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
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:

$X \leq Y$
Left-truncated survival data

- Lifetime $Y \rightarrow$ 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 \( \lambda \), 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 \((46 - 72.2)/\sqrt{68.3} = -3.16\), which indicates that the null hypothesis should be rejected in favour of a smaller hazard rate.

<table>
<thead>
<tr>
<th>Entry</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta )</td>
<td>( \nu + 1 )</td>
</tr>
<tr>
<td>1</td>
<td>782</td>
</tr>
<tr>
<td>1</td>
<td>1020</td>
</tr>
<tr>
<td>1</td>
<td>856</td>
</tr>
<tr>
<td>1</td>
<td>915</td>
</tr>
<tr>
<td>1</td>
<td>863</td>
</tr>
<tr>
<td>1</td>
<td>906</td>
</tr>
<tr>
<td>1</td>
<td>955</td>
</tr>
<tr>
<td>1</td>
<td>943</td>
</tr>
<tr>
<td>1</td>
<td>943</td>
</tr>
<tr>
<td>1</td>
<td>837</td>
</tr>
<tr>
<td>1</td>
<td>966</td>
</tr>
<tr>
<td>1</td>
<td>936</td>
</tr>
<tr>
<td>1</td>
<td>919</td>
</tr>
<tr>
<td>1</td>
<td>852</td>
</tr>
<tr>
<td>1</td>
<td>1073</td>
</tr>
<tr>
<td>1</td>
<td>925</td>
</tr>
<tr>
<td>1</td>
<td>967</td>
</tr>
<tr>
<td>0</td>
<td>806</td>
</tr>
<tr>
<td>0</td>
<td>969</td>
</tr>
<tr>
<td>0</td>
<td>923</td>
</tr>
</tbody>
</table>

Here \( \delta = 1 \) if subject died during study, \( \delta = 0 \) otherwise; \( \nu + 1 \) is 1 + age in months at entry into study; \( \lambda^* \) 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

- **Left-truncated data**
  \[ \{(X_j, Y_j); j = 1, \ldots, n\} \]
  subject to \( X_j \leq Y_j \)

  i.i.d. from \( P(X \leq x, Y \leq y \mid X \leq Y) \)

- **Quasi-independence assumption** (Tsai, 1990)
  \[ H_0: \Pr(X = x, Y = y \mid X \leq Y) \propto dF_X(x)dF_Y(y) \]
Estimating survival

- **Target**: \( S_Y(t) = P(Y > t) \)

- **Product-limit representation**
  \[
  S_Y(t) = \prod_{u \leq t} \left\{ 1 - P(Y = u \mid Y \geq u) \right\}
  \]

- **Hazard function**
  \[
  P(Y = u \mid Y \geq u) = \frac{P(Y = u)}{P(Y \geq u)} = \frac{P(Y = u, X \leq u)}{P(Y \geq u, X \leq u)}
  \]

Quasi-independence
Estimating survival

- **Product-limit estimator of** $S_Y(y)$
  
  \[
  \hat{S}_Y(y) = \prod_{u \leq y} \left\{ 1 - \frac{\sum_{i=1}^{n} I(X_i < u, Y_i = u)}{\sum_{i=1}^{n} I(X_i < u, Y_i \geq u)} \right\}
  \]

  
  \[
  S_Y(t) = \prod_{u \leq t} \{ 1 - P(Y = u \mid Y \geq u) \}
  \]
Product-limit Estimates of $S_Y(y)$

869 month (AGE = 72 )

10 year survival = 61 %

1094 month (AGE = 91 )
Testing quasi-independence

\[ H_0 : \ Pr(X = x, Y = y \mid X \leq Y) \propto dF_x(x)dF_y(y) \]

Available test statistics:

1. Chen et al. (1996 JASA)
   - Based on the conditional Pearson-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-independence

\[ H_0 : \quad \Pr(X = x, Y = y | X \leq Y) \propto dF_X(x)dF_Y(y) \]

is rejected at 5\% level

*Table 4 of Emura and Wang (2010).*

Tests of quasi-independence for the Channing House data.

<table>
<thead>
<tr>
<th>Test</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logrank test (L_{\rho=1})</td>
<td>0.048</td>
</tr>
<tr>
<td>Tsai test</td>
<td>0.043</td>
</tr>
<tr>
<td>Marting &amp; Betensky test</td>
<td>0.040</td>
</tr>
</tbody>
</table>
• 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)
  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 \( \pi(x, y) \equiv \Pr(X \leq x, Y > y \mid X \leq Y) \)

- Copula model (Chaieb et al, 2006):
  \[
  \pi(x, y) = \phi_{\alpha}^{-1} \left[ \phi_{\alpha} \{ F_X(x) \} + \phi_{\alpha} \{ S_Y(y) \} \right] / c
  \]

- Clayton copula:
  \[
  \phi_{\alpha}(t) = \frac{t^{-(\alpha-1)} - 1}{(\alpha - 1)}
  \]

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

  \[
  = \frac{1}{c} \left[ F_X(x)^{-(\alpha-1)} + S_Y(y)^{-(\alpha-1)} - 1 \right]^{-1/\alpha-1}
  \]
Copula model for dependent truncation

\[
\pi(x, y) \equiv \Pr(X \leq x, Y > y \mid X \leq Y)
\]

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

Chaieb et al (2006): Plug in

\[
\hat{\pi}(x, y) = \frac{1}{n} \sum_{j=1}^{n} I( X_j \leq x, Y_j > y )
\]

⇒ Get the estimator of (α, c, F_X, S_Y)
Estimating equation

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

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

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

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

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

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

\[ \iff \phi_\alpha(c \hat{\pi}(t, t)) = \phi_\alpha(F_X(t)) + \phi_\alpha(S_Y(t)) \]

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

\[(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \Rightarrow t_1 < \cdots < t_{2n-1} < t_{2n}\]
Solving: \( \phi_{\alpha}(c \hat{\tau}(t,t)) = \phi_{\alpha}(F_X(t)) + \phi_{\alpha}(S_Y(t)) \)

At \( t_1 = 1^{st} \) Entry, nobody die \( \Rightarrow \) \( S_Y(t_1) = 1 \)

\[
\phi_{\alpha}(F_X(t_1)) = \phi_{\alpha}(c \hat{\tau}(t_1,t_1)) - \phi_{\alpha}(S_Y(t_1)) = \phi_{\alpha}(c / 3) - \phi_{\alpha}(1) = \phi_{\alpha}(c / 3)
\]

\( \therefore F_X(t_1) = c / 3 \)
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 \((X_3, Y_3) = (4, 6)\).

<table>
<thead>
<tr>
<th>1st Entry</th>
<th>(\hat{\pi}(t_j, t_j))</th>
<th>(\phi_\alpha{c\hat{\pi}(t_j, t_j)})</th>
<th>(\phi_\alpha{\hat{F}_X(t_j)})</th>
<th>(\phi_\alpha{\hat{S}_Y(t_j)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t_1 = X_1 = 1)</td>
<td>( \frac{1}{3} )</td>
<td>(\phi_\alpha\left(\frac{c}{3}\right))</td>
<td>(\phi_\alpha\left(\frac{c}{3}\right))</td>
<td>0</td>
</tr>
<tr>
<td>(t_2 = X_2 = 2)</td>
<td>( \frac{2}{3} )</td>
<td>(\phi_\alpha\left(\frac{2c}{3}\right))</td>
<td>(\phi_\alpha\left(\frac{2c}{3}\right))</td>
<td>0</td>
</tr>
<tr>
<td>(t_3 = Y_1 = 3)</td>
<td>( \frac{1}{3} )</td>
<td>(\phi_\alpha\left(\frac{c}{3}\right))</td>
<td>(\phi_\alpha\left(\frac{2c}{3}\right))</td>
<td>(\phi_\alpha\left(\frac{c}{3}\right) - \phi_\alpha\left(\frac{2c}{3}\right))</td>
</tr>
<tr>
<td>(t_4 = X_3 = 4)</td>
<td>( \frac{2}{3} )</td>
<td>(\phi_\alpha\left(\frac{2c}{3}\right))</td>
<td>(2\phi_\alpha\left(\frac{2c}{3}\right) - \phi_\alpha\left(\frac{c}{3}\right))</td>
<td>(\phi_\alpha\left(\frac{c}{3}\right) - \phi_\alpha\left(\frac{2c}{3}\right))</td>
</tr>
<tr>
<td>(t_5 = Y_2 = 5)</td>
<td>( \frac{1}{3} )</td>
<td>(\phi_\alpha\left(\frac{c}{3}\right))</td>
<td>(2\phi_\alpha\left(\frac{2c}{3}\right) - \phi_\alpha\left(\frac{c}{3}\right))</td>
<td>(2\phi_\alpha\left(\frac{c}{3}\right) - 2\phi_\alpha\left(\frac{2c}{3}\right))</td>
</tr>
<tr>
<td>Last die (t_6 = Y_3 = 6)</td>
<td>( \frac{0}{3} )</td>
<td>(\phi_\alpha(0))</td>
<td>Undetermined</td>
<td>Undetermined</td>
</tr>
</tbody>
</table>
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** \(\rightarrow\) **Explicit formula**

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

\[
\hat{F}_X(t) = \phi_{\alpha}^{-1}\left( \sum_{j: t_1 < X_j \leq t} \left[ \phi_{\alpha} \left\{ c \frac{\tilde{R}(X_j)}{n} \right\} - \phi_{\alpha} \left\{ c \frac{\tilde{R}(X_j) - 1}{n} \right\} \right] + \phi_{\alpha} \left( \frac{c}{n} \right) \right)
\]
Proposed vs. Chaieb et al.

1) Proposed

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

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

2) Chaieb et al. (2006)

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

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

$$U_c(\alpha, c) = \phi_\alpha\{\hat{F}_X(x_0)\}$$

$$= \phi_\alpha\left\{c \frac{R(x_0, x_0 +)}{n}\right\} - \phi_\alpha\{S_Y(x_0)\}$$

$$= \phi_\alpha\left\{c \frac{R(x_0, x_0 +)}{n}\right\} + \sum_{j: Y_j < x_0} \left[\phi_\alpha\left\{c \frac{\tilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c \frac{\tilde{R}(Y_j) - 1}{n}\right\}\right]$$

$$= \phi_\alpha\left\{c \frac{\tilde{R}(x_0)}{n}\right\} + \sum_{j: Y_j < x_0} \left[\phi_\alpha\left\{c \frac{\tilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c \frac{\tilde{R}(Y_j) - 1}{n}\right\}\right],$$

where the last equation is equivalent to Equation (6). $\square$
Extension for right-censoring

- **Left - truncation + Right - censoring**

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

\[ X_j = \text{Entry age} \]

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

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

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

$(X_j, Z_j, \delta_j); \ j = 1,...,n$ subject to $X_j \leq Z_j$
Estimated survival function

- Estimated survival at age 72: 869 months
- Estimated survival at age 91: 1094 months

Survival function estimated using:
- Product limit estimator
- Copula-based estimator

10-year survival:
- 0.2 survival probability at age 72
- 0.2 survival probability at age 91

Age at death (month):
- 850 to 1100 months
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). The R depend.truncation package [http://cran.r-project.org/web/packages/depend.truncation/index.html](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.
• Hyde J (1977) Testing survival under right censoring and left truncation. Biometrieka 64: 225-230
• Tsai WY (1990) Testing the association of independence of truncation time and failure time. Biometrieka 77: 169-177