Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Shu et al., J Biom Biostat 2018, 9:4
DOI: 10.4172/2155-6180.1000410

Research Article Open Access

Study on Outlier Detection Method in Survival Analysis: Weibull Regression Outlier Model

Chang Shu[1,2], Tingting Qin[2], Xiaoping Chen[3*#] and Ping Yin[2*#]

[1]*Hubei Province for the Clinical Medicine Research Center of Hepatic Surgery, Key Laboratory of Organ Transplantation, Ministry of Education and Ministry of Public Health, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China*
[2]*Department of Epidemiology and Biostatistics and State Key Laboratory of Environment Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China*
[3]*HUST, Hubei Province for the Clinical Medicine Research Center of Hepatic Surgery; Key Laboratory of Organ Transplantation, Ministry of Education and Ministry of Public Health, Hepatic Surgery Centre at Tongji Hospital, Tongji Medical College, Wuhan 430000, China*
[#]*Authors equally contributed to this work and should be considered as co-corresponding authors*

Abstract

Background: This study intends to construct Weibull regression outlier model with outlier and using Bayesian method to get parameters estimation and statistical inference. The proposal model may contribute to further thoroughly and systematically complement and implement of outlier detection methods in survival analysis and fully excavate and utilize the survival data.

Method: We construct the Weibull regression outlier model by introducing an n-dimensional shift vector as an outlier indicator to the traditional Weibull regression model. The Bayesian method is used for parameters estimation and MCMC method is used for statistical inference. The prior for γ is conditional Laplace distribution and the point estimation of γ is posterior median. According to confidence interval criterion, the components of γ whose 50% confidence interval contained 0 are shrank to 0. Then the nonzero components of γ are supposed to be outliers.

Results: The results of simulation study and real example study show that the proposal models are not sensitive to censor rate of data and the ratio of outlier would slightly influence the accuracy of proposal models. The estimations of coefficient of outlier models are robust.

Conclusion: The outliers in survival data may contain the new information related to the prognosis of disease which has not been discovered yet. By the proposal WROM, we could achieve outlier detection and parameter robust estimation at the same time.

Keywords: Survival analysis; Outlier; Bayesian method; Weibull distribution; Lasso

Background

Outlier is an inevitable problem since people came to understanding the world by gleaning the data from it. In Hample's survey [1], he proposed the theory that the proportions of outlier in datasets are usually ranging from 1% to 10% or even more. As we known, outlier would negatively affects our results [2]. The rough method is deleting the suspect data points by some classic rejection rules [3], such as Thompson's rule, Grubbs rule, Dixon rule, and so on. Even if the suspect data point could be a normal data point with low probability. Because people believe that compare to deleting a normal data, keeping an outlier might do more harm than good [4]. However, with the in-depth understanding of outlier, the hidden or neglected value of outlier have been realized gradually. Sometimes hidden behind the outlier is unexplored, entirely new knowledge which may help us to discovery new things and push forward the development of science. Hawkins's definition of outlier is relevant and widely accepted. He thinks outlier is the subject which differs from other subjects so greatly that we have to suspect it may be produced by different mechanism [5]. The statisticians had made great efforts to mine the outlier. For now, the method to mine the outlier could been divided into five types: statistical-based [6,7], depth-based, distant-based [8], density-based [9] and cluster-based. Each method has its specific use of scope, but all above five methods are only applicable in complete data.

The outlier problem may also exists in survival data. Intuitively, the subject whose survival time is considerably longer or shorter than predicted may be considered as the outlier. The new information behind these outliers would be unexplored protective or risk factors related to diseases prognosis. In a manner of speaking, the outlier in survival data may help us have a deeper understanding of the diseases.

In practical work, especially in medical study, the survival data is usually incomplete. It may contain censored data, that is, for a variety of reasons, we can't observe the failure event occurred, and we only obtain part of the survival information. It was clear that the outlier detection methods mentioned above are not suitable for this situation. In addition, the censored data would conceal up the outlier in part, and the information for outlier detection in survival data would be more covert.

***Corresponding authors:** Ping Yin, Department of Epidemiology and Biostatistics and State Key Laboratory of Environment Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, Tel: 8613098818499; E-mail: pingyin2000@126.com

Xiaoping Chen, Hepatic Surgery Center, Hubei Province for the Clinical Medicine Research Center of Hepatic Surgery, Key Laboratory of Organ Transplantation, Ministry of Education and Ministry of Public Health, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, Tel: +86 27-83663500; E-mail: chenxp@medmail.com.cn

Received August 27, 2018; **Accepted** September 17, 2018; **Published** September 21, 2018

Citation: Shu C, Qin T, Chen X, Yin P (2018) Study on Outlier Detection Method in Survival Analysis: Weibull Regression Outlier Model. J Biom Biostat 9: 410. doi: 10.4172/2155-6180.1000410

During the last century, statisticians attempted to detect the outlier in survival data by using the order statistic of survival time to construct certain statistic. Failing to account for covariance related to survival time, these methods are rough. In recent years, scholars attempted to mine the outliers on the basis of survival model which contain the covariance. Nardi and Schemper [10] presented a more rational method. In general, we can evaluate the predicted effect of COX model by comparing the estimated survival rate and 0.5. According to this idea, they formed Log-odds residual and normal deviation residual respectively and utilized the distribution information of the residuals to detect outliers. Eo [11] utilized the residual to explore the outlier in quantile regression model.

She and Owen [12] innovatively presented a method for outlier detection in Means Shift Outlier Model. They used the nonconvex penalty function to get the shrinkage estimation of shift vector. Then the non-zero element of shift vector prompt the subject would be an outlier. The advantage of this method is not only the ability to deal with multiple outliers, but it could also give the robust parameter estimation at the same time. We found that, in most parametric survival model, there is a linear relationship between covariance and function of hazard function. Shall we construct an outlier model based on parametric survival model and use the nonconvex penalty function to detect outlier? This is the main subject we discussed in this study.

Methods

Model

Let T denote the survival time, assume that, t_i (i=1,…, n) are independent and follow a Weibull distribution with scale parameter λ and shape parameter ω. Then for the ith subject (i=1,…, n), the survival function and hazard function are as follows:

$$S(t_i|\lambda,\omega)=1-F(t_i|\lambda,\omega)=exp(-\lambda t_i^{\omega})$$

$$h(t_i)=\frac{f(t_i|\lambda,\omega)}{S(t_i|\lambda,\omega)}=\lambda\omega t_i^{\omega-1}$$

Let C denote the censor time and δ is indicator of censor. If the failure event is observed then=1, otherwise=0. Assume that, the censoring mechanism is random. Then the observed time $y=(y_1,\ldots,y_n)$ are:

$$y_i=\begin{cases} t_i & \delta_i=1 \\ c_i & \delta_i=0 \end{cases} \quad i=1,\ldots,n$$

Let us suppose that the outlier is a result of shifted of λ, then we constructed the outlier model as follow:

$$\lambda=exp(\boldsymbol{X}'\beta+\gamma)$$

Where $\boldsymbol{X}$ is the covariance matrices, $\boldsymbol{\beta}$ is the unknown regression coefficient and $\boldsymbol{\gamma}$ is the unknown outlier indicator vector. Then the likelihood function of Weibull regression outlier model (WROM) is:

$$L(\beta,\gamma,\omega|y,x,\delta)=\prod_{i=1}^{n}\left[\omega exp(X_i'\beta+\gamma_i)y_i^{\omega-1}exp(-\lambda exp(X_i'\beta+\gamma_i)y_i^{\omega-1})\right]^{\delta_i}\times$$

$$\left[exp(-exp(X_i'\beta+\gamma_i)y_i^{\omega})\right]^{1-\delta_i}$$

Parameter estimation and outlier detection

The most common methods of parameter estimation of traditional Weibull regression model (WRM) are maximum likelihood estimation, EM algorithm, Bayes, and so on. However, the likelihood function of WROM is very complicated and the unknown parameters are high dimensional. If we adopt maximum likelihood estimation or EM algorithm, we have to face the complicated calculation and statistical inference problem. Therefore, we adopt Bayes method to parameter estimation and MCMC method to statistical inference.

For the priors of β and ω we refered to traditional Weibull regression model. We assume that β_i(i=1,…, p) are independent and follow a flat normal distribution, ω follows a flat Gama distribution when we know little about β and ω [13-15]. The discussion focuses on γ.

The outlier should be a small part of dataset. Obviously, we can assume that γ is sparse, that is the most elements of γ are zero and non-zero elements of γ prompt the subjects could be outliers. It is noteworthy that our interests are the non-zero elements of γ. We expect the appropriate prior which make every element of γ have high probability to be zero and every element of γ have the same probability to be non-zero. Finally, we can using the sparse solution of γ to achieve outlier detection. This thinking is similar to variable selection. In other words, we could draw lessons from variable selection method to outlier detection.

Bayesian LASSO [16] could get sparse solution as well as robust parameter estimation, therefore we employed Bayesian LASSO to obtain solution of γ. We could write the hierarchical Bayesian LASSO model as follow:

$$\boldsymbol{\gamma}\,|\,\sigma^2,\tau_1^2,\ldots,\tau_n^2\sim\mathcal{N}_n(0_n\sigma^2,D_\tau)$$

$$D_\tau=diag(\tau_1^2,\ldots,\tau_n^2)$$

$$\tau_1^2,\ldots,\tau_n^2\sim\prod_{j=1}^{n}\frac{\alpha^2}{2}exp(-\lambda^2\tau_j^2/2)d\tau_j^2$$

$$\sigma^2\sim\pi(\sigma^2)d\sigma^2$$

After integrating the $\tau_1^2,\ldots,\tau_n^2$, the prior of γ is as follow:

$$\pi(\boldsymbol{\gamma}|\sigma^2)=\prod_{i=1}^{n}\frac{\alpha}{2\sqrt{\sigma^2}}exp(-\alpha|\gamma_i|/\sigma^2)$$

Where σ^2 and α^2 are hyper-parameters. The former guarantee the posterior of γ is unimodal and the latter control the shrink degree of γ. We could place a hyper-prior upon them (inverse Gamma distribution for σ^2 and Gamma distribution for α^2), then the uncertainty of these two hyper-parameters would be taken into account during parameter estimation. The joint posterior distribution of parameters is as follow:

$$P(\beta,\gamma|y,x,\delta)\propto L(\beta,\gamma|y,x,\delta)\times\pi(\beta)\times\pi(\gamma)\times\pi(\omega)$$

Different from ordinary LASSO, Bayesian LASSO can't shrink parameter to 0 directly. But Bayesian LASSO provides interval estimates that we can used to outlier detection. According to Confidence Interval Criterion proposed by Li and Lin [17] , if the 50% CI of γ_i contain 0, we let it be 0. On the contrary, if the 50% CI of γ_i don't contain 0, we identified it as outlier.

Results

Simulation study

To assess the effectiveness of the proposed model, we conducted simulation study and employed R, M and S to evaluate the accuracy of the outlier detection. Definitions of these three indexs are as follows:

$$R=\frac{\text{correctly identified outlier}+\text{correctly identified normal data}}{\text{sample size}}$$

$$M = \frac{\text{outlier be identified as normal data}}{\text{true outlier}}$$

$$S = \frac{\text{normal data be identified as outlier}}{\text{true normal data}}$$

Apparently, for desired outlier detection method, R should approximate to 100% and S and M should be close to 0%. The effectiveness of parameter estimation would be evaluated by Mean, SD and MSE of the parameters. To fully evaluate the effectiveness of parameter estimation, we simulated three different situations.

$\hat{\boldsymbol{\theta}}_1$: Modelling all the simulated data with WROM;

$\hat{\boldsymbol{\theta}}_2$: After deleting the outliers that identified by WROM, modelling the remainder data with WROM;

$\hat{\boldsymbol{\theta}}_3$: Modelling all the simulated data with WRM.

The procedure of simulated data generating is as follow. Firstly, let all the γ_i to be 0. Then we randomly choose p components of them, half of which are set to be 5 and the other half of which are set to be -5. Then we generated the covariance $X=(x_1,x_2)$ from uniform distribution between [0,1] and binomial distribution with probability of success of 0.5 respectively. Next, we set β=(0.5,1) and ω=1.5. The survival time T would generated by the proposed WROM. The censored time C was generated by a similar Weibull distribution. If $t_i \le c_i$, then let δ_i=1, and $y_i=t_i$; if $t_i>c_i$, let δ_i=0 and $y_i=c_i$. In simulation study, we consider several different situations. N set to be 500 and 1000, the censored rate set to be 20% and 40%, and proportion of outlier set to be 10% and 20%. In every situation, we repeat simulation 500 times.

The results of outlier detection were showed in Table 1. In all situations R was higher than 96%, and S and M were very low. When sample size is constant, R slightly decreased with proportion of outlier, S and M don't show significant change. The outlier detection effect improved with the rise of sample size. We also found that censored rate has negligible impact on outlier detection effect. To some extent, the higher proportion of outlier means we get less normal data and this will influence the ability of model to differentiate normal data and outlier. Along with sample size increase, we get more information of normal data, and then the model can better identify the normal data and outlier. The censored had been taken into account during the parameter estimate, so it has little impact on outlier detection.

Table 2 shows the results of parameter estimation of simulation

N	Proportion of outlier	Censored rate=20%			Censored rate=40%		
		S	M	R	S	M	R
500	10%	2.23	3.07	97.60	2.33	3.14	97.59
	20%	3.38	3.49	96.60	3.48	3.53	96.50
1000	10%	1.49	2.53	98.40	1.50	2.04	98.34
	20%	2.25	2.74	97.66	2.37	3.00	97.50

Table 1: The results of outlier detection of different situation (%).

Proportion of outlier	Censored rate		β_1			β_2			ω		
			Mean	SD	MSE	Mean	SD	MSE	Mean	SD	MSE
N=500											
10%	20%	$\hat{\theta}_1$	0.496	0.148	0.026	0.996	0.141	0.024	1.505	0.130	0.024
		$\hat{\theta}_2$	0.502	0.139	0.019	1.001	0.132	0.020	1.497	0.120	0.023
		$\hat{\theta}_3$	0.464	0.183	0.048	0.836	0.191	0.055	1.317	0.143	0.058
	40%	$\hat{\theta}_1$	0.504	0.162	0.028	1.004	0.145	0.026	1.504	0.179	0.033
		$\hat{\theta}_2$	0.498	0.141	0.022	1.002	0.132	0.021	1.498	0.163	0.027
		$\hat{\theta}_3$	0.456	0.198	0.053	0.866	0.221	0.061	1.313	0.159	0.060
20%	20%	$\hat{\theta}_1$	0.497	0.169	0.035	0.997	0.165	0.033	1.496	0.158	0.026
		$\hat{\theta}_2$	0.501	0.153	0.023	1.001	0.154	0.022	1.501	0.148	0.022
		$\hat{\theta}_3$	0.522	0.202	0.049	0.866	0.215	0.051	1.306	0.161	0.065
	40%	$\hat{\theta}_1$	0.496	0.173	0.034	1.004	0.174	0.034	1.498	0.173	0.030
		$\hat{\theta}_2$	0.503	0.154	0.024	1.000	0.155	0.023	1.500	0.165	0.027
		$\hat{\theta}_3$	0.535	0.207	0.052	0.864	0.211	0.057	1.313	0.166	0.065
N=1000											
10%	20%	$\hat{\theta}_1$	0.495	0.146	0.027	0.997	0.151	0.025	1.503	0.136	0.019
		$\hat{\theta}_2$	0.502	0.134	0.022	1.000	0.141	0.019	1.501	0.114	0.017
		$\hat{\theta}_3$	0.469	0.189	0.041	0.833	0.187	0.053	1.318	0.122	0.051
	40%	$\hat{\theta}_1$	0.503	0.152	0.029	1.002	0.153	0.026	1.496	0.156	0.025
		$\hat{\theta}_2$	0.498	0.136	0.024	0.999	0.140	0.021	1.499	0.131	0.017
		$\hat{\theta}_3$	0.454	0.197	0.048	0.857	0.220	0.059	1.311	0.160	0.060
20%	20%	$\hat{\theta}_1$	0.503	0.170	0.024	0.997	0.165	0.029	1.505	0.155	0.024
		$\hat{\theta}_2$	0.500	0.156	0.021	1.003	0.149	0.021	1.498	0.133	0.018
		$\hat{\theta}_3$	0.526	0.198	0.051	0.851	0.216	0.056	1.296	0.155	0.068
	40%	$\hat{\theta}_1$	0.498	0.174	0.034	1.005	0.168	0.031	1.502	0.169	0.029
		$\hat{\theta}_2$	0.500	0.166	0.025	1.001	0.150	0.025	1.499	0.151	0.023
		$\hat{\theta}_3$	0.541	0.208	0.055	0.859	0.216	0.059	1.306	0.157	0.064

Table 2: The results of parameter estimation of simulation study.

study. The results of $\hat{\theta}_3$ are the worst, the point estimations are far away from the true value and the SD and MSE are larger. The results of $\hat{\theta}_1$ and $\hat{\theta}_2$ are well matched. Therefore, we think WROM could provide a robust parameter estimation.

Moreover, we used one set simulated data to draw the survival curves. True Curve was composed of all normal data, WRM was composed of all simulated data, and WROM was composed of "clean" data which was obtained by deleting the identified outlier.

We can deduce from the Figures 1 and 2 that WRM were different from True Curve, and these differences would increases with the proportion of outlier. That is, outlier would indeed have negative effect on our statistical inference. On the other hand, WROM almost overlapped with True Curve. This also illustrate that outlier detection effect of WROM is satisfied.

Real example

The real data stems from German Breast Cancer Study (GBCS), the download URL is http://www.umass.edu/statdata/statdata/data/gbcs.txt. This study had enrolled 686 breast cancer patients whose age is less than 65 years. 246 of the patients only received three courses of chemotherapy and other 440 of the patients received hormone treatment at least 2 years after three courses of chemotherapy. The endpoint of this study is tumor recurrence, tumor metastasis or death. The effectiveness of the therapy is evaluated by progression-free survival (RFS). Sought to explore the outlier in the data and compare the prognosis of two groups, we fitted the data with WROM. Besides therapy group, the covariance also include age at diagnose, size of tumor, grade of tumor, number of lymphatic metastasis, level of progesterone receptor and level of estrogen receptor. Referring to literature [18], the assignment of each covariance are as follow:

Therapy group (X_1): hormone treatment=1, chemotherapy=2.

Age at diagnose (X_2): less than 45=1, between 45 and 60=2, over 60=3.

Tumor size (X_3): smaller than 20 mm=1, between 21 mm and 30 mm=2, over 30 mm=3.

Grade of tumor (X_4): poor differentiation=1, moderately differentiation=2, well differentiation=3.

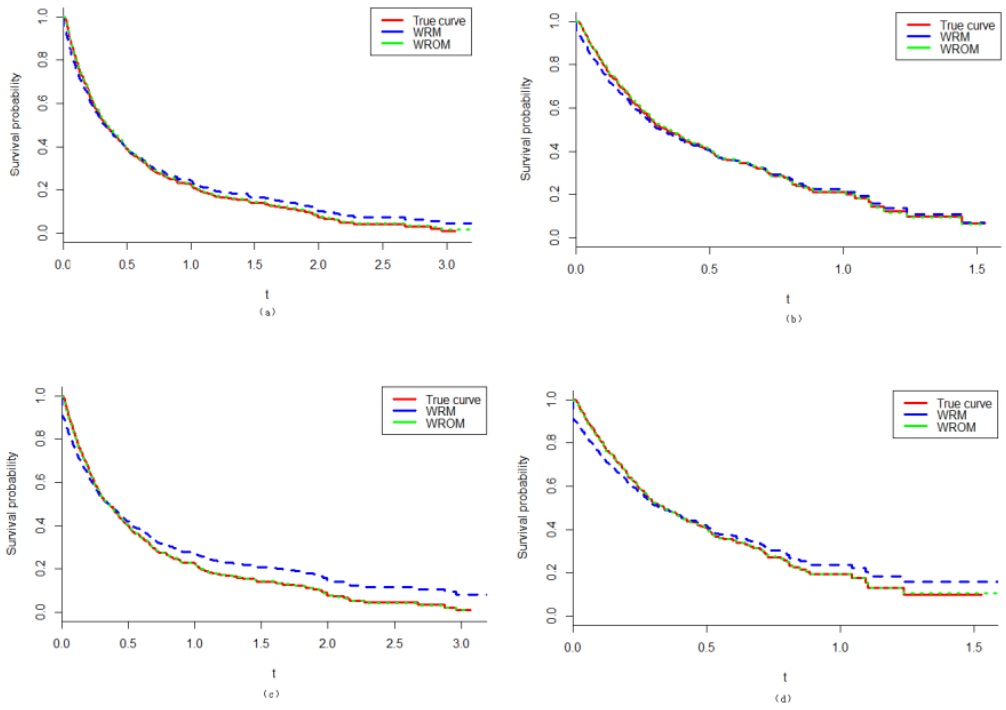
Number of lymphatic metastasis (X_5): less than 3=1, between 4 and 9=2, more than 10=3.

Level of progesterone receptor (X_6): negative (less than 20 fmol mg^{-1})=1, positive (over 20 fmol mg^{-1})=2.

Level of estrogen receptor (X_7): negative (less than 20 fmol mg^{-1})=1, positive (over 20 fmol mg^{-1})=2.

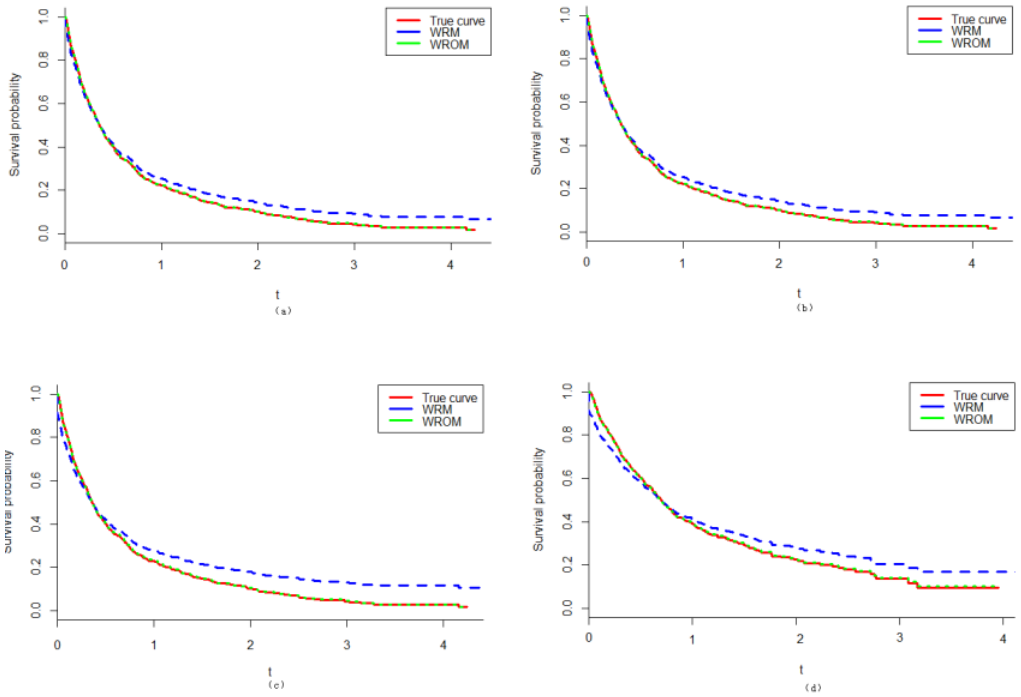
In real example study, we also modeled four different models. Firstly, we fitted the whole data with WROM. After deleting the identified outlier, we fitted the "clean" data with WROM again. In addition, we also fitted the whole data and "clean" data with WRM respectively. During the modeling, we simultaneously constructed three Markov chains. The iterations time is 30000 and first 3000 iterations is burn-out.

By WROM, there are 19.1% (131/686) of patients identified as outlier. In whole data, the median PFS of chemotherapy group



a:ratio of outlier:10%,censored rate:20%; b:ratio of outlier:10%,censored rate:40%; c:ratio of outlier:20%,censored rate:20%; d:ratio of outlier:20%,censored rate:40%

Figure 1: Survival Curves of different situation in simulation study (n=500).



a:ratio of outlier:10%,censored rate:20%; b:ratio of outlier:10%,censored rate:40%; c:ratio of outlier:20%,censored rate:20%; d:ratio of outlier:20%,censored rate:40%

Figure 2: Survival Curves of different situation in simulation study (n=1000).

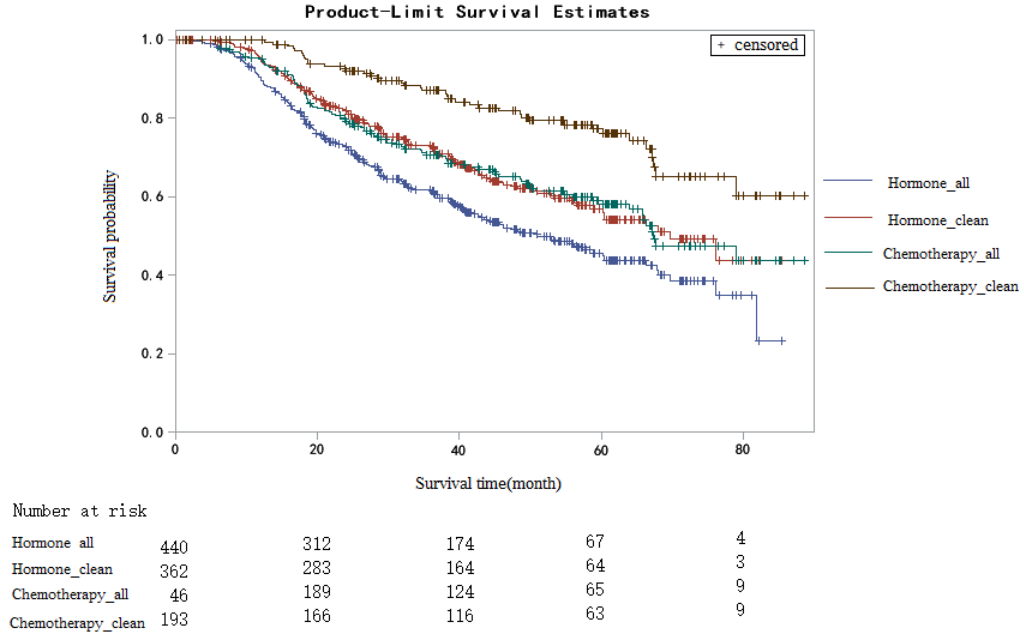


Figure 3: The survival curves of whole data and "clean" datarespectively.

and hormone group is 56.36 months (SE=1.84) and 50.42 months (SE=1.54) respectively. In "clean" data, the median PFS of two group is 67.02 months (SE=1.68) and 55.19 months (SE=1.46) respectively. Figure 3 show the survival curves of whole data and "clean" data. We compared the survival curves of whole data and "clean" data by Log-Rank test, the results indicated that after deleting the outlier, the survival curves significant changed (hormone group: χ^2=10.8090, P=0.001 and chemotherapy group: χ^2=15.5676, P>0.0001).

From the history trace plots and GR statistic line (Figures 1-3) of every parameter, we can tell the MCMC method were convergent. The

Parameters	Whole data				"Clean" data			
	Mean	SD	MC error	95% CI	Mean	SD	MC error	95% CI
WROM								
Therapy group	-0.92	0.13	0.0024	-1.173, -0.671	-1.05	0.14	0.0028	-1.329, -0.782
Age at diagnose	-0.28	0.08	0.0013	-0.431, -0.122	-0.32	0.09	0.0016	-0.491, -0.147
Tumor size	-0.12	0.08	0.0012	-0.260, 0.032	-0.09	0.08	0.0014	-0.249, 0.063
Grade of tumor	-0.77	0.08	0.0018	-0.935, -0.607	-0.90	0.09	0.0021	-1.084, -0.723
Number of lymphatic metastasis	0.59	0.07	0.0011	0.447, 0.735	0.75	0.08	0.0012	0.598, 0.902
Level of progesterone receptor	-1.48	0.13	0.0031	-1.741, -1.232	-1.75	0.15	0.0039	-2.039, -1.471
Level of estrogen receptor	-0.42	0.13	0.0029	-0.668, -0.169	-0.44	0.14	0.0033	-0.719, -0.173
Shape parameter	1.32	0.06	0.0015	1.213, 1.430	1.49	0.07	0.0021	1.353, 1.621
WRM								
Therapy group	-0.83	0.13	0.0025	-1.268, -0.762	-0.96	0.14	0.0031	-1.229, -0.697
Age at diagnose	-0.28	0.08	0.0016	-0.475, -0.160	-0.27	0.08	0.0015	-0.433, -0.110
Tumor size	-0.05	0.07	0.0014	-0.235, 0.060	-0.13	0.08	0.0014	-0.282, 0.028
Grade of tumor	-0.64	0.09	0.0017	-0.951, -0.618	-0.85	0.09	0.0021	-1.034, -0.669
Number of lymphatic metastasis	0.44	0.07	0.0012	0.491, 0.788	0.65	0.08	0.0012	0.497, 0.796
Level of progesterone receptor	-1.25	0.13	0.0032	-1.805, -1.273	-1.62	0.14	0.0033	-1.893, -1.345
Level of estrogen receptor	-0.33	0.13	0.0032	-0.647, -0.118	-0.42	0.14	0.0033	-0.689, -0.152
Shape parameter	1.10	0.05	0.0013	1.211, 1.407	1.39	0.07	0.0019	1.266, 1.522

Table 3: The parameter estimation of WROM and WRM.

parameter estimate results were presented in Table 3. When modelling WROM, the results of whole data matched those of "clean" data. On the contrary, the results of WRM changed significantly. This tell us that WRM is sensitive to outlier and WROM's results is robust even when the data contains outliers. The results also reveal that additional hormone therapy cannot prolong PFS of breast cancer patients.

We used the DIC criterion to evaluate fitting effects of four models. Among four models, WRM with whole data has the highest DIC (3245) and WROM with "clean" data has the lowest DIC (2742). When fitting whole data, WROM's fitting effect is better than WRM (3026 VS 3245). And when fitting "clean" data, the fitting effect of WROM and WRM are close (2742 VS 2802). This means, when outlier exists, fitting data with WRM which does not take into account the outlier is not appropriate.

Conclusion

This paper proposed a method to construct Weibull regression outlier model to detect outlier in survival data which survival time follow a Weibull distribution, as well as to get parameter robust estimation. Both simulation study and real example study indicate that WROM could well detect potential outlier and provide robust reliable parameter estimation.

In this paper, we assumed the survival time is from Weibull distribution. However, in our practical work, there is a wide range of survival time distribution and it is not easy to determine which distribution the survival time followed. Can we imitate the thinking of WROM to construct other Bayesian parameter survival model or Bayesian Semi-parameter survival model? This deserve further study. Although WROM can identified outliers, it still fails to provide exploration of the cause of outliers. This needs domain-specific analyses.

Declaration

Availability of data and materials

The dataset supporting the conclusions of this article is available in http://www.umass.edu/statdata/statdata/data/gbcs.txt.

Author's Contribution

P.Y. and X.C. designed study; C.S. and T.Q. performed simulation study and analyzed data; C.S. wrote the paper.

Acknowledgement

This work was supported by the special funds of Health research 2013, MOH, China (No. 201302009), National Natural Science Foundation of China (No. 81573262) and the Fundamental Research Funds for the Central Universities, HUST (No. 2016YXZD042).

References

1. Hampel F, Ronchetti E, Rousseeuw P (1986) Robust statistics: the approach based on influence functions. New York: Wiley, pp: 1-502.
2. Weisberg S (2005) Applied linear regression (3rdedn), John Wiley & Sons, pp: 1-368.
3. Ferguson, Thomas S (1961) Rules for rejection of outliers. Revue Inst Int De Stat 3: 29-43.
4. Beckman RJ, Cook RD (1983) Outlier.....s. Technometrics 2: 119-149.
5. Hawkins DM (1980) Identification of outliers. London: Chapman and Hall.
6. Taylan P, Yerlikaya-Özkurt F, Weber G (2014) An approach to the mean shift outlier model by Tikhonov regularization and conic programming. Intell Data Anal 18: 79-94.
7. Ferreira CS, Mattos TB, Balakrishnan N (2016) Mean-shift outliers model in skew scale-mixtures of normal distributions. J Stat Comput Simul 86: 2346-2361.
8. Ha J, Seok S, Lee J (2015) A precise ranking method for outlier detection. J Inf Sci 324: 88-107.
9. Evans K, Love T, Thurston SW (2015) Outlier identification in model-based cluster analysis. J Classif 32: 63-84.
10. Nardi A, Schemper M (1999) New residuals for Cox regression and their application to outlier screening. Biometrics 55: 523-529.
11. Eo SH, Hong SM, Cho HJ (2014) Identification of outlying observations with quantile regression for censored data. Cornell University Library.
12. She Y, Owen AB (2011) Outlier detection using nonconvex penalized regression. J Am Stat Assoc 106: 626-639.
13. Kostoulas P, Nielsen SS, Browne WJ (2010) A Bayesian Weibull survival model for time to infection data measured with delay. Pre Vet Med 94: 191-201.

14. Li H, Yuan R, Peng W (2011) Bayesian inference of Weibull distribution based on probability encoding method. International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineerin, IEEE 1: 365-369.

15. Kundu D, Mitra D (2016) Bayesian inference of Weibull distribution based on left truncated and right censored data. Comput Stat Data Anal 99: 38-50.

16. Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103: 681-686.

17. Li Q, Lin N (2010) The Bayesian elastic net. Bayesian Anal 5: 151-170.

18. Sauerbrei W, Royston P, Bojar H (1999) Modelling the effects of standard prognostic factors in node-positive breast cancer. German Breast Cancer Study Group (GBSG). Br J Cancer 79: 1752-1760.