Estimation of the association for bivariate interval censored data with copulas

Kris Bogaerts & Emmanuel Lesaffre

I-Biostat
Katholieke Universiteit Leuven & Universiteit Hasselt

18-20 September 2008
Missing Information in Survival Data Beyond Right Censoring
1. Tooth emergence
2. Association measures
3. Copula approach
4. Conclusions
Tooth emergence

**Definition**

*Emergence time of a tooth* is the chronological age of a child at which that tooth appears in the mouth.
Tooth emergence

Introduction

Definition

Emergence time of a tooth is the chronological age of a child at which that tooth appears in the mouth.

Usefulness

- Diagnosis of certain growth disturbances: supplement other maturity indicators
- Forensic dentistry: estimation of chronological age of a child with unknown birth records
- Dental practice: help in diagnosis and treatment planning
Signal Tandmobiel® Study

- Longitudinal dental study (Flanders)
- 4468 children followed-up: 7-12 years old (1996-2001)
- Annual dental examinations: trained dentists
- Permanent tooth emergence recorded
- Oral health behaviour
Left-Interval-Right Censored Data

0: Not Emerged; 1: Emerged

Left Censoring

\[
\begin{align*}
1 & \quad 1 & \quad 1 & \quad 1 & \quad 1 & \quad 1 & \quad 1 & \quad 1 \\
\times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times \\
\text{\textit{t}}_{1} & & & & & & & \\
\end{align*}
\]

Interval Censoring

\[
\begin{align*}
0 & \quad 0 & \quad 1 & \quad 1 & \quad 1 & \quad 1 & \quad 1 & \quad 1 \\
\times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times \\
\text{\textit{t}}_{i} & \quad \text{\textit{t}}_{i+1} & & & & & & \\
\end{align*}
\]

Right Censoring

\[
\begin{align*}
0 & \quad 0 & \quad 0 & \quad 0 & \quad 0 & \quad 0 & \quad 0 & \quad 0 \\
\times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times & \quad \times \\
\text{\textit{t}}_{6} & & & & & & & \\
\end{align*}
\]
Dental research questions

- “How strongly are the emergence times of maxillar first premolars related?”
Dental research questions

1. "How strongly are the emergence times of maxillary first premolars related?"
2. "Does the association between the emergence times of maxillary first premolars differ for boys and girls?"
3. "Does this association also depend on oral health behavior?"

Not fitting a fully parametric model
- Flexibly estimate the marginal distributions
- Parametrically model the association

Copulas

K. Bogaerts & E. Lesaffre (I-Biostat)
Dental research questions

- “How strongly are the emergence times of maxillary first premolars related?”
- “Does the association between the emergence times of maxillary first premolars differ for boys and girls?”
Dental research questions

- “How strongly are the emergence times of maxillary first premolars related?”
- “Does the association between the emergence times of maxillary first premolars differ for boys and girls?”
- “Does this association also depend on oral health behaviour?”
Dental research questions

- "How strongly are the emergence times of maxillary first premolars related?"
- "Does the association between the emergence times of maxillary first premolars differ for boys and girls?"
- "Does this association also depend on oral health behaviour?"

- Not fitting a fully parametric model
- Flexibly estimate the marginal distributions
- Parametrically model the association
- Copulas
Survival copula

Definition survival copula

Let $\tilde{C}(u, v)$ be the survival copula of $X$ and $Y$ from $[0, 1] \times [0, 1]$ into $[0, 1]$ given by

$$\tilde{C}(u, v) = u + v - 1 + C(1 - u, 1 - v).$$

We have then

$$S(x, y) = \tilde{C}(S_{F_1}(x), S_{F_2}(y)).$$
Association measures

- **Kendall’s τ**

\[ \tau = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1. \]

- **Spearman’s correlation**

\[ \rho_S = 12 \int_0^1 \int_0^1 u \cdot v dC(u, v) - 3 \]

\[ = 12 \int_0^1 \int_0^1 C(u, v) dudv - 3. \]
Survival copula

Copulas

- **Clayton copula**
  \[(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}\]
  with \(\theta \in [-1, \infty) \setminus \{0\}\)

- **Normal copula**
  \[\Phi_P(\Phi^{-1}(u), \Phi^{-1}(v))\]

- **Plackett copula**
  \[
  \frac{[1 + (\theta - 1)(u + v)] - \sqrt{[1 + (\theta - 1)(u + v)]^2 - 4uv\theta(\theta - 1)}}{2(\theta - 1)}
  \]
  with \(\theta \neq 1\)
Association measures

Association measures for the 3 copulas

- **Clayton copula**
  \[ \tau = \frac{\theta}{\theta + 2} \]

- **Normal copula**
  \[ \tau = \frac{2}{\pi} \arcsin \rho \]
  \[ \rho_S = \frac{6}{\pi} \arcsin \frac{1}{2} \rho \]

- **Plackett copula**
  \[ \rho_S = \frac{\theta + 1}{\theta - 1} - \frac{2\theta}{(\theta - 1)^2} \ln \theta \]
Estimation method

- \( S(t_1, t_2) = \hat{C}_\alpha(S_1(t_1), S_2(t_2)) \) with \( \alpha \) the copula parameter
Estimation method

- \( S(t_1, t_2) = \tilde{C}_\alpha(S_1(t_1), S_2(t_2)) \) with \( \alpha \) the copula parameter
- Extension from Sun et al. (2006) who estimated association with Clayton copula

Extension to allow covariates in the marginal distribution by fitting an AFT model with a flexible error distribution (Komárek et al. (2005))

Extension by allowing the copula parameter \( \alpha \) to depend on covariates

Clayton and Plackett copula

\[
\log(\alpha_i) = \gamma T Z_i
\]

Normal copula

\[
\frac{1}{2} \log \left( \frac{1+\alpha_i}{1-\alpha_i} \right) = \gamma T Z_i
\]
Estimation method

- \( S(t_1, t_2) = \tilde{C}_\alpha(S_1(t_1), S_2(t_2)) \) with \( \alpha \) the copula parameter
- Extension from Sun et al. (2006) who estimated association with Clayton copula
- Extension to allow covariates in the marginal distribution by fitting an AFT model with a flexible error distribution (Komárek et al. (2005))
Estimation method

- \( S(t_1, t_2) = \tilde{C}_\alpha(S_1(t_1), S_2(t_2)) \) with \( \alpha \) the copula parameter

- Extension from Sun et al. (2006) who estimated association with Clayton copula

- Extension to allow covariates in the marginal distribution by fitting an AFT model with a flexible error distribution (Komárek et al. (2005))

- Extension by allowing the copula parameter \( \alpha \) to depend on covariates
  - Clayton and Plackett copula
    \[
    \log(\alpha_i) = \gamma^T Z_i
    \]
  - Normal copula
    \[
    \frac{1}{2} \log \frac{1+\alpha_i}{1-\alpha_i} = \gamma^T Z_i
    \]
Marginal distributions

- \( \log(T_i) = \beta x + \gamma \varepsilon_i \)
- \( \varepsilon_i \sim \sum c_j \Phi(\cdot | \mu_j, \sigma) \)
Marginal distributions

- \( \log(T_i) = \beta x + \gamma \varepsilon_i \)
- \( \varepsilon_i \sim \sum c_j \Phi(\cdot \mid \mu_j, \sigma) \)
- penalty
Marginal distributions (cont’d)
Two stage estimation method

- Two stage estimation method (Shih & Louis (1995))
- Full log-likelihood $\log L(\alpha, S_1, S_2)$
Two stage estimation method

- Two stage estimation method (Shih & Louis (1995))
- Full log-likelihood log $L(\alpha, S_1, S_2)$
- First estimate marginal distributions $S_1$ and $S_2 \Rightarrow \hat{S}_1$ and $\hat{S}_2$
Two stage estimation method

- Two stage estimation method (Shih & Louis (1995))
- Full log-likelihood $\log L(\alpha, S_1, S_2)$
- First estimate marginal distributions $S_1$ and $S_2$ $\Rightarrow \hat{S}_1$ and $\hat{S}_2$
- Plug-in estimates from the marginals in the likelihood $\log L(\alpha, S_1, S_2)$ and estimate copula parameters in $\log L(\alpha, \hat{S}_1, \hat{S}_2)$
Variance estimation

- Not easy
- Theory of inference functions (Godambe (1991))
  very complicated derivatives
  \[ \Rightarrow \quad \text{Bootstrap } M \text{ samples of size } n \text{ with replacement} \]
Simulation

Simulation set-up

- 3 different copulas: Clayton, normal and Plackett
- 2 different marginal distributions: log-normal and mixture of two log-normals
- 2 censoring schemes
  1. 10% left, 70% interval and 20% right censoring
  2. 10% left, 50% interval and 40% right censoring
- 2 sample sizes: 100 and 500
- Dependence on covariates for the copula parameter
  1. no
  2. categorical covariate
  3. continuous covariate
- 2 settings: low ($\tau = 0.15$) and high ($\tau = 0.60$) association
Simulation (cont’d)

Simulation results
- For all settings small bias for the corresponding association measure
- No clear effect of censoring scheme, marginal distribution or setting
- Variance estimation works fine in the settings tested
Dental application

“Does the association between the emergence times of the contralateral teeth 14 and 24 differ for boys and girls?”
Dental application

“Does the association between the emergence times of the contralateral teeth 14 and 24 differ for boys and girls?”

- Marginally fit AFT model with gender as a covariate and a flexible error term distribution
- Fitted the Clayton, normal and Plackett copula
Best fitting copula

- Choose best fitting copula
- AIC

Plackett copula was the best fitting copula

Significant difference in association between boys ($\rho_S = 0.81$) and girls ($\rho_S = 0.86$)

No guarantee that best fitting copula fits well
Best fitting copula

- Choose best fitting copula
- AIC
- Plackett copula was the best fitting copula
- Significant difference in association between boys ($\rho_S = 0.81$) and girls ($\rho_S = 0.86$)
Best fitting copula

- Choose best fitting copula
- AIC
- Plackett copula was the best fitting copula
- Significant difference in association between boys ($\rho_S = 0.81$) and girls ($\rho_S = 0.86$)
- No guarantee that best fitting copula fits well
Goodness-of-fit

- Few research done, especially for interval censored data
- In the absence of covariates, compare local cross ratio function between copula approach and PNM approach
- A better fitting copula exists
Penalised normal mixture

- On the log-scale: fixed fine grid

Bogaerts and Lesaffre, proc. JSM, 2003
Penalised normal mixture

- On the log-scale: fixed fine grid
- At each grid point, an uncorrelated bivariate normal density

Bogaerts and Lesaffre, proc. JSM, 2003
Penalised normal mixture

- On the log-scale: fixed fine grid
- At each grid point, an uncorrelated bivariate normal density
- Weighted sum of normal distributions

Bogaerts and Lesaffre, proc. JSM, 2003
Penalised normal mixture

- On the log-scale: fixed fine grid
- At each grid point, an uncorrelated bivariate normal density
- Weighted sum of normal distributions
- Penalty

Bogaerts and Lesaffre, proc. JSM, 2003
Penalised normal mixture

- On the log-scale: fixed fine grid
- At each grid point, an uncorrelated bivariate normal density
- Weighted sum of normal distributions
- Penalty

Bogaerts and Lesaffre, proc. JSM, 2003
Penalised normal mixture

- On the log-scale: fixed fine grid
- At each grid point, an uncorrelated bivariate normal density
- Weighted sum of normal distributions
- Penalty

Bogaerts and Lesaffre, proc. JSM, 2003
Plug-in approach

- In the population version of the local or global association measure, replace the true density, CDF, survival function, etc. by their estimated counterpart derived from the PNM estimate.
Plug-in approach

- In the population version of the local or global association measure, replace the true density, CDF, survival function, etc. by their estimated counterpart derived from the PNM estimate.
- For instance

\[
\hat{\tau} = 4 \cdot \sum_i \sum_j \sum_k \sum_l \hat{c}_{ij} \hat{c}_{kl} \Phi \left( \frac{\mu_i - \mu_k}{\sqrt{2\sigma_1}} \right) \Phi \left( \frac{\nu_j - \nu_l}{\sqrt{2\sigma_2}} \right) - 1
\]
Dental application 2

“Does this association also depend on oral health behaviour?”

- Use dmft-index at age 7 as a marker of oral health behaviour during the first seven years
- dmft-index = sum of all decayed, missing due to caries or filled deciduous teeth
Dental application 2 (cont’d)

![Graph showing the Spearman's rho for boys and girls against the dmft-index](image)

- **Boys**
- **Girls**
If censoring scheme is reasonable and correct copula is chosen, method works fine.
Conclusions

- If censoring scheme is reasonable and correct copula is chosen, method works fine.
- Further research needed for goodness-of-fit tools