|\theta_1-\theta_{1,0}| \geq \delta/\sqrt{2}\right] \cup \left[|\theta_2-\theta_{2,0}| \geq \delta/\sqrt{2}\right]\right\} \\
&= \mathbf{E}_0(Y)\cdot\sup\left\{h_2(\theta_2) : \left[|\theta_2-\theta_{2,0}| \geq \delta/\sqrt{2}\right]\right\} + \left(1-\mathbf{E}_0(Y)\right)\cdot\sup\left\{h_1(\theta_1) : \left[|\theta_1-\theta_{1,0}| \geq \delta/\sqrt{2}\right]\right\} \\
&< \mathbf{E}_0(Y)\cdot h_2(\theta_{2,0}) + (1-\mathbf{E}_0(Y))\cdot h_1(\theta_{1,0}) = \mathbf{P}_0 \log f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}_0},
\end{aligned} \tag{A.8}$$

where, the last strict inequality follows from the arguments presented earlier in (A.7). Similarly, for a small enough $\delta > 0$, write the sets, $B_\delta = \left\{(\boldsymbol{\beta},\boldsymbol{\theta}) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \geq \delta^2\right\}$, $B_{\delta,1} = \left\{\boldsymbol{\beta} \in \mathbb{K} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \geq \delta/\sqrt{2}\right\}$, and $B_{\delta,2} = \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} :\ \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \geq \delta/\sqrt{2}\right\}$. Then,

$$\sup\left\{\mathbf{P}_0 g_{1,\boldsymbol{\beta},\boldsymbol{\theta}} : (\boldsymbol{\beta},\boldsymbol{\theta}) \in B_\delta\right\} = \sup\left\{\mathbf{P}_0 \log f_{\boldsymbol{\beta}} + \mathbf{P}_0 \log f_{\boldsymbol{\beta}_0,\boldsymbol{\theta}} : (\boldsymbol{\beta},\boldsymbol{\theta}) \in B_\delta\right\}$$

$$\leq \sup\left\{\mathbf{P}_0 \log f_\beta : (\beta,\theta) \in B_\delta\right\} + \sup\left\{\mathbf{P}_0 \log f_{\beta_0,\theta} : (\beta,\theta) \in B_\delta\right\}$$
$$\leq \sup\left\{\mathbf{P}_0 \log f_\beta : (\beta,\theta) \in B_{\delta,1} \cup B_{\delta,2}\right\} + \sup\left\{\mathbf{P}_0 \log f_{\beta_0,\theta} : (\beta,\theta) \in B_{\delta,1} \cup B_{\delta,2}\right\}$$
$$= \sup\left\{\mathbf{P}_0 \log f_\beta : \beta \in B_{\delta,1}\right\} + \sup\left\{\mathbf{P}_0 \log f_{\beta_0,\theta} : \theta \in B_{\delta,2}\right\}$$
$$< \mathbf{P}_0 \log f_{\beta_0} + \mathbf{P}_0 \log f_{\beta_0,\theta_0} = \mathbf{P}_0 g_{1,\beta_0,\theta_0}, \tag{A.9}$$

following (A.5) and (A.8). Now, we can write, $\mathbf{P}_0 g_{2,\beta,\theta} = \mathbf{P}_0 \log \widetilde{f}_{\beta,\theta}(\widetilde{Y},\mathbf{X})$, where, $\widetilde{f}_{\beta,\theta}(\widetilde{y},\mathbf{x})$ denotes the joint density of $(\widetilde{Y},\mathbf{X})$ with respect to a dominating measure ν, when the underlying parameter is (β,θ). Also, $\{\widetilde{f}_{\beta,\theta}(\widetilde{y},\mathbf{x}) : \beta \in \mathbb{K}, \theta \in \Theta\}$ constitutes a family of distributions indexed by the parameters (β,θ). Thus, Lemma 5.35 of van der Vaart (1998) also implies

$$\mathbf{P}_0 g_{2,\beta,\theta} = \mathbf{P}_0 \log\log \widetilde{f}_{\beta,\theta}(\widetilde{Y},\mathbf{X}) \leq \mathbf{P}_0 \log \widetilde{f}_{\beta_0,\theta_0}(\widetilde{Y},\mathbf{X}) = \mathbf{P}_0 g_{2,\beta_0,\theta_0}, \tag{A.10}$$

for any $(\beta,\theta) \in \mathbb{K}\times\Theta$. It should be noted that we have not used strict inequality in (A.10). Combining (A.9) and (A.10), we can easily obtain the desired identifiability result. This completes the verification of all required conditions for ensuring consistency as stated in Theorem 5.7 of van der Vaart (1998) and the proof is completed. □

Lemma 5.2. *Suppose, assumptions (A3) and (A6)(i) hold. Then, there exists a constant $c_1 \in (0,1)$, such that,*

$$c_1 < h_{\beta,\theta}(\mathbf{x}) < 1 - c_1, \quad \textit{for all } \beta \in \mathbb{K},\ \theta \in \Theta \textit{ and all } \mathbf{x}. \tag{A.11}$$

Proof of Lemma 5.2. From the definition of $h_{\beta,\theta}(\mathbf{x})$ (cf. (3.4)), and the above mentioned assumptions, it follows that, either $h_{\beta,\theta}(\mathbf{x}) \in (\theta_1, 1-\theta_2) \subset (\delta_1, 1-\delta_1)$ or $h_{\beta,\theta}(\mathbf{x}) \in (1-\theta_2, \theta_1) \subset (1-\delta_2, \delta_2)$, since $\theta_1 + \theta_2 \neq 1$. We can choose c_1 appropriately to ensure that the claim follows. □

Proof of Lemma 3.1. In order to prove this result, we follow Theorem 3.1 of Newey and McFadden (1994), which provides sufficient conditions for asymptotic normality of extremum estimators. We verify these sufficient conditions in order to complete the proof. The primary requirement is consistency of the CMLE's $(\widehat{\beta}_n, \widehat{\theta}_n)$, which has been shown in Lemma 5.1. Secondly, assumptions (A1) and (A3) ensure that $\beta_0 \in \text{int}(\mathbb{K})$ and $\theta_0 \in \text{int}(\Theta)$. The next step is to ensure that, $(\beta,\theta) \mapsto M_n(\beta,\theta)$ is twice continuously differentiable in a small enough neighbourhood of (β_0,θ_0).

In order to show this, note that, assumption (A1) and (A4) ensure that $|\mathbf{x}^T\beta|$ is uniformly bounded on the support of $\mathbf{X}$ and for all $\beta \in \mathbb{K}$. Thus, assumption (A6)(i) ensures that there exists a $c_0 \in (0,1)$, such that $c_0 < \psi(\mathbf{x}^T\beta) < 1 - c_0$, for all $\mathbf{x} \in \mathcal{X}$ and all $\beta \in \mathbb{K}$. Similarly, Lemma 5.2 ensures we can find a $c_1 \in (0,1)$, such that, $c_1 < h_{\beta,\theta}(\mathbf{x}) < 1 - c_1$, for all $\beta \in \mathbb{K}$, $\theta \in \Theta$ and all $\mathbf{x} \in \mathcal{X}$. Now, we consider the expressions for first order partial derivatives of $g_{i,\beta,\theta}$, $i = 1,2$, provided in (B.2) and (B.4) respectively. The existence of the these derivatives will be ensured if the terms $\psi(\mathbf{x}^T\beta)$ and $h_{\beta,\theta}(\mathbf{x})$ remain bounded away from 0 and 1, for each (β,θ) in a neighbourhood of (β_0,θ_0) and for all $(y,\widetilde{y},\mathbf{x}) \in \mathcal{X}$. The above argument on the lower and upper bounds for $\psi(\mathbf{x}^T\beta)$ and $h_{\beta,\theta}(\mathbf{x})$ imply that these derivatives will exists in a neighbourhood of (β_0,θ_0). Similarly, the expressions provided in Appendix B also show that all second order partial derivatives of $g_{i,\beta,\theta}$, $i = 1,2$, will exist in a neighbourhood of (β_0,θ_0), under the same conditions. Continuity of the second order partial derivatives is ensured by assumption (A6)(ii). Next, we can write,

$$H_n(\beta,\theta) \equiv \begin{pmatrix} \frac{\partial}{\partial\beta} M_n(\beta,\theta) \\ \frac{\partial}{\partial\theta_1} M_n(\beta,\theta) \\ \frac{\partial}{\partial\theta_2} M_n(\beta,\theta) \end{pmatrix} = f_n \cdot \mathbb{P}_{n_1} h_{1,\beta,\theta}(Y,\widetilde{Y},\mathbf{X}) + (1-f_n)\cdot \mathbb{P}_{n_2} h_{2,\beta,\theta}(\widetilde{Y},\mathbf{X}), \quad \text{where,} \tag{A.12}$$

$h_{i,\beta,\theta}$, $i = 1,2$, are defined in (B.2) and (B.4). Using (B.2), (1.1) and (1.3) it can be checked that $\mathbf{P}_0 h_{i,\beta_0,\theta_0} = \mathbf{0}$, for each $i = 1,2$. Hence, $\mathbb{P}_{n_i} h_{i,\beta_0,\theta_0}$, $i = 1,2$, are sample means of mean zero i.i.d. random vectors. Write,

$$\boldsymbol{\Sigma}_{1,0} = \mathbf{Var}_0\left(h_{1,\beta_0,\theta_0}(Y,\widetilde{Y},\mathbf{X})\right) \quad \text{and} \quad \boldsymbol{\Sigma}_{2,0} = \mathbf{Var}_0\left(h_{2,\beta_0,\theta_0}(\widetilde{Y},\mathbf{X})\right).$$

The stated assumptions ensure that $\boldsymbol{\Sigma}_{i,0}$, $i = 1,2$, will exist. Then, using multivariate CLT for i.i.d. random vectors it follows that,

$$\sqrt{n}H_n(\beta_0,\theta_0) = \sqrt{n}\left[f_n\cdot(\mathbb{P}_{n_1} - \mathbf{P}_0)h_{1,\beta_0,\theta_0} + (1-f_n)\cdot(\mathbb{P}_{n_2} - \mathbf{P}_0)h_{2,\beta_0,\theta_0}\right]$$
$$= \sqrt{f_n}\cdot\mathbb{G}_{n_1} h_{1,\beta_0,\theta_0} + \sqrt{1-f_n}\cdot\mathbb{G}_{n_2} h_{2,\beta_0,\theta_0} \xrightarrow{d} N_{p+2}\left(\mathbf{0},\ \boldsymbol{\Sigma}_0\right),$$

where, $\boldsymbol{\Sigma}_0 = f\cdot\boldsymbol{\Sigma}_{1,0} + (1-f)\cdot\boldsymbol{\Sigma}_{2,0}$, and the detailed expressions for $\boldsymbol{\Sigma}_{1,0}$ and $\boldsymbol{\Sigma}_{2,0}$ are provided in (B.8). This follows due to assumption (A5) and also by using the independence of $\mathbb{G}_{n_1} h_{1,\beta_0,\theta_0}$ and $\mathbb{G}_{n_2} h_{2,\beta_0,\theta_0}$.

For the next step, we need to show that the Hessian of $M_n(\beta,\theta)$ converges uniformly in a neighbourhood of (β_0,θ_0), to a continuous limit matrix, which we denote by $J(\beta,\theta)$. Since there are finitely many elements in the Hessian, it is enough to show uniform convergence for each element of the Hessian matrix. For simplicity, we illustrate the argument for the following term of the Hessian matrix of M_n (corresponding to second order partial derivative w.r.t. β_j and β_k). We can write,

$$\frac{\partial^2}{\partial\beta_j\beta_k} M_n(\beta,\theta) = f_n\cdot\mathbb{P}_{n_1}\left[\frac{\partial^2}{\partial\beta_j\beta_k} g_{1,\beta,\theta}(Y,\widetilde{Y},\mathbf{X})\right] + (1-f_n)\cdot\mathbb{P}_{n_2}\left[\frac{\partial^2}{\partial\beta_j\beta_k} g_{2,\beta,\theta}(\widetilde{Y},\mathbf{X})\right]$$

$$= f_n \cdot \mathbb{P}_{n_1}\left[X_j X_k \cdot \left[\frac{\psi^{(2)}(\mathbf{X}^T\boldsymbol{\beta})\cdot\{Y-\psi(\mathbf{X}^T\boldsymbol{\beta})\}}{\psi(\mathbf{X}^T\boldsymbol{\beta})\cdot\{1-\psi(\mathbf{X}^T\boldsymbol{\beta})\}} - \left(\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta})\right)^2 \cdot \left\{\frac{Y}{\left(\psi(\mathbf{X}^T\boldsymbol{\beta})\right)^2} + \frac{(1-Y)}{\left(1-\psi(\mathbf{X}^T\boldsymbol{\beta})\right)^2}\right\}\right]\right]$$
$$+ (1-f_n)\cdot\mathbb{P}_{n_2}\left[X_j(1-\theta_1-\theta_2)\left\{X_k\psi^{(2)}(\mathbf{X}^T\boldsymbol{\beta})\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{Y},\mathbf{X}) + \psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta})\cdot\frac{\partial}{\partial\beta_k}\,\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{Y},\mathbf{X})\right\}\right], \tag{A.13}$$

where, $\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})$ is defined in (B.1). As per assumptions (A1) and (A3), $\boldsymbol{\beta}_0 \in \text{int}(\mathbb{K})$ and $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$. Hence, there exists a small enough $\delta > 0$, such that, $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| < 2\delta\} \subset \mathbb{K}$ and $\{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| < 2\delta\} \subset \Theta$. Consider the δ-radius closed neighbourhood of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ and write $N_\delta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le \delta\}$ and $N_\delta(\boldsymbol{\theta}) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le \delta\}$. Note, $N_\delta(\boldsymbol{\beta}_0)$ and $N_\delta(\boldsymbol{\theta}_0)$ are compact sets in $\mathbb{R}^p$ and $\mathbb{R}^2$ respectively. Further, if we consider the above second order partial derivative expressions and use assumption (A6), then we can claim that,

$$\xi_{i,\boldsymbol{\beta},\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x}) \equiv \frac{\partial^2}{\partial\beta_j\beta_k}\, g_{i,\boldsymbol{\beta},\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x}), \quad i=1,2,$$

are continuous at each $(\boldsymbol{\beta},\boldsymbol{\theta}) \in N_\delta(\boldsymbol{\beta}_0)\times N_\delta(\boldsymbol{\theta}_0)$, for every fixed choice of $(y,\widetilde{y},\mathbf{x})$. Further, note that as $h_{\boldsymbol{\beta},\boldsymbol{\theta}}(\mathbf{x})$ and $\psi(\mathbf{x}^T\boldsymbol{\beta})$ are uniformly bounded away from 0 and 1, we can claim that there exists a constant $K_1 \in (0,\infty)$ (not depending on $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and δ), such that

$$\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in N_\delta(\boldsymbol{\beta}_0)\times N_\delta(\boldsymbol{\theta}_0)}\left|\frac{\partial}{\partial\beta_k}\,\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})\right| \le \sup_{\boldsymbol{\theta}\in\Theta}|1-\theta_1-\theta_2|\cdot\sup_{\boldsymbol{\beta}\in\mathbb{K}}\left|x_k\cdot\psi^{(1)}(\mathbf{x}^T\boldsymbol{\beta})\right|\cdot K_1\left[\widetilde{y}+(1-\widetilde{y})\right]$$
$$\le 2K_1|x_k|\cdot\sup_{\boldsymbol{\beta}\in\mathbb{K},\mathbf{x}\in\mathcal{X}}|\psi^{(1)}(\mathbf{x}^T\boldsymbol{\beta})| \le K_2|x_k|,$$

where, $K_2 = 2K_1\cdot\sup_{\boldsymbol{\beta}\in\mathbb{K},\mathbf{x}\in\mathcal{X}}|\psi^{(1)}(\mathbf{x}^T\boldsymbol{\beta})|$, and the supremum in this expression is bounded, because it is taken over a bounded set (cf. assumptions (A1) and (A4)) and also because $\psi^{(1)}(\cdot)$ is a probability density function. Thus, $K_2 \in (0,\infty)$. Due to same reasons we can claim that, $\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})$ is also uniformly bounded (over $(\boldsymbol{\beta},\boldsymbol{\theta})$ for each fixed $(\widetilde{y},\mathbf{x})$) by a constant term. Thus, using the line arguments followed earlier and the derivative expression provided in Appendix B, we can claim that the class $\{\xi_{i,\boldsymbol{\beta},\boldsymbol{\theta}} : (\boldsymbol{\beta},\boldsymbol{\theta}) \in \mathbb{K}\times\Theta\}$, will have an envelope function of the form, $C_i\cdot|x_jx_k|$, for some $C_i \in (0,\infty)$, for each $i=1,2$.

Now, Cauchy–Schwarz inequality implies, $\mathbf{P}_0|X_iX_j| = \mathbf{E}|X_iX_j| \le \sqrt{\mathbf{E}(X_i^2)\cdot\mathbf{E}(X_j^2)} < \infty$, which follows from the finite variance assumption (A4). Hence, these classes have L_1-integrable (w.r.t. $\mathbf{P}_0$) envelope functions, and the functions $\xi_{i,\boldsymbol{\beta},\boldsymbol{\theta}}$ are measurable as they are continuous in $(y,\widetilde{y},\mathbf{x})$. The same argument works if we focus on the class $\{\xi_{i,\boldsymbol{\beta},\boldsymbol{\theta}} : (\boldsymbol{\beta},\boldsymbol{\theta}) \in N_\delta(\boldsymbol{\beta}_0)\times N_\delta(\boldsymbol{\theta}_0)\}$. Now, using standard bracketing entropy arguments (see Example 19.8 of van der Vaart, 1998), we can claim that

$$\sup\left\{\left(\mathbb{P}_{n_1}-\mathbf{P}_0\right)\left[\frac{\partial^2}{\partial\beta_j\partial\beta_k}\, g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}\right] : (\boldsymbol{\beta},\boldsymbol{\theta}) \in N_\delta(\boldsymbol{\beta}_0)\times N_\delta(\boldsymbol{\theta}_0)\right\} = o_{\mathbf{P}_0}(1),$$

and a similar claim holds if we use $\mathbb{P}_{n_2}$ and $g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}$ (in place of $\mathbb{P}_{n_1}$ and $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}$ respectively). Now,

$$\left|f_n\cdot\mathbb{P}_{n_1}\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}\right] + (1-f_n)\cdot\mathbb{P}_{n_2}\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}\right] - f\cdot\mathbf{P}_0\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}\right] - (1-f)\cdot\mathbf{P}_0\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}\right]\right|$$
$$\le f_n\cdot\left|(\mathbb{P}_{n_1}-\mathbf{P}_0)\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}\right]\right| + (1-f_n)\cdot\left|(\mathbb{P}_{n_2}-\mathbf{P}_0)\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}\right]\right| + |f_n-f|\cdot\left|\mathbf{P}_0\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}\right]\right|$$
$$+ |f_n-f|\cdot\left|\mathbf{P}_0\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}\right]\right|.$$

The first two terms converge in the expression on the r.h.s. above converge in probability ($\mathbf{P}_0$) to zero, while

$$\left|\mathbf{P}_0\left[\frac{\partial^2}{\partial\beta_j\beta_k}g_{i,\boldsymbol{\beta},\boldsymbol{\theta}}\right]\right| \le C_i\cdot\mathbf{P}_0|X_jX_k| \in (0,\infty), \quad \text{for } i=1,2,$$

if we use the envelope function $C_i|x_jx_k|$, found earlier. Thus, assumption (A5) will ensure that the last two terms converge to zero as $n\to\infty$. Now, if we consider the expressions of all other second order partial derivatives of $g_{i,\boldsymbol{\beta},\boldsymbol{\theta}}$ provided in Appendix B, then using similar arguments we will be able to show that these derivatives are continuous in $(\boldsymbol{\beta},\boldsymbol{\theta})$ in a compact neighbourhood around $(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)$ and further the corresponding classes will have $\mathbf{P}_0$-integrable envelope function, either of the form $C|x_jx_k|$ or a constant envelope function. A similar argument as above will ensure that the corresponding terms of the Hessian of $M_n(\boldsymbol{\beta},\boldsymbol{\theta})$ will converge uniformly to a limiting value in this neighbourhood. Since the arguments are similar, we skip the details. Finally, we can claim that there exists a $\delta > 0$, such that

$$\sup\left[\left\|\begin{pmatrix}\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}M_n(\boldsymbol{\beta},\boldsymbol{\theta}) & \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}^T}M_n(\boldsymbol{\beta},\boldsymbol{\theta})\\ \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\beta}^T}M_n(\boldsymbol{\beta},\boldsymbol{\theta}) & \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}M_n(\boldsymbol{\beta},\boldsymbol{\theta})\end{pmatrix} - J(\boldsymbol{\beta},\boldsymbol{\theta})\right\| : (\boldsymbol{\beta},\boldsymbol{\theta}) \in N_\delta(\boldsymbol{\beta}_0)\times N_\delta(\boldsymbol{\theta}_0)\right] = o_{\mathbf{P}_0}(1),$$

where $J(\boldsymbol{\beta},\boldsymbol{\theta})$ is the limit of the Hessian of $M_n(\boldsymbol{\beta},\boldsymbol{\theta})$, and $\|\cdot\|$ denotes the Frobenius norm for matrices. We write $J_0 = J(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)$ and the expression for each element of J_0 (cf. (B.7)) is provided in Appendix B. It can be shown that $J_0 = -\boldsymbol{\Sigma}_0^{-1}$, which will ensure the existence of J_0. This completes the verification of all required conditions in Theorem 3.1 of Newey and McFadden (1994) and thereby we can claim that the stated distribution convergence result holds. □

Proof of Theorem 3.2. Define the quantities,

$$\left.\begin{aligned} &M_{1,R}(\boldsymbol{\eta}) = \mathbf{P}_0 g_{1,R,\boldsymbol{\eta}}(Y,\widetilde{Y},\mathbf{X}), \quad M_{2,R}(\boldsymbol{\eta}) = \mathbf{P}_0 g_{2,R,\boldsymbol{\eta}}(Y,\widetilde{Y},\mathbf{X}), \quad \text{and} \\ &M_R(\boldsymbol{\eta}) = f \cdot M_{1,R}(\boldsymbol{\eta}) + (1-f)\cdot M_{2,R}(\boldsymbol{\eta}), \quad \text{for all } \boldsymbol{\eta} \in \Delta. \end{aligned}\right\} \tag{A.14}$$

The first step is to prove consistency of the CMLE $\widehat{\boldsymbol{\eta}}_n$ defined in (3.2). Firstly note that, using the arguments provided in proof of Lemma 5.1, we can show, $\left\{ g_{i,\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}} : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta}^{(r)} \in \Theta \right\}$, $r = 1,\ldots,R$, and $i = 1,2$, can be shown to be $\mathbf{P}_0$ Donsker classes of functions, which also satisfy, $\sup\left\{ |\mathbf{P}_0 g_{i,\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}| : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta}^{(r)} \in \Theta \right\} < \infty$, for $i = 1,2$, (following (A.2)). Also, the map, $\mathbf{x} \mapsto \mathbf{1}(\mathbf{x} \in G_r)$, is an uniformly bounded map. As a result, using Corollary 9.32(v) of Kosorok (2008), it now follows that the class of functions,

$$\left\{ \mathbf{1}(\mathbf{x} \in G_r) \cdot g_{1,\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(y,\widetilde{y},\mathbf{x}) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta}^{(r)} \in \Theta \right\}, \quad r = 1,\ldots,R,$$

is a $\mathbf{P}_0$-Donsker class, and the same claim holds if $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(y,\widetilde{y},\mathbf{x})$ is replaced by $g_{2,\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(\widetilde{y},\mathbf{x})$. Since sums of Donsker classes are Donsker (see Corollary 9.32(i) of Kosorok, 2008), it now follows

$$\sup\left\{ (\mathbb{P}_{n_1} - \mathbf{P}_0) \sum_{r=1}^{R} \mathbf{1}(\mathbf{X} \in G_r) \cdot g_{1,\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(Y,\widetilde{Y},\mathbf{X}) : \boldsymbol{\beta} \in \mathbb{K}, \boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(R)} \in \Theta \right\} = o_{\mathbf{P}_0}(1),$$

and the same argument holds if $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(y,\widetilde{y},\mathbf{x})$ is replaced by $g_{2,\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(\widetilde{y},\mathbf{x})$. Thus, $\{g_{1,R,\boldsymbol{\eta}} : \boldsymbol{\eta} \in \Delta\}$ and $\{g_{2,R,\boldsymbol{\eta}} : \boldsymbol{\eta} \in \Delta\}$ are both $\mathbf{P}_0$-Donsker classes. The proof for well-separatedness of $M_R(\boldsymbol{\eta})$ (cf. (A.14)) proceeds along the same lines as done in Lemma 5.1, for the function $M(\boldsymbol{\beta},\boldsymbol{\theta})$. We skip the details. Further, $M_{n,R}(\widehat{\boldsymbol{\eta}}_n) \geq M_{n,R}(\boldsymbol{\eta}_0) - o_{\mathbf{P}_0}(1)$, from the definition of the CMLE in (3.2). This proves consistency of $\widehat{\boldsymbol{\eta}}_n$. Next, note that following (B.2) and (B.4), we can write

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}}\, g_{1,R,\boldsymbol{\eta}}(y,\widetilde{y},\mathbf{x}) &= \sum_{r=1}^{R} \mathbf{1}(\mathbf{x} \in G_r)\cdot \mathbf{x} \cdot \psi^{(1)}(\mathbf{x}^T\boldsymbol{\beta}) \cdot \left[\frac{y}{\psi(\mathbf{x}^T\boldsymbol{\beta})} - \frac{(1-y)}{\{1-\psi(\mathbf{x}^T\boldsymbol{\beta})\}} \right], \\
\frac{\partial}{\partial \theta_1^{(j)}}\, g_{1,R,\boldsymbol{\eta}}(y,\widetilde{y},\mathbf{x}) &= \mathbf{1}(\mathbf{x} \in G_j)\cdot (1-y) \cdot \left[\frac{\widetilde{y}}{\theta_1^{(j)}} - \frac{(1-\widetilde{y})}{\left(1-\theta_1^{(j)}\right)} \right], \\
\frac{\partial}{\partial \theta_2^{(j)}}\, g_{1,R,\boldsymbol{\eta}}(y,\widetilde{y},\mathbf{x}) &= \mathbf{1}(\mathbf{x} \in G_j)\cdot y \cdot \left[\frac{(1-\widetilde{y})}{\theta_2^{(j)}} - \frac{\widetilde{y}}{\left(1-\theta_2^{(j)}\right)} \right], \\
\frac{\partial}{\partial \boldsymbol{\beta}}\, g_{2,R,\boldsymbol{\eta}}(\widetilde{y},\mathbf{x}) &= \sum_{r=1}^{R} \mathbf{1}(\mathbf{x} \in G_r)\cdot \mathbf{x} \cdot \left(1-\theta_1^{(r)}-\theta_2^{(r)}\right) \cdot \psi^{(1)}(\mathbf{x}^T\boldsymbol{\beta}) \cdot \chi_{\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(\widetilde{y},\mathbf{x}), \\
\frac{\partial}{\partial \theta_1^{(j)}}\, g_{2,R,\boldsymbol{\eta}}(\widetilde{y},\mathbf{x}) &= \mathbf{1}(\mathbf{x} \in G_j)\cdot \{1-\psi(\mathbf{x}^T\boldsymbol{\beta})\} \cdot \chi_{\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(\widetilde{y},\mathbf{x}), \\
\frac{\partial}{\partial \theta_2^{(j)}}\, g_{2,R,\boldsymbol{\eta}}(\widetilde{y},\mathbf{x}) &= -\mathbf{1}(\mathbf{x} \in G_j)\cdot \psi(\mathbf{x}^T\boldsymbol{\beta}) \cdot \chi_{\boldsymbol{\beta},\boldsymbol{\theta}^{(r)}}(\widetilde{y},\mathbf{x}).
\end{aligned}$$

Note, following (1.1) and (2.3) we can write,

$$\begin{aligned}
&\mathbf{P}_0 \left[\sum_{r=1}^{R} \mathbf{1}(\mathbf{X} \in G_r)\cdot \mathbf{X} \cdot \psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0) \cdot \left\{ \frac{Y}{\psi(\mathbf{X}^T\boldsymbol{\beta}_0)} - \frac{(1-Y)}{\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\}} \right\} \right] = \mathbf{0}_{p\times 1}, \\
&\mathbf{P}_0 \left[\sum_{r=1}^{R} \mathbf{1}(\mathbf{X} \in G_r)\cdot \mathbf{X} \cdot \left(1-\theta_{1,0}^{(r)}-\theta_{2,0}^{(r)}\right) \cdot \psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0) \cdot \chi_{\boldsymbol{\beta}_0,\boldsymbol{\theta}_0^{(r)}}(\widetilde{Y},\mathbf{X}) \right] = \mathbf{0}_{p\times 1}, \\
&\mathbf{P}_0 \left[\mathbf{1}(\mathbf{X} \in G_j)\cdot (1-Y) \cdot \left\{ \frac{\widetilde{Y}}{\theta_{1,0}^{(j)}} - \frac{(1-\widetilde{Y})}{\left(1-\theta_{1,0}^{(j)}\right)} \right\} \right] = 0,
\end{aligned}$$

and similarly, the expected value (under $\mathbf{P}_0$) of the other derivatives found above, at the true parameter values will be equal to zero. Hence,

$$\begin{aligned}
&\sqrt{n}\cdot \left[\frac{\partial}{\partial \boldsymbol{\eta}}\, M_{n,R}(\boldsymbol{\eta})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \right] \\
&= \sqrt{n}\cdot \left[f_n \cdot \left(\mathbb{P}_{n_1} - \mathbf{P}_0\right) \frac{\partial}{\partial \boldsymbol{\eta}}\, g_{1,R,\boldsymbol{\eta}}(Y,\widetilde{Y},\mathbf{X})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} + (1-f_n)\cdot \left(\mathbb{P}_{n_2} - \mathbf{P}_0\right) \frac{\partial}{\partial \boldsymbol{\eta}}\, g_{2,R,\boldsymbol{\eta}}(Yt,\mathbf{X})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \right] \\
&= \sqrt{f_n}\cdot \mathbb{G}_{n_1}\left(\frac{\partial}{\partial \boldsymbol{\eta}}\, g_{1,R,\boldsymbol{\eta}}(Y,\widetilde{Y},\mathbf{X})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \right) + \sqrt{1-f_n}\cdot \mathbb{G}_{n_2}\left(\frac{\partial}{\partial \boldsymbol{\eta}}\, g_{2,R,\boldsymbol{\eta}}(\widetilde{Y},\mathbf{X})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \right).
\end{aligned}$$

is a weighted sum of independent mean-zero random vectors, with each vector converging in distribution to a Gaussian limiting random vector. Note that, the stated assumptions will ensure that

$$\Gamma_{1,\boldsymbol{\eta}_0} = \mathbf{Var}_0\left(g_{1,R,\boldsymbol{\eta}}(Y,\widetilde{Y},\mathbf{X})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \right) \quad \text{and} \quad \Gamma_{2,\boldsymbol{\eta}_0} = \mathbf{Var}_0\left(g_{2,R,\boldsymbol{\eta}}(\widetilde{Y},\mathbf{X})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \right), \tag{A.15}$$

exists. The detailed expressions for the elements of Γ_{1,η_0} and Γ_{2,η_0} are provided in (B.9) and (B.10) respectively. Hence, using the CLT we can claim,

$$\sqrt{n}\cdot\left[\frac{\partial}{\partial\boldsymbol{\eta}}\,M_{n,R}(\boldsymbol{\eta})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}\right]\xrightarrow{d}N_{p+2R}\left(\mathbf{0},\ f\cdot\Gamma_{1,\eta_0}+(1-f)\cdot\Gamma_{2,\eta_0}\right). \tag{A.16}$$

The Hessian of $M_{n,R}(\boldsymbol{\eta})$ will converge in probability, uniformly in a neighbourhood of the true parameter $\boldsymbol{\eta}_0$, to a limit matrix $J_R(\boldsymbol{\eta})$. This can be shown using similar steps as in the proof of Theorem 3.2. Write, $J_{0,R}=J_R(\boldsymbol{\eta}_0)$, and it can be shown that $J_{0,R}=-f\cdot\Gamma_{1,\eta_0}-(1-f)\cdot\Gamma_{2,\eta_0}$, following the arguments provided in Appendix B. Thus, Theorem 3.1 of Newey and McFadden (1994) can be now used to complete the proof. □

Proof (*Outline of Proof of Corollary 3.4*)**.** The technical arguments for the proof of Corollary 3.4(i) proceed in the same manner as in the proof of Lemma 3.1 (for the non-differential case). The primary difference is the altered definition of the map $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}$ which does not change the Donsker property for the class of functions $\{g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}:(\boldsymbol{\beta},\boldsymbol{\theta})\in\Delta\}$. The remaining arguments follow similarly and the absence of the θ_2 (or θ_1) component changes the asymptotic covariance matrix of the CMLE's. The proof for the differential case (with $R>1$) will follow along the same lines. We skip the routine details. □

Appendix B. Expression of matrices used in Theorem 3.2 and Lemma 3.1

We begin with expressions for second order partial derivatives of $g_{i,\boldsymbol{\beta},\boldsymbol{\theta}}$, $i=1,2$ (cf. (3.4)). For simplicity of notation, define the following function, for all $(\widetilde{y},\mathbf{x})$,

$$\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})\equiv\frac{\widetilde{y}}{h_{\boldsymbol{\beta},\boldsymbol{\theta}}(\mathbf{x})}-\frac{(1-\widetilde{y})}{\{1-h_{\boldsymbol{\beta},\boldsymbol{\theta}}(\mathbf{x})\}}. \tag{B.1}$$

Then it can be shown that,

$$h_{1,\boldsymbol{\beta},\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x})=\begin{pmatrix}\frac{\partial}{\partial\boldsymbol{\beta}}\,g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x})\\ \frac{\partial}{\partial\theta_1}\,g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x})\\ \frac{\partial}{\partial\theta_2}\,g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x})\end{pmatrix}=\begin{pmatrix}\mathbf{x}\cdot\psi^{(1)}(\mathbf{x}^T\boldsymbol{\beta})\cdot\left[\frac{y}{\psi(\mathbf{x}^T\boldsymbol{\beta})}-\frac{(1-y)}{\{1-\psi(\mathbf{x}^T\boldsymbol{\beta})\}}\right]\\ (1-y)\cdot\left[\frac{\widetilde{y}}{\theta_1}-\frac{(1-\widetilde{y})}{(1-\theta_1)}\right]\\ y\cdot\left[\frac{(1-\widetilde{y})}{\theta_2}-\frac{\widetilde{y}}{(1-\theta_2)}\right]\end{pmatrix} \tag{B.2}$$

Note, the first p and last two components of $h_{1,\boldsymbol{\beta},\boldsymbol{\theta}}$ depend on $\boldsymbol{\beta}$, θ_1 and θ_2 respectively. Hence, Hessian matrix of $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}(y,\widetilde{y},\mathbf{x})$ will have a block-diagonal structure. Write, $J_1(\boldsymbol{\beta},\boldsymbol{\theta})$ to denote the $(p+2)\times(p+2)$ expected Hessian matrix of $g_{1,\boldsymbol{\beta},\boldsymbol{\theta}}$ at any $(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathbb{K}\times\Theta$. Then, it can be shown that,

$$\begin{aligned}&J_1(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)=J_{1,0}\\&=\begin{pmatrix}-\mathbf{E}\left[\mathbf{X}\mathbf{X}^T\cdot\frac{\left(\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0)\right)^2}{\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\cdot\left\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}}\right] & \mathbf{0}_{p\times1} & \mathbf{0}_{p\times1}\\ \mathbf{0}_{1\times p} & -\frac{\mathbf{E}\left[1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right]}{\theta_{1,0}(1-\theta_{1,0})} & 0\\ \mathbf{0}_{1\times p} & 0 & -\frac{\mathbf{E}\left[\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right]}{\theta_{2,0}\{1-\theta_{2,0}\}}\end{pmatrix}.\end{aligned} \tag{B.3}$$

Also, we can write (cf. (B.1)),

$$h_{2,\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})=\begin{pmatrix}\frac{\partial}{\partial\boldsymbol{\beta}}\,g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})\\ \frac{\partial}{\partial\theta_1}\,g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})\\ \frac{\partial}{\partial\theta_2}\,g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})\end{pmatrix}=\begin{pmatrix}\mathbf{x}\cdot(1-\theta_1-\theta_2)\cdot\psi^{(1)}(\mathbf{x}^T\boldsymbol{\beta})\cdot\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})\\ \{1-\psi(\mathbf{x}^T\boldsymbol{\beta})\}\cdot\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})\\ -\psi(\mathbf{x}^T\boldsymbol{\beta})\cdot\chi_{\boldsymbol{\beta},\boldsymbol{\theta}}(\widetilde{y},\mathbf{x})\end{pmatrix}. \tag{B.4}$$

Let $J_2(\boldsymbol{\beta},\boldsymbol{\theta})$ denote the $(p+2)\times(p+2)$ expected Hessian matrix of $g_{2,\boldsymbol{\beta},\boldsymbol{\theta}}$ at any $(\boldsymbol{\beta},\boldsymbol{\theta})$. Let $(J_2(\boldsymbol{\beta},\boldsymbol{\theta}))_{(k,l)}$ denote the (k,l)-th element of $J_2(\boldsymbol{\beta},\boldsymbol{\theta})$. Note, $J_2(\boldsymbol{\beta},\boldsymbol{\theta})$ is a symmetric matrix. Define,

$$c(\boldsymbol{\theta}_0)=(1-\theta_{1,0}-\theta_{2,0})\quad\text{and}\quad h^{\star}_{\boldsymbol{\beta}_0,\boldsymbol{\theta}_0}(\mathbf{x})=h_{\boldsymbol{\beta}_0,\boldsymbol{\theta}_0}(\mathbf{x})\{1-h_{\boldsymbol{\beta}_0,\boldsymbol{\theta}_0}(\mathbf{x})\},\quad\text{for all }\mathbf{x}. \tag{B.5}$$

Then, we can write,

$$\begin{aligned}&J_2(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)=J_{2,0}\\&=\mathbf{E}\left[\frac{1}{h^{\star}_{\boldsymbol{\beta}_0,\boldsymbol{\theta}_0}(\mathbf{X})}\cdot\begin{pmatrix}-\mathbf{X}\mathbf{X}^T\cdot c^2(\boldsymbol{\theta}_0)\left(\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0)\right)^2 & -\mathbf{X}\cdot c(\boldsymbol{\theta}_0)\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0)\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\} & \mathbf{X}\cdot c(\boldsymbol{\theta}_0)\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0)\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\\ & -\left\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2 & \psi(\mathbf{X}^T\boldsymbol{\beta}_0)\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\}\\ & & -\left\{\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2\end{pmatrix}\right].\end{aligned} \tag{B.6}$$

This implies,

$$J(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = J_0 = f \cdot J_{1,0} + (1-f) \cdot J_{2,0}, \quad \text{(cf. (B.3) and (B.6))}, \tag{B.7}$$

is the Hessian of $M(\boldsymbol{\beta}, \boldsymbol{\theta})$ (cf. (A.1)) at $(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$. Now consider the expression for $\boldsymbol{\Sigma}_{1,0}$ and $\boldsymbol{\Sigma}_{2,0}$. It can be checked,

$$\boldsymbol{\Sigma}_{1,0} = \mathbf{Var}\left(h_{1,\boldsymbol{\beta}_0,\boldsymbol{\theta}_0}(Y, \widetilde{Y}, \mathbf{X})\right) = -J_{1,0}, \quad \text{and} \quad \boldsymbol{\Sigma}_{2,0} = \mathbf{Var}\left(h_{2,\boldsymbol{\beta}_0,\boldsymbol{\theta}_0}(Y, \widetilde{Y}, \mathbf{X})\right) = -J_{2,0}, \tag{B.8}$$

where $J_{1,0}$ and $J_{2,0}$ are defined in (B.3) and (B.6) respectively. Now we consider the expressions for $\Gamma_{1,\boldsymbol{\eta}_0}$ and $\Gamma_{2,\boldsymbol{\eta}_0}$, provided in (A.15). Note, $\Gamma_{1,\boldsymbol{\eta}_0}$ is a $(p+2R)\times(p+2R)$ dimensional symmetric block-diagonal matrix, with entries described below,

$$\left(\Gamma_{1,\boldsymbol{\eta}_0}\right)_{i,j} = \begin{cases} \mathbf{E}\left[X_i X_j \cdot \dfrac{\left(\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0)\right)^2}{\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\}}\right] & \text{if } 1 \le i,j \le p, \\ \dfrac{\mathbf{E}\left[\left\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}\cdot \mathbf{1}(\mathbf{X}\in G_r)\right]}{\theta_{1,0}^{(r)}\cdot\{1-\theta_{1,0}^{(r)}\}} & \text{if } i=j=p+2r-1,\ r=1,\dots,R, \\ \dfrac{\mathbf{E}\left[\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\cdot \mathbf{1}(\mathbf{X}\in G_r)\right]}{\theta_{2,0}^{(r)}\cdot\{1-\theta_{2,0}^{(r)}\}} & \text{if } i=j=p+2r,\ r=1,\dots,R, \\ 0 & \text{o.w.} \end{cases} \tag{B.9}$$

Similarly, $\Gamma_{2,\boldsymbol{\eta}_0}$ is a $(p+2R)\times(p+2R)$ dimensional symmetric matrix with entries,

$$\left(\Gamma_{2,\boldsymbol{\eta}_0}\right)_{i,j} = \begin{cases} \mathbf{E}\left[X_i X_j \cdot \left(\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0)\right)^2 \cdot \sum_{r=1}^{R} \mathbf{1}(\mathbf{X}\in G_r)\cdot \dfrac{c^2\left(\theta_0^{(r)}\right)}{h^{\star}_{\boldsymbol{\beta}_0,\theta_0^{(r)}}(\mathbf{X})}\right] & \text{if } 1 \le i,j \le p, \\ \mathbf{E}\left[\mathbf{1}(\mathbf{X}\in G_r)\cdot \dfrac{\left(1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right)^2}{h^{\star}_{\boldsymbol{\beta}_0,\theta_0^{(r)}}(\mathbf{X})}\right] & \text{if } i=j=p+2r-1,\ r=1,\dots,R, \\ \mathbf{E}\left[\mathbf{1}(\mathbf{X}\in G_r)\cdot \dfrac{\left(\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right)^2}{h^{\star}_{\boldsymbol{\beta}_0,\theta_0^{(r)}}(\mathbf{X})}\right] & \text{if } i=j=p+2r,\ r=1,\dots,R, \\ -\mathbf{E}\left[\mathbf{1}(\mathbf{X}\in G_r)\cdot \dfrac{\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\cdot\left\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}}{h^{\star}_{\boldsymbol{\beta}_0,\theta_0^{(r)}}(\mathbf{X})}\right] & \text{if } i=p+2r-1,\ j=p+2r, \text{ for } r=1,\dots,R, \\ c\left(\theta_0^{(r)}\right)\cdot\mathbf{E}\left[X_i\cdot\mathbf{1}(\mathbf{X}\in G_r)\cdot \dfrac{\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0)\cdot\left\{1-\psi(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}}{h^{\star}_{\boldsymbol{\beta}_0,\theta_0^{(r)}}(\mathbf{X})}\right] & \text{if } 1\le i\le p,\ j=p+2r-1,\ r=1,\dots,R, \\ -c\left(\theta_0^{(r)}\right)\cdot\mathbf{E}\left[X_i\cdot\mathbf{1}(\mathbf{X}\in G_r)\cdot \dfrac{\psi^{(1)}(\mathbf{X}^T\boldsymbol{\beta}_0)\cdot\psi(\mathbf{X}^T\boldsymbol{\beta}_0)}{h^{\star}_{\boldsymbol{\beta}_0,\theta_0^{(r)}}(\mathbf{X})}\right] & \text{if } 1\le i\le p,\ j=p+2r,\ r=1,\dots,R, \\ 0 & \text{o.w.} \end{cases} \tag{B.10}$$

where, $c(\boldsymbol{\theta}_0)$ and $h^{\star}_{\boldsymbol{\beta}_0,\theta_0^{(r)}}(\mathbf{x})$ are defined in (B.5).

References

Abrevaya, J., Hausman, J., 1999. Semiparametric estimation with mismeasured dependent variables: An application to duration models for unemployment spells. Ann. d'Écon. Stat. (55/56), 243–275.

Alix-Garcia, J., Millimet, D., 2023. Remotely incorrect? Accounting for nonclassical measurement error in satellite data on deforestation. J. Assoc. Environ. Resour. Econ..

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press.

Black, D., Sanders, S., Taylor, L., 2003. Measurement of higher education in the census and current population survey. J. Amer. Statist. Assoc. 98 (463), 545–554.

Bollinger, C.R., David, M.H., 1997. Modeling discrete choice with response error: Food stamp participation. J. Amer. Statist. Assoc. 92 (439), 827–835.

Bollinger, C.R., David, M.H., 2001. Estimation with response error and nonresponse: Food-stamp participation in the SIPP. J. Bus. Econom. Statist. 19 (2), 129–141.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics. Vol. 5, first ed. Elsevier, pp. 3705–3843.

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement error in nonlinear models, second ed. In: Monographs on Statistics and Applied Probability, vol. 105, Chapman & Hall/CRC, Boca Raton, FL, p. xxviii+455, A modern perspective.

Chatterjee, A., Bandyopadhyay, T., Adhya, S., 2016. Pseudo-likelihood and bootstrapped pseudo-likelihood inference in logistic regression model with misclassified responses. Preprint available at https://www.isid.ac.in/~statmath/2016/isid201612.pdf.

Cheng, G., Huang, J.Z., 2010. Bootstrap consistency for general semiparametric M-estimation. Ann. Statist. 38 (5), 2884–2915.

Cochran, W.G., 1977. Sampling Techniques, third ed. John Wiley.

Copas, J.B., 1988. Binary regression models for contaminated data. J. R. Stat. Soc. Ser. B Stat. Methodol. 50 (2), 225–265, With discussion.

Demidenko, E., 2004. Mixed Models. In: Wiley Series in Probability and Statistics, Wiley-Interscience (John Wiley & Sons), Hoboken, NJ, p. xviii+704, Theory and applications.

Edwards, J.K., Cole, S.R., Troester, M.A., Richardson, D.B., 2013. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. Am. J. Epidemiol. 177, 904–912.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Statist. 7 (1), 1–26.

Fahrmeir, L., Kaufmann, H., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Statist. 13 (1), 342–368.

Gart, J.J., Zweifel, J.R., 1967. On the bias of various estimators of the logit and its variance with application of quantal bioassay. Biometrika 54, 181–187.

Gilbert, P.B., Yu, X., Rotnitzky, A., 2014. Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. Stat. Med. 33 (6), 901–917.

Gine, E., 1997. Lectures on some aspects of the bootstrap. In: Lectures on Probability Theory and Statistics (Saint-FLour, 1996). In: Lecture Notes in Math., vol. 1665, Springer, Berlin, pp. 37–151.

Gouriéroux, C., Monfort, A., 1981. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. J. Econometrics 17 (1), 83–97.

Haldane, J.B.S., 1956. The estimation and significance of the logarithm of a ratio of frequencies. Ann. Hum. Genet. 20 (4), 309–311.

Hausman, J., 2001. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. J. Econ. Perspect. 15 (4), 57–67.

Hausman, J.A., Abrevaya, J., Scott-Morton, F.M., 1998. Misclassification of the dependent variable in a discrete-response setting. J. Econometrics 87 (2), 239–269.

Hug, S., 2010. The effect of misclassifications in probit models: Monte Carlo simulations and applications. Polit. Anal. 18 (1), 78.

Katz, J.N., Katz, G., 2010. Correcting for survey misreports using auxiliary information with an application to estimating turnout. Am. J. Polit. Sci. 54 (3), 815–835.

Kosorok, M.R., 2008. Introduction to Empirical Processes and Semiparametric Inference. In: Springer Series in Statistics, Springer, New York, p. xiv+483.

Kothari, B., Bandyopadhyay, T., 2011. Can India's "literate" read? Int. Rev. Educ. 56 (5), 705–728.

Kreider, B., Pepper, J., 2008. Inferring disability status from corrupt data. J. Appl. Econometrics 23 (3), 329–349.

Lyles, R.H., Tang, L., Superak, H.M., King, C.C., Celentano, D.D., Lo, Y., Sobel, J.D., 2011. Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. Epidemiology 22 (4), 589–597.

Lynch, A.G., Marioni, J.C., Tavaré, S., 2007. Numbers of copy-number variations and false-negative rates will be underestimated if we do not account for the dependence between repeated experiments. Am. J. Hum. Genet. 81 (2), 418–419.

Magder, L.S., Hughes, J.P., 1997. Logistic regression when the outcome is measured with uncertainty. Am. J. Epidemiol. 146 (2), 195–203.

Meyer, B.D., Mittag, N., 2017. Misclassification in binary choice models. J. Econometrics 200 (2), 295–311.

Meyer, B.D., Mittag, N., Goerge, R.M., 2022. Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. J. Hum. Resour. 57 (5), 1605–1644.

Neuhaus, J.M., 1999. Bias and efficiency loss due to misclassified responses in binary regression. Biometrika 86 (4), 843–855.

Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Handbook of Econometrics, vol. 4, (Supplement C), Elsevier, pp. 2111–2245.

Neyman, J., 1938. Contribution to the theory of sampling human populations. J. Amer. Statist. Assoc. 33 (201), 101–116.

Nguimkeu, P., Denteh, A., Tchernis, R., 2019. On the estimation of treatment effects with endogenous misreporting. J. Econometrics 208 (2), 487–506.

Poterba, J.M., Summers, L.H., 1995. Unemployment benefits and labor market transitions: A multinomial logit model with errors in classification. Rev. Econ. Stat. 77 (2), 207–216.

Pratt, J.W., 1981. Concavity of the log likelihood. J. Amer. Statist. Assoc. 76 (373), 103–106.

Rekaya, R., Smith, S., Hay, E.H., Farhat, N., Aggrey, S.E., 2016. Analysis of binary responses with outcome-specific misclassification probability in genome-wide association studies. Appl. Clin. Genet. 9, 169–177.

Roy, S., Banerjee, T., Maiti, T., 2005. Measurement error model for misclassified binary responses. Stat. Med. 24 (2), 269–283.

Savoca, E., 2011. Accounting for misclassification bias in binary outcome measures of illness: The case of post-traumatic stress disorder in male veterans. Sociol. Methodol. 41 (1), 49–76.

Smith, S., Hay, E.H., Farhat, N., Rekaya, R., 2013. Genome wide association studies in presence of misclassified binary responses. BMC Genet. 14 (1), 124.

van de Geer, S.A., 2000. Empirical Processes in M-Estimation. In: Cambridge Series in Statistical and Probabilistic Mathematics, vol. 6, Cambridge University Press, Cambridge, p. xii+286.

van der Vaart, A.W., 1998. Asymptotic Statistics. In: Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3, Cambridge University Press, Cambridge, p. xvi+443.

Wang, W., Scharfstein, D., Tan, Z., MacKenzie, E.J., 2009. Causal inference in outcome-dependent two-phase sampling designs. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (5), 947–969.

Wellner, J.A., Zhan, Y., 1996. Bootstrapping Z-estimators. Technical report, Department of Statistics, University of Washington, Technical Report 308.

Yang, S., Ding, P., 2020. Combining multiple observational data sources to estimate causal effects. J. Amer. Statist. Assoc. 115 (531), 1540–1554.

Yi, G.Y., 2017. Statistical Analysis with Measurement Error Or Misclassification. In: Springer Series in Statistics, Springer, New York, p. xxvii+479, Strategy, method and application, With a foreword by Raymond J. Carroll.

Zawistowski, M., Sussman, J.B., Hofer, T.P., Bentley, D., Hayward, R.A., Wiitala, W.L., 2017. Corrected ROC analysis for misclassified binary outcomes. Stat. Med. 36 (13), 2148–2160.