Causal temperature-mortality relationship using time-series data

Jaemin Oh[1] and Yeonseung Chung[2]

[1,2]Department of Mathematical Sciences, KAIST

October 24, 2022

Abstract

Temperature-mortality relationship has been analyzed by using regional time series data under the DLNM framework. There has been many concerns about causal interpretation on the temperature-mortality relationship, because of unmeasured confounders, model misspecification, and mixing of the design and the analysis stages. In this paper, we used the potential outcome framework to deal with latter two issues, and obtained the consistent result compared to the previous studies. This work shines a light on the possibility of causal interpretation on 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](Gasparrini lancet, 김윤희 교수님 EHP more?). Usually, the relationship has been analyzed by using regional time series data. There are difficulties in analyzing the relationship with time series data, such as temporal trend of outcomes, and the existence of delayed effect. These difficulties have been addressed by the DLNM framework [3] which used the quasi-Poisson regression [15] to estimate coefficients. It produces 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. Such regression analyses have some advantages: lagged effect can be easily identified; it summarizes the information of the curve representing relative risk (RR) by spline coefficients, so meta-analysis is possible with considering the covariance of effects at different temperatures. These advantages make regression approach popular in environmental epidemiology, especially for the topic of temperature-mortality relationship.

However, there has been a concern about causal interpretation on the results obtained by regression analysis. Obviously, the primary issue is the existence of unmeasured confounders that cannot be solved without collecting additional data. Additionally, in recent debate on air pollution study [6], two aspects of regression analysis were pointed out: mixing of design stage and analysis stage that uses outcomes in confounding bias adjustment [14] and model misspecification problem. In fact, these problems remain the same in the context of the temperature mortality relationship; the DLNM framework mixes design stage and analysis stage, since it fits additional spline for each year to adjust for temporal confounding; it is susceptible to model misspecification [4] since we don't know the exact placement of knots and the exact degrees of freedom. Therefore, these two problems should be solved first to make causal interpretation possible.

As a solution, we suggests to use Rubin causal model (RCM) [7] known as the potential outcome framework. It separates the design stage and the analysis stage, and does not need any parametric model to outcome generating process, so it is more free to model misspecification. The potential outcome framework was first introduced to analyze the data of randomized experiments [13], but now widely used in observational studies [16], and even in time series data [1]. In this paper, we used potential outcome framework to estimate logRR curve of the ambient temperature and compared the result to the DLNM framework.

2 Method

The state of a region at time t can be described by (Y_t, W_t, C_t) where Y_t is a vector of outcome, W_t is daily mean temperature, and C_t is a vector of confounders.

2.1 Conceptual framework

What would be the value of Y_t if $W_t = w'$ had been observed instead of $W_t = w$? The outcome of this counterfactual imagination is called potential outcome and written as $Y_t(w')$. If we know true value of $Y_t(w)$ and $Y_t(w')$, we can get the relative risk of w against w' by

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

However, we never know the true $Y_t(w)$, due to its counterfactual nature. This is called the fundamental problem of causal inference [7].

There has been many studies to address this problem. In (marginally) randomized experiment, $\mathbb{E}[Y(w)]$ can be estimated from observed data [13]. In observational studies, one can estimate causal estimand

$\mu(w) = \mathbb{E}[Y(w)]$ by preprocessing the data to approximate randomization e.g., inverse probability weighting, standardization, matching [12]. Among those techniques, common assumptions that makes it possible to estimate the causal estimand are below:

Assumption 1 (Consistency)

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

Positivity assumption says all treatments are possible for each confounder. Weak unconfoundedness assumption says conditional on current confounders, potential outcomes are already determined. With these, the causal estimand can be calculated as

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

This is called "inverse probability weighting" (IPW). Thus, a natural estimator of the causal estimand is

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

Note that the first equality comes from the interated expectation formula and consistency assumption, the second equality comes from weak unconfoundedness assumption, and the last equality is due to the definition of $p(w|C_t)$. By positivity assumption, we can divide by $p(w|C_t)$.

Still, we need to estimate $p(w|C_t)$ since it is unknown to us in general. When the treatment is binary, $p(w|C_t)$ is called propensity score, and it is used to adjust for confounding bias [12]. 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 logit model to data. For categorical or continuous treatment, GPS can be estimated by fitting ordered probit model or boosting.

PS and GPS have two nice properties [8, 12]. The first one is 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.

The main reason why we use inverse probability weighting by GPS is to achieve covariate balance. In randomized experiment, distribution of covariates are similar across each treated group. However, we are now dealing with observational data so we are looking forward to achieve covariate balance on pseudo-population generated by imposing appropriate weights to each observation. One criteria for covariate balance is the absolute correlation (AC) [19]. If treatment and covariates are independent, then their correlation must be zero. So, small value of AC can be an one possible evidence of covariate balance. Usually, AC with < 0.1 is considered as acceptible.

In fact, matching method to adjust confounding bias is available for continuous treatment data [17]. However, in our case, extreme temperatures are rarely observed, so matching estimator can heavily depend on such observations and can lead exaggerated effect estimate. Instead, in IPW method, too large weight is trimmed so it will produce less exaggerated estimate compared to the matching method.

3 Application

The data has daily mean temperature in the first decimal place, and all-cause mortality count during the period from 1997 to 2018 across 36 regions in South Korea.

For $i = 1, \ldots, N$ and $t = 1, \ldots, T$, let $(Y_{i,t}, W_{i,t}, C_{i,t})$ be information of i-th region at time t. Y is daily total death count, W is daily mean temperature and C is (year, month, the week of year, the day of year). We rounded daily mean temperature to integer value.

To compare estimate obtained above to the previous studies, we fitted DLNM to our data. For temperature dimension, we used quadratic B-spline, and placed knots at 10th, 75th, 90th quantiles. For lag dimension, we considered 21 lags, used natural B-spline, and placed 3 knots at equally spaced log values. For temporal trend adjustment, we fitted additional natural B-spline with 8 degrees of freedom for each year.

3.1 Single series estimate

In this subsection in which we concentrate on single time series, we will drop i to simpify the notation. The first stage is design stage to adjust for time confounding. We adjusted confounding bias by stabilized inverse probability weighting [18] that assign weights

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

to each obsevation. Here, we don't know true probability densities, so we instead used estimated values. One can use normal density as denominator $p(W_t)$, but we used relative frequency as the denominator since weighting method with relative frequency achieved lower absolute correlation in this application. See the second row and third row of Table 1. We trimmed weights bigger than 10 by 10, because some estimated weights were too large so effect estimate was heavily dependent to those observations.

To estimate the numerator $p(W_t|C_t)$, we assumed that the conditional distribution of treatment given covariates is a normal distribution, i.e.

$$W|C \sim N(m(C), \sigma(C)^2)$$

where $m(C)$, and $\sigma(C)$ are some functions of C. Since they are unknown in general, we should estimate them. The mean $\hat{m}(C)$ is estimated by boosting, and the standard deviation $\hat{\sigma}(C)$ is estimated by boosting residuals [8, 19]. Hyperparameters such as depth of tree(3), shrinkage(0.1), and the number of trees(20) are determined heuristically to minimize the absolute correlation.

We calculated absolute correlation (AC) [19] to see whether the covariate balance is achieved. 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.

	year	month	the week of year	the day of 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 adjustment; the second row is AC after adjustment, with relative frequency of temperature as marginal probability; the third row is AC after adjustment, with normal assumption on marginal distribution of temperature.

After covariate balance is acheived (here, AC$<$ 0.1), we estimated the causal effect by Horvitz-Thompson estimator with stabilized weights,

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

With $\hat{\mu}(w)$, we calculaated logRR curve by $\log \hat{\mu}(w) - \log \hat{\mu}(20)$. Uncertainty of logRR curve is quantified by Moving Block Bootstrapping (MBB) [11]. In the bootstrap procedure, we did not fit gps model repeatedely for each bootstrap sample since we need to re-sample the pseudo-population itself that acheives covariate balance.

In fact, asymptotic confidence interval can be calculated if we concentrate on risk difference instead of risk ratio. However, there is no known result about asymptotic distribution of risk ratio. The reason why we care about the relative risk is 1) consistency to the previous studies 2) that the population sizes differ across regions. If we consider the risk difference, then additional suitable normalizing step will be required.

3.2 Pooling 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 outcome of one region is not affected by other regions' potential outcomes and treatments.

For each region i, we estimated $\hat{\mu}_i(w)$ in the previous section. To obtain 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 estimation of $V(\tau)$ remains. We used R package 'mixmeta' to obtain pooled estimate, which is the weighted average of region specific effect estimate where weight is inversely proportional to the variance $S_i + V$ of estimator. Precision of pooled estimator is sum of precisions of region specific estimator. Confidence interval is obtained by adding (or subtracting) 1.96 times of pooled standard deviation $\hat{\sigma}$ to pooled logRR curve.

3.3 Result

In figure 1, the upper left panel is a logRR curve obtained under the DLNM framework; the upper right panel is obtained by applying potential outcome framework; the lower left panel is smoothed version of upper right pannel (kernel: Gaussian, bandwidth: 6); the lower right pannel is a logRR curve without adjusting 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 temporal trend, the difference between two panels comes from autocorrelation of 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 most cold temperature, we can see that the confidence interval is narrow compared to the most hot 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.

In the lower left pannel, 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 potential outcome framework can adjust temporal confounding bias in some extent. Moreover, we don't know what is the true logRR curve, but we may insist that the logRR curve obtained from potential outcome framework is more general in the sense that it becomes similar to the curve of DLNM after kernel smoothing.

4 Discussion

In 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 the potential outcome framework. In this work, the causal link between the ambient temperature and mortality is found under the potential outcome framework. Consistency between new approach and existing regression method reinforces usefulness of regression method, and adds an evidence of causal relationship found by previous studies. However, there are some limitations to overcome.

4.1 Limitations

In this framework, we could not do several analyses that the DLNM framework can do: lagged effect analysis, theoretical confidence interval, and meta-regression.

The DLNM framework produces exposure-response surface with treatment dimension and lag dimension. Therefore, we can easily measure the lagged effect of any treatment level after obtaining the surface. In contrast, we estimated the effect of the treatment by using the current outcome variable. This can produce only exposure-response curve, since we don't use any information about lagged treatment or lagged outcome. Thus our approach does not have ability to measure the lagged effect of the treatment. There is a method to measure the lagged effect of the treatment in a single time series [2]. However, this paper is for binary (sequential) treatment so it is hard to be directly applied to continuous treatment case, because of the curse of dimensionality. With maximum lag L, the number of combinations of possible treatment path is 2^{L+1} for binary case but there are more than 50 categories of daily mean temperature so the number of treatment path exceeds 50^{L+1}. Even if $L = 2$, the number of possible combinations of treatment assignment for 3 days exceeds the length of series 8000. Note that we used $L = 21$ for the DLNM framework. Since we did not account for lagged effect, someone may say to us that the calculated logRR curve does not represent overall effect but represents instant effect. However, due to high autocorrelation of daily mean temperature during short period, we assumed that treatment history would have

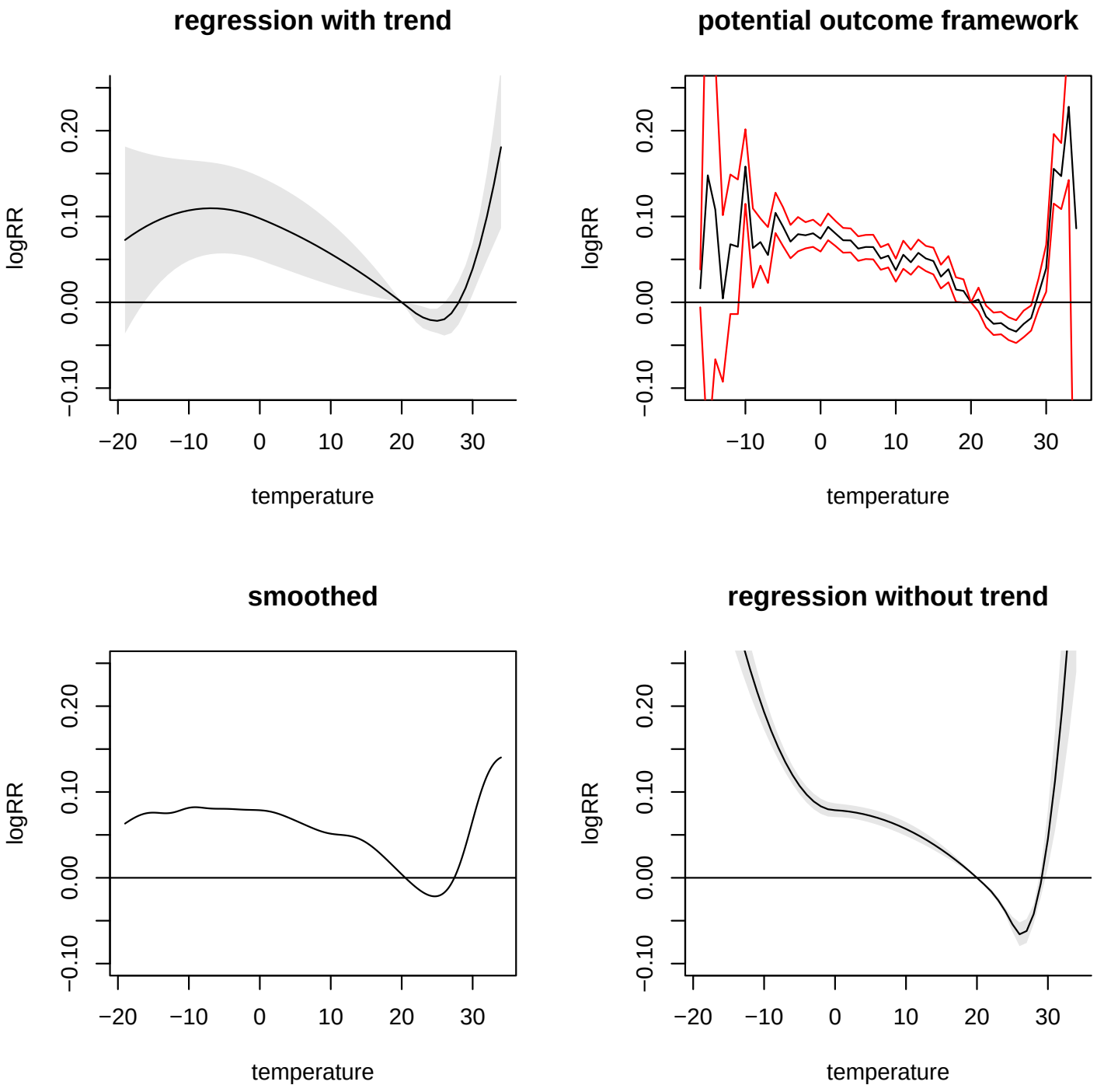

Figure 1: Estimated overall effect. For extreme hot or cold temperature, estimated effects have quite large uncertainties.

been almost the same during short period, so our effect estimate represents overall effect.

Traditional regression method quantifies its uncertainties based on asymptotic normality, because it is fitted by maximizing quasi-likelihood function. However, in our case, we quantified the uncertainty of estimated logRR curve by moving block bootstrap to include information about temporal correlation. This method has a crucial disadvantage. When we have only one observation for treatment w, its uncertainty based on MBB is zero. It contradicts to our intuition that larger sample size gives lower uncertainty. We used MBB method because there is no other useful way to estimate standard error of logRR curve. The difficulty comes from the definition of RR. It is a ratio between two effect estimate, not the difference. Moreover, the estimated standard error of pooled logRR curve is not similar to the one from the DLNM framework. This is because, the precision of overall effect was obtained by summing up all precisions from each region, but some regions did not have observations of extremal temperature. This leads us to relatively low precision of overall effect estimates in extreme temperature compared to the precision obtained from the DLNM framework.

It is well known that the temperature-mortality relationship is heterogeneous across regions. (citation) The heterogeneity was usually explained by meta-regression that has spline coefficients obtained from the DLNM framework as response variable, and regional level variables such as latitude as meta-predictors. Intuitively, spline coefficients represents a curve. So meta-regression with coefficients means analyzing the heterogeneity of curve itself across regions. However, we just pooled the effect estimates from each region by naively taking weighted average for each treatment value. This approach has two disadvantages compared to the meta-regression. The first one is, it does not use the information near that treatment value, since we don't have the covariance matrix of estimated effects. Another is, our approach reflects heterogeneity across regions by the simple random effect meta-analysis model, but cannot explain this heterogeneity by regional level predictors since meta-regression technique is hard to be applied to our effect estimates.

In addition to the weekness compared to the regression method stated above, our method has several disadvantages to be addressed: incorrect treatment assignment mechanism, high variance, and violation of assumptions.

When the treatment assignment is conditionally randomized, we can use inverse probability weighting to generate a pseudo-population that treatment assignment is marginally randomized. I would say in the context of temperature-mortality relationship, daily mean temperature is conditionally randomized. Because from the viewpoint of the Earth, the "assignment mechanism" of daily mean temperature is heavily dependent on the meridinal altitude. Moreover, the meridinal altitude is able to be predicted almost perfectly by the date. So, we can say that conditional on the date, daily mean temperature is

randomly determined where the randomness comes from cloud, rain, air mass, typhoon, CO2 emission, global warming, etc. It would be very good if we can include those factors into our GPS model to adjust for confounding bias of such meteorological variables, but it is impossible because of practical issue. Rather, by exploiting the fact that there are so many factors that may have influence on daily mean temperature, we gave normal assumption on the distribution of daily mean temperature based on the central limit theorem.

We estimated logRR curve without any modeling assumption after confounding adjustment. Also, the estimator does not borrow information near given treatment value. This makes the estimate soley rely on observed values. However, we have few observations for extreme temperature. So few observations play an important role in effect estimating procedure, and it means our approach may have high variance and low bias. It is the same for previous studies that there are only few extreme temperature observations, but they assumed parametric model to the outcome generating process so effect estimates at extreme temperature is a result of borrowing information near the temperature point, and it makes extreme cases play somewhat shrinked role compared to our approach, which means lower variance and higher bias.

There is a possibility of violation on two key assumptions in the potential outcome framework. The first one is unconfoundedness that potential outcome and treatment are independent given measured confounders. This assumption may be violated when there is an unmeasured confounder, since unmeasured confounder can change the distribution of potential outcome. In our case, we cannot measure everything related to temperature-mortality relationship so it is plausible to think unconfoundedness assumption is violated. The second one is positivity (overlap) assumption that any treatment has positive probability of being assigned for each confounder. In our case, the confounder is time and there is always some chance of extreme temperature because of catastrophic events in theory. However those events rarely happen in reality, so stochastic positivity violation [20] can happen. The reason is we don't have enough sample size. To address this issue, we made normal assumption on gps, and trimmed very small probability to 0.1 to ensure stability.

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, Francesco Sera, Ana Maria Vicedo-Cabrera, Veronika Huber, Shilu Tong, Micheline de Sousa Zanotti Stagliorio Coelho, Paulo Hilario Nascimento Saldiva, Eric Lavigne, Patricia Matus Correa, Nicolas Valdes Ortega, Haidong Kan, Samuel Osorio, Jan Kyselý, Aleš Urban, Jouni J K Jaakkola, Niilo R I Ryti, Mathilde Pascal, Patrick G Goodman, Ariana Zeka, Paola Michelozzi, Matteo Scortichini, Masahiro Hashizume, Yasushi Honda, Magali Hurtado-Diaz, Julio Cesar Cruz, Xerxes Seposo, Ho Kim, Aurelio Tobias, Carmen Iñiguez, Bertil Forsberg, Daniel Oudin Åström, Martina S Ragettli, Yue Leon Guo, Chang fu Wu, Antonella Zanobetti, Joel Schwartz, Michelle L Bell, Tran Ngoc Dang, Dung Do Van, Clare Heaviside, Sotiris Vardoulakis, Shakoor Hajat, Andy Haines, and Ben Armstrong. Projections of temperature-related excess mortality under climate change scenarios. *The Lancet Planetary Health*, 1(9):e360–e367, 2017.

[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] 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.

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

[14] Donald B. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808 – 840, 2008.

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

[16] 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.

[17] Xiao Wu, Fabrizia Mealli, Marianthi-Anna Kioumourtzoglou, Francesca Dominici, and Danielle Braun. Matching on generalized propensity scores with continuous exposures, 2018.

[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] 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.

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