Causal Inference in Observational Studies with Clustered Data

Meng Wu and Recai M. Yucel

Department of Epidemiology and Biostatistics
SUNY-Albany

January 11, 2018

Overall Goal

Study causal inference under the framework of potential outcome in non-randomized settings such as administrative or survey data, with complex structures due to correlated observational units. Our overall research goal is motivated by:

Challenges:

- Association does not imply causation
- While causal inference has been increasingly popular in observational studies, its comprehensive treatment in clustered data is somewhat lacking
- Further complications occur due to measurement error as well as missing data (not this presentation)

Outline

- Previous work
- Causal inference in clustered data
- Our methods
  - ACE estimated by random intercept linear mixed model and IPW by standard logistic regression.
  - ACE estimated by random intercept linear mixed model and IPWby random intercept logistic regression.
- Simulation study
- Application
- Discussion

Potential outcome framework

- Potential outcome framework describes the nature of causal effect. If exposure or treatment A is not present, what outcome B would have been.
- Originally proposed by Neyman (1923), then extended by Rubin (1974) to more general settings with implication for observational data.
- The framework is known as Neyman-Rubin Causal Model or Rubin Causal Model (RCM) .

Causal effect under RCM

- Given a dichotomous treatment, each subject has two potential outcomes. One is potentially realized under treatment and the other one is under control.
- Individual causal effect is the difference between the two potential outcomes.
- Fundamental issue of causal inference: only one potential outcome can be observed at a time for the same subject.
- The missing potential outcome makes it impossible to identify individual causal effect.
- Solution: average causal effect (ACE) at the population level.

Dual-modeling strategy (Robins and Rotnitzky, 1995)

- When treatment is not assigned randomly, estimates of potential outcomes are affected by selection bias.
- Probability of treatment assignment given confounders (known as the propensity scores) can be used to remove selection bias.
- Robin and Rotnitzky developed a dual-modeling strategy using propensity scores and showed double robustness in the estimation of outcome.
- This method adjusted residuals in the potential outcome model by inverse probability weighting (IPW).

Other dual-modeling methods

- Coefficient adjustment by IPW.
- Use of inverse propensity as a predictor in the regression models (Bang and Robins, 2005).
- Classification of propensities to create dummy variables and include the dummy variables into the regression models (Schafer, 2008).

Estimation of ACE variance (Schafer and Kang, 2008)

- A simple data setting: let y_1 denote the potential outcome under treatment and y_0 denote the potential outcome under control

$$\hat{ACE} = \hat{\mu_1} - \hat{\mu_0} = a^T\theta$$

where $\hat{\mu_1} = E(y_1) = x^T\beta_1$,$\hat{\mu_0} = E(y_0) = x^T\beta_0$, $a = (0, 0, -1, 1)^T$ and $\theta = (\beta_0, \beta_1, \mu_0, \mu_1)^T$.

- The OLS estimates $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\mu}_0, \hat{\mu}_1)$ from the linear regression can be treated as a solution of a set of joint estimation equations $\sum_{i=1}^{N} \varphi(\theta) = 0$.

- By the central limiting theory and Taylor approximation,

$$\hat{\theta} \approx N(\theta, \hat{J}(\phi(\hat{\theta}))^{-1} V(\phi(\hat{\theta}))(\hat{J}(\phi(\hat{\theta}))^{-1})^T), \quad (1)$$

where $\hat{J}(\varphi(\hat{\theta})) = E\left(\frac{\partial\varphi(\theta)}{\partial\theta}\right)$, $V(\varphi(\theta)) = E(\varphi(\theta)\varphi(\theta)^T)$. Variance of $\hat{ACE}$ can then be estimated by:

$$\hat{V}(\hat{ACE}) = \frac{1}{N} a^T \hat{A}^{-1} \hat{B} (A^{-1})^T a \quad (2)$$

where $B = E(\varphi\varphi^T), A = \hat{J}(\varphi(\hat{\theta}))$.

Notations and settings

- Data Structure

  T_{ij1} - Treatment assignment for subject j in cluster i.

  T_{ij0} - Control group assignment for subject j in cluster i.

  x_{ij} - Observed covariates for subject j in cluster i.

  y_{ij1} - Potential outcome for subject j in cluster i had the subject been assigned to treatment group.

  y_{ij0} - Potential outcome for subject j in cluster i had the subject been assigned to control group.

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

- Units within the same cluster are correlated, but observations from different clusters are independent.

Definitions

- Individual Causal Effect.

$$CE_{ij} = y_{ij1} - y_{ij0}$$

- Average Causl Effect (ACE)

$$ACE = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(y_{ij1} - y_{ij0}) \tag{3}$$

- Fundamental problem: CE_{ij} can not be obtained as only one potential outcome can be observed at a time.
- While it is impossible to compute individual causal effect directly, we can conduct inference on ACE.

Assumptions

- Exchangeability: treatment groups are comparable and outcome is independent of the treatment assignment.
  - This assumption does not hold in observational studies due to confounding effect.
  - Conditional exchangeability: the outcome is independent of treatment assignment conditional on confounding variables.
  - Treatment assignment needs to be modeled.
- Positivity: no unobserved confounders for each treatment group.
- Stable Unit Treatment Value Assumption (SUTVA)
  - Consistency: the treatment effect is the same for all the units.
  - No interference: the potential outcomes of an unit are not affected by the treatment assignment of other units.

Models

- Potential outcome: linear mixed-effects model with random intercept

$$y_{ij} = \alpha_i + x_{ij}^T \beta + \varepsilon_{ij}, \tag{4}$$

where α_i is the cluster effect and assumed to be distributed as N(0, σ_α^2)

- Treatment assignment
  - Standard logistic regression: no clustering effect on treatment assingment

$$\pi_{ij} = (1 + exp(-z_{ij}^T \gamma))^{-1} \tag{5}$$

where z_{ij} is the covariates that are associated with treatment assignment.
  - Random intercept logistic regression: with clustering effect

$$\pi_{ij} = (1 + exp(-z_{ij}^T \gamma - \zeta_i))^{-1} \tag{6}$$

where ζ_i is the random intercept with an independent and identical distribution $N(0, \sigma^2)$.

Method 1: ACE estimated by random-intercept linear mixed model and IPW by standard logistic regression

- Potential outcome models:

$$y_{ij0} = \alpha_{i0} + x_{ij0}^T \beta_0 + \epsilon_{ij0}, \qquad y_{ij1} = \alpha_{i1} + x_{ij1}^T \beta_1 + \epsilon_{ij1} \tag{7}$$

- Coefficients in equation (7) are estimated using data from control group and treatment group, respectively.

$$\hat{ACE} = E(y_{ij1}) - E(y_{ij0}) = \hat{\mu}_1 - \hat{\mu}_0 + \hat{\alpha}_1 - \hat{\alpha}_0 + \hat{\epsilon}_1 - \hat{\epsilon}_0 \tag{8}$$

where

$$\hat{\mu}_1 = \frac{1}{N}\sum_i^m \sum_j^{n_i} \{x_{ij}^T \hat{\beta}_1\}, \hat{\mu}_0 = \frac{1}{N}\sum_i^m \sum_j^{n_i} \{x_{ij}^T \hat{\beta}_0\}, \hat{\alpha}_1 = \frac{1}{N}\sum_i^m \sum_j^{n_i} \{\hat{\alpha}_{i1}\}, \hat{\alpha}_0 = \frac{1}{N}\sum_i^m \sum_j^{n_i} \{\hat{\alpha}_{i0}\},$$

$$\hat{\epsilon}_1 = \frac{\sum_i^m \sum_j^{n_i} T_{ij} \hat{\pi}_{ij}^{-1} (y_{ij} - x_{ij}^T \hat{\beta}_1 - \hat{\alpha}_{i1})}{\sum_i^m \sum_j^{n_i} T_{ij} \hat{\pi}_{ij}^{-1}}, \hat{\epsilon}_0 = \frac{\sum_i^m \sum_j^{n_i} (1 - T_{ij})(1 - \hat{\pi}_{ij})^{-1} (y_{ij} - x_{ij}^T \hat{\beta}_0 - \hat{\alpha}_{i0})}{\sum_i^m \sum_j^{n_i} (1 - T_{ij})(1 - \hat{\pi}_{ij})^{-1}}.$$

Method 1: ACE estimated by Random intercept linear mixed model and IPW by standard logistic regression

- Note that

$$\hat{ACE} = \hat{\mu}_1 - \hat{\mu}_0 + \hat{\alpha}_1 - \hat{\alpha}_0 + \hat{\epsilon}_1 - \hat{\epsilon}_0 = a^T\theta \tag{9}$$

where $a^T = (0, 0, 0, -1, 1, -1, 1, -1, 1), \hat{\theta} = (\hat{\gamma}, \hat{\beta}_0, \hat{\beta}_1, \hat{\mu}_0, \hat{\mu}_1, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\epsilon}_0, \hat{\epsilon}_1)^T$.

- $\hat{\theta}$ can be thought as the solution of a set of joint estimation equations $\sum_{i}^{m}\sum_{j}^{n_i}\varphi_{ij}(\theta) = 0$, where

$$\varphi_{ij\gamma} = (T_{ij} - \pi_{ij})z_{ij}, \varphi_{ij\beta_0} = (1 - T_{ij})x_{ij}^T(\bar{y}_{i0} - x_{ij0}^T\beta_0),$$
$$\varphi_{ij\beta_1} = T_{ij}x_{ij}^T(\bar{y}_{i1} - x_{ij1}^T\beta_1), \varphi_{ij\mu_0} = x_{ij0}^T\beta_0 - \mu_0,$$
$$\varphi_{ij\mu_1} = x_{ij1}^T\beta_1 - \mu_1, \varphi_{ij\alpha_0} = \bar{y}_{ij0} - \bar{x}_{i0}^T\beta_0 - \alpha_0,$$
$$\varphi_{ij\alpha_1} = \bar{y}_{ij1} - \bar{x}_{i1}^T\beta_1 - \alpha_1, \varphi_{ij\epsilon_0} = (1 - T_{ij})(1 - \pi_{ij})^{-1}(y_{ij} - x_{ij}^T\beta_0 - \alpha_{i0} - \epsilon_0),$$
$$\varphi_{ij\epsilon_1} = T_{ij}\pi_{ij}^{-1}(y_{ij} - x_{ij}^T\beta_1 - \alpha_{i1} - \epsilon_1).$$

- ACE variance can be estimated by equation (2).

Method 1: ACE estimated by Random intercept linear mixed model and IPW by standard logistic regression

- Jacobian matrix A is a 9 by 9 lower triangle matrix:

$$\begin{bmatrix} A_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{33} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{42} & 0 & A_{44} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{53} & 0 & A_{55} & 0 & 0 & 0 & 0 \\ 0 & A_{62} & 0 & 0 & 0 & A_{66} & 0 & 0 & 0 \\ 0 & 0 & A_{73} & 0 & 0 & 0 & A_{77} & 0 & 0 \\ A_{81} & A_{82} & 0 & 0 & 0 & A_{86} & 0 & A_{88} & 0 \\ A_{91} & 0 & A_{93} & 0 & 0 & 0 & A_{97} & 0 & A_{99} \end{bmatrix}$$

Method 1: ACE estimated by Random intercept linear mixed model and IPW by standard logistic regression

- We derive the elements in Jacobian matrix A as:

$$\hat{A}_{11} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\hat{pi}_{ij}(1-\hat{pi}_{ij})z_{ij}z_{ij}^T, \quad \hat{A}_{22} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-T_{ij})x_{ij}^T x_{ij}, \quad \hat{A}_{33} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}T_{ij}x_{ij}^T x_{ij},$$

$$\hat{A}_{42} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}x_{ij}, \quad \hat{A}_{44} = -1, \quad \hat{A}_{53} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}x_{ij}, \quad \hat{A}_{55} = -1, \quad \hat{A}_{62} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\bar{x}_{i0}, \quad \hat{A}_{66} = -1,$$

$$\hat{A}_{73} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\bar{x}_{i1}, \quad \hat{A}_{77} = -1, \quad \hat{A}_{81} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-T_{ij})\hat{pi}_{ij}(1-\hat{pi}_{ij})^{-1}(y_{ij}-x_{ij}\hat{\beta}_0-\alpha_0-\epsilon_0)z_{ij}^T,$$

$$\hat{A}_{91} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}T_{ij}\hat{pi}_{ij}^{-1}(1-\hat{pi}_{ij})(y_{ij}-x_{ij}\hat{\beta}_1-\alpha_1-\epsilon_1)z_{ij}^T, \quad \hat{A}_{82} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-T_{ij})(1-\hat{pi}_{ij})^{-1}x_{ij},$$

$$\hat{A}_{93} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}T_{ij}\hat{pi}_{ij}^{-1}x_{ij}, \quad \hat{A}_{86} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-T_{ij})(1-\hat{pi}_{ij})^{-1},$$

$$\hat{A}_{97} = T_{ij}\hat{pi}_{ij}^{-1}, \quad \hat{A}_{88} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-T_{ij})(1-\hat{pi}_{ij})^{-1}, \quad \hat{A}_{99} = T_{ij}\hat{pi}_{ij}^{-1}$$

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- Cluster-specific effects are explicitly allowed in the model for the probability of treatment assignment as described in equation (6)
- The formula for $\hat{ACE}$ is the same as equation (8), but with different parameters in the re-written equation $\hat{ACE} = a^T\theta$:

$$\hat{ACE} = \hat{\mu}_1 - \hat{\mu}_0 + \hat{\alpha}_1 - \hat{\alpha}_0 + \hat{\epsilon}_1 - \hat{\epsilon}_0 = a^T\theta \tag{10}$$

where
$a^T = (0,0,0,0,-1,1,-1,1,-1,1), \hat{\theta} = (\hat{\gamma}, \hat{\sigma}^2, \hat{\beta}_0, \hat{\beta}_1, \hat{\mu}_0, \hat{\mu}_1, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\epsilon}_0, \hat{\epsilon}_1)^T$.

- $\hat{\theta}$ can be thought as the solution of a set of joint estimation equations $\sum_i^m \sum_j^{n_i} \varphi_{ij}(\theta) = 0$, where the added parameters γ and σ^2 are from treatment assignment model.

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- Accordingly, the φ is expressed as:

$$\varphi_{ij\gamma} = T_{ij} z_{ij} - \frac{1}{n_i} \frac{\int_{-\infty}^{\infty} \sum_{j}^{n_i} z_{ij} \frac{e^{z_{ij}^T \gamma + \zeta_i}}{1+e^{z_{ij}^T \gamma + \zeta_i}} e^{h_i(\gamma,\zeta)} d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)} d\zeta},$$

$$\varphi_{ij\sigma^2} = -\frac{1}{2\sigma^2 n_i} + \frac{1}{2\sigma^4 n_i} \frac{\int_{-\infty}^{\infty} \alpha^2 e^{h_i(\gamma,\zeta)} d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)} d\zeta},$$

$$\varphi_{ij\beta_0} = (1 - T_{ij}) x_{ij}^T (\bar{y}_{i0} - x_{ij0}^T \beta_0)$$

$$\varphi_{ij\beta_1} = T_{ij} x_{ij}^T (\bar{y}_{i1} - x_{ij1}^T \beta_1), \varphi_{ij\mu_0} = x_{ij0}^T \beta_0 - \mu_0, \varphi_{ij\mu_1} = x_{ij1}^T \beta_1 - \mu_1,$$

$$\varphi_{ij\alpha_0} = \bar{y}_{ij0} - \bar{x}_{i0}^T \hat{\beta}_0 - \alpha_0, \varphi_{ij\alpha_1} = \bar{y}_{ij1} - \bar{x}_{i1}^T \beta_1 - \alpha_1,$$

$$\varphi_{ij\epsilon_0} = (1 - T_{ij})(1 - \pi_{ij})^{-1} (y_{ij} - x_{ij}^T \beta_0 - \alpha_{i0} - \epsilon_0,$$

$$\varphi_{ij\epsilon_1} = T_{ij} \pi_{ij}^{-1} (y_{ij} - x_{ij}^T \beta_1 - \alpha_{i1} - \epsilon_1).$$

- ACE variance is estimated using equation (2).

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- Jacobian matrix A in this method is a 10 by 10 lower triangle matrix:

$$\begin{bmatrix} A_{11} & A_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{33} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{44} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{53} & 0 & A_{55} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{64} & 0 & A_{66} & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{73} & 0 & 0 & 0 & A_{77} & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{84} & 0 & 0 & 0 & A_{88} & 0 & 0 \\ A_{91} & A_{92} & A_{93} & 0 & 0 & 0 & A_{97} & 0 & A_{99} & 0 \\ A_{101} & A_{102} & 0 & A_{104} & 0 & 0 & 0 & A_{108} & 0 & A_{1010} \end{bmatrix}$$

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- We derive the elements in Jacobian matrix A as:

$$\hat{A}_{11} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\{-\frac{1}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta}\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} Z_{ij}z_{ij}^T\frac{e^{z_{ij}^T\gamma+\zeta_i}}{(1+e^{z_{ij}^T\gamma+\zeta_i})^2}e^{h_i(\gamma,\zeta)}d\zeta + (\frac{\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} Z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta})^2\},$$

$$\hat{A}_{21} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\frac{1}{2\sigma^4}\{-\frac{\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} Z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}\zeta^2 d\zeta}{\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta} + \frac{\int_{-\infty}^{\infty}\zeta^2 e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} Z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta}\},$$

$$\hat{A}_{91} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} T_{ij}\hat{\pi}_{ij}(1-\hat{pi}_{ij})^{-1}(y_{ij}-x_{ij}\hat{\beta}_1-\alpha_1-\epsilon_1)z_{ij}^T,$$

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- We derive the elements in Jacobian matrix A as:

$$\hat{A}_{101} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} T_{ij}\hat{\pi}_{ij}^{-1}(1-\hat{pi}_{ij})(y_{ij} - x_{ij}\hat{\beta}_1 - \alpha_1 - \epsilon_1)z_{ij}^T,$$

$$\hat{A}_{12} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\{-\frac{\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} Z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}\frac{\zeta^2}{2\sigma^4}d\zeta}{\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta},$$

$$+\frac{\int_{-\infty}^{\infty}\frac{\zeta^2}{2\sigma^4}e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} Z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta}\},$$

$$\hat{A}_{22} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\{\frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\frac{\int_{-\infty}^{\infty}\zeta^2 e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta} + \frac{1}{2\sigma^4}(\frac{\int_{-\infty}^{\infty}\frac{\zeta^4}{2\sigma^4}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta},$$

$$-\frac{\int_{-\infty}^{\infty}\zeta^2 e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}\frac{\zeta^2}{2\zeta^4}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}e^{h_i(\gamma,\zeta)}d\zeta})\},$$

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- We derive the elements in Jacobian matrix A as:

$$\hat{A}_{92} = 0, \quad \hat{A}_{102} = 0, \quad \hat{A}_{33} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1 - T_{ij})x_{ij}^{T}x_{ij}, \quad \hat{A}_{53} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} x_{ij},$$

$$\hat{A}_{73} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} \bar{x}_{i0}, \quad \hat{A}_{93} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1 - T_{ij})(1 - \hat{\pi}_{ij})^{-1}x_{ij},$$

$$\hat{A}_{44} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} T_{ij})x_{ij}^{T}x_{ij}, \quad \hat{A}_{64} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} x_{ij},$$

$$\hat{A}_{84} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} \bar{x}_{i1}, \quad \hat{A}_{104} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} T_{ij}\hat{\pi}_{ij}^{-1}x_{ij}, \quad \hat{A}_{55} = -1, \quad \hat{A}_{66} = -1, \quad \hat{A}_{77} = -1,$$

$$\hat{A}_{97} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1 - T_{ij})(1 - \hat{\pi}_{ij})^{-1}, \quad \hat{A}_{88} = -1, \quad \hat{A}_{108} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} T_{ij}\hat{\pi}_{ij})^{-1},$$

$$\hat{A}_{99} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1 - T_{ij})(1 - \hat{\pi}_{ij})^{-1}, \quad \hat{A}_{1010} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} T_{ij}\hat{\pi}_{ij}^{-1},$$

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- Approximation of integrals in the computation of ACE variance

$$f_i(1) = \int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)} d\zeta, \quad f_i(2) = \int_{-\infty}^{\infty} \zeta^2 e^{h_i(\gamma,\zeta)} d\zeta, \quad f_i(3) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_i} Z_{ij} \frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}} e^{h_i(\gamma,\zeta)} d\zeta,$$

$$f_i(4) = \int_{-\infty}^{\infty} \frac{\zeta^2}{2\sigma^4} e^{h_i(\gamma,\zeta)} d\zeta, \quad f_i(5) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_i} Z_{ij} z_{ij}^T \frac{e^{z_{ij}^T\gamma+\zeta_i}}{(1+e^{z_{ij}^T\gamma+\zeta_i})^2} e^{h_i(\gamma,\zeta)} d\zeta,$$

$$f_i(6) = \int_{-\infty}^{\infty} \frac{\zeta^4}{2\sigma^4} e^{h_i(\gamma,\zeta)} d\zeta, \quad f_i(7) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_i} Z_{ij} \frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}} e^{h_i(\gamma,\zeta)} \frac{\zeta^4}{2\sigma^4} d\zeta,$$

$$f_i(8) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_i} Z_{ij} \frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}} e^{h_i(\gamma,\zeta)} \zeta^2 d\zeta$$

where $k_i = \sum_{j=1}^{n_i} T_{ij}$ and $h_i(\gamma,\zeta) = k_i\zeta - \frac{\zeta^2}{2\sigma^2} - \sum_{j=1}^{n_i} ln(1+e^{\gamma z_{ij}+\zeta})$.

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- For example, to approximate $f_i(1)$, we need to minimum $h_i(\gamma, \zeta) = -k_i\zeta + \frac{\zeta^2}{2\sigma^2} + \sum_{j=1}^{n_i} ln(1 + e^{\gamma z_{ij} + \zeta})$.

$$\frac{dh_i}{d\zeta} = -k_i + \frac{\zeta}{\sigma^2} + e^{\zeta}\sum_{j=1}^{n_i}\frac{B_j}{1 + B_j e^{\zeta}}, \frac{d^2 h_i}{d\zeta^2} = \frac{1}{\sigma^2} + e^{\zeta}\sum_{j=1}^{n_i}\frac{B_j}{(1 + B_j e^{\zeta})^2},$$

  where $B_j = e^{\gamma z_{ij}}$.
- The second derivative is positive,the function has a unique minimum.
- Newton-Raphson algorithm can be used to solve $\frac{dh_i}{d\zeta} = 0$.

$$\zeta_{s+1} = \zeta_s - (\frac{dh_i}{d\zeta})(\frac{d^2 h_i}{d\zeta^2})^{-1} \quad (11)$$

Method 2: ACE estimated by Random intercept linear mixed model and IPW by random intercept logistic regression

- Starting from zero, if ζ_{max} is the limiting point of the iterations, the integral $f_i(1)$ can be approximated by

$$\int_{-\infty}^{\infty} e^{h(\zeta)} d\zeta \approx \sqrt{2}\hat{\nu}_h \sum_{k=1}^{K} w_k exp[\zeta_k^2 + h(\zeta_{max} + \sqrt{2}\hat{\nu}_h \zeta_k)] \tag{12}$$

  where $\hat{\nu}_h = (-\frac{d^2 h_i}{d\zeta^2}|_{\zeta=\zeta_{max}})^{-\frac{1}{2}}$, K and α_k are Gauss-Hermite abscissas and weights.

- Note: integral $f_i(2)$ and $f_i(8)$ are not a unimodal function, approximation needs to be split into two intervals (-1, 0) and (0,1).
- All the integrals can be approximated by this algorithm.

Simulation

Performance criteria

The performance of our methods over 1000 samples from a population described in the next slide the following criteria:

- Average bias
- Root-Mean-Square error (RMSE)
- Standard deviation of ACE estimates
- Average of standard error estimates
- The percent coverage rate of nominal 95% confidence intervals.

Data generation

- First, simulate a set of three covariates from independent normal distributions with varying means and variances: $x_1 \sim N(4,4)$, $x_2 \sim N(4,4)$ and $x_3 \sim N(12,4)$.
- Next, treatment membership T is simulated from a binomial distribution with membership probability

$$\pi_{ij} = (1 + exp(-(-3 + x_{1ij} + 3x_{2ij} - x_{3ij} + \zeta_i)^{-1},$$

where ζ_i is a cluster random-effect assumed to follow a normal distribution $N(0, \sigma^2)$, independently across the clusters $i = 1, 2, \ldots, m$. In the first scenario, we work with a fixed value of σ^2 which is varied in Scenario 2.

- Based on the covariates as well as the treatment assignment,we simulate outcome

$$y_{ij} = 18 + 2x_{1ij} + 3x_{2ij} + 0.8x_{3ij} + 4t_{ij} + \epsilon_{ij} + \alpha_i,$$

where ϵ_{ij} and α_i refer to residual error term and random-effects, respectively. They are further assumed to be independent and normally distributed:

$$\epsilon_{ij} \sim N(0, \sigma^2), \alpha_i \sim N(0, \sigma_1^2), \forall i, j.$$

Data generation

- We simulated a population of 3,000,000 observational units that are grouped under 3000 clusters.
- Scenario 1: clusters between 50 and 150, and number observations within clusters between 40 and 100 (σ^2=0.5 and σ_1^2=0.59).
- Scenario 2: ICC ranged from 0.12 to 0.85 for the outcome model and 0.08 to 0.3 for the treatment assignment (100 clusters and 60 units within each cluster).
- Scenario 3: replace x_2 and x_3 with asymmetric covariates distributed as x_k^2 and $lnN(2, 0.36)$, same ICC settings as scenario 2.

Table 1: *Performance of Methods for Estimating ACE and Standard Error from Data with Various Sample Size: Average Bias (Bias), Root-Mean-Square Error(RMSE), Standard Deviation of ACE Estimates (SD), Average of SE Estimates (SE), Percent Coverage Rate of Nominal 95% Confidence Intervals (CR)*

	Method 1. IPW by Logistic Model (ACE=4.0)						Method 2. IPW by Mixed Logistic Model (ACE=4.0)				
	Bias	RMSE	SD	SE	CR		Bias	RMSE	SD	SE	CR
						m = 150					
100	2.67	2.99	1.35	1.28	33		2.60	2.90	1.28	3.10	76
90	2.57	3.05	1.65	1.37	38		1.81	2.18	1.21	1.53	69
80	2.39	3.42	2.45	1.44	36		2.08	2.38	1.16	2.04	72
70	2.85	3.39	1.84	1.34	32		1.63	2.43	1.81	2.02	77
60	2.68	3.55	2.32	1.48	28		2.06	2.47	1.37	1.67	65
50	2.88	3.60	2.16	1.54	35		1.71	2.86	2.28	2.16	68
40	2.92	3.63	2.15	1.69	36		2.18	2.77	1.70	1.84	59
						m = 100					
100	2.48	3.30	2.17	1.53	38		2.60	2.90	1.28	3.11	76
90	2.45	3.11	1.92	1.57	44		1.48	2.24	2.48	2.24	75
80	2.84	3.42	1.90	1.54	33		2.06	2.47	1.37	1.67	65
70	2.80	3.27	1.70	1.62	41		2.11	2.62	1.56	2.62	79
60	2.52	3.49	2.42	1.75	43		1.73	2.61	1.98	2.45	75
50	2.84	3.54	2.12	1.82	46		2.21	2.67	1.50	2.69	74
40	3.10	3.84	2.27	1.87	39		2.22	2.88	1.83	3.04	74
						m = 50					
100	2.61	3.30	2.02	1.87	50		2.21	2.67	1.50	2.69	74
90	2.86	3.78	2.47	1.80	42		2.11	2.62	1.56	2.56	79
80	3.04	3.41	1.55	1.88	44		1.73	2.61	1.98	2.45	75
70	3.14	3.69	1.94	1.89	42		1.60	2.68	2.14	5.23	92
60	2.95	4.13	2.89	2.06	46		2.11	2.62	1.56	2.56	79
50	3.33	3.86	1.94	2.22	48		2.45	2.93	1.61	4.71	88
40	3.18	3.95	2.34	2.44	56		2.71	3.05	1.40	4.87	93

Table 2: *Performance of Methods for Estimating ACE and Standard Error Based on Various Intra Class Correlations: Average Bias (AB), Root-Mean-Square Error(RMSE), Standard Deviation of ACE Estimates (SD), Average of SE Estimates (SE), Percent Coverage Rate of Nominal 95% Confidence Intervals (CR)*

	Method 1. IPW by Logistic Model (ACE=4.0)						Method 2. IPW by Mixed Logistic Model (ACE=4.0)				
ICC1	Bias	RMSE	SD	SE	CR		Bias	RMSE	SD	SE	CR
						ICC.2=0.08					
0.12	2.68	3.19	1.72	1.74	53		2.42	2.78	1.36	5.35	86
0.40	2.95	3.44	1.77	1.65	41		2.00	2.76	1.90	8.93	92
0.59	3.01	3.49	1.77	1.66	39		1.84	2.70	1.99	7.62	92
0.85	3.16	3.60	1.73	1.55	34		2.05	2.62	1.63	4.45	90
						ICC.2=0.30					
0.12	3.13	3.40	1.32	1.45	32		1.60	2.73	2.21	2.79	85
0.40	3.52	3.68	1.06	1.34	21		2.08	2.60	1.57	2.96	87
0.59	3.52	3.68	1.06	1.35	21		2.21	2.81	1.74	3.23	87
0.85	3.26	3.59	1.52	1.42	30		1.67	2.74	2.17	3.57	86

Table 3: *Performance of Methods for Estimating ACE and Standard Error from Samples with Asymmetric Covariates: Average Bias (Bias), Root-Mean-Square Error(RMSE), Standard Deviation of ACE Estimates (SD), Average of SE Estimates (SE), Percent Coverage Rate of Nominal 95% Confidence Intervals (CR)*

	Method 1. IPW by Logistic Model (ACE=4.0)						Method 2. IPW by Mixed Logistic Model (ACE=4.0)				
ICC1	Bias	RMSE	SD	SE	CR		Bias	RMSE	SD	SE	CR
						ICC.2=0.08					
0.12	1.41	2.38	1.91	1.61	83		1.23	2.00	1.58	1.64	79
0.40	0.97	2.34	2.12	1.70	85		0.98	1.86	1.58	1.70	80
0.59	1.41	2.03	1.46	1.66	87		0.26	2.18	2.19	1.68	95
0.85	1.50	2.27	1.70	1.66	82		-0.25	2.42	2.43	1.63	93
						ICC.2=0.30					
0.12	1.37	1.73	1.05	1.38	82		1.22	1.79	1.31	1.40	87
0.40	1.24	1.82	1.32	1.46	85		-0.04	2.05	2.06	1.68	92
0.59	1.48	1.83	1.09	1.37	86		-0.06	1.49	1.49	1.47	95
0.85	1.37	2.21	1.74	1.52	82		-0.21	2.51	2.52	1.77	94

Prenatal care

- Prenatal care plays an important role for the well-being of pregnant women and their babies
- Studies have shown that inadequate prenatal care is significantly associated with low birth weight (Donaldson, 1984; Scholl, 1987; Hueston, 1995; Pedraza 2013; Loftus, 2015)
- Although the benefit of adequate care on reduction of incidents of low birth weight has been widely discussed in literature, the explicit quantitative effect of inadequate care on birth weight for those full term babies has rarely been reported
- Our goal is to investigate to what extent the inadequate prenatal care affects birth weight

Data source

- New York State 2009 vital records, collected separately from the five boroughs of New York City and the rest of state
- 54,880 birth records with gestational age greater than 37 weeks
- Mothers are covered by New York State Medicaid program
- 120 hospitals considered as clusters
- Less than 9 visits are grouped as inadequate care based on Kessner index (Kessner, 1973)

Table 4: *Descriptive Statistics for Confounding Variables*

Binary Variable	Inadequate Care		Adequate Care	
	N	%	*N*	%
Previous Low Birth Weight* * *	177	0.6%	201	0.8%
Preexisting Hypertension	252	0.8%	205	0.8%
pregnancy Induced Diabetes	1,094	3.6%	846	3.4%
Adverse Event	718	2.4%	612	2.5%
less than HS Education* * *	9,804	32.4%	9,563	38.9%
SSI Eiligible	366	1.2%	342	1.4%
Black* * *	6,233	20.6%	6,644	27.0%
Hispanic	10,376	34.3%	8,559	34.8%
White* * *	10,270	33.9%	6,877	28.0%
No Previous Live Births* * *	14,853	49.0%	11,190	45.5%
Smoking Status	3,942	13.0%	3,250	13.2%
Continous Variable	Mean	SD	Mean	SD
Gestational Age	39.4	0.93	39.4	0.97
Mother's BMI* * *	35.6	26.2	33.6	21
Mother's Age	26.1	5.7	25.6	5.9

† Note: adverse event includes abruptio placenta or Eclampsia or infection or pregnancy induced hypertension

† *** Significantly different between two groups.

Results

- Our simulation experiments show that method 2 is a preferable choice when the clustering effect on treatment assignment is not ignorable.
- Receiving inadequate care would reduce birth weight of 24.1 gram on average. The estimated standard error of the ACE is 4.7.

- The dual model strategy in our methods has potential to address selection bias into the treatment group
- Ignoring the unobservable cluster-specific effect on treatment assignment leads to dismal performance
- Correct specification of both models is ideal but not realistic
- We have found that both methods may fail under some circumstances such as extreme large ICCs or small sample size. This may be caused by the outliers in the predicted propensity scores. These outliers can lead to extreme values in the estimation equations as well as integral approximation.
- We recommend the estimation of ACE should be within each cluster if the clustering effect is large.