Measurement error effects on bias and variance in two-stage regression, with application to air pollution epidemiology

Chris Paciorek

Department of Statistics, University of California, Berkeley

and

Adam Szpiro

Department of Biostatistics, University of Washington

Funded by NIEHS and US EPA.

July 29, 2012

Environmental exposure and measurement error

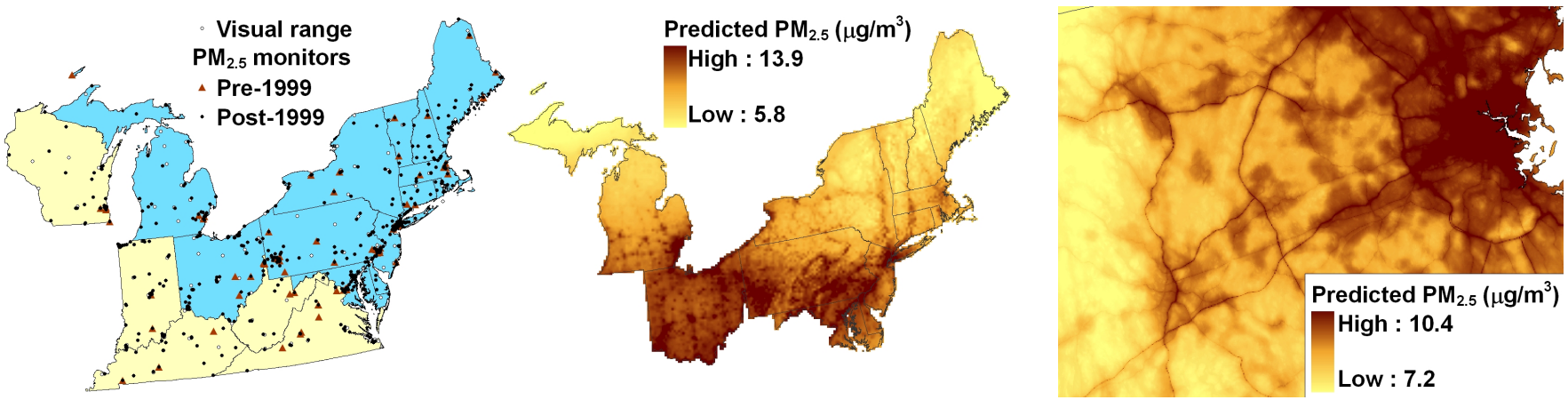
- Consider a health analysis that focuses on the association of exposure, X, with a health outcome, Y:

$$Y = X\beta_x + Z\beta_z + \epsilon$$

- In environmental, occupational, and other contexts, X is not known with certainty and often not even measured directly.
- Some strategies for estimating exposure:
  - Central site measurements
  - Spatial prediction
  - Exposure regression based on various covariates
  - Deterministic modeling
  - Remote sensing proxies

Spatial modeling example

Exposure data locations (left) and $PM_{2.5}$ predictions (northeast US (center) and greater Boston (right))



[Yanosky et al. (2009), Environmental Health Perspectives]

Types of measurement error

Traditional measurement error types for observed exposure:

- Berkson error: unmeasured variability in true exposure
- Classical error: noise in the observed exposure

Extension to correlated, heteroscedastic errors in the context of modeled exposure (Szpiro et al. 2011, Biostatistics):

- Berkson-like error: missing components of true exposure
- Classical-like error: noise in estimating exposure

Implications for $\hat{\beta}_x$:

1. Berkson and Berkson-like error increase variance but do not induce bias (in a linear model).
    - Caveat: Berkson-like error can cause bias in some circumstances.
2. Classical and classical-like error induce bias and affect variance.

Probabilistic framework

What is random in this context (exposure data => exposure model => exposure predictions => health model)?

1. Random instrument error?
2. Random exposure surfaces in space-time (e.g., due to random weather)?
3. Random societal structure (i.e., random sources of pollution)?
4. Random exposure data locations (monitor placement)?
5. Random health outcomes conditioning on individuals in study?
6. Random sampling of individuals in a study?

Our proposal

- Treat exposure surfaces as fixed and monitor placement (exposure data locations) as random.
  - Exposure is in principle predictable, but not so in practice
  - Long-term average air pollution can be viewed as deterministic:
    - Frequentist interpretation: "how would the results have changed if I had measured the system differently?", not "how would they have changed if the spatial pattern of air pollution were different?".
- Statistical implications:
  - "Random X" regression/spatial modeling.
  - Nonparametric bootstrap (resample monitor locations and associated observations) follows naturally
  - Don't assume a true exposure model; this produces a new source of measurement error

"Random X" regression

White (1980; Econometrica, over 12,000 citations!) shows that "random X" regression gives consistent estimation and $\hat{\beta}$ is asymptotically normal.

- Random X regression is not unbiased in finite samples.
- $E(\hat{\beta})$ may not exist in finite samples.
- Sandwich estimation of Var($\hat{\beta}$).

Note that we'll use this framework in our analysis of the exposure model, so the "Random X" is the exposure covariates, not the exposure in the health model.

A basic two-stage model

Basic health model:

$$Y = X\beta_x + Z\beta_z + \epsilon$$

Exposure decomposition (data generating model):

$$X(s) = \phi(s) + \eta; \quad \eta \sim \mathcal{N}(0, \sigma_\eta^2)$$

Let $R(s)$ (our 'Random X') be a set of exposure covariates and spatial basis functions. We DEFINE γ as the projection of $\phi(s)$ onto $R(s)$ with respect to the spatial distribution of health study participants, $G(s)$:

$$\gamma = \text{argmin}_\xi \int (\phi(s) - R(s)\xi)^2 dG(s)$$

This gives us the following exposure model:

$$X(s) = R(s)^\top \gamma + U_{\text{BL}}(s) + \eta$$

Measurement error decomposition

- We estimate γ with $\hat{\gamma}$ by OLS regression using the exposure data, assuming exposure locations come from the spatial distribution, $G(s)$.
- Building on Szpiro et al. (2011), we have the following decomposition of exposure error:

$$\begin{aligned} U(s) &= X(s) - R(s)\hat{\gamma} \\ &= \underbrace{X(s) - \phi(s)}_{\text{Berkson}} + \underbrace{\phi(s) - R(s)\gamma}_{\text{Berkson-like}} + \underbrace{R(s)\gamma - R(s)\hat{\gamma}}_{\text{Classical-like}} \end{aligned}$$

Berkson-like error

- $U_{BL}(s) = \phi(s) - R(s)\gamma$ is the difference between the potentially predictable variation in exposure and the projection of that variation onto the chosen basis functions.
  - This difference is heteroscedastic.
- Does the Berkson-like error cause bias in estimating β_x?
  - If we knew γ, we could use White (1980) to show that this error does not induce bias in $\hat{\beta}_x$, because U_{BL} is orthogonal to our 'estimated' exposure, $R(s)\gamma$).
- However, there are some complications...

Compatibility conditions to avoid bias from Berkson-like error

1. Design your study such that $G(s) = H(s)$: locations/covariates of people and exposure data should 'match'.
   - γ is based on $G(s)$, the distribution of study participants.
   - $\hat{\gamma}$ estimates $\gamma^* = \text{argmin}_\xi \int (\phi(s) - R(s)\xi)^2 dH(s)$, which is based on $H(s)$, the distribution of exposure data locations.
   - If $G(s) \neq H(s)$, then $\gamma^* \neq \gamma$, which induces bias.
2. Include spatially-structured components of the health confounders, Z, in the exposure model.
   - Why? This ensures that the Berkson-like error term, $U_{\text{BL}}(s)$, is orthogonal to all the terms in the health model.
   - This is similar to the need to include covariates in regression calibration.

Classical-like error

- $U_{\text{CL}}(s) = R(s)\gamma - R(s)\hat{\gamma}$ is the contribution of exposure model estimation error to measurement error in the health model.
    - This difference is heteroscedastic and correlated.
- This error could induce severe bias if our exposure predictions, $R(s)\hat{\gamma}$, are very noisy.
- We assess the impacts of classical-like error based on a Taylor series approximation for $\hat{\beta}_x$ as a function of $\hat{\gamma} - \gamma$.

Approximate bias and variance

- Define $w(s) = R(s)\gamma$ and $\hat{w}(s) = R(s)\hat{\gamma}$ as the predictable exposure and its estimator.
- Focus on asymptotics w.r.t. the number of exposure observations.
- Consistency based on first-order Taylor expansion:

$$\hat{\beta}_x \xrightarrow{\mathrm{d}} \mathcal{N}\left(\beta_x, \beta_x^2 \frac{\int w(s_1)w(s_2)\mathrm{Cov}(\hat{w}(s_1), \hat{w}(s_2))dG(s_1)dG(s_2)}{(\int (w(s)^\top w(s))^2 dG(s))^2}\right).$$

- Relative bias based on second-order Taylor expansion:

$$-\frac{\int w(s)E(\hat{w}(s) - w(s))dG(s)}{\int (w(s)^\top w(s))^2 dG(s)} - \frac{\int \mathrm{Var}(\hat{w}(s))dG(s)}{\int (w(s)^\top w(s))^2 dG(s)} + \cdot$$
$$2\frac{\int w(s_1)w(s_2)\mathrm{Cov}(\hat{w}(s_1), \hat{w}(s_2))dG(s_1)dG(s_2)}{(\int (w(s)^\top w(s))^2 dG(s))^2}$$

where the first term involves the bias of $\hat{\gamma}$ (which occurs because we are in the 'random X' setting).

Towards a practical strategy

To minimize bias and account for uncertainty in $\hat{\beta}_x$ induced by exposure error, we suggest:

1. Try to have the distribution of exposure observations (in geographic space and covariate space) roughly match the distribution of health participants.
    - This will minimize bias from the Berkson-like error.
2. Try to avoid overly-parameterized exposure models to minimize bias from classical-like error from the exposure estimation (see Szpiro talk tomorrow)
    1. (Optionally) Correct for the bias using our asymptotically-derived bias estimator.
3. Use the nonparametric bootstrap (resampling exposure observations and health observations) to estimate $\text{Var}(\hat{\beta}_x)$.
    - Note that the nonparametric bootstrap is fully consistent with our probabilistic framework.

Simulation Results

- Data: 50000 health observations, 250 exposure observations, exposure variation based on fitted models from previous NHS work
- Exposure model: land-use covariates plus 25 spatial basis functions
- Health model: logistic regression

	Relative bias % (sim. s.e.)		
	oracle (B-L error only)	no bias corr'n	with bias corr'n
Full exposure model	2.6% (1.0)	-0.8 % (1.0)	2.1% (1.0)
Small-scale spatial var'n + covariates	2.8% (1.7)	-4.1% (1.6)	1.9% (1.7)
Small-scale spatial variation only	11.8% (3.0)	-10.0% (2.8)	13.0% (3.7)
	Coverage %		
	oracle (B-L error only)	no bias corr'n (boot)	with bias corr'n (boot)
Full exposure model	95.7%	94.8%	95.2%
Small-scale spatial var'n + covariates	95.5%	95.7%	96.1%
Small-scale spatial variation only	95.2%	96.9%	96.9%

Conclusions/Implications

- More work remains to wrestle with impact of nonlinear health models.
- Given our results for classical-like error, we hypothesize that health studies based on limited exposure data may be seriously biased: bias can be quantified as the ratio of uncertainty in exposure predictions to true variation in exposure.
- Exposure and health stages should be considered jointly to better understand and minimize the measurement error impact.
- If one uses a deterministic model to predict exposure, we are in trouble in terms of quantifying the measurement error implications.

Conclusions/Implications (f

- In pollution studies that examine multiple exposures (e.g., pollutants), these issues will be particularly important, as the amount of measurement error for the different pollutants is likely to differ.
  - This framework can help to understand the effects of measurement error in that context: missing components of variability in one exposure can play the role of unmeasured confounders!
  - See talk by Adam Szpiro in Session 230, tomorrow at 2 pm (which includes other interesting and related measurement error talks).