

Pseudo-observations for bivariate survival data

Yael Travis-Lumer¹ Micha Mandel¹ Rebecca Betensky²

¹Statistics and Data Science, The Hebrew University of Jerusalem, Israel

²Department of Biostatistics, New York University, New York

ISDSA

May 30, 2024

*Supported by the Drs. Eva & Shelby Kashket Memorial Fellowship, and
by the ISRAEL SCIENCE FOUNDATION (grant No. 1147/20)

Outline

- 1 Introduction
 - Bivariate survival data
 - The univariate pseudo-observations approach
- 2 Bivariate pseudo-observations
 - Proposed approach
 - Theoretical results
 - Simulation study
 - Data analysis
- 3 Summary

Outline

- 1 Introduction
 - Bivariate survival data
 - The univariate pseudo-observations approach
- 2 Bivariate pseudo-observations
 - Proposed approach
 - Theoretical results
 - Simulation study
 - Data analysis
- 3 Summary

Bivariate survival data - motivating example

- Consider time to blindness in diabetic patients (2 eyes)

Bivariate survival data - motivating example

- Consider time to blindness in diabetic patients (2 eyes)
- For each eye, time to blindness is either observed or censored

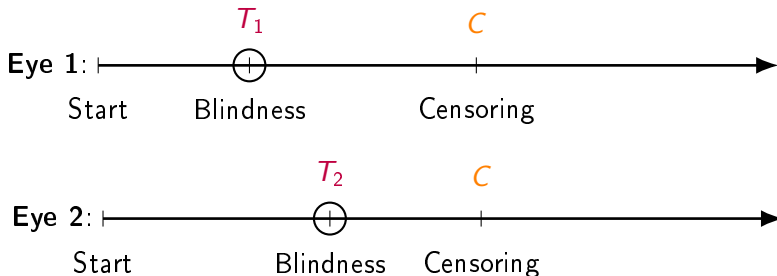
Bivariate survival data - motivating example

- Consider time to blindness in diabetic patients (2 eyes)
- For each eye, time to blindness is either observed or censored
- **Four** possible scenarios:

Bivariate survival data - motivating example

- Consider time to blindness in diabetic patients (2 eyes)
- For each eye, time to blindness is either observed or censored
- **Four** possible scenarios:

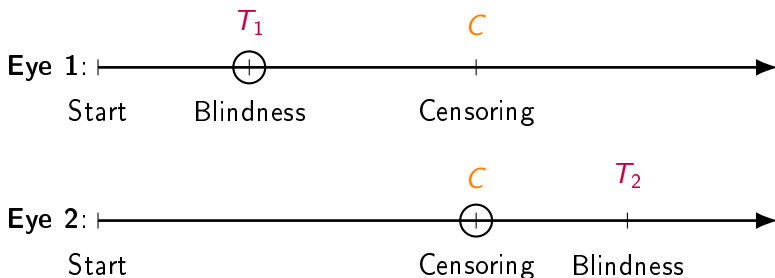
A. Double event (1,1)



Bivariate survival data - motivating example

- Consider time to blindness in diabetic patients (2 eyes)
- For each eye, time to blindness is either observed or censored
- **Four** possible scenarios:

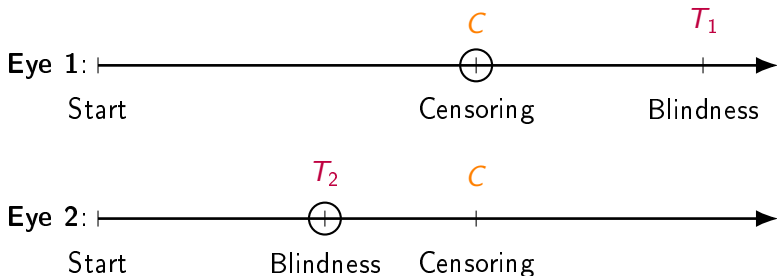
B. Single event (1,0)



Bivariate survival data - motivating example

- Consider time to blindness in diabetic patients (2 eyes)
- For each eye, time to blindness is either observed or censored
- **Four** possible scenarios:

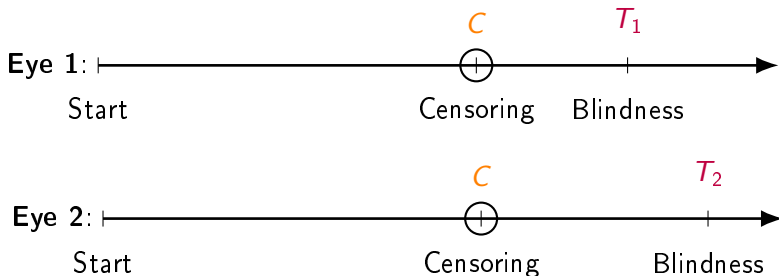
C. Single event (0,1)



Bivariate survival data - motivating example

- Consider time to blindness in diabetic patients (2 eyes)
- For each eye, time to blindness is either observed or censored
- **Four** possible scenarios:

D. No event (0,0)

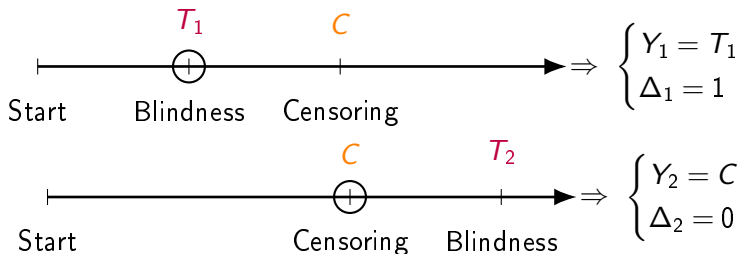


Bivariate right censored data - definition

- (T_1, T_2) are the bivariate survival times
- (C_1, C_2) are the bivariate censoring times
- $Y_1 = \min(T_1, C_1)$ and $Y_2 = \min(T_2, C_2)$ are the observed times
- $\Delta_1 = I(T_1 \leq C_1)$ and $\Delta_2 = I(T_2 \leq C_2)$ are the corresponding indicators

Bivariate right censored data - definition

- (T_1, T_2) are the bivariate survival times
- (C_1, C_2) are the bivariate censoring times
- $Y_1 = \min(T_1, C_1)$ and $Y_2 = \min(T_2, C_2)$ are the observed times
- $\Delta_1 = I(T_1 \leq C_1)$ and $\Delta_2 = I(T_2 \leq C_2)$ are the corresponding indicators



Bivariate right censored data - definition

- (T_1, T_2) are the bivariate survival times
- (C_1, C_2) are the bivariate censoring times
- $Y_1 = \min(T_1, C_1)$ and $Y_2 = \min(T_2, C_2)$ are the observed times
- $\Delta_1 = I(T_1 \leq C_1)$ and $\Delta_2 = I(T_2 \leq C_2)$ are the corresponding indicators

Observed data

We observe n i.i.d. copies $\{(Y_{1i}, \Delta_{1i}, Y_{2i}, \Delta_{2i}, Z_i)\}$,
 $Z_i \in \mathbb{R}^p$ are the covariates.

Notation and goal

- $S_{T_1, T_2}(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ is the bivariate survival function
- $S_{T_1, T_2}(t_1, t_2 | Z) = P(T_1 > t_1, T_2 > t_2 | Z)$ is the covariate adjusted bivariate survival

Notation and goal

- $S_{T_1, T_2}(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ is the bivariate survival function
- $S_{T_1, T_2}(t_1, t_2 | Z) = P(T_1 > t_1, T_2 > t_2 | Z)$ is the covariate adjusted bivariate survival

Goal

Estimate $S_{T_1, T_2}(t_1, t_2 | Z)$ using pseudo-observations

Outline

- 1 Introduction
 - Bivariate survival data
 - The univariate pseudo-observations approach
- 2 Bivariate pseudo-observations
 - Proposed approach
 - Theoretical results
 - Simulation study
 - Data analysis
- 3 Summary

Univariate pseudo-observations - background

- Let T be a univariate survival time (subject to censoring)

Univariate pseudo-observations - background

- Let T be a univariate survival time (subject to censoring)
- Let $\theta = E[f(T)]$

Univariate pseudo-observations - background

- Let T be a univariate survival time (subject to censoring)
- Let $\theta = E[f(T)]$
- Consider a Generalized Linear Model (GLM):

$$E[f(T) | Z] = g^{-1}(\beta^T Z)$$

Univariate pseudo-observations - background

- Let T be a univariate survival time (subject to censoring)
- Let $\theta = E[f(T)]$
- Consider a Generalized Linear Model (GLM):

$$E[f(T) | Z] = g^{-1}(\beta^T Z)$$

- Note that $\theta = E[E[f(T) | Z]]$

Univariate pseudo-observations - background

- Let T be a univariate survival time (subject to censoring)
- Let $\theta = E[f(T)]$
- Consider a Generalized Linear Model (GLM):

$$E[f(T) | Z] = g^{-1}(\beta^T Z)$$

- Note that $\theta = E[E[f(T) | Z]]$

Goal

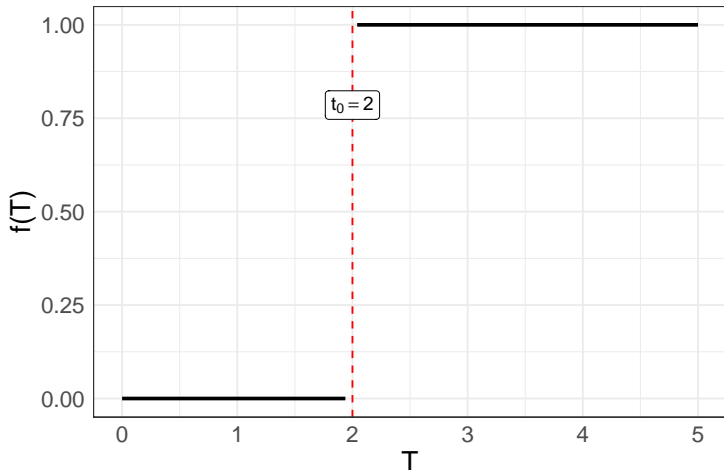
Estimate β in $E[f(T) | Z] = g^{-1}(\beta^T Z)$

Example 1 - survival at t_0

- Let $f(T) = 1(T > t_0)$ be an indicator function

Example 1 - survival at t_0

- Let $f(T) = 1(T > t_0)$ be an indicator function



Example 1 - survival at t_0

- Let $f(T) = 1(T > t_0)$ be an indicator function
- $\theta = E[f(T)] = E[1(T > t_0)] = P(T > t_0) = S(t_0)$ is the survival probability at t_0

Example 1 - survival at t_0

- Let $f(T) = 1(T > t_0)$ be an indicator function
- $\theta = E[f(T)] = E[1(T > t_0)] = P(T > t_0) = S(t_0)$ is the survival probability at t_0

Assumption

$$E[f(T) | Z] = P(T > t_0 | Z) = S(t_0 | Z) = g^{-1}(\beta^T Z),$$

where for example

Example 1 - survival at t_0

- Let $f(T) = 1(T > t_0)$ be an indicator function
- $\theta = E[f(T)] = E[1(T > t_0)] = P(T > t_0) = S(t_0)$ is the survival probability at t_0

Assumption

$$E[f(T) | Z] = P(T > t_0 | Z) = S(t_0 | Z) = g^{-1}(\beta^T Z),$$

where for example

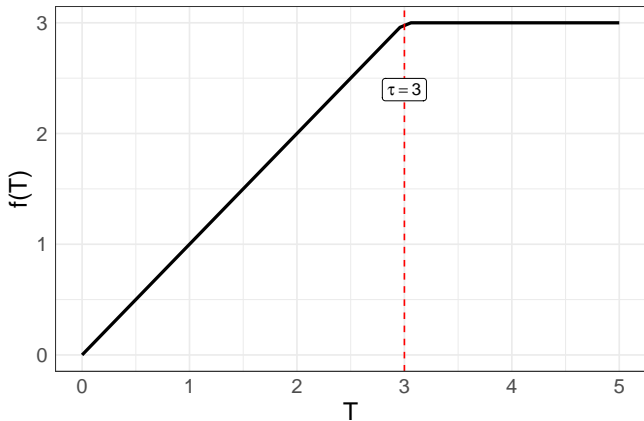
- 1 $g(x) = \log\left(\frac{x}{1-x}\right)$ is the **logit** link
- 2 $g(x) = \Phi^{-1}(x)$ is the **probit** link
- 3 $g(x) = \log(-\log(x))$ is the **cloglog** link

Example 2 - restricted mean

- Let $f(T) = \min(T, \tau)$, for some $\tau > 0$

Example 2 - restricted mean

- Let $f(T) = \min(T, \tau)$, for some $\tau > 0$



Example 2 - restricted mean

- Let $f(T) = \min(T, \tau)$, for some $\tau > 0$
- $\theta = E[f(T)] = E[\min(T, \tau)]$ is the τ -restricted mean

Example 2 - restricted mean

- Let $f(T) = \min(T, \tau)$, for some $\tau > 0$
- $\theta = E[f(T)] = E[\min(T, \tau)]$ is the τ -restricted mean

Assumption

$E[f(T) | Z] = E[\min(T, \tau) | Z] = g^{-1}(\beta^T Z)$, where for example

Example 2 - restricted mean

- Let $f(T) = \min(T, \tau)$, for some $\tau > 0$
- $\theta = E[f(T)] = E[\min(T, \tau)]$ is the τ -restricted mean

Assumption

$E[f(T) | Z] = E[\min(T, \tau) | Z] = g^{-1}(\beta^T Z)$, where for example

- 1 $g(x) = x$ is the **identity** link
- 2 $g(x) = \log(x)$ is the **log** link

Pseudo-observations - motivation

Goal

Estimate β in $E[f(T) | Z] = g^{-1}(\beta^T Z)$

Pseudo-observations - motivation

Goal

Estimate β in $E[f(T) | Z] = g^{-1}(\beta^T Z)$

Problem - T is not fully observed

How can we fit such GLMs to censored data?

Pseudo-observations - motivation

Goal

Estimate β in $E[f(T) | Z] = g^{-1}(\beta^T Z)$

Problem - T is not fully observed

How can we fit such GLMs to censored data?

Solution (Anderson, 2003)

Replace $f(T)$ with pseudo-observations.

Pseudo-observations - definition

Let $\hat{\theta}$ be a consistent estimator of $\theta = E[f(T)]$

Pseudo-observations - definition

Let $\hat{\theta}$ be a consistent estimator of $\theta = E[f(T)]$ (e.g. $\hat{\theta} = \hat{S}_{KM}(t_0)$).

Pseudo-observations - definition

Let $\hat{\theta}$ be a consistent estimator of $\theta = E[f(T)]$ (e.g. $\hat{\theta} = \hat{S}_{KM}(t_0)$).

Definition

The pseudo-observation for $f(T_i)$ is

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$$

* $\hat{\theta}^{-i}$ is the estimator applied to the sample of size $n-1$ not containing subject i .

Pseudo-observations - cont'd

The regression model

$$g(E[f(T)|Z]) = \beta^T Z$$

The pseudo-observations

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$$

Pseudo-observations - cont'd

The regression model

$$g(E[f(T)|Z]) = \beta^T Z$$

The pseudo-observations

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$$

Model fitting

Use $\hat{\theta}_i$ as the response in the regression model $g(\hat{\theta}_i) = \beta^T Z_i$.
Estimate β by solving the [estimating equations](#).

Pseudo-observations - cont'd

The regression model

$$g(E[f(T)|Z]) = \beta^T Z$$

The pseudo-observations

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$$

Model fitting

Use $\hat{\theta}_i$ as the response in the regression model $g(\hat{\theta}_i) = \beta^T Z_i$.
Estimate β by solving the [estimating equations](#).

Theorem (Overgaard et al., 2017)

Independent censoring $\implies \hat{\beta}$ is consistent and asymptotically normal.

Outline

- 1 Introduction
 - Bivariate survival data
 - The univariate pseudo-observations approach
- 2 Bivariate pseudo-observations
 - Proposed approach
 - Theoretical results
 - Simulation study
 - Data analysis
- 3 Summary

Pseudo-observations for bivariate survival data

- T_1 and T_2 are the bivariate failure times
- The censoring times (C_1, C_2) are independent of both (T_1, T_2) and Z
- $\theta = E[f(T_1, T_2)]$

Pseudo-observations for bivariate survival data

- T_1 and T_2 are the bivariate failure times
- The censoring times (C_1, C_2) are independent of both (T_1, T_2) and Z
- $\theta = E[f(T_1, T_2)]$

Examples

- 1 $f(T_1, T_2) = I(T_1 > t_1^0, T_2 > t_2^0) \implies \theta = S_{T_1, T_2}(t_1^0, t_2^0)$
- 2 $f(T_1, T_2) = \min(T_1, T_2) \implies \theta = E[\min(T_1, T_2)]$
- 3 $f(T_1, T_2) = \min(T_1, \tau) \implies \theta = E[\min(T_1, \tau)]$

Bivariate pseudo-observations - cont'd

- Assume the regression model $E[f(T_1, T_2) | Z] = g^{-1}(\beta^T Z)$
- Let $\hat{\theta}$ be a consistent estimator of $\theta = E[f(T_1, T_2)]$
- The **pseudo-observation** for $f(T_{1i}, T_{2i})$ is

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$$

- Use $\hat{\theta}_i$ as the response in the regression model $\hat{\theta}_i = g^{-1}(\beta^T Z)$

Main example

- Consider $f(T_1, T_2) = I(T_1 > t_1^0, T_2 > t_2^0)$ which corresponds to $\theta = S_{T_1, T_2}(t_1^0, t_2^0)$

Main example

- Consider $f(T_1, T_2) = I(T_1 > t_1^0, T_2 > t_2^0)$ which corresponds to $\theta = S_{T_1, T_2}(t_1^0, t_2^0)$
- Model $E[f(T_1, T_2) | Z] = S_{T_1, T_2}(t_1^0, t_2^0 | Z)$ by $g^{-1}(\beta^T Z)$, where g is the logit link function $g(x) = \log\left(\frac{x}{1-x}\right)$

Main example

- Consider $f(T_1, T_2) = I(T_1 > t_1^0, T_2 > t_2^0)$ which corresponds to $\theta = S_{T_1, T_2}(t_1^0, t_2^0)$
- Model $E[f(T_1, T_2) | Z] = S_{T_1, T_2}(t_1^0, t_2^0 | Z)$ by $g^{-1}(\beta^T Z)$, where g is the logit link function $g(x) = \log\left(\frac{x}{1-x}\right)$
- Let $\hat{\theta}$ be a consistent estimator of $S_{T_1, T_2}(t_1^0, t_2^0)$

Main example

- Consider $f(T_1, T_2) = I(T_1 > t_1^0, T_2 > t_2^0)$ which corresponds to $\theta = S_{T_1, T_2}(t_1^0, t_2^0)$
- Model $E[f(T_1, T_2) | Z] = S_{T_1, T_2}(t_1^0, t_2^0 | Z)$ by $g^{-1}(\beta^T Z)$, where g is the logit link function $g(x) = \log\left(\frac{x}{1-x}\right)$
- Let $\hat{\theta}$ be a consistent estimator of $S_{T_1, T_2}(t_1^0, t_2^0)$
- Calculate the pseudo-observations $\{\hat{\theta}_i\}_{i=1}^n$ and use them as the response in the regression model

Main example

- Consider $f(T_1, T_2) = I(T_1 > t_1^0, T_2 > t_2^0)$ which corresponds to $\theta = S_{T_1, T_2}(t_1^0, t_2^0)$
- Model $E[f(T_1, T_2) | Z] = S_{T_1, T_2}(t_1^0, t_2^0 | Z)$ by $g^{-1}(\beta^T Z)$, where g is the logit link function $g(x) = \log\left(\frac{x}{1-x}\right)$
- Let $\hat{\theta}$ be a consistent estimator of $S_{T_1, T_2}(t_1^0, t_2^0)$
- Calculate the pseudo-observations $\{\hat{\theta}_i\}_{i=1}^n$ and use them as the response in the regression model

Question

What estimators of $S_{T_1, T_2}(t_1^0, t_2^0)$ should we use?

Nonparametric estimators of $S_{T_1, T_2}(t_1, t_2)$

We consider two nonparametric estimators of the joint survival function:

- 1 The simple estimator of Lin and Ying (1993)
- 2 The well-known Dabrowska estimator (1988)

Nonparametric estimators of $S_{T_1, T_2}(t_1, t_2)$

We consider two nonparametric estimators of the joint survival function:

- 1 The simple estimator of Lin and Ying (1993)
- 2 The well-known Dabrowska estimator (1988)

Notes

- The estimator of Lin and Ying assumes a univariate censoring variable and is an IPCW estimator
- The Dabrowska estimator is more general but more complicated
- Both estimators are consistent

Outline

- 1 Introduction
 - Bivariate survival data
 - The univariate pseudo-observations approach
- 2 **Bivariate pseudo-observations**
 - Proposed approach
 - **Theoretical results**
 - Simulation study
 - Data analysis
- 3 Summary

Theoretical results

- Consider our main example

$$S_{T_1, T_2}(t_1^0, t_2^0 | Z) = \frac{\exp(\beta^T Z)}{1 + \exp(\beta^T Z)},$$

where we use the bivariate pseudo-observations that are based on the Lin and Ying estimator.

Theoretical results

- Consider our main example

$$S_{T_1, T_2}(t_1^0, t_2^0 | Z) = \frac{\exp(\beta^T Z)}{1 + \exp(\beta^T Z)},$$

where we use the bivariate pseudo-observations that are based on the Lin and Ying estimator.

Result

Based on Overgaard et al. (2017), we prove that $\hat{\beta}$ is **consistent** and **asymptotically normal** $\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} \mathcal{N}(0, M^{-1}\Sigma(M^{-1})^T)$.

Outline

- 1 Introduction
 - Bivariate survival data
 - The univariate pseudo-observations approach
- 2 **Bivariate pseudo-observations**
 - Proposed approach
 - Theoretical results
 - **Simulation study**
 - Data analysis
- 3 Summary

- Goal: estimate $S_{T_1, T_2}(t_1, t_2 | Z)$

- **Goal:** estimate $S_{T_1, T_2}(t_1, t_2 | Z)$
- We assume the logistic regression model

$$S_{T_1, T_2}(t_1, t_2 | Z) = \frac{\exp(\beta^T Z)}{1 + \exp(\beta^T Z)}$$

- **Goal:** estimate $S_{T_1, T_2}(t_1, t_2 | Z)$
- We assume the logistic regression model

$$S_{T_1, T_2}(t_1, t_2 | Z) = \frac{\exp(\beta^T Z)}{1 + \exp(\beta^T Z)}$$

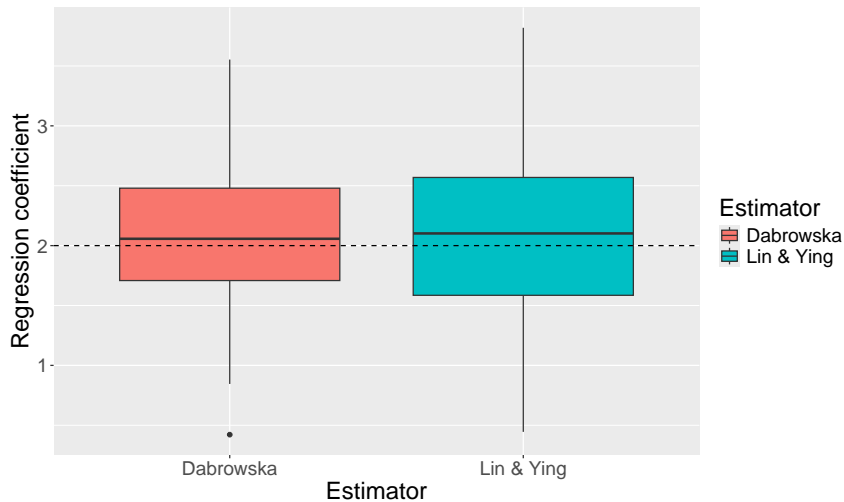
- We simulate $Z \sim U(0.5, 1.5)$, and bivariate failure times generated from the assumed logistic model

- **Goal:** estimate $S_{T_1, T_2}(t_1, t_2 | Z)$
- We assume the logistic regression model

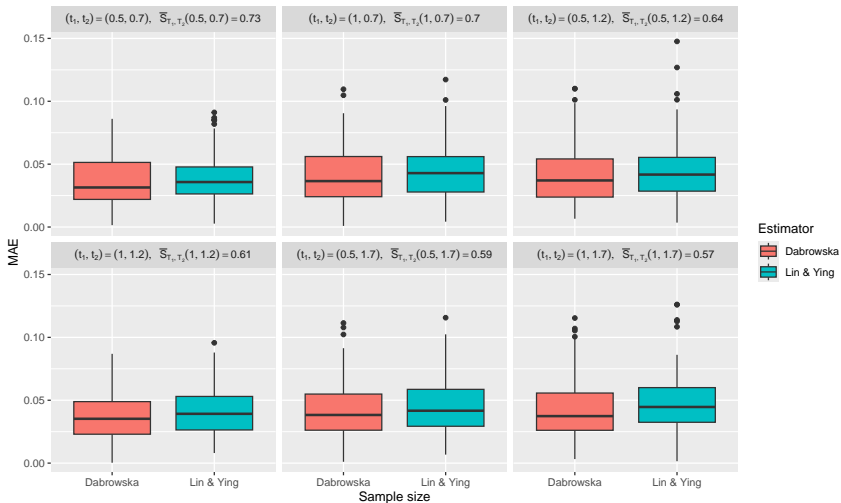
$$S_{T_1, T_2}(t_1, t_2 | Z) = \frac{\exp(\beta^T Z)}{1 + \exp(\beta^T Z)}$$

- We simulate $Z \sim U(0.5, 1.5)$, and bivariate failure times generated from the assumed logistic model
- We used a univariate censoring variable, $K = 6$ simultaneous time points, $n = 200$, and 100 simulations

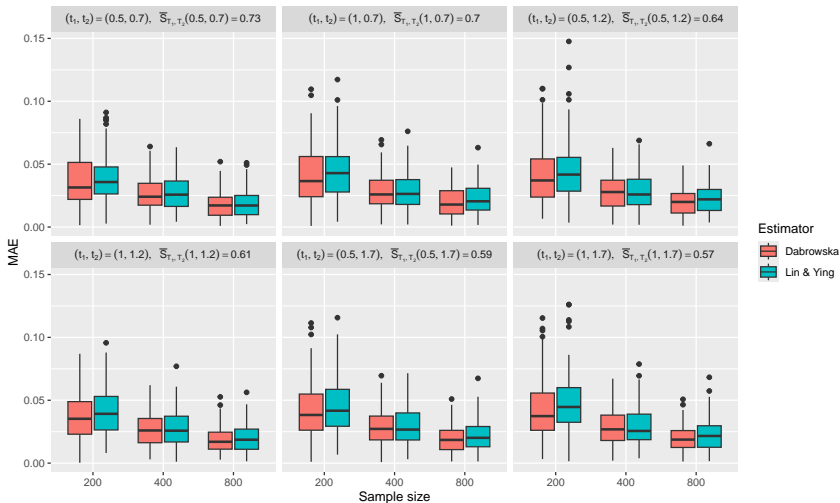
Bivariate logistic data - $\hat{\beta}_1$



Bivariate logistic data - mean absolute error (MAE)



Bivariate logistic data - analysis by sample size



Outline

- 1 Introduction
 - Bivariate survival data
 - The univariate pseudo-observations approach
- 2 Bivariate pseudo-observations
 - Proposed approach
 - Theoretical results
 - Simulation study
 - Data analysis
- 3 Summary

Diabetic retinopathy study

- 197 patients with diabetic retinopathy
- one eye of each patient is randomly assigned to laser treatment (T_1)
- time to blindness is measured from initiation of treatment
- covariates include age at diagnosis of diabetes and risk score (between 6-12)

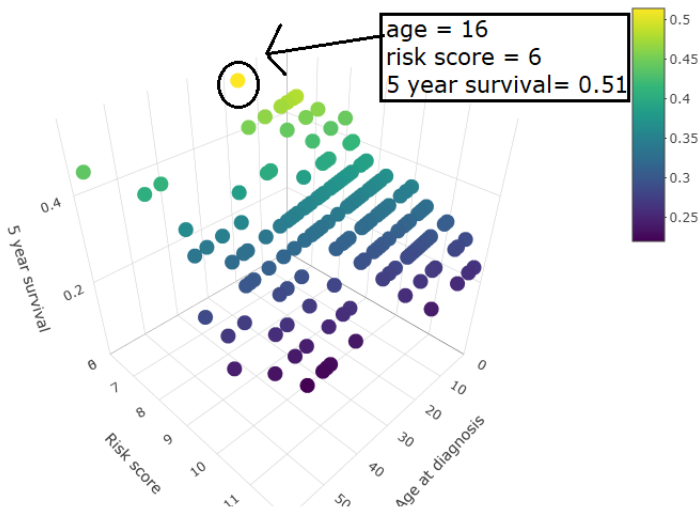
Diabetic retinopathy study

- 197 patients with diabetic retinopathy
- one eye of each patient is randomly assigned to laser treatment (T_1)
- time to blindness is measured from initiation of treatment
- covariates include age at diagnosis of diabetes and risk score (between 6-12)
- We estimate $S_{T_1, T_2}(5, 5 | Z)$ using the bivariate pseudo-observations approach based on the Dabrowska estimator

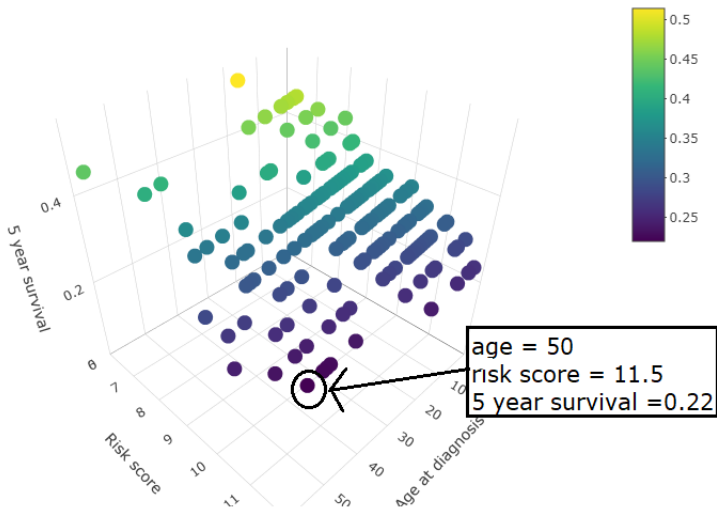
5 year survival of both eyes - $\hat{S}_{T_1, T_2}(5, 5 | Z)$



5 year survival of both eyes - $\hat{S}_{T_1, T_2}(5, 5 | Z)$



5 year survival of both eyes - $\hat{S}_{T_1, T_2}(5, 5 | Z)$ (Dabrowska)



Summary

Summary

- **Approach:** we generalized the **pseudo-observations approach** to **bivariate survival data**.
- **Theory:** For the Lin and Ying estimator, we proved that $\hat{\beta}$ is **consistent** and **asymptotically normal**.
- **Simulations:** The approach manages to **correctly estimate** β and $S_{T_1, T_2}(t_1^0, t_2^0 | Z)$.
- **General framework:** Our approach can be applied to **general quantities**, providing that a consistent estimator exists.
- **Generalizations:** other regression models, multivariate parameter θ , multivariate failure times (T_1, T_2, \dots, T_d)

Thank you!

Thank you!
Questions?

The estimating equations

The model

$$g(E[f(T_i) | Z_i]) = \beta^T Z_i$$

The estimating equations

The model

$$g(E[f(T_i) | Z_i]) = \beta^T Z_i$$

The estimating equations

$$\sum_{i=1}^n \left(\frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i) \right)^T (\hat{\theta}_i - g^{-1}(\beta^T Z_i)) = \sum_{i=1}^n U_i(\beta) = 0$$

The estimating equations

The model

$$g(E[f(T_i) | Z_i]) = \beta^T Z_i$$

The estimating equations

$$\sum_{i=1}^n \left(\frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i) \right)^T (\hat{\theta}_i - g^{-1}(\beta^T Z_i)) = \sum_{i=1}^n U_i(\beta) = 0$$

Estimation based on $K > 1$ time points

If the model is fitted with $K > 1$ time points, use generalized estimating equations with a working covariance matrix V

Back to [pseudo-observations](#)