Treatment Effects Estimation with Unmeasured Confounding Variables

Namhwa Lee, Shujie Ma (Advisor)

Department of Statistics, UC-Riverside

August 5, 2024

OUTLINE

1. Introduction
2. Running Example and Setup
3. Estimation
4. Numerical Results
    - Air pollution on mental health
5. Conclusion

Introduction

Background

- The goal of much observational research is to identify factors that have a causal effect on outcomes.
- Many observational studies extend beyond a single time point and frequently incorporate repeated measures.
  - Healthcare interventions $\rightarrow$ Progression of a chronic disease
  - Environmental exposures $\rightarrow$ Health problems
  - Training program $\rightarrow$ Better performance

Introduction

Binary Treatment:

- $D_{ij} = 1$ if i-th subject at time period j was received a treatment.
- $D_{ij} = 0$ otherwise.

Potential outcomes:

- $Y_{ij}(1)$: outcome if i-th subject at time period j is exposed.
- $Y_{ij}(0)$: outcome if i-th subject at time period j is not exposed.
- Consistency:
  $Y_{ij}(D_{ij}) = Y_{ij}$ if subject i at time j receives treatment D_{ij}.

Introduction

Treatment effects

- Treatment effects for *i*-th subject at time *j*: $Y_{ij}(1) - Y_{ij}(0)$
- Average treatment effects (ATE): $\mathbb{E}[Y_{ij}(1) - Y_{ij}(0)]$
- A possible estimator of ATE:

$$\mathbb{E}[Y_{ij}(1)|D_{ij} = 1, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i] - \mathbb{E}[Y_{ij}(0)|D_{ij} = 0, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i]$$

Introduction

Treatment effects

- Selection bias: In general, for $d = 0, 1$,

$$\mathbb{E}[Y_{ij}(d)] \neq \mathbb{E}[Y_{ij}(d) | D_{ij} = d, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i]$$

- Requires the assumption of **no unmeasured confounders** (Hirano and Imbens, 2001).
- **No unmeasured confounders** (Rosenbaum and Rubin, 1983)

$$Y_{ij}(0), Y_{ij}(1) \perp D_{ij} \big| \boldsymbol{X}_{ij}, \boldsymbol{Z}_i$$

Running example

Air pollution on mental health dataset

- CitieS-Health Barcelona Panel Study (Gignac et al., 2022).
- Comprises 3,333 observations from 286 distinct participants, collected in Barcelona, Spain.
- There are various variables:
  - Environmental variables, Meteorological variables, Self-reported survey data, Results from the Stroop test
- **Goal**: Identifying causal effects of *short-term exposure* to air pollution on mental health.

Running example

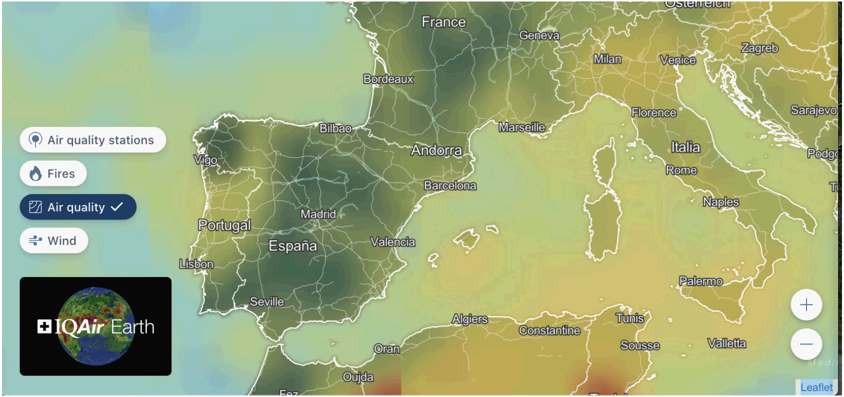

Figure 1: Air pollution map. Image: IQAir

Running Example

Running example

- Outcome (Y_{ij}): Mental health test score of i-th participant at time period j.
- Treatment (D_{ij}): Binary air pollution index when the i-th participant is conducting the j-th test.
- Confounding variable
  - Time-invariant ($\boldsymbol{Z}_i$): gender, education, smoking behavior.
  - Time-variant ($\boldsymbol{X}_{ij}$): stress level, temperature, humidity,

Running Example

Running example

- Highly likely that unmeasured confoundings exist.
- Integrated variable of unmeasured confoundings (U_i):
  - Size of nearby green and blue area
  - Urban transportation intensity
  - Individual characteristics

Research question

- How can we estimate the treatment effects with U_i?

Setup

Repeated measured data

- For each individual i, we have $\{Y_{ij}, D_{ij}, \boldsymbol{Z}_i, \boldsymbol{X}_{ij}, U_i\}$ where
  - Y_{ij}: observed outcome variable.
  - D_{ij}: binary treatment assignment variable.
  - $\boldsymbol{Z}_i \in \mathbb{R}^{p_1}$: time-invariant covariates.
  - $\boldsymbol{X}_{ij} \in \mathbb{R}^{p_2}$: time-varying covariates.
  - $U_i \in \mathbb{R}$: incorporated unmeasured confounding.

Setup

Assumptions

- To identify the treatment effects, we assume

(A1) Conditional ignorability: $Y_{ij}(0), Y_{ij}(1) \perp D_{ij} \big| \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, U_i$.

(A2) Positivity: $0 < P(D_{ij} = 1 | \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, U_i) < 1$.

(A3) Consistency: $Y_{ij} = D_{ij} Y_{ij}(1) + (1 - D_{ij}) Y_{ij}(0)$.

- Under assumptions $(A1) - (A3)$, ATE can be identified as

$$\mathbb{E}\left[Y_{ij}(1) - Y_{ij}(0)\right] = \mathbb{E}\Big[\mathbb{E}\left[Y_{ij} \big| \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i, D_{ij} = 1\right]\Big] - \mathbb{E}\Big[\mathbb{E}\left[Y_{ij} \big| \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i, D_{ij} = 0\right]\Big].$$

Setup

Model formulation

- For $i = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, n_i$:

$$\begin{aligned} E(Y_{ij} | D_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, U_i) &= \begin{pmatrix} 1 & D_{ij} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \end{pmatrix} \boldsymbol{X}_{ij}^* + \begin{pmatrix} 1 & D_{ij} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} U_i \qquad (1) \\ &= \boldsymbol{\beta}_1^\top \boldsymbol{X}_{ij}^* + D_{ij} \boldsymbol{\beta}_2^\top \boldsymbol{X}_{ij}^* + \alpha_1 U_i + D_{ij} \alpha_2 U_i \end{aligned}$$

  where

  $\boldsymbol{X}_{ij}^* = (1, \boldsymbol{Z}_i^\top, \boldsymbol{X}_{ij}^\top)^\top \in \mathbb{R}^{p_1+p_2+1}$,
  $\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1+p_2+1}$: intercept and the main effects.
  $\boldsymbol{\beta}_2 \in \mathbb{R}^{p_1+p_2+1}$: treatment, and interaction effects between D_{ij} and $(\boldsymbol{Z}_i, \boldsymbol{X}_{ij})$.

- This model can address the situation where both observed and unobserved confounders affect outcomes differently depending on treatment assignments.

Setup

Model formulation

- For identification purposes, we reparameterize (1) by defining $b_i := \alpha_1 U_i$, and $\omega := \alpha_2/\alpha_1$:

$$E(Y_{ij}|D_{ij}, b_i, \boldsymbol{Z}_i, \boldsymbol{X}_{ij}) = \boldsymbol{\beta}_1^\top \boldsymbol{X}_{ij}^* + D_{ij}\boldsymbol{\beta}_2^\top \boldsymbol{X}_{ij}^* + \left(1 + \omega D_{ij}\right) b_i, \quad (2)$$

- b_i: An incorporated effect of unmeasured confoundings.
- ω: The impact of b_i on the outcome could vary depending on D_{ij}.

Setup

Three-stage model

- **(Stage 1)** For individual unit i at j-th repeated measuerment:

$$Y_{ij} = \boldsymbol{\beta}_1^\top \boldsymbol{X}_{ij}^* + D_{ij}\boldsymbol{\beta}_2^\top \boldsymbol{X}_{ij}^* + \big(1 + \omega D_{ij}\big) b_i + \epsilon_{ij} \tag{3}$$

  where $\epsilon_{ij} \sim N(0, \sigma^2)$.

- **(Stage 2)** For the treatment assignment,

$$P(D_{ij} = 1 | \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, b_i) = \frac{\exp(\boldsymbol{\eta}^\top \boldsymbol{X}_{ij}^* + \xi b_i)}{1 + \exp(\boldsymbol{\eta}^\top \boldsymbol{X}_{ij}^* + \xi b_i)} \tag{4}$$

  where $\boldsymbol{\eta} \in \mathbb{R}^{p_1+p_2+1}, \xi \in \mathbb{R}$, and further assume $D_{ij} \perp D_{ij'} | \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, b_i$ for $j \neq j'$.

SETUP

Three-stage model

- **(Stage 3)** The integrated effect of unmeasured confoundings,

$$b_i \sim N(0, \sigma_b^2)$$

independently with each other, with ϵ_{ij}, and with observed covariates $(\boldsymbol{X}_{ij}, \boldsymbol{Z}_i)$.

SETUP

Treatment effects

- Under the three-stage model and $(A1)-(A3)$, treatment effects are defined as follows:

$$\mathbb{E}\left[Y_{ij}(1) - Y_{ij}(0) \middle| b_i, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i\right] = \boldsymbol{\beta}_2^\top \boldsymbol{X}_{ij}^* + \omega b_i, \tag{5}$$

$$\mathbb{E}\left[Y_{ij}(1) - Y_{ij}(0) \middle| \boldsymbol{X}_{ij}, \boldsymbol{Z}_i\right] = \boldsymbol{\beta}_2^\top \boldsymbol{X}_{ij}^*, \tag{6}$$

$$\mathbb{E}\left[Y_{ij}(1) - Y_{ij}(0)\right] = \boldsymbol{\beta}_2^\top \mathbb{E}\left[\boldsymbol{X}_{ij}^*\right]. \tag{7}$$

Estimation

Incomplete-data problem

- For each individual across the repeated measurements, we have

$$\{y_{ij}, d_{ij}, \boldsymbol{z}_i, \boldsymbol{x}_{ij}, b_i\}.$$

- Due to the unobservable nature of b_i, the observed dataset is

$$\{y_{ij}, d_{ij}, \boldsymbol{z}_i, \boldsymbol{x}_{ij}\}.$$

- Incomplete-data problem involving latent or missing variable.
- Need to estimate $\boldsymbol{\theta} := (\boldsymbol{\beta}, \sigma, \omega, \boldsymbol{\eta}, \xi, \sigma_b)$ with incomplete data.

Estimation

EM algorithm

- A widely used tool for incomplete data problems (Dempster et al., 1977).
  - Linear mixed effect models (Laird and Ware, 1982)
  - Mixture models (McLachlan et al., 2004)
- We will treat b_i as latent.
- Advantage: We can estimate b_i through $\mathbb{E}(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$.

Estimation

E-step

- A conditional expectation of the complete data likelihood given the observed data and the current estimate of the parameters is determined (Wu, 1983).
- With our setup,

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right) := \mathbb{E}\left[\ell_c(\boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{d}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}^{(k)}\right]. \tag{8}$$

where $\boldsymbol{\theta}^{(k)}$ be the current estimate of the parameters, and the complete-data log likelihood,

$$\ell_c(\boldsymbol{\theta}) := \sum_{i=1}^{m} \log f(\boldsymbol{y}_i, \boldsymbol{d}_i, b_i, \boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}). \tag{9}$$

ESTIMATION

M-step

- Find updates of each parameter estimates that maximize (8).

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q^*\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \omega, \boldsymbol{\eta}, \xi, \sigma_b)$.

ESTIMATION

Approximation method

- Challenge when employing the EM algorithm:
  - Unable to obtain closed-form expressions for $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$.
- How to address this challenge?
  - Compute an approximation of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ using Laplace's method.

Estimation

Laplace's approximation method with an extra positive factor

Suppose that $g : \mathbb{R}^p \to \mathbb{R}$ is a smooth function and scalar function of b_i with a unique minimum at $\tilde{b}_i$. If there is an additional factor $h(\cdot)$ which is a smooth and positively valued function, we can employ an alternative Laplace's approximation:

$$\int h(b_i)\exp\{-Ng(b_i)\}db_i \approx \left(\frac{2\pi}{N}\right)^{p/2} \frac{h(\tilde{b}_i)e^{-Ng(\tilde{b}_i)}}{|H(g)(\tilde{b}_i)|^{1/2}} \tag{10}$$

where $H(g)(\tilde{b}_i)$ is a Hessian matrix of g evaluated at $\tilde{b}_i$ (Butler, 2007).

Estimation

Laplacian-Variant EM algorithm

- Non-linear model for the treatment assignment
  $\rightarrow$ No explicit expression.
- Existing EM algorithm cannot be applied.
- Laplacian-Variant EM algorithm iterations:

  (Step 0) Find a set of initial values, $\boldsymbol{\theta}^{(0)}$.
  (Step 1) Derive $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$.
  (Step 2) Approximate $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \approx Q^*(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$.
  (Step 3) Find $\boldsymbol{\theta}^{(k+1)}$ that maximizes $Q^*(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$.

- Repeat (Step 2) - (Step 3) until the algorithm converges.

Numerical Results

Air pollution on mental health dataset

- The average treatment effect (ATE) indicates that good PM2.5 levels causes a modest improvement in cognitive performance.
- This suggests that short-term exposure to high PM2.5 level has a slight negative impact on mental cognitive health.

Table 1: Estimates of Average Treatment Effect

Estimate	Mean	95% C.I.	90% C.I.	P-value
0.043	0.043 (0.024)	(-0.002, 0.091)	(0.004, 0.082)	0.078

Numerical Results

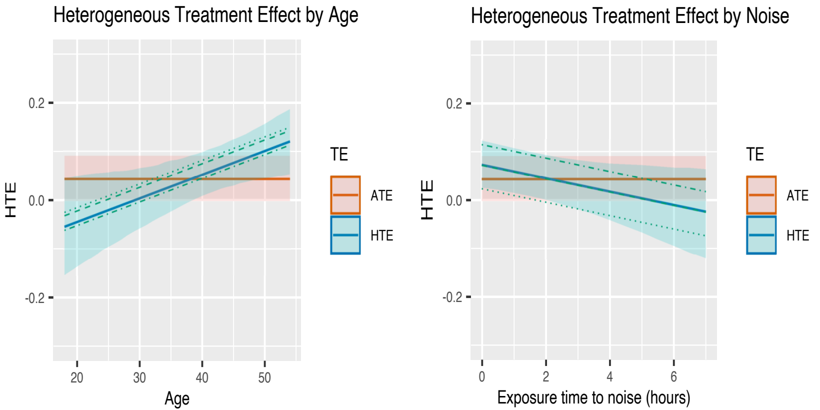

Figure 2: HTE Plot by Age and Exposure time to noise

Conclusion

To sum up

- We propose a method that alleviates the no unmeasured confounders assumption by incorporating the integrated effect of unmeasured confoundings (b_i).
- We propose Laplacian-Variant EM algorithm as a key estimation tool, treating b_i as latent.
- We can jointly estimate parameters in the three-stage model including nonlinear relationships between treatment assignments and observed and unobserved confounders.

Conclusion

Discussion

- Can incorporate univariate b_i in the model due to the identifiability issue.
- Aimed to identify the causal effect of short-term exposure to air pollution on mental health.
- However, we had to use the annual average (long-term) standard as a criterion

References I

Azevedo-Filho, A. and Shachter, R. D. (1994). Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In *Uncertainty proceedings 1994*, pages 28–36. Elsevier.

Butler, R. W. (2007). *Saddlepoint approximations with applications*, volume 22. Cambridge University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Gignac, F., Righi, V., Toran, R., Paz Errandonea, L., Ortiz, R., Mijling, B., Naranjo, A., Nieuwenhuijsen, M., Creus, J., and Basagaña, X. (2022). Cities-health barcelona panel study results.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2:259–278.

References II

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

McLachlan, G. J., Krishnan, T., and Ng, S. K. (2004). The em algorithm. Technical report, Papers.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.

Stroop Test

Stroop test

- A psychological assessment tool used to gauge cognitive processing speed and selective attention.
- Participants are presented with a list of color names written in different colors.
- Task is to name the ink color of each word as quickly and accurately as possible, regardless of the actual word.
- Assessing test performance involves analyzing the participants' response times and accuracy under different conditions.

Appendix: TE

Assumptions

- Under assumptions $(A1)-(A3)$, ATE can be identified as

$$
\begin{aligned}
\tau &= \mathbb{E}\Big[\mathbb{E}\big[Y_{ij}(1) - Y_{ij}(0)\big|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[Y_{ij}(1)\big|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i\big]\Big] - \mathbb{E}\Big[\mathbb{E}\big[Y_{ij}(0)\big|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[Y_{ij}(1)\big|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i, D_{ij} = 1\big]\Big] - \mathbb{E}\Big[\mathbb{E}\big[Y_{ij}(0)\big|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i, D_{ij} = 0\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[Y_{ij}\big|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i, D_{ij} = 1\big]\Big] - \mathbb{E}\Big[\mathbb{E}\big[Y_{ij}\big|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, , U_i, D_{ij} = 0\big]\Big].
\end{aligned}
$$

- Assumptions:

(A1) Conditional ignorability: $Y_{ij}(0), Y_{ij}(1) \perp D_{ij}\big|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, U_i$.

(A2) Positivity: $0 < P(D_{ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{Z}_i, U_i) < 1$.

(A3) Consistency: $Y_{ij} = D_{ij}Y_{ij}(1) + (1 - D_{ij})Y_{ij}(0)$.

Appendix: TE

Treatment effects

- Under the three-stage model and $(A1) - (A3)$,

$$\mathbb{E}\left[Y_{ij}(1) \middle| D_{ij} = 1, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, b_i\right] = (\boldsymbol{\beta}_1^\top + \boldsymbol{\beta}_2^\top)\boldsymbol{X}_{ij}^* + (1+\omega)b_i,$$
$$\mathbb{E}\left[Y_{ij}(0) \middle| D_{ij} = 0, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i, b_i\right] = \boldsymbol{\beta}_1^\top \boldsymbol{X}_{ij}^* + b_i.$$

- The CATE given both observed covariates and integrated unobserved effects would be:

$$\mathbb{E}\left[Y_{ij}(1) - Y_{ij}(0) \middle| b_i, \boldsymbol{X}_{ij}, \boldsymbol{Z}_i\right] = \boldsymbol{\beta}_2^\top \boldsymbol{X}_{ij}^* + \omega b_i. \tag{11}$$

Appendix: TE

Treatment effects

- The HTE given the observed covariates would be:

$$\mathbb{E}\left[Y_{ij}(1) - Y_{ij}(0) \middle| \boldsymbol{X}_{ij}, \boldsymbol{Z}_i\right] = \boldsymbol{\beta}_2^\top \boldsymbol{X}_{ij}^* \tag{12}$$

- The ATE would be

$$\mathbb{E}\left[Y_{ij}(1) - Y_{ij}(0)\right] = \boldsymbol{\beta}_2^\top \mathbb{E}\left[\boldsymbol{X}_{ij}^*\right] \tag{13}$$

Appendix: Estimation

Estimates of b_i

- From Laplacian-Variant EM algorithm, we can obtain $\hat{b}_i := \mathbb{E}(b_i | \boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i^*, \hat{\boldsymbol{\theta}})$ as follows:

$$\mathbb{E}\left[b_i | \boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \hat{\boldsymbol{\theta}}\right] = \tilde{b}_i + \frac{1}{\hat{\sigma}^2} \tilde{b}_i' \left\{ \boldsymbol{y}_i - \tilde{\boldsymbol{x}}_i \hat{\boldsymbol{\beta}} - (\mathbf{1} + \hat{\omega} \boldsymbol{d}_i) \tilde{b}_i \right\}^\top (\mathbf{1} + \hat{\omega} \boldsymbol{d}_i) + \sum_{j=1}^{n_i} \left\{ d_{ij} - \hat{\mu}_2(\boldsymbol{x}_{ij}^*, \tilde{b}_i) \right\} \hat{\xi} \tilde{b}_i' - \frac{1}{\hat{\sigma}_b^2} \tilde{b}_i \tilde{b}_i'$$

where $\tilde{b}_i$ is the same in (Case 1), $\tilde{b}_i' = \frac{\partial}{\partial t} \tilde{b}_i(t)$, and

$$\hat{\mu}_2(\boldsymbol{x}_{ij}^*, \tilde{b}_i) = \frac{\exp(\hat{\boldsymbol{\eta}}^\top \boldsymbol{x}_{ij}^* + \hat{\xi} \tilde{b}_i)}{1 + \exp(\hat{\boldsymbol{\eta}}^\top \boldsymbol{x}_{ij}^* + \hat{\xi} \tilde{b}_i)}$$

Appendix: Estimation

Stage 1

- Updates for $\boldsymbol{\beta}$ is the solution of

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q_1^*\left(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(k)}\right) = 0$$

and, the update $\boldsymbol{\beta}^{(k+1)}$ is calculated as

$$\boldsymbol{\beta}^{(k+1)} = \left(\sum_{i=1}^{m} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{X}}_i\right)^{-1} \sum_{i=1}^{m} \tilde{\boldsymbol{X}}_i^\top \left\{\boldsymbol{Y}_i - \left(\mathbf{1} + \omega^{(k)} \boldsymbol{D}_i\right) \mu_i\left(\boldsymbol{\theta}^{(k)}\right)\right\} \tag{14}$$

where $\mu_i\left(\boldsymbol{\theta}^{(k)}\right)$ is an approximated value of posterior mean of b_i, $\mathbb{E}\left(b_i | \boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}^{(k)}\right)$.

APPENDIX: ESTIMATION

Stage 1

- Updates for ω and σ^2 are

$$\omega^{(k+1)} = \left\{\sum_{i=1}^{m} \mathbf{1}^\top \boldsymbol{D}_i \delta_i(\boldsymbol{\theta}^{(k)})\right\}^{-1} \left\{\sum_{i=1}^{m} \left(\boldsymbol{Y}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}\right)^\top \boldsymbol{D}_i \mu_i(\boldsymbol{\theta}^{(k)})\right\} - 1, \tag{15}$$

$$\begin{aligned} \sigma^{2,(k+1)} &= \frac{1}{N}\sum_{i=1}^{m} \left(\boldsymbol{y}_i - \tilde{\boldsymbol{x}}_i\boldsymbol{\beta}\right)^\top \left(\boldsymbol{y}_i - \tilde{\boldsymbol{x}}_i\boldsymbol{\beta}\right) \\ &\quad - \frac{2}{N}\sum_{i=1}^{m} \left(\boldsymbol{y}_i - \tilde{\boldsymbol{x}}_i\boldsymbol{\beta}\right)^\top (\mathbf{1} + \omega \boldsymbol{d}_i) \mu_i(\boldsymbol{\theta}^{(k)}) \\ &\quad + \frac{1}{N}\sum_{i=1}^{m} (\mathbf{1} + \omega \boldsymbol{d}_i)^\top (\mathbf{1} + \omega \boldsymbol{d}_i) \delta_i(\boldsymbol{\theta}^{(k)}), \end{aligned} \tag{16}$$

where $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^{n_i}$, and N is the total number of records ($N = \sum_{i=1}^{m} n_i$).

Appendix: Estimation

Stage 2

- No explicit solution for the updates of $\boldsymbol{\theta}_2 = (\boldsymbol{\eta}, \xi)$.

$$g_1 = g_1(\boldsymbol{\eta}, \xi) := \frac{\partial}{\partial \boldsymbol{\eta}} Q_2^*(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(k)}) = 0,$$
$$g_2 = g_2(\boldsymbol{\eta}, \xi) := \frac{\partial}{\partial \xi} Q_2^*(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(k)}) = 0.$$

- A numerical optimization method such as Newton Raphson's algorithm should be used.
- The first and second derivative of Q_2^* should be calculated.

APPENDIX: ESTIMATION

Stage 2

- Let $g(\boldsymbol{\eta}, \xi)$ be the first derivative of Q_2^*:

$$g(\boldsymbol{\eta}, \xi) = \begin{bmatrix} g_1(\boldsymbol{\eta}, \xi) \\ g_2(\boldsymbol{\eta}, \xi) \end{bmatrix}$$

- Let $\boldsymbol{J}$ be a Jacobian matrix:

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial g_1}{\partial \boldsymbol{\eta}} & \frac{\partial g_1}{\partial \xi} \\ \frac{\partial g_2}{\partial \boldsymbol{\eta}} & \frac{\partial g_2}{\partial \xi} \end{bmatrix}.$$

APPENDIX: ESTIMATION

Stage 2

- Then, the Newton's method finding maximizers of Q_2^* with respect to $\boldsymbol{\theta}_2$ would be

$$\begin{bmatrix} \boldsymbol{\eta}_{t+1} \\ \xi_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\eta}_t \\ \xi_t \end{bmatrix} - \boldsymbol{J}^{-1}_{(\boldsymbol{\eta}_t, \xi_t)} \, g(\boldsymbol{\eta}_t, \xi_t) \tag{17}$$

  where $g(\boldsymbol{\eta}_t, \xi_t)$ and $\boldsymbol{J}_{(\boldsymbol{\eta}_t, \xi_t)}$ refer to the previously defined functions evaluated at $\boldsymbol{\eta} = \boldsymbol{\eta}_t$ and $\xi = \xi_t$.

- The final values from (17) will be the updates for $\boldsymbol{\eta}$ and ξ.

Appendix: Estimation

Stage 3

- Updates for σ_b^2

$$\sigma_b^{2,(k+1)} = \frac{1}{m}\sum_{i=1}^{m} \delta_i(\boldsymbol{\theta}^{(k)}). \tag{18}$$

where $\delta_i\left(\boldsymbol{\theta}^{(k)}\right)$ is an approximated value of $\mathbb{E}\left(b_i^2|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}^{(k)}\right)$.

Appendix: Approximation

Where do we require approximations?

- The incomplete-data log likelihood:

$$\ell_{obs}(\boldsymbol{\theta}) = \int \sum_{i=1}^{m} \log f(\boldsymbol{y}_i, \boldsymbol{d}_i, b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) db_i. \tag{19}$$

- Posterior moments of b_i:

$$\mathbb{E}(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i) = \int b_i f(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i) db_i \tag{20}$$

$$\mathbb{E}(b_i^2|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i) = \int b_i^2 f(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i) db_i \tag{21}$$

- Integral in $Q_2(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$:

$$\int \log\left\{1 + \exp(\boldsymbol{\eta}^T \boldsymbol{x}_{ij}^* + \xi b_i)\right\} f(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i) db_i \tag{22}$$

Appendix: Approximation

Laplace's approximation method

- The conditional density of b_i given the observed data can be written as

$$f(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i) = \frac{f(\boldsymbol{y}_i|\boldsymbol{d}_i, b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f(\boldsymbol{d}_i|b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f(b_i)}{\int f(\boldsymbol{y}_i|\boldsymbol{d}_i, b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f(\boldsymbol{d}_i|b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f(b_i) db_i}. \tag{23}$$

- Choices of h and g for the Laplace's approximation (10) would be

$$h(b_i) = s(b_i) f(\boldsymbol{d}_i|b_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$$
$$g(b_i) = -\frac{1}{n}\left\{ \log m(b_i) + \log f(\boldsymbol{y}_i|\boldsymbol{d}_i, b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) + \log f(b_i) \right\}$$

where $s(\cdot)$ and $m(\cdot)$ are smooth and positive functions.

Appendix: Approximation

Approximating posterior moments

- We cannot choose $s(b_i) = b_i$ because h should be positive.
- Computes Laplace's approximation for the moment generating function (Azevedo-Filho and Shachter, 1994).
- Dividing (Case 2) by (Case 1) will result in the approximated value of posterior moment generating function (mgf):

$$\mathbb{E}[e^{tb_i}|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i] = \frac{\int e^{tb_i} f(\boldsymbol{y}_i|\boldsymbol{d}_i, b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f(\boldsymbol{d}_i|b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f(b_i) db_i}{\int f(\boldsymbol{y}_i|\boldsymbol{d}_i, b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f(\boldsymbol{d}_i|b_i, \boldsymbol{x}_i, \boldsymbol{z}_i) f(b_i) db_i} \approx \frac{\mathit{Case\ 2}}{\mathit{Case\ 1}}$$

- This approximation of the ratio is justified by Tierney and Kadane (1986).

Appendix: Approximation

Approximating posterior moments

- By the definition of expectation and variance from the mgf,

$$\mathbb{E}\left(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\right) = \frac{\partial}{\partial t} \log \mathbb{E}\left(e^{tb_i}|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\right)\Big|_{t=0},$$

$$Var\left(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\right) = \frac{\partial^2}{\partial t^2} \log \mathbb{E}\left(e^{tb_i}|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\right)\Big|_{t=0}.$$

- The posterior second moment can be obtained by computing

$$\mathbb{E}\left(b_i^2|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\right) = Var\left(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\right) + \left\{\mathbb{E}\left(b_i|\boldsymbol{y}_i, \boldsymbol{d}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\right)\right\}^2.$$