Pergamon

Nonlinear Analysis, Theory, Methods & Applications, Vol. 30, No. 5, pp. 3201–3206, 1997
Proc. 2nd World Congress of Nonlinear Analysts

0362-546X/97 $17.00 + 0.00

PII: S0362-546X(97)00414-8

ON TESTING INDEPENDENCE WITH RIGHT TRUNCATED DATA

ÜLKÜ GÜRLER
Department of Industrial Engineering, Bilkent University,
06533 Bilkent, Ankara, TURKEY.
E-mail:ulku@bilkent.edu.tr.

Key words and phrases: Right truncation,Test of independence, Reverse hazard, Nonparametric estimation.

1. INTRODUCTION

Inference with bivariate data gained considerable interest recently. See eg.[1],[10],[12]. All of these studies however consider estimation of the bivariate distribution function under various bivariate censoring models. Recently Gürler[7,8] considered estimation of the bivariate distribution and the hazard functions under truncation/censoring models. The purpose of this study is to investigate procedures for testing the independence of the components of the bivariate vector for truncated data. To this end, further properties of the bivariate functionals introduced in Gürler[8] are elaborated. Two alternative methods for hypothesis testing are suggested and some large sample properties are derived. The procedures suggested in this paper are applicable to left/right truncated and left truncated right censored data. However to keep the presentation simple we confine the discussion to the right truncated case. Also, to avoid technicalities, it is assumed that all the univariate and the bivariate distribution functions are absolutely continuous admitting densities.

2. PRELIMINARIES

In bivariate truncation model, the triplets (Y_i, X_i, T_i), $i = 1, \ldots, n$ are observed for which $(Y_i \leq T_i)$. The interest is in the pair of random variables (Y, X) with distribution function $F(y, x)$. Here, T is a random variable, which is assumed to be independent of (Y, X), with d.f. G. The marginal d.f.'s of Y and X are denoted by F_Y and F_X respectively. For the identifiability of F, it is assumed that $a_{F_Y} \leq a_G$ and $b_{F_Y} \leq b_G$ (see [16]). Let $\bar{F}(y, x) = P(Y > y, X > x)$ be the bivariate survival function and for a univariate function g, such that $0 \leq g(x) \leq 1$, let $\bar{g}(x) = 1 - g(x)$. The distribution of the observed variables (Y, X, T) are given by:

$$\begin{aligned} H_{Y,X,T}(y, x, t) &= P(Y \leq y, X \leq x, T \leq t \mid Y \leq T) \\ &= \alpha^{-1} \int_0^t F(y \wedge u, x)\, dG(u) \end{aligned}$$

where $\alpha = P(Y \leq T)$, and $y \wedge u = \min(y, u)$. This joint distribution function can be further elaborated to result in:

$$h_{Y,X}(y, x) = \alpha^{-1} \bar{G}(y) f(y, x) \tag{1}$$

$$C(y) = \alpha^{-1} \bar{G}(y) F_Y(y) = H_Y(y) - H_T(y) \tag{2}$$

where $h_{Y,X}$ and f denote the bivariate densities of H anf F. Considering the foregoing relations, Gürler[8] suggests the following estimator

$$\bar{F}_n(y, x) = \frac{1}{n} \sum_i \frac{F_{Y,n}(Y_i)}{C_n(Y_i)} I(Y_i > y, X_i > x)$$

where

$$\begin{aligned} F_{Y,n}(y) &= \prod_{i:Y_i > y} [1 - s(Y_i)/nC_n(Y_i)] \\ s(u) &= \#\{i : Y_i = u\} \\ nC_n(u) &= \#\{i : Y_i \leq u \leq T_i\} \end{aligned}$$

The estimator $F_{Y,n}(y)$ is suggested by Lynden-Bell[10] and the large sample properties of it for left truncation model are studied in in [17],[2],[6],[14]. Gürler[8] establishes the following results:

$$F_{Y,n}(y) - F_Y(y) = \xi_{1,n}(y) + \epsilon_{1,n}(y) \quad (3)$$

$$\bar{F}_{Y,X,n}(y) - F_{Y,X}(y,x) = \xi_{2,n}(y,x) + \epsilon_{2,n}(y,x) \quad (4)$$

where

$$sup_{y>a>a_F}|\epsilon_{1,n}(y)| = sup_{y>a>a_F}|\epsilon_{2,n}(y,x)| = O(\log^3 n/n)$$

and $\epsilon_{1,n}(y), \epsilon_{2,n}(y,x)$ as defined in Gürler[8] are mean zero i.i.d. variables.

Bivariate reverse-hazard in right truncation model:
Following the representation of Dabrowska[4], of the bivariate d.f. in terms of the three component bivariate hazard vector, Gürler[8] suggested a bivariate reverse-hazard vector $\lambda(y,x)$ for right truncated data as described below. For any bivariate function $\phi(u,v)$, differentiable in both components, let $\phi(\partial u, v)$ denote the partial derivative w.r.t. to first argument with similar notation applying to the other component. Then

$$\begin{aligned}\lambda(u,v) &= \{F(\partial u,\partial v)/F(u,v)\ ,\ F(\partial u,v)/F(u,v)\ ,\ F(u,\partial v)/F(u,v)\} \\ &\equiv \{\lambda_{12}(u,v),\lambda_1(u,v),\lambda_2(u,v)\} \end{aligned} \quad (5)$$

Let $R(y,x) = -\log F(y,x)$, then

$$F(y,x) = F_X(x)F_Y(y)exp\{-\Lambda(y,x)\} \quad (6)$$

where

$$\begin{aligned}\Lambda(y,x) &= \int_y^{b_{F_Y}}\int_x^{b_{F_X}} R(du,dv) \\ &= \int_y^{b_{F_Y}}\int_x^{b_{F_X}} [\lambda_{12}(u,v) - \lambda_1(u,v)\lambda_2(u,v)]dudv\end{aligned}$$

$$\lambda_{12}(u,v) = \frac{H_{Y,X}(\partial u,\partial v)}{C_2(u,v)}$$

$$\lambda_1(u,v) = \frac{H_{Y,X}(\partial u,v)}{C_2(u,v)}$$

$$\lambda_2(u,v) = \frac{C_2(u,\partial v)}{C_2(u,v)}.$$

where

$$\begin{aligned}C_2(y,x) &= F^*_{Y,X}(y,x) - H_{T,X}(y-,x) \\ &= \alpha^{-1}[1-G(y)]F(y,x)\end{aligned} \quad (7)$$

with empirical counterpart given by:

$$\begin{aligned}C_{2,n}(u,v) &= n^{-1}\#\{i : Y_i \le u \le T_i\ ,\ X_i \le v\} \\ &= n[H_{n,Y,X}(u,v) - H_{n,T,X}(u-,v)].\end{aligned}$$

where $H_{n,Y,X}$ and $H_{n,T,X}$ are the empirical bivariate d.f.'s. For these step functions, let Δ refer to the difference operator. Then an estimator for the reverse- hazard is obtained as follows:

$$\begin{aligned}\lambda_n(u,v) &= \{\frac{H_{n,Y,X}(\Delta u,\Delta v)}{C_{2,n}(u,v)}, \frac{H_{n,Y,X}(\Delta u,v)}{C_{2,n}(u,v)}, \frac{C_{2,n}(u,\Delta v)}{C_{2,n}(u,v)}\} \\ &\equiv \{\lambda_{12,n}(u,v),\lambda_{1,n}(u,v),\lambda_{2,n}(u,v)\}\end{aligned} \quad (8)$$

Observe that

$$\begin{aligned} C_{2,n}(Y_i, X_j) &= n^{-1}\sum_{k=1}^{n} I(Y_k \le Y_i \le T_k\ ,\ X_k \le X_j) \\ H_{n,Y,X}(Y_i, X_j) &= n^{-1}\sum_{k=1}^{n} I(Y_k \le Y_i,\ X_k \le X_j) \\ C_{2,n}(Y_i, \Delta X_j) &= n^{-1}\sum_{k=1}^{n} I(Y_k \le Y_i \le T_k\ ,\ X_k = X_j) \\ H_{n,Y,X}(\Delta Y_i, X_j) &= n^{-1}\sum_{k=1}^{n} I(Y_k = Y_i,\ X_k \le X_j) \end{aligned}$$

and assuming no ties in the data,

$$\begin{aligned} C_{2,n}(Y_i, \Delta X_j) &= n^{-1} I(Y_j \le Y_i \le T_j) \\ H_{n,Y,X}(\Delta Y_i, X_j) &= n^{-1} I(X_i \le X_j) \end{aligned}$$

Hence

$$H_{n,Y,X}(\Delta Y_i, X_j) C_{2,n}(Y_i, \Delta X_j) = n^{-2} I(Y_j \le Y_i \le T_j, X_i \le X_j)$$

Then the following estimator for $\Lambda(y, x)$ is obtained

$$\begin{aligned} -\Lambda_n(y,x) &= \sum_{i=1}^{n} \frac{I(Y_i > y, X_i > x)}{nC_n(Y_i, X_i)} \\ &- \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{I(Y_i > y, X_i > x) I(Y_j \le Y_i \le T_j, X_i \le X_j)}{n^2 C_n^2(Y_i, X_j)} \\ &= \sum_{i=1}^{n} \frac{I(Y_i > y, X_i > x)}{nC_n(Y_i, X_i)} - \sum_{i=1}^{n} \frac{I(Y_i > y, X_i > x)}{n^2 C_n^2(Y_i, X_i)} \\ &- \sum_{i=1}^{n}\sum_{j\ne}^{n} \frac{I(Y_i > y, X_i > x) I(Y_j \le Y_i \le T_j, X_i \le X_j)}{n^2 C_n^2(Y_i, X_j)} \\ &\equiv I - II - III \end{aligned}$$

3. TESTING INDEPENDENCE OF Y AND X

The foregoing discussions in Section 2 lead to the following approaches for testing the independence of Y and X. Let

$$r(y,x) = \frac{C_2(y,x)}{C(y)} = \frac{F(y,x)}{F_Y(y)}$$

and

$$S_1(y,x) = \bar{F}(y,x) - \bar{F}_Y(y)\bar{r}(y,x)$$

which can be estimated by

$$S_{1,n}(y,x) = \bar{F}_n(y,x) - \bar{F}_{Y,n}(y)\bar{r}_n(y,x)$$

where

$$\bar{r}_n(y,x) = 1 - \frac{C_{2,n}(y,x)}{C_n(y)} = \frac{\#\{i : Y_i \le y \le T_i\ ,\ X_i > x\}}{\#\{i : Y_i \le y \le T_i\}}$$

Under the hypothesis of independence, $r(y,x) = F_X(x)$ and $S_1(y,x) = 0$. Therefore test statistics can be based on the functionals of $S_{1,n}(y,x)$, such as $\sum_{i=1}^{k} S_{1,n}(y_i,x_i)$ or $\sup_{1\le i\le k} S_{1,n}(y_i,x_i)$, where $(y_i,x_i), i = 1,..,k$ would be pre-selected points on R^2. Similarly, from the foregoing discussions on the reverse-hazard, we observe that if Y and X are independent, $\Lambda(y,x) = 0$. Hence an alternative approach could be to use $\sum_{i=1}^{k} \Lambda_{1,n}(y_i,x_i)$ or $\sup_{1\le i\le k} \Lambda_{1,n}(y_i,x_i)$ as a test statistics. A related approach for the censored data can be found in [13] and the references therein. Comparison of these tests and their behavior relating to power, unbiasedness, relative efficiency etc. are subject to further investigation. However, we present below the following large sample results, which would lead to establish such properties. Using the results of [16],[17], following Lemmas are obtained. Proofs of these results and the theorems below, which are not presented here for space considerations can found in Gürler[9].

LEMMA 1:

$$\text{a}-) \quad E[\frac{1}{nC_n(Y_i,X_i)}|Y_i = y, X_i = x] = 1/nC(y,x)\{1-[1-C(y,x)]^n\}$$

$$\text{b}-) \quad E[\frac{1}{n^2C_n^2(Y_i,X_i)}|Y_i = y, X_i = x] = I_n[C(y,x)]/nC(y,x)$$

$$\text{c}-) \quad E[\frac{I(Y_j \le Y_i \le T_j, X_i \le X_j}{n^2C_n^2(Y_i,X_j)}|Y_i = y, X_j = x)]$$
$$= \{1 - I_n[C(y,x)] - [1-C(y,x)]^n/[n(n-1)C^2(y,x)]$$

where

$$nI_n[C(y,x)] = \frac{1}{n} + \sum_{i=0}^{n-1} \frac{i}{n-i}[1-C(y,x)]^i - [1-C(y,x)]^n[1 - n\sum_{i=1}^{n}\frac{1}{i}]$$

LEMMA 2:

$$\begin{aligned} E[-\Lambda_n(y,x)] &= \int_{u>y}\int_{v>x}[\lambda_{12}(u,v) - \Lambda_1(u,v)\Lambda_2(u,v)]dudv + B_n(y,x) \\ &= \int_{u>y}\int_{v>x}[\frac{h_{Y,X}(u,v)}{C_2(u,v)} - \frac{1}{C_2^2(u,v)}H_{Y,X}(\partial u,v)H_{Y,X}(u,\partial v)]dudv + B_n(y,x) \end{aligned}$$

where

$$\begin{aligned} B_n(y,x) &= \sum_{u>y}\sum_{v>x}\frac{\{I_n[C(y,x)] - [1-C(y,x)]^n\}}{C_2(u,v)}H_{Y,X,n}(\Delta u,\Delta v) \\ &- \sum_{u>y}\sum_{v>x}\frac{\{I_n[C(y,x)] + [1-C(y,x)]^n\}}{C_2^2(u,v)}H_{Y,X,n}(\Delta u,v)H_{Y,X,n}(u,\Delta v) \end{aligned}$$

Let

$$\begin{aligned} I_i(u) &= I(Y_i \le u \le T_i) \\ I_i(u,v) &= I(Y_i \le u \le T_i, X_i \le v) \end{aligned}$$

$$U_n(u,v) = \sum_{i=1}^{n}\sum_{j\ne i}^{n}[I_i(u,v)I_j(u,v) - C^2(u,v)]$$

LEMMA 3: Let $r_n^*(y,x)$ be the Hajek projection of $r_n(y,x)$. Then,

$$r(y,x) = r_n^*(y,x) + \epsilon_n(y,x)$$

where, for $\delta > 3/2$

$$\sup_{a_F<y;0<x<\infty}|\epsilon_n(y,x)| = O(\log^\delta n/n)$$

and

$$\begin{aligned} r_n^*(y,x) &= \frac{C_2(y,x)}{C(y)} + \frac{C^{-2}(y)}{n}\sum[C(y)I_i(y,x) - C_2(y,x)I_i(y)] \\ &\equiv \frac{C_2(y,x)}{C(y)} + \xi_{3,n}(y,x) \end{aligned}$$

The above Lemmas, together with the results of Gürler(1996) lead to the following theorems.

THEOREM 1: Let $a_F < y < y$ and suppose $\int \bar{G}^{-2}(u)F(du) < \infty$. Then,

$$\begin{aligned} \Lambda_n(y,x) - \Lambda(y,x) &= \int_{u>y}\int_{v>x} \frac{U_n(u,v)}{C_2^4(u,v)} H_{Y,X}(\partial u, v)C(u,\partial v)dudv \\ &+ \int_{u>y}\int_{v>x} \frac{C_{2,n}(u,v) - C_2(u,v)}{C_2^2(u,v)} h_{Y,X}(uu,v)dudv + R_n(y,x) \\ &\equiv \xi_{\Lambda,n}(y,x) + R_{1,n}(y,x) \end{aligned}$$

where

$$\sup_{(y,x)\in T_{ab}} |R_n(y,x)| = O(\log^3 n/n)$$

Note that $E[\xi_{\Lambda,n}(y,x)] = 0$ and $\xi_{\Lambda,n}(y,x)$ is a sum of identical but not independent random variables. In fact it is a sum of two U-Statistics and this enables us to study the large sample properties, which is still subject to further investigation. The representations in (3) and (4) now leads to a corresponding representation for the statistic $S_{1,n}(y,x)$, from which the results for the functionals of it can be derived.

THEOREM 2: Under the conditions of Theorem 1,

$$\begin{aligned} S_{1,n}(y,x) - S_1(y,x) &= \xi_{2,n}(y,x) - \bar{F}(y)\xi_{3,n}(y,x) + R_{2,n}(y,x) \\ &\equiv \gamma_n(y,x) + R_{2,n}(y,x) \end{aligned}$$

where

$$\sup_{(y,x)\in T_{ab}} |R_{2,n}(y,x)| = O(\log^3 n/n)$$

The $\gamma_n(y,x)$ above is again in the form of sum of mean zero i.i.d. random variables, for which standard large sample results apply.

REFERENCES:

1. AKRITAS, M.G., Nearest Neighbor Estimation of a Bivariate Distribution Under Random Censoring, *Annals of Statistics*, 22, 1299-1327, (1994).
2. CHAO, M. T., LO, S-H., Some Representations of the Nonparametric Maximum Likelihood Estimators with Truncated Data, *Annals of Statistics*, 16, 661-668 (1988).
3. CHEN K., CHAO M-T., and LO S-H., On Strong Uniform Consistency of the Lynden-Bell Estimator for Truncated Data, *Annals of Statistics*, 23, 440-449 (1995) .
4. DABROWSKA, D.M. , Kaplan-Meier Estimate on the Plane, *Annals of Statistics*, 16, 1475-1489 (1988).
5. DABROWSKA, D.M., Nonparametric Regression With Censored Covariates (to appear) (1995). *Journal of Multivariate Analysis.*
6. GIJBELS I., and WANG, J.L., Strong Representations of the Survival Function Estimator for Truncated and Censored Data with Applications,*Journal of Multivariate Analysis*, 47, 210-229, (1993).
7. GÜRLER, Ü., Bivariate Distribution and Hazard Functions When a Component is Randomly Truncated, To appear *Journ. Mult. Analy.*,(1996).

8. GÜRLER, Ü., Bivariate Estimation with Right Truncated Data , *J.A.S.A.*, 91, 1152-1165, (1996).

9. GÜRLER, Ü., On Testing Independence with Right Truncated Data, *IEOR Research Report- 9620*, Depart. of Ind. Eng., Bilkent Univ., Ankara-TURKEY,(1996).

10. LIN, D.Y. and YING, Z., A Simple Nonparametric Estimator of the Bivariate Survival Function Under Univariate Censoring,*Biometrika*, 80, 573-581(1993).

11. LYNDELL-BELL, D. A Method of Allowing for Known Observational Selection in Small Samples Applied to 3CR Quasars, *Monthly Notices of the Royal Astronomy Society*, 155, 95-118, (1971).

12. PRENTICE, R.L., CAI, J., Covariance and Survivor Function Estimation Using Censored Multivariate Failure Time Data,*Biometrika*, 79, 495-512(1992).

13. PONS, O., A test of Independence Between Two Censored Survival Times, *Scand. Jour. Statist.*, 13, 173-185(1986)

14. STUTE, W.,Almost Sure Representations of the Product-Limit Estimator for Truncated Data, *Annsals of Statistics*, 21, 146-156(1993).

15. STUTE, W., Consistent Estimation Under Random Censorship When Covariables are Present, *Journal of Multivariate Analysis*, 45, 89-103 (1993).

16. UZUNOGULLARI,Ü., and WANG, J-L. On the Hajek Projection of the Kernel hazard Estimators, *Sankhya*, (1993)

17. WOODROOFE, M., Estimating a Distribution Function With Truncated Data, *Annals of Statistics*, 13, 163-177(1985).