BIOS 791 Miniproject 1

Taeim Kwon

The paper chosen for this miniproject was Maximum Likelihood Estimation for Semiparametric Regression Models with Panel Count Data[1]. The panel count data is often seen in clinical trials, econometrics, and epidemiological studies. It has the event of interest measured intermittently, meaning the exact times of recurrent events are not observed. This leads to some practical issues of analyzing panel count data. One of the issues is that correlations would exist within subjects due to repeated measures. The previous studies have used the proportional means model to study panel count data with time-independent covariates or the marginal distributions using the models, but there were some limitations such as inefficiency and unstableness in parameter estimation. Additionally, including only time-independent covariates would have more restrictions on the models and fail to detect time-dynamics which often exist by the nature of data [2].

Thus, the authors of the paper proposed nonhomogeneous Poisson processes with random effects to study how one or more types of recurrent events are affected by time-dependent covariates. It differs from the other studies as their model with random effects would address the within-subject correlations of recurrent events and try to capture the effect of time-dependent covariates on recurrent events. They employed an EM algorithm to ensure a closed form of the estimator for the cumulative baseline intensity function, and the parameter estimators derived from the estimating equations were unique. Finally, based on the empirical process theory, they showed the estimators are consistent, as well as asymptotically normal and efficient.

For methods, a nonhomogeneous Poisson process with intensity function was specified to be $N_{ki}(t)$ for subject $i = 1, ..., n$, type of recurrent events $k = 1, ..., K$ and time t with a set of some time-dependent covariates $X_i(t)$ where the intensity function was given by

$$\lambda_{ki}(t) = \lambda_k(t) exp(\beta_k^T X_i(t) + b_{ki}^T Z_i(t) + \xi_i^T \tilde{Z}_i(t)).$$

Here β_k is a vector of the parameters of interest in d-dimensional space, b_{ki} is a random effect in p-dimensional space, and ξ_i is a random effect shared by the K types of events in q-dimensional space. Let b_{ki} and ξ_i be normal with mean 0 and covariance matrices Σ_k and Ψ, respectively, assuming they are mutually independent. The increments $\Delta_{kij} = N_{ki}(U_{kij}) - N_{ki}(U_{ki,j-1})$ where U_{kij} represents the examination time at $j = 1, ..., m_{ki}$ for subject i and type of events k. Since $\Delta_{kij}|X_i(t), b_{ki}, \xi_i \sim \text{Poisson}(\mu)$, where μ is the integral of $\lambda_{ki}(t)$ over $U_{ki,j-1}$ to U_{kij}, and b_{ki} and ξ_i follows the multivariate normal distribution,

the likelihood is proportional to

$$\prod_{i=1}^{n}\left(\int_{\xi_i}\phi(\xi_i;\Psi)\prod_{k=1}^{K}\int_{b_{ki}}\phi(b_{ki};\Sigma_k)\prod_{j=1}^{m_{ki}}\frac{\left[\int_{U_{ki,j-1}}^{U_{kij}}\exp\left\{\beta_k^\top X_i(t)+b_{ki}^\top Z_i(t)+\xi_i^\top\tilde{Z}_i(t)\right\}\,d\Lambda_k(t)\right]^{\Delta_{kij}}}{\Delta_{kij}!}\right.$$
$$\left.\times\exp\left[-\int_0^{C_{ki}}\exp\left\{\beta_k^\top X_i(t)+b_{ki}^\top Z_i(t)+\xi_i^\top\tilde{Z}_i(t)\right\}\,d\Lambda_k(t)\right]\,db_{ki}\,d\xi_i\right)$$

Then, they used an EM algorithm to find nonparametric maximum likelihood estimators for $\hat{\beta}_k, \hat{\Lambda}_k, \hat{\Sigma}_k, \hat{\Psi}$ after letting $\Lambda_k(t)=\sum_{i=1}^{m_k}\lambda_{kl}I(t_{kl}\le t)$ where $l=1,...,m_k$ and introducing independent Poisson random variables W_{kil}. The Newton-Raphson method was used to maximize the conditional expectation of a complete-data log likelihood given the observed data for the M-step. The Gaussian quadrature approximation was employed to find the conditional expectations for the E-step. After setting initial values for the parameters, they iterated the algorithm until convergence. They then proposed the prediction of future events from the event history. Moreover, three regularity conditions followed by two theorems were given to show the nonparametric maximum likelihood estimators $\theta=(\beta_k,\Sigma_k,\Psi)$ and Λ_k are consistent, and asymptotically normal and efficient. The first condition ensures boundedness of θ_0, the second condition is met under linear independence of covariates, and the third condition entails that examination times are positively distanced. The profile loglikelihood was used to calculate the covariance matrix which attains the semiparametric efficiency bound. Some concepts related to the proofs of the theorems and learned in class include Glivenko-Cantelli and Donsker theorems with the bounded inverse theorem in Banach space. This section will be further studied in the miniproject 2.

Theorem 1. Under Conditions 1–3, $\|\hat{\theta}-\theta_0\|\to 0$ almost surely, where $\|\cdot\|$ is the Euclidean norm. In addition, $\sum_{k=1}^{K}\sup_{t\in[0,\tau_k]}\left|\hat{\Lambda}_k(t)-\Lambda_{k0}(t)\right|\to 0$ almost surely.

Theorem 2. Under Conditions 1–3, $n^{1/2}(\hat{\theta}-\theta_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

In their simulation studies, two types of recurrent events with intensity functions were examined for $n=200, 400$ or 800. The parameter estimators were well converged in the EM algorithm for $n=400$ and 800 and were nearly unbiased with true variation. Additionally, a separate univariate analysis for the two types of events was done using the algorithm, and they found that both the separate analysis and joint analysis worked well, although the joint analysis was more efficient than the separate analysis. Overall, the results were shown to perform well and efficiently.

They applied their methods to a skin cancer trial [3] which had two types of recurrent events, basal cell carcinoma and squamous cell carcinoma, to examine treatment effect. The result was not statistically significant, but the risk of both basal cell carcinoma and squamous cell carcinoma decreased by treatment. They skin cancer trial [3], though, found that there was a significant treatment effect only on basal cell carcinoma because the standard two-sample t-test that is not suitable for panel count data was used in the study.

It should be noted in the paper random effects were incorporated in the model and that theoretical problems involved in random effects were were well tackled to improve some practical issues. Having fixed effects only in the model would not be a good idea to get the prediction of future events as unobserved heterogeneity which would affect the prediction cannot be explained by the fixed effects. Instead, random effects help in using past event history to get the better prediction of future events. They also found parameter estimation was efficient, and random effects could account for the dependence of recurrent events. Furthermore, the proportional means model, which is different from the proportional intensity model in that the proportional means model is for population-average effects, can be derived from their EM algorithm. Lastly, they raised a point that can be further studied under the dependence of the examination times on the recurrent event process, which can be explained by a proportional intensity model for the examination times sharing random effects with their model.

References

[1] Zeng D. and Lin D. Maximum likelihood estimation for semiparametric regression models with panel count data. *Biometrika*, 108(4):947–960, 2020.

[2] Guo Y., Sun D., and Sun J. Inference of a time-varying coefficient regression model for multivariate panel count data. *Journal of Multivariate Analysis*, 192:105047, 2022.

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