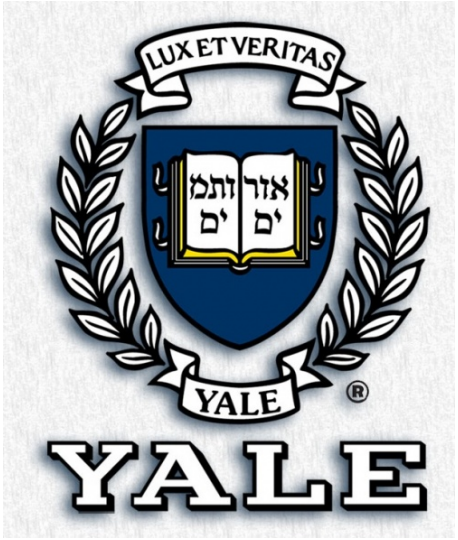
Accounting for measurement uncertainty in environmental preterm studies

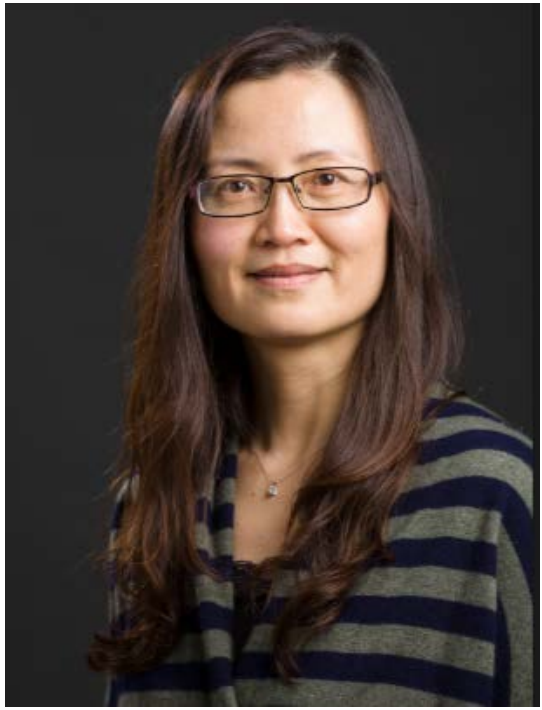
Shuangge (Steven) Ma

Department of Biostatistics, Yale University



How it started

- An environmental epidemiologic study conducted in Lanzhou City, which belongs to the west region of China, sits on the yellow river, and is (unfortunately) highly polluted.
- PI: Dr. Yawei Zhang, Department of Environmental Health Sciences, Yale School of Public Health.
- The scientific question of interest: is air pollution associated with preterm birth and birth defects?
- Evidences in the literature are conflicting.



Environment International 76 (2015) 71–77

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Data collection

Ambient air pollutant PM_{10} and risk of preterm birth in Lanzhou, China

Nan Zhao [a,1], Jie Qiu [b,1], Yaqun Zhang [c,1], Xiaochun He [b], Min Zhou [b], Min Li [d], Xiaoying Xu [b], Hongmei Cui [b], Ling Lv [b], Xiaojuan Lin [b], Chong Zhang [b], Honghong Zhang [b], Ruifeng Xu [b], Daling Zhu [b], Ru Lin [b], Tingting Yao [b], Jie Su [b], Yun Dang [b], Xudong Han [b], Hanru Zhang [b], Haiya Bai [b], Ya Chen [b], Zhongfeng Tang [b], Wendi Wang [b], Yueyuan Wang [b], Xiaohui Liu [b], Bin Ma [b], Sufen Liu [b], Weitao Qiu [b], Huang Huang [a], Jiaxin Liang [a], Qiong Chen [e], Min Jiang [f], Shuangge Ma [a], Lan Jin [g], Theodore Holford [a], Brian Leaderer [a], Michelle L. Bell [g], Qing Liu [b,*], Yawei Zhang [a,**]

- For details, refer to Zhao et al. (2015, Environmental International) and references therein.
- Briefly, the study was based on the Gansu Provincial Maternity & Child Care Hospital (arguably the largest and best in Gansu).
- Samples were collected between 2010 and 2012. Follow-up is still ongoing.
- A total of 10,542 pregnant women participated in the study, with a participation rate of 73.4%.
- Exploratory analysis did not suggest any obvious selection bias.

Data collection

Table 2
Distributions of selected characteristics between cases and control.

Characteristics	Cases (n = 677)	Controls (n = 8292)	P value
	N (%)	N (%)	
Maternal age (years)			
<30	394 (58.2)	5300 (63.9)	.003
≥30	283 (41.8)	2992 (36.1)	
Highest education level			
<College	356 (52.6)	2987 (36.0)	<.001
≥College	321 (47.4)	5305 (64.0)	
Family monthly income (RMB per capita)			
<3000	456 (67.4)	4705 (56.7)	<.001
≥3000	221 (32.6)	3587 (43.3)	
Employment during pregnancy			
No	368 (54.4)	3786 (45.7)	<.001
Yes	309 (45.6)	4506 (54.3)	
Pre-pregnancy BMI			
≤18.5	133 (19.7)	1708 (20.6)	.006
18.5–24.0	448 (66.2)	5502 (69.1)	
≥24.0	96 (14.2)	851 (10.3)	
Smoking during pregnancy			
No	532 (78.6)	6732 (81.2)	.10
Yes	145 (21.4)	1560 (18.8)	
Season of conception			
Fall	176 (26.0)	2280 (27.5)	.29
Winter	169 (25.0)	1882 (22.7)	
Spring	151 (22.3)	1723 (20.8)	
Summer	181 (26.7)	2407 (29.0)	
Pre-history of preterm			
No	644 (95.1)	8267 (99.7)	<.001
Yes	33 (4.9)	25 (0.3)	
Parity			
Primiparous	452 (66.8)	6282 (75.8)	<.001
Multiparous	225 (33.2)	2010 (24.2)	
C-section			
No	366 (54.1)	5307 (64.0)	<.001
Yes	311 (46.0)	2985 (36.0)	
Preeclampsia			
No	612 (90.4)	8152 (98.3)	<.001
Yes	65 (9.6)	140 (1.7)	
Cooking fuel			
Gas or electricity	538 (79.5)	7173 (86.5)	<.001
Biomass or coal	41 (6.1)	196 (2.4)	
Others	98 (14.5)	923 (11.1)	

Statistical analysis in Zhao et al. (2015)

- Analysis was conducted by Ms. Nan Zhao (a PhD student in EHS supervised by Zhang and Ma) and S. Ma.

- Outcome: PB (preterm birth) is defined as delivery prior to 37 completed weeks of gestation.
- Covariates: PM10 (a major measure of air pollution level) + a few confounder (as suggested in the literature).

- Analysis: summary statistics + logistic regressions.
- Papers published! Nan's dissertation approved! So what may be wrong?

A closer look at the outcome variable

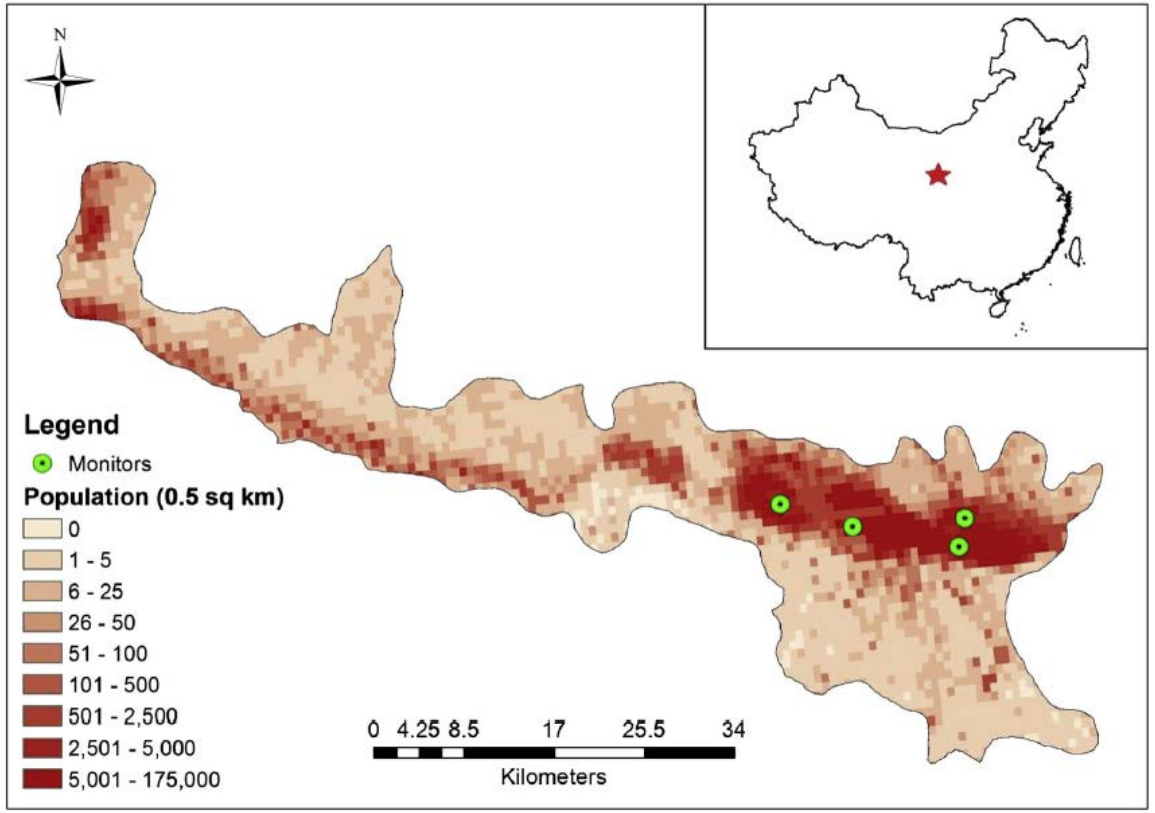
- Dichotomization is not efficient. 37.1 and 36.9 weeks are not that different. [This is a very old problem.]

- A more unique problem: data on gestation days were collected using questionnaires (recall) and are not expected to be 100% accurate.

- A conversation with the PI and an OBGYN doctor:
  "if the questionnaire says 255 days, what is the real date?"
  " … do not know. Perhaps somewhere between 248 and 262 days."
  " … all 15 days equally likely?"
  "perhaps not."

A closer look at PM10

- The daily PM10 level is calculated as a weighted average of home and work levels.
- Sensors are very expensive.
- What was done: four sensors for the whole city. The PM10 level for a specific location is calculated as the weighted average of the four sensor values, with weights inversely proportional to the distances to the sensors.



Similar imputation has been done in the literature.

But, is there any problem?

Our strategy for the outcome variable

- Instead of analyzing the dichotomized PB status, we will analyze the continuous gestation days.

- As we do not have full confidence in the observed values, for each observation, we will create an interval – we are highly confident that the real gestation days will fall in this interval.
- The creation of interval censored data has been previously considered in some survival studies.
- Some characteristics of our interval censoring are different from the "standard" ones.

Our strategy for PM10

Consider a location without a sensor:

- Without additional data, we can only assume that the true PM10 and computed (imputed) values are not systematically different.
- Variation:
  - Our conceptual model:

$$True\ PM10 - imputed\ PM10 = random\ error$$

  - This random error has been neglected in the existing analysis.
- Our strategy: bring it back.

Methods

Consider a semiparametric model

$$Y = m(\beta' X) + \varepsilon,$$

where

- Y is the gestation days, and X contains PM10 and all confounders;
- $m(.)$ is monotone increasing but otherwise unspecified;
- $||\beta||$=1 for identifiability;
- the dimension of X is larger than one. There is at least one continuous component;
- the random error ε satisfies “standard” moment conditions.

An iterative estimation procedure

1. For each observation not on a sensor spot, get a “plausible guess” of the PM10 value (D1);
2. Initialize the estimate of β;
3. For each observation, get a “plausible guess” of the value of Y (D2);
4. With the current estimate of β and current values of Y, obtain the nonparametric estimate of $m(.)$ (D3);
5. With the current estimate of $m(.)$ and current values of Y, obtain the estimate of β (D4);
6. Repeat Steps 3-5 until convergence. The value of β at convergence provides one estimate of the regression coefficient vector.
7. Repeat Steps 1-6 multiple times, and aggregate the β estimates.

The overall iteration, D1, and D2 share a similar flavor with multiple imputation.

- Without knowing the true values of Y and PM10, we treat them as missing, for which imputation is a simple and effective tool.

D3 is a constrained nonparametric estimation.

D4 is straightforward.

Get a plausible guess of PM10 value (D1)

For a location without a sensor, compute the PM10 value as the weighted average of the four sensor measurements. The weights are inversely proportional to the distances.

Our conceptual model:

$$True\ PM10 = computed\ PM10 + random\ error$$

Make the simple assumption that the random error has a normal distribution with mean zero. (May need a truncation to avoid negative values).

We then only need to be concerned with the variance:

- The intrinsic variation associated with the sensor, which can be obtained from the manufacturer.
- The variation associated with imputation, which can be obtained by comparing the observed value of one sensor against that computed using the other three sensors.

Assume that the two components are additive.

Get a plausible guess of the value of Y (D2)

For the observed value Y, denote the imputed value as $\tilde{Y}$.

- $\tilde{Y} \in [Y - c, Y + c]$. That is, the observed value is not "too off". The value of c is determined based on expert opinions (prefixed). In our analysis, we take a conservative value with $c = 10$.

- $\tilde{Y} \sim N(m(\beta' X), \sigma^2)$. That is, the regression model determines the most likely value. σ^2 is also determined based on expert opinions (prefixed).

Estimate the function $m(.)$ (D3)

[*Slides copied from those prepared by Ms. Yinjun Zhao.*]

- Let $U \sim unif[0,1]$ and $\xi = m(U)$, then

$$F_\xi(t) = Pr\{\xi \leq t\} = Pr\{m(U) \leq t\} = Pr\{U \leq m^{-1}(t)\} = m^{-1}(t) \quad (1)$$

- On the other hand, the CDF of ξ can be written as

$$F_\xi(t) = \int_{-\infty}^{t} f_\xi(s)ds = \int_{-\infty}^{t} \frac{1}{Nh_d} \sum_{i=1}^{N} k_d \left[\frac{\xi_i - s}{h_d}\right] ds \qquad (2)$$

where ξ_i are observations of ξ, $i = 1, 2, \cdots, N$.

- By (1) and (2)

$$m^{-1}(t) = \int_{-\infty}^{t} \frac{1}{Nh_d} \sum_{i=1}^{N} k_d \left[\frac{m(U_i) - s}{h_d} \right] ds \tag{3}$$

where $U_i = i/N$ are observations of U, $i = 1, 2, \cdots, N$.

- Given the sample $\{(z_i, Y_i)\}$, $z_i \in [0, 1]$, $i = 1, 2, \cdots, N$, and $Y = m(z)$,

$$\hat{m}^{-1}(t) = \int_{-\infty}^{t} \frac{1}{Nh_d} \sum_{i=1}^{N} k_d \left[\frac{\hat{m}(i/N) - s}{h_d} \right] ds \tag{4}$$

where $\hat{m}(i/N)$ is classical Nadaraya-Watson estimator, i.e.

$$\hat{m}(i/N) = \left\{ \sum_{j=1}^{N} k_r \left[\frac{z_j - (i/N)}{h_r} \right] Y_i \right\} / \left\{ \sum_{j=1}^{N} k_r \left[\frac{z_j - (i/N)}{h_r} \right] \right\} \tag{5}$$

Estimate β (D4)

As the estimate of $m^{-1}()$ is available, this is a simple least squares problem.

Inference for β

Two considerations in computing variance:

- When there is no measurement uncertainty in X and Y, this is a semiparametric estimation problem. We propose inference using the weighted bootstrap (which adds random exp(1) weights to observations).
- Multiple imputations are conducted. We adopt the existing formula for combining multiple estimates and generating variance estimation.

Computational considerations

Computational cost: affordable. A few minutes on a regular laptop.

Convergence:

- observed in all simulations, within a small number of iterations.
- has not been rigorously proved. Convergence with imputation-based approaches is not an easy problem. Further complication is brought by the nonparametric estimation.

Simulation

We have conducted 20+ sets of simulations.

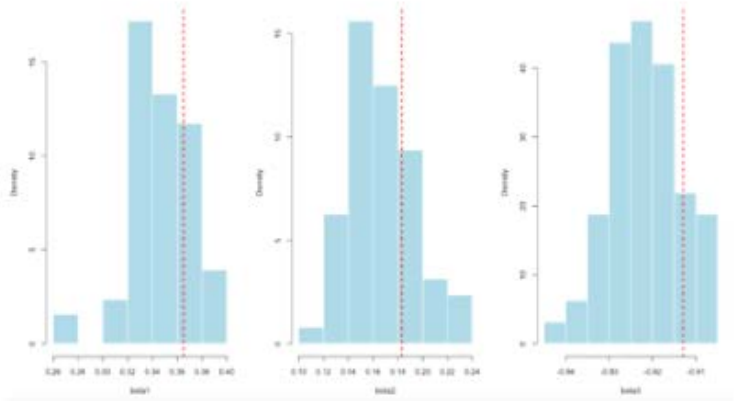
One example:

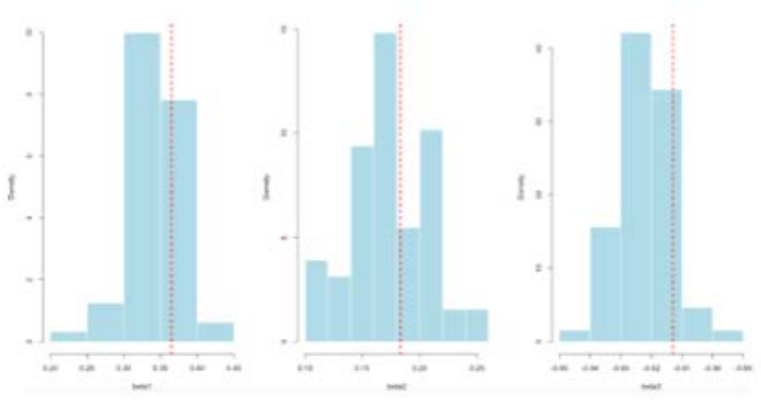
- $m(.)$ is the logit function (multiplied by a constant)
- Three correlated covariates: two normal and one uniform.
- The observed X_1 and Y are equal to the true values plus random errors and then truncated (little impact).
- Sample size=1600; Data-dependent window selection in nonparametric estimation.

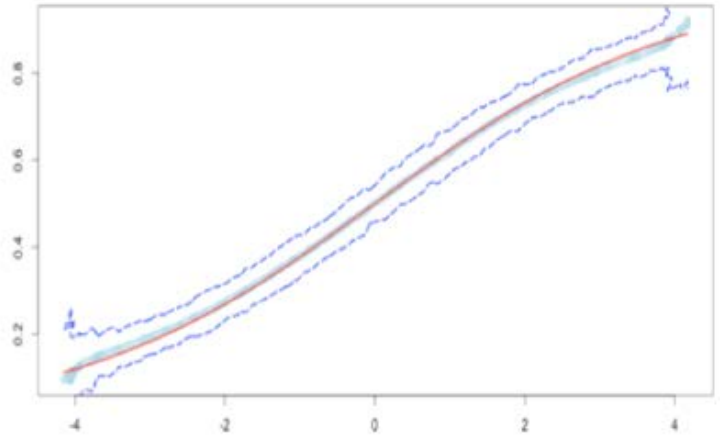
Results and comparison with the single index model (SIM)

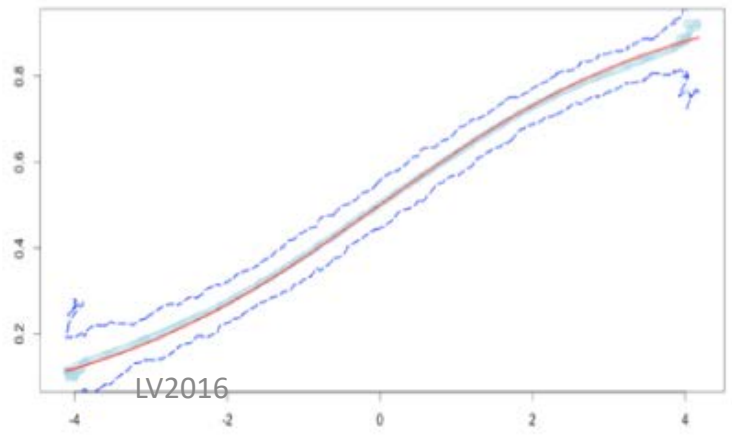
My Model	β_1	β_2	β_3
mean	0.3471	0.1681	-0.9219
sd	0.0242	0.0261	0.0079
root mse	0.0299	0.0298	0.0119

SIM	β_1	β_2	β_3
mean	0.3437	0.1735	-0.9218
sd	0.0307	0.0323	0.009
root mse	0.0371	0.0334	0.0126









Data analysis

	pm10	mage	parity	season	passivesmoke
Estimate	-0.715	-0.15	-0.146	0.162	-0.046
CI	(-0.787,-0.642)	(-0.168,-0.123)	(-0.278,-0.008)	(-0.01,0.199)	(-0.161,0.159)
	college	**income**	**cook**	**tempreture**	
Estimate	0.167	-0.285	-0.424	-0.357	
CI	(0.015,0.306)	(-0.406,-0.209)	(-0.518,-0.149)	(-0.4,-0.335)	

Main finding: a significantly negative relationship between PM10 and gestation days.

Logistic regression leads to the same qualitative conclusion. But the quantitative findings (relative contributions of covariates) are different.

Remarks

- Measurement uncertainty (as we have considered) is not rare in epidemiologic studies. However, little attention has been paid to this problem.
- The proposed approach is easy to implement. A tradeoff is that theoretical investigation is difficult.
- There are a few model and parameter assumptions, which can be potentially relaxed.

Acknowledgements

Thanks so much to conference organizers and all of you!