REAL-TIME CAUSAL INFERENCE

BY KWEKU OPOKU-AGYEMANG

Center for Effective Global Action
Department of Agricultural and Resource Economics
University of California, Berkeley
Berkeley, California, U.S.A.
and Cornell Tech, New York, U.S.A.
kweku@berkeley.edu

Abstract: The paper highlights several areas on statistical measurement and causal inference in common real-time data environments. Treatment effects under real-time randomization within data streams are found estimable using controlled and natural experiments motivated by real-time regression analyses. A bias occurs as a result of ignoring concept drift when classical regression statistics are naïvely applied to real-time experimental data. An algorithm performs difference-in-difference estimation for real-time program evaluations. A new Problem of Causal Inference is introduced for real-time data environments. The paper closes with brief implications.

Keywords: Causal inference; Real-Time Data; Randomized Experiments; Natural Experiments.

JEL: C10, C18.

Acknowledgments: I thank several people at the Center for Effective Global Action, the Department of Agricultural and Resource Economics, the Berkeley Institute for Data Science, the Berkeley Institute for Transparency in Social Science, the Development Engineering Seminar and the Mechanical Engineering Department, all of the University of California, Berkeley as well as the Working Group in African Political Economy for various informal discussions while writing the paper. The author is solely responsible for this article and its implications: the usual disclaimer applies.

1. Introduction

This paper discusses approaches of measurement and causal inference within the context of real-time data, which is information delivered immediately after collection. Such data is increasingly ubiquitous in a rapidly growing segment of the knowledge economy that collects vast amounts of consumer and firm information. Examples include networked sensors in smart phone devices, social media platforms, personalized mobile health technologies and ride-sharing initiatives, all of which are based on the premise of nearly constantly generating data. To understand how to analyze such information, the paper is motivated by important results in data streams analyses in computer science and engineering (Zhu and Shaha, 2002, Babcock et al 2002, Nadungodage et al 2011), but that have not yet been harnessed within the econometrics of causal inference literature. In such settings, there is little to no delay in the timeliness of the data as well as the statistical and computational problems of data measurement and analysis and experimenters must collect data and create model parameters instantaneously. Real-time variation has not yet been discussed in program evaluation to the best of my knowledge.

While helping close this gap in the literature, the paper builds on a growing number of methods are arising to analyze high-dimensional statistical data environments (e.g. Belloni and Chernozhukov (2009); Belloni, Chen, Chernozhukov, and Hansen (2012); Belloni, Chernozhukov, Hansen, and Fernandez-Val (2013a); Belloni, Chernozhukov. and Hansen (2013b); also see Liran and Einav (2014); King (2014) and Varian (2014)). Real-time data may also be important for environments based on continuous time (e.g. Dragoman and Dragoman 2004) and technology-driven data collection (Couper, 2005) as well as experimental settings where study subjects may behave in unintended ways that require researcher reactions (Chassang et al, 2010). Finer measurements at the level of real-time settings are further motivated by instances where being surveyed at baseline can by itself significantly affect parameter estimates (Zwane et al, 2011). Other potential applications include networks (e.g. Fowler and Christakis (2008)), mobile telephony (e.g. Kamenica, Mullanaithan and Thaler (2011) that have not yet been extended to real-time settings as well as emerging research areas such as precision medicine and other health care innovations (see Stern, Alexander,

and Chandra, 2017). We provide simple approaches for real-time experimentation that are broadly reconciliable with datastreams which rapidly provide data for experimental and other program evaluations.

Real-time data may be observed whereby the arrival of new data always replaces the old data (as assumed in the motivation section) or always co-exists the old data (as assumed in the full analysis later). We first assume the existence of randomized treatments T_i that may vary in real-time. We also assume that the real-time assignment probabilities do not depend on the potential outcomes in the vein of the Rubin Causal Model (RCM), also called the Neyman-Rubin model (Rubin, (1974); also note Splawa-Neyman (1990), Imbens, Angrist and Rubin (1991); Imbens and Wooldrige (2009)).

The paper now proceeds. We first illustrate how treatment effects may be studied in real-time data. Let $\mathcal{D}_i$ or $D[i]$ denote the value of datastream $\mathcal{D}$ at timepoint i. For timepoints i through j inclusive, the corresponding stream is $\mathcal{D}\left\{(X_i, T_i, Y_i)\right\}_{i=1}^{n}[i...j]$. The stream is summarized as $\mathcal{D}[i...j]$. A stream with stream identifier i is denoted $\left[\mathcal{D}\left\{(X_i, T_i, Y_i)\right\}_{i=1}^{n}\right]^i$.

We let the real-time data be represented in the following form:

$$\mathcal{D} = \{(X_i, T_i, Y_i)\}$$

which are independent identically distributed triples (X_i, T_i, Y_i)

We let $X \in \mathbb{R}^p$ represent the subject characteristics measured at baseline; $T \in \mathcal{T}$ refers to the binary or continuous treatment received by the subjects; and the outcome is summarized by $Y \in \mathbb{R}$. The real-time data is generated according to the linear form $Y_i = \alpha + \beta T_i + \epsilon_i$ for a constant α and standard error term ϵ_i. The details of causal inference cross-sectional data modeling are in Rubin (1978) and Imbens and Rubin (2015).

For a window of interest $[p, q]$, we represent the statistics of interest by

$$stat\left(\mathcal{D}_j^{i_1}, \mathcal{D}_j^{i_2}, ..., \mathcal{D}_j^{i_k}, j \in [p, q]\right)$$

or

$$stat\left(\left(\mathcal{D}\left\{\left(X_i,T_i,Y_i\right)\right\}_{i=1}^{n}\right)_j^{i_1},\left(\mathcal{D}\left\{\left(X_i,T_i,Y_i\right)\right\}_{i=1}^{n}\right)_j^{i_2},...,\left(\mathcal{D}\left\{\left(X_i,T_i,Y_i\right)\right\}_{i=1}^{n}\right)_j^{i_k}\right)$$

A treatment rule is the map $\pi : \mathbb{R}^p \longrightarrow \mathcal{T}$ so that a subject observed with $X = x$ receives treatment $\pi(x)$. The potential outcome under treatment $t \in \mathcal{T}$ is $Y^*(t)$so that the potential outcome under π is defined as $Y^*(\pi) = Y^*\{\pi(X)\}$. The marginal mean outcome is $E\{Y^*(\pi)\}$ and the conditional density of T given X is $p(t|X)$. These definitions are standard.

We generate the real-time treatment effects as follows. The average treatment effect in real time over $[p,q]$ is given by the difference between the expectations of real-time outcomes attributable to treatment t_1 and the expectations of real-time outcomes attributable to control t_0 over the same window.

$$\mathbb{E}\left(\mathcal{Y}(t_1)_j^{i_1},\mathcal{Y}(t_1)_j^{i_2},...,\mathcal{Y}(t_1)_j^{i_k},j\in[p,q]\right)$$

$$-\mathbb{E}\left(\mathcal{Y}(t_0)_j^{i_1},\mathcal{Y}(t_0)_j^{i_2},...,\mathcal{Y}(t_0)_j^{i_k},j\in[p,q]\right)$$

where the real-time outcomes attributable to the treatment are documented as

$$\mathcal{Y}(t)_j^{i_1},\mathcal{Y}(t)_j^{i_2},...,\mathcal{Y}(t)_j^{i_k}$$

and the real-time outcomes attributable to the control are given by

$$\mathcal{Y}(c)_j^{i_1},\mathcal{Y}(c)_j^{i_2},...,\mathcal{Y}(c)_j^{i_k}.$$

If the above can be computed, we identify this formula as the Rubin Causal Model in real-time. Unit-level real-time causal effects are

$$\mathbb{E}\{\left(\mathcal{Y}(t_1)_j^{i_1},\mathcal{Y}(t_1)_j^{i_2},...,\mathcal{Y}(t_1)_j^{i_k}\right)$$

$$-\left(\mathcal{Y}\left(t_0\right)_j^{i_1},\mathcal{Y}\left(t_0\right)_j^{i_2},...,\mathcal{Y}\left(t_0\right)_j^{i_k},\right)\}$$

for $j \in [p,q]$. The real-time Population Average Treatment effect on the treated is defined by

$$\mathbb{E}\{\left(\mathcal{Y}\left(t_1\right)_j^{i_1},\mathcal{Y}\left(t_1\right)_j^{i_2},...,\mathcal{Y}\left(t_1\right)_j^{i_k}\right)$$

$$-\left(\mathcal{Y}\left(t_0\right)_j^{i_1},\mathcal{Y}\left(t_0\right)_j^{i_2},...,\mathcal{Y}\left(t_0\right)_j^{i_k}\right)\}|\left(t_0\right)_j^{i_1},\left(t_0\right)_j^{i_2},...,\left(t_0\right)_j^{i_k}$$

for $j \in [p,q]$.

The focus on real-time environments requires detailed statistical definitions on time. The definitions are in the vein of Zhu and Shasha (2002). A *timepoint* is the smallest unit of time over which data collection occurs (e.g. a second). A *basic window* is a consecutive subsequence over which a digest of data is maintained incrementally (e.g. a few minutes). A *sliding window* is a user-defined sub-sequence of basic windows over which the user requires statistics (e.g. a hour). To illustrate, a health scientist may generalize the analysis of Green and MacDonald (1981) on the static effect of intravenous glucose on body temperature to real-time data by asking, "what are the real-time effects of intravenous glucose T on body temperature outcome Y for treatment patients relative to control patients for the last twelve hours?"

Assuming a length of a sliding window w and current timepoint t, for a specific sliding window q, we compute $stat\left(\mathcal{D}\left(X_i,T_i,Y_i\right),q\left(k\right)\right)$in the subsequence $\mathcal{D}\left(X_i,T_i,Y_i\right)[t-w+1..t]$.

Consider the stream $\mathcal{D}_i, i = 1,\ldots,w$. We will assume that we are in a position to monitor the following descriptive statistics in the real-time data:

- Single stream statistics (e.g. average, standard deviation, best fit slope).
- Correlation coefficients:

$$corr\left(\mathcal{X},\mathcal{Y}\right)$$

- Correlation coefficients for all three variables $\mathcal{Y},\mathcal{T},\mathcal{X}$:

$$corr\left(\mathcal{Y},\mathcal{T},\mathcal{X},\right)$$

- Autocorrelation (the correlation of the series with itself at an earlier time)
- Beta: the sensitivity of a stream $\mathcal{T}$ to the values of another stream $\mathcal{Y}$ (or weighted collection of streams).

$$beta\left(\mathcal{X},\mathcal{Y}\right)$$

To compute statistics over a sliding window, we maintain a synopsis data structure for the stream to compute the statistics in a timely manner. Our framework subdivides the sliding windows equally into shorter basic windows, to facilitate the efficient elimination of old data and the incorporation of new data. Digests are kept for both basic and sliding windows.

Let w be the size of a sliding window, b be the size of a basic window, k be the number of basic windows within a sliding window. The data within a sliding window is $\mathcal{D}\left(X_i,T_i,Y_i\right)\left[t-w+1..t\right]$. From the above, $w=kb$.

Let $\mathcal{D}\left(X_i,T_i,Y_i\right)\left[0\right],...,\mathcal{D}\left(X_i,T_i,Y_i\right)\left[k-1\right]$ represent a sequence of basic windows, whereby

$$\mathcal{D}\left(X_i,T_i,Y_i\right)\left[i\right]=\mathcal{D}\left(X_i,T_i,Y_i\right)\left[\left(t-w\right)+ib+1..\left(t-w\right)+\left(i+1\right)b\right]$$

The new basic window is $\mathcal{D}\left(X_i,T_i,Y_i\right)\left[k\right]$ and $\mathcal{D}\left(X_i,T_i,Y_i\right)\left[0\right]$ is the expiring basic window. The $j-$th value in the basic window $\mathcal{D}\left(X_i,T_i,Y_i\right)\left[i\right]$ is $\mathcal{D}\left(X_i,T_i,Y_i\right)\left[i;j\right]$.

The size of the basic window is important since in some cases the researcher may report all statistics for basic window i before basic window $i+1$ completes (at which point the researcher must compute statistics for window $i+1$.

The moving average and moving standard deviations are as follows. The information to be maintained for the moving average is

$$\sum \left(\mathcal{D}\left(X_i, T_i, Y_i \right) \left[t - w + 1 \ldots t \right] \right).$$

For each basic window $\mathcal{D}\left(X_i, T_i, Y_i \right) [i]$, we may maintain the digest

$$\sum \left(\mathcal{D}\left(X_i, T_i, Y_i \right) [i] \right) = \sum_{j=1}^{b} \mathcal{D}\left(X_i, T_i, Y_i \right) [i; \quad j].$$

After b new datapoints become available, we compute the sum over the new basic window $\mathcal{D}\left(X_i, T_i, Y_i \right) [k]$. The sum over the sliding window may be updated as:

$$\sum_{new} \left(\mathcal{D}\left(X_i, T_i, Y_i \right) \right) = \sum_{old} \left(\mathcal{D}\left(X_i, T_i, Y_i \right) \right)$$

$$+ \sum \mathcal{D}\left(X_i, T_i, Y_i \right) [k] - \sum \mathcal{D}\left(X_i, T_i, Y_i \right) [0]$$

Assume a series of covariates $\boldsymbol{s}^x$ as well as a series of error terms, $\boldsymbol{s}^\epsilon$. For causality, we assume strict exogeneity of the treatment series, so that $\boldsymbol{E}\left[\boldsymbol{s}^\epsilon | \boldsymbol{s}^x \right] = \mathbf{0}$ where $s^x = [\alpha, s^t]$, which is based on an assumed random assignment of T_i and of the treatment series s^t, which is implemented in a randomized manner and arrives at random. We will use covariate balance in real-time to verify the exogeneity of treatment streams throughout the real-time data.

2. Covariate Balance in Real-Time

We discuss cases where causal inference is based on background variables such as sociodemographic indicators or pure experiments, based on the standard assumptions of the Rubin Causal Model. Hansen and Bowers (2008) use Fisher's randomization inference when confirming covariate balance in static data, which may be done with stratification (e.g. Jin and

Rubin 2008) to ensure covariate balance. We focus this approach of discussing identification on real-time settings.

The index $i_k = (1, \dots n)_k$ refers to assignment units for each window k. The treatment assignment of the ith cluster of observation units at the k-th window is z_{ik}. The total number of x-values for observation units in cluster i is x_{ik}. The size of the cluster is c_{ik}

The observed difference between treatment and control groups in window k is $d_{pk}(\boldsymbol{z},\boldsymbol{x})$, so that:

$$d_{pk}(\boldsymbol{z},\boldsymbol{x}) = \left[\frac{\boldsymbol{z_k}^{\boldsymbol{t}}\boldsymbol{x_k}}{\boldsymbol{z_k^t c_k}}\right] - \left[\frac{\boldsymbol{(1-z_k)}^{\boldsymbol{t}}\boldsymbol{x_k}}{\boldsymbol{(1-z_k)}^{t}\boldsymbol{c_k}}\right]$$

and $d_{pk}(\boldsymbol{z},\boldsymbol{x})$ refers to the difference in treatment group averages and control group averages at each window k. If A_k is the set of treatment assignments from which the actual assignment $\boldsymbol{z_k}$ is selected, we have the following two-sided randomization $p-$value attaching to the hypothesis of nonselection on x as:

$$= \boldsymbol{P}\left(|d_{pk}(\boldsymbol{Z_k},\boldsymbol{x_k})| > |d_{pk}(\boldsymbol{Z_k},\boldsymbol{x_k})|\right) + \frac{1}{2}\boldsymbol{P}\left(|d_{pk}(\boldsymbol{Z_k},\boldsymbol{x_k})| = |d_{pk}(\boldsymbol{Z_k},\boldsymbol{x_k})|\right)$$

The random vector $\boldsymbol{Z_k}$ is uniformly distributed on possible treatment assignments A_k for all k. If simple random samples of n_{tb} are selected from n_b clusters for a treatment stream within blocks b, with the rest assigned to a control stream, we have:

$$d_k(\boldsymbol{z},\boldsymbol{x}) = \left\{\left(\frac{1}{\sum h_b \bar{c}_b}\right)\left[\sum_{b=1}^{B}\boldsymbol{Z}_b^t\boldsymbol{x}_b - \sum_{b=1}^{B} n_{tb}\left(\frac{\mathbf{1}^t\boldsymbol{x}_b}{n_b}\right)\right]\right\}_k$$

$$\mathbb{E}\left(d_k(\boldsymbol{Z},\boldsymbol{x})\right) = \mathbb{E}\left(d_k(\boldsymbol{Z},\boldsymbol{v})\right) = 0,$$

for all k.

The variance and covariance are given per window as follows:

$$\operatorname{Var}\left(d\left(\boldsymbol{Z},\boldsymbol{x}\right)\right)=\left\{\left(\frac{1}{\sum h_b\bar{c}_b}\right)^2\sum_{b=1}^{B}h_b\bar{c}_b\frac{s^2\left(\boldsymbol{x}_b\right)}{\bar{c}_b}\right\}_k$$

$$\operatorname{Cov}\left(d_k\left(\boldsymbol{Z},\boldsymbol{x}\right),\left(d_k\left(\boldsymbol{Z},\boldsymbol{v}\right)\right)\right)=\left\{\left(\frac{1}{\sum h_b\bar{c}_b}\right)^2\sum_{b=1}^{B}h_b\bar{c}_b\frac{s_k\left(\boldsymbol{x}_b;\boldsymbol{v}_b\right)}{\bar{c}_b}\right\}_k$$

for all k.

Consider $b\in B$ strata, from which simple random samples of $n_{t1}\dots n_{tB}$ clusters are selected into the treatment group from $n_1\dots n_B$ total clusters. As a random variable, the adjusted difference of treatment and control averages at every window k is given as:

$$d_k\left(\boldsymbol{Z},\boldsymbol{x}\right)=\left\{\sum_{b=1}^{B}w_b\left[\frac{\boldsymbol{Z}_b^t\boldsymbol{x}_b}{c_{tb}}-\frac{\left(\mathbf{1}-\boldsymbol{Z}_b^t\boldsymbol{x}_b\right)}{\left(c-c_{tb}\right)}\right]\right\}_k$$

$$=\left\{\sum_{b=1}^{B}w_bh_b^{-1}\bar{c}_b^{-1}\boldsymbol{Z}_b^t\boldsymbol{x}_b-\sum_{b=1}^{B}w_bc_b^{-1}\left(n_b-n_{tb}\right)^{-1}\mathbf{1}_b^t\boldsymbol{x}_b\right\}_k$$

3. Regressions with Exogenous Real-Time Experimental Treatments

Guided by the average treatment effects in the first section, we start with the Ordinary Least Square regressions, assuming exogenous treatments. However, although it is necessary to continuously update the regression model parameters while receiving new data, the tremendous data volume renders it impossible to scan the entire data stream multiple times to regenerate the regression model parameters. The Incremental Mathematical Stream Regression approach (Nadungodage, Xia, Li, Lee, and Ge (2011)) recalculates the regression function parameters based on the latest data in the current window and the synopsis of previous data. The Approximate Stream Regression approach approximately estimates the regression parameters by using a weighted average of the parameters obtained from current window data and the regression parameters of pervious windows.

We now assume the existence of *concept drift:* meaning that regression coefficients can change over time. The Approximate Stream Regression adapts to the incidence of concept drift, automatically updating regression coefficients to current values without having to rebuild the entire model.

Assuming the model with the following functional form:

$$Y = \beta T + \varepsilon$$

If we let β_{k-1} refer to the coefficient over window $k-1$, then according to the Exponential Weighted Moving Average (Brown and Meyer, 1961), we obtain the following,

$$\beta_k = (1-\sigma)\,\beta'_k + \sigma\beta_{k-1}$$

such that σ is a smoothing factor and constant value where both $\sigma \leq 1$ and $(1-\sigma) \leq 1$. Also, β'_k is a parameter vector for the kth window (calculated considering only the data records of the kth window). Also, β_{k-1} is the parameter vector for the $(k-1)$th window (calculated considering all data records seen up to $(k-1)$th window). Also, β_k is the refined parameter vector for the kth window calculated considering all data records seen up to the kth window. Substituting and expanding the Right Hand Side,

$$Y_k = \{(1-\sigma)\,\beta'_k + \sigma\beta_{k-1}\}\,T_k + \varepsilon_k$$

$$Y_k = \{(1-\sigma)\,\beta'_k\}\,T_k + \{\sigma\beta_{k-1}\}\,T_k + \varepsilon_k$$

$$Y_k = \left\{(1-\sigma)\,\beta'_k + \cdots + (1-\sigma)\,(\sigma)^j\,\beta_{k-j} + (\sigma)^k\,\beta_1\right\}T_k + \varepsilon_k$$

$$Y_k = (1-\sigma)\,\beta'_k T_k + \cdots + (1-\sigma)\,(\sigma)^j\,\beta_{k-j}T_k + (\sigma)^k\,\beta_1 T_k + \varepsilon_k$$

Assume exogeneity in real-time; $cov\,(\varepsilon, (T)) = 0$

Theorem 1. *The bias of the simple experimental coefficient when used in real-time environments where concept drift is relevant is given by:*

$$\left\{(1-\sigma)\,\beta_k' + \cdots + (1-\sigma)\,(\sigma)^j\,\beta_{k-j} + (\sigma)^k\,\beta_1\right\} - \left\{\left(T_k'T_k\right)^{-1}T_k'Y_k\right\} > 0$$

Proof. The expression is always positive due to the existence of concept drift. Since time is strictly increasing, concept drift is always positive which leads to larger estimates in real-time data. The bias is the difference in estimates.

□

4. Experiments with Difference-in-Difference Estimators in Real-Time

Difference-in-difference estimators are used to compute treatment effects when time trends have the potential to confound treatment impacts (see Imbens and Wooldridge, 2009, also see Bertrand, Duflo, and Mullainathan 2004 for clustered standard errors). We provide a real-time version of difference-in-difference estimation. Given the importance of time factors and concept drift, we use the Approximate Stream Regression approach. The difference-in-difference estimator produces causal effects, assuming that parallel trend assumptions hold.

$$Y_k = \alpha_k + \beta_k T_k + \gamma_k t_k + \delta_k\,(T_k \cdot t_k) + \varepsilon_k, \text{ for all } k.$$

α_k : constant term for window k

β_k : treatment group specific effect for window k, which accounts for average permanent differences between treatment and control

γ_k : time trend (in real-time) common to treatment and control groups in window k

δ_k : true effect of treatment in window k

We make the following points:

We are interested in an unbiased estimate $\hat{\delta}$, whereby $E\left[\hat{\delta}\right] = \delta$. The error term is zero on average: $E\left[\varepsilon_k\right] = 0$. The error term is not correlated with the other variables (exogeneity): $cov\left(\varepsilon_k, (T_k)\right) = 0$ and $cov\left(\varepsilon_k, (t_k)\right) = 0$.

Parallel trend assumption: $cov\left(\varepsilon_k, (T_k \cdot t_k)\right) = 0$. This identifying assumption is based on the premise that the real-time average change in the control group in real-time represents the counterfactual change in the treatment group if there were no treatment exposure received by the real-time data. We consider the average change in real-time outcomes for the treated in the absence of treatment to equal the average change in outcome for the non-treated.

From the difference-in-difference regression equation, also note that

$$E\left[Y_0^T\right] = \alpha + \beta$$

$$E\left[Y_1^T\right] = \alpha + \beta + \gamma + \delta$$

$$E\left[Y_0^C\right] = \alpha$$

$$E\left[Y_1^C\right] = \alpha + \gamma$$

From the Exponential Weighted Moving Average, we obtain the following coefficients for the difference-in-difference equation:

$$\delta_k = (1 - \theta)\,\delta_k' + \theta\delta_{k-1}$$

whereby θ is a smoothing factor and constant value where both $\theta \leq 1$ and $(1 - \theta) \leq 1$. Also, δ_k' is a parameter vector for the kth window (calculated considering only the data records of the kth window). Also, δ_{k-1} is the parameter vector for the $(k-1)$th window (calculated considering all data records seen up to $(k-1)$th window). Also, δ_k is the refined

parameter vector for the kth window calculated considering all data records seen up to the kth window. We similarly define the following:

$$\gamma_k = (1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1}$$

and

$$\beta_k = (1-\tau)\,\beta_k' + \tau\beta_{k-1}$$

Expanding

$$\delta_k = (1-\theta)\,\delta_k' + \cdots + (1-\theta)\,(\theta)^j\,\delta_{k-j} + (\theta)^k\,\delta_1$$

$$\gamma_k = (1-\vartheta)\,\gamma_k' + \cdots + (1-\vartheta)\,(\vartheta)^j\,\gamma_{k-j} + (\vartheta)^k\,\gamma_1$$

$$\beta_k = (1-\tau)\,\beta_k' + \cdots + (1-\tau)\,(\tau)^j\,\beta_{k-j} + (\tau)^k\,\beta_1$$

The refined value is a weighted combination of the previous values and the current value. By varying the smoothing factor τ, we can adjust the importance of the historical data. There is a negative relationship between the size of the smoothing factor and the weight assigned to the current data, as well as a positive relationship between the size of the smoothing factor and the degree to which the historical data affects the regression parameters. For equal weights between current values and past historical data, $\tau = \frac{1}{2}$.

Depending on the characteristics of the data stream, we may specify a desired half-life for the current window. Then the smoothing constant θ may be calculated in the following manner using the following half-life window principle:

$$\theta^k = e^{-\lambda k}$$

The value of λ can be calculated by $e^{-\lambda t_{\frac{1}{2}}} = \frac{1}{2}$, whereby $t_{\frac{1}{2}}$ is the desired approximate half-life time of a particular set of records. Thus, $\lambda = \frac{\ln(2)}{t_{\frac{1}{2}}}$ and $\theta = e^{\left[\frac{\ln(2)}{t_{\frac{1}{2}}}\right]}$.

For example, if we need to reduce the impact of the records in the kth window approximately by half, when we proceed to process the $(k+5)$th window, then $t_{\frac{1}{2}} = 5$ and $e^{\left[\frac{\ln(2)}{5}\right]}$.

If concept drift is present, recent samples will better reflect the status of the data stream. For such non-statiionary data we can use a smaller half-life value, so that θ will be adjusted to asign a higher weight on recent samples and generate more accurate regression model.

The difference-in-difference regression equation is given as:

$$Y_k = \alpha_k + \left\{(1-\tau)\,\beta'_k + \cdots + (1-\tau)\,(\tau)^j\,\beta_{k-j} + (\tau)^k\,\beta_1\right\} T_k$$

$$+ \left\{(1-\vartheta)\,\gamma'_k + \cdots + (1-\vartheta)\,(\vartheta)^j\,\gamma_{k-j} + (\vartheta)^k\,\gamma_1\right\} t_k$$

$$+ \left\{(1-\theta)\,\delta'_k + \cdots + (1-\theta)\,(\theta)^j\,\delta_{k-j} + (\theta)^k\,\delta_1\right\} (T_k \cdot t_k) + \varepsilon_k, \text{ for all } k$$

Theorem 2. The real-time difference-in-difference estimator is unbiased.

Proof.

$$\delta_{RTDD} = \bar{Y}^T_{1k} - \bar{Y}^T_{0k} - \left(\bar{Y}^C_{1k} - \bar{Y}^C_{0k}\right)$$

The proof is in the Appendix.

□

5. Algorithms

We provide simple algorithms to implement the approaches, proposing simple scripts to perform measurement and causal inference instantaneously without any delay in timeliness. The procedures are summarized in Algorithms 1-3. These approaches allow for the estimation of Average Causal Effects using the different data environments. To verify the

randomization of treatments in practice, real-time covariate balance may be utilized, so that real-time covariates are considered balanced when they are not statistically different from one another. However, some data may be constrained in this regard (e.g. if covariate variables are not available and the treatment is randomly implemented, as assumed). The real-time covariate balance may support the real-time parallel trend observations in the difference-in-difference implementation.

Algorithm 1 General Method of Simple Estimation of Average Causal Effects in Real-Time Data with Concept Drift (ASR)

Require: $k \geq 0$
Ensure: All parameter vectors for the current window of interest and the previous window
1: **while** real-time covariate variables balance **do**
2: Solve $(Y_k = \beta_k T_k + \varepsilon_k)$
3: **end while**

Algorithm 2 General Method of Difference-in-Difference Estimation in Real-Time Data

Require: $k \geq 0$
Ensure: All parameter vectors for the current window of interest and the previous window
1: **while** parallel trends assumption holds and real-time covariate variables balance **do**
2: Solve $(Y_k = \alpha_k + \beta_k T_k + \gamma_k t_k + \delta_k (T_k \cdot t_k) + \varepsilon_k)$
3: **end while**

6. A Secondary Problem of Causal Inference in Real-Time

To measure the causal effect of study Treatment A versus study Treatment B, the investigator requires the outcome for the identical study individual in both alternative futures. Since it is impossible to see both potential outcomes at once, one of the potential outcomes is always missing. This observation is known as the Fundamental Problem of Causal Inference (Holland, 1986).

We have considered treatment states and control states to be distinct in the real-time data. If the treatment states and control states may overlap in the real-time environment however, then we can observe the value of treated and non-treated outcomes on the same variable when different states are relevant. We limit our discussion to a single treatment

and control for simplicity and clarity. It is theoretically possible that an individual will experience a treatment in a way (state) at a particular point in time that is the same as not experiencing the treatment. Similarly, an individual may not experience a treatment in a way that is the theoretically the same as experiencing the treatment. We call this the *secondary problem of causal inference in real-time*. This dilemma may be one reason why some experiments may not appear to be replicable in real-time data.

Definition. The Secondary Problem in real-time occurs in the event where

$$\left((t_1)_j^{i_k}, j \in [p,q]\right) = \left((t_0)_j^{i_k}, j \in [p,q]\right).$$

7. Conclusion

This paper focuses on experimentation in real-time data environments. In the simple experimental case, we assumed that that no new variables emerged in the real-time data. If necessary, the introduction of new variables may also be considered and added to gain plausible treatment effects. Also real-time fixed effects and window dummies are compatible with the regressions as well as clustered standard errors. Estimation of treatment effects in observational data collected in real-time may benefit from estimated dose-response functions (e.g. Imbens, 2000), and other aspects of the program evaluation toolkit, but generalized to focus on real-time information. In this spirit of such research, collaborations between economists and computer and other scientists as well as engineers may be beneficial to policy makers interested in exploiting non-traditional data structures, of which real-time information is only one kind.

The difference-in-difference regression equation is given by expanding,

$$Y_k = \alpha_k + \beta_k T_k + \gamma_k t_k + \delta_k \left(T_k \cdot t_k\right) + \varepsilon_k, \text{ for all } k.$$

$$Y_k = \alpha_k + \left\{(1-\tau)\,\beta_k' + \tau\beta_{k-1}\right\} T_k$$

$$+ \left\{(1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1}\right\} t_k + \left\{(1-\theta)\,\delta_k' + \theta\delta_{k-1}\right\} \left(T_k \cdot t_k\right) + \varepsilon_k, \text{ for all } k$$

Taking expectations and substituting yields the following:

$$E\left[Y_0^T\right]_k = \alpha_k + \left\{(1-\tau)\,\beta_k' + \tau\beta_{k-1}\right\}$$

$$= \alpha_k + \left\{(1-\tau)\,\beta_k' + \cdots + (1-\tau)\,(\tau)^j\,\beta_{k-j} + (\tau)^k\,\beta_1\right\}$$

$$E\left[Y_1^T\right]_k = \alpha_k + \left\{(1-\tau)\,\beta_k' + \tau\beta_{k-1}\right\} + (1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1} + (1-\theta)\,\delta_k' + \theta\delta_{k-1}$$

$$= \alpha_k + \left\{(1-\tau)\,\beta_k' + \cdots + (1-\tau)\,(\tau)^j\,\beta_{k-j} + (\tau)^k\,\beta_1\right\}$$

$$+ \left\{(1-\vartheta)\,\gamma_k' + \cdots + (1-\vartheta)\,(\vartheta)^j\,\gamma_{k-j} + (\vartheta)^k\,\gamma_1\right\}$$

$$+ \left\{(1-\theta)\,\delta_k' + \cdots + (1-\theta)\,(\theta)^j\,\delta_{k-j} + (\theta)^k\,\delta_1\right\}$$

$$E\left[Y_0^C\right]_k = \alpha_k$$

$$E\left[Y_1^C\right]_k = \alpha_k + (1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1}$$

$$= \alpha_k + \left\{(1-\vartheta)\,\gamma_k' + \cdots + (1-\vartheta)\,(\vartheta)^j\,\gamma_{k-j} + (\vartheta)^k\,\gamma_1\right\}$$

The difference-in-difference estimator is therefore unbiased:

$$\delta_{RTDD_k} = E\left[\bar{Y}_{1k}^T\right] - E\left[\bar{Y}_{0k}^T\right] - \left(E\left[\bar{Y}_{1k}^C\right] - E\left[\bar{Y}_{0k}^C\right]\right)$$

Proof.

$$\left(\alpha_k + \left\{(1-\tau)\,\beta_k' + \tau\beta_{k-1}\right\} + (1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1} + (1-\theta)\,\delta_k' + \theta\delta_{k-1}\right)$$

$$-\left(\alpha_k + \left\{(1-\tau)\,\beta_k' + \tau\beta_{k-1}\right\}\right) - \left(\alpha_k + (1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1}\right)$$

$$-\alpha_k + (1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1} - \left((1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1}\right)$$

$$\delta_{RTDD_k} = \left((1-\vartheta)\,\gamma_k' + \vartheta\gamma_{k-1}\right) = \delta_k$$

Q.E.D.

□

References

.

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Eects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431-442.

Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002, June). Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 1-16). ACM.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica* 80, 2369-2429.

Belloni, A., and V. Chernozhukov (2009): "Computational Complexity of MCMC-based Estimators in Large Samples," *Annals of Statistics* 37, 2011-2055.

Belloni, A., C. Hansen, and I. Fernandez-Val (2013a). "Program Evaluation with High-Dimensional Data," MIT Working Paper.

Belloni, A., V. Chernozhukov. and C. Hansen (2013b): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls (with an Application to Abortion and Crime)," *Review of Economic Studies*, forthcoming.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004): "How much should we trust differences-in-differences estimates?" *Quarterly Journal of Economics* 119, 249-275.

Brown, R. G. and Meyer, R. F. 1961, "The fundamental theorem of exponential smoothing," *Operations Research*, 9(5), 673-685.

Couper, M. P. (2005): "Technology Trends in Survey Data Collection," *Social Science Computer Review*, 23, 486-501.

Dragoman, D. and M. Dragoman (2004): *Quantum-Classical Analogies* (Springer, Berlin).

Einav, L. and J. Levin (2014): "The Data Revolution and Economic Analysis," in *Innovation Policy and the Economy*, Volume 14, ed. by J. Lerner and S. Stern, forthcoming.

Fowler, J. H. and N. A. Christakis (2008): "Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis over 20 years in the Framingham Heart Study," *BMJ: British Medical Journal*, 337, 1-9.

Green, J. Hilary, and I. A. Macdonald (1981). "The influence of intravenous glucose on body temperature." *Quarterly Journal of Experimental Physiology* 66(4), 465-473.

Hansen, Ben B. and Jake Bowers (2008). "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science*, 23, 219–236.

Holland, P. W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945-960.

Hui, Xiaonan, Jinhai Zhou, Chen Xu, Shilie Zheng, Hao Chi, Xiaofeng Jin, and Xianmin Zhang (2013). "A real-time detection and self-control phase-sensitive OTDR distributed sensor system." In *12th International Conference on Optical Communications and Networks (ICOCN)*.

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

Imbens, Guido W., and Donald B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, New York.

Imbens, Guido W., and Jeffrey M. Wooldridge (2009). "Recent developments in the econometrics of program evaluation." *Journal of Economic Literature* 47(1), 5-86.

Jin, H. and D. B. Rubin (2008): "Principal Stratication for Causal Inference with Extended Partial Compliance," *Journal of the American Statistical Association*, 103, 101-111.

Kamenica, E., S. Mullainathan and R. H. Thaler. (2011): "Helping Consumers Know Themselves," *American Economic Review: Papers & Proceedings* 101, 417-422.

King, G. (2011): "Ensuring the Data Rich Future of the Social Sciences, *Science,* 331, 719-721.

Nadungodage, Chandima Hewa, Yuni Xia, Fang Li, Jaehwan John Lee, and Jiaqi Ge (2011). "StreamFitter: a real time linear regression analysis system for continuous data streams." In *International Conference on Database Systems for Advanced Applications*, pp. 458-461. Springer Berlin Heidelberg, 2011.

Neyman, J. 1923 [1990]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5, 465–472.

Rubin, D. (1974): "Estimating Causal Impacts of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66,688-701.

Stern A. D., B. A. Alexander, and A, Chandra (2017). "The Economics of Precision Medicine." *Science*, forthcoming.

Varian, H. R. (2014): "Big Data: New Tricks for Ecoometrics," *Journal of Economic Perspectives*, 28(2): 3-28.

Van der Wal, C. H., A. C. J. ter Haar, F. K. Wilhelm, R. N. Schouten, C. J. P. M. Harmans, T. P. Orlando, S. Lloyd, and J. E. Mooij, (2000): "Quantum Superposition of Macroscopic Persistent-Current States," *Science* 290, 773-777.

Zhu, Yunyue, and Dennis Shasha. "Statstream: Statistical monitoring of thousands of data streams in real time." *Proceedings of the 28th international conference on Very Large Data Bases.* VLDB Endowment, 2002.

Zwane, A. P., J. Zinman, E. Van Dusene, W. Paiente, C. Null, E. Miguel, M. Kremer, D. S. Karlan, R. Hornbeck, X. Giné, E. Duflo, F. Devoto, B. Crepon, and A. Banerjee, (2011): "Being Surveyed Can Change Later Behavior and Related Parameter Estimates," *Proceedings of the National Academy of Sciences,* 108, 1821–1826.