Keio University

The Problem of Treating Imputed Data as Observed Data When We Estimate the Effect of Exposure to Particulate Matter

Tomoshige Nakamura*

Mihoko Minami

BOSTON UNIVERSITY/KEIO UNIVERSITY WORKSHOP 2016
Probability and Statistics

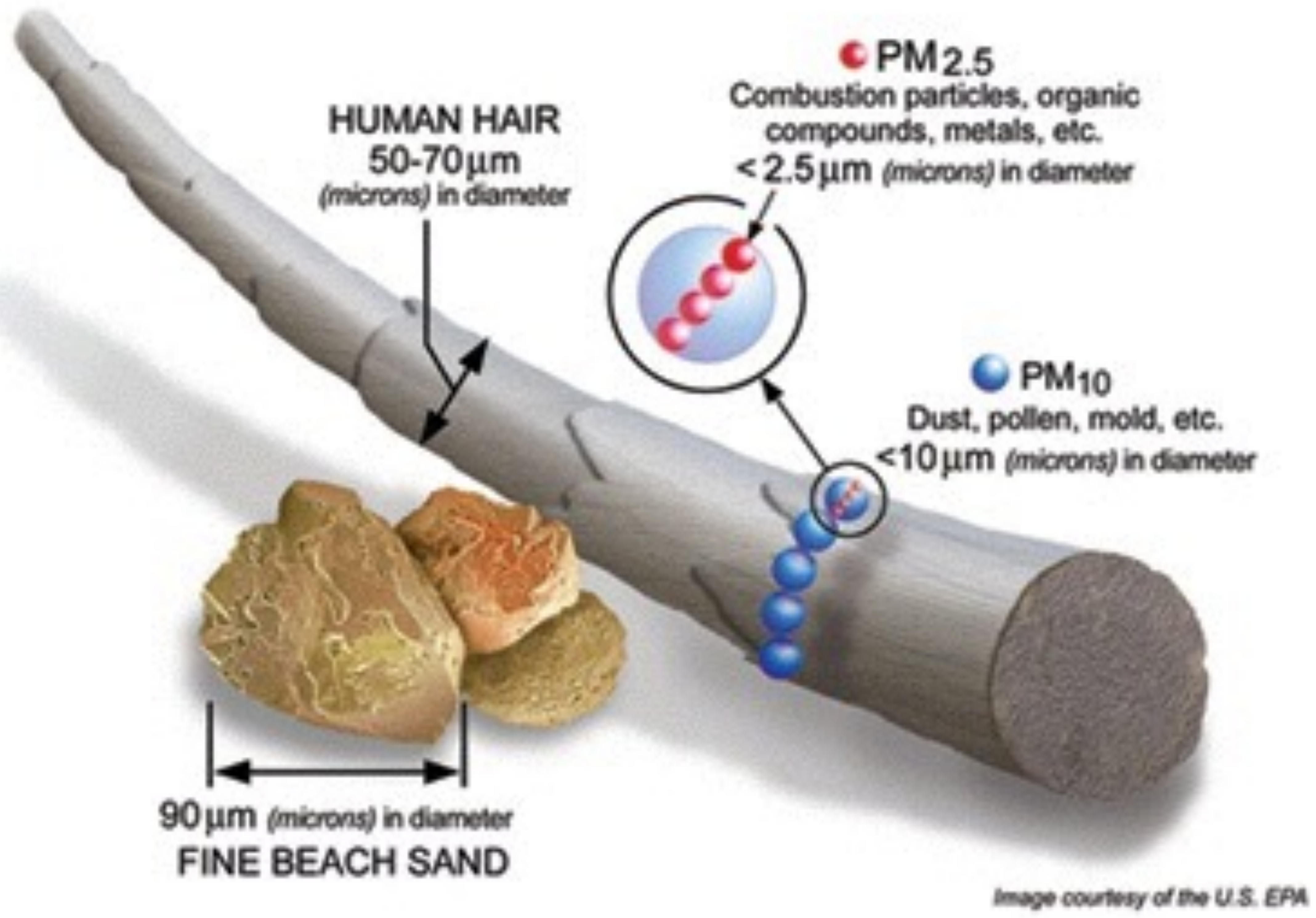
Boston University — August 15-19, 2016

About This Research

- We consider the problem of estimating the effect of Particulate Matters to our health, when **<u>the number of monitoring stations is limited</u>**

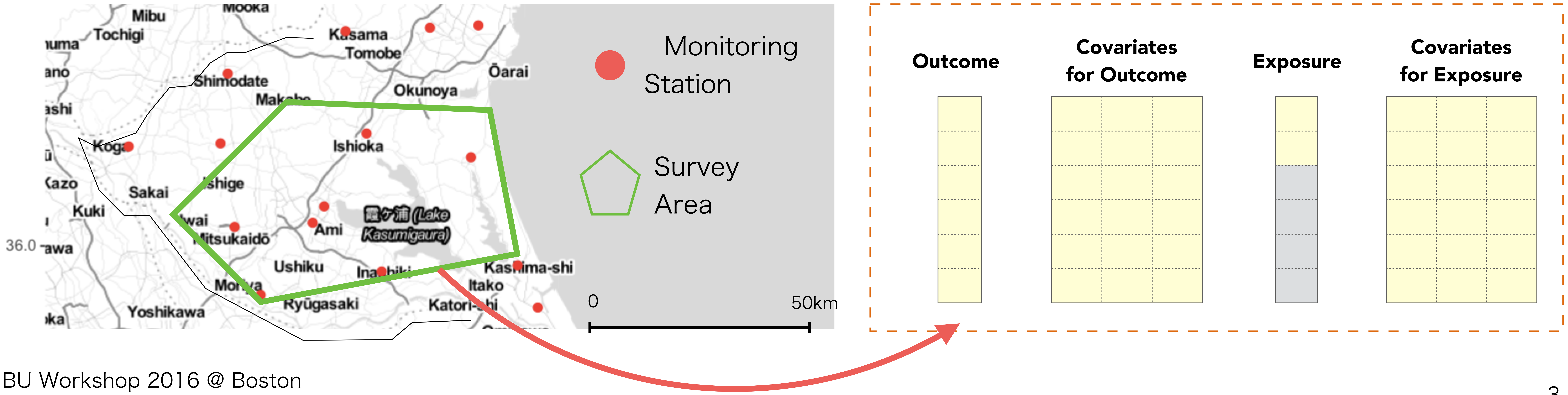
<u>Particulate Matter</u>

- Particulate Matter is Complex mixture of extremely small particles and liquid droplets, and are widely studied and concerned the relationship with various diseases.
- Estimating the effect of Particulate Matter exposure to our health is one of the active research topic of environmental epidemiology.



About This Research

- When we estimate the effect of exposure, we often use the community health survey data, and need the information of exposure at survey areas.
- Left figure shows the survey conducted at some area of Japan, we can see the number of monitoring stations in survey areas are limited.



The aim of this research

- In many of researches for particulate matter exposure, **missing data problem are not paid enough attention**, and simple regression model are used to fill the missing of exposure
- Then, the effect of exposure is estimated as if they were observed

Notation!

- The another reason why regression imputation is used is that "the analyst for exposure assessment and the analyst for estimating the effect exposure are not same one".
- So, the unified bayesian approach is not the realistic choice to analysis.

Procedure using for estimating the effect of exposure in common research.

- The procedure of estimating the effect of exposure using regression imputation can be decompose to 3 steps.

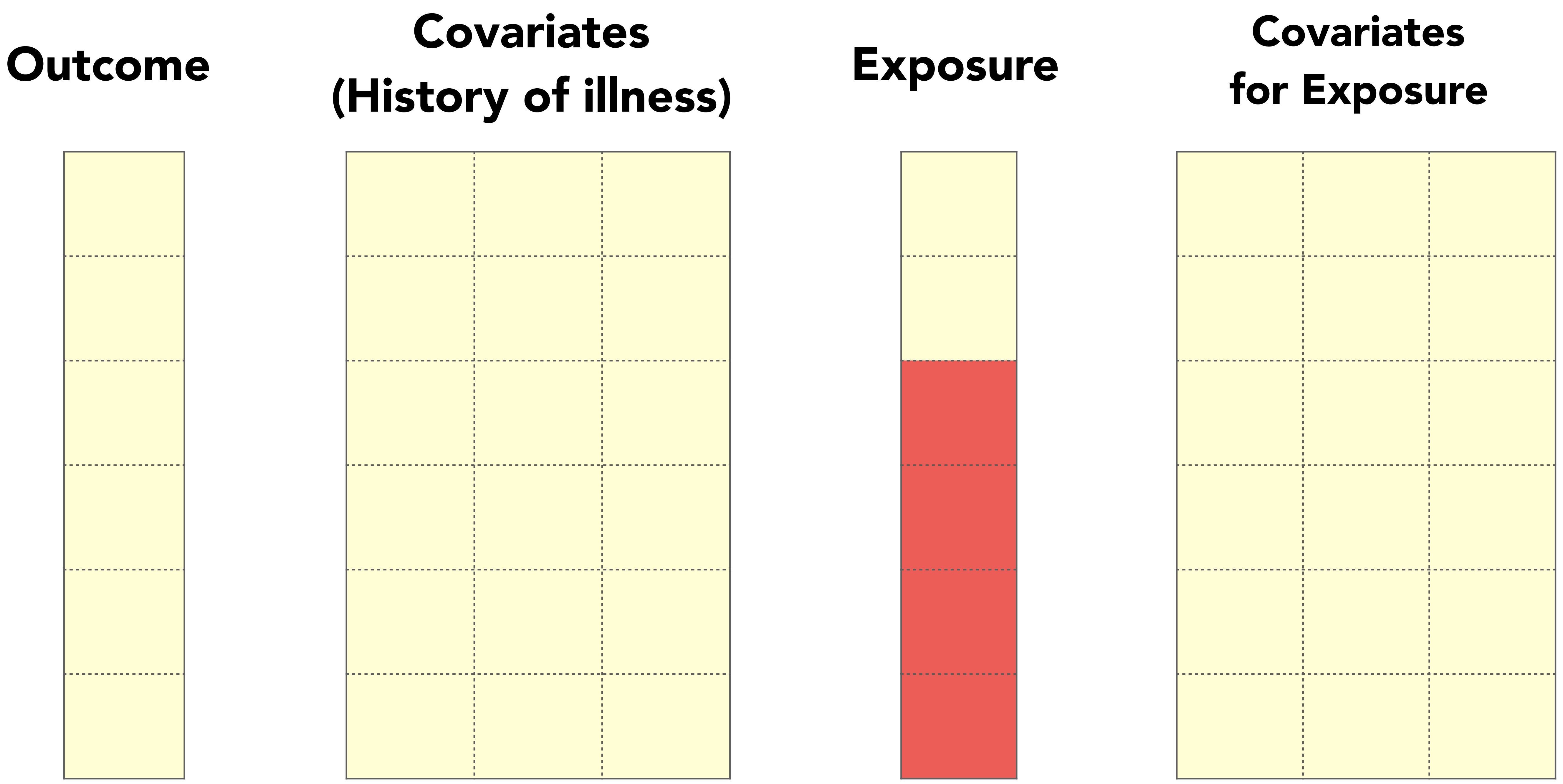
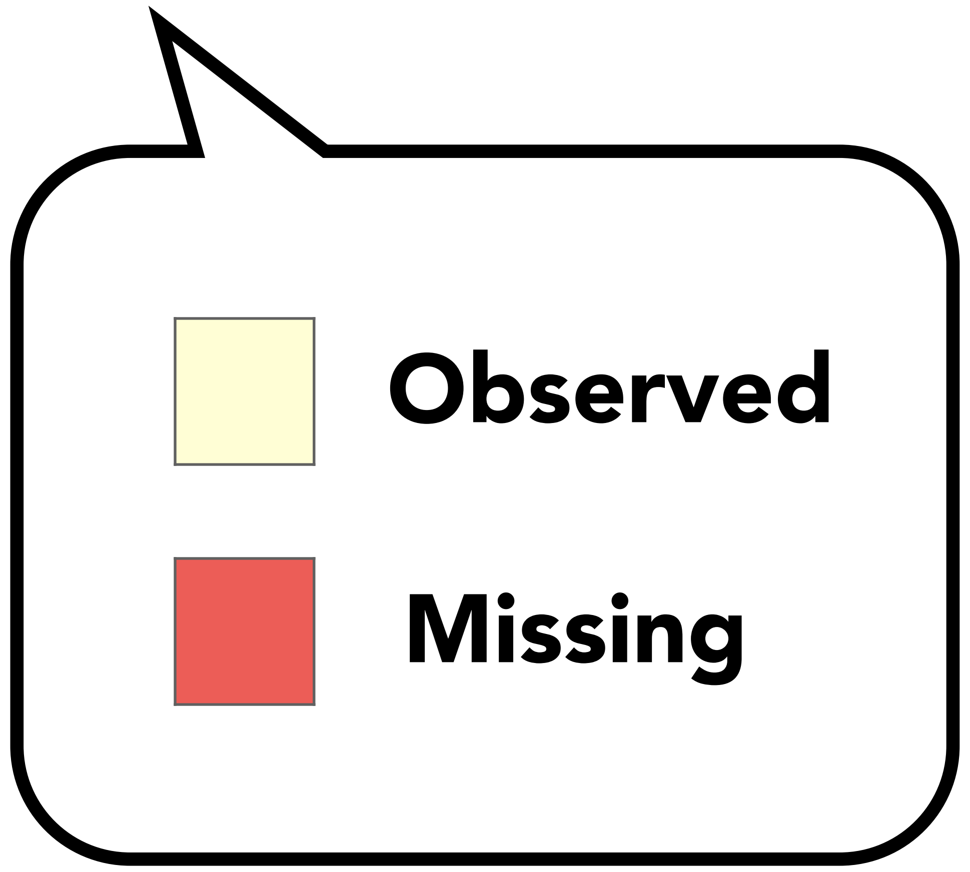


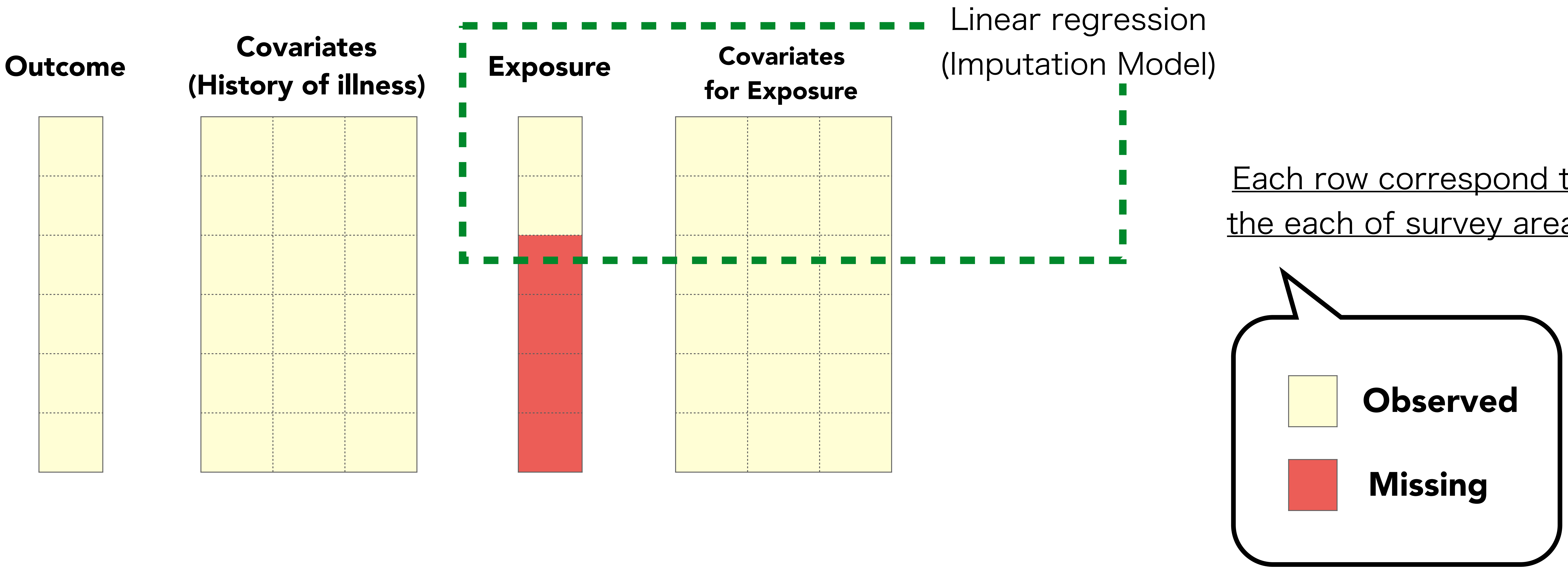
Fig: dataset obtained from the survey.

<u>Each row correspond to the each of survey areas</u>



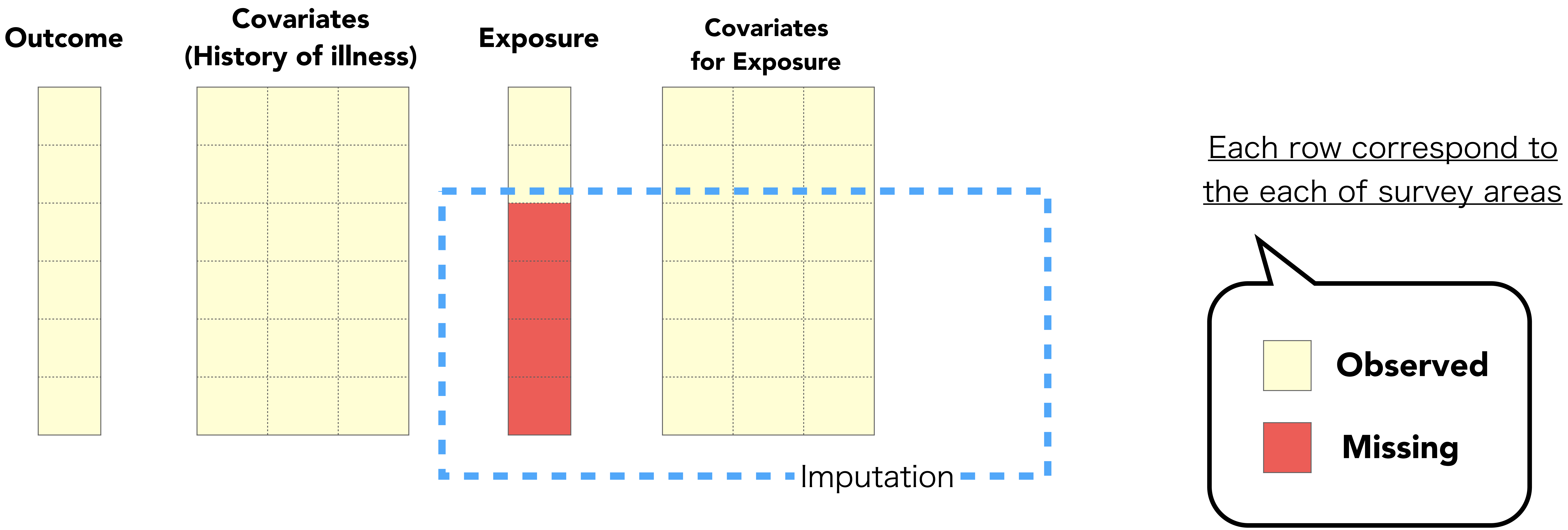
Procedure using for estimating the effect of exposure in common research.

- Step1 (constructiong the model for exposure) : fitting linear regression model to the data observed at monitoring stations, and construct the prediction model for exposure.



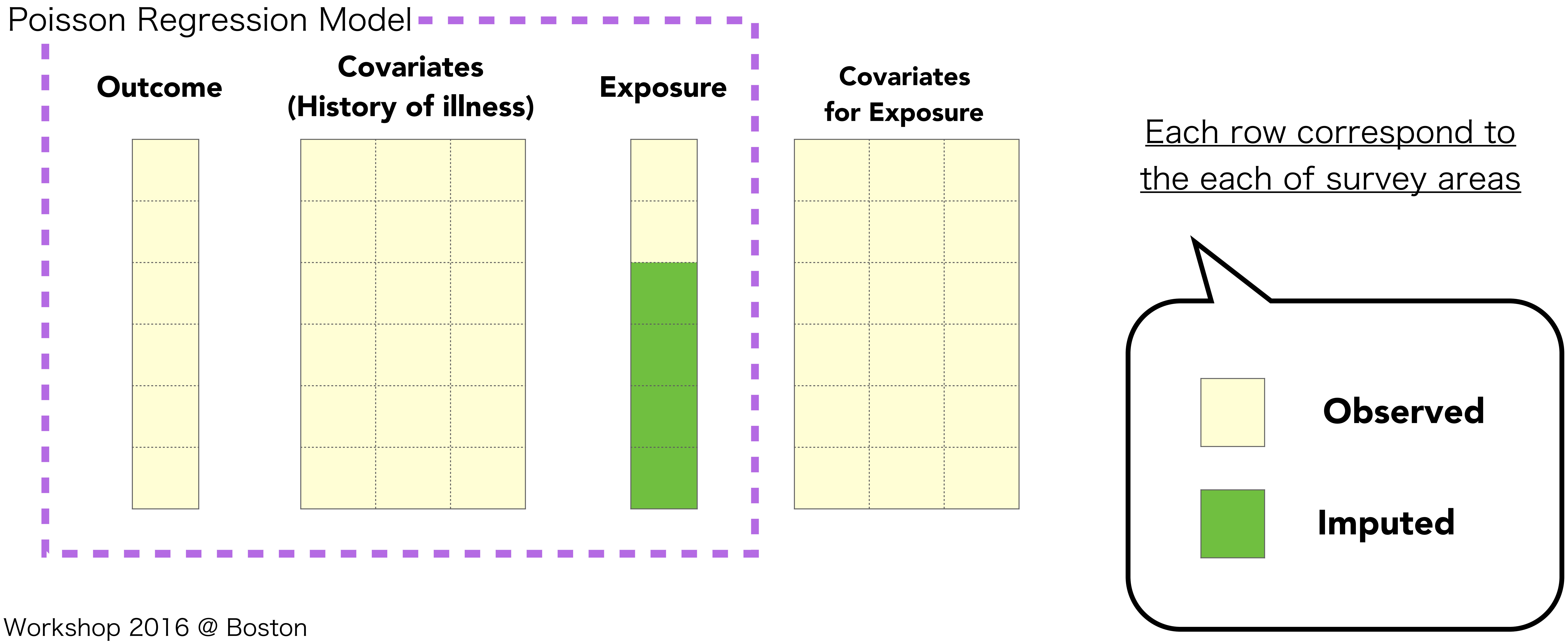
Procedure using for estimating the effect of exposure in common research.

- Step2 (filling the missing) : using the model constructed at first step and covariates, we predict the exposure of sub-survey area which cannot be observed exposure.
- Then fill the missing by predicted values of model.



Procedure using for estimating the effect of exposure in common research.

- Step3 (Estimating the effect) Then poisson regression model is fitted to regression imputed data, and estimate the effect of exposure



The aim of this research

From the missing data analysis context

- If we treat the imputed values as if they are observed, the consistency of estimator may be violated, the variance of estimator will be underestimated !!

Purpose

A. We try to <u>clarify the problem of using regression imputation</u> from analytical and practical points of view, and organize the important points of matter when we use regression imputation in practical analysis.

B. To develop the method to estimate the effect of exposure, avoiding the problem caused by the limited number of monitoring stations.

Today I'm going to talk…

Focus on…

A. We try to clarify the problem of using regression imputation from analytical and practical points of view, and organize the important points of matter when we use regression imputation in practical analysis.

At first, let us consider the problem of regression imputation from analytical point of view under the simple setting.

Analytical point of view - Setting

Setting

we consider the case of using linear regression model for imputation and poisson regression model to estimate the effect …

- This setting, can be formulated as the problem of estimating the regression coefficients β by fitting model-(2) to data generated by Model-(1).

(1) Generating scheme of data

$$Y_i|X_i \sim \mathrm{Possion}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0^* + X_i\beta_1^*$$

$$X_i|\mu_i \sim \mathrm{Normal}(\mu_i, \sigma^2)$$

$$\mu_i \ : \text{mean of exposure}$$

(2) Model for analysis

$$Y_i|\mu_i \sim \mathrm{Possion}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \mu_i\beta_1$$

$$\mu_i \ : \text{known}$$

use mean of X as covariates

Assume to know the true mean of exposure

under this setting, consider the properties of estimators

Analytical point of view - Consistency of regression coefficient estimator.

The estimator for $\beta = (\beta_0, \beta_1)^T$ has following property

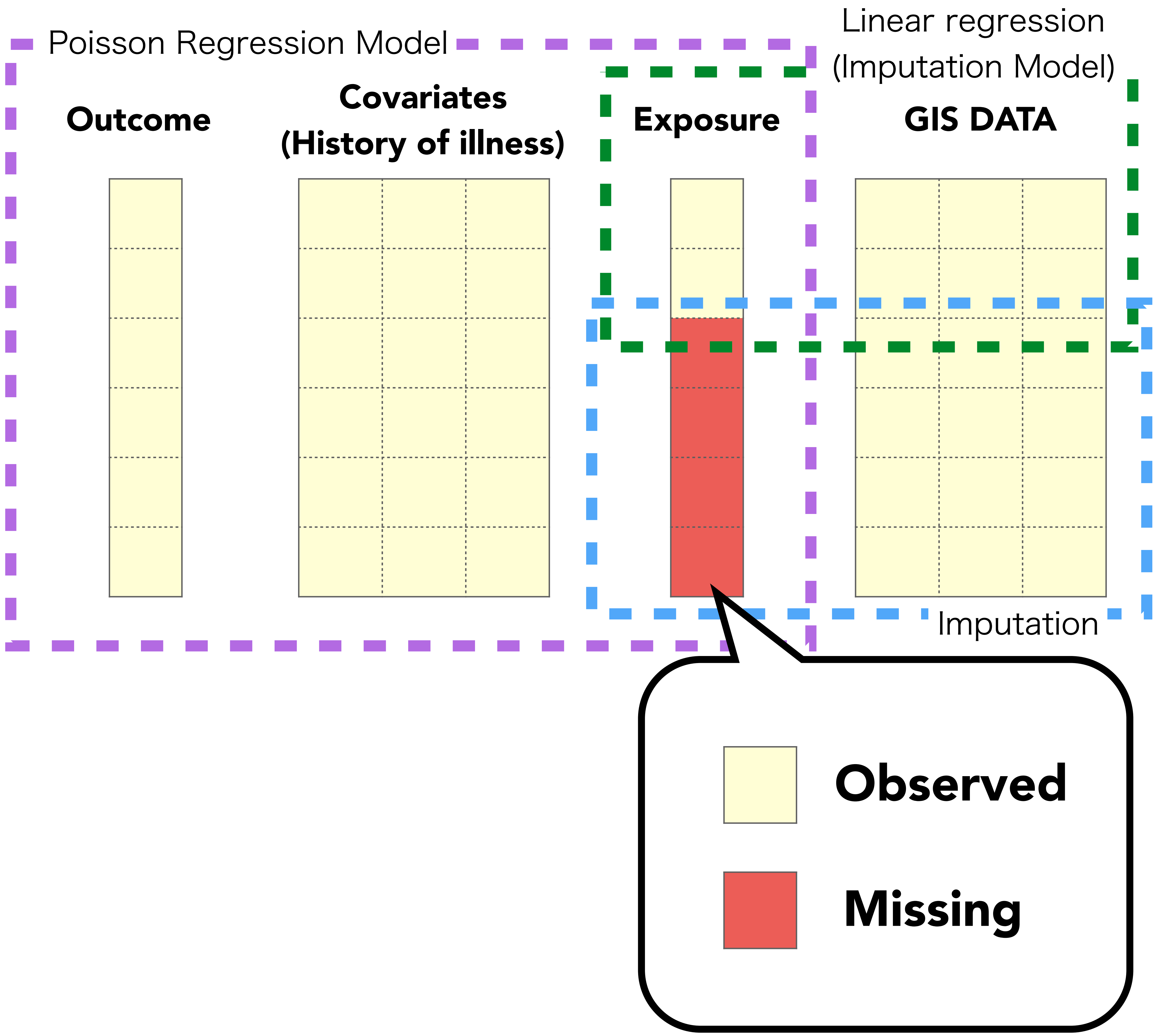
$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \xrightarrow{P} \begin{pmatrix} \beta_0^* + \dfrac{\sigma^2}{2}\beta_1^{*2} \\ \beta_1^* \end{pmatrix}$$

- $\hat{\beta}_0$ is not a consistent estimator
- $\hat{\beta}_1$ is consistent estimator

We are interested in only, $\hat{\beta}_1$
so the inconsistency of intercept is not a significant problem

Analytical point of view - Result

- If we <u>know the true mean of exposure</u>, the estimator for effect of exposure will be consistent.
- When we don't know the true mean of exposure, the estimator of exposure will be **inconsistent**
- In both cases, **asymptotic variance will be smaller** than when we treat imputed values properly.



What is the practitioner want to know is

- Now we showed, the inference based on regression imputation method is invalid.
- However, what is the practitioner want to know is..

(A) <u>How much the bias will be occur when we use the regression imputation ?</u>

(B) <u>When the large bias will be occur ?</u>

➡ <u>To make clear these things,</u>

<u>we perform the simulation of regression imputation under previous settings</u>

Practical point of view - simulation setting

- We visualize the inconsistency of estimator and underestimation of asymptotic variance.

Procedure of Simulation

1. Generate 160 size data containing 70% missing exposure.
2. Construct the Linear regression model for imputation using data obtained at monitoring station.
3. Impute the missing values of exposure by linear predictor.
4. Fit poisson regression model to regression imputed data, and estimate the coefficients.
5. iterate 1-3 procedure , 1000 times

Practical point of view - simulation setting (procedure)

Step(1) : Generating the data

$$Y_i|X_i \sim \mathrm{Possion}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0^* + X_i\beta_1^*$$

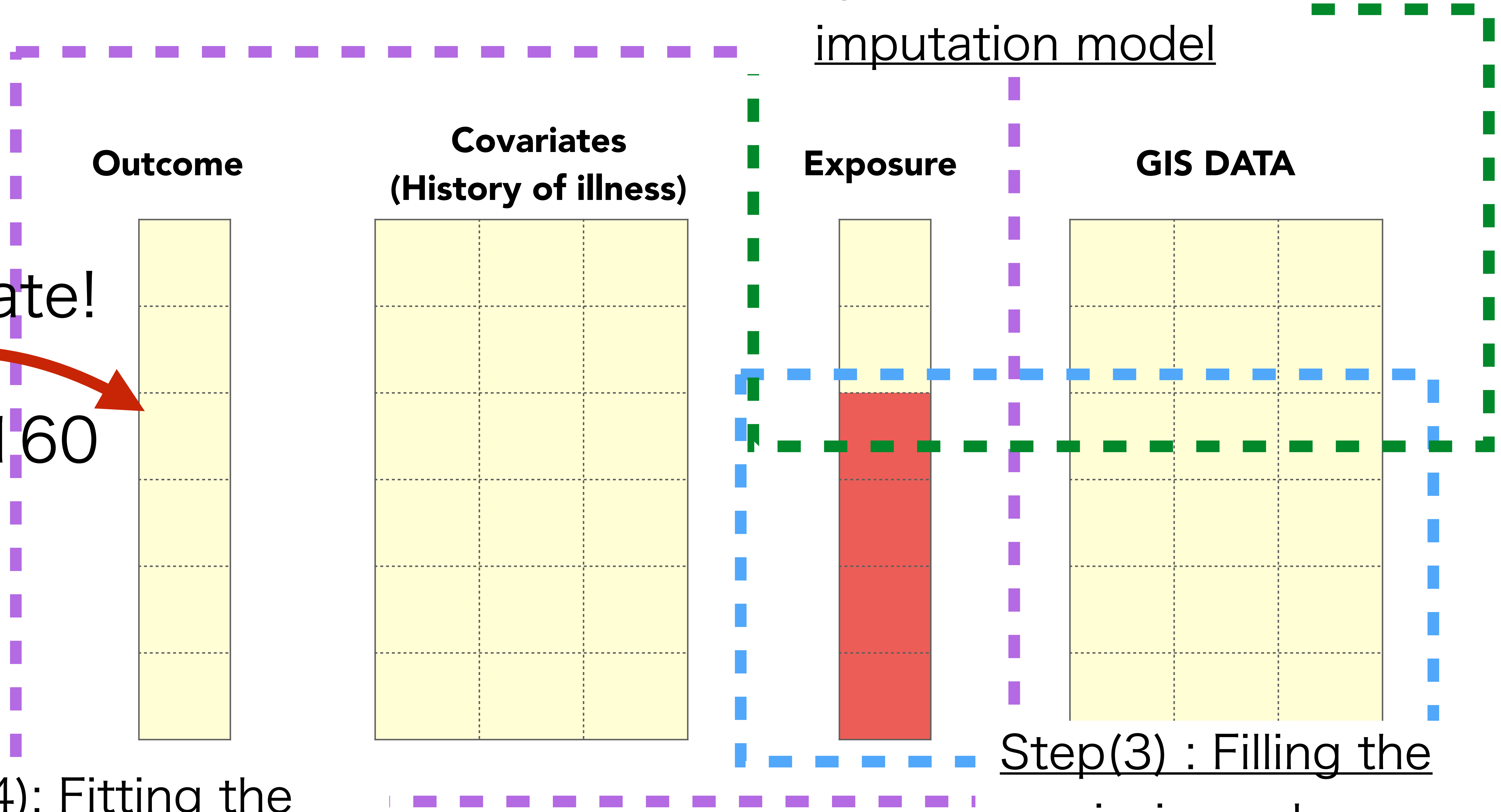
$$X_i|\mu_i \sim \mathrm{Normal}(\mu_i, \sigma^2)$$

$$\mu_i \; : \text{fixed}$$

generate!

size : 160

Step(2) Construct the imputation model



Step(3) : Filling the missing values

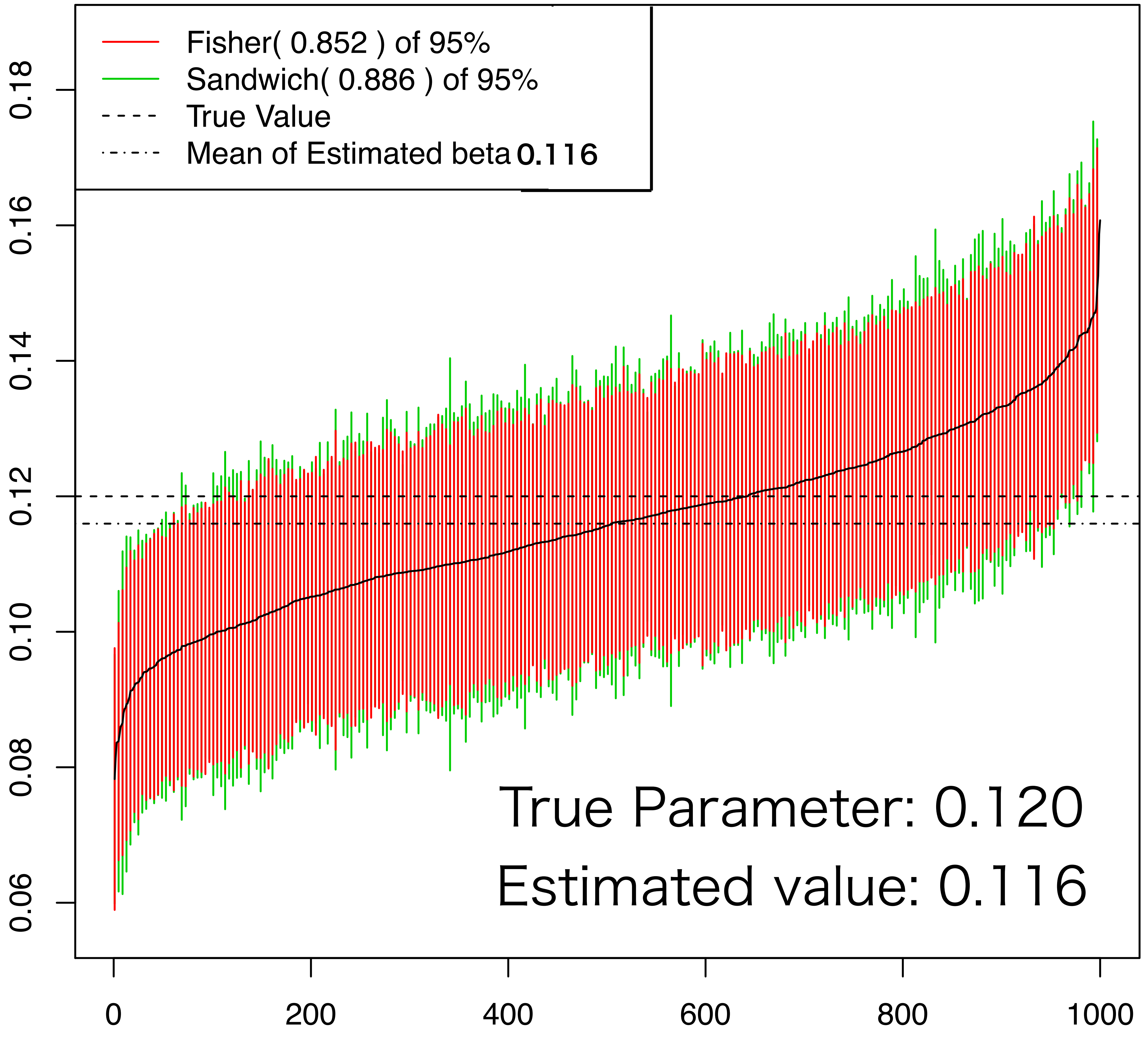
Step(4): Fitting the Poisson Regression Model

Step(5): Estimate the Coefficient of Exposure and its 95% CI

iterate 1000 times

Result - Fitting Poisson Regression Model to Regression Imputed Data.

- Compute 95% Confidence interval for coefficient of poisson regression model, 1000 times. (sort by coef)
- Black Solid Line : Estimated coefficients
- Red Band : 95%CIs based on Fisher Information (85,2% contain the true value)
- Green Band : 95%CIs based on Sandwich Estimator (88.6% contain the true value)
- Mean of Estimated value does not consistent to true value. 95% CIs based on fisher information and Sandwich Estimator is shorter but not too much.



Result - analytical & practical point of view (simple setting).

By Analysis	· The estimator of the effect of exposure is **inconsistent** · Asymptotic variance is **underestimated**
By Simulation	· **If** we can **specify the mean model** for exposure **properly** · Average of estimates is approximately equal to the true parameter value. · **Underestimation** of asymp. variance matters little

Result

Using the regression imputation is **not appropriate in analytically**,
but is **not unacceptable in practice**

Result - analytical & practical point of view.

By Analysis

- The estimator of the effect of exposure is **inconsistent**
- Asymptotic variance is **underestimated**

By Simulation

- **If** we can <u>**specify the mean model**</u> for exposure <u>**properly**</u>
- Average of estimates is approximately equal to the true parameter value.
- **Underestimation** of asymp. variance matters little, when $R^2 > 0.6$

Result

Using the regression imputation is <u>**not appropriate in analytically**</u>,
but is <u>**not unacceptable in practice**</u>

Simulation (practical settings) - Procedure

Let me consider the data generated from following model.

Model for Outcome

$$Y_i | X_i \sim \text{Poisson}(\lambda_i)$$

$$\log \lambda_i = \beta_0 + x_i \beta_1 + \sum_{k=2}^{4} z_{ik} \beta_k$$

i = 1,2, ⋯, 160

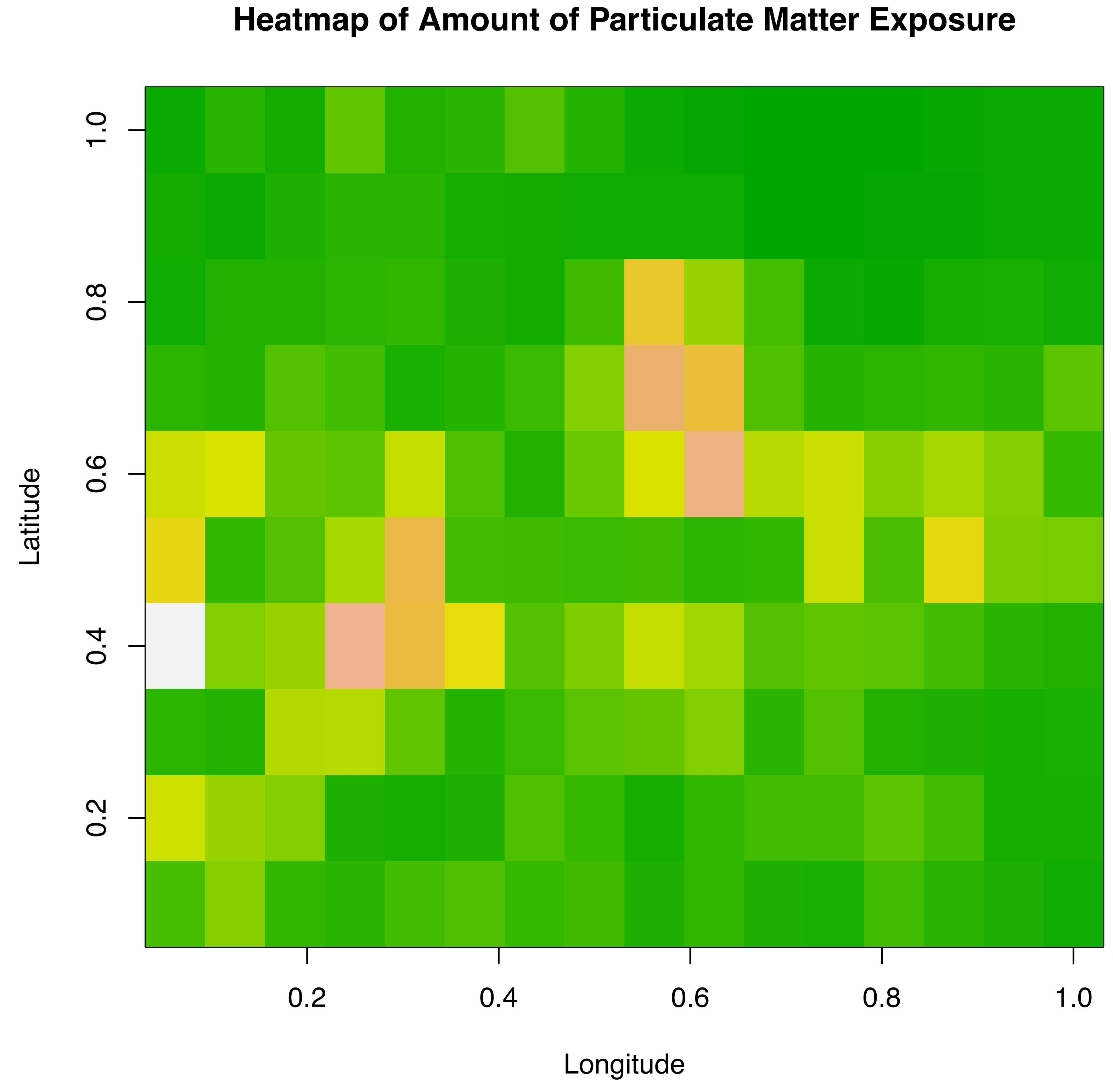
Model for Exposure

$$X_i | \mu_i \sim \text{Normal}(\mu_i, \sigma^2)$$

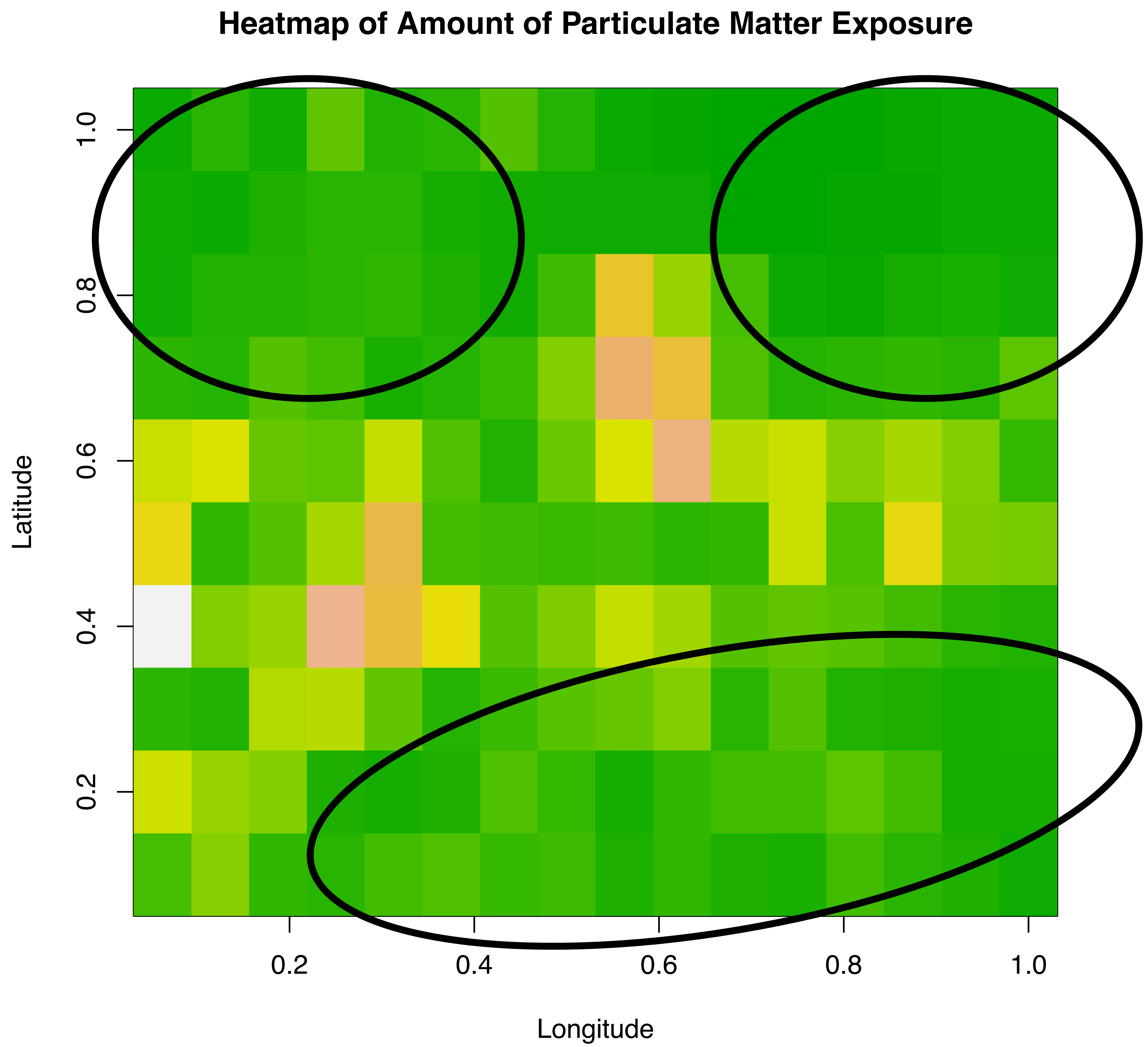
$$\mu_i = \alpha_0 + \exp\left(\sum_{j=1}^{4} w_{ij} \alpha_j + \gamma_{i0} + \sum_{j=1}^{4} w_{ij} \gamma_{ij} + \phi_i \right)$$

where, α_j are fixed effect, and γ_{ij} are the random effect at each site, and ϕ_i is spatial component.

One of the realization from model for exposure



Simulation (practical settings) - Procedure

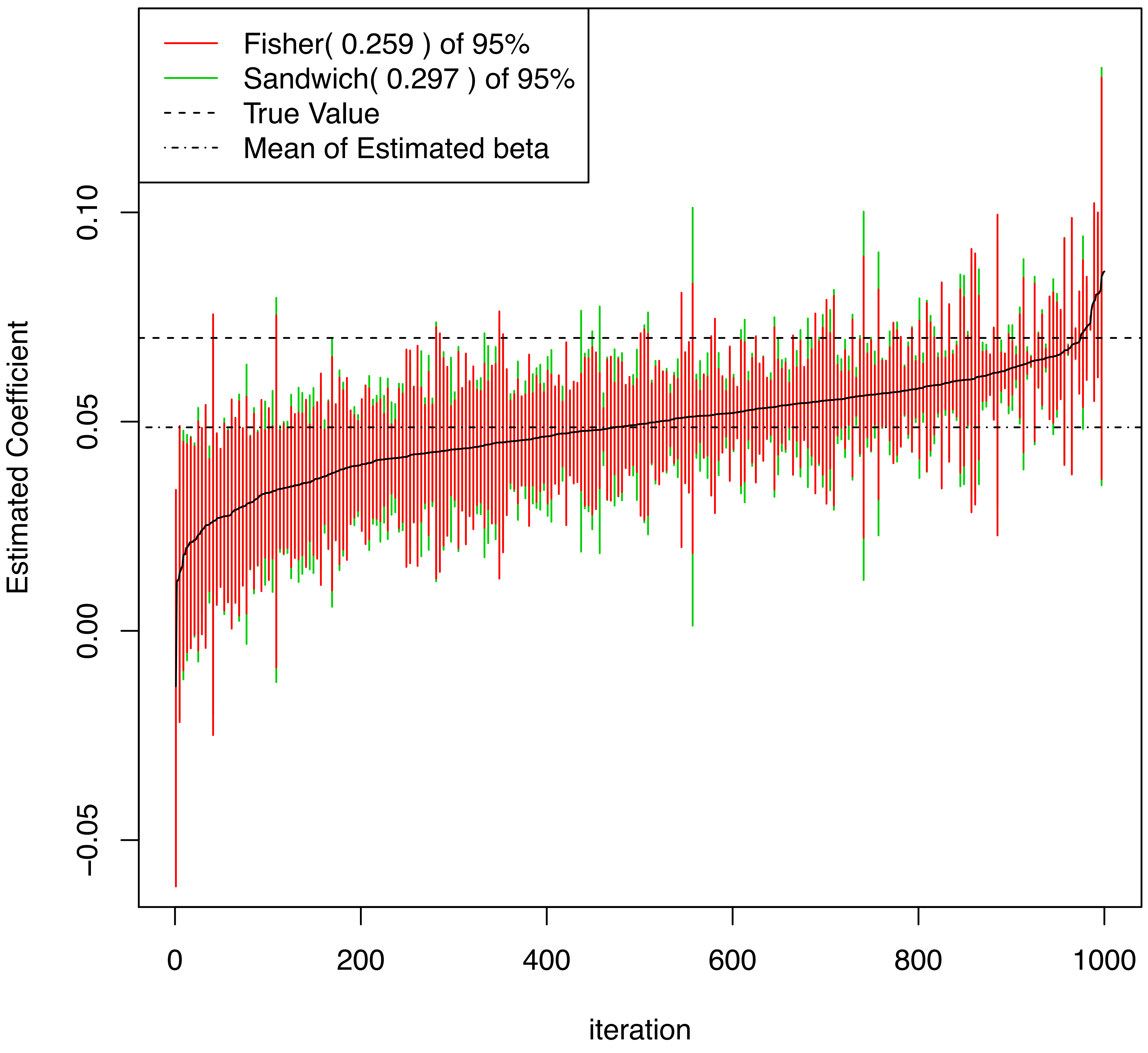


- In general, monitoring stations are not established with low particulate matter concentration.
- At areas surrounded by black solid line, we assume that we cannot observe the exposure information.
- Then, we use regression imputation and estimate the effect of exposure β (same as previous simulation).

Iterate these procedure 1000 times.

Simulation (practical settings) - Result

- Compute 95% Confidence interval for coefficient of poisson regression model, 1000 times. (sort by coef)
- Red Band : 95%CIs based on Fisher Information (25.9% contain the true value)
- Green Band : 95%CIs based on Sandwich Estimator (29.7% contain the true value)
- Black Line : Estimated coefficients
- This figure shows, the consistency of estimator is violated, highly biased, and when there is a spatial correlation, the estimated value of exposure will be smaller than true value.



Conclusion

<u>Simple Settings</u>

using regression imputation to fill the missing does not cause the serious problem

<u>Realistic Settings</u>

if we ignore the effect of spatial structure and random effect of each sites, **the effect of exposure will be underestimated!!**

In real, for harmful effect, the negative impact of misjudgment

conclusion	There is a effect		There is no effect
	↓	$\ll$	↓
In real	There is no effect		There is a effect
	safe!		dangerous!

Conclusion and Discussion

Conclusion

Using regression imputation to estimate the effect of exposure causes very serious problem, so it is **unacceptable in practice !!!!**

Ongoing work

We now show the underestimation of the effect only by simulation, so next we show them by analytically

we develop the method alternative to regression imputation that can make robust inference under the mild model misspecification for exposure.

References - 1

- Kioumourtzoglou, M.A., Schwartz, J. D., Weisskopf, M. G., Melly, S. J., Wang, Y., Dominici, F., Zanobetti, A. (2016). ``Long-term PM2.5 Exposure and Neurological Hospital Admissions in the Northeastern United States.'' Environmental Health Perspectives, 124(1), 23-29.
- Madrigano J, Kloog I, Goldberg R, Coull BA, Mittleman MA, Schwartz J. (2013). ``Long-term Exposure to PM2.5 and Incidence of Acute Myocardial Infarction.'' Environ Health Perspect 121:192-196;
- Yongping Hao, Lina Balluz, Heather Strosnider, Xiao Jun Wen, Chaoyang Li, and Judith R. Qualters (2015). ``Ozone, Fine Particulate Matter, and Chronic Lower Respiratory Disease Mortality in the United States'', American Journal of Respiratory and Critical Care Medicine, Vol. 192, No. 3, pp. 337-341.
- Cynthia A. Garcia , Poh-Sin Yap , Hye-Youn Park , Barbara L. Weller (2016)., ``Association of long-term PM2.5 exposure with mortality using different air pollution exposure models: impacts in rural and urban California'', International Journal of Environmental Health Research , Vol. 26, Iss. 2
- U.S. EPA.(2009) ``2009 Final Report: Integrated Science Assessment for Particulate Matter.'' U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-08/139F.
- Rubin, D. B. (1976). ``Inference and missing data''. Biometrika, 63, 581–592.
- Cesaroni G, Badaloni C, Gariazzo C, Stafoggia M, Sozzi R, Davoli M, Forastiere F. (2013). ``Long-Term Exposure to Urban Air Pollution and Mortality in a Cohort of More than a Million Adults in Rome.'' Environ Health Perspect 121:324-331;

references - 2

- Nakamura, T. Minami, M.(2015). Prediction of PM10 and PM2.5 Concentration Using Land Use Data and Spatial Correlation". In JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association. 2509-2514.
- Peter J. Huber. (1967), The behavior of maximum likelihood estimates under nonstandard conditions," in Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1967), 221-233
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.
- Tsiatis, A. A. (2006). Semiparametric theory and missing data. New York: Springer.
- Vaart, A. W. (1998). Asymptotic statistic. Cambridge: Cambridge University Press.
- Wood, S. N. (2006), Generalized additive models: An introduction with R. Boca Raton, FL: Chapman & Hall/CRC.