A Semiparametric Regression Model for Panel Count Data: Some Progress

Jon A. Wellner
University of Washington

with Ying Zhang,
University of Central Florida, Statistics
and Hao Liu,
University of Washington, Biostatistics

Second Biostatistics Symposium,
Seattle, November 20, 2000

Email: jaw@stat.washington.edu

*http: //www.stat.washington.edu/
jaw/jaw.research.html*

Outline.

1. **Introduction:**
   A semiparametric regression model for panel count data.

2. **Maximum Pseudo-likelihood and Maximum Likelihood estimators.**

3. **Properties of the estimators when the Poisson assumption fails.**

4. **Efficiency comparisons: when should we avoid the pseudo-MLE?**

5. **Problems.**

1. Introduction:

A Semiparametric Regression Model for Panel Count Data

I. Model for the Counting Process

A. Mean structure: $E\{\mathbb{N}(t)|Z\} = e^{\theta' Z}\Lambda(t)$,
Λ monotone non-decreasing.

B. Poisson process assumption:

$(\mathbb{N}|Z) \sim$ non-homogeneous Poisson process .

C. Parameters of interest: (θ, Λ)
(or just θ).

D. Study estimators when the Poisson assumption B **fails**.

II. Observation Process and Covariate Distribution:

A. $(K, \underline{T}_K | Z) \sim G(\cdot|Z)$ conditionally independent of $(\mathbb{N}|Z)$;
K is the (random) number of observation times of the process $\mathbb{N}$;
$\underline{T}_K$ is a vector of ordered observation times:

$$0 = T_{K,0} < T_{K,1} < \ldots < T_{K,K}\,.$$

B. $Z \sim H$ on R^d.

C. No assumptions about G or H.

III. Data and Primary Goal:

A. Data:

$$X = \Big(Z, K, \underline{T}_K, \mathbb{N}(T_{K,1}), \ldots, \mathbb{N}(T_{K,K})\Big)$$
$$\equiv (Z, K, \underline{T}_K, \underline{\mathbb{N}}_K) .$$

We observe $X_1, \ldots, X_n$ i.i.d. as X.

B. Pictures!

C. Based on $X_1, \ldots, X_n$ i.i.d. as X, estimate (θ, Λ).

2. Maximum Pseudo-likelihood and Maximum Likelihood estimators

A. Pseudo-likelihood: use the marginal distributions of $\mathbb{N}$,

$$P(\mathbb{N}(t) = k|Z) = \frac{\Lambda(t|Z)^k}{k!}\exp(-\Lambda(t|Z))$$

and *ignore dependence* between $\mathbb{N}(t_1)$, $\mathbb{N}(t_2)$ to obtain the pseudo-likelihood:

$$\begin{aligned} & l_n^{ps}(\theta, \Lambda) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{K_i} \Big\{ \mathbb{N}^{(i)}(T_{K_i,j}^{(i)}) \log \Lambda(T_{K_i,j}^{(i)}) \\ &\quad + \mathbb{N}^{(i)}(T_{K_i,j}^{(i)})\theta' Z_i - e^{\theta' Z_i}\Lambda(T_{K_i,j}^{(i)}) \Big\} . \end{aligned}$$

Then

$$(\widehat{\theta}_n^{ps}, \widehat{\Lambda}_n^{ps}) \equiv \text{argmax}_{\theta,\Lambda}\, l_n^{ps}(\theta, \Lambda)\,.$$

Implement in two steps:

$$\widehat{\Lambda}_n^{ps}(\cdot, \theta) \equiv \text{argmax}_{\Lambda}\, l_n^{ps}(\theta, \Lambda)\,,$$

and define

$$l_n^{ps,profile}(\theta) \equiv l_n^{ps}(\theta, \widehat{\Lambda}_n^{ps}(\cdot, \theta))\,.$$

Then

$$\widehat{\theta}_n^{ps} = \text{argmax}_{\theta}\, l_n^{ps,profile}(\theta)\,,$$

and

$$\widehat{\Lambda}_n^{ps} = \widehat{\Lambda}_n^{ps}(\cdot, \widehat{\theta}_n^{ps})\,.$$

Let $t_1 < \ldots < t_m$ denote the ordered distinct observation time points in the collection of all observations times,
$\{T^{(i)}_{K_i,j},\ j = 1, \ldots, K_i,\ i = 1, \ldots, n\}$, and set

$$\begin{aligned}
w_l &= \sum_{i=1}^{n} \sum_{j=1}^{K_i} 1_{[T^{(i)}_{K_i,j} = t_l]}\,, \\
\overline{N}_l &= \frac{1}{w_l} \sum_{i=1}^{n} \sum_{j=1}^{K_i} \mathbb{N}^{(i)}_{K_i,j} 1_{[T^{(i)}_{K_i,j} = t_l]}\,, \\
\overline{A}_l(\theta, Z) &= \frac{1}{w_l} \sum_{i=1}^{n} \sum_{j=1}^{K_i} \exp(\theta' Z^{(i)})\, 1_{[T^{(i)}_{K_i,j} = t_l]}\,.
\end{aligned}$$

$$\widehat{\Lambda}_n^{ps}(\cdot,\theta) = \text{left-derivative of Greatest Convex Minorant of } \{(\sum_{l\le i} w_l \overline{A}_l(\theta,Z), \sum_{l\le i} w_l \overline{N}_l)\}_{i=1}^m$$

$$= \max_{i\le l} \min_{j\ge l} \frac{\sum_{i\le p\le} w_p \overline{N}_p}{\sum_{i\le p\le} w_p \overline{A}_p(\theta,Z)} \quad \text{at} \quad t_l\,,$$

which is easy to compute.

B. Maximum likelihood: use the independence of the increments of $\mathbb{N}$, $\Delta\mathbb{N}(s,t] \equiv \mathbb{N}(t) - \mathbb{N}(s)$, and the Poisson distribution of increments,

$$
\begin{aligned}
&P(\Delta\mathbb{N}(s,t] = k|Z) \\
&\quad = \frac{[\Delta\Lambda((s,t]|Z)]^k}{k!} \exp(-\Delta\Lambda((s,t]|Z))
\end{aligned}
$$

to obtain the log-likelihood:

$$
\begin{aligned}
&l_n(\theta,\Lambda) \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\Big\{\Delta\mathbb{N}^{(i)}((T^{(i)}_{K_i,j-1}, T^{(i)}_{K_i,j}]) \\
&\qquad\qquad \cdot \log \Delta\Lambda((T^{(i)}_{K_i,j-1}, T^{(i)}_{K_i,j}]) \\
&\qquad + \Delta\mathbb{N}^{(i)}((T^{(i)}_{K_i,j-1}, T^{(i)}_{K_i,j}])\theta' Z_i \\
&\qquad - e^{\theta' Z_i}\Delta\Lambda((T^{(i)}_{K_i,j-1}, T^{(i)}_{K_i,j}])\Big\} .
\end{aligned}
$$

Then

$$(\widehat{\theta}_n, \widehat{\Lambda}_n) \equiv \text{argmax}_{\theta,\Lambda}\, l_n(\theta, \Lambda)\,.$$

Implement this maximization in two steps (profile likelihood):

$$\widehat{\Lambda}_n(\cdot, \theta) \equiv \text{argmax}_{\Lambda}\, l_n(\theta, \Lambda)\,,$$

and define

$$l_n^{profile}(\theta) \equiv l_n(\theta, \widehat{\Lambda}_n(\cdot, \theta))\,.$$

Then

$$\widehat{\theta}_n = \text{argmax}_{\theta}\, l_n^{profile}(\theta)\,,$$

and

$$\widehat{\Lambda}_n = \widehat{\Lambda}_n(\cdot, \widehat{\theta}_n)\,.$$

Computation of the (profile) "estimator" $\widehat{\Lambda}_n(\cdot, \theta)$ is **hard**, but possible:
iterative convex minorant algorithm.

3. Properties of the Estimators when the Poisson Assumption Fails

Theorem 1. If assumption A holds, then (under further integrability and boundedness hypotheses):

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightarrow_d Z \sim N_d\left(0, A^{-1}B\left(A^{-1}\right)'\right),$$

and

$$\sqrt{n}(\widehat{\theta}_n^{ps} - \theta_0) \rightarrow_d Z^{ps} \sim N_d\left(0, (A^{ps})^{-1}B^{ps}\left((A^{ps})^{-1}\right)'\right)$$

where

$$\begin{aligned} B &= Em^*(\theta_0, \Lambda_0; X)^{\otimes 2} \\ &= E\left\{\sum_{j,j'=1}^{K} C_{j,j'}(Z)\left[Z - \frac{E\left(Ze^{\theta_0' Z}|K, \underline{T}_{K,j,j'}\right)}{E\left(e^{\theta_0' Z}|K, \underline{T}_{K,j,j'}\right)}\right]\right. \\ &\qquad \left.\left[Z - \frac{E\left(Ze^{\theta_0' Z}|K, \underline{T}_{K,j,j'}\right)}{E\left(e^{\theta_0' Z}|K, \underline{T}_{K,j,j'}\right)}\right]'\right\},\end{aligned}$$

$$
\begin{aligned}
A &= -\dot{S}_{11}(\theta_0,\Lambda_0) + \dot{S}_{21}(\theta_0,\Lambda_0)[\mathbf{h}^*] \\
&= E\left\{ \sum_{j=1}^{K} \Delta\Lambda_{0Kj} e^{\theta_0' Z} \left[Z - \frac{E\left(Ze^{\theta_0' Z} | K, \underline{T}_{K,j,j-1}\right)}{E\left(e^{\theta_0' Z} | K, \underline{T}_{K,j,j-1}\right)} \right] \right.
\end{aligned}
$$

$$
C_{j,j'}(Z) = \text{Cov}\left[\Delta N_{Kj}, \Delta N_{Kj'} | Z, K, \underline{T}_K\right] .
$$

$$
\begin{aligned}
B^{ps} &= E m^{*ps}(\theta_0,\Lambda_0; X)^{\otimes 2} \\
&= E\left\{ \sum_{j,j'=1}^{K} C^{ps}_{j,j'}(Z) \left[Z - \frac{E\left(Ze^{\theta_0' Z} | K, T_{K,j}\right)}{E\left(e^{\theta_0' Z} | K, T_{K,j}\right)} \right] \right. \\
&\qquad \left. \left[Z - \frac{E\left(Ze^{\theta_0' Z} | K, T_{K,j'}\right)}{E\left(e^{\theta_0' Z} | K, T_{K,j'}\right)} \right]' \right\} , \\
A^{ps} &= -\dot{S}^{ps}_{11}(\theta_0,\Lambda_0) + \dot{S}^{ps}_{21}(\theta_0,\Lambda_0)[\mathbf{h}^*] \\
&= E\left\{ \sum_{j=1}^{K} \Lambda_{0Kj} e^{\theta_0' Z} \left[Z - \frac{E\left(Ze^{\theta_0' Z} | K, T_{K,j}\right)}{E\left(e^{\theta_0' Z} | K, T_{K,j}\right)} \right] \right. \\
C^{ps}_{j,j'}(Z) &= \text{Cov}\left[N_{Kj}, N_{Kj'} | Z, K, T_{K,j,j'}\right] ,
\end{aligned}
$$

If the Poisson process assumption B holds,

$$
A = B = I(\theta) ,
$$

and $\widehat{\theta}$ is (asymptotically) efficient.

4. Efficiency comparisons: when should we avoid the pseudo-MLE?

Scenario 1: Supppose that:

- $\Lambda_0(t) = \lambda t$.
- $\mathbb{N}$ is a Poisson process.
- $(K, \underline{T}_K)$ independent of Z.
- $(\underline{T}_K | K) \sim$ order statistics of K i.i.d. $U[0, M]$ random variables.
- $K \sim$ one of:
  - (a) Degenerate at k_0.
  - (b) (Shifted) Poisson(γ).
  - (c) Discrete zeta(α).

$$ARE(pseudo, mle) = \frac{[E(K/2)]^2}{E\left\{\frac{K}{K+1}\right\} E\left\{\frac{K(2K+1)}{6}\right\}}$$

Scenario 2: Supppose that:

- $\Lambda_0(t) = \lambda t$.
- $\mathbb{N}$ is a Mixed-Poisson
  (= Negative-Binomial process).
- $(K, \underline{T}_K)$ independent of Z.
- $(\underline{T}_K | K) \sim$ order statistics of
  K i.i.d. $U[0, M]$ random variables.
- $K \sim$ one of:
  - (a) Degenerate at k_0.
  - (b) (Shifted) Poisson(γ).
  - (c) Discrete zeta(α).

$$
\begin{aligned}
&ARE(pseudo, mle)(NegBin) \\
&= \frac{\left(1 + a \frac{E\left(\frac{K}{K+2}\right)}{E\left(\frac{K}{K+1}\right)}\right)}{\left(1 + a \frac{E\left(\frac{K(3K+1)}{12}\right)}{E\left(\frac{K(2K+1)}{6}\right)}\right)} \\
&\quad \cdot ARE(pseudo, mle)(Poisson) .
\end{aligned}
$$

where $a \equiv q/p = \lambda M/\gamma$.

5. Problems.

- Algorithms for computation of the MLE $\widehat{\theta}$ of θ.

□ Pseudo MLE of Λ, no covariates: Sun and Kalbfleisch (1995);

□ MLE of Λ, no covariates: Zhang and Wellner (2000);

□ pseudo MLE $\widehat{\theta}^{ps}$ of θ: Zhang (1999), (2000).

- Further efficiency comparisons when $(K, \underline{T}_K)$ is dependent on Z.

- Other compromise / hybrid estimators?

- Inference about Λ: likelihood ratio based confidence intervals for Λ (Banerjee and Wellner (2000))

- MLE's for Mixed Poisson Process $(\mathbb{N}|Z)$ (e.g. Negative-Binomial)?

6. Selected References.

Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica* **5**, 279 - 290.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28**, to appear.

Zhang, Y. (1998). Estimation for Counting Processes Based on Incomplete Data. Unpublished Ph.D. dissertation, University of Washington.

Zhang, Y. (1999). A semiparametric pseudo likelihood estimation for panel count data. Proceedings, ASA Statistical Computing Section, 1-9.