Underestimation of standard errors in regression analysis for pollution exposure assessment using multi-source data

Tomoshige Nakamura* Mihoko Minami†

Abstract

We are interested in investigating the effect of particulate matter exposure on human heath in Japan using community health survey data. However, in Japan, the number of the monitoring stations around the survey area is very limited and the observations on these measurements at local community area are not generally available. When particulate matter concentrations are not observed in survey area, Land Use Regression (LUR) is often used to fill the missing values.

In general, if we use regression imputation to fill the missing values, the inference based on regression imputed data might be wrong. For example, the consistency of estimator may be violated, and the variance of estimator may be underestimated. So, in our research, we try to clarify the problem using regression imputation when we estimate the effect of particulate matter.

Key Words: regression imputation method, pollution exposure assessment, missing data problem

1. Introduction

In this research, we consider to estimate the effect of particulate matter exposure on our health using community health survey data and particulate matter concentration data observed at monitoring stations around the survey area. Particulate matter is a complex mixture of extremely small particules and liquid droplets, and the relationship between particulate matter exposure and variace diseases is highly concerned (e.g. athma, lung cancer, cardiovascular diseases, see Cynthia et al.,2016; Kioumourtzoglou et al.,2016; Madriano et al.,2013;). For example, ESCAPE Study investigates the long-term effects of exposure to air pollution on human health in Europe.

When we estimate the effect of particulate matter exposure, we need the information of particulate matter concentration in survey area. In U.S., measurement of particulate matter concentration were started in the early of 2000s. However, in Asia, especially in Japan, particulate matter concentration are started to be observed at many sites in past few years. For these reasons, the number of monitoring stations is limited.

The figure1 shows the survey conducted at some area of Japan, the area surrounded by the green solid line represents community health survey area, and red dots represents the monitoring stations for particulate matter(PM10) exposure. We can see the number of monitoring stations in the survey area limited, for example, in the left above segment of suvey area, there is no monitoring station. So, the data we obtain from the survey containing the missing values

When we treat the data containing the missing values, we must pay enough attention to the missing data mechanisms(Rubin, 1987), and how to deal the missings. For example, if missing mechanism is Missing Completely at Random, then we can ignore the missing values when we estimate the effect. But if missing mechanism is Missing at Random, then

*Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

†Department of Mathematics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

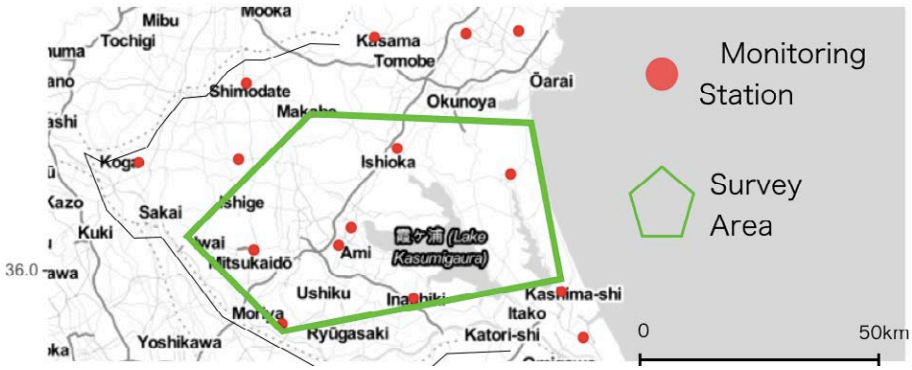

Figure 1: The survey area and established monitoring stations around the survey area.

we must consider the distribution of missing values given observed data. If we misspecify the missing mechanism, the statistical inference under that assumption might be wrong.

However, in many reseaches of estimating particulate matter effect, these problems are not paid enough attention and inverse distance weighting method(IDW)[a], or Land Use Regression(LUR) are used to fill the missing values of exposure. In this article, we focus on the case using LUR to fill the missing. LUR is the method to estimate the spatial outcome at non-observed area by linear regression using GIS information and meteorological variables as predictor, and widely used to fill the missing values of particulate matter exposure(Vlachogianni, 2011).

In general, when we treat imputed values as they were observed, then the inference based on imputed data is not always proper (Rubin 1987). For example, from missing data analysis constext, the estimated coefficient obtained by fitting regression model to imputed data may not be guaranteed its consistency, and the estimated variance will be underestimated.

The purpose of this paper is to clarify the problem of using regression imputation to fill the missing values of exposure when we estimate the effect of particulate matter exposure. In section 2, we describe the procedure that used to estimate the effect of particulate matter exposure in common research, then we show the consistency of estimator of the effect obtained under the previous settings and we illustrate the variance of that estimator will be underestimated through simulation. In section 3, we summarize the section 2, and present the difficulty of estimating the effect of particulate matter exposure in realistic situation and our on-going work.

2. The problem of using regression imputation when we estimate the effect of particulate matter exposure

In this section, we first describe the procedure of estimating the effect of particulate matter exposure in common reseach. In section 2.2, we show the consistency of estimator for effect of exposure under the procedure using in common research, and in section 2.3, we describe the variance underestimation of estimator through simulation study.

analytically, and we show the underestimation of variance by simulation under this setting.

2.1 The procedure of estimating the effect of exposure in common research

In common research, following three steps are used to estimate the effect of particulate matter.

(1) Constructing the model for particulate matter exposure: In this step, constructing prediction model for particulate matter exposure by fitting linear regression model to observed data obtained at monitoring stations. In this model, GIS information and meteorological variables are used as predictor.

(2) Filling missing exposure: In this step, missing values of exposure are filled by the predicted values computed by the model constructed in step (1). Then dataset without missing values are obtained, hereafter referred to it as "imputed data".

(3) Estimating the effect : In this step, the effect is estimated by fitting regression model to imputed dataset .

In section 2.2, we show the property of estimator obtained by using this procedure.

2.2 The consistency of regression coefficient estimator

We consider the consistency of estimator of the effect of particulate matter obtained by above mentioned procedure. In step(3) of the procedure, Logistic regression model or Poisson regression model and its extension are often used. So we focus on the case using Poisson regression model to estimate the effect.

The procedure used in common research can be formulated as the problem of estimating the regression coefficient β_1 by fitting model (FM) to data generated by (DG). For the sake of simplicity, we use only two variables as a predictor, intercept and X_i. X_i is corresponding the variable observed at site i which we want to estimate the effect on outcome Y. In our case, X_i is the exposure at site i. We denote the mean of X_i as μ_i.

(DG)) Generating scheme of data

$$\begin{aligned} Y_i|X_i &\sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0^* + X_i\beta_1^* \\ X_i|\mu_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &: \text{ mean of exposure} \end{aligned} \tag{1}$$

(FM) Fitting model

$$\begin{aligned} Y_i|X_i &\sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + \mu_i\beta_1 \\ \mu_i &: \text{ mean of exposure} \end{aligned} \tag{2}$$

In practical situation, we don't know μ_i, so we need to estimate it. But, to avoid the complexity, we assume to know μ_i, and investigate the property of estimator under this assumption. Under these settings, the estimator for regression coefficient $\beta = (\beta_0, \beta_1)^T$, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$, has following property.

Theorem 1. The estimator for regression coefficient, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$, has following property.

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \xrightarrow{P} \begin{pmatrix} \beta_0^* + \frac{\sigma^2}{2}\hat{\beta}_1^{*2} \\ \beta_1^* \end{pmatrix} \tag{3}$$

Theorem1 shows $\hat{\beta}_0$ is not consistent estimator for true paramete value β_0^*, and $\hat{\beta}_1$ is consistent estimator for β_1^*. In practical situation, we are not interested in $\hat{\beta}_0$, but only interested in $\hat{\beta}_1$, so, when we use Poisson regression model to estimate the effect of exposure and if we can assume to know μ_i, we can obtain the consistent estimator for the effect. This result can be extend to the case of Poisson regression model contains other variables. In such a case estimated coefficient $\hat{\beta}_1$ is a consistent estimator as well.

As a result, if we can specify the mean model for exposure properly, we can obtain onsistent estimator for the effect of exposure, when we use regression imputation method to fill the missing values.

2.3 Variance underestimation for regression coefficient estimator

In section 2.2, we showed the consistency of β_1. In this section, we focus on the variance of $\hat{\beta}_1$. When we treat imputed values as if they were observed, variance of $\hat{\beta}_1$ calculate as follows (Wood,2006).

$$\text{Var}(\hat{\beta}_1) = \{(Z^T W Z)^{-1}\}_{2,2}, \qquad \text{where } \{Z\}_i = (1, \mu_i)^T \tag{4}$$

However, $\text{Var}(\hat{\beta}_1)$ is not a true variance of $\hat{\beta}_1$ because fitting model (2) is different from data generating scheme (1). So, we perform the simulation and visualize the degree of underestimation of (4) under the realistic model and parameter settings.

In this simulation, we assume (DG-S) for generating scheme of data, and fit model (FM-S) to the data generated by (DG-S) and compute estimator $\hat{\beta}_1$ and its 95% confidence interval based on the variance (4). The procedure of simulation is the following.

1. Generating simulation data with sample size 160 from model (DG-S)
2. Fit the model (FM-S) to simulation data and comple $\hat{\beta}_1$ and 95% confidence interval based on (4).
3. Iterate (1), (2), 1000 times.

(DG-S)) Generating scheme of data

$$\begin{aligned} Y_i|X_i &\sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= 0.5 + 0.08 \times X_i + 0.05 \times Z_{i1} + 0.04 \times Z_{i2} - 0.03 \times Z_{i3} \\ X_i|\mu_i &\sim \text{Normal}(\mu_i, 2.5^2) \\ \mu_i &= 17 + \exp(w_{i1} \times 0.5 - w_{i1} \times 0.2 + w_{i3} \times 0.4 + w_{i4} \times 0.5) \end{aligned} \tag{5}$$

where $(w_{1j}, w_{2j}, \cdots, w_{160j})^T \sim \text{MVN}(0, \Sigma)$, $Z_{ij} \sim \text{Nornal}(0, 1)$, and Σ is some covariance matrix between site i and site j.

(FM-S) Fitting model

$$\begin{aligned} Y_i|X_i &\sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + \beta_1 \times \mu_i + \gamma_1 \times Z_{i1} + \gamma_2 \times Z_{i2} - \gamma_3 \times Z_{i3} \end{aligned} \tag{6}$$

Figure2 shows 95% confidence intervals sorted by estimated coefficients, red bands represents the 95% confidence interval for each estimator, green solid line represents the true parameter value, and black dashed line represents the average of estimated coefficients.

As shown in figure2, the green solid line and the black dashed line are overlapping, so the estimated coefficient $\hat{\beta}_1$ may be unbiased estimator for $\hat{\beta}_1$.

Also, figure2 shows 84.9 % of 95% confidence intervals contains the true value, this shows the underestimation of variance of $\hat{\beta}_1$ based on (4).

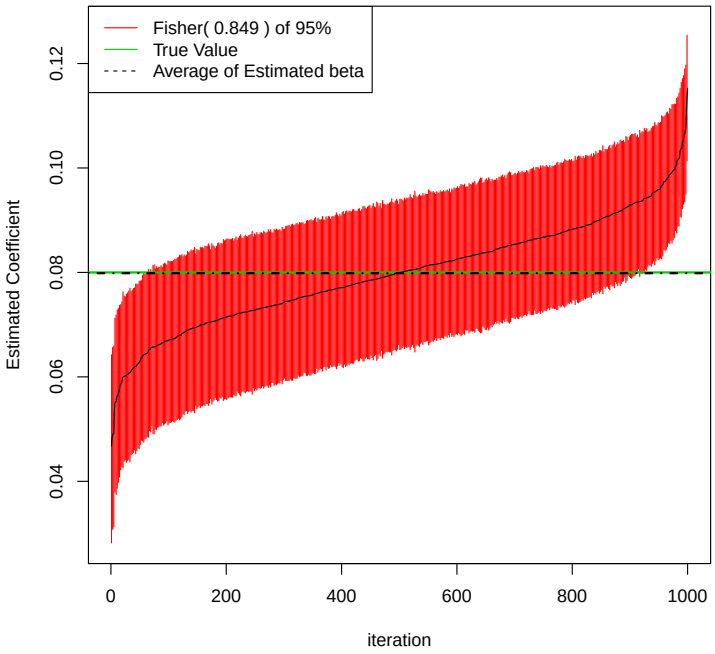

Figure 2:

As a result of this simulation, estimated variance of $\hat{\beta}_1$ become smaller than that is properly estimated. So when we use regression imputation method and if we treat imputed values as if they were observed, statistical infernece might become wrong.

3. Discussion

In this article, we describe the problem of using regression imputation method to fill the exposure missing when we estimate the effect of particulate matter. As a result of section2, when we use linear regression model for imputation and use poisson regressiom model to estimate the effect, estimator for the effect is consistent, however, variance of its estimator might be underestimated.

By the underestimation of variance, confidence interval of estimator become shorter, and then the inference performed with these confidence intervals would result in misleading conclusion.

To make matters worse, the consistency of regression coefficient estimator is guaranteed only when we can know the true mean μ_i, or we can specify the true model for particulate matter exposure. When the number of monitoring station is limited, constructing the model for spatial distribution of particulate matter concentration in whole survey area is not easy task. From these discussion, using regression imputation method to fill the missing values is not approvable when the number of monitoring stations is limited. To avoid these problems, we have to consider to use bayesian model containtg missing values as parameters to estimate the effect.

Anyway, even if we use bayesian approach, to construct the model for partculate matter concentration is critical task to estimate the effect of particulate matter. So, we try to construct precise prediction model for partculate matter exposure by using spatio-temporal

model.

REFERENCES

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

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 B. A., Mittleman M. A., Schwartz J. (2013), "Long-term Exposure to PM2.5 and Incidence of Acute Myocardial Infarction." *Environmental Health Perspectives*, 121:192-196;

Rob, B., Ole, R., Massimo, S., Zorana J. A., el al.(2013), "Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project", *Lancet*, 360, 1233 - 1242

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, New York: Wiley.

Vaart, A. W. (1998). *Asymptotic statistic*. Cambridge: Cambridge University Press.

Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., Kukkonen, J., (2011), "Evaluation of a multiple regression model for the forecasting of the concentrations of NOX and PM10 in Athens and Helsinki", Sci Total Environ, 409 (8), pp. 1559-1571

Wang, M., Beelen, R., Bellander, T., Birk, M., Cesaroni G., et al. (2014), "Performance of multi-city land use regression models for nitrogen dioxide and fine particles", *Environmental Health Perspectives* 122, 843-849;

Wood, S. N. (2006), *Generalized additive models: An introduction with R*, Boca Raton, FL: Chapman & Hall/CRC.

Yongping, H., Lina, B., Heather, S., Xiao J. W., Chaoyang L., and Judith R. Q. (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.