SCHOOL OF STATISTICS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

NONPARAMATRIC HYPOTHESIS TESTING IN CLUSTERED SURIVAL MODEL

By:
John D. Eustaquio and Erniel B. Barrios

UPSS Working Paper No. 2014-05
October 2014

School of Statistics
Ramon Magsaysay Avenue
U.P. Diliman, Quezon City
Telefax: 928-08-81
Email: updstat@yahoo.com

ABSTRACT

We developed a nonparametric test procedure based on bootstrap in testing for the presence of clustering in survival data. Assuming a model that incorporates the clustering effect into the Cox Proportional Hazards model, simulation studies indicate that the procedure is correctly-sized and powerful in a reasonably wide range of scenarios. The test procedure for the presence of clustering over time is also robust to model misspecification. With large number of clusters, the test is powerful even if the data is highly heterogeneous and even if there is misspecification error.

Keywords: bootstrap confidence interval, survival analysis, clustered data, backfitting algorithm, generalized additive models, nonparametric bootstrap

MSC Codes: 62F40, 62N05, 91C20, 62G08

1. Introduction

With the influx of big data mostly in customer-oriented companies, data mining techniques became a common tool for gathering insights and identify patterns that are important decision making instrument towards achieving the company's goals. The company aims to develop strategies that will initially attract customers and then engage them to patronize the products afterwards. This is because in highly competitive sectors, customer recruitment usually entails more costs (engagement incentives) relative to creating promotions geared towards customer retention. Lock engagement within a specific period of time (e.g., loyalty cards, telecommunication, credit cards, etc.) is commonly practiced, this can be

attractive to customers provided that some incentives are available in return. The value of the offer and the lock-up period can be determined from the lifetime profitability of the customer.

The prominence of customer valuation goes along with the growth of usage data together with profile data of customers. The companies deal with customer valuation by searching for patterns from customer behavior. Information on the likely engagement of a customer can be inferred from the historical record of customer transactions, these are information are important to customer relationship management. A credit card company may waive annual fees for their client under the assumption that the customer is expected to be profitable within a year. A subscriber of a telecommunication service be given incentives like free mobile devices or discounts on their bill for as long as they continue with their engagement with the service provider for some duration.

Thus, estimation of lifetime profitability of the customer could provide the crucial inputs in customer relationship management (CRM). Churn of customers is crucial event for telecommunication, financial, other services, and utility sectors. Customer churn is the propensity of customers to cease doing business with a company in a given period of time. The goal of CRM is to characterize customers who would likely churn and to motivate these customers to continue doing business with the company.

Churn is affected by several factors, some may manifest as clustering effect, i.e., members of the same cluster will more likely similar characteristics that affect their churn probabilities. For example, geographical location may affect strength of telecommunication signals, directly influencing churn among a group (cluster) of subscribers

Gauran (2012), proposed the following additive model to characterize the dynamics of customer churn:

$$\log\left[h_i^{(k)}(t)\right] = \lambda_k + log[\lambda_i(t)] + f(x_i) + \varepsilon_i(t) \qquad (1)$$

The logarithm of the response variable within the k[th] cluster on time t, $log\left[h_i^{(k)}(t)\right]$ is expressed as a linear function of λ_k, the cluster-specific random intercept; $log[\lambda_i(t)]$, the baseline risk of the i[th] observation at time t; $f(x_i)$ the smooth function of the covariate of the i[th] observation; and $\varepsilon_i(t)$, the error term. To estimate this model, Gauran (2012) assumed that the propensities to churn of the observations belonging in the same cluster are similar. It is of great interest to know whether there is indeed a clustering effect or that the effect is can be assumed to be similar across clusters.

We developed nonparametric procedure (based on bootstrap) to verify the clustering assumption on Model (1). This will allow drawing of inference without the need to make strong distributional assumptions and without much demand for analytic methods for the determination of the sampling distribution of statistics involved in model (Mooney and Duval, 1993). This is also true for the part of the robustness of the validity of the statistical analysis (Davison and Hinkley, 1997).

2. Inference in Survival Models

The Cox proportional hazard model also known as the Cox's regression model is very popular in survival analysis. The model allows quantification of the relationship between the failure times and a set of explanatory variables, given in the following model (Cox, 1972)

$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)exp(\beta_0\boldsymbol{Z})$, where λ_0 is an unknown baseline hazard function, Z is a p-dimensional covariate, and β_0 is a vector of regression coefficients.

Qin and Jing (2001) evaluated the effectiveness of empirical likelihood (EL) method for the Cox's regression model by comparing it with normal approximation (NA) method using simulation studies. In Qin and Jing (2001), $\lambda_0(s)ds$ where *s* is any given time, is used in one of the formulae for the construction of the confidence region. However, λ_0 is usually not known in the Cox's model and is needed to be estimated.

On the other hand, bootstrap based inference is broadly used if the asymptotic distribution of a test statistic is difficult, if not impossible to derive analytically. These cases arise from having nuisance parameters that affect a test statistic, for instance, due to unobserved heteroskedasticity. Asymptotically pivotal test statistics, bootstrapping procedures often outperform first order asymptotic approximations because of its faster convergence (Beran, 1988). Cai et al (2000) recommend residual based bootstrap approach for the computation of the test statistic when comparing the residual sum of squares from parametric and semiparametric functional regressions.

Herwartz (2009) proposed a factor based bootstrap approach for the hypothesis testing procedure to verify if the functional coefficient is constant. They showed that the approach can cope with various forms of heteroskedasticity as it preserves the relationship between the error term variance and the corresponding regressors. They have also shown that in the framework of semiparametric regressions the factor based bootstrap may be more advantageous than wild, pairs or residual based bootstrap inference.

3. Methodology

Gauran (2012) proposed a nonparametric estimation method in survival analysis and illustrated that it can provide better characterization or churning probabilities in a clustered customer survival data. The functional form of the effect of the covariates is relaxed so that varying effect of the covariates across clusters can be easily accounted into model.

Given *n* clusters with *m* elements observed for *T* time points, the postulated model:

$$\mathbf{log}\left[\boldsymbol{h}_{\boldsymbol{i}}^{(\boldsymbol{k})}(\boldsymbol{t})\right] = \boldsymbol{\lambda_k} + \boldsymbol{log}[\lambda_i(\boldsymbol{t})] + \boldsymbol{f}(\boldsymbol{x_i}) + \boldsymbol{\varepsilon_i}(\boldsymbol{t}) \qquad (2)$$

where $i = 1,2,\dots,m, t = 1,2,\dots,T$

n = total number of clusters

$h_i^{(k)}(t)$ = i[th] value of the response variable within the k[th] cluster at time t

λ_k = cluster-specific random intercept

$\lambda_i(t)$ = baseline risk of the i[th] observation at time t

x_i = value of the covariate of the i[th] observation within the k[th] cluster

$f(x_i)$ = smooth function of x_i (nonparametric)

The model is assumed to include additive components, hence, the backfitting algorithm is used in the estimation. The main interest is on the hypothesis testing about λ_k, i.e., we test the hypothesis $H_o: \lambda_1 = \lambda_2 = \cdots = \lambda_n$ versus H_1: $\lambda_k is\ different\ for\ at\ least\ one\ cluster$. The components are estimated sequentially, the most important one to be estimated first, until all terms in the model are estimated. The backfitting algorithm is known to provide good estimates among the model terms estimated early on in the iterative process (Santos and Barrios, 2012). Thus, the cluster-specific term which is crucial to the hypothesis testing procedure is estimated first. The backfitting estimation procedure is iterated until there are no more significant information for λ_k left on the residuals.

3.1 *Estimating the Clustered Survival Model*

As proposed by Gauran (2012), the modified Cox Proportional Hazard Model was transformed into an additive model so that the components can be estimated separately with the suitable estimation method imbedded in the backfitting algorithm. The clustered survival model is estimated as follows:

1. The function of the covariates, $f(x_i)$ and the baseline risk, $log[\lambda_i(t)]$ are ignored first and the cluster specific term, λ_k is estimated. Since λ_k is a random effect, a mixed effects model is used to estimate $\widehat{\lambda_k}$.
2. The partial residual $\widehat{e_i}$ is computed as $\left(\log\left[h_i^{(k)}(t)\right] - \widehat{\lambda_k}\right)$. At this point, the partial residual contains information on $f(x_i)$ and $log[\lambda_i(t)]$.
3. Ignore $f(x_i)$ further, the partial residual $\widehat{e_i}$ is used to estimate $log[\lambda_i(t)]$. New residuals are computed as $\left(\log\left[h_i^{(k)}(t)\right] - \widehat{\lambda_k} - \widehat{log[\lambda_i(t)]}\right)$. This residual is a function of $f(x_i)$ alone.
4. The residuals in Step 3 is used to estimate $f(x_i)$ using smoothing splines. Finally, the residual is computed as $\left(\log\left[h_i^{(k)}(t)\right] - \widehat{\lambda_k} - \widehat{log[\lambda_i(t)]} - \widehat{f(x_i)}\right)$. The estimate for component to be estimated in the next step is left out for the computation of the residual to ensure that the residual will contain information on that term alone.
5. Iterate until convergence.

3.2 *Hypothesis Testing Procedure for the Presence of Cluster Effect*

Given the clustered survival data, we test the following hypotheses:

$$\mathbf{H_o}: \lambda_1 = \lambda_2 = \cdots = \lambda_k \qquad (3)$$

$$\mathbf{H_a}: \boldsymbol{\lambda_i} \neq \boldsymbol{\lambda_j} \ for\ at\ least\ one\ pair\ i \neq j$$

The following procedure is proposed for testing the hypotheses in (3) assuming Model 2.

Algorithm:

1. Estimate the clustered survival Model in (2) ignoring clustering first, i.e., assume $\mathbf{H_o}$ is true.
2. Estimate λ_0, the homogenous random intercept under $\mathbf{H_o}$.
3. Generate *r* bootstrap resamples of size *n* from the residuals after Steps 1 and 2.
4. Estimate the model for each of the resamples.
5. Collect $\hat{\lambda}_k$'s from each resamples.
6. Sort $\hat{\lambda}_k$ (i. e $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3$) in either ascending or descending order.
7. Compute the percentiles from Step 6 to construct a (1-α) % confidence interval.
8. Reject the null hypothesis that there is no clustering effect at α% level of significance if at least one of the intervals in Step 7 fail to contain λ_0.

The above procedure makes use of the different survival data for each cluster to represent the variation in terms of the cluster-specific component of the model. Resampling to produce the nonparametric estimate of Cox Proportional Hazards model (with random intercept) allows us to approximate the sampling distribution of the cluster-specific parameter estimator to evaluate clustering effect. This bootstrap method is less sensitive to model misspecification.

The proposed procedure for testing constant random effect across all clusters is applicable to any model that will fill the clustered survival data. The comparison of clustering effect across all clusters, however, is based only on one parameter (λ_k) in the model.

3.3 *Simulation Studies*

A simulation study is designed to evaluate the power and size of the test procedure for clustering effect in a survival model. Table 1 shows the simulation boundaries employed in the study.

Table 1. Boundaries of Simulation Study

1	Number of Clusters	Small – 3 Large – 10
2	Distribution of the error term	ε ~ N(0,1) ε ~ N(0,10)
3	Misspecification *m* in the model	m = 1, m = 5
4	Distribution of $\delta_i(t)$	$\delta_i(t) \sim Exp(1)$
5	Distribution of x_i	x_i ~ Weibull (1,2)
6	Contribution of the terms in the model	a. Equal Contribution b. Dominating Cluster Effect c. Dominating Baseline Risk d. Dominating Covariate
7	Distribution of λ_k	(see table 2)

Gauran (2012) noted that the overall behavior of the model is confounded with the baseline hazard function. Without loss of generality, $log[\lambda_i(t)]$ is assumed to follow $\gamma e^{-\gamma t} + \delta_i(t)$ where $\gamma = 0.5$. The scale parameter of the exponential distribution is the constant churn rate. Even if the number of observations (cluster size) in each cluster is constant, the number of clusters is allowed to vary (small=3, large=10). These simulation settings aims to account the fact that adding clusters not individual observations leads to an increase in efficiency of estimates of models (Arceneaux and Nickerson, 2009).

The contribution of the three components λ_k, x_i and $log[\lambda_i(t)]$ are also considered to verify if the test procedure would still yield correct inference regardless of the contribution of λ_k on the total variation in the model.

Clusters are assumed to follow Poisson distribution with varying means as indicated in Table 2. These settings are included to account for various cases spread of the cluster distributions. Settings N and O were included to measure the size of the test.

Table 2. Mean of the Poisson Distributed Cluster-Specific Component

Scenario		Number of Clusters	Cluster Label									
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	With Clustering Effect	Small k=3	5	6	7	----------						
B			5	7	9	----------						
C			5	9	13	----------						
D			5	5	8	----------						
E			5	5	10	----------						
F		Large k=10	5	6	7	8	9	10	11	12	13	14
G			5	7	9	11	13	15	17	19	21	23
H			5	5	5	5	5	5	5	5	5	7
I			5	5	5	5	5	5	5	5	5	10
J			5	5	5	5	5	5	6	6	6	6
K			5	5	5	5	5	5	10	10	10	10
L			5	5	5	7	7	7	7	9	9	9
M			5	5	5	6	6	6	9	9	9	9
N	Without Clustering Effect	Small	5	5	5	----------						
O		Large	5	5	5	5	5	5	5	5	5	5

Presence of misspecification error is simulated by multiplying the error term with some large constant to inflate their variance. A multiplier value *m* =5 in the error term represents cases of model misspecification.

Finally, the covariate x_i is generated from Weibull distribution with shape parameter of 1 and scale parameter of 2. The Weibull distribution was chosen because it is used to model a variety of lifetime distributions analogous to the characteristics of the covariates (Gauran, 2012).

4. Results and Discussion

For each of the simulation boundaries stated above, the results were presented by size of the cluster (n=10 and n=30 for small and large cluster size respectively). Misspecification error (m=5) and the variance of the error term are also included in summarizing the results.

4.1 Power of the Test

The power of the proposed test is computed for the scenarios with small or large number of clusters where at least one of the means of the cluster-specific term is different from the other. For the computation of the power, scenarios where the null hypothesis is indeed false were generated and the proportion of cases that resulted to the rejection of the null hypothesis is considered in computing power of the test.

Small Number of Clusters

Five major scenarios were considered in measuring the power of the test for small number of clusters (k=3). Ranging from those with the distribution of the cluster-specific terms that have comparable means up to those with one cluster with cluster-specific term far away from the mean of the other two clusters.

Mean Increments by 1 for Each Succeeding Cluster

Table 3 shows that the power of the test is at least 0.90 for cases where there is no misspecification error and variance of the error term is small. In cases where there is

misspecification and the variance of the error term is high, power of the test is also near 0.8. The test is relatively more powerful in most cases, especially when the sample size is large.

Table 3. Power of the test when the mean increments by 1 for each succeeding cluster

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.93	0.98	0.85	0.93
			m = 5	0.91	0.94	0.84	0.83
	α = 0.05		m = 1	0.92	0.95	0.85	0.89
			m = 5	0.90	0.94	0.82	0.81
	α = 0.01		m = 1	0.91	0.94	0.79	0.86
			m = 5	0.87	0.93	0.81	0.80

Mean Increments by 2 for Each Succeeding Cluster

When the means of the cluster specific term are moderately far from each other, the power of the test are mostly above 0.9. The power is fairly robust to misspecification error when the variance of the error term is low. When there is high variance in error term and misspecification is present, the power dropped significantly but is maintained above 0.8.

Table 4. Power of the test when the mean increments by 2 for each succeeding cluster

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.96	0.98	0.87	0.95
			m = 5	0.92	0.96	0.85	0.87
	α = 0.05		m = 1	0.95	0.98	0.86	0.91
			m = 5	0.91	0.95	0.83	0.84
	α = 0.01		m = 1	0.94	0.97	0.82	0.90
			m = 5	0.90	0.94	0.81	0.81

Mean Increments by 4 for Each Succeeding Cluster

When clusters are distinctly different from each other, the test yield very high power approaching one in almost all cases. The test also exhibited robustness to misspecification error and with large variability in the error term.

Table 5. Power of the test when the mean increments by 4 for each succeeding cluster

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.99	1.00	0.97	0.98
			m = 5	0.94	0.99	0.91	0.95
	α = 0.05		m = 1	0.98	0.99	0.97	0.98
			m = 5	0.94	0.98	0.90	0.94
	α = 0.01		m = 1	0.97	0.98	0.95	0.96
			m = 5	0.94	0.97	0.90	0.94

Two Clusters have the same effect and the other is slightly different

For the case where 2 among the 3 clusters have the same mean for the cluster-specific term, the power of the test is still fairly high, even with misspecification error. The difference in cluster size also did not result to drastic changes in power of the test.

Table 6. Power of the test when two clusters have the same effect and the other is slightly different

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.82	0.83	0.79	0.80
			m = 5	0.78	0.82	0.76	0.76
	α = 0.05		m = 1	0.79	0.79	0.75	0.78
			m = 5	0.75	0.77	0.73	0.75
	α = 0.01		m = 1	0.77	0.79	0.74	0.74
			m = 5	0.71	0.75	0.66	0.67

Two Clusters have the same effect and the other is quite different

The power of the test for most of the cases where the means of two out of the three cluster-specific term are the same and the other one is quite far from the other two are very high. Still consistent with the literature, the lowest power of the test for this scenario is observed when there is misspecification and the variability in error term is high.

Table 7. Power of the test when two clusters have the same effect and the other is quite different

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.93	0.97	0.88	0.89
			m = 5	0.90	0.91	0.81	0.85
	α = 0.05		m = 1	0.91	0.97	0.87	0.89
			m = 5	0.89	0.90	0.81	0.85
	α = 0.01		m = 1	0.90	0.94	0.87	0.87
			m = 5	0.86	0.88	0.79	0.84

Large Number of Clusters

For the scenario with fairly large number of clusters (k=10), 8 scenarios were considered in assessing the power of the test. Ranging from those with the distribution of the cluster-specific term that have really close means, to those with one cluster that has the mean of the cluster-specific term very far away from the mean of the other 9 clusters.

Mean Increments by 1 for Each Succeeding Cluster

As the mean of the cluster-specific term is increased by one for each cluster in the case where there is a large number of clusters, the power is close to 1 in most cases except for the scenario where there is misspecification and large variability is error term.

Table 8. Power of the test when the mean increments by 1 for each succeeding cluster

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	1.00	1.00	1.00	1.00
			m = 5	1.00	1.00	0.91	0.96
	α = 0.05		m = 1	0.99	1.00	0.99	0.99
			m = 5	0.97	0.99	0.91	0.96
	α = 0.01		m = 1	0.97	0.99	0.97	0.98
			m = 5	0.97	0.97	0.91	0.96

Mean Increments by 2 for Each Succeeding Cluster

When the mean of the cluster-specific terms is increased by 2 for each of the succeeding clusters, the power of the test is equal to 1 in most cases. Even when there is misspecification, the power of the test is at least 0.95.

Table 9. Power of the test when the mean increments by 2 for each succeeding cluster

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	1.00	1.00	1.00	1.00
			m = 5	1.00	1.00	0.98	0.99
	α = 0.05		m = 1	1.00	1.00	1.00	1.00
			m = 5	1.00	1.00	0.96	0.98
	α = 0.01		m = 1	1.00	1.00	0.99	1.00
			m = 5	1.00	1.00	0.95	0.96

All clusters have the same effect except for one that is slightly different

As can be seen in Table 10, the power of test is higher than 0.5 in most cases when all the clusters have the same effect except for one that is slightly different. The power increases when there is no misspecification and the variability of the error term is low.

Table 10. Power of the test when all clusters have the same effect except for one that is slightly different

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,1)	
Sample Size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.71	0.71	0.57	0.61
			m = 5	0.50	0.66	0.44	0.47
	α = 0.05		m = 1	0.70	0.72	0.58	0.62
			m = 5	0.47	0.65	0.42	0.43
	α = 0.01		m = 1	0.66	0.70	0.56	0.59
			m = 5	0.48	0.66	0.43	0.43

All clusters have the same effect except for one that is quite different

When the mean of the cluster-specific term is the same for most of the clusters while the mean of one cluster is really far, the power is close to 0.9 in most cases except for the scenario where there is misspecification error and large error variability. While power declines in cases with misspecification error and large error variability, this is still high in the vicinity of 0.85.

Table 11. Power of the test when all clusters have the same effect except for one that is quite different

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.92	0.93	0.88	0.89
			m = 5	0.90	0.90	0.83	0.86
	α = 0.05		m = 1	0.90	0.91	0.87	0.89
			m = 5	0.87	0.89	0.82	0.86
	α = 0.01		m = 1	0.89	0.90	0.86	0.88
			m = 5	0.87	0.88	0.82	0.85

Clusters have two inherent grouping effect that are close to each other

In reference to table 12, the power of test is low in most cases when the cluster means are almost similar for all clusters. Consistent with the results of the other scenarios,

the power increases when there is no misspecification error and when error variability is low.

Table 12. Power of the test when the clusters have two inherent grouping effect that are close to each other

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,1)	
Sample Size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.73	0.79	0.56	0.68
			m = 5	0.51	0.71	0.49	0.48
	α = 0.05		m = 1	0.74	0.74	0.62	0.64
			m = 5	0.50	0.65	0.40	0.50
	α = 0.01		m = 1	0.74	0.75	0.58	0.61
			m = 5	0.47	0.62	0.43	0.46

Clusters have two inherent grouping effect that are far from each other

When cluster-specific mean for large number cluster groups into two values very far from each other, the power of the test is close to 0.9 in most cases except for the scenario where there is misspecification error and large error variability. In those cases where there is misspecification and the variability is high, the power of the test decreased but is still over 0.8.

Table 13. Power of the test when the clusters have two inherent grouping effect that are far from each other

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,1)	
Sample Size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.94	0.99	0.87	0.90
			m = 5	0.89	0.93	0.78	0.88
	α = 0.05		m = 1	0.88	0.98	0.85	0.87
			m = 5	0.82	0.87	0.83	0.85
	α = 0.01		m = 1	0.87	0.94	0.82	0.87
			m = 5	0.86	0.82	0.76	0.81

Clusters have three inherent grouping effect that are close to each other

Table 14 shows that the power of the test is close to 0.7 when the means of the cluster-specific terms are similar, in fact, some of the clusters even have the same mean. In cases where there is misspecification error and the variance of the error term is high, the power of the test is around 0.6. The test is also relatively more powerful in most cases with large sample size.

Table 14. Power of the test when the clusters have three inherent grouping effect that are close to each other

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,1)	
Sample Size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.88	0.87	0.80	0.81
			m = 5	0.74	0.79	0.78	0.75
	α = 0.05		m = 1	0.81	0.86	0.72	0.76
			m = 5	0.73	0.80	0.72	0.71
	α = 0.01		m = 1	0.78	0.77	0.68	0.71
			m = 5	0.68	0.73	0.69	0.68

Clusters have three inherent grouping effect that are far from each other

The power is at least 0.6 for most of the cases where the means of the cluster-specific term are almost the same for most of the cluster while the other clusters have equal means, though their mean is quite far from the other group of clusters. Still consistent with the literature, the lowest power of the test for this scenario is attained when there is misspecification error and the error variability is high.

Table 15. Power of the test when the clusters have three inherent grouping effect that are far from each other

Distribution of error term				ε ~ N(0,1)		ε ~ N(0,1)	
Sample Size				n=10	n=30	n=10	n=30
Level of Significance	α = 0.1	*misspecification (m)*	m = 1	0.85	0.86	0.63	0.66
			m = 5	0.69	0.78	0.49	0.59
	α = 0.05		m = 1	0.80	0.84	0.64	0.71
			m = 5	0.57	0.76	0.46	0.55

	α = 0.01		m = 1	0.73	0.83	0.65	0.69
			m = 5	0.59	0.75	0.46	0.47

4.2 Size of the Test

The proposed test is also assessed for its size by using two scenarios, one with small and the other one with large number of clusters where the cluster-specific term all comes from the Poisson distribution with the same means. To evaluate correct sizing of the test, settings where the null hypothesis is true, i.e., there is no clustering effect, were simulated. If not more than 100*α of the results rejected the null hypothesis, the test is considered correctly-sized.

Small number of clusters

In the cases where there is really no clustering effect in small number of clusters (k=3), the test is correctly sized in all cases at 10% level of significance even if there is misspecification error and the data is highly heterogenous. The size of the test for this scenario is still correct even if the cluster size is small (n=10).

For small number of clusters, at 5% level of significance, the test is still correctly sized for most cases where the cluster–specific term comes from the same distribution in all of the three clusters. The test is incorrectly sized when the baseline risk dominated the model and when the cluster size is small. This is due to the fact that when there are more observations in a clusters, there would be larger chance of having extreme values on those clusters that would tend to influence the test to reject the null hypothesis. Also, the test is incorrectly-sized when the covariates have the largest contribution in the model and there

is misspecification error. This could be attributed to the fact that misspecification error contributes in the increase of heterogeneity of the data.

Table 16. Size of the test for the settings when there are small number of clusters (k=3)

Level of Significance				α = 0.10				α = 0.05				α = 0.01			
Distribution of error term				ε ~ N(0,1)		ε ~ N(0,10)		ε ~ N(0,1)		ε ~ N(0,10)		ε ~ N(0,1)		ε ~ N(0,10)	
Sample size				n=10	n=30	n=10	n=30	n=10	n=30	n=10	n=30	n=10	n=30	n=10	n=30
Dominating Term	Equal	misspecification (m)	m = 1	6	3	5	9	0	2	3	5	0	0	0	1
			m = 5	6	7	9	10	4	2	5	4	1	2	1	3
	Cluster Effect		m = 1	3	7	6	6	0	3	1	5	0	0	0	0
			m = 5	6	6	9	6	3	1	4	2	0	1	1	1
	Baseline Risk		m = 1	3	7	3	8	2	6	0	8	0	2	0	3
			m = 5	4	10	5	9	1	8	1	7	1	5	1	5
	Covariates		m = 1	7	9	9	10	5	4	5	6	0	3	1	3
			m = 5	9	7	5	10	3	6	1	7	0	3	1	4

Large number of clusters

The test is correctly-sized for most of the cases where there are large number of clusters and the distribution of the cluster-specific term is the same across all ten clusters with 10% level of significance. When there is a small number of clusters, the test became incorrectly-sized when the baseline risk of the function of the covariates dominates the model and there is misspecification error in a highly heterogeneous data.

In reference to table 17, the test is correctly sized when the cluster-specific term dominates the model or the model components have similar contributions. In case where the baseline risk dominates the model, the test is incorrectly sized when the cluster size is large at 5% level of significance. Also consistent with the results of the other scenarios, the

test is incorrectly-sized when there is misspecification error and/or the cluster size is large for the cases where the function of the covariates dominates the model.

In cases where the mean of the cluster-specific term comes from the same distribution, with large number of clusters, the test is correctly sized in most of the cases even at 1% level of significance.

Table 17. Size of the test for the settings when the number of clusters is large (k=10)

Level of Significance				α = 0.10				α = 0.05				α = 0.01			
Distribution of error term				$\varepsilon \sim N(0,1)$		$\varepsilon \sim N(0,10)$		$\varepsilon \sim N(0,1)$		$\varepsilon \sim N(0,10)$		$\varepsilon \sim N(0,1)$		$\varepsilon \sim N(0,10)$	
Sample size				n=10	n=30	n=10	n=30	n=10	n=30	n=10	n=30	n=10	n=30	n=10	n=30
Dominating Term	***Equal***	**misspecification (m)**	*m = 1*	8	9	6	9	2	4	3	3	0	0	0	2
			m = 5	7	8	7	10	2	5	4	8	0	3	1	5
	Cluster Effect		*m = 1*	9	5	9	8	5	3	5	5	0	0	0	0
			m = 5	8	6	8	9	4	4	3	5	0	1	1	3
	Baseline Risk		*m = 1*	9	10	4	10	3	6	1	7	0	4	0	6
			m = 5	5	8	8	13	3	6	4	9	1	5	1	3
	Covariates		*m = 1*	7	7	10	10	2	5	5	6	1	2	1	4
			m = 5	6	9	7	15	4	7	5	10	1	5	1	8

5. Conclusions

The proposed test can correctly identify whether clustering effect is present or not. Simulation studies indicate that the procedure is correctly sized and powerful in a reasonably wide range of data scenarios. The power of the test is higher (close to 1 or equal to 1) for distant alternative values, regardless of the number of clusters. In cases where the clusters are near each other for as long as there is at least one cluster different from the rest, the test still resulted in high power. The test procedure for constant cluster effect over time is also robust to misspecification errors, especially when the cluster size is large. In survival data characterized by large number of clusters, the test is powerful even if the data is highly heterogenous and/or there is misspecification error.

References

Arceneaux, K. and Nickerson, D. (2009). Modeling Certainty with Clustered Data: A comparison of Methods. *Political Analysis*, 17, 177-190.

Beran, R., (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83, 687–697.

Cai, Z., Fan, J., Yao, Q., (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95, 941–956.

Cox, D. R. (1972). Regression model and life tables (with discussion). *Journal of Royal Statistical Society B*, 34,187–220.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and their Application. Cambridge: Cambridge University Press.

Gauran, I. M. (2012). Nonparametric Modeling of Clustered Customer Survival Data. MS Thesis, University of the Philippines.

Herwartz, H. and Xu, F. (2009). A new approach to bootstrap inference in functional coefficient models. *Computational Statistics and Data Analysis*, 53, 2155–2167.

Mooney,C.Z. and Duval,R.D. (1993). Bootstrapping: A Nonparametric Approach to Statistical Inference. CA: SAGE Publications, Inc.

Qin, G. and Jing, B. (2001). Empirical likelihood for Cox regression model under random censorship. *Communications in Statistics-Simulation and Computation*, 30,79–90.

Santos, E. and Barrios, E. (2012). Decomposition of Multicollinear Data and Time Series using Backfitting and Additive Models. Communications in Statistics - Simulation and Computation, 41:8, 1693-1710.