Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 230-242
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Maximum Likelihood Estimation for the Pooled Repeated Partly Interval-Censored Observations Logistic Regression Model

Naghmeh Daneshi, Jong Sung Kim

Fariborz Maseeh Department of Mathematics and Statistics, Portland State University, Portland, OR, USA
Email: dnaghmeh@pdx.edu, jong@pdx.edu

How to cite this paper: Daneshi, N. and Kim, J.S. (2021) Maximum Likelihood Estimation for the Pooled Repeated Partly Interval-Censored Observations Logistic Regression Model. *Open Journal of Statistics*, **11**, 230-242.
https://doi.org/10.4236/ojs.2021.111012

Received: December 22, 2020
Accepted: February 23, 2021
Published: February 26, 2021

Abstract

Often in longitudinal studies, some subjects complete their follow-up visits, but others miss their visits due to various reasons. For those who miss follow-up visits, some of them might learn that the event of interest has already happened when they come back. In this case, not only are their event times interval-censored, but also their time-dependent measurements are incomplete. This problem was motivated by a national longitudinal survey of youth data. Maximum likelihood estimation (MLE) method based on expectation-maximization (EM) algorithm is used for parameter estimation. Then missing information principle is applied to estimate the variance-covariance matrix of the MLEs. Simulation studies demonstrate that the proposed method works well in terms of bias, standard error, and power for samples of moderate size. The national longitudinal survey of youth 1997 (NLSY97) data is analyzed for illustration.

Keywords

EM Algorithm, Longitudinal Studies, Louis' Method, Partly Interval-Censored Failure Time Data, Pooled Repeated Observations

1. Introduction

In longitudinal studies, subjects who are likely to progress to a new state during the study are monitored over time. For example, in clinical trials, subjects 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. However, others miss their follow-up visits, and they may learn that the event of interest had already occurred when they came back. The event times

for these patients are censored within the corresponding person-specific time intervals. Although there are multiple follow-up visiting intervals for each subject, researchers often use one particular interval that contains the true unknown failure time unless they had accurately determined the failure time. This is known as "partly interval-censored failure time data". There are quite a few research works based on partly interval-censored data such as [1] [2] [3] and [4] among others.

Another commonly available data type in longitudinal studies is called pooled repeated observations. Subjects have multiple follow-up visits as usual. From every visit, a subject obtains a binary outcome for the event of interest. 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. [5] and [6] 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 on 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.

Now, we define pooled repeated partly interval-censored data. We have pooled repeated observations, but some binary outcomes and covariates are incomplete. They can only be determined with certain unknown probabilities within the corresponding specific follow-up visits. In this case, the analysis of such data requires a new method that combines a model that handles pooled repeated observations without censoring and a method that deals 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 extend the work of [7], who employed conditional expected score test (CEST) to determine the presence of 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 and the response is pooled, repeated, and partly interval-censored. We assume that the missing data is missing at random (MAR). In MAR data, there might be systematic differences between the observed and missing data, but the differences can be explained by the observed data. EM algorithm was originally developed to handle MAR data.

The organization of this paper is as follows. In Section 2, we present a logistic regression model for pooled repeated partly interval-censored data. In Section 3, we provide the details of computation of the MLEs of the regression parameter via EM algorithm and the variance estimation through the missing information principle. Section 4 displays the simulation study results. Section 5 illustrates an application to a real data set. Finally, Section 6 briefly summarizes what we have

achieved and also discusses potential extensions of our work.

2. Model

We consider a case of longitudinal studies, where subjects are at risk of an event of interest 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.

Let T_i be the time subject i experiences the event of interest, $i=1,\cdots,n$. At the beginning of the study, every subject is assigned to the same follow-up visits, t_j, $j=1,\cdots,M$. Let y_{ij} be the indicator of whether or not subject i has had the event of interest in the jth interval given a subject was event-free through t_{j-1} and x_{ij}, the covariate at time t_{j-1}. Since we are interested in modeling a binary outcome, we use a logit link to model the probability of event as in [7].

$$\operatorname{logit}\left(p_{ij}\right)=\log\left(p_{ij}\Big/\left(1-p_{ij}\right)\right)=\alpha+\beta' x_{ij}, \tag{1}$$

where

$$p_{ij}=P\left(y_{ij}=1\,|\,x_{ij},T_i>t_{j-1}\right). \tag{2}$$

We construct the full (complete) log-likelihood, assuming as if there were no missing visits while subjects are in the study.

$$l=\sum_{i=1}^{n}\sum_{j=1}^{M_i}\left[-\log\left(1+\exp\left(\alpha+\beta' x_{ij}\right)\right)+y_{ij}\left(\alpha+\beta' x_{ij}\right)\right], \tag{3}$$

where M_i is the index of the last time subject i was in the study.

3. Methods

3.1. Parameter Estimation

Assume that the i^{th} subject missed visits after time t_{L_i} and came back at t_{R_i}. L_i is the index of the last time 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. Then $y_{iL_i}=0$, $y_{iR_i}=1$, and y_{ij} is missing for $L_i+1\le j\le R_i-1$. For the subjects who do not miss visits, $L_i+1=R_i$. Whenever subjects miss visits, their covariate value, x_{ij}, is also missing. We use the EM algorithm ([8]) to estimate the parameters.

E-step: For individuals whose failure times are interval-censored, we need to estimate both y_{ij} and x_{ij} in the expression (3) for $j\in\{L_i+1,\cdots,R_i-1\}$.

x_{ij} could be continuous or categorical ([9]). We assume that x_{ij} has a linear growth curve with fixed effects to incorporate a real data, NLSY97. That is,

$$x_{ij}=\theta_{0i}+\theta_{1i}t_{j-1}+\varepsilon_{ij}, \tag{4}$$

where $\varepsilon_{ij}\sim N\left(0,\sigma_\varepsilon^2\right)$, $cov\left(\varepsilon_{ij},\varepsilon_{ij'}\right)=0, j\neq j'$. We estimate x_{ij} by $\hat{x}_{ij}=\hat{\theta}_{0i}+\hat{\theta}_{1i}t_{j-1}$ for $L_i+1\le j\le R_i-1$, where $\hat{\theta}_{0i}$ and $\hat{\theta}_{1i}$ are least squares es-

timators.

If x_{ij} is ordinal, we assign numbers to corresponding categories. Then we again assume linear growth curve with fixed effects to estimate the missing x_{ij}'s. Let n_c be the number of categories for this ordinal variable. For each individual *i*, the observed x_{ij}'s are used in model (4) to compute $\hat{\theta}_{0i}$ and $\hat{\theta}_{1i}$. Then we compute $\hat{x}_{ij} = \hat{\theta}_{0i} + \hat{\theta}_{1i} t_{j-1}$ as usual.

Next, we create $n_c - 1$ thresholds in order to uniquely assign $\hat{x}_{ij}$ into one of the n_c categories. Note that $\hat{x}_{ij} \sim N\left(\theta_{0i} + \theta_{1i} t_{j-1}, \sigma^2_{\hat{x}_i}\right)$ and $\hat{x}_i = \left\{\hat{x}_{i,L_i+1}, \cdots, \hat{x}_{i,R_i-1}, x_{i,R_i}\right\}$. We use the quantiles of this normal distribution to define the thresholds. Since we need to compute $\hat{\sigma}^2_{\hat{x}_i}$ to define thresholds, we need at least three distinct observed covariate values, x_{ij}'s for each subject, otherwise, $\hat{\sigma}^2_{\hat{x}_i}$ would be undefined due to the zero degrees of freedom.

The observed ordinal covariates for some subjects do not include the entire ordinal categories. Therefore, the ordinal logistic regression model does not work for estimating ordinal covariates. In **Appendix 2**, we provide a detailed rationale for choosing fixed effects model, its extension in a general setting, and challenges with random effects model.

$$\begin{aligned}
&\hat{y}_{ij} = E\left[y_{ij} \mid Y_i, \hat{x}_i, \alpha, \beta, T_i > t_{j-1}\right] \\
&= P\left[T_i = t_j \mid t_{L_i} < T_i \le t_{R_i}, Y_i, \hat{x}_i, \alpha, \beta\right] \\
&= \begin{cases}
\dfrac{\hat{p}_{ij}}{\hat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1}\left[\hat{p}_{ik}\prod_{o=L_i+1}^{k-1}\left(1-\hat{p}_{io}\right)\right] + \prod_{o=L_i+1}^{R_i-1}\left(1-\hat{p}_{io}\right)} & \text{if } j = L_i+1, \\
\dfrac{\hat{p}_{ij}\prod_{o=L_i+1}^{j-1}\left(1-\hat{p}_{io}\right)}{\hat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1}\left[\hat{p}_{ik}\prod_{o=L_i+1}^{k-1}\left(1-\hat{p}_{io}\right)\right] + \prod_{o=L_i+1}^{R_i-1}\left(1-\hat{p}_{io}\right)} & \text{if } j \in \left\{L_i+2, \cdots, R_i-1\right\}, \\
\dfrac{\prod_{o=L_i+1}^{j-1}\left(1-\hat{p}_{io}\right)}{\hat{p}_{i,L_i+1} + \sum_{k=L_i+2}^{R_i-1}\left[\hat{p}_{ik}\prod_{o=L_i+1}^{k-1}\left(1-\hat{p}_{io}\right)\right] + \prod_{o=L_i+1}^{R_i-1}\left(1-\hat{p}_{io}\right)} & \text{if } j = R_i,
\end{cases}
\end{aligned} \tag{5}$$

where

$$\hat{p}_{ij} = \frac{\exp\left(\alpha + \beta'\hat{x}_{ij}\right)}{1 + \exp\left(\alpha + \beta'\hat{x}_{ij}\right)} \tag{6}$$

This is an extension of a geometric-type experiment, where the probability of success (progression) changes at each follow-up visit, t_j, $L_i + 1 \le j \le R_i$.

M-step: We find the values of α and β that maximize the expected value of log-likelihood in Equation (3), conditioned on the observed data. Therefore, we have

$$\left(\hat{\alpha}, \hat{\beta}\right) = \arg\max l_{\alpha,\beta} \mid \hat{y}_{ij}, \hat{x}_{ij}, \tag{7}$$

where $\hat{y}_{ij} = y_{ij}, \hat{x}_{ij} = x_{ij}$, if uncensored.

Expressions (5)-(7) are repeated until convergence. As there are no closed forms for $\hat{\alpha}$ and $\hat{\beta}$, we used an optimization package optim in R to obtain $\left(\hat{\alpha}, \hat{\beta}\right)$.

3.2. Variance Estimation

We apply Louis' method for variance estimation using the notation in [10]. Following the missing information principle, we compute the observed information by subtracting the missing information from the complete information.

$$\frac{-\partial^2 \log P(\theta \mid W)}{\partial \theta^2} = -\int_v \frac{\partial^2 \log P(\theta \mid W,V)}{\partial \theta^2} P(V \mid \theta, W)\,\mathrm{d}V - Var\left(\frac{-\partial \log P(\theta \mid W,V)}{\partial \theta}\right), \tag{8}$$

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

The details of the expression (8) are provided in **Appendix 3**.

4. Simulation Study

4.1. Data Simulation

We considered $n = 300$ subjects who have $M = 7$ follow-up visits each. We generated covariates as follows:

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

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

$x1_{ij}$ 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.

First, we generate $n = 300$ subjects who have complete follow-up visits. This makes the original complete data (OC), pooled repeated data. We randomly choose n_1 subjects out of these. This makes the exact data (E), a proper subset of the OC. For the remaining $n_2 = n - n_1$ subjects, we randomly designate some of their follow-up visits missing. This makes the pooled repeated interval-censored observations. The observed data (O), which is the pooled repeated partly interval-censored data, is the mix of pooled repeated data (E) and pooled repeated 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 the exact data, we have $R_i = L_i + 1$ and for the pooled repeated interval-censored data, $R_i \geq L_i + 2$. Then for $j = 1, \cdots, L_i$, we have $y_{ij} = 0$ and for $j = R_i, \cdots, M$, we have $y_{ij} = 1$. y_{ij} is missing for $j = L_i + 1, \cdots, R_i - 1$ in the pooled repeated interval-censored data. y_{ij} is 1 when the i^{th} subject at risk at the j^{th} visit experiences the event of interest in the j^{th} interval.

We computed the bias and variance for original complete data, exact data, and observed data based on $B = 1500$ replications. In addition, we investigated the power of our test.

4.2. Results

We first considered the case where there was only one attribute 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 1**. For all the different combinations of n_1 and n_2, the proposed estimator based on the observed data produces a smaller bias and a smaller variance than that based on the exact data alone. In particular, for the case of (250, 50), containing E 84% (250) and only 16% (50) pooled-repeated interval-censored data, the proposed estimator produces a smaller bias and a smaller variance than that based on E alone. We also notice that the more exact data we have, the smaller bias and variance we get. These results have a quite similar pattern to those in [3], who employed a proportional hazards model with partly interval-censored data. [6] 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. In order to see if bootstrap would be of help, we also ran simulations with various pairs of n_1 and n_2 to compare the bootstrap variance with the variances for the O, E, and OC. We considered two covariates; one is continuous and the other is ordinal. **Table 2** shows the results. For all pairs of n_1 and n_2, the bootstrap variance for the O is smaller than that for OC, which is supposed to be the smallest. This is, bootstrapping suffers from substantial underestimation. Therefore, we do not recommend it for this setting. Another issue is that it is time-consuming.

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 1** and **Figure 2** 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 1**, 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. [11] achieved approximately 80% power in

Table 1. Results for 1-dimensional β, $\beta^{true} = 3.6$, B: Bias, σ^2 : variance.

(n_1, n_2)	B_E	B_O	B_{OC}	σ_E^2	σ_O^2	σ_{OC}^2
(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 2. Estimated variance, boot: bootstrap.

(n_1, n_2)	Parameter	σ_O^2	σ_{OC}^2	σ_E^2	σ_{Boot}^2
(50, 250)	β_1	0.7312	0.1514	1.2726	0.1244
	β_2	0.0152	0.0049	0.0287	0.0035
(100 ,200)	β_1	0.4723	0.1617	0.5491	0.1157
	β_2	0.0108	0.0050	0.0160	0.0022
(150, 150)	β_1	0.2765	0.1463	0.3130	0.1175
	β_2	0.0076	0.0043	0.0087	0.0025
(200, 100)	β_1	0.1848	0.1442	0.2153	0.1391
	β_2	0.0064	0.0049	0.0077	0.0028
(250, 50)	β_1	0.1522	0.1342	0.1604	0.1274
	β_2	0.0053	0.0050	0.0059	0.0039
(270, 30)	β_1	0.1677	0.1621	0.1696	0.1497
	β_2	0.0052	0.0049	0.0055	0.0026
(290, 10)	β_1	0.1483	0.1461	0.1509	0.1321
	β_2	0.0047	0.0046	0.0048	0.0029

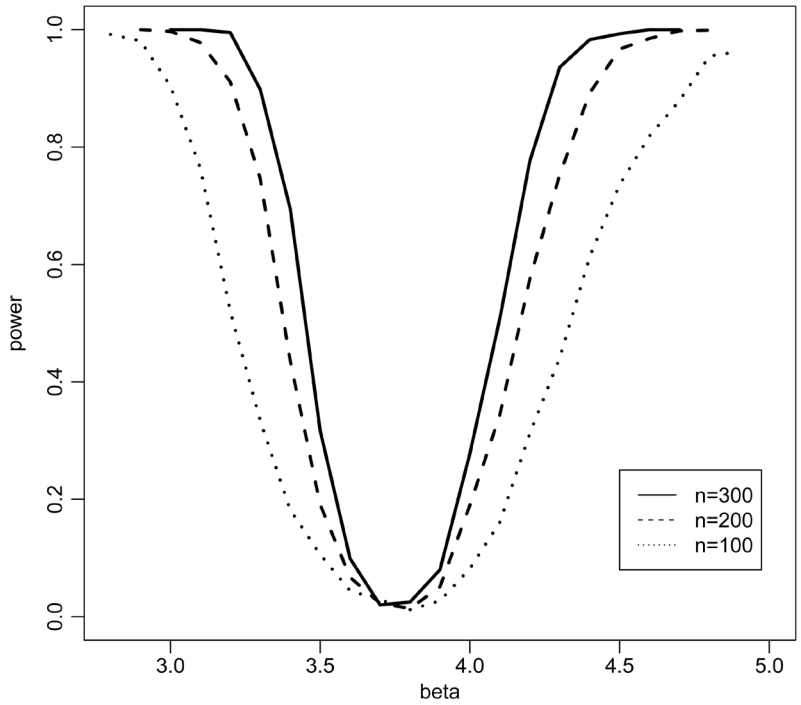

Figure 1. Power of the test for one-dimensional β.

detecting the effect size of 0.75 for the proportional hazards model with a sample of size 300 using current status data. Considering that pooled repeated partly interval-censored data has more information than current status data, we fully

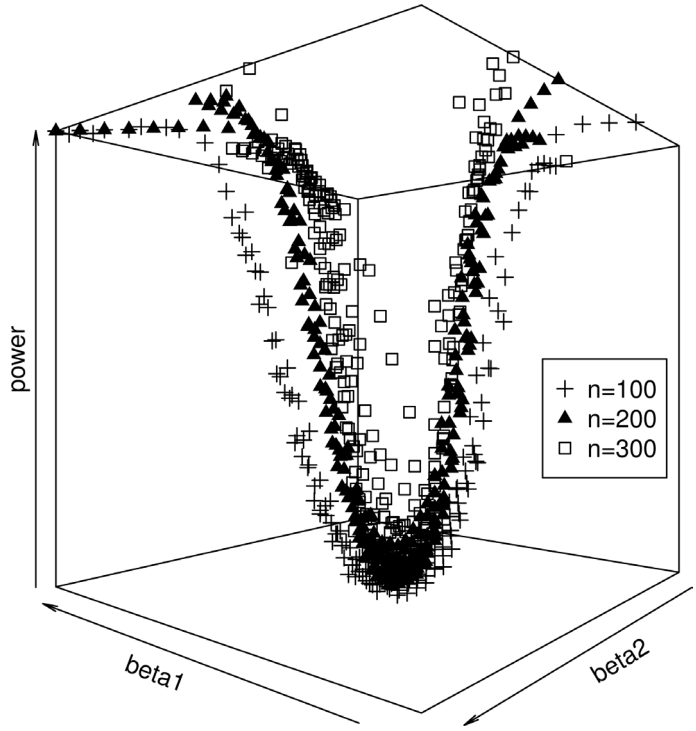

Figure 2. Power of the test for multidimensional β.

Table 3. The 95% coverage probabilities.

(n_1, n_2)	β_1	β_2	Joint
(50, 250)	0.878	0.853	0.835
(100, 200)	0.883	0.874	0.866
(150, 150)	0.899	0.886	0.881
(200, 100)	0.910	0.906	0.903
(250, 50)	0.947	0.931	0.936

agree with this better power result. The 95% coverage probabilities for different proportions of pooled repeated partly interval-censored data are shown in **Table 3**.

In summary, even a small amount of pooled repeated interval-censored data within O does make our statistical inference more accurate and more powerful.

5. Analysis of NLSY97 Data

For more than 4 decades, the National Longitudinal Surveys (NLS) data have served as an important tool for economists, sociologists, and other researchers. The NLSY97 is a nationally representative sample of approximately 9000 youths who were 12 to 18 years old as of December 31, 1996. The NLSY97 is designed to document the transition from school to work and into adulthood. It collects extensive information about youths' labor market behavior and educational experiences over time. In addition to educational and labor market experiences, the

NLSY97 contains detailed information on many other topics. Some of the areas included in the data are criminal behavior, alcohol, and drug use. For the purpose of illustration of our methods, we use the NLSY97 data from 1997 to 2013 ([12]). We illustrate how to analyze the effects of covariates that may affect an adolescent's smoking behavior.

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

$$y_{ij} = \begin{cases} 1, & \text{a frequent smoker} \\ 0, & \text{not a frequent smoker.} \end{cases} \tag{9}$$

Exact observations (E) are available in approximately 87.5% of those analyzed. The 1st covariate, $x1_{ij}$, is the number of days an individual drank alcohol in the last 30 days. The 2nd covariate, $x2_{ij}$, is an individual's self-evaluation of "general state of health". $x2_{ij}$ is defined as: 1 = excellent, 2 = very good, 3 = good, 4 = fair, and 5 = poor. The covariate effects are estimated by the EM algorithm in Section 0. The standard errors of these estimators are computed by Louis' method in Section 0. The results are shown in **Table 4**. Fixing an individual's self-evaluated health level as the subject drinks alcohol one more day during the past 30 days, the log of odds of becoming a frequent smoker increases by 0.1 (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 log of odds of becoming a frequent smoker increases by 0.19 (s.e. = 0.015).

Additionally, we analyzed only E from O in order to see how much smaller the pooled repeated interval-censored data can help make the analysis more accurate. Another rationale for this is some practitioners often analyze only E due to the unavailability of software. The results are shown in **Table 5**. The parameter estimates are very close to those from O. However, the estimated standard errors are much larger than those from O. 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 E alone and O. 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, O provides us with much stronger evidence for the

Table 4. The results of NLSY97 analysis using the observed data.

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

Table 5. The results of NLSY97 analysis using only the exact 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

effect. Therefore, this data analysis reaffirms that even a small amount of pooled repeated interval-censored portion of O increases the sensitivity of the test.

6. Discussion

We focused on developing a method to estimate the regression parameters and the variance-covariance matrix of those estimators for the pooled repeated partly interval-censored data logistic regression model. We employed the EM algorithm 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. To our knowledge, this is the first extensive power study for the pooled repeated partly interval-censored data logistic regression model. 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 (0.45) for the regression parameter of interest. However, one needs a much smaller size, only around 200, for a bit larger effect size (0.55).

There are several potential extensions of our methods. Our methods can also be used when the predetermined follow-up visits were person-dependent. Our methods can be extended to handle correlated covariates by employing a ridge regression model ([13]), variable selections by lasso regression ([14]), and multiple progression states due to the fact that the likelihood factors into a distinct term for each interval ([15]).

Last but not least, we note that there are challenges in including either left-censoring or right-censoring. Refer to **Appendix 1** for details.

Acknowledgements

The authors appreciate Dr. Alexis Dinno for introducing the data.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gao, F., Zeng, D.L. and Lin, D.Y. (2017) Semiparametric Estimation of the Accelerated Failure Time Model with Partly Interval-Censored Data. *Biometrics*, **73**, 1161-1168. https://doi.org/10.1111/biom.12700

[2] Zhao, X.Q., Zhao, Q., Sun, J.G. and Kim, J.S. (2008) Generalized Log-Rank Tests for Partly Interval-Censored Failure Time Data. *Biometrical Journal*, **50**, 375-385. https://doi.org/10.1002/bimj.200710419

[3] Kim, J.S. (2003) Maximum Likelihood Estimation for the Proportional Hazards Model with Partly Interval-Censored Data. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **65**, 489-502. https://doi.org/10.1111/1467-9868.00398

[4] Huang, J. (1999) Asymptotic Properties of Nonparametric Estimation Based on

Partly Interval-Censored Data. *Statistica Sinica*, **9**, 501-519.

[5] Adrienne Cupples, L., D'Agostino, R.B., Anderson, K. and Kannel, W.B. (1988) Comparison of Baseline and Repeated Measure Covariate Techniques in the Framingham Heart Study. *Statistics in Medicine*, **7**, 205-218. https://doi.org/10.1002/sim.4780070122

[6] D'Agostino, R., Lee, M.L., Belanger, A., Cupples, L.A., Anderson, K. and Kannel, W.B. (1990) Relation of Pooled Logistic Regression to Time Dependent Cox Regression Analysis: The Framingham Heart Study. *Statistics in Medicine*, **9**, 1501-1515. https://doi.org/10.1002/sim.4780091214

[7] Finkelstein, D.M., Wang, R., Ficociello, L.H. and Schoenfeld, D.A. (2010) A Score Test for Association of a Longitudinal Marker and an Event with Missing Data. *Biometrics*, **66**, 726-732. https://doi.org/10.1111/j.1541-0420.2009.01326.x

[8] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1997) Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *Journal of the Royal Statistical Society*: *Series B* (*Methodologica*), **39**, 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[9] Masyn, K.E., Petras, H. and Liu, W. (2014) Growth Curve Models with Categorical Outcomes. In: Bruinsma, G. and Weisburd, D., Eds., *Encyclopedia of Criminology and Criminal Justice*, Springer, New York, 2013-2025. https://doi.org/10.1007/978-1-4614-5690-2_404

[10] Tanner, M.A. (1996) Tools for Statistical Inference. 3rd Edition, Springer-Verlag, New York. https://doi.org/10.1007/978-1-4612-4024-2

[11] Mongoué-Tchokoté, S. and Kim, J.S. (2008) New Statistical Software for the Proportional Hazards Model with Current Status Data. *Computational Statistics and Data Analysis*, **52**, 4272-4286. https://doi.org/10.1016/j.csda.2008.02.007

[12] Bureau of Labor Statistics, U.S. Department of Labor (2015) 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. Columbus.

[13] Hoerl, A., Kennard, R. and Baldwin, K. (1975) Ridge Regression: Some Simulations. *Communications in Statistics*, **4**, 105-123.

[14] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[15] Allison, P.D. (2010) Survival Analysis Using SAS: A Practical Guide. SAS Institute, Cary.

[16] Wulfsohn, M.S. and Tsiatis, A. (1997) A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, **53**, 330-339. https://doi.org/10.2307/2533118

Appendix 1. Right and Left Censoring in the Model

In some special cases, the visiting time of some subjects in the data may have either right or left censoring. If a subject has not failed at the last visit ($y_{iL_i}=0$) 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 $R_i=M$. As NLSY predetermined M for all subjects, M plays the role of ∞.

One may want to impute the covariate, x_{ij} and reponse, y_{ij} according to the procedures in Section 3.1. Unfortunately, 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 the event of interest, then both y_{ij} and x_{ij} are missing for $j=1,\cdots,k-1$, and $y_{ik}=1$. Therefore, the covariate, x_{ij} and response, y_{ij} should be estimated for $j\le k-1$ at E-step. We merely have $L_i=0$, $R_i=k$, and two observed covariate values x_{i0} and x_{ik}. 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.

Appendix 2. Imputation of Covariates

In Section 3.1, we assumed that covariates have a linear growth curve with fixed effects. This was motivated by NLSY97 data. In NLSY97, follow-up interviews were relatively far apart (1 year). Additionally, some individuals had no change in their covariate values, e.g., some individuals had no drinking throughout the study. This motivated us to assume that for a given individual, the covariate values are uncorrelated at different follow-up visits, *i.e.*, $cov\left(\varepsilon_{ij},\varepsilon_{ij'}\right)=0, j\ne j'$.

If the follow-up time intervals are relatively short and there are no constant covariate values for any individual over time, one may adopt a linear growth curve with fixed effects and autocorrelated errors. That is,

$$x_{ij}=\theta_{0i}+\theta_{1i}t_{j-1}+\varepsilon_{ij}, \tag{10}$$

where ε_{ij} is an autoregressive process with lag 1, AR (1), $\varepsilon_{ij}\sim N\left(0,\sigma_\varepsilon^2\right)$, $cov\left(\varepsilon_{i,j},\varepsilon_{i,j+1}\right)=\rho,\rho\ne 0$.

[7] and [16] assumed random effects. In a linear growth curve with random effects, all subjects have the same growth curve distribution, which depends on time points and it is correlated within the same subject. The least squares estimators for this model are the same for all subjects. For example, assume that we get $\hat{\theta}_0$ and $\hat{\theta}_1$ for a random effect model. If two subjects i^* and i^{**} have missing covariates at a given time point j, then they will have the same estimated covariate $\hat{x}_{ij}=\hat{\theta}_0+\hat{\theta}_1 t_j$ for $i=i^*,i^{**}$. This may cause a substantial amount of bias.

Appendix 3. Formulas for Computing the Variance in Section 3.2

The variance estimation is based on Y_i, the observed binary outcomes for subject i and the variability of missing response, y_{ij} conditioned on x_i, where $x_i = \left\{x_{i1}, x_{i2}, \cdots, x_{i,L_i}, \hat{x}_{i,L_i+1}, \cdots, \hat{x}_{i,R_i-1}, x_{i,R_i}\right\}$. Let $\theta = \left(\alpha, \vec{\beta}\right)$, and $z_i = \left(1, x_i\right)$. Then the complete information matrix in (8) can be computed by

$$\sum_i z_i z_i^{\mathrm{T}} \frac{\exp\left(\theta^{\mathrm{T}} z_i\right)}{\left(1+\exp\left(\theta^{\mathrm{T}} z_i\right)\right)^2}. \tag{11}$$

The missing information in (8) is computed by Monte Carlo simulations.

$$\begin{aligned} &Var\left(\frac{-\partial \log P\left(\theta \mid W, V\right)}{\partial \theta}\right) \\ &= E\left(\frac{-\partial \log P\left(\theta \mid W, V\right)}{\partial \theta}\right)^2 - \left[E\left(\frac{-\partial \log P\left(\theta \mid W, V\right)}{\partial \theta}\right)\right]^2, \end{aligned} \tag{12}$$

where

$$E\left(\frac{-\partial \log P\left(\theta \mid W, V\right)}{\partial \theta}\right) \approx \frac{1}{B}\sum_{b=1}^{B}\left[\frac{-\partial \log P\left(\theta \mid W, V_b\right)}{\partial \theta}\right]$$

and

$$E\left(\frac{-\partial \log P\left(\theta \mid W, V\right)}{\partial \theta}\right)^2 \approx \frac{1}{B}\sum_{b=1}^{B}\left[\frac{-\partial \log P\left(\theta \mid W, V_b\right)}{\partial \theta}\right]^2.$$