Open Journal of Statistics, 2014, 4, 313-317
Published Online June 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.44031

Statistical Diagnosis for Random Right Censored Data Based on Kaplan-Meier Product Limit Estimate

Shuling Wang[1], Xiaohong Deng[1], Lin Zheng[2]

[1]Department of Fundamental Course, Air Force Logistics College, Xuzhou, China
[2]School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China
Email: 155328313@qq.com

Received 20 April 2014; revised 15 May 2014; accepted 2 June 2014

Abstract

In this work, we consider statistical diagnostic for random right censored data based on K-M product limit estimator. Under the definition of K-M product limit estimator, we obtain that the relation formula between estimators. Similar to complete data, we define likelihood displacement and likelihood ratio statistic. Through a real data application, we show that our proposed procedure is validity.

Keywords

Random Right Censorship, Kaplan-Meier Product-Limit Estimator, Empirical Likelihood, Outliers, Influence Analysis

1. Introduction

Statistical diagnosis developed in the mid-1970s, which is a new statistical branch.

In the course of development of the past 40 years, most scholars have studied the data into a convenient and effective statistical model. For example, the diagnosis and influence analysis of linear regression model has been fully developed (R. D. Cook and S. Weisberg [1], Bocheng Wei, Guobin Lu & Jianqing Shi [2]); The varing coefficient model is a useful extension of classical linear model. Regarding the varying coefficient model, especially for the B-spline estimation of parameter, diagnosis and influence analysis have some results (Cai, Z., Fan, J., Li, R. [3], Fan, J., Zhang, W. [4]). However, all the above results are obtained under the uncensored case. In many applications, some of the responses and/or covariants may not be observed, but are censored. For censored data, the usual statistical techniques for complete data situations are not readily applicable. Because there are too many hypothesis, it is easy to lose information.

How to cite this paper: Wang, S.L., Deng, X.H. and Zheng, L. (2014) Statistical Diagnosis for Random Right Censored Data Based on Kaplan-Meier Product Limit Estimate. *Open Journal of Statistics*, **4**, 313-317.
http://dx.doi.org/10.4236/ojs.2014.44031

As we all known, the distribution function of a random variable X contains all of the probabilistic information about X. Hence this paper tries to use non-parametric maximum likelihood estimate (NPMLE) [5] of distribution function in follow-up study.

The rest of the paper is organized as follows. The right censoring and K-M product limit estimator is introduced in Section 2. Outlier diagnosis and influence analysis are presented in Section 3. An example is given to illustrate our results in Section 4.

2. Right Censoring and Kaplan-Meier Product Limit Estimator

Here, the distribution of a real-valued random variables X_i is of direct interest. For each X_i there is a $Y_i \in R$. This Y_i may be random. If $X_i \le Y_i$ we observe X_i, otherwise X_i is censored to (Y_i, ∞). We say that X_i is right censored by Y_i. Let $Z_i = \min\{X_i, Y_i\}$ and $\delta_i = 1_{X_i \le Y_i}$. For example, X_i could be survival time after an operation, with Y_i the time from the operation to the end of the study.

The idea of the K-M product limit estimator is given by the conditional probability. Let $t_i \le t_{i+1}$:

$$\begin{aligned} F(t_i) &= 1 - S(t_i) = 1 - P(T > t_i) = 1 - P(T > t_i, T > t_{i-1}) = 1 - P(T > t_i \mid T > t_{i-1}) P(T > t_{i-1}) \\ &= 1 - P(T > t_i \mid T > t_{i-1}) P(T > t_{i-1} \mid T > t_{i-2}) \cdots P(T > t_0 = 0) \end{aligned}$$

We assume that at the start of the study all subjects were alive, so $P(T > T_0 = 0) = 1$. The conditional probability is

$$P(T > t_i, T > t_{i-1}) = \frac{r_i - d_i}{r_i} \quad P(T > t_i, T > t_{i-1}) = \frac{r_i - d_i}{r_i},$$

where r_i is the number of subjects at risk in the study at the time t_i, and d_i is the number of subject dying at time t_i. The Kaplan-Meier estimator of CDF is

$$\hat{F} = 1 - \prod_{i \mid t_i \le t} \frac{r_i - d_i}{r_i}.$$

3. Statistical Diagnostic

For complete data, diagnostic measures of outlier contain case deletion and mean shift, influence statistics contain Cook's distance, W-K statistic, covariance ratio statistic, AP statistic, likelihood distance and so on. Similarly, we derive several diagnostic measures for right censored data.

3.1. Outlier Diagnosis

Let $\hat{F}(i)$ is the K-M product limit estimator of distribution function after case deletion, then there are lemma

$$\hat{F}(i)\hat{F} - \hat{F}(i) = \begin{cases} \dfrac{d_i}{r_i} \displaystyle\prod_{\substack{j \mid t_j \le t \\ (j \ne i)}} \dfrac{r_j - d_j}{r_j}, & t_i \le t \\ 0, & t_i > t \end{cases}.$$

Proof: By the definition of K-M product limit estimator, there are

$$\hat{F} = 1 - \prod_{j \mid t_j \le t} \frac{r_j - d_j}{r_j}$$

and

$$\hat{F}_{(i)} = 1 - \prod_{\substack{j \mid t_j \le t \\ (j \ne i)}} \frac{r_j - d_j}{r_j}$$

Since, when $t_i \le t$, there are

$$\hat{F}-\hat{F}(i)=\left(1-\prod_{j|t_j\le t}\frac{r_j-d_j}{r_j}\right)-\left(1-\prod_{\substack{j|t_j\le t\\(j\neq i)}}\frac{r_j-d_j}{r_j}\right)$$

$$=\prod_{\substack{j|t_j\le t\\(j\neq i)}}\frac{r_j-d_j}{r_j}-\prod_{j|t_j\le t}\frac{r_j-d_j}{r_j}$$

$$=\left(1-\frac{r_i-d_i}{r_i}\right)\prod_{\substack{j|t_j\le t\\(j\neq i)}}\frac{r_j-d_j}{r_j}=\frac{d_i}{r_i}\prod_{\substack{j|t_j\le t\\(j\neq i)}}\frac{r_j-d_j}{r_j}\quad .$$

when $t_i>t$, $\hat{F}-\hat{F}_{(i)}=0$ is obviously.

From the lemma, we can construct the relation formula between estimators, which is the foundation of discussion.

3.2. Influence Analysis

3.2.1. Likelihood Displacement

The likelihood function is defined as

$$L(F)=\prod_{i=1}^{n}F(\{Z_i\})^{\delta_i}\,F((Z_i,\infty))^{1-\delta_i}$$

where $\delta_i=1_{X_i\le Y_i}$, $Z_i=\min(X_i,Y_i)$, which can be computed from the Z_i and δ_i without knowing the Y_i from uncensored X_i.

Likelihood displacement is the method for measuring influence, which is advanced by Cook and Weisberg in 1982, which is advanced from the view of data fitting. Considering the influence of deleting the i-th case. Then, the likelihood displacement can be expressed as follows

$$LD_i(F)=2\left[L(F)-L(F_{(i)})\right].$$

3.2.2. Likelihood Ratio Statistic

For complete data, there is likelihood ratio statistic. Similarly, we define the likelihood ratio statistic for censored data based on K-M product limit estimator as follows

$$R_i=\frac{L(F_{(i)})}{L(F)}=\frac{\prod_{j=1,j\neq i}^{n}F_{(i)}(\{Z_j\})^{\delta_j}\,F_{(i)}((Z_j,\infty))^{1-\delta_j}}{\prod_{i=1}^{n}F(\{Z_i\})^{\delta_i}\,F((Z_i,\infty))^{1-\delta_i}}$$

4. Numerical Studies

(**Vicious Tumour Data**) In this section, we consider an example as the illustration for the above results. Considering a clinical research trial data (see Andersen [6]). There are 205 cancer patients who have been treated in Odense university hospital and tracked until the end of 1977. The survival time of some individuals due to death or end of the trial for other reasons were censored. Wang Qihua [7] ultized a linear semi-parametric model to fit these test data. Wang Shuling *et al.* [8] ultized a nonparametric regression model with random right censorship to fit the data of 126 female patients. Now we consider the first twenty data, calculate the likelihood function by MATLAB and obtain $L(F)=0.0000003$. The originality data and the other results are in following **Table 1**.

Where $L(F_{(i)})$ is likelihood function after deleting the i-th case, the results of LD_i and R_i are as follows.

Table 1. The originality data and the value of $L\left(F_{(i)}\right)$.

Z_i	δ_i	$L\left(F_{(i)}\right)$
0.0099	0	0.0000003
0.0232	0	0.0000003
0.0279	1	0.00000102
0.0295	1	0.00000102
0.0355	0	0.00000043
0.0386	1	0.00000086
0.0469	1	0.00000086
0.0667	1	0.00000086
0.0817	1	0.00000086
0.0826	0	0.00000077
0.0833	1	0.00000074
0.0858	1	0.00000074
0.0869	1	0.00000074
0.0872	1	0.00000074
0.0982	1	0.00000074
0.1055	1	0.00000074
0.1156	1	0.00000074
0.1252	1	0.00000074
0.1271	1	0.00000074
0.1312	1	0.00000074

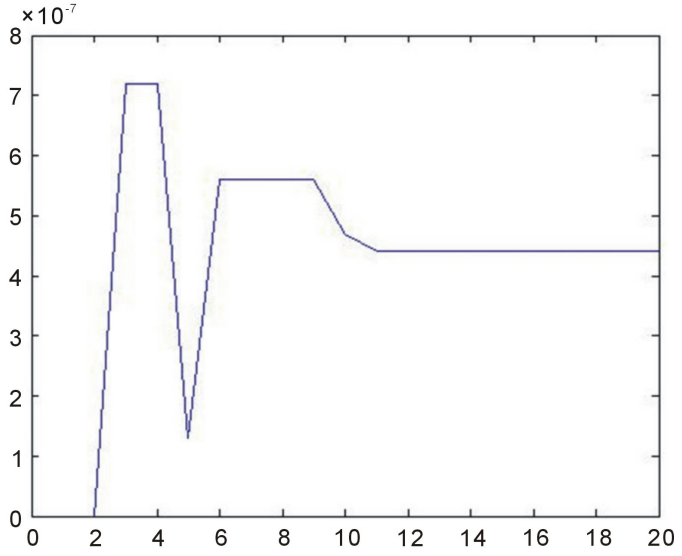

Figure 1. The value of LDi.

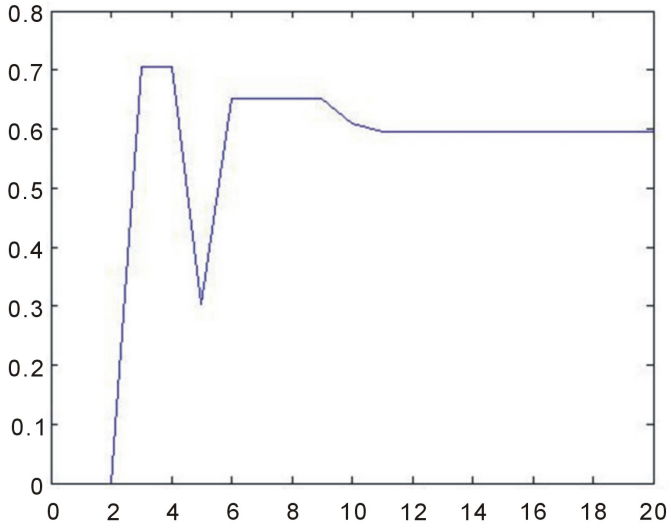

Figure 2. The value of Ri.

Figure 1 and **Figure 2** show that the first, second, third, fourth and fifth data are outliers. Indeed, this result is similar to Wang Shuling *et al.* [8].

References

[1] Cook, R.D. and Weisberg, S. (1982) Residuals and Influence in Regression. Chapman and Hall, New York.

[2] Wei, B.C., Lu, G.B. and Shi, J.Q. (1990) Statistical Diagnostics. Publishing House of Southeast University, Nanjing.

[3] Cai, Z., Fan, J. and Li, R. (2000) Efficient Estimation and Inferences for Varying-Coefficient Models. *Journal of American Statistical Association*, **95**, 888-902. http://dx.doi.org/10.1080/01621459.2000.10474280

[4] Fan, J. and Zhang, W. (2008) Statistical Methods with Varying Coefficient Models. *Statistics and Its Interface*, **1**, 179-195. http://dx.doi.org/10.4310/SII.2008.v1.n1.a15

[5] Owen, A. (2001) Empirical Likelihood. Chapman and Hall, New York. http://dx.doi.org/10.1201/9781420036152

[6] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) Statistical Models Based on Counting Processes. Springer-Verlag, New York.

[7] Wang, Q.H. (2006) Analysis of Survival Data. Science Press, Beijing.

[8] Wang, S.L., Feng, Y. and Liu, X.B. (2010) Statistical Diagnostics of Nonparametric Regression Model with Random Right Censorship. *Journal of Hefei University of Technology* (*Natural Science*), **33**, 470-473.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.