MODELING THE PROBABILITY OF TUBERCULOSIS CONVERSION FROM ECOLOGICAL MOMENTARY ASSESSMENT OF SOCIAL PATTERNS

by

CHUNLA HE

(Under the Direction of Stephen L. Rathbun)

ABSTRACT

Transmission of *Mycobacterium tuberculosis (M. tuberculosis)* relies on prolonged contacts with people infected with *M. tuberculosis*. The Community Health Study of Social Networks and Tuberculosis (COHSONET), an ongoing study initiated by Whalen, aims to evaluate the effects of social contacts on the risk of *M. tuberculosis* conversion through Ecological Momentary Assessment (EMA). This dissertation offers the linear probability model as an alternative to the logistic regression, to describe the risk of *M. tuberculosis* conversion as a function of proportions of time participants spent in different location contexts, a surrogate variable for social contacts. To restrict the predictive values from the linear probability model in a meaningful interval [0, 1], we propose two constrained optimization approaches in the current dissertation: the constrained ordinary least squares (OLS) and constrained adaptive LASSO. Within the constrained parameter space, both constrained OLS and constrained adaptive LASSO estimators are asymptotically consistent and asymptotically normal given all parameter estimates lying within the boundary of parameter space. Other than that, the constrained adaptive LASSO is an oracle procedure, and thus has consistent model selection. Intensive simulations

demonstrate that both constrained OLS and constrained adaptive LASSO estimators are asymptotically consistent because their empirical mean biases tend to approach zero with an increased sample size. Moreover, the constrained OLS estimators (MLEs) perform as well as maximum likelihood estimators and bias-reduced penalized maximum likelihood estimators(PMLEs) in the logistic regression when all parameters are in the interior of the boundary. In particular, the constrained OLS in the linear probability model outperforms both MLE and PMLE in the logistic regression model when some parameters close to the boundary of the parameter space. The constrained adaptive LASSO appears to have better performance than the constrained OLS in the linear probability model when some parameters lie well on the boundary of the parameter space.

INDEX WORDS: Social contact patterns, *M. tuberculosis*, Ecological Momentary Assessment, linear probability model, logistic regression model, constrained, ordinary least squares, adaptive LASSO, maximum likelihood estimators, penalized maximum likelihood estimators.

MODELING THE PROBABILITY OF TUBERCULOSIS CONVERSION FROM ECOLOGICAL MOMENTARY ASSESSMENT OF SOCIAL PATTERNS

by

CHUNLA HE

B.Med., Southern Medical University, China, 2009

M.S., Southern Medical University, China, 2012

M.S., The University of Georgia, 2016

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2017

MODELING THE PROBABILITY OF TUBERCULOSIS CONVERSION FROM ECOLOGICAL MOMENTARY ASSESSMENT OF SOCIAL PATTERNS

by

CHUNLA HE

Major Professor: Stephen L. Rathbun
Committee: Christopher Whalen
Hanwen Huang
Ye Shen

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2017

ACKNOWLEDGEMENTS

There are many debts to gratefully acknowledge. My first thanks must go to the University of Georgia, the Graduate School, the Department of Epidemiology and Biostatistics made this program possible. Foremost, I am deeply indebted to my academic advisor, Dr. Stephen Rathbun. Dr. Rathbun has been all one could hope for in an advisor. Over all the years at UGA, Dr. Rathbun's expertise in statistics has hone my skills in researching and being an independent biostatistician. At the end of my Ph.D. program, I am only beginning to realize how much has been offered and taught. Without his guidance and encouragement, I could not finish such a complex work.

My dissertation committee have been instrumental throughout the study and research process. Dr. Christopher Whalen not only generously offered me the opportunity to start my PhD program at UGA, but also allowed me to use his wonderful research data for my dissertation. Also, he helped cultivate my interest in doing research in the field of public health. My thanks also go to Dr. Ye Shen and Dr. Hanwen Huang. Their excellent courses in LASSO and generalized linear models contribute directly to the development of my dissertation. I am grateful to have been a student of them, whose knowledge and insights continually widen my horizons.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

1.1 INTRODUCTION 1

1.2 MOTIVATING DATASET 6

1.3 DESIGN-BASED INFERENCE FOR THE DOMAIN MEANS 7

1.4 MISSING DATA MODELING 9

1.5 MODEL-BASED MEASUREMENT ERROR 12

1.6 DESIGN-BASED ALGORITHM 15

1.7 COHSONET STUDY OF RISK OF TUBERCULOSIS CONVERSION 17

1.8 DISSERTAION OUTLINE 24

1.9 REFERENCES 25

2 CONSTRAINED LINEAR PROBABILITY MODEL IN DETEMRING THE PROBABILITY OF TUBERCULOSIS CONVERSION FROM ECOLOGICAL ASSESSMETN OF SOCIAL PATTERNS 32

2.1 INTRODUCTION 32

2.2 CONSTRAINED LINEAR PROBABILITY MODEL 35
2.3 SIMULATIONS 55
2.4 APPLICATION OF COHSONET DATA 68
2.5 DISCUSSION 77
2.6 REFERENCES 79
3 CONSTRAINED ADAPTIVE LASSO FOR ESTIMATING THE PROBABILITY OF TUBERCULOSIS CONVERSION FROM ECOLOGICAL MOMENTARY ASSESSMENT OF SOCIAL PATTERNS88
3.1 INTRODUCTION 86
3.2 CONSTRAINED ADAPTIVE LASSO 89
3.3 SIMULATIONS 96
3.4 APPLICATION OF COHSONET DATA 109
3.5 DISCUSSION 116
3.6 REFERENCES 117
4 CONCLUSIONS 123
4.1 SUMMARY 123
4.2 FUTURE RESEARCH 126
4.3 REFERENCES 128

LIST OF TABLES

Page

1.1 Mean levels of random effects describing the periodic pattern of the probability of answering phone calls (n=288) 18

1.2 Parameter estimates in the constrained linear probability model over all location contexts (n=189) 22

1.3 Parameter estimates in the constrained linear probability model over reduced location contexts (n=189) 24

2.1 Simulation results for Scenario A with known location contexts: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals 58

2.2 Simulation results for Scenario B with known location contexts: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals 62

2.3 Comparison of simulation results between Scenarios A and C: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals; Ratio, empirical variance of known location contexts versus estimated location contexts 66

2.4 Parameter estimates using different approaches over all location contexts (n=189). 95% CI: 95% confidence interval 74

2.5 Parameter estimates based on reduced location contexts (n=189). 95% CI: 95% confidence interval 76

3.1 Simulation results for Scenario A with known location contexts: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals; % To 0, percentage of estimates falling on zero; % To 1, percentage of estimates falling on one 99

3.2 Simulation results for Scenario B with known location contexts: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals; % To 0, percentage of estimates falling on zero; % To 1, percentage of estimates falling on one 103

3.3 Comparison of simulation results between Scenarios A and C: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals; % To 0, percentage of estimates falling on zero; % To 1, percentage of estimates falling on one; Ratio, empirical mean ratio of variance of bias.....107

3.4 Parameter estimates of constrained adaptive LASSO and constrained OLS in the linear probability model over all location contexts (n=189). 95% CI: 95% confidence interval.....114

3.5 Parameter estimates of constrained adaptive LASSO and constrained OLS in the linear probability model based on the reduced location contexts (n=189). 95% CI: 95% confidence interval 116

LIST OF FIGURES

Page

1.1 Random effects modeling of probability of answering the phone calls (n=288)....19

1.2 Mean proportions of time spent at each location context (n=288) 20

2.1 An illustration of tangent cone.....43

2.2 Effect of constraints on the probability density function for a parameter with expected value μ_N 50

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Tuberculosis is a potentially fatal contagious disease that is caused by the bacterium *Mycobacterium tuberculosis* (*M. tuberculosis*). The World Health Organization (WHO) estimated that 9.0 million people developed tuberculosis and 1.5 million died from the disease in 2013. Tuberculosis threatens the health of people all over the world, and is most prevalent in resource-limited countries such as the developing countries in Africa. It is estimated that 2.9 million new cases of tuberculosis occur per year on the sub-continent and that the incidence of tuberculosis is 125 cases per 100,000 population, 25 times the rate found in the US.

Tuberculosis is transmitted through *M. tuberculosis* bacteria carried on droplets of secretions emitted by actively infected patients via exhaling, coughing, sneezing, or talking. Susceptible individuals sharing the same environment can inhale droplets that may leads to one of three different clinical outcomes: complete clearance of the pathogen, latent tuberculosis infection, or progression to primary active disease (Bhatt and Salgame, 2007; Roach et al., 2002). *M. tuberculosis* infection is unlikely to be transmitted through a casual contact in a single incident. Transmission of *M. tuberculosis* requires prolonged contact with an infectious case (Houk et al., 1968; Kenyon et al., 1996). The spread of *M. tuberculosis* from person to person is highly dependent upon the frequency and nature of

contacts between infected and susceptible members of the population (Horby et al., 2011; Rehkopf et al., 2015).

In the late 1950s, Wells and Riley initiated the first study to evaluate the impact of indoor environment on tuberculosis transmission, in which guinea pigs were exposed to air from a tuberculosis ward and infection rates were measured under controlled conditions (Riley et al., 1962; Sultan et al., 1960). The transmission dynamics of index cases of tuberculosis within households is now well understood (Guwatudde et al., 2003; Lienhardt et al., 2003a; Lienhardt et al., 2003b; Reichler et al., 2002; Whalen et al., 2011). Knowledge concerning the dynamics outside the household, however, is less complete. From previous reports on tuberculosis outbreaks, *M. tuberculosis* can be transmitted in community venues such as clinics, hospitals (Dooley et al., 1992), bars (Classen et al., 1999; Yaganehdoost et al., 1999), and homeless shelters (Barnes et al., 1996). Although it is known that *M. tuberculosis* transmission does occur in settings outside household, the magnitude of transmission in these locations is not well quantified.

Traditionally, self-report surveys have been conducted to measure contact patterns relevant to the transmission of tuberculosis. The nature of social contacts is affected by demographic factors, the living and working environment, socio-cultural norms and individual lifestyle choices; all of which vary by place and time. Self-reported social contacts data suffer from a variety of biases, recall bias in particular. Study designs based on retrospective self-report data cannot assess complex and temporally dynamic psychological, behavioral, and physiological processes in the natural environment. Additionally, most social contact surveys have been conducted in developed western countries, yet the majority of the world's population lives in less developed countries

where family structures, socio-cultural norms, population mobility and the living and work environment may differ in important ways from western countries (Horby et al., 2011; Mossong et al., 2008). It is well acknowledged that infectious diseases such as tuberculosis and HIV are prevalent in developing countries in Africa. Contact networks may be very dense in countries where tuberculosis is prevalent, which makes it more difficult to quantify the nature of transmission due to rapid accumulation of tuberculosis cases. There is therefore a strong need to determine contact networks in developing countries. Whalen attempted to address this knowledge gap by conducting an Ecological Momentary Assessment (EMA) of contact patterns in the Rubaga Division of Kampala, Uganda.

Ecological momentary assessment (EMA) is a method of data collection originating from the behavioral sciences that enables evaluation of within-person patterns of behavior and experience via repeated sampling of participants' behaviors and experiences in their natural environments using diaries or surveys completed one or more times per day (Shiffman, Stone and Hufford, 2008). EMA data can be collected through the use of electronic devices such as personal digital assistants (PDAs) or mobile phones. EMA minimizes recall bias by requiring participants to immediately respond to random prompts regarding current as opposed to past states or record specific events on a daily basis in their natural environments (Stone and Shiffman, 2002). EMA is widely used in behavioral medicine research including studies of dieting (Carels et al., 2004) and smoking (Shiffman et al., 2002). EMA has also been employed to study the effects of momentary psychological states on chronic pain (Feldman, Downey and Schaffer-Neitz, 1999), and hypertension (Kamarck et al., 2007). So far, a few studies using EMA have

been conducted to evaluate the impacts of momentary (Barta, Tennen and Kiene, 2010; Mustanski, 2007) and contextual factors including alcohol use (Barta et al., 2008; Yang et al., 2015) in HIV prevention behaviors. To the best of our knowledge, however, EMA has never been used to evaluate the effects of social contacts on the transmission dynamics of tuberculosis.

EMA methods offer a number of important advantages over the traditional self-reported surveys as well as other longitudinal approaches (Shiffman et al., 2008; Wray, Merrill and Monti, 2014). One crucial merit of EMA methods is that it can characterize changes in dynamic processes occurring over relatively short periods of time. In particular, EMA data can capture the dynamic interplay between various situations, environments, and behavior (Shiffman et al., 2008). In addition, EMA can avoid biases inherent in retrospective recall of momentary states. In spite of the versatility and flexibility of EMA methods for measuring momentary phenomena, they are associated with some limitations. EMA methods obviously depend on subjects' compliance with instructions to respond to prompts. Numerous repeated assessments according to EMA protocols over long periods of time may yield reduced compliance over time, which can pose significant challenges for the analysis of EMA data. In particular, non-compliance cannot only result in missing data, but introduce bias in the data that are collected as well. Moreover, EMA studies of contact patterns involving signal-contingent reports (e.g., participants are given the same number of phone calls over the study period) could have other limitations because respondents may be more or less likely to answer the phone calls in specific environments, so that the collected data might not be missing at random. A comprehensive understanding of the nature of missing data in EMA is needed when

choosing an analytic approach for EMA data. In this regard, measurement error in the current study does not only come from the random sampling process, but may also result from statistical approaches chosen for aggregating the EMA data.

The objective of this prospectus is to propose statistical models to describe the risk of *M. tuberculosis* conversion as a function of social contact patterns as described by proportions of time spent in different location contexts (e.g., home, friend's or relative's homes, work places, bars, markets, etc.). Logistic regression is the most popular statistical model for estimating the probability of binary response. However, it is not considered here because the logistic regression score equations are nonlinear in the predictors. The classical approach to handling measurement errors in logistic regression treats the observed values of the predictors as being equal to the true value plus a random error. In the current study, however, errors in the observed values are due to both random sampling as well random effects modeling, which causes uncertainty in the preciseness of parameter estimates in the logistic regression model. Therefore, this dissertation considers linear probability models instead, which ensures the unbiasedness of estimating equations and hence consistency of parameter estimators.

The remainder of this chapter is organized as follows: the motivating data for this dissertation will be briefly introduced in Section 1.2. We will then discuss statistical inference for design-unbiased domain means in Section 1.3. Section 1.4 presents a brief review of random effects model for estimating the probability that participants responded to the phone calls. In Section 1.5, a brief review of model-base measurement error models will be given. In section 1.6, we will demonstrate how to obtain designed-unbiased parameters in the constrained linear regression model. In section 1.7, we will

present the random effects model modeling results as exploratory analysis of the constrained linear probability model. This chapter will conclude with an outline of the contents for the rest of the dissertation.

1.2 MOTIVATIVATING DATASET

The motivating dataset for this dissertation is from the Community Health Study of Social Networks and Tuberculosis (COHSONET), Whalen's ongoing study on transmission dynamics of *M tuberculosis* in the Rubaga Division of Kampala, Uganda. Whalen hypothesizes that some locations place subjects at greater risk than others FOR *M tuberculosis* transmission. So one aim of his study is to identify the source location contexts of transmission in a community where the tuberculosis infection occurs, and the types of human interaction which have the potential to facilitate the spread of the disease.

Individuals aged between 15 and 45 years and were free of *M. tuberculosis* infection (i.e., TST < 5 mm) were eligible for inclusion in the COHSONET study. A random sample of cohort members was followed up for one year in an attempt to closely monitor the social contact patterns through EMA. Participants were prompted to answer a set of questions concerning the location and surrounding environment when the calls were answered. Sampling times when the phone calls were made were randomly generated via a self-correcting point process. The self-correcting point process is a special type of a point process with conditional intensity

$$\lambda(\mathrm{t}|\mathcal{F}_t) = \exp\{\alpha_0 + \alpha_1(t - \alpha_2 N(t))\}, \qquad t \in [0, T] \tag{1.1}$$

where α_0 , α_1, and α_2 are constants (Isham and M., 1979; Ogata and Vere-Jones, 1984; Vere-Jones and Ogata, 1984), and $\alpha_1,\ \alpha_2 > 0$. This point process is a self-correcting in

the sense that if the number of events strays from the target $1/\alpha_2$, then the assessment rate compensates to force this difference back towards zero. The baseline intensity is $\exp\{\alpha_0\}$. The parameters α_0 and α_2 govern the mean number of phone calls made per day, while α_1 controls the variability of the number of calls per day and the regularity of the spacing of the assessment times. Note that the self-correcting point process generates more regularly spaced assessment times and less variation in numbers of assessments per day than the Poisson process, reducing burden on the study subjects. In the COHSONET study, α_0, $\alpha_1, and\ \alpha_2$ were set equal to -0.602, 3, and 1.825, respectively, targeting 200 random assessments per year.

The TST test was used to assess whether participants contracted *M. tuberculosis* at the end of one-year cohort study according to standardized definitions of skin test conversion (Menzies 1999). Cohort members who tested negative at baseline but had a TST > 10 mm after one year and increment more than 6 mm from baseline are considered to be *M. tuberculosis* converters.

1.3 DESIGNED-BASED INFERENCE FOR THE DOMAIN MEAN

The objective of this dissertation is to describe the risk of *M. tuberculosis* conversion as a function of proportions of time participants spent in different location contexts. More specifically, the risk of *M. tuberculosis* conversion is modeled as a function of the domain means, the population proportions of time participants spent in different location contexts over a one-year period of EMA study, where the populations are comprised of the set of all times in the one-year study interval of each participant. More formally, let $x_i(t)$ denote a vector of indicator variables, whose *j*-th element takes the value one if the

subject *i* is in location *j* at time *t*, and the value zero if otherwise. Then the vector of domain means (e.g., proportions of time participants spent in different location contexts) is

$$X_i(T) = \frac{1}{T}\int_0^T x_i(t)dt. \qquad (1.2)$$

where the integral is over the study period $[0, T]$. Evaluation of integral requires that the time-varying covariates $x_i(t)$ be known functions of time. Unless the subjects are observed 24 hours per day 7 days per week, these domain means are unknown. If, however, participants are sampled at times realized from a known probability-based sampling design, then design-unbiased estimators of the domain means may be obtained. The remainder of this section describes how the domain means can be estimated based on the EMA sampling strategy.

Suppose that the subjects are sampled from a known temporal point process. A temporal point process models the occurrences of recurrent events over time. Let the measure *N*(t) represent the number of events in (0, t]. The behavior of a temporal point process *N*(.) is typically modeled by specifying its conditional intensity λ(t), which refers to the rate at which events occur per unit time conditional on the prior history of the point process. Assume that the time-varying covariates $x_i(t)$ are sampled according to a temporal point process with conditional intensity

$$\lambda(t) = \lim_{\delta \to 0} \frac{E\{N[t,\ t+\delta] | \mathcal{F}_t\}}{\delta}, \qquad t \in [0, T] \qquad (1.3)$$

where $T \subset \mathbb{R}$ is the sampling interval, *N*[t, t + δ] denotes the number of events in the interval [t, t + δ], and $\mathcal{F}_t$, the smallest σ-algebra generated by $\{N(u, t]; 0 < u \leq t\}$, denotes the entire history of the point process *N* up to time t. In the COHSONET study,

each participant was sampled at times realized from a self-correcting point process with intensity (1.1).

If the time-varying covariates are sampled according to a point process with known intensity $\lambda_i(t)$, and $\lambda_i(t) > 0$ for almost all $t \in T$, then a design-unbiased estimator can be obtained from

$$\hat{X}_i(T_i) = \frac{1}{T}\sum_{t\in N_i}\frac{x_i(t)}{\lambda_i(t)} \tag{1.4}$$

where N_i denotes the set of times at which assessments were made for subject *i*. The above estimator is design unbiased in the sense that its expected value equal to $X_i(T)$ under the probability model induced by the sampling design. The variance-covariance matrix of the design-unbiased estimator may be computed using methods similar to those found in the proof of Theorem 1 of Ogata (1978) yielding

$$var\{\hat{X}_i(T)\} = \frac{1}{T^2}\int_0^T \frac{x_i(t)x_i(t)^T}{\lambda_i(t)}dt \ .$$

A design-unbiased estimator for $var\{\hat{X}_i(T)\}$ is

$$var\{\hat{X}_i(T)\} = \frac{1}{T^2}\sum_{t\in N_i}\frac{x_i(t)x_i(t)^T}{{\lambda_i}^2(t)}. \tag{1.5}$$

1.4 MISSING DATA MODELING

The methods described in Section 1.3 require that subjects respond to all calls. However, subjects only responded to approximately 63.7% of calls in the COHSONET study. Given the substantial amount of missing data, there is potential bias in estimates of model parameters describing the impact of location contexts on risk of *M. tuberculosis* conversion. The only information available for unanswered calls is the time and date at which each call was made. Therefore, it is only feasible to describe the pattern of answered phone calls as a function of calling times. Let $p_i(t)$ denote the probability that a call at time t is answered by subject i. Let $Z_i(t) = 1$ if a call is answered at time t by

subject i, and $Z_i(t) = 0$ if otherwise. Assume that $Z_i(t), t \in N_i$, are independently sampled from a Bernoulli distribution with thinning function $p_i(t)$, where N_i denotes the set of times at which calls are made to subject i, a realization of a point process with intensity $\lambda_i(t)$. Then the set of answered calls ${N_i}^*$ is a realization of a thinned point process with intensity $\lambda_i(t)p_i(t)$(Cressie 1991). Assume that the data are missing at random, the design-unbiased estimators in (1.5) may be replaced with corrected estimators

$$\hat{X}_i(T) = \frac{1}{T}\sum_{t\in {N_i}^*} \frac{x_i(t)}{\lambda_i(t)p_i(t)}\,. \tag{1.6}$$

In the COHSONET study, exploratory data analysis indicated that the missing data pattern depended on the time of day, a pattern that is likely to vary among study participants. The location contexts in which participants spend their time are also likely to be a function of time of day, a function that may also vary among study participants. Therefore, a random effects model is proposed for the thinning point process, which can be described through a logit link

$$log\frac{p_i(t)}{1-p_i(t)} = {\alpha_i}^T z_i(t). \tag{1.7}$$

Parameters $\alpha_i's$ are assumed to be independently sampled from a normal distribution with mean μ and variance-covariance matrix $\sum$.

Rathbun and Shiffman (2016) proposed a method for fitting mixed effects models for the impact of partially-observed covariates on recurrent events data. They reviewed various methods for fitting such models, which include approaches for fitting models with gamma frailties (Lawless, 1987; Thall, 1988), Laplace approximations to the likelihood (Breslow and Clayton, 1993) and maximum hierarchical likelihood (Lee and Nelder, 1996), etc. None of these methods produce consistent estimators if the sampling interval is small. The Expectation-Maximization (EM) algorithm can be used to obtain consistent

estimators in the mixed effects model. In most instances, however, it is challenging to determine the E-step in the mixed effects model because the conditional expectation of the complete data log-likelihood is an intractable integral (Steele, 1996). Steele (1996) used a second-order Laplace approximation for computation of conditional expectations within the E-step. Rathbun and Shiffman's (2016) model is an extension of Steele's (1996) modified EM algorithm for fitting generalized linear mixed models to recurrent events data with incompletely observed time-varying covariates.

We assume that thinning function in (1.7) is a periodic as follows

$$log\frac{p_i(t)}{1-p_i(t)} = \sum_{k=1}^{K} u_{ik}\cos(\frac{2\pi kt}{\tau} + \phi_{ik}),$$

where u_{ik} denotes the amplitude, τ represents the period set to 1 (day), and ϕ_{ik} denotes the phase. This model can be reparameterized as

$$log\frac{p_i(t)}{1-p_i(t)} = \gamma_{i0} + \sum_{k=1}^{K}\left\{\gamma_{i1k}\cos\left(\frac{2\pi kt}{\tau}\right) + \gamma_{i2k}\sin(\frac{2\pi kt}{\tau})\right\}, \qquad (1.8)$$

where the amplitude is

$$u_{ik} = \sqrt{\gamma_{i1k}^2 + \gamma_{i2k}^2},$$

and the phase is

$$\phi_{ik} = -tan^{-1}(\frac{\gamma_{i1k}}{\gamma_{i2k}}).$$

Alternatively, we can also write the random effects logistic function as a cubic spline function, which takes the form

$$log\frac{p_i(t)}{1-p_i(t)} = \alpha_{i0} + \alpha_{i1}t + \alpha_{i2}t^2 + \alpha_{i3}t^3 + \sum_{k=1}^{K} u_{ik}(t-K_k)_+^3$$

The function

$$(t-K_k)_+^3 = \begin{cases} 0, & for\ t < K_k \\ (t+K_k)^3, & for\ t > K_k \end{cases}$$

represents a broken cubic line with a knot K_k. Here, K denotes the number of knots.

1.5 MODEL-BASED MEASUREMENT ERROR

Since the maximum likelihood estimators (MLEs) in the logistic regression have a bias of order $O(n^{-1})$ (Firth 1993) and its score equations are nonlinear functions of the design-unbiased estimators, the logistic regression model may yield biased parameter estimates, and thus is abandoned in the current study. Instead, we propose a linear probability model for determining the risk *M. tuberculosis* conversion

$$E\left(Y_i \middle| X_i(T)\right) = \Pr\left(Y_i = 1 \middle| X_i(T)\right) = \beta^T X_i(T); \qquad i = 1, \dots, n. \tag{1.9}$$

where $0 \leq \beta_j \leq 1 \; for \; j = 1, \dots, p$, and Y_i is a binary variable denoting whether a subject contracts *M. tuberculosis* $(i.e., Y_i = 1)$ or not $(i.e., Y_i = 0)$ at the end of the study, and $X_i(T)$ is a vector of variables corresponding to the proportions of time subject i spent in the different location contexts. Note that the linear probability model considered here does not contain an intercept term.

Hellevik (2009) demonstrated that linear probability model is a compelling alternative to logistic regression model in many situations with a binary dependent variable. The major advantage of the linear probability model is its interpretability. In the current context, the regression coefficient β_j represents the risk of *M. tuberculosis* conversion if the participants spent 100% of time in location j; $j = 1, \dots, p$. Ordinary least squares (OLS) estimator is a popular parameter estimation in linear regression model, however, direct application of it in the data with a dichotomous outcome can produce estimated probabilities outside the unit interval [0, 1]. To obtain a meaningful modeling results, we

need to restrict the parameter estimates in the linear probability model within the unit interval.

In the current context, proportions of time participants spent in different location contexts were not observed, but were estimated using the design-unbiased estimators (expression (1.6)). Ignoring measurement errors in the predictors can cause bias in the parameter estimation, a loss of power for detecting relationships among variables of interest, and therefore masks the features of the data.

Let X denote a predictor that cannot be observed exactly in a study (e.g., the proportions of time participants spent in each location), Y denote the response variable, and $\hat{X}$ denote an observed value of X which is subject to measurement error. If the conditional distribution of Y given both X and $\hat{X}$ is the same as that of Y given $\bar{X}$ (i.e., $f_{Y|X\hat{X}} = f_{Y|X}$), then $\hat{X}$ is a surrogate for X. There are a number of ways to relate the measured $\hat{X}$ to the true variable X. In practice, two simple error structures are commonly used: classical measurement error model and Berkson measurement error model.

Classical measurement error model (Carroll et al., 2006b). The standard statistical model for the case in which $\hat{X}$ is a surrogate for X is the additive model $\hat{X} = X$+U, where U has mean zero and is independent of X. This is often called the classical error model. Since E($\hat{X} \mid X$) = X, $\hat{X}$ is unbiased measurement of X. Therefore, the classical error model is an independent, unbiased, additive measurement error model. It is worth noting that not all measuring methods can yield unbiased measurements. However, it is often possible to obtain an unbiased measurement via calibration.

Berkson measurement error model (Berkson, 1950). If X varies around $\hat{X}$, then the statistical model $X = \hat{X} + \mathrm{U}$ is appropriate. Here, U has mean zero and is independent of $\hat{X}$. This is the so called Berkson error model. For this model, we can see that $\mathrm{E}(X \mid \hat{X}) = \hat{X}$, and $\hat{X}$ is called an unbiased Berkson predictor of X. This error model is well suited to the experimental situations in which the observed variable $\hat{X}$ is under rigorous control, and the true experimental condition X is an error-free variable and varies around $\hat{X}$. Consequently, the unbiased Berkson error model is hardly suitable for sampling design or direct measurement. Nonetheless, it is possible to calibrate the biased measurement so that the assumption of the Berkson error model is satisfied.

One reason that the classical and Berkson error models are popular is due to the fact that many error structures can be transformed from one to the other. Suppose that $\hat{X}^*$ is a surrogate for X, $\hat{X}^* = \gamma_1 + \gamma_x X + U^*$, where U^* is independent of X, then the transformed variable $\hat{X} = \frac{\hat{X}^* - \gamma_1}{\gamma_x}$ satisfies the classical error model $\hat{X} = X + \mathrm{U}$ (where $\mathrm{U} = \frac{U^*}{\gamma_x}$). Alternatively, the surrogate $\hat{X}^*$ can be transformed to an unbiased additive Berkson error structure through the transformation $\hat{X} = \mathrm{E}(X|\hat{X}^*)$. The transformation that changes an uncalibrated surrogate $\hat{X}^*$ into a classical error model is called error calibration, while the one that maps $\hat{X}^*$ into a Berkson error model is called regression calibration (Carroll et al., 2006a).

In spite of appealing properties of classical and Berkson error models, the measurement error in the current study does not arise from additive errors in the measurement process. On the contrary, it originates from the sampling error inherent to the probability-based

sampling design used to sample the times when phone calls should be made, and from the random effect modeling procedure to estimate the probability of answering a phone call for subject i at time t. Therefore, more complicated measurement error models should be considered.

1.6 DESIGN-BASED ALGORITHM

The goal of this dissertation is to estimate the effects of the proportions of time participants spent in each location context on the risk of *M. tuberculosis* conversion. Consider substituting design-unbiased estimators $\widehat{X}_l(T)$ in expression (1.6) into the linear probability model in (1.9) to obtain

$$E\big(Y_i\big|X_i(T)\big) = \Pr\big(Y_i = 1\big|X_i(T)\big) = \beta^T \widehat{X}_l(T); \qquad i = 1, \dots, n. \tag{1.10}$$

In the current context, direct application of ordinary least squares (OLS) for parameter estimation appears to be inappropriate due to the following reasons. In the first place, the linear probability model in (1.10) violates the homoscedasticity assumption because its variance is not constant. Goldberger suggested estimating the parameters using weighted least squares (WLS) to achieve homoscedasticity (Goldberger, 1964). In the second place, since the response variable Y_i is a binary variable, we need to ensure the probabilities (1.10) lie within the unit interval [0, 1]. Therefore, the linear probability model in (1.10) is subjected to the constraints $0 \le \beta_j \le 1 \; for \; j = 1, \dots, p$. In addition, we proposed the constrained WLS as an alternative option to constrained OLS in an attempt to address the issue of heteroscedasticity,

$$L(\beta) = \sum_{i=1}^{n} w_i [y_i - \beta^T X_i(T)]^2 \,, \tag{1.11}$$

where $w_i = \frac{1}{\pi_i(1-\pi_i)}$, $\pi_i = \Pr\big(Y_i = 1\big|X_i(T)\big)$, and $0 \le \beta_j \le 1 \; for \; j = 1, \dots, p$.

The predictors $X_i(T)$ in expression (1.11) should be replaced with the design-unbiased estimators $\widehat{X}_\iota(T)$ since they are unobservable. In the current study, the error inherent in $\widehat{X}_\iota(T)$ is attributed to sampling error arises from the probability-model used to select random sampling times as well as from the random effect modeling. As a result, substitution $\widehat{X}_\iota(T)$ for $X_i(T)$ may result in biased estimating equations because

$$E\{\widehat{X}_\iota(A_i)\,\widehat{X}_\iota(A_i)^{\,T}\} = X_i(T)X_i(T)^T + \frac{1}{T^2}\int_0^T \frac{x_i(t)x_i(t)^T}{\lambda_i(t)}dt.$$

However, we can fix this problem by replacing the bias quadratic term $\widehat{X}_\iota(A_i)\,\widehat{X}_\iota(A_i)^{\,T}$ with the design-unbiased estimators

$$\frac{1}{T^2}\sum_{s\neq t\in N_i}\frac{x_i(s)x_i(t)^T}{\lambda_i(s)\lambda_i(t)^T}\,.$$

To facilitate the computation, we can use the following equation to substitute the above quantity,

$$\widehat{X}_\iota(T)\,\widehat{X}_\iota(T)^{\,T} - \frac{1}{T^2}\sum_{t\in N_i}\frac{x_i(t)x_i(t)^T}{\lambda_i^2(t)}\,. \qquad (1.12)$$

To obtain the unbiased estimates of β using the constrained WLS is equivalent to find the solution which minimizes

$$L^*(\beta) = \sum_{i=1}^n w_i\{{y_i}^2 - 2y_i\widehat{X}_\iota(T)^{\,T}\beta + \beta^T[\widehat{X}_\iota(T)\,\widehat{X}_\iota(T)^{\,T} - \frac{1}{T^2}\sum_{t\in N_i}\frac{x_i(t)x_i(t)^T}{\lambda_i^2(t)}]\beta\}. \quad (1.13)$$

where $0 \le \beta_j \le 1\ for\ j = 1, \dots, p.$

Since direct application of OLS or WLS without any restrictions may yield nonsensible predictive values, we will employ constrained optimization algorithms to obtain meaningful estimates in the current study. So far, a wide variety of optimization algorithms including trust-region (Moré, Garbow and Hillstrom, 1981; Moré and Sorensen, 1983), conjugation gradient (Beale, 1972), Newton-Raphson (Koziel and

Yang, 2011), Nelder-Mead Simplex (Nelder and Mead, 1965), interior point (Mehrotra, 1992), and quasi-Newton methods (Fletcher and Powell, 1963), have been proposed and developed statistical packages for solving the optimized parameters. Each optimization technique requires a continuous objective function. The majority of optimization algorithms require continuous first- and/or second-order derivatives of the objective function, but some of them are derivative free such as Simplex method. In spite of a wide availability of constrained optimization algorithms, no single one is invariably superior to others. In the current study, we employed the dual quasi-Newton technique (Powell, 1982a, b) to obtain constrained OLS and WLS estimates. All data analyses were performed in SAS 9.4 (SAS Inst., Cary, NC). The NLP procedure as well as PROC IML language were used to constrained parameter estimates.

1.7 COHSONET STUDY OF RISK OF TUBERCULOSIS CONVERSION

We illustrate results of random effects modeling of the probability of answering a phone call at selected times here using the COHSONET data. The random effects model (1.8) with $K = 4$ was used to describe the periodic pattern of answering phone calls. Estimates of parameters in the random effects model for missing data patterns were obtain using a Fortran program available in the supplementary material of Rathbun and Shiffman (2016). Each subject was scheduled to receive 200 phone calls over one-year study interval. Observations from subjects receiving too few phone calls may not be reliable, and thus we considered only subjects received more than 30 phone calls. Consequently, a total of 288 subjects were included in the mixed effects analysis. Table 1.1 presents the estimated mean levels of the random effects in the random effects model. The large standard deviation in the intercept indicates that there is considerable variation among

subjects in the rates at which they answered the phone calls. Large standard deviations in the remaining terms would suggest that there is considerable variation among subjects in the patterns in which calls were answered. Figure 1.1 plots the expected probability of answering phone calls as a function time as estimated from Steele's (1996) modified EM algorithm as obtained from the mean μ of the random effects γ. On average, participants were most likely to respond to the phone calls in the early morning (i.e., 7:00 am - 8:00 am). There appeared to be an increasing trend between 9:00am and 7:00pm and a sharp decreasing trend between 8:00pm to 11:00pm, with subjects being most likely to answer phone calls at 7:00pm, and least likely to answer phone calls at the end of a day.

Table 1.1 Mean levels of random effects describing the periodic pattern of the probability of answering phone calls (n=288).

Variable	Estimate	Standard Deviation	P-Value
Intercept	0.565	0.402	0.159
sin(1*2πt/24)	0.894	0.444	0.044
cos(1*2πt/24)	-0.626	0.589	0.288
sin(2*2πt/24)	-0.651	0.475	0.171
cos(2*2πt/24)	-1.327	0.288	0.000
sin(3*2πt/24)	-0.105	0.239	0.660
cos(3*2πt/24)	-1.044	0.223	0.000
sin(4*2πt/24)	-0.236	0.087	0.007
cos(4*2πt/24)	-0.288	0.102	0.005

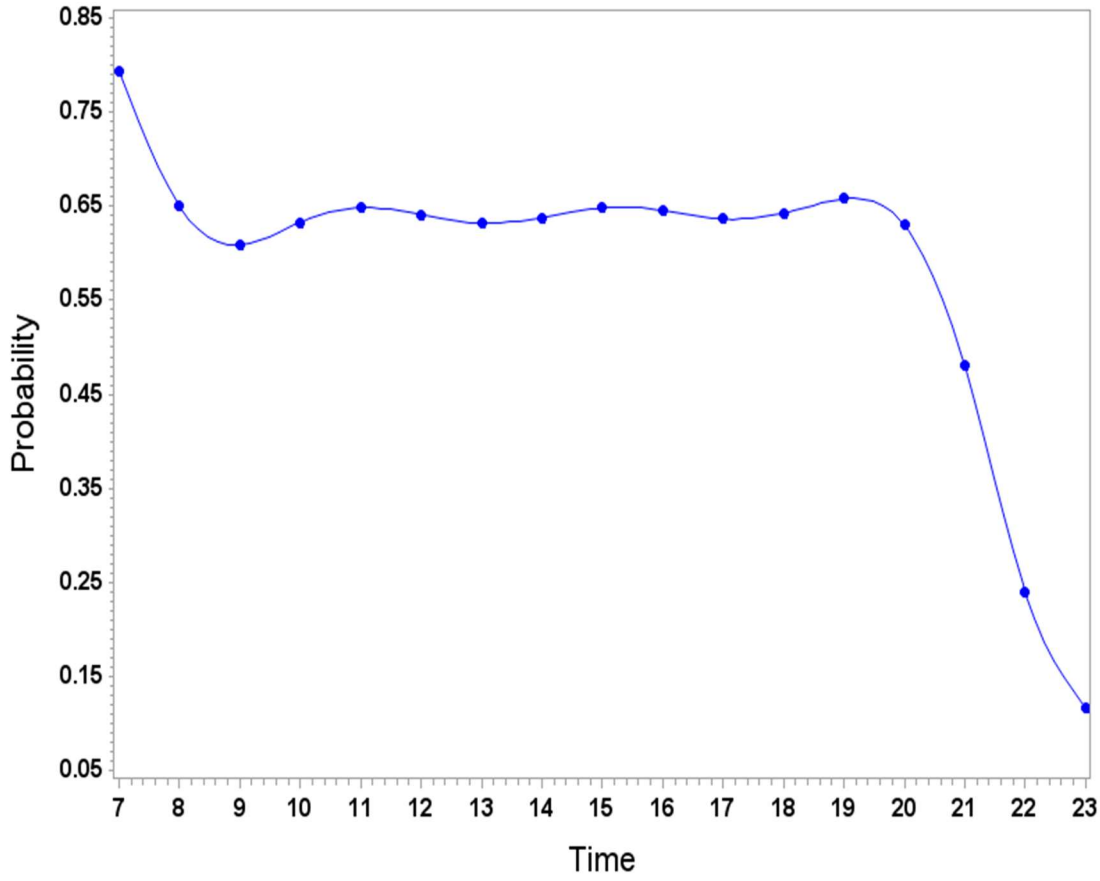

Figure 1.1 Random effects modeling of probability of answering the phone calls (n=288).

To adjust for effects of missing data, the design-unbiased estimators in expression (1.6) was used to estimate the proportions of time each participant spent in each location context (Figure 1.2). On average, participants spent the most time at homes (i.e. 32.4%), followed by work places (32.1%), public transports (7.1%), and shopping centers (4.1%). It seems that participants in the COHSONET study rarely spent time at women groups, gyms/recreations, clubs, schools, neighbors' homes and hospital (less than 1%).

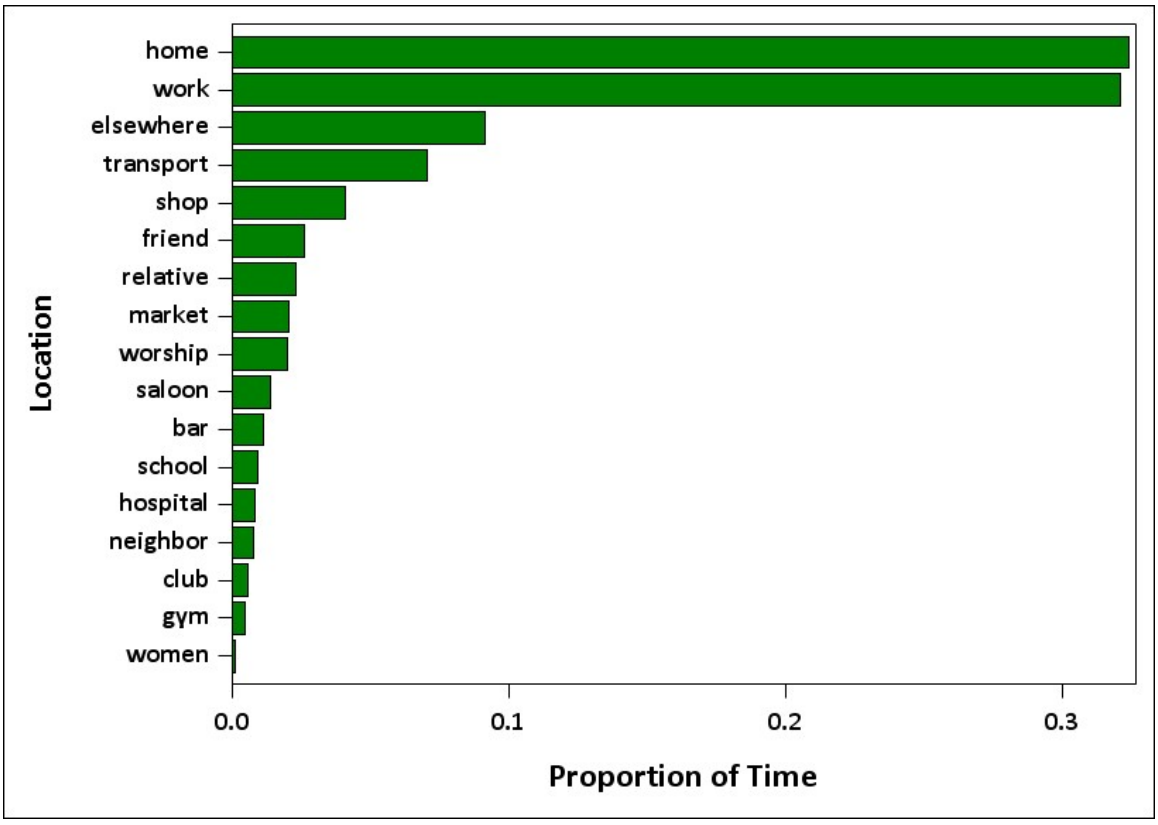

Figure 1.2 Mean proportions of time spent at each location context (n=288).

We used the constrained linear probability model to estimate the effects of the proportions of time participants spent in each location context on the risk of *M. tuberculosis* conversion. Note that the statistical modeling was based upon complete cases with both *M. tuberculosis* conversion information and proportions of time spent in each location context. Therefore, only 189 subjects with complete information were included in the constrained linear probability models. We concerned that the biased quadratic term $\widehat{X}_\iota(T)\ \widehat{X}_\iota(T)^{\ T}$ in the normal equation may lead to substantial difference in the parameter estimates, so we evaluated modeling results in both cases: (1) Biased estimator in which no manipulation was made to correct the potential biasness due to the

biased quadratic term; (2) Bias-corrected estimator in which we substituted the biased quadratic term with the design-unbiased estimators in expression (1.12).

Table 1.2 presents parameter estimates in the linear probability model over all location contexts using constrained OLS and WLS approaches. In general, there appeared to be no substantial difference between biased estimates and bias-corrected estimates of the constrained OLS: both estimates indicated that there was almost no risk of *M. tuberculosis* conversion if people spent all their time at such location contexts as work places, clubs, saloons, gyms, hospitals, women groups, market places, and neighbors' homes. On the contrary, the risk of *M. tuberculosis* conversion would be increased by 100% if a person spent all their time at schools, worship centers, bars, and shopping and trading centers. Among constrained OLS biased estimates with values within the interval [0, 1], it appeared that the bias-corrected estimates tended to shrink them toward the closest boundary of constraints. Regarding constrained WLS estimates, there appeared to be no obvious relationship between biased and bias-corrected estimates. The constrained WLS bias-corrected approach tended to shrink more parameters toward the boundary. With respect to the bias-corrected estimate, the constrained OLS and WLS yielded the same or close estimates among most of location contexts except at friend's home and public transport. For example, the constrained OLS revealed that friend's homes were very dangerous location, while the constrained WLS indicated that the risk of contracting M. tuberculosis only increased by 13.74% if people spent all the time at friend's homes. Finally, we failed to find any substantial associations in the biased estimates between the constrained OLS and WLS approach.

Table 1.2 Parameter estimates in the constrained linear probability model over all location contexts (n=189)

	Constrained OLS		Constrained WLS	
Location context	**Biased estimator#**	**Bias-corrected estimator#**	**Biased estimator#**	**Bias-corrected estimator#**
Home	0.2905	0.2722	0.3913	0.4291
Friend's home	0.5472	1.0000	0.4413	0.1374
Relative's home	0.0139	0.0000	0.2517	0.0000
Work	0.0000	0.0000	0.1551	0.0992
School	1.0000	1.0000	0.7560	0.6994
Worship center	1.0000	1.0000	0.9629	1.0000
Club/Association	0.0000	0.0000	0.0000	0.0000
Bar	1.0000	1.0000	0.9983	1.0000
Saloon	0.0000	0.0000	0.1470	0.0000
Gym/Recreation	0.0000	0.0000	0.0000	0.0000
Hospital/Clinic	0.0000	0.0000	0.0000	0.0000
Shopping/Trading center	1.0000	1.0000	0.8824	1.0000
Public transport	0.5007	0.3245	0.6370	0.8346
Women group	0.0000	0.0000	0.0167	0.0000
Market place	0.0000	0.0000	0.0000	0.0000
Neighbor's home	0.0000	0.0000	0.0885	0.0000
Elsewhere	0.6434	0.8614	0.7186	0.8077

Biased estimator: no manipulation was made on the biased quadratic term $\widehat{X}_{\iota}(T)\ \widehat{X}_{\iota}(T)^{\ T}$;
Bias-corrected estimator: the biased quadratic term was replaced with the design-unbiased quantity in expression (1.12).

We were concerned that including all location contexts in the linear probability model might lead to an overparameterized model. Therefore, we fitted another statistical modeling based on reduced location contexts by pooling friend's, relative's, and neighbor's homes into one location context as "Other home", and pooling the rest of locations other than home, other home, work place, and public transport as "Elsewhere". Based on results shown as in Table 1.3, the greatest number of location contexts were found with zero risk of *M. tuberculosis* conversion through the constrained OLS bias-corrected estimators. The biased estimator and bias-corrected estimator yielded similar estimates at homes in the constrained OLS. Nonetheless, parameter estimates at other homes and public transports were different between two estimators of the constrained OLS. In contrast, the biased estimator and bias-corrected estimator of constrained WLS appeared to yield very similar parameter estimates when the number of location contexts were reduced. Moreover, constrained OLS biased estimates tended to be similar to both biased and bias-corrected estimates of the constrained WLS.

Table 1.3 Parameter estimates in the constrained linear probability model over reduced location contexts (n=189).

	Constrained OLS		Constrained WLS	
Location context	**Biased estimator#**	**Bias-corrected estimator#**	**Biased estimator#**	**Bias-corrected estimator#**
Home	0.2575	0.2481	0.3653	0.3814
Other home*	0.3022	0.0000	0.4161	0.2475
Work	0.0000	0.0000	0.1577	0.1063
Public transport	0.4129	0.0000	0.5742	0.6993
New elsewhere*	0.6823	0.9432	0.6559	0.7305

* Other home is a pooled location of "Friend's home", "Relative's home", and "Neighbor's home"; New elsewhere is a pooled location of "School", "Worship center", "Club/Association", "Bar", "Saloon", "Gym/Recreation", "Hospital/Clinic", "Shopping/Trading center", "Women group", "Market place" and "Elsewhere" in table 1.2.

Biased estimator: no manipulation was made on the biased quadratic term $\widehat{X}_i(T)\,\widehat{X}_i(T)^T$; Bias-corrected estimator: the biased quadratic term was replaced with the design-unbiased quantity in expression (1.12).

1.8 DISSERTATION OUTLINE

This dissertation aims to build statistical models to describe the risk of *M. tuberculosis* conversion as a function of proportions of time participants spent in different location contexts. We offered the linear probability model as an alternative to the logistic regression due to potential bias from the parameter estimate in the logistic regression. So as to restrict the parameter estimates to lie on the interval [0,1], we will propose two different constrained optimization approaches. In chapter 2, we will demonstrate the asymptotic properties of the constrained OLS and WLS in the linear probability model, and employ simulation studies to investigate the finite-sample properties of the proposed approaches and compare their performance with the logistic regression model. Since the constrained OLS or WLS in chapter 2 does not allow model selection, we will propose

the constrained adaptive LASSO as an alternative optimization approach. Similar to chapter 2, we will demonstrate the asymptotic properties of constrained adaptive LASSO estimates, and use simulation studies to explore the finite-sample properties.

1.9 REFERENCES

Global Tuberculosis Report 2014. World Health Organization Geneva.

Barnes PF, el-Hajj H, Preston-Martin S, Cave MD, Jones BE, Otaya M, Pogoda J, and Eisenach KD. 1996. Transmission of tuberculosis among the urban homeless. JAMA 275(4):305-307.

Barta WD, Portnoy DB, Kiene SM, Tennen H, Abu-Hasaballah KS, and Ferrer R. 2008. A daily process investigation of alcohol-involved sexual risk behavior among economically disadvantaged problem drinkers living with HIV/AIDS. AIDS Behav 12(5):729-740.

Barta WD, Tennen H, and Kiene SM. 2010. Alcohol-involved sexual risk behavior among heavy drinkers living with HIV/AIDS: negative affect, self-efficacy, and sexual craving. Psychol Addict Behav 24(4):563-570.

Beale EML. 1972. Numerical Methods for Nonlinear Optimization. London: Academic Press.

Berkson J. 1950. Are there two regressions? Journal of American Statistical Association 45:164-180.

Bhatt K, and Salgame P. 2007. Host innate immune response to Mycobacterium tuberculosis. J Clin Immunol 27(4):347-362.

Breslow NE, and Clayton DG. 1993. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association 88(421):9-25.

Carels RA, Douglass OM, Cacciapaglia HM, and O'Brien WH. 2004. An ecological momentary assessment of relapse crises in dieting. J Consult Clin Psychol 72(2):341-348.

Carroll RJ, Midthune D, Freedman LS, and Kipnis V. 2006a. Seemingly unrelated measurement error models, with application to nutritional epidemiology. Biometrics 62(1):75-84.

Carroll RJ, Ruppert D, Stefanski LA, and Crainiceanu CM. 2006b. Measurement Error in Nonlinear Models: A Modern Perspective. New York: Chapman and Hall/CRC

Classen CN, Warren R, Richardson M, Hauman JH, Gie RP, Ellis JH, van Helden PD, and Beyers N. 1999. Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. Thorax 54(2):136-140.

Cressie N. 1991. Statistics for Spatial Data. New York: Wiley.

Dooley SW, Villarino ME, Lawrence M, Salinas L, Amil S, Rullan JV, Jarvis WR, Bloch AB, and Cauthen GM. 1992. Nosocomial transmission of tuberculosis in a hospital unit for HIV-infected patients. JAMA 267(19):2632-2634.

Feldman SI, Downey G, and Schaffer-Neitz R. 1999. Pain, negative mood, and perceived support in chronic pain patients: a daily diary study of people with reflex sympathetic dystrophy syndrome. J Consult Clin Psychol 67(5):776-785.

Firth D. 1993. Bias reduction of maximum likelihood estimates. Biometrika 80(1):27-38.

Fletcher R, and Powell MJD. 1963. A Rapidly Convergent Descent Method for Minimizatio. Computer Journal 6:163-168.

Goldberger AS. 1964. Econometric Theory. New York: John Wiley.

Guwatudde D, Nakakeeto M, Jones-Lopez EC, Maganda A, Chiunda A, Mugerwa RD, Ellner JJ, Bukenya G, and Whalen CC. 2003. Tuberculosis in household contacts of infectious cases in Kampala, Uganda. Am J Epidemiol 158(9):887-898.

Horby P, Pham QT, Hens N, Nguyen TT, Le QM, Dang DT, Nguyen ML, Nguyen TH, Alexander N, Edmunds WJ et al. . 2011. Social contact patterns in Vietnam and implications for the control of infectious diseases. PLoS One 6(2):e16965.

Houk VN, Baker JH, Sorensen K, and Kent DC. 1968. The epidemiology of tuberculosis infection in a closed environment. Arch Environ Health 16(1):26-35.

Isham V, and M. W. 1979. A self-correcting point process. Stochastic Processes and their Applications 8(3):335-347.

Kamarck TW, Muldoon MF, Shiffman SS, and Sutton-Tyrrell K. 2007. Experiences of demand and control during daily life are predictors of carotid atherosclerotic progression among healthy men. Health Psychol 26(3):324-332.

Kenyon TA, Valway SE, Ihle WW, Onorato IM, and Castro KG. 1996. Transmission of multidrug-resistant Mycobacterium tuberculosis during a long airplane flight. N Engl J Med 334(15):933-938.

Koziel S, and Yang XS. 2011. Computational Optimization, Methods and Algorithms. Poland: Springer-Verlag Berlin Heidelberg.

Lawless JF. 1987. Regression Methods for Poisson Process Data. Journal of the American Statistical Association 82(399):808-815.

Lee Y, and Nelder JA. 1996. Hierarchical generalized linear models. Journal of the Royal Statistical Society B 58:619-678.

Lienhardt C, Fielding K, Sillah J, Tunkara A, Donkor S, Manneh K, Warndorff D, McAdam KP, and Bennett S. 2003a. Risk factors for tuberculosis infection in sub-Saharan Africa: a contact study in The Gambia. Am J Respir Crit Care Med 168(4):448-455.

Lienhardt C, Sillah J, Fielding K, Donkor S, Manneh K, Warndorff D, Bennett S, and McAdam K. 2003b. Risk factors for tuberculosis infection in children in contact with infectious tuberculosis cases in the Gambia, West Africa. Pediatrics 111(5 Pt 1):e608-614.

Mehrotra S. 1992. On the Implementation of a Primal-Dual Interior Point Method. SIAM J Optim 2(4):575–601.

Moré JJ, Garbow BS, and Hillstrom KE. 1981. Testing Unconstrained Optimization Software. ACM Transactions on Mathematical Software 7:17--41.

Moré JJ, and Sorensen DC. 1983. Computing a Trust-Region Step. SIAM Journal on Scientific and Statistical Computing 4:553--572.

Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, Massari M, Salmaso S, Tomba GS, Wallinga J et al. . 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5(3):e74.

Mustanski B. 2007. The influence of state and trait affect on HIV risk behaviors: a daily diary study of MSM. Health Psychol 26(5):618-626.

Nelder JA, and Mead R. 1965. A Simplex Method for Function Minimization. Computer Journal 7:308-313.

Ogata Y, and Vere-Jones D. 1984. Inference for earthquake models: a self-correcting model. Stoch Proc Appl 17:337-347.

Powell MJD. 1982a. Extensions to Subroutine VF02AD. In: R. F. Drenick and F. Kozin e, editor. Systems Modeling and Optimization, Lecture Notes in Control and Information Sciences Berlin-Heidelberg-New York: Springer-Verlag.

Powell MJD. 1982b. VMCWD: A Fortran Subroutine for Constrained Optimization. DAMTP 1982/NA4. Cambridge, England.

Rehkopf D, Furumoto-Dawson A, Kiszewski A, and Awerbuch-Friedlander T. 2015. Spatial Spread of Tuberculosis through Neighborhoods Segregated by Socioeconomic Position: A Stochastic Automata Model. Discrete Dynamics in Nature and Society 2015.

Reichler MR, Reves R, Bur S, Thompson V, Mangura BT, Ford J, Valway SE, Onorato IM, and Contact Investigation Study G. 2002. Evaluation of investigations conducted to detect and prevent transmission of tuberculosis. JAMA 287(8):991-995.

Riley RL, Mills CC, O'Grady F, Sultan LU, Wittstadt F, and Shivpuri DN. 1962. Infectiousness of air from a tuberculosis ward. Ultraviolet irradiation of infected air: comparative infectiousness of different patients. Am Rev Respir Dis 85:511-525.

Roach DR, Bean AG, Demangel C, France MP, Briscoe H, and Britton WJ. 2002. TNF regulates chemokine induction essential for cell recruitment, granuloma formation, and clearance of mycobacterial infection. J Immunol 168(9):4620-4627.

Shiffman S, Gwaltney CJ, Balabanis MH, Liu KS, Paty JA, Kassel JD, Hickcox M, and Gnys M. 2002. Immediate antecedents of cigarette smoking: an analysis from ecological momentary assessment. J Abnorm Psychol 111(4):531-545.

Shiffman S, Stone AA, and Hufford MR. 2008. Ecological momentary assessment. Annual Review of Clinical Psychology 4:1-32.

Steele BM. 1996. A modified EM algorithm for estimation in generalized mixed models. Biometrics 52(4):1295-1310.

Stone AA, and Shiffman S. 2002. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 24(3):236-243.

Sultan L, Nyka W, Mills C, O'Grady F, Wells W, and Riley RL. 1960. Tuberculosis disseminators. A study of the variability of aerial infectivity of tuberculous patients. Am Rev Respir Dis 82:358-369.

Thall PF. 1988. Mixed Poisson Likelihood Regression Models for Longitudinal Interval Count Data. Biometrics 44(1):197-209.

Vere-Jones D, and Ogata Y. 1984. On the moments of a self-correcting process. J Appl Prob 21:335-342.

Whalen CC, Zalwango S, Chiunda A, Malone L, Eisenach K, Joloba M, Boom WH, and Mugerwa R. 2011. Secondary attack rate of tuberculosis in urban households in Kampala, Uganda. PLoS One 6(2):e16137.

Wray TB, Merrill JE, and Monti PM. 2014. Using Ecological Momentary Assessment (EMA) to Assess Situation-Level Predictors of Alcohol Use and Alcohol-Related Consequences. Alcohol Res 36(1):19-27.

Yaganehdoost A, Graviss EA, Ross MW, Adams GJ, Ramaswamy S, Wanger A, Frothingham R, Soini H, and Musser JM. 1999. Complex transmission dynamics of clonally related virulent Mycobacterium tuberculosis associated with barhopping by predominantly human immunodeficiency virus-positive gay men. J Infect Dis 180(4):1245-1251.

Yang C, Linas B, Kirk G, Bollinger R, Chang L, Chander G, Siconolfi D, Braxton S, Rudolph A, and Latkin C. 2015. Feasibility and Acceptability of Smartphone-Based Ecological Momentary Assessment of Alcohol Use Among African American Men Who Have Sex With Men in Baltimore. JMIR Mhealth Uhealth 3(2):e67.

CHAPTER 2

CONSTRAINED LINEAR PROBABILITY MODEL IN DETERMINNING THE PROBABILITY OF TUBERCULOSIS CONVERSION FROM ECOLOGICAL MOMENTATY ASSESSMENT OF SOCIAL PATTERNS

2.1 INTRODUCTION

Tuberculosis is an infectious disease that threatens the health of people all over the world, and it is most prevalent in resource-limited countries such as the developing countries in Africa (WHO 2014). Although transmission in certain common locations such as households, bars, and neighborhood clinics has been investigated (Dooley et al. 1992; Reichler et al. 2002; Yaganehdoost et al. 1999), the relative risks of transmission as a function of locations where susceptible people spend their time has yet to be quantified. The Community Health Study of Social Networks and Tuberculosis (COHSONET), is an ongoing study in Uganda aiming to evaluate the effects of social networks on the transmission dynamics of *Mycobacterium tuberculosis*. Disease-free participants are followed up for one year in an attempt to estimate risk of contracting tuberculosis as a function of the proportions of time spent in different location contexts (e.g., home, work, etc.). Time spent in different settings is quantified using Ecological Momentary Assessment (EMA), a method in the behavioral sciences that enables evaluation of subjects' emotional states and environments through repeated sampling in their every-day environments using electronic devices (Shiffman et al. 2008; Stone and Shiffman 2002).

In the COHSONET study, participants are contacted on their cell phones at randomly selected times to answer questions regarding their current location contexts at the times of each call. By selecting sampling times generated from a known probability-based sampling design, design-unbiased estimates (Cassel et al. 1977) of the proportions of time participants spend in each setting can be obtained.

In the COHSONET study, the outcome variable indicating if participants have contracted *M. tuberculosis* is dichotomous, suggesting the application of logistic regression to predict this outcome as a function of location contexts. Since logistic regression is nonlinear, however, maximum likelihood estimators (MLEs) have a bias of order $O(n^{-1})$ (Firth 1993), a bias that has been observed in a number of practical applications (Firth 1993; Gart and Zwifel 1967; Hirji et al. 1987; Park and Park 2003; Wagler 2011). Moreover, MLEs behave poorly in presence of separation (Albert and Anderson 1984; Heinze and Schemper 2002; Hirji et al. 1987; Kolassa 1997; Lesaffre and Albert 1989; Zorn 2005). Separation occurs when one or more covariates in a logistic regression model perfectly predicts the binary outcome, which is associated with infinite coefficients and standard errors. Firth (1993) proposed a method to reduce bias in the parameter estimates in the logistic regression model through penalizing the likelihood using the Jeffreys invariant prior. A body of evidence demonstrated superiority of penalized maximum likelihood based models at a small to moderate sample size, and in existence of separation (Heinze and Schemper 2002; Wagler 2011; Zorn 2005). We suspect that a small portion of participants are likely to spend most of time in some uncommon settings (e.g., bar) but which may tend to have very high/low risk of contracting *M. tuberculosis*. In this case, penalized likelihood estimators in the logistic regression should be more

appropriate than MLEs for determining the risk of *M. tuberculosis* infection as a function of proportions of time spent at each location

Linear probability model is another option for regression models with dichotomous dependent variable. There is evidence suggesting that the ordinary least squares (OLS) estimator in the linear probability model performs as well as or even outperform maximum likelihood estimates in the logistic regression model in certain situations (Deke 2014; Hellevik 2009; Pedroza and Troung 2016). Several arguments contribute to decreasing popularity of linear probability model with binary dependent outcome. One concern of using linear regression model is obtaining meaningless predicted probability which falls outside the unit interval [0, 1]. Another limitation is that the OLS estimator in the linear regression model violate assumptions of heterogeneity and normality as in classic linear regression model. In spite of these drawbacks, the linear probability model has a striking merit of ease of interpretation. The coefficient estimates in the linear probability model can be directly interpreted as the mean marginal effect of covariates on the outcome.

Over the years, several researchers have been attracted to the linear probability model and have proposed different approaches to address problems as aforementioned. Goldberger (1964) suggested estimating the probability by OLS and then re-estimating the model by weighted least squares (WLS) to resolve the issue of heteroscedasticity (Goldberger 1964). Unfortunately, Goldberger's method fails to prevent production of inappropriate predicted probability, and thus the WLS estimation breaks down. To fix this problem, Goldfeld and Quandt (1972) proposed only using those observations having OLS estimates between 0 and 1 to do the WLS estimation (Goldfeld and Quandt 1972).

Hensher and Johnson (1981) proposed bounding the weights and assigning negative weights a constant value (Hensher and Johnson 1981). Mullahy (1990) proposed a quasi-generalized least squares estimator which is a generalization of the Goldfeldt-Quandt and Hensher-Johnson estimators (Mullahy 1990). However, one annoying limitation of these methods is that different options of weighting strategies may lead to different conclusions and none of them guarantee predictive probability falls inside the unite interval. In this paper, we propose using constraints to force estimates within a meaningful range.

The primary purpose of this paper is to compare the performance of constrained linear probability model with logistic regression model in situations with a binary response. In consideration of potential heterogeneity, we calculated and compared the behaviors of both inequality constrained least squares to weighted least squares estimators. Moreover, we studied the performance of bias-reduced penalized maximum likelihood in the logistic regression.

The contents are distributed in five sections including the introduction. Section 2.2 is devoted to describing the constrained linear regression model and asymptotic properties of constrained estimators. In section 2.3, we demonstrate results of the simulation studies. We apply the proposed methods to the COHSONET study in Section 2.4. In section 2.5, we give a discussion of the results.

2.2 CONSTRAINED LINEAR PROBABILITY MODEL

The aim of COHSONET study is to estimate the risk of *M. tuberculosis* conversion as a function of proportions of time participants spent in different location contexts. Suppose that n subjects are randomly sampled and each subject i is observed over a set of times

belonging to the Borel set $T \subset \mathbb{R}$ with Lebesgue measure $|T| < \infty$. In the current context, the sampling domain T is the union of finite number of disjoint time intervals corresponding to the set of times when the phone called is made, and $|T|$ is the total length of time subject i is observed. Let $X_i(T)$ denote a $p \times 1$ vector corresponding to the proportions of time participants spent in each location context. The elements of the vector $X_i(T)$ are between zero and one, and add up to one. Let Y_i denote a Bernoulli dependent variable corresponding to *M. tuberculosis* conversion status which is observed over the time interval $[0, T]$. Hence, the linear probability model considered is

$$E\big(Y_i \big| X_i(T)\big) = \Pr\big(Y_i = 1 \big| X_i(T)\big) = \beta^T X_i(T); \qquad i = 1, \dots, n. \tag{2.1}$$

where $0 \leq \beta_j \leq 1 \; for \; j = 1, \dots, p$. Note that the linear probability model considered here does not contain an intercept term. The interpretation of coefficients in the linear probability model is straightforward. The coefficient β_j represents the risk of *M. tuberculosis* conversion if the participants spent 100% of time in location j; $j = 1, \dots, p$. Note that the constraints on the parameters and the predictors imply the $0 \leq \beta^T X_i(T) \leq 1$ for all subjects $i = 1,2, \cdots, n$.

2.2.1 Fully Observed Location Contexts

To motivate the proposed methods of statistical inference, we temporarily assume in this subsection that the vectors of location context proportions $X_i(T)$ are known for all participants $i = 1, \dots, n$. Define the objective function of the OLS estimator for the constrained linear probability model in expression (2.1) as

$$Q(\beta) = \sum_{i=1}^{n} [y_i - \beta^T X_i(T)]^2 \; . \tag{2.2}$$

Note that the OLS estimator is unbiased for finite samples, a property that is not shared by logistic regression whose maximum likelihood estimators has bias of order $O(n^{-1})$

(Firth 1993). To overcome issue of heteroskedasticity, we also consider estimating parameters β in the inequality constrained linear probability model using constrained WLS, whose objective function is defined as

$$Q(\beta) = \sum_{i=1}^{n} w_i[y_i - \beta^T X_i(T)]^2 , \quad (2.3)$$

where $w_i = [\pi_i(1 - \pi_i)]^{-1}$.

Since the linear probability model in (2.1) is subject to inequality constraints, the typical optimization strategies such as Newton-Raphson algorithm are not applicable here. A wide variety of optimization approaches including sequential quadratic programming (SQP) (Nocedal and WrighT 2006), trust-region (Moré et al. 1981; Moré and Sorensen 1983), conjugation gradient (Beale 1972), Newton-Raphson (Koziel and Yang 2011), Nelder-Mead Simplex (Nelder and Mead 1965), interior point (Mehrotra 1992), and quasi-Newton methods (Fletcher and Powell 1963) have been proposed and statistical packages have been developed for each approach. Each optimization approach requires a continuous objective function, a requirement that is satisfied by both OLS and WLS estimators. The majority of optimization techniques require continuous first- and/or second-order derivatives of the objective function, but some of them are derivative free such as Simplex method. In spite of a wide availability of constrained optimization approaches, no single one is invariably superior to others. In the current study, we employed the dual quasi-Newton technique (Powell 1982a; Powell 1982b) to obtain optimized parameter estimates for the constrained linear probability model.

2.2.2 Partially Observed Location Contexts

The OLS and WLS objective functions require that the vectors of location contexts $X_i(T)$ be known for all participants. More formally, the location contexts $X_i(T)$are comprised of

the population proportions of time participants spent in each location context over the period of the EMA study, where the populations are comprised of the set of all times in the one-year study interval of each participant. Let $x_i(t)$ denote a vector of indicator variables, whose *j*-th element takes the value one if the subject *i* is in location *j* at time *t*, and the value zero if otherwise. Then the vector of population proportions is equal to the domain means

$$X_i(T) = \frac{1}{T}\int_0^T x_i(t)dt,$$

where the integral is over the study interval $[0, T]$. Evaluation of integral requires that the time-varying covariates $x_i(t)$ be known functions of time. Unless the subjects are observed 24 hours per day for 7 days per week, these domain means are unknown. If, however, EMA samples are collected at times realized from a known probability-based sampling design, then design-unbiased estimators of the domain means may be obtained.

Suppose that for subject i, the time-varying covariates $x_i(t)$ are sampled according to a temporal point process *N*(.) with conditional intensity

$$\lambda(t) = \lim_{\delta \to 0} \frac{E\{N[t,\ t+\delta] | \digamma_t\}}{\delta}, \qquad t \geq 0$$

where *N*[t, t + δ] denotes the number of events in the time interval [t, t + δ], and $\digamma_t$, the smallest σ-algebra generated by $\{N(u, t]; 0 < u \leq t\}$, represents the history of the point process *N*(.) up to time t.

If the time-varying covariates are sampled according to a point process $N_i(t)$ with known intensity $\lambda_i(t)$, and $\lambda_i(t) > 0$ for all $t \geq 0$ except on a set of measure zero, then a

design-unbiased estimator of a domain mean is

$$\hat{X}_i(T) = \frac{1}{T}\sum_{t\in N_i}\frac{x_i(t)}{\lambda_i(t)}, \tag{2.4}$$

where N_i is the set of times at which assessments were made for subject *i*. This estimator is design unbiased in the sense that its expected value is $X_i(T)$ under the probability model induced by the sampling design. To investigate the large-sample inferential properties of the estimator $\hat{X}_i(T)$, we can also write expression (2.4) as

$$\hat{X}_i(T) = \frac{1}{T}\int_0^T \frac{x_i(t)}{\lambda_i(t)}\, dN_i(t). \tag{2.5}$$

Since the location contexts $X_i(T)$ are unobservable in the current study, the design-unbiased estimators $\hat{X}_i(T)$ will be substituted into the OLS and WLS objective functions in expressions (2.2) and (2.3) to obtain the proposed parameter estimators.

To demonstrate that the design-unbiased estimator $\hat{X}_i(T)$ is consistent for the domain means $X_i(T)$ as $T \to \infty$, we consider the following assumptions (Rathbun 1996):

(A.1) The conditional intensity $\lambda(\mathrm{t})$ is greater than zero except on a set of measures zero.

(A.2) $X_i(T)$ is finite for all $T > 0$.

(A.3) $\frac{1}{T}\int_0^T \frac{x(t)x^T(t)}{\lambda(t)}dt < \infty$ for all $T > 0$.

Assumption (A.1) is necessary to demonstrate that $\hat{X}_i(T)$ is design-unbiased. Assumptions (A.2) and (A.3) allow us to bound the difference of $X_i(T) - \hat{X}_i(T)$. Given a counting process $N_i(t)$, the martingale $M_i(t)$ is defined as the difference of and its integrated intensity $\wedge_i\,(t)$

$$M_i(t) = N_i(t) - \wedge_i\,(t).$$

Note that the difference between $X_i(T)$ and $\hat{X}_i(T)$is a zero-mean martingale. The following theorem proves that the design-unbiased estimators $\hat{X}_i(T)$ converge in probability to the domain means $X_i(T)$ as $T \to \infty$.

THEOREM 1. *Suppose Assumptions (A.1) - (A.3) are satisfied, then*

$$X_i(T) - \hat{X}_i(T) \xrightarrow{P} 0$$

PROOF. The boundedness of $\lambda(t)$ and $x(t)$ implies that

$$\bar{X}_i(T) - \hat{X}_i(T) = (\frac{x_i(t)}{\lambda_i(t)} * M_i)_T = \frac{1}{T}\int_0^T \frac{x_i(t)}{\lambda_i(t)} dM_i(t),$$

is a zero-mean, $\mathcal{F}_T$ −martingale. Under Assumptions (A.1) - (A.3),

$$sup_T E\left\{\left(\frac{x_i(t)}{\lambda_i(t)} * M_i(t)\right)_T^2\right\} < \infty.$$

Hence, $(\frac{x_i(t)}{\lambda_i(t)} * M_i(t))_T$ is a square-integral function, and the process

$$\langle \frac{x_i(t)}{\lambda_i(t)} * M_i(t) \rangle_T = \frac{1}{T^2}\int_0^T \frac{x_i(t)x_i(t)^T}{\lambda_i(t)} dt$$

is the quadratic variation of $(\frac{x_i(t)}{\lambda_i(t)} * M_i(t))_T$ (Kallianpur 1980). Since the square of a square-integral martingale $\left(\frac{x_i(t)}{\lambda_i(t)} * M_i(t)\right)_T^2$ is dominated by its quadratic variation, we can apply Lenglart's inequality here (Karr 1986):

$$Pr\left\{\left|\frac{1}{T}\int_0^T \frac{x_i(t)}{\lambda_i(t)} dM_i(t)\right| \geq \varepsilon\right\} \leq \frac{\eta}{\varepsilon^2} + Pr\left\{\frac{1}{T^2}\int_0^T \frac{x_i(t)x_i(t)^T}{\lambda_i(t)} dt \geq \eta\right\},$$

for any $\varepsilon, \eta > 0$, and finite study interval $[0,\ T]$. By Chebyshev's inequality,

$$Pr\left\{\frac{1}{T^2}\int_0^T \frac{x_i(t)x_i(t)^T}{\lambda_i(t)} dt \geq \eta\right\}$$

$$\leq \frac{1}{\eta^2 T^2} E\left\{\int_0^T \frac{x_i(t)x_i(t)^T}{\lambda_i(t)} dt\right\}$$

$$\leq \frac{1}{\eta^2 T^2} E\left\{\sup_{t\in T} \frac{x_i(t)x_i(t)^T}{\lambda_i(t)}\right\}$$

By Assumptions (A.1) - (A.3), the right hand side of the above expression converges to zero as $T \to \infty$, which complete the proof that

$$X_i(T) - \hat{X}_i(T) \xrightarrow{P} 0. \blacksquare$$

2.2.3 Asymptotic Properties of Constrained Estimators

To demonstrate the asymptotic behavior of an estimator, a typical assumption is that the true value of the parameter lies in the interior of a parameter space. This assumption is convenient, but in the current context, the true parameter value may lie on the boundary of a parameter space in our constrained optimization problem. The asymptotic behaviors of estimators in constrained regression models have been examined by many investigators. Liew (1976) studied the asymptotic as well as small sample properties of inequality constrained least squares (ICLS) estimates (Liew 1976). However, inference based on Liew's method may be misleading because his expression for the covariance matrix of the ICLS estimator is contingent upon knowing which constraints are active and which are not. Geweke (1986) points out that this variance matrix is ill-posed, since in practice it is hardly possible to know which constraints will be active ahead of time (Geweke 1986). Self and Liang (1987) considered the asymptotic behaviors of maximum likelihood estimators on the boundary relying on the Chernoff regularity conditions (Chernoff 1954; Self and Liang 1987). Let $\langle x, y\rangle$ denote the standard scalar product of two vectors $x, y \in R^P$. And by $\|x\| = \langle x, x\rangle^{1/2}$ denote the Euclidean norm of vector. By

$$dist(x, S) = inf_{z\in S}||x - z||,$$

we denote the distance from a point $x \in R^P$ to the set S.

Chernoff conditions require that the parameter space $S \in R^P$ be approximated at θ_0 by a cone C_S with vertex at θ_0. More specifically, Self and Liang (1987) require that

$$\inf_{x \in C_S} || x - y || = o(\|y - \theta_0\|) \quad \forall y \in S$$

and

$$\inf_{y \in S} || x - y || = o(\|x - \theta_0\|) \quad \forall x \in C_S .$$

Recall that a cone is a set C in R^P has the property that $x \in C$ implies $\lambda(x - \theta_0) + \theta_0 \in C$, for all $\lambda \geq 0$.

Geyer (1994) pointed out that the above definition is closely related to various tangent cones used in the optimization literature. The limit sets

$$T_S(x) = \limsup_{t \downarrow 0} \frac{S - x}{t},$$

and

$$\bar{T}_S(x) = \liminf_{t \downarrow 0} \frac{S - x}{t}$$

are respectively called contingent (Bouligand) and inner cones to S at $x \in S$. A vector ϑ lies in the contingent tangent cone if and only if there exist a sequence t_n decreasing to 0 and a sequence x_n in S converging to x such that $\frac{x_n - x}{t_n} \to \vartheta$. Similarly, a vector ϑ lies in the inner tangent cone if and only if for every sequence t_n decreasing to 0 there exists a sequence x_n in S converging to x such that $\frac{x_n - x}{t_n} \to \vartheta$. By definition of limits superior and inferior, we have $\bar{T}_S(x) \subset T_S(x)$. Both contingent and inner cones are closed. The parameter set S is Chernoff regular at a point $x \in S$ if $\bar{T}_S(x) = T_S(x)$ (Geyer 1994). Figure 2.1 illustrates a tangent cone which is Chernoff regular at a point x, from which we can observe that contingent and inner cones coincide with each other.

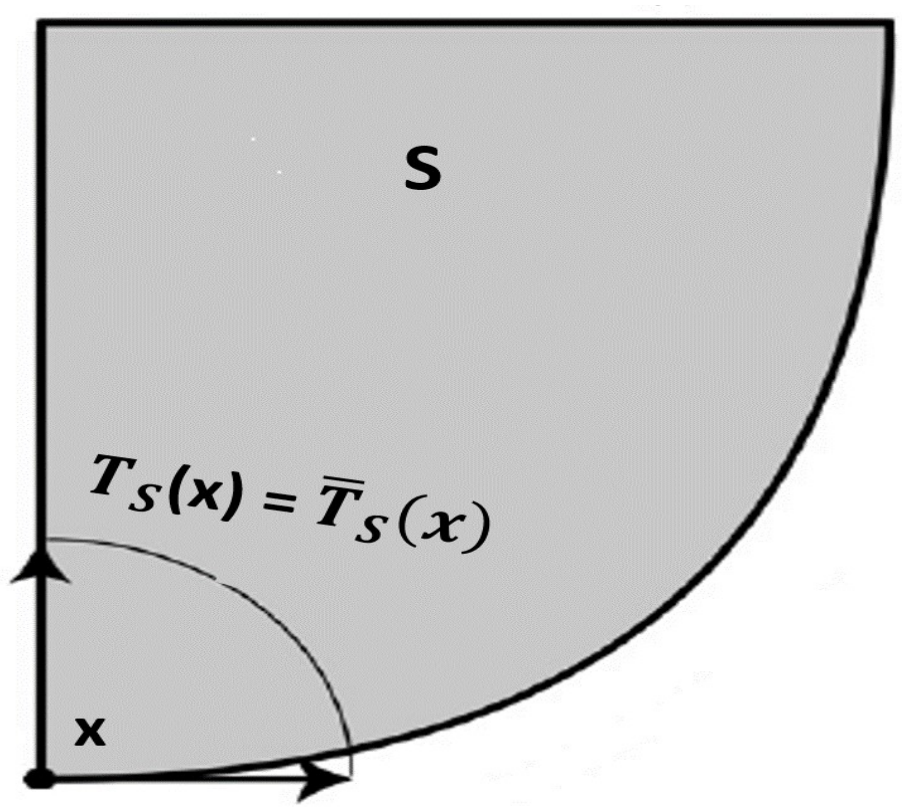

Figure 2.1 An illustration of tangent cone

Let Y denote a random variable sampled from a probability distribution indexed by a parameter $\theta = (\theta_1, \dots, \theta_p)$, where θ takes values in a parameter space $S \subset R^P$, with some θ_j possibly lying on the boundary of S. Suppose that $Y_1, \dots, Y_n$ are independent sample of Y. Let $\{f(\cdot;\theta): \theta \in S\}$ denote a family of real-valued functions on Y such that $E\{f(Y, X; \theta)\} < \infty$, where X are predictors of Y which can be either known or random variables. Let $\hat{\theta}_n$denote the M-estimator, obtained by maximizing an objective function

$$F_n(\theta) = \sum_{i=1}^{n} f(Y_i, X_i; \theta)$$

subject to the constraint that $\hat{\theta}_n \in S$.

Inspired by Self and Liang (1987), and Wong et. Al (2016), we invoke the following assumptions throughout the paper.

(B.1) The first three derivatives of $f(\cdot;\theta)$ with respect to each θ_j $(j = 1, \dots, p)$ exist on the intersection of a neighborhood N of the true parameter value θ_0 and S. If θ_j is on the boundary, then the derivatives are taken from the appropriate sides.

(B.2) The first derivative of $F_n(\theta_0)$ satisfies

$$n^{-1}U_n(\theta_0) = n^{-1}\left[\frac{\partial}{\partial\theta_0}F_n(\theta_0)\right] \to 0,$$

with probability one, as $n \to \infty$.

(B.3) The Hessians

$$n^{-1}H_{\boldsymbol{n}}(\theta) = n^{-1}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}F_n(\theta)\right],$$

$n^{-1}H_n(\theta) \to -I(\theta)$ with probability one, as $n \to \infty$, and $I(\theta)$ is positive definite.

(B.4) There exists a function $M(Y)$ such that

$$\left|\frac{\partial^3}{\partial\theta_j\theta_k\theta_l}f(Y,X;\theta)\right| < M(Y)$$

for all Y in the support of $f(Y,X;\theta)$and all θ on the intersection of neighborhood $\mathbb{N}$ of θ_0 and S and $E\{M(Y)\} < \infty$.

(B.5) The intersections of S and the closure of the neighborhood $\mathbb{N}$ centered about θ_0 constitute closed subsets of R^P.

(B.6) The model is identifiable.

Note that Assumptions (B.2) and (B.3) are not the same as in Wong et al. (2016). The revisions allow a more general result. For example, Wong et al. (2016) assume that $E\left[\frac{\partial}{\partial\theta_0}F_n(\theta_0)\right] = 0$ which together with independence implies (B.2) by the law of large

numbers. While this unbiasedness condition is still satisfied by the elements of the objective function with error-free predictors, it is not satisfied if we replace the error-free predictors with observed variables which are subject to measurement errors. However, assumption (B.2) remains satisfied provided that observed variables converges to the true variables (i.e., $\hat{X}_i(T) \to X_i(T)$) as $n \to \infty$. Assumption (B.3) is modified as well so as to ensure that the constrained maximizer is unique. The following theorem demonstrates that $\hat{\theta}_n$ is a consistent estimator for θ_0:

THEOREM 2. *Under Assumption (B.1) - (B.6), as $n \to \infty$ there exists a sequence of points $\hat{\theta}_n$ in the parameter set S, at which local maxima of $F_n(\theta)$ occur, and that converges to θ_0 in probability. Moreover, $n^{1/2}(\hat{\theta}_n - \theta_0) = O_P(1)$.*

PROOF. Take any $\delta > 0$, let $N(\theta_0, \delta)$ denote the neighborhood of θ_0. Since the intersection of S and the closure of δ is closed, $F_n(\theta)$ must have a local maximum on this set. The consistency of $\hat{\theta}_n$ may be shown by proving that $F_n(\theta) < F_n(\theta_0)$ with probability tending towards one for all θ in S that are at a distance δ from θ_0. Similar to arguments by Lehmann (1983, pp. 429-432), we expand $n^{-1}F_n(\theta)$ about θ_0 through the Taylor series (Lehmann 1998),

$$n^{-1}F_n(\theta) - n^{-1}F_n(\theta_0) = T_1 + T_2 + T_3,$$

where

$$T_1 = n^{-1}U_n(\theta_0)^T(\theta - \theta_0),$$

$$T_2 = (2n)^{-1}(\theta - \theta_0)^T H_n(\theta_0)(\theta - \theta_0),$$

$$T_3 = \frac{1}{6}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}(\theta_j - \theta_{j0})(\theta_k - \theta_{k0})(\theta_l - \theta_{l0})\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta_j\theta_k\theta_l}f(X_i;\theta^*),$$

and θ^* lies between θ and θ_0. To prove $F_n(\theta) - F_n(\theta_0) < 0$ for all θ in S that are at a distance δ from θ_0 with probability approaching one, we need to show that for any sufficiently small δ, the maximum of T_2 is negative, while T_1 and T_3 are smaller than T_2. By Assumption (B.2), $|n^{-1}U_n(\theta_0)| < \delta^2$ and hence $|T_1| < p\delta^3$ with probability tending to one. For T_2, consider

$$2T_2 = -(\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0) + (\theta - \theta_0)^T\{n^{-1}H_n(\theta_0) + I(\theta_0)\}(\theta - \theta_0).$$

Analogous to the argument for T_1, the absolute value for the second term is less than $p^2\delta^3$ with probability tending to one. The first term is a nonrandom quadratic form of $(\theta - \theta_0)$. This can be reduced to a diagonal form $\sum \lambda_i \xi_i^2$ by an orthogonal transformation, which becomes $\sum \xi_i^2 = \delta^2$ within the closed set $S \cap b(\theta_0, \delta)$, so that $\sum \lambda_i \xi_i^2 \leq \lambda_1 \delta^2$. Combining the first and second terms, there exists $c > 0$ such that $T_2 < -c\delta^2$ with probability approaching one. By assumption (B.4), we have

$$|\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3}{\partial\theta_j\theta_k\theta_l} f(X_i; \theta^*)| < 2m$$

and hence $|T_3| < b\delta^3$ with probability tending to one in the closed set $S \cap b(\theta_0, \delta)$, where $b = p^3 m/3$. Combining three inequalities, we have $\max(T_1 + T_2 + T_3) < -c\delta^2 + (b + p)\delta^3$. This quantity is less than zero if $\delta < c/(b + p)$, which proves the local maxima θ is consistent for estimating θ_0 with a closed set S.

The proof of root-n consistency (i.e., $n^{1/2}(\hat{\theta}_n - \theta_0) = O_P(1)$) follows Lemma 1 of Chernoff (1954). By taking the Taylor series expansion of $n^{-1}F_n(\theta)$ over the close set S, we have

$$n^{-1}F_n(\hat{\theta}_n) = n^{-1}F_n(\theta_0) + n^{-1}U_n'(\theta_0)(\hat{\theta}_n - \theta_0) + 2n^{-1}(\hat{\theta}_n - \theta_0)' H_n(\theta_0)(\theta_\Omega - \theta_0)$$
$$+|\hat{\theta}_n - \theta_0|^3 \cdot O_P(1) \qquad (2.6)$$

Since $\hat{\theta}_n$ is consistent for θ_0, we have that for any $\epsilon > 0$ there exists a sequence of $c_{n\epsilon} \to 0$ and a K_ϵ such that with probability greater than $1 - \epsilon$

$$\left|\hat{\theta}_n - \theta_0\right| < c_{n\epsilon},$$

$$|n^{-1}U_n(\theta_0)| < \frac{K_\epsilon}{\sqrt{n}},$$

and

$$\sum_{j=1}^{p}\sum_{k=1}^{p}(H_{njk}(\theta_0) + I_{njk}(\theta_0))^2 < c_{n\varepsilon}$$

and the third term is less than $\left|\hat{\theta}_S - \theta_0\right|^3 K_\epsilon$. Provided that these inequalities are satisfied, there exists K_ϵ^* such that expression (2.6) is less than

$$-\frac{1}{2}\left(\hat{\theta}_S - \theta_0\right)' I(\theta_0)\left(\hat{\theta}_S - \theta_0\right) + K_\epsilon^*\left(\frac{\left|\hat{\theta}_S - \theta_0\right|}{\sqrt{n}} + c_{n\epsilon}\left|\hat{\theta}_S - \theta_0\right|^2\right) < 0.$$

The theorem follows since otherwise $\frac{|\hat{\theta}_S - \theta_0|}{\sqrt{n}}$ will diverge which contradicts the requirement that expression (2.6) is less than zero with probability tending to one as $n \to \infty$. ■

The next theorem describes the asymptotic distribution for the M-estimator $\hat{\theta}$, some of which may lie on the boundary of its parameter space.

THEOREM 3. *Let Z be a random variable with a multivariate Gaussian distribution with mean θ and covariance matrix $I^{-1}(\theta_0)$, where θ is restricted to lie in $C_S - \theta_0$. Let G denote the distribution of the M-estimator $\hat{\theta}$ of θ based on a single realization of Z when $\theta = 0$, then under conditions (B.1) – (B.6), the limiting distribution is*

$$n^{1/2}\left(\hat{\theta}_n - \theta_0\right) \to_d G.$$

PROOF. The proof of the limiting distribution of $n^{1/2}(\hat{\theta}_n - \theta_0)$ depends on two approximations. First, S is approximated by a cone C_S in which $F_n(\theta)$ is maximized. This can be proved through the root-n consistency of $\hat{\theta}_n$ and the definition of approximating cones shown previously. The second approximation relies on the argument in Lemma 1 by Self and Liang (1987). For θ restricted in $C_S - \theta_0$, the M-estimators $\hat{\theta}_n$ satisfy

$$\hat{\theta}_n - \theta_0 = Z_n + o_P(n^{-1/2}),$$

where $Z_n = n^{-1}I^{-1}(\theta_0)U_n(\theta_0)$. Moreover, the root-n consistency of the M-estimators indicate that $\hat{\theta}_n - \theta_0 = O_P(n^{-1/2})$. For large n, we thus have

$$2\left(n^{-1}F_n(\hat{\theta}_n) - n^{-1}F_n(\theta_0)\right)$$

$$= 2n^{-1}U_n(\theta_0)^T(\hat{\theta}_n - \theta_0) + n^{-1}(\hat{\theta}_n - \theta_0)^T H_n(\theta_0)(\hat{\theta}_n - \theta_0) + O_P(n^{-1/2})$$

$$= 2n^{-1}U_n(\theta_0)^T(\hat{\theta}_n - \theta_0) - (\hat{\theta}_n - \theta_0)^T I(\theta_0)(\hat{\theta}_n - \theta_0)$$

$$+ (\hat{\theta}_n - \theta_0)^T\{n^{-1}H_n(\theta_0) + I(\theta_0)\}(\hat{\theta}_n - \theta_0) + O_P(n^{-3/2})$$

Assumptions B.3 and B.4 together imply that the elements of $n^{-1}H_n(\theta_0) + I(\theta_0)$ converge in distribution to a normal distribution as n increases so that $n^{-1}H_n(\theta_0) + I(\theta_0)$=$O_P(n^{-1/2})$. Moreover, since $\hat{\theta}_n - \theta_0$ is $O_P(n^{-1/2})$, the third term in this expression is $O_P(n^{-1/2})$. Therefore, the expression above equals

$$-\{(\hat{\theta}_n - \theta_0)^T I(\theta_0)(\hat{\theta}_n - \theta_0) - 2n^{-1}U_n(\theta_0)^T(\hat{\theta}_n - \theta_0) + n^{-2}U_n^T(\theta_0)I^{-1}(\theta_0)U_n(\theta_0)\}$$

$$+ n^{-2}U_n^T(\theta_0)I^{-1}(\theta_0)U_n(\theta_0) + O_P(n^{-3/2})$$

$$= -\{n^{-1}U_n^T(\theta_0)I^{-1}(\theta_0) - (\hat{\theta}_n - \theta_0)\}^T I(\theta_0)\{n^{-1}U_n^T(\theta_0)I^{-1}(\theta_0) - (\hat{\theta}_n - \theta_0)\}$$

$$+ n^{-2}U_n^T(\theta_0)I^{-1}(\theta_0)U_n(\theta_0) + O_P(n^{-3/2})$$

$= -\{Z_n - (\hat{\theta}_n - \theta_0)\}^T I(\theta_0)\{Z_n - (\hat{\theta}_n - \theta_0)\} + n^{-2}U_n^T(\theta_0)I^{-1}(\theta_0)U_n(\theta_0) +$

$O_P(n^{-3/2})$.

Letting $u = n^{1/2}(\hat{\theta}_n - \theta_0)$, we then have

$\Psi_n(u) = 2\left(F_n(\theta_0 + n^{1/2}u) - F_n(\theta_0)\right)$

$= -\{n^{1/2}Z_n - u\}^T I(\theta_0)\{n^{1/2}Z_n - u\} + 2n^{-1}U_n^T(\theta_0)I^{-1}(\theta_0)U_n(\theta_0) + O_P(n^{1/2})$

$\rightarrow_d -(W-u)^T I(\theta_0)(W-u) + W^T I(\theta_0)W = \Psi(u)$

where W~N$(0, I^{-1}(\theta_0))$. As justified in Theorem 1, $\Psi_n(u)$ is dominated by the first quadratic term, thus u maximizes $\Psi_n(u)$ in $C - \theta_0$when $\hat{\theta}_n$ maximizes $F_n(\hat{\theta}_n)$ over the cone C. Since the limit law of W is multivariate Gaussian with mean 0 and covariance matrix $I^{-1}(\theta_0)$, the proof is completed. ■

Self and Liang (1987) demonstrated that the M-estimator $\hat{\theta}$ has an asymptotic normal distribution if the true value θ_0 is the interior point of S. Otherwise, non-normal asymptotic behavior of optimized estimators can appear if θ_0 belongs to the boundary of the set of feasible solutions (Roese-Koerner et al. 2012; Self and Liang 1987). Figure 2.2 illustrates the behaviors of probability density function (pdf) with versus without constraints for a single parameter. The dashed red curve represents the pdf under normal distribution, while the solid black line denotes the pdf restricted to the interval [0, 1]. Based on Figure 2.2 (A) (Figure 2.2 (C)), we can see that the constrained pdf is not normally distributed any more in the presence of constraints, and that we tend to obtain overestimated (underestimated) mean value (i.e., μ_T) when the expected mean value $(i.e., \mu_N)$ is closed to the lower (upper) bound of constraints. On the contrary, the

expected mean value and constrained mean value are almost the same when the true value falls within the interior of the parameter space.

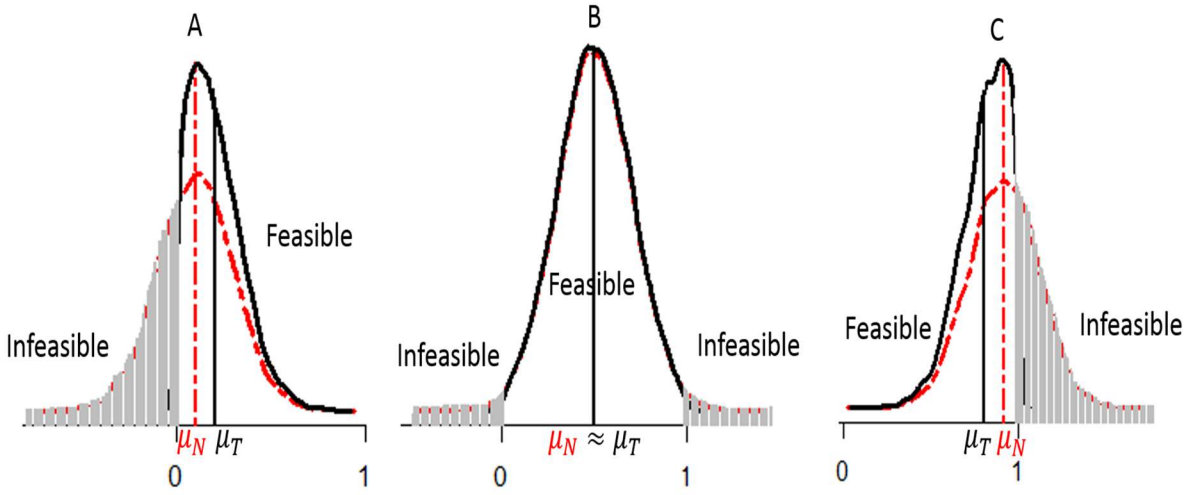

Figure 2.2 Effect of constraints on the probability density function for a parameter with expected value μ_N.

Theorems 2 and 3 justify consistency and limiting distribution of local maxima when some estimators may lie on the boundary of the parameter space. However, it is not uncommon that the M-estimators calculated by an iterative optimization routine can be trapped in a locally optimal solution if the constrained objective function not convex. That is, root-n consistent sequence of local maximizers may have different asymptotic distribution as the global maximizers. Geyer (1994) and Shapiro (2000) argued that Chernoff regularity is not sufficient for asymptotic equivalence of root-n consistent local M-estimators to hold (Geyer 1994; Shapiro 2000). Assuming that the parameter set S is approximated by a cone and is closed, Shapiro suggested that nearly convexity and prox-regularity of S is required for obtaining the asymptotic equivalence of the local and global maximum estimators on the boundary of a parameter space (Shapiro 2000). By

Shapiro (2000), the set S is nearly convex if at a point $x_0 \in S$, there exist a neighborhood $\mathbb{N}$ of x_0 and a function $k(x, x')$ tending to zero as $x \to x_0$, $x' \to x_0$, such that

$$dist\big(x' - x, T_S(x)\big) \leq k(x, x')\|x' - x\| \quad \forall x, x' \in S \cap \mathbb{N}.$$

The set S is prox-regular if at a point $x_0 \in S$, there exist a neighborhood $\mathbb{N}$ of x_0 and a positive constant K such that

$$dist\big(x' - x, T_S(x)\big) \leq K\|x' - x\|^2 \quad \forall x, x' \in S \cap \mathbb{N}.$$

Therefore, the asymptotic properties for the M-estimators $\hat{\theta}$ in the current study implied in Theorems 2 and 3 depend on propositions as follows.

Proposition 1. The parameter space S in the current study is nearly convex and prox-regular.

Recall that a parameter space is combinations of all possible values for different parameters contained in a particular model. Since the candidate values for parameters in the inequality constrained models are restricted in a closed unit cube which is convex, it follows that the parameter set S in the current study is nearly convex and prox-regular in terms of Shapiro (2000).

The proof of convexity of the objective function in the current study requires the following lemma.

Lemma 1. *If a function F is strictly convex, and $S \in R^P$ is closed convex set, then the optimization problem*

$$min\, F(y) \qquad s.t.\ y \in S$$

has a unique solution if it has any solutions in S.

PROOF. Assume both x_1 and x_2 are optimal solution to $F(x)$ which is subject to $x \in S$, then we have $\frac{x_1+x_2}{2} \in S$ due to the convexity of set S. By strictly convexity of F, we have

$$F\left(\frac{x_1+x_2}{2}\right) < \frac{1}{2}F(x_1) + \frac{1}{2}F(x_2) = F(x_1) = F(x_2)$$

This contradicts that x_1 and x_2 are both optimal solutions, and hence follows that the optimization problem has a unique solution if function F is strictly convex, and $S \in R^P$ is closed convex set. Clearly, the OLS estimator in (2.2) is strictly convex, so parameter estimation based on it has a unique optimized solution on the unit cube.

So far, we have demonstrated that the local and global optimized solutions are the same in current paper, and thus Self and Liang's method can be applied to show the asymptotic behaviors for the parameter estimates in the constrained linear regression models. Next, we will demonstrate that Assumptions (B.1) to (B.6) are satisfied for the objective function defined as the negative value of (2.2) for the constrained linear probability model.

Preposition 2. Assuming linear independence of vectors X_i, assumptions (B.1) to (B.6) are satisfied in the constrained linear probability model with the objective function $F_n(\theta) = -\sum_{i=1}^{n}[Y_i - \theta^T X_i]^2$.

PROOF. The identifiability of the model is straightforward, and thus assumption (B.6) is satisfied. Assumption (B.5) is satisfied as well since the parameter space S is the unit cube, a closed convex set. For each θ_j $(j = 1, \dots, p)$ on the intersection of neighborhood $\mathbb{N}$ of the true parameter value θ_0 and S, we obtain the first three derivative of $F_n(\theta)$,

$$\frac{\partial}{\partial\theta}F_n(\theta) = \sum_{i=1}^{n}(Y_i - \theta^T X_i)X_i,$$

$$\frac{\partial}{\partial\theta\theta^T}F_n(\theta) = -\sum_{i=1}^{n} X_i {X_i}^T,$$

and obviously the third derivative of θ is zero. Hence, assumption (B.1) is satisfied, and assumption (B.4) follows immediately. Assumptions (B.2) follows directly from

$$E\left[\frac{\partial}{\partial\theta_0}f(X_i,\theta_0)\right] = E\left[Y_i {X_i}^T - {\theta_0}^T X_i {X_i}^T\right] = 0,$$

and the law of large numbers since the terms in $n^{-1}U_n(\theta_0)$ are bounded between -1 and 1. Furthermore, $n^{-1}H_n(\theta_0)$ is the mean of i.i.d. random bounded variables X^TX, so the convergence of $n^{-1}H_n(\theta)$ follows from the law of large numbers. Convergence of $n^{-1}H_n(\theta)$to a positive definite matrix $I(\theta)$is ensured since the predictors are linearly independent. This completes the proof.

Proposition 2 together with Theorem 3 demonstrates that the constrained OLS estimators $\hat{\theta}$ are approximately distributed as a multivariate normal random vector with variance-covariance matrix:

$$(X^TX)^{-1}X^TVX(X^TX)^{-1}$$

restricted to lie on the unit cube, where V is a diagonal matrix with diagonal elements $V_{ii} = \hat{\theta}^T X_i(1-\hat{\theta}^T X_i)$ $(i = 1, \dots, n)$. Assumption (B.2) is not satisfied for the objective function with a weighting term (i.e., $F_n(\theta) = -\sum_{i=1}^{n} w_i[y_i - \theta^T X_i(T)]^2$), which implies that estimates it yields biased estimators.

In the current study, the location contexts X_i are unobservable. We demonstrate that assumptions (B.1) to (B.6) are also satisfied when X_i is substituted by $\hat{X}_i$. Here, we assume that $T_n \to \infty \; and \; n \to \infty$, so that $X_i \xrightarrow{P} \hat{X}_i$ as demonstrated in Theorem 1.

Preposition 3. The assumptions (B.1) to (B.6) are also satisfied when unobserved X_i are replaced with design-unbiased estimators $\hat{X}_i$ in the constrained linear probability model.

PROOF. Similar to justification of Preposition 3, assumptions (B.1) and (B.4) - (B.6) are satisfied when X_i are replaced with design-unbiased estimators $\hat{X}_i$ in the objective function in the current study. To demonstrate that assumption (B.2) is satisfied, take

$$\frac{\partial}{\partial\theta_0}\sum_{i=1}^{n} f\left(Y_i, \hat{X}_i, \theta_0\right)$$

$$= \frac{\partial}{\partial\theta_0}\sum_{i=1}^{n} f(Y_i, X_i, \theta_0) + \left[\frac{\partial}{\partial\theta_0}\sum_{i=1}^{n} f\left(Y_i, \hat{X}_i, \theta_0\right) - \frac{\partial}{\partial\theta_0}\sum_{i=1}^{n} f(Y_i, X_i, \theta_0)\right],$$

we can see that the first term $\frac{\partial}{\partial\theta_0}\sum_{i=1}^{n} f(X_i, \theta_0)$ goes to zero under Preposition 2, and the second term also goes to zero since the function f is continuous at X_i by the continuity theorem found in Theorem 1.1 in the book written by (Boos and Stefanski 2013). Therefore, assumption (B.2) is satisfied here. Similarly, assumption (B.3) follows from the continuity Theorem.

2.2.4 Confidence interval estimation

It is challenging to obtain the confidences intervals in the presence of constraints in the parameter space. The standard procedure for obtaining confidence intervals is not satisfactory under a constrained parameter space because it does not take into account the information regarding the constraints. Methods for constructing confidence bounds in a constrained parameter space have been widely investigated (Mandelkern 2002; Roe and Woodroofe 2001; Wang 2008; Zhang and Woodroofe 2003). Among various options for setting confidence intervals under boundary constraints, the Bayesian credible interval stands out because it yields the shortest expected length for confidence intervals in most cases (Wang 2008).

Analogous to the frequentist confidence interval, the Bayesian approach delineates a region which contains a large fraction of the posterior mass of a parameter. One approach for obtaining this is the region of the highest posterior density (HPD) (Box and Tiao 1992), which treats the boundary constraints as its prior information and describing it with uniform distribution (Koch 1990). The practical implementation of HPD relies on Markov chain Monte Carlo (MCMC) techniques (Chen and Shao 1999). One appealing feature of the HPD confidence region is that it does not require the confidence region to be equal-tailed, so it performs well even when the parametric function is asymmetric (Liu et al. 2015; Vexler et al. 2016). Additionally, Tian et al. (2011) demonstrate that the HPD credible region is asymptotically unbiased for parameters with normal distribution (Tian et al. 2011). In the current study, we calculated HPD credible intervals for parameter estimates obtained from the constrained linear probability model.

2.3 SIMULATIONS

Simulations were carried out so as to mimic the properties of the COHSONET data. The constrained linear probability model (expression (2.1)) was used to generate independent observations of the Bernoulli random variable Y_i, representing the TB conversion indicator, where the elements of the $p\times1$ vector of parameters are constrained to lie between zero and one. The elements of the $p\times1$ vector of covariates $X_{i.}$ give the proportions of time participants spent in in each of the p location contexts and so are constrained to lie between zero and one and to sum up to one. Therefore, the vectors $X_{i.}$ were independently sampled from a Dirichlet distribution to fulfill such constraints.

The simulations aim to compare the performance of the constrained linear probability model to the logistic regression model. To be specific, we seek to evaluate the behavior of ordinary least square (OLS) and weighted least square (WLS) estimators in the constrained linear probability model, and maximum likelihood estimate (MLE) and bias-reduced penalized maximum likelihood estimate (PMLE) in the logistic regression model. We conducted simulations of 3 different scenarios: (A) An ideal setting where Dirichlet means are not closed to zero; (B) One location context with a Dirichlet mean close to zero; and (C) The location contexts $X_{i.}$ are not directly observed but estimated from data. For each scenario, we explored the impacts of setting a parameter close to the boundary of the parameter space. To explore the impact of sample size, sample sizes were set to 100, 300, and 1000. Each simulation was replicated 1000 times. We compared empirical mean bias, empirical standard deviation (SD) and percent coverage of nominal 95% confidence intervals (CR).

2.3.1 Scenario A: known proportions, no small Dirichlet means

Under this scenario, $p = 3$ with Dirichlet means for location contexts of 0.25, 0.35, and 0.40, respectively. We fixed the parameters for the first two location contexts to $\beta_1 = 0.20$, and $\beta_2 = 0.70$. For the last location context, we took β_3 equal to 0.01, to 0.50, and 0.75, so as to investigate the potential impacts of setting this parameter at different location within the constrained parameter space on proposed statistical approaches.

Simulation results for Scenario A are presented in Table 2.1. When β_3 lies well within the interior of parameter space (i.e., $\beta_3 = 0.5\ or\ 0.75$), PMLE appeared to perform the best with respect to empirical standard deviation, while OLS yielded the smallest empirical bias. Nevertheless, the s empirical standard deviations of the OLS estimates

were not much larger than those of the PMLE. Percentage coverage of 95% confidence intervals for OLS and MLE estimators were good, but PMLE had 100% coverage throughout. Increasing the sample size tended to reduce empirical bias and standard deviation of OLS, MLE, and PMLE estimates. As expected, WLS estimates had smaller standard deviations of parameter estimates than OLS estimates. However, WLS yielded large biases regardless of sample size, a result which is consistent with the observation that the derivative of the weighted sum of squared residuals with respect to the model parameter β has non-zero expectation and hence yields an asymptotically biased estimating equation. Moreover, WLS tended to shrink parameters toward 0.5 because the empirical mean bias of WLS estimates tended to be positive when the true parameter value is less than 0.5, and tended to be negative when the true parameter value is greater than 0.5.

When β_3 lies close to the boundary of the parameter space (i.e., $\beta_3 = 0.01$), OLS method appeared to perform best with respect to coverage of nominal 95% confidence intervals and empirical mean bias. PMLE was successful in reducing bias relative to MLE method in the logistic regression model. WLS continued to perform poorly with respect to the empirical mean bias and percentage coverage of nominal 95% confidence intervals, especially for the larger sample sizes.

Table 2.1 Simulation results for Scenario A with known proportions: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals.

			Constrained Linear Probability Model						Logistic Regression Model					
			OLS			WLS			MLE			PMLE		
n	Parameter	True Value	Bias	SD	CR (%)	Bias	SD	CR (%)	Bias	SD	CR (%)	Bias	SD	CR (%)
100	β_1	0.2	-0.0050	0.1091	91.3	0.1355	0.0861	92.5	0.0184	0.1092	93.9	0.0291	0.1057	100.0
	β_2	0.7	-0.0034	0.1031	92.7	-0.0674	0.0612	99.4	0.0327	0.0956	91.6	0.0226	0.0935	100.0
	β_3	0.01	0.0177	0.0377	99.5	0.1349	0.0592	59.7	0.0523	0.0351	38.8	0.0050	0.0361	30.0
300	β_1	0.2	-0.0037	0.0626	93.8	0.1402	0.0448	49.5	0.0128	0.0615	94.3	0.0168	0.0609	100.0
	β_2	0.7	0.0023	0.0560	96.1	-0.0709	0.0319	94.2	0.0361	0.0516	90.5	0.0326	0.0513	92.3
	β_3	0.01	0.0082	0.0236	98.6	0.1484	0.0443	8.1	0.0498	0.0204	3.9	0.0050	0.0020	1.8
1000	β_1	0.2	-0.0021	0.0338	94.6	0.1417	0.0232	0.2	0.0113	0.0334	93.7	0.0125	0.0333	100.0
	β_2	0.7	0.0005	0.0324	95.6	-0.0744	0.0180	29.9	0.0331	0.0298	79.3	0.0321	0.0297	100.0
	β_3	0.01	0.0030	0.0136	98.7	0.1535	0.0226	0.0	0.0491	0.0110	0.0	0.0040	0.0111	0.0
100	β_1	0.2	0.0006	0.1199	92.4	0.1321	0.0786	95.6	0.0208	0.1026	93.7	0.0304	0.1004	100.0
	β_2	0.7	0.0011	0.1052	92.0	-0.0922	0.0588	97.6	-0.0028	0.0976	94.1	-0.0092	0.0952	100.0
	β_3	0.5	-0.0001	0.1027	92.2	0.0001	0.0535	100.0	0.0002	0.1033	93.2	0.0000	0.1002	100.0
300	β_1	0.2	-0.0017	0.0699	93.3	0.1375	0.0424	55.2	0.0160	0.0582	94.7	0.0194	0.0578	100.0
	β_2	0.7	0.0047	0.0576	95.1	-0.0929	0.0315	88.3	0.0017	0.0537	95.3	-0.0005	0.0532	100.0
	β_3	0.5	0.0003	0.0569	94.7	0.0007	0.0291	100.0	0.0002	0.0577	95.1	0.0000	0.0571	100.0
1000	β_1	0.2	-0.0001	0.0379	94.1	0.1573	0.1029	0.7	0.0157	0.0319	92.1	0.0168	0.0318	100.0
	β_2	0.7	0.0010	0.0325	95.2	-0.0842	0.0687	4.8	-0.0013	0.0303	95.5	-0.0020	0.0302	100.0
	β_3	0.5	-0.0001	0.0321	94.3	0.0146	0.0855	96.9	0.0000	0.0326	94.9	0.0000	0.0325	100.0

Table 2.1 (Continued)

Table 2.1 (Continued)

			Constrained Linear Probability Model						Logistic Regression Model					
			OLS			WLS			MLE			PMLE		
n	Parameter	True Value	Bias	SD	CR (%)	Bias	SD	CR (%)	Bias	SD	CR (%)	Bias	SD	CR (%)
100	β_1	0.2	-0.0020	0.1187	93.7	0.1300	0.0770	96.4	0.0171	0.0992	94.6	0.0268	0.0973	100.0
	β_2	0.7	-0.0003	0.1025	93.0	-0.0927	0.0591	98.1	-0.0034	0.0986	94.2	-0.0095	0.0959	100.0
	β_3	0.75	0.0017	0.0896	94.1	-0.0112	0.0555	94.8	-0.0004	0.0828	95.4	0.0000	0.0811	100.0
300	β_1	0.2	-0.0009	0.0699	93.7	0.1376	0.0418	56.3	0.0156	0.0570	94.2	0.0190	0.0567	100.0
	β_2	0.7	0.0036	0.0561	95.6	-0.0929	0.0317	82.6	0.0021	0.0543	95.7	-0.0001	0.0538	100.0
	β_3	0.75	0.0000	0.0512	94.2	-0.0117	0.0304	39.2	-0.0004	0.0477	94.1	0.0000	0.0473	100.0
1000	β_1	0.2	-0.0010	0.0384	93.8	0.1438	0.0574	0.5	0.0142	0.0316	92.5	0.0152	0.0315	100.0
	β_2	0.7	0.0010	0.0320	94.9	-0.0922	0.0375	6.2	0.0001	0.0309	95.4	-0.0005	0.0308	100.0
	β_3	0.75	-0.0001	0.0280	96.0	-0.0115	0.0348	100.0	-0.0005	0.0261	95.5	0.0000	0.0260	100.0

2.3.2 Scenario B: known proportions & a small Dirichlet mean

We consider the scenario where $p = 4$ with Dirichlet means for each location contexts of 0.25, 0.35, 0.35, and 0.05. We fixed the marginal effects of the first three location contexts to $\beta_1 = 0.20$, $\beta_2 = 0.50$, and $\beta_3 = 0.75$. For the last location context, we took β_4 equal to 0.01, 0.50 and 0.99, to investigate the potential impacts of parameters closed to the boundary.

Simulation results for scenario B are shown as Table 2.2. When β_4 lies well within the interior of parameter space (i.e, $\beta_4 = 0.5$), similar to Section 2.3.1, PMLE appeared to perform the best with respect to empirical standard deviation and percentage coverage of 95% confidence intervals; OLS estimates yielded the smallest empirical bias; and WLS estimates tended to shrink parameters toward 0.5. Increasing the sample size appeared to reduce bias and empirical standard deviations of OLS, MLE, and PMLE estimates. WLS continued to perform poorly irrespective of sample size.

When β_4 lies close to the lower boundary of the parameter space (i.e., β_4=0.01), OLS appeared to perform best with respect to empirical mean bias and coverage of nominal 95% confidence intervals, especially when the sample size is large. PMLE appeared to perform better than OLS in a smaller sample with respect to the coverage of nominal 95% confidence intervals. However, a close examination revealed that OLS estimates yielded empirical standard deviations competitive with PMLE in small samples. Therefore, OLS is superior to PMLE when some parameters lie closed to the boundary. As expected, PMLE was successful in reducing bias relative to MLE in the logistic regression model. Additionally, we can see that WLS approach performed poorly with

respect to empirical mean bias, and coverage of nominal 95% confidence intervals, especially for the larger sample sizes. When β_4 lies close to the upper bound of the parameter space (i.e., β_4=0.99), the performance of all approaches appeared to be slightly improved and achieved similar results as β_4=0.01.

Table 2.2 Simulation results for Scenario B with known proportions & a small Dirichlet mean: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals.

			Constrained Linear Probability Model						Logistic Regression Model					
			OLS			WLS			MLE			PMLE		
n	Parameter	True Value	Bias	SD	CR (%)	Bias	SD	CR (%)	Bias	SD	CR (%)	Bias	SD	CR (%)
100	β_1	0.20	-0.0032	0.1198	90.7	0.1294	0.0818	95.2	0.0188	0.1048	93.3	0.0274	0.1018	100.0
	β_2	0.50	-0.0040	0.1098	92.4	0.0008	0.0587	99.7	0.0033	0.1113	94.4	0.0002	0.1095	97.6
	β_3	0.75	-0.0055	0.1060	92.5	-0.1134	0.0644	95.1	-0.0064	0.0952	94.0	-0.0163	0.0932	100.0
	β_4	0.01	0.0899	0.1718	99.1	0.1559	0.1784	99.3	0.1135	0.1629	80.9	0.1429	0.1564	100.0
300	β_1	0.20	0.0009	0.0711	92.7	0.1393	0.0442	54.5	0.0182	0.0605	93.4	0.0215	0.0600	100.0
	β_2	0.50	-0.0013	0.0600	96.1	0.0017	0.0310	100.0	0.0031	0.0618	96.2	0.0024	0.0610	100.0
	β_3	0.75	-0.0013	0.0581	94.3	-0.1156	0.0336	54.5	-0.0044	0.0521	94.6	-0.0077	0.0517	96.5
	β_4	0.01	0.0453	0.0872	99.1	0.1611	0.1190	93.6	0.0829	0.0798	56.9	0.0956	0.0817	100.0
1000	β_1	0.20	-0.0020	0.0377	94.5	0.1407	0.0345	0.3	0.0132	0.0320	92.6	0.0142	0.0319	93.3
	β_2	0.50	-0.0022	0.0347	94.3	0.0013	0.0238	99.8	0.0009	0.0356	94.3	0.0007	0.0355	100.0
	β_3	0.75	-0.0012	0.0329	94.2	-0.1177	0.0238	0.0	-0.0049	0.0296	93.5	-0.0059	0.0295	100.0
	β_4	0.01	0.0227	0.0457	98.5	0.1798	0.0659	41.8	0.0747	0.0391	8.9	0.0789	0.0396	7.1
100	β_1	0.20	-0.0003	0.1201	91.4	0.1301	0.0810	94.8	0.0188	0.1038	93.4	0.0284	0.1013	100.0
	β_2	0.50	-0.0010	0.1104	93.0	-0.0001	0.0583	99.8	-0.0005	0.1111	93.9	-0.0014	0.1093	97.4
	β_3	0.75	-0.0030	0.1055	91.9	-0.1146	0.0642	94.8	-0.0105	0.0950	94.5	-0.0183	0.0929	100.0
	β_4	0.50	-0.0020	0.2931	91.5	-0.0012	0.2172	96.4	-0.0018	0.2747	93.9	-0.0017	0.2425	100.0
300	β_1	0.20	0.0016	0.0712	93.0	0.1391	0.0436	55.6	0.0180	0.0595	93.4	0.0216	0.0591	100.0
	β_2	0.50	0.0007	0.0614	95.4	0.0004	0.0315	100.0	0.0008	0.0626	95.8	0.0008	0.0618	100.0
	β_3	0.75	0.0000	0.0586	94.2	-0.1172	0.0339	52.1	-0.0069	0.0527	94.6	-0.0096	0.0523	96.8
	β_4	0.50	-0.0003	0.1893	90.0	-0.0001	0.1061	99.4	-0.0002	0.1798	92.8	-0.0002	0.1698	100.0
1000	β_1	0.20	-0.0014	0.0379	94.5	0.1425	0.0459	0.4	0.0140	0.0317	92.6	0.0151	0.0316	93.1
	β_2	0.50	-0.0015	0.0350	93.5	0.0010	0.0327	99.6	-0.0014	0.0356	93.7	-0.0014	0.0355	100.0
	β_3	0.75	-0.0006	0.0329	94.3	-0.1186	0.0294	0.1	-0.0070	0.0297	93.3	-0.0078	0.0297	100.0
	β_4	0.50	0.0000	0.0973	94.3	-0.0003	0.0590	99.1	0.0003	0.0971	94.8	0.0002	0.0952	100.0

Table 2.2 (continued)

Table 2.2 (continued)

			Constrained Linear Probability Model						Logistic Regression Model					
			OLS			WLS			MLE			PMLE		
n	Parameter	True Value	Bias	SD	CR (%)	Bias	SD	CR (%)	Bias	SD	CR (%)	Bias	SD	CR (%)
100	β_1	0.20	0.0018	0.1217	91.4	0.1288	0.0819	95.1	0.0153	0.1043	93.2	0.0273	0.1022	100.0
	β_2	0.50	0.0008	0.1099	92.6	-0.0015	0.0583	99.8	-0.0048	0.1113	93.8	-0.0028	0.1074	97.6
	β_3	0.75	-0.0004	0.1057	91.1	-0.1143	0.0651	94.5	-0.0117	0.0971	93.6	-0.0181	0.0944	100.0
	β_4	0.99	-0.0098	0.1750	99.1	-0.0165	0.1814	99.9	-0.0121	0.1670	79.8	-0.0152	0.1614	100.0
300	β_1	0.20	0.0029	0.0725	92.7	0.1376	0.0441	57.6	0.0158	0.0602	93.5	0.0200	0.0598	100.0
	β_2	0.50	0.0020	0.0607	95.5	-0.0011	0.0315	99.9	-0.0019	0.0627	96.2	-0.0013	0.0618	100.0
	β_3	0.75	0.0010	0.0582	94.2	-0.1178	0.0341	50.9	-0.0075	0.0533	94.1	-0.0098	0.0528	96.3
	β_4	0.99	-0.0047	0.0875	99.0	-0.0162	0.1183	94.1	-0.0083	0.0792	57.3	-0.0097	0.0812	100.0
1000	β_1	0.20	-0.0009	0.0380	94.2	0.1426	0.0550	0.7	0.0118	0.0316	92.6	0.0131	0.0316	93.4
	β_2	0.50	-0.0011	0.0351	93.7	0.0005	0.0398	99.4	-0.0040	0.0360	94.0	-0.0038	0.0358	100.0
	β_3	0.75	0.0002	0.0328	93.5	-0.1182	0.0339	0.2	-0.0067	0.0300	92.7	-0.0074	0.0299	100.0
	β_4	0.99	-0.0023	0.0452	99.2	-0.0184	0.0954	42.4	-0.0072	0.0390	10.4	-0.0076	0.0040	8.1

2.3.3 Scenario C: estimated proportions, no small Dirichlet means

The objective of this scenario is to assess the impact of replacing known proportions of time spent in each location context by a given participant with estimated proportions. The simulation settings for scenario C with respect to Dirichlet means and regression coefficients were identical to those in scenario A. As in scenario A, proportions of time spent in each location context were independently sampled from a Dirichlet distribution. The frequencies at which participants were observed at the location contexts were then independently sampled from a multinomial distribution with parameters set according to the realization of the Dirichlet distribution and sample size generated from a Poisson distribution with mean 200, the targeted size of phone calls by the COHSONET study. Sample proportions computed from the realization of the multinomial distribution were then used to estimate the proportions of time spent in each location context as generated from the Dirichlet distribution. OLS and PMLE estimates were then obtained using estimated location contexts, and results were compared to OLS and PMLE estimates obtained using the known location contexts realized from the Dirichlet distribution.

Table 2.3 presents the simulation results for OLS and PMLE estimates under the known and estimated location contexts. We can see that the empirical mean bias for OLS estimate in the constrained linear probability model was small under all simulation settings. The empirical bias of PMLE estimate was bigger than that of OLS estimates in the presence of a parameter close to the boundary. There appears to be no significant difference in the empirical mean bias of OLS estimate between models using known and estimated proportions. However, bias of PMLE modeling from estimated proportions appears to be slightly bigger than that from known proportions. Increasing sample size reduced empirical mean bias as well as standard deviations in both OLS and PMLE estimates. In contrast, changes in sample sizes do not appear to

substantially impact the coverage rates of OLS estimates. PMLE tends to perform better than OLS when all parameters are in the interior of the constrained boundary, but it does not behave as well as OLS when the true value of a parameter is near the boundary. Based on the ratio of empirical variance of known location contexts versus estimated location contexts, we failed to find any significant differences in the estimates obtained from known or estimated location contexts.

Table 2.3 Comparison of simulation results between Scenarios A and C: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals; Ratio, empirical variance of known location contexts versus estimated location contexts.

			Constrained Linear Probability Model (OLS)							Logistic Regression Model (PMLE)						
			Known Proportions			Estimated Proportions				Known Proportions			Estimated Proportions			
n	Parameter	True Value	Bias	SD	CR (%)	Bias	SD	CR (%)	Ratio	Bias	SD	CR (%)	Bias	SD	CR (%)	Ratio
100	β_1	0.2	-0.0050	0.1091	91.3	-0.0042	0.1091	91.4	0.999	0.0291	0.1057	100.0	0.0294	0.1057	100.0	1.001
	β_2	0.7	-0.0034	0.1031	92.7	-0.0051	0.1031	92.8	1.000	0.0226	0.0935	100.0	0.0207	0.0937	100.0	0.996
	β_3	0.01	0.0177	0.0377	99.5	0.0187	0.0382	99.1	0.972	0.0050	0.0361	30.0	0.0601	0.0364	29.3	0.982
300	β_1	0.2	-0.0037	0.0626	93.8	-0.0030	0.0624	94.4	1.005	0.0168	0.0609	100.0	0.0172	0.0607	100.0	1.005
	β_2	0.7	0.0023	0.0560	96.1	0.0005	0.0559	96.2	1.004	0.0326	0.0513	92.3	0.0305	0.0513	92.8	0.997
	β_3	0.01	0.0082	0.0236	98.6	0.0092	0.0241	98.5	0.955	0.0050	0.0206	1.8	0.0529	0.0207	1.8	0.976
1000	β_1	0.2	-0.0021	0.0338	94.6	-0.0015	0.0338	95.4	1.000	0.0125	0.0333	100.0	0.0129	0.0332	100.0	1.000
	β_2	0.7	0.0005	0.0324	95.6	-0.0014	0.0325	95.3	0.993	0.0321	0.0297	100.0	0.0300	0.0299	100.0	0.987
	β_3	0.01	0.0030	0.0136	98.7	0.0041	0.0140	98.5	0.944	0.0040	0.0111	0.0	0.0505	0.0111	0.0	0.987
100	β_1	0.2	0.0006	0.1199	92.4	0.0022	0.1201	92.5	0.997	0.0304	0.1004	100.0	0.0317	0.1006	100.0	0.994
	β_2	0.7	0.0011	0.1052	92.0	0.0002	0.1049	91.9	1.006	-0.0092	0.0952	100.0	-0.0101	0.0949	100.0	1.005
	β_3	0.5	-0.0001	0.1027	92.2	-0.0002	0.1025	92.4	1.004	0.0000	0.1002	100.0	0.0000	0.1000	100.0	1.004
300	β_1	0.2	-0.0017	0.0699	93.3	-0.0003	0.0698	93.1	1.001	0.0194	0.0578	100.0	0.0205	0.0579	100.0	0.997
	β_2	0.7	0.0047	0.0576	95.1	0.0036	0.0577	94.7	0.998	-0.0005	0.0532	100.0	-0.0016	0.0533	100.0	0.995
	β_3	0.5	0.0003	0.0569	94.7	0.0004	0.0568	94.8	0.100	0.0000	0.0571	100.0	0.0014	0.0570	100.0	1.002
1000	β_1	0.2	-0.0001	0.0379	94.1	0.0014	0.0379	94.0	1.001	0.0168	0.0318	100.0	0.0179	0.0319	100.0	0.996
	β_2	0.7	0.0010	0.0325	95.2	0.0000	0.0325	95.2	0.998	-0.0020	0.0302	100.0	-0.0029	0.0303	100.0	0.996
	β_3	0.5	-0.0001	0.0321	94.3	-0.0001	0.0320	94.7	1.005	0.0000	0.0325	100.0	0.0000	0.0324	100.0	1.005

Table 2.3 (Continued)

Table 2.3 (Continued)

			Constrained Linear Probability Model (OLS)							Logistic Regression Model (PMLE)						
			Known Proportions			Estimated Proportions				Known Proportions			Estimated Proportions			
n	Parameter	True Value	Bias	SD	CR (%)	Bias	SD	CR (%)	Ratio	Bias	SD	CR (%)	Bias	SD	CR (%)	Ratio
100																
	β_1	0.2	-0.0020	0.1187	93.7	0.0002	0.1186	93.7	1.001	0.0268	0.0973	100.0	0.0285	0.0974	100.0	0.997
	β_2	0.7	-0.0003	0.1025	93.0	-0.0009	0.1024	93.7	1.003	-0.0095	0.0959	100.0	-0.0100	0.0958	100.0	1.002
	β_3	0.75	0.0017	0.0896	94.1	0.0009	0.0894	94.1	1.004	0.0000	0.0811	100.0	-0.0011	0.0810	100.0	1.003
300																
	β_1	0.2	-0.0009	0.0699	93.7	0.0010	0.0700	93.3	0.996	0.0190	0.0567	100.0	0.0205	0.0570	100.0	0.990
	β_2	0.7	0.0036	0.0561	95.6	0.0030	0.0563	95.5	0.995	-0.0001	0.0538	100.0	-0.0007	0.0539	100.0	0.994
	β_3	0.75	0.0000	0.0512	94.2	-0.0001	0.0511	94.2	1.002	0.0000	0.0473	100.0	-0.0007	0.0473	100.0	1.000
1000																
	β_1	0.2	-0.0010	0.0384	93.8	0.0009	0.0384	94.0	0.999	0.0152	0.0315	100.0	0.0167	0.0316	100.0	0.996
	β_2	0.7	0.0010	0.0320	94.9	0.0006	0.0320	95.2	0.998	-0.0005	0.0308	100.0	-0.0009	0.0309	100.0	0.995
	β_3	0.75	-0.0001	0.0280	96.0	-0.0002	0.0279	96.0	0.941	0.0000	0.0260	100.0	-0.0006	0.0260	100.0	1.003

2.4 APPLICATION OF THE COHSONET DATA

The proposed approaches are illustrated using data from the COHSONET study which was designed to investigate the impact of social contact patterns on the risk of *M. tuberculosis* conversion. We hypothesize that the proportions of time participants spent in different location contexts may be regarded as a surrogate variable for social contact patterns. To evaluate the social contact patterns, a cohort of individuals aged between 15 and 45 years and were free of *M. tuberculosis* infection at baseline were enrolled in the COHSONET study. Participants were prompted to answer a set of questions concerning the location and surrounding environment at the times when calls were answered during a one-year follow-up period. Sampling times when the phone calls were made were randomly generated from a self-correcting point process.

The conditional intensity for a self-correcting point process takes the form

$$\lambda(\mathrm{t}|\mathcal{F}_t) = \exp\{\alpha_0 + \alpha_1(t - \alpha_2 N(t))\}, \qquad t \in [0, T]$$

where α_0 , α_1, and α_2 are constants (Isham and M. 1979; Ogata and Vere-Jones 1984; Vere-Jones and Ogata 1984), and $\alpha_1,\ \alpha_2 > 0$. This point process is a self-correcting in the sense that if the number of events strays from the target $1/\alpha_2$, then the assessment rate compensates to force this difference back towards zero. The baseline intensity is $\exp\{\alpha_0\}$. The parameters α_0 and α_2 govern the mean number of phone calls made per day, while α_1 controls the variability of the number of calls per day and the regularity of the spacing of the assessment times. Note that the self-correcting point process generates more regularly spaced assessment times and less variation in numbers of assessments per day than the Poisson process, reducing burden on the study subjects. Sampling from the

self-correcting point process guarantees collection of representative samples from which design-unbiased estimates of the distributions of time participants spent in different location contexts may be obtained. In the COHSONET study, α_0= -0.602, $\alpha_1 = 3$, and α_2 = 1.825 targeting 200 random assessments per year.

In the COHSONET study, only 63.7% of phone calls were answered. Given the substantial amount of missing data, there is potential for bias in estimates of model parameters describing the impact of location contexts on risk of *M. tuberculosis* conversion. The only information available for unanswered calls is the time and date at which each call was made. Therefore, it is only feasible to describe the pattern of answered phone calls as a function of calling times. Suppose $p_i(t)$ is the probability that a call at time t is answered by subject i. Let $Z_i(t) = 1$ if a call is answered at time t by subject i, and $Z_i(t) = 0$ if otherwise. Assume that $Z_i(t), t \in N_i$, are independently sampled from a Bernoulli distribution with thinning function $p_i(t)$, where N_i denotes the set of times at which calls are made to subject i, a realization of a point process with intensity $\lambda_i(t)$. Then the set of answered calls N_i^* is a realization of a thinned point process with intensity $\lambda_i(t)p_i(t)$(Cressie 1991). Assume that the data are missing at random, the design-unbiased estimators in (2.4) may be replaced with corrected estimators

$$\hat{X}_i(T) = \frac{1}{T}\sum_{t\in N_i^*}\frac{x_i(t)}{\lambda_i(t)p_i(t)} \ . \qquad (2.7)$$

Exploratory data analysis suggested that the missing data pattern depended on the time of day, a pattern that is likely to vary among study participants. The location contexts in which participants spend their time are also likely to be a function of time of day, a

function that may also vary among study participants. We assume that the thinning function is periodic, as described through its logit transformation,

$$log\frac{p_i(t)}{1-p_i(t)} = \sum_{k=1}^{K} u_{ik}\cos(\frac{2\pi kt}{\tau} + \phi_{ik})$$

where u_{ik} denotes the amplitude, τ represents the period set to 1 (day), and ϕ_{ik} denotes the phase. The model may be reparameterized by writing

$$log\frac{p_i(t)}{1-p_i(t)} = \gamma_{i0} + \sum_{k=1}^{K}\left\{\gamma_{1i}\ cos\left(\frac{2\pi kt}{\tau}\right) + \gamma_{2ik}sin\left(\frac{2\pi kt}{\tau}\right)\right\},$$

where the amplitude is $u_{ik} = \sqrt{\gamma_{1ik}^2 + \gamma_{2i}^2}$ and the phase is $\phi_{ik} = -tan^{-1}(\gamma_{1ik}/\gamma_{2i}\ )$. As to describe variation among participants' missing data patterns, the parameter vectors γ_i are assumed to be independently sampled from a multivariate normal distribution with mean μ and variance-covariance matrix Σ.

Laplace approximations to the likelihood (Breslow and Clayton 1993) and maximum hierarchical likelihood (Lee and Nelder 1996), both lead to inconsistent estimates when the sampling domain is small (Rathbun and Shiffman 2016). The Expectation-Maximization (EM) algorithm can produce consistent estimates in the random effects model regardless of sampling domains. Nevertheless, it remains challenge to compute the E-step in the random effects model because the conditional expectation is an intractable integral. Steele (1996) proposed using a second-order Laplace approximation for computation of conditional expectations within the E-step (Steele 1996) for generalized linear mixed models. We implemented Steele's (1996) method for parameter estimation in the random effects model using FORTRAN code available in the supplementary material of Rathbun and Shiffman (2016).

In current study, only the 288 subjects who responded to more than 30 phone calls over the study are eligible for inclusion in the data analysis. In the random effects modelling of probability of answering phone calls as a function of time, we set $k = 4$, a value which we found well captured the periodic patterns of answering the phone calls in EMA data. Figure 1.1 in Chapter 1 plots the expected probability of answering phone calls as a function time as estimated from Steele's (1996) modified EM algorithm as obtained from the mean μ of the random effects γ. On average, participants were most likely to respond to the phone calls in the early morning (i.e., 7:00 am -8:00 am). There appeared to be an increasing trend between 9:00am and 7:00pm and a sharp decreasing trend between 8:00pm to 11:00pm, with subjects being most likely to answer phone calls at 7:00pm, and least likely to answer phone calls at the end of a day.

As illustrated in Figure 1.2 concerning the estimated proportions of time participants spent in different location contexts, participants spent the most time at homes (i.e. 32.4%), followed by work places (32.1%), public transports (7.1%), and shopping centers (4.1%). It seems that participants in the COHSONET study rarely spent time at women groups, gyms/recreations, clubs, schools, neighbors' homes and hospitals (less than 1%). Comparisons of the proposed approaches were conducted based upon complete cases with both *M. tuberculosis* conversion information and proportions of time spent in each location context. In the rest of this section, a total of 189 subjects with complete information were included in the statistical modeling.

Table 2.4 presents parameter estimates from inequality constrained linear probability models using OLS and WLS methods, as well as from logistic regression models using classical MLE and bias-reduced PMLE approaches based upon all location contexts

being observed. The fitted models suggested that schools, worship centers, bars, and shopping centers, were the most dangerous location contexts with estimated risks for *M. tuberculosis* conversion close to 100% if participants had spent all of their time at these public locations. Risk of *M. tuberculosis* conversion was also high in public transport and friends' homes, with an estimated risk of approximately 50% for persons spending all of their time at these locations. In contrast, people who spent their time in gym/recreation, women groups, market places and neighbors' homes were least likely to become infected with *M. tuberculosis*.

Across all presented approaches, OLS in the constrained linear probability model yielded the largest number of estimates on the boundary. On the contrary, both MLE and PMLE methods in the logistic regression were least likely to produce parameter estimates on the boundary. In the constrained linear probability model, the OLS estimate trended towards zero when the corresponding WLS estimate was less than 0.25, while it trended towards one when the WLS estimated being greater than 0.75. Otherwise, the WLS estimates were shrunk towards the midpoint of the constrained interval by the OLS approach. The simulations suggested that constrained WLS estimates were shrunk toward 0.5. This also appeared to be the case for the analysis of the COHSONET data. For the most part, constrained WLS estimates tended to be closer to 0.5 than constrained OLS estimates. As compared to MLE estimates in the logistic regression model, PMLE appeared to force estimates towards the midpoint of the constrained parameter space. However, there appeared to be no significant differences in the parameter estimates between MLE and PMLE approaches since their 95% confidence intervals were almost the same. In addition, the length of the 95% confidence intervals appeared to be shortest in the OLS

estimates among all approaches, which indicated that the OLS estimates were subject to smallest uncertainty. Furthermore, the constrained OLS estimates tended to shrink MLE and PMLE estimates towards the closest boundary. For example, if the MLE or PMLE estimate was less than 0.5, then the OLS estimate tended to shrink it towards zero. Otherwise, the MLE and PMLE estimates were shrunk toward one.

Table 2.4 Parameter estimates using different approaches over all location contexts (n=189). 95% CI: 95% confidence interval.

	Constrained linear Probability Model				Logistic Regression Model			
	OLS		WLS		MLE		PMLE	
Location Context	**Estimate**	**95% CI**	**Estimate**	**95% CI**	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Home	0.2905	(0.0000-0.6022)	0.3913	(0.0360-0.7466)	0.3214	(0.0684-0.7292)	0.3252	(0.0684-0.7292)
Friend's home	0.5472	(0.0525-1.0000)	0.4413	(0.0000-0.9477)	0.6697	(0.0000-1.0000)	0.6850	(0.0000-1.0000)
Relative's home	0.0139	(0.0000-0.9372)	0.2517	(0.0000-0.9436)	0.0589	(0.0000-0.9997)	0.1081	(0.0000-0.9997)
Work	0.0000	(0.0000-0.3144)	0.1551	(0.0000-0.4715)	0.0450	(0.0065-0.2238)	0.0523	(0.0065-0.2375)
School	1.0000	(0.0609-1.0000)	0.7560	(0.0562-1.0000)	0.9657	(0.0004-1.0000)	0.9506	(0.0004-1.0000)
Worship center	1.0000	(0.0767-1.0000)	0.9629	(0.0742-1.0000)	0.9299	(0.0073-1.0000)	0.8953	(0.0073-1.0000)
Club/Association	0.0000	(0.0000-0.9473)	0.0000	(0.0000-0.9477)	0.1341	(0.0000-1.0000)	0.3216	(0.0000-1.0000)
Bar	1.0000	(0.0726-1.0000)	0.9983	(0.0721-1.0000)	0.9864	(0.0294-1.0000)	0.9746	(0.0294-1.0000)
Saloon	0.0000	(0.0000-0.9412)	0.1470	(0.0000-0.9444)	0.0967	(0.0000-1.0000)	0.1984	(0.0000-1.0000)
Gym/Recreation	0.0000	(0.0000-0.9479)	0.0000	(0.0000-0.9482)	0.0000	(0.0000-0.9669)	0.0000	(0.0000-0.9972)
Hospital/Clinic	0.0000	(0.0000-0.9448)	0.0000	(0.0000-0.9456)	0.0336	(0.0000-1.0000)	0.1288	(0.0000-1.0000)
Shopping/Trading center	1.0000	(0.0717-1.0000)	0.8824	(0.0649-1.0000)	0.9599	(0.0033-1.0000)	0.9507	(0.0033-1.0000)
Public transport	0.5007	(0.0298-0.9716)	0.6370	(0.0627-1.0000)	0.8195	(0.0029-0.9998)	0.7860	(0.0029-0.9998)
Women group	0.0000	(0.000-0.9497)	0.0167	(0.0000-0.9498)	0.0005	0.0000-1.0000)	0.0653	(0.0000-1.0000)
Market place	0.0000	(0.0000-0.9410)	0.0000	(0.0000-0.9424)	0.0538	(0.0000-1.0000)	0.0936	(0.0000-1.0000)
Neighbor's home	0.0000	(0.0000-0.9472)	0.0885	(0.0000-0.9480)	0.0139	(0.0000-1.0000)	0.0426	(0.0000-1.0000)
Elsewhere	0.6434	(0.0781-1.0000)	0.7186	(0.0818-1.0000)	0.8964	(0.0373-0.9995)	0.8711	(0.0374-0.9995)

† Parameter estimates and 95%CI from the logistic regression models are converted using the formula $e^{\alpha}/(1+e^{\alpha})$as to be comparable to those in the linear probability models.

The results in Table 2.4 tended to yield wide confidence intervals, covering most of the parameter space. Hence, the full model including all location contexts might be regarded as over-parameterized because over-parameterized models often yield estimates with high levels of uncertainty. Next, we performed statistical analyses based on reduced location contexts by pooling friend's, relative's, and neighbor's homes into one location context as "Other home", and pooling the remaining locations except for home, other home, work place, and public transport into "Elsewhere". As shown in Table 2.5, the estimated risk of *M. tuberculosis* conversion at home dropped about 3 to 5 percent compared to the modeling results in Table 2.4. People remained at zero risk of contracting tuberculosis if they spent all the time at work places according to OLS estimates in the constrained linear probability model. Nonetheless, the estimated risk of *M. tuberculosis* conversion at work places increased substantially based upon WLS, MLE and PMLE estimates. Similar to results in Table 2.4, WLS estimates tended to be closer to 0.5 compared to OLS estimates in the linear probability model, and PMLE tended to yield estimates closer to the midpoint relative to MLE method in the logistic regression. However, it appeared that MLE and PMLE produced narrower confidence intervals than those of constrained OLS in the constrained linear probability model.

Table 2.5 Parameter estimates based on reduced location contexts (n=189). 95% CI: 95% confidence interval.

	Constrained linear Probability Model				Logistic Regression Model			
	OLS		WLS		MLE		PMLE	
Location Context	**Estimate**	**95% CI**	**Estimate**	**95% CI**	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Home	0.2575	(0.0000-0.5566)	0.3653	(0.0272-0.7033)	0.2722	(0.0588-0.6624)	0.2816	(0.0588-0.6624)
Other home*	0.3022	(0.0000-0.9360)	0.4161	(0.0000-0.9422)	0.2960	(0.0002-0.9988)	0.3187	(0.0002-0.9988)
Work	0.0000	(0.0000-0.3018)	0.1577	(0.0000-0.4588)	0.0458	(0.0074-0.2104)	0.0490	(0.0074-0.2175)
Public transport	0.4129	(0.0000-0.9372)	0.5742	(0.0603-1.0000)	0.6072	(0.0016-0.9991)	0.6122	(0.0016-0.9991)
New elsewhere*	0.6823	(0.2136-1.0000)	0.6559	(0.1766-1.0000)	0.8126	(0.2184-0.9855)	0.8006	(0.2147-0.9856)

† Parameter estimates and 95%CI from the logistic regression models are converted using the formula $e^{\alpha}/(1+e^{\alpha})$as to be comparable to those in the linear probability models.

* Other home is a pooled location of "Friend's home", "Relative's home", and "Neighbor's home";

New elsewhere is a pooled location of "School", "Worship center", "Club/Association", "Bar", "Saloon", "Gym/Recreation", "Hospital/Clinic", "Shopping/Trading center", "Women group", "Market place" and "Elsewhere" in table 2.4.

2.5 DISCUSSION

When analyzing binary dependent variable, the logistic regression model is commonly used. In this paper, we propose using constrained linear probability model as an alternative to the logistic regression model for fitting a binary outcome. We compared the performance of four different approaches in modeling dichotomous dependent variable: the OLS and WLS approaches in constrained linear probability models, and classical MLE and bias-reduced PMLE methods in logistic regression models. We conducted intensive simulations demonstrating that the OLS estimates in the constrained linear probability model were superior to both classical MLE and bias-reduced PMLE estimates in logistic regression models with respect to empirical mean bias, especially when one or some parameters of interest are closed to the boundary of parameter space. Moreover, the simulation study confirmed that the proposed OLS estimates in the constrained linear probability model are consistent and asymptotically unbiased.

The bias-reduced PMLE approach in the logistic regression model is based on penalizing the likelihood using the Jeffreys invariant prior (Firth 1993), and intends to reduce bias in parameter estimates relative to the classical MLE method, especially for small samples. As expected, simulation results revealed that the empirical bias from the PMLE method was always smaller than the MLE method in the logistic regression, and increasing the sample size tended to narrow the difference between them.

Although weighted least squares has been intensively studied in resolving the problem of heteroscedasticity in the linear probability model (Goldberger 1964; Goldfeld and Quandt 1972; Hensher and Johnson 1981; Mullahy 1990), it performs poorly in the constrained linear

probability model. As suggested by the observation that the derivative of the weighted sum of squared residuals with respect to the model parameter β has non-zero expectation and thus hence yields an asymptotically biased estimating equation, we obtained substantial bias in the WLS estimates in the simulation study. Therefore, when constrains are posed on the linear regression model, it becomes problematic in applying WLS approach.

In the COHSONET study, the true proportions of time participants spent in different location contexts were unavailable, we instead substituted them with design-unbiased estimators in the proposed statistical models. The current methods did not take into account sources of variation due to sampling and random effects modeling because it is challenging to obtain estimates of the variation due to these sources. Nonetheless, simulation studies based on estimated location contexts suggested that such errors were ignorable, so we can obtain similar parameter estimates using the design-unbiased estimators as using the true location contexts in the COHSONET data.

To sum up, OLS approach is a robust approach in the linear probability model which subjected to unit interval constraints. Estimates from the constrained OLS are consistent and asymptotically unbiased. In spite of these advantages, one limitation of constrained OLS is that it does not perform model selection. However, it may be essential to identify significant subset of location contexts from a great number of candidates. The constrained adaptive lasso, which enjoys the merits of model selection and asymptotically consistency (Wong et al. 2016; Zou 2006), can serve be a promising direction in the future research.

2.6 REFERENCES

Albert A, and Anderson JA. 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71:1-10.

Beale EML. 1972. Numerical Methods for Nonlinear Optimization. London: Academic Press.

Boos DD, and Stefanski LA. 2013. Asymptotic Normality and Some Basic Asymptotic Results. Essential Statistical Inference Theory and Methods. New York: Springer. p 14.

Box GEP, and Tiao GC. 1992. Bayesian Inference in Statistical Analysis. New York: John Wiley & Sons.

Breslow NE, and Clayton DG. 1993. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association 88(421):9-25.

Cassel CM, Sarndal CE, and Wretman JH. 1977. Foundations of Inference in Survey Sampling. New York: Wiley.

Chen MH, and Shao QM. 1999. Monte Carlo Estimation of Bayesian Credible and HPD Intervals. Journal of Computational and Graphic Statistics 8:69-92.

Chernoff H. 1954. On the Distribution of the Likelihood Ratio. Annals of Mathematical Statistic 25:573-578.

Cressie N. 1991. Statistics for Spatial Data. New York: Wiley.

Deke J. 2014. Using the Linear Probability Model to Estimate Impacts on Binary Outcomes in Randomized Controlled Trials. Evaluation Technical Assistance Brief for OAH & ACYF Teenage Pregnancy Prevention Grantees.

Dooley SW, Villarino ME, Lawrence M, Salinas L, Amil S, Rullan JV, Jarvis WR, Bloch AB, and Cauthen GM. 1992. Nosocomial transmission of tuberculosis in a hospital unit for HIV-infected patients. JAMA 267(19):2632-2634.

Firth D. 1993. Bias reduction of maximum likelihood estimates. Biometrika 80(1):27-38.

Fletcher R, and Powell MJD. 1963. A Rapidly Convergent Descent Method for Minimizatio. Computer Journal 6:163-168.

Gart JJ, and Zwifel JR. 1967. On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika 54(1/2):181-187.

Geweke J. 1986. Exact Inference in the Inequality Constrained Normal Linear Regression Model. Journal of Applied Econometrics, 1:127-141.

Geyer CJ. 1994. On the Asymptotics of Constrained M-Estimation. The Annals of Statistics 22(4):1993-2010.

Goldberger AS. 1964. Econometric Theory. New York: John Wiley.

Goldfeld SM, and Quandt RE. 1972. Nonlinear Methods in Econometrics. Amsterdam: North-Holland.

Heinze G, and Schemper M. 2002. A solution to the problem of separation in logistic regression. Stat Med 21(16):2409-2419.

Hellevik O. 2009. Linear versus logistic regression when the dependent variable is a dichotomy. Quality & Quantity 43(1):59-74.

Hensher DA, and Johnson LW. 1981. Applied Discrete Choice Modeling. London: Croom Helm.

Hirji KF, Mehta CR, and Patel NR. 1987. Computing Distributions for Exact Logistic Regression. Journal of the American Statistical Association 82(400):1110-1117.

Isham V, and M. W. 1979. A self-correcting point process. Stochastic Processes and their Applications 8(3):335-347.

Kallianpur G. 1980. Stochastic Filtering Theory. New York: Springer.

Karr AF. 1986. Point Processes and their Statistical Inference. New York: Marcel Dekker.

Koch KR. 1990. Bayesian inference with geodetic applications. Berlin Heidelberg, New York: Springer.

Kolassa JE. 1997. Infinite parameter estimates in logistic regression, with application to approximate conditional inference. Scandinavian Journal of Statistics 24(523–530).

Koziel S, and Yang XS. 2011. Computational Optimization, Methods and Algorithms. Poland: Springer-Verlag Berlin Heidelberg.

Lee Y, and Nelder JA. 1996. Hierarchical generalized linear models. Journal of the Royal Statistical Society B 58:619-678.

Lehmann EL. 1998. Theory of Point Estimation. 2nd ed. New York: Springer-Verlag.

Lesaffre E, and Albert A. 1989. Partial separation in logistic discrimination. Journal of the Royal Statistical Society 51:109 –116.

Liew CK. 1976. Inequality Constrained Least-Squares Estimation. Journal of the American Statistical Association 71:746-751.

Liu Y, Gelman A, and Zheng T. 2015. Simulation-efficient shortest probability intervals. Statistics and Computing 25:809-819.

Mandelkern M. 2002. Setting confidence intervals for bounded parameters. Statistical Science 17:149-172.

Mehrotra S. 1992. On the Implementation of a Primal-Dual Interior Point Method. SIAM J Optim 2(4):575–601.

Moré JJ, Garbow BS, and Hillstrom KE. 1981. Testing Unconstrained Optimization Software. ACM Transactions on Mathematical Software 7:17--41.

Moré JJ, and Sorensen DC. 1983. Computing a Trust-Region Step. SIAM Journal on Scientific and Statistical Computing 4:553--572.

Mullahy J. 1990. Weighted least squares estimation of the linear probability model revisited. Economics Letters.

Nelder JA, and Mead R. 1965. A Simplex Method for Function Minimization. Computer Journal 7:308-313.

Nocedal J, and WrighT SJ. 2006. Numerical optimization. New York: Springer Science & Business Media.

Ogata Y, and Vere-Jones D. 1984. Inference for earthquake models: a self-correcting model. Stoch Proc Appl 17:337-347.

Park D, and Park S. 2003. Small sample properties of parametric and nonparametric estiamtors in quantal bioassay. Inter-Stat (London) 5:1-13.

Pedroza C, and Troung VTT. 2016. Performance of models for estimating absolute risk difference in multicenter trials with binary outcome. BMC Medical Research Methodology 16(113).

Powell MJD. 1982a. Extensions to Subroutine VF02AD. In: R. F. Drenick and F. Kozin e, editor. Systems Modeling and Optimization, Lecture Notes in Control and Information Sciences Berlin-Heidelberg-New York: Springer-Verlag.

Powell MJD. 1982b. VMCWD: A Fortran Subroutine for Constrained Optimization. DAMTP 1982/NA4. Cambridge, England.

Rathbun SL. 1996. Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. Journal of Statistical Planning and Inference 51:55-74.

Rathbun SL, and Shiffman S. 2016. Mixed Effects Models for Recurrent Events Data with Partially Observed Time-Varying Covariates: Ecological Momentary Assessment of Smoking. Biometrics 72(1):46-55.

Reichler MR, Reves R, Bur S, Thompson V, Mangura BT, Ford J, Valway SE, Onorato IM, and Contact Investigation Study G. 2002. Evaluation of investigations conducted to detect and prevent transmission of tuberculosis. JAMA 287(8):991-995.

Roe BP, and Woodroofe MB. 2001. Setting confidence belts. Physical Review 60:3009-3015.

Roese-Koerner L, Devaraju B, Sneeuw N, and Schuh W-D. 2012. A stochastic framework for inequality constrained estimation. Journal of Geodesy 86(11):1005-1018.

Self SG, and Liang K. 1987. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio TestsUnder Nonstandard Conditions. Journal of the American Statistical Association 82(398):605-610.

Shapiro A. 2000. On the Asymptotics of Constrained Local M-Estimators The Annals of Statistics 28(3):948-960.

Shiffman S, Stone AA, and Hufford MR. 2008. Ecological momentary assessment. Annual Review of Clinical Psychology 4:1-32.

Steele BM. 1996. A modified EM algorithm for estimation in generalized mixed models. Biometrics 52(4):1295-1310.

Stone AA, and Shiffman S. 2002. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 24(3):236-243.

Tian L, Wang R, Cai T, and Wei LJ. 2011. The Highest Confidence Density Region and Its Usage for Joint Inferences about Constrained Parameters. Biometrics 67:604-610.

Vere-Jones D, and Ogata Y. 1984. On the moments of a self-correcting process. J Appl Prob 21:335-342.

Vexler A, Zou L, and Hutson AD. 2016. Data-Driven Confidence Interval Estimation Incorporating Prior Information with an Adjustment for Skewed Data. THE AMERICAN STATISTICIAN 70(3):243-249.

Wagler A. 2011. Bias Reduced Logistic Dose-Response Models. Journal of Biopharmaceutical Statistics 21:405–422.

Wang H. 2008. Confidence intervals for the mean of a normal distribution with restricted parameter space. Journal of Statistical Computation and Simulation 78(9):829-841.

WHO. 2014. Global Tuberculosis Report 2014. Geneva.

Wong KY, Goldberg Y, and Fine JP. 2016. Oracle Estimation of Parametric Models under Boundary Constraints. Biometrics 72:1173–1183.

Yaganehdoost A, Graviss EA, Ross MW, Adams GJ, Ramaswamy S, Wanger A, Frothingham R, Soini H, and Musser JM. 1999. Complex transmission dynamics of clonally related virulent Mycobacterium tuberculosis associated with barhopping by predominantly human immunodeficiency virus-positive gay men. J Infect Dis 180(4):1245-1251.

Zhang T, and Woodroofe M. 2003. Credible and confidence sets for restricted parameter spaces. Journal of Statistical Planning and Inference 115(2):479-490.

Zorn C. 2005. A Solution to Separation in Binary Response Models. Political Analysis 13:157–170.

Zou H. 2006. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101(476):1418-1429.

CHAPTER 3

CONSTRAINED ADAPTIVE LASSO APPROACH FOR ESTIMATING THE PROBABILITY OF TUBERCULOSIS CONCERSION FROM ECOLOGICAL MODEMNTAR ASSESSMENT OF SOCIAL PATTERNS

3.1 INTRODUCTION

Tuberculosis is an infectious disease that threatens the health of people all over the world, and it is most prevalent in resource-limited countries such as the developing countries in Africa (WHO 2014). Transmission of *Mycobacterium tuberculosis* requires prolonged contact with infectious people (Houk et al. 1968; Kenyon et al. 1996). Therefore, the social contact patterns regarding how and where susceptible members interact with infected patients may determine whether tuberculosis is transmitted or not (Horby et al. 2011; Rehkopf et al. 2015).

The Community Health Study of Social Networks and Tuberculosis (COHSONET), is an ongoing study in Uganda aiming to evaluate the effects of social networks on the transmission dynamics of *M. tuberculosis*. In the COHSONET study, social contacts data were collected through Ecological Momentary Assessment (EMA), a method in the behavioral sciences that enables evaluation of subjects' emotional states and environments through repeated sampling in their every-day environments using electronic devices (Shiffman et al. 2008; Stone and Shiffman 2002). The proportions of time patients spent in different location contexts can serve as a surrogate variable for social contacts, which can be design-unbiased estimators (Cassel et al.

1977) as long as the EMA samples are generated from a known probability-based sampling design.

In the COHSONET study, the outcome variable indicating if participants have become infected with *M. tuberculosis* is dichotomous, suggesting the applicability of logistic regression to predict this outcome as a function of location contexts. Since logistic regression is nonlinear, however, maximum likelihood estimators (MLEs) have a bias of order $O(n^{-1})$ (Firth 1993), a bias that has been observed in a number of practical applications (Firth 1993; Gart and Zwifel 1967; Hirji et al. 1987; Park and Park 2003; Wagler 2011). Linear probability model is an appealing option for dichotomous dependent variable due to its ease of implementation, interpretability and good predictive performance. There is evidence suggesting that the ordinary least squares (OLS) estimator in the linear probability model performs as well as or even outperform maximum likelihood estimates in the logistic regression model in certain situations such as modeling the treatment effect in the randomized controlled trial (Deke 2014; Hellevik 2009; Pedroza and Troung 2016). In most cases, however, direct application of OLS estimates in the linear probability model is ill-posed due to the possibility of yielding predicted probabilities fall outside the unit interval zero and one. Therefore, a mathematical approach which can restrict the predicted values into the unit interval is required by the linear probability model.

In the classical regression problem, there are generally two approaches for estimation and model selection when there are a great number of parameters: variable selection methods such as stepwise regression that do not perform regularization, and regularization methods such as l_2 penalty which do not lead to variable selection. The LASSO (Least Absolute Shrinkage and Selection Operator), which places an l_1 penalty on the coefficients, is a method combining the

advantages of both approaches and thus can perform model selection and regularization at the same time (Tibshirani 1996). One appealing feature of LASSO is that it can enhance the predictive accuracy and interpretability of the statistical model by simultaneously applying model selection and regularization. LASSO deliberately introduces bias into the estimation by adding an l_1 penalty to the objective function so as to reduce the variability of the estimate. In general, LASSO estimators yield smaller mean square error than the OLS estimates, especially in the presence of multicollinearity. Although the LASSO was originally defined for penalized least squares, it can be easily extended to a wide variety of statistical models including generalized linear models, generalized estimating equations, M-estimators, and constrained linear regression models (James et al. 2013; Tibshirani 1997).

Although the LASSO performs variable selection, it does not satisfy the oracle property (Fan and Li 2001; Zou 2006); that is, the LASSO estimator does not perform as well as an oracle estimator under which parameters taking the value zero are known in advance. The adaptive LASSO which contains a weighting term in the l_1 penalty, on the other hand, produces oracle estimates (Zhang and Lu 2007). Specifically, the adaptive lasso enjoys the following properties: (1) selection consistency, if a true parameter value belongs to the prespecified set, the corresponding estimators converges to the true value with probability tending to one, and (2) shares the same asymptotic distribution as the oracle estimator.

The current study aims to use a constrained linear probability model to fit the probability of *M. tuberculosis* conversion as a function of proportions of time participants spent in different location contexts. Huang et al. (2013) proposed using the constrained LASSO to restrict the parameter estimate to be in a unit interval to ensure that the outcome variable, probabilities

associated with the presence of a certain peptide, falls within a reasonable prediction region, as well as to identify an optimal subset of candidate proteins that significantly contributing to the presence of a specific peptide (Huang et al. 2013). To overcome the potential issue of inconsistent model selection in the constrained LASSO, we employed the constrained adaptive LASSO proposed by Wong et al. (2016), to estimate the probability of *M. tuberculosis* conversion as a function of proportions of time participants spent in each location context using a linear probability model.

The rest of paper is structured as follows. In section 3.2, we demonstrate oracle properties for constrained adaptive LASSO procedure. Section 3.3 is devoted to comparing simulations results between different statistical approaches. We present application results of proposed methods to the COHSONET study in section 3.4. We conclude with a discussion of this work in section 3.5.

3.2 CONSTRAINED ADAPTIVE LASSO

The aim of the COHSONET study is to estimate the risk of *M. tuberculosis* conversion as a function of the proportions of time participants spent in different location contexts. Suppose that n subjects are randomly sampled and each subject i is observed over a set of times belonging to the Borel set $T \subset \mathbb{R}$ with Lebesgue measure $|T| < \infty$. Let $X_i(T)$ denote a $p \times 1$ vector corresponding to the proportions of time participants spent in each location context during a study interval [0, T] so that the elements of the vector $X_i(T)$ are constrained to lie between zero and one and sum to one. Let Y_i denote a Bernoulli dependent variable corresponding to *M. tuberculosis* conversion status which is observed over the time interval $[0, T]$. The probability of *M. tuberculosis* conversion is described by the linear probability model

$$E\left(Y_i\middle|X_i(T)\right) = \Pr\left(Y_i = 1\middle|X_i(T)\right) = \beta^T X_i(T); \qquad i = 1, \dots, n. \tag{3.1}$$

where $0 \leq \beta_j \leq 1 \; for \; j = 1, \dots, p$. Note that the linear probability model considered here does not contain an intercept term. The interpretation of coefficients in the linear probability model is straightforward. The coefficient β_j represents the risk of *M. tuberculosis* conversion if the participants spent 100% of time in location j; $j = 1, \dots, p$.

The location contexts $X_i(T)$ are comprised of the population proportions of time participants spent in each location over the period of the EMA study, where the population is a set of randomly selected times in a one-year study interval for each participant. In practice, however, it is not practical to observe the location contexts in which participants spend their time at all points in time during study interval [0, T]. Therefore, the true location contexts $X_i(T)$ are unknown. In practice, if the time-varying covariates are sampled according to a point process $N_i(t)$ with known intensity $\lambda_i(t)$, and $\lambda_i(t) > 0$ for all $t \geq 0$ except on a set of measure zero, then we can obtain design-unbiased estimators of the location contexts $X_i(T)$ from

$$\hat{X}_i(T) = \frac{1}{T}\sum_{t \in N_i} \frac{x_i(t)}{\lambda_i(t)}, \tag{3.2}$$

where N_i is the set of times at which assessments were made for subject i. This estimator is design unbiased in the sense that the expected value $\hat{X}_i(T)$ equal to $X_i(T)$ under the probability model induced by the sampling design. Given boundedness of $X_i(T)$ and $\lambda(t)$, design-unbiased estimators $\hat{X}_i(T)$ converge to the true location contexts $X_i(T)$ with the probability of one when $T \to \infty$ as $n \to \infty$ (see Theorem 1 in Section 2.2.2). For ease of demonstration, we demonstrate the limiting distribution of constrained adaptive LASSO estimators assuming the true location contexts $X_i(T)$ are known. Given the continuity of the objective function, LASSO estimators

based on estimated location contexts $\hat{X}_i(T)$ will share the same limiting distribution provided that $\hat{X}_i(T) \to X_i(T)$ as $T \to \infty$ with $n \to \infty$.

3.2.1 Oracle properties of adaptive LASSO under boundary constraints

The purpose of this study is to fit the linear probability model under the boundary constraints using the adaptive LASSO. Let $Y_1, \dots, Y_n$ denote i.i.d. random vectors of random variable sampled from a parametric family of probability distributions with parameter θ, where θ takes values in a parameter space $S \subset R^P$, with θ_j possibly lying on the boundary of S. In our application, S is the $p-$dimensional unit cube. Let $\{f(\cdot;\theta): \theta \in S\}$ denote a family of real-valued functions on Y such that $E\{f(Y;\theta)\} < \infty$. Let $\tilde{\theta}_n$ denote the M-estimator, obtained by maximizing an objective function

$$F_n(\theta) = \sum_{i=1}^{n} f(Y_i;\theta).$$

Similar to Lehmann (1998), Self and Liang (1987), and Wong et al. (2016), we invoke the following assumptions throughout the paper.

(1) The first three derivatives of $F_n(\theta)$ with respect to each θ_j $(j = 1, \dots, p)$ exist on the intersection of a neighborhood N of the true parameter value θ_0 and S. If θ_j is on the boundary, then the derivatives are taken from the appropriate sides.

(2) The first derivative of $F_n(\theta)$ satisfies

$$n^{-1}U_n(\theta_0) = n^{-1}\left[\frac{\partial}{\partial\theta_0}F_n(\theta_0)\right] \to 0,$$

with probability one, as $n \to \infty$.

(3) The Hessians

$$n^{-1}H_{\boldsymbol{n}}(\theta) = n^{-1}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}F_n(\theta)\right],$$

$n^{-1}H_n(\theta) \to -I(\theta)$ with probability one, as $n \to \infty$, and $I(\theta)$ is positive definite.

(4) On the intersection of neighborhood $\mathbb{N}$ of θ_0 and S, n^{-1} times the absolute value of the third derivative of $l_n(\theta)$ is bounded by $M(Y)$ with $E\{M(Y)\} < \infty$ for some function $M(.)$.

(5) The intersections of S and the closure of the neighborhood $\mathbb{N}$ centered about θ_0 constitute closed subsets of R^P.

(6) The model is identifiable.

Note that Assumptions (2) and (3) are not the same as those in Wong et al. (2016). The revisions allow a more general result. For example, they assume that $E\left[\frac{\partial}{\partial\theta_0}F_n(\theta_0)\right] = 0$ which together with independence implies (2) by the law of large numbers (Wong et al. 2016). While this unbiasedness condition is still satisfied by the elements of the objective function with error-free predictors, it is not satisfied if we substitute the observed covariates which are subject to measurement errors for error-free true predictors. However, assumption (2) remains satisfied provided that observed variables converges to the true variables $\hat{X}_i(T) \to X_i(T)$ as $n \to \infty$. Assumption (3) is modified as well so as to ensure that the constrained maximizer is unique.

Let $\{\theta_j^1, \dots, \theta_j^k\}$ denote the set of values possibly on the boundary of parameter space S for each θ_j $(j = 1, \dots, p)$. Inspired by Wong et al. (2016), we construct the objective function for constrained adaptive LASSO as:

$$\Phi_n(\theta) = F_n(\theta) - \gamma_n \sum_{j=1}^{p}\sum_{k=1}^{k_j}\widetilde{w}_j^k\left|\theta_j - \theta_j^k\right|, \qquad (3.3)$$

where $\widetilde{w}_j^k = \left|\tilde{\theta}_j - \theta_j^k\right|^{-1}$ $(k = 1, \dots, k_j; j = 1, \dots, p)$ with $\tilde{\theta}_j$ being the M-estimator of θ_j, and γ_n is a tuning parameter. Recall that the standard LASSO does not have any constraints but is

constructed to shrink estimators towards zero. Under the constrained LASSO, estimators are shrunk towards the closest boundary. What makes it adaptive is that the weights are based on estimates obtained without any shrinkage. Obviously, the standard adaptive LASSO is a special case of constrained adaptive LASSO procedure in which the candidate set of boundary constraints has only one value with $k_j = 1$ and $\theta_j^1 = 0$.

Let θ_0 be the true parameter value, $\boldsymbol{\mathcal{A}}$ denote the set of parameters away from the boundary, and $\boldsymbol{\mathcal{B}}$ be the set parameters that lie on the boundary. Write $\theta_0 = (\theta_{01}^T, \theta_{02}^T)^T$, where θ_{01} and θ_{02} are $p_1 -$ and $p_2 -$dimensional vectors corresponding to indices in $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$, respectively, similarly followed by $\hat{\theta}_n = (\hat{\theta}_{n1}^T, \hat{\theta}_{n2}^T)^T$. In a proof similar to that of Theorem 1 of Wong et al. (2016), it can be proved that the constrained adaptive estimates $\hat{\theta}_n$ are root-n consistent under assumptions (1) - (6) with $n^{-1/2}\gamma_n \to 0$ and $\gamma_n \to \infty$ as $n \to \infty$. That is, the probability that the constrained adaptive estimates equal the true value tends to one irrespective of whether or not any of the true parameters lie on the boundary. Also, Theorem 1 in Wong et al. (2016) demonstrates that constrained adaptive LASSO estimates of $\hat{\theta}_1$, a subset of parameter estimates not on the boundary, have an asymptotically normal distribution with mean θ_{01} and variance-covariance matrix $I_{11}^{-1}(\theta_{01})$within the constrained parameter space S, where $I_{11}(\theta_{01})$ is the upper left $p_1 \times p_1$ submatrix of the information matrix (Wong et al. 2016). In our example, the variance-covariance matrix $I_{11}^{-1}(\theta_{01})$ equals

$$(X_{p1}^T X_{p1})^{-1} X_{p1}^T V_{p1} X_{p1} (X_{p1}^T X_{p1})^{-1},$$

where $X_{p1}$is a vector of predictors away from the boundary, and V_{p1} is a diagonal matrix with diagonal elements $V_{ii} = \hat{\theta}_1{}^T X_{ip}\ (1 - \hat{\theta}_1{}^T X_{ip}\ )$ $(i = 1, \dots, n)$.

3.2.2 Coordinate decent algorithm and global maximization

To obtain the constrained adaptive LASSO solution we use the coordinate decent algorithm (Wright 2015). We start with an initial estimator (e.g., the constrained OLS estimate) for the parameters in the objective function. At each subsequent iteration, we fix all but one parameter and optimize the objective function with respect to that specific parameter. If the optimized estimate for the parameter is achieved outside the boundary of the parameter space, the estimate is set to be on the closet boundary. The entire optimization process is repeated until the resulting estimators converge at which the algorithm terminates.

Note that estimators obtained by an iterative routine such as coordinate decent can be trapped in a locally optimal solution if the corresponding objective function is not convex (Shapiro 2000). To avoid achieving inconsistent estimates, it is required that the corresponding problem meets the conditions of "nearly convexity" and "prox-regularity". In the current study, the parameter space is restricted in a closed unit cube which is convex (and hence, nearly convex) and prox-regular as defined by Shapiro (2000). Therefore, the maximizers obtained in the current study are global maximizers.

The tuning parameter is a key element in the penalized function as in expression (3.3), which determines whether the model selection is consistent or not. In the current study, the tuning parameter γ_n was selected via minimizing the Bayesian information criterion (BIC) (Schwarz 1978; Wang et al. 2007)

$$BIC(\gamma) = \log\left(\hat{\sigma}_\gamma^2\right) + \frac{\log(n)\, df(\gamma)}{n}$$

where $\hat{\sigma}_\gamma^2 = \sum_{i=1}^{n}(y_i - X_i^T\hat{\beta}_\gamma)^2/n$. The BIC is a consistent model selector in the sense that it identifies the true model from a set of candidate models with a probability approaches to 1 in large samples (Zhang et al. 2010).

3.2.3 Confidence interval estimation

It is challenging to obtain the confidences intervals in the presence of constraints in the parameter space. The standard procedure obtaining symmetric confidence intervals is not satisfactory for the constrained parameter space because it does not take into account the information regarding the constraints. Methods for constructing confidence bounds in a constrained parameter space have been widely investigated (Mandelkern 2002; Roe and Woodroofe 2001; Wang 2008; Zhang and Woodroofe 2003). Among various options for setting confidence intervals under boundary constraints, the Bayesian credible interval stands out because it yields the shortest expected length for a given confidence interval $1 - \alpha$ in most cases (Wang 2008).

Analogous to the frequentist confidence interval, the Bayesian approach delineates a region which contains a large fraction of the posterior mass of a parameter. One approach for obtaining this is the region of the highest posterior density (HPD) (Box and Tiao 1992), which treats the boundary constraints as its prior information and describing it with uniform distribution (Koch 1990). The practical implementation of HPD relies on Markov chain Monte Carlo (MCMC) techniques (Chen and Shao 1999). One appealing feature of the HPD confidence region is that it does not require the confidence region to be equal-tailed, so it performs well even when the parametric function is asymmetric (Liu et al. 2015; Vexler et al. 2016). Additionally, Tian et al. (2011) demonstrate that the HPD credible region is asymptotically unbiased for parameters with

normal distribution (Tian et al. 2011). In the current study, we calculated HPD credible intervals for parameter estimates obtained from the constrained linear probability model.

3.3 SIMULATIONS

Simulations were carried out so as to mimic the properties of the COHSONET data. The linear probability model (i.e., expression (2.1)) was used to generate independent observations of the Bernoulli random variable Y_i , representing the *M. tuberculosis* conversion indicator, where the elements of the $p\times1$ vector of parameters are constrained to lie between zero and one. The elements of the $p\times1$ vector of covariates $X_{i.}$ provide the proportions of time participants spent in in each of the p location context and thus are restricted to lie between zero and one and to sum up to one. Therefore, the vectors $X_{i.}$ were independently sampled from a Dirichlet distribution to fulfill such restrictions.

The objective of the simulations is to compare the performance of the adaptive LASSO to ordinary least squares (OLS) estimates in the linear probability model under the boundary constraints. We performed simulations of three different scenarios: (A) An ideal setting where Dirichlet means are not close to zero; (B) One location context with a Dirichlet mean close to zero; and (C) The location contexts $X_{i.}$ are not directly observed but estimated from data. Under each scenario, we evaluated the impacts of setting a parameter on the boundary of the parameter space. To explore the impact of sample size, sample sizes were set to 100, 300, and 1000. Each simulation was replicated 1000 times. We compared empirical mean bias, empirical standard deviation (SD), percentage coverage of nominal 95% confidence intervals (CR), and percentage of estimates falling on each of the boundaries zero and one.

3.3.1 Scenario A: known proportions, no small Dirichlet means

Under this scenario, the number of location contexts p was fixed at 4 with Dirichlet means of 0.15, 0.20, 0.30, and 0.35, respectively. The parameters for the first three location contexts were held constant at $\beta_1 = 0.50$, $\beta_2 = 0.20$, and $\beta_3 = 0.80$. For the last location context, we took β_4 equal to 0.0, to 0.5, and 1.0, so as to investigate the potential impacts of setting a parameter on the boundary of the parameter space on proposed statistical approaches.

Simulation results for Scenario A are presented in Table 3.1. When β_4 lies on the lower bound of the parameter space (i.e., $\beta_4 = 0.0$), the performance of the constrained OLS approach appeared to be not inferior to or even better than the constrained adaptive LASSO with respect to estimating parameters known to fall in the interior of the parameter space. Regarding the parameter β_4, whose true value was zero, the constrained adaptive LASSO appeared to be superior to the constrained OLS with respect the percentage of estimates of zero. The constrained OLS yielded zero estimates of β_4 at zero regardless of sample size. In contrast, approximately a half of replicate adaptive LASSO took values of zero for β_4. Since the percentage taking the value zero appeared to not increase with increasing sample size, this suggested that the tuning parameter selected based on BIC was small. However, further exploration found that similar results were obtained using the tuning parameter selected from the fivefold cross-validation (results not presented here). As expected, empirical biases of constrained adaptive LASSO and constrained OLS estimates both trended toward zero as the sample size increases. At the sample size of 1000, the empirical mean biases and standard deviations of the constrained adaptive LASSO were almost the same or even smaller relative to those of the constrained OLS. As

expected, the empirical bias for the constrained adaptive LASSO and constrained OLS trended towards zero with increasing sample size.

When β_4 lies on the upper bound of the constraints (i.e., $\beta_4 = 1.0$), the constrained OLS approach tended to perform better than the constrained adaptive LASSO for samples of up to 300 subjects. However, the constrained adaptive LASSO appeared to be superior to the constrained OLS estimates with respect to the empirical mean bias at a sample size of 1000. Similar to $\beta_4$at the lower bound, the bias of constrained adaptive LASSO and OLS estimates both trended towards zero as the sample size increases. Unlike situation of β_4 equal to 0.0 in which constrained OLS yielded 0% of estimates of zero, a small proportion of β_4 estimates from the constrained OLS were set exactly equal to one.

When β_4 lies within the interior of the parameter space (i.e., $\beta_4 = 0.5$), the empirical mean bias was lower under the constrained adaptive LASSO than under the constrained OLS. The constrained OLS appeared to perform better than the constrained adaptive LASSO with respect to percentage coverage of nominal 95% confidence intervals for parameters with relatively small Dirichlet means, while the constrained adaptive LASSO seemed to yield higher percentage coverage of nominal 95% confidence interval for the parameter with a relatively large Dirichlet mean. Furthermore, neither of these two approaches shrunk the estimates towards the boundary when the parameter lies well within the interior of the parameter space.

Table 3.1 Simulation results for Scenario A with known proportions: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals; % To 0, percentage of estimates falling on zero; % To 1, percentage of estimates falling on one.

			Constrained Adaptive LASSO					Constrained OLS				
n	**Parameter**	**True Value**	**Bias**	**SD**	**CR (%)**	**% To 0**	**% To 1**	**Bias**	**SD**	**CR (%)**	**% To 0**	**% To 1**
100	β_1	0.5	-0.0106	0.1734	91.9			-0.0132	0.1697	92.5		
	β_2	0.2	0.0132	0.1263	91.1			0.0006	0.1234	91.3		
	β_3	0.8	-0.0052	0.1063	92.0			-0.0049	0.1043	92.9		
	β_4	0.0	0.0268	0.0412	99.2	48.6	0.0	0.0246	0.0391	99.7	0.0	0.0
300	β_1	0.5	0.0000	0.0981	94.1			-0.0020	0.1003	93.5		
	β_2	0.2	0.0018	0.0736	93.8			-0.0029	0.0718	94.8		
	β_3	0.8	-0.0024	0.0607	95.4			-0.0033	0.0625	94.1		
	β_4	0.0	0.0140	0.0213	99.3	51.8	0.0	0.0131	0.0207	99.3	0.0	0.0
1000	β_1	0.5	-0.0020	0.0541	95.3			-0.0008	0.0541	95.3		
	β_2	0.2	-0.0023	0.0409	93.6			-0.0025	0.0390	95.3		
	β_3	0.8	-0.0003	0.0333	94.7			-0.0013	0.0336	94.4		
	β_4	0.0	0.0081	0.0118	99.0	50.0	0.0	0.0077	0.0117	98.6	0.0	0.0
100	β_1	0.5	0.0012	0.1746	93.0			-0.0087	0.1731	93.3		
	β_2	0.2	0.0033	0.1323	90.5			0.0059	0.1281	95.5		
	β_3	0.8	0.0045	0.1041	92.7			0.0005	0.1040	93.6		
	β_4	0.5	-0.0001	0.1087	94.0	0.0	0.0	-0.0005	0.1127	93.2	0.0	0.0
300	β_1	0.5	-0.0010	0.0991	94.6			0.0015	0.1029	93.9		
	β_2	0.2	-0.0005	0.0792	93.9			-0.0001	0.0778	94.5		
	β_3	0.8	0.0017	0.0608	94.3			0.0008	0.0633	92.3		
	β_4	0.5	-0.0003	0.0617	95.4	0.0	0.0	0.0004	0.0627	94.9	0.0	0.0
1000	β_1	0.5	-0.0003	0.0551	94.7			-0.0001	0.0568	95.0		
	β_2	0.2	-0.0017	0.0443	93.6			-0.0010	0.0424	95.2		
	β_3	0.8	0.0018	0.0335	94.7			-0.0003	0.0333	95.8		
	β_4	0.5	0.0000	0.0340	95.1	0.0	0.0	0.0038	0.0350	93.6	0.0	0.0

Table 3.1 (Continued)

Table 3.1 (Continued)

n	Parameter	True Value	Constrained Adaptive LASSO					Constrained OLS				
			Bias	SD	CR (%)	% To 0	% To 1	Bias	SD	CR (%)	% To 0	% To 1
100	β_1	0.5	0.0046	0.1719	92.2			0.0005	0.1697	92.2		
	β_2	0.2	0.0087	0.1328	89.3			0.0121	0.1295	95.9		
	β_3	0.8	0.0031	0.0997	93.1			0.0048	0.0968	92.2		
	β_4	1.0	-0.0026	0.0393	99.5	0.0	50.7	-0.0023	0.0373	99.7	0.0	7.3
300	β_1	0.5	0.0007	0.0978	94.2			0.0040	0.0987	93.7		
	β_2	0.2	0.0026	0.0794	92.2			0.0036	0.0787	94.4		
	β_3	0.8	0.0015	0.0556	95.4			0.0036	0.0573	92.8		
	β_4	1.0	-0.0014	0.0210	99.6	0.0	49.6	-0.0014	0.0203	99.3	0.0	5.8
1000	β_1	0.5	0.0004	0.0546	94.4			0.0012	0.0550	95.0		
	β_2	0.2	0.0008	0.0445	91.0			0.0012	0.0431	95.4		
	β_3	0.8	0.0017	0.0302	96.3			0.0020	0.0300	95.6		
	β_4	1.0	-0.0007	0.0105	99.2	0.0	50.2	-0.0007	0.0104	98.5	0.0	2.8

3.3.2 Scenario B: known proportions & a small Dirichlet mean

Under this scenario, we consider $p = 4$ with Dirichlet means for the first three location contexts of $0.35, 0.32, 0.30$, and a small Dirichelet mean for the last location of 0.03. We set the parameters for the first three location contexts at $\beta_1 = 0.80$, $\beta_2 = 0.20$, and $\beta_3 = 0.50$. For the last location context, we took β_4 equal to 0.0, 0.5 and 1.0, aiming to investigate the potential impacts of parameters on the boundary of the parameter space.

Simulation results under this scenario are shown in Table 3.2. The empirical mean bias and especially the empirical standard deviation for both constrained adaptive LASSO and constrained OLS estimates were much bigger in the location context with small a Dirichlet mean. The empirical standard deviation of estimate of the low Dirichlet mean group appeared to be smaller around the boundary (i.e., $\beta_4 = 0.0$ or 1.0) relative to within the interior (i.e., $\beta_4 = 0.5$) of the parameter space.

When β_4 lies on the lower boundary of the parameter space (i.e., $\beta_4 = 0.0$), the constrained adaptive LASSO approach appeared to perform better than the constrained OLS with respect to empirical mean bias at large samples. Nonetheless, we failed to find any significant difference between two approaches when the sample size is small. The constrained OLS estimate was inferior to constrained adaptive LASSO with respect to the percentage of estimates of β_4 at zero. In contrast, the constrained adaptive LASSO approach led more than 50% of estimates for β_4 well on the lower bound for the constraints. As expected, the empirical mean bias for constrained adaptive LASSO and constrained OLS trended towards zero as the sample size increases.

Unlike β_4 on the lower bound, the constrained OLS yielded a substantial number of β_4 estimates well on the upper bound (i.e., $\beta_4 = 1.0$). Moreover, the constrained adaptive LASSO appeared to be superior to the constrained OLS estimates with respect to the empirical mean bias when $\beta_4 = 1.0$. In general, there were no significant difference in the performance of constrained adaptive LASSO relative to constrained OLS when β_4 lies well on the upper bound (i.e., $\beta_4 = 1.0$) or within the interior (i.e., $\beta_4 = 0.5$) of the parameter space. However, it appeared that the constrained adaptive LASSO was more likely to yield estimates of zero and one when the true value of β_4 equal to 0.5 at small sample sizes (i.e., n=100 or 300).

Table 3.2 Simulation results for Scenario B with known proportions & a small Dirichlet mean: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals; % To 0, percentage of estimates falling on zero; % To 1, percentage of estimates falling on one.

			Constrained Adaptive LASSO					Constrained OLS				
n	**Parameter**	**True Value**	**Bias**	**SD**	**CR (%)**	**% To 0**	**% To 1**	**Bias**	**SD**	**CR (%)**	**% To 0**	**% To 1**
100	β_1	0.8	0.0021	0.1139	94.6			0.0012	0.0941	93.6		
	β_2	0.2	0.0097	0.1069	91.8			0.0071	0.1009	93.1		
	β_3	0.5	0.0011	0.0922	94.3			-0.0017	0.1171	93.5		
	β_4	0.0	0.1330	0.2352	96.9	56.7	2.1	0.1400	0.2511	98.2	0.0	3.5
300	β_1	0.8	-0.0006	0.0678	94.5			-0.0013	0.0549	94.2		
	β_2	0.2	0.0001	0.0589	94.6			-0.0007	0.0593	93.3		
	β_3	0.5	-0.0025	0.0562	93.2			0.0008	0.0675	95.2		
	β_4	0.0	0.0708	0.1152	99.3	51.7	0.0	0.0692	0.1199	98.8	0.0	0.0
1000	β_1	0.8	-0.0003	0.0362	95.9			0.0002	0.0309	94.8		
	β_2	0.2	0.0009	0.0327	94.1			-0.0010	0.0319	94.7		
	β_3	0.5	0.0008	0.0286	95.6			-0.0011	0.0361	95.9		
	β_4	0.0	0.0345	0.0537	99.0	51.6	0.0	0.0406	0.0608	98.8	0.0	0.0
100	β_1	0.8	0.0010	0.1159	94.8			0.0038	0.0943	93.5		
	β_2	0.2	0.0041	0.1066	91.4			0.0086	0.1014	93.6		
	β_3	0.5	-0.0003	0.0932	94.6			-0.0001	0.1174	93.5		
	β_4	0.5	-0.0005	0.3555	67.1	15.3	15.7	-0.0007	0.3566	92.8	0.0	14.2
300	β_1	0.8	0.0000	0.0675	94.6			-0.0001	0.0552	94.3		
	β_2	0.2	-0.0002	0.0589	94.9			0.0005	0.0598	93.6		
	β_3	0.5	-0.0020	0.0566	93.4			0.0015	0.0675	95.4		
	β_4	0.5	0.0021	0.2395	89.3	2.5	2.5	-0.0005	0.2356	91.0	0.0	1.7
1000	β_1	0.8	-0.0003	0.0362	95.8			0.0008	0.0311	94.4		
	β_2	0.2	0.0008	0.0329	94.4			-0.0004	0.0320	94.4		
	β_3	0.5	0.0014	0.0286	95.6			-0.0003	0.0361	95.7		
	β_4	0.5	-0.0003	0.1203	94.9	0.0	0.0	0.0018	0.1235	94.5	0.0	0.0

Table 3.2 (continued)

Table 3.2 (continued)

n	Parameter	True Value	Constrained Adaptive LASSO					Constrained OLS				
			Bias	SD	CR (%)	% To 0	% To 1	Bias	SD	CR (%)	% To 0	% To 1
100	β_1	0.8	0.0001	0.1144	94.4			0.0056	0.0934	93.4		
	β_2	0.2	-0.0015	0.1018	92.5			0.0098	0.1020	93.1		
	β_3	0.5	-0.0052	0.0941	94.3			0.0028	0.1179	93.2		
	β_4	1.0	-0.0140	0.2407	96.8	2.4	54.6	-0.0143	0.2504	98.7	0.0	47.3
300	β_1	0.8	0.0001	0.0671	94.5			0.0009	0.0556	94.0		
	β_2	0.2	0.0007	0.0587	95.0			0.0017	0.0599	93.5		
	β_3	0.5	-0.0022	0.0560	93.4			0.0025	0.0675	95.2		
	β_4	1.0	-0.0065	0.1101	99.3	0.0	54.3	-0.0072	0.1182	99.1	0.0	48.3
1000	β_1	0.8	-0.0002	0.0361	96.1			0.0013	0.0309	94.4		
	β_2	0.2	0.0013	0.0329	94.1			0.0002	0.0317	94.8		
	β_3	0.5	0.0012	0.0285	95.2			0.0002	0.0359	95.9		
	β_4	1.0	-0.0034	0.0526	99.5	0.0	51.1	-0.0035	0.0542	99.3	0.0	34.2

3.3.3 Scenario C: estimated proportions, no small Dirichlet means

The objective of this scenario is to assess the impact of replacing known proportions of time spent in each location context by a given participant with estimated proportions. The simulation settings for scenario C with respect to Dirichlet means and regression coefficients were identical to those in scenario A. As in scenario A, proportions of time spent in each location context were independently sampled from a Dirichlet distribution. The frequencies at which participants were observed at the location contexts were then independently sampled from a multinomial distribution with parameters set according to the realization of the Dirichlet distribution and sample size generated from a Poisson distribution with mean 200, the targeted size of phone calls by the COHSONET study. Sample proportions computed from the realization of the multinomial distribution were then used to estimate the proportions of time spent in each location context as generated from the Dirichlet distribution. Constrained adaptive LASSO estimates were then obtained based on estimated location contexts, and results were compared to constrained adaptive LASSO estimates obtained using the true location contexts realized from the Dirchlet distribution.

Table 3.3 presents the simulation results of constrained adaptive LASSO under Scenario C versus Scenario A under the known and estimated location contexts. It appeared that the constrained adaptive LASSO estimates yielded smaller empirical mean biases under known location contexts relative to estimated location context when the sample size is sufficiently big. The ratio of variance of constrained adaptive LASSO estimates modeling from known versus estimated location contexts always varied around 1, which revealed that uncertainty due to variation from sampling and statistical modeling did not pose a substantial impact on the

parameter estimates in the constrained adaptive LASSO estimates. Moreover, there appeared to be equal likelihood for constrained adaptive LASSO under either known or estimated location contexts to shrink parameter estimates to the boundary (i.e., $\beta_4 = 0.0 \text{ or } 1.0$).

Table 3.3 Comparison of simulation results between Scenarios A and C: Bias, empirical mean difference of estimate and true parameter value; SD, empirical standard deviation of bias; CR, percentage coverage of nominal 95% confidence intervals; % To 0, percentage of estimates falling on zero; % To 1, percentage of estimates falling on one; Ratio, empirical mean ratio of variance of bias.

			Known Proportions					Estimated Proportions					
n	Parameter	True Value	Bias	SD	CR (%)	% To 0	% To 1	Bias	SD	CR (%)	% To 0	% To 1	Ratio
100	β_1	0.5	-0.0106	0.1734	91.9			-0.0029	0.1712	92.5			1.025
	β_2	0.2	0.0132	0.1263	91.1			-0.0009	0.1275	89.6			0.982
	β_3	0.8	-0.0052	0.1063	92.0			-0.0019	0.1018	93.0			1.091
	β_4	0.0	0.0268	0.0412	99.2	48.6	0.0	0.0261	0.0394	99.7	51.7	0.0	1.095
300	β_1	0.5	0.0000	0.0981	94.1			-0.0018	0.0963	94.3			1.039
	β_2	0.2	0.0018	0.0736	93.8			0.0030	0.0743	93.0			0.983
	β_3	0.8	-0.0024	0.0607	95.4			-0.0074	0.0601	94.4			1.021
	β_4	0.0	0.0140	0.0213	99.3	51.8	0.0	0.0153	0.0224	99.0	48.7	0.0	0.909
1000	β_1	0.5	-0.0020	0.0541	95.3			0.0006	0.0534	94.1			1.025
	β_2	0.2	-0.0023	0.0409	93.6			0.0002	0.0401	94.4			1.038
	β_3	0.8	-0.0003	0.0333	94.7			-0.0053	0.0326	95.1			1.043
	β_4	0.0	0.0081	0.0118	99.0	50.0	0.0	0.0089	0.0120	98.5	45.2	0.0	0.977
100	β_1	0.5	0.0012	0.1746	93.0			0.0031	0.1744	92.4			1.003
	β_2	0.2	0.0033	0.1323	90.5			-0.0001	0.1333	90.3			0.986
	β_3	0.8	0.0045	0.1041	92.7			0.0021	0.1025	92.3			1.032
	β_4	0.5	-0.0001	0.1087	94.0	0.0	0.0	-0.0001	0.1090	93.5	0.0	0.0	0.994
300	β_1	0.5	-0.0010	0.0991	94.6			0.0006	0.0998	93.5			0.986
	β_2	0.2	-0.0005	0.0792	93.9			0.0024	0.0771	93.8			1.057
	β_3	0.8	0.0017	0.0608	94.3			-0.0036	0.0601	95.2			1.024
	β_4	0.5	-0.0003	0.0617	95.4	0.0	0.0	0.0003	0.0620	94.2	0.0	0.0	0.989
1000	β_1	0.5	-0.0003	0.0551	94.7			0.0031	0.0543	94.4			1.030
	β_2	0.2	-0.0017	0.0443	93.6			0.0011	0.0436	93.4			1.030
	β_3	0.8	0.0018	0.0335	94.7			-0.0027	0.0331	95.1			1.023
	β_4	0.5	0.0000	0.0340	95.1	0.0	0.0	0.0010	0.0344	95.0	0.0	0.0	0.981

Table 3.3 (Continued)

Table 3.3 (Continued)

n	Parameter	True Value	Known Proportions					Estimated Proportions					Ratio
			Bias	SD	CR (%)	% To 0	% To 1	Bias	SD	CR (%)	% To 0	% To 1	
100	β_1	0.5	0.0046	0.1719	92.2			0.0090	0.1636	93.5			1.103
	β_2	0.2	0.0087	0.1328	89.3			0.0044	0.1354	87.7			0.963
	β_3	0.8	0.0031	0.0997	93.1			0.0006	0.0959	93.1			1.081
	β_4	1.0	-0.0026	0.0393	99.5	0.0	50.7	-0.0024	0.0359	99.9	0.0	49.2	1.198
300	β_1	0.5	0.0007	0.0978	94.2			0.0037	0.0974	94.0			1.008
	β_2	0.2	0.0026	0.0794	92.2			0.0073	0.0782	92.3			1.029
	β_3	0.8	0.0015	0.0556	95.4			-0.0026	0.0559	96.5			0.991
	β_4	1.0	-0.0014	0.0210	99.6	0.0	49.6	-0.0015	0.0209	99.2	0.0	46.0	1.013
1000	β_1	0.5	0.0004	0.0546	94.4			0.0057	0.0518	95.0			1.110
	β_2	0.2	0.0008	0.0445	91.0			0.0036	0.0436	92.5			1.039
	β_3	0.8	0.0017	0.0302	96.3			-0.0008	0.0300	96.1			1.011
	β_4	1.0	-0.0007	0.0105	99.2	0.0	50.2	-0.0008	0.0111	98.9	0.0	49.1	0.894

3.4 APPLICATION OF COHSONET DATA

The proposed approaches are illustrated using data from the COHSONET study which was designed to investigate the impact of social contact patterns on the risk of *M. tuberculosis* conversion. We hypothesize that the proportions of time participants spent in different location contexts may be regarded as a surrogate variable for social contact patterns. To evaluate the social contact patterns, a cohort of individuals aged between 15 and 45 years and were free of *M. tuberculosis* infection at baseline were enrolled in the COHSONET study. Participants were prompted to answer a set of questions concerning their location and surrounding environment at the times when calls were answered during a one-year follow-up period. Sampling times when the phone calls were made were randomly generated from a self-correcting point process.

The conditional intensity for a self-correcting point process takes the form

$$\lambda(\mathrm{t}|\mathcal{F}_t) = \exp\{\alpha_0 + \alpha_1(t - \alpha_2 N(t))\}, \qquad t \in [0, T]$$

where α_0 , α_1, and α_2 are constants (Isham and M. 1979; Ogata and Vere-Jones 1984; Vere-Jones and Ogata 1984), and $\alpha_1,\ \alpha_2 > 0$. This point process is a self-correcting in the sense that if the number of events strays from the target $1/\alpha_2$, then the assessment rate compensates to force this difference back towards zero. The baseline intensity is $\exp\{\alpha_0\}$. The parameters α_0 and α_2 govern the mean number of phone calls made per day, while α_1 controls the variability of the number of calls per day and the regularity of the spacing of the assessment times. The self-correcting point process generates more regularly spaced assessment times and less variation in numbers of assessments per day

than the Poisson process, reducing burden on the study subjects. In the COHSONET study, α_0= -0.602, $\alpha_1 = 3$, and α_2 = 1.825 targeting 200 random assessments per year.

In the COHSONET study, only 63.7% of phone calls were answered. Given the substantial amount of missing data, there is potential for bias in estimates of model parameters describing the impact of location contexts on risk of TB conversion. The only information available for unanswered calls is the time and date at which each call was made. Therefore, it is only feasible to describe the pattern of answered phone call as a function of calling times. Let $p_i(t)$ denote the probability that a call at time t is answered by subject i. Let $Z_i(t) = 1$ if a call is answered at time t by subject i, and $Z_i(t) = 0$ if otherwise. Assume that $Z_i(t), t \in N_i$, are independently sampled from a Bernoulli distribution with thinning function $p_i(t)$, where N_i denotes the set of times at which calls are made to subject i, a realization of a point process with intensity $\lambda_i(t)$. Then the set of answered calls N_i^* is a realization of a thinned point process with intensity $\lambda_i(t)p_i(t)$(Cressie 1991). Assume that the data are missing at random, the design-unbiased estimators in (3.2) may be replaced with corrected estimators

$$\hat{X}_i(T) = \frac{1}{T}\sum_{t\in N_i^*}\frac{x_i(t)}{\lambda_i(t)p_i(t)} \quad . \tag{3.4}$$

Exploratory data analysis suggested that the missing data pattern depended on the time of day, a pattern that is likely to vary among study participants. The location contexts in which participants spend their time are also likely to be a function of time of day, a function that may also vary among study participants. We assume that the thinning function is periodic, as described through its logit transformation,

$$log\frac{p_i(t)}{1-p_i(t)} = \sum_{k=1}^{K} u_{ik}\cos(\frac{2\pi kt}{\tau} + \phi_{ik})$$

where u_{ik} denotes the amplitude, τ represents the period set to 1 (day), and ϕ_{ik} denotes the phase. The model may be reparameterized by writing

$$log\frac{p_i(t)}{1-p_i(t)} = \gamma_{i0} + \sum_{k=1}^{K}\left\{\gamma_{1i}\ cos\left(\frac{2\pi kt}{\tau}\right) + \gamma_{2ik}sin\left(\frac{2\pi kt}{\tau}\right)\right\},$$

where the amplitude is $u_{ik} = \sqrt{\gamma_{1ik}^2 + \gamma_{2ik}^2}$ and the phase is $\phi_{ik} = -tan^{-1}(\gamma_{1ik}/\gamma_{2i}\ )$. As to describe variation among participants' missing data patterns, the parameter vectors γ_i are assumed to be independently sampled from a multivariate normal distribution with mean μ and variance-covariance matrix Σ.

Laplace approximations to the likelihood (Breslow and Clayton 1993) and maximum hierarchical likelihood (Lee and Nelder 1996), both lead to inconsistent estimates when the sampling domain is small (Rathbun and Shiffman 2016). The Expectation-Maximization (EM) algorithm can produce consistent estimates in the random effects model regardless of sampling domains. Nevertheless, it remains challenge to compute the E-step in the random effects model because the conditional expectation is an intractable integral. Steele (1996) proposed using a second-order Laplace approximation for computation of conditional expectations within the E-step (Steele 1996) for generalized linear mixed models. We implemented Steele's (1996) method for parameter estimation in the random effects model using FORTRAN code available in the supplementary material of Rathbun and Shiffman (2016).

In current study, only subjects who responded to more than 30 phone calls over the study are eligible for inclusion in the data. A total of 288 subjects who received more than 30 phone calls were included in the random effects modeling. Previous investigation revealed that participants were most likely to respond to the phone calls in the early morning (i.e., 7:00 am -8:00 am), and least likely to answer phone calls at the end of a day (Figure 1.1). In addition, there appeared to be an increasing trend in the patterns of answering phone calls between 9:00am and 7:00pm.

We found that on average participants spent the most time at homes (i.e. 32.4%), followed by work places (32.1%), public transports (7.1%), and shopping centers (4.1%). It seems that participants in the COHSONET study rarely spent time at women groups, gyms/recreations, clubs, schools, neighbors' homes and hospitals (less than 1%) (Figure 1.2). To evaluate different approaches in the linear probability model, only complete cases containing both *M. tuberculosis* conversion information and proportions of time spent in each location context available were included, and thus observations from a total of 189 participants were included in the statistical modeling in the rest of this section.

Table 3.4 presents parameter estimates of constrained adaptive LASSO and constrained OLS in the linear probability model over all location contexts. Adaptive LASSO estimates falling on either boundary are treated as known, and so confidence intervals are not computed (Wong et al. 2016). Based on the outputs from the constrained adaptive LASSO approach, we can see that a total of seven location contexts were of zero risk of contracting *M. tuberculosis* including the relatives' homes, work places, clubs/associations, saloons, gyms/recreations, women groups and neighbors' homes,

while there appeared to be 100% risk of infection with *M. tuberculosis* for people spending all of their time at four different location contexts including schools, worship centers, bars and shopping/trading centers. Similar conclusions were implied by the constrained OLS estimates except in hospitals/clinics, where the constrained OLS yielded an estimate of zero compared to the constrained adaptive LASSO estimate of 0.3693. Moreover, the widths of 95% confidence intervals in the constrained adaptive LASSO appeared to be narrower than those in the constrained OLS estimates. Therefore, the constrained adaptive LASSO estimates were subject to less uncertainty relative to the constrained OLS estimates.

Table 3.4 Parameter estimates of constrained adaptive LASSO and constrained OLS in the linear probability model over all location contexts (n=189). 95% CI: 95% confidence interval.

	Constrained Adaptive LASSO		Constrained OLS	
Location Context	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Home	0.2023	(0.0000-0.4463)	0.2905	(0.0000-0.6022)
Friend's home	0.3089	(0.0000-0.9437)	0.5472	(0.0525-1.0000)
Relative's home	0.0000	----	0.0139	(0.0000-0.9372)
Work	0.0000	----	0.0000	(0.0000-0.3144)
School	1.0000	----	1.0000	(0.0609-1.0000)
Worship center	1.0000	----	1.0000	(0.0767-1.0000)
Club/Association	0.0000	----	0.0000	(0.0000-0.9473)
Bar	1.0000	----	1.0000	(0.0726-1.0000)
Saloon	0.0000	----	0.0000	(0.0000-0.9412)
Gym/Recreation	0.0000	----	0.0000	(0.0000-0.9479)
Hospital/Clinic	0.3693	(0.0000-0.9475)	0.0000	(0.0000-0.9448)
Shopping/Trading center	1.0000	----	1.0000	(0.0717-1.0000)
Public transport	0.7257	(0.0806-1.0000)	0.5007	(0.0298-0.9716)
Women group	0.0000	----	0.0000	(0.000-0.9497)
Market place	0.1029	(0.0000-0.9387)	0.0000	(0.0000-0.9410)
Neighbor's home	0.0000	----	0.0000	(0.0000-0.9472)
Elsewhere	0.8927	(0.1501-1.0000)	0.6434	(0.0781-1.0000)

Since results in Table 3.4 tended to yield wide confidence intervals, covering most of the parameter space, the full model including all location contexts might be regarded as over-parameterized. Therefore, we fit models based on the reduced location contexts by pooling friend's, relative's, and neighbor's homes into one location context as "Other home", and pooling the remaining locations expect for home, other home, work place, and public transport into "Elsewhere". Based on Table 3.5, the estimated risk of *M. tuberculosis* conversion at home and public transport appeared to decrease as compared to the modeling results in Table 3.4. In particular, the estimated risk of contracting *M. tuberculosis* at public transport reduced by about 25% using the constrained adaptive LASSO approach. People remained at zero risk of infecting *M. tuberculosis* at work place in terms of both constrained OLS and constrained adaptive LASSO estimates. In general, there were no significant difference between constrained adaptive LASSO and constrained OLS estimates under the reduced location contexts setting in terms of the 95% confidence intervals.

Table 3.5 Parameter estimates of constrained adaptive LASSO and constrained OLS in the linear probability model based on the reduced location contexts (n=189). 95% CI: 95% confidence interval.

	Constrained Adaptive LASSO		Constrained OLS	
Location Context	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Home	0.2212	(0.0000-0.5174)	0.2575	(0.0000-0.5566)
Other home*	0.5330	(0.0568-1.0000)	0.3022	(0.0000-0.9360)
Work	0.0000	----	0.0000	(0.0000-0.3018)
Public transport	0.4816	(0.0114-0.9518)	0.4129	(0.0000-0.9372)
New elsewhere*	0.6767	(0.2385-1.0000)	0.6823	(0.2136-1.0000)

* Other home is a pooled location of "Friend's home", "Relative's home", and "Neighbor's home"; New elsewhere is a pooled location of "School", "Worship center", "Club/Association", "Bar", "Saloon"," Gym/Recreation", "Hospital/Clinic", "Shopping/Trading center", "Women group", "Market place" and "Elsewhere" in table 3.4.

3.5 DISCUSSION

The proposed adaptive LASSO for the constrained linear probability model is offered as an alternative to the constrained OLS method. The simulation study reveals that the constrained adaptive LASSO is a competitive alternative to the constrained OLS in the linear probability model. In particular, the constrained adaptive LASSO performs better than constrained OLS in the presence of some parameters lying on the boundary of the parameter space.

In the COHSONET study, the true proportions of time participants spent in different location contexts were unavailable, so we instead substituted them with design-unbiased estimators in the proposed statistical models. Since the estimated proportions of time participants spent in different location contexts are subjected to variation sourced from

random sampling and missing data modeling, it is challenging to obtain estimate of variation due to these sources. Nonetheless, simulation studies based on estimated covariates suggested that such errors had little impact on the parameter estimates, so we can obtain unbiased parameter estimates using the design-unbiased estimators in the COHSONET data.

In summary, the constrained adaptive LASSO is a robust approach in the linear probability model under constraints that parameters lie in the unit interval. It has good performance regardless of the assumption of heteroscedasticity. Moreover, constrained adaptive LASSO estimators are consistent and asymptotically unbiased.

3.6 REFERENCES

Box GEP, and Tiao GC. 1992. Bayesian Inference in Statistical Analysis. New York: John Wiley & Sons.

Breslow NE, and Clayton DG. 1993. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association 88(421):9-25.

Cassel CM, Sarndal CE, and Wretman JH. 1977. Foundations of Inference in Survey Sampling. New York: Wiley.

Chen MH, and Shao QM. 1999. Monte Carlo Estimation of Bayesian Credible and HPD Intervals. Journal of Computational and Graphic Statistics 8:69-92.

Cressie N. 1991. Statistics for Spatial Data. New York: Wiley.

Deke J. 2014. Using the Linear Probability Model to Estimate Impacts on Binary Outcomes in Randomized Controlled Trials. Evaluation Technical Assistance Brief for OAH & ACYF Teenage Pregnancy Prevention Grantees.

Fan J, and Li R. 2001. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association 96:1348-1360.

Firth D. 1993. Bias reduction of maximum likelihood estimates. Biometrika 80(1):27-38.

Gart JJ, and Zwifel JR. 1967. On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika 54(1/2):181-187.

Hellevik O. 2009. Linear versus logistic regression when the dependent variable is a dichotomy. Quality & Quantity 43(1):59-74.

Hirji KF, Mehta CR, and Patel NR. 1987. Computing Distributions for Exact Logistic Regression. Journal of the American Statistical Association 82(400):1110-1117.

Horby P, Pham QT, Hens N, Nguyen TT, Le QM, Dang DT, Nguyen ML, Nguyen TH, Alexander N, Edmunds WJ et al. . 2011. Social contact patterns in Vietnam and implications for the control of infectious diseases. PLoS One 6(2):e16965.

Houk VN, Baker JH, Sorensen K, and Kent DC. 1968. The epidemiology of tuberculosis infection in a closed environment. Arch Environ Health 16(1):26-35.

Huang T, Gong H, Yang C, and He Z. 2013. ProteinLasso: A Lasso Regression Approach to Protein Inference Problem in Shotgun Proteomics. Computational Biology and Chemistry 43:46-54.

Isham V, and M. W. 1979. A self-correcting point process. Stochastic Processes and their Applications 8(3):335-347.

James GM, Paulson C, and Rusmevichientong P. 2013. Penalized and Constrained Regression. University of Southern California.

Kenyon TA, Valway SE, Ihle WW, Onorato IM, and Castro KG. 1996. Transmission of multidrug-resistant Mycobacterium tuberculosis during a long airplane flight. N Engl J Med 334(15):933-938.

Koch KR. 1990. Bayesian inference with geodetic applications. Berlin Heidelberg, New York: Springer.

Lee Y, and Nelder JA. 1996. Hierarchical generalized linear models. Journal of the Royal Statistical Society B 58:619-678.

Liu Y, Gelman A, and Zheng T. 2015. Simulation-efficient shortest probability intervals. Statistics and Computing 25:809-819.

Mandelkern M. 2002. Setting confidence intervals for bounded parameters. Statistical Science 17:149-172.

Ogata Y, and Vere-Jones D. 1984. Inference for earthquake models: a self-correcting model. Stoch Proc Appl 17:337-347.

Park D, and Park S. 2003. Small sample properties of parametric and nonparametric estiamtors in quantal bioassay. Inter-Stat (London) 5:1-13.

Pedroza C, and Troung VTT. 2016. Performance of models for estimating absolute risk difference in multicenter trials with binary outcome. BMC Medical Research Methodology 16(113).

Rathbun SL, and Shiffman S. 2016. Mixed Effects Models for Recurrent Events Data with Partially Observed Time-Varying Covariates: Ecological Momentary Assessment of Smoking. Biometrics 72(1):46-55.

Rehkopf D, Furumoto-Dawson A, Kiszewski A, and Awerbuch-Friedlander T. 2015. Spatial Spread of Tuberculosis through Neighborhoods Segregated by Socioeconomic Position: A Stochastic Automata Model. Discrete Dynamics in Nature and Society 2015.

Roe BP, and Woodroofe MB. 2001. Setting confidence belts. Physical Review 60:3009-3015.

Schwarz G. 1978. Estimating the dimension of a model. Ann Stat 6:461-464.

Shapiro A. 2000. On the Asymptotics of Constrained Local M-Estimators The Annals of Statistics 28(3):948-960.

Shiffman S, Stone AA, and Hufford MR. 2008. Ecological momentary assessment. Annual Review of Clinical Psychology 4:1-32.

Steele BM. 1996. A modified EM algorithm for estimation in generalized mixed models. Biometrics 52(4):1295-1310.

Stone AA, and Shiffman S. 2002. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 24(3):236-243.

Tian L, Wang R, Cai T, and Wei LJ. 2011. The Highest Confidence Density Region and Its Usage for Joint Inferences about Constrained Parameters. Biometrics 67:604-610.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58:267–288.

Tibshirani R. 1997. The lasso Method for Variable Selection in the Cox Model. Statistics in Medicine 16:385-395.

Vere-Jones D, and Ogata Y. 1984. On the moments of a self-correcting process. J Appl Prob 21:335-342.

Vexler A, Zou L, and Hutson AD. 2016. Data-Driven Confidence Interval Estimation Incorporating Prior Information with an Adjustment for Skewed Data. THE AMERICAN STATISTICIAN 70(3):243-249.

Wagler A. 2011. Bias Reduced Logistic Dose-Response Models. Journal of Biopharmaceutical Statistics 21:405–422.

Wang H. 2008. Confidence intervals for the mean of a normal distribution with restricted parameter space. Journal of Statistical Computation and Simulation 78(9):829-841.

Wang H, Li R, and Tsai CL. 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94(3):553-568.

WHO. 2014. Global Tuberculosis Report 2014. Geneva.

Wong KY, Goldberg Y, and Fine JP. 2016. Oracle Estimation of Parametric Models under Boundary Constraints. Biometrics 72:1173–1183.

Wright SJ. 2015. Coordinate descent algorithms. Mathematical Programming 151(1):3-34.

Zhang HH, and Lu W. 2007. Adaptive Lasso for Cox's proportional hazards model. Biometrika 94(3):691–703.

Zhang T, and Woodroofe M. 2003. Credible and confidence sets for restricted parameter spaces. Journal of Statistical Planning and Inference 115(2):479-490.

Zhang Y, Li R, and Tsai CL. 2010. Regularization Parameter Selections via Generalized Information Criterion. Journal of the American Statistical Association 105(489):312-323.

Zou H. 2006. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101(476):1418-1429.

CHAPTER 4

CONCLUSIONS

4.1 SUMMARY

Ecological Momentary Assessment (EMA) captures real-time states and behaviors in subjects' natural environments. The time-stamped EMA technique has several advantages over the traditional methods such as reducing recall biases and establishing the temporal ordering of events (Shiffman et al. 2008; Stone and Shiffman 1994; Stone and Shiffman 2002). The implementation of EMA study relies on portable electronic devices such as personal digital assistants and smart phones. Technological advancement has made EMA possible in more and more contexts. In recent years, EMA has been applied to study the prevention behaviors of infectious diseases such as HIV (Barta et al. 2008; Cook et al. 2016).

In the Community Health Study of Social Networks and Tuberculosis (COHSONET), Whalen's ongoing studies on transmission dynamics of *M tuberculosis* in the Rubaga Division of Kampala, Uganda, EMA was employed to investigate social contact patterns among *M. tuberculosis* free subjects at baseline and for one year of follow up. We treated proportions of time patients spent in different location contexts as a surrogate variable for social contacts, and estimated them through design-unbiased estimators (Cassel et al.

1977) since the EMA samples were generated from a known probability-based sampling design.

This dissertation aims to model the risk of *M. tuberculosis* conversion as a function of proportions of time participants spent in different location contexts. Logistic regression is the most commonly method being applied to the data with a dichotomous outcome. However, the maximum likelihood estimates (MLEs) of the logistic regression model have a bias of order $O(n^{-1})$ (Firth 1993). Although a penalized MLE approach has been proposed by Firth (1993) to reduce bias in MLE, we did not favor the logistic regression model because it is nonlinear. We were concerned that the use of design-unbiased predictors (i.e., proportions of time participants spent in different location contexts) might lead to unpredictable uncertainty in the parameter estimates from the logistic regression model. Therefore, we proposed a linear probability model, which we hoped would reduce the uncertainty related to the logistic regression model, as an alternative option. One salient issue that confronts us in the linear probability model is that direct application of any estimating approaches such as ordinary least squares (OLS) may produce meaningless results that the predictive value of the outcome is outside the interval of zero to one.

In order to address this limitation, we proposed two different approaches which constrain the predicted values from the linear probability model to lie on the unit interval between zero and one. The first approach that we proposed was the constrained OLS in Chapter 2, the implementation of which leans on the quasi-Newton technique (Powell 1982a; Powell 1982b). Inspired by previous work concerning optimized estimators under constraints

(Self and Liang 1987; Shapiro 2000; Wong et al. 2016), we demonstrated that the constrained estimators are root-n consistent and have a unique asymptotic distribution. Since heteroscedasticity is a potential threat in the linear probability model, we also investigated constrained weighted least squares (WLS) estimates. However, both theoretical demonstration and simulations studies suggested that the constrained WLS yielded biased parameter estimates due to a nonlinear, parameter-dependent weighting term.

The second approach, the constrained adaptive LASSO in Chapter 3, was more desirable due to its capabilities in model selection, regularization as well as imposing boundary constraints on the parameter estimation. There is evidence that the adaptive LASSO is an oracle procedure. That is, the model selection is consistent, and parameter estimation is consistent as well as asymptotically normal. Following Wong et al. (2016), the constrained adaptive LASSO estimates are consistent and approach a limiting distribution in the constrained parameter space. Particularly, estimates of the constrained adaptive LASSO are asymptotically normal if none of parameter estimates falls on the boundary of the constraints.

All proposed approaches were compared through a set of simulation studies. Overall, the constrained OLS performed as well as the constrained adaptive LASSO except for parameters lying on the constrained boundaries. It appeared that the constrained OLS was poor at forcing estimates to the boundary. However, the constrained OLS was not inferior to the constrained adaptive LASSO if the location context with a small Dirichlet mean but with strong impact on the risk of dependent variable (e.g., the true parameter value

for this location context equal to 1.0). In addition, simulation studies revealed that the constrained OLS in the linear probability model was a competitive option for the logistic regression. The constrained OLS estimates tended to be less biased than the MLE and penalized MLE estimates in the logistic regression. In particular, the logistic regression model had a poor performance in the presence of parameters close to the boundary of the parameter space. Furthermore, simulation studies indicated that the potential variations due to random sampling of the EMA data as well as from the random effects modeling of the probability of answering the phone calls might not pose substantial threats on the parameter estimation in both logistic regression and constrained linear probability models.

4.2 FUTURE RESEARCH

Together with its ease of interpretability, this dissertation demonstrated that the constrained linear probability model is competitive alternative to the logistic regression model. The proposed constrained OLS and adaptive LASSO estimates in the linear probability model are not only less biased than the MLE and penalized MLE estimates in the logistic regression model, but they also perform far better when some parameters in the model have values close to or lie on the boundary of the constrained parameter space. In spite of these appealing merits of the proposed constrained approaches, there are further areas of research remain untouched in this dissertation.

4.2.1 Measurement error model

In the COHSONET study, the true proportions of time participants spent in different location contexts were not available, so we replaced them with design-unbiased estimators. The calculation of design-unbiased estimators relied on two assumptions: (1) The sampling times of the EMA data were randomly selected from a self-correcting point process with known intensities; (2) The made phone calls were not responded randomly, and thus a random effects model proposed by Rathbun and Shiffman (2016) was well suited to predict the probability of answering a phone call (Rathbun and Shiffman 2016). Therefore, parameter estimates in the proposed statistical models are subjected variation due to sampling as well as random effects modeling. Although simulations studies indicated these two sources of variation had minimal impacts on parameter estimates, quantifying that impact remains a future direction of research.

4.2.2 Estimation of confidence interval bounds

The construction of confidence intervals is beyond the scope of the current work, but it is crucial to obtain unbiased confidence intervals in the field of statistical modeling. According to Wong et al. (2016), confidence intervals can be constructed based upon the standard (i.e., not constrained) estimations if all parameter estimates lie within the interior of the boundaries. While for modelling results containing some parameter estimates on the boundary, treating these parameters as known and establishing the confidence intervals based upon the rest of parameters which are within the interior of the constraints (Wong et al. 2016). However, confidence intervals built upon Wong et al.'s assumption could lead to estimates beyond the boundary, especially when the sample size

is small, or a parameter estimate is close to the boundary, or the interval of the boundary is narrow, or any combination of the listed situations. To address this issue, the current dissertation used the highest posterior density (HPD) (Box and Tiao 1992) to obtain confidence regions. In spite of its good performance among parameters with asymmetric distributions, the HPD confidence regions have been criticized by some researchers for being too dependent on the parametric assumption (Roese-Koerner et al. 2012; Vexler et al. 2016; Zhou and Reiter 2010). Zhou and Reiter (2010) demonstrated that the HPD estimators are biased if the parametric assumptions are not correct. In consideration of the under-coverage rate of 95% confidence intervals in the proposed constrained estimators, we suspect that HPD confidence regions might not be suitable here. Therefore, it can be a future research direction to develop optimal confidence intervals for parameter estimates under certain constraints.

4.3 REFERENCES

Barta WD, Portnoy DB, Kiene SM, Tennen H, Abu-Hasaballah KS, and Ferrer R. 2008. A daily process investigation of alcohol-involved sexual risk behavior among economically disadvantaged problem drinkers living with HIV/AIDS. AIDS Behav 12:729-740.

Box GEP, and Tiao GC. 1992. Bayesian Inference in Statistical Analysis. New York: John Wiley & Sons.

Cassel CM, Sarndal CE, and Wretman JH. 1977. Foundations of Inference in Survey Sampling. New York: Wiley.

Cook PF, McElwain CJ, and radley-Springer LAB. 2016. Brief report on ecological momentary assessment: everyday states predict HIV prevention behaviors. BMC Res Notes 9(9).

Firth D. 1993. Bias reduction of maximum likelihood estimates. Biometrika 80(1):27-38.

Powell MJD. 1982a. Extensions to Subroutine VF02AD. In: R. F. Drenick and F. Kozin e, editor. Systems Modeling and Optimization, Lecture Notes in Control and Information Sciences Berlin-Heidelberg-New York: Springer-Verlag.

Powell MJD. 1982b. VMCWD: A Fortran Subroutine for Constrained Optimization. DAMTP 1982/NA4. Cambridge, England.

Rathbun SL, and Shiffman S. 2016. Mixed Effects Models for Recurrent Events Data with Partially Observed Time-Varying Covariates: Ecological Momentary Assessment of Smoking. Biometrics 72(1):46-55.

Roese-Koerner L, Devaraju B, Sneeuw N, and Schuh W-D. 2012. A stochastic framework for inequality constrained estimation. Journal of Geodesy 86(11):1005-1018.

Self SG, and Liang K. 1987. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio TestsUnder Nonstandard Conditions. Journal of the American Statistical Association 82(398):605-610.

Shapiro A. 2000. On the Asymptotics of Constrained Local M-Estimators The Annals of Statistics 28(3):948-960.

Shiffman S, Stone AA, and Hufford MR. 2008. Ecological momentary assessment. Annual Review of Clinical Psychology 4:1-32.

Stone AA, and Shiffman S. 1994. Ecological momentary assessment (EMA) in behavioral medicine. Annals of Behavioral Medicine 16:199-206.

Stone AA, and Shiffman S. 2002. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 24(3):236-243.

Vexler A, Zou L, and Hutson AD. 2016. Data-Driven Confidence Interval Estimation Incorporating Prior Information with an Adjustment for Skewed Data. The American Statistician 70(3):243-249.

Wong KY, Goldberg Y, and Fine JP. 2016. Oracle Estimation of Parametric Models under Boundary Constraints. Biometrics 72:1173–1183.

Zhou X, and Reiter JP. 2010. A Note on Bayesian Inference afterMultiple Imputation. The American Statistician 64:159-163.