Semiparametric Regression Analysis of Panel Count Data: A Practical Review

Sy Han Chiou[1], Chiung-Yu Huang[2], Gongjun Xu[3], and Jun Yan[4]

[1]Department of Mathematical Sciences, University of Texas at Dallas, U.S.A.
[2]Department of Epidemiology and Biostatistics, University of California at San Francisco, U.S.A.
[3]Department of Statistics, University of Michigan, U.S.A.
[4]Department of Statistics, University of Connecticut, U.S.A.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/insr.12271

Summary

Panel count data arise in many applications when the event history of a recurrent event process is only examined at a sequence of discrete time points. In spite of the recent methodological developments, the availability of their software implementations has been rather limited. Focusing on a practical setting where the effects of some time-independent covariates on the recurrent events are of primary interest, we review semiparametric regression modeling approaches for panel count data that have been implemented in R package `spef`. The methods are grouped into two categories depending on whether the examination times are associated with the recurrent event process after conditioning on covariates. The reviewed methods are illustrated with a subset of the data from a skin cancer clinical trial.

Key words: Counting process; estimating equation; frailty; maximum likelihood; recurrent event

1 Introduction

Panel count data is a special kind of event history data where the occurrence of recurrent events is observed only at a sequence of discrete time points, as opposed to being observed continuously in time. In contrast to conventional recurrent event data, where the exact occurrence times of the events are known, panel count data only have the count of events in each "panel" between successive examination times points (Kalbfleisch and Lawless, 1985). Panel count data frequently arise in many fields such as clinical trials, epidemiological studies, and engineering, when continuous follow-up to obtain exact event times of each subject is infeasible or too costly. The term "panel count" in econometrics refers to longitudinal or clustered count data (e.g., Riphahn et al., 2003; Croissant et al., 2008; Hsiao, 2014); although somewhat related, it is to be distinguished from the context of event history data as we focus on here.

The goal of this article is to review regression analysis for panel count data with a focus on methods that are available in the R environment (R Core Team, 2017). Many statistical methods have been developed to analyze panel count data, but quality controlled software implementation remains rather limited. In their recently published book on panel count data analysis, Sun and Zhao (2013) noted the absence of actively maintained software packages at the time of writing their book (Sun and Zhao, 2013, p.222). Two R packages for panel count data are publicly available at this time. Package `spef` (Chiou et al., 2017) provides multiple methods in a unified interface, with an earlier version presented in Wang and Yan (2011). Package `PCDSpline` (Yao and Wang, 2014) is an implementation of the gamma frailty model of Yao et al. (2016). Instead of providing a comprehensive review of all existing methods, we focus on semiparametric regression models with time-independent covariates as implemented in the `spef` package; methods and software for handling time-varying covariates have been much less developed (Huang et al., 2010). Covariate effects on the recurrent events are of primary interest. Non-parametric estimation is possible with `spef` package by specifying an intercept-only model. We give more details on methods that are available in `spef` package

and that were not treated in detail in Sun and Zhao (2013). The illustration code will help readers who need to analyze a panel count dataset to obtain some quick insights easily.

One challenge in practical panel count data analysis is that the examination process or the follow-up time may be informative about the recurrent event process even after conditioning on available covariates. For example, patients with higher tumor recurrence rates may have more frequent clinical examinations as they may require more medical attention (Li et al., 2011; Sun and Zhao, 2013). Another example is in labor progression of women giving childbirth, if each 1 cm increment of cervical dilation is treated as a recurrent event, then women with faster cervix dilation may have more frequent vaginal examinations (Ma and Sundaram, 2017). Informative examination times are often encountered in panel count data, and falsely treating informative examination times as noninformative could result in biased regression coefficient estimation and misleading conclusions. Similar situations may arise where the follow-up time is informative. Therefore, we grouped the methods into two categories depending on whether or not informative examinations or follow-up times can be accommodated.

This article is organized as follows. A subset of the data from a skin tumor clinical trial is introduced in Section 2 to demonstrate the structure and graphical features of panel count data. Notations of observed data and some of the most popular semiparametric models are presented in Section 3. Methods under the assumption of noninformative and informative examination/censoring times are reviewed in Section 4 and Section 5, respectively, illustrated with the skin tumor data. The performances of the implemented methods under different settings in a simulation study are reported in Section 6. A discussion concludes in Section 7.

2 Skin Cancer Chemoprevention Trail

We illustrate the usage of the `spef` package with a skin cancer prevention study (Bailey et al., 2010). The whole dataset is available in Sun and Zhao (2013, Table A.3.) and is included in

the `spef` package under the name `skinTumor`. The study was a randomized, double-blind, placebo-controlled phase-3 clinical trial conducted at the University of Wisconsin Comprehensive Cancer Center. The primary objective was to determine whether the application of difluoromethylornithine (DFMO) as a chemoprevention agent would lead to a significant reduction in the occurrence of new skin tumors. The study consisted of 290 patients with a history of skin tumor. These patients were randomly assigned into two groups: a treatment group with oral DFMO at a daily dose of 0.5 gram/m^2 and a placebo group with matching dosage. At each examination time during the follow-up, the number of newly developed skin tumors were counted, measured, and removed. Comprehensive analysis of the whole data can be found in recent publications (e.g., Li et al., 2011; Sun and Zhao, 2013; Chiou et al., 2017).

For illustration propose, we only use a subset of `skinTumor` containing 73 patients who enrolled in the study after the age of 70 years because some methods with bootstrapping are computationally demanding for large samples. Of the 73 patients, 40 were male and 41 were in the treatment group. The average number of examination times was 8.9 in this subset of patients, with three quartiles being 7, 9, and 10. The average number of skin tumors developed for each patient in this subset throughout the study was 2.9 (median = 3). We named this subset `skiTum` and used this name in the sequel. To view the structure of panel count data, we show the data for one patient (with id 10):

```
library(spef)
data(skinTumor)
skiTum <- subset(skinTumor, age >= 70)
subset(skiTum, id == 10, select = c(id, time, count, dfmo, priorTumor))

##     id time count dfmo priorTumor
## 95  10  183     1    0         16
## 96  10  366     0    0         16
```

```
## 97  10  569     0    0         16
## 98  10  757     0    0         16
## 99  10  940     1    0         16
## 100 10 1011     0    0         16
## 101 10 1024     0    0         16
```

The patient with id 10 was followed for 1024 days from the enrollment, examined 7 times on days after enrollment as shown in variable `time`, with the corresponding number of tumors in variable `count`. This patient was assigned to the placebo group (`dfmo = 0`) and had 16 skin tumors prior to enrollment. Treatment indicator (`dfmo`) and prior tumor counts (`priorTumor`) will be used as covariates in the regression model for the tumor occurrences in this study. Following Wang and Yan (2011), we display the data in a tile plot that shows not only the panel count but also the examination times of each subject using package `ggplot2` (Wickham, 2009):

```
library(ggplot2)
ggplot(skiTum, aes(time, as.factor(id), width = 25, height = 1)) +
    geom_tile(aes(fill = count)) + theme_bw() +
    theme(axis.text.y = element_blank(), axis.ticks = element_blank()) +
    facet_grid(dfmo ~ ., scales = "free_y", as.table = FALSE,
               labeller = labeller(dfmo = function(x) paste("DFMO =", x))) +
    scale_fill_gradient(low = "grey", high = "black") +
    scale_x_continuous(breaks = seq(0, 2000, 200)) +
    labs(fill = "Count") + xlab("Time in days") + ylab("Patient")
```

[Figure 1 about here.]

Figure 1 presents the resulting tile plot. It appears that patients in the treatment group have slightly more examinations than those in the placebo group, which might indicate informative examination times.

All the models in the sequel have the same model formula specified via `PanelSurv`, which is similar to the `Surv` function in the `survival` package (Therneau, 2015). We consider models with two covariates: `dfmo` and `priorTumor`. For better interpretation of the baseline function, we center `priorTumor` by its median 3:

```
skiTum$priorTumor <- skiTum$priorTumor - 3
fm <- PanelSurv(id, time, count) ~ dfmo + priorTumor
```

The major function to fit regression models for panel count data in the `spef` package is `panelReg`, which takes the model formula as an input and returns an object of class `panelReg`.

3 Notation and Regression Models

For subject i, $i = 1, \ldots, n$, let $N_i(t)$ be counting process of recurrent events of interest. Suppose that the event counts are only observable at K_i discrete random time points, $0 = t_{i0} < t_{i1} < t_{i2} < \ldots < t_{iK_i} \leq \tau$, where t_{ij} is the jth examination time, K_i is a positive integer-valued random variable, and τ is the longest follow-up time in the data. Let G be the time grid formed by all distinctive examination times: $0 < s_1 < \cdots < s_g = \tau$, where g is the number of distinctive examination times. A subject-specific, time-independent covariate vector X_i is observed and its effect on the occurrence of the events is or primary interest. The observed data are independent and identically distributed copies of $\{t_{ij}, K_i, N_i(t_{ij}), X_i; j = 1, \ldots, K_i\}, i = 1, \ldots, n$. Let $n_{ij} = N_i(t_{ij}) - N_i(t_{ij-1})$ be the number of events in the time interval $(t_{ij-1}, t_{ij}]$ and $m_i = N_i(Y_i)$ be the total number of events during the follow-up, where $Y_i = t_{iK_i}$ is the last examination time. Additionally, there could be a censoring or follow up time C_i, which may or may not equal to the last observation time Y_i. As in recurrent event settings, the censoring time C_i's are always observed unlike in the case of standard right-censored survival data. Both the examination times and the follow-up time can potentially be informative about the event process after conditioning on the covariates.

Earlier models for recurrent event processes characterize the intensity function (Gail et al., 1980; Prentice et al., 1981; Andersen and Gill, 1982). To introduce the common models, we drop the index i for ease of notation. Let $\mathrm{d}N(t) = N\{(t+\mathrm{d}t)^-\} - N(t^-)$. The intensity function is defined as the event occurrence rate conditional on the whole event history

$$\lambda(t) = \lim_{\Delta \to 0+} \frac{1}{\Delta} \Pr[\mathrm{d}N(t) = 1 | \mathcal{H}(t^-)],$$

where $\mathcal{H}(t^-) = \{N(u) : 0 \leq u < t\}$ is the event history up to t. The Cox-type intensity model incorporates covariate X in the intensity function (Andersen and Gill, 1982)

$$\lambda(t; X) = \lambda_0(t) \exp(X^\top \beta), \tag{1}$$

where $\lambda_0(t)$ is nonnegative baseline intensity function, and β is a vector of regression coefficients for covariate vector X.

In practice, the Cox-type intensity model in Model (1) might be inadequate and difficult to verify (Lin et al., 2000). In contrast to Model (1), recent approaches characterize the rate function $r(t)$ of $N(t)$ defined by $E\{\mathrm{d}N(t)\} = r(t)\,\mathrm{d}t$ and the mean function $\mu(t) = \int_0^t r(s)\,\mathrm{d}s$ (Nelson, 1988; Pepe and Cai, 1993; Lawless and Nadeau, 1995; Lin et al., 2000). Unlike the intensity function, the rate or mean function does not completely specify the stochastic nature of $N(t)$; they are, respectively, sometimes referred to as the marginal intensity and cumulative intensity function. Covariates can be incorporated in the form of proportional rates model

$$r(t; X) = r_0(t) \exp(X^\top \beta), \tag{2}$$

for some nonnegative baseline rate function $r_0(t)$, or proportional means model

$$\mu(t; X) = \mu_0(t) \exp(X^\top \beta), \tag{3}$$

for some nondecreasing baseline mean function $\mu_0(t)$. Since we consider time-independent

covariate so far, Model (2) and Model (3) are equivalents.

A commonly used modification to Model (1) and (3) is to introduce a positive frailty variable or random effect. Specifically, conditional on a frailty Z and covariate vector X, the proportional intensity model becomes

$$\lambda(t; X, Z) = Z\lambda_0(t)\exp(X^\top\beta),$$

and the proportional means model becomes

$$\mu(t; X, Z) = Z\mu_0(t)\exp(X^\top\beta). \tag{4}$$

For identification purpose, it is often assumed that $\mathsf{E}(Z|X) = 1$. The frailty is useful in allowing over-dispersion in the count (e.g., Hua et al., 2014) or dependence between $N(\cdot)$ and the examination or censoring times (e.g., Huang et al., 2006; He et al., 2009).

The baseline intensity function $\lambda_0(t)$ and the baseline mean function $\mu_0(t)$ are often left completely unspecified and estimated nonparametrically. Since $\mu_0(t)$ and the cumulative baseline intensity $\Lambda_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s$ are nondecreasing functions, they can be specified by monotone splines (Ramsay, 1988). The motonone spline specification offers a good compromise between flexibility and computational advantage, so it has been adopted by many authors in various settings (Lu et al., 2009; Hua and Zhang, 2012; Deng et al., 2015; Hua et al., 2014; Yao et al., 2016). An implementation of monotone splines is available in `R` package `splines2` for this purpose.

A recent accelerated mean model (Xu et al., 2017; Chiou et al., 2017) has rate function

$$r(t; X, Z) = Zr_0\{t\exp(X^\top\beta)\}\exp(X^\top\beta), \tag{5}$$

where the distribution of frailty Z is unspecified beyond $\mathsf{E}(Z|X) = 1$. This model formulation is different from the Cox-type specifications and it connects to the accelerated failure time

(AFT) models in that, unconditional on Z, $\mu(t;X) = \mathsf{E}\{N(t)|X\} = \mu_0\{t\exp(X^\top\beta)\}$. The covariate effects modify the time scale of the cumulative mean function and have a direct marginal interpretation. For example, if X is a treatment indicator, then the expected number of events by time t among the treated subjects ($X = 1$) equals the expected number of events by time te^β in the control group ($X = 0$).

4 Noninformative Examination/Censoring Times

We first consider the situation where the examination times and the censoring time are noninformative for the event process. That is, conditional on the covariates, the examination/censoring times and the event process are independent. The conditional independence assumption allows one to treat the examination/censoring times as if they were fixed instead of random.

4.1 Likelihood-Based Approaches

The non-homogeneous Poisson process has been studied first, in which case the Cox-type intensity model (1) and the proportional means model (3) coincide. So we consider Model (3) only. From the independent increments of Poisson processes, the log likelihood function is

$$L(\beta,\mu_0) = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\left\{n_{ij}\log\mu_0(t_{ij}) + n_{ij}X_i^\top\beta - \mu_0(t_{ij})\exp(X_i^\top\beta)\right\}.$$

Parameter estimation of β depends on the specification of $\mu_0(t)$. If $\mu_0(t)$ is unspecified, the nonparametric maximum likelihood estimator (MLE) of $\mu_0(t)$ is the non-decreasing step function that jumps only at the times of the grid G of distinct examination times (Wellner and Zhang, 2000). The MLE of $(\beta,\mu_0(t))$, denoted by $(\hat\beta_n,\hat\mu_n(t))$, can be obtained from a computationally intensive iterative procedure (Wellner and Zhang, 2007).

To reduce the computation complexity in obtaining MLE, Lu et al. (2009) specified

$\log \mu_0(t)$ by monotone B-splines $\log \mu_0(t) = \sum_{i=1}^{\kappa} \alpha_i B_i(t)$, where $B_i(t)$, $i = 1, \ldots, \kappa$, are the B-spline basis functions with κ degrees of freedom. The degrees of freedom, κ, is typically chosen to be $\lceil g^{1/3} \rceil + 1$ where $\lceil \cdot \rceil$ is the ceiling function and g is the number of distinctive examination times as defined in Section 3. The MLE of (β, α), denoted by $(\hat{\beta}_n, \hat{\alpha}_n)$, can then be found from a constrained optimization for any given K. Lu et al. (2009) show that under certain regularity conditions $\hat{\beta}_n$ is consistent, asymptotically normal, and asymptotically as efficient as that obtained when $\mu_0(t)$ is unspecified. For the skin tumor example, this method is called by setting `method = "MLs"` in the `panelReg` function from `spef` package:

```
panelReg(fm, data = skiTum, method = "MLs", se = "Bootstrap",
         control = list(R = 50))

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "MLs", se = "Bootstrap",
##     control = list(R = 50))
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## dfmo       -0.2375    0.7886   0.3114 -0.762     0.45
## priorTumor  0.0806    1.0839   0.0187  4.319  1.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The standard errors of the regression coefficient estimates were obtained from bootstrap with 50 replicates by setting `se = "Bootstrap"` and, in `control`, `"R = 50"`. The implementation of monotone splines in the `spef` package was based on the methods proposed in Ramsay (1988). The same model can also be fit with `PCDReg.nf` function from the `PCDSpline` package (Yao and Wang, 2014). The `PCDSpline` package further allows a gamma frailty to account for within-subject dependence (Yao et al., 2016).

A less efficient but simpler approach to obtain the regression coefficient estimate is to maximize the following pseudo-likelihood based on the Poisson distribution of each $N(t_{ij})$ ignoring within-subject dependence

$$L_p(\beta, \mu_0) = \sum_{i=1}^{n} \sum_{j=1}^{K_i} N(t_{ij}) \log \mu_0(t_{ij}) + N(t_{ij}) X_i^\top \beta - \mu_0(t_{ij}) \exp(X_i^\top \beta).$$

The estimator of β with an unspecified $\mu_0(t)$ (Zhang, 2002) can be obtained by setting `method = "MPL"`:

```
(fit.MPL <- panelReg(fm, data = skiTum, method = "MPL", se = "Bootstrap",
                     control = list(R = 50)))

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "MPL", se = "Bootstrap",
##     control = list(R = 50))
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## dfmo       -0.2320    0.7929   0.3543 -0.655     0.51
## priorTumor  0.0880    1.0920   0.0173  5.095  3.5e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimator of β when $\mu_0(t)$ is specified by monotone B-splines (Lu et al., 2009) can be obtained by setting `method = "MPLs"`:

```
(fit.MPLs <- panelReg(fm, data = skiTum, method = "MPLs", se = "Bootstrap",
                      control = list(R = 50)))

##
```

```
## Call:
## panelReg(formula = fm, data = skiTum, method = "MPLs", se = "Bootstrap",
##     control = list(R = 50))
##
##               coef exp(coef) se(coef)     z Pr(>|z|)
## dfmo       -0.2402    0.7864   0.3637 -0.66     0.51
## priorTumor  0.0874    1.0913   0.0176  4.97  6.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hua et al. (2014) considered Model (4) with Z assumed to be a gamma variable with mean 1 and variance σ^2. Under the working assumption that $N(\cdot)$ is a non-homogeneous Poisson process, the full likelihood after integrating Z out has a closed-form in terms of β and μ_0. By approximating $\mu_0(t)$ with monotone splines with parameter vector α, they estimate α and β after fixing σ^2 at a method of moment estimate based on pseudolikelihood estimator from Zhang (2002) and Wellner and Zhang (2007).

The estimated baseline mean function for the aforementioned methods can be accessed from the `baseline` component in the object returned from the `panelReg` call. The `spef` package provides a utility function for its graphcial presentation through the generic function `plot`. For example, the estimated baseline mean function from `method = "MPL"` and `method = "MPLs"` can be plotted as follows:

```
plot(fit.MPLs, lwd = 1.5, main="")
plot(fit.MPL, add = TRUE, lty = 2, lwd = 1.5)
legend("topleft", c("MPL", "MPLs"), bty = "n", lty = 1:2, lwd = 1.5)
```

[Figure 2 about here.]

Figure 2 shows the overlaid estimated curves from the two methods. They are interpreted as the mean function for patients in the placebo group with 3 prior tumors. Baseline function estimates from other methods in the sequel, if available, can be a accessed similarly.

4.2 Estimating Equation Approaches

Sun and Wei (2000) allow dependence among the event process, examination time process, and the censoring time through covariates if the latter two follow a proportinal means model and a proportional hazards model, respectively. Define the examination time process $H_i(t) = \tilde{H}_i\{\min(t, C_i)\} = \sum_{j=1}^{K_i} I(t_{ij} \leq t)$. Assume that the mean function of $\tilde{H}_i(t)$ has the form

$$\mu_i^H(t) = \mu_0^H(t) \exp(X_i^\top \gamma), \tag{6}$$

where $\mu_0^H(t)$ is a completely unspecified function and γ is a regression coefficient vector. Further assume that covariate effects on the censoring time can be specified by a Cox proportional hazards model for C_i,

$$\lambda_i^C(t|X_i) = \lambda_0^C(t) \exp(X_i^\top \eta), \tag{7}$$

where $\lambda_0^C(t)$ is a completely unspecified baseline hazard function and η is a regression coefficient vector. The covariates are assumed to have been centered by theie means in the derivation of the method.

Sun and Wei (2000) proposed estimating equations by considering $\int N_i(t)\, \mathrm{d}H_i(t)$. Under the model specifications for $\mu_i^H(t)$ and $\lambda_i^C(t)$,

$$\mathsf{E}\left\{\int N_i(t)\, \mathrm{d}H_i(t)\right\} = \exp\{X_i^\top(\beta+\gamma)\} \int \mu_0(t) S_i(t)\, \mathrm{d}\mu_0^H(t),$$

where $S_i(t) = \exp\{-\int_0^t \lambda_0^C(s)\, \mathrm{d}s + X_i^\top \eta\}$. Therefore, if γ and η are known, β can be estimated

from the following estimating equation

$$\sum_{i=1}^{n} X_i \exp\{-X_i^\top(\beta+\gamma)\} \int \frac{N_i(t)}{S_i(t)} \mathrm{d}H_i(t) = 0. \tag{8}$$

The unknown quantities in the equation can be replaced with their estimates: γ can be estimated from estimating equations for proportional rates models (Lawless and Nadeau, 1995); η can be estimated from partial score equations (Kalbfleisch and Prentice, 2011); and the baseline hazard $\lambda_0^C(t)$ can be estimated as in a standard survival analysis. Sun and Wei (2000) established the consistency and asymptotic normality of the resulting estimator requiring the correct specification of the models for the examination times and the censoring time. The estimator of β can be obtained by setting `method = "EE.SWc"`:

```
panelReg(fm, data = skiTum, method = "EE.SWc", se = "Bootstrap",
         control = list(R = 50))

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "EE.SWc", se = "Bootstrap",
##     control = list(R = 50))
##
##              coef exp(coef) se(coef)    z Pr(>|z|)
## dfmo        0.842     2.320    0.622 1.35     0.18
## priorTumor 0.110     1.116    0.028 3.92    9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When $\eta = 0$, in which case the censoring time does not depend on covariates, the estimator can be obtained by setting `method = "EE.SWb"`:

```
panelReg(fm, data = skiTum, method = "EE.SWb", se = "Bootstrap",
         control = list(R = 50))

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "EE.SWb", se = "Bootstrap",
##     control = list(R = 50))
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## dfmo       -0.0830    0.9203   0.3306 -0.251      0.8
## priorTumor  0.1333    1.1426   0.0235  5.674  1.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An even simpler version of Sun and Wei (2000) assuming independent examination and censoring by setting $\gamma = \eta = 0$ can be obtained by setting `method = "EE.SWa"`.

Hu et al. (2003) proposed a more efficient estimating equation that extends the method of Lawless and Nadeau (1995) for recurrent event analysis. Define $h_i(t) = H_i(t) - H_i(t^-)$ for each i so that $h_i(t) = 1$ if t is an examination time of subject i and $h_i(t) = 0$ otherwise. Assume that $E\{h_i(t)\} > 0$ for each $t \in \mathcal{T}$ where $\mathcal{T} \subset (0, \tau]$ is the collection of all observed examination times on a grid. Conditioning on the examination times, Hu et al. (2003) proposed a natural estimating equation for β

$$\sum_{i=1}^{n} \sum_{j=1}^{K_i} w(t_{ij}) \left\{ X_i - \frac{\sum_{k=1}^{n} I(C_k \geq t_{ij}) X_k \exp(X_k^\top \beta) o_k(t_{ij})}{\sum_{k=1}^{n} I(C_k \geq t_{ij}) \exp(X_k^\top \beta) o_k(t_{ij})} \right\} n_{ij} = 0, \tag{9}$$

where $w(\cdot)$ is a known, possibly data dependent weight function and $o_k(t)$ indicates whether subject k has an observation at time t. The estimating equation (9) was constructed under the assumption that there is more than one subject with the same examination time. Thus,

this method cannot be applied to scenarios where all examination times are distinct, which implies $o_k(t_ij) = 1$ when $k = i$ and 0 otherwise. Solution to the conditional estimating equations (9) with $w(t) = 1$ can be obtained by setting `method = "EE.HSWc"`:

```
panelReg(fm, data = skiTum, method = "EE.HSWc", se = "Bootstrap",
         control = list(R = 50))

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "EE.HSWc", se = "Bootstrap",
##     control = list(R = 50))
##
##              coef exp(coef) se(coef)     z Pr(>|z|)
## dfmo       0.0483    1.0494   0.2322 0.208    0.840
## priorTumor 0.0524    1.0538   0.0236 2.221    0.026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To allow covariate effects on the examination times in a proportional means Model (6), Hu et al. (2003) proposed an estimating equation unconditional on the examination times

$$\sum_{i=1}^{n}\sum_{j=1}^{K_i} w(t_{ij})\left[X_i - \frac{\sum_{k=1}^{n} I(C_k \geq t_{ij}) X_i \exp\{X_k^\top(\beta+\gamma)\}}{\sum_{k=1}^{n} I(C_k \geq t_{ij}) \exp\{X_k^\top(\beta+\gamma)\}}\right] n_{ij} = 0, \tag{10}$$

where γ needs to be replaced with an estimate as in solving (8). In contrast to (8), this equation does not require model specification of the censoring time. See Section 5.4.3 of Sun and Zhao (2013) for more discussion on comparison of the estimating equation approaches. Solution to the marginal estimating equations (10) can be obtained by setting `method = "EE.HSWm"`:

```
panelReg(fm, data = skiTum, method = "EE.HSWm", se = "Bootstrap",
         control = list(R = 50))

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "EE.HSWm", se = "Bootstrap",
##     control = list(R = 50))
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## dfmo       -0.1652    0.8478   0.3130 -0.528   0.6000
## priorTumor  0.0539    1.0554   0.0167  3.223   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since panel counts are similar to longitudinal data, Hua and Zhang (2012) applied generalized estimating equations (GEE) (Liang and Zeger, 1986) to marginal Model (3) with $\log \lambda(t)$ approximated by monotone splines with parameters α as in Lu et al. (2009). The panel counts from subject i form a vector $\mathcal{N}_i = \{N_i(t_{i1}), \ldots, N_i(t_{i,K_i})\}^\top$, with mean vector $\boldsymbol{\mu}_i = \{\mu(t_{i1}; X_i), \ldots, \mu(t_{iK_i}; X_i)\}^\top$. The GEE has the form

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \theta} V_i^{-1} (\mathcal{N}_i - \boldsymbol{\mu}_i) = 0, \tag{11}$$

where $\theta^\top = (\beta^\top, \alpha^\top)$, and V_i is a $K_i \times K_i$ working covariance matrix of $\mathcal{N}_i$. Hua and Zhang (2012) used a two-iterative algorithm to solve for θ. First, a Newton–Raphson update is applied to solve (11); second, the estimate of α is projected to a legitimate space via quadratic programming such that the resulting splines is monotone nondecreasing. Flexible choices of the working covariance matrix V_i's can lead to higher efficiency in estimation and robustness to overdispersion.

5 Informative Examination/Censoring Times

5.1 Frailty Methods

One way to allow informative examination times after conditioning on covariates is to introduce a frailty, or random effect that is shared by both the recurrent event process and the examination time process. Huang et al. (2006) considered Model (4), which allows the examination times to be associated with the event process through the frailty after conditioning on the covariates. The approach of Huang et al. (2006) is especially appealing in that there is no need to specify the distribution of the frailty, or models for the examination process and the censoring time. The estimation procedure takes advantage of the fact that, conditional on $\{Z_i, X_i, K_i, Y_i\}$, the unobserved K_i examination times are order statistics of independent and identically distributed random variables with distribution function

$$F_i(t) = \frac{\mu(t; X_i, Z_i)}{\mu(Y_i; X_i, Z_i)} = \frac{\mu_0(t)}{\mu_0(Y_i)}.$$

This formulation suggests that the estimation of $F(t)$ does not involve X_i and Z_i. Let $\Phi(t) = \mu_0(t)/\mu_0(\tau)$, where τ is still the longest follow-up time. A nonparametric estimator of $F(t)$ is obtained by maximizing

$$\prod_{i=1}^{n} \prod_{j=1}^{K_i} \left[\frac{\Phi(T_{i,j}) - \Phi(T_{i,j-1})}{\Phi(Y_i)} \right]^{n_{ij}},$$

which is mathematically equivalent to the likelihood constructed from a set of independently interval-censored and right-truncated data. Therefore, the maximization of the likelihood can be implemented by the Turnbull's self-consistency algorithm (Turnbull, 1976). When computational performance is of concern, the squared extrapolation method of Varadhan and Roland (2008) can be adopted to accelerate the maximization. Then $\Lambda(\tau)$ and β are

obtained from solving

$$n^{-1}\sum_{i=1}^{n} w_i \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left[m_i \Phi(Y_i)^{-1} - \mu_0(\tau)\exp(X_i^\top \beta) \right] = 0,$$

where w_i is a weight function and $\Phi(\cdot)$ is replaced with its estimate. This approach with $w_i = 1$ is requested by setting `method = "HWZ"`:

```
panelReg(fm, data = skiTum, method = "HWZ", se = "Bootstrap",
                    control = list(R = 50))

## [1] "Warning: SE based on 32 converged bootstrap samples"
##
## Call:
## panelReg(formula = fm, data = skiTum, method = "HWZ", se = "Bootstrap",
##     control = list(R = 50))
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## dfmo       -0.3069    0.7358   0.3117 -0.984     0.32
## priorTumor  0.0771    1.0801   0.0198  3.900  9.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A warning message indicates that not all of the 50 bootstrap converged. The reported bootstrap standard errors are based on those that converged.

Alternative approaches specify models for the examination times and the censoring time. Extending the estimation strategies of Sun and Wei (2000), Sun et al. (2007) investigated a similar semiparametric model with Z_i^α in place of Z_i in Model (4), where Z_i is an unobserved multiplicative frailty introduced into Model (6) for the examination times. He et al.

(2009) used two frailties to introduce dependence among the three models (3), (6), and (7) beyond covariate effects. Specifically, one frailty enters all three models while the other enters Model (3) and (7). Model parameters are estimated through a three-step estimation procedure. This method imposes a distributional assumption on the underlying random effect and requires the examination process to be a nonhomogeneous Poisson process, which is needed in an EM algorithm in handling the parameters and frailties in the model for the examination process. Zhao et al. (2013) proposed a more general model which replaces Z in Model (4) with $f(Z)$, where Z is a multiplicative frailty introduced into Model (6) as in Sun et al. (2007), and f is a positive, completely unspecified link function. They relaxed the Poisson assumption for the examination process. The methods of He et al. (2009) and Zhao et al. (2013) are presented in detail in Sun and Zhao (2013, Sections 6.2–6.3).

5.2 Augmented Estimating Equations

Wang et al. (2013) approached the problem by treating the unobserved event times as missing data. Consider the time grid G in Section 3, let $\mathbb{N}_{ij} = N_i(s_j) - N_i(s_{j-1})$ be the number of events occurred in $(s_{j-1}, s_j]$. Only summations of $\mathbb{N}_{ij}$'s over those subintervals whose union coincides with an observation window are observed. Regardless of the examination times, if $\mathbb{N}_{ij}$'s were observed, under conditional independent censoring, Model (3) suggests a set of complete-data estimating equations:

$$\sum_{i=1}^{n} \Big[\mathbb{N}_{ij} - \lambda_j \exp(X_i^\top \beta)\Big] r_{ij} = 0, \quad j = 1, \ldots, G,$$
$$\sum_{i=1}^{n}\sum_{j=1}^{G} \Big[\mathbb{N}_{ij} - \lambda_j \exp(X_i^\top \beta)\Big] X_i r_{ij} = 0,$$

where $\lambda_j = \Lambda(s_j) - \Lambda(s_{j-1})$ is the baseline mean number of events occurring in interval $(s_{j-1}, s_j]$, and $r_{ij} = I(s_j \le C_i)$ is the at-risk indicator. The model parameters are estimated by an Expectation-Solving (ES) algorithm (Elashoff and Ryan, 2004), an analog of the EM

algorithm for estimating equations without specifying the full likelihood. The algorithm iterates between imputing the values of $\mathbb{N}_{ij}$'s and solving the conditional expected version of the complete-data estimating equations given the observed data. This method is called by setting `method = "AEE"`:

```
panelReg(fm, data = skiTum, method = "AEE", se = "Bootstrap",
         control = list(R = 50))

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "AEE", se = "Bootstrap",
##     control = list(R = 50))
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## dfmo       -0.2624    0.7692   0.3069 -0.855     0.39
## priorTumor  0.0805    1.0839   0.0193  4.181  2.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the case of informative censoring, the number of events between the last examination time Y_i and τ is also treated as missing and imputed using a working model; see (Wang et al., 2013) for more details. This method is requested by setting `method = "AEEX"`:

```
panelReg(fm, data = skiTum, method = "AEEX", se = "Bootstrap",
         control = list(R = 50))

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "AEEX", se = "Bootstrap",
```

```
##     control = list(R = 50))
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## dfmo       -0.2953    0.7443   0.3011 -0.981     0.33
## priorTumor  0.0761    1.0791   0.0191  3.979  6.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.3 Accelerated Mean Model

Chiou et al. (2017) estimated the parameters of the accelerated mean model (5) by a profile estimating equation approach. Specifically, consider the transformed times $t_{ij}^*(\beta) = t_{ij}\exp(X_i^\top\beta)$ and censoring time $Y_i^*(\beta) = Y_i\exp(X_i^\top\beta)$, $i = 1,\ldots,n$. Conditional on (Z_i, X_i, K_i, Y_i), the unobserved K_i examination times on the transformed scale $t_{ij}^*(\beta)$ are order statistics of independent and identically distributed random variables with distribution function $\mu_0(t)/\mu_0(Y_i^*(\beta))$. Let $\Phi(t) = \mu_0(t)/\mu_0(\tau_\beta)$, where $\tau_\beta = \tau\sup_i\exp(X_i^\top\beta)$. For given β, Φ can be estimated with the same method of Huang et al. (2006) except that the estimate depends on β. Define $\hat{\Phi}_n(t;\beta)$ as the resulting estimator. Then, β is estimated by solving the estimating equation

$$\sum_{i=1}^{n} X_i\left[m_i\hat{\Phi}_n^{-1}\{Y_j^*(\beta);\beta\} - \frac{1}{n}\sum_{j=1}^{n} m_j\hat{\Phi}_n^{-1}\{Y_j^*(\beta);\beta\}\right] = 0.$$

In our implementation, this equation is solved with a gradient-free spectral method (Barzilai and Borwein, 1988; La Cruz et al., 2006). The accelerated mean model is called by setting `method = "AMM"`. Since fitting this model is much more computing intensive than other methods, we timed this call:

```
system.time(fit.AMM <- panelReg(fm, data = skiTum, method = "AMM",
                                se = "smBootstrap", control = list(R = 50)))

##     user   system  elapsed
## 2484.372   25.064 2508.341

fit.AMM

##
## Call:
## panelReg(formula = fm, data = skiTum, method = "AMM", se = "smBootstrap",
##     control = list(R = 50))
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## dfmo       -0.0650    0.9370   0.2211 -0.294   0.7700
## priorTumor  0.0906    1.0948   0.0344  2.630   0.0085 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The standard errors was obtained from a smoothed bootstrap procedure proposed in Chiou et al. (2017) by setting `se = "smBootstrap"`. The standard bootstrap procedure to obtain the standard errors is still available by setting `se = "Bootstrap"`.

6 Simulation

We extended the simulation studies in Huang et al. (2006) and Wang et al. (2013) to provide a thorough comparison among the estimators discussed in this paper. Since the regression coefficient in the accelerated mean model is interpreted differently than those in the proportional means model, we focus here on the comparison of the regression coefficient estimates

in the proportional means model. We generated recurrent events from a Poisson process with mean model specified in Model (4) for $t \in [0, \tau]$ with $\tau = 10$. The baseline mean function was set to be $\mu_0(t) = 2t$. Two mutually independent covariates, X_{i1} and X_{i2} were generated from the Bernoulli distribution with rate 0.5 and the standard normal distribution, respectively. The regression coefficients were set to be $\beta = (\beta_1, \beta_2)^\top = (0.5, 1)^\top$. The subject-specific frailty Z_i had three configurations: 1) fixed at constant 1; 2) generated from a gamma distribution with mean 1 and variance 0.5; or 3) generated from a uniform distribution over $[0, 2]$. The sample size n had two levels, 100 and 200.

We considered three scenarios depending on how examination times associate with recurrent events:

- Scenario 1: examination times and recurrent events are independent. The number of examinations, K_i, was generated from a district uniform distribution on $\{1, \ldots, 6\}$ and the distinct examination times $t_{i1}, \ldots, t_{iK_i}$, were the order statistics of K_i independent and identically distributed uniform distribution over $[0, 10]$.

- Scenario 2: examination times and recurrent events are independent conditioning on the covariates. If $X_{i1}X_{i2} > 0$, then the number of examinations, K_i, was generated from a district uniform distribution on $\{1, \ldots, 8\}$ and the distinct examination times were the order statistics of K_i independent and identically distributed exponential distribution with mean 2; otherwise, K_i and $t_{i1}, \ldots, t_{iK_i}$ were generated in the same fashion as in Scenario 1.

- Scenario 3: examination times are informative about the recurrent events after conditioning on the covariates. If $X_{i1}X_{i2} > 0$ and $Z_i > 1$, then K_i and $t_{i1}, \ldots, t_{iK_i}$ were generated as in the case of $X_{i1}X_{i2} > 0$ in Scenario 2; otherwise, they were generated in the same fashion as in Scenario 1.

Under the study designs, Scenario 3 reduces to Scenario 1 when Z_i was fixed at 1 but the two scenarios are different otherwise. In Scenario 2 when the examination times and recurrent

events are independent conditioning on covariate, subjects with $X_{i1} = 1$ and $X_{i2} > 0$ are more likely to be examined more frequently. In Scenario 3 when the examination times are informative about the recurrent events, the design implies a positive association between the underlying recurrent event process and the examination time process; subjects with $X_{i1} = 1$, $X_{i2} > 0$ and $Z_i > 1$ have a higher event rate and tend to be examined more frequently. Since examination times were generated from continuous probability distributions for all three scenarios, `EE.HSWc` estimator was excluded from the study as the `EE.HSWc` estimator is not applicable to scenarios when there are no ties in examination times. The standard errors were estimated using the standard bootstrap procedure by setting `se = "Bootstrap"` with `R = 200` bootstrap samples. For each configuration, 1000 datasets were generated and analyzed. The timing results were obtained on a Linux machine with 2 GHz CPU.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

Table 1 presents the results under Scenario 1. All estimators are virtually unbiased. The empirical standard errors and the estimated standard errors from the standard bootstrap procedure agree closely for all estimators, suggesting that the bootstrap procedure provides valid inference. The estimating-equation-based estimators were fastest to compute, but they appear to have higher standard errors than other estimators. All estimators had higher standard errors in the case of gamma frailty, which has high variance than the case of uniform frailty. The empirical coverage percentages are mostly reasonably close to the nominal level of 95%, with a closer agreement with larger sample size (results for $n = 200$ not shown).

Table 2 summarizes the results under Scenario 2. No estimator except those based on estimating equations show noticeable bias. The substantial bias and, consequently, the low coverage rate of the confidence intervals from the estimating equation approaches are due to

their misspecification of the examination time process. The other estimators do not require specification of the examination time process, which might not be of primary interest. They appear to have similar results regarding bias and standard errors. Among them, the `AEE` estimator is the fastest and has the smallest standard errors, albeit the advantage in standard error is small.

Table 3 summarizes the results under Scenario 3. Under this setting of informative examination times, the only unbiased estimators appear to be the `HWZ` estimator and the `AEEX` estimator, with comparable standard errors. This is explained by the rationals on which they are derived. Their coverage rates of the confidence intervals were a bit lower than the nominal rate for the continuous regression coefficient, and the agreement improves as the sample size becomes $n = 200$ (results not shown). The `AEEX` estimator is twice as fast as the `HWZ` estimator.

7 Discussion

Nonparametric estimation of the mean cumulative function or mean rate function (e.g., Sun and Zhao, 2013, Chapter 3–4) plays an important role in many methods for semiparametric regression models. Estimation of semiparametric approaches often involves an alternate iteration between updating the estimate of β and updating the estimate of $\mu_0(t)$, the latter of which is often based on nonparametric estimation given β. For example, the MLE and MPLE of Wellner and Zhang (2007) are based on the one-sample nonparametric MLE and MPLE of Wellner and Zhang (2000). The method of Huang et al. (2006) does not require alternate iteration in estimating the parameters of Model (4) because of the special structure of this model. When the idea is adapted to the accelerated mean Model (5) of Chiou et al. (2017), nonparametric estimation given the parametric part becomes necessary in an alternate iteration procedure. Some nonparametric estimation methods with self-consistent algorithm (Hu et al., 2009a,b) have not been, but could be combined with a parametric

estimation procedure to form a semiparametric approach. For methods implemented in the `spef` package, nonparametric estimation can be requested by setting right hand of the model formula to be intercept only; for example, `PanelSurv(id, time, count) ~ 1`. In addition, the baseline function estimates can be plotted with the generic `plot` function as illustrated in Section 4 and 5.

The scope of this review is limited to available implementations of semiparametric regression models with time-independent covariates. A wide range of topics on panel count data have been studied, many of which have been reviewed by Sun and Zhao (2013). Examples are nonparametric comparison (Zhang, 2006), semiparametric transformation models (Li et al., 2010), multivariate panel count data analysis (He et al., 2008; Li et al., 2011; Zhang et al., 2013; Li et al., 2015), measurement errors (Kim, 2007), mixed recurrent event and panel count data analysis (Zhu et al., 2013), varying-coefficient models (He et al., 2017), incorporation of observation history in regression (Li et al., 2010; Deng et al., 2015), and so on. Some topics are worth investigating; for example, adapting the semiparametric regression with time-dependent covariates for recurrent event data (Huang et al., 2010) to panel count data. The unavailability of cutting-edge methods to practitioners calls for user-friendly, quality controlled software implementation as reproductive statistical research gains sharpened focus.

References

Andersen, P. K. and R. D. Gill (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics 10*, 1100–1120.

Bailey, H. H., K. Kim, A. K. Verma, K. Sielaff, P. O. Larson, S. Snow, T. Lenaghan, J. L. Viner, J. Douglas, and N. E. Dreckschmidt (2010). A randomized, double-blind, placebo-controlled phase 3 skin cancer prevention study of α-difluoromethylornithine in subjects with previous history of skin cancer. *Cancer Prevention Research 3*(1), 35–47.

Barzilai, J. and J. M. Borwein (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis 8*(1), 141–148.

Chiou, S., G. Xu, J. Yan, and C.-Y. Huang (2017). Semiparametric estimation of the accelerated mean model with panel count data under informative examination times. *Biometrics*. To appear.

Chiou, S. H., X. Wang, and J. Yan (2017). *spef: Semiparametric Estimating Functions*. R package version 1.0-6.

Croissant, Y., G. Millo, et al. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software 27*(2), 1–43.

Deng, S., L. Liu, and X. Zhao (2015). Monotone spline-based least squares estimation for panel count data with informative observation times. *Biometrical Journal 57*(5), 743–765.

Elashoff, M. and L. Ryan (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics 13*(1), 48–65.

Gail, M. H., T. J. Santner, and C. C. Brown (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics 36*, 255–266.

He, X., X. Feng, X. Tong, and X. Zhao (2017). Semiparametric partially linear varying coefficient models with panel count data. *Lifetime Data Analysis 23*(3), 439–466.

He, X., X. Tong, and J. Sun (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Analysis 15*(2), 177–196.

He, X., X. Tong, J. Sun, and R. J. Cook (2008). Regression analysis of multivariate panel count data. *Biostatistics 9*(2), 234–248.

Hsiao, C. (2014). *Analysis of Panel Data*. Cambridge university press.

Hu, X., J. Sun, and L.-J. Wei (2003). Regression parameter estimation from panel counts. *Scandinavian Journal of Statistics 30*(1), 25–43.

Hu, X. J., S. W. Lagakos, and R. A. Lockhart (2009a). Generalized least squares estimation of the mean function of a counting process based on panel counts. *Statistica Sinica 19*, 561.

Hu, X. J., S. W. Lagakos, and R. A. Lockhart (2009b). Marginal analysis of panel counts through estimating functions. *Biometrika 96*(2), 445–456.

Hua, L. and Y. Zhang (2012). Spline-based semiparametric projected generalized estimating equation method for panel count data. *Biostatistics 13*(3), 440–454.

Hua, L., Y. Zhang, and W. Tu (2014). A spline-based semiparametric sieve likelihood method for over-dispersed panel count data. *Canadian Journal of Statistics 42*(2), 217–245.

Huang, C.-Y., J. Qin, and M.-C. Wang (2010). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics 66*(1), 39–49.

Huang, C.-Y., M.-C. Wang, and Y. Zhang (2006). Analysing panel count data with informative observation times. *Biometrika 93*(4), 763–775.

Kalbfleisch, J. and J. F. Lawless (1985). The analysis of panel data under a markov assumption. *Journal of the American Statistical Association 80*(392), 863–871.

Kalbfleisch, J. D. and R. L. Prentice (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kim, Y.-J. (2007). Analysis of panel count data with measurement errors in the covariates. *Journal of Statistical computation and Simulation 77*(2), 109–117.

La Cruz, W., J. Martínez, and M. Raydan (2006). Spectral residual method without gradient information for solving large-scale nonlinear systems of equations. *Mathematics of Computation 75*(255), 1429–1448.

Lawless, J. F. and C. Nadeau (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics 37*(2), 158–168.

Li, N., D.-H. Park, J. Sun, and K. Kim (2011). Semiparametric transformation models for multivariate panel count data with dependent observation process. *Canadian Journal of Statistics 39*(3), 458–474.

Li, N., L. Sun, and J. Sun (2010). Semiparametric transformation models for panel count data with dependent observation process. *Statistics in Biosciences 2*(2), 191–210.

Li, Y., X. He, H. Wang, B. Zhang, and J. Sun (2015). Semiparametric regression of multivariate panel count data with informative observation times. *Journal of Multivariate Analysis 140*, 209–219.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

Lin, D., L. Wei, I. Yang, and Z. Ying (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B 62*(4), 711–730.

Lu, M., Y. Zhang, and J. Huang (2009). Semiparametric estimation methods for panel count data using monotone b-splines. *Journal of the American Statistical Association 104*(487), 1060–1070.

Ma, L. and R. Sundaram (2017). Analysis of gap times based on panel count data with informative observation times and unknown start time. *Journal of the American Statistical Association*. In press.

Nelson, W. (1988). Graphical analysis of system repair data. *Journal of Quality Technology 20*(1), 24–35.

Pepe, M. S. and J. Cai (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association 88*(423), 811–820.

Prentice, R. L., B. J. Williams, and A. V. Peterson (1981). On the regression analysis of multivariate failure time data. *Biometrika 68*(2), 373–379.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science 3*, 425–441.

Riphahn, R. T., A. Wambach, and A. Million (2003). Incentive effects in the demand for health care: A bivariate panel count data estimation. *Journal of Applied Econometrics 18*(4), 387–405.

Sun, J., X. Tong, and X. He (2007). Regression analysis of panel count data with dependent observation times. *Biometrics 63*(4), 1053–1059.

Sun, J. and L. J. Wei (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society: Series B 62*(2), 293–302.

Sun, J. and X. Zhao (2013). *Statistical Analysis of Panel Count Data*. New York: Springer.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B 38*(3), 290–295.

Varadhan, R. and C. Roland (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics 35*(2), 335–353.

Wang, X., S. Ma, and J. Yan (2013). Augmented estimating equations for semiparametric panel count regression with informative observation times and censoring time. *Statistica Sinica 23*, 359–381.

Wang, X. and J. Yan (2011). Fitting semiparametric regressions for panel count survival data with an R package spef. *Computer Methods and Programs in Biomedicine 104*(2), 278–285.

Wellner, J. A. and Y. Zhang (2000). Two estimators of the mean of a counting process with panel count data. *Annals of Statistics 28*, 779–814.

Wellner, J. A. and Y. Zhang (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics 35*(5), 2106–2142.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Xu, G., S. H. Chiou, C.-Y. Huang, M.-C. Wang, and J. Yan (2017). Joint scale-change models for recurrent events and failure time. *Journal of the American Statistical Association 112*, 794–805.

Yao, B. and L. Wang (2014). *PCDSpline: Semiparametric regression analysis of panel count data using monotone splines*. R package version 1.0.

Yao, B., L. Wang, and X. He (2016). Semiparametric regression analysis of panel count data allowing for within-subject correlation. *Computational Statistics & Data Analysis 97*, 47–59.

Zhang, H., H. Zhao, J. Sun, D. Wang, and K. Kim (2013). Regression analysis of multivariate panel count data with an informative observation process. *Journal of Multivariate Analysis 119*, 71–80.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika 89*(1), 39–48.

Zhang, Y. (2006). Nonparametric k-sample tests with panel count data. *Biometrika 93*(4), 777–790.

Zhao, X., X. Tong, and J. Sun (2013). Robust estimation for panel count data with informative observation times. *Computational Statistics & Data Analysis 57*(1), 33–40.

Zhu, L., X. Tong, H. Zhao, J. Sun, D. K. Srivastava, W. Leisenring, and L. L. Robison (2013). Statistical analysis of mixed recurrent event data with application to cancer survivor study. *Statistics in Medicine 32*(11), 1954–1963.

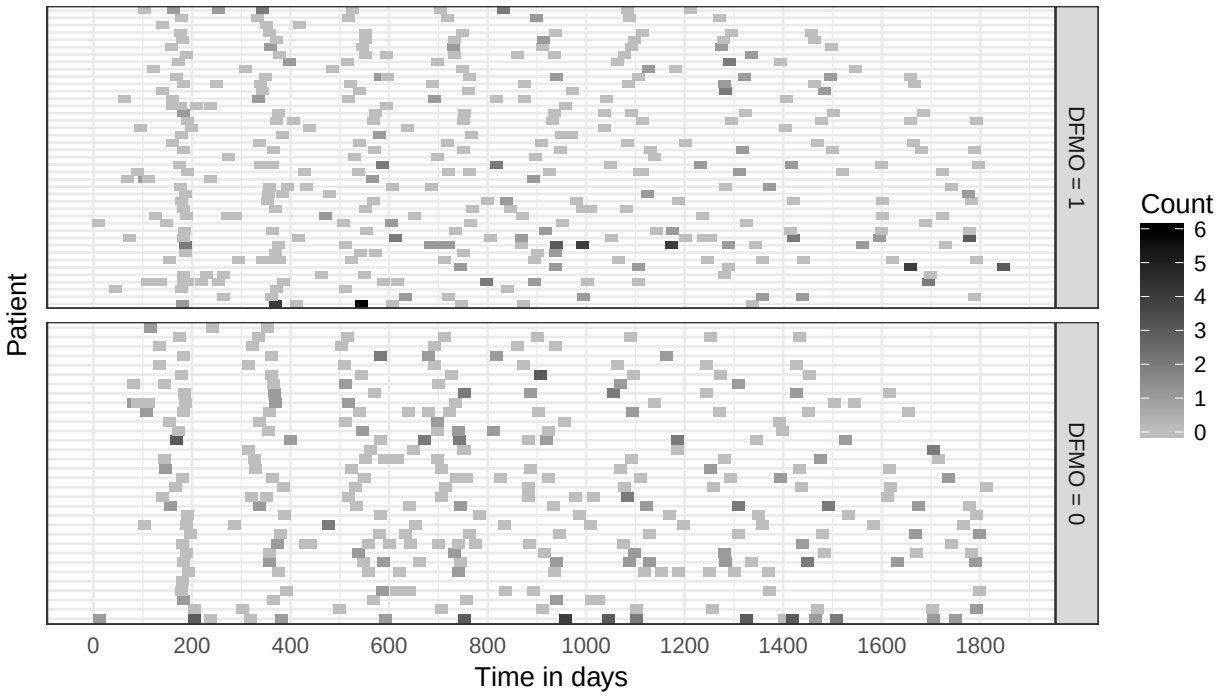

Figure 1: Tile plot of the skin tumor data. Each tile represents an examination time. Darker grays mean larger number of tumor since the last visit.

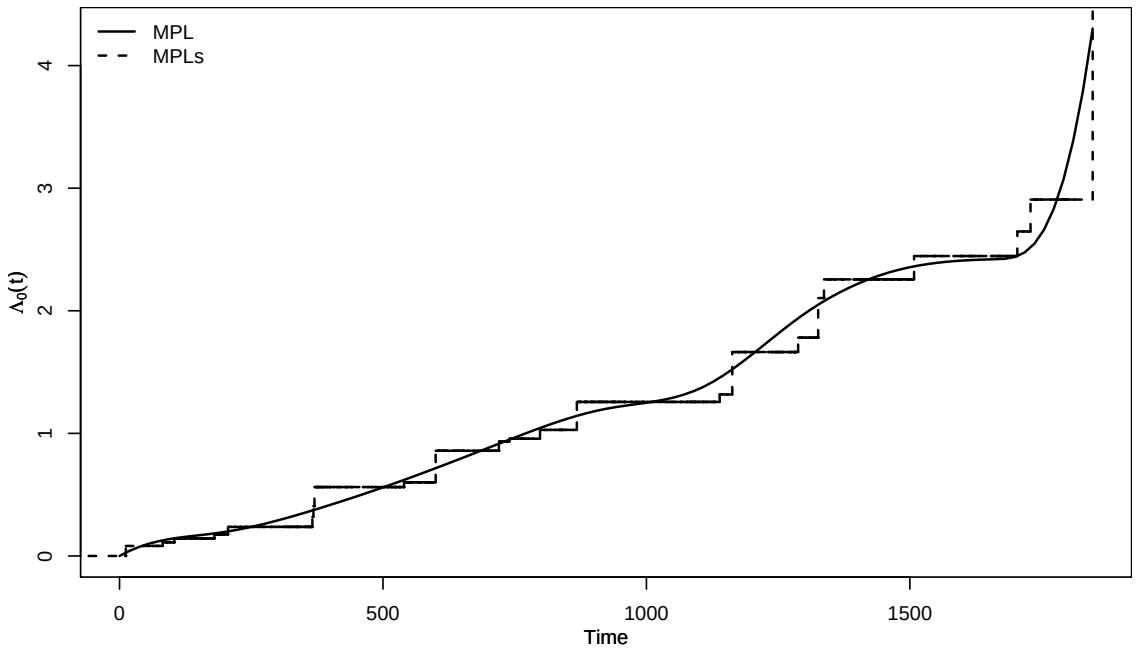

Figure 2: Estimated baseline mean function from the MPL method and the MPLs method.

Table 1: Simulation results for Scenario 1 where the examination times and the recurrent events are independent with $n = 100$. Column bias is the average bias; ESE is the empirical standard error; ASE is the average standard error based on the standard bootstrap; CP is the empirical coverage probability (%); time is the average time in seconds used in both point estimation and bootstrap variance estimation.

	bias		ESE		ASE		CP (%)		Time
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	
					$Z = 1$				
MLs	−0.001	0.001	0.037	0.019	0.038	0.021	95.3	96.9	588.7
MPL	−0.001	0.002	0.043	0.023	0.043	0.024	94.9	96.5	237.9
MPLs	−0.001	0.001	0.042	0.023	0.044	0.023	96.3	96.4	543.9
EE.SWc	−0.007	0.005	0.205	0.101	0.202	0.103	96.2	96.6	72.1
EE.SWb	0.005	0.006	0.149	0.087	0.159	0.087	96.1	96.1	47.7
EE.SWa	0.005	0.007	0.129	0.078	0.137	0.076	96.4	95.2	3.4
EE.HSWm	−0.005	0.011	0.241	0.135	0.242	0.128	95.2	93.8	65.6
HWZ	−0.001	0.001	0.046	0.022	0.046	0.023	94.0	93.8	1227.9
AEE	−0.001	0.002	0.037	0.019	0.039	0.021	95.7	96.8	176.3
AEEX	−0.002	−0.002	0.044	0.021	0.046	0.024	95.1	95.8	375.4
					$Z \sim$ gamma distribution				
MLs	0.007	−0.007	0.206	0.126	0.195	0.107	94.7	90.3	676.0
MPL	0.010	−0.007	0.215	0.127	0.198	0.107	93.8	90.4	264.8
MPLs	0.009	−0.007	0.216	0.129	0.202	0.110	93.8	90.3	578.4
EE.SWc	−0.002	0.010	0.310	0.148	0.297	0.148	94.0	95.9	66.3
EE.SWb	0.012	0.007	0.216	0.113	0.227	0.121	95.5	96.0	43.7
EE.SWa	0.004	0.007	0.205	0.108	0.210	0.113	94.9	96.4	3.0
EE.HSWm	0.013	−0.010	0.304	0.179	0.310	0.166	95.9	92.8	60.9
HWZ	0.007	−0.007	0.201	0.124	0.190	0.113	93.3	91.3	1053.2
AEE	0.007	−0.007	0.205	0.125	0.194	0.110	94.8	91.2	237.1
AEEX	−0.005	−0.011	0.200	0.122	0.192	0.112	94.6	91.5	362.5
					$Z \sim$ uniform distribution				
MLs	−0.008	−0.005	0.177	0.106	0.171	0.096	94.1	90.6	674.3
MPL	−0.009	−0.008	0.183	0.110	0.175	0.097	94.4	90.8	266.5
MPLs	−0.008	−0.007	0.187	0.118	0.179	0.099	95.2	90.2	581.7
EE.SWc	−0.006	0.004	0.305	0.139	0.274	0.139	92.4	95.6	66.8
EE.SWb	−0.007	0.001	0.195	0.111	0.210	0.112	96.3	95.1	44.1
EE.SWa	−0.014	0.002	0.174	0.105	0.194	0.103	97.0	94.9	3.1
EE.HSWm	0.002	0.001	0.308	0.176	0.297	0.163	94.7	93.4	61.3
HWZ	−0.010	−0.005	0.177	0.109	0.164	0.098	92.6	90.1	1070.1
AEE	−0.009	−0.005	0.176	0.111	0.169	0.095	93.7	90.4	235.0
AEEX	−0.013	−0.009	0.173	0.107	0.168	0.095	94.6	90.1	366.7

Table 2: Simulation results for Scenario 2 where the examination times and the recurrent events are conditionally independent given covariates with $n = 100$. Column bias is the average bias; ESE is the empirical standard error; ASE is the average standard error based on the standard bootstrap; CP is the empirical coverage probability (%); time is the average time in seconds used in both point estimation and bootstrap variance estimation.

	bias		ESE		ASE		CP (%)		Time
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	
					$Z = 1$				
MLs	−0.001	0.001	0.043	0.022	0.045	0.023	96.0	96.4	600.9
MPL	−0.002	0.002	0.052	0.026	0.053	0.028	95.1	96.7	243.4
MPLs	−0.004	0.001	0.050	0.025	0.052	0.027	95.8	96.5	599.5
EE.SWc	−0.229	−0.193	0.372	0.164	0.362	0.158	79.1	68.5	65.8
EE.SWb	−0.833	−0.347	0.195	0.097	0.197	0.105	1.7	8.9	43.5
EE.SWa	−0.338	−0.139	0.147	0.080	0.149	0.081	39.7	59.6	3.1
EE.HSWm	−1.527	−0.399	0.291	0.164	0.292	0.158	0.0	29.8	60.8
HWZ	−0.008	0.000	0.060	0.028	0.059	0.029	94.6	95.7	1120.5
AEE	−0.001	0.001	0.043	0.022	0.045	0.024	95.2	96.5	236.7
AEEX	−0.020	−0.007	0.055	0.027	0.056	0.029	94.3	96.8	474.5
					$Z \sim$ gamma distribution				
MLs	0.010	−0.007	0.202	0.119	0.192	0.104	94.4	90.9	728.1
MPL	0.025	−0.002	0.200	0.122	0.195	0.105	94.5	90.9	282.0
MPLs	0.004	−0.006	0.203	0.124	0.199	0.108	94.7	90.5	684.1
EE.SWc	−0.277	−0.202	0.464	0.207	0.405	0.187	79.3	71.4	66.9
EE.SWb	−0.828	−0.346	0.247	0.127	0.256	0.134	10.5	27.1	43.9
EE.SWa	−0.333	−0.139	0.216	0.110	0.221	0.115	68.1	77.5	3.1
EE.HSWm	−1.500	−0.409	0.349	0.207	0.340	0.198	1.5	40.4	61.4
HWZ	0.010	−0.006	0.212	0.128	0.198	0.115	92.9	91.7	1081.3
AEE	0.014	−0.006	0.199	0.119	0.190	0.103	94.4	90.9	341.3
AEEX	−0.007	−0.016	0.203	0.123	0.192	0.108	94.8	90.8	512.6
					$Z \sim$ gamma distribution				
MLs	−0.007	0.001	0.177	0.109	0.168	0.094	94.5	91.1	673.2
MPL	0.014	0.007	0.185	0.110	0.171	0.093	94.8	90.4	264.8
MPLs	−0.003	0.004	0.187	0.112	0.176	0.098	94.9	90.7	638.4
EE.SWc	−0.272	−0.192	0.442	0.205	0.392	0.188	80.2	75.4	61.7
EE.SWb	−0.808	−0.345	0.239	0.128	0.242	0.128	8.8	25.4	40.9
EE.SWa	−0.327	−0.138	0.202	0.103	0.204	0.109	64.6	77.1	3.5
EE.HSWm	−1.492	−0.400	0.349	0.209	0.332	0.183	0.8	41.2	57.3
HWZ	−0.017	0.001	0.179	0.117	0.172	0.094	93.0	89.7	989.9
AEE	−0.002	0.002	0.174	0.108	0.166	0.090	94.5	89.5	301.4
AEEX	−0.030	−0.009	0.168	0.112	0.169	0.096	95.0	90.2	463.4

Table 3: Simulation results for Scenario 3 where the examination times are informative about the recurrent events after conditioning on covariates with $n = 100$. Column bias is the average bias; ESE is the empirical standard error; ASE is the average standard error based on the standard bootstrap; CP is the empirical coverage probability (%); time is the average time in seconds used in both point estimation and bootstrap variance estimation.

	bias		ESE		ASE		CP (%)		Time
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	
				$Z \sim$ gamma distribution					
MLs	−0.147	−0.038	0.199	0.117	0.188	0.102	86.7	87.7	594.9
MPL	−0.164	−0.048	0.210	0.115	0.190	0.101	84.5	87.3	221.6
MPLs	−0.172	−0.048	0.208	0.118	0.192	0.103	85.5	86.6	496.9
EE.SWc	−0.248	−0.110	0.318	0.153	0.290	0.147	81.6	85.8	64.9
EE.SWb	−0.379	−0.155	0.215	0.113	0.225	0.119	59.9	74.7	42.6
EE.SWa	−0.206	−0.078	0.192	0.100	0.208	0.109	86.1	88.7	2.9
EE.HSWm	−0.723	−0.191	0.314	0.186	0.309	0.166	38.5	70.6	58.8
HWZ	−0.003	−0.006	0.212	0.125	0.199	0.112	91.7	91.2	990.9
AEE	−0.144	−0.038	0.198	0.116	0.189	0.100	88.8	90.7	225.6
AEEX	−0.015	−0.014	0.206	0.122	0.191	0.105	92.6	89.9	419.1
				$Z \sim$ uniform distribution					
MLs	−0.143	−0.032	0.181	0.113	0.173	0.096	86.7	89.7	633.3
MPL	−0.169	−0.039	0.186	0.114	0.177	0.097	83.3	88.8	246.3
MPLs	−0.175	−0.040	0.189	0.115	0.180	0.099	83.0	89.3	547.7
EE.SWc	−0.258	−0.130	0.318	0.150	0.292	0.147	79.1	83.2	64.7
EE.SWb	−0.453	−0.183	0.218	0.104	0.216	0.115	45.2	66.5	43.9
EE.SWa	−0.221	−0.087	0.185	0.099	0.193	0.103	80.6	87.9	3.1
EE.HSWm	−0.861	−0.229	0.323	0.179	0.303	0.165	21.0	65.7	59.8
HWZ	−0.008	0.001	0.181	0.123	0.171	0.112	93.4	91.5	1010.1
AEE	−0.142	−0.032	0.181	0.112	0.171	0.105	86.6	90.6	241.7
AEEX	−0.022	−0.010	0.174	0.116	0.167	0.097	94.6	90.9	434.9