Conference at Shobi University, Japan
February 20-21, 2014

Algorithms for estimating survival function under dependent left-truncation

- with applications to elderly residents' lifetime analysis.

Takeshi Emura
Graduate Institute of Statistics,
National Central University, Taiwan

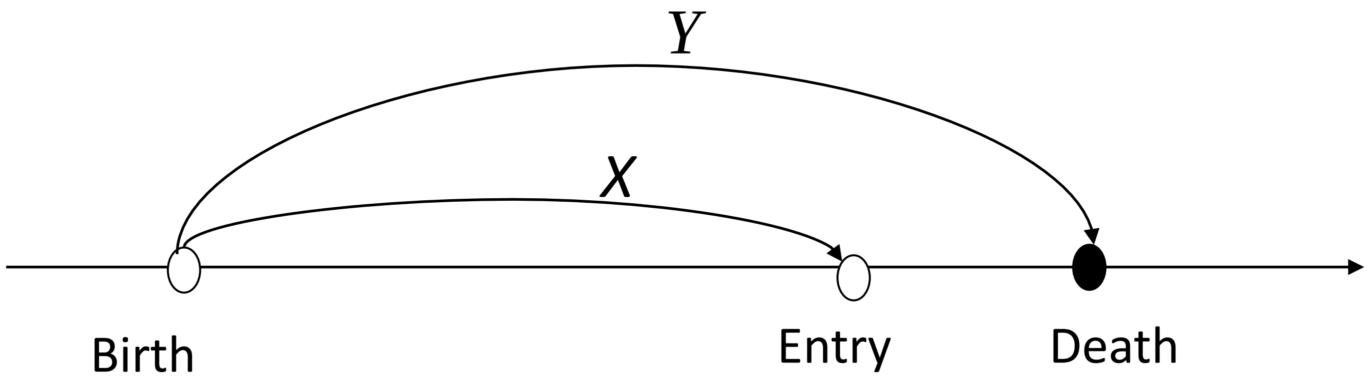
Outlines

- Left-truncated survival data - review

  (elderly residents’ survival in Channing house)

- Product-limit estimator - review
- Copula-based estimator - review
- Proposed algorithm
- Data analysis
- Conclusion

Left-truncated survival data



Channing House data (Hyde, 1977, 1980)

Channing house is a retirement center in California

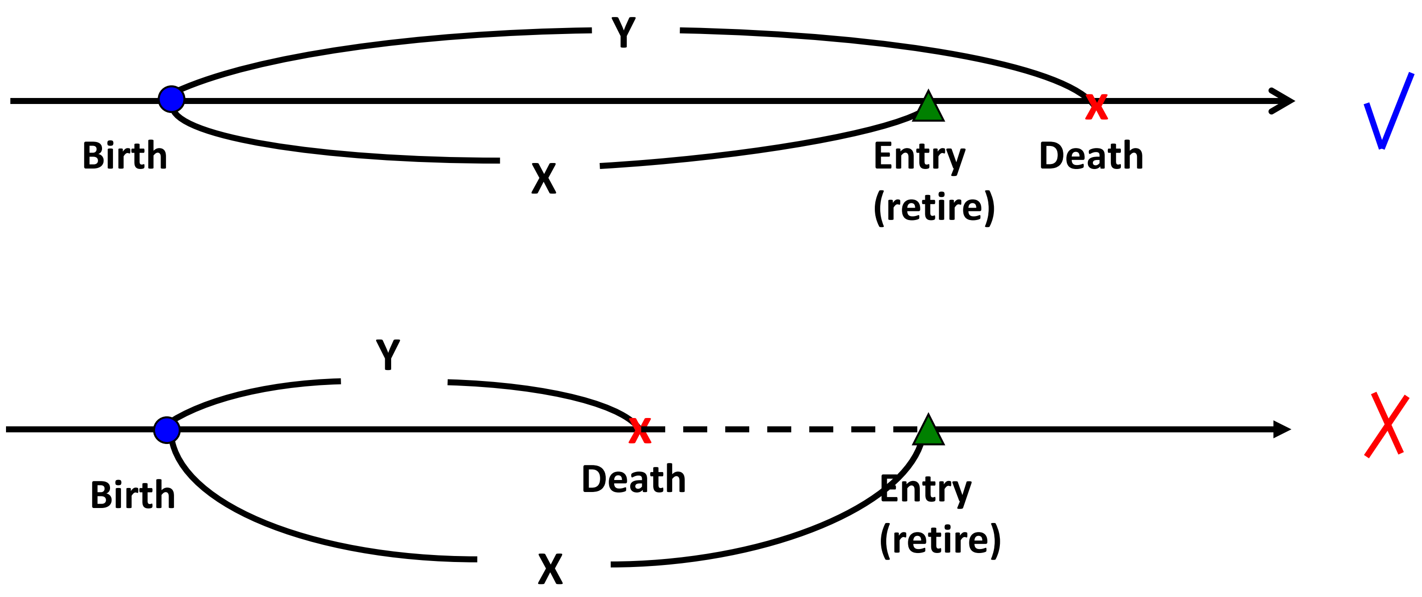
- n = 97 elderly residents in the Channing house
- Age at entry = X
- Age at death = Y (possibly right-censored)

Left-truncation criterion:

$$X \le Y$$

Left-truncated survival data

- Lifetime Y → left-truncated by X



Linear interpolation in the 1958 Commissioners Standard Ordinary Mortality Table for Male Lives was used to generate the distribution F, and hence h_t, for each month. The table stops at age 100, so that the data must be artificially censored at 1200 months. This does not affect the data in Table 1, but it does mean that $E(\lambda^*)$ is finite.

The observed number of deaths was 46, and the expected number was 72·2. The estimated variance was 68·3. The value of the statistic is thus $(46-72{\cdot}2)/\sqrt{68{\cdot}3} = -3{\cdot}16$, which indicates that the null hypothesis should be rejected in favour of a smaller hazard rate.

Entry Death

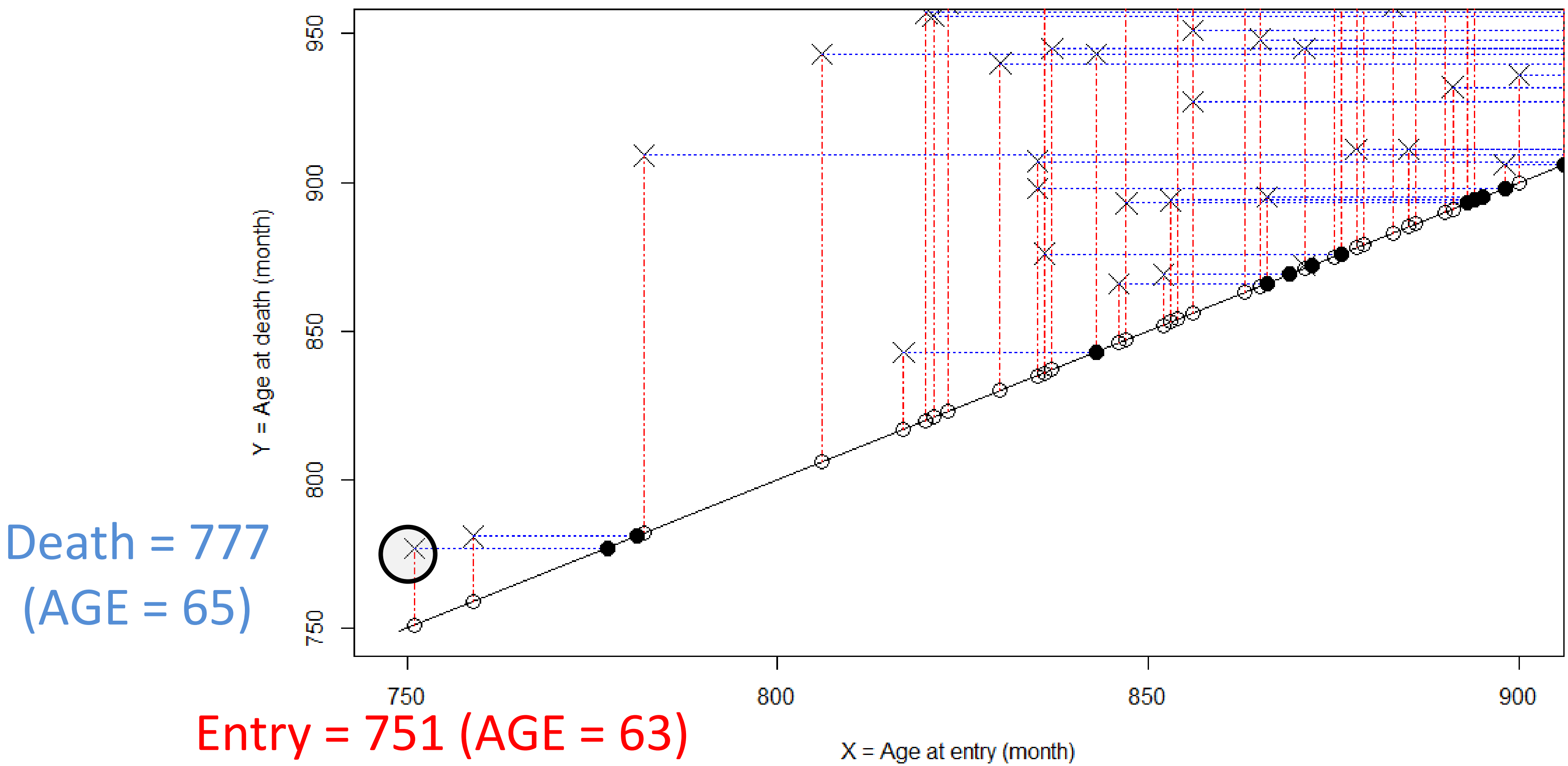
Table 1. *Survival data for males in a retirement community*

δ	$\nu+1$	λ^*	δ	$\nu+1$	λ^*	δ	$\nu+1$	λ^*	δ	$\nu+1$	λ^*	δ	$\nu+1$	λ^*
1	782	909	0	865	1002	1	854	989	0	1016	1153	0	959	972
1	1020	1128	0	953	1031	0	890	1027	0	969	1106	1	921	993
1	856	969	0	871	945	0	1041	1044	0	900	936	1	836	876
1	915	957	0	982	1006	0	978	1005	0	898	906	1	919	993
1	863	983	0	883	959	0	836	973	0	846	866	1	751	777
1	906	1012	0	817	843	0	938	1064	1	964	1029	1	906	966
1	955	1055	0	875	1012	0	886	1023	1	984	1053	1	835	907
1	943	1025	0	821	956	0	876	1013	1	1046	1080	1	946	1031
1	943	1043	0	936	1073	0	955	977	1	871	872	1	759	781
1	837	945	0	971	1107	0	960	1047	1	847	893	0	909	914
1	966	1009	0	830	940	0	843	943	0	962	966	1	962	998
1	936	971	0	885	911	0	856	951	1	853	894	1	984	1022
1	919	1033	0	894	1031	0	847	984	1	967	985	1	891	932
1	852	869	0	893	996	0	1027	1058	1	1063	1094	1	835	898
1	1073	1139	0	866	895	0	988	1045	1	856	927	1	1039	1060
1	925	1036	0	878	1015	0	953	953	1	865	948	1	1010	1044
1	967	1085	0	820	957	0	978	1018	1	1051	1059	0	823	960
0	806	943	0	1007	1043	0	981	1118	1	1010	1012			
0	969	1001	0	879	1016	0	926	970	1	878	911			
0	923	1060	0	956	1093	0	1036	1070	1	1021	1094			

Here $\delta = 1$ if subject died during study, $\delta = 0$ otherwise; $\nu+1$ is 1 + age in months at entry into study; λ^* is age in months when last seen in study.

If the discrete version is viewed as an approximation for the continuous case, and if it is assumed that the actual hazard rate is a multiple c of the hazard rate corresponding to F, then the ideas of § 4 can be applied. A 90 % confidence interval for c is (0·500, 0·812), and the approximate median unbiased estimate of c is 0·637.

I would like to thank Dr Rupert Miller for his support and sound advice, and Dr Bradley Efron for some helpful comments. Dr Walter Bortz kindly permitted me to use the data in



- Hyde (1980) assumed:

  knowing the person's entry age will provide no additional information about prospects for survival
- This means $X \perp Y$

  (X: Age at entry Y: Age at death)

Left-truncated survival data

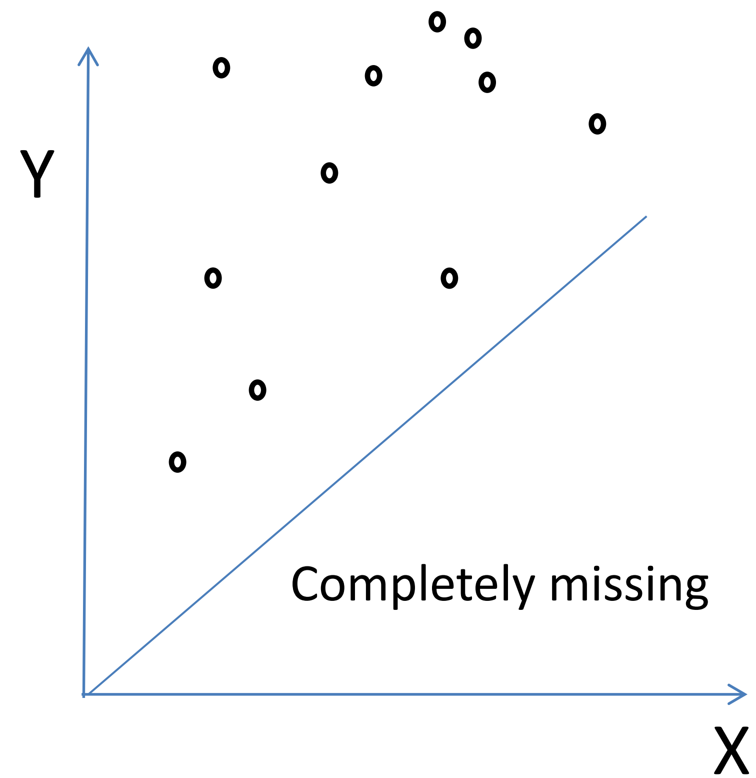
- <u>Left-truncated data</u>

$$\{(X_j, Y_j);\ j = 1, \ldots, n\}$$

subject to $X_j \le Y_j$

$$\Uparrow$$

i.i.d. from $\mathrm{P}(X \le x, Y \le y \mid X \le Y)$



- <u>Quasi-independence assumption</u> (Tsai, 1990)

$$\mathrm{H}_0:\ \ \Pr(X = x, Y = y \mid X \le Y) \propto dF_X(x)\, dF_Y(y)$$

Estimating survival

- Target : $S_Y(t) = P(Y > t)$
- Product-limit representation

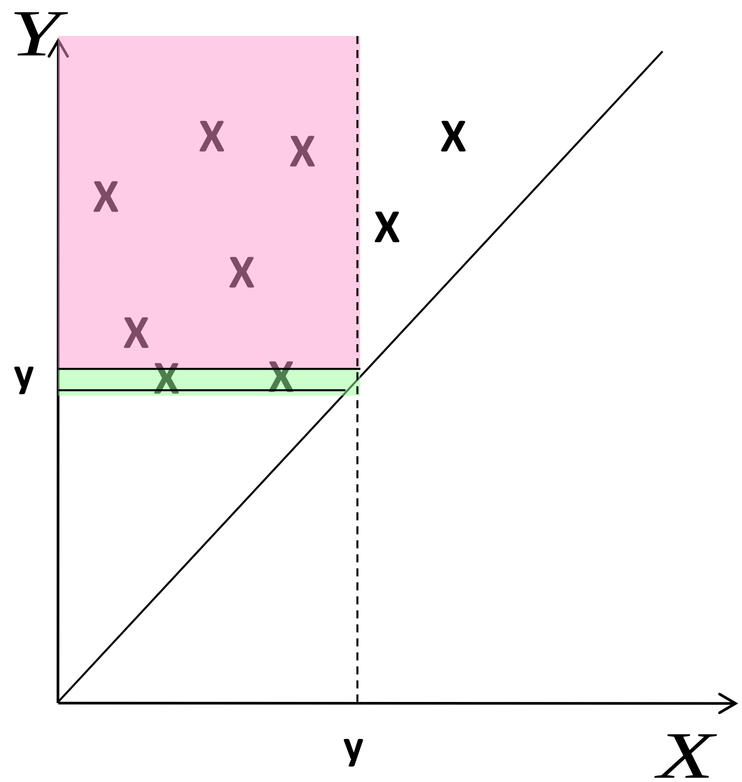
$$S_Y(t) = \prod_{u \le t} \{ 1 - P(Y = u \mid Y \ge u) \}$$

- Hazard function

$$P(Y = u \mid Y \ge u) = \frac{P(Y = u)}{P(Y \ge u)}$$

$$= \frac{P(Y = u, X \le u)}{P(Y \ge u, X \le u)}$$

Quasi-independence



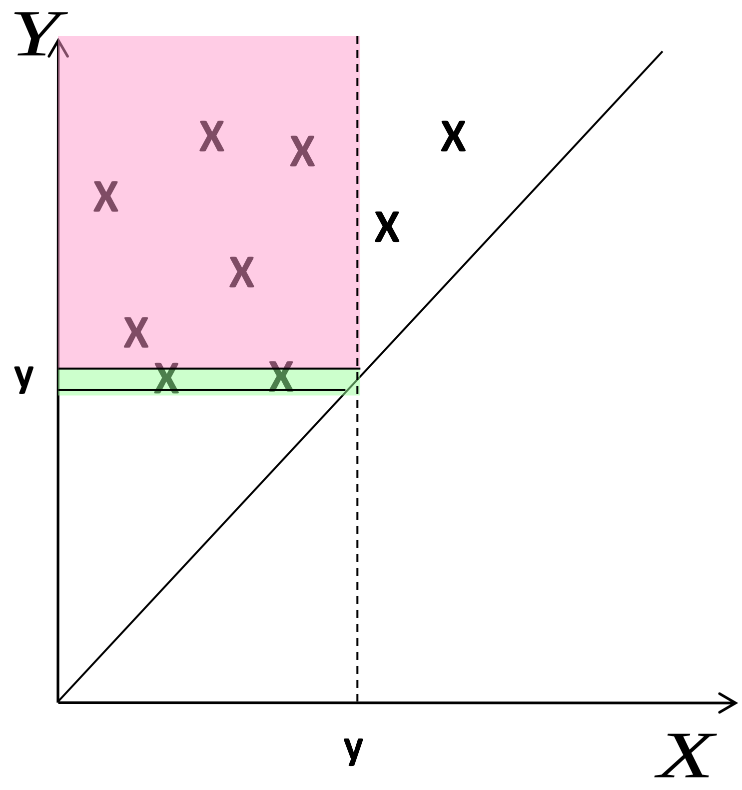
Estimating survival

- Product-limit estimator of $S_Y(y)$

  (Lynden-Bell 1971)

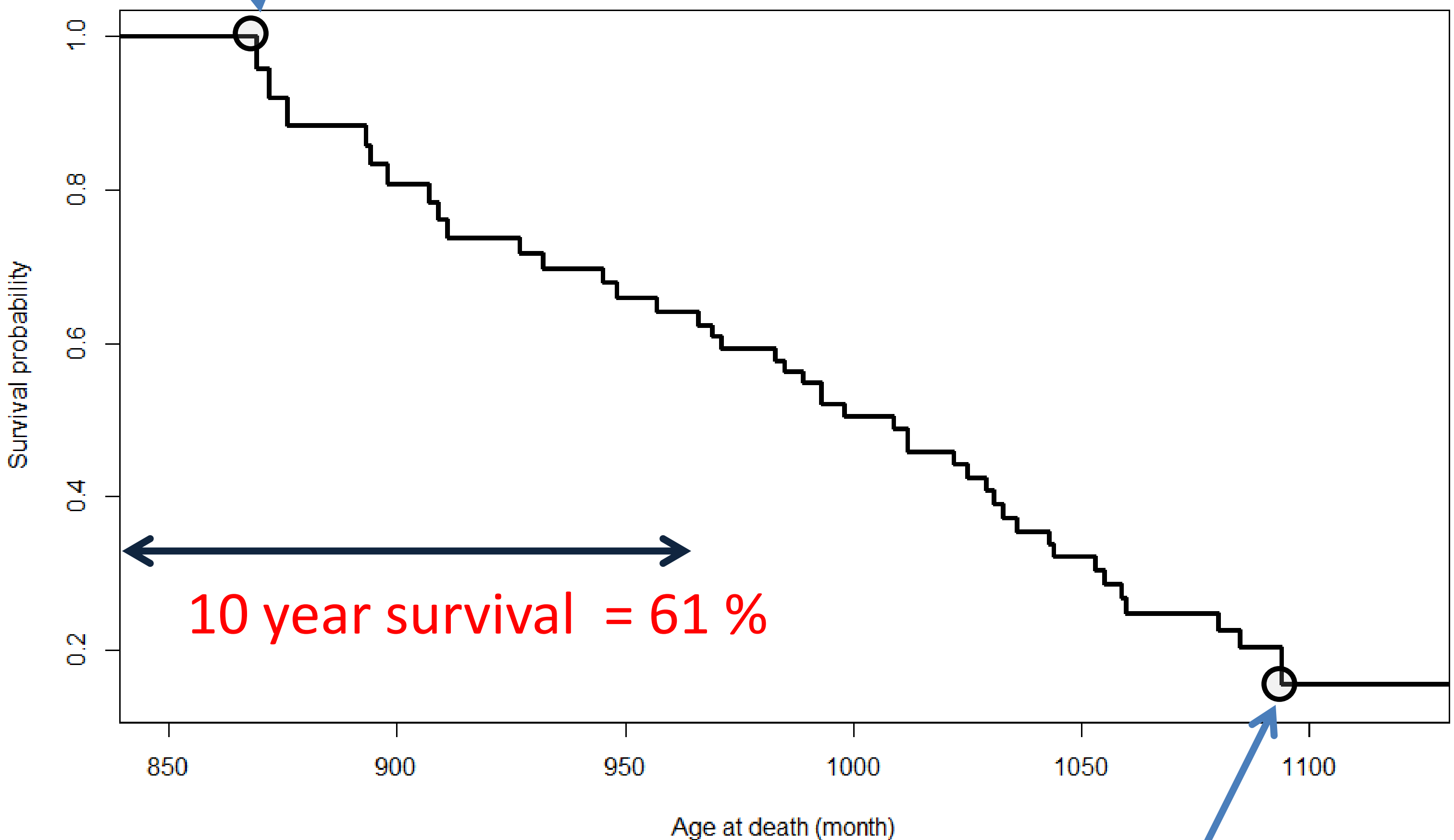
$$\hat{S}_Y(y) = \prod_{u \le y} \left\{ 1 - \frac{\sum_{i=1}^{n} I(X_i < u, Y_i = u)}{\sum_{i=1}^{n} I(X_i < u, Y_i \ge u)} \right\}$$

$$\mathrm{S}_Y(t) = \prod_{u \le t} \{ 1 - P(Y = u \mid Y \ge u) \}$$



Product-limit Estimates of $S_Y(y)$

869 month
(AGE = 72)



1094 month
(AGE = 91)

Testing quasi-independence

Testing quasi-independence

$$\mathrm{H}_0: \quad \Pr(X = x, Y = y \mid X \le Y) \propto dF_X(x)dF_Y(y)$$

Available test statistics:

1. Chen et al. (1996 JASA)

   - Based on the conditional Peason-correlation

2. Tsai (1990 Biometrika); Martin & Betensky (2005 JASA)

   -Based on the conditional Kendall's tau

3. Emura & Wang (2010 JMVA) - Based on weighted-logrank test

   (Optimal weight choice)

Testing quasi-indpendence

$$H_0: \ \Pr(X = x, Y = y \mid X \le Y) \propto dF_X(x)dF_Y(y)$$

is rejected at 5 % level

Table 4 of Emura and Wang (2010).

Tests of quasi-independence for the Channing House data.

	Logrank test $L_{\rho=1}$	Tsai test	Marting & Betensky test
P-value	0.048	0.043	0.040

- Quasi-independence is questionable

  → Product-limit estimates of survival probability may be biased

- In Channing house data, the truncation (entry age) may be informative on survival.

  → Motivate Copula modeling for dependent truncation

  Chaieb, Rivest, Abdous (2006, Biometrika)

  Beaudoin and Lakhal-Chaieb (2008 Stat. Med.),

  Emura and Wang (2010, JMVA)

  Emura, Wang and Hung (2011 Sinica)

  Emura and Wang (2012, JMVA)

  Ding (2012 Lifetime Data Analysis)

Copula model for dependent truncation

Let $\pi(x, y) \equiv \Pr(X \le x, Y > y \mid X \le Y)$

$\bullet$ Copula model (Chaieb et al, 2006):

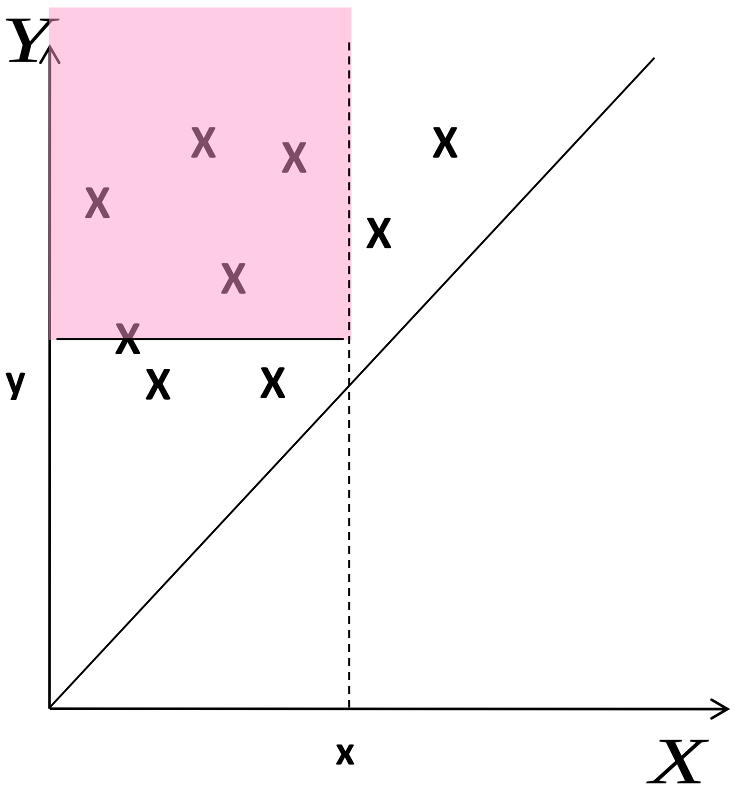
$$\pi(x, y) = \phi_\alpha^{-1}[\phi_\alpha\{F_X(x)\} + \phi_\alpha\{S_Y(y)\}]/c$$

$\bullet$ Clayton copula:

$$\phi_\alpha(t) = (t^{-(\alpha-1)} - 1)/(\alpha - 1)$$

$$\Rightarrow \Pr(X \le x, Y > y \mid X \le Y)$$

$$= (1/c)[F_X(x)^{-(\alpha-1)} + S_Y(y)^{-(\alpha-1)} - 1]^{-\frac{1}{\alpha-1}}$$



Copula model for dependent truncation

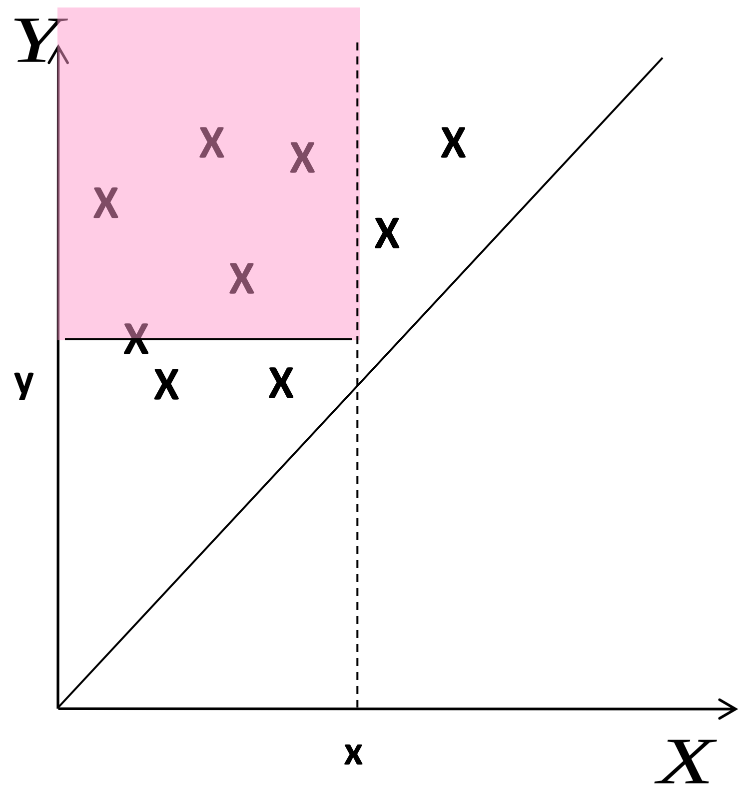
$$\pi(x, y) \equiv \Pr(X \le x, Y > y \mid X \le Y)$$

$$\pi(x, y)$$
$$= \phi_\alpha^{-1}[\phi_\alpha\{F_X(x)\} + \phi_\alpha\{S_Y(y)\}]/c$$

Chaieb et al (2006): Plug in

$$\hat{\pi}(x, y) \equiv \frac{1}{n}\sum_{j=1}^{n} \mathbf{I}(X_j \le x, Y_j > y)$$

$\Rightarrow$ Get the estimator of (α, c, F_X, S_Y)



Estimating equation

$$\hat{\pi}(t,t) = \phi_\alpha^{-1}[\phi_\alpha\{F_X(t)\} + \phi_\alpha\{S_Y(t)\}]/c$$
$$\Leftrightarrow \phi_\alpha\big(c\hat{\pi}(t,t)\big) = \phi_\alpha(F_X(t)) + \phi_\alpha(S_Y(t))$$

Chaieb et al. (2006) use some algebraic techniques of Rivest and Wells (2001, JMVA) to get solutions:

$$\hat{S}_Y(t) = \phi_\alpha^{-1}\left(-\sum_{j:Y_j \le t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(Y_j)-1}{n}\right\}\right]\right)$$

$$\hat{F}_X(t) = \phi_\alpha^{-1}\left(-\sum_{j:X_j > t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(X_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_j)-1}{n}\right\}\right]\right)$$

Estimating equation

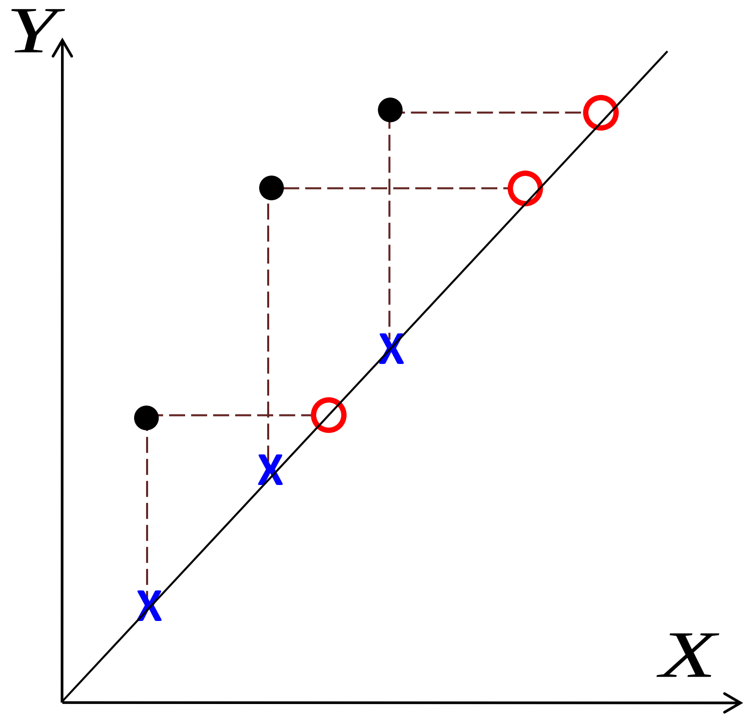
$$\hat{\pi}(t,t) = \phi_\alpha^{-1}[\phi_\alpha\{F_X(t)\} + \phi_\alpha\{S_Y(t)\}]/c$$
$$\Leftrightarrow \phi_\alpha\left(c\hat{\pi}(t,t)\right) = \phi_\alpha(F_X(t)) + \phi_\alpha(S_Y(t))$$

In this research:

I propose an new algorithm to solve the estimating equation

- New algorithm is easier to understand (straightforward derivation)
- New algorithm yields the same solution as Chaieb et al. (2006)

Proposed algorithm

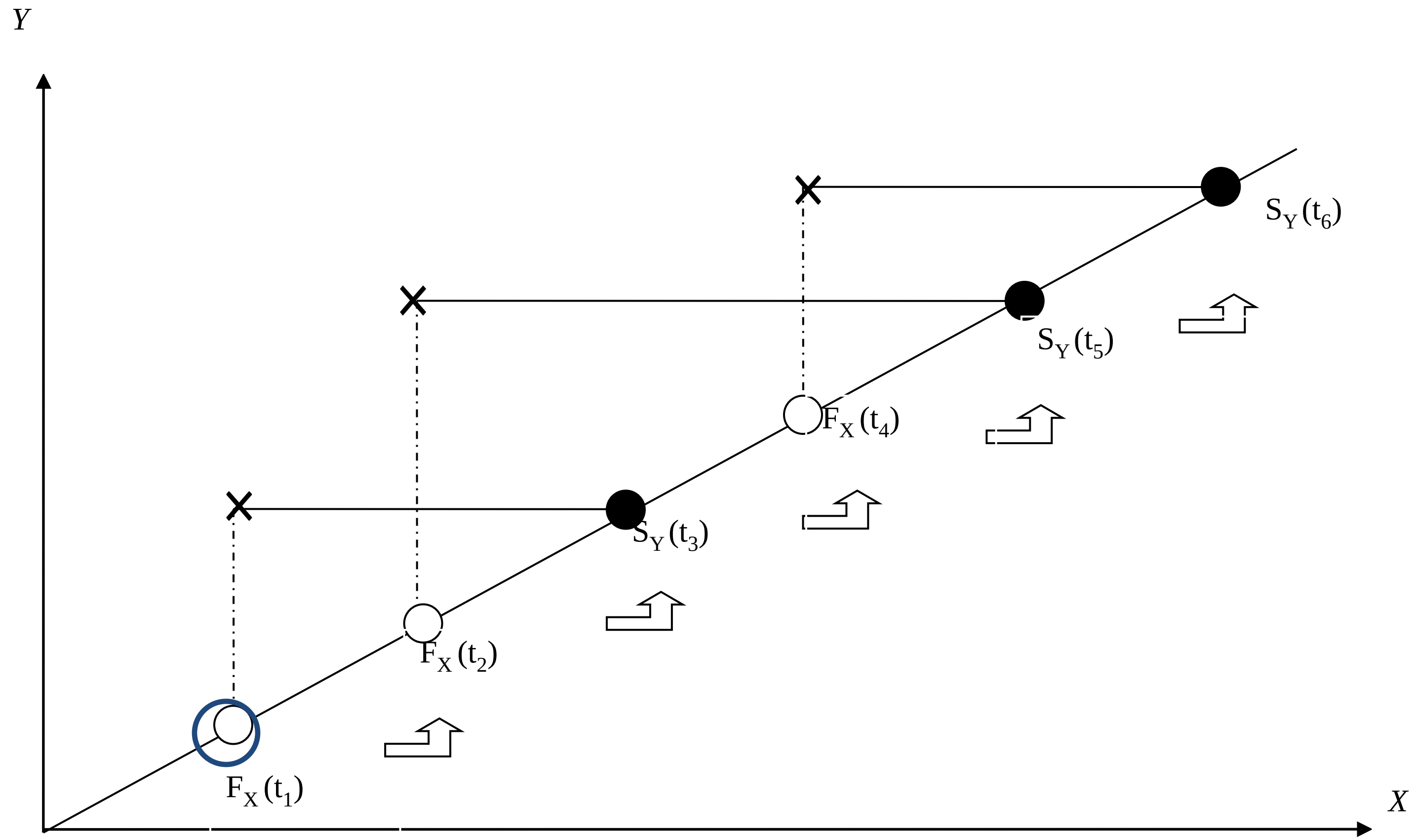


Last die

$$(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \Rightarrow t_1 < \cdots < t_{2n-1} < t_{2n}$$

1st Entry

Solving: $\phi_\alpha\left(c\hat{\pi}(t,t)\right) = \phi_\alpha(F_X(t)) + \phi_\alpha(S_Y(t))$



At t_1 = 1st Entry, nobody die → $S_Y(t_1) = 1$

$$\rightarrow \phi_\alpha(F_X(t_1)) = \phi_\alpha\left(c\hat{\pi}(t_1,t_1)\right) - \phi_\alpha(S_Y(t_1)) = \phi_\alpha\left(c/3\right) - \phi_\alpha(1) = \phi_\alpha\left(c/3\right)$$

$$\therefore F_X(t_1) = c/3$$

Table 1: Results of performing Step 1 of the proposed algorithm for a small dataset: $(X_1, Y_1) = (1, 3)$, $(X_2, Y_2) = (2, 5)$ and $(X_3, Y_3) = (4, 6)$.

1st Entry	$\hat{\pi}(t_j, t_j)$	$\phi_\alpha\{c\hat{\pi}(t_j, t_j)\}$	$\phi_\alpha\{\hat{F}_X(t_j)\}$	$\phi_\alpha\{\hat{S}_Y(t_j)\}$
$t_1 = X_1 = 1$	$\frac{1}{3}$	$\phi_\alpha\left(\frac{c}{3}\right)$	$\phi_\alpha\left(\frac{c}{3}\right)$	0
$t_2 = X_2 = 2$	$\frac{2}{3}$	$\phi_\alpha\left(\frac{2c}{3}\right)$	$\phi_\alpha\left(\frac{2c}{3}\right)$	0
$t_3 = Y_1 = 3$	$\frac{1}{3}$	$\phi_\alpha\left(\frac{c}{3}\right)$	$\phi_\alpha\left(\frac{2c}{3}\right)$	$\phi_\alpha\left(\frac{c}{3}\right) - \phi_\alpha\left(\frac{2c}{3}\right)$
$t_4 = X_3 = 4$	$\frac{2}{3}$	$\phi_\alpha\left(\frac{2c}{3}\right)$	$2\phi_\alpha\left(\frac{2c}{3}\right) - \phi_\alpha\left(\frac{c}{3}\right)$	$\phi_\alpha\left(\frac{c}{3}\right) - \phi_\alpha\left(\frac{2c}{3}\right)$
$t_5 = Y_2 = 5$	$\frac{1}{3}$	$\phi_\alpha\left(\frac{c}{3}\right)$	$2\phi_\alpha\left(\frac{2c}{3}\right) - \phi_\alpha\left(\frac{c}{3}\right)$	$2\phi_\alpha\left(\frac{c}{3}\right) - 2\phi_\alpha\left(\frac{2c}{3}\right)$
Last die $t_6 = Y_3 = 6$	$\frac{0}{3}$	$\phi_\alpha(0)$	Undetermined	Undetermined

Solution of the proposed algorithm

- 1) Under the Clayton $\phi_\alpha(t) = (t^{-(\alpha-1)} - 1)/(\alpha - 1)$

$$\hat{F}_X(X_1) = 1/3, \quad \hat{F}_X(X_2) = 2/3, \quad \hat{F}_X(X_3) = 1,$$

$$\hat{S}_Y(Y_{(1)}) = 2/3, \quad \hat{S}_Y(Y_{(2)}) = 1/3, \quad \hat{S}_Y(Y_{(3)}) = \text{undetermined.}$$

- 2) Under the quasi-independence $\phi_\alpha(t) = -\log(t)$

$$\hat{F}_X(X_1) = 1/4, \quad \hat{F}_X(X_2) = 1/2, \quad \hat{F}_X(X_3) = 1,$$

$$\hat{S}_Y(Y_{(1)}) = 1/2, \quad \hat{S}_Y(Y_{(2)}) = 1/4, \quad \hat{S}_Y(Y_{(3)}) = \text{undetermined.}$$

Proposed method

Proposed algorithm → Explicit formula

$$\hat{S}_Y(t) = \phi_\alpha^{-1}\left(-\sum_{j:\,Y_j \le t} \left[\phi_\alpha\left\{ c\,\frac{\tilde{R}(Y_j)}{n} \right\} - \phi_\alpha\left\{ c\,\frac{\tilde{R}(Y_j)-1}{n} \right\} \right] \right)$$

$$\hat{F}_X(t) = \phi_\alpha^{-1}\left(\sum_{j:\,t_1 < X_j \le t} \left[\phi_\alpha\left\{ c\,\frac{\tilde{R}(X_j)}{n} \right\} - \phi_\alpha\left\{ c\,\frac{\tilde{R}(X_j)-1}{n} \right\} \right] + \phi_\alpha\left(\frac{c}{n} \right) \right)$$

Proposed vs. Chaieb et al.

1) Proposed

$$\hat{S}_Y(t) = \phi_\alpha^{-1}\left(-\sum_{j:Y_j \le t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(Y_j)-1}{n}\right\}\right]\right)$$

$$\hat{F}_X(t) = \phi_\alpha^{-1}\left(\sum_{j:\ t_1 < X_j \le t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(X_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_j)-1}{n}\right\}\right] + \phi_\alpha\left(\frac{c}{n}\right)\right)$$

2) Chaieb et al. (2006)

$$\hat{S}_Y(t) = \phi_\alpha^{-1}\left(-\sum_{j:Y_j \le t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(Y_j)-1}{n}\right\}\right]\right)$$

$$\hat{F}_X(t) = \phi_\alpha^{-1}\left(-\sum_{j:X_j > t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(X_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_j)-1}{n}\right\}\right]\right)$$

Two estimators (Proposed vs. Chaieb et al.) are the same

***Theorem 1**: The proposed estimating equation (10) is equivalent to the estimating equation (6) of Chaieb et al. (2006) under* $x_0 \in [X_{(n)}, t_{2n-1}]$.

Proof: Note that $\phi_\alpha\{\hat{F}_X(t_{2n-1})\} = \phi_\alpha\{\hat{F}_X(x_0)\}$ since there is no jump for X beyond $x_0 \in [X_{(n)}, t_{2n-1}]$. Thus, the estimating equation (10) becomes

$$\begin{aligned}
U_c(\alpha, c) &= \phi_\alpha\{\hat{F}_X(x_0)\} \\
&= \phi_\alpha\left\{c\frac{R(x_0, x_0+)}{n}\right\} - \phi_\alpha\{S_Y(x_0)\} \\
&= \phi_\alpha\left\{c\frac{R(x_0, x_0+)}{n}\right\} + \sum_{j:Y_j \le x_0}\left[\phi_\alpha\left\{c\frac{\widetilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\widetilde{R}(Y_j)-1}{n}\right\}\right] \\
&= \phi_\alpha\left\{c\frac{\widetilde{R}(x_0)}{n}\right\} + \sum_{j:Y_j < x_0}\left[\phi_\alpha\left\{c\frac{\widetilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\widetilde{R}(Y_j)-1}{n}\right\}\right],
\end{aligned}$$

where the last equation is equivalent to Equation (6).□

Extension for right-censoring

- Left - truncation + Right - censoring

$$Z_j = \min(Y_j, C_j) = \min(\text{ Age at death , withdrawal })$$

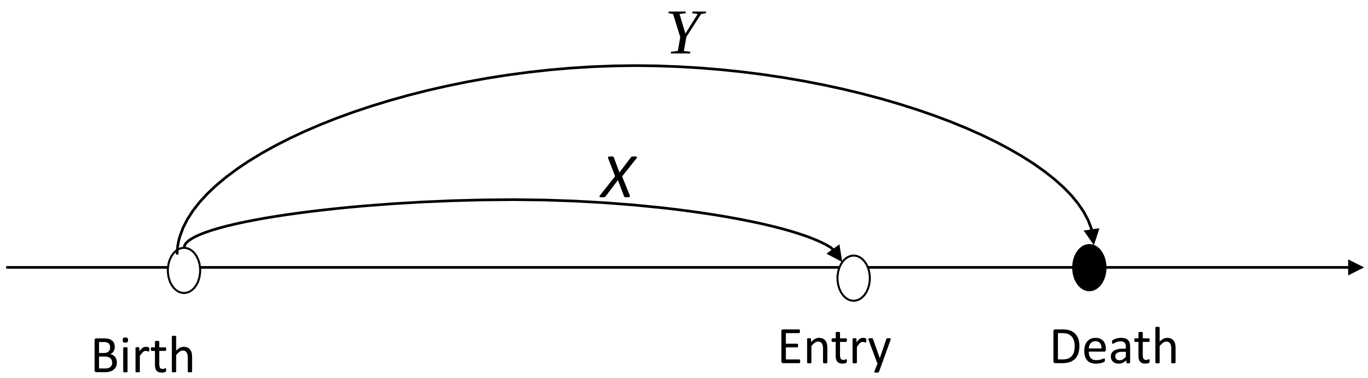
$$X_j = \text{Entry age}$$

$$\delta_j = \begin{cases} 1 & \text{die during the study : } Y_j \le C_j \\ 0 & \text{withdraw from the study: } Y_j > C_j \end{cases}$$

$$\text{Data : } \{(X_j, Z_j, \delta_j);\ j = 1, \ldots, n\} \ \text{ subject to } X_j \le Z_j$$

$$\Rightarrow \hat{S}_Y(t) = \phi_\alpha^{-1}\left(-\sum_{j;\, z_j \le t,\, \delta_j = 1} \left[\phi_\alpha\left\{ c^* \frac{\tilde{R}(Z_j)}{n\hat{S}_C(Z_j)} \right\} - \phi_\alpha\left\{ c^* \frac{\tilde{R}(Z_j) - 1}{n\hat{S}_C(Z_j)} \right\} \right] \right)$$

Data analysis



Channing House data (Hyde, 1977, 1980)

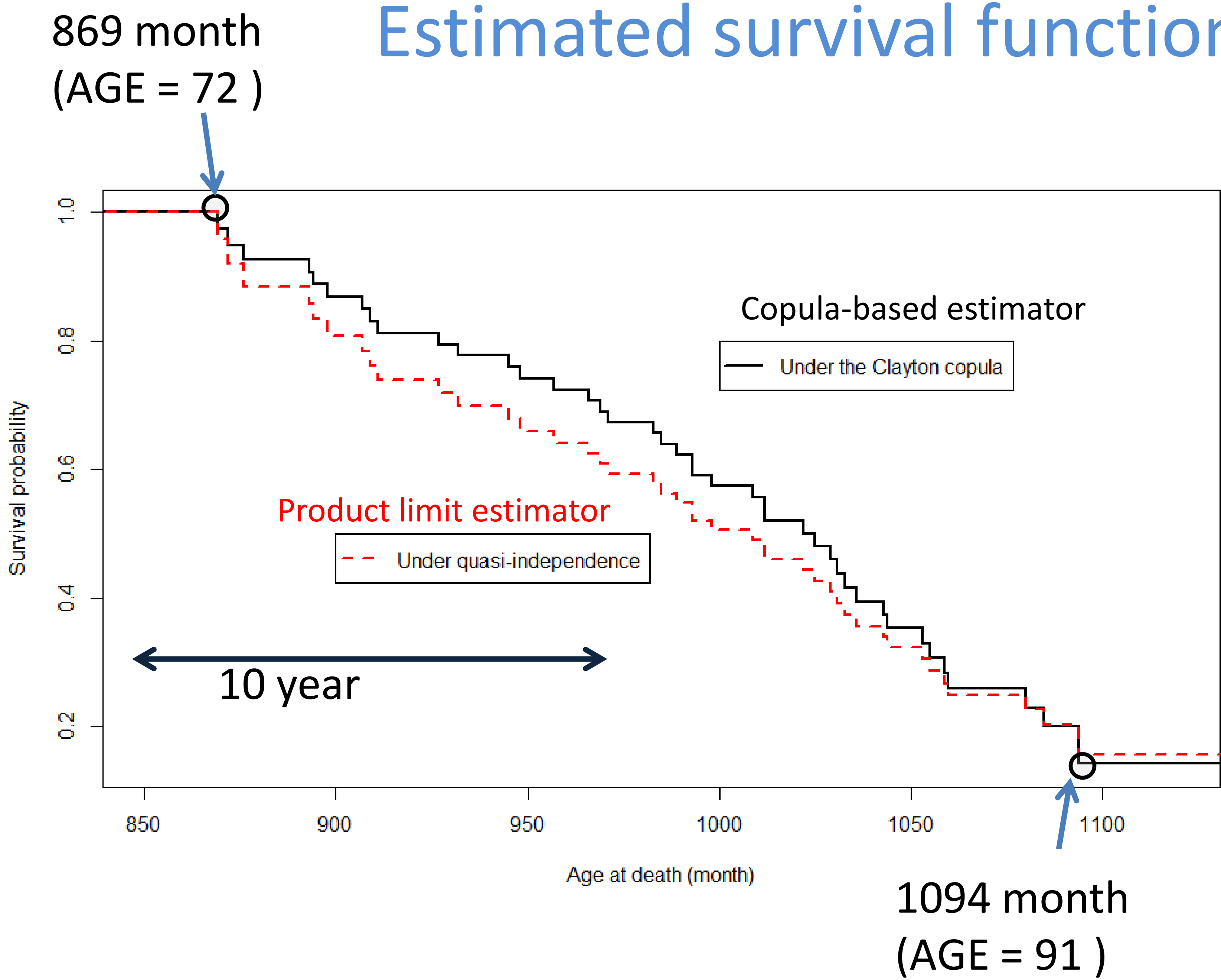
n = 97 elderly residents in the Channing house

- Age at entry = X
- Age at death = Z = min (Y, C)
  (possibly right-censored by withdrawal)

Data:

$$(X_j, Z_j, \delta_j);\ j = 1, \ldots, n \quad \text{subject to } X_j \le Z_j$$

Estimated survival function



Interpretation

- 10 year survival probability

  68.9% (Copula-based estimator)

  60.9% (Product limit estimator)

  Product-limit estimator substantially underestimate the benefit of the Channing house

- 20 year survival probability

  20.0% (Copula-based estimator)

  20.3% (Product limit estimator)

  No difference in long-term survivorship

Summary

- The product-limit estimator for survival function relies on the quasi-independence on left-truncation (Tsai 1990)
- Quasi-independence is rejected for the Channing house data (Emura and Wang 2010).
- Copula models relax the quasi-independence by introducing the dependence between survival and truncation (Chaieb et al., 2006).
- In this research, I propose a new algorithm to solve the estimating equation of Chaieb et al. (2006). R depend.truncation package http://cran.r-project.org/web/packages/depend.truncation/index.html
- Copula approaches yield high survival probability for elderly residents in the Channing house residents than the product limit estimator does.

References

- Chen CH, Tsai WY, Chao WH (1996) The product-moment correlation coefficient and linear regression for truncated data. Journal of the American Statistical Association 91, 1181-1186
- Chaieb LL, Rivest LP, Abdous B (2006) Estimating survival under a dependent truncation. Biometrika 93: 655-69
- Ding AA (2012) Copula identifiability conditions for dependent truncated data model. Lifetime Data Analysis 18 (4): 397-407
- Emura T, Wang W (2010) Testing quasi-independence for truncation data. Journal of Multivariate Analysis 101: 223-239
- Emura T, Wang W, Hung HN (2011) Semi-parametric inference for copula models for truncated data. Statistica Sinica 21: 349-367
- Emura T, Wang W (2012) Nonparametric maximum likelihood estimation for dependent truncation data based on copulas. Journal of Multivariate Analysis 110: 171-188
- Hyde J (1977) Testing survival under right censoring and left truncation. Biometrika 64: 225-230
- Hyde J (1980) Survival analysis with incomplete observations, In Biostatistics Casebook, R. G. Miller, B. Efron, B. W. Brown, and L. E. Moses, eds. New York; John Wiley and Son, 31-46
- Lynden-Bell D (1971) A method of allowing for known observational selection in small samples applied to 3RC quasars. Mon Nat R Astron Soc Lett 155: 95-118
- Martin EC, Betensky RA (2005) Testing quasi-independence of failure and truncation via conditional Kendall's tau. Journal of the American Statistical Association 100: 484-92
- Rivest LP, Wells MT (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. Journal of Multivariate Analysis 79: 138-55
- Tsai WY (1990) Testing the association of independence of truncation time and failure time. Biometrika 77: 169-177