

# Causal temperature-mortality relationship using time-series data

Jaemin Oh<sup>1</sup>

<sup>1</sup>Department of Mathematical Sciences, KAIST

October 18, 2022

## Abstract

In this paper, we analyzed the causal relationship between the ambient temperature and all-cause mortality in South Korea using the regional time-series data.

## 1 Introduction

Identifying the causal effect of the ambient temperature on human health is an important research topic due to global warming and climate change. The best way to identify the causal effect would be observing how the change of the ambient temperature affects the change of total death count in one region under Ceteris Paribus condition. Note that, perfect Ceteris Paribus condition is only possible when we can access to the parallel world that another treatment would have been assigned while other variables are the same. However, it is definitely impossible. Therefore, we must identify the causal effect by analyzing time series data which is a record of date, daily mean temperature, and daily total death count in a region. Analysis of this time series data is a challenging task, because of several reasons: the ambient temperature is continuous treatment, so the size of the effect may be differ for each temperature; the effect of the ambient temperature may be delayed; serial correlation of daily mean temperature makes difficult to separate instant and delayed effect; daily death count has a temporal trend, it increases as the population size increases. In previous studies (Gasparrini, Guo, etc), these difficulties were addressed by the combination of the DLNM framework [3] and the quasi-Poisson regression [14] that produces exposure-response surface  $\mu : (w, l) \mapsto \mu(w, l)$  where  $w$  is the ambient temperature and  $l$  is a time lag.

The DLNM framework has many advantages: it summarizes the information of the RR curve by spline coefficients, so multivariate meta-analysis is possible; lagged effect can be easily identified; well developed R-package is available [4]. These advantages make the DLNM framework popular in environmental epidemiology, especially for the topic of temperature-mortality relationship. However, model selection in the DLNM framework is quite complicated so we may say that it is susceptible to model misspecification [5]. It eliminates temporal trend of death count by fitting additional spline for each year, and it requires determination of degrees of freedom for each spline base. Also we never know the exact degrees of freedom and placement of knots for spline bases of the ambient temperature. Those values can be determined based on the previous studies, or Akaike information criterion [3].

There was a debate on air pollution study that usual regression approach to analyze the health effect of particulate matter does not inform causality very well [6]. One reason is, as stated above, model selection problem and another reason is mixing of design stage and analysis stage. Rubin said that final outcome data should not be used in design stage without some exceptions [13]. Here, in the context of the DLNM framework, design stage is fitting splines for each year to adjust confounding bias, and analysis stage is fitting the cross basis to the data. The DLNM framework is done by doing Poisson regression to the data, stated concerns about regression analysis is still valid in temperature-mortality relationship.

A model that separates design and analysis stage is Rubin causal model (RCM) [7]. It is about how to compare counterfactual outcomes of another parallel world, without access to the parallel world under suitable assumptions. The potential outcome framework was first introduced to analyze the data of randomized experiments [12], but now widely used in observational studies [15]. Furthermore, the potential outcome framework is more free to the problem of model misspecification than the traditional approach, because it is in some sense "semi-parametric" that does not require parametric model for outcome generating process. [1] So in this paper, we used potential outcome framework to estimate logRR curve of the ambient temperature.

Usually, the potential outcome framework is applied to individual level data. For example, analyzing treatment effect of new medicine. However, it is hard to earn individual level data for temperature-mortality relationship, to say that the true cause of death was the ambient temperature, and to analyze short-term effect on individual level. Therefore we used regional data instead of individual level data.

## 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 a treatment, and  $C_t$  is a vector of confounders.

In regression settings, we usually have fitted the model:

$$g(\mathbb{E}[Y_t|W_t, C_t]) = s(W_t)\beta + \eta(C_t)\gamma$$

where  $g$  is a link function and  $(s, \gamma)$  are bases of function space, such as spline bases. However, as described before, regression approaches inform weak causality because of mixing of design and analysis stages and model selection problem. To deal with this problem, we used potential outcome framework with nonparametric effect estimator.

### 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

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

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 [12]. 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. [11] 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,  $Y_t(W_t) = Y_t$ .*

Without consistency assumption, we cannot know potential outcomes at all.

#### Assumption 2 (Positivity)

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

**Assumption 3 (Weak Unconfoundedness)**

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

This assumption says that, conditional on current confounders, potential outcomes are already determined. With this, the causal estimand can be calculated as

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

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 iterated 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 [11]. 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, 11]. 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 needed.

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) [18]. 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 acceptable.

In fact, matching method to adjust confounding bias is available for continuous treatment data [16]. 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

(DATA DESCRIPTION)

For  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , 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). In the next subsection in which we concentrate on single time series, we will drop  $i$  to simplify the notation.

#### 3.1 Single series estimate

The first stage is design stage to adjust for time confounding. We adjusted confounding bias by stabilized inverse probability weighting [17] by weighting each observation by

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

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. (report the difference) 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, 18]. Hyperparameters such as depth of tree, shrinkage, and the number of trees are determined to minimize the absolute correlation. (report them)

We calculated absolute correlation (AC) [18] 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.

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

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

With  $\hat{\mu}(w)$ , we calculated logRR curve by  $\log \hat{\mu}(w) - \log \hat{\mu}(20)$ . Uncertainty of logRR curve is quantified by Moving Block Bootstrapping (MBB) [10]. In the bootstrap procedure, we did not fit gps model repeatedly for each bootstrap sample since we need to re-sample the pseudo-population itself that achieves 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

(Assumption 4) - independence of regions

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

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.

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 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 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 studies to analyze the short term relationship between the ambient temperature and the all-cause mortality using the potential outcome framework.

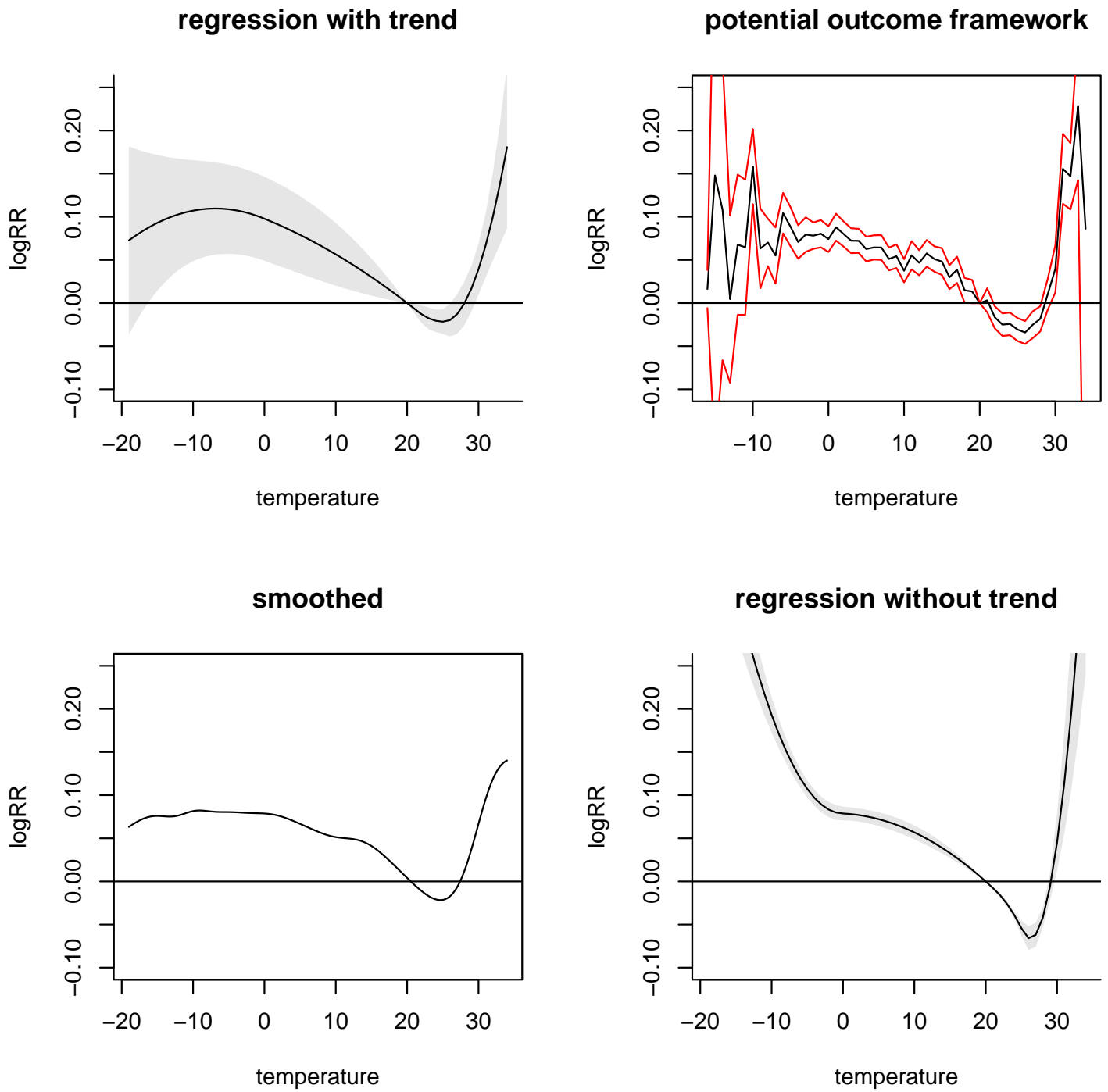


Figure 1: Estimated overall effect. For extreme hot or cold temperature, estimated effects have quite large uncertainties.



## 4.1 Limitation

When estimating the effect of extreme temperature, we heavily relied on very few observations of extreme cases. 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 estimated by borrowing information near the temperature point, and it makes extreme cases play somewhat shrinked role.

Since the gps model does not contain other meteorological confounders, it may not adjust for other confounding bias. We are just looking forward to bias be adjusted by weighting. This naive anticipation can be proven after collecting more region level data such as economical status, particulate matter, the ratio of elderly, ratio of green land, existence of sea or river nearby, etc.

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.

There may be a positivity issue, because of serial correlation. Theoretically, there is always some chance of extreme temperature because of catastrophic events. However that event rarely happens in reality, so stochastic positivity violation [19] can happen. To address this issue, we made normal assumption on gps, but the probability of temperature that is far from  $m(C)$  is still very small.

When pooling the effect estimates from each region, we just naively took weighted average. However, it is well known that the temperature-mortality relationship is heterogeneous across regions. The heterogeneity was usually explained by meta-regression that has spline coefficients as response variable, and regional level variables as meta-predictors. 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 that are high dimensional.

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, in 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, 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 assumed normal distribution on the distribution of daily mean temperature based on the central limit theorem.

## 4.2 Future work

The primary issue to overcome is the curse of dimensionality. Refer to [2], causal time series analysis is investigating the relationship between treatment paths and potential outcomes, not just a single treatment and corresponding potential outcomes. For binary treatment case, the number of combinations of possible treatment path is  $2^{L+1}$ , where  $L$  is the maximum lag. However, in the context of temperature-mortality relationship, the treatment is continuous so it is hard to consider treatment path even after categorizing the treatment to integers. In our work, there are 50 categories of daily mean temperature so the number of treatment path would have been  $50^{L+1}$  that is greater than the length of series for  $L = 2$ .

Even though it is log scale, the estimated effects for cold and hot temperature are quite large compared to previous studies. I think they are somewhat exaggerated because we used current temperature only to estimate lagged effect ( $Y_{t+l} \leftarrow W_t$ ), but previous studies used previous temperature observations to fit the current outcome ( $Y_t \leftarrow (W_t, \dots, W_{t-L})$ ). So our estimated lagged effect must be big due to serial correlation of temperature. For example, imagine that daily mean temperature at time  $t$  was  $W_t$ . For  $t'$  near  $t$ , say  $t' \in (t - L, t + L)$ ,  $W_{t'}$  would have been not much differ to  $W_t$ . This leads ( $Y_{t+l} \leftarrow W_t$ ) to be exaggerated. This is kind of small sample size issue that observations are not enough to average out the effects of temperatures after  $t$ . One way to remove the bias is divide by suitable coefficient [2], but there is no standard method to determine the coefficient. Further, even if we just use a uniform coefficient, there are too many possible combinations of treatment path (we are now dealing with continuous treatment) so the coefficient will be zero.

## 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. Distributed lag linear and non-linear models in r: The package dlnm. *J. Stat. Softw.*, 43(8):1–20, July 2011.
- [5] Antonio Gasparrini. Modelling Lagged Associations in Environmental Time Series Data: A Simulation Study. *Epidemiology*, 27(6), 2016.
- [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] Hans R. Kunsch. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, 17(3):1217 – 1241, 1989.
- [11] 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.
- [12] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.
- [13] Donald B. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808 – 840, 2008.
- [14] R. W. M. WEDDERBURN. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3):439–447, 12 1974.
- [15] 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.
- [16] Xiao Wu, Fabrizia Mealli, Marianthi-Anna Kioumourtzoglou, Francesca Dominici, and Danielle Braun. Matching on generalized propensity scores with continuous exposures, 2018.

- [17] 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.
- [18] 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.
- [19] Paul N Zivich, Stephen R Cole, and Daniel Westreich. Positivity: Identifiability and estimability. 2022.