이 학 박 사 학 위 논 문

H-likelihood Approach for Incomplete Data

계층적 가능도를 이용한 불완전 자료 분석

2022년 8월

서울대학교 대학원

통계학과

한 정 섭

H-likelihood Approach for Incomplete Data
계층적 가능도를 이용한 불완전 자료 분석

지도교수 임 요 한

이 논문을 이학박사 학위논문으로 제출함
2022년 6월

서울대학교 대학원
통계학과
한 정 섭

한정섭의 이학박사 학위논문을 인준함
2022년 6월

위 원 장	정 성 규	(인)
부위원장	임 요 한	(인)
위 원	이 영 조	(인)
위 원	Myunghee Cho Paik	(인)
위 원	하 일 도	(인)

H-likelihood Approach for Incomplete Data

By

Jeongseop Han

A Thesis
Submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
August, 2022

ABSTRACT

H-likelihood Approach for Incomplete Data

Jeongseop Han
The Department of Statistics
The Graduate School
Seoul National University

The h-likelihood has been proposed as an extension of Fisher's likelihood to allow the maximum likelihood estimation for statistical models including unobserved latent variables of recent interest. However, the current h-likelihood approach does not allow maximum likelihood estimators (MLEs) of variance components as Henderson's joint likelihood (1959) does not in linear mixed models. In this thesis, we discuss how to form the canonical scale for the h-likelihood in order to facilitate joint maximization for MLEs of whole parameters.

To show the usefulness of the h-likelihood for analyzing incomplete data, various types of unobserved latent variables are examined; missing data, random effect and censored data. As we shall see, a statistical model for unobserved latent variables may not be

identifiable based on the observed data. Thus, we also present how to make robust inferences against various assumptions on statistical models.

Keywords: Canonical scale, Censored data, Imputation, Laplace approximation, Maximum likelihood estimation, Missing data, Random effect, Robust inference.

Student Number: 2014 − 21213

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this thesis, the use of the h-likelihood for incomplete data is discussed. First, maximum likelihood (ML) estimation and maximum likelihood imputation based on the h-likelihood are examined. In missing data problem, model assumptions on the missing mechanism are not identifiable from the observed data. Thus, robust inference against model misspecification and outliers are presented by using the h-likelihood. However, finding canonical scale may not be available in complex models such as crossed and correlated random effect models. To obtain MLEs for fixed parameters in general statistical models, the enhanced Laplace approximation (ELA) method is proposed. After obtaining the MLEs of fixed parameters, the ML imputation can also be obtained by using the weak canonical scale. Finally, the use of the generalized extreme value distribution in analyzing extremely high censored data is investigated.

1.1 Maximum Likelihood Imputation

Missing data are prevalent in statistical problems, but ignoring them can lead to erroneous results (Little and Rubin, 2019; Kim and Shao, 2021). Imputation is a popular technique for dealing with missing data. However, if imputed data are treated as observed, the use of the standard statistical procedure could result in erroneous inference, giving a biased estimator with an underestimated standard error estimator. Multiple imputation has been proposed by Rubin (1987) to address the uncertainty associated with imputation. However, it requires the self-consistency conditions (Wang and Robins, 1998; Meng, 1994; Yang and Kim, 2016), which may not necessarily hold. An alternative method by Kim (2011) is fractional imputation.

ML estimation of Fisher (1922) is widely accepted in estimating fixed parameters. Missing data can be viewed as unobserved random parameters (Lee et al., 2017) so that imputation can be viewed as a prediction of random parameters, namely missing data. It necessitates an extension of the Fisher likelihood to statistical models that include unobserved random variables (Berger and Wolpert, 1984; Butler, 1986). Lee and Nelder (1996) intended an extension of ML estimation to models with unobserved random parameters via h-likelihood, defined on a particular scale of random parameters in the linear predictor. However, they confronted severe objections due to difficulties as Bayarri et al. (1988) showed that ML estimation of extended likelihood often provides nonsensical estimation for both fixed and random parameters. Furthermore, Firth (2006) noted that the linear predictor to form the

h-likelihood might not be necessarily well defined. All the counterexamples against the h-likelihood, for examples in Little and Rubin (2002), are associated with a wrong choice of scale to form h-likelihood. Little and Rubin (2019) described the current status of h-likelihood "Unlike maximization of the marginal likelihood of Fisher (1922), maximization of an extended likelihood does not generally give consistent estimates of the parameters (Breslow and Lin, 1995) ... Lee and Nelder (2001) and Lee et al. (2006) propose maximizing a "modification" ... which is the correct ML approach. For more details, see Lee and Nelder (2009) and the discussion, particularly Meng (2009)." The success of h-likelihood approach looks coincidental, so that Meng (2009) tried a rigorous theoretical justification for the use of h-likelihood by showing its Bartlett identities. But he ended up highlighting the difficulty caused by the difference between fixed and random parameter estimations. Thus, the benefit of using h-likelihood has not been well accepted yet. This chapter establishes the original aim of the h-likelihood whose maximization without any modification provides correct ML estimation and ML imputation by giving rigorous justifications.

Lee et al. (2006) defined h-likelihood precisely, but they have not fully exploited its usefulness. For example, an immediate drawback of the current h-likelihood is that it does not allow MLEs of variance components as Henderson's Henderson (1959) joint likelihood does not. So Lee et al. (2017) use a modification to obtain MLEs of variance components, etc. We need to reformulate the h-likelihood in a thoroughly consistent way to avoid modification. Jacobian terms do not play any role in Fisher's (1922) ML es-

timation of fixed parameters. However, in models with random parameters, as we shall show, Jacobian terms play a key role in ML estimation. This property has not been well known yet in literature. We clarify the role of the Jacobian term in defining h-likelihood. Currently, the h-likelihood has been defined mainly for random effect models, where linear predictors are defined (Lee and Nelder, 1996). To illustrate our proposal for a much wider class of models, we consider the imputation problem, which does not require a linear predictor, as noted by Firth (2006), and encounters difficulties in ML estimation of random parameters, as noted by Meng (2009). We clarify that the definitions of canonical scale and canonical function are keys to leading valid ML estimation on both fixed and random parameters without any modification in h-likelihood.

In Section 2.1, we describe the basic setup for missing data problem. In Section 2.2, we define the h-likelihood by using canonical scale and canonical function in terms of Jacobian term. Moreover, properties of MLEs for fixed and random parameters by using the h-likelihood are examined. In Section 2.3, we propose the weak canonical scale based on the Laplace approximation. The weak canonical scale can give proper ML imputation when the canonical scale is unknown. In Section 2.4, we propose the ML imputation by using the MLE for random parameters. Illustrative examples in Section 2.5 show the usefulness of the h-likelihood in the missing data problem.

1.2 Robust Imputation under Missing at Random

Missing data is a fundamental problem in statistics. Ignoring missing data may lead to biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings (Dong and Peng, 2013). However, missing mechanism may not be fully identified based on the observation (Molenberghs et al., 2008). Therefore, several assumptions are proposed for the generating process of missing data. The following three assumptions are widely accepted in analyzing missing data: missing-completely-at-random (MCAR), missing-at-random (MAR), and missing-not-at-random (MNAR). Under the MCAR assumption, missing mechanism does not depend on any observation. Therefore, statistical inference is available based on observation only but the MCAR assumption is often unrealistic. On the other hand, even though parameter estimation can be made easily by h-likelihood approach (Lee et al., 2017), the MNAR assumption is not necessarily useful as the model assumption since the missing mechanism is not identifiable by observation (Molenberghs et al., 2008; van Buuren, 2018). Among assumptions on the missing mechanism, the MAR assumption is widely used.

In the presence of missing data, the imputation method and the weighting method are frequently employed to estimate the parameters (Kim and Shao, 2021). Imputation is widely used to handle item nonresponse because it ensures that analysis results from different users are consistent. By appropriately incorporating observed auxiliary variables into the imputation model, imputa-

tion can reduce nonresponse bias and achieve efficient estimation. Popular methods of imputation include multiple imputation (Rubin, 1978) and fractional imputation (Kim, 2009).

Recently, Han et al. (2022a) proposed the ML imputation based on the h-likelihood. By using the h-likelihood, simple joint maximization directly gives estimation of fixed parameters and imputation of missing data. However, the correct specification of model may be difficult in the presence of missing data while any imputation method uses an imputation model, either implicitly or explicitly. How to make the imputation method less dependent on the imputation model is an important practical problem.

There are two main approaches in implementing a robust imputation method. One approach is to use a flexible model, nonparametric or semiparametric, to develop a robust imputation method. Nonparametric kernel regression imputation of Cheng (1994), semiparametric Gaussian mixture model imputation of Sang (2020), and the random forest imputation of Dagdoug et al. (2021) are examples of the robust imputation method using flexible models. The other approach is to use the propensity score (PS) model explicitly into the parameter estimation step for imputation to get doubly robust estimation. Doubly robust estimation has been investigated widely in the literature. For examples, see Bang and Robins (2005), Cao et al. (2009), Kim and Haziza (2014), Han and Wang (2013), and Chen and Haziza (2017).

In this chapter, we consider the second approach further and consider an extension of doubly robust estimation by establishing sufficient conditions for the asymptotic equivalence between the

imputation method and the weighting method based on the PS model. The imputation method gives a consistent estimator if the outcome regression (OR) model is correctly specified, whereas the weighting method gives a consistent estimator if the PS model is correctly specified. Under this equivalence, we can obtain double robustness as both the imputed estimator and the weighted estimator are consistent under the OR model and the PS model, respectively. Consequently, the internal bias calibration (IBC) condition proposed by Firth and Bennett (1998) can be applied to the imputation problem in the context of missing data. Based on the IBC condition, the estimating equation for regression coefficients takes the form of weighted least squares. We will demonstrate that the IBC condition can be achieved by introducing statistical models on mean and dispersion in view of the double hierarchical generalized linear model (DHGLM) proposed by Lee and Nelder (2006) in modeling approach. Given DHGLM, the h-likelihood permits MLEs of fixed parameters as well as ML imputation of random parameters, namely random effects and missing data.

Compared to the likelihood-based approach, Wang and Kim (2021) recently proposed obtaining the PS weight using the projection method relative to the Kullback-Leibler (KL) divergence in the information projection theory. The KL-divergence-based projection method is well-defined because it permits the moment-type constraint. To generalize the KL-divergence while maintaining the moment-type constraint, Eguchi (2021) proposed the γ-power divergence. The information projection approach based on the γ-power divergence gives a more general form of the optimal solu-

tion which contains additional scale parameter γ. Furthermore, the statistical model derived from γ-power divergence produces robust inferences against outliers. This robustness is also available within the framework of the DHGLM, as we shall see.

The structure of the chapter is as follows: In Section 3.1, ideas of the double robustness and the IBC condition with the basic setup are introduced. In Section 3.2, we examine how to obtain the imputation estimator based on the IBC condition. In Section 3.3, we present the use of the γ-power divergence to enlarge the class of propensity score models. In Section 3.4, we examine the IBC condition in modeling approach, especially DHGLM. Robust inference against outliers is also discussed. Simulation study in Section 3.5 shows the usefulness of the proposed method. All required evidences are presented in the Appendix.

1.3 Enhanced Laplace Approximation

Lee and Nelder (1996) proposed the use of the h-likelihood for making inferences about statistical models with latent variables which are widely used in various fields. Consider a hierarchical generalized linear model (HGLM) with $\mathrm{E}(y|z) = \mu$, $\mathrm{var}(y|z) = \phi V(\mu)$, and the linear predictor

$$\eta = g(\mu) = X\beta + L(\Sigma)z,$$

where $V(\mu)$ is the variance function, β indicates fixed effects, z indicates latent variables, namely random effects, and $\tau = (\phi, \Sigma)$ are dispersion parameters. The h-likelihood of the HGLM is written

as

$$H(\theta, z) = f_\theta(y, z) = f_\theta(y \mid z) f(z).$$

The h-likelihood consists of three objects: the observed data y, fixed unknown parameters $\theta = (\beta, \tau)$, and unobserved latent variables z. The marginal likelihood can be used to estimate the fixed parameters θ by integrating out the latent variables from the h-likelihood:

$$L_m(\theta) = f_\theta(y) = \int H(\theta, z) dz. \tag{1.1}$$

To make inferences about the random effects z, Lee et al. (2017) proposed the use of the predictive likelihood:

$$L_p(z|y; \theta) = f_\theta(z \mid y) = f_\theta(y, z)/f_\theta(y) = H(\theta, z)/L_m(\theta),$$

which is analogous to the use of a Bayesian posterior under a flat prior on θ.

In random effects models, the h-likelihood can be explicitly written, whereas the marginal and predictive likelihoods often involve intractable integration. The Gauss-Hermite quadrature can be used for the integral shown in (1.1). However, this formulation becomes numerically difficult as the dimension of integration increases (Hedeker and Gibbons, 2006). Instead, in random effects models, Lee and Nelder (2001) proposed the use of the Laplace approximation (LA) (Tierney and Kadane, 1986), which is widely used and has been implemented by various packages (Rue et al., 2009; Kristensen et al., 2016; Lee and Noh, 2018). Recently, Perry (2017) proposed a fast moment-based method for random effects models, which does not allow correlated random effects and is restricted to nested random effects models. Thus, this method cannot be used for crossed random effects models. In this chapter, for

the maximum likelihood (ML) estimation, we exploit an alternative expression of the marginal likelihood:

$$L_m(\theta) = H(\theta, z)/L_p(z \mid y; \theta). \tag{1.2}$$

For the log-likelihoods we use $h(\theta, z) = \log H(\theta, z)$, $\ell_m(\theta) = \log L_m(\theta)$, and $\ell_p(z|y;\theta) = \log L_p(z|y;\theta)$.

Lee and Nelder (2001) extended the restricted likelihood (Patterson and Thompson, 1971) for normal linear mixed models to HGLMs, which is important for estimating the dispersion parameter τ. However, there is no theoretical justification that the current approximate maximum likelihood estimator (MLE) and restricted maximum likelihood estimator (REMLE), which are based on the LA, are asymptotically equivalent to the true MLE and REMLE. Furthermore, how their consistent variance estimators could be obtained remains ambiguous. In this chapter, we propose the use of an integrated likelihood as a new restricted likelihood and introduce the enhanced LA (ELA), which provides the MLE, REMLE, and their consistent variance estimators.

1.4 Accelerated Failure Time Random Effect Model with GEV Distribution

In survival analysis, accelerated failure time (AFT) model has been widely used as an alternative to Cox's proportional hazard (PH) model. The main advantage of the AFT model is its direct interpretation between survival time and covariates (Ha et al., 2002, 2017). To enjoy this property, robustness against the misspecification of the distributional assumption should be guaranteed. The

robustness against the misspecification of the distributional assumption in the AFT model was presented in Ha et al. (Ha et al., 2002). However, investigating the robustness about more general cases including highly censored survival data has not been studied yet. The generalized extreme value (GEV) distribution with three parameters (location, scale and shape) allows a flexible modeling for skewed, heavy-tailed, and heavily censored data (Roy and Dey, 2014; Bladt and Albrecher, 2021). Clustered survival data allow correlation among individual survival times within the same cluster and they are often encountered in various clustered clinical studies such as a multi-center clinical trial, a dental study of teeth or implants, a pair or family study, and study of recurrent or multiple events (Hougaard, 2000; Ha et al., 2017). Random effects are useful to model such dependence. In this chapter, we are interested in the analysis of heavily censored clustered survival data. Thus, we propose an AFT random effect model with GEV distribution to allow a robust inference against heavily censored clustered data. The model inference is based on the h-likelihood (Lee and Nelder, 1996). Unlike the classical likelihood which only contains fixed parameters, the h-likelihood is constructed to have both fixed parameters and random parameters (Lee and Nelder, 1996). This chapter is organized as follows. In Section 5.1, we describe the AFT random effect model with GEV distribution. In Section 5.2, we derive the estimation procedure based on h-likelihood. The proposed method is demonstrated using simulation study in Section 5.3 and is illustrated with a practical example data set in Section 5.4. Technical details are given in Appendix.

Chapter 2

Maximum Likelihood Imputation

Chapter Summary

Maximum likelihood (ML) estimation is widely used in statistics. The h-likelihood has been proposed as an extension of Fisher's likelihood to statistical models including unobserved latent variables of recent interest. Its advantage is that the joint maximization gives ML estimators (MLEs) of both fixed and random parameters with their standard error estimates. However, the current h-likelihood approach does not allow MLEs of variance components as Henderson's joint likelihood does not in linear mixed models. In this chapter, we show how to form the h-likelihood in order to facilitate joint maximization for MLEs of whole parameters. We also show the role of the Jacobian term which allows MLEs in the presence of unobserved latent variables. To obtain MLEs for fixed parameters, intractable integration is not necessary. As an illustration, we show one-shot ML imputation for missing data by treating them as realized but unobserved random parameters. We show that the h-likelihood bypasses the expectation step in the expectation-maximization (EM) algorithm and allows single ML imputation instead of multiple imputations. We also discuss the difference in predictions in random effects and missing data.

2.1 Basic Setup

Assume that we have a study variable Y with dominating measure μ and a covariate vector $\boldsymbol{X}$. The study variable Y is subject to missingness and the covariates are always observed. Assume further that there are n independent and identically distributed realizations of $(\boldsymbol{X}, Y, \delta)$, denoted by $\{(\boldsymbol{x}_i, y_i, \delta_i) : i = 1, \ldots, n\}$, where δ_i is the missingness indicator defined by $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. We are interested in estimating $\eta = \mathrm{E}(Y)$ from the observed data.

Under existence of missing data, an imputation estimator of η can be written as

$$\hat{\eta}_{\mathrm{I}} = \frac{1}{n} \sum_{i=1}^{n} \{\delta_i y_i + (1 - \delta_i)\hat{y}_i\}$$

where $\hat{y}_i$ is the imputed value of y_i. To predict realized values y_i of unobserved missing data, we consider a frequentist approach using the ML imputation. The current procedure for ML imputation can be described as follows:

Step 1: Estimate ψ by maximizing the observed likelihood

$$\begin{aligned} L_m(\psi) &= f_\psi\left(y_{\mathrm{obs}}, \delta \mid \boldsymbol{x}\right) \\ &= \int f_\psi(y_{\mathrm{obs}}, y_{\mathrm{mis}}, \delta \mid \boldsymbol{x}) dy_{\mathrm{mis}}, \end{aligned} \tag{2.1}$$

where $f_\psi(y_{\mathrm{obs}}, y_{\mathrm{mis}}, \delta|\boldsymbol{x})$ is the joint density function of $(y_{\mathrm{obs}}, y_{\mathrm{mis}}, \delta)$ given $\boldsymbol{x}$ with fixed unknown parameter ψ and $(y_{\mathrm{obs}}, y_{\mathrm{mis}})$ is the observed and missing part of the complete data $y_{\mathrm{com}} = (y_1, \ldots, y_n)$, respectively.

Step 2: For each i with $\delta_i = 0$, obtain a predictor of y_i

$$\begin{aligned} \hat{y}_i &= \int y f(y \mid \boldsymbol{x}_i, \delta_i = 0; \hat{\psi}) d\mu(y) \\ &= \mathrm{E}_{\hat{\psi}} \left(Y_i \mid \boldsymbol{x}_i, \delta_i = 0 \right), \end{aligned} \tag{2.2}$$

where $\hat{\psi}$ is the MLE of ψ obtained from Step 1.

We use subscript m in the observed likelihood in (2.1) to emphasize that the likelihood is developed from the marginal density of the observed data. Robins and Wang (2000) and Kim and Shao (2021) present some asymptotic properties of the imputation estimator under ML imputation. The above two-step imputation procedure, however, is computationally involved as the ML estimation of the fixed parameter ψ is often based on the iterative procedure such as EM algorithm (Dempster et al., 1977). However, such a conditional mean imputation in (2.2) does not necessarily give the best prediction in terms of maximizing the predictive distribution. For example, if y is categorical, the conditional mean is not necessarily categorical.

In this chapter, instead of using the conditional mean imputation in (2.2), we propose using conditional mode of the h-likelihood given by

$$\hat{y}_{\text{mis}} = \arg\max_{y_{\text{mis}}} H(\hat{\psi}, y_{\text{mis}})$$

in the next section. In many practical situations, the conditional mode imputation is attractive as it respects the "maximum likelihood" principle by treating the unobserved y values as realized random parameters. By treating y_{mis} as the random parameters and applying the usual ML procedure, we can obtain imputed values, namely ML imputation, that adhere to the frequentist princi-

ple to the greatest extent possible. An immediate practical advantage is that one-shot imputation directly allows the ML estimation of fixed parameters. For one-shot imputation to be meaningful, as we shall show, it estimates the canonical function to predict future (or missing) variable, which resolves summarizability problem raised by Meng (2009).

Naively treating the missing observations as unknown parameters will be subject to biased estimation, which is well known as pointed out by Neyman and Scott (1948). Thus, we employ a technique known as h-likelihood (Lee and Nelder, 1996), to circumvent this issue and obtain valid inferences. Yun et al. (2007) studied the h-likelihood approach to estimate fixed parameters in missing data problems. We introduce the ML imputation of missing data and conduct a more systematic investigation, elucidating the mysteries of h-likelihood in general.

2.2 H-likelihood

In this chapter, we rearrange the indices as $\delta_i = 1$ for $i = 1, \ldots, n_{\text{obs}}$ and $= 0$ for $i = n_{\text{obs}} + 1, \ldots, n$ where $n_{\text{obs}} = \sum_{i=1}^{n} \delta_i$, i.e., the first n_{obs} responses are observed and remaining $n_{\text{mis}} = n - n_{\text{obs}}$ responses are not observed. Missing data can be viewed as prediction of future data which are not observed yet. By treating y_{mis} as random parameters, the complete-data log-likelihood is an

extended log-likelihood

$$\begin{aligned}\ell_e(\psi, y_{\text{mis}}) &= \log L_e(\psi, y_{\text{mis}}) \\ &= \log f_\psi(y_{\text{obs}}, y_{\text{mis}}, \delta \mid \boldsymbol{x}) \\ &= \sum_{i=1}^{n_{\text{obs}}} \log f_\psi(y_i, \delta_i = 1 \mid \boldsymbol{x}_i) \\ &\quad + \sum_{i=n_{\text{obs}}+1}^{n} \log f_\psi(y_{\text{mis},i}, \delta_i = 0 \mid \boldsymbol{x}_i).\end{aligned}$$

Extended likelihood principle (Bjørnstad, 1996) states that $L_e(\psi, y_{\text{mis}})$ carries all the information in the data about unknown parameters ψ and y_{mis}.

Lee and Nelder (1996) proposed the h-likelihood for ML estimation on both fixed and random parameters. Due to a Jacobian term, unlike a transformation of fixed parameter ψ, a nonlinear transformation of random parameter $v = g(y_{\text{mis}})$ changes the extended likelihood

$$L_e(\psi, v) = L_e(\psi, y_{\text{mis}}) \left| \frac{\partial y_{\text{mis}}}{\partial v} \right|.$$

Here if the joint maximization of $L_e(\psi, v)$ gives the MLE of ψ, that of $L_e(\psi, y_{\text{mis}})$ cannot give the MLE of ψ. It means that specifying the scale of a random parameter in defining the h-likelihood is important to obtain MLEs via its maximization. In this chapter, we elaborate on how to use the Jacobian term to form such an h-likelihood.

Following Lee et al. (2017), the predictive likelihood of random parameter v can be defined as

$$L_p(v \mid \mathcal{D}; \psi) \equiv f_\psi(v \mid \mathcal{D}, \boldsymbol{x}) = f_\psi(v, \mathcal{D} \mid \boldsymbol{x}) / f_\psi(\mathcal{D} \mid \boldsymbol{x}),$$

where $\mathcal{D} = \{y_{\text{obs}}, \delta\}$ and subscript p is used to emphasize the predictive likelihood for v. Thus, the marginal likelihood is expressed as

$$L_m(\psi) = \frac{L_e(\psi, v)}{L_p\left(v \mid \mathcal{D}; \psi\right)}. \tag{2.3}$$

Given ψ, let

$$\tilde{v} = \tilde{v}(\psi, \mathcal{D}, \boldsymbol{x}) = \arg\max_v L_e(\psi, v) = \arg\max_v L_p\left(v \mid \mathcal{D}; \psi\right) \tag{2.4}$$

be the common mode of the extended likelihood and the predictive likelihood. Note that the common mode $\tilde{v}(\psi, \mathcal{D}, \boldsymbol{x})$ is a function of both parameter and data. However, we denote it as $\tilde{v}$ for notational convenience. Evaluating the marginal likelihood in (2.3) at $v = \tilde{v}$ leads to

$$L_m(\psi) = \frac{L_e(\psi, \tilde{v})}{L_p(\tilde{v} \mid \mathcal{D}; \psi)}. \tag{2.5}$$

If both $L_e(\psi, v)$ and $L_p(v|\mathcal{D}; \psi)$ are explicitly available, at least at the mode $\tilde{v}$, the MLE for ψ is immediately obtained from (2.5). However, both are not often available.

Definition 2.2.1. If a scale $v = g(y_{\text{mis}})$ satisfies

$$L_e(\psi, \tilde{v}) \propto L_m(\psi), \tag{2.6}$$

the v-scale is called the *canonical scale* and the mode $\tilde{v}$ is called the *canonical function*. The extended likelihood defined on the canonical scale v is called the *h-likelihood*,

$$H(\psi, v) = L_e(\psi, v).$$

By combining (2.5) and (2.6), $L_p(\tilde{v}|\mathcal{D}; \psi)$ does not depend on ψ if v-scale is canonical, i.e. information neutral with respect to ψ at the mode $\tilde{v}$.

Here, we emphasize defining the h-likelihood with different parametrization of a random parameter. Let $\hat{\zeta}$ be the MLE of $\zeta = k(\psi)$ under the transformation $k(\cdot)$. Then, the MLE $\hat{\psi} = k^{-1}(\hat{\zeta})$ is invariant with respect to the transformation. Similarly, the MLE of a parameter from the h-likelihood is transformation invariant. That is, we can treat v as if it is a fixed parameter after defining the h-likelihood in the sense that

$$H(\psi, y_{\text{mis}}) = H\{\psi, g^{-1}(y_{\text{mis}})\} = H(\psi, v) \tag{2.7}$$

(Lee and Nelder, 2005). Here, we denote $H(\psi, y_{\text{mis}})$ the h-likelihood in terms of y_{mis} as (2.7), whereas $L_e(\psi, y_{\text{mis}})$ indicates the extended likelihood in which the canonical scale is yet unknown. From (2.7), the conditional mode of y_{mis} is defined by

$$\tilde{y}_{\text{mis}} = \arg\max_{y_{\text{mis}}} H(\psi, y_{\text{mis}}) = g^{-1}(\tilde{v}). \tag{2.8}$$

If the transformation $g(\cdot)$ is not linear, we get

$$\tilde{y}_{\text{mis}} \neq \arg\max_{y_{\text{mis}}} L_e(\psi, y_{\text{mis}}).$$

Thus, under the canonical condition (2.6), MLEs of both fixed and random parameters can be obtained by maximizing $H(\psi, v) = L_e(\psi, v)$.

Lee et al. (2017) gave a correct definition of canonical scale above, but have not exploited it to form the h-likelihood. We now state a sufficient condition for the canonical property in (2.6) as follows.

Proposition 2.2.1. If a transformation $v = g(y_{\text{mis}})$ with bijective, differentiable function $g(\cdot)$ satisfies

$$\left|\frac{\partial v}{\partial y_{\text{mis}}}\right|_{v=\tilde{v}} \propto L_p(\tilde{y}_{\text{mis}} \mid \mathcal{D}; \psi),$$

where $\tilde{y}_{\text{mis}} = g^{-1}(\tilde{v})$ and $\tilde{v}$ is defined in (2.4), the canonical property in (2.6) is satisfied.

Proposition 2.2.1 gives further interpretation about Definition 2.2.1.

$$L_m(\psi) = \frac{L_e(\psi, \tilde{y}_{\text{mis}})}{L_p(\tilde{y}_{\text{mis}} \mid \mathcal{D}; \psi)} \propto L_e(\psi, \tilde{y}_{\text{mis}}) \left| \frac{\partial y_{\text{mis}}}{\partial v} \right|_{v=\tilde{v}} = L_e(\psi, \tilde{v}) = H(\psi, \tilde{v}). \tag{2.9}$$

Moreover, it shows how the canonical scale allows ML estimation. Now, we first study the ML estimation of the fixed parameter using h-likelihood.

2.2.1 MLE of Fixed Parameter

Equation (2.9) characterizes the canonical scale which allows the ML estimation.

Theorem 2.2.1. Suppose that the predictive likelihood $L_p(y_{\text{mis}}|\mathcal{D}; \psi)$ is unimodal with respect to y_{mis}. Then, there exists the canonical scale to form the h-likelihood.

Theorem 2.2.1 states a sufficient condition for the existence of a canonical scale. When an explicit form of the canonical scale is not available, we present a way of defining a weak canonical scale based on the Laplace approximation in Section 2.3. For now, we assume that an explicit form of the canonical scale $v = g(y_{\text{mis}})$ is known. The following theorem shows how to obtain the MLE of fixed parameter and also its variance estimator.

Theorem 2.2.2. (i) The MLE of ψ can be obtained by solving the score equation

$$\frac{\partial \ell_m}{\partial \psi} = \frac{\partial}{\partial \psi} h(\psi, \tilde{v}) = \frac{\partial h}{\partial \psi}\Big|_{v=\tilde{v}} = 0,$$

where $h = \log H(\psi, v)$ and $\ell_m = \ell_m(\psi) = \log L_m(\psi)$.

(ii) The variance estimator of the MLE can be obtained from the Hessian matrix of the h-likelihood as

$$\hat{I}^{\psi\psi} = I^{\psi\psi}\big|_{\psi=\hat{\psi}},\ I^{\psi\psi} = \left(-\frac{\partial^2 \ell_m}{\partial\psi\partial\psi^T}\right)^{-1},$$

where the definition of $I^{\psi\psi}$ is in Appendix.

To compare the h-likelihood approach with the EM algorithm, note that

$$\frac{\partial \ell_m(\psi)}{\partial \psi} = \mathrm{E}_\psi\left\{\frac{\partial}{\partial\psi}\ell_e(\psi, y_{\mathrm{mis}})\Big|\mathcal{D}, \boldsymbol{x}\right\}.$$

This equality is called the mean score theorem (Louis, 1982). The EM algorithm (Dempster et al., 1977) obtains the solution to $\partial\ell_m(\psi)/\partial\psi = 0$ by

$$\psi^{(t+1)} \leftarrow \text{solve } \mathrm{E}_{\psi^{(t)}}\left\{\frac{\partial}{\partial\psi}\ell_e(\psi, y_{\mathrm{mis}})\Big|\mathcal{D}, \boldsymbol{x}\right\} = 0. \tag{2.10}$$

The h-likelihood approach gives the MLE of the fixed parameter without requiring the E-step in (2.10) which is often computationally intensive.

2.2.2 MLE of Random Parameter

If we let y_{mis} be the unobserved part of the data, the missing data problem becomes a prediction problem. To understand Meng's point in Meng (2009), assume that y_{obs} and y_{mis} are independent and the scale $v = g(y_{\mathrm{mis}})$ is the canonical scale. Prediction of future data can be viewed as missing data problem where $y_{t+1}, \ldots, y_{t+n_{\mathrm{mis}}}$ are future data at the present time $t = n_{\mathrm{obs}}$. Meng (2009) showed that

$$\hat{v} - v = g\left(\hat{y}_{\mathrm{mis}}\right) - g\left(y_{\mathrm{mis}}\right) = g'\left(\tilde{y}_{\mathrm{mis}}\right)\left(\hat{y}_{\mathrm{mis}} - y_{\mathrm{mis}}\right) + R_{n_{\mathrm{obs}}},$$

where

$$R_{n_{\text{obs}}} = O_p(1) \text{ and } g'(\tilde{y}_{\text{mis}})(\hat{y}_{\text{mis}} - y_{\text{mis}}) = O_p(1).$$

Meng (2009) claimed that $\hat{v} - v$ is not summarizable because of the *nonnegligiblity* of the remainder term $R_{n_{\text{obs}}}$, i.e., consistency and asymptotic normality for the MLE $\hat{v}$ from the h-likelihood are not guaranteed.

Now we investigate the summarizability properties of the MLE $\hat{v}$. In missing data problem, the ML estimation of random parameter can be called the ML imputation. Let ψ_0 be the true value of ψ. As MLE $\hat{\psi}$ is estimating ψ_0 and similarly the MLE $\hat{y}_{\text{mis}}$ predicts a realized value of y_{mis} by estimating the conditional mode $y_{\text{mis},0} = \tilde{y}_{\text{mis}}(\psi_0, \mathcal{D}, \boldsymbol{x})$ in (2.8), which is a function of data and unknown parameter ψ_0. This clarifies the summarizability problem raised by Meng (2009); while $\hat{y}_{\text{mis}} - y_{\text{mis}}$ is not summarizable, $\hat{y}_{\text{mis}} - y_{\text{mis},0}$ is summarizable as in Theorem 2.2.3 below. Note that

$$y_{\text{mis}} - \hat{y}_{\text{mis}} = y_{\text{mis},0} - \hat{y}_{\text{mis}} + \varepsilon,$$

where $\varepsilon = y_{\text{mis}} - y_{\text{mis},0}$. In missing data problem, $\varepsilon = O_p(1)$. In view of predicting unobservable future (or missing) random variable, we estimate ε as null. Then, $\hat{y}_{\text{mis}}$ is estimating $y_{\text{mis},0}$ to predict y_{mis}. Thus, we obtain

$$\text{var}_\psi\left(\hat{y}_{\text{mis}} - y_{\text{mis}}\right) = \text{var}_\psi(\hat{y}_{\text{mis}} - y_{\text{mis},0}) + \text{var}_\psi(\varepsilon|\mathcal{D}, \boldsymbol{x}).$$

The first term is the variance due to estimating $y_{\text{mis},0}$ by $\hat{y}_{\text{mis}}$ and the second term is the variance due to the unidentifiable error term ε. The second term may decrease with a better imputation model, but it does not decrease with larger sample size. Moreover, to

obtain a standard error for prediction of $y_{\rm mis}$, we need to estimate the conditional variance of ε by using

$$\mathrm{var}_\psi(\varepsilon \mid \mathcal{D}, \boldsymbol{x}) = \mathrm{var}_\psi(y_{\rm mis} - y_{\rm mis,0} \mid \mathcal{D}, \boldsymbol{x}) = \mathrm{var}_\psi(y_{\rm mis} \mid \mathcal{D}, \boldsymbol{x}).$$

Here, we are interested in estimating $\mathrm{var}(\hat{y}_{\rm mis} - y_{\rm mis})$. Thus, we write the h-likelihood with respect to $y_{\rm mis}$ as $h = h(\psi, y_{\rm mis}) = h\{\psi, g^{-1}(v)\}$. Note that

$$\frac{\partial \tilde{y}_{\rm mis}^{\rm T}}{\partial \psi} = -I_{\psi y_{\rm mis}} I_{y_{\rm mis} y_{\rm mis}}^{-1}$$

and the variance estimator of $\hat{\psi}$ is $\hat{I}^{\psi\psi}$ by Theorem 2.2.2, where $I_{\psi y_{\rm mis}} = -\partial^2 h/\partial\psi\partial y_{\rm mis}^{\rm T}|_{y_{\rm mis}=\tilde{y}_{\rm mis}}$ and $I_{y_{\rm mis} y_{\rm mis}} = -\partial^2 h/\partial y_{\rm mis}\partial y_{\rm mis}^{\rm T}|_{y_{\rm mis}=\tilde{y}_{\rm mis}}$. Then, by using the delta method, we have the asymptotic normality of $\hat{y}_{\rm mis}$ as follows.

Theorem 2.2.3. Under regularity conditions in Appendix, we have

$$\sqrt{n_{\rm obs}}\,(\hat{y}_{\rm mis} - y_{\rm mis,0}) \xrightarrow{\rm d} \mathrm{N}\,(0, V)\,,$$

where $V = \lim_{n_{\rm obs}\to\infty} n_{\rm obs} \hat{I}_{y_{\rm mis} y_{\rm mis}}^{-1} \hat{I}_{y_{\rm mis}\psi} \hat{I}^{\psi\psi} \hat{I}_{\psi y_{\rm mis}} \hat{I}_{y_{\rm mis} y_{\rm mis}}^{-1}$ and $\hat{I}_{\psi y_{\rm mis}}$, $\hat{I}_{y_{\rm mis} y_{\rm mis}}$ are evaluated at $\psi = \hat{\psi}$. The variance of $\hat{y}_{\rm mis} - y_{\rm mis,0}$ can be estimated as

$$\begin{aligned} \widehat{\mathrm{var}}\,(\hat{y}_{\rm mis} - y_{\rm mis,0}) &= \mathrm{var}_{\hat{\psi}}\,(\hat{y}_{\rm mis} - y_{\rm mis,0}) \\ &= \hat{I}_{y_{\rm mis} y_{\rm mis}}^{-1} \hat{I}_{y_{\rm mis}\psi} \hat{I}^{\psi\psi} \hat{I}_{\psi y_{\rm mis}} \hat{I}_{y_{\rm mis} y_{\rm mis}}^{-1}. \quad (2.11) \end{aligned}$$

If $\mathrm{E}_\psi(\varepsilon) = 0$, $\hat{y}_{\rm mis}$ is an asymptotically unbiased estimator of $y_{\rm mis}$. However, the assumption $\mathrm{E}_\psi(\varepsilon) = 0$ is coming from model assumption which may not be identifiable by observed data. Now, to discuss the estimation of the variance due to the model error

ε, suppose that there exists a normalizing transformation $z = k(v) = k\{g(y_{\text{mis}})\} = k \circ g(y_{\text{mis}}) = r(y_{\text{mis}})$ with $r(\cdot) = k \circ g(\cdot)$ such that $L_p(z|\mathcal{D};\psi)$ is from the normal density with mean $\tilde{z} = \arg\max_z L_p(z|\mathcal{D};\psi)$ and covariance matrix I_{zz}^{-1}, where $I_{zz} = -\partial^2 h(\psi,z)/\partial z \partial z^{\mathrm{T}}|_{z=\tilde{z}}$. Then, it gives the h-likelihood

$$h\left(\psi,z\right) = \ell_m(\psi) + \frac{1}{2}\log\left|\frac{1}{2\pi}I_{zz}\right| - \frac{1}{2}\left(z-\tilde{z}\right)^{\mathrm{T}} I_{zz}\left(z-\tilde{z}\right).$$

Here, $\tilde{z} = \mathrm{E}_\psi(z|\mathcal{D},\boldsymbol{x}) = r(\tilde{y}_{\text{mis}})$ provided by the normality of the predictive likelihood $L_p(z|\mathcal{D};\psi)$. This leads to $\mathrm{E}_\psi(\varepsilon) = \mathrm{E}_\psi(z - z_0) = 0$,

$$\mathrm{var}_\psi\left(\hat{z}-z\right) = \mathrm{var}_\psi\left(\hat{z}-z_0\right) + \mathrm{E}_\psi\left\{\mathrm{var}_\psi\left(z_0 - z \mid \mathcal{D},\boldsymbol{x}\right)\right\}$$

and $\widehat{\mathrm{var}}(z_0 - z|\mathcal{D},\boldsymbol{x}) = \hat{I}_{zz}^{-1}$, where $\hat{z} = r(\hat{y}_{\text{mis}})$ and $z_0 = r(y_{\text{mis},0}) = \mathrm{E}_{\psi_0}(z|\mathcal{D},\boldsymbol{x})$. This gives

$$\begin{aligned}
\widehat{\mathrm{var}}\left(\hat{z}-z\right) &= \widehat{\mathrm{var}}\left(\hat{z}-z_0\right) + \widehat{\mathrm{var}}\left(z_0 - z \mid \mathcal{D},\boldsymbol{x}\right) \\
&= \hat{I}_{zz}^{-1}\hat{I}_{z\psi}\hat{I}^{\psi\psi}\hat{I}_{\psi z}\hat{I}_{zz}^{-1} + \hat{I}_{zz}^{-1} \\
&= \hat{I}^{zz}.
\end{aligned}$$

Therefore, if a normalizing transformation exists, the h-likelihood gives not only MLEs of both fixed and random parameters, but also their corresponding variance estimators. Moreover, if y_{mis} itself satisfies normal approximation well, then, we can have a reasonable variance estimator from the Hessian matrix of h-likelihood

$$\begin{aligned}
\widehat{\mathrm{var}}\left(\hat{y}_{\text{mis}} - y_{\text{mis}}\right) &= \widehat{\mathrm{var}}\left(\hat{y}_{\text{mis}} - y_{\text{mis},0}\right) + \widehat{\mathrm{var}}\left(y_{\text{mis},0} - y_{\text{mis}} \mid \mathcal{D},\boldsymbol{x}\right) \\
&= \hat{I}_{y_{\text{mis}}y_{\text{mis}}}^{-1}\hat{I}_{y_{\text{mis}}\psi}\hat{I}^{\psi\psi}\hat{I}_{\psi y_{\text{mis}}}\hat{I}_{y_{\text{mis}}y_{\text{mis}}}^{-1} + \hat{I}_{y_{\text{mis}}y_{\text{mis}}}^{-1} \\
&= \hat{I}^{y_{\text{mis}}y_{\text{mis}}}.
\end{aligned}$$

Thus, $\hat{y}_{\text{mis},i} \pm 1.96\sqrt{\hat{I}_{ii}^{y_{\text{mis}}y_{\text{mis}}}}$ is 95% predictive interval of $y_{\text{mis},i}$, where $\hat{I}_{ii}^{y_{\text{mis}}y_{\text{mis}}}$ is the ith diagonal element of $\hat{I}^{y_{\text{mis}}y_{\text{mis}}}$. The length of predictive interval is $O_p(1)$ and coverage probability becomes exact as $n \to \infty$ (Lee and Kim, 2016). However, in practice, the normalizing transformation is not known. Thus, in general, for the prediction of y_{mis}, Lee and Kim (2016, 2020) proposed to use the predictive distribution after eliminating ψ defined as

$$f(y_{\text{mis}} \mid \mathcal{D}, \boldsymbol{x}) = \int f_{\psi}(y_{\text{mis}} \mid \mathcal{D}, \boldsymbol{x}) c(\psi) d\psi, \tag{2.12}$$

where $c(\psi)$ is the confidence density (Schweder and Hjort, 2016). By using the predictive likelihood (2.12), we can account for the uncertainty caused by estimating ψ. Via simulation studies, Lee and Kim (2016) showed that resulting predictive interval maintains the stated coverage probability well as n grows.

From Theorem 2.2.2, MLE $\hat{\psi}$ from the marginal likelihood can be obtained by

$$\frac{\partial \ell_m(\psi)}{\partial \psi} = \frac{\partial h(\psi, \tilde{v})}{\partial \psi} = 0$$

and ML imputation $\hat{y}_{\text{mis}} = g^{-1}(\hat{v})$ of $y_{\text{mis}} = g^{-1}(v)$ from the predictive likelihood can be obtained by

$$\frac{\partial \ell_p(v \mid \mathcal{D}; \tilde{\psi})}{\partial v} = \frac{\partial h(\tilde{\psi}, v)}{\partial v} = 0,$$

where $\tilde{\psi}$ is solution to $\partial h(\psi, v)/\partial \psi = 0$. In contrast to the EM algorithm, the h-likelihood provides not only the ML estimation for fixed parameters from $H(\psi, \tilde{v})$, but also ML imputation on random parameters from $H(\tilde{\psi}, v)$ as in Figure 2.1. Moreover, the necessary standard error estimates are also given straightforwardly.

$$\text{Find } \tilde{v} \text{ by solving } \frac{\partial h(\psi,v)}{\partial v} = 0 \quad \left(\begin{array}{rcl} H(\psi,\tilde{v}) &=& L_p(\tilde{v} \mid \mathcal{D};\psi) L_m(\psi) \\ &\propto& L_m(\psi) \\ \\ H(\tilde{\psi},v) &=& L_p(v \mid \mathcal{D};\tilde{\psi}) L_m(\tilde{\psi}) \\ &\propto& L_p(v \mid \mathcal{D};\tilde{\psi}) \end{array} \right) \quad \text{Find } \tilde{\psi} \text{ by solving } \frac{\partial h(\psi,v)}{\partial \psi} = 0$$

Figure 2.1: Estimation procedure of the h-likelihood.

Example 2.2.1. Suppose that n variables are generated from the exponential distribution with mean θ_0 but only the first $n-1$ variables are observed, i.e., $n_{\text{obs}} = n-1$ and $y_{\text{mis}} = y_n$ is not observed. In this example, the extended likelihood defined on y_{mis}-scale is

$$\ell_e(\theta, y_{\text{mis}}) = -n \log \theta - \frac{(n-1)\bar{y}_{\text{obs}} + y_{\text{mis}}}{\theta}.$$

Note that y_{mis}-scale is not canonical but $v = \log y_{\text{mis}}$ is a canonical scale which gives

$$h(\theta, v) = \ell_e(\theta, y_{\text{mis}}) + \log \left| \frac{\partial y_{\text{mis}}}{\partial v} \right| = -n \log \theta - \frac{(n-1)\bar{y}_{\text{obs}} + e^v}{\theta} + v$$

and

$$h(\theta, y_{\text{mis}}) = -n \log \theta - \frac{(n-1)\bar{y}_{\text{obs}} + y_{\text{mis}}}{\theta} + \log y_{\text{mis}}.$$

Here, the canonical function of y_{mis} is $\tilde{y}_{\text{mis}} = \theta$ which gives the MLE $\hat{\theta} = \bar{y}_{\text{obs}}$ and ML imputation $\hat{y}_{\text{mis}} = \hat{\theta} = \bar{y}_{\text{obs}}$. In this example, the MLE of θ, $\hat{\theta}$, satisfies the asymptotic normality

$$\sqrt{n_{\text{obs}}} \left(\hat{\theta} - \theta_0 \right) \xrightarrow{\text{d}} \text{N}\left(0, \theta_0^2\right).$$

By Theorem 2.2.3, the ML imputation for y_{mis}, $\hat{y}_{\text{mis}}$, satisfies the asymptotic normality

$$\sqrt{n_{\text{obs}}} \left(\hat{y}_{\text{mis}} - y_{\text{mis},0} \right) \xrightarrow{\text{d}} \text{N}\left(0, \theta_0^2\right),$$

where $y_{\text{mis},0} = \theta_0$ and $\widehat{\text{var}}(\hat{y}_{\text{mis}} - y_{\text{mis},0}) = n_{\text{obs}}^{-1}\hat{\theta}^2$, i.e., (2.11) gives valid variance estimator of $\hat{y}_{\text{mis}} - y_{\text{mis},0}$. Moreover,

$$\hat{I}^{y_{\text{mis}}y_{\text{mis}}} = \hat{\theta}^2\left(1 + \frac{1}{n_{\text{obs}}}\right) = \widehat{\text{var}}(\hat{y}_{\text{mis}} - y_{\text{mis}}).$$

Here $\widehat{\text{var}}(\hat{y}_{\text{mis}} - y_{\text{mis}}) = \widehat{\text{var}}\,(\hat{y}_{\text{mis}} - y_{\text{mis},0}) + \widehat{\text{var}}(y_{\text{mis}}|y_{\text{obs}}) = \hat{\theta}^2/n_{\text{obs}} + \hat{\theta}^2$. Thus, the h-likelihood gives a correct ML imputation. In this example, $\tilde{y}_{\text{mis},0} = \theta$ is a function of parameter only so that $\hat{y}_{\text{mis}} - y_{\text{mis},0}$ is summarizable. But y_{mis} is not identifiable since $\varepsilon = O_p(1)$ with $\text{E}_\theta(\varepsilon) = 0$. Asymptotically correct probability statement on y_{mis} can be made based on predictive interval whose length is $O_p(1)$.

Example 2.2.2. Consider a one-way mixed model

$$y_{ij} = \mu + u_i + \epsilon_{ij},\ i = 1, \ldots, q,\ j = 1, \ldots, n,$$

where random effects u_i are iid $\text{N}(0, \lambda^2)$, ϵ_{ij} are iid $\text{N}(0, \sigma^2)$ and u_i and ϵ_{ij} are independent. Henderson's (1959) joint likelihood is the current h-likelihood of Lee and Nelder (1996)

$$\begin{aligned}\ell_e(\theta, u) &= \sum_{i,j}\left\{-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(y_{ij} - \mu - u_i)^2\right\} \\ &\quad + \sum_i\left(-\frac{1}{2}\log 2\pi\lambda^2 - \frac{1}{2\lambda^2}u_i^2\right), \qquad (2.13)\end{aligned}$$

where $\theta = (\mu, \sigma^2, \lambda^2)$. However, joint maximization of (2.13) cannot give the MLEs of variance components σ^2 and λ^2. Consider a v-scale

$$v_i = \left\{-\frac{\partial^2\ell_e(\theta, u)}{\partial u_i^2}\right\}^{0.5} u_i = \left(\frac{\sigma^2 + n\lambda^2}{\sigma^2\lambda^2}\right)^{0.5} u_i,$$

which leads to the extended likelihood

$$\begin{aligned}\ell_e(\theta, v) &= \ell_e(\theta, u) + \log\left|\frac{\partial u}{\partial v}\right| \\ &= -\frac{N-q}{2}\log 2\pi\sigma^2 - \frac{q}{2}\log 2\pi\left(\sigma^2 + n\lambda^2\right) \\ &\quad -\frac{1}{2\sigma^2}\sum_{i,j}\left\{y_{ij} - \mu - \left(\frac{\sigma^2\lambda^2}{\sigma^2+n\lambda^2}\right)^{0.5} v_i\right\}^2 \\ &\quad -\frac{\sigma^2}{2(\sigma^2+n\lambda^2)}\sum_i v_i^2 - \frac{q}{2}\log 2\pi,\end{aligned}$$

where $N = qn$. Since $\ell_e(\theta, \tilde{v}) = \ell_m(\theta)$, where

$$\tilde{v}_i = \tilde{v}_i(\theta, y_i) = \frac{n\lambda^2(\bar{y}_{i\cdot} - \mu)}{\{\sigma^2\lambda^2(\sigma^2+n\lambda^2)\}^{0.5}},$$

$y_i = (y_{i1}, \ldots, y_{in})$ and $\bar{y}_{i\cdot} = n^{-1}\sum_{j=1}^n y_{ij}$, we have h-likelihood $h = \ell_e(\theta, v)$, whose simple maximization gives MLEs of the whole parameters θ. Also, it gives the best linear unbiased predictors for realized but unobserved random parameters

$$\hat{u}_i = \tilde{u}_i\left(\hat{\theta}, y_i\right) = \widehat{\mathrm{E}(u_i \mid y_i)},\ i = 1, \ldots, q,$$

where

$$\tilde{u}_i(\theta, y_i) = \left\{\frac{\sigma^2\lambda^2}{\sigma^2+n\lambda^2}\right\}^{\frac{1}{2}} \tilde{v}_i(\theta, y_i) = \frac{n\lambda^2}{\sigma^2+n\lambda^2}(\bar{y}_{i\cdot} - \mu) = \mathrm{E}_\theta(u_i \mid y_i).$$

In this example, the target of $\hat{u}_i$ is

$$u_{i0} = \tilde{u}_i(\theta_0, y_i) = \mathrm{E}_{\theta_0}(u_i \mid y_i),$$

where $\theta_0 = (\mu_0, \sigma_0^2, \lambda_0^2)$ is the true value of θ. If the MLE $\hat{\theta}$ converges to θ_0,

$$\widehat{\mathrm{var}}\left(\hat{u}_i - u_i\right) = \widehat{\mathrm{var}}\left(\hat{u}_i - u_{i0}\right) + \widehat{\mathrm{var}}\left(u_i - u_{i0} \mid y_i\right) \xrightarrow{\mathrm{P}} 0$$

as $(q, n) \to \infty$. Thus, in this example, we have a consistent estimator of unobserved random parameter u_i, i.e., u_i is identifiable with $\varepsilon = u_i - u_{i0} = o_p(1)$. This can also be shown that

$$\begin{aligned}
\lim_{(q,n)\to\infty} \hat{u}_i &= \lim_{(q,n)\to\infty} u_{i0} \\
&= \lim_{(q,n)\to\infty} \frac{n\lambda_0^2}{\sigma_0^2 + n\lambda_0^2}(\bar{y}_{i\cdot} - \mu_0) \\
&= \lim_{(q,n)\to\infty} \frac{n\lambda_0^2}{\sigma_0^2 + n\lambda_0^2}(u_i + \bar{\epsilon}_{i\cdot}) = u_i,
\end{aligned}$$

where $\bar{\epsilon}_{i\cdot} = n^{-1}\sum_{j=1}^{n} \epsilon_{ij}$. Model assumptions on u_i can also be checkable: for various model checking plots, see Lee et al. (2017). Furthermore, if different model assumptions on $f_\psi(u)$ lead to an identical h-likelihood, then it leads to equivalent inferences for identifiable random effects (Lee and Nelder, 2006). In missing data problem with $\varepsilon = y_{\text{mis}} - y_{\text{mis},0} = O_p(1)$, model assumptions $f_\psi(y_{\text{mis}}|\mathcal{D}, \boldsymbol{x})$ cannot be checkable from the observed data (Molenberghs et al., 2008).

Since u_i itself is the normalizing transformation in this example, variances can be estimated as

$$\begin{aligned}
\hat{I}^{-1}_{u_i u_i} &= \left(-\frac{\partial^2 h}{\partial u_i^2}\right)^{-1}\Big|_{\theta=\hat{\theta}} = \frac{\hat{\sigma}^2\hat{\lambda}^2}{\hat{\sigma}^2 + n\hat{\lambda}^2} = \widehat{\text{var}}\,(u_i \mid y_i) = \widehat{\text{var}}\,(u_i - u_{i0} \mid y_i)\,, \\
\hat{I}^{u_i u_i} &= \hat{I}^{-1}_{u_i u_i} + \frac{\partial \tilde{u}_i}{\partial \theta^{\mathrm{T}}} \widehat{\text{var}}\left(\hat{\theta}\right) \frac{\partial \tilde{u}_i}{\partial \theta}\Big|_{\theta=\hat{\theta}} = \widehat{\text{var}}\,(\hat{u}_i - u_i)\,, \\
\hat{I}^{\theta\theta} &= \widehat{\text{var}}\left(\hat{\theta}\right).
\end{aligned}$$

Thus, proper MLEs of both fixed and random parameters and their variance estimators can be obtained by the maximization of the newly defined h-likelihood, which differs from the joint likelihood of Henderson (1959). Asymptotically correct probability statement

on u_i can be made from the predictive interval whose length is $o_p(1)$. For more details about general random effect models, see Paik et al. (2015), Lee et al. (2017), and Lee and Kim (2020).

2.3 Scale for Joint Maximization

When the canonical scale is unknown, Lee et al. (2017) proposed the use of the Laplace approximation to give an approximate MLE (Tierney and Kadane, 1986), which has been implemented by various packages (Kristensen et al., 2016; Ha et al., 2019). In this section, we study how to form an h-likelihood with a weak canonical scale whose joint maximization provides approximate MLEs obtained by the Laplace approximation. Given y_{mis}-scale, consider a b-scale with $b = g_1(y_{\text{mis}})$. Let Ω_b be the support of b taking a rectangle form $\Omega_b = \prod_{i=n_{\text{obs}}+1}^{n}[l_i, u_i]$, where l_i and u_i are permitted to take the value of $-\infty$ and ∞ with boundary set $\partial\Omega_b$, $\xi = (\psi, b)$ and $f_\psi(b)$ be the density function of b. Meng (2009) studied the regularity conditions for the first and second Bartlett identities of an extended likelihood $\ell_e(\psi, b)$.

Theorem 2.3.1 (Meng, 2009)**.** (i) If $f_\psi(b) = 0$ for any $b \in \partial\Omega_b$, the first Bartlett identity holds.

$$\mathrm{E}_\psi\left[\frac{\partial}{\partial\xi}\ell_e(\psi, b)\right] = 0. \tag{2.14}$$

(ii) Furthermore, if $\partial f_\psi(b)/\partial b = 0$ for any $b \in \partial\Omega_b$, the second Bartlett identity holds.

$$\mathrm{E}_\psi\left[\left(\frac{\partial}{\partial\xi}\ell_e(\psi, b)\right)\left(\frac{\partial}{\partial\xi}\ell_e(\psi, b)\right)^{\mathrm{T}}\right] + \mathrm{E}_\psi\left[\frac{\partial^2}{\partial\xi\partial\xi^{\mathrm{T}}}\ell_e(\psi, b)\right] = O. \tag{2.15}$$

Corollary below gives an easy way of having a b-scale to satisfy Bartlett identities.

Corollary 2.3.1. Let $\Omega_b = \mathbb{R}^{n_{\text{mis}}}$. If $\mathrm{E}_\psi(b_i) < \infty$ for all i, the b-scale satisfies Bartlett identities.

The second Bartlett identity (2.15) guarantees that the predictive likelihood $L_p(b|\mathcal{D};\psi)$ is unimodal with respect to b even though $L_p(y_{\text{mis}}|\mathcal{D};\psi)$ may not be unimodal. From Theorem 2.2.1, if we have such an extended likelihood $L_e(\psi, b)$ there exists the canonical scale $v = g(b)$ to form the h-likelihood. But, the explicit form of $g(\cdot)$ for the canonical scale may not be known. In this case, we may consider an approximation of canonical scale based on the Laplace approximation, which is widely used to obtain an approximate MLE of fixed parameter, $\hat{\psi}^{\text{Lap}}$ (Raudenbush et al., 2000; Lee et al., 2017).

Definition 2.3.1. Suppose that b-scale satisfies the Bartlett identities and $\ell_e(\psi, b)$ is the corresponding extended log-likelihood. Now, consider a w-scale defined as

$$w = g_2(b) = \tilde{\Omega}_{bb}^{\frac{1}{2}} b, \tag{2.16}$$

where $\tilde{b} = \tilde{b}(\psi, \mathcal{D}, \boldsymbol{x}) = \arg\max_b \ell_e(\psi, b)$ and $\tilde{\Omega}_{bb} = -\partial^2 \ell_e(\psi, b)/\partial b \partial b^{\mathrm{T}}|_{b=\tilde{b}}$. Here, we call w-scale *weak canonical* and

$$H = L_e(\psi, w) = L_e(\psi, b) \left| \frac{\partial b}{\partial w} \right|$$

the h-likelihood with weak canonical scale w.

By the above definition, weak canonical scale also satisfies Bartlett identities in (2.14) and (2.15) since the transformation

(2.16) is linear. Furthermore, we have

$$\tilde{w} = \tilde{w}(\psi, \mathcal{D}, x) = \arg\max_{w} L_e(\psi, w) = g_2\{\tilde{b}(\psi, \mathcal{D}, x)\}$$

since $\tilde{b}$ is the mode of $L_e(\psi, b)$ and the transformation $g_2(\cdot)$ is linear. Note that the joint maximization of the h-likelihood with weak canonical scale gives the approximate MLE for ψ based on the Laplace approximation as follows.

$$\hat{L}_m(\psi) = L_e\left(\psi, \tilde{b}\right)\left|\frac{1}{2\pi}\tilde{\Omega}_{bb}\right|^{-\frac{1}{2}} \propto L_e\left(\psi, \tilde{b}\right)\left|\frac{\partial b}{\partial w}\right|_{w=\tilde{w}} = L_e\left(\psi, \tilde{w}\right).$$

This weak canonical scale does not require the existence of linear predictor. In HGLMs, a scale satisfying additivity in the linear predictor is called a weak canonical scale (Lee et al., 2017), which satisfies Corollary 2.3.1. In Appendix, we show how to compute the standard error estimate of the approximate MLE obtained from $\ell_e(\psi, \tilde{w}) = \log L_e(\psi, \tilde{w})$.

2.4 ML Imputation

In this section, we propose the ML imputation via h-likelihood.

Definition 2.4.1. With the canonical scale $v_i = g(y_{\text{mis},i})$ and the canonical function $\tilde{v}_i(\psi, \mathcal{D}, \boldsymbol{x})$, the ML imputation gives imputed values

$$\hat{y}_{\text{mis},i} = g^{-1}(\hat{v}_i),\ \hat{v}_i = \tilde{v}_i\left(\hat{\psi}, \mathcal{D}, \boldsymbol{x}\right). \tag{2.17}$$

Theorem 2.2.3 implies that the MLE of a random parameter is a consistent estimator of the canonical function. Based on the ML imputation (2.17), we propose to use the estimator

$$\bar{y}_{\text{ML}} = \frac{1}{n}\left(\sum_{i=1}^{n_{\text{obs}}} y_i + \sum_{i=n_{\text{obs}}+1}^{n} \hat{y}_{\text{mis},i}\right)$$

as an estimator of $\eta = \mathrm{E}(Y)$. If the canonical scale is unknown, the ML imputation based on the weak canonical scale can be used. Weak canonical scale always exists and is known. This scale gives the estimator of η as

$$\bar{y}_{\mathrm{ML}}^{\mathrm{Lap}} = \frac{1}{n}\left(\sum_{i=1}^{n_{\mathrm{obs}}} y_i + \sum_{i=n_{\mathrm{obs}}+1}^{n} \hat{y}_{\mathrm{mis},i}^{\mathrm{Lap}}\right),$$

where $\hat{y}_{\mathrm{mis}}^{\mathrm{Lap}} = g^{-1}(\hat{w})$, $\hat{w} = \tilde{w}(\hat{\psi}^{\mathrm{Lap}}, \mathcal{D}, x)$ and $g = g_2 \circ g_1$. From Theorem 2.2.1 and the definition of the weak canonical scale (2.16), we see that the canonical scale is a linear transformation of the weak canonical scale w. Given ψ, MLEs of random parameters are invariant with respect a linear transformation (Lee and Nelder, 2005) and

$$\begin{aligned}
\hat{\ell}_m(\psi) - \ell_m(\psi) &= \ell_p(y_{\mathrm{mis}} \mid \mathcal{D}; \psi) - \hat{\ell}_p(y_{\mathrm{mis}} \mid \mathcal{D}; \psi) \\
&= \{\ell_p(v \mid \mathcal{D}; \psi) + \log|\partial v/\partial y_{\mathrm{mis}}|\} \\
&\quad - \left\{\hat{\ell}_p(v \mid \mathcal{D}; \psi) + \log|\partial v/\partial y_{\mathrm{mis}}|\right\} \\
&= \ell_p(v \mid \mathcal{D}; \psi) - \hat{\ell}_p(v \mid \mathcal{D}; \psi).
\end{aligned}$$

Thus, the ML imputation under weak canonical scale is valid in the sense that

$$\hat{y}_{\mathrm{mis}}^{\mathrm{Lap}} - \hat{y}_{\mathrm{mis}} = O_p\left(\left|\hat{\psi}^{\mathrm{Lap}} - \hat{\psi}\right|\right),$$

where $\hat{\ell}_p(y_{\mathrm{mis}}|\mathcal{D};\psi) = \log \hat{L}_p(y_{\mathrm{mis}}|\mathcal{D};\psi)$ and $\hat{L}_e(y_{\mathrm{mis}}|\mathcal{D};\psi) = L_e(\psi, y_{\mathrm{mis}})/\hat{L}_m(\psi)$. Recently, Han and Lee (2022) developed the enhanced Laplace approximation (ELA) to obtain the MLE $\hat{\psi}$ generally. Thus, the ML imputation can be always implemented even when the canonical scale is not known by using a weak canonical scale from the ELA.

Given the MLE $\hat{\psi}$, all the results on the ML imputation in Section 2.2.2 hold.

Under missing at random (MAR) of Rubin (1976), the h-likelihood becomes

$$h = \log f_\theta(y_{\text{obs}} \mid \boldsymbol{x}) + \log f_\theta(y_{\text{mis}} \mid \boldsymbol{x}) + \log f_\rho(\delta \mid \boldsymbol{x}) + \log \left| \frac{\partial y_{\text{mis}}}{\partial v} \right|,$$

where θ is the parameter for the response model and ρ is the parameter associate with the missing mechanism. Under MAR assumption, the canonical function of v depends only on θ and $\boldsymbol{x}$ to give ML imputed values $\hat{y}_{\text{mis},i} = \tilde{y}_{\text{mis},i}(\hat{\theta}, \boldsymbol{x}_i)$, $\tilde{y}_{\text{mis},i}(\theta, \boldsymbol{x}_i) = g^{-1}\{\tilde{v}_i(\theta, \boldsymbol{x}_i)\}$.

Example 2.4.1. Little and Rubin (2019) considered censored exponential model, where $y_{\text{com}} = (y_{\text{obs}}, y_{\text{mis}})$ are independent exponential random variables with mean θ and the missing mechanism is set to $\delta = I(Y \leq c)$ with known c. Here the missing mechanism is not ignorable and the complete-data likelihood is

$$\ell_e\left(\theta, y_{\text{mis}}\right) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n_{\text{obs}}} y_i - \frac{1}{\theta} \sum_{i=n_{\text{obs}}+1}^{n} y_{\text{mis},i}.$$

They noted that joint maximization of the complete-data likelihood provides nonsensical modes $(n_{\text{obs}}\bar{y}_{\text{obs}} + n_{\text{mis}}c)/n$ for θ and c for $y_{\text{mis},i}$, where $\bar{y}_{\text{obs}} = \sum_{i=1}^{n_{\text{obs}}} y_i/n_{\text{obs}}$ is the sample mean based on the observed responses. Now we know that MLEs (modes) should be obtained from the h-likelihood. Yun et al. (2007) found the canonical scale $v_i = \log\left(y_{\text{mis},i} - c\right)$ to form the h-likelihood

$$\begin{aligned} h &= \ell_e\left(\theta, y_{\text{mis}}\right) + \log \left| \frac{\partial y_{\text{mis}}}{\partial v} \right| \\ &= -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n_{\text{obs}}} y_i + \sum_{i=n_{\text{obs}}+1}^{n} \left\{ -\frac{1}{\theta}\left(c + e^{v_i}\right) + v_i \right\}. \end{aligned}$$

The canonical function of v is $\tilde{v}_i(\theta) = \log\theta$ which gives

$$h\left\{\theta, \tilde{v}(\theta)\right\} = -n_{\text{obs}}\log\theta - \frac{1}{\theta}\sum_{i=1}^{n_{\text{obs}}} y_i - \frac{n_{\text{mis}}c}{\theta} - n_{\text{mis}} = \ell_m(\theta) - n_{\text{mis}} \propto \ell_m(\theta).$$

This gives the true MLE $\hat{\theta} = \bar{y}_{\text{obs}} + n_{\text{mis}}c/n_{\text{obs}}$ and the ML imputed values $\hat{y}_{\text{mis},i} = \hat{\theta} + c > c$ to lead that

$$\bar{y}_{\text{ML}} = \frac{1}{n}\left(\sum_{i=1}^{n_{\text{obs}}} y_i + \sum_{i=n_{\text{obs}}+1}^{n} \hat{y}_{\text{mis},i}\right) = \hat{\theta}$$

and $\widehat{\text{var}}(\bar{y}_{\text{ML}}) = \widehat{\text{var}}(\hat{\theta}) = \hat{I}^{\theta\theta} = \hat{\theta}^2/n_{\text{obs}}$. Little and Rubin (2019) used the EM algorithm. With the E-step

$$\text{E}_\theta(y_{\text{mis},i} | y_{\text{mis},i} > c) = \theta + c,$$

the M-step gives

$$\theta^{(t+1)} = \frac{1}{n}\left[\sum_{i=1}^{n_{\text{obs}}} y_i + n_{\text{mis}}\left\{\theta^{(t)} + c\right\}\right].$$

Thus, the EM algorithm gives the identical MLE $\hat{\theta}$. But, the EM algorithm does not provide the variance estimator directly.

To examine the performance of the ML imputation, we set about 22% of responses as unobserved and compare three estimators $\bar{y}_{\text{com}} = \sum_{i=1}^{n} y_i/n$, $\bar{y}_{\text{obs}} = \sum_{i=1}^{n} \delta_i y_i/n_{\text{obs}}$, and $\bar{y}_{\text{ML}}$ using random samples from exp(2) distribution. The estimator $\bar{y}_{\text{com}}$ is considered as a benchmark since it cannot be used in practice. In Figure 2.2, it is shown that the proposed method works well. Moreover, $\bar{y}_{\text{obs}}$ shows a non-negligible bias in amount $n_{\text{mis}}c/n_{\text{obs}} \approx 0.86$ since the missing mechanism is not ignorable.

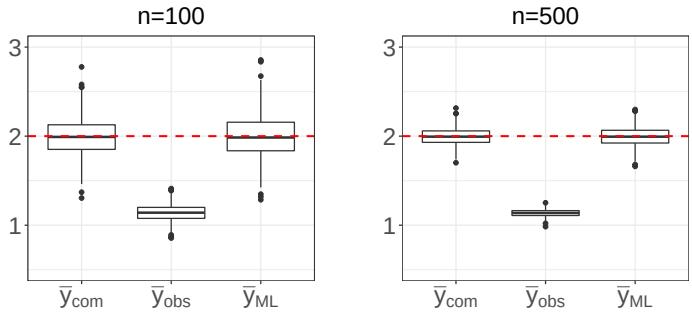


Figure 2.2: Boxplots of estimators in exponential mean model with $c = 3$. Dotted line indicates the true value of η.

2.5 Illustrative Examples

2.5.1 Normal Regression Model

Consider a normal regression model $Y|x \sim \mathrm{N}(\beta_0+\beta_1 x, \sigma^2)$ with response probability model $\mathrm{logit}\{\mathrm{P}_\rho(\delta = 1|x)\} = \rho_0+\rho_1 x+\rho_2 x^2$ under a MAR assumption. Here, y_{mis}-scale itself satisfies the Bartlett identities but it is canonical scale only for (β_0, β_1). Thus, the joint maximization of $\ell_e(\theta, y_{\mathrm{mis}})$ cannot give the MLE of σ^2, where $\theta = (\beta_0, \beta_1, \sigma^2)$. However, v-scale defined by $v_i = y_{\mathrm{mis},i}/\sigma$ is the canonical scale with canonical function $\tilde{v}_i(\theta, x_i) = (\beta_0 + \beta_1 x_i)/\sigma$ for $i = n_{\mathrm{obs}} + 1, \ldots, n$. Then, the canonical function of y_{mis} is $\tilde{y}_{\mathrm{mis},i}(\theta, x_i) = \beta_0 + \beta_1 x_i = \mathrm{E}_\theta(y_{\mathrm{mis},i}|x_i)$ and the ML imputed values are $\hat{y}_{\mathrm{mis},i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Moreover,

$$\hat{I}^{-1}_{y_{\mathrm{mis},i} y_{\mathrm{mis},i}} = \left(-\frac{\partial^2 h}{\partial y^2_{\mathrm{mis},i}} \right)^{-1} \Big|_{\theta=\hat{\theta}} = \hat{\sigma}^2 = \widehat{\mathrm{var}}\left(y_{\mathrm{mis},i} \mid \mathcal{D}, \boldsymbol{x}\right).$$

Since

$$\tilde{y}_{\mathrm{mis},i}(\theta, x_i) = \mathrm{E}_\theta(y_{\mathrm{mis},i}|x_i),$$

the MLEs can also be obtained by the EM algorithm.

For a simulation study, we generate $n = 100$ and $n = 500$ samples with $\theta = (1, 2, 1)$, $\rho = (1, 2, 0.3)$ and $x \sim \mathrm{U}(-1, 1)$. From Figure 2.3, we can see that $\bar{y}_{\mathrm{obs}}$ is positively biased because the covariate x increases both $\mathrm{E}_\theta(Y|x)$ and $\mathrm{P}_\rho(\delta = 1|x)$. Also, the performance of $\bar{y}_{\mathrm{ML}}$ is almost same as $\bar{y}_{\mathrm{com}}$.

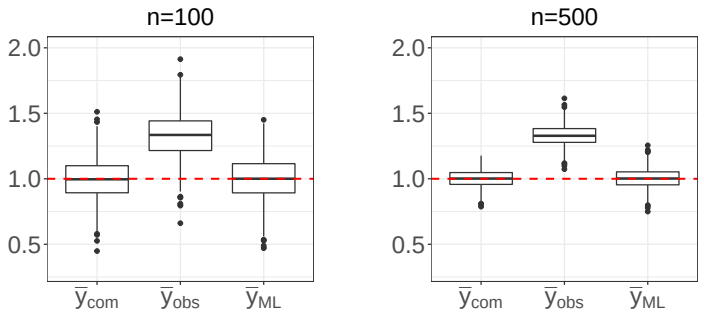


Figure 2.3: Boxplots of estimators in normal regression model. Dotted line indicates the true value of η.

2.5.2 Exponential Regression Model

Consider an exponential regression model with mean $\mathrm{E}_\beta(Y|x) = \exp(\beta_0 + \beta_1 x)$, $\beta = (\beta_0, \beta_1)$ and the MAR mechanism as the Example 2.5.1. In this example, $v = \log y_{\mathrm{mis}}$ scale is the canonical scale which also satisfies Bartlett identities by Corollary 2.3.1. Here the canonical function of $y_{\mathrm{mis},i}$ is $\tilde{y}_{\mathrm{mis},i} = \exp(\beta_0 + \beta_1 x_i) = \mathrm{E}_\beta(y_{\mathrm{mis},i}|x_i)$ and the ML imputed values are $\hat{y}_{\mathrm{mis},i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$. Moreover,

$$\hat{I}^{-1}_{y_{\mathrm{mis},i} y_{\mathrm{mis},i}} = \left(-\frac{\partial^2 h}{\partial y^2_{\mathrm{mis},i}} \right)^{-1} \Big|_{\theta=\hat{\theta}} = \hat{y}^2_{\mathrm{mis},i} = \widehat{\mathrm{var}}\left(y_{\mathrm{mis},i} \mid \mathcal{D}, \boldsymbol{x} \right).$$

Figure 2.4 shows simulation results with β and ρ being the same as in Example 2.5.1. Compared to $\bar{y}_{\text{com}}$, $\bar{y}_{\text{ML}}$ gives almost the same performances, whereas $\bar{y}_{\text{obs}}$ is biased.

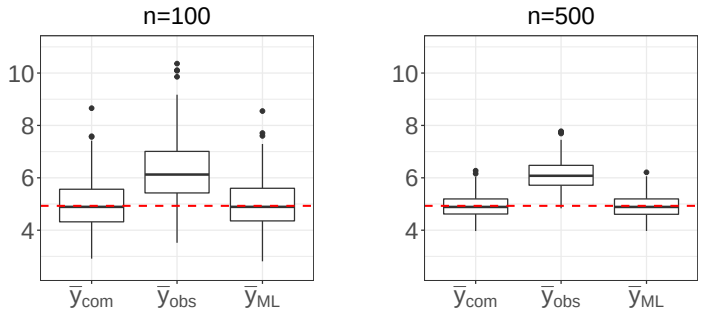


Figure 2.4: Boxplots of estimators in exponential regression model. Dotted line indicates the true value of η.

2.5.3 Tobit Regression Model

Suppose that responses are generated from the normal regression model in Example 2.5.1. In addition, missing data are created by $y_{\text{mis}} > c$ at a known censoring point c. The extended likelihood

$$\begin{aligned}\ell_e(\theta, y_{\text{mis}}) &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_{\text{obs}}}\left(y_i - \tilde{x}_i^{\mathrm{T}}\beta\right)^2 \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=n_{\text{obs}}+1}^{n}\left(y_{\text{mis},i} - \tilde{x}_i^{\mathrm{T}}\beta\right)^2,\end{aligned}$$

where $\theta = (\beta, \sigma^2)$, $\beta = (\beta_0, \beta_1)$ and $\tilde{x} = (1, x)$. Here a b-scale

$$b_i = g_1(y_{\text{mis},i}) = \log(y_{\text{mis},i} - c),$$

satisfies Bartlett identities by Corollary 2.3.1 but it is not canonical. Now, consider a w-scale with $w_i = g_2(b_i) = \tilde{\Omega}_{b_i b_i}^{0.5} b_i$ by (2.16).

Then, we have the approximate MLE $\hat{\theta}^{\text{Lap}}$ and approximate ML imputed values $\hat{y}_{\text{mis}}^{\text{Lap}}$ by jointly maximizing $\ell_e(\theta, w)$. However, the exact marginal log-likelihood is available in Tobit regression model.

$$\begin{aligned}\ell_m(\theta) &= -\frac{n_{\text{obs}}}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_{\text{obs}}}\left(y_i - \tilde{x}_i^{\text{T}}\beta\right)^2 \\ &\quad + \sum_{i=n_{\text{obs}}+1}^{n}\log\left\{\Phi\left(\frac{\tilde{x}_i^{\text{T}}\beta - c}{\sigma}\right)\right\}.\end{aligned}$$

This means that explicit form of the predictive likelihood $L_e(b_i|y_{\text{obs}};\theta)$ is available to give the canonical scale

$$v_i = L_e\left(\tilde{b}_i \mid y_{\text{obs}};\theta\right) b_i, \tag{2.18}$$

where

$$\tilde{b}_i = \log\left\{\tilde{x}_i^{\text{T}}\beta - c + \sqrt{\left(\tilde{x}_i^{\text{T}}\beta - c\right)^2 + 4\sigma^2}\right\} - \log 2.$$

Thus, all MLEs are computed directly by simple maximization of the h-likelihood.

In the simulation study, we examine the performance of ML imputations by using two estimators $\bar{y}_{\text{ML}}$ using the MLE and $\bar{y}_{\text{MI}}^{\text{Lap}}$ using the approximate MLE. From (2.18), we see that both b and w are linear transformations of v. Thus, approximate ML imputation works well as approximate MLE does. Given MLE for fixed parameters, weak canonical scale gives an exact ML imputation.

For simulation, we set $\theta = (1, 3, 1)$, $c = 3$ and $x_i = -1+2i/n$ for $i = 1, \ldots, n$. In Figure 2.5, we see that the difference between $\bar{y}_{\text{ML}}$ and $\bar{y}_{\text{ML}}^{\text{Lap}}$ is negligible because $\hat{\theta}$ and $\hat{\theta}^{\text{Lap}}$ are very close. Therefore, we can use the weak canonical scale and approximate MLE when canonical scale is unknown.

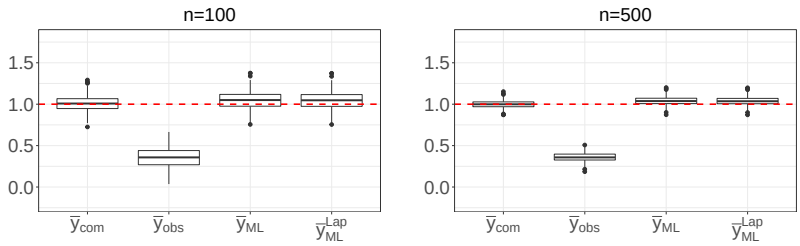


Figure 2.5: Boxplots of estimators in Tobit regression model. Dotted line indicates the true value of η.

2.6 Conclusion

Firth (2006) and Meng (2009) raised two important reservations about the use of the h-likelihood. Firth (2006) noted that the linear predictor in HGLM may not be well-defined to form the h-likelihood. Lee et al. (2006) resolved his question by defining the canonical scale. Meng (2009) claimed the asymptotic theory for the prediction of the future data would be impossible because the consistency cannot be achieved for the predicted values from the h-likelihood. In this chapter, we have answered their queries on the h-likelihood in the context of imputation for missing data. Specifically, we have shown that prediction becomes an estimation of canonical function of the h-likelihood whose consistent estimation and asymptotic normality can be justifiable. We further showed that standard errors of prediction can be directly obtained from the h-likelihood.

Little and Rubin (2019) pointed out that the current h-likelihood procedure achieves the correct ML estimation by modifying h-likelihood. In this chapter, we achieve the true ML approach via h-

likelihood without any modification by reformulating the h-likelihood. We present the meaning of the canonical scale and canonical function in detail, which allow ML estimation of fixed parameters and ML imputation of random parameters, namely missing data. The Jacobian term is a key to finding the canonical scale.

The ML imputation using the h-likelihood estimates the conditional mode, rather than the conditional mean of the missing value. We call this conditional mode imputation the ML imputation for the random parameters. The h-likelihood used for ML imputation provides an efficient algorithm because resampling procedure for multiple imputations or expectation steps in EM algorithm is not compulsory.

Appendix: Supplementary Materials for "Maximum Likelihood Imputation"

A1 Regularity Conditions

In this chapter, we assume the following regularity conditions in developing the proposed method.

(R1) Let $\psi_0 = \arg\max_\psi \mathrm{E}_\psi\{\ell_m(\psi)\}$ be the true value of ψ. Here, the number of fixed parameters does not depend on n_{obs}. Then, the MLE $\hat{\psi} = \arg\max_\psi \ell_m(\psi)$ satisfies the asymptotic normality with mean ψ_0 and variance $\mathcal{I}_0^{-1} = \mathcal{I}^{-1}(\psi_0)$, where

$$\mathcal{I}(\psi) = \lim_{n_{\mathrm{obs}}\to\infty} \frac{1}{n_{\mathrm{obs}}} \left(-\frac{\partial^2 \ell_m(\psi)}{\partial\psi\partial\psi^{\mathrm{T}}} \right) \Big|_{\psi=\psi_0}$$

is the expected Fisher information.

(R2) The support of missing values

$$\Omega_{y_{\text{mis}}} = \left\{ y_{\text{mis}} \in \mathbb{R}^{n_{\text{mis}}} : \prod_{i=n_{\text{obs}}+1}^{n} f_\psi \left(y_{\text{mis},i}, \delta_i = 0 \mid \boldsymbol{x}_i \right) > 0 \right\} \subset \mathbb{R}^{n_{\text{mis}}}$$

does not depend on fixed parameter ψ.

A2 Proofs

A2.1 Proof of Theorem 3.1

Proof. By assumption, there exists $\tilde{y}_{\text{mis}} = \arg\max_{y_{\text{mis}}} \ell_e(\psi, y_{\text{mis}})$. Now, consider a v-scale defined by

$$v_i = g(y_{\text{mis},i}) = \left\{ L_p \left(\tilde{y}_{\text{mis}} \mid \mathcal{D}; \psi \right) \right\}^{1/n_{\text{mis}}} y_{\text{mis},i}, \ i = n_{\text{obs}} + 1, \ldots, n,$$

with the predictive likelihood $L_p(y_{\text{mis}} \mid \mathcal{D}; \psi) = f_\psi(y_{\text{mis}} \mid \mathcal{D}, \boldsymbol{x})$. Here, the transformation $g(\cdot)$ is bijective and differentiable since it is linear. The predictive likelihood on v-scale is also well-defined with the Jacobian term

$$L_p(v \mid \mathcal{D}; \psi) = L_p(y_{\text{mis}} \mid \mathcal{D}; \psi) \left| \frac{\partial y_{\text{mis}}}{\partial v} \right|, \ \left| \frac{\partial v}{\partial y_{\text{mis}}} \right|_{v=\tilde{v}} = L_p(\tilde{y}_{\text{mis}} \mid \mathcal{D}; \psi),$$

where $\tilde{v}_i = g(\tilde{y}_{\text{mis},i})$. Note that $\tilde{v}$ is also the mode of $L_p(v|\mathcal{D}; \psi)$ since $\tilde{y}_{\text{mis}}$ is the mode of $L_p(y_{\text{mis}}|\mathcal{D}; \psi)$ and the transformation $g(\cdot)$ is linear. Therefore, there exists a canonical scale which satisfies (12). □

A2.2 Proof of Theorem 3.2

Proof. Let v-scale be the canonical scale and $\tilde{v} = \tilde{v}(\psi, \mathcal{D}, \boldsymbol{x})$. Then, the h log-likelihood can be written as

$$h \left(\psi, \tilde{v} \right) = \ell_m \left(\psi \right) + c,$$

where c is a constant which is free of ψ. Then, we can prove the first equality

$$\frac{\partial}{\partial\psi}h\left(\psi,\tilde{v}\right)=\frac{\partial h}{\partial\psi}\Big|_{v=\tilde{v}}+\frac{\partial\tilde{v}^{\mathrm{T}}}{\partial\psi}\frac{\partial h}{\partial v}\Big|_{v=\tilde{v}}=\frac{\partial h}{\partial\psi}\Big|_{v=\tilde{v}}=\frac{\partial\ell_m}{\partial\psi},$$

where $h=h(\psi,v)$ and $\ell_m=\ell_m(\psi)$. To show the second equality, recall that

$$\frac{\partial h}{\partial v}\Big|_{v=\tilde{v}}=0. \tag{2.19}$$

By differenciating (2.19) with respect to ψ,

$$\frac{\partial^2 h}{\partial\psi\partial v^{\mathrm{T}}}\Big|_{v=\tilde{v}}+\frac{\partial\tilde{v}^{\mathrm{T}}}{\partial\psi}\left\{\frac{\partial^2 h}{\partial v\partial v^{\mathrm{T}}}\right\}_{v=\tilde{v}}=O.\ \Rightarrow\ \frac{\partial\tilde{v}^{\mathrm{T}}}{\partial\psi}=-I_{\psi v}I_{vv}^{-1}|_{v=\tilde{v}}. \tag{2.20}$$

Therefore, from (2.20), we can prove the required result.

$$\begin{aligned}
\frac{\partial\ell_m}{\partial\psi} &= \frac{\partial}{\partial\psi}h\left(\psi,\tilde{v}\right).\\
\Rightarrow -\frac{\partial^2\ell_m}{\partial\psi\partial\psi^{\mathrm{T}}} &= -\frac{\partial^2 h}{\partial\psi\partial\psi^{\mathrm{T}}}\Big|_{v=\tilde{v}}-\frac{\partial\tilde{v}^{\mathrm{T}}}{\partial\psi}\frac{\partial^2 h}{\partial v\partial\psi^{\mathrm{T}}}\Big|_{v=\tilde{v}}\\
&= I_{\psi\psi}-I_{\psi v}I_{vv}^{-1}I_{v\psi}\\
&= \left(I^{\psi\psi}\right)^{-1}.
\end{aligned}$$

Here,

$$\begin{aligned}
\begin{pmatrix} I^{\psi\psi} & I^{\psi v}\\ I^{v\psi} & I^{vv}\end{pmatrix} &= \begin{pmatrix} I_{\psi\psi} & I_{\psi v}\\ I_{v\psi} & I_{vv}\end{pmatrix}^{-1},\\
\begin{pmatrix} I_{\psi\psi} & I_{\psi v}\\ I_{v\psi} & I_{vv}\end{pmatrix} &= \begin{pmatrix} -\partial^2 h/\partial\psi\partial\psi^{\mathrm{T}} & -\partial^2 h/\partial\psi\partial v^{\mathrm{T}}\\ -\partial^2 h/\partial v\partial\psi^{\mathrm{T}} & -\partial^2 h/\partial v\partial v^{\mathrm{T}}\end{pmatrix}_{v=\tilde{v}}.
\end{aligned}$$

□

A2.3 Proof of Corollary 4.1

Proof. It suffices to show that the case $n_{\text{mis}} = 1$. If $\mathrm{E}_\psi(b) < \infty$, then

$$\lim_{|b|\to\infty} b f_\psi(b, \delta = 0 \mid \boldsymbol{x}) = 0 \;\Rightarrow\; \lim_{|b|\to\infty} f_\psi(b, \delta = 0 \mid \boldsymbol{x}) = 0.$$

Since f_ψ is continuous, $f_\psi(b, \delta = 0 \mid \boldsymbol{x}) = 0$ for $b \in \partial\Omega_b = \{-\infty, \infty\}$. Moreover, f_ψ is bounded since f_ψ is a density function of a continuous random variable whose support is the whole real line with finite mean. This guarantees that f_ψ is uniformly continuous which implies

$$\begin{aligned}
& \lim_{|b|\to\infty} f'_\psi(b, \delta = 0 \mid \boldsymbol{x}) \\
= \; & \lim_{|b|\to\infty} \lim_{t\to 0} \frac{f_\psi(b + t, \delta = 0 \mid \boldsymbol{x}) - f_\psi(b, \delta = 0 \mid \boldsymbol{x})}{t} \\
= \; & \lim_{t\to 0} \lim_{|b|\to\infty} \frac{f_\psi(b + t, \delta = 0 \mid \boldsymbol{x}) - f_\psi(b, \delta = 0 \mid \boldsymbol{x})}{t} \\
= \; & 0,
\end{aligned}$$

i.e., $f'_\psi(b, \delta = 0 \mid \boldsymbol{x}) = 0$ for $b \in \partial\Omega_b = \{-\infty, \infty\}$. Then, the first and second Bartlett identities hold by the result of Theorem 4.1. □

A2.4 Score and Hessian of $\hat{\ell}_m(\psi)$ and $\ell_e(\psi, \tilde{w})$

Proof. By the definition of $\hat{\ell}_m(\psi)$, the score and Hessian can be expressed as

$$\begin{aligned}
\frac{\partial}{\partial \psi_j}\hat{\ell}_m(\psi) &= \frac{\partial}{\partial \psi_j}\ell_e(\psi, b)\Big|_{b=\tilde{b}} - \frac{1}{2}\,\mathrm{tr}\left\{\left(I^b_{bb}\right)^{-1}\left(\frac{\partial}{\partial \psi_j}I^b_{bb}\right)\right\}, \\
-\frac{\partial^2}{\partial \psi_j \partial \psi_k}\hat{\ell}_m(\psi) &= I^b_{\psi_j\psi_k} - I^b_{\psi_j b}\left(I^b_{bb}\right)^{-1} I^b_{b\psi_k} \\
&\quad + \frac{1}{2}\,\mathrm{tr}\left\{\left(I^b_{bb}\right)^{-1}\left(\frac{\partial^2}{\partial \psi_j \partial \psi_k}I^b_{bb}\right)\right. \\
&\quad \left. - \left(I^b_{bb}\right)^{-1}\left(\frac{\partial}{\partial \psi_j}I^b_{bb}\right)\left(I^b_{bb}\right)^{-1}\left(\frac{\partial}{\partial \psi_k}I^b_{bb}\right)\right\},
\end{aligned}$$

for $1 \le j,\ k \le p$, where $I^b_{xy} = -\frac{\partial^2}{\partial x \partial y^{\mathrm{T}}}\ell_e(\psi, b)\big|_{b=\tilde{b}}$. On the other hand, with $\ell_e = \ell_e(\psi, w)$,

$$\begin{aligned}
\frac{\partial}{\partial \psi}\ell_e(\psi, \tilde{w}) &= \frac{\partial \ell_e}{\partial \psi}\Big|_{w=\tilde{w}}, \\
\left\{-\frac{\partial^2}{\partial \psi \partial \psi^{\mathrm{T}}}\ell_e(\psi, \tilde{w})\right\}^{-1} &= I^{\psi\psi}_e,
\end{aligned}$$

where

$$\begin{pmatrix} I^{\psi\psi}_e & I^{\psi w}_e \\ I^{w\psi}_e & I^{ww}_e \end{pmatrix} = \begin{pmatrix} I_{e,\psi\psi} & I_{e,\psi w} \\ I_{e,w\psi} & I_{e,ww} \end{pmatrix}^{-1}$$

with

$$\begin{pmatrix} I_{e,\psi\psi} & I_{e,\psi w} \\ I_{e,w\psi} & I_{e,ww} \end{pmatrix} = \begin{pmatrix} -\frac{\partial^2 \ell_e}{\partial \psi \partial \psi^{\mathrm{T}}} & -\frac{\partial^2 \ell_e}{\partial \psi \partial w^{\mathrm{T}}} \\ -\frac{\partial^2 \ell_e}{\partial w \partial \psi^{\mathrm{T}}} & -\frac{\partial^2 \ell_e}{\partial w \partial w^{\mathrm{T}}} \end{pmatrix}_{w=\tilde{w}}.$$

□

Chapter 3

Robust Imputation under Missing at Random

Chapter Summary

Imputation is a popular technique for handling item nonresponse. By properly incorporating the observed auxiliary variables, imputation can reduce the nonresponse bias and obtain efficient estimation. Among various imputation methods, an advantage of the maximum likelihood (ML) imputation is that one-shot imputation allows the maximum likelihood estimator (MLE) of fixed parameter. However, correct specification of statistical model may be difficult in the presence of missing data. How to find a robust imputation method that is less sensitive to the failure of the assumed model is an important practical problem in the missing data literature. If the missing mechanism is missing-at-random, doubly robust estimator gives useful estimator since the consistency of the estimator is guaranteed either the outcome regression (OR) model or the propensity score (PS) model is correctly specified. To obtain the doubly robust estimator, the internal bias calibration (IBC) condition is presented. Moreover, we examine the IBC condition in modeling approach. Correct specification of the outcome model or propensity score model is equivalent to that of mean or dispersion in double hierarchical generalized linear model. In addition, we discuss how to allow robust inference against outliers. Simulation study shows that the proposed method allows robust inference against not only the violation of various model assumptions, but also outliers.

3.1 Basic Setup

Suppose that we are interested in estimating the parameter of interest η^* defined though

$$\mathrm{E}\left\{U\left(\eta^*; Y\right)\right\} = 0,$$

where $U(\eta; y)$ is the given estimating function. Suppose further that there are n independently and identically distributed realizations of $(\boldsymbol{X}, Y, \delta)$, denoted by $\{(\boldsymbol{x}_i, y_i, \delta_i) : i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ is an vector of observed covariates and δ_i is the missingness indicator defined by

$$\delta_i = \begin{cases} 1, \text{ if } y_i \text{ is observed,} \\ 0, \text{ otherwise.} \end{cases}$$

Without nonresponse, a consistent estimator of η^* is obtained by solving

$$\hat{U}_n(\eta) \equiv \frac{1}{n} \sum_{i=1}^{n} U(\eta; y_i) = 0.$$

Under nonresponse, one way to estimate η^* is to use the expected estimating equation

$$\frac{1}{n} \sum_{i=1}^{n} \left[\delta_i U(\eta; y_i) + (1 - \delta_i)\, \mathrm{E}\left\{U(\eta; Y_i) \mid \boldsymbol{x}_i, \delta_i = 0\right\}\right] = 0. \quad (3.1)$$

To compute the conditional expectation in (3.1), we often employ the MAR assumption of Rubin (1976). That is,

$$f(y \mid \boldsymbol{x}, \delta = 0) = f(y \mid \boldsymbol{x}) = f(y \mid \boldsymbol{x}, \delta = 1).$$

Under the MAR assumption, we can have the conditional expectation in (3.1) as

$$\mathrm{E}\left\{U(\eta; Y) \mid \boldsymbol{x}, \delta = 0\right\} = \mathrm{E}\left\{U(\eta; Y) \mid \boldsymbol{x}\right\} = \mathrm{E}\left\{U(\eta; Y) \mid \boldsymbol{x}, \delta = 1\right\}.$$

Thus, it suffices to estimate $\bar{U}(\eta; \boldsymbol{x}) = \mathrm{E}\{U(\eta; Y)|\boldsymbol{x}, \delta = 1\}$ from the set of respondents under the MAR assumption. The conditional expectation in (3.1) is based on the model for $[y|\boldsymbol{x}]$, which is often called the outcome regression (OR) model. On the other hand, another approach uses a model for $[\delta|\boldsymbol{x}]$, which is often called the propensity score (PS) model.

To compute the conditional expectation $\bar{U}(\eta; \boldsymbol{x})$, we employ the OR model $f(y|\boldsymbol{x}; \boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$. Under the MAR assumption, we can estimate $\boldsymbol{\theta}$ by maximizing

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})$$

with respect to $\boldsymbol{\theta}$ and then η^* can be estimated by the imputed estimating equation

$$\hat{U}_I(\eta) \equiv \frac{1}{n} \sum_{i=1}^{n} \left[\delta_i U(\eta; y_i) + (1 - \delta_i) \, \mathrm{E}\left\{ U(\eta; Y_i) \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}} \right\} \right] = 0. \tag{3.2}$$

To compute the conditional expectation in (3.2), Kim (2011) proposed the fractional imputation method. Consistency of the solution $\hat{\eta}_I$ to (3.2) is based on the assumption that the regression outcome model $f(y|\boldsymbol{x}; \boldsymbol{\theta})$ is correctly specified.

To protect against model misspecification, one can utilize a propensity score model for $\mathrm{P}(\delta = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}; \boldsymbol{\phi})$ and apply

$$\hat{U}_{\mathrm{DR}}(\eta) \equiv \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\delta_i}{\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})} U(\eta; y_i) + \left(1 - \frac{\delta_i}{\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})} \right) \mathrm{E}\left\{ U(\eta; Y_i) \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}} \right\} \right] = 0 \tag{3.3}$$

as an estimating equation for η, where $\hat{\boldsymbol{\phi}}$ is the maximizer of

$$\ell(\boldsymbol{\phi}) = \sum_{i=1}^{n} \left[\delta_i \log \pi(\boldsymbol{x}_i; \boldsymbol{\phi}) + (1 - \delta_i) \log\{1 - \pi(\boldsymbol{x}_i; \boldsymbol{\phi})\} \right].$$

Now, let $\hat{\pi}_i = \pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})$. Since

$$\hat{U}_{\mathrm{DR}}(\eta) - \hat{U}_n(\eta) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\delta_i}{\hat{\pi}_i}\right)\left[U(\eta; y_i) - \mathrm{E}\left\{U(\eta; Y_i) \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}\right\}\right], \tag{3.4}$$

the right side of (3.4) is approximately unbiased to zero if either the OR model $f(y|\boldsymbol{x}; \boldsymbol{\theta})$ or the PS model $\mathrm{P}(\delta = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}; \boldsymbol{\phi})$ is correctly specified. Thus, the estimating equation $\hat{U}_{\mathrm{DR}}(\eta)$ in (3.3) gives a doubly robust (DR) estimator. From (3.4), we can achieve $\hat{U}_{\mathrm{DR}}(\eta) = \hat{U}_I(\eta)$ if

$$\sum_{i=1}^{n}\delta_i\left(\frac{1}{\hat{\pi}_i} - 1\right)\left[U(\eta; y_i) - \mathrm{E}\left\{U(\eta; Y_i) \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}\right\}\right] = 0. \tag{3.5}$$

We can view (3.5) as a key condition to get a doubly robust imputation in the sense of Kim and Haziza (2014). Condition (3.5) is called the internal bias calibration (IBC), which was originally termed by Firth and Bennett (1998) in the context of design-consistent estimation of the model parameters under complex sampling. The imputation estimating equation in (3.2) satisfying the IBC condition (3.5) is called internally bias calibrated. The IBC condition is a sufficient condition for double robustness. How to find the imputed estimator satisftying the IBC condition under a more general class of OR models and PS models is our main research problem. We will address this issue in the next section.

3.2 Semiparametric Outcome Regression Model

The model assumption based on the estimating equation such as $\mathrm{E}\{U(\eta; Y)\} = 0$ is regarded as a semiparametric model. Thus, instead of making parametric model assumption for $[y \mid \boldsymbol{x}]$, it makes

sense to relax the parametric model assumptions for developing DR imputation. For the OR model, we now assume that there exists $b_1(\boldsymbol{x}), \ldots, b_L(\boldsymbol{x})$ such that

$$\mathrm{E}\{U(\eta; Y) \mid \boldsymbol{x}\} \in \mathrm{span}\{b_0(\boldsymbol{x}), b_1(\boldsymbol{x}), \cdots, b_L(\boldsymbol{x})\} := \mathcal{H} \qquad (3.6)$$

for all η, where $b_0(\boldsymbol{x}) \equiv 1$. Assumption (3.6) can be called the semiparametric OR model. If $U(\eta; y) = \eta - y$, model (3.6) reduces to the usual regression model

$$\mathrm{E}\{U(\eta; Y) \mid \boldsymbol{x}\} = \sum_{k=0}^{L} \beta_k b_k(\boldsymbol{x}).$$

However, finding an imputation estimating equation using (3.6) is tricky as the vector of regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_L)^{\mathrm{T}}$ since $\boldsymbol{\beta} = \boldsymbol{\beta}(\eta)$ is a function of η. Thus, even if we can obtain $\hat{\boldsymbol{\beta}}(\eta)$ from the normal equation, finding the solution to the imputed estimating equation

$$\hat{U}_I(\eta) \equiv \frac{1}{n} \sum_{i=1}^{n} \left[\delta_i U(\eta; y_i) + (1 - \delta_i) \hat{\mathrm{E}}\{U(\eta; Y_i) \mid \boldsymbol{x}_i\} \right] = 0 \quad (3.7)$$

is not feasible in general, where

$$\hat{\mathrm{E}}\{U(\eta; Y) \mid \boldsymbol{x}\} = \sum_{k=0}^{L} \hat{\beta}_k(\eta) b_k(\boldsymbol{x})$$

and $\hat{\beta}_k(\eta)$ satisfies

$$\sum_{i=1}^{n} \delta_i \left\{ U(\eta; y_i) - \sum_{k=0}^{L} \hat{\beta}_k(\eta) b_k(\boldsymbol{x}_i) \right\} h(\boldsymbol{x}_i) = 0$$

for any $h(\boldsymbol{x}) \in \mathcal{H}$ and for all η.

To avoid the difficulty of finding $\hat{\boldsymbol{\beta}}(\eta)$ and solving the imputed estimating equation in (3.7), Wang and Kim (2021) proposed the

use of the information projection technique for self-efficient PS estimation. The basic idea is to find the PS weights $\omega(\boldsymbol{x};\boldsymbol{\phi}) = 1/\pi(\boldsymbol{x};\boldsymbol{\phi})$ which satisfies the self-efficiency property

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i\omega(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})U(\eta;y_i) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i U(\eta;y_i) + (1-\delta_i)\hat{\mathrm{E}}\{U(\eta;Y_i)\mid \boldsymbol{x}_i\}\right] \tag{3.8}$$

holds for all η, where the parameters $\hat{\boldsymbol{\phi}}$ are estimated by the calibration equation

$$\sum_{i=1}^{n}\delta_i\omega(\boldsymbol{x}_i;\boldsymbol{\phi})\boldsymbol{b}_i = \sum_{i=1}^{n}\boldsymbol{b}_i, \tag{3.9}$$

where $\boldsymbol{b}_i = (b_0(\boldsymbol{x}_i), b_1(\boldsymbol{x}_i), \ldots, b_L(\boldsymbol{x}_i))^{\mathrm{T}}$. Wang and Kim (2021) proved that the PS weights in (3.13) satisfying the calibration condition in (3.9) satisfies the self-efficiency property in (3.8). Once $\omega(\boldsymbol{x};\hat{\boldsymbol{\phi}})$ satisfying (3.8) is obtained, we can use

$$\sum_{i=1}^{n}\delta_i\hat{\omega}_i U(\eta;y_i) = 0 \tag{3.10}$$

to obtain the solution to the imputed estimating equation in (3.7), where $\hat{\omega}_i = \omega(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})$. If $U(\eta;y) = \eta - y$, the estimating equation (3.10) gives an estimator

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\delta_i\hat{\omega}_i y_i &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta} + \frac{1}{n}\sum_{i=1}^{n}\delta_i\hat{\omega}_i\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right\} + \frac{1}{n}\sum_{i=1}^{n}\delta_i(\hat{\omega}_i - 1)\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\end{aligned}$$

for any $\boldsymbol{\beta}$. Thus, the imputed estimating equation satisfying the self-efficiency property (3.8) can be derived as

$$\hat{\mathrm{E}}\{U(\eta;Y_i)\mid \boldsymbol{x}_i\} = \eta - \boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{IBC}},$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{IBC}}$ is the solution which satisfies the IBC condition

$$\sum_{i=1}^{n} \delta_i(\hat{\omega}_i - 1)\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right) = 0. \tag{3.11}$$

Then, we can have a DR imputed estimator

$$\hat{\eta}_{\mathrm{IBC}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{IBC}}\right\}. \tag{3.12}$$

By using the information projection approach, Wang and Kim (2021) presented the PS weight model

$$\omega(\boldsymbol{x};\boldsymbol{\phi}) = 1 + \frac{n_{\mathrm{mis}}}{n_{\mathrm{obs}}}\exp\left\{\boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}\right\} \tag{3.13}$$

where $n_{\mathrm{obs}} = \sum_{i=1}^{n}\delta_i$ is the number of observed outcomes, $n_{\mathrm{mis}} = n - n_{\mathrm{obs}}$ is the number of nonresponses, and $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_L)^{\mathrm{T}}$ is the vector of parameters in the PS weight model. Note that the PS weight model (3.13) can be equivalently represented as the logistic (log-odds) PS model

$$\log\left\{\frac{\pi(\boldsymbol{x};\boldsymbol{\phi})}{1-\pi(\boldsymbol{x};\boldsymbol{\phi})}\right\} = \log\left(\frac{n_{\mathrm{obs}}}{n_{\mathrm{mis}}}\right) - \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}. \tag{3.14}$$

In the next sections, we address how to obtain the robust imputed estimator against model misspecification of the PS model and outliers in the OR model.

3.3 Misspecification of Propensity Score Model

Wang and Kim (2021) proposed the PS weight model (3.13) based on the information projection approach. Indeed, the authors' approach is based on the Kullback-Leibler (KL) divergence. In this section, we examine how to enlarge the class of PS weight models by using the γ-power divergence.

Eguchi (2021) presented the γ-power divergence as a generalization of KL-divergence to enlarge the class of statistical models by introducing an additional scale parameter γ.

Definition 3.3.1 (Eguchi, 2021)**.** Let q and s be two probability density functions. Given $\gamma > 0$, the divergence

$$\begin{aligned} D_\gamma(q\|s) &= \frac{1}{\gamma(\gamma+1)} \log \int q(\boldsymbol{x})^{\gamma+1} d\boldsymbol{x} \\ &\quad -\frac{1}{\gamma} \log \int q(\boldsymbol{x}) s(\boldsymbol{x})^{\gamma} d\boldsymbol{x} \\ &\quad +\frac{1}{\gamma+1} \log \int s(\boldsymbol{x})^{\gamma+1} d\boldsymbol{x}. \end{aligned} \tag{3.15}$$

is called the γ-power divergence.

Similar to the KL-divergence, $D_\gamma(q\|s) \geq 0$ for all q, s and equality holds if and only if $q = s$. At $\gamma = 0$, γ-power divergence is defined as the KL-divergence

$$D_0(q\|s) = \lim_{\gamma \to 0} D_\gamma(q\|s) = D_{\mathrm{KL}}(q\|s).$$

Following theorem gives the PS weight model with respect to the γ-power divergence.

Lemma 3.3.1. Based on the γ-power divergence, the information projection approach gives the PS weight model

$$\omega(\boldsymbol{x}; \boldsymbol{\phi}, \gamma) = 1 + \frac{n_{\mathrm{mis}}}{n_{\mathrm{obs}}} \left(1 + \gamma \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\phi}\right)^{1/\gamma}. \tag{3.16}$$

Similar to the case of the KL-divergence, the parameter $\boldsymbol{\phi}$ in (3.16) is estimated by solving the calibration equation (3.9).

Recall that the PS weight model (3.13) induces the logistic PS model. On the other hand, the PS weight model (3.16) induces the

power-odds model

$$\left\{\frac{\pi(\boldsymbol{x};\boldsymbol{\phi})}{1-\pi(\boldsymbol{x};\boldsymbol{\phi})}\right\}^{\gamma}=\left(\frac{n_{\rm obs}}{n_{\rm mis}}\right)^{\gamma}\frac{1}{\{1+\gamma\boldsymbol{b}(\boldsymbol{x})^{\rm T}\boldsymbol{\phi}\}}. \tag{3.17}$$

Guerrero and Johnson (1982) proposed the power-odds model (3.17) to generalize the logistic regression model.

Remark 3.3.1. Eguchi (2021) independently derived the power-odds model (3.17) based on the γ-power divergence for robust inference against mislabeled binary outcome. Suppose that we observe the mislabeled data $\delta^{(m)}=1-\delta$ instead of δ. Let

$$\begin{aligned}\tau_0 &= \mathrm{P}(\delta^{(m)}=1\mid\delta=0,\boldsymbol{x}),\\ \tau_1 &= \mathrm{P}(\delta^{(m)}=0\mid\delta=1,\boldsymbol{x})\end{aligned}$$

be mislabel probabilities (Hung et al., 2018). Then,

$$\mathrm{P}(\delta^{(m)}=1\mid\boldsymbol{x})=\tau_0\,\mathrm{P}(\delta=0\mid\boldsymbol{x})+(1-\tau_1)\,\mathrm{P}(\delta=1\mid\boldsymbol{x}). \tag{3.18}$$

Thus, the robust inference against the mislabeled binary outcome is equivalent to the robust inference against the model misspecification when the true model is (3.18).

To examine the robustness of the power-odds model (3.17), let

$$\begin{aligned}\pi_{\rm log}(\boldsymbol{x}) &= \frac{\exp\{\boldsymbol{b}(\boldsymbol{x})^{\rm T}\boldsymbol{\phi}\}}{1+\exp\{\boldsymbol{b}(\boldsymbol{x})^{\rm T}\boldsymbol{\phi}\}},\\ \pi_{\rm pow}(\boldsymbol{x}) &= \frac{\{1+\gamma\boldsymbol{b}(\boldsymbol{x})^{\rm T}\boldsymbol{\phi}\}^{1/\gamma}}{1+\{1+\gamma\boldsymbol{b}(\boldsymbol{x})^{\rm T}\boldsymbol{\phi}\}^{1/\gamma}}.\end{aligned}$$

Then, for $\delta^{(m)}=1$ and all $\gamma>0$,

$$\left|\delta^{(m)}-\pi_{\rm log}(\boldsymbol{x})\right|\geq\left|\delta^{(m)}-\pi_{\rm pow}(\boldsymbol{x})\right|.$$

This shows how the power-odds model (3.17) allows the robust inference against mislabeled binary outcome.

Wang and Kim (2021) showed the asymptotic normality of the estimator obtained by the PS weighted estimating equation (3.10). In the following corollary, we generalize their result to the imputed estimator based on the γ-power divergence.

Corollary 3.3.1. Let

$$\hat{\eta}_\gamma = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_\gamma\right\}, \tag{3.19}$$

where $\hat{\boldsymbol{\beta}}_\gamma$ is the solution to

$$\sum_{i=1}^{n}\delta_i\boldsymbol{b}_i\left\{\hat{\omega}_i(\gamma)-1\right\}(y_i-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}) = \boldsymbol{0}$$

and $\hat{\omega}_i(\gamma) = \omega(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}}_\gamma,\gamma)$. Under the MAR assumption, let $\boldsymbol{\beta}^*$ be the probability limit of $\hat{\boldsymbol{\beta}}_\gamma$. If the condition $\mathrm{E}(Y|\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta}^*$ is satisfied, then

$$\sqrt{n}\left(\hat{\eta}_\gamma - \eta^*\right) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, V_\gamma\right),$$

where

$$V_\gamma = \mathrm{var}\{\boldsymbol{b}(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{\beta}^*\} + \mathrm{E}[\delta\{\omega(\boldsymbol{X};\boldsymbol{\phi}^*,\gamma)\}^2\,\mathrm{var}(Y\mid\boldsymbol{X})] \tag{3.20}$$

and $\boldsymbol{\phi}^*$ is the probability limit of $\hat{\boldsymbol{\phi}}_\gamma$.

In estimating γ, we propose to choose γ which minimizes the variance of $\hat{\eta}_\gamma$, V_γ. Note that the first term in V_γ does not depend on γ. Therefore, it suffices to find γ which minimizes the second term in V_γ.

Theorem 3.3.1. Let

$$d_i(\gamma) = \delta_i\left\{\hat{\omega}_i(\gamma)\right\}^2\left(y_i - \boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\zeta}}_\gamma\right)^2,$$

where

$$\hat{\boldsymbol{\zeta}}_\gamma = \left\{ \sum_{i=1}^{n} \delta_i \frac{\partial \omega(\boldsymbol{b}_i; \boldsymbol{\phi}, \gamma)}{\partial \boldsymbol{\phi}} \boldsymbol{b}_i^{\mathrm{T}} \Big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}_\gamma} \right\}^{-1} \left\{ \sum_{i=1}^{n} \delta_i \frac{\partial \omega(\boldsymbol{b}_i; \boldsymbol{\phi}, \gamma)}{\partial \boldsymbol{\phi}} y_i \Big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}_\gamma} \right\}.$$

If the OR model is correctly specified, as $n \to \infty$,

$$\bar{d}_\gamma = \frac{1}{n} \sum_{i=1}^{n} d_{\gamma,i} \xrightarrow{\mathrm{P}} \mathrm{E}[\delta \{\omega(\boldsymbol{X}; \boldsymbol{\phi}^*, \gamma)\}^2 \, \mathrm{var}(Y \mid \boldsymbol{X})].$$

Note that both $\hat{\boldsymbol{\zeta}}_\gamma$ and $\hat{\boldsymbol{\phi}}_\gamma$ depend on γ. To reduce the effect of estimation error in determining the tuning parameter γ, we propose to find $\hat{\gamma}$ by minimizing $\bar{d}_\gamma$ with the K-fold cross-validation.

3.4 Outliers in Outcome Regression Model

So far, we examine the IBC condition which leads to robust inference against misspecification of the OR model or PS model. We also derive the PS weight model based on the γ-power divergence. In this section, we discuss how to allows robust inference against outliers in outcome, Y.

In the presence of outliers, one may use the t-distribution (Lange et al., 1989) to allow the robust inference against outliers. Eguchi (2021) independently derived the t-distribution based on the γ-power divergence. However, it is not straightforward how to extend the IBC condition to the t-distribution. Instead, consider the following random effect model, namely DHGLM (Lee and Nelder, 2006)

$$Y \mid \boldsymbol{x}, u \sim \mathrm{N}\left\{\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})u\right\}, \; u \sim \text{Inv-gamma}(\alpha+1, \alpha), \quad (3.21)$$

where

$$\mu(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\beta}^* \text{ and } \; \sigma^2(\boldsymbol{x}) = \frac{\sigma_0^2}{\omega(\boldsymbol{x}) - 1}.$$

We can see that correct specification of models $\mathrm{E}(Y|\boldsymbol{x})$ and $\mathrm{P}(\delta = 1|\boldsymbol{x})$ is equivalent to that of mean $\mu(\boldsymbol{x})$ and dispersion $\sigma^2(\boldsymbol{x})$ in DHGLM, respectively. For missing mechanism, we extend the definition of the MAR assumption as

$$(Y, u) \perp \delta \mid \boldsymbol{x}$$

to maintain the property that δ only depends on $\boldsymbol{x}$ (Ibrahim and Molenberghs, 2009). Then, the random effect model (3.21) induces the marginal distribution of $Y|\boldsymbol{x}$ as t-distribution. Here, the constraint $\mathrm{E}(u) = 1$ guarantees that $\mathrm{var}(Y|\boldsymbol{x}) = \sigma^2(\boldsymbol{x})$ does not depend on the degrees of freedom of resulting t-distribution (Lee and Nelder, 2006).

Under the model (3.21), the ML imputation method of Han et al. (2022a) gives an imputed estimator

$$\hat{\eta}_{\mathrm{D}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{D}}\right\}, \tag{3.22}$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{D}}$ is the maximum h-likelihood estimator (Lee et al., 2017) obtained by solving

$$\sum_{i=1}^{n}\delta_i \boldsymbol{b}_i \frac{\hat{\omega}_i - 1}{\tilde{u}_i}(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{0}. \tag{3.23}$$

Here, $\hat{\omega}_i$ is an estimator of the PS weight and

$$\tilde{u}_i = \frac{(\hat{\omega}_i - 1)(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta})^2/\sigma_0^2 + 2\alpha}{3 + 2\alpha}.$$

In (3.23), if the ith observation $(\delta_i y_i, \delta_i \boldsymbol{x}_i)$ has large residual,

$$\left|y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right|/\tilde{u}_i \to 0 \quad \text{as} \quad \left|y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}\right| \to \infty.$$

This shows how the robustness against outliers in Y can be achieved. Moreover, $\tilde{u}_i \to 1$ as $\alpha \to \infty$, i.e., the model (3.21) satisfies the

IBC condition if there is no outlier in Y. In Appendix, we present how to construct the h-likelihood of the model (3.21).

In summary, we can allow various types of robustness by using the DHGLM in modeling approach. For $\hat{\eta}_{\mathrm{D}}$, robust inference against outliers in OR model is allowed by introducing a random effect u in dispersion. The estimator $\hat{\eta}_{\mathrm{D}}$ comes the estimator $\hat{\eta}_{\mathrm{IBC}}$ when there is no outlier. Moreover, let

$$\hat{\eta}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\boldsymbol{b}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right\}, \tag{3.24}$$

be the regression imputed estimator, where $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is the solution to $\sum_{i=1}^{n}\delta_i\boldsymbol{b}_i(y_i - \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{0}$ (Han et al., 2022a). The consistency of estimator $\hat{\eta}_{\mathrm{ML}}$ holds only when the OR model is correctly specified, i.e., double robustness is no longer guaranteed. The presented estimators have the following relationship:

$$\hat{\eta}_{\mathrm{D}} \underset{\tilde{u}_i=1}{\rightarrow} \hat{\eta}_{\mathrm{IBC}} \underset{\hat{\omega}_i=2}{\rightarrow} \hat{\eta}_{\mathrm{ML}}.$$

Here, $\hat{\omega}_i = 2$ indicates that

$$\hat{\mathrm{P}}(\delta_i = 1 \mid \boldsymbol{x}_i) = \hat{\mathrm{P}}(\delta_i = 0 \mid \boldsymbol{x}_i) = 0.5,$$

i.e., the missing mechanism is the MCAR which does not account the PS model in estimating $\boldsymbol{\beta}$. In $\hat{\eta}_{\mathrm{IBC}}$, the use of power-odds model allows robust inference against the misspecification of the PS model compared to the log-odds PS model by enlarging the class of PS models.

3.5 Simulation Study

In this section, we conduct simulation studies to compare various methods. $\bar{y}_{\mathrm{com}}$ is also considered as a benchmark. If a model de-

rived from the γ-power divergence is used, γ is chosen to maximize the likelihood in modeling approach or minimize the variance of corresponding estimator in estimating equation approach with 10-fold cross validation to avoid the overfitting. In all cases, covariates $\boldsymbol{b}(\boldsymbol{x}) = (1, x_1, x_2)$ are generated as $x_1 \sim \mathrm{U}(0,1)$, $x_2 \sim \exp(1)$, and parameters in propensity score models are set to satisfy around 70% of responses which are observed. The performance of various estimators is examined in terms of the following quantities: (i) bias= $\eta - \bar{\eta}$, $\bar{\eta} = \sum_{t=1}^{T} \hat{\eta}^{(t)}/T$, (ii) SD= $\{\sum_{t=1}^{T}(\hat{\eta}^{(t)} - \bar{\eta})^2/(T-1)\}^{0.5}$, (iii) bias/SD, and (iv) RMSE= $\{\sum_{t=1}^{T}(\hat{\eta}^{(t)} - \eta)^2/T\}^{0.5}$, where $T = 500$ is the number of iterations.

3.5.1 Robustness against Model Misspecification

In this section, we examine the performance of proposed methods. For $\hat{\eta}_{\mathrm{D}}$, we consider the power-odds PS model. For $\hat{\eta}_{\mathrm{IBC}}$, we consider two PS models: power-odds PS model ($\hat{\eta}_{\mathrm{pow}}$) and log-odds PS model ($\hat{\eta}_{\log}$). Given covariates $\boldsymbol{b}(\boldsymbol{x})$, Y and δ are generated as follows.

> OM1 (Outcome regression Model 1): $Y|\boldsymbol{x}$ follows normal distribution with mean $\mathrm{E}(Y|\boldsymbol{x}) = 1+0.2x_1+0.2x_2$ and variance 1. Under OM1, $\eta^* = 1.4$.
>
> OM2: $Y|\boldsymbol{x}$ follows exponential distribution with mean $\mathrm{E}(Y|\boldsymbol{x}) = 1 + 0.2x_1 + 0.2x_2 + e^{x_1} + x_2^2$. Under OM2, $\eta^* = 2.9 + 0.5e^2$.
>
> PM1 (Propensity score Model 1): $\delta|\boldsymbol{x}$ follow Bernoulli distribution with
>
> $$\frac{\mathrm{P}(\delta = 1 \mid \boldsymbol{x})}{\mathrm{P}(\delta = 0 \mid \boldsymbol{x})} = \exp(-0.1 + 0.5x_1 + 0.5x_2),$$

i.e., log-odds model (logistic model) is considered.

PM2: $\delta|\boldsymbol{x}$ follow Bernoulli distribution with the power-odds

$$\frac{\mathrm{P}(\delta = 1 \mid \boldsymbol{x})}{\mathrm{P}(\delta = 0 \mid \boldsymbol{x})} = \frac{6}{1 + x_1 + x_2}.$$

Compared to the OM1, $\mathrm{E}(Y|\boldsymbol{x})$ is not correctly specified in OM2. Moreover, $\mathrm{var}(Y|\boldsymbol{x})$ is proportional to $\mathrm{E}(Y|\boldsymbol{x})$.

OM1PM1	$n = 500$					$n = 1000$				
	$\bar{y}_{\mathrm{com}}$	$\hat{\eta}_{\mathrm{D}}$	$\hat{\eta}_{\mathrm{pow}}$	$\hat{\eta}_{\mathrm{log}}$	$\hat{\eta}_{\mathrm{ML}}$	$\bar{y}_{\mathrm{com}}$	$\hat{\eta}_{\mathrm{D}}$	$\hat{\eta}_{\mathrm{pow}}$	$\hat{\eta}_{\mathrm{log}}$	$\hat{\eta}_{\mathrm{ML}}$
bias	-0.0057	-0.0038	-0.0038	-0.0038	-0.0039	0.0015	0.0012	0.0012	0.0012	0.0012
SD	0.0450	0.0545	0.0545	0.0546	0.0544	0.0324	0.0391	0.0391	0.0391	0.0390
bias/SD	-0.1267	-0.0698	-0.0697	-0.0696	-0.0710	0.0471	0.0297	0.0299	0.0302	0.0307
RMSE	0.0453	0.0546	0.0546	0.0547	0.0545	0.0324	0.0391	0.0391	0.0391	0.0390

Table 3.1: Simulation results under the OM1PM1 case.

OM2PM2	$n = 500$					$n = 1000$				
	$\bar{y}_{\mathrm{com}}$	$\hat{\eta}_{\mathrm{D}}$	$\hat{\eta}_{\mathrm{pow}}$	$\hat{\eta}_{\mathrm{log}}$	$\hat{\eta}_{\mathrm{ML}}$	$\bar{y}_{\mathrm{com}}$	$\hat{\eta}_{\mathrm{D}}$	$\hat{\eta}_{\mathrm{pow}}$	$\hat{\eta}_{\mathrm{log}}$	$\hat{\eta}_{\mathrm{ML}}$
bias	0.0140	-0.0054	0.0031	0.0469	0.0003	0.0139	-0.0028	0.0088	0.0567	0.0050
SD	0.4448	0.5524	0.5560	0.5964	0.5586	0.3028	0.4173	0.4240	0.4759	0.4288
bias/SD	0.0314	-0.0097	0.0056	0.0787	0.0005	0.0459	-0.0068	0.0208	0.1191	0.0117
RMSE	0.4446	0.5519	0.5555	0.5977	0.5580	0.3029	0.4169	0.4237	0.4788	0.4284

Table 3.2: Simulation results under the OM2PM2 case.

Based on simulation results in Table 3.1 and 3.2, we can check that proposed estimators $\hat{\eta}_{\mathrm{D}}$ and $\hat{\eta}_{\mathrm{pow}}$ are consistent even though $\mathrm{E}(Y|\boldsymbol{x})$ is not correctly specified in OM2.

Recall that the pow-odds model covers the log-odds model. Thus, we also consider the following PS model to examine the performance of estimators when both outcome model and PS model are incorrectly specified.

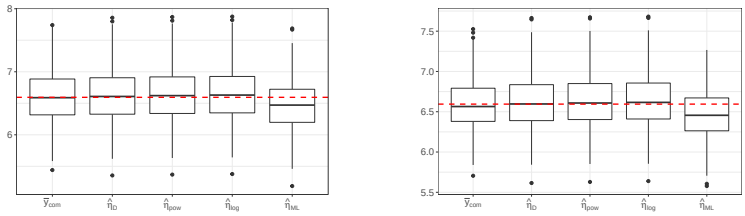
PM3: $\delta|\boldsymbol{x}$ follow Bernoulli distribution with

$$\frac{\mathrm{P}(\delta = 1 \mid \boldsymbol{x})}{\mathrm{P}(\delta = 0 \mid \boldsymbol{x})} = \exp(-0.1 + 0.5x_1^2 + 0.5x_2^2).$$

Note that the functional form of PM3 is the log-odds model but covariates are (x_1^2, x_2^2), not (x_1, x_2). Therefore, the consistency is no longer guaranteed for all estimators under OM2PM3.

OM2PM3	$n = 500$					$n = 1000$				
	$\bar{y}_{\mathrm{com}}$	$\hat{\eta}_{\mathrm{D}}$	$\hat{\eta}_{\mathrm{pow}}$	$\hat{\eta}_{\mathrm{log}}$	$\hat{\eta}_{\mathrm{ML}}$	$\bar{y}_{\mathrm{com}}$	$\hat{\eta}_{\mathrm{D}}$	$\hat{\eta}_{\mathrm{pow}}$	$\hat{\eta}_{\mathrm{log}}$	$\hat{\eta}_{\mathrm{ML}}$
bias	0.0171	0.0296	0.0419	0.0503	-0.1326	-0.0043	0.0163	0.0287	0.0366	-0.1291
SD	0.4312	0.4508	0.4512	0.4521	0.4095	0.3059	0.3249	0.3250	0.3258	0.2893
bias/SD	0.0397	0.0656	0.0928	0.1114	-0.3237	-0.0140	0.0502	0.0882	0.1123	-0.4461
RMSE	0.4311	0.4513	0.4527	0.4544	0.4300	0.3056	0.3250	0.3260	0.3275	0.3165

Table 3.3: Simulation results under the OM2PM3 case.



(a) (b)

Figure 3.1: Boxplots of estimators in OM2PM3: (a) for $n = 500$ and (b) for $n = 1000$.

Table 3.3 shows the simulation result of various estimators. We can see that proposed method $\hat{\eta}_{\mathrm{D}}$ and $\hat{\eta}_{\mathrm{pow}}$ give smaller bias than existing methods $\hat{\eta}_{\mathrm{log}}$ and $\hat{\eta}_{\mathrm{ML}}$. Moreover, even though there is no outlier, $\hat{\eta}_{\mathrm{D}}$ gives smaller bias and variance compared to $\hat{\eta}_{\mathrm{pow}}$.

3.5.2 Robustness against Outliers

In this section, we consider OM3PM2 model with

$$\text{OM3: } Y \mid \boldsymbol{x} \sim \text{Exp}\left(1 + 0.2x_1 + 0.2x_2\right)$$

to check the performance of various estimators. Note that all estimators are consistent under OM3 if there is no outlier. After data are generated, additional noise generated from $\text{U}(0, 100)$ is added to 20% of observed outcomes.

Table 3.4 and Figure 3.2 shows performances of various estimators. Compared to $\hat{\eta}_{\text{pow}}$, $\hat{\eta}_{\text{log}}$, and $\hat{\eta}_{\text{ML}}$, only $\hat{\eta}_{\text{D}}$ reduces the bias due to the outliers. Moreover, $\hat{\eta}_{\text{D}}$ gives comparable results to the benchmark $\bar{y}_{\text{com}}$.

OM3PM2	$n = 500$					$n = 1000$				
	$\bar{y}_{\text{com}}$	$\hat{\eta}_{\text{D}}$	$\hat{\eta}_{\text{pow}}$	$\hat{\eta}_{\text{log}}$	$\hat{\eta}_{\text{ML}}$	$\bar{y}_{\text{com}}$	$\hat{\eta}_{\text{D}}$	$\hat{\eta}_{\text{pow}}$	$\hat{\eta}_{\text{log}}$	$\hat{\eta}_{\text{ML}}$
bias	6.7189	6.8475	9.9148	9.9179	9.9154	6.7791	6.9004	10.0342	10.0352	10.0335
SD	0.8931	0.9847	1.3061	1.3077	1.3052	0.6850	0.7549	0.9917	0.9940	0.9918
bias/SD	7.5229	6.9542	7.5909	7.5842	7.5969	9.8960	9.1404	10.1187	10.0957	10.1166
RMSE	6.7779	6.9178	10.0003	10.0035	10.0008	6.8136	6.9415	10.0830	10.0842	10.0823

Table 3.4: Simulation results under the OM3PM2 case with outliers.

3.6 Conclusion

In this chapter, we investigate the conditions under which the consistency of estimators is guaranteed when the study variable is only partially observed. To obtain the doubly robust imputed estimator, we propose the IBC condition that ensures the equivalence between the imputation method and the weighting method

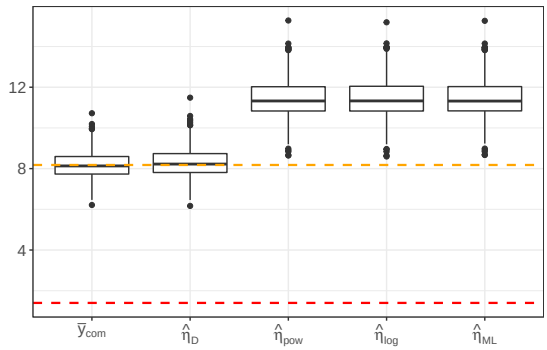


Figure 3.2: Boxplots of estimator in the presence of outliers in Y and δ under OM3PM2 when $n = 1000$. Red dotted line indicates η^* and orange dotted line indicates the average of $\bar{y}_{\text{com}}$.

by means of the PS weight. An interesting point of the IBC condition is that estimating equation for regression coefficient in OR model is estimated by using the PS as weights. In estimating the PS weight, the log-odds model can be used, which can also be derived by using the information projection approach with the KL divergence.

Eguchi presented the γ-power divergence to generalize the KL divergence. We show how the IBC condition can be achieved in terms of the γ-power divergence. The PS model induced by the γ-power divergence is equivalent to the power-odds model. The power-odds PS model allows more general class of PS models compared to the log-odds PS model.

One aspect of Eguchi's γ-power divergence is that the γ-power divergence allows robust inferences against outliers and γ-power divergence for the outcome regression model induces the t-distribution.

To extend the IBC condition while allowing the robustness against outliers, we introduce the DHGLM of Lee and Nelder (2006) in the modeling approach by incorporating a random effect into the dispersion model since achieving the IBC condition with the t-distribution is not straightforward. In the DHGLM, double robustness can be understood as a correct specification of the mean and variance of the DHGLM. Based on the DHGLM framework, we can have ML estimation for fixed parameters as well as ML imputation for random parameters, namely random effect and missing data. Advantage of modeling approach is efficient algorithm for estimation of fixed parameters and imputation of random parameters.

Appendix

Proof of Lemma 3.3.1

Proof. Let

$$\mathcal{L} = \left\{ f_0 : \int \boldsymbol{b}(\boldsymbol{x}) f_0(\boldsymbol{x}) d\boldsymbol{x} = \frac{1}{1-p} \left[\mathrm{E}\left\{ \boldsymbol{b}(\boldsymbol{X}) \right\} - p \int \boldsymbol{b}(\boldsymbol{x}) f_1(\boldsymbol{x}) d\boldsymbol{x} \right] \right\} \tag{3.25}$$

be the linear family, where $f_k(\boldsymbol{x}) = f(\boldsymbol{x}|\delta = k)$, $k = 0, 1$ and $p = \mathrm{P}(\delta = 1)$. By using the information projection method, we want to find $\tilde{f}_0$ which minimizes $D_\gamma(f_0 \| f_1)$ given f_1, i.e.,

$$\tilde{f}_0 = \arg\min_{f_0 \in \mathcal{L}} D_\gamma(f_0 \| f_1). \tag{3.26}$$

If $f_0 \in \mathcal{L}$ and $g_0 \in \mathcal{L}$

$$h_0 = t f_0 + (1-t) g_0 \in \mathcal{L} \tag{3.27}$$

for all $t \in [0,1]$. In information projection theory, the line of the form (3.27) is called m-geodesic line. Moreover, the γ-power divergence has the dually flat structure with m-geodesic line. Thus, the information projection (3.26) is not only well-defined, but it tells us how to estimate f_1.

Given the empirical distribution $\hat{P}_1(\boldsymbol{x}) = n_{\text{obs}}^{-1}\delta_i I(\boldsymbol{x} = \boldsymbol{x}_i)$, $\hat{f}_1$ which minimizes the γ-power divergence $D_\gamma(\hat{P}_1 \| f_1)$ under constraint $\mathcal{L}_1 = \{f_1 : \sum_{i=1}^n \delta_i f_1(\boldsymbol{x}_i) = 1\}$ is given by $\hat{f}_1(\boldsymbol{x}_i) = n_{\text{obs}}^{-1}$ for $\{\boldsymbol{x}_i : \delta_i = 1\}$. Moreover, $\tilde{f}_0(\boldsymbol{x}_i; \boldsymbol{\phi}, \gamma)$ which minimizes the γ-power divergence $D_\gamma(f_0 \| \hat{f}_1)$ under constraint $\tilde{f}_0 \in \mathcal{L}$ is given as

$$\tilde{f}_0(\boldsymbol{x}_i; \boldsymbol{\phi}, \gamma) = \left(1 + \gamma \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\phi}\right)^{1/\gamma} \hat{f}_1(\boldsymbol{x}_i), \tag{3.28}$$

for $\delta_i = 1$. Note that

$$\frac{1}{\mathrm{P}(\delta = 1 \mid \boldsymbol{x})} = 1 + \frac{1-p}{p}\frac{f_0(\boldsymbol{x})}{f_1(\boldsymbol{x})}.$$

By combining $\hat{p} = n_{\text{obs}}/n$ and (3.28), we have the PS model

$$\omega(\boldsymbol{x}_i; \boldsymbol{\phi}, \gamma) = 1 + \frac{n_{\text{mis}}}{n_{\text{obs}}}\left(1 + \gamma \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\phi}\right)^{1/\gamma}$$

for $\delta_i = 1$. □

Proof of Theorem 3.3.1

Proof. Given γ, let

$$\begin{aligned} \tilde{\eta}_\gamma(\boldsymbol{\phi}) &= \frac{1}{n}\sum_{i=1}^n \delta_i \omega(\boldsymbol{x}_i; \boldsymbol{\phi}, \gamma) y_i, \\ U_\gamma(\boldsymbol{\phi}) &= \frac{1}{n}\sum_{i=1}^n \delta_i \omega(\boldsymbol{x}_i; \boldsymbol{\phi}, \gamma) \boldsymbol{b}_i - \frac{1}{n}\sum_{i=1}^n \boldsymbol{b}_i. \end{aligned}$$

Then, $\hat{\eta}_\gamma = \tilde{\eta}_\gamma(\hat{\boldsymbol{\phi}}_\gamma)$ and $U_\gamma(\hat{\boldsymbol{\phi}}_\gamma) = \mathbf{0}$. Since we use the estimate $\hat{\boldsymbol{\phi}}_\gamma$ instead of fixed, true value $\boldsymbol{\phi}^*$, summands $\{\delta_i \omega(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}}_\gamma, \gamma) y_i\}$ are no

longer independent. Note that Talyor expansions of $\hat{\eta}_\gamma = \tilde{\eta}_\gamma(\hat{\boldsymbol{\phi}}_\gamma)$ and $\mathbf{0} = U_\gamma(\hat{\boldsymbol{\phi}}_\gamma)$ at $\boldsymbol{\phi}^*$ are

$$\begin{aligned}
\hat{\eta}_\gamma &= \tilde{\eta}_\gamma(\hat{\boldsymbol{\phi}}_\gamma) \\
&= \tilde{\eta}_\gamma(\boldsymbol{\phi}^*) + \left\{ \frac{\partial \tilde{\eta}_\gamma(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} \right\}^{\mathrm{T}} (\hat{\boldsymbol{\phi}}_\gamma - \boldsymbol{\phi}^*) \\
&\quad + o_p(n^{-1/2}), \qquad (3.29) \\
\mathbf{0} &= U_\gamma(\hat{\boldsymbol{\phi}}_\gamma) \\
&= U_\gamma(\boldsymbol{\phi}^*) + \left\{ \frac{\partial U_\gamma(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} \right\}^{\mathrm{T}} (\hat{\boldsymbol{\phi}}_\gamma - \boldsymbol{\phi}^*) \\
&\quad + o_p(n^{-1/2}). \qquad (3.30)
\end{aligned}$$

By combining (3.29) and (3.30) we can express $\hat{\eta}_\gamma$ as

$$\begin{aligned}
\hat{\eta}_\gamma &= \tilde{\eta}_\gamma(\boldsymbol{\phi}^*) + \left\{ \frac{\partial \tilde{\eta}_\gamma(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} \right\}^{\mathrm{T}} (\hat{\boldsymbol{\phi}}_\gamma - \boldsymbol{\phi}^*) + o_p(n^{-1/2}), \\
&= \frac{1}{n}\sum_{i=1}^{n} \delta_i \omega(\boldsymbol{x}_i; \boldsymbol{\phi}^*, \gamma) y_i - \frac{1}{n} \hat{\boldsymbol{\zeta}}_\gamma^{\mathrm{T}} \left\{ \sum_{i=1}^{n} \delta_i \omega(\boldsymbol{x}_i; \boldsymbol{\phi}^*, \gamma) \boldsymbol{b}_i - \sum_{i=1}^{n} \boldsymbol{b}_i \right\} \\
&\quad + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{ \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\zeta}}_\gamma + \delta_i \omega(\boldsymbol{x}_i; \boldsymbol{\phi}^*, \gamma)(y_i - \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\zeta}}_\gamma) \right\} + o_p(n^{-1/2}).
\end{aligned}$$

Under the assumption $\mathrm{E}(Y|\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta}^*$, $\mathrm{E}(\hat{\boldsymbol{\zeta}}_\gamma) = \mathrm{E}\{\mathrm{E}(\hat{\boldsymbol{\zeta}}_\gamma|\boldsymbol{x})\} = \boldsymbol{\beta}^*$ and

$$\mathrm{var}\{Y - \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta}^* \mid \boldsymbol{x}\} = \mathrm{E}[\{Y - \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta}^*\}^2 \mid \boldsymbol{x}]$$

which gives

$$\begin{aligned}
& \mathrm{E}[\delta\{\omega(\boldsymbol{x}; \boldsymbol{\phi}^*, \gamma)\}^2 \,\mathrm{var}\{Y - \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta}^*\} \mid \boldsymbol{x}] \\
= \quad & \mathrm{E}[\delta\{\omega(\boldsymbol{x}; \boldsymbol{\phi}^*, \gamma)\}^2 \,\mathrm{E}[\{Y - \boldsymbol{b}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta}^*\}^2 | \boldsymbol{x}]].
\end{aligned}$$

Linearization technique makes summands be asymptotically independent which gives simple variance estimator, especially the second term in V_γ. □

Construction of the h-likelihood

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma_0^2, \alpha)^{\mathrm{T}}$ be vector of fixed parameters, $\boldsymbol{u}$ be vector of random effects and $\boldsymbol{y}_{\mathrm{mis}}$ be the vector of nonresponses. In model (3.21), $\boldsymbol{\xi} = \log(\boldsymbol{u})$ scale is canonical for $\boldsymbol{\beta}$ and σ_0^2, but not for α. Also, $\boldsymbol{y}_{\mathrm{mis}}$-scale is not canonical for σ_0^2. Instead, consider a $\boldsymbol{w}$-scale defined as

$$w_i = \left\{ \frac{\omega_i - 1}{\sigma_0^2} e^{-\xi_i} \right\}^{0.5} y_{\mathrm{mis},i}.$$

On the $\boldsymbol{\xi}$-scale and $\boldsymbol{w}$-scale, the extended likelihood is

$$\begin{aligned} \ell_e(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{w}) &= \ell_e(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y}_{\mathrm{mis}}) + \log \left| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\xi}} \right| + \log \left| \frac{\partial \boldsymbol{y}_{\mathrm{mis}}}{\partial \boldsymbol{w}} \right| \qquad (3.31) \\ &= \sum_{i=1}^{n} \delta_i \left[-\frac{1}{2} \log \left\{ \frac{2\pi\sigma_0^2}{\omega_i - 1} \right\} - \frac{1}{2} \log u_i - \frac{1}{2\sigma_0^2} \frac{\omega_i - 1}{u_i} \left(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2 \right] \\ &\quad + \sum_{i=1}^{n} (1 - \delta_i) \left[-\frac{1}{2} \log 2\pi - \frac{1}{2\sigma_0^2} \frac{\omega_i - 1}{u_i} \left(y_{\mathrm{mis},i} - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2 \right] \\ &\quad + \sum_{i=1}^{n} \left\{ (\alpha + 1) \log(\alpha) - \log \Gamma(\alpha + 1) - (\alpha + 1) \log u_i - \frac{\alpha}{u_i} \right\}, \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. Furthermore, consider $\boldsymbol{v}$-scale given as

$$v_i = (1.5 + \alpha)^{0.5} \xi_i.$$

Then, all random parameters $\boldsymbol{v}$ and $\boldsymbol{y}_{\mathrm{mis}}$ are canonical to fixed parameters $\boldsymbol{\theta}$, i.e., joint maximization of the h-likelihood gives

MLEs of $(\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{y}_{\text{mis}})$. In this case, the h-likelihood becomes

$$\begin{aligned} h(\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{y}_{\text{mis}}) &= \sum_{i=1}^{n} \delta_i \left[-\frac{1}{2} \log \left\{ \frac{2\pi\sigma_0^2}{\omega_i - 1} \right\} - \frac{v_i}{2} - \frac{\omega_i - 1}{2\sigma_0^2} e^{-v_i} \left(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2 \right] \\ &\quad + \sum_{i=1}^{n} (1 - \delta_i) \left[-\frac{1}{2} \log 2\pi - \frac{\omega_i - 1}{2\sigma_0^2} e^{-v_i} \left(y_{\text{mis},i} - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2 \right] \\ &\quad + \sum_{i=1}^{n} \left\{ -(\alpha + 1) v_i - \alpha e^{-v_i} - \frac{1}{2} \log(1.5 + \alpha) \right\}, \\ &\quad + n(\alpha + 1) \log(\alpha) - n \log \Gamma(\alpha + 1). \end{aligned} \tag{3.32}$$

For $\delta_i = 1$, canonical function of v_i (or, the mode of the h-likelihood with respect to v_i given $\boldsymbol{\theta}$) is

$$e^{-\tilde{v}_i} = \frac{3 + 2\alpha}{(\omega_i - 1)(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta})^2 / \sigma_0^2 + 2\alpha}.$$

Moreover, canonical function of $y_{\text{mis},i}$ is

$$\tilde{y}_{\text{mis},i} = \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta}.$$

By the property of the h-likelihood, joint maximization of $(\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{y}_{\text{mis}})$ gives the MLE $\hat{\boldsymbol{\beta}}_{\text{D}}$ and $\hat{\sigma}_0^2$ by solving

$$\sum_{i=1}^{n} \delta_i \boldsymbol{b}_i \frac{\omega_i - 1}{\tilde{u}_i} \left(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta} \right) = \mathbf{0}, \tag{3.33}$$

$$\sum_{i=1}^{n} \delta_i \left\{ -\frac{1}{2\sigma_0^2} + \frac{1}{2(\sigma_0^2)^2} \frac{\omega_i - 1}{\tilde{u}_i} \left(y_i - \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2 \right\} = 0, \tag{3.34}$$

where $\tilde{u}_i = \exp(\tilde{v}_i)$. In the model (3.21), $\hat{\alpha}$, MLE of α, can be obtained by the joint maximization of the h-likelihood (3.32) even though there is no explicit form of $\hat{\alpha}$. Alternatively, the moment-based estimator for α can be used. Note that

$$\mathrm{E}\left(\tilde{u}_i\right) = \mathrm{E}\left\{\mathrm{E}\left(\tilde{u}_i \mid \boldsymbol{x}_i\right)\right\} = \frac{1 + 2\alpha}{3 + 2\alpha}.$$

Then, α can be estimated by solving

$$\bar{\tilde{u}} = \frac{1}{n_{\text{obs}}} \sum_{i=1}^{n} \delta_i \tilde{u}_i = \frac{1+2\alpha}{3+2\alpha}.$$

Chapter 4

Enhanced Laplace Approximation

Chapter Summary

The Laplace approximation (LA) has been proposed as a method for approximating the marginal likelihood of statistical models with latent variables. However, the approximate maximum likelihood estimators (MLEs) based on the LA are often biased for binary or spatial data, and the corresponding Hessian matrix underestimates the standard errors of these approximate MLEs. A higher-order approximation has been proposed; however, it cannot be applied to complicated models such as correlated random effects models and does not provide consistent variance estimators. In this chapter, we propose an enhanced LA (ELA) that provides the true MLE and its consistent variance estimator. We study its relationship to the variational Bayes method. We also introduce a new restricted maximum likelihood estimator (REMLE) for estimating dispersion parameters. The results of numerical studies show that the ELA provides a satisfactory MLE and REMLE, as well as their variance estimators for fixed parameters. The MLE and REMLE can be viewed as posterior mode and marginal posterior mode under flat priors, respectively. Some comparisons are also made with Bayesian procedures under different priors.

4.1 Review of the LA

Throughout the chapter, we impose the following regularity conditions:

R1. The parameter space Θ is convex.

R2. All likelihoods are smooth and unimodal with respect to θ.

The LA to the marginal likelihood $L_m(\theta)$ is

$$\hat{L}_m(\theta) = H(\theta, \tilde{z}) \left| \frac{1}{2\pi} \tilde{\Omega}_{zz} \right|^{-\frac{1}{2}},$$

where $\tilde{z} = \arg\max_z h(\theta, z) = \arg\max_z \ell_p(z|y;\theta)$ and

$$\tilde{\Omega}_{zz} = -\frac{\partial^2}{\partial z \partial z^{\mathrm{T}}} h(\theta, z)\Big|_{z=\tilde{z}} = -\frac{\partial^2}{\partial z \partial z^{\mathrm{T}}} \ell_p(z|y;\theta)\Big|_{z=\tilde{z}}.$$

According to (1.2), the LA to $L_m(\theta)$ can be defined as

$$\hat{L}_m(\theta) = H(\theta, \tilde{z}) / \hat{L}_p(\tilde{z}|y;\theta),$$

This formulation can be viewed as the use of an approximate predictive likelihood $\hat{L}_p(z|y;\theta)$ in (1.2), based on the normal distribution

$$z \mid y \sim \mathrm{N}\left(\tilde{z}, \tilde{\Omega}_{zz}^{-1}\right). \tag{4.1}$$

This gives

$$\hat{\ell}_m(\theta) = \log \hat{L}_m(\theta) = h(\theta, \tilde{z}) - \hat{\ell}_p(\tilde{z}|y;\theta) = h(\theta, \tilde{z}) - \frac{1}{2} \log \left| \frac{1}{2\pi} \tilde{\Omega}_{zz} \right|.$$

Thus, the LA is exact when the predictive likelihood is normal. Let $\hat{\theta}$ be the MLE and $\hat{\theta}^{\mathrm{L}}$ be the approximate MLE, which are modes of $\ell_m(\theta)$ and $\hat{\ell}_m(\theta)$, respectively. As the sample size $n \to \infty$, if $\hat{\theta} \xrightarrow{\mathrm{P}} \theta_0$ and

$$\ell_m(\theta) - \hat{\ell}_m(\theta) \xrightarrow{\mathrm{P}} 0, \text{ uniformly in } \theta, \tag{4.2}$$

then $\hat{\theta}^{\mathrm{L}} \xrightarrow{\mathrm{P}} \theta_0$. However, in general, it is difficult to justify that the LA $\hat{\ell}_m(\theta)$ satisfies the uniform convergence condition (4.2). Let θ_0^{L} be the probability limit of $\hat{\theta}^{\mathrm{L}}$. If $\sqrt{n}(\hat{\theta}^{\mathrm{L}} - \theta_0^{\mathrm{L}}) = O_p(1)$, then

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{L}} - \theta_0^{\mathrm{L}}\right) \xrightarrow{\mathrm{d}} \mathrm{N}\left\{0, \mathcal{G}^{-1}\left(\theta_0^{\mathrm{L}}\right)\right\}, \tag{4.3}$$

where $\tilde{\mathcal{G}}(\theta) = \tilde{\mathcal{H}}(\theta)\tilde{\mathcal{K}}^{-1}(\theta)\tilde{\mathcal{H}}(\theta)$, $\tilde{\mathcal{H}}(\theta) = \mathrm{E}\{-\partial^2\hat{\ell}_m(\theta)/\partial\theta\partial\theta^{\mathrm{T}}\}$, $\tilde{\mathcal{K}}(\theta) = \mathrm{var}\{\partial\hat{\ell}_m(\theta)/\partial\theta\}$ and $\mathcal{G}(\theta_0^{\mathrm{L}}) = \lim_{n\to\infty} n^{-1}\tilde{\mathcal{G}}(\theta_0^{\mathrm{L}})$. Kristensen et al. (2016) and Lee et al. (2017) proposed the use of the inverse Hessian matrix of $\hat{\ell}_m(\theta)$ as a variance estimator of $\hat{\theta}^{\mathrm{L}}$. Ogden (2017) provided regularity conditions that guarantee asymptotic equivalence between the Hessian matrix of $\hat{\ell}_m(\theta)$ and that of $\ell_m(\theta)$. However, these conditions are hardly satisfied. As mentioned in Bologa et al. (2021), the Bayesian approach views the approximate MLE $\hat{\theta}^{\mathrm{L}}$ as an approximate mode of the posterior distribution under a flat prior on θ. Pauli et al. (2011) further showed that

$$\sqrt{n}\left(\theta - \hat{\theta}^{\mathrm{L}}\right) \mid y \xrightarrow{\mathrm{d}} \mathrm{N}\left\{0, \mathcal{H}^{-1}\left(\theta_0^{\mathrm{L}}\right)\right\},$$

where

$$\mathcal{H}\left(\theta_0^{\mathrm{L}}\right) = \lim_{n\to\infty} \frac{1}{n}\left\{-\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\hat{\ell}_m(\theta)\Big|_{\theta=\theta_0^{\mathrm{L}}}\right\}.$$

Thus, the variance estimators presented by Kristensen et al. (2016) and Lee et al. (2017) can be viewed as estimating the variance of the approximate Bayesian posterior mode $\hat{\theta}^{\mathrm{L}}$; see the numerical study of Bologa et al. (2021). In addition, Jin and Lee (2022) investigated the frequentist sandwich variance estimator (4.3) of the approximate MLE $\hat{\theta}^{\mathrm{L}}$.

Assume that d is the dimension of the integral in (1.1). The LA is valid in the sense that $\ell_m(\theta) - \hat{\ell}_m(\theta) = o_p(1)$ when $d = o(n^{1/3})$

(Shun and McCullagh, 1995; Ogden, 2021); thus, the LA may not be suitable for crossed random effects models with $d = O(n^{1/2})$ and correlated random effects models with $d = O(n)$. Furthermore, the performance of the LA is often unsatisfactory for binary outcomes (Shun, 1997). Thus, Shun and McCullagh (1995) proposed the use of the second-order LA in the exchangeable binary array model for salamander mating data. Shun (1997) investigated parameter estimation based on the second-order LA. However, due to the complexity of the approximation, the author could compute only some selected terms. Noh and Lee (2007) showed how to compute all the terms in the second-order LA and developed a REML estimation procedure for salamander mating data. However, the second-order LA can be applied to a limited class of models due to the complexity of the approximation. Furthermore, even if the second-order LA is applicable, the approximation is often slow because a considerable number of terms must be computed.

In summary, (i) $\ell_m(\theta) - \hat{\ell}_m(\theta) \neq o_p(1)$ as d increases, and (ii) even if $\ell_m(\theta) - \hat{\ell}_m(\theta) = o_p(1)$, the approximate MLE $\hat{\theta}^{\mathrm{L}}$ may not be the MLE $\hat{\theta}$. Furthermore, (iii) it is not known how to obtain a consistent variance estimator for $\hat{\theta}$. (iv) It is also of interest to have REMLEs for dispersion parameters. A general higher-order LA may not be sufficient for resolving these problems.

4.2 ELA

Assume that $q(z)$ is an arbitrary density function with $\int q(z)dz = 1$ that has the same support as the predictive likelihood $L_p(z|y;\theta)$.

Next, from (1.1) the marginal likelihood is defined as

$$L_m(\theta) = \int H(\theta, z)dz = \int \frac{H(\theta, z)}{q(z)} q(z)dz.$$

Thus, we can approximate the marginal likelihood as

$$\tilde{L}_B(\theta) = \frac{1}{B}\sum_{b=1}^{B} \frac{H(\theta, Z_b)}{q(Z_b)},$$

where Z_b are iid samples from $q(z)$. Since $H(\theta, Z_b)/q(Z_b)$ can be viewed as iid samples with the mean $L_m(\theta)$, $\tilde{L}_B(\theta)$ is a consistent estimator of $L_m(\theta)$, i.e., as $B \to \infty$,

$$\tilde{L}_B(\theta) \xrightarrow{\mathrm{P}} L_m(\theta).$$

The variational Bayes method has been proposed for approximating $\ell_m(\theta)$ (Kingma and Welling, 2013). For any $q(z)$,

$$\begin{aligned} \ell_m(\theta) &= \int \log\left\{\frac{H(\theta, z)}{q(z)}\right\} q(z)dz + R \\ &\geq \int \log\left\{\frac{H(\theta, z)}{q(z)}\right\} q(z)dz = \ell_v(\theta; q), \end{aligned}$$

where

$$R = \int \log\left\{\frac{q(z)}{L_p(z \mid y; \theta)}\right\} q(z)dz \geq 0,$$

and $\ell_v(\theta; q)$ is referred to as the evidence lower bound (ELBO). The marginal log-likelihood in (1.1) can be approximated by maximizing the ELBO

$$\hat{\ell}_v(\theta) = \max_q \ell_v(\theta; q).$$

In the variational Bayes methods, $q(z)$ is often assumed to have a normal density $\mathrm{N}(\mu, \Gamma)$ with an arbitrary mean μ and arbitrary covariance matrix Γ. In general, the ELBO is not a tight lower

bound, i.e., $\ell_m(\theta) - \hat{\ell}_v(\theta) > 0$ since $R > 0$. To address this issue, Burda et al. (2016) modified the ELBO as follows:

$$\tilde{\ell}_{v,B}(\theta; \mu, \Gamma) = \mathrm{E}_{\mu,\Gamma}\left\{\log \tilde{L}_B(\theta)\right\},$$

where Z_b are iid samples from $\mathrm{N}(\mu, \Gamma)$. The authors used the see-saw algorithm: (i) given θ, update (μ, Γ) by maximizing $\tilde{\ell}_{v,B}(\theta; \mu, \Gamma)$ and (ii) given (μ, Γ), update θ by maximizing $\tilde{\ell}_{v,B}(\theta; \mu, \Gamma)$. In correlated random effects models with $d = n$, estimating μ and Γ is not straightforward. The ELBO has been studied to approximate the marginal log-likelihood. However, the main interest of this chapter is how to obtain the true MLE $\hat{\theta}$ and its consistent variance estimator in general cases.

According to the expression (1.2), if the value of $L_p(z^*|y;\theta)$ is known at any point z^*, it is immediate that $L_m(\theta) = H(\theta, z^*)/L_p(z^*|y;\theta)$. However, in general, $L_p(z|y;\theta)$ is not known for all z. Recall that the LA approximates the predictive likelihood $L_p(z|y;\theta)$ at $\tilde{z}$ by $\hat{L}_p(\tilde{z}|y;\theta)$ as

$$\hat{L}_m(\theta) = H(\theta, \tilde{z})/\hat{L}_p(\tilde{z}|y;\theta).$$

Since $\ell_m(\theta) - \hat{\ell}_m(\theta) = \hat{\ell}_p(\tilde{z}|y;\theta) - \ell_p(\tilde{z}|y;\theta)$, the accuracy of the LA is the same as that of the predictive likelihood $\hat{L}_p(z|y;\theta)$. Let

$$\hat{L}_B(\theta) = \frac{1}{B}\sum_{b=1}^{B} \hat{L}_m(\theta; Z_b),$$

where $\{Z_b : b = 1, \ldots, B\}$ are iid samples from $\mathrm{N}(\tilde{z}, \tilde{\Omega}_{zz}^{-1})$ and

$$\hat{L}_m(\theta; Z) = H(\theta, Z)/\hat{L}_p(Z|y;\theta).$$

The LA is $\hat{L}_B(\theta)$ with $B = 1$ at $Z_b = \tilde{z}$. We call $\hat{L}_B(\theta)$ the *ELA* when $q(z)$ is the density function of $\mathrm{N}(\tilde{z}, \tilde{\Omega}_{zz}^{-1})$. In the Appendix,

we show that if the true predictive likelihood $L_p(z|y;\theta)$ is normal, then, for all $B \geq 1$

$$\hat{L}_B(\theta) = L_m(\theta). \tag{4.4}$$

If $\hat{L}_p(z|y;\theta)$ is close to the true $L_p(z|y;\theta)$, we expect that $\hat{L}_B(\theta)$ provides an accurate estimate of $L_m(\theta)$ for small values of B. As the LA provides an accurate approximation of $L_m(\theta)$, the use of $\mathrm{N}(\tilde{z}, \tilde{\Omega}_{zz}^{-1})$ as $q(z)$ is preferred. Burda et al. (2016) improved the variational method by exploiting the expression (1.1) of the marginal likelihood. The ELA further improves the variational method by using the alternative expression (1.2).

Theorem 4.2.1. Let $\hat{\ell}_B(\theta) = \log \hat{L}_B(\theta)$ and $\hat{\theta}_B^{\mathrm{ELA}} = \arg\max_\theta \hat{\ell}_B(\theta)$. Under regularity conditions R1 and R2, as $B \to \infty$,

$$\hat{\theta}_B^{\mathrm{ELA}} \xrightarrow{\mathrm{P}} \hat{\theta}.$$

Now, we study how to obtain a consistent estimator for the information matrix

$$I(\theta) = -\frac{\partial^2 \ell_m(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}.$$

Let $\hat{I}_B = I_B(\hat{\theta}_B^{\mathrm{ELA}})$, where

$$\begin{aligned} I_B(\theta) \;=\; & \left[\sum_{b=1}^{B}\left\{w(\theta, Z_b)\frac{\partial h(\theta, Z_b)}{\partial\theta}\right\}\right]\left[\sum_{b=1}^{B}\left\{w(\theta, Z_b)\left(\frac{\partial h(\theta, Z_b)}{\partial\theta}\right)^{\mathrm{T}}\right\}\right] \\ & -\sum_{b=1}^{B}\left[w(\theta, Z_b)\left\{\frac{\partial h(\theta, Z_b)}{\partial\theta}\left(\frac{\partial h(\theta, Z_b)}{\partial\theta}\right)^{\mathrm{T}} + \frac{\partial^2 h(\theta, Z_b)}{\partial\theta\partial\theta^{\mathrm{T}}}\right\}\right] \end{aligned}$$

and $w(\theta, Z_b) = \hat{L}_m(\theta, Z_b)/\sum_{t=1}^{B}\hat{L}_m(\theta, Z_t)$. Then, we have the following theorem.

Theorem 4.2.2. As $B \to \infty$, $\hat{I}_B \xrightarrow{\mathrm{P}} I(\hat{\theta})$.

According to Theorem 4.2.2, the variance of the MLE $\hat{\theta}$ can be consistently estimated by

$$\widehat{\text{var}}\left(\hat{\theta}\right) = \hat{I}_B^{-1}.$$

4.3 Restricted Likelihood

For cases in which τ and β are orthogonal, Cox and Reid (1987) proposed the use of an adjusted profile likelihood for the dispersion parameters τ based on the marginal likelihood $L_m(\theta)$:

$$\hat{R}(\tau) = L_m\left(\tau, \tilde{\beta}\right)\left|\frac{1}{2\pi}\tilde{\Omega}_{\beta\beta}\right|^{-\frac{1}{2}},$$

where $\tilde{\beta} = \tilde{\beta}(\tau) = \arg\max_\beta L_m(\beta, \tau)$ and $\tilde{\Omega}_{\beta\beta} = \{-\partial^2 \ell_m(\beta, \tau)/\partial\beta\partial\beta^{\mathrm{T}}\}|_{\beta=\tilde{\beta}}$. Barndorff-Nielsen (1987) noted that the Cox-Reid adjusted profile likelihood is the LA to the integrated likelihood

$$R(\tau) = \int L_m(\tau, \beta)d\beta = \hat{R}(\tau)(1 + O_p(n^{-1})).$$

Under the flat conditional prior $\pi(\beta|\tau) = 1$, Sweeting (1987) noted that the integrated likelihood becomes the marginal posterior density of τ:

$$R(\tau) = \int L_m(\tau, \beta)\pi(\beta|\tau)d\beta = \hat{R}(\tau)(1 + O_p(n^{-1})).$$

Barndorff-Nielsen (1983) derived the magic formula to determine $f_\tau(\hat{\tau}|\hat{\beta})$ for the MLEs $\hat{\theta} = (\hat{\beta}, \hat{\tau})$. Under the parameter orthogonality of τ and β, Cox and Reid (1987) showed that

$$f_\tau(\hat{\tau}|\hat{\beta}) = \hat{R}(\tau)(1 + O_p(n^{-1})).$$

Thus, we can view the Cox-Reid result as a case in which the conditional likelihood can be applied to eliminate nuisance fixed parameters. Note that

$$R(\tau) = f_\tau(\hat{\tau}|\hat{\beta})(1 + O_p(n^{-1})).$$

Thus, we propose to call, in this chapter, the integrated likelihood, namely the marginal posterior under $\pi(\beta|\tau) = 1$,

$$R(\tau) = \int L_m(\tau, \beta) d\beta = \int \int H(\tau, \beta, z) dz d\beta$$

the restricted likelihood. With the ELA, $R(\tau)$ can always be computed, as shown below, whereas $f_\tau(\hat{\tau}|\hat{\beta})$ is hardly available. The use of $R(\tau)$ does not require parameter orthogonality of Cox and Reid (1987), which would be hard to verify in general random effects models. From a frequentist perspective, the use of the integrated likelihood to eliminate the nuisance parameters has been examined for predicting unobserved latent variables z by Lee and Kim (2016).

When the marginal likelihood $\ell_m(\theta)$ is not available, Lee and Nelder (2001) proposed the use of the extended restricted likelihood

$$\hat{r}(\tau) = \log \hat{R}(\tau) = h(\tau, \tilde{\beta}, \tilde{z}) - \frac{1}{2} \log \left| \frac{1}{2\pi} \tilde{\Omega}_{\psi\psi} \right|,$$

where $\psi = (\beta, z)$, $\tilde{\psi} = \arg\max_\psi h(\beta, \tau, z)$ and $\tilde{\Omega}_{\psi\psi} = \{-\partial^2 h(\beta, \tau, z)/\partial\psi\partial\psi^{\mathrm{T}}\}|_{\psi=\tilde{\psi}}$. In this chapter, we refer to $\hat{r}(\tau) = \log \hat{R}(\tau)$ as the approximate restricted log-likelihood. Similar to (4.1), the restricted likelihood $R(\tau)$ can be approximated by using the approximate predictive likelihood $\hat{L}_p(\psi|y;\tau)$ based on a normal distribution

$$\psi \mid y \sim \mathrm{N}\left(\tilde{\psi}, \tilde{\Omega}_{\psi\psi}^{-1}\right).$$

Thus, Lee and Nelder's (2001) extended restricted likelihood $\hat{R}(\tau;\psi) = H(\tau,\psi)/\hat{L}_p(\psi|y;\tau)$ is the LA to $R(\tau)$. In normal linear mixed models, $R(\tau) = \hat{R}(\tau) = f_\tau(\hat{\tau}|\hat{\beta})$ becomes the restricted (or residual) likelihood of Patterson and Thompson (1971): see Chapter 5 of Lee et al. (2017).

We explore how to use the ELA to obtain the REMLE. Let

$$\hat{R}_B(\tau) = \frac{1}{B}\sum_{b=1}^{B} \hat{R}(\tau;\psi_b),$$

where $\{\psi_b : b = 1,\ldots,B\}$ are iid samples from $\mathrm{N}(\tilde{\psi}, \tilde{\Omega}^{-1}_{\psi\psi})$. Then, it is immediate that

$$\hat{r}_B(\tau) = \log \hat{R}_B(\tau) \xrightarrow{\mathrm{P}} r(\tau) = \log R(\tau)$$

as $B \to \infty$. Moreover, let $J(\tau) = -\partial^2 r(\tau)/\partial\tau\partial\tau^{\mathrm{T}}$ and $\hat{J}_B = J_B(\hat{\tau}_B^{\mathrm{ELA}})$, where

$$\begin{aligned}
\hat{\tau}_B^{\mathrm{ELA}} &= \arg\max_{\tau} \hat{R}_B(\tau), \\
J_B(\tau) &= \left[\sum_{b=1}^{B}\left\{\zeta(\tau,\psi_b)\frac{\partial h(\tau,\psi_b)}{\partial\tau}\right\}\right]\left[\sum_{b=1}^{B}\left\{\zeta(\tau,\psi_b)\left(\frac{\partial h(\tau,\psi_b)}{\partial\tau}\right)^{\mathrm{T}}\right\}\right] \\
&\quad -\sum_{b=1}^{B}\left[\zeta(\tau,\psi_b)\left\{\frac{\partial h(\tau,\psi_b)}{\partial\tau}\left(\frac{\partial h(\tau,\psi_b)}{\partial\tau}\right)^{\mathrm{T}} + \frac{\partial^2 h(\tau,\psi_b)}{\partial\tau\partial\tau^{\mathrm{T}}}\right\}\right],
\end{aligned}$$

and $\zeta(\tau,\psi_b) = \hat{R}_m(\tau,\psi_b)/\sum_{t=1}^{B}\hat{R}_m(\tau,\psi_t)$. Then, we have the following theorem.

Theorem 4.3.1. Let $\hat{\tau} = \arg\max_\tau r(\tau)$ be the REMLE of τ. As $B \to \infty$,

(i) $\hat{\tau}_B^{\mathrm{ELA}} \xrightarrow{\mathrm{P}} \hat{\tau}$,

(ii) $\hat{J}_B \xrightarrow{\mathrm{P}} J(\hat{\tau})$.

Thus, the variance estimator of the REMLE $\hat{\tau}$ can be consistently estimated by $\widehat{\text{var}}(\hat{\tau}) = \hat{J}_B^{-1}$. The second-order LA is applicable to only a limited class of models; for example, it cannot be applied to models with correlated random effects. The current version of the second-order LA in the dhglm in R (Lee and Noh, 2018) allows only crossed models with two independent random effects. However, the ELA is applicable to any statistical models with latent variables, as illustrated below.

4.4 Salamander Mating Data

In this chapter, we investigate how to obtain the frequentist MLE and REMLE, as well as their variance estimators. From a Bayesian perspective, the MLE and its variance estimator for $\theta = (\beta, \tau)$ are the posterior mode and its variance under a flat prior on θ, whereas the REMLE and its variance estimator for τ are the marginal posterior mode and its variance under a flat conditional prior on $\beta|\tau$. Here, we investigate the performance of the MLE, REMLE, and their variance estimators, based on the ELA, through numerical studies.

McCullagh and Nelder (1989) presented the salamander mating data. Three experiments were conducted to collect these data: two experiments were performed with the same salamanders in the summer and fall of 1986, and the third experiment was conducted in the fall of the same year using different salamanders. The salamander data are difficult to analyse as crossed models are required for binary data with correlated random effects. The Gauss-Hermite quadrature cannot be used due to the large value of d. Here, we

use the ELA for the analysis. We use simulation studies with $T = 200$ replications to evaluate the performance of various methods based on the following quantities: (i) Est: $\bar{\theta} = \sum_{t=1}^{T} \hat{\theta}^{(t)}/T$, (ii) SE: $\sum_{t=1}^{T} \widehat{\text{s.e.}}(\hat{\theta}^{(t)})/T$ and (iii) SD: $\{\sum_{t=1}^{T}(\hat{\theta}^{(t)} - \bar{\theta})^2/(T-1)\}^{1/2}$, where $\hat{\theta}^{(t)}$ is an estimate at the tth replication. To evaluate the performance of the point estimation, we compare the Est and true value of the fixed parameters. The similarity between the SE and the SD indicates the performance of the variance estimation.

4.4.1 Summer Data

Shun (1997) and Noh and Lee (2007) investigated the data that were collected during the summer to show how the second-order LA can be applied. The authors fitted a crossed model with $d = O(n^{1/2})$. For $i = 1, \ldots, I = 20$ and $j = 1, \ldots, J = 20$, let $y_{ij} \in \{0, 1\}$ be the binary outcome that indicates whether mating was successful for the ith female and the jth male. Each female was paired with six males for mating, generating in 120 observations. The authors considered the following random effects model:

$$\text{logit}\, \text{P}\left(y_{ij} = 1 \mid z_i^f, z_j^m\right) = x_{ij}^{\text{T}}\beta + \sigma_f z_i^f + \sigma_m z_j^m,$$

where $z_i^f \sim \text{N}(0, 1)$ and $z_j^m \sim \text{N}(0, 1)$ are female random effects and male random effects, respectively, which are assumed to be independent of each other. The covariates x_{ij} include an intercept, the main effects Trtf and Trtm, and their interaction Trtf·Trtm, where Trtf (Trtm) = 0, 1 for Rough Butt salamanders and Whiteside salamanders, respectively.

The simulation results are presented in Table 4.1. Here $\hat{\ell}_m$ ($\hat{\ell}_m^s$) represents the approximate MLE and $\hat{r}$ ($\hat{r}^s$) represents the approx-

Table 4.1: Simulation results for the summer data.

Method	Intercept	Trtf	Trtm	Trtf×Trtm	σ_f	σ_m
True value	1.06	-3.05	-0.72	3.77	1.22	1.22
MQL	0.78	-2.36	-0.51	2.87	0.86	0.88
PQL	0.85	-2.51	-0.57	3.05	0.94	0.96
CPQL	1.25	-3.48	-0.90	4.33	1.09	1.04
D&M	1.09	-3.15	-0.83	4.04	1.29	1.32
$\hat{\ell}_m$	0.93	-2.82	-0.60	3.21	1.04	1.00
$\hat{\ell}_m^s$	0.98	-2.94	-0.63	3.64	1.19	1.20
$\hat{r}$	1.15	-3.21	-0.79	3.82	1.26	1.27
SE ($\hat{r}$)	0.83	1.08	0.96	1.12	0.34	0.35
SD ($\hat{r}$)	0.97	1.54	0.92	1.54	0.61	0.69
$\hat{r}^s$	1.05	-3.02	-0.69	3.72	1.23	1.24
SE ($\hat{r}^s$)	0.70	0.90	0.83	0.97	0.30	0.29
SD ($\hat{r}^s$)	0.62	0.87	0.66	0.92	0.48	0.49
$\hat{r}_2$	1.11	-3.11	-0.84	3.85	1.11	1.18
$\hat{r}_{10}$	0.99	-3.09	-0.73	3.78	1.25	1.25
$\hat{r}_{50}$	1.07	-3.02	-0.72	3.77	1.21	1.23
SE ($\hat{r}_{50}$)	0.48	0.75	0.65	0.96	0.27	0.28
SD ($\hat{r}_{50}$)	0.51	0.80	0.57	0.89	0.38	0.42

imate REMLE calculated using the first-order (second-order) LA. $\hat{r}$ and $\hat{r}^s$ are the HL(1,1) and HL(2,2), respectively, of Noh and Lee (2007) with the approximate MLE of β and the approximate REMLE of τ maximizing $\hat{\ell}_m$ ($\hat{\ell}^s_m$) and $\hat{r}$ ($\hat{r}^s$), respectively. The authors also examined the performance of the penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) methods of Breslow and Clayton (1993) and Drum and McCullagh's (1993) method (D&M). Note that the PQL method has large biases in estimating the dispersion parameters (Lee and Nelder, 1996; Noh and Lee, 2007). Breslow and Lin (1995) derived a correction factor for the PQL (CPQL) to remove the asymptotic bias. Noh and Lee (2007) noted that the approximate REMLE $\hat{r}^s$, based on the second-order LA, produced the least bias in estimating θ among the existing methods at the time. Table 4.1 shows that the REMLEs $\hat{r}$ and $\hat{r}^s$ perform better than the MLEs $\hat{\ell}_m$ and $\hat{\ell}^s_m$. $\hat{r}_B$ is the ELA estimation based on B random samples, where the MLE of β and the REMLE of τ maximize $\hat{\ell}_B$ and $\hat{r}_B$, respectively. $\hat{r}_B$ with $B \geq 10$ improves the approximate REMLE $\hat{r}$ based on the first-order LA and $\hat{r}_{50}$ improves the approximate REMLE $\hat{r}^s$ based on the second-order LA. The ELA is considerably easier to implement than $\hat{r}^s$. To evaluate the performance of variance estimators, we compare $\hat{r}$, $\hat{r}^s$, and $\hat{r}_B$. We observe that $\hat{r}$ underestimates the SD. The SE of $\hat{r}^s$ and $\hat{r}_{50}$ well estimate the SDs of the mean parameters; however, for σ_f and σ_m, both $\hat{r}^s$ and $\hat{r}_{50}$ underestimate the SD. This underestimation of the ELA vanishes as n increases, as discussed below.

4.4.2 Pooled Data

For the pooled data from the three experiments, for which $k = 1, 2, 3$, Karim and Zeger (1992) considered the following model:

$$\text{logit}\left\{\text{P}\left(y_{ijk} = 1 \mid z_i^f, z_j^m\right)\right\} = x_{ijk}^{\text{T}}\beta + \Sigma_{f,k}^{1/2} z_i^f + \Sigma_{m,k}^{1/2} z_j^m,$$

where $z_i^f = (z_{i1}^f, z_{i2}^f, z_{i3}^f)^{\text{T}} \sim \text{N}(0, I)$ and $z_j^m = (z_{j1}^m, z_{j2}^m, z_{j3}^m)^{\text{T}} \sim \text{N}(0, I)$ are independent,

$$\Sigma_f = \begin{pmatrix} \sigma_{f_1}^2 & \rho_f \sigma_{f_1} \sigma_{f_2} & 0 \\ \rho_f \sigma_{f_1} \sigma_{f_2} & \sigma_{f_2}^2 & 0 \\ 0 & 0 & \sigma_{f_2}^2 \end{pmatrix}, \; \Sigma_m = \begin{pmatrix} \sigma_{m_1}^2 & \rho_m \sigma_{m_1} \sigma_{m_2} & 0 \\ \rho_m \sigma_{m_1} \sigma_{m_2} & \sigma_{m_2}^2 & 0 \\ 0 & 0 & \sigma_{m_2}^2 \end{pmatrix},$$

and $\Sigma_{f,k}^{1/2}$ and $\Sigma_{m,k}^{1/2}$ are the kth rows of $\Sigma_f^{1/2}$ and $\Sigma_m^{1/2}$, respectively. Here, $\Sigma_{f,k}^{1/2} z_i^f$ and $\Sigma_{m,k}^{1/2} z_j^m$ with $k = 1, 2$ represent correlated random effects. For the pooled data, an additional covariate indicating the season (0=summer and 1=fall) is included. In terms of the dispersion parameters, $\sigma_{f_1}^2$ ($\sigma_{m_1}^2$) is the variance in the summer and $\sigma_{f_2}^2$ ($\sigma_{m_2}^2$) is the variance in the fall for female (male) salamanders. Moreover, ρ_f (ρ_m) describes the correlation resulting from the same salamander being selected in the first two experiments. The second-order LA cannot be applied since the random effects are correlated. Among frequentist methods, for correlated random effects models, the PQL of Breslow and Clayton (1993) and $\hat{r}$ of Lee and Nelder (2001) can be applied. Breslow and Clayton (1993) applied the PQL method under the constraints $\sigma_{m_1} = \sigma_{m_2}$ and $\rho_m = 1$. Karim and Zeger (1992) used the Gibbs sampler to analyse the results from a Bayesian perspective.

Table 4.2 shows the estimation results for the pooled data obtained by various methods. It is well known that the PQL has

Table 4.2: Estimates of θ for the pooled data. The values in the parentheses are the estimated standard errors.

Method	β_0	β_1	β_2	β_3	β_4	σ_{f_1}	σ_{f_2}	ρ_f	σ_{m_1}	σ_{m_2}	ρ_m
Gibbs	1.48	-0.62	-3.13	-0.76	3.90	1.39	1.17	-0.15	1.12	1.42	0.96
	(0.64)	(0.54)	(0.62)	(0.62)	(0.72)						
PQL	1.18	-0.50	-2.43	-0.62	3.01	1.04	0.79	-0.15	0.95	0.95	1
	(0.49)	(0.41)	(0.44)	(0.46)	(0.52)						
$\hat{r}$	1.53	-0.63	-3.23	-0.79	4.02	1.49	1.12	-0.05	0.90	1.44	0.72
	(0.58)	(0.53)	(0.56)	(0.53)	(0.59)	(0.38)	(0.37)	(0.24)	(0.43)	(0.33)	(0.16)
$\hat{r}_{50}$	1.50	-0.63	-3.16	-0.76	3.90	1.46	1.12	-0.13	0.95	1.40	1.00
	(0.60)	(0.51)	(0.56)	(0.57)	(0.61)	(0.46)	(0.31)	(0.38)	(0.37)	(0.34)	(0.02)

large bias in binary data. For the ELA, we set $B = 50$ for the point estimation and $B = 1000$ for the standard error estimation. The approximate REMLE calculated using $\hat{r}$ differs from the true REMLE calculated using the ELA $\hat{r}_{50}$ when estimating ρ_m. The Gibbs sampler uses a flat prior for the mean parameters β and noninformative priors $\pi(\Sigma_f) \propto |\Sigma_f|^{-2}$ and $\pi(\Sigma_m) \propto |\Sigma_m|^{-2}$ for the dispersion parameters. This approach gives results similar to $\hat{r}_{50}$, which are marginal posterior modes under flat priors. For the hypotheses

$$H_0 : \ \rho_m = 1, \ H_1 : \ \rho_m \neq 1,$$

the ELA gives the likelihood ratio test $2\{\hat{\ell}_{50}(\hat{\theta}) - \hat{\ell}_{50}(\hat{\theta}^0)\} = 0.1022$, where $\hat{\theta}^0$ is the REMLE under the null hypothesis. Thus, we cannot reject H_0. This result indicates why the estimates of ρ_m are often close to 1 in Table 4.2. Thus, we consider a submodel with a shared random effects model in which $z_{j2}^m = \gamma_m z_{j1}^m$ for some γ_m.

Table 4.3 shows that the estimation performance of the ELA is better than that of $\hat{r}$ for all θ. In particular, $\hat{r}$ severely underestimates the standard errors. The ELA improves the point estimation and the standard error estimation. As shown in Tables 4.1 and 4.3, the SE obtains better estimates of the SD for the pooled data with $n = 360$ than for the summer data with $n = 120$. This result implies that the ELA provides consistent standard error estimators for the REMLEs.

4.5 Rongelap Spatial Data

Diggle et al. (1998) presented the Rongelap data, available at the geoRglm in R (Christensen and Ribeiro Jr, 2017), which were

Table 4.3: Simulation results for the pooled data.

		β_0	β_1	β_2	β_3	β_4	σ_{f_1}	σ_{f_2}	ρ_f	σ_m	γ_m
True value		1.50	-0.65	-3.20	-0.75	3.90	1.45	1.10	-0.15	1.00	1.50
$\hat{r}$	Est	1.69	-0.73	-3.48	-0.92	4.28	1.68	1.31	-0.12	1.24	1.58
	SE	0.65	0.58	0.59	0.58	0.63	0.37	0.35	0.23	0.37	0.59
	SD	0.75	0.67	0.72	0.67	0.84	0.61	0.39	0.18	0.49	0.73
$\hat{r}_{50}$	Est	1.53	-0.75	-3.10	-0.72	3.80	1.50	1.26	-0.14	1.06	1.55
	SE	0.61	0.55	0.56	0.58	0.62	0.52	0.37	0.44	0.45	0.80
	SD	0.64	0.52	0.52	0.60	0.60	0.43	0.30	0.41	0.40	0.75

obtained by the Marshall Islands National Radiological Survey, to determine whether Rongelap Island is safe with respect to radionuclide contamination. The data include gamma-ray counts y_i of radionuclide concentrations over time t_i at the spatial location s_i for $i = 1, \ldots, n = 157$ different locations on Rongelap Island. Diggle et al. (1998) considered the following Poisson random effects model:

$$y_i \mid z \sim \text{Poi}\left(t_i \lambda_i\right), \ \log \lambda_i = \beta_0 + \Sigma_i^{1/2} z, \tag{4.5}$$

where $z = (z_1, \ldots, z_n)^{\mathrm{T}} \sim \mathrm{N}(0, I)$, $\Sigma_i^{1/2}$ is the ith row of $\Sigma^{1/2}$ and the (i, j)th element of Σ is

$$\Sigma_{ij} = \exp\left\{\phi - \exp(\alpha)\|s_i - s_j\|_2\right\}, \tag{4.6}$$

where $\|s_i - s_j\|_2$ is the distance between the ith location and the jth location.

The integrated nested Laplace approximation (INLA) in R (Rue et al., 2009) is a widely used Bayesian procedure for fitting spatial data. Given the prior $\pi(\theta)$, the INLA approximates the posterior $\pi(\theta|y) \propto L_m(\theta)\pi(\theta)$ as $\hat{\pi}(\theta|y) \propto \hat{L}_m(\theta)\pi(\theta)$ based on the LA. Then, the INLA uses the approximate elementwise marginal posteriors

$$\hat{\pi}(\theta_k \mid y) = \int \hat{\pi}(\theta \mid y) d\theta_{-k}, \tag{4.7}$$

where $\theta_{-k} = (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots)$. Instead of (4.6), the INLA uses the following parametrization:

$$\Sigma_{ij} = \exp\left\{-\log 2\pi - \alpha - 2\xi - \exp(\alpha)\|s_i - s_j\|_2\right\}, \tag{4.8}$$

where $\phi = -\log 2\pi - \alpha - 2\xi$. The covariance model (4.6) is referred to as an exponential covariance function, whereas model (4.8) is

the Matérn covariance function, which is adopted by the INLA (Moraga, 2019). Under Gaussian priors for β_0, ξ, and α, the INLA provides the mean, mode, and standard deviations using random samples from the marginal posterior (4.7).

Although the responses are counts and thus not binary, since $d = n$, the LA may not be suitable. In addition, the second-order LA cannot be used due to the correlated random effects. We fitted the original Poisson random effects model (4.5), but it showed a severe lack-of-fit, with a scaled deviance of 6.466 for 0.717 degrees of freedom. If there is no lack-of-fit, the scaled deviance follows the chi-squared distribution with computed degrees of freedom. Bivand et al. (2015) proposed the overdispersed Poisson model for $y_i|z$:

$$c_i \mid z \sim \text{Poi}(\lambda_i),\ \log \lambda_i = \beta_0 + \Sigma_i^{1/2} z, \tag{4.9}$$

where $c_i = y_i/t_i$. The authors fitted the model (4.9) by using the INLA. Note that under the model (4.9), we have an overdispersed Poisson random effects model with $\text{E}(y_i|z) = t_i\lambda_i = \mu_i$, $\text{var}(y_i|z) = t_i^2\lambda_i = t_i\mu_i > \mu_i$ and overdispersion parameters $t_i > 1$. Lee et al. (2017) showed that the use of the model (4.9) is equivalent to the use of the extended quasi-likelihood (Lee and Nelder, 2000) for fitting an overdispersed Poisson model with $y_i|z$. The overdispersed Poisson model (4.9) has a scaled deviance of 120.1 with 146.9 degrees of freedom, confirming no lack-of-fit. Thus, the overdispersed Poisson model (4.9) achieves a better fit than the original Poisson model (4.5).

For the ELA, $B = 200$ is selected to fit β_0, $B = 1000$ is selected to fit τ and $B = 2000$ is selected to estimate the standard error.

Table 4.4: Estimates of the parameters according to the Rongelap data under the model (11). The values in parentheses are the estimated standard errors.

Method	β_0	ϕ	α	ξ
$\hat{r}$	1.966 (0.129)	-3.051 (0.355)	-2.708 (0.827)	1.961 (0.203)
$\hat{r}_B$	1.983 (0.102)	-3.325 (0.932)	-2.489 (1.424)	1.988 (0.724)
$INLA$	2.005 (0.116)	·	-1.822 (0.722)	1.886 (0.524)
$INLA^*$	1.990 (0.436)	·	-1.674 (0.722)	1.770 (0.524)

The estimation results of the Rongelap data with model (4.9) are presented in Table 4.4. For the point estimates, we consider both the posterior mean ($INLA$) and posterior mode ($INLA^*$) of the INLA output. The INLA provides a posterior standard deviation (PSD) for samples from the marginal posterior distribution as a standard error estimation. Since the Bayesian approach is not invariant with respect to the transformation of parameters, we do not report on ϕ for the INLA. However, ML estimation is invariant with respect to transformation; thus, we present the ELA result of ξ obtained by using the delta method. The REMLEs calculated by the ELA are marginal posterior modes under flat priors; thus, the difference between the ELA and the INLA would be caused by the use of different priors, although these differences are not significant.

We perform a simulation study with model (4.9). To reduce the complexity of using the extended quasi-likelihood method, we use a Poisson random effects model by setting $t_i = 1$. According to Table 4.5, the point estimates of β_0 are similar for all the eval-

Table 4.5: Simulation results for the Rongelap data.

		β_0	ϕ	α	ξ
True value		1.980	-3.000	0.100	0.531
$\hat{r}$	Est	1.976	-3.023	0.178	0.504
	SE	0.050	0.341	0.534	0.318
	SD	0.048	0.416	0.688	0.430
$\hat{r}_B$	Est	1.977	-3.014	0.119	0.528
	SE	0.051	0.476	0.740	0.442
	SD	0.049	0.437	0.728	0.444
$INLA$	Est	1.986	·	0.051	0.673
	PSD	0.087	·	0.681	0.602
	SD	0.051	·	0.675	0.598
$INLA^*$	Est	1.988	·	0.037	0.632
	PSD	0.087	·	0.681	0.602
	SD	0.051	·	0.627	0.595

uated methods. In terms of the standard error estimates, the LA $\hat{r}$ underestimates the SD of the estimators. The ELA provides accurate REMLEs. We report the INLA results to highlight the differences caused by the use of different priors. The INLA computes the PSDs using samples from the marginal posteriors, whereas the standard error estimates of the REMLEs are computed using the Hessian matrix without resampling. In summary, different priors could yield different dispersion parameter estimates.

4.6 Conclusion

The LA and the variational Bayes method have been proposed as methods for approximating the marginal likelihood. However, resulting approximate MLEs and REMLEs could be often biased for binary or spatial data. Furthermore, a consistent variance estimation method is not available. With the ELA, the MLE, REMLE, and their consistent variance estimators can be obtained in general for statistical models with unobserved latent variables. The results of numerical studies confirm that the ELA provides satisfactory MLE and REMLE for a wide variety of models. Furthermore, the MLE and REMLE are Bayesian posterior modes and marginal posterior modes, respectively, under flat priors. Thus, we can have both frequentist and Bayesian interpretations from ML and REML analyses.

Appendix: Proofs

Proof of (4.4)

Proof. Suppose that the true predictive likelihood $L_p(z|y;\theta)$ is from a normal distribution. Let m and S be mean and covariance matrix of normal distribution of which predictive log-likelihood is

$$\ell_p\left(z \mid y;\theta\right) = -\frac{1}{2}\log|2\pi S| - \frac{1}{2}(z-m)^{\mathrm{T}} S^{-1}(z-m).$$

Then, $\tilde{z} = m$ and $\tilde{\Omega}_{zz} = S^{-1}$ since

$$\begin{aligned} \frac{\partial}{\partial z}\ell_p\left(z \mid y;\theta\right) &= -S^{-1}(z-m), \\ \frac{\partial}{\partial z \partial z^{\mathrm{T}}}\ell_p\left(z \mid y;\theta\right) &= -S^{-1}. \end{aligned}$$

Thus, we have $\hat{L}_p(z|y;\theta) = L_p(z|y;\theta)$ for all z which gives $\hat{L}_m(\theta) = L_m(\theta)$. Moreover,

$$\hat{L}_B(\theta) = \frac{1}{B}\sum_{b=1}^{B}\frac{H(\theta, Z_b)}{\hat{L}_p(Z_b \mid y;\theta)} = \frac{1}{B}\sum_{b=1}^{B}\frac{H(\theta, Z_b)}{L_p(Z_b \mid y;\theta)} = L_m(\theta),$$

for all $B \geq 1$. □

Proof of Theorem 4.2.1

Proof. Note that there exists a constant $M > 0$ such that

$$\hat{L}_m(\theta; Z) = \frac{H(\theta, Z)}{\hat{L}_p(Z \mid y;\theta)} \leq \frac{H(\theta, \tilde{z})}{\hat{L}_p(Z \mid y;\theta)} \leq M \tag{4.10}$$

with probability one, i.e., $\hat{L}_m(\theta; Z)$ is bounded with probability one. By the law of large numbers, we have

$$\hat{L}_B(\theta) = \frac{1}{B}\sum_{b=1}^{B}\frac{H(\theta, Z_b)}{\hat{L}_p(Z_b \mid y;\theta)} \xrightarrow{\mathrm{P}} \int \frac{H(\theta, z)}{\hat{L}_p(z \mid y;\theta)}\hat{L}_p(z \mid y;\theta)dz = L_m(\theta)$$

as $B \to \infty$ for all θ. Then, from the Theorem 2.7 of Newey and McFadden (1994), we can conclude that $\hat{\theta}_B^{\mathrm{ELA}} \xrightarrow{\mathrm{P}} \theta$. □

Proof of Theorem 4.2.2

Proof. Note that the Hessian matrix of the marginal log-likelihood can be expressed as

$$\frac{\partial^2 \ell_m(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} = -\left\{\frac{1}{L_m(\theta)}\frac{\partial L_m(\theta)}{\partial\theta}\right\}\left\{\frac{1}{L_m(\theta)}\left(\frac{\partial L_m(\theta)}{\partial\theta}\right)^{\mathrm{T}}\right\} + \frac{1}{L_m(\theta)}\frac{\partial^2 L_m(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}. \tag{4.11}$$

By introducing an arbitrary density function $q(z)$, we have

$$\begin{aligned}
\frac{\partial L_m(\theta)}{\partial\theta} &= \int \frac{\partial h(\theta,z)}{\partial\theta}\frac{H(\theta,z)}{q(z)}q(z)dz, \\
\frac{\partial^2 L_m(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} &= \int \left\{\frac{\partial h(\theta,z)}{\partial\theta}\left(\frac{\partial h(\theta,z)}{\partial\theta}\right)^{\mathrm{T}} + \frac{\partial^2 h(\theta,z)}{\partial\theta\partial\theta^{\mathrm{T}}}\right\}\frac{H(\theta,z)}{q(z)}q(z)dz.
\end{aligned}$$

Recall that

$$\frac{\partial \ell_m(\theta)}{\partial\theta} = \frac{1}{L_m(\theta)}\frac{\partial L_m(\theta)}{\partial\theta} = \frac{1}{L_m(\theta)}\int \frac{\partial h(\theta,z)}{\partial\theta}\hat{L}_m(\theta,z)\hat{L}_p(z \mid y;\theta)dz \tag{4.12}$$

and $\hat{L}_B(\theta) \xrightarrow{\mathrm{P}} L_m(\theta)$ as $B \to \infty$. By assumption of unimodality, there exists $\{\hat{\theta}_B^{\mathrm{ELA}}, \hat{\theta}\} \in \Theta_1 \subset \Theta$ such that

$$\sup_{\theta\in\Theta_1}\left|\frac{\partial h(\theta,z)}{\partial\theta}\right| \le M_1, \ \sup_{\theta\in\Theta_1}\left|\frac{\partial^2 h(\theta,z)}{\partial\theta\partial\theta^{\mathrm{T}}}\right| \le M_2$$

given $M_1, M_2 > 0$ for all z. Moreover, $w(\theta, Z)$ is bounded provided by (4.10). Then,

$$\frac{1}{B}\sum_{b=1}^{B}\frac{\partial h(\theta,Z_b)}{\partial\theta}\hat{L}_m(\theta,Z_b) \xrightarrow{\mathrm{P}} \frac{\partial L_m(\theta)}{\partial\theta}. \tag{4.13}$$

By using the Slutsky's theorem, we have

$$\frac{\frac{1}{B}\sum_{b=1}^{B}\frac{\partial h(\theta,Z_b)}{\partial\theta}\hat{L}_m(\theta,Z_b)}{\frac{1}{B}\sum_{t=1}^{B}\hat{L}_m(\theta,Z_t)} = \sum_{b=1}^{B}\frac{\partial h(\theta,Z_b)}{\partial\theta}w(\theta,Z_b) \xrightarrow{\mathrm{P}} \frac{\partial \ell_m(\theta)}{\partial\theta}$$

as $B \to \infty$. Similar to (4.13), we also have

$$\frac{1}{B}\sum_{b=1}^{B}\left\{\frac{\partial h(\theta, Z_b)}{\partial \theta}\left(\frac{\partial h(\theta, Z_b)}{\partial \theta}\right)^{\mathrm{T}} + \frac{\partial^2 h(\theta, Z_b)}{\partial \theta \partial \theta^{\mathrm{T}}}\right\}\hat{L}_m(\theta, Z_b) \xrightarrow{\mathrm{P}} \frac{\partial^2 L_m(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}$$

which implies

$$\sum_{b=1}^{B}\left\{\frac{\partial h(\theta, Z_b)}{\partial \theta}\left(\frac{\partial h(\theta, Z_b)}{\partial \theta}\right)^{\mathrm{T}} + \frac{\partial^2 h(\theta, Z_b)}{\partial \theta \partial \theta^{\mathrm{T}}}\right\} w(\theta, Z_b) \xrightarrow{\mathrm{P}} \frac{1}{L_m(\theta)}\frac{\partial^2 L_m(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}. \tag{4.14}$$

By combining (4.13) and (4.14), we have

$$I_B(\theta) \xrightarrow{\mathrm{P}} I(\theta) = -\frac{\partial^2 \ell_m(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}$$

as $B \to \infty$ for $\theta \in \Theta_1$. By definition, Θ_1 contains $\hat{\theta}_B^{\mathrm{ELA}}$ and $\hat{\theta}$. Also, $\hat{\theta}_B^{\mathrm{ELA}}$ converges to $\hat{\theta}$ as shown in Theorem 4.2.1. In conclusion, $\hat{I}_B$ converges to $\hat{I}$ as $B \to \infty$ which proves the Theorem 4.2.2. □

Chapter 5

AFT Random Effect Model with GEV Distribution

Chapter Summary

Generalized extreme value (GEV) distribution is widely used for analyzing extreme events. For analyzing heavily censored data we suggest the use of GEV distribution by treating uncensored observations as extreme events. We are interested in the analysis of heavily censored clustered survival data. The correlation among clustered survival times can be modeled via random effects. In this chapter, we propose the use of an accelerated failure time (AFT) random effect model with GEV distribution to directly describe the relationship between survival time and covariates. The performance of the proposed method is evaluated via simulation study, which shows that the estimated regression parameters are robust even when not only data are heavily censored but also distributional assumption on the error distribution is violated. The proposed method is illustrated with a real data example.

5.1 Model

5.1.1 GEV Distribution

A random variable Y is said to be distributed as GEV(μ, σ, ζ) distribution if

$$\mathrm{P}\left(Y \leq y\right) = \exp\left\{-M(y)\right\},$$

where

$$M(y) = \left\{1 + \zeta\left(\frac{y-\mu}{\sigma}\right)\right\}_{+}^{-\frac{1}{\zeta}}.$$

Here, $\mu \in \mathbb{R}$, $\sigma \in (0, \infty)$, and $\zeta \in \mathbb{R}$ are location, scale, and shape parameters, respectively, and $a_+ = \max(0, a)$. By permitting three parameters, the GEV distribution is useful to model skewed, heavy-tailed, and heavily censored data Bladt and Albrecher (2021). In particular, the shape parameter ζ controls the tail behavior of the GEV distribution Roy and Dey (2014). Special cases of the GEV are the Gumbel, Fréchet, and reversed Weibull distribution by taking $\zeta = 0$, $\zeta > 0$, and $\zeta < 0$, respectively. Here, the case $\zeta = 0$ is interpreted as $\zeta \to 0$. Note that the GEV distribution belongs to the location family, i.e.,

$$Y \sim \mathrm{GEV}(0, \sigma, \zeta) \;\Leftrightarrow\; Y + \mu \sim \mathrm{GEV}(\mu, \sigma, \zeta), \quad \text{for all} \;\; \mu \in \mathbb{R}.$$

5.1.2 AFT Random Effect Model with GEV Distribution

Consider the clustered survival data, where the size of each cluster or subject can be different. Let T_{ij} be survival time (i.e., time-to-event) for the jth observation of the ith subject (or cluster) and let C_{ij} be the corresponding censoring time ($i = 1, \ldots, q$;

$j = 1, \ldots, n_i$). Here, q is the number of clusters, n_i is the number of individuals in the ith cluster (i.e. cluster size), and $n = \sum_{i=1}^{q} n_i$ is the total sample size. In multi-center clinical trials, n_i is the number of patients in the ith center and n is the total number of patients coming from all q centers. Similarly, in the dental study n_i is the number of existing teeth in a mouth of the ith subject and n is the total number of teeth of all q subjects. Typically, a correlation among T_{ij}'s can be induced by the clustering. In the bivariate data, $n_i = 2$ for all i. Note that random effects are useful for modelling such dependence among T_{ij}'s.

The proposed AFT random effect model with GEV distribution can be written as

$$\log(T_{ij}) = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \tag{5.1}$$

where $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$ is a $p \times 1$ vector of covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients corresponding to $\boldsymbol{x}_{ij}$, $v_i \sim \mathrm{N}(0, \alpha)$ is a common random effect representing the unobserved subject effect of the ith subject, $\varepsilon_{ij} \sim \mathrm{GEV}(0, \sigma, \zeta)$ and all these random quantities are independent. The model 5.1 is an extension of GEV-AFT model Roy and Dey (2014) to the random effects model which can be viewed as a linear mixed model under the log-transformation of survival time T_{ij}. Here, we make the following two assumptions (Ha et al., 2002):

> Assumption 1: Given v_i, the pairs (T_{ij}, C_{ij}) are conditionally independent for $j = 1, \ldots, n_i$ and T_{ij} and C_{ij} are also conditionally independent in each pair.
>
> Assumption 2: Given v_i, $\{C_{ij}, j = 1, \ldots, n_i\}$ are condition-

ally noninformative for T_{ij}.

The observed random variables for the AFT model are given by

$$Y_{ij} = \min\{\log\ T_{ij}, \log\ C_{ij}\} \quad \text{and} \quad \delta_{ij} = I(T_{ij} \leq C_{ij}),$$

where δ_{ij} is censoring indicator and $I(\cdot)$ denotes the indicator function. The h-likelihood (Ha et al., 2002) for the AFT model 5.1 under Assumptions 1 and 2 is defined as

$$h(\boldsymbol{\theta}, \boldsymbol{v}) = \sum_{i=1}^{q} h_i(\boldsymbol{\theta}, v_i) \ \text{ with } \ h_i(\boldsymbol{\theta}, v_i) = \sum_{j=1}^{n_i} \ell_{1ij} + \ell_{2i}, \qquad (5.2)$$

where

$$\begin{aligned} \ell_{1ij} &= \delta_{ij}\{-\log\sigma + (1+\zeta)\log M_{ij} - M_{ij}\} \\ &\quad +(1-\delta_{ij})\log\{1-\exp(-M_{ij})\}, \\ \ell_{2i} &= -\frac{1}{2}\log(2\pi\alpha) - \frac{1}{2\alpha}v_i^2. \end{aligned}$$

Here, ℓ_{1ij} is the logarithm of the conditional density function for (Y_{ij}, δ_{ij}) given v_i and ℓ_{2i} is the logarithm of the density function for v_i. Moreover, $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma, \zeta, \alpha)^{\mathrm{T}}$ is fixed parameters, $\mu_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + v_i$ is the linear predictor, and

$$M_{ij} = M(y_{ij}) = \left\{1 + \zeta\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right)\right\}_{+}^{-1/\zeta}.$$

5.2 Estimation Procedure

To obtain the maximum likelihood estimator of $\boldsymbol{\theta}$, we need to obtain the marginal likelihood $\ell(\boldsymbol{\theta})$, by integrating out the random effect $\boldsymbol{v}$,

$$\ell(\boldsymbol{\theta}) = \log \int_{\Omega_{\boldsymbol{v}}} \exp\{h(\boldsymbol{\theta}, \boldsymbol{v})\}\, d\boldsymbol{v},$$

where $\Omega_{\boldsymbol{v}} = \prod_{i=1}^{q} \Omega_{v_i}$ with $\Omega_{v_i} = \{v_i : \sigma + \zeta(y_{ij} - \mu_{ij}) > 0\}$ is the support of the random effect $\boldsymbol{v}$. However, obtaining the explicit form of the marginal likelihood $\ell(\boldsymbol{\theta})$ is generally intractable. Moreover, in AFT random effect model with GEV distribution, the support of the random effect depends on the fixed parameters $\boldsymbol{\theta}$. Instead, an adjusted profile h-likelihood (Ha et al., 2017) $p_{\boldsymbol{v}}(h)$ based on the Laplace approximation can be used to approximate $\ell(\boldsymbol{\theta})$ as follows:

$$\ell\left(\boldsymbol{\theta}\right) \approx p_{\boldsymbol{v}}\left(h\right) = h\left(\boldsymbol{\theta}, \tilde{\boldsymbol{v}}\right) - \frac{1}{2}\log\det\left(\frac{1}{2\pi}\boldsymbol{H}_{\boldsymbol{vv}}\right)_{\boldsymbol{v}=\tilde{\boldsymbol{v}}},$$

where $\boldsymbol{H}_{\boldsymbol{vv}} = -\partial^2 h(\boldsymbol{\theta}, \boldsymbol{v})/\partial \boldsymbol{v} \partial \boldsymbol{v}^{\mathrm{T}}$, and *logdet* is the logarithm of the determinant. Here,

$$\tilde{\boldsymbol{v}} = \tilde{\boldsymbol{v}}(\boldsymbol{\theta}) = \arg\max_{v} h(\boldsymbol{\theta}, \boldsymbol{v})$$

is the mode of the h-likelihood given $\boldsymbol{\theta}$. Compared to other random effect models, it is unusual that the support of the random effect $\Omega_{\boldsymbol{v}}$ depends on the fixed parameters $\boldsymbol{\theta}$. However, the Laplace approximation evaluates the integral at the near of the mode. Therefore, we can still make inference about $\boldsymbol{\theta}$ by using $p_{\boldsymbol{v}}(h)$. Moreover, to obtain accurate estimates of the fixed parameters, we decompose the whole fixed parameters $\boldsymbol{\theta}$ into two parts: regression coefficients $\boldsymbol{\beta}$ and dispersion parameters $\boldsymbol{\phi} = (\sigma, \zeta, \alpha)^{\mathrm{T}}$. Then, $\boldsymbol{\beta}$ is estimated (Ha et al., 2017) from $p_{\boldsymbol{v}}(h)$ and $\boldsymbol{\phi}$ from $p_{\boldsymbol{\psi}}(h)$ given by

$$p_{\boldsymbol{\beta}}(\ell) \approx p_{\boldsymbol{\psi}}(h) = h(\boldsymbol{\phi}, \tilde{\boldsymbol{\psi}}) - \frac{1}{2}\log\det\left(\frac{1}{2\pi}\boldsymbol{H}_{\boldsymbol{\psi\psi}}\right)_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}},$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{v}^{\mathrm{T}})^{\mathrm{T}}$, $\tilde{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} h(\boldsymbol{\phi}, \boldsymbol{\psi})$, and $\boldsymbol{H}_{\boldsymbol{\psi\psi}} = -\partial^2 h(\boldsymbol{\phi}, \boldsymbol{\psi})/\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^{\mathrm{T}}$. Here, $\hat{\boldsymbol{\phi}}$ obtained from $p_{\boldsymbol{\psi}}(h)$ is

called the restricted maximum likelihood (REML) estimator (Lee et al., 2017).

In summary, the fitting algorithm is given as follows:

Step 0: Set initial values $\hat{\boldsymbol{\theta}}^{(0)}$. Then, for $t = 1, 2, \ldots$, repeat **Steps** 1-3 until the maximum absolute difference between $\hat{\boldsymbol{\theta}}^{(t)}$ and $\hat{\boldsymbol{\theta}}^{(t-1)}$ is less than 10^{-4}.

Step 1: Compute $\hat{\boldsymbol{v}}^{(t)} = \arg\max_{\boldsymbol{v}} h(\boldsymbol{\theta}, \boldsymbol{v})$ given $\hat{\boldsymbol{\theta}}^{(t-1)}$.

Step 2: Compute $\hat{\boldsymbol{\beta}}^{(t)} = \arg\max_{\boldsymbol{\beta}} p_{\boldsymbol{v}}(h)$ given $\hat{\boldsymbol{v}}^{(t)}$ and $\hat{\boldsymbol{\phi}}^{(t-1)}$.

Step 3: Compute $\hat{\boldsymbol{\phi}}^{(t)} = \arg\max_{\phi} p_{\boldsymbol{\psi}}(h)$ given $\hat{\boldsymbol{\psi}}^{(t)}$.

After the convergence has occurred, the variance of $\hat{\boldsymbol{\beta}}$ can be estimated as

$$\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}\right) = \left(-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} p_{\boldsymbol{v}}(h)\right)^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \boldsymbol{v}=\hat{\boldsymbol{v}}},$$

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} p_{\boldsymbol{v}}(h) \approx \left[\frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} - \frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{v}^{\mathrm{T}}}\left(\frac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{v}^{\mathrm{T}}}\right)^{-1}\frac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right]_{\boldsymbol{v}=\hat{\boldsymbol{v}}},$$

where $\hat{\boldsymbol{v}} = \tilde{\boldsymbol{v}}(\hat{\boldsymbol{\theta}})$. For more details about the computation of h, $p_{\boldsymbol{v}}(h)$, and $p_{\boldsymbol{\psi}}(h)$, see Appendix A.

5.3 Simulation Study

The simulation study is conducted to evaluate the performance of the proposed method, based on 500 replications of simulated data. In particular, the robustness of the AFT random effect model with GEV distribution against the distributional assumption is

also studied by comparing the AFT random effect model with normal distribution. For $i = 1, \ldots, q$ and $j = 1, \ldots, n_i$, we generate

$$\log(T_{ij}) = \mu_{ij} + v_i + \varepsilon_{ij} \text{ with } \mu_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}, \quad (5.3)$$

where $(\beta_0, \beta_1, \beta_2) = (0.5, 1.5, -1.5)$, $x_{ij1}, \sim \mathrm{N}(0, 1)$, $x_{ij2} \sim$ Bernoulli(0.5), and $v_i \sim \mathrm{N}(0, 2)$. We consider sample sizes $(q, n_i) = (100, 15)$ and $(200, 15)$ for all i. The true errors ε_{ij} are generated from four distributions

(C1) GEV: $\varepsilon_{ij} \sim \mathrm{GEV}\,(0, \sigma = 1.2, \zeta = -0.7)$.

(C2) N: $\varepsilon_{ij} \sim \mathrm{N}\,(0, \lambda = 1.22)$.

(C3) T: $\varepsilon_{ij} \sim t$-distribution where degrees of freedom is 5.

(C4) LG: $\varepsilon_{ij} \sim \log\{\mathrm{Gamma}\,(1.5, 5)\}$.

In case C1, we investigate the performance of the proposed method when the distribution assumption is correct. C2 investigates the robustness of the proposed method by comparing with the normal error distribution. For robustness against the misspecification of distributional assumption, we consider a t-distribution (C3) as a heavy-tailed distribution and a LG distribution (C4) as a skewed distribution. Censoring times C_{ij} are generated from an uniform distribution with a parameter empirically determined to achieve the stated censoring rate, about 50% and 90%. We fit the following two models under (5.3),

$$\begin{aligned} &\text{GEV-AFT model } (\mathrm{M_{GEV}}) : \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{GEV}\,(0, \sigma, \zeta)\,, \\ &\text{Normal-AFT model } (\mathrm{M_N}) : \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{N}\,(0, \lambda)\,. \end{aligned}$$

From $B = 500$ replications of simulated data, the performance of the estimates of the regression coefficients $\hat{\beta}_j$, $j = 1, 2$ is evaluated by (i) closeness between the mean $\bar{\beta}_j = \sum_{b=1}^{B} \hat{\beta}_j^{(b)}/B$ and β_j^{true}, (ii) closeness between the mean of estimates of standard error (SE) of $\hat{\beta}_j$, $\widehat{\text{SE}}(\hat{\beta}_j)$, and the standard deviation (SD) of $\hat{\beta}_j$ defined by $\sum_{b=1}^{B}(\hat{\beta}_j^{(b)} - \bar{\beta}_j)^2/(B-1)$, and (iii) maintenance of the empirical coverage probability (CP) for a nominal 95% confidence interval for β_j; this observation stems from the fact that the estimated standard error is close enough to the standard deviation of the estimates. For dispersion parameters, the mean and standard deviations of estimates are also presented.

The simulation results are summarized in Table 5.1 and 5.2. When the true error distribution is GEV, the proposed GEV-AFT model overall works well in terms of biases of $\hat{\boldsymbol{\theta}}$. In addition, the estimated SE of $\hat{\beta}_j$ $(j = 1, 2)$ is close to the empirical SD, which is the estimate of $\{\text{var}(\hat{\beta}_j)\}^{1/2}$. When the censoring rate is high (90%), the proposed GEV-AFT model gives better agreement with the nominal value of 0.95 for CPs of $\boldsymbol{\beta}$ in all cases, compared to the Normal-AFT model.

We observe that the estimates $(\hat{\sigma}, \hat{\zeta}, \hat{\alpha})$ of all dispersion parameters are close to their true values even if the censoring rate is extremely high. As expected, we see that the biases and variations (SEs and SDs) tend to decrease as the sample size increases. On the other hand, the Normal-AFT model shows some underestimation for the absolute magnitude of β_j $(j = 1, 2)$ and variance estimation, leading to substantially lower CPs.

When the error distribution is normal, the proposed method

gives robust estimation results comparable to the normal AFT model. We see that the normal AFT model by Ha et al. (Ha et al., 2002) provides reasonable results under 50% censoring. This indicates that the Normal-AFT model leads to biased results when the censoring rate is high. Moreover, the estimation performance of the proposed method still shows robust results for the estimated regression parameters when the true error distribution is T or LG. As expected, the Normal-AFT model gives severely biased results, particularly for the variance of the random effect α under the skewed LG distribution.

In summary, the simulation results suggest that the proposed method is indeed reasonable and gives robust estimation results for the regression parameters in all cases compared to the results of normal AFT model.

5.4 Real Data Analysis: COHRI Data

In this section, we analyze the consortium for oral health-related informatics (COHRI) data which is highly censored, correlated survival data. The COHRI data consist of de-identified electronic dental records of $q = 5,336$ subjects (baseline patient age between 16-90) with about 6 years of follow-up, derived from the AxiUm database at Creighton University School of Dentistry Stark et al. (2010). This AxiUM consortium allowed for information exchange of medical and dental electronic records, primarily for research. Here, the survival time is the time until tooth-loss. The number of teeth per subject n_i varied from 1 to 30, with mean 12.35 and median 12. In particular, only about 7% (4,593 observations) among

$n = \sum_{i=1}^{5,336} n_i = 65,890$ observations were lost during the follow-up, clearly indicating that observations are heavily censored with about 93%.

We consider the 14 covariates of interest as follows:

- Mobility (0-5 scale),
- BOP; proportion of tooth-sites that bled when probed (%),
- Plaque; proportion of tooth-sites stained with bacterial plaque (%),
- PDmean; mean pocket depth for that tooth,
- CALmean; mean clinical attachment level for that tooth,
- Crown; tooth has crown (0 = yes, 1 = no),
- Filled; tooth has filled (0=yes, 1=no),
- Decayed; tooth has decayed (0=yes, 1=no),
- D.F.sites; the number of decayed and filled sites,
- Age (in years),
- Gender (0=female, 1=male),
- Diabetes (0=yes, 1=no),
- Tobacco; use of tobacco (0=yes, 1=no),
- Molar; inspected tooth is molar (0=no, 1=yes).

The mean (standard deviation) of age is 58.4 (18.1) years, with 49.5% males, 9.3% with diabetes, and 22.2% are smokers. Summary statistics of these covariates are presented in Table 5.3. For molars, 7.3% were lost during the follow-up, while for the non-molars, it is 6.8%, again implying that observations are heavily censored.

The results of fitting GEV and Normal AFT models using the h-likelihood are presented in Table 5.4. First, the GEV-AFT model shows that all covariates are significant at level 5%. For example, the tooth-loss time in the non-smoker group (Tobacco=no) is significantly increased by a factor of $\exp(1.148) = 3.15$, as compared to the smoker group (Tobacco=yes), while the loss time in the Molar group (Molar=yes) is significantly decreased by a factor of $\exp(-0.262) = 0.77$, as compared to the non-Molar group (Molar=no). The variance of random effect $\hat{\alpha} = 2.399$ is somewhat large, which account for a correlation among survival times. Next, we find that the Normal-AFT model also gives significant results except for four covariates (Plaque, Age, Gender, and Diabetes). The Normal-AFT model shows that most of the estimates of the regression coefficient are smaller, which confirms the underestimation from simulation results of Table 5.2. Moreover, the simulation results also indicate that the true distribution of the survival time until tooth-loss seems to be skewed. In Table 5.4, the Normal-AFT model has wider confidence intervals than the GEV-AFT model.

For the selection between the GEV-AFT and Normal-AFT models, two Akaike information criteria (AIC (Lee et al., 2017; Ha et al., 2017)) are considered, the marginal AIC (mAIC (Ha

et al., 2012)) and conditional AIC (cAIC (Vaida and Blanchard, 2005)). The mAIC and cAIC are, respectively, defined by

$$\begin{aligned} \text{mAIC} &= -2p_{\boldsymbol{v}}(h) + 2\,\text{df}_{\text{m}}, \\ \text{cAIC} &= -2\ell_1 + 2\,\text{df}_{\text{c}}\,. \end{aligned}$$

Here, $\ell_1 = \sum_{i,j} \ell_{1ij}$ is given in (5.2), df_{m} is the number of fixed parameters, and $\text{df}_{\text{c}} = \text{trace}(\boldsymbol{H}^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}\boldsymbol{H}^{*}_{\boldsymbol{\psi}\boldsymbol{\psi}})|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}},\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}$ with $\boldsymbol{H}^{*}_{\boldsymbol{\psi}\boldsymbol{\psi}} = -\partial^2\ell_1/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\text{T}}$. The mAIC selects a better marginal model between the two AFT models, whereas the cAIC selects a better subject-specific model. The value of smaller AIC indicates a better model. With the COHRI dataset, in the GEV-AFT model mAIC=47839.13 and cAIC=31861.8, and in the Normal-AFT model mAIC=48037.67 and cAIC=35138.24. Thus, both mAIC and cAIC indicate that the proposed gives better marginal and subject-specific models compared to the Normal-AFT model.

5.5 Conclusion

In this chapter, we propose the use of the AFT random effect model with GEV distribution to analyze heavily censored clustered data. Usually, it is assumed that the distribution of error term in the AFT model is normal distribution. However, in simulation study, we have shown that the assumption of normal distribution for error term does not give valid estimates when censoring rate is extremely high. We have also demonstrated via simulation and real data example that the proposed method gives robust estimation results for the model parameters even when the censoring rate is extremely high and distributional assumption of error term is

violated.

Appendix A: Computations for Estimation Procedures

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be a model matrix for the fixed effect $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{Z} = \text{diag}\left(\mathbb{1}_{n_1}, \cdots, \mathbb{1}_{n_q}\right) \in \mathbb{R}^{n\times q}$ be a model matrix for the random effect $\boldsymbol{v} \in \mathbb{R}^q$, $\boldsymbol{\delta} = \left(\delta_{11}, \cdots, \delta_{qn_q}\right)^{\mathrm{T}} \in \mathbb{R}^n$ be a vector of censoring indicators, and $\boldsymbol{M} = \left(M_{11}, \cdots, M_{qn_q}\right)^{\mathrm{T}} \in \mathbb{R}^n$. Here, $\mathbb{1}$ is a vector of ones with corresponding length. Then, the h-likelihood (5.2) can be written in matrix form as follows:

$$\begin{aligned} h(\boldsymbol{\theta}, \boldsymbol{v}) &= \boldsymbol{\delta}^{\mathrm{T}} \left(-\log\sigma \mathbb{1}_n + (1+\zeta)\log \boldsymbol{M} - \boldsymbol{M}\right) \\ &\quad + \left(\mathbb{1}_n - \boldsymbol{\delta}\right)^{\mathrm{T}} \log\left(\mathbb{1}_n - e^{-\boldsymbol{M}}\right) - \frac{q}{2}\log(2\pi\alpha) - \frac{1}{2\alpha}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}. \end{aligned}$$

Here, $\log \boldsymbol{M} = (\log M_{11}, \ldots, \log M_{qn_q})^{\mathrm{T}}$ and $e^{-\boldsymbol{M}} = (e^{-M_{11}}, \ldots, e^{-M_{qn_q}})^{\mathrm{T}}$.

Appendix A.1: Computation of $p_{\boldsymbol{v}}(h)$

For estimating $\boldsymbol{\beta}$, we propose to use the adjusted profile likelihood $p_{\boldsymbol{v}}(h)$. Based on the h-likelihood (5.4), we can compute the $p_{\boldsymbol{v}}(h)$ as follows.

$$p_{\boldsymbol{v}}(h) = h(\boldsymbol{\beta}, \tilde{\boldsymbol{v}}) - \frac{1}{2}\log\det\left(\frac{1}{2\pi}\boldsymbol{H}_{\boldsymbol{vv}}\right)\Big|_{\boldsymbol{v}=\tilde{\boldsymbol{v}}},$$

where $\tilde{\boldsymbol{v}} = \arg\max_{\boldsymbol{v}} h(\boldsymbol{\theta}, \boldsymbol{v})$,

$$\boldsymbol{H}_{\boldsymbol{vv}} = -\frac{\partial^2 h(\boldsymbol{\theta}, \boldsymbol{v})}{\partial \boldsymbol{v}\partial \boldsymbol{v}^{\mathrm{T}}} = \frac{1}{\sigma^2}\boldsymbol{Z}^{\mathrm{T}}\,\text{diag}(\boldsymbol{W})\boldsymbol{Z} + \frac{1}{\alpha}\boldsymbol{I}_q,$$

and

$$\begin{aligned}
\boldsymbol{W} &= \boldsymbol{K} \circ \boldsymbol{M}^{1+\zeta}, \\
\boldsymbol{K} &= \boldsymbol{M}^{1+\zeta} \circ \boldsymbol{B} - (1+\zeta)\boldsymbol{A} \circ \boldsymbol{M}^{\zeta}, \\
\boldsymbol{A} &= \boldsymbol{\delta} \circ \left(\frac{1+\zeta}{\boldsymbol{M}} - \mathbb{1}_n \right) + (\mathbb{1}_n - \boldsymbol{\delta}) \circ \frac{1}{e^{\boldsymbol{M}} - \mathbb{1}_n}, \\
\boldsymbol{B} &= \boldsymbol{\delta} \circ \frac{1+\zeta}{\boldsymbol{M}^2} + (\mathbb{1}_n - \boldsymbol{\delta}) \circ \frac{e^{\boldsymbol{M}}}{\left(e^{\boldsymbol{M}} - \mathbb{1}_n\right)^2}.
\end{aligned}$$

Together with $\boldsymbol{H}_{\boldsymbol{vv}}$, following quantities provide the estimation of the variance of $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned}
\boldsymbol{H}_{\boldsymbol{\beta\beta}} &= -\frac{\partial^2 h(\boldsymbol{\theta}, \boldsymbol{v})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \frac{1}{\sigma^2} \boldsymbol{X}^{\mathrm{T}} \operatorname{diag}(\boldsymbol{W}) \boldsymbol{X}, \\
\boldsymbol{H}_{\boldsymbol{\beta v}} &= -\frac{\partial^2 h(\boldsymbol{\theta}, \boldsymbol{v})}{\partial \boldsymbol{\beta} \partial \boldsymbol{v}^{\mathrm{T}}} = \frac{1}{\sigma^2} \boldsymbol{X}^{\mathrm{T}} \operatorname{diag}(\boldsymbol{W}) \boldsymbol{Z},
\end{aligned}$$

where $\operatorname{diag}(\boldsymbol{W})$ is the diagonal matrix of which diagonal elements are $\boldsymbol{W}$.

Appendix A.2: Computation of $p_{\boldsymbol{\psi}}(h)$

For estimating $\boldsymbol{\phi}$, we propose to use the adjusted profile likelihood $p_{\boldsymbol{\psi}}(h)$. Based on the h-likelihood (5.4), we can compute the $p_{\boldsymbol{\psi}}(h)$ as follows.

$$p_{\boldsymbol{\psi}}(h) = h(\boldsymbol{\phi}, \tilde{\boldsymbol{\psi}}) - \frac{1}{2} \log \det \left(\frac{1}{2\pi} \boldsymbol{H}_{\boldsymbol{\psi\psi}} \right) \Big|_{\boldsymbol{\psi} = \tilde{\boldsymbol{\psi}}},$$

where $\tilde{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} h(\boldsymbol{\phi}, \boldsymbol{\psi})$ and

$$\boldsymbol{H}_{\boldsymbol{\psi\psi}} = \begin{pmatrix} \boldsymbol{H}_{\boldsymbol{\beta\beta}} & \boldsymbol{H}_{\boldsymbol{\beta v}} \\ \boldsymbol{H}_{\boldsymbol{\beta v}}^{\mathrm{T}} & \boldsymbol{H}_{\boldsymbol{vv}}. \end{pmatrix}$$

Appendix B: Tables

Table 5.1: Simulation study of fitting two models, GEV (M_{GEV}) and Normal (M_N) AFT random effect models. Simulation data are generated under various error distributions (F_ε) with GEV, Normal (N), t, and log-gamma (LG) with 50% censoring rate; q is the number of clusters and n_i is cluster size. True values for parameters: regression parameters $\beta_1 = 1.5, \beta_2 = -1.5$; variance of normal random effect $\alpha = 2$.

				$\hat{\beta}_1$				$\hat{\beta}_2$				$\hat{\sigma}$		$\hat{\zeta}$		$\hat{\alpha}$		$\hat{\lambda}$	
F_ε	q	n_i	Fitted Model	Mean	SD	SE	CP	Mean	SD	SE	CP	Mean	SD	Mean	S.D	Mean	SD	Mean	SD
GEV	100	15	M_{GEV}	1.495	0.043	0.043	0.944	-1.495	0.079	0.078	0.942	1.276	0.039	-0.646	0.028	2.130	0.417	-	-
			M_N	1.514	0.046	0.042	0.922	-1.513	0.085	0.077	0.930	-	-	-	-	2.104	0.456	1.463	0.084
	200	15	M_{GEV}	1.493	0.032	0.030	0.934	-1.498	0.056	0.055	0.940	1.277	0.030	-0.650	0.022	2.110	0.300	-	-
			M_N	1.512	0.034	0.030	0.912	-1.518	0.060	0.054	0.900	-	-	-	-	2.068	0.282	1.460	0.063
N	100	15	M_{GEV}	1.500	0.039	0.042	0.971	-1.489	0.078	0.076	0.943	1.108	0.030	-0.271	0.026	2.506	0.649	-	-
			M_N	1.475	0.045	0.040	0.893	-1.453	0.078	0.072	0.875	-	-	-	-	1.903	0.289	1.136	0.063
	200	15	M_{GEV}	1.505	0.030	0.030	0.929	-1.501	0.054	0.054	0.956	1.106	0.022	-0.271	0.020	2.409	0.507	-	-
			M_N	1.464	0.028	0.028	0.698	-1.455	0.053	0.050	0.837	-	-	-	-	1.897	0.207	1.137	0.039
t	100	15	M_{GEV}	1.474	0.055	0.048	0.881	-1.474	0.095	0.088	0.911	1.331	0.068	-0.371	0.043	2.177	0.603	-	-
			M_N	1.493	0.049	0.043	0.917	-1.496	0.084	0.078	0.924	-	-	-	-	2.028	0.557	1.567	0.137
	200	15	M_{GEV}	1.473	0.042	0.034	0.818	-1.473	0.070	0.063	0.890	1.340	0.051	-0.375	0.039	2.161	0.416	-	-
			M_N	1.494	0.036	0.030	0.886	-1.494	0.062	0.055	0.927	-	-	-	-	1.955	0.219	1.955	0.219
LG	100	15	M_{GEV}	1.534	0.042	0.040	0.848	-1.539	0.071	0.072	0.910	1.089	0.037	-0.518	0.033	1.980	0.304	-	-
			M_N	1.579	0.046	0.042	0.524	-1.585	0.075	0.074	0.798	-	-	-	-	18.025	9.745	1.114	0.071
	200	15	M_{GEV}	1.537	0.029	0.028	0.760	-1.540	0.053	0.051	0.880	1.093	0.025	-0.521	0.023	1.995	0.213	-	-
			M_N	1.585	0.032	0.029	0.200	-1.589	0.058	0.053	0.599	-	-	-	-	16.675	6.655	1.117	0.048

Table 5.2: Simulation study of fitting two models, GEV (M_{GEV}) and Normal (M_N) AFT random effect models. Simulation data are generated under various error distributions (F_ε) with GEV, Normal (N), t, and log-gamma (LG) with 90% censoring rate; q is the number of clusters and n_i is cluster size. True values for parameters: regression parameters $\beta_1 = 1.5, \beta_2 = -1.5$; variance of normal random effect $\alpha = 2$.

				$\hat{\beta}_1$				$\hat{\beta}_2$				$\hat{\sigma}$		$\hat{\zeta}$		$\hat{\alpha}$		$\hat{\lambda}$	
F_ε	q	n_i	Fitted Model	Mean	SD	SE	CP	Mean	SD	SE	CP	Mean	SD	Mean	S.D	Mean	SD	Mean	SD
GEV	100	15	M_{GEV}	1.468	0.095	0.094	0.918	-1.452	0.177	0.175	0.938	1.183	0.044	-0.686	0.065	2.006	0.493	-	-
			M_N	1.369	0.096	0.084	0.720	-1.357	0.185	0.155	0.841	-	-	-	-	1.514	0.414	1.653	0.116
	200	15	M_{GEV}	1.487	0.070	0.070	0.967	-1.492	0.108	0.131	0.967	1.179	0.031	-0.683	0.049	1.983	0.346	-	-
			M_N	1.364	0.071	0.060	0.374	-1.349	0.120	0.109	0.717	-	-	-	-	1.486	0.360	1.655	0.076
N	100	15	M_{GEV}	1.457	0.099	0.099	0.932	-1.459	0.164	0.177	0.958	1.151	0.083	-0.356	0.046	1.996	0.323	-	-
			M_N	1.402	0.104	0.088	0.746	-1.405	0.167	0.158	0.886	-	-	-	-	1.672	0.409	1.124	0.154
	200	15	M_{GEV}	1.455	0.062	0.070	0.910	-1.471	0.118	0.126	0.955	1.158	0.068	-0.348	0.032	1.978	0.236	-	-
			M_N	1.399	0.066	0.062	0.649	-1.414	0.121	0.111	0.865	-	-	-	-	1.665	0.296	1.113	0.102
t	100	15	M_{GEV}	1.391	0.115	0.092	0.692	-1.379	0.184	0.169	0.861	1.188	0.069	-0.606	0.103	1.682	0.497	-	-
			M_N	1.382	0.112	0.085	0.662	-1.373	0.184	0.156	0.829	-	-	-	-	1.553	0.458	1.704	0.170
	200	15	M_{GEV}	1.377	0.076	0.064	0.498	-1.371	0.124	0.119	0.807	1.186	0.054	-0.607	0.084	1.546	0.280	-	-
			M_N	1.373	0.074	0.060	0.456	-1.365	0.120	0.110	0.769	-	-	-	-	1.453	0.277	1.694	0.119
LG	100	15	M_{GEV}	1.473	0.103	0.093	0.892	-1.466	0.175	0.169	0.939	1.026	0.042	-0.625	0.082	2.357	0.914	-	-
			M_N	1.644	0.122	0.103	0.698	-1.635	0.122	0.103	0.877	-	-	-	-	39.348	20.511	1.435	0.101
	200	15	M_{GEV}	1.468	0.072	0.065	0.875	-1.463	0.119	0.118	0.946	1.024	0.034	-0.626	0.063	2.214	0.632	-	-
			M_N	1.646	0.083	0.072	0.450	-1.643	0.138	0.127	0.775	-	-	-	-	38.260	17.058	1.439	0.065

Table 5.3: Summary statistics of numeric and categorical covariates for COHRI data

Covariate (numeric)	Min	Median	Mean	Max	SD	Covariate (categorical)	0 (%)	1 (%)
Mobility	0.00	0.00	0.04	3.00	0.24	Crown	81.6	18.4
BOP	0.00	0.00	11.74	100.00	22.04	Filled	28	72
Plaque	0.00	0.00	19.05	100.00	28.28	Decayed	74	26
PDmean	0.00	2.50	2.51	12.00	0.72	Gender	50.5	49.5
CALmean	0.00	2.50	2.64	15.00	0.93	Diabetes	90.7	9.3
D.F.sites	0.00	2.00	2.20	5.00	1.72	Tobacco	77.8	22.2
Age	16.00	61.00	58.41	90.00	18.08	Molar	61	39

Table 5.4: Results of fitting the GEV (M_{GEV}) and Normal (M_N) AFT random effect models for COHRI data. LB and UB, lower and upper bounds of 95% confidence interval of regression parameter

	M_{GEV}				M_N			
Variable	Estimate	SE	LB	UB	Estimate	SE	LB	UB
Mobility	-0.618	0.012	-0.642	-0.594	-0.548	0.043	-0.633	-0.464
BOP	-0.004	0.0003	-0.0046	-0.0034	-0.004	0.001	-0.005	-0.002
Plague	0.001	0.0003	0.0004	0.0016	0.001	0.001	-0.0002	0.003
PDmean	-0.192	0.009	-0.210	-0.174	-0.277	0.039	-0.355	-0.200
CALmean	-0.277	0.007	-0.291	-0.263	-0.200	0.032	-0.262	-0.138
Crown	-0.342	0.037	-0.415	-0.269	-0.310	0.062	-0.431	-0.188
Filled	-0.796	0.043	-0.880	-0.712	-0.816	0.045	-0.903	-0.728
Decayed	0.591	0.028	0.536	0.646	0.385	0.037	0.313	0.458
D.F.sites	-0.262	0.012	-0.286	-0.238	-0.223	0.015	-0.252	-0.193
Age	-0.008	0.001	-0.010	-0.006	-0.005	0.006	-0.016	0.005
Gender	-0.069	0.017	-0.102	-0.036	-0.055	0.163	-0.375	0.265
Diabetes	0.280	0.025	0.231	0.329	0.176	0.254	-0.322	0.675
Tobacco	1.148	0.018	1.113	1.183	1.104	0.191	0.729	1.479
Molar	-0.262	0.023	-0.307	-0.217	-0.480	0.036	-0.551	-0.409
σ	1.352	-	-	-	-	-	-	-
ζ	-0.726	-	-	-	-	-	-	-
α	2.399	-	-	-	13.265	-	-	-
λ	-	-	-	-	2.217	-	-	-

Bibliography

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–973.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Barndorff-Nielsen, O. E. (1987). Discussion on Parameter orthogonality and approximate conditional inference (by D. R. Cox and N. Reid). *Journal of the Royal Statistical Society: Series B*, **49**, 18–20.

Bayarri, M. J., DeGroot, M. H. and Kadane, J. B. (1988). *Statistical Decision Theory and Related Topics IV. Vol. 1, eds S.S. Gupta and J.O. Berger*. New York: Springer.

Berger, J. O. and Wolpert, R. (1984). *The Likelihood Principle*. Hayward: Institute of Mathematical Statistics Monograph Series.

Bivand, R. S., Gomez-Rubio, V. and Rue, H. (2015). Spatial data analysis with R-INLA with some extensions. *Journal of Statistical Software*, **63**, 1–31.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and likelihood principle. *Journal of the American Statistical Association*, **91**, 791–806.

Bladt, M. and Albrecher, H. (2021). Trimmed extreme value estimators for censored heavy-tailed data. *Electronic journal of statistics*, **15**, 3112–3136.

Bologa, C. G., Pankratz, V. S., Unruh, M. L., Roumelioti, M. E., Shah, V., Shaffi, S. K., Arzhan, S., Cook, J. and Argyropoulos, C. (2021). High performance implementation of the hierarchical likelihood for generalized linear mixed models: an application to estimate the potassium reference range in massive electronic health records datasets. *BMC Medical Research Methodology*, **21**, 151.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Burda, Y., Grosse, R. and Salakhutdinov, R. (2016). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Butler, R. W. (1986). Predictive likelihood inference with applications. *Journal of the Royal Statistical Society: Series B*, **48**, 1–38.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, **96**, 723–734.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, **104**, 439–453.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, **89**, 81–87.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B*, **49**, 1–39.

Christensen, O. F. and Ribeiro Jr, P. J. (2017). geoRglm: A package for generalised linear spatial models, *R package version 0.9-11*.

Dadgoug, M., Goga, C. and Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, `https://doi.org/10.1080/01621459.2021.1987250`.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**, 1–37.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based

geostatistics. *Journal of the Royal Statistical Society: Series C*, **47**, 299–350.

Dong, Y. and Peng, C. J. Y. (2013). Principled missing data methods for researchers. *SpringerPlus*, **2**, 222.

Drum, M. L. and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrika*, **49**, 677–689.

Eguchi, S. (2021). Pythagoras theorem in information geometry and applications to generalized linear models. *Handbook of statistics*, **45**, 15–42.

Firth, D. (2006). Invited discussion (Seconder of the Vote of Thanks) of 'Double hierarchical generalized linear models' by Lee and Nelder. *Journal of the Royal Statistical Society: Series C*, **55**, 168–170.

Firth, D. and Bennett, K. E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B*, **60**, 3–21.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A*, **222**, 309–368.

Guerrero, V. M. and Johnson, R. A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika*, **69**, 309–314.

Ha, I., Lee, Y. and Song, J. (2002). Hierarchical-likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, **8**, 163–176.

Ha, I., Noh, M. and Lee, Y. (2012). frailtyHL: a package for fitting frailty models with h-likelihood. *R Journal*, **4**, 307–320.

Ha, I., Jeong, J. and Lee, Y. (2017). *Statistical Modeling of Survival Data with Random Effects*. Springer.

Ha, I., Noh, M., Kim, J. and Lee, Y. (2019). frailtyHL: Frailty models via hierarchical likelihood. *R package version 2.3*. (Available from `https://CRAN.R-project.org/package=frailtyHL`).

Han, J. and Lee, Y. (2022). Enhanced Laplace approximation. *Manuscript prepared*.

Han, P. and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, **100**, 417–430.

Han, J., Lee, Y. and Kim, J. K. (2022). ML imputation via h-likelihood. *Manuscript prepared*.

Han, J., Ha, I., Lee, Y. and Bandyopadhyay, D. (2022). Fitting heavily censored data by an accelerated failure time random effect model with generalized extreme value distribution using hierarchical likelihood. *Manuscript prepared*.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Wiley-Interscience.

Henderson, C. R., Kempthorne, O., Searle, S. R. and Von Krosigk, C. M. (1959). The estimation of genetic and environmental trends from records subject to culling. *Biometrics*, **15**, 192–218.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer.

Hung, H., Jou, Z. Y. and Huang, S. Y. (2018). Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, **74**, 145–154.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *TEST*, **18**, 1–44.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, **76**, 243–263.

Jin, S. and Lee, Y. (2022). Standard error estimates in hierarchical generalized linear models. *Manuscript prepared*.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics*, **48**, 681–694.

Kim, J. K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, **19**, 145–158.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, **98**, 119–132.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**, 375–394.

Kim, J. K. and Shao, J. (2021). *Statistical Methods for Handling Incomplete Data (2nd ed.)*. CRC press.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. and Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, **70**, 1–21.

Lange, K. L. and Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution, *Journal of the American Statistical Association*, 84:881–896.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalised linear models (with discussion). *Journal of the Royal Statistical Society: Series B*, **58**, 619–678.

Lee, Y. and Nelder, J. A. (2000). The relationship between double exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex ratio data. *Journal of the Royal Statistical Society: Series C*, **49**, 413–419.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalized linear models: a synthesis of generalised linear models, random effects models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y. and Nelder, J. A. (2005). Likelihood for random-effect models (with discussion). *Statistics and Operations Research Transactions*, **29**, 141–164.

Lee, Y. and Nelder, J. A. (2006). Fitting via alternative random effect models. *Statistics and Computing*, **16**, 69–75.

Lee, Y. and Nelder, J. A. (2006). Double hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series C*, **55**, 139–185.

Lee, Y. and Nelder, J. A. (2009). Likelihood inference for models with unobservables: Another view. *Statistical Science*, **24**, 255–269.

Lee, Y. and Noh, M. (2018). dhglm: Double hierarchical generalized linear models. *R package version 2.0*. (Available from `https://CRAN.R-project.org/package=dhglm`).

Lee, Y. and Kim, G. (2016). H-likelihood predictive intervals for unobservables. *International Statistical Review*, **84**, 487–505.

Lee, Y. and Kim, G. (2020). Properties of h-Likelihood estimators in clustered data. *International Statistical Review*, **88**, 380–395.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall/CRC.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2017). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood (2nd ed.)*. Chapman & Hall/CRC.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data (2nd ed.)*. John Wiley & Sons.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data (3rd ed.)*. John Wiley & Sons.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **44**, 226–233.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd ed.)*. Chapman & Hall, London.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, **9**, 538–573.

Meng, X. L. (2009). Decoding the h-likelihood. *Statistical Science*, **24**, 280–293.

Molenberghs, G., Beunckens, C. and Kenward, M. G. (2008). Every missingness not at random has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B*, **70**, 371–388.

Moraga, P. (2019). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Chapman & Hall/CRC.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, **4**, 2111–2245.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1–32.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, **57**, 896–915.

Ogden, H. E. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika*, **104**, 153–164.

Ogden, H. E. (2021). On the error in Laplace approximations of high-dimensional integrals. *Stat*, **10**, e380.

Paik, C. M., Lee, Y. and Ha, I. (2015). Frequentist inference on random effects based on summarizability. *Statistica Sinica*, **25**, 1107–1132.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pauli, F., Racugno, W. and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, **21**, 149–164.

Perry, P. O. (2017). Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society: Series B*, **79**, 267–291.

Raudenbush, S. W., Yang, M. and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, **9**, 141–157.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika*, **87**, 113–124.

Roy, V. and Dey, D. (2014). Propriety of posterior distributions arising in categorical and survival models under generalized extreme value distribution. *Statistica Sinica*, **24**, 699–722.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D. B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, 20–34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, **71**, 319–392.

Sang, H., Kim, J. K. and Lee, D. (2020). Semiparametric fractional imputation using Gaussian mixture models for multivariate missing data. *Journal of the American Statistical Association*, `https://doi.org/10.1080/01621459.2020.1796358`.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood and Probability. Statistical Inference with Confidence Distributions*. Cambridge University Press.

Shun, Z. (1997). Another look at the salamander mating data: A modified Laplace approximation approach. *Journal of the American Statistical Association*, **92**, 341–349.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B*, **57**, 749–760.

Stark, P. C., Kalenderian, E., White, J. M., e.a. (2010). Consortium for oral health-related informatics: improving dental research, education, and treatment. *Journal of dental education*, **74**, 1051–1065.

Sweeting, T. J. (1987). Discussion on Parameter orthogonality and approximate conditional inference (by D. R. Cox and N. Reid). *Journal of the Royal Statistical Society: Series B*, **49**, 20–21.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Vaida, F and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press LLC.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, **85**, 935–948.

Wang, H. and Kim, J. K. (2021). Information projection approach to propensity score estimation for handling selection bias under missing at random. *arXiv:2104.13469 [stat.ME]*.

Wang, Z., Shi, J. Q. and Lee, Y. (2017). Extended t-process regression models. *Journal of Statistical Planning and Inference*, **189**, 38–60.

Yang, S. and Kim, J. K. (1994). A note on multiple imputation for method of moments estimation. *Biometrika*, **103**, 244–251.

Yun, S., Lee, Y. and Kenward, M. G. (2007). Using h-likelihood for missing observations. *Biometrika*, **94**, 905–919.

국문초록

계층적 가능도는 고정된 모수만을 취급하던 기존의 가능도를 확장하여 관측되지 않은 잠재 변수를 포함하는 통계 모형에 대해 최대 가능도 추정을 허락하기 위해 제안되었다. 하지만, 기존의 계층적 가능도는 분산 성분을 포함한 모든 추정의 대상에 대해 최대 가능도 추정을 허락하지 못한다는 한계가 있었다. 본 학위논문에서는 계층적 가능도의 정준 척도의 성질을 살펴본 뒤, 이를 바탕으로 모든 모수들의 최대 가능도 추정량을 얻는 방법에 대하여 논의하였다.

불완전 자료의 예로는 결측 자료, 변량 효과, 중도 절단 자료 등이 있다. 이러한 불완전 자료에 대하여, 계층적 가능도를 이용한 통계적 추론의 유용성을 살펴보았다. 하지만, 관측되지 않은 잠재 변수에 대한 통계 모형의 경우, 관측된 자료로부터 항상 식별 가능하지 않을 수 있다. 따라서, 본 학위논문에서는 통계 모형에 사용되는 다양한 가정들에 대해 로버스트한 추론을 허락하는 방법도 함께 제시하였다.

주요어 : 정준 척도, 중도 절단 자료, 대치법, 라플라스 근사, 최대 가능도 추정, 결측 자료, 변량 효과, 로버스트 추론.

학 번 : 2014 – 21213