Cluster sampling approach for multivariate survival data analysis

Qiyue He[1*], Pengkun Liang[1], Dinghao Wang[1]

[1]Department of Mathematics and Statistics, University of Calgary,
2500 University Drive NW, Calgary, AB, Canada T2N1N4
[*]Correspondence: qiyue.he@ucalgary.ca

Abstract

Clustered data is quite common in clinical trails that either an individual is recorded multiple survival events or some individuals within a family are grouped into a cluster. Previous works provide different approaches including Poisson regression and survival modeling. In this report, we follow the paper by Pingfu Fu and J. Sunil Rao and carry out some simulation and real data analysis to compare the modified method in this paper and usual estimation.

Keywords: small sample; variance estimation; survival analysis; clustered data

1 Introduction

In medical research, clustered survival events can occur quite often. We usually come across a scenario where an individual is subject to experiencing repeat events (e.g., recurrent of a diseasse) over time. For example, a child is diagnosed with chronic lung disease (CLD) for a period of time. After recovery, the child will be susceptible to repeat occurrences of CLD in the future [1]. In this case, each cluster represents a child, and the time to the start of each CLD period is a series of observations in each cluster. For clustered dataset, to make the statistical inference and estimate the unknown parameters, Anderson and Gill proposed a AG model [2], in which each subject is treated as a multi-event counting process with essentially independent increments. Prentice, Williams and Peterson proposed a conditional PWP model [3], while Wei, Lin and Weisfeld presented a WLW model [4] in which they ignored the correlation and obtained the estimated coefficients, followed by fix of the variance of estimated coefficients.

With the increasing number of research interest in the regression modelling of multivariate clustered survival data [5][6], Segal and Neuhaus [7] derived a way to use GEE Poisson regression techniques to analyze such data, which combined the generalized linear regression model with the generalized estimating equation machinery [8] to do the point estimation. To evaluate the performance of the parameter estimation, robust inference is highly recommended, which is handled by using sandwich estimators for estimated variance of the estimated regression parameters.

Survey sampling is another area where clustered events are quite common. In this scenario, the design effects approach [9] which is based on sample survey techniques, has been used. In order to take the correlation among observations within each cluster into consideration, one can either transform the data by a design effect and then apply standard methods assuming independence, or one can firstly apply standard methods assuming independence, and then adjust the variances of the parameter estimates by design effect. Rao and Scott [9][10], Bieler and Williamns has applied relative work in non-survival settings [11]. In this paper, we will use the design effect method under maximum likelihood estimation of Poisson distribution and GEE Poisson regression to show that the design effect approach also performs well for the clustered survival data.

The rest of the paper is organized as follows: theoretical methods for survival model estimation and model diagnostic are introduced in Section 2. Simulation studies under different cluster size, different number of clusters, different censoring rate, and different values of index for the positive stable distribution are presented in Section 3. Real data analysis and different model comparisons are conducted in Section 4. Conclusions and future work are mentioned in Section 5.

2 Theoretical Method

Assume that we have a sample of observed time data denoted as $(T_{ijk}, \delta_{ijk}, x_{ijk})$, where for observation $k = 1, 2, ..., n_{ij}$ of idividual (cluster) $j = 1, 2, ..., m_i$ of treatment group $i = 1, 2, ..., G$, T_{ijk} represents the observed time, δ_{ijk} is a censoring indicator taking the value 1 if T_{ijk} is uncensored and 0 otherwise, and x_{ijk} is a p-dimensional vector of covariates.

We follow the assumption mentioned in Segal and Neuhaus's paper [7]. In our case, we assume that the observed survival time T_{ijk} follows the Weibull distribution with shape parameter α and scale parameter $\lambda = 1$. i.e., $T_{ijk} \sim \text{Weibull}(\alpha, 1)$.

2.1 Estimate the Survival Function by Using GEE Poisson Regression Model

By Cox's proportional hazards model, we can get the hazard function for the kth observation of the jth individual in the ith treatment group involves covariates x_{ijk},

$$h_{ijk}(t) = h_0(t)\exp(\beta^{\mathrm{T}} x_{ijk}) \tag{1}$$

where β is a p-dimensional vector of regression parameters, and $h_0(t)$ is the baseline hazard function, i.e. $h_0(t) = \alpha t^{\alpha-1}$. Thus, the cumulative hazard function for the kth observation of the jth individual in the ith treatment group is

$$\begin{aligned}
H_{ijk}(t) &= \int_0^t h_{ijk}(s)ds \\
&= \int_0^t h_0(s)\text{exp}(\beta^{\mathrm{T}} x_{ijk})ds \\
&= \text{exp}(\beta^{\mathrm{T}} x_{ijk}) \int_0^t h_0(s)ds \\
&= \text{exp}(\beta^{\mathrm{T}} x_{ijk})H_0(t)
\end{aligned} \tag{2}$$

where $H_0(t)$ is the cumulative baseline hazard function. Then, we can get the survival function for the kth observation of the jth individual in the ith treatment group as

$$\begin{aligned}
S_{ijk}(t) &= \text{exp}(-H_{ijk}(t)) \\
&= \text{exp}\Big\{ - \text{exp}(\beta^{\mathrm{T}} x_{ijk})H_0(t)\Big\}
\end{aligned} \tag{3}$$

and the probability density function of the observed survival time for the kth observation of the jth individual in the ith treatment group is

$$\begin{aligned}
f_{ijk}(t) &= h_{ijk}(t)S_{ijk}(t) \\
&= \Bigg(h_0(t)\text{exp}(\beta^{\mathrm{T}} x_{ijk})\Bigg)\Bigg(\text{exp}\Big\{ - \text{exp}(\beta^{\mathrm{T}} x_{ijk})H_0(t)\Big\}\Bigg) \\
&= h_0(t)\text{exp}\Big(\beta^{\mathrm{T}} x_{ijk} - \text{exp}(\beta^{\mathrm{T}} x_{ijk})H_0(t)\Big)
\end{aligned} \tag{4}$$

Under the standard assumption of independent censoring, the likelihood for the kth observation of the jth individual in the ith treatment group is

$$\begin{aligned}
L_{ijk}(\alpha,\beta) &= f_{ijk}(t)^{\delta_{ijk}} S_{ijk}(t)^{1-\delta_{ijk}} \\
&= \Big\{h_0(t)\exp\big(\beta^{\mathrm{T}}x_{ijk} - \exp(\beta^{\mathrm{T}}x_{ijk})H_0(t)\big)\Big\}^{\delta_{ijk}} \exp\Big\{-\exp(\beta^{\mathrm{T}}x_{ijk})H_0(t)\Big\}^{1-\delta_{ijk}} \\
&= \Big\{h_0(t)\exp(\beta^{\mathrm{T}}x_{ijk})\Big\}^{\delta_{ijk}} \exp\Big\{-\exp(\beta^{\mathrm{T}}x_{ijk})H_0(t)\Big\} \\
&= \Big\{\frac{h_0(t)}{H_0(t)}H_0(t)\exp(\beta^{\mathrm{T}}x_{ijk})\Big\}^{\delta_{ijk}} \exp\Big\{-\exp(\beta^{\mathrm{T}}x_{ijk})H_0(t)\Big\} \\
&= \mu_{ijk}^{\delta_{ijk}}\exp(-\mu_{ijk})\Big(\frac{h_0(t)}{H_0(t)}\Big)^{\delta_{ijk}}
\end{aligned} \tag{5}$$

where $\mu_{ijk} = H_0(t)\exp(\beta^{\mathrm{T}}x_{ijk})$. Since δ_{ijk} can only take the value 0 or 1, equation (5) can be further transformed into

$$L_{ijk}(\alpha,\beta) = \frac{\mu_{ijk}^{\delta_{ijk}}\exp(-\mu_{ijk})}{\delta_{ijk}!}\Big(\frac{h_0(t)}{H_0(t)}\Big)^{\delta_{ijk}} \tag{6}$$

In equation (6), the term

$$\frac{h_0(t)}{H_0(t)} = \frac{\alpha t^{\alpha-1}}{t^{\alpha}} = \frac{\alpha}{t} \tag{7}$$

does not include β. Therefore, when we take the derivative of log-likelihood function for the maximum likelihood estimation (MLE) of β, we can ignore the term $\Big(\frac{h_0(t)}{H_0(t)}\Big)^{\delta_{ijk}}$, and consider

$$L_{ijk}(\alpha,\beta) \propto \frac{\mu_{ijk}^{\delta_{ijk}}\exp(-\mu_{ijk})}{\delta_{ijk}!} \tag{8}$$

In equation (8), the term $\frac{\mu_{ijk}^{\delta_{ijk}}\exp(-\mu_{ijk})}{\delta_{ijk}!}$ can be thought of as a random variable of δ_{ijk} following the Poisson distribution with mean μ_{ijk}. Since $\mu_{ijk} = H_0(t)\exp(\beta^{\mathrm{T}}x_{ijk})$, we can further get

$$\log(\mu_{ijk}) = \log(H_0(t)) + \beta^{\mathrm{T}}x_{ijk} \tag{9}$$

which suggests that the estimation of β can be obtained by using the generalized linear regression model with Poisson family and offset $\log(H_0(t)) = \log(t^\alpha) = \alpha\log(t)$.

Then, we consider two cases:

- Case 1: $\alpha = 1$ suggests that the observed survival time follows a exponential distribution, which is a special case of Weibull distribution. Under this case, there is no other unknown parameters in $L_{ijk}(\alpha, \beta)$, and thus β is the only parameter vector that we are going to estimate. Based on equation (9), Segal and Neuhaus [7] used the GEE Poisson regression model to estimate β for the clustered dataset.

- Case 2: $\alpha > 0$ and $\alpha \neq 1$ suggests that the observed survival time follows a general Weibull distribution. Under this case, apart from β, there is another parameter α in $L_{ijk}(\alpha, \beta)$, which is a shape parameter in Weibull distribution. Therefore, we need to do the iteration during the estimation process, in which we update single parameter β or α at each time. Segal and Neuhaus [7] derived a way to estimate α

$$\hat{\alpha} = \frac{\sum_{i=1}^{G}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}}\delta_{ijk}}{\sum_{i=1}^{G}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}}\log(t_{ijk})(\hat{\mu}_{ijk} - \delta_{ijk})} \tag{10}$$

  Details of implementation is shown as follows:

  - **STEP 1:** We set an initial value of α, i.e., $\alpha = \alpha^{(0)}$.
  - **STEP 2:** Update β. For general Weibull distribution, at observed time T_{ijk}, we can get $H_0(T_{ijk}) = T_{ijk}^{\alpha}$. Based on equation (9), we can use GEE Poisson regression model to get the updated $\hat{\beta}^*$. Then, we set $\beta = \hat{\beta}^*$.
  - **STEP 3:** Calculate updated μ_{ijk}.

$$\begin{aligned} \mu_{ijk} &= H_0(t)\exp(\beta^{\mathrm{T}}x_{ijk}) \\ &= T_{ijk}^{\alpha}\exp(\beta^{\mathrm{T}}x_{ijk}) \end{aligned} \tag{11}$$

  - **STEP 4:** Update α. Based on equation (10), we can get the updated $\hat{\alpha}^*$, and then we set $\alpha = \alpha^*$

Repeat STEP 2, STEP 3, and STEP 4 until predetermined convergence condition has reached.

2.2 Estimate the Survival Function by Using Our New Approach

Based on the same parametric assumption mentioned in Section 2.1, from equation (8), we can make the assumption that $\delta_{ijk} \sim \text{Poisson}(\mu_{ijk})$, and $\text{P}(\delta_{ijk}|x_{ijk}, \beta) = \frac{\mu_{ijk}^{\delta_{ijk}} e^{-\mu_{ijk}}}{\delta_{ijk}!}$. To estimate β, based on the idea of the maximum likelihood estimation and equation (8), we can get

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{G} \prod_{j=1}^{m_i} \prod_{k=1}^{n_{ij}} L_{ijk}(\alpha, \beta) \\ &\propto \prod_{i=1}^{G} \prod_{j=1}^{m_i} \prod_{k=1}^{n_{ij}} \mu_{ijk}^{\delta_{ijk}} \exp(-\mu_{ijk}) \end{aligned} \tag{12}$$

and thus the log likelihood function is

$$\begin{aligned} l(\alpha, \beta) &\propto \sum_{i=1}^{G} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} \delta_{ijk} \text{log}(\mu_{ijk}) - \mu_{ijk} \\ &\propto \sum_{i=1}^{G} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} \delta_{ijk} \Big\{ \text{log}\big(H_0(T_{ijk})\big) + \beta^T x_{ijk} \Big\} - \sum_{i=1}^{G} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} H_0(T_{ijk}) \text{exp}(\beta^T x_{ijk}) \end{aligned} \tag{13}$$

$$U(\alpha, \beta) = \frac{\partial l(\alpha, \beta)}{\partial \beta} \tag{14}$$

Similar as case 2 mentioned in Section 2.1, since there are 2 parameters in the log-likelihood function, we need to do the iterations during the estimation process, in which we update single parameter β or α at each time. Details of implementation is shown as follows:

- **STEP 1:** We set an initial value of α, i.e., $\alpha = \alpha^{(0)}$.

- **STEP 2:** Update β. For general Weibull distribution, at observed time T_{ijk}, we can get $H_0(T_{ijk}) = T_{ijk}^{\alpha}$. Based on equation (13), we can further get the score function as

$$\begin{aligned}
U(\alpha,\beta) &= \frac{\partial l(\alpha,\beta)}{\partial \beta} = 0 \\
\Longrightarrow \sum_{i=1}^{G}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} \delta_{ijk} x_{ijk} &- \sum_{i=1}^{G}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} H_0(T_{ijk})\exp(\beta^T x_{ijk}) x_{ijk} = 0 \\
\Longrightarrow \sum_{i=1}^{G}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} \delta_{ijk} x_{ijk} &- \sum_{i=1}^{G}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} T_{ijk}^{\alpha}\exp(\beta^T x_{ijk}) x_{ijk} = 0
\end{aligned} \tag{15}$$

To get the updated estimation β^*, in our new approach, we use the Newton - Raphson method to solve this problem. Then, we set $\beta = \beta^*$.

- **STEP 3:** Calculate updated μ_{ijk}.

$$\begin{aligned}
\mu_{ijk} &= H_0(t)\exp(\beta^{\mathrm{T}} x_{ijk}) \\
&= T_{ijk}^{\alpha}\exp(\beta^{\mathrm{T}} x_{ijk})
\end{aligned} \tag{16}$$

- **STEP 4:** Update α. Based on equation (10), we can get the updated $\hat{\alpha}^*$, and then we set $\alpha = \alpha^*$

Repeat STEP 2, STEP 3, and STEP 4 until predetermined convergence condition has reached.

2.3 Model Diagnostic

2.3.1 Modified Sandwich Estimate

Since for clustered data, the inverse of the information matrix is no longer a valid estimate of the variance $\hat{\beta}$ [12], Binder proposed a "modified sandwich estimate" and showed that it converges to the true variance of $\hat{\beta}$ when the number of the clusters goes to infinity.

By using Taylor series, we can get

$$U(\hat{\beta}) = U\big(E(\hat{\beta})\big) + \frac{\partial U\big(E(\hat{\beta})\big)}{\partial \hat{\beta}}\big(\hat{\beta} - E(\hat{\beta})\big) + o\big(\hat{\beta} - E(\hat{\beta})\big) \tag{17}$$

Then, by Delta Method, the variance of $\hat{beta}$ is

$$\hat{V}(\hat{\beta}) = (I^{-1})V_U(I^{-1})^{\mathrm{T}} \tag{18}$$

where $V_U = \hat{V}[U(\hat{\beta})]$. Binder gave conditions under which equation (18) consistently estimates the asymptotic variance of $\hat{\beta}$.

To derive the cluster covariance matrix of $U(\hat{\beta})$, firstly, we linearize $U(\hat{\beta})$, and then apply a betwee-cluster variance estimator for the linearized statistic. Let

$$Z_{ijk} = x_{ijk}^{\mathrm{T}}\hat{r}_{ijk} \tag{19}$$

where $\hat{r_{ijk}} = \delta_{ijk} - \mu_{ijk}$. Then

$$Z_{ij} = \sum_{k=1}^{n_{ij}} Z_{ijk} \tag{20}$$

The mean square matrix of the associated between-cluster within treatment group is

$$S_z = \sum_i m_i S_{zi}, \tag{21}$$

where m_i denotes the number of clusters in treatment group i and

$$S_{zi} = \sum_j (Z_{ij} - \bar{Z}_i)(Z_{ij} - \bar{Z_i})^{\mathrm{T}}/(m_i - 1) \tag{22}$$

depicting the $p \times p$ matrix of sample mean squares and cross products from treatment group i with

$$\bar{Z}_i = \sum_j \frac{Z_{ij}}{m_i} \tag{23}$$

Following equation (18), the estimated variance for β is given by $\hat{V}(\beta) = I^{-1}S_z(I^{-1})^{\mathrm{T}}$.

2.3.2 Efficiency

To check the underestimation or overestimation of var($\hat{\beta}$), efficiency quantity is used. Let B be the number of simulations, $\beta_i,\ i = 1, 2, ..., p$ be the true value of the coefficients, $\hat{\beta}_{ij},\ i = 1, 2, ..., p,\ j = 1, 2, ..., B$ be the estimated of $\hat{\beta}_i$ in iteration j, and $\hat{\sigma}^2_{i,j}$ be the variance estimate of $\hat{\beta}_i$ in jth simulation after correction which accounts for the correlation of survival times within each cluster. Then, the biases of the variance estimate is evaluated by the following efficiency quantity:

$$r_i = \frac{\hat{\bar{\sigma}}_i^2}{(m(\tilde{\beta}_i))^2 + var(\tilde{\beta}_i)^2} \tag{24}$$

where

$$\begin{aligned}\hat{\bar{\sigma}}_i^2 &= \sum_j \hat{\sigma}^2_{ij}/B \\ \tilde{\beta}_i &= (\hat{\beta}_{i,1} - \beta_i, ..., \hat{\beta}_{i,B} - \beta_i)^{\mathrm{T}}\end{aligned} \tag{25}$$

and $m(\tilde{\beta}_i)$ is the sample mean of $\tilde{\beta}_i$ and $var(\tilde{\beta}_i)$ is the sample variance of $\tilde{\beta}_i$. If $r_i > 1$, then the variance is empirically overestimated, if $r_i < 1$, then the variance is empirically underestimated.

3 Simulation Study

3.1 Positive stable distribution

Firstly, we give a basic introduction of positive stable distribution which is used as part of the data generating process for simulation part. We set

$$\gamma = |1 - i\tan(\pi\alpha/2)|^{-1/\alpha} \tag{26}$$

The earliest example we could find of this formular is in Feller's paper [13], and then Hougard [14] gave the same formula, along with the Laplace transformation:

$$L(s) = E[\exp(-sX)] = \exp(-s^{\alpha}) \tag{27}$$

Based on Samorodnitsky and Taqqu's parameterization [15] where, Weron [16] gave the characteristic function for $\alpha \neq 1$,

$$\varphi(t) = E[\exp(itX)] = \exp\left(i\delta t - \gamma^{\alpha}|t|^{\alpha}\left(1 - i\beta\text{sign}(t)\tan\frac{\pi\alpha}{2}\right)\right) \tag{28}$$

Follow Weron's paper, in our case, we set $\beta = 1$ and $\delta = 0$, we get the characteristic function:

$$\varphi(t) = \exp\left[-\gamma^{\alpha}|t|^{\alpha}\left(1 - i\text{sign}(t)\tan\frac{\pi\alpha}{2}\right)\right] \tag{29}$$

So under Weron's work, we use "**rstable( )**" function in R to generate random varaibles from positive stable distribution with parameters $\alpha = \alpha$, $\beta = 1$, $\gamma = \gamma(\alpha)$, $\delta = 0$ and $pm = 1$.

3.2 Random Data Generating Process

In our case, T_{ijk} is the observed survival time of observation k for individual j in treatment group i conditional on an observed covariate ζ_j, where ζ_j follows positive stable distribution with index α. In other words, we can define another random variable Y_{ijk} to be Weibull distributed with scale parameter $\exp(\beta^T x_{ijk})$ and shape parameter a, then T_{ijk} $= Y_{ijk}\zeta_j^{-1/a}$. Under this data generation, T_{ijk} 's within a cluster are multivariate Weibull with Weibull margins having scale $\exp(\alpha\beta^T x_{ij})$ and shape αa and the correlation between $\log(T_{ijk})$ and $\log(T_{ijl}$) is just $1 - \alpha^2$ for $k \neq l$).

In order to examine the performance of the newly proposed approach for parameter estimation, we vary the cluster size $n_{ij} = 5$, 10 and the number of clusters $m_i = 20$, 50. For simplicity, set shape parameter of Weibull distribution as $a = 2$ and there is only one regression coefficient $\beta = 0$. The index of the positive stable distribution α varies from 0.3 to 0.7 and the censoring percentage is 10% and 20%.

3.3 Simulation Results

For each setting, we compared the performance of our proposed new model with GEE Poisson regression model in Table 1-4.

$m_i = 20$, cencored rate = 10%	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
$\hat{a}$(new approach), scale parameter					
$n_{ij} = 5$	2.21830	2.20807	2.03110	2.02778	2.01692
$n_{ij} = 10$	2.08743	2.00567	2.00872	2.01378	2.02597
$\hat{a}$(GEE Poisson regression), scale parameter					
$n_{ij} = 5$	2.02240	2.01455	2.02730	2.03038	2.01730
$n_{ij} = 10$	2.01523	1.99890	2.01076	2.012516	2.02344
$\hat{\beta}$(new approach)					
$n_{ij} = 5$	0.03657	0.00365	0.00827	0.00136	-0.01071
$n_{ij} = 10$	0.01215	0.02254	-0.01391	0.00965	-0.00187
$\hat{\beta}$(GEE Poisson regression)					
$n_{ij} = 5$	0.02101	0.00104	0.01035	-0.00196	-0.00184
$n_{ij} = 10$	0.01329	0.01761	0.01317	0.01165	-0.00014
variance of $\hat{\beta}$(robust)					
$n_{ij} = 5$	0.03657	0.03901	0.04957	0.05074	0.04618
$n_{ij} = 10$	0.02039	0.02254	0.02602	0.02363	0.02598
variance of $\hat{\beta}$(modified sandwich estimate)					
$n_{ij} = 5$	0.00281	0.01929	0.02355	0.02444	0.02954
$n_{ij} = 10$	0.01193	0.01175	0.01317	0.01369	0.01249

Table 1: Simulation results under $m_i = 20$, censored rate =10%. In this table, a is the shape parameter of Weibull distribution, β is the regression parameter, α is the index of positive stable distribution, n_{ij} is the number of observations in each cluster, m_i is the number of clusters.

$m_i = 20$, cencored rate = 20%	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
$\hat{a}$(new approach), scale parameter					
$n_{ij} = 5$	2.02528	2.01447	2.02286	2.02401	2.02608
$n_{ij} = 10$	2.00999	2.00556	2.01922	2.00692	2.02819
$\hat{a}$(GEE Poisson regression), scale parameter					
$n_{ij} = 5$	2.02821	2.02139	20.2697	2.02444	2.02964
$n_{ij} = 10$	2.01294	2.00642	2.02096	2.00759	2.03124
$\hat{\beta}$(new approach)					
$n_{ij} = 5$	0.01224	0.00319	-0.00412	-0.00345	0.01355
$n_{ij} = 10$	-0.00039	-0.00702	-0.0085	-0.01581	0.02036
$\hat{\beta}$(GEE Poisson regression)					
$n_{ij} = 5$	0.02865	0.01783	0.00851	0.01017	0.01414
$n_{ij} = 10$	0.00135	-0.00068	0.00048	-0.00707	0.02058
variance of $\hat{\beta}$(robust)					
$n_{ij} = 5$	0.05146	0.05306	0.05646	0.05787	0.05736
$n_{ij} = 10$	0.02785	0.02698	0.02404	0.02969	0.02422
variance of $\hat{\beta}$(modified sandwich estimate)					
$n_{ij} = 5$	0.02679	0.02614	0.04663	0.04442	0.03824
$n_{ij} = 10$	0.02551	0.02355	0.01893	0.01594	0.01295

Table 2: Simulation results under $m_i = 20$, censored rate =20%. In this table, a is the shape parameter of Weibull distribution, β is the regression parameter, α is the index of positive stable distribution, n_{ij} is the number of observations in each cluster, m_i is the number of clusters.

$m_i = 50$, cencored rate = 10%	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
$\hat{a}$(new approach), scale parameter					
$n_{ij} = 5$	2.00606	1.99975	2.00881	2.02004	2.01537
$n_{ij} = 10$	2.01353	2.00944	2.00795	2.01151	1.99929
$\hat{a}$(GEE Poisson regression), scale parameter					
$n_{ij} = 5$	2.00039	2.00067	2.00885	2.02158	2.01269
$n_{ij} = 10$	2.00715	2.01011	2.00341	2.00711	1.99846
$\hat{\beta}$(new approach)					
$n_{ij} = 5$	0.00998	0.00169	-0.00144	-0.00223	-0.01259
$n_{ij} = 10$	-0.00612	-0.00407	0.00343	0.01696	0.00879
$\hat{\beta}$(GEE Poisson regression)					
$n_{ij} = 5$	0.00289	0.00192	0.00901	0.00789	-0.00447
$n_{ij} = 10$	0.00328	-0.00301	0.00993	0.01287	-0.00104
variance of $\hat{\beta}$(robust)					
$n_{ij} = 5$	0.02324	0.0222	0.01958	0.01581	0.01801
$n_{ij} = 10$	0.01101	0.01099	0.01076	0.00797	0.00879
variance of $\hat{\beta}$(modified sandwich estimate)					
$n_{ij} = 5$	0.01154	0.02176	0.01582	0.01711	0.01594
$n_{ij} = 10$	0.00581	0.00573	0.01264	0.00959	0.00889

Table 3: Simulation results under $m_i = 50$, censored rate =10%. In this table, a is the shape parameter of Weibull distribution, β is the regression parameter, α is the index of positive stable distribution, n_{ij} is the number of observations in each cluster, m_i is the number of clusters.

$m_i = 50$, cencored rate = 20%	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
$\hat{a}$(new approach), scale parameter					
$n_{ij} = 5$	2.00376	2.01785	2.01143	2.00899	1.99813
$n_{ij} = 10$	2.01238	2.00765	2.00098	2.00162	1.98957
$\hat{a}$(GEE Poisson regression), scale parameter					
$n_{ij} = 5$	2.00113	2.01973	2.01542	2.01214	1.99958
$n_{ij} = 10$	2.014494	2.00974	2.00137	2.00178	1.98936
$\hat{\beta}$(new approach)					
$n_{ij} = 5$	-0.00215	-0.00035	0.00058	0.00143	-0.00684
$n_{ij} = 10$	-0.00249	-0.00242	0.00651	-0.00140	0.00056
$\hat{\beta}$(GEE Poisson regression)					
$n_{ij} = 5$	0.00574	0.00036	-0.00110	0.00442	-0.00551
$n_{ij} = 10$	0.00020	-0.00311	0.00258	0.00635	0.00016
variance of $\hat{\beta}$(robust)					
$n_{ij} = 5$	0.01761	0.01978	0.02158	0.02215	0.02028
$n_{ij} = 10$	0.00941	0.01056	0.00454	0.01097	0.00959
variance of $\hat{\beta}$(modified sandwich estimate)					
$n_{ij} = 5$	0.02657	0.02188	0.01882	0.01703	0.01609
$n_{ij} = 10$	0.01860	0.01149	0.00958	0.00877	0.00788

Table 4: Simulation results under $m_i = 50$, censored rate =20%. In this table, a is the shape parameter of Weibull distribution, β is the regression parameter, α is the index of positive stable distribution, n_{ij} is the number of observations in each cluster, m_i is the number of clusters.

3.4 Conclusions based on simulations

- Several trends

  As we can see, the accuracy of estimations does depend on some variates.

First of all, both models perform better estimation as the sample size grows(either increases m_i or n_{ij}). In the meanwhile we can also detect a decrease in variance from both methods, which again indicates the accuracy goes up with sample size.

Secondly, with the sample size and censored rate fixed we can test the influence of α, the index of positive stable distribution. If we have greater value of α we tend to have higher variance and bias of the estimated $beta$, which can be explained by the property of positive stable distribution (see Weron's [16] proof).

Moreover, the higher censored rate is,the higher the variance of β which is mainly due to loss of information. But further tests should be done to collect more evidence.

- Comparisons between two models

  The main purpose of these experiments is to compare the performance of these two models. Besides the common trends with varying setting, both two models provide significant estimators for β and a. We may conclude that both models are good enough and there is little difference between two estimations, and again there should be further experiments.

  However, the main idea of this paper is that the author developed a modified sandwich estimate for the variance (see 2.3.1) which performs better than usual robust variance. In fact, we can see an overall dominance of this method under various conditions compared to usual robust estimations and it support the author's main idea in this paper. Especially when the sample size is small, the modified sandwich gives much better estimation and it is quite meaningful given the fact that mostly we don't have a large sample in practice.

4 Real Data(CGD) Analysis and Model Comparison

4.1 Data discription

In order to check how the new method works on the real data situation, we run these methods on the CGD data, the well-known Chronic Granulomatous Disease(CGD) dataset. CGD is a group of inherited rare disorders of the immune function characterized by recurrent pyogenic infections which usually present early in life and may lead to death in childhood. Phagocytes from CGD patients ingest microorganisms normally but fail to kill them, primarily due to the inability to generate a respiratory burst dependent on the production of superoxide and other toxic oxygen metabolites.

The data used in Pingfu Fu's paper (our main paper) is a little different from that used by Fleming and Harrington. In this part, we still use the data in Fleming and Harrington's paper, which is a build-in dataset in R. In this dataset, there are 128 individuals (clusters) and we don't consider the treatment group i. We use **tstop** representing the observed survival time T_{jk}; **status** representing the censored indicator δ_{jk}, $1 =$ the interval ends with an infection, $0 =$ the interval does not end with an infection; **treat** representing the only covariates, which could take two values "**0 = placebo**" and "**1 = rIFN-g(gamma interferon)**".

4.2 Model description

We use our proposed model to estimate the covariate coefficient β and the shape parameter a of Weibull distribution, and compare the results with different previous published models, i.e., Segal and Neuhaus's GEE Poisson regression model, and Cox Proportional hazard rates model learned in STAT 633 course. In the end, we use var$(\hat{\beta})$ obtained by new approach and use Wald test to check whether gamma interferon is an important macrophage activating factor for CGD patients.

Note that, in this real data analysis, assume observed survival time T_{jk} follows a Weibull distribution with shape parameter $\alpha = 1$ and scale parameter $\lambda = 1$ for Cox model and Segal's model. However, we do not fix α for the our proposed new model. As the results, we'll get both estimators for covariate coefficient β and shape parameter α.(We still set $\lambda = 1$ to simplified the calculation)

4.3 Results Comparison

Comparison results are shown in Table 5.

methods	$\hat{\beta}$	$se(\hat{\beta})$	$\lvert\hat{\beta}\rvert/se(\hat{\beta})$
New approach	-1.9587	0.5391	3.633
Segal & Neuhaus	-0.829	0.2445	3.391
Cox model	-1.0953	0.2610	4.196

Table 5: Results Comparison

Since the data used in the main paper is a little bit different from our data, the estimator of $\hat{\beta}$ is also a little bit different. Compared with Segal Neuhaus 's GEE Poisson regression model and Cox proportional hazard rates model, our method is still effective to detect significance of the treatment effect(gamma interferon) though the coefficient is underestimated since the index from the positive stable distribution is between 0 and 1. Meanwhile, we can get the estimator of α :

$$\hat{\alpha} = 0.0958$$

Although we find that the new method have the largest variance after our experiment, it's still a reasonable method to obtain the modified robust variance for hypothesis test. Consider a set of hypothesis:

$$H_0 : \beta = 0, \quad H_a : \beta \neq 0$$

By Wald test, test statistics is

$$\chi_w^2 = \frac{(\hat{\beta} - \beta)^2}{se(\hat{\beta})^2} = 3.633^2 = 13.1986 \sim X_1^2$$

the degree of freedom for chi-squared distribution is df = 1. Under 95% confidence level, since $\chi_w^2 = 13.1986 > \chi_1^2 = 3.84$, it is significant to reject null hypothesis, i.e., there is no evidence that gamma interferon is an important macrophage activating factor for CGD patients. Compared with the wald test statistics obtained by other two models, we can make the conclusion that our new proposed model is a reasonable method leading to the correct hypothesis test conclusion.

5 Conclusions and Future Work

Based on the simulation study and real data analysis, we find the new approach based on the iteration of Newton-Rahpson method is very effective. This new method could reduce the bias of estimator significantly, and it's very stable compared with other approaches (the new estimator's sample variance is smaller).

In the simulation case, when the cluster size is large enough, the new approach's variance tends to be small and the bias of these estimator is also converge to 0. According to the results given in the third part, both model performs well when the cluster size and number of clusters is large, and the new approach gives a modified robust estimator of variance. In the end, We confirmed the effective of the new method using our output.

Our current work is based on the assumption that the scale parameter of Weibull distribution equals 1, which means the time of interest T_{ijk} follows $f(t) = \alpha\lambda^\alpha t^{\alpha-1}e^{-(\lambda t)^\alpha}$ with $\lambda = 1$. And this assumption could easily lead to the log-linear model we discussed in part 2.

The future study could look into the case that the scale parameter $\lambda \neq 1$. In this case, the log-linear model that we used will not hold anymore. Meanwhile, the update formula for α will contents more nuisance parameters. More theoretical work is needed to deal with this case. For example, in the simulation study, the $T'_{ijk}s$ within a cluster are multivariate Weibull with Weibull margins having scale $exp(\alpha\beta' x_{ijk})$ and shape αa. Then, after we derive the $\hat{\beta}$ from the regression model, we'll need a formula to reparameterize it to get the true parameter.

References

[1] Norton K. I. et al,. (2001). Chronic radiographic lung changes in children with vertically transmitted HIV-I infection. American Journal of Roentgenology 176, 1553-1558.

[2] Andersen P. K., & Gill R. D. (1982). Cox's regression model for counting processes: a large sample study. The Annals of Statistics 10, 1100-1120.

[3] Prentice R. L., Williams B. J., Peterson A. V. (1981). On the regression analysis of multivariate failure time data. Biometrika 68, 373-379.

[4] Wei L. J., Lin, D.Y., & Weissfeld L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. Journal of the American Statistical Association 84, 1065-1073.

[5] Liang, K.-Y., Self, S. G. and Chang, Y.-c. 'Modeling marginal hazards in multivariate failure-time data', Journal of the Royal Statistical Society, Series B, in press.

[6] Lee, E. W., Wei, L. J. and Ying, Z. 'Linear regression analysis for highly stratified failure time data', Journal of the American Statistical Association, in press. 4

[7] Segal M. R., & Neuhaus J. M. (1993). Robust inference for multivariate survival data. Statistics in Medicine 12, 1019-1031.

[8] Liang K. Y., & Zeger S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22.

[9] Rao, J., & Scott, A. J. (1999). A simple method for analysing overdispersion in clustered Poisson data. Statistics in Medicine 48, 577-585.

[10] Rao, J., & Scott, A. J. (1992). A simple method for the analysis of clustered binary data. Biometrics 48, 577-585.

[11] Bieler G. S., & Williams R. L. (1995). Cluster sampling techniques in quantal response teratology and developmental toxicity studies. Biometrics 51, 764-776.

[12] Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review 51, 279-292.

[13] William Feller (1991). An Introduction to Probability Theory and Its Applications, Volume 2, 2nd Edition

[14] PHILIP HOUGAARD (1986). A class of multivanate failure time distributions. Biometrika, 73(3), 671–678.

[15] Samorodnitsky, Gennady; Taqqu, Murad S (1994). Levy Measures of Infinitely Divisible Random Vectors and Slepian Inequalities. Ann. Probab. 22(4), 1930-1956.

[16] Weron (2001). Levy-stable distributions revisited: tail index > 2 does not exclude the Levy-stable regime Rafał. International Journal of Modern Physics, 12(2), 209-223.