Summer 7-16-2019

Estimation of Association Between a Longitudinal Marker and Interval-Censored Progression Times

Naghmeh Daneshi
Portland State University

Estimation of Association Between a Longitudinal Marker and Interval-Censored Progression Times

by

Naghmeh Daneshi

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Mathematical Sciences

Dissertation Committee:
Jong Sung Kim, Chair
Robert Fountain
Subhash Kochar
Alexis Dinno
Wayne Wakeland

Portland State University
2019

Abstract

In longitudinal studies, we observe the subjects who are likely to progress to a new state during the study time. For example, in clinical trials the stage of a progressing disease is recorded at each follow-up visit. The primary goal is to estimate the relationship between the attributes and the subject's progression state. In such studies, some subjects complete all their follow-up visits and their progression state are observed without any missingness. However, others miss their follow-up visits and when they come back, they learn that they have progressed to a new state. In this case, not only are their progression states at each follow-up interval-censored, but their time-dependent covariates are incomplete. In such studies, the observations are missing at random (MAR).

The event of interest, i.e., progression, may have several possible patterns. In some studies, we might be studying progression to only one new state. For example, we are interested in studying the attributes that affect an individual's progression from being a non-smoker to a frequent smoker. Another example would be the patients who are believed to have high risk for developing diabetes, are monitored for advancing to type 2 diabetes. In other studies, the event of interest involves multiple stages. Examples of these studies include several stages of cancer, or different stages of smoking (non-smoker, light smoker, intermittent smoker, heavy smoker, etc.). These states are chronological.

The times of observation, i.e., follow-up interview visits, are pre-specified for these studies. At each time point, the attributes are measured and recorded. Since the

study continues over time, it is common for some subjects to miss their follow-up visits. In this case not only the outcome (event of interest) is censored, but their time-dependent attributes are incomplete. In this case, both outcome and attributes need to be estimated for the missed visits.

We are interested in studying the time-dependant covariates' effect on the progression. Expectation-maximization (EM) algorithm is used for estimating the parameters. The variance-covariance matrix of the maximum likelihood estimator (MLE) is calculated using the missing information principles. Simulation studies revealed that the proposed method works well in terms of variance, bias, and power in the samples of moderate sizes.

When we are estimating the association between longitudinal covariates and an event, we may run into the large number of attributes, which are explanatory but could be highly correlated. Using the usual maximum likelihood estimation method leads to inaccurate parameter estimates. Additionally, the estimators have large variance. Elliot, et al., [**14**] proposed Mixed Ridge Regression when the outcome of the process is continuous. This method applies ridge regression to a linear mixed effects longitudinal model. In our proposed model, the longitudinal outcome is binary. We apply ridge penalization (based on the L^2 norm) to our model to get more accurate parameter estimates.

Another important aspect in building a good predictive model is variable selection. Sometimes there are many attributes in a dataset. These attributes are not necessarily correlated. We are interested in choosing a smallest best subset of them for inference. We perform the variable selection by adding the LASSO penalization (based on the L^1 norm) to the likelihood to be able to simultaneously choose the appropriate covariates and estimate the covariate effects.

Lastly, the preliminary model is extended to the case when there are more than one progression states in the model. These progressions are chronological and assumed to be non-time-reversible. Missing pattern is more complex than that for one progression state case, but the rest of the procedures are pretty similar to those for one progression state case.

Dedicated to My beloved Mom, Ms. Fariba Maghsoodlou and My late grandparents, Mr. Manoocher Maghsoodlou and Mrs. Ezat Ebrahimi

Acknowledgments

I would like to thank my advisor, Dr. Jong Sung Kim, for introducing me to research. Dr. Kim has always been very supportive and helping me with my research topic and stimulating great discussions. He is a great educator. I have learnt a lot from him through taking various classes from him and also doing research.

Additionally, many thanks to my dissertation committee, Dr. Robert Fountain, Dr. Subhash Kochar, Dr. Alexis Dinno, and Dr. Wayne Wakeland for taking the time to read my work and provide helpful feedback. I specially appreciate Dr. Dinno for introducing the data.

I would like to thank Dr. Fariborz Maseeh, the donor of the Eugene Enneking Doctoral Fellowship for providing a great opportunity for me to focus on my research. I was honored to be the recipient of this award several times.

I thank all of the other faculty that taught me the various aspects of statistical science.

Lastly, I would like to thank my beloved mom, Ms. Fariba Maghsoodlou who has always been a role model to me and has supported me all my life. Words are not enough to appreciate her. I thank my late grandparents, Mr. Manoochehr Maghsoodlou and Mrs. Ezat Ebrahimi who were very supportive and always emphasized

the importance of education. Their help and supports mean a lot to me. I wouldn't have accomplished this without my mom and grandparents.

Table of Contents

List of Tables

List of Figures

1

Introduction

In the analysis of longitudinal studies, subjects who are likely to progress to a new state during the study, are monitored over time. For example, in clinical trials, patients who are at high risk of a certain disease are monitored and have follow-up visits. Some subjects complete all of their follow-up visits and their failure times are recorded accurately. But others may miss their follow-up visits and when they come back, they learn that they have progressed to a new state, i.e., the event of interest has already occurred. The event for these patients is censored within the person-specific time interval. This is known as "partly interval-censored failure time data". Although there are multiple follow-up visiting intervals, for each subject, researchers use one particular interval that is only known to contain the true unknown failure time, unless they had accurately determined the failure time. There are quite a few research works based on partly interval-censored data such as [**18**], [**46**], [**26**], and [**23**] among others.

Another commonly available data type in longitudinal studies is called pooled repeated observations (PRO). In such studies, subjects have multiple follow-up visits as usual. A subject obtains a binary outcome for the event from every visit. All those repeated binary outcomes are pooled together to develop a model to analyze the effects of time-dependent covariates on the occurrence of the event. [**8**] and [**9**] pooled such repeated observations with binary outcomes for the event of interest into a single sample. Then they used logistic regression model to estimate the effects of

the risk factors to the occurrence of the event. Each observation interval is considered a mini follow-up study in which the current risk factors are updated to predict events in the interval. Once an individual has an event in a particular interval, all subsequent intervals from that individual are excluded from the analysis.

In the present study, we define pooled repeated observations for partly interval-censored data. We have pooled repeated observations but some binary outcomes are incomplete, and can only be determined with certain unknown probabilities within some specific follow-up visits. In this case, the analysis of such data would need a method that combines some models that handle pooled repeated observations without censoring and methods that deal with partly or completely interval-censored data.

The main goal of this study is to estimate the effects of the time-dependent covariates on the occurrence of the event of interest (e.g., progression to a disease, becoming a frequent smoker, etc.). We extended the work of [**16**], who employed conditional expected score test (CEST) to develop a test for association of a longitudinal marker and an event with missing binary outcomes, to the estimation problem when the event of interest has a single progression state or double progression states and the response is pooled, repeated and partly interval-censored.

2

Longitudinal Data Analysis in Practice

2.1. Rao Score Test

The Rao score test for testing the hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is

$$\left[\frac{l'(\theta_0)}{\sqrt{nI(\theta_0)}}\right]^2 \rightarrow \chi^2_{(1)} \text{in distribution} \tag{1}$$

where $l(\theta_0)$ is the log-likelihood function evaluated under the null hypothesis, $l'(\theta_0) = \frac{\partial l(\theta_0)}{\partial \theta_0}$, and $I(\theta_0)$ is the Fisher's information.

2.2. CEST for Association of Longitudinal Markers and Interval-Censored Event Times

One of the applications of longitudinal studies of subjects for the occurrence of an event of interest is in clinical trials. Patients who are at high risk of progressing to a disease, are monitored over time. These patients may miss their follow-up visits and the disease has progresses when they return. The progression of the disease is unobserved and has happened during a time interval (It is interval-censored). For these patients, the markers, e.g., lab tests are missing during the interval-censored follow-up visits. [**16**] proposed a conditional expected score test (CEST) to see whether there is an association between a longitudinal covariate and patient‘s failure time.

Let T_i be the time that patient i has an event, $i = 1, \ldots, n$. t_j is the clinic visits at which patients are monitored, $j = 1, \ldots, M$. Let δ_{ij} be the indicator of whether or not patient i has had an event (failure) since previous visit at t_{j-1}. Y_{ij} is the indicator

of whether or not patient i was in follow-up (and at risk) at visit t_j. z_{ij} is patient i's covariate, measured at time t_{j-1} (during the jth interval).

The response variable, δ_{ij} is modeled using the logit link. It indicates failure (progression of the disease) as a function of the covariate, z_{ij} which was measured at time t_{j-1} (during the jth interval).

$$\text{logit}P_\beta(z_{ij}) = \log \frac{P_\beta(z_{ij})}{1 - P_\beta(z_{ij})} = \alpha + \beta' z_{ij} \tag{2}$$

where $P_\beta(z_{ij})$ is the conditional probability of patient i having an event during the jth interval and was event-free through t_{j-1} and z_{ij} is the value of covariate at this time.

The complete log-likelihood is

$$l = \sum_{i=1}^{n} \sum_{j=1}^{M} [-Y_{ij} \log(1 + \exp(\alpha + \beta' z_{ij})) + \delta_{ij}(\alpha + \beta' z_{ij})] \tag{3}$$

The numerator of score test of $H_0 : \beta = 0$, assuming that there are no missing (or censored) observations and the data are complete, is:

$$U = l' = \frac{\partial}{\partial \beta} l = \sum_{i=1}^{n} \sum_{j=1}^{M} (\delta_{ij} z_{ij} - Y_{ij} z_{ij} \widehat{p}) \tag{4}$$

where $\widehat{p}$ is the maximum likelihood estimator(MLE) of $P = P_0 = \frac{\exp(\alpha)}{1+\exp(\alpha)}$ under $H_0 : \beta = 0$ and it is computed by

$$\widehat{p} = \frac{\sum_i \sum_j \delta_{ij}}{\sum_i \sum_j Y_{ij}} \tag{5}$$

In order to use the CEST for $H_0 : \beta = 0$, it is assumed that event's data could be censored due to missing visits. Assume that ith subject missed visits after time

t_{L_i} and came back at t_{R_i}, where L_i is the index of the last time that patient i was visited and was event-free (did not fail) and R_i is the index of the event's time. Then δ_{ij} is missing during the $L_{i+1} \leq j \leq R_{i-1}$. Let $\delta_{i0} = \sum_{j=1}^{M} \delta_{ij}$ be the indicator of whether subject has ever failed or left follow-up without having an event. Then $\delta_{i0} = 1$ indicates that subject i failed while they were in follow-up. This means that the last time patient i was observed is the event's time. $\delta_{i0} = 0$ indicates that subject i has not failed by last visit, i.e., the event's time is right-censored. Furthermore, whenever subjects miss a visit, the value of their covariate z_{ij} is missing as well. [**43**] suggested that z_{ij} has linear growth curve with random effects and used the predicted values that are found using this model to impute missing values [**5**].

The CEST can be derived from (4) using a principle (noted by [**13**]) that the score based on incomplete data is the expected value of complete data, conditioned on observed data.

$$\partial L^y(\phi) = E_\phi(\partial L^x(\phi)|y) \tag{6}$$

where $\partial L^x(\phi)$ is the Fisher score function based on complete data x and $\partial L^y(\phi)$ is the score function based on the incomplete data y.

The following four steps are required to obtain the score test for incomplete data:

Step 1. Select a model.

Step 2. Find the score test for the complete data.

Step3. Find the conditional expected value of the score test (obtained in step 2) given the incomplete data, under the null hypothesis.

Step 4. The test statistic obtained in step 3 is the CEST.

To drive the test for $H_0 : \beta = 0$ using CEST we need:

$$
\begin{aligned}
& E[\delta_{ij}|Y,\delta,\beta=0] \\
& = P[T_i = t_j | t_{L_i} < T_i \leq t_{R_i}, Y, \widehat{x}_{ij}, \alpha, \beta] \\
& = \frac{p(1-p)^{j-1}}{\sum_{k=L_i+1}^{R_i} p(1-p)^{k-1}}, k \in (\, L_i, R_i]\} \\
& = w_{d_{ij}}
\end{aligned} \tag{7}
$$

$w_{r_{ij}} = 1 - \sum_{l=1}^{j-1} w_{d_{il}}$ is the probability that patient i has had an event on or after the visit t_j given Y,δ, and $\beta = 0$.

The MLE of $P = \frac{\exp(\alpha)}{1+\exp(\alpha)}$ under the incomplete data, assuming $H_0 : \beta = 0$ is

$$
\widehat{p} = \frac{\sum_i \sum_j w_{d_{ij}}}{\sum_i \sum_j w_{r_{ij}}} \tag{8}
$$

The test statistic of CEST for $H_0 : \beta = 0$ can be computed by

$$
U = \sum_{i=1}^{n} \sum_{j=1}^{M} (w_{d_{ij}} - \widehat{p} w_{r_{ij}}) z_{ij} \tag{9}
$$

Furthermore, [**15**] extended this model to the events that have with two progression states. Assume that the data on progression is completely observed for all the subjects ($i = 1, \ldots, n$). It is also assumed that the occurrence of progression is before or at death. One application of such model arises in cancer clinical trials, where the observed deaths during the follow-up visits are nearly always related to cancer, [**15**].

Let y_{ij} be the indicator of whether patient i is in follow-up at time j. Therefore, $y_{ij} = 1$ until the patient dies or leaves the follow-up, after which $y_{ij} = 0$. ρ_{ij} is the indicator of whether or not patient i has had the first evidence of progression at time j. It is also assumed that $\rho_{i0} = 0$, which implies that the patient i showed the first sign of progression after entering the study and at time $j = 0$, before the study

started, he was progression-free. w_{ij} is the indicator of whether patient i is at risk for progression at time j. Note that $w_{ij} = y_{ij}[1 - \sum_{k=1}^{j-1} \rho_{ik}]$. $w_{ij} = 1$ implies that Patient i was at risk for progression at time j. Therefore, at visits $k = 1, \ldots, j-1$ the patient has not progressed. Therefore, $\rho_{ik} = 0$ for $k = 1, \ldots, j-1$. Also, we have $y_{ij} = 1$ since the patient has not died until time j and is in follow up at time j.Let δ_{ij} be the indicator for whether or not patient i has died at time j. $j = 1, \ldots, M$ are the times at which a progression was assessed, a death was recorded, or a patient was censored (all times of clinic visits, deaths, or censoring). X_i is the value of the covariate (e.g., treatment indicator) for patient i.

[**15**] used a generalized person-exam risk model and then logistic link was used to relate the covariates to the events (progression, death, etc.). For generalized person-exam risk model, the observations that are over multiple times, are combined into a single sample. The logit link to model progression as a function of covariate x at time j is

$$\text{logit}P_j(x) = \log \frac{P_j(x)}{1 - P_j(x)} = \alpha_j + \beta' x \tag{10}$$

where $P_\beta(z_{ij})$ is the conditional probability of observing progression at jth follow-up time given that the individual was free of progression through time $j-1$ and x is the value of covariate for treatment.

Additionally, progression is assumed to be non-time-reversible and patients are only monitored for death after their progression has occurred. Further, it is assumed that patients who die of the disease, e.g., cancer, progressed before their death. The hazard function of death at the kth time given that progression has happened at time j, is modeled by

$$\text{logit}q_{jk}(x) = \log \frac{q_{jk}(x)}{1 - q_{jk}(x)} = \theta_{jk} + \gamma' x \tag{11}$$

where q_{jk} is the probability of observing death at time k, given that death has happened after progression at time j.

To test for treatment effect (β) on progression and mortality, it is assumed that $\beta = \gamma$. When there are short grouping intervals between exams, pooling repeated observations (PRO) is asymptotically equivalent to grouped proportional hazards model for a time-dependent covariate. Under the PRO model, the contribution of the ith subject to the log-likelihood is

$$\begin{aligned} l_i = & \sum_{j=1}^{M} [-w_{ij} \log(1 + \exp(\alpha_j + \beta' x_i)) + \rho_{ij}(\alpha_j + \beta' x_i)] \\ & \sum_{j=1}^{M} [\sum_{k=j}^{M} -y_{ik} \log(1 + \exp(\theta_{jk} + \beta' x_i)) + \delta_{ik}(\theta_{jk} + \beta' x_i)] \rho_{ij} \end{aligned} \tag{12}$$

Assuming that no one has missed the follow-up visits and the data is fully observed, the numerator of the score test for the hypothesis $H_0 : \beta = 0$ is obtained by taking the first derivative of the log-likelihood with respect to β and then letting $\beta = 0$. The test statistic is

$$U = \sum_{i=1}^{n} [\sum_{j=1}^{M} [\rho_{ij} - w_{ij} \widehat{p}_j] x_i + \sum_{j=1}^{M} (\sum_{k=j}^{M} [\delta_{ik} - y_{ik} \widehat{q}_{jk}]) \rho_{ij}] \tag{13}$$

where $w_{ij} = y_{ij}[1 - \sum_{k=0}^{j-1} \rho_{ik}]$. The estimated probability progression at time j is $\widehat{p}_j = \frac{\sum_i \rho_{ij}}{\sum_i w_{ij}}$ and the estimated probability of death at time $k \geq j$ is $\widehat{q}_{jk} = \frac{\sum_i \rho_{ij} \delta_{ik}}{\sum_i y_{ik} \rho_{ij}}$.

To derive the CEST for this model, consider the case where subject i (who has an individual risk of progression) has missed the follow-up visits after the time t_{L_i}

and comes back at time t_{R_i}. Further, assume that the subject did not have evidence of progression at the last time that they were seen, i.e., $\rho_{iL_i} = 0$. They have come back, showing that disease has progressed, $\rho_{iR_i} = 1$. In this case, ρ_{ij} is missing for $L_i + 1 \leq j \leq R_i$. In this case, the progression is censored into intervals by missed visits. Since the progression is an absorbing state, if a patient misses their follow-up visits and they are still progression-free when they return at time k, it is assume that $\rho_{ij} = 0$, for $j \leq k$.

Similar to [**16**], CEST can be applied to test for the covariate effect. The CEST method is based on a principle by Efron: "The score based on incomplete data is the expected value of the complete data score conditional on the observed data" [**12**]. The CEST for the incomplete data is obtained by taking the conditional expected value of the complete data score (13), conditioned on the observed data. Note that under $H_0 : \beta = 0$, (ρ_{ij}, w_{ij}) and (δ_{ik}, y_{ik}) are independent of x_i. Therefore, the contribution of subject i to the score function is:

$$\sum_{j=1}^{M} E[\rho_{ij} - w_{ij}\widehat{p}_j|(\, L_i, R_i]\,]x_i + \sum_{j=1}^{M} E[\sum_{k=j}^{M}(\delta_{ik} - y_{ik}\widehat{q}_{jk})x_i\rho_{ij}|(\, L_i, R_i]\,] \tag{14}$$

Under $H_0 : \beta = 0$, the expected values, conditional on observed data are obtained as:

$$\begin{aligned} &E[\rho_{ij}|\rho, y, w, \delta_{ik} = 0, \beta = 0] \\ &= \frac{p_j \prod_{r=1}^{j-1}(1-p_r) \times q_{jk} \prod_{s=j}^{k-1}(1-q_{js})}{\sum_{m=L_i+1}^{R_i} p_m \prod_{r=1}^{m-1}(1-p_r) \times q_{mk} \prod_{s=m}^{k-1}(1-q_{ms})}, j \in (\, L_i, R_i]\} \end{aligned} \tag{15}$$

and

$$E(\rho_{ij}\delta_{ik}|\text{data}) = E(\rho_{ij}|\text{data}) \times \delta_{ik} = \widehat{\rho}_{ij}\delta_{ik} \tag{16}$$

The test statistic of CEST for the data with interval-censored progression times under $H_0 : \beta = 0$ is

$$U = \sum_{i=1}^{n} \sum_{j=1}^{M} [(\widehat{\rho}_{ij} - \widehat{w}_{ij}\widehat{p}_j)x_i + \sum_{k=j}^{M} [\delta_{ik} - y_{ik}\widehat{q}_{jk}]x_i\widehat{\rho}_{ij}] \tag{17}$$

where $\widehat{p}_j = \frac{\sum_i \widehat{\rho}_{ij}}{\sum_i \widehat{w}_{ij}}$, $\widehat{w}_{ij} = y_{ij}[1 - \sum_{k<j} \widehat{\rho}_{ik}]$, and $\widehat{q}_{jk} = \frac{\sum_i \widehat{\rho}_{ij}\delta_{ik}}{\sum_i y_{ik}\widehat{\rho}_{ij}}$. This is estimated iteratively by applying EM algorithm.

3

Pooled Repeated Observations Logistic Regression Model With One Progression State and Partly Interval-Censored Data

We consider a case of longitudinal studies, where subjects are at risk of an event of interest (e.g., progression to a new state) and have follow-up visits. Some subjects make complete follow-up visits, but others miss some of their follow-up appointments and come back after the event of interest has occurred. Whenever they miss a visit, both their binary outcome of the event of the interest and covariates are missing. Our proposed model estimates the effects of time-dependent covariates on the event of interest.

Since we are interested in modeling a binary outcome, we use a logit link to model the probability of event as in [**16**]. Let T_i be the time subject i experiences the event of interest, $i = 1, \ldots, n$. Each subject i has several visits. Let t_j be the time for follow-up visits, when subjects are monitored; $j = 1, \ldots, M$. y_{ij} is the indicator of whether or not subject i has had an event since the previous visit at t_{j-1}, and x_{ij} is subject i's covariate measurement at time t_{j-1} (to be used for the jth interval).

$$\text{logit}(P_{ij}) = \log(P_{ij}/(1 - P_{ij})) = \alpha + \beta' x_{ij}, \tag{18}$$

where

$$P_{ij} = P(y_{ij} = 1 | x_{ij}, T_i > t_{j-1}). \tag{19}$$

P_{ij} is the probability of observing an event in the jth interval given a subject was event-free through t_{j-1} and x_{ij}, the covariate at time t_{j-1}.

We construct the full (complete) log-likelihood, assuming all observations were exact.

$$l = \sum_{i=1}^{n} \sum_{j=1}^{M_i} [-\log(1 + \exp(\alpha + \beta' x_{ij})) + y_{ij}(\alpha + \beta' x_{ij})] \tag{20}$$

3.1. Parameter Estimation Using the EM Algorithm

3.1.1. Notation. Assume that the i^{th} subject has missed their follow-up visits after time t_{L_i} and came back at t_{R_i}. Then $y_{iL_i} = 0$, $y_{iR_i} = 1$, and y_{ij} is missing for $L_i + 1 \leq j \leq R_i - 1$. L_i is the index of the last time that subject i made the visit and was event-free. R_i is the index of the first time subject i was observed with the event of interest. M_i is the index of the last time subject i was observed. Whenever subjects miss visits, their corresponding covariate value, x_{ij}, is also missing. We use the EM algorithm [**12**] to estimate the parameters.

3.1.2. EM Algorithm. E-step: We need to estimate y_{ij} and x_{ij} in the expression (20) for $j \in \{L_i + 1, \ldots, R_i - 1\}$ for individuals whose failure times are interval-censored.

x_{ij} could be continuous or categorical ([**32**]). We assume that x_{ij} has a linear growth curve with fixed effects for the sake of simplicity whereas [**46**] and [**16**] assumed random effects.

$$x_{ij} = \theta_{0i} + \theta_{1i} t_{j-1} + \epsilon_{ij}, \tag{21}$$

where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $cov(\epsilon_{ij}, \epsilon_{ij'}) = 0, j \neq j'$.

We estimate x_{ij} by $\widehat{x}_{ij} = \widehat{\theta}_{0i} + \widehat{\theta}_{1i} t_{j-1}$ for $L_i + 1 \leq j \leq R_i - 1$, where $\widehat{\theta}_{0i}$ and $\widehat{\theta}_{1i}$ are least squares estimators.

$$
\begin{aligned}
&E[y_{ij}|Y, \widehat{x}_{ij}, \alpha, \beta, T_i > t_{i,j-1}] \\
&= P[T_i = t_j | t_{L_i} < T_i \leq t_{R_i}, Y, \widehat{x}_{ij}, \alpha, \beta] = \\
&\begin{cases}
\frac{\widehat{p}_{ij}}{\widehat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1} [\widehat{p}_{ik} \prod_{o=L_i+1}^{k-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\
\text{if } j = L_i + 1, \\
\frac{\widehat{p}_{ij} \prod_{o=L_i+1}^{j-1} (1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1} [\widehat{p}_{ik} \prod_{o=L_i+1}^{k-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\
\text{if } j \in \{L_i + 2, ..., R_i - 1\}, \\
\frac{\prod_{o=L_i+1}^{j-1} (1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1} [\widehat{p}_{ik} \prod_{o=L_i+1}^{k-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\
\text{if } j = R_i,
\end{cases}
\end{aligned}
\tag{22}
$$

where

$$
\widehat{p}_{ij} = \frac{\exp(\alpha + \beta' \widehat{x}_{ij})}{1 + \exp(\alpha + \beta' \widehat{x}_{ij})} \tag{23}
$$

If x_{ij} is ordinal, we can still assume linear growth curve with fixed effects to estimate the missing x_{ij}'s. Let nc be the number of categories for this ordinal variable. For each individual i, the observed x_{ij}'s are used in model 21) to compute $\widehat{\theta}_{0i}$ and $\widehat{\theta}_{1i}$. Then we compute $\widehat{x}_{ij} = \widehat{\theta}_{0i} + \widehat{\theta}_{1i} t_{j-1}$ as usual.

Next, we create $nc-1$ thresholds in order to uniquely assign $\widehat{x}_{ij}$ into one of the nc categories. Note that $\widehat{x}_{ij} \sim N(\theta_{0i} + \theta_{1i} t_{j-1}, \sigma_{\widehat{x}}^2)$. We use the quantiles of this normal distribution to define the thresholds. There is limitation in using the linear growth curve for ordinal data. We need to compute $\widehat{\sigma}_{\widehat{x}}^2$ to define thresholds. This requires at least three observed covariate values, x_{ij}'s for each subject as otherwise, $\widehat{\sigma}_{\widehat{x}}^2$ would be

undefined due to the zero degrees of freedom.

M-step: We find the values of α and β that maximize the expected value of log-likelihood in conditioned on the missing data. Therefore, we have

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\max l_{\alpha,\beta} | \widehat{y}_{ij}, \widehat{x}_{ij}. \tag{24}$$

Expressions (22) - (24) are repeated until convergence. As there are no closed forms for $\widehat{\alpha}$ and $\widehat{\beta}$, we used an optimization package `optim` in R.

3.2. Variance Estimation Using Louis' Method

Louis' method is one of the approaches to obtain the variance of parameters that are obtained by the EM algorithm. It uses the missing information principle [**36**] and [**30**]. We use the Louis' method for variance estimation using the notation in [**39**]. This method uses the missing information principle,
Observed Information = Complete Information - Missing Information.

$$\begin{aligned} \frac{-\partial^2 \log P(\theta|W)}{\partial \theta^2} = & - \int_z \frac{\partial^2 \log P(\theta|W,V)}{\partial \theta^2} P(V|\theta, W) dZ \\ & - Var(\frac{-\partial \log P(\theta|W,V)}{\partial \theta}), \end{aligned} \tag{25}$$

where W: observed data, i.e., partly interval-censored pooled repeated observations, V: latent data, the true unknown counterpart of the interval-censored portion of W, $\theta|W$: observed posterior, and $\theta|W,V$: augmented posterior.

3.3. Power

We compute the power of the test $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$ using the asymptotic normality of maximum likelihood estimators. Note that for some fairly strong conditions, $\theta^{(t)} \to \theta_{MLE}$ as $t \to \infty$, where $\theta^{(t)}$ is the parameter estimate of an EM algorithm at t^{th} step.

Using the above theorem, we have:

$$(\widehat{\beta} - \beta)'[I(\widehat{\beta})]^{-1}(\widehat{\beta} - \beta) \to \chi^2_{length(\widehat{\beta})}$$

where $\widehat{\beta}$ is the maximum likelihood estimator of β.

We compute the power by simulating B copies of this test and power is calculated by R/B where R is the number of times where H_0 was rejected in the B tests [**34**].

3.4. Simulation Study

For illustrating this method, we considered $n = 300$ subjects who have $M = 7$ number of follow-up visits each. We generated two time-dependent covariates as follows:

$$x1_{ij} \sim N(5.8 + 0.3t_{j-1}, 0.1).$$

$$x2_{ij} \sim \ N(0.4 + 0.15t_{j-1}, 0.1).$$

$x1_{ij}$ represents a continuous covariate with larger values and faster growth rate over time, while $x2_{ij}$ represents one with smaller values and slower growth rate over time.

We first generate $n = 300$ subjects who have complete follow-up visits. This makes original complete data (OC). We randomly choose n_1 subjects out of these. This makes exact data (E). For the remaining $n_2 = n - n_1$ subjects, we randomly designate some of their follow-up visits missing. This makes interval-censored data.

The observed data (O), also known as partly interval-censored data, is the combination of exact and interval-censored data. We considered several values for n_1 and n_2 to cover different proportions of exact data.

We randomly sampled L_i and R_i for each patient. Note that for exact data we have $R_i = L_i + 1$ and for interval-censored data $R_i \geq L_i + 2$. Then for $j = 1, \ldots, L_i$ we have $y_{ij} = 0$ and for $j = R_i, \ldots, M$, $y_{ij} = 1$. y_{ij} is missing for $j = L_i + 1, \ldots, R_i - 1$ in the interval-censored data. y_{ij} was 1 when the i^{th} subject was in the follow-up visit and was at risk at the j^{th} visit and it was 0 otherwise. The Figure 3.2 demonstrates the partly interval-censored data, using y_{ij}.

We computed the bias and variance, using the Louis's method in 3.2 for original complete data, exact data, and observed data based on $B = 1500$ replications. In addition, we investigated the power of our test using 3.3.

We first considered the case where there was only one attribute ($x1_{ij}$) in the model. The EM algorithm (Section 3.1) was used for the parameter estimation. The variance of the parameter estimator was calculated using Louis' method (Section 3.2).

The results are shown in Table 3.1. For all the different combinations of n_1 and n_2, the proposed estimator based on the observed data produce a smaller bias and a smaller variance than that based on the exact data alone. In particular, for the case of (250, 50), that is the one that contains 16% interval-censored data; the proposed estimator produces a smaller bias and a smaller variance than that based on the exact data of size 250 alone. We also notice that the more exact data we have, smaller the bias and variance. These results have a quite similar pattern to those in [**26**], who employed a proportional hazards model with partly interval-censored data. D'Agostino et al. (1990) [**9**] notes that pooled repeated observations logistic regression is close to the time-dependent covariate Cox regression analysis. Therefore, this simulation result coincides with what we expected.

y_{ij}	visit times (j)						
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
	[1,] 0	0	0	0	1	1	1
	[2,] 0	0	0	1	1	1	1
	[3,] 0	0	0	0	0	NA	1
	[4,] 0	0	1	1	1	1	1
	[5,] 0	0	0	NA	NA	NA	1
	[6,] 0	0	NA	NA	NA	NA	1
patient (i)	[7,] 0	0	0	NA	NA	NA	1
	[8,] 0	0	0	NA	1	1	1
	[9,] 0	0	0	0	NA	NA	1
	[10,] 0	0	NA	NA	NA	1	1
	[11,] 0	0	0	0	1	1	1
	[12,] 0	0	NA	1	1	1	1
	[13,] 0	0	NA	NA	1	1	1
	[14,] 0	0	NA	NA	NA	NA	1
	[15,] 0	0	NA	NA	1	1	1

FIGURE 3.1. Partly interval-censored data.

(n_1, n_2)	B_E	B_O	B_{OC}	σ^2_E	σ^2_O	σ^2_{OC}
(250,50)	0.559	0.241	0.021	0.043	0.028	0.017
(200,100)	0.624	0.326	0.023	0.056	0.031	0.022
(150,150)	0.769	0.457	0.025	0.059	0.034	0.022
(100,200)	0.812	0.608	0.022	0.065	0.038	0.023
(50,250)	0.838	0.809	0.023	0.078	0.044	0.026

TABLE 3.1. Results for 1-dimensional β, $\beta^{true} = 3.6$, B:Bias, σ^2:variance, E:exact data, O:observed data, and OC: original complete data.

Next, we computed the power of the test $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$. We considered both one-dimensional covariate and two-dimensional covariates. we considered 3 sample sizes (100, 200, and 300) and for each of these sample sizes we ran $B = 1000$ replications of the test. The power was calculated as the proportion of times H_0 was rejected at 5% level of significance. Both Figure 3.3 and Figure 3.4 show the powers for different values of β_0 and different sample sizes. The power curves are symmetric for all the different sample sizes. As a sample size increases or the parameter values are farther apart from the true parameter value (i.e., an effect size increases), the corresponding power increases. From Figure 3.3, with a sample of size $n = 300$, one can achieve 80% power for the effect size of 0.45. Moreover, for the effect size of 0.55, a sample of size $n = 200$ is enough to achieve 80% power. [**34**] achieved about 80% power in detecting the effect size of 0.75 for the proportional hazards model with a sample of size 300 current status data. Considering that pooled repeated observations partly interval-censored data has more information than current status data, our better power result is correct.

In summary, even a small amount of interval-censored data portion of an original partly interval-censored data set does help our statistical inference to be more accurate and more powerful.

3.5. Right and Left Censoring

In some special cases, the visiting time of some subjects in the data may have right and/or left censoring. If a subject has not progressed at the last interview ($y_{iL_i} = 0$) and does not come back for the proceeding interview visits, then the subject's time to the event of interest is right-censored. In this case $L_i = M_i$ and $R_i = \infty$. We can impute the covariates x_{ij} using the linear growth curve in Section 3.1 and estimate y_{ij} for the missed visits, $j = L_i + 1, \ldots, M$. The E-step to estimate y_{ij} is given by

$$
\begin{aligned}
& E[y_{ij}|Y, \widehat{x}_{ij}, \alpha, \beta, T_i > t_{i,j-1}] \\
& = P[T_i = t_j | t_{L_i} < T_i \leq t_{R_i}, Y, \widehat{x}_{ij}, \alpha, \beta] \\
& = \frac{\widehat{p}_{ij} \prod_{o=L_i+1}^{j-1} (1 - \widehat{p}_{io})}{\widehat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{M} [\widehat{p}_{ik} \prod_{o=L_i+1}^{k-1} (1 - \widehat{p}_{io})]},
\end{aligned} \tag{26}
$$

where

$$
\widehat{p}_{ij} = \frac{\exp(\alpha + \beta' \widehat{x}_{ij})}{1 + \exp(\alpha + \beta' \widehat{x}_{ij})}.
$$

However, extrapolating the covariates x_{ij} for $j > L_i$ using the linear growth curve in Section 3.1 may well increase bias and variance.

If a subject's first visit is at time k and the subject shows the symptoms of event of interest, then y_{ij} is missing for $j = 1, \ldots, k-1$, and $y_{ik} = 1$ (left censoring). In this case, the covariate, x_{ij}, and response, y_{ij}, should be estimated for $j \leq k-1$ at E-step. We merely have $L_i = 0$, $R_i = k$, and one observed covariate value x_{ij}. Therefore, we cannot fit the subject-dependent growth curve to estimate the covariates at the missed visits.

In summary, there is no merit to include individuals whose event-times are either left-censored, or right -censored when fitting a logistic regression model with pooled repeated observations.

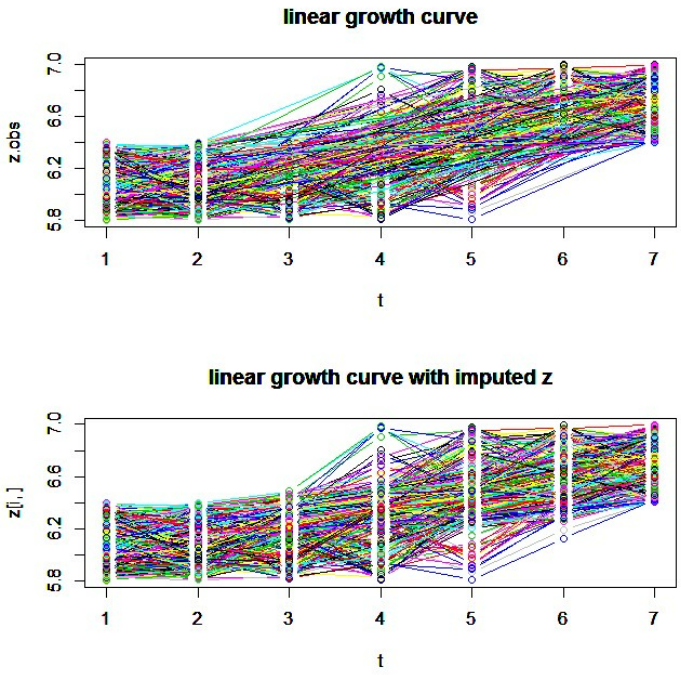

FIGURE 3.2. Linear growth curve.

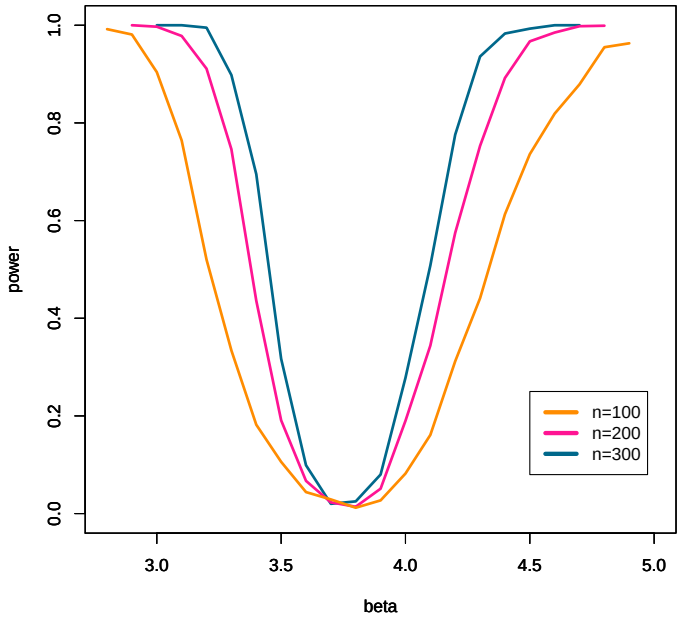

FIGURE 3.3. Power of the test for one-dimensional β.

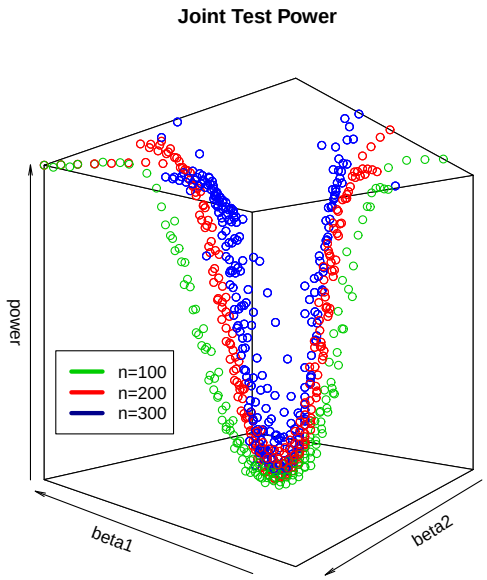

FIGURE 3.4. Power of the test for multidimensional β.

4

Variable Selection for Pooled Repeated Observations Logistic Regression Model With Partly Interval-Censored Data

4.1. Introduction to Variable Selection

When we work with a large data set, there are many variables that can be used for prediction. However, not all these variables predict the response variable efficiently. Additionally, having too many predictors decreases the accuracy of our predictions. Our goal is to choose the best subset of the data to use for prediction and making inferences. This is done through variable selection techniques. They help us remove the redundant covariates from the data.

4.2. Model

Consider the problem where there are many covariates in the model and we are interested in selecting a subset of them (variable selection) and estimating the covariate effects based on this subset model. Selecting meaningful variables in multivariate models, increases prediction accuracy. There are various techniques for variable selection. However, classical variable selection methods, such as the forward selection and the backward elimination methods are time-consuming, unstable, and sometimes unreliable for making inferences, [**37**]. One of the well-known techniques for variable selection is the least absolute shrinkage and selection operator (LASSO). It is one of the well-known shrinkage methods. LASSO was introduced by Tibshirani [**40**]. The LASSO estimator for the linear model $Y = X\beta + \varepsilon$ is defined by

(27)
$$\hat{\beta}_{LASSO} = \arg\min_{\beta}(\beta - \hat{\beta})' X' X (\beta - \hat{\beta})$$
$$\text{subject to} \|\beta\| \leq \lambda$$
$$(L^1 \text{norm penalization})$$

.

$\|.\|$ is the Euclidean norm (not squared). λ is called the tuning parameter and it controls the amount of shrinkage that is applied to the β estimates. The main feature of the LASSO method is that it sets some of the coefficients to 0 and shrinks the others. This method simultaneously selects variables and estimates parameters of the model, [**37**]. Since the LASSO method has good computational properties, its solution path is predictable. If we are analyzing the complete data (data without any censoring), the LASSO estimate of the vector of parameters, β is the solution to

(28)
$$\hat{\beta}_{LASSO} = \arg\max_{\beta} (l(\beta) + \lambda\|\beta\|).$$

However, in the case of partly interval-censored data with pooled repeated observations, some of the response variables, y_{ij} are missing. In this case, we need to use an imputation method, e.g., the EM algorithm to iteratively, impute the missing observations, y_{ij} for the censored times (E-step), and then use the imputed values to update the expected value of the likelihood, conditioned on the missing data (M-step).

The E-step, when there is only one progression, will be computed similar to Equation (22) in the previous chapter.

$$E[y_{ij}|Y,\widehat{x}_{ij},\alpha,\beta,T_i > t_{i,j-1}]$$

$$= P[T_i = t_j | t_{L_i} < T_i \leq t_{R_i}, Y, \widehat{x}_{ij}, \alpha, \beta] =$$

$$(29) \qquad \begin{cases} \frac{\widehat{p}_{ij}}{\widehat{p}_{i,L_i+1}+\sum_{k=L_i+2}^{R_i-1}[\widehat{p}_{ik}\prod_{o=L_i+1}^{k-1}(1-\widehat{p}_{io})]+\prod_{o=L_i+1}^{R_i-1}(1-\widehat{p}_{io})} \\ \text{if } j = L_i+1, \\ \frac{\widehat{p}_{ij}\prod_{o=L_i+1}^{j-1}(1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1}+\sum_{k=L_i+2}^{R_i-1}[\widehat{p}_{ik}\prod_{o=L_i+1}^{k-1}(1-\widehat{p}_{io})]+\prod_{o=L_i+1}^{R_i-1}(1-\widehat{p}_{io})} \\ \text{if } j \in \{L_i+2,...,R_i-1\}, \\ \frac{\prod_{o=L_i+1}^{j-1}(1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1}+\sum_{k=L_i+2}^{R_i-1}[\widehat{p}_{ik}\prod_{o=L_i+1}^{k-1}(1-\widehat{p}_{io})]+\prod_{o=L_i+1}^{R_i-1}(1-\widehat{p}_{io})} \\ \text{if } j = R_i, \end{cases}$$

where

$$(30) \qquad \widehat{p}_{ij} = \frac{\exp(\alpha+\beta'\widehat{x}_{ij})}{1+\exp(\alpha+\beta'\widehat{x}_{ij})}$$

However, the optimization on M-step will be updated. We will add the the LASSO penalty term (L^1 norm) to shrink the parameters β. Note that the intercept, α is not included in the penalty term, [**33**].

$$(31) \qquad (\widehat{\alpha},\widehat{\beta}) = \arg\max\left(l_{\alpha,\beta}+\lambda\|\beta\|\right)|\widehat{y}_{ij},\widehat{x}_{ij}.$$

If there are groups of highly correlated variables in the data, the LASSO method tends to choose only one variable from each group. When categorical predictors (factors) are present in the data, the LASSO solution only selects individual dummy variables instead of whole factors. The group LASSO is an extension of the LASSO penalty and can overcomes this problem. The group LASSO is used on pre-defined

groups of variables. Furthermore, the group LASSO estimator for the logistic regression is shown to be statistically consistent, even if the number of predictors is much larger than sample size, [**33**]. The group LASSO estimator for the complete data is:

$$\hat{\beta}_{LASSO} = \arg\max_{\beta} \left(l(\beta) + \lambda \sum_{l=1}^{L} \|\beta_l\|\right). \tag{32}$$

Instead of setting individual β's to zero, this method sets a group of coefficients, β_l to zero. However, this method selects a larger number of groups than it is necessary, which causes some noisy variables to be included in the model, [**37**]. Additionally, the groups of coefficients are pre-defined in this method. Therefore, unless there are reasonable ways (prior information, previous literature, etc.) of grouping the variables, the usual LASSO can be used for variable selection.

4.2.1. Choice of λ. There are several criteria for choosing the penalty parameter, λ, including the prediction error and cross-validation. We would use the criterion that was suggested by [**40**]. The prediction error (PE) is estimated over a grid of values of λ from 0 to 1. λ is the value that yields the lowest PE, where

$$PE = E_{(Y,X)}[\sum_{i=1}^{n} \sum_{j=1}^{M_i} (Y_{ij} - \hat{Y}_{ij})^2] \tag{33}$$

The tuning parameter, λ controls the amount of penalization. However, as λ gets larger, more coefficients, β are set to zero.

4.3. EM Algorithm for Variable Selection Via LASSO and Group LASSO

We will introduce an EM based algorithm that can be used for pooled repeated partly interval-censored data. This algorithm finds the LASSO estimator, β_{LASSO} and the tuning parameter, λ, such that the prediction error is minimized.

EM algorithm for LASSO variable selection in pooled repeated partly interval-censored data

Step 1 Use the exact data (data without any censoring) to draw the LASSO trace. Use the plot to get a plausible range for λ.

Step 2 Define the initial value for the parameters, $\beta^{(0)}$.

Step 3 Run the EM algorithm for all the values in the range of λ from Step 1

Step 3.1(E Step) Estimate $\widehat{y}_{ij}$ for $L_i < j \leq R_i$ using (22).

Step 3.2(M Step) Find $\hat{\beta}^{(r)}_{LASSO}$ using (32).

Repeat steps (3.1)-(3.2) until $\hat{\beta}^{(r)}_{LASSO}$ converges.

Step 4 Find the prediction error, $PE^{(r)}$ using $\hat{\beta}^{(r)}_{LASSO}$.

$\hat{\beta}_{LASSO}$ is found by

$\hat{\beta}_{LASSO} = min_{\hat{\beta}^{(r)}_{LASSO}} PE^{(r)}$

Next, we would like to introduce an algorithm for variable selection using the group LASSO. If there are some reasonable ways for grouping the covariates, one can use this method. Otherwise, we use the usual method. [**33**] has introduced an algorithms for the logistic group LASSO based on block co-ordinate descent minimization. We have adopted [**33**]'s algorithm to modify the maximization step of the EM algorithm in (3.1). This will provide an EM based variable selection algorithm via group LASSO.

Let G be the number of groups for group LASSO. Note that each group, g can contain one or more covariates and these variables are pre-specified. Additionally, we do not penalize the intercept. Let β_{-g} be the parameter vector β when setting β_g to 0. For example, let $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T$ and group g be $\beta_g = (\beta_1, \beta_3)^T$. Then, $\beta_{-g} = (0, \beta_2, 0, \beta_4, \beta_5)^T$. df_g is the degrees of freedom for gth predictor. For a continuous variable, $df = 1$ and for a categorical variable with c levels, we have

$df = c - 1$. Let $s(df_g) = df_g^{1/2}$. Also, $||\overrightarrow{u}||_2^2 = \sum_{i=1}^{n} u_i^2$ and $\overrightarrow{u} \subseteq \mathbb{R}^n$.

EM algorithm for group LASSO variable selection in pooled repeated partly interval-censored data

Step 1 Define the initial value for the parameters, $\beta^{(0)}$.

Step 2 Run the EM algorithm.

Step 2.1(E Step) Estimate $\widehat{y}_{ij}$ for $L_i < j \leq R_i$ using (22).

Step 2.2(M Step) $\widehat{\beta} = \arg\max_{\beta}[l(\beta) - \lambda \sum_{g=1}^{G} s(df_g)||\beta_g||_2]$.

Loop over groups, for $g = 1, \ldots, G$.

Step 2.2.1 Find $\widehat{p}_{ij(\beta_{-g})} = \frac{\exp(\alpha + \widehat{\beta}_{-g} x_{ij})}{1+\exp(\alpha + \widehat{\beta}_{-g} x_{ij})}$.

Step 2.2.2 If $||X_g^T(y_{ij} - \widehat{p}_{ij(\beta_{-g})})||_2 \leq \lambda s(df_g)$ then $\beta_g \leftarrow 0$

else $\beta_g \leftarrow \arg\max_{\beta_g}[l(\beta) - \lambda \sum_{g=1}^{G} s(df_g)||\beta_g||_2]$.

Repeat step 2 until $\widehat{\beta}$ converges.

4.4. Simulation

For illustrating this method, we considered $n = 300$ subjects who have $M = 7$ follow-up visits for each subject. The partly interval-censored data is simulated similar to Section 3.4. For variable selection, we generated five time-dependent covariates as follows:

$$x1_{ij} \sim N(5.8 + 0.3t_{j-1}, 0.1).$$

$$x2_{ij} \sim N(0.4 + 0.15t_{j-1}, 0.1).$$

$$x3_{ij} \sim \text{Binomial}(\text{size} = 5, \text{prob} = 1/2).$$

$$x4_{ij} \sim \text{Gamma}(\text{shape} = 3, \text{rate} = 4).$$

$$x5_{ij} \sim \text{Uniform}(0, 1).$$

$x3_{ij}$ represents an original covariate, while the rest of simulated time-dependent covariates are continuous. Note that in NLSY97, we had an ordinal covariate that represented an individual's self-evaluation of general state of health and it had 5 levels.

We will use the EM algorithm in Section 4.3 for variable selection via LASSO. Additionally, we need an estimation of the tuning parameter, λ. This is done by minimizing the prediction error. In linear models, you can write $\widehat{y_{ij}}$ as a function of λ. However, in our model, $\widehat{\beta}$ does not have a closed form. Therefore, we cannot write the prediction error in closed form and need to minimize it numerically. Finding a proper estimation of λ is essential, since overestimating it will cause more coefficients to be zero than necessary, [**45**]. The graph of the LASSO trace illustrates the convergence path for various values of λ, [**27**] and [**1**].

Figure 4.1 shows the path for LASSO trace and prediction error for different values of λ using the simulated data. The simulated data supports the importance of not shrinking too many covariates for model selection via LASSO.

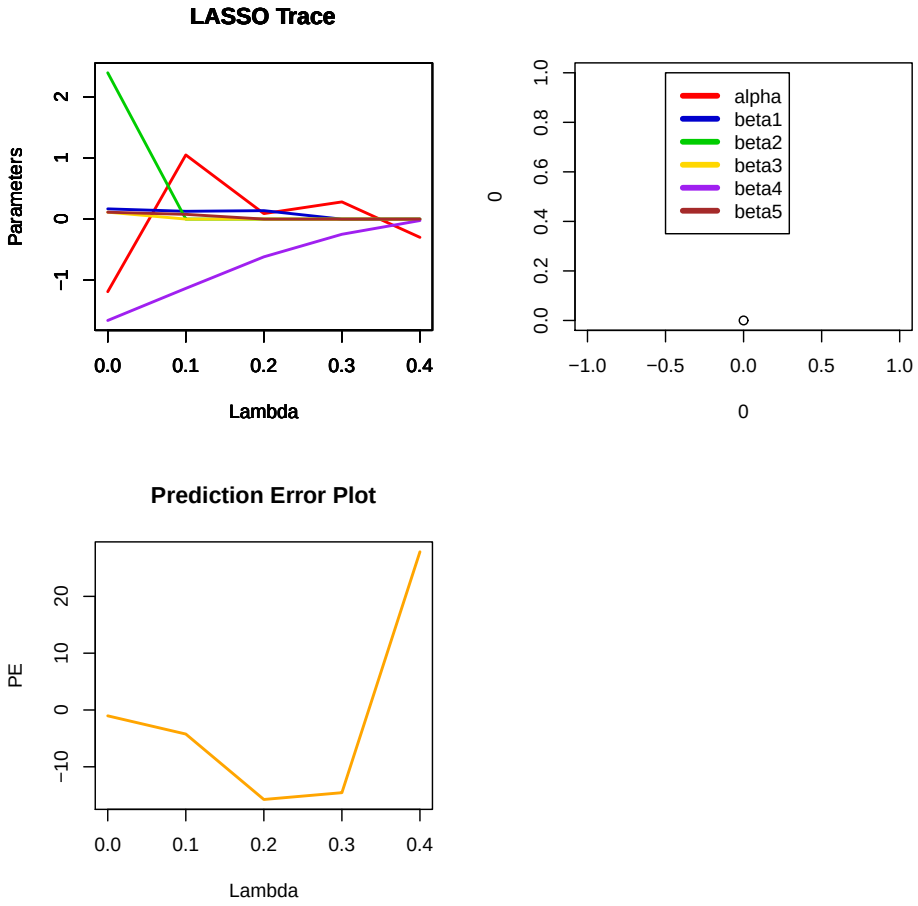

FIGURE 4.1. The path for LASSO trace and prediction error.

5

Pooled Repeated Observations Ridge Logistic Regression Model With Partly Interval-Censored Data

5.1. Introduction

Often, when one tries to build a regression model with many covariates, they may find out that there exists a relationship between some of the predicting variables. This situation is referred to as "multicollinearity" [**35**]. In this case, the ordinary least squares (OLS) estimators of the regression parameters may not work properly. Consequences of using usual estimators when we have multicollinearity are 1) Estimators may not be very accurate and can be far away from true estimators, and 2) Large variances for the estimators [**31**]. (Since we need to compute the inverse of $X'X$ matrix where X is the design matrix.) [**44**] has studied the effects of multicollinearity considering various collinearity patterns.

There are several methods for parameter estimation in case of multicollinearity. However, a method that has been used extensively, is called "Ridge Regression". Ridge regression was introduced by Hoerl and Kennard [**21**] and it is the most commonly used method to combat multicollinearity. It adds a L^2 norm penalization on parameters. The ridge estimator, $\hat{\beta}_{Ridge}$ for the ill-conditioned linear model $Y = X\beta + \varepsilon$ is the solution to

$$\hat{\beta}_{Ridge} = \arg\min_{\beta}(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) \tag{34}$$

$$\text{subject to} \|\beta\|^2 \leq \lambda$$

(L^2norm penalization)

.

The ridge estimator is part of the shrinkage estimators, since it shrinks the least-squares estimator toward the origin. The resulting estimates from applying ridge regression, $\hat{\beta}_{Ridge}$ will reduce the prediction variance, even though they are biased.

Some alternatives to ridge regression have been introduced in literature. Jensen and Ramirez [**24**] proposed surrogate regression in 2010. In this method the design matrix, X is modified to $X_\lambda = P_1(Diag(\zeta_i^2 + \lambda I)^{\frac{1}{2}})Q'$, where P_1,Q, and ζ_i are obtained by applying the singular value decomposition (SVD) on the $X'X$ matrix. The surrogate estimator, $\widehat{\beta}_{surrogate}$ is the solution to the normal equation $X'_\lambda X_\lambda \beta = X'_\lambda Y$. We conducted simulations [**10**] to compare the performance of the ridge and surrogate estimators. The results show that surrogate regression model has smaller AIC and SSE than those of ridge regression model, although the difference is negligible. Ridge and Surrogate estimators use a constant, λ as the tuning parameter, to modify the ordinary least square (OLS) estimators to get a better estimate. We would like a smaller value for λ since a larger λ would lead to a larger bias and also a larger value of mean squared error (MSE), $MSE = Variance + Bias^2$. Although [**24**] showed that $SSE(\beta_{ridge}) < SSE(\beta_{surrogate})$ and surrogate estimator has a smaller variance inflation factor (VIF) and condition number, our simulations revealed the SSEs do not differ significantly for small values of λ.

Double penalized estimators were introduced by Shen and Gao [**38**] to simultaneously combat separation and multicollinearitry in multiple logistic regression. The double penalized likelihood function adds a second penalty term to Firth's penalized likelihood function [**17**] by including a ridge parameter which forces the parameters to spherical restrictions, $\beta_{DP} = \arg\max[l(\beta) + \frac{1}{2}\log \mid A \mid -\lambda\|\beta\|^2]$, where A is the Fisher information matrix for the log-likelihood function, $l(\beta)$. However, the simulation that [**38**] carried out, did not induce collinearity and it only induced separation in the data. [**20**] carried out simulations to investigate the performance of double-penalized estimator in presence of multicollinearity. Although the differences were small, the iterative ridge estimator had a better behavior than the Shen and Gao's double-penalized estimator, when the problems of collinearity appear in the data.

5.2. Model

Now we consider the case when there are several covariates, x_{ij} in the model and there is high degree of correlation among these predictor variables, i.e., some of these covariates are correlated. This situation is called "Multicollinearity" in the literature. Using the regular method for parameter estimation in this case leads to estimates that are far from true values and have large variance.

However, since there are large numbers of informative covariates in model, which have high degrees of correlation, we would like to apply the ridge estimation technique by adding the L^2 norm penalty to the model in Section 3.1. If we have complete data (no censoring), the ridge estimator is the solution to the equation

$$\hat{\beta}_{Ridge} = \arg\max_{\beta} (l(\beta) + \lambda\|\beta\|^2). \tag{35}$$

where $l(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{M_i} [-\log(1+\exp(\alpha+\beta' x_{ij})) + y_{ij}(\alpha+\beta' x_{ij})]$, is the complete log-likelihood, $\|.\|$ is the Euclidean norm of the parameter vector, and λ is the tuning parameter.

In the above model, we can expect to have partly interval-censored pooled repeated observations (PRO) regression model where some subjects' event of interest (progression to a new state) and time-varying covariate values were missing within a time interval due to that fact that these subjects missed their follow-up visits. In this case, we use the expectation-maximization (EM) algorithm to iteratively, impute the missing observations, y_{ij} for the censored times (E-step), and then use the imputed values to update the expected value of the penalized likelihood, conditioned on the missing data (M-step).

The E-step, when there is only one progression, will be computed similar to the previous sections.

$$
\begin{aligned}
&E[y_{ij}|Y, \widehat{x}_{ij}, \alpha, \beta, T_i > t_{i,j-1}] \\
&= P[T_i = t_j | t_{L_i} < T_i \leq t_{R_i}, Y, \widehat{x}_{ij}, \alpha, \beta] = \\
&\begin{cases}
\frac{\widehat{p}_{ij}}{\widehat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1} [\widehat{p}_{ik} \prod_{o=L_i+1}^{k-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\
\text{if } j = L_i + 1, \\
\frac{\widehat{p}_{ij} \prod_{o=L_i+1}^{j-1} (1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1} [\widehat{p}_{ik} \prod_{o=L_i+1}^{k-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\
\text{if } j \in \{L_i + 2, ..., R_i - 1\}, \\
\frac{\prod_{o=L_i+1}^{j-1} (1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1} [\widehat{p}_{ik} \prod_{o=L_i+1}^{k-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\
\text{if } j = R_i,
\end{cases}
\end{aligned}
\tag{36}
$$

where

$$\widehat{p}_{ij} = \frac{\exp(\alpha + \beta'\widehat{x}_{ij})}{1 + \exp(\alpha + \beta'\widehat{x}_{ij})} \tag{37}$$

However, the optimization in M-step will be updated. We will add the the ridge penalty term (L^2 norm) to shrink the parameters β. Note that the intercept, α is not included in the penalty term.

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\max\left(l_{\alpha,\beta} + \lambda\|\beta\|^2\right)|\widehat{y}_{ij}, \widehat{x}_{ij}. \tag{38}$$

The expected value in the above equation does not have a closed form, therefore, we can use the law of large numbers and Monte Carlo techniques to approximate it. Furthermore, the above equation does not have a closed form for β. However, we can use the optimization packages that are available for **R** software, to solve for the parameter vector β. To increase the convergence rate of the optimization, we can provide the Jacobian, and Hessian of the equation.

5.2.1. Choice of λ. The value of the tuning parameter, λ is unknown and it needs to be estimated in order to perform the analysis. [**22**]'s criterion was to choose the value of λ that minimizes the mean squared error of the ridge estimator. Cross validation is another criterion that can be used, however it is very time-consuming, [**6**]. [**14**] also suggested to iteratively update the value of λ at each iteration by minimizing the mean squared error (MSE) in the M-step of the EM algorithm. When the response variable is continuous (linear models), the estimated parameters have a closed form and can be written as a function of λ. In this case, MSE can be written as a function of λ. However, in the case of binary responses (generalized linear models), the parameter estimates may not have a closed form and minimizing MSE becomes very inefficient.

Besides the traditional methods that are mentioned for choosing the tuning parameter, λ in ridge regression, another criterion is inspection of the ridge trace plot. In this method the estimated parameters, $\hat{\beta}_{Ridge}$ are plotted for various values of λ. In this method we inspect the ridge trace to find out for what value of λ the ridge trace starts stabilizing.

5.3. Simulation Study

We conducted simulations to study the proposed model. We considered $n = 300$ subjects who have $M = 7$ follow-up visits for each subject. The partly interval-censored data is simulated similar to Section 3.4. For illustrating covariates in the presence of multicollinearity, we generated five time-dependent covariates as follows:

$$x1_{ij} \sim N(5.8 + 0.3t_{j-1}, 0.1).$$

$$x2_{ij} \sim N(0.4 + 0.15t_{j-1}, 0.1).$$

$$x3_{ij} = x1_{ij} + 2 * x2_{ij} + \text{random noise}$$

$$x4_{ij} \sim \text{Gamma}(\text{shape} = 3, \text{rate} = 4).$$

Since some of the covariates are highly correlated, the estimation method in Section 3.1 does not work. The optimization in M-step cannot be performed in the absence of the penalty term and hence, the algorithm does not converge. Therefore, the starting value of λ for the ridge trace should be greater than zero, for example $\lambda^{initial} = 0.001$. We plot the ridge trace for values of $\lambda > 0$. Ridge method uses λ, to modify the M-step in Section 3.1 to get parameter estimates in presence of multicollinearity. Note that we would like a smaller value for λ, since a larger λ would lead to a larger bias and also a larger MSE ($MSE = Variance + Bias^2$).

Figure 5.1 shows the ridge trace versus different values of λ for the simulated data. It suggests that using a value of $\lambda \approx 0.03$ is appropriate. Table 5.1 shows the parameter estimates for this tuning parameter.

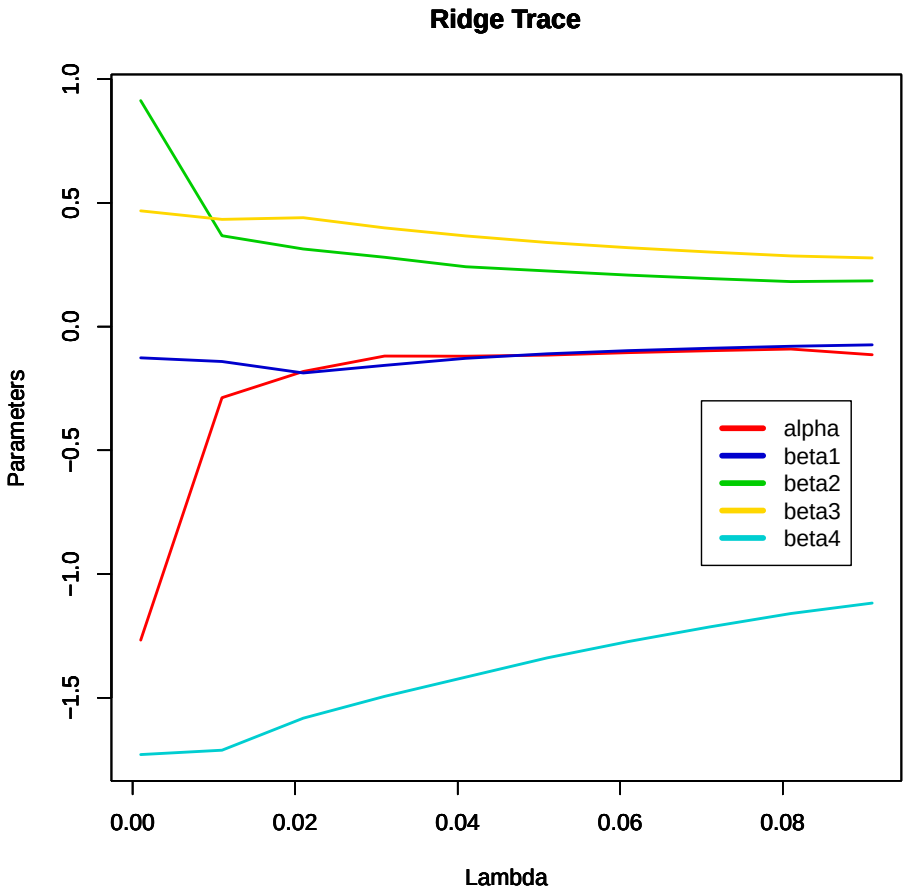

FIGURE 5.1. The ridge trace.

Parameter	α	β_1	β_2	β_3	β_4
Estimate	-0.18	-0.19	0.31	0.44	-1.55

TABLE 5.1. Simulation Results for Pooled Repeated Observations Ridge Logistic Regression

6

Pooled Repeated Observations Logistic Regression Model With Partly Interval-Censored Data for Two Progression States

There are situations where the event of interest have several states that are chronological. Consider the model in Section 3.1. That model accounts for the case, when we are attempting to predict one progression state from time-varying covariates. We would like to extend that model to the situation when there are at least two progression stages. First, consider the model when there are two states. These states are chronological, which means that the 1^{st} progression occurs before the 2^{nd} progression. The two events of interest in this model are the first and second progressions. Some subjects complete all of their follow-up visits and their progression times are determined accurately and we have completely observed the data for these subjects.However, since the study is happening over time, others miss some of their follow-up visits and when they come back, they learn that either one of their events of interest (or both) has already occurred. The events for the latter subjects is censored within the person-specific time interval. This is known as partly interval-censored failure time data.

We define pooled repeated observations partly interval-censored data for two progression states. We have pooled repeated observations but some binary outcomes are incomplete. They can only be determined with certain unknown probabilities within some specific follow-up visits. To analyze this model, we need to combine some models that handle pooled repeated observations without censoring and methods that deal with partly or completely interval-censored data.

The main goal of this study is to estimate the effects of the time-dependent covariates on the occurrence of the events of interest (e.g., different states of progression to a disease, e.g., different stages of cancer, different stages from being a non-smoker to becoming a heavy smoker, etc.). We extend the model in Chapter (3) to the problem of estimating the covariate effects, when the events of interest consist of two progression states and the responses are pooled, repeated and partly interval-censored.

[**7**] proposed a model to estimate the regression coefficients of a multi-state model where the transition direction is irreversible and [**15**] proposed a CEST to test for the treatment effect of a two-state model with binary responses. The treatment effect in [**15**] was not time-dependent. Our model extends the model in [**15**] to estimate the time-varying covariate effects on the occupance of the first and second progressions.

6.1. Notation

We are using the following notation for the model where there are two chronological events of interest, i.e., progressions. Let Y_{ij} be the indicator of whether or not the subject i has had the 1^{st} event of interest at time j. (Assume that the subject i showed the first sign of the 1st progression after entering the study and at time $j = 0$, before the study starts, they were disease-free, i.e., $Y_{i0} = 0$). Let Z_{ij} be the indicator for whether or not subject i has had the 2^{nd} event of interest at time j. $j = 1, \ldots, M$ are the times at which the 1st progression was assessed, the 2nd progression was recorded, or a subject was censored. x_{ij} is the value of the time-dependent covariate for subject i, during the jth visit.

L_i is the index of the last time that subject i was visited and was event-free (did not experience the first event of interest, $Y_{iL_i} = 0$). R_i is the index of the first event's time (first time when subject i is observed with the 1^{st} progression, $Y_{iR_i} = 1$). It is the index of the first time that subject i was observed with an event. M_i is the index of

the last time that subject i was observed in the study. L_i^* is the index of the last time that subject i was visited and did not experience the second event of interest, given that the first event of interest has happened. i.e., $Z_{ik} = 0 \mid Y_{ij} = 1, k > j$. R_i^* is the index of the 2^{nd} event's time (first time that subject i is observed with the 2^{nd} event of interest, given the first event has previously happened. i.e., $Z_{ik} = 1 \mid Y_{ij} = 1, k > j$.

Let T_i be the time that subject i has experienced the 1^{st} event of interest, $i = 1, \ldots, n$. T_i^* is the time that subject i has the 2^{nd} event of interest, $i = 1, \ldots, n$. Lastly, t_j is the pre-specified follow-up visits at which subjects are monitored, $j = 1, \ldots, M$.

6.2. Model

Since we are interested in modeling binary outcomes, we use a logit link to model the probability of the event as in [**11**]. We model the first progression by

$$\text{logit}(p_{ij}) = \log(p_{ij}/(1 - p_{ij})) = \alpha + \beta' x_{ij}, \tag{39}$$

where

$$p_{ij} = P(y_{ij} = 1 | x_{ij}, T_i > t_{j-1}). \tag{40}$$

p_{ij} is the probability of observing the 1^{st} progression at the jth interval, given a subject was free of progression through t_{j-1} and x_{ij} is the covariate at time t_{j-1} for individual i.

The second progression by

$$\text{logit}(q_{ik}) = \log(q_{ik}/(1 - q_{ik})) = \theta + \gamma' x_{ik}, \tag{41}$$

where

$$q_{ik} = P(z_{ik} = 1 | y_{ij} = 1, k > j, x_{ik}, T_i^* > t_{k-1}). \tag{42}$$

q_{ik} is the probability of observing the 2^{nd} progression at the kth visit for subject i, given the 2^{nd} progression happens after the 1^{st} progression at time j, where $k > j$ and x_{ij} is the covariate at time t_{j-1}.

We construct the full (complete) log-likelihood, assuming all observations were exact.

$$l = \sum_{i=1}^{n} \sum_{j=1}^{M_i} \{[-\log(1+\exp(\alpha+\beta' x_{ij})) + y_{ij}(\alpha+\beta' x_{ij})] + y_{ij}[-\log(1+\exp(\theta+\gamma' x_{ij})) + z_{ij}(\theta+\gamma' x_{ij})]\} \tag{43}$$

We can rewrite the log-likelihood in (43) as

$$l = \sum_{i=1}^{n} \sum_{j=1}^{M_i} \{l_{ij}^{(1)} + y_{ij} l_{ij}^{(2)}\} \tag{44}$$

where $l_{ij}^{(1)}$ is the contribution of subject i at time j to the likelihood for the 1^{st} progression and $l_{ij}^{(2)}$ is the contribution of subject i at time j to the likelihood for the 2^{nd} progression. Note that as long as $y_{ij} = 0$, there is no contribution from $l_{ij}^{(2)}$ to the likelihood. Starting from $j = L_i + 1$, the contribution of $l_{ij}^{(2)}$ starts. Figure 6.1 demonstrates the contribution of the second progression to the log-likelihood.

6.3. Different Patterns for the Partly Interval-Censored Events

We are studying two chronological events and we are dealing with partly interval censored observations. First thing to consider is that there are particular combinations at which the binary responses are defined. Table 6.1 demonstrates the possible scenarios. Since the second progression cannot happen prior to the first progression, $Y_{ij} = 0$ implies that $Z_{ik} = 0$, for $k > j$. Figure 6.2 demonstrates the outcome for a subject who has complete observations.

For the subjects, who miss some of their follow-up visits and have incomplete data, there are several patterns at which the events are censored.

Case 1. The first progression, y_{ij} is interval-censored and the second progression z_{ij} is completely observed. In this case y_{ij} is missing for $j = L_i + 1, \ldots, R_i - 1$, $L_i^* + 1 = R_i^*$, and $R_i \leq L_i^*$.

Case 2. The second progression, z_{ij} is interval-censored and the first progression y_{ij} is completely observed. In this case z_{ij} is missing for $j = L_i^* + 1, \ldots, R_i^* - 1$, $L_i + 1 = R_i$, and $R_i \leq L_i^*$.

Case 3. Both progressions, y_{ij} and z_{ij} are interval-censored. This can happen in two ways. Figure 6.3 displays the possible scenarios.

Case 3.1. The progressions are censored within two mutually exclusive intervals. In this case y_{ij} is missing for $j = L_i + 1, \ldots, R_i - 1$ and z_{ij} is missing for $j = L_i^* + 1, \ldots, R_i^* - 1$, where $R_i < L_i^* + 1$. Figure 6.4 demonstrates these two separates intervals.

Case 3.2. The progressions are censored within the same intervals. In this case y_{ij} and z_{ij} are missing for $j = L_i + 1, \ldots, R_i - 1$. Additionally, we have $L_i = L_i^*$ and $R_i = R_i^*$. Figure 6.5 demonstrates these two intervals.

Figure 6.6 shows the partly interval-censored data for two progression, with an example of all the above patterns for the interval censoring of two events.

6.4. Parameter Estimation Using EM Algorithm

We use EM algorithm [**12**] to estimate the parameters when we have partly interval-censored data. Whenever subjects miss their follow-up visits, their covariate x_{ij} is missing as well as the responses, y_{ij}, z_{ij}, or both. x_{ij} could be continuous or categorical. We assume that x_{ij} has a linear growth curve with fixed effects.

$$x_{ij} = \eta_{0i} + \eta_{1i} t_{j-1} + \epsilon_{ij} \tag{45}$$

where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $cov(\epsilon_{ij}, \epsilon_{ij'}) = 0, j \neq j'$. We estimate x_{ij} by $\widehat{x}_{ij} = \widehat{\eta}_{0i} + \widehat{\eta}_{1i} t_{j-1}$ for $L_i + 1 \leq j \leq R_i - 1$, where $\widehat{\eta}_{0i}$ and $\widehat{\eta}_{1i}$ are least squares estimators.

E-step At E-step, we need to estimate y_{ij} and z_{ij} for the missed visits. However, the E-step would be different for each of the interval-censoring patterns in (6.3).

Case 1. We need to estimate y_{ij} for $j = L_i + 1, \ldots, R_i$.

$$E[y_{ij}|Y, Z, \widehat{x}_{ij}, \alpha, \beta, \theta, \gamma, t_{L_i} < T_i \leq t_{R_i}, T_i^* = t_{R_i^*}, k = R_i^* > j]$$

$$= P[T_i = t_j | t_{L_i} < T_i \leq t_{R_i}, T_i^* = t_{R_i^*}, k = R_i^* > j, Y, Z, \widehat{x}_{ij}, \alpha, \beta, \theta, \gamma] =$$

$$\begin{cases} \frac{\widehat{p}_{ij}}{\widehat{p}_{i,L_i+1} + \sum_{w=L_i+2}^{R_i-1} [\widehat{p}_{iw} \prod_{o=L_i+1}^{w-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\ \text{if } j = L_i + 1, \\ \frac{\widehat{p}_{ij} \prod_{o=L_i+1}^{j-1} (1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1} + \sum_{w=L_i+2}^{R_i-1} [\widehat{p}_{iw} \prod_{o=L_i+1}^{w-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\ \text{if } j \in \{L_i + 2, ..., R_i - 1\}, \\ \frac{\prod_{o=L_i+1}^{j-1} (1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1} + \sum_{w=L_i+2}^{R_i-1} [\widehat{p}_{iw} \prod_{o=L_i+1}^{w-1} (1-\widehat{p}_{io})] + \prod_{o=L_i+1}^{R_i-1} (1-\widehat{p}_{io})} \\ \text{if } j = R_i, \end{cases} \tag{46}$$

where

$$\widehat{p}_{ij} = \frac{\exp(\alpha + \beta' \widehat{x}_{ij})}{1 + \exp(\alpha + \beta' \widehat{x}_{ij})} \tag{47}$$

Case 2. We need to estimate z_{ik} for $k = L_i^* + 1, \ldots, R_i^*$.

$$E[z_{ik}|Y, Z, \widehat{x}_{ij}, \alpha, \beta, \theta, \gamma, T_i = t_{R_i}, t_{L_i^*} < T_i^* \leq t_{R_i^*}, , k > j = R_i]$$

$$= P[T_i^* = t_k | T_i = t_{R_i}, t_{L_i^*} < T_i^* \leq t_{R_i^*}, k > j = R_i, Y, Z, \widehat{x}_{ij}, \alpha, \beta, \theta, \gamma] =$$

$$\begin{cases} \frac{\widehat{q}_{ik}}{\widehat{q}_{i,L_i^*+1} + \sum_{w=L_i^*+2}^{R_i^*-1} [\widehat{q}_{iw} \prod_{o=L_i^*+1}^{w-1} (1-\widehat{q}_{io})] + \prod_{o=L_i^*+1}^{R_i^*-1} (1-\widehat{q}_{io})} \\ \text{if } k = L_i^* + 1, \\ \frac{\widehat{q}_{ik} \prod_{o=L_i^*+1}^{j-1} (1-\widehat{q}_{io})}{\widehat{q}_{i,L_i^*+1} + \sum_{w=L_i^*+2}^{R_i^*-1} [\widehat{q}_{iw} \prod_{o=L_i^*+1}^{w-1} (1-\widehat{q}_{io})] + \prod_{o=L_i^*+1}^{R_i^*-1} (1-\widehat{q}_{io})} \\ \text{if } k \in \{L_i^* + 2, ..., R_i^* - 1\}, \\ \frac{\prod_{o=L_i^*+1}^{j-1} (1-\widehat{q}_{io})}{\widehat{q}_{i,L_i^*+1} + \sum_{w=L_i^*+2}^{R_i^*-1} [\widehat{q}_{iw} \prod_{o=L_i^*+1}^{w-1} (1-\widehat{q}_{io})] + \prod_{o=L_i^*+1}^{R_i^*-1} (1-\widehat{q}_{io})} \\ \text{if } k = R_i^*, \end{cases} \tag{48}$$

where

$$\widehat{q}_{ik} = \frac{\exp(\theta + \gamma' \widehat{x}_{ik})}{1 + \exp(\theta + \gamma' \widehat{x}_{ik})} \tag{49}$$

Case 3.1. When both progressions are censored in mutually exclusive intervals, we need to estimate y_{ij} for $j = L_i + 1, \ldots, R_i$ and z_{ik} for $k = L_i^* + 1, \ldots, R_i^*$, where $k > j$. This can be done by using formulae (46) and (48). Figure 6.4 demonstrates the indices of the responses that need to be estimated at E-step.

Case 3.2. When both progressions are censored within the same interval ($L_i = L_i^*$ and $R_i = R_i^*$), due to the time hierarchy between the two progressions ($k > j$), we cannot use the same formulae as in (46) and (48) for E-step. In this scenario, y_{ij} needs to be estimated for $j = L_i + 1, \ldots, R_i - 1$ and z_{ik} for $k = L_i^* + 2, \ldots, R_i^*$. The E-step in this is as follows.

$$E[y_{ij}|Y,Z,\widehat{x}_{ij},\alpha,\beta,\theta,\gamma,t_{L_i}<T_i\leq t_{R_i},T_i^*=t_{R_i^*},k=R_i^*>j]$$

$$=P[T_i=t_j|t_{L_i}<T_i\leq t_{R_i},T_i^*=t_{R_i^*},k=R_i^*>j,Y,Z,\widehat{x}_{ij},\alpha,\beta,\theta,\gamma]=$$

$$\begin{cases} \frac{\widehat{p}_{ij}}{\widehat{p}_{i,L_i+1}+\sum_{w=L_i+2}^{R_i-1}[\widehat{p}_{iw}\prod_{o=L_i+1}^{w-1}(1-\widehat{p}_{io})]+\prod_{o=L_i+1}^{R_i-1}(1-\widehat{p}_{io})} \\ \text{if } j=L_i+1, \\ \frac{\widehat{p}_{ij}\prod_{o=L_i+1}^{j-1}(1-\widehat{p}_{io})}{\widehat{p}_{i,L_i+1}+\sum_{w=L_i+2}^{R_i-1}[\widehat{p}_{iw}\prod_{o=L_i+1}^{w-1}(1-\widehat{p}_{io})]+\prod_{o=L_i+1}^{R_i-1}(1-\widehat{p}_{io})} \\ \text{if } j\in\{L_i+2,...,R_i-1\}, \end{cases} \tag{50}$$

where

$$\widehat{p}_{ij}=\frac{\exp(\alpha+\beta'\widehat{x}_{ij})}{1+\exp(\alpha+\beta'\widehat{x}_{ij})} \tag{51}$$

and

$$E[z_{ik}|Y,Z,\widehat{x}_{ij},\alpha,\beta,\theta,\gamma,T_i=t_{R_i},t_{L_i^*}<T_i^*\leq t_{R_i^*},,k>j=R_i]$$

$$=P[T_i^*=t_k|T_i=t_{R_i},t_{L_i^*}<T_i^*\leq t_{R_i^*},k>j=R_i,Y,Z,\widehat{x}_{ij},\alpha,\beta,\theta,\gamma]=$$

$$\begin{cases} \frac{\widehat{q}_{ik}\prod_{o=L_i^*+1}^{j-1}(1-\widehat{q}_{io})}{\widehat{q}_{i,L_i^*+1}+\sum_{w=L_i^*+2}^{R_i^*-1}[\widehat{q}_{iw}\prod_{o=L_i^*+1}^{w-1}(1-\widehat{q}_{io})]+\prod_{o=L_i^*+1}^{R_i^*-1}(1-\widehat{q}_{io})} \\ \text{if } k\in\{L_i^*+2,...,R_i^*-1\}, \\ \frac{\prod_{o=L_i^*+1}^{j-1}(1-\widehat{q}_{io})}{\widehat{q}_{i,L_i^*+1}+\sum_{w=L_i^*+2}^{R_i^*-1}[\widehat{q}_{iw}\prod_{o=L_i^*+1}^{w-1}(1-\widehat{q}_{io})]+\prod_{o=L_i^*+1}^{R_i^*-1}(1-\widehat{q}_{io})} \\ \text{if } k=R_i^*, \end{cases} \tag{52}$$

where

$$\widehat{q}_{ik}=\frac{\exp(\theta+\gamma'\widehat{x}_{ik})}{1+\exp(\theta+\gamma'\widehat{x}_{ik})} \tag{53}$$

Figure 6.5 displays the time intervals for which the 1^{st} and 2^{nd} progressions need to be estimated at E-step.

M-step We find the values of α, β, θ, and γ that maximize the expected value of log-likelihood in equation (44), conditioned on the missing data. Therefore, we have

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{\theta}, \widehat{\gamma}) = \arg\max l_{\alpha,\beta,\theta,\gamma} | \widehat{y}_{ij}, \widehat{z}_{ij}, \widehat{x}_{ij}. \tag{54}$$

Expressions for E-step (46), (48), (50), and (52) - (54) are repeated until convergence. As there are no closed forms for $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\theta}$, $\widehat{\gamma}$, we use an optimization package `optim` in R.

6.5. Variance Estimation

We use Louis' method for variance estimation, which is based on the missing information principle. Using the notation in [**39**], we have

$$\begin{aligned} \frac{-\partial^2 \log P(\zeta|W)}{\partial \zeta^2} = & - \int_z \frac{\partial^2 \log P(\zeta|W,V)}{\partial \zeta^2} P(V|\zeta, W) dZ \\ & -Var(\frac{-\partial \log P(\zeta|W,V)}{\partial \zeta}), \end{aligned} \tag{55}$$

where W is the observed data, i.e., partly interval-censored pooled repeated observations for two progressions, V is the latent data, the true unknown counterpart of the interval-censored portion of W, $\zeta|W$ is the observed posterior, and $\zeta|W,V$ is the augmented posterior.

Since there is no closed form for the missing information, the variance in equation (55) was computed using Monte Carlo simulation.

Here, there are four parameters, α and β for the first progression, and the parameters for the second progression are θ and γ. Let Σ be the variance-covariance matrix of all the parameters, Σ_1 be the variance-covariance matrix of the parameters of the

1^{st} progression, and Σ_2 be the variance-covariance matrix of the parameters of the 2^{nd} progression. Furthermore, let $I_{observed}$ be the observed information matrix of all the parameters, $I_{1,observed}$ be the observed information matrix of the parameters of 1^{st} progression, and $I_{2,observed}$ be the observed information matrix of the parameters of 2^{nd} progression. We know that $\Sigma = I^{-1}_{observed}$ and by applying Louis' method we can obtain $I_{observed}$ using $I_{observed} = I_{complete} - I_{missing}$. Therefore, we have

$$\Sigma = I^{-1}_{observed} = \begin{bmatrix} I_{1,observed} & 0 \\ 0 & I_{2,observed} \end{bmatrix}^{-1} = \begin{bmatrix} I^{-1}_{1,observed} & 0 \\ 0 & I^{-1}_{2,observed} \end{bmatrix} \tag{56}$$

since $\frac{\partial^2(.)}{\partial\alpha\theta} = \frac{\partial^2(.)}{\partial\alpha\gamma} = \frac{\partial^2(.)}{\partial\beta\theta} = \frac{\partial^2(.)}{\partial\beta\gamma} = 0$ which implies that $cov(\alpha, \theta) = cov(\alpha, \gamma) = cov(\beta, \theta) = cov(\beta, \gamma)$, i.e., the parameters of the first progression are independent of the parameters of the second progression. Note that the binary responses of the two progressions are dependent due to the chronological hierarchy between them.

The above result is very computationally efficient, because we can partition the high dimensional observed information matrix $I_{observed}$ and find the inverses of the two partitioned pieces in order to obtain the variance-covariance matrix Σ.

6.6. Simulation Study

Furthermore, we applied these method to the simulated data. We simulated $n = 700$ subjects and each of them had $M = 10$ follow-up visits. We randomly selected some subjects. Some of the follow-up visits of these subjects were randomly assigned missing according to the pattern in Section 6.3. We sampled L_i, R_i, L_i^*, and R_i^* using a discrete uniform distribution over appropriate visit indices. This makes interval-censored data. The observed data , also known as partly interval-censored data, is the combination of exact and interval-censored data. We assumed that the simulated data

only contains interval-censored times. Additionally, we generated $x1_{ij}$. It represents a continuous covariate.

$$x1_{ij} \sim N(5.8 + 0.3t_{j-1}, 0.1).$$

We computed the parameter estimates using the EM algorithm in Section 6.4, and variance using the Louis's method in Section 6.5 for the observed data. In addition, we investigated the power of our tests.

6.7. Results

We first simulated the data using Section 6.6. The EM algorithm in Section 6.4 was used for the estimation of the two progressions. The variances of the parameter estimates were calculated using Louis' method, described in Section 6.5. Furthermore, we computed the percent relative bias using $|\frac{estimate-true}{true}| \times 100\%$ as in [**25**]. The results are summarized in Table 6.2. The baseline parameter of the first progression, α was estimates as $\widehat{\alpha} = -22.92$ with the standard error of 0.57. The EM estimate of covariate effect for the first progression, β was $\widehat{\beta} = 3.20$. The standard error was 0.08. The percent relative biases for α and β were 0.36 and 6.62, respectively.

The estimated baseline parameter of the second progression was $\widehat{\theta} = -16.73$ with $S.E. = 3.29$. The estimate of covariate effect for the second progression, γ was $\widehat{\gamma} = 2.03$ with $S.E. = 0.4$. Comparing the covariate effect on first and second progression, shows it has a positive effect on the log-odds of the progressions. However, it affects the log-odds of the first progression at a faster pace than the second progression. Additionally, the parameters of the second progression have larger standard deviations in comparison to the parameters of the first progression. Note that for each individual i, more observations (follow-up visits) are used for the calculations of the variance.

If we have complete observations,

j	1	2	3	4	5	6	7	8	9	10	…	M
y_{ij}	0	0	0	1	1	1	1	1	1	1	…	1
z_{ij}	0	0	0	0	0	0	1	1	1	1	…	1

Contribution of $l_{ij}^{(2)}$ starts.

If we have interval-censored observations,

j	1	2	3	4	5	6	7	8	9	10	…	M
y_{ij}	0	0	0	NA	NA	1	1	1	1	1	…	1
z_{ij}	0	0	0	0	0	0	0	NA	1	1	…	1

Contribution of $l_{ij}^{(2)}$ starts.

FIGURE 6.1. Contribution of the 2nd progression to the log-likelihood function.

Y_{ij}	$Z_{ik}, k > j$	*Note*
0	0	
0	1	**Impossible**
1	0	
1	1	

TABLE 6.1. Outcomes of the Two Progressions

j	1	2	3	4	5	6	7	8	9	10	….	M
y_{ij}	0	0	0	1	1	1	1	1	1	1	….	1
$z_{ik}\ (k > j)$	0	0	0	0	0	0	1	1	1	1	….	1

FIGURE 6.2. Complete data for two progressions.

Parameter	*Estimate*	*S.E.*	*PercentRelativeBias*
α	-22.92	0.57	0.36
β	3.20	0.08	6.62
θ	-16.73	3.29	4.59
γ	2.03	0.40	1.52

TABLE 6.2. Simulation Results for Two Progressions Using One Covariate

Next, the power of tests $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$ and $H_0 : \gamma = \gamma_0$ vs. $H_1 : \gamma \neq \gamma_0$. We considered 4 sample sizes (300, 500, 600, and 700)and for each of these sample sizes we ran $B = 1000$ replications of these tests. The power was calculated for each of the sample sizes by R/B, where R is the number of times when H_0 was rejected in the B replications. Furthermore, we considered 5% level of significance for all of the power simulations. Figure 6.7 shows the power for different values of β_0, γ_0, and different sample sizes. As sample size increases, the bias decreases and the power increases. The power curve was symmetric for all different sample sizes. Sample size of $n = 300$ had the largest bias (0.4 for β and 1.7 for γ).Power was the lowest for this sample size. The power curves of sample sizes of 600 and 700 were very close to each other. As the sample size increases or the parameter values are farther apart from the true parameter value (i.e., an effect size increases), the corresponding power increases. Note that as the number of subjects (n) increases, the parameter estimates are more accurate and the EM algorithm converges faster.

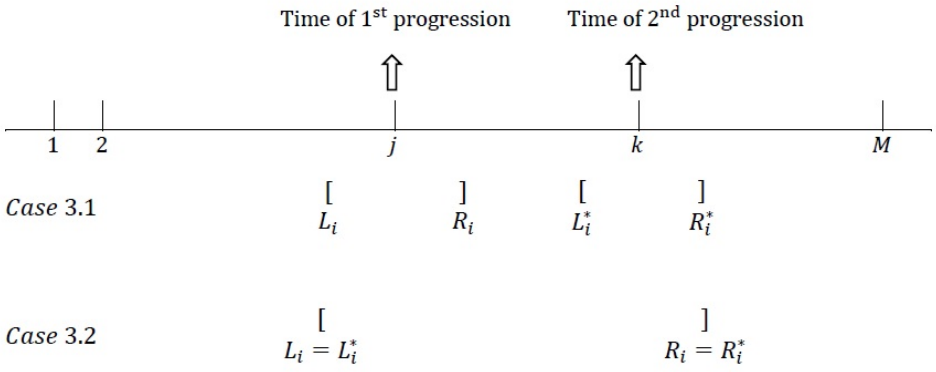

FIGURE 6.3. Both progressions are interval-censored. (Case 3)

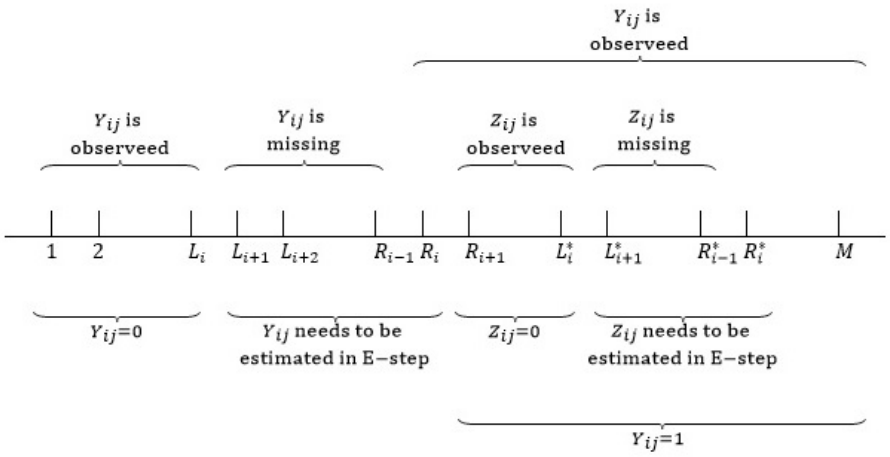

FIGURE 6.4. The two events are censored in different intervals. (Case 3.1)

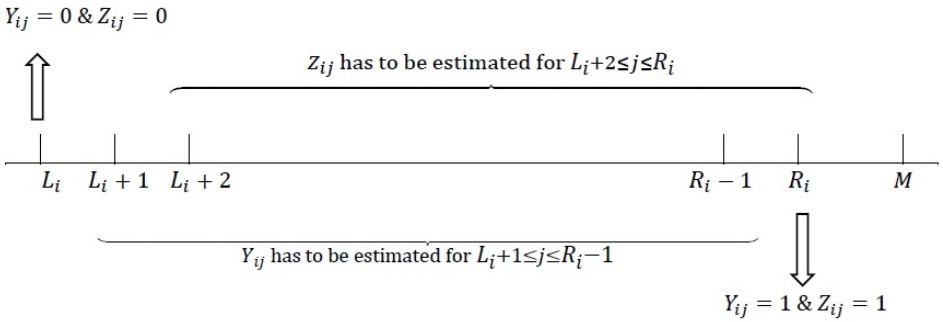

FIGURE 6.5. The two events are censored within the same interval. (Case 3.2)

j	1	2	3	4	5	6	7	8	9	10	M	**Case 1**
y_{ij}	0	0	0	NA	NA	1	1	1	1	1	1	
$z_{ik}\ (k > j)$	0	0	0	0	0	0	1	1	1	1	1	
j	1	2	3	4	5	6	7	8	9	10	M	**Case 2**
y_{ij}	0	0	0	0	1	1	1	1	1	1	1	
$z_{ik}\ (k > j)$	0	0	0	0	0	0	NA	NA	NA	1	1	
j	1	2	3	4	5	6	7	8	9	10	M	**Case 3.1**
y_{ij}	0	0	0	NA	NA	1	1	1	1	1	1	
$z_{ik}\ (k > j)$	0	0	0	0	0	0	0	NA	1	1	1	
j	1	2	3	4	5	6	7	8	9	10	M	**Case 3.2**
y_{ij}	0	0	0	NA	NA	NA	NA	1	1	1	1	
$z_{ik}\ (k > j)$	0	0	0	NA	NA	NA	NA	1	1	1	1	

FIGURE 6.6. Partly interval-censored data for two progressions.

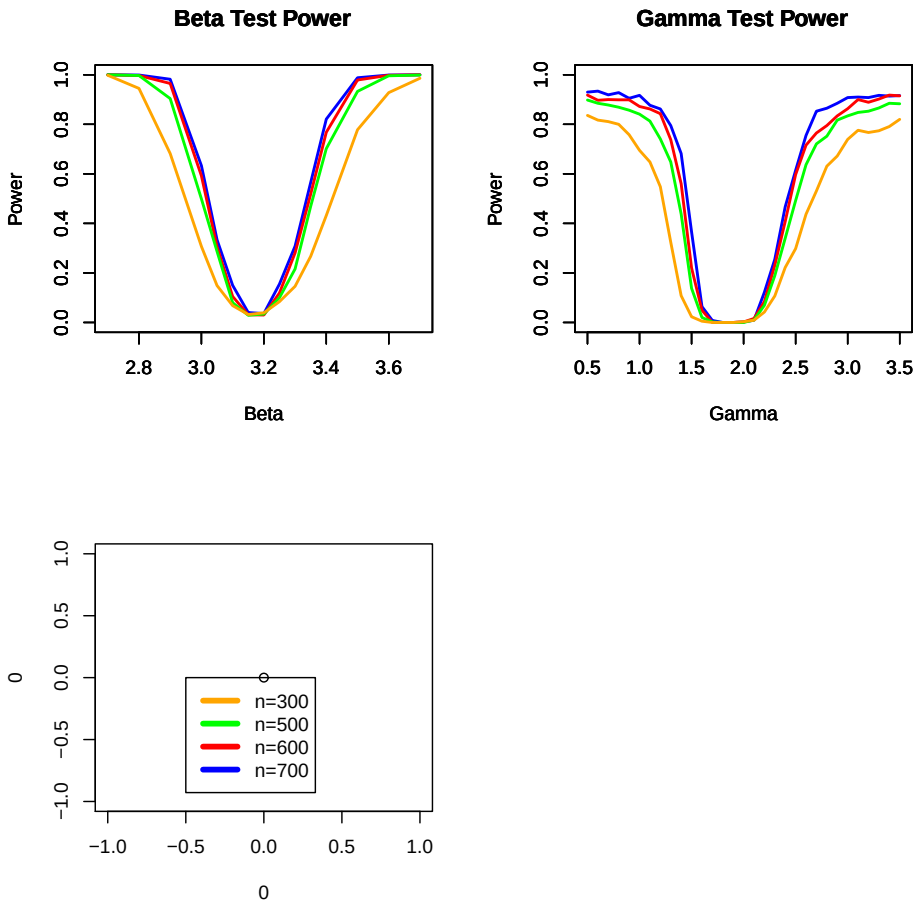

FIGURE 6.7. Power of the test for two peogressions

7

Analysis of NLSY97 Data

7.1. Analysis of One Progression

The National Longitudinal Survey of Youth 1997 (NLSY97) is an ongoing study of subjects who were ages 12-18 in 1997. For the purpose of illustration of our methods, we use the NLSY97 data from 1997 to 2013 ([**41**]). We illustrate how to analyze the effects of covariates that may affect an adolescent's risk of progressing in categories of tobacco smoking.

There are 8984 subjects in the data set. We analyze the 1822 subjects who had not ever smoked a single cigarette at the beginning of the study (1997) but by the end of 2013 became frequent smokers (smoking for more than 10 days in a month). The response variable is defined as

$$y_{ij} = \begin{cases} 1, & \text{a frequent smoker} \\ 0, & \text{not a frequent smoker.} \end{cases} \tag{57}$$

Exact observations are available in about 87.5%. The 1^{st} covariate, $x1_{ij}$, is the number of days an individual drank alcohol in the last 30 days. The 2^{nd} covariate, $x2_{ij}$ is an individual's self-evaluation of "general state of health". $x2_{ij}$ is defined as: 1 being excellent, 2 being very good, 3 being good, 4 being fair, and 5 being poor. The covariate effects are estimated by the EM algorithm in Section 3.1. The standard errors of these estimators are estimated by Louis' method in Section 3.2. The results are shown in Table 7.1. Fixing an individual's self-evaluated health level, as the

subject drinks alcohol one more day during the past 30 days, the odds of becoming a frequent smoker increases by 2.72 (s.e.=0.002). Furthermore, by fixing an individual's amount of drink, as the subject's health level rises (i.e., gets worse) by one unit, the odds of becoming a frequent smoker increases by 1.21 (s.e.=0.015).

Additionally, we analyzed only exact part of the observed data in order to see how much smaller interval-censored portion of the data can help to make the analysis more accurate. Also, some practitioners may only analyze the exact data in practice, due to the unavailability of software. The results are shown in Table 7.2. The parameter estimates are very close to those from the observed data. However, the estimated standard errors are much larger than those from the observed data. This is consistent with the simulation results in Section 3.2. The Wald test statistic for testing $(\beta_1, \beta_2) = (0, 0)$ is quite large for both the exact part of the data alone and the entire observed data. Therefore, the p-values are nearly 0. Though both tests tell us that the covariates have a statistically significant effect on adolescent's smoking behavior, the whole (observed) data provides us with a much stronger evidence for the effect. Therefore, this data analysis reaffirms that using even a small interval-censored portion of the observed data increases the sensitivity of the test.

7.2. Variable Selection in NLSY97

The National Longitudinal Survey of Youth 1997 (NLSY97) from 1997 to 2013 is used to illustrate how to select covariates that may affect an adolescent's smoking behavior. This is the data set that was used in Section 7.1. We analyze 1822 subjects who did not smoke at the beginning of the study (1997) but by the end of 2013 became frequent smokers (smoking for more than 10 days in a month). The response variable is defined as

$$y_{ij} = \begin{cases} 1, & \text{a frequent smoker} \\ 0, & \text{not a frequent smoker.} \end{cases} \tag{58}$$

The 1^{st} covariate, $x1_{ij}$, is the number of days an individual drank alcohol in the last 30 days. The 2^{nd} covariate, $x2_{ij}$ is an individual's self-evaluation of "general state of health". $x2_{ij}$ is defined as: 1 being excellent, 2 being very good, 3 being good, 4 being fair, and 5 being poor. The 3^{rd} covariate, $x3_{ij}$ is whether an individual has ever been suspended from school since the last interview. $x3_{ij}$ is binary with 1 being yes and 0 being no. The 4^{th} covariate, $x4_{ij}$ is whether an individual has ever sold illegal drugs (marijuana (pot, grass), hashish (hash), etc.) since the last interview visit. It is a binary covariate with 1 being yes and 0 being no. Lastly, the 5^{th} covariate, $x5_{ij}$ is whether an individual has ever sold hard illegal drugs (heroin, cocaine, LSD, etc.) since the last interview. $x5_{ij}$ is binary with 1 being yes and 0 being no. The variable selection and estimation of covariate effects via LASSO are done by the EM algorithm in Section 4.3. The results are shown in Table 7.3.

As an individual drinks alcohol one more day during the past 30 days, the odds of becoming a frequent smoker increases by 2.90, fixing the subject's other attributes. Furthermore, as a subject's health level rises (i.e., gets worse) by one unit, the odds of becoming a frequent smoker increases by 1.004, fixing the individual's amount of drink and other attributes. If an individual has sold hard illegal drugs during a given year, the odds of becoming a frequent smoker increases by 3.29. The covariate effects for suspension from school and selling illegal drugs since the last interview visit, were estimated as 0. Therefore, $x3_{ij}$ and $x4_{ij}$ are excluded from the final model.

7.3. NLSY97 Data Analysis Using PRO Ridge Model

The National Longitudinal Survey of Youth 1997 (NLSY97) from 1997 to 2013 is used to illustrate how to estimate the covariates that may affect an adolescent's smoking behavior in the presence of multicollinearity. We analyze the 1822 subjects who did not smoke at the beginning of the study (1997) but by the end of 2013 became frequent smokers (smoking for more than 10 days in a month). The response variable is defined as

$$y_{ij} = \begin{cases} 1, & \text{a frequent smoker} \\ 0, & \text{not a frequent smoker.} \end{cases} \tag{59}$$

The 1^{st} covariate, $x1_{ij}$, is the number of days an individual drank alcohol in the last 30 days. The 2^{nd} covariate, $x2_{ij}$ is an individual's self-evaluation of "general state of health". $x2_{ij}$ is defined as: 1 being excellent, 2 being very good, 3 being good, 4 being fair, and 5 being poor. The 3^{rd} covariate, $x3_{ij}$ is whether an individual has ever been suspended from school since the last interview. $x3_{ij}$ is binary with 1 being yes and 0 being no. The 4^{th} covariate, $x4_{ij}$ is the number of times that an individual sell illegal drugs since the last interview. $x3_{ij}$ and $x4_{ij}$ are highly correlated.

The estimation of covariate effects using the ridge model are done by the EM algorithm in Section 5.2. The results are shown in Table 7.4.

As an individual drinks alcohol one more day during the past 30 days, the odds of becoming a frequent smoker increases by 1.11, fixing the subject's other attributes. Furthermore, as a subject's health level rises (i.e., gets worse) by one unit, the odds of becoming a frequent smoker increases by 1.002, fixing the individual's amount of drink and other attributes. If an individual is suspended from school since the last interview visit, the log of odds of becoming a frequent smoker increases by 1.16.

Lastly, if an individual sells illegal drugs for one more day during the past year, the log of odds of becoming a frequent smoker increases by 1.005.

7.4. Analysis of Two Progression States

The National Longitudinal Survey of Youth 1997 (NLSY97) from 1997 to 2013 is used to illustrate how to analyze the effects of covariates that may affect an adolescent's smoking behavior when there are two stages for smoking behavior. We analyze the 1822 subjects who did not smoke at the beginning of the study (1997) but by the end of 2013 became frequent smokers (smoking for more than 10 days in a month). The first state is becoming an intermittent (light) smoker (smoking for more than 6 days in a month). The response variable for the first progression is defined as

$$y_{ij} = \begin{cases} 1, & \text{an intermittent smoker} \\ 0, & \text{not an intermittent smoker.} \end{cases} \tag{60}$$

The second state is becoming a frequent smoker (smoking for more than 10 days in a month) after being an intermittent smoker. The response variable for the second progression is defined as

$$z_{ij} = \begin{cases} 1, & \text{a frequent smoker} \\ 0, & \text{not a frequent smoker.} \end{cases} \tag{61}$$

The 1^{st} covariate, x_{ij}, is the number of days an individual drank alcohol in the last 30 days. The covariate effects for the 1^{st} and 2^{nd} progressions are estimated by the EM algorithm in Section 6.4. The standard errors of these estimators are estimated by Louis method in Section 6.5. The results are shown in Table 7.5.

As a subject drinks alcohol one more day during the past 30 days, the log of odds of becoming an intermittent smoker increases by 1.12 (s.e.=0.002). Furthermore, as a subject drinks alcohol one more day during the past 30 days, the log of odds of becoming an frequent smoker, given that the individual was previously an intermittent smoker increases by 1.02 (s.e.=0.003). Drinking alcohol during the past 30 days has a larger effect on becoming an intermittent smoker than a frequent smoker. However, the covariate effects' variation is larger for the second progression.

$\hat{\alpha}$	$se(\hat{\alpha})$	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$\hat{\beta}_2$	$se(\hat{\beta}_2)$
-2.36	0.041	0.103	0.002	0.19	0.015

TABLE 7.1. The Results of NLSY97 Analysis Using the Observed Data

$\hat{\alpha}$	$se(\hat{\alpha})$	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$\hat{\beta}_2$	$se(\hat{\beta}_2)$
-2.35	0.067	0.102	0.004	0.18	0.028

TABLE 7.2. The Results of NLSY97 Analysis Using Only the Exact Data

$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
-1.985	1.064	0.004	4.95×10^{-07}	-5.68×10^{-08}	0.191

TABLE 7.3. The Results of NLSY97 Analysis Using LASSO

$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
-1.864	0.094	0.002	0.146	0.005

TABLE 7.4. The Results of NLSY97 Analysis Using Ridge Model

$\hat{\alpha}$	$s.e(\hat{\alpha})$	$\hat{\beta}$	$s.e(\hat{\beta})$	$\hat{\theta}$	$s.e(\hat{\theta})$	$\hat{\gamma}$	$s.e(\hat{\gamma})$
-1.71	0.019	0.115	0.002	1.128	0.022	0.019	0.003

TABLE 7.5. The Results of NLSY97 Analysis for Two States of Smoking

8

Discussion

In this dissertation, we focused on developing a method to estimate the regression parameters and the variance-covariance matrix of those estimators for the pooled repeated observations logistic regression model with partly interval-censored data under various conditions. The EM algorithm was employed to estimate the parameters; and missing information principle to estimate the variance-covariance matrix of those estimators. Monte Carlo simulation demonstrates acceptable levels of bias, standard error, and power.

In Chapter 3, we introduced the PRO logistic regression model with one progression state and partly interval-censored data. Maximum likelihood estimation, based on the EM algorithm was employed to estimate the effects of the time-dependent covariates. The variance of the MLE's was computed using the Louis' method. The simulation results suggest that in practice, one needs a sample of size around 300 to achieve an 80% power of the test to detect a very small effect size (.45) for the regression parameter of interest, but needs a much smaller size, only around 200, for a bit larger effect size (.55).

For the same model, we employed the LASSO method for variable selection. The tuning parameter was estimated using the minimum prediction error criteria. We introduced an EM-based algorithm to perform this method in practice. LASSO trace was used to illustrate the algorithm for a simulated data. Additionally, ridge penalty was used to estimate the regression parameters in the presence of multicollinearity among the covariates. An alternative to combat multicollinearity is a principle

components analysis. [**2**] estimates the parameters of logistic regression using principle components. Ridge method shrinks the collinear attributes whereas principle components just throws them out. Additionally, the tuning parameter, λ can take on any positive value, but the possible values for the tuning parameter in principle components analysis are limited.

In Chapter 6, the PRO logistic regression is extended to the model with two chronological progressions. The interval-censored portion of the data can contain different patterns of missingness. The EM was used to estimate the effects of the time-dependent covariates on the first and the second progressions. Missing information principle was used to estimate the variance-covariance matrix of those estimators. The Monte Carlo simulations showed a similar pattern to the model with one progression. The estimated parameters of the second progression have a higher variance compared to the first progression. We expect this result since there are less observations that contribute to the second progression's likelihood. Additionally, for a given covariate, we need a larger effect size for the second progression to reach a fixed power, compared to the effect size of the first progression. In simulation studies in Section 6.6 and analysis of NLSY97 in Section 7.4, we used the same covariates for modeling the two progressions. However, we do not have to use the same covariates for the two progressions. One could use additional covariates for the second progression and estimate them.

A limitation of using these models is that they can only be used when the progressions are not time-reversible. For example, in AIDS studies, once a patient progresses from being HIV positive to having AIDS, they stay in this new state and cannot go back to the previous state. However, there are applications when transition states are time-reversible and progression can happen multiple time during longitudinal studies, e.g., pregnancy. Our current model. can be easily extended to estimate the effect of

the attributes in models with time-reversible progression states. E-step and M-step would be more complicated.

For those subjects who miss follow-up visits, not only are their event times interval-censored, but also their time-dependent measurements are incomplete. In Section 3.1, we used a linear growth curve with fixed effects to estimate the incomplete covariates. If in certain applications we know that some covariates have a faster growing rates over time, one may consider other types of growth curves (e.g., exponential growth curve) to estimate the missing attributes.

The methods in this dissertation were applied to the situation where the visit times were equally spaced,. However, they can also be used when the predetermined follow-up visits were not evenly distributed in time. If more specific information is provided, one can employ a linear growth curve model with random effects and other models to handle the missing covariates. The model in Chapter (6) can be extended to multiple progression states by the fact that the likelihood factors into a distinct term for each interval [**3**].

Another extension to the PRO logistic regression with partly interval-censored data is to incorporate prior information of the parameters in the analysis. Making Bayesian inferences will allow us to have more robust results. Since we are investigating the binary responses for this model, a logit link was used. Posterior distribution based on the parameters of a logistic regression model does not have a closed form. However, it can be derived numerically. [**19**] gives an example of such derivations for a logistic regression model. Those ideas can be applied to PRO logistic model to take a Bayesian approach to the problem. However, this Bayesian data analysis requires selecting appropriate priors. This is essential to avoid misleading results. Additionally, the posterior distributions can be heavily affected by the choice of the

priors. Unless some adequate prior information of the parameters exists, the classical approach is preferred. The prior information could be obtained from previous literature describing the subjects, or upon investigator's knowledge of the subject.

In this dissertation we assumed fixed effects for modeling the binary outcomes in the log-likelihood, (20) and (43). These likelihoods can be extended in modeling person-specific random intercept and slopes terms to account for the within person correlation as in [**14**]. Repeatedly measuring an outcome on each individual over time can induce correlation among the observations. We may assume mixed effects model in the logit link (18). Then the model in Chapter (3) will be updated to

$$\text{logit}(P_{ij}) = \log(P_{ij}/(1-P_{ij})) = \alpha + \beta' x_{ij} + b_i' z_i, \tag{62}$$

where b_i is the person-specific random effects with $b_i \sim MVN(0, D$ and z_i is the corresponding random effects design matrix. The complete log-likelihood, assuming all observations were exact will be

$$l = \sum_{i=1}^{n} \sum_{j=1}^{M_i} [-\log(1+\exp(\alpha + \beta' x_{ij} + b_i' z_i)) + y_{ij}(\alpha + \beta' x_{ij} + b_i' z_i)] \tag{63}$$

The parameter estimation for this PRO logistic mixed effects model with partly interval-censored observations, requires an additional steps to the EM algorithm in Section 3.1. At E-step we also need to estimate the b_i, the parameter of the random effect and at M-step, D needs to be estimated as well.

This mixed effects model can also be applied to the model in Chapter (4) for variable selection using LASSO. [**29**] proposes an efficient L^1 regulized logistic regression and [**4**] applies LASSO to longitudinal logistic model with mixed effects, assuming that the observations are fully observed. [**28**] introduces an algorithm for estimating

the parameters of random effects longitudinal data with continuous response variable. We may have to add the L^1 norm penalty to the likelihood in (63) and modify the variable selection algorithms.

Lastly, the LASSO variable selection technique does not work in the presence of multicollinearity in the data, [**40**]. If there is a group of highly correlated predictors in the model, LASSO tends to select only one predictor among them. There are some methods that combine the two techniques, LASSO for variable selection and ridge for multicollinearity to avoid the drawback of the LASSO and improve the variable selection. The elastic net (Enet) is a mixture of LASSO and ridge penalty. The mixture of minimum concave penalty (MCP) and ridge penalty, Mnet, has been developed for linear regression. Snet is the mixture of smoothly clipped absolute deviation (SCAD) and ridge penalty. [**42**] proposed new algorithms for the analysis of count data regression with highly correlated biomarkers. Such mixture algorithms can be applied to the PRO logistic regression model with partly interval-censored data in Chapters (3) and (6). These methods can help to simultaneously, improve the variable selection method and combat multicollinearity.

References

[1] S. ADHIKARI, F. LECCI, J. T. BECKER, B. W. JUNKER, L. H. KULLER, O. L. LOPEZ, AND R. J. TIBSHIRANI, *High-dimensional longitudinal classification with the multinomial fused lasso*, Statistics in Medicine, 38 (2019), pp. 2184–2205.

[2] A. M. AGUILERA, M. ESCABIAS, AND M. J. VALDERRAMA, *Using principal components for estimating logistic regression with high-dimensional multicollinear data*, Computational Statistics and Data Analysis, 50 (2006), pp. 1905–1924.

[3] P. D. ALLISON, *Survival analysis using SAS: a practical guide*, SAS Institute, 2012.

[4] A. ARRIBAS-GIL, R. D. L. CRUZ, E. LEBARBIER, AND C. MEZA, *Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators*, Biometrics, 71 (2015), pp. 333–343.

[5] S. B. BULL, C. MAK, AND C. M. GREENWOOD, *A modified score function estimator for multinomial logistic regression in small samples*, Computational Statistics and Data Analysis, 39 (2002), pp. 57–74.

[6] S. L. CESSIE AND J. C. V. HOUWELINGEN, *Ridge estimators in logistic regression*, Applied Statistics, 41 (1992), p. 191.

[7] B. CHEN, G. Y. YI, AND R. J. COOK, *Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process*, Statistics in Medicine, 29 (2010), pp. 205–218.

[8] L. A. Cupples, R. B. DAgostino, K. Anderson, and W. B. Kannel, *Comparison of baseline and repeated measure covariate techniques in the framingham heart study*, Statistics in Medicine, 7 (1988), pp. 1175–1189.

[9] R. B. Dagostino, M.-L. Lee, A. J. Belanger, L. A. Cupples, K. Anderson, and W. B. Kannel, *Relation of pooled logistic regression to time dependent cox regression analysis: The framingham heart study*, Statistics in Medicine, 9 (1990), pp. 1501–1515.

[10] N. Daneshi and J. S. Kim, *A comparison between surrogate and ridge estimators in linear regression*, Technical Report, (2015).

[11] N. Daneshi and J. S. Kim, *Maximum likelihood estimation for the pooled repeated observations logistic regression model with partly interval-censored data*, Submitted, (2019).

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society. Series B (Methodological), 39 (1977), pp. 1–38.

[13] B. Efron, *The two sample problem with censored data*, Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.; Univ. of Calif. Press, (1967), pp. 831–853.

[14] M. Eliot, J. Ferguson, M. P. Reilly, and A. S. Foulkes, *Ridge regression for longitudinal biomarker data*, The International Journal of Biostatistics, 7 (2011), pp. 1–11.

[15] D. M. Finkelstein and D. A. Schoenfeld, *A joint test for progression and survival with interval-censored data from a cancer clinical trial*, Statistics in Medicine, 33 (2014), pp. 1981–1989.

[16] D. M. Finkelstein, R. Wang, L. H. Ficociello, and D. A. Schoenfeld, *A score test for association of a longitudinal marker and an event with missing data*, Biometrics, 66 (2010), pp. 726–732.

[17] D. Firth, *Bias reduction of maximum likelihood estimates*, Biometrika, 80 (1993), pp. 27–38.

[18] F. Gao, D. Zeng, and D.-Y. Lin, *Semiparametric estimation of the accelerated failure time model with partly interval-censored data*, Biometrics, 73 (2017), pp. 1161–1168.

[19] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, CRC Press, 2013.

[20] F. Godnez-Jaimes, G. Ramrez-Valverde, R. Reyes-Carreto, F. Ariza-Hernandez, and E. Barrera-Rodriguez, *Collinearity and separated data in the logistic regression model*, Agrociencia, 46 (2012), pp. 411–425.

[21] A. Hoerl and R. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970).

[22] A. Hoerl, R. Kennard, and K. Baldwin, *Ridge regression: Some simulations*, Communications in Statistics - Simulation and Computation, 4 (1975), pp. 105–123.

[23] J. Huang, *Asymptotic properties of nonparametric estimation based on partly interval-censored data*, Statistica Sinica, 9 (1999), pp. 501–519.

[24] D. Jensen and D. Ramirez, *Surrogate models in ill-conditioned systems*, Journal of Statistical Planning and Inference, 140 (2010), pp. 2069–2077.

[25] T. H. Jung, P. Peduzzi, H. Allore, T. C. Kyriakides, and D. Esserman, *A joint model for recurrent events and a semi-competing risk in the presence of multi-level clustering*, Statistical Methods in Medical Research, (2018), pp. 1–15.

[26] J. S. Kim, *Maximum likelihood estimation for the proportional hazards model with partly interval-censored data*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65 (2003), pp. 489–502.

[27] K. Koh, S.-J. Kim, and S. Boyd, *An interior-point method for large-scale l1-regularized logistic regression*, Journal of Machine Learning Research, 8 (2007), pp. 1519–1555.

[28] N. M. Laird and J. H. Ware, *Random-effects models for longitudinal data*, Biometrics, 38 (1982), pp. 963–974.

[29] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng, *Efficient l1 regularized logistic regression*, American Association for Artificial Intelligence, (2006), pp. 401–408.

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

[31] R. L. Mason, R. F. Gunst, and J. T. Webster, *Regression analysis and problems of multicollinearity*, Communications in Statistics, 4 (1975), pp. 277–292.

[32] K. E. Masyn, H. Petras, and W. Liu, *Growth curve models with categorical outcomes*, Encyclopedia of Criminology and Criminal Justice, (2014), pp. 2013–2025.

[33] L. Meier, S. V. D. Geer, and P. Buhlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70 (2008), pp. 53–71.

[34] S. Mongoue-Tchokote and J.-S. Kim, *New statistical software for the proportional hazards model with current status data*, Computational Statistics and Data Analysis, 52 (2008), pp. 4272–4286.

[35] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, Wiley, 2013.

[36] T. Orchard and M. A. Woodbury, *A missing information principle: Theory and applications*, Sixth Berkeley Symposium, (1972).

[37] J. Ra and K.-J. Rhee, *Efficiency of selecting important variable for longitudinal data*, Psychology, 05 (2014), pp. 6–11.

[38] J. Shen and S. Gao, *A solution to separation and multicollinearity in multiple logistic regression*, Journal of Data Science, 6 (2008), pp. 515–531.

[39] M. A. Tanner, *Tools for statistical inference*, Springer-Verlag New York, 2012.

[40] R. Tibshirani, *Regression shrinkage and selection via the lasso: a retrospective*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58 (1996), pp. 267–288.

[41] B. United States Department of Labor Statistics, *National Longitudinal Survey of Youth 1997 cohort, 1997-2013 (rounds 1-16)*, Produced by the National Opinion Research Center, the University of Chicago and distributed by the Center for Human Resource Research, The Ohio State University., 2015.

[42] Z. Wang, S. Ma, M. Zappitelli, C. Parikh, C.-Y. Wang, and P. Devarajan, *Penalized count data regression with application to hospital stay after pediatric cardiac surgery*, Statistical Methods in Medical Research, 25 (2016), pp. 2685–2703.

[43] M. S. Wulfsohn and A. A. Tsiatis, *A joint model for survival and longitudinal data measured with error*, Biometrics, 53 (1997), pp. 330–339.

[44] W. Yoo, R. Mayberry, S. Bae, K. Singh, U. P. He, and J. W. Lillard, *A study of effects of multicollinearity in the multivariable analysis*, International Journal of Applied Science and Technology, 4 (2014), pp. 9–19.

[45] H. H. Zhang and L. Wenbin, *Adaptive lasso for cox's proportional hazards model*, Biometrika, 94 (2007), pp. 691–703.

[46] X. Zhao, Q. Zhao, J. Sun, and J. S. Kim, *Generalized log-rank tests for partly interval-censored failure time data*, Biometrical Journal, 50 (2008), pp. 375–385.