Causal inference on temperature-mortality relationship: comparing the distributed lag nonlinear model and Rubin causal model

Jaemin Oh[1] and Yeonseung Chung[2]

[1,2]Department of Mathematical Sciences, KAIST, Daejeon, South Korea

November 6, 2022

Abstract

The temperature-mortality relationship has been analyzed by using regional time series data under the DLNM framework. There have been many concerns about the causal interpretation of the temperature-mortality relationship, because of unmeasured confounders, model selection problems, and mixing of the design stage & the analysis stage. In this article, we used Rubin causal model (RCM) to deal with the last two issues, and obtained a consistent result compared to the previous studies. This work shines a light on the possibility of the causal interpretation of the temperature-mortality relationship analyzed so far.

1 Introduction

Due to global warming and climate change, analyzing the effect of the ambient temperature on human health is an important research topic [5, 10, 19]. Usually, the analysis of the relationship has conducted in a regional time series setting, and it was challenging because of the temporal trend in the time series and the existence of the delayed effect. These difficulties have been addressed by the distributed lag nonlinear model (DLNM) framework [3] producing nonlinear exposure-response surface $\mu : (w, l) \mapsto \mu(w, l) \in \mathbb{R}$ where w is the ambient

temperature and l is a time lag. In general, regression methods containing the DLNM framework have some advantages: identification of lagged effects and low dimensional summary of estimated nonlinear exposure-response curve derived from the surface. These advantages make the regression approach popular in environmental epidemiology, especially for the topic of the temperature-mortality relationship.

However, there has been a concern about the causal interpretation on the results obtained by regression analysis. In the recent debate on air pollution study [6] that uses similar tools to analyze time-series data, two aspects were pointed out: mixing of the design stage & the analysis stage, and the model selection problem. Rubin said, in the observational study, design stage that adjusts confounding bias and the analysis stage that relates treatments and outcomes should be separated to approximate the gold standard of causal inference, the randomized experiment [15]. However, regression methods mix two stages by including confounders as regressors to control their confounding bias, e.g., the DLNM framework eliminates seasonality by fitting additional spline basis for time. Moreover, the DLNM framework is susceptible to model selection problem [4] since we don't know the exact placement of knots and the exact degrees of freedom. Therefore, we have to solve these issues first to make causal interpretation possible, together with collecting more data to remove the existence of unmeasured confounders.

As a solution, we suggest using Rubin causal model (RCM) [7] known as the potential outcome framework. It separates the design stage and the analysis stage, and does not need to select any parametric model for the outcome-generating process. RCM was first introduced to analyze the data of randomized experiments [14], but now widely used in observational studies [17], and even in time series data [1]. In this paper, we used RCM to estimate the log of relative risk (logRR) curve of the ambient temperature and compared the result to the DLNM framework.

2 Dataset

The data is composed of daily mean temperature in the first decimal place, and daily all-cause death counts during the period from 1997 to 2018 across 36 regions in South Korea. Figure 1 presents daily mean temperature and daily all-cause deaths in Seoul between 2010 and 2018. The daily mean temperature shows apparent seasonality and the peaks are increasing due to global warming. The number of deaths shows annual seasonality with an increase in cold

seasons and a decrease in warm seasons. It shows a long-term increasing trend also.

3 Method

For $i = 1, \ldots, N(= 36)$ and $t = 1, \ldots, T(= 8054)$, the information of i-th region at time t can be described by $(Y_{i,t}, W_{i,t}, C_{i,t})$ where Y is the all-cause death counts, W is the mean temperature, and C is the vector of the year, the month of the year, the week of the year, and the day of the year. Hereafter, we consider a fixed region and drop the subscript i to simplify the notation.

3.1 Distributed Lag Nonlinear Model

In the DLNM framework, the temperature-mortality association is described by the equations

$$Y_t \sim quasi - Possion(\lambda_t),$$
$$\log(\lambda_t) = \alpha + cb(W_t, \ldots, W_{t-L}; \beta) + \eta(C_t; \gamma).$$

where λ_t is the mean of the Poisson distribution with overdispersion (namely, Quasi-Possion [16]), cb is the cross-basis function with specified lag L, and η is the spline basis. By maximizing the quasi-likelihood function, we get estimates $\hat{\alpha}, \hat{\beta}$ and $\hat{\gamma}$. Here, we emphasize that including η in the model is equivalent to eliminating temporal trend in all-cause deaths to see the short-term variation, and fitting cb function is estimating temperature effect, thus the design stage & the analysis stage are mixed.

3.2 Rubin Causal Model

Now we introduce the notation of potential outcomes. $Y_t(w)$ refers to the potential outcome variable at time t that would have been observed under the treatment value $W_t = w$. $Y_t(w')$ is the potential outcome variable that would have been observed by the counterfactual imagination that $W_t = w'$ had been observed instead of $W_t = w$. Observed outcome Y_t is equal to the potential outcome under observed treatment value, $Y_t(W_t)$. This is called consistency assumption 1. See the reference [12] for a more detailed explanation of the definition.

We might be interested in the individual risk ratio

$$\frac{Y_t(w)}{Y_t(w')}$$

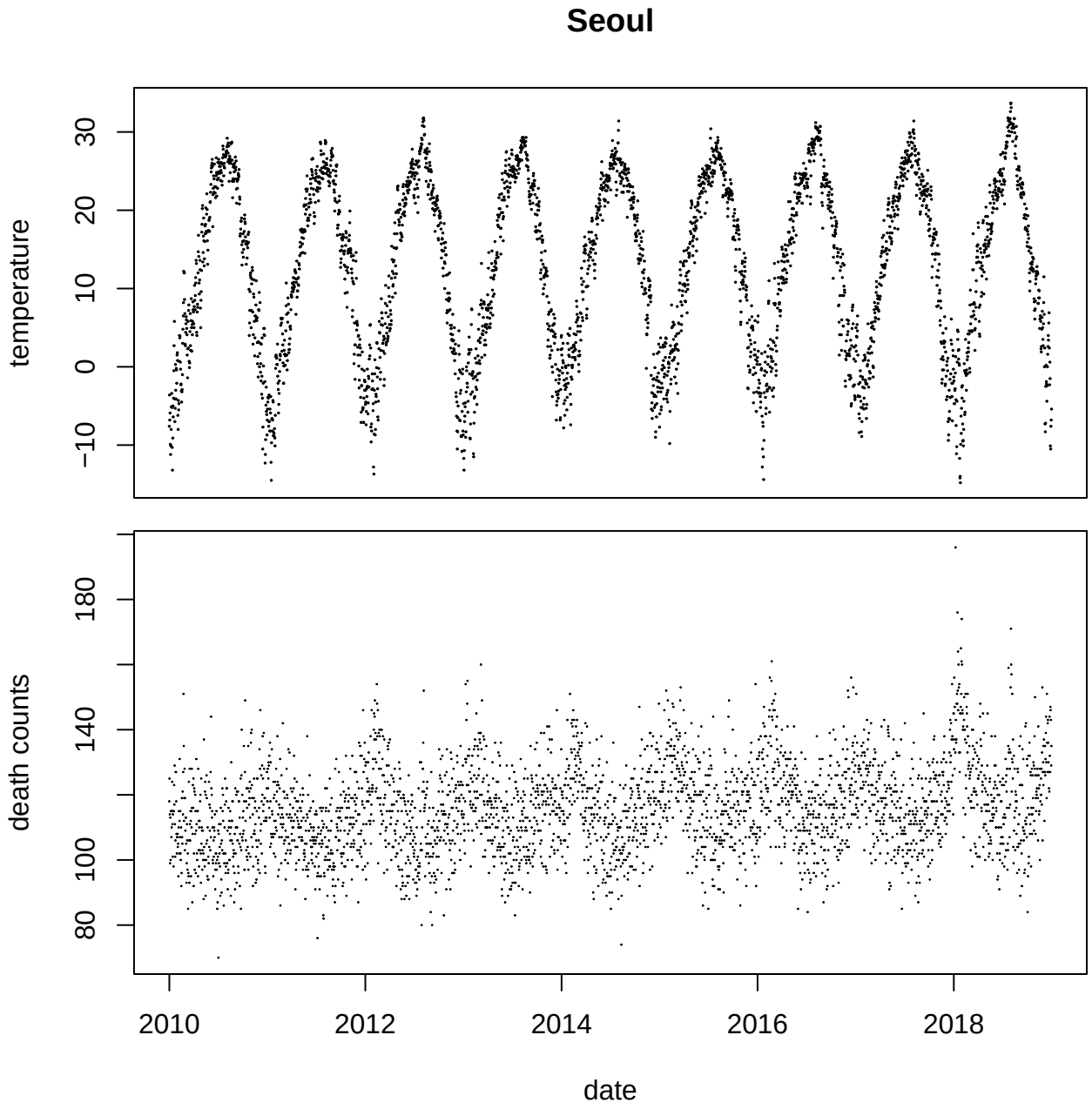

Figure 1: Daily times series of mean temperature (in Celcius) and all-cause death in Seoul from 2010 to 2018

if we know true values of $Y_t(w)$ and $Y_t(w')$. However, we never know the true potential outcomes of unmeasured treatment, due to its counterfactual nature. This is called the fundamental problem of causal inference [7]. Instead, we concentrate on the average risk ratio

$$\frac{\mathbb{E}\left[Y_t(w)\right]}{\mathbb{E}\left[Y_t(w')\right]}.$$

There have been many studies to estimate $\mathbb{E}[Y_t(w)]$. In a marginally randomized experiment, $\mu(w) = \mathbb{E}[Y(w)]$ can be estimated from observed data [14]. In observational studies, one can estimate causal estimand $\mu(w)$ by preprocessing the data to approximate randomization e.g., inverse probability weighting, standardization, matching [13]. Note that we dropped the subscript t to indicate a more general situation than the time-series setting. Throughout those techniques, the fundamental assumptions that make it possible to estimate the causal estimand $\mu(w)$ are below:

Assumption 1 (Consistency)

The potential outcome for observed treatment is equal to the observed outcome. That is, $Y_t(W_t) = Y_t$.

Assumption 2 (Positivity)

Discrete treatment: For all w *and* C_t, $p(w|C_t) = Pr\left(W_t = w|C_t\right) \in (0, 1)$.

Continuous treatment: For all w *and* C_t, $p(w|C_t) > 0$ *where* $p(w|C_t)$ *is a conditional density.*

Assumption 3 (Weak Unconfoundedness)

For all w, $Y_t(w) \perp W_t|C_t$.

The positivity assumption says all treatments are possible for each confounder. The weak unconfoundedness assumption, also known as "weak ignorability" or "selection on observables" in different contexts, says conditional on current confounders, the assignment mechanism is random to potential outcomes. Under these three assumptions, the causal estimand can be calculated as

$$\begin{aligned}
\mathbb{E}\left[Y_t \frac{1_{(W_t=w)}}{p(w|C_t)}\right] &= \mathbb{E}\left[\mathbb{E}\left(Y_t(w)\frac{1_{(W_t=w)}}{p(w|C_t)}|C_t\right)\right] \\
&= \mathbb{E}\left[Y_t(w)\frac{\mathbb{E}\left(1_{(W_t=w)}|C_t\right)}{p(w|C_t)}\right] \\
&= \mathbb{E}\left[Y_t(w)\right] = \mu(w),
\end{aligned}$$

where p is a mass or density function for discrete or continuous treatment respectively. The first equality comes from the iterated expectation formula and assumption 1, the second equality comes from assumption 3, and the third equality is due to the definition of $p(w|C_t)$. Thanks to the assumption 2, we can divide by $p(w|C_t)$. This is called "inverse probability weighting" (IPW) which is used to approximate marginally randomized experiments in which all treatments have the same probability of occurrence from conditionally randomized experiments. Therefore, a natural estimator of the causal estimand is

$$\hat{\mu}(w) = \frac{1}{T}\sum_{t=1}^{T} Y_t \frac{1_{(W_t=w)}}{p(w|C_t)}.$$

Still, we need to estimate $p(w|C_t)$ since it is generally unknown to us. When the treatment is binary, $p(w|C_t)$ is called propensity score, and it is used to adjust for confounding bias [13]. The propensity score can be extended to "generalized propensity score" (GPS) for categorical or continuous treatment [9]. For binary treatment, one can estimate propensity score by fitting the logit model to data. For categorical or continuous treatment, GPS can be estimated by fitting the ordered probit model or boosting.

PS and GPS have two nice properties [8,13]. The first one is the balancing property, which means that conditional on the same PS (or GPS), treatment and covariates are independent. The second one is PS-unconfoundedness, which means that conditional independence of potential outcome and treatment given PS. PS-unconfoundedness is implied by balancing property and unconfoundedness. These properties are the basis of propensity score-based matching methods. But in this paper, we used inverse probability weighting so these properties are not necessary to be explained more.

The main reason why we use inverse probability weighting by GPS is to achieve covariate balance. In randomized experiments, the distributions of covariates are similar across each treated group. However, we are now dealing with observational data which is not randomized. So, we generate a pseudo-population by imposing appropriate weights on each observation and look forward to the pseudo-population achieving covariate balance. One criterion for covariate balance is the absolute correlation (AC) [20]. If treatment and covariates are independent, then their correlation must be zero. So, a small value of AC can be evidence of covariate balance. Usually, AC with < 0.1 is considered acceptible.

4 Application

We set the reference temperature by 20°C.

4.1 Distributed Lag Nonlinear Model

We fitted the DLNM to our data to obtain region-specific effect estimates. Based on the previous study [5], we used quadratic B-spline and placed knots at 10th, 75th, and 90th quantiles for temperature dimension, we considered $L = 21$ lags, used natural B-spline, and placed 3 knots at equally spaced values in the log scale for lag dimension for lag dimension, we fitted additional natural B-spline of dates with 8 degrees of freedom for each year to eliminate seasonality, and indicator variables of the day of the week were included to control its effect.

With multiple effect estimates from various regions, we pooled those estimates by multivariate meta-analysis. We modeled effect estimates $\hat{\beta}_i$ from i-th region as a mixed-effect model

$$\hat{\beta}_i \sim N_m(\beta, S_i + V)$$

where N_m denotes m-dimensional multivariate normal distribution, S_i and V are with-in and between study error covariances, respectively, and β is the true aggregated effect. We used R package 'mixmeta' to estimate β and its confidence interval. See the upper left panel of figure 2.

4.2 Rubin Causal Model

We rounded daily mean temperatures to integer values. The first stage is the design stage to adjust for temporal confounding. The weight for confounding adjustment was We weighted each observation by

$$q_t = \min\left\{\frac{\hat{p}(W_t)}{\hat{p}(W_t|C_t)}, 10\right\},$$

where the first term is called stabilized inverse probability weight [18] that approximates a marginally randomized experiment in which each treatment has its relative frequency as the probability of assignment. To estimate $p(W_t|C_t)$, first we need to think about how daily mean temperature is determined. It would be very complicated, but the main factor should be the meridional altitude which can be predicted very well by the date. There might be more predictors than the date, but we considered their influence as a Gaussian error based on the central limit theorem. Therefore, we assumed the conditional distribution of treatment conditioned on covariates is a normal distribution, and estimated its mean and variance, i.e.

$$W_t|C_t \sim N(m(C_t), \sigma(C_t)^2)$$

where $m(C_t)$, and $\sigma(C_t)$ are some functions of C_t. The estimated mean $\hat{m}(C_t)$ was obtained by regressing W_t on C_t with boosting, and the estimated standard deviation $\hat{\sigma}(C)$ was calculated

by boosting residuals [8, 20]. Hyperparameters such as depth of tree(3), shrinkage(0.1), and the number of trees(20) were determined heuristically to minimize the absolute correlation. As an estimate of $p(W_t)$, we used relative frequency. One may use normal density with sample mean and sample variance too, but the weighting method with relative frequency achieved a lower absolute correlation in this application. See the second and third rows of Table 1. We trimmed weights bigger than 10 by 10, because some untrimmed weights were too large so effect estimate was heavily dependent on those observations.

We calculated absolute correlation (AC) [20] to see whether the covariate balance is achieved (here, AC < 0.1). Let c_t be a component of C_t, then absolute correlation with weight q_t is the absolute value of Pearson correlation coefficient between c_t and W_t, regarding each observation as q_t observations with the same values. See the table 1.

	year	month	week of the year	day of the year
1	0.01405874	0.25223312	0.25127380	0.25115386
2	0.03168787	0.08356729	0.08222320	0.08186817
3	0.03033376	0.12613061	0.12417815	0.12356290

Table 1: Absolute correlation (AC) before/after adjustment by IPW. The first row is AC before the adjustment; the second row is AC after adjustment, with the relative frequency of temperature as marginal probability; the third row is AC after adjustment, with a normal assumption on the marginal distribution of temperature.

The second stage is the analysis stage which relates treatments and outcomes with weights. We estimated the causal effect by the Horvitz-Thompson estimator with trimmed stabilized weights,

$$\hat{\mu}(w) = \frac{\sum_{t=1}^{T} q_t Y_t 1_{(W_t=w)}}{\sum_{t=1}^{T} q_t 1_{(W_t=w)}}.$$

To be consistent with previous studies and standardize different population sizes across regions, we calculated logRR curve by $\log \hat{\mu}(w) - \log \hat{\mu}(20)$ instead of risk difference. However, there is no known result about the distribution of the estimator of the risk ratio. So we measured the uncertainty of logRR curve by Moving Block Bootstrapping (MBB) [11] with 2000 bootstrap samples, 20 blocks, and each block is length of 400. In the bootstrap procedure, we did not fit the GPS model repeatedly for each bootstrap sample since we need to re-sample the pseudo-population itself that achieves covariate balance.

To pool the estimates, we assumed $X_i = (Y_{i,t}, W_{i,t}, C_{i,t})_{t=1}^{T}$ for $i = 1, \ldots, N$ are independent and potential outcomes of one region are independent of other regions' potential outcomes and treatments.

We estimated $\hat{\mu}_i(w)$ for each region i previously. To obtain the aggregated estimate, we assumed

$$\hat{\mu}_i(w) = \mu(w) + \epsilon_i + \tau,$$

where $\epsilon_i \sim N(0, s_i)$ is within study error and $\tau \sim N(0, v)$ is between study error. s_i is estimated by bootstrap, so the estimation of v remains. We estimated v, and aggregated estimates from 36 regions by taking the weighted average of region-specific effect estimates where weight is inversely proportional to the variance $s_i + v$ of the estimator,

$$\hat{\mu}(w) = \sum_{i=1}^{N} \frac{\hat{\mu}_i(w)}{s_i + \hat{v}} / \sum_{i=1}^{N} \frac{1}{s_i + \hat{v}}.$$

The precision of pooled estimator is the sum of the precisions of region-specific estimators,

$$\frac{1}{\hat{\sigma}^2} = \sum_{i=1}^{N} \frac{1}{s_i + \hat{v}}.$$

Confidence interval is obtained by

$$\hat{\mu}(w) \pm 1.96\hat{\sigma}.$$

See the upper right panel of figure 2.

We could do multivariate meta-analysis like the DLNM framework with bootstrap covariance, but its computation cost was too high, so we just considered univariate meta-analysis.

4.3 Result

In figure 2, the upper left panel is a logRR curve obtained under the DLNM framework; the upper right panel is obtained by applying RCM; the lower left panel is smoothed version of the upper right panel (kernel: Gaussian, bandwidth: 6); the lower right panel is a logRR curve without adjusting the temporal trend under the DLNM framework.

The estimated logRR curve of the lower right panel has exaggerated values at extreme temperatures, compared to the upper left panel. Since the model of the lower right panel does not consider a temporal trend, the difference between the two panels comes from the autocorrelation of the outcome variable.

The logRR curve of the upper right panel is spiky because we estimated it by model-free method, so it heavily depends on the observations. For the coldest temperature, we can see

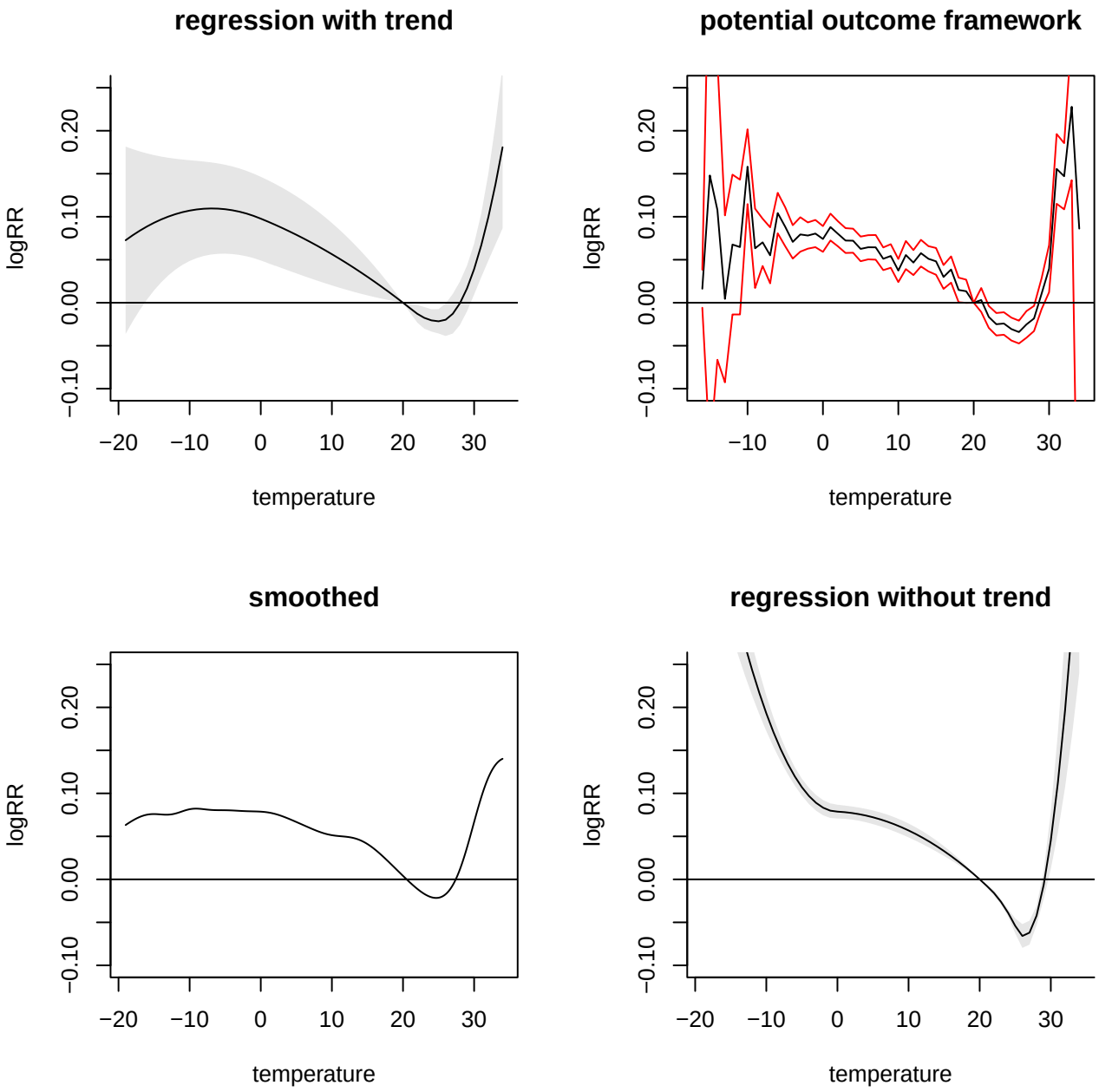

Figure 2: Estimated overall effect. For extreme hot or cold temperatures, estimated effects have quite large uncertainties.

that the confidence interval is narrow compared to the hottest temperature. This is because there is only one such observation, so uncertainty captured by bootstrap is due to the variation of effect estimate at the reference temperature. Moreover, the standard error of pooled logRR curve at extreme temperatures seems larger than the one from the DLNM framework, because such extreme cases did not happen in some regions.

In the lower-left panel, we applied kernel smoothing to our estimate to remove spikes. The smoothed curve and the curve of the upper left panel have similar values compared to the curve of the lower right panel. From this point of view, we may say RCM can adjust temporal confounding bias to some extent. Moreover, we don't know what is the true logRR curve, but we may insist that the logRR curve obtained from RCM is more general in the sense that it becomes similar to the curve of DLNM after kernel smoothing.

5 Discussion

In this article, the DLNM framework and Rubin causal model were compared to discover a causal link between ambient temperature and all-cause mortality. To the extent of my knowledge, there has been no study analyzing the short-term relationship between the ambient temperature and the all-cause mortality using RCM. Two pooled results obtained via two different methods showed similar effects of the ambient temperature, after smoothing. This similarity or consistency added evidence of a causal relationship found by previous studies.

In RCM, we could not do two things that the DLNM framework can do: identification of the lagged effect, and distribution-based confidence interval. The exposure-response surface that the DLNM framework produces makes us possible to identify lagged effect of the ambient temperature. In contrast, RCM cannot measure the lagged effect of the ambient temperature, since it does not use any information about lagged temperature. Indeed, there is a method to measure the lagged effect of binary treatment in RCM [2], but the curse of dimension inhibits the direct application of that method to our case. The DLNM framework provides the confidence interval based on asymptotic normality followed by maximizing the quasi-likelihood function. Unfortunately, there is no exact or asymptotic distribution of our estimator in our case, due to the difficulty derived from the definition of RR, which is the ratio between two effects, not the difference.

In addition to the weakness compared to the regression method stated above, our method has disadvantages of itself to be addressed: incorrect treatment assignment mechanism, and

violation of assumptions. We predicted daily mean temperature with only date, despite the existence of other factors such as cloud, rain, air mass, typhoon, CO2 emission, global warming, etc, because the primary factor that determines daily mean temperature is the meridional altitude. So the assignment mechanism we assumed was theoretically incorrect, but can be improved by including other factors that are able to predict the temperature in the GPS model. There is a possibility of violation on two key assumptions 2,3 in RCM. The first one is the positivity assumption that any treatment has a positive probability of being assigned for each confounder. In our case, the confounder is time and there is always a nonzero probability of extreme temperature because of catastrophic events. However those events rarely happen in reality, so stochastic positivity violation [21] can happen due to the small sample size. To fix the issue, we made a normal assumption on GPS and trimmed a very small probability to 0.1 to ensure stability. The second one is (weak) unconfoundedness which may be violated when there is an unmeasured confounder. In our case, we cannot measure everything related to the temperature-mortality relationship so it is plausible to think the assumption 3 is violated. To overcome this issue, we should carefully think about confounding structure and collect more potential confounders.

References

[1] Joshua D. Angrist, Òscar Jordà, and Guido M. Kuersteiner. Semiparametric estimates of monetary policy effects: String theory revisited. *Journal of Business & Economic Statistics*, 36(3):371–387, 2018.

[2] Iavor Bojinov and Neil Shephard. Time series experiments and causal estimands: Exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682, 2019.

[3] A. Gasparrini, B. Armstrong, and M. G. Kenward. Distributed lag non-linear models. *Statistics in Medicine*, 29(21):2224–2234, 2010.

[4] Antonio Gasparrini. Modelling Lagged Associations in Environmental Time Series Data: A Simulation Study. *Epidemiology*, 27(6), 2016.

[5] Antonio Gasparrini, Yuming Guo, Masahiro Hashizume, Eric Lavigne, Antonella Zanobetti, Joel Schwartz, Aurelio Tobias, Shilu Tong, Joacim Rocklöv, Bertil Forsberg, Michela Leone, Manuela De Sario, Michelle L Bell, Yue-Liang Leon Guo, Chang-Fu Wu, Haidong Kan, Seung-Muk Yi, Micheline de Sousa Zanotti Stagliorio Coelho, Paulo Hi-

lario Nascimento Saldiva, Yasushi Honda, Ho Kim, and Ben Armstrong. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *Lancet*, 386(9991):369–375, July 2015.

[6] Gretchen T. Goldman and Francesca Dominici. Don't abandon evidence and process on air pollution policy. *Science*, 363(6434):1398–1400, 2019.

[7] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[8] Guido Imbens and Keisuke Hirano. The propensity score with continuous treatments. 2004.

[9] Guido W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

[10] Yoonhee Kim, Ho Kim, Antonio Gasparrini, Ben Armstrong, Yasushi Honda, Yeonseung Chung, Chris Fook Sheng Ng, Aurelio Tobias, Carmen Íñiguez, Eric Lavigne, Francesco Sera, Ana M. Vicedo-Cabrera, Martina S. Ragettli, Noah Scovronick, Fiorella Acquaotta, Bing-Yu Chen, Yue-Liang Leon Guo, Xerxes Seposo, Tran Ngoc Dang, Micheline de Sousa Zanotti Stagliorio Coelho, Paulo Hilario Nascimento Saldiva, Anna Kosheleva, Antonella Zanobetti, Joel Schwartz, Michelle L. Bell, and Masahiro Hashizume. Suicide and ambient temperature: A multi-country multi-city study. *Environmental Health Perspectives*, 127(11):117007, 2019.

[11] Hans R. Kunsch. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, 17(3):1217 – 1241, 1989.

[12] Hernán MA and Robins JM. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

[13] PAUL R. ROSENBAUM and DONALD B. RUBIN. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983.

[14] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[15] Donald B. Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007.

[16] R. W. M. WEDDERBURN. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3):439–447, 12 1974.

[17] X. Wu, D. Braun, J. Schwartz, M. A. Kioumourtzoglou, and F. Dominici. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. *Science Advances*, 6(29):eaba5692, 2020.

[18] Stanley Xu, Colleen Ross, Marsha A. Raebel, Susan Shetterly, Christopher Blanchette, and David Smith. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, 13(2):273–277, 2010.

[19] Xiaofang Ye, Rodney Wolff, Weiwei Yu, Pavla Vaneckova, Xiaochuan Pan, and Shilu Tong. Ambient temperature and morbidity: A review of epidemiological evidence. *Environmental Health Perspectives*, 120(1):19–28, 2012.

[20] Ghosh D. Zhu Y, Coffman DL. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, 3(1):25–40, 2015.

[21] Paul N Zivich, Stephen R Cole, and Daniel Westreich. Positivity: Identifiability and estimability. 2022.