Empirical likelihood inference for the panel count data with informative observation process

Faysal Satter, Yichuan Zhao
Georgia State University

1. Panel Count Data

- Arise in long-term event-history or longitudinal studies.
- It may infeasible or unrealistic to monitor the study subjects continuously.
- They are only observed at discrete time points within the study period.
- Only the total number of events occurred between two time points is known instead of actual time of the events.
- Such datasets are commonly known as panel count data.

2. Example

Patient ID	Size	Months (0 – 10 – 20 – 30)
		Placebo group
1	3	1 0 . . . .
2	1	2 0 . . 0 . . . .
3	1	1 . . . . . . 0 . . . .
4	1	5 . . 0 . . . . . 0 0 . . . .
5	1	4 0 . . 0 . 1 0 . . 0 . . . .
6	1	1 . . 0 . . . . . . 0 . . . 0 . . . .
7	1	1 . 0 . . . . 0 . 2 . . . 3 . 0 . . . .
8	1	1 . . 0 . . . . 0 . . . 0 . . . .
9	3	1 . . . . 2 . . . . 0 . 0 . . . . . 0 . . . .
10	3	1 . . 0 . . . 0 . . 6 . . . . 3 . . . 0 . . . . 0 . . . .
11	1	1 0 . 8 . . . . 0 . 0 . . 0 . . 8 . . 0 . . . 8 . . . .

3. Underlying Processes

- The recurrent event process: Controls the number of events occurred between two time points.
- The observation process: Controls the observation times for each subject.

4. Model Setup

Denote:
$\boldsymbol{Y_i(t)}$: The cumulative number of event occurrences before or at time t,
$\boldsymbol{O_i(t)}$: The total observation before or at time t.
$\boldsymbol{Z_i(t)}$: A vector of covariates.

One can observe the dataset:

$$\{O_i(t), Z_i(t), Y_i(T_{i,1}), \cdots, Y_i(T_{i,M_i});\ 0 \le t, T_{i,K_i} \le C_i, i = 1, \cdots, n\},$$

i.e., we only have panel count data on $Y_i(t)$'s.

Define $\mathcal{F}_{it} = \{O_i(s), 0 \le s < t, i = 1, \cdots, n\}$ as the history or filtration of the observation process O_i up to time $t-$.

The observation process $O_i(t)$ follows the proportional rate model

$$E\{dO_i(t)|Z_i(t)\} = e^{\gamma Z_i(t)}\lambda_0(t)dt,$$

where γ is a vector of unknown parameters and $\lambda_0(\cdot)$ is an unspecified baseline rate function.

we model the conditional mean function of $Y_i(t)$ given $Z_i(t)$ and $\mathcal{F}_{it}$ as

$$E\{Y_i(t)|Z_i(t), \mathcal{F}_{it}\} = g\{\mu_0(t)e^{\beta_1' Z_i(t) + \beta_2' H(\mathcal{F}_{it})}\},$$

where $g(\cdot)$ is a known twice continuously differentiable and strictly increasing function, $\mu_0(t)$ denotes an unknown arbitrary function of t, β_1 and β_2 are vectors of unknown regression parameters, and $H(\cdot)$ is a vector of known functions of $\mathcal{F}_{it}$.

5. Empirical Likelihood

Advantages:

- EL enjoys parametric likelihood benefits
- Parametric assumption not required
- Confidence region shaped by data only
- The confidence interval is range respective, transformation invariant, Bartlett correctable .
- Perform better when sample size is small
- Estimation of variance is not needed, as the studentization is done internally

Define $W_{ni}(\beta;\hat{\gamma}) = \int_0^\tau W(t)\{X_i(t) - \hat{E}_X(t;\beta,\hat{\gamma})\}d\hat{M}_i(t;\beta,\hat{\gamma}) - \int_0^\tau \frac{W(t)\hat{R}(t;\beta,\hat{\gamma})}{S^{(0)}(t,\hat{\gamma})}d\hat{M}_i^*(t;\hat{\gamma}) - \hat{P}(\beta,\hat{\gamma})\hat{D}^{-1}\int_0^\tau\{Z_i(t) - \bar{Z}(t;\hat{\gamma})\}d\hat{M}_i^*(t;\hat{\gamma})$.

Let $p = (p_1, p_2, \cdots, p_n)$ be a probability vector, i.e., $\sum_{i=1}^n p_i = 1$, and $p_i \ge 0$ for all i. Then the EL ratio, evaluated at true parameter value β_0 is defined as:
$R(\beta_0) = sup\{\prod_{i=1}^n np_i : \sum_{i=1}^n p_i W_{ni}(\beta_0;\hat{\gamma}) = 0, p_i \ge 0, \sum_{i=1}^n p_i = 1\}$.
The empirical log-likelihood ratio at β is given by

$$l(\beta) = -2logR(\beta) = 2\sum_{i=1}^n log\{1 + \lambda' W_{ni}(\beta;\hat{\gamma})\}$$

where $\lambda = (\lambda_1, \cdots, \lambda_p)'$ is the solution to

$$\sum_{i=1}^n \frac{W_{ni}(\beta;\hat{\gamma})}{1 + \lambda' W_{ni}(\beta;\hat{\gamma})} = 0$$

Theorem 1: Under the regularity conditions stated in the Appendix, $l(\beta_0)$ converges in distribution to χ_p^2 as $n \to \infty$, where χ_p^2 is a chi-square distribution with p degrees of freedom.

6. Simulation Result

	$\tau = 1$			
	β_1		β_2	
(β_1, β_2)	NA	EL	NA	EL
				$n = 30$
(0.1, 0.1)	0.888 (2.012)	0.900 (2.063)	0.848 (0.237)	0.871 (0.266)
(0.3, 0)	0.904 (2.455)	0.911 (2.451)	0.849 (0.326)	0.870 (0.324)
(0, 0.1)	0.896 (2.056)	0.901 (2.128)	0.848 (0.246)	0.867 (0.262)
(0.3, 0.1)	0.890 (1.931)	0.915 (1.985)	0.842 (0.223)	0.879 (0.287)
(0.6, 0.2)	0.870 (1.535)	0.872 (1.588)	0.802 (0.172)	0.818 (0.188)
				$n = 70$
(0.1, 0.1)	0.926 (1.351)	0.925 (1.393)	0.886 (0.161)	0.899 (0.187)
(0.3, 0)	0.941 (1.637)	0.944 (1.668)	0.887 (0.224)	0.899 (0.228)
(0, 0.1)	0.919 (1.381)	0.922 (1.419)	0.877 (0.164)	0.886 (0.174)
(0.3, 0.1)	0.923 (1.300)	0.924 (1.343)	0.878 (0.151)	0.902 (0.164)
(0.6, 0.2)	0.914 (1.034)	0.920 (1.126)	0.822 (0.115)	0.854 (0.134)

7. Bladder Cancer Dataset

Two treatment groups: placebo (47 patients) and theitepa (38 patients).
Let β_1, β_2 and β_3 represent the efects of the treatment, the size of the largest tumor, and the number of initial tumors, respectively. In addition, α is the effect of the observation or visit process.
Also, we assume $H(\mathcal{F}_{it}) = O_i(t-)$.

	NA		EL	
	CI	Length	CI	Length
β_1	(-2.487, -0.996)	1.491	(-2.497, -0.979)	1.518
β_2	(-0.314, 0.111)	0.425	(-0.342, 0.103)	0.445
β_3	(0.157, 0.416)	0.259	(0.161, 0.436)	0.275
α	(0.006, 0.011)	0.093	(0.002, 0.099)	0.097