A Counterfactual Approach to Quantify the Causal Effect of Fine Particulate Matter on Mortality

Hengfang Wang    Zhulin He    Zhengyuan Zhu

Iowa State University
CSSM

04/21/2017
1 Motivation

2 Structural Mean Models

3 Structural Nested Mean Models

4 PM$_{2.5}$ Case

5 Simulation Setup

6 Future Work
Some cities in developing countries are suffering high concentration of fine particular matter (here, we mean PM$_{2.5}$), like Beijing in China, New Delhi in India and so on.

Donald J. Trump decided to reduce the fund of EPA and swore to bloom they manufacturing industry, which might bring this problem to the USA.

Global issue.
Research reveals that short-term PM$_{2.5}$ exposure increases risk of cardiovascular and respiratory diseases.

There’s statistical analysis with Medicare cohorts to establish the association between PM$_{2.5}$ and mortality in people aged 65 and over.

Challenge: hard to quantify the causal effect of PM$_{2.5}$ due to other confounding effects such as age, medical history and meteorological variables.
Characteristics of Our Study

- The treatment (PM$_{2.5}$ exposure density): assigned to a geographical area, continuous.
- The outcomes (death or not): observed at individual level, binary.
- The treatment changes over time and is confounded with meteorological variables.
- The treatment and some confounding factors are spatially and temporally correlated.
- Confounders like other causes of death might exist.
1 Motivation

2 Structural Mean Models

3 Structural Nested Mean Models

4 PM$_{2.5}$ Case

5 Simulation Setup

6 Future Work
The counterfactual concept, also called potential outcome.

Intuitively, suppose $Y^a$ and $Y^{a*}$ represent the potential outcomes for an individual if s/he received the treatment or control, respectively.

The causal effect for the individual could be expressed as $Y^a - Y^{a*}$.

In most cases, an individual can only receive one treatment, only one of the potential outcomes could be observed.
Let $a^* = 0$ for notational convenience. Average causal effects can be defined in terms of comparisons of average potential outcomes,

$$RD = E(Y^a \mid L = l, A = a) - E(Y^0 \mid L = l, A = a)$$ or $$RR = \frac{E(Y^a \mid L = l, A = a)}{E(Y^0 \mid L = l, A = a)}.$$ (1) (2)

where $RD$ denotes relative difference, $RR$ denotes relative risk.
Structural Mean Models (SMMs) parameterize average causal effects in subjects receiving level $a$ of treatment as

$$
g\{E(Y^a | L = l, A = a)\} - g\{E(Y^0 | L = l, A = a)\} = \gamma^*(l, a; \psi^*), \tag{3}
$$

where

- $g(\cdot)$ is a known link function
- $\gamma^*(l, a; \psi)$ is a known function, satisfying $\gamma^*(l, 0; \psi) = 0$ for all $l$ and $\psi$

Typically, parameterization: $\gamma^*(l, a; 0) = 0$ for all $a$ and $l$ will make:

$$\psi^* = 0 \rightarrow \text{no treatment effect}$$
Structural Mean Models (SMMs)

Some examples of SMMs:

- **Additive or linear SMM** uses the identity link $g(x) = x$:

  $$E(Y^a \mid L = l, A = a) - E(Y^0 \mid L = l, A = a) = (\psi_0^* + \psi_1^* l) a, \tag{4}$$

- **Multiplicative or loglinear SMM** uses the log link $g(x) = \log(x)$,

  $$\frac{E(Y^a \mid L = l, A = a)}{E(Y^0 \mid L = l, A = a)} = \exp\{(\psi_0^* + \psi_1^* l) a\},$$

- **Logistic SMM** uses the logit link $g(x) = \text{logit}(x)$,

  $$\frac{\text{odds}(Y^a = 1 \mid L = l, A = a)}{\text{odds}(Y^0 = 1 \mid L = l, A = a)} = \exp\{(\psi_0^* + \psi_1^* l) a\},$$
One can use a SMM to construct a variable $U^*(\psi)$ whose mean value (within certain groups) equals the mean outcome that can be seen that it had treatment removed from that subset.

- If $g(\cdot)$ is the identity link,
  \[ U^*(\psi) = Y - \gamma^*(L, A; \psi), \]

- If $g(\cdot)$ is the log link,
  \[ U^*(\psi) = Y \exp\{-\gamma^*(L, A; \psi)\}, \]

- If $g(\cdot)$ is the logit link,
  \[
  U^*(\psi) = \expit\left[\logit\left\{E(Y \mid L, A)\right\} - \gamma^*(L, A; \psi)\right] \\
  = g^{-1}\left[g\{E(Y \mid L, A)\} - \gamma^*(L, A; \psi)\right] \\
  \\
  \text{where } \expit(x) = \frac{\exp(x)}{1+\exp(x)} \text{ is the inverse function of logit}(\cdot) \\
  \\
  \text{Then} \\
  E\{U^*(\psi^*) \mid L, A\} = E(Y^0 \mid L, A). \tag{5} \]
Ignorability Assumptions

- The required no unmeasured confounders assumption for the identification of the parameter $\psi^*$ indexing SMMs can be formulated as

$$A \perp\!\!\!\!\!\!\perp Y^0 \mid L,$$

meaning that $A$ is conditionally independent of $Y^0$ given $L$.

- The assumption above indicates that the treatment assignment is independent of potential outcome given the covariates.
Estimation under Ignorability

- The SMM together with the ignorability assumption (6) implies that

\[
E \{ U^* (\psi^*) \mid L, A \} = E (Y^0 \mid A, L) = E (Y^0 \mid L) = E \{ U^* (\psi^*) \mid L \}.
\]

- Estimation of $\psi^*$ in a SMM can thus be based on solving estimating equations:

\[
\sum_{i=1}^{n} \left[ d^* (A_i, L_i) - E \{ d^* (A_i, L_i) \mid L_i \} \right] \cdot \left[ U_i^* (\psi) - E \{ U_i^* (\psi) \mid L_i \} \right] = 0,
\]

where essentially set $\hat{\text{Cov}} (U^* (\psi), d^* (A, L)) = 0$, where $d^* (A, L)$ is usually an arbitrary function.
For instance, for model (8) (linear SMM),

\[
E(Y^a \mid L = l, A = a) - E(Y^0 \mid L = l, A = a) = (\psi_0^* + \psi_1^* l)a,
\]

the choice \(d^*(A_i, L_i) = (1, L_i)'A_i\) results in estimating equations

\[
\sum_{i=1}^{n} \binom{1}{L_i} \{A_i - E(A_i \mid L_i)\} \cdot [Y_i - E(Y_i \mid L_i) - (\psi_0 + \psi_1 L_i)\{A_i - E(A_i \mid L_i)\}] = 0,
\]
A locally efficient estimator of $\psi^*$ [under the SMM together with the ignorability assumption (6)] can be attained by setting

$$d^*(A, L) = E\left\{ \frac{\partial U^*(\psi^*)}{\partial \psi} \mid A, L \right\},$$

when the variance of $U^*(\psi^*)$ given $A, L$ is constant.

Local here means that the efficiency is only attained when this constant variance assumption is met and models for all conditional expectations involved in (7) are correctly specified.
Motivation

Structural Mean Models

Structural Nested Mean Models

PM$_{2.5}$ Case

Simulation Setup

Future Work
Suppose that measurements on exposure and confounders are collected at time point $t_0$ and that outcome measurements are recorded at fixed later time points $t_1, \ldots, t_{K+1}$.

- For a variable $X$, $X_k$ denote the level of the variable that one obtains at time $t_k$.
- $\overline{X}_k = \{X_0, X_1, \ldots, X_k\}$ denotes the history of $X$ through $t_k$.
- $\underline{X}_k \equiv \{X_k, \ldots, X_{K+1}\}$ denotes the future of the variable.
Notation Setup

- $A_k$ : the treatment provided at time $t_k$, $k = 0, \ldots, K$
- $L_k$ : other covariates measured at time $t_k$, $k = 0, \ldots, K$
- $Y_k$, the outcome measured at time $t_k$, $k = 1, \ldots, K + 1$, is part of $L_k$.
- We presume the variables are ordered $L_0, A_0, L_1, A_1$, etc.
Notation Setup

- Let $Y_{m}^{\bar{a}_{m-1}}$ denote the outcome that would be seen at time $t_{m}$ in a given individual’s treatment received history.
- $Y_{m}^{\bar{a}_{m-1}}$ is a potential outcome, via the consistency assumption that $Y_{m}^{\bar{A}_{m-1}} = Y_{m}^{\bar{a}_{m-1}}$ if $\bar{A}_{m-1} = \bar{a}_{m-1}$.
- Assume that treatment at or after $t_{m}$ cannot affect outcome at times up to $t_{m}$; thus, $Y_{m}^{\bar{a}_{m-1},a_{m}} = Y_{m}^{\bar{a}_{m-1},\bar{a}_{m}}$ for $a_{m} \neq \bar{a}_{m}$.
- Causal effects (short term) can now be defined as comparisons of potential outcomes $Y_{K+1}^{(\bar{a}_{K-1},a_{K})}$ for the same group of subjects for different treatment histories (only different at time $t_{K}$ (or between $t_{K}$ and $t_{K+1}$)) $\bar{a}_{K} = (\bar{a}_{K-1}, a_{K})$, $\bar{a}_{K}^{\dagger} = (\bar{a}_{K-1}, a_{K}^{\dagger})$, where $a_{K} \neq a_{K}^{\dagger}$.
Structural nested mean models (SNMMs) simulate the sequential removal of an amount of treatment at $t_m$ on subsequent average outcomes, after having removed the effects of all subsequent treatments.

Contrasts of $Y_{m+1}^{\bar{a}_m}$ and $Y_{m+1}^{\bar{a}_{m-1},0}$ conditionally on treatment and covariate histories through $t_m$ as

$$g\left\{ E\left( Y_{m+1}^{\bar{a}_m} \mid L_m = \bar{l}_m, A_m = \bar{a}_m \right) \right\} - g\left\{ E\left( Y_{m+1}^{\bar{a}_{m-1},0} \mid L_m = \bar{l}_m, A_m = \bar{a}_m \right) \right\} = \gamma^*_m(\bar{l}_m, \bar{a}_m; \psi^*)$$

for each $m = 0, \ldots, K$ and $(\bar{l}_m, \bar{a}_m)$, where $\gamma^*_m(\bar{l}_m, \bar{a}_m; \psi)$ 1-dimensional function.
Typically, the parameterization is chosen to be $\gamma^*_m(\tilde{l}_m, \bar{a}_m; 0) = 0$ for all $\tilde{l}_m, \bar{a}_m$ so that $\psi = 0$ encodes the null hypothesis of no treatment effect.

Under the SNMM, as in the previous section, it is possible to define a transformation $U^*_m(\psi^*)$ of $Y_{m+1}$, whose mean value equals the mean that would be observed if treatment were suspended from time $t_m$ onward, in the sense that

$$E\{U^*_m(\psi^*) \mid \tilde{L}_m, \bar{A}_{m-1} = \bar{a}_{m-1}, A_m\}$$

$$= E(Y_{m+1}^{\bar{a}_{m-1}, 0} \mid \tilde{L}_m, \bar{A}_{m-1} = \bar{a}_{m-1}, A_m),$$

for $m = 0, \ldots, K$. 

$$ (9) $$
If $g(\cdot)$ is the identity link, $U_m^*(\psi)$ is a scalar with

$$Y_{m+1} - \gamma_{m,m+1}^* (\bar{L}_m, \bar{A}_m; \psi),$$

for $m = 0, \ldots, K$

If $g(\cdot)$ is the log link, $U_m^*(\psi)$ is a scalar with

$$Y_{m+1} \exp \left\{ \gamma_{m,m+1}^* (\bar{L}_m, \bar{A}_m; \psi) \right\},$$

for $m = 0, \ldots, K$
If \( g(\cdot) \) is the logit link, \( U^*_m(\psi) \) is a scalar with under the SNMM we have that

\[
\text{expit}\left[ \logit\left\{ E(Y_{m+1} \mid \bar{L}_m, \bar{A}_m) \right\} - \gamma^*_{m,m+1}(\bar{L}_m, \bar{A}_m; \psi) \right],
\]

for \( m = 0, \ldots, K \)
Sequential Ignorability

- The assumption of ignorable treatment assignment can be generalised to sequential treatments as follows:

\[ A_m \perp \perp Y_{m+1}^{a_{m-1},0} \mid L_m, A_{m-1} = a_{m-1}, \]  

for \( m = 0, \ldots, K \).

- The assumption above indicates that treatment assignment at time \( t_m \) is independent of potential outcome given the history of covariates and assignments.
Estimation under Sequential Ignorability

- This assumption together with identity (9) imply that

$$E\left\{ U_m(\psi^*) \mid \bar{L}_m, \bar{A}_m \right\} = E\left\{ U_m(\psi^*) \mid \bar{L}_m, \bar{A}_{m-1} \right\}$$

for all $m$ under a SNMM.

- The parameter $\psi^*$ indexing a SNMM can therefore be estimated by solving

$$\sum_{i=1}^{n} \sum_{m=0}^{K} \left[ d_m(\bar{L}_{im}, \bar{A}_{im}) - E\left\{ d_m(\bar{L}_{im}, \bar{A}_{im}) \mid \bar{L}_{im}, \bar{A}_{i,m-1} \right\} \right] (11)$$

$$\times \left[ U_{im}(\psi) - E\left\{ U_{im}(\psi) \mid \bar{L}_{im}, \bar{A}_{i,m-1} \right\} \right] = 0,$$

where $d_m(\bar{L}_{im}, \bar{A}_{im}), m = 0, \ldots, K$ is an arbitrary $p \times 1$-dimensional function, with $p$ the dimension of $\psi$. 
Estimation under Sequential Ignorability

When the previous outcome is included in the confounder history (i.e., $\bar{L}_{im}$ includes $Y_{im}$) and there is homoscedasticity [i.e., when the conditional variance of $U_{im}(\psi^*)$ given $\bar{L}_{im}, \bar{A}_{im}$ is constant for $m = 0, \ldots, K$], then local semiparametric efficiency under the SNMM is attained upon choosing

$$d_m(\bar{L}_{im}, \bar{A}_{im}) = E\left\{ \frac{\partial U_m(\psi^*)}{\partial \psi} \mid \bar{L}_{im}, \bar{A}_{im} \right\}.$$
1 Motivation

2 Structural Mean Models

3 Structural Nested Mean Models

4 PM2.5 Case

5 Simulation Setup

6 Future Work
Basic Setup

- $V$: individual-level information, such as age, gender
- $\bar{L}_t = (L_0, L_1, \cdots, L_t)$: meteorological information up to time $t$, where $t = 1, \ldots, T$
- $\bar{A}_t = (A_0, A_1, \cdots, A_t)$: PM$_{2.5}$ exposure information up to time $t$, where $t = 1, \ldots, T$
- $a^*$: baseline PM$_{2.5}$ exposure density
- $Y_{t+1, a}$: individual potential outcomes at time $t + 1$ if the individual received a sequential set of PM$_{2.5}$ exposure density as $\bar{A}_t = (\bar{a}_{t-1}, a)$ up to time $t$
- $D$: an unmeasured set of disease that cause death
- $C$: a set of unobserved confounders that cause death
Basic Setup

- Directed acyclic graph (DAG) for our case:

```
   L
  ↓
  A
  ↓
V
  ↓
  D
  ↑
  C
  ↓
  Y
```
Causal Effects

- Short-term risk difference (RD) for density $A_t = a_t$ and mortality outcome $Y_{ij}$ at $t + 1$:

$$RD(a|V, \bar{L}_t, \bar{A}_t = \bar{a}_t) = E(Y_{t+1}^{\bar{a}_t-1,a}|V, \bar{L}_t, \bar{A}_t = \bar{a}_t) - E(Y_{t+1}^{\bar{a}_t-1,a^*}|V, \bar{L}_t, \bar{A}_t = \bar{a}_t)$$  \hspace{1cm} (12)

- Short-term relative risk (RR) under the above setting:

$$RR(a|V, \bar{L}_t, \bar{A}_t = \bar{a}_t) = \frac{E(Y_{t+1}^{\bar{a}_t-1,a}|V, \bar{L}_t, \bar{A}_t = \bar{a}_t)}{E(Y_{t+1}^{\bar{a}_t-1,a^*}|V, \bar{L}_t, \bar{a}_t)}$$  \hspace{1cm} (13)
Assumptions

- Consistency assumption: if $\bar{A}_t = \bar{a}_t$, then $Y_{t+1} = Y_{\bar{a}_t}^{\bar{a}_t}$.
- Potential outcome at time $t+1$ is not affected by the treatment received at or after time $t+1$.
- Sequential ignorance: $\{A_t, L_t\} \perp \perp Y_{\bar{a}_t-1}^{\bar{a}_t-1, a^*_t} | L_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}$ for all $t$ and $s$, where $Y_{\bar{a}_t-1}^{\bar{a}_t-1, a^*_t} = (Y_{\bar{a}_t-1, a^*_t-1}, Y_{\bar{a}_t-1, a^*_t, a^*_s}, \ldots, Y_{\bar{a}_t-1, a^*_t, a^*_s})$.
- Positivity assumptions: $0 < f(A_t | L_t, \bar{A}_{t-1}) < 1$ and $0 < f(A_t, L_t | L_{t-1}, \bar{A}_{t-1}) < 1$, where $f$ is the density function.
- Additional independence assumption: $V \perp \perp \{L_t, \bar{A}_t\}$ for all $t$. 
Structural Nested Mean Model (SNMM)

Under the aforementioned assumptions, to estimate the short-term causal effect of PM$_{2.5}$ density on mortality, we propose a SNMM as

$$\gamma_t(v, \bar{l}_t, \bar{a}_t; \psi) = g\{E(Y_{t+1}^{\bar{a}_t-1, a_t} \mid V = v, \bar{L}_t = \bar{l}_t, \bar{A}_t = \bar{a}_t)\}$$

$$- g\{E(Y_{t+1}^{\bar{a}_t-1, a^*} \mid V = v, \bar{L}_t = \bar{l}_t, \bar{A}_t = \bar{a}_t)\}$$

for all $t$, where

- $g$ : link function (here, we assign logit link)
- $\gamma_t(v, \bar{l}_t, \bar{a}_t; \psi)$: a known function of $v, \bar{l}_t$ and $\bar{a}_t$, smooth in a parameter vector $\psi$. $\gamma_t$ also satisfies that $\gamma_t(v, \bar{l}_t, \bar{a}_{t-1}, a_t = a^*; \psi) = 0$ for all $v, \bar{l}_t, \bar{a}_{t-1}$, and $\psi$. 
Spatial temporal model of PM$_{2.5}$ and meteorological variables

- According to the SNMMs and the additional independence assumption, the expected potential outcome at time $t + 1$ can be expressed as

$$E(Y_{t+1}^{\bar{a}_{t-1}, \bar{a}^*} | V = v, \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1})$$

$$= E_{A_t, L_t | \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}} \left( g^{-1} \left( g(E(Y_{t+1}^{\bar{a}_{t-1}, a_t} | V = v, \bar{L}_{t} = \bar{l}_{t}, \bar{A}_{t} = \bar{a}_{t})) - \gamma_t(v, \bar{l}_{t-1}, L_t, \bar{a}_{t-1}, A_t; \psi) \right) \right).$$

- We need information for $E(Y_{t+1}^{\bar{a}_t} | V = v, \bar{L}_{t} = \bar{l}_{t}, \bar{A}_{t} = \bar{a}_{t})$, and the distribution of $A_t, L_t | \bar{L}_{t-1}, \bar{A}_{t-1}$.
Adjustment for confounding causes of death

- These and many other causes of death could confound the causal effect of PM$_{2.5}$ exposure density on mortality.
- We propose a ratio function to adjust for the effects of confounding causes on mortality. The ratio function is defined as a ratio between two longitudinal baseline potential outcomes, i.e.

$$r_{t+1,s}(V) = \frac{E(Y_{t+1,a^*} \mid V, \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1})}{E(Y_{t+1,a^*} \mid V, \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1})}, \text{ for all } t \text{ and some small } s.$$
Adjustment for confounding causes of death

Accordingly, the expected baseline potential outcome

\[ E(\bar{Y}_{t+1+s}^{\bar{a}_{t-1}, a^*} | V = v, \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}) \]

can be expressed as

\[ E(\bar{Y}_{t+1+s}^{\bar{a}_{t-1}, a^*} | V = v, \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}) \]
\[ = r_{t+1,s}(V) \times \]

\[ E_{A_t, L_t | \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}} \left( g^{-1} \left( g(E(\bar{Y}_{t+1}^{\bar{a}_{t-1}, a_t} | V = v, \bar{L}_t = \bar{l}_t, \bar{A}_t = \bar{a}_t)) \right) \right. \]
\[ \left. - \gamma_t(v, \bar{l}_{t-1}, L_t, \bar{a}_{t-1}, A_t; \psi) \right) \]
Adjustment for confounding causes of death

- Accessing \( r_{t+1,s}(V) \) can be challenging.
- Use ratios of all-cause-mortality rates to approximate \( r_{t+1,s}(V) \). If the proportions of baseline-PM\( _{2.5} \)-mortality in all-cause-mortality (\( p_{t+1} \) and \( p_{t+1+s} \)) do not change over time, i.e \( p_{t+1} = p_{t+1+s} \), \( r_{t+1,s}(V) \) can be expressed as

\[
\frac{E(Y_{t+1+s,a^*} \mid V, \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1})}{E(Y_{t+1,a^*} \mid V, \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1})} = \frac{p_{t+1+s} M_{t+1+s}(V)}{p_{t+1} M_{t+1}(V)} = \frac{M_{t+1+s}(V)}{M_{t+1}(V)},
\]

where \( M_{t+1}(V) \) and \( M_{t+1+s}(V) \) are all-cause-mortality rates with respect to individual information at time \( t + 1 \) and \( t + 1 + s \) respectively.
Estimation Setup

- $Y_{ij,t}$: the observed outcome for individual $j$ in area $i$ at time $t$.
- $V_{ij} = (V_{ij,1}, V_{ij,2})$: represents the individual’s information (gender, age)
- $L_{i,t} = (L_{i,t,1}, L_{i,t,2})$: the meteorological information (precipitation, wind speed) in area $i$ at time $t$
- $A_{i,t}$: PM$_{2.5}$ density in area $i$ at time $t$ respectively.
- $\hat{A}_{i,t}$ estimator for marginal expectation of $A_{i,t}$
- $\hat{L}_{i,t}$: estimator for marginal expectation of $L_{i,t}$
Estimation Setup

- $\mu_t(V_{ij}, L_{i,t}, A_{i,t}; \eta)$: a parametric model for $E(Y_{ij,t+1} | V_{ij}, L_{i,t}, A_{i,t})$
- $D_{ij,t}$ could be expressed as $D_{ij,t} = (1, L'_{i,t}, A_{i,t}, V'_{ij})^T$
- $w_{ij}$: the weight that individual $j$ in area $i$ was selected into the study. Under the sequential ignorance assumption, we have

$$w_{i,t+s}^* = \frac{P(Y_{ij,t+1+s, \overline{A}_{i,t-1}, \overline{L}_{i,t-1})P(A_{i,t}, L_{i,t})}{P(Y_{ij,t+1+s, \overline{A}_{i,t-1}, \overline{L}_{i,t-1}, A_{i,t}, L_{i,t})} = \frac{P(A_{i,t}, L_{i,t})}{P(A_{i,t}, L_{i,t} | \overline{A}_{i,t-1}, \overline{L}_{i,t-1})} = w_{i,t}^*$$

(15)

for small $s$. 

Wang, He & Zhu (ISU)
Estimating Equation

- Under the aforementioned assumptions with the SNMM, and assuming $\mu_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \eta)$ is correctly specified, we could consistently estimate all related parameters by solving the estimating equations

$$\sum_i \sum_j \sum_t w_{ij} w_{i,t+s}^* (A_{i,t} - \hat{A}_{i,t}) r_{t+1,s}(V_{ij}) \left(g^{-1}\left(g(\mu_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \eta)) - \gamma_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \psi)\right)\right) = 0,$$

$$\sum_i \sum_j \sum_t w_{ij} w_{i,t+s}^* (L_{i,t} - \hat{L}_{i,t}) r_{t+1,s}(V_{ij}) \left(g^{-1}\left(g(\mu_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \eta)) - \gamma_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \psi)\right)\right) = 0,$$

$$\sum_i \sum_j w_{ij} w_{i,t}^* D_{ij,t}^T \{Y_{ij,t+1} - \mu_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \eta)\} = 0,$$

for $s = 0, 1$. 

Wang, He & Zhu (ISU)
1. Motivation

2. Structural Mean Models

3. Structural Nested Mean Models

4. PM$_{2.5}$ Case

5. Simulation Setup

6. Future Work
Estimating Equation Simplification

- \( w_{i,t+s}^* = w_{i,t}^* \), for \( s = 0, 1 \) (although notation abuse...)
- \( w_{ij} = 1 \), for all \( i \) and \( t \)
- \( \mu_t(\cdot) \) can be a GLM of its covariates, i.e.,

\[
\begin{align*}
\mu_t &= E(Y_{ij,t+1}) \\
g(\mu_t) &= \zeta_{ij,t} \\
\zeta_{ij,t} &= (V_{i,j}, \bar{L}_{i,t}, \bar{A}_{i,t}) \eta_t
\end{align*}
\]

Take the link function as logit function, which is the same form of \( g(\cdot) \).

To compute \( w_{i,t}^* \), we assume the following model:

\[
A_{i,t} = \alpha_i A_{i,t-1} + \beta_i' L_{i,t} + c_i + \epsilon_{i,t}
\]

where \( \epsilon_{i,t} \sim i.i.dN(0, \sigma^2_1) \)
A simplified version is:

\[
\sum_i \sum_j \sum_t w_{i,t}^*(A_{i,t} - \hat{A}_{i,t}) r_{t+1,s}(V_{ij}) \left( g^{-1} \left( \zeta_{ij,t} - \gamma_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \psi) \right) \right) = 0,
\]

\[
\sum_i \sum_j \sum_t w_{i,t}^*(L_{i,t} - \hat{L}_{i,t}) r_{t+1,s}(V_{ij}) \left( g^{-1} \left( \zeta_{ij,t} - \gamma_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \psi) \right) \right) = 0,
\]

\[
\sum_i \sum_j w_{i,t}^* D_{ij,t}^T \{ Y_{ij,t+1} - \mu_t(V_{ij}, \bar{L}_{i,t}, \bar{A}_{i,t}; \eta_t) \} = 0,
\]

where \( s = 0, 1 \), and

\[
g\{ E(Y_{t+1}^{\bar{a}_{t-1}, a_t} | V = v, \bar{L}_t = \bar{l}_t, \bar{A}_t = \bar{a}_t) \} - g\{ E(Y_{t+1}^{\bar{a}_{t-1}, a_*} | V = v, \bar{L}_t = \bar{l}_t, \bar{A}_t = \bar{a}_t) \} = \gamma_t^* (\bar{a}_{i,t}, \bar{l}_{i,t}, v_{ij}; \psi)
\]

\[
= (\psi_0 + \psi_1 a_{i,t} + \psi_2 l_{i,t,1} + \psi_3 l_{i,t,2} + \psi_4 v_{ij,1} + \psi_5 v_{ij,2}) a_{i,t},
\]

where \( \psi \) is a vector with 6 elements.
Even a clever housewife cannot cook a meal without rice.
The simulation data should be similar to real data.
Real data hasn’t arrived yet.
Find some data to feed ourselves.
PM$_{2.5}$ & Meteorological Data

- PM$_{2.5}$ data from EPA:
  [https://www.epa.gov/outdoor-air-quality-data](https://www.epa.gov/outdoor-air-quality-data)

- Meteorological data (wind speed, precipitation) from NOAA:
  [https://www.ncdc.noaa.gov/cdo-web/](https://www.ncdc.noaa.gov/cdo-web/)

- Raw data:
  - Area: site level data
  - Time period: daily data
  - Missingness
  - Misalignment...

- Cleaned data:
  - Area: Massachusetts county level data
  - Time period: 2009 ~ 2013 monthly data
Simulation Setup

Individual Level Data

- County level population data from USCB: [https://www.census.gov/geo/reference/centersofpop.html](https://www.census.gov/geo/reference/centersofpop.html)

- Assuming the population is same for each month during 2009 ~ 2013 within each county

- Population construction:

<table>
<thead>
<tr>
<th></th>
<th>65 ~ 74</th>
<th>75 ~ 84</th>
<th>85+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>3%</td>
<td>1.9%</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

- Gender within each group:

<table>
<thead>
<tr>
<th></th>
<th>65 ~ 74</th>
<th>75 ~ 84</th>
<th>85+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>47%</td>
<td>45%</td>
<td>43%</td>
</tr>
<tr>
<td>Female</td>
<td>53%</td>
<td>55%</td>
<td>57%</td>
</tr>
</tbody>
</table>
Individual Level Mortality Data

- Mortality data from CDC:  
  https://wonder.cdc.gov/controller/datarequest/D76
- Raw data:
  - Area: county level data
  - Time period: monthly data
  - Missingness
- Solution: fit a generalized linear model with logit link based on individual level, meteorological variables and PM$_{2.5}$ exposure density, then 'predict' each individual’s death condition based on the complete covariates we obtained previously.
Estimation

- The estimation procedure has been illustrated before
- The data we generated contains around $1.2 \times 10^7$ cases
- $\hat{\psi} = (1.214, 2.044, -0.159, 2.650, 0.100, 0.800)$
- Recall

$$g\{E(Y_{t+1}^{\bar{a}_{t-1},a_t} | V = v, \bar{L}_t = \bar{l}_t, \bar{A}_t = \bar{a}_t)\} - g\{E(Y_{t+1}^{\bar{a}_{t-1},a^*_t} | V = v, \bar{L}_t = \bar{l}_t, \bar{A}_t = \bar{a}_t)\} = \gamma^*(\bar{a}_i,t, \bar{l}_i,t, v_{ij}; \psi)$$

$$= \psi_0 + \psi_1 a_i,t + \psi_2 l_i,t,1 + \psi_3 l_i,t,2 + \psi_4 v_{ij},1 + \psi_5 v_{ij},2) a_i,t,$$
1 Motivation

2 Structural Mean Models

3 Structural Nested Mean Models

4 PM_{2.5} Case

5 Simulation Setup

6 Future Work
Variance Estimation

- The estimating equations above could be denoted as
  \[ U(\psi) = \sum_i \sum_j \sum_t U_{ijt}(\psi) = 0. \]
- \( \nabla U(\psi) \): the gradient of \( U(\psi) \) with respect to \( \psi \)
- The estimated variance of \( \hat{\psi} \) is obtained as
  \[
  \hat{\text{Var}}(\hat{\psi}) = \left( \nabla U(\hat{\psi}) \right)^{-1} V(U(\hat{\psi})) \left( \nabla U(\hat{\psi})^T \right)^{-1}.
  \]
RR and RD

- Using the delta method, we can also obtain estimators and corresponding variance estimators of RD and RR in previous slides respectively.
- Develop hypothesis tests to examine the significance of RD and RR based on large sample theory.
Dose-Response Curves

With a specified spatial model on PM$_{2.5}$ density and meteorological information, we can also compute and plot a dose-response curve of individuals given covariates $V$ by estimating the expected potential outcome

$$E(Y_{t+1}^a|V) = E \left\{ E(Y_{t+1}^a|V, L_t, A_t)|V \right\} = E_{A_t, L_t|V} E(Y_{t+1}^a|V, L_t, A_t).$$ (18)

Under the additional independence assumption the above equation can be simplified as

$$E_{A_t, L_t|V} E(Y_{t+1}^a|V, L_t, A_t) = E_{A_t, L_t} E(Y_{t+1}^a|V, L_t, A_t),$$ (19)

which can be computed based on our model.
We can even try cases with measurement error:

\[
\begin{align*}
L & \rightarrow A \\
A & \rightarrow A^* \\
A^* & \rightarrow D \\
V & \rightarrow D \\
C & \rightarrow D \\
D & \rightarrow Y
\end{align*}
\]

where \( A^* \) is individual-level PM\(_{2.5} \) exposure density.
References


