Wald-Difference-in-Differences Estimation without Individual-Level Treatment Data

Takumi Hattori[1], **Kohsuke Kubota**[1], **Keiichi Ochiai**[1]

[1]NTT DOCOMO, INC.

{takumi.hattori.zw, kousuke.kubota.xt, ochiaike}@nttdocomo.com

Abstract

In-store advertising, such as digital signage and in-store posters, is a crucial advertising method that influences customer purchasing behavior. While their effectiveness is typically evaluated by displaying ads on a store-by-store basis and comparing the purchasing behavior of those exposed to ads with those who are not, obtaining ad exposure data for individual customers is costly, making it challenging to conduct accurate causal inference with individual-level treatment variables. A common approach to address this issue is to perform causal inference considering non-compliance, setting visitors to stores implementing an ad campaign as the treatment group and similar customers who have visited comparable stores as the control group. In this setting, a popular estimator is the ratio of two Difference-in-Differences (DID) estimates: one for the outcome variable and another for the treatment variable. However, previous studies assumed that the DID estimate for the treatment variable is known from public data, which is not always the case. To overcome this limitation, we propose a method for estimating causal effects utilizing the fact that, for binary treatment variables, the DID estimate of the treatment variable represents the change in the proportion of compliers in the treatment group. Our method leverages Gaussian Mixture Model to estimate the proportion. This approach allows for the estimation of the treatment effect on the compliers even in advertising strategies where ad exposure data for individual customers is unobserved.

Introduction

Understanding the impact of in-store advertising, for example, digital signage, on customer purchasing behavior is increasingly important in retail practice (Brakus, Schmitt, and Zarantonello 2009; Puccinelli et al. 2009). In-store advertising draws customer attention to promoted products and sometimes outperforms traditional marketing strategies, including discounts (Han, Chandukala, and Li 2022). By conducting experimental store-by-store advertising treatments, Sachse, Oetzel, and Klapper (2023) revealed that store advertising significantly increases sales of promoted items. Estimating the impact of in-store advertising on key business metrics, e.g., customer transaction amounts, is essential for improving advertising and enhancing its value.

Despite its importance, it is challenging to evaluate the effects of in-store advertising through causal inference by classifying customers into treatment and control groups based on ad exposure due to the difficulty in obtaining data on whether customers have seen the advertisement. Although widely used methods for acquiring ad exposure data for individual customers include questionnaire (Reicks, Splett, and Fishman 1997) and eye-tracking (Huddleston et al. 2015), conducting such surveys for each advertising campaign poses a significant burden.

A common approach to causal inference without ad exposure data for individual customers is to consider non-compliance by using store visits as a proxy for ad exposure. In this context, the treatment group is defined as visitors to stores that implement a campaign, and the control group consists of customers similar to the treatment group who visited comparable stores. To estimate the causal effect on the compliers, the Wald-Difference-in-Differences (DID) estimator is commonly used (De Chaisemartin and d'Haultfoeuille 2018). This estimator is the ratio of the DID estimate of the outcome variable to the DID estimate of the treatment variable. For binary treatment variables, the DID estimate of the treatment variable represents the change in the proportion of units in the treatment group that received the treatment (hereafter referred to as the treatment ratio). Thus, the Wald-DID estimator is calculated by dividing the DID estimate of the treatment variable by the treatment ratio.

Although previous studies typically estimate Wald-DID by assuming that the treatment ratio is known from public data, it is not always feasible to derive this value from publicly available information. For example, Adena et al. (2015) used the trend of public radio listening rates when evaluating the impact of listening to political radio programs on voting results, using the Wald-DID estimator. However, public data representing the probability of noticing in-store advertising, which is a store-specific campaign, generally does not exist. Therefore, the calculation of the Wald-DID estimator for in-store advertising campaigns requires an estimation of the treatment ratio for those campaigns.

To address this issue, we propose a method to accurately calculate the Wald-DID estimator without individual-level treatment variables by estimating the treatment ratio. The process involves: creating pairs from treatment and control groups using covariate matching, estimating ATT for each pair and clustering treatment group units via Gaussian Mixture Models (GMM) (Reynolds et al. 2009) to estimate the treatment ratio. The key idea behind this approach is that

ATTs for compliers and non-compliers come from distinct distributions, with non-compliers expected to have a lower mean given a positive advertising effect. Under certain conditions, the estimated treatment ratio represents the DID estimate of the treatment variable. Through simulated experiments, we demonstrate that this method can estimate the causal effect on those exposed to advertising with high accuracy, without requiring individual-level treatment variables.

The contributions of this paper are as follows:

- We extended the Wald-DID method by leveraging GMM to estimate the proportion of the compliers when individual-level treatment data is unavailable.
- We evaluated the effectiveness of the proposed method by verifying the accuracy of causal effect estimation using simulation data.

Problem Setting

This study estimates the Local Average Treatment Effect (LATE) on the compliers in the treatment group when individual-level treatment variables are unavailable. LATE measures the treatment effect for the compliers, individuals who always follow the given assignment (Angrist and Imbens 1995; Angrist, Imbens, and Rubin 1996).

Notation

Following the notation of De Chaisemartin and d ' Haultfoeuille (2020), we define both individual-level and group level variables in this order. We define individuals as units and consider a set of N units $I = \{I_1, ..., I_N\}$, classified into treatment ($G = 1$) or control ($G = 0$) groups. We consider two time steps ($T = 0$ for pre-treatment and $T = 1$ for post-treatment). We assume that treatment occurs only when $G = 1$ and $T = 1$. For any individual-level variable Z for unit I_i in the group $G = g$ at the period $T = t$, we use the notation Z_{igt}. $D_{igt} \in \{0, 1\}$ is a binary treatment variable (e.g., exposure to advertising), assumed to be unobserved. Y_{igt} is an outcome. $\mathbf{X_{igt}}$ represents d-dimensional covariates.

Next, we define group level variables. We denote any variable Z for the group $G = g$ at the period $T = t$ as Z_{gt}. We denote the proportion of units that received the treatment as γ_{gt}. Since treatment can only occur when $G = 1$ and $T = 1$, the realized values of γ_{gt} are as follows:

$$\gamma_{gt} = \begin{cases} \gamma & \text{if } (g,t) = (1,1), \text{ where } 0 < \gamma \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For all (g, t), we define the group-level averages of treatment and potential outcomes as follows:

$D_{gt} = \frac{1}{N_{gt}} \sum_{n=1}^{N_{gt}} D_{igt}$, $Y_{gt} = \frac{1}{N_{gt}} \sum_{n=1}^{N_{gt}} Y_{igt}$,

where N_{gt} represents the number of observed units.

Assumptions

We outline the assumptions required for identifying causal effects in our method. Our approach uses the Wald-DID estimator from De Chaisemartin and d'Haultfoeuille (2018) to estimate LATE, and therefore follows the same assumptions as that paper. These assumptions include those common in standard DID (common trends, stable treatment effect, homogeneous treatment effect) as well as additional ones necessary for addressing non-compliance (De Chaisemartin and d ' Haultfoeuille 2018). Below, we list the assumptions that become necessary when considering non-compliance:

Assumption 1 (Fuzzy design)

$$\begin{aligned} &E[D_{11}] \geq E[D_{10}] \quad \text{and} \\ &E[D_{11}] - E[D_{10}] \geq E[D_{01}] - E[D_{00}]. \end{aligned} \tag{2}$$

This assumption requires a larger proportion of units in the treatment group to receive the treatment than the control group from pre-treatment to post-treatment periods. In our study, treatment can only occur when $G = 1$ and $T = 1$, ensuring this assumption holds.

Assumption 2 (Stable percentage of treated units in the control group)

$$0 < E[D_{01}] = E[D_{00}] < 1. \tag{3}$$

This assumption means that the proportion of units receiving the treatment in the control group does not change between pre-treatment and post-treatment periods. Similar to Assumption 1, this assumption also holds given the problem setting we are dealing with in this study.

Assumption 3 （Treatment participation equation）

$$D_{igt} = 1\{V \geq v_{gt}\}, \text{ with } V \perp\!\!\!\perp T | G, \tag{4}$$

where the latent indicator V represents a unit's susceptibility to treatment, with its threshold for treatment participation v_{gt} varying by the period and the group (Vytlacil 2002). This ensures monotonicity, meaning treatment status transitions uniformly within each group. Assumptions 1 and 3 imply no units in the treatment group revert to a non-treated state.

Identification

Next, we define the estimand and the estimator for our study, focusing on compliers in the treatment group, termed *switchers*, defined as $S = \{I_i | D_{10} = 0, D_{11} = 1\}$. This subset transitioned from untreated at $T = 0$ to treated at $T = 1$. The estimand, LATE for the switchers, is expressed as follows: $\Delta = E[Y_{11}(1) - Y_{11}(0)|S]$.

The Wald-DID estimator is defined as $W_{DID} = DID_Y / DID_D$. For any variable Z_{gt}, DID_Z is expressed by the following equation:$DID_Z = (E[Z_{11}] - E[Z_{10}]) - (E[Z_{01}] - E[Z_{00}])$. By satisfying the Assumptions in the previous section, the estimator becomes unbiased for the estimand (De Chaisemartin and d ' Haultfoeuille 2018).

This study uniquely assumes that the individual-level treatment variable D_{igt} is unobserved which is typically used for W_{DID} estimation. Here, we demonstrate that W_{DID} can be estimated using γ_{gt} instead of DID_D. By referring to the equation 1, γ_{11} is equal to DID_D, in our problem setting where the treatment variable is binary and the treatment occurs only when $G = 1$ and $T = 1$.

$$\begin{aligned} DID_D &= (E[D_{11}] - E[D_{10}]) - (E[D_{01}] - E[D_{00}]) \\ &= (\gamma_{11} - \gamma_{10}) - (\gamma_{01} - \gamma_{00}) \\ &= \gamma_{11}. \end{aligned} \tag{5}$$

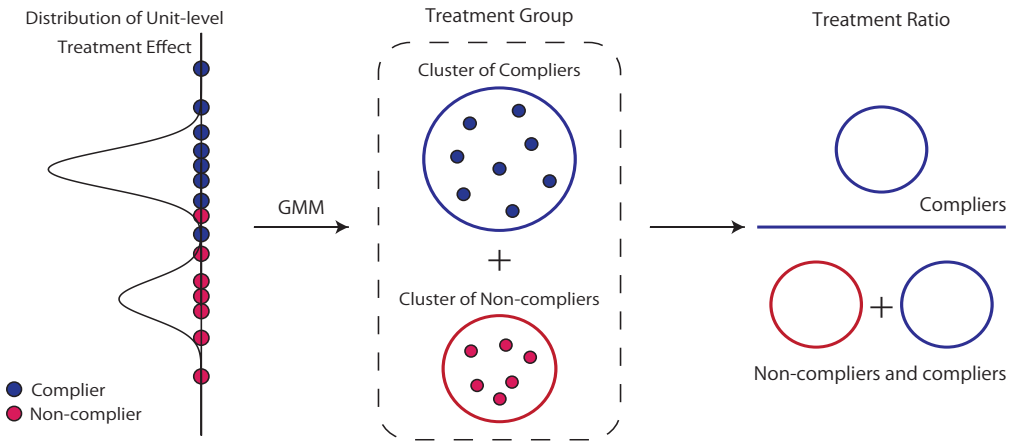

Figure 1: The ATT distribution differs between units who viewed the advertisement and those who did not. We cluster the treatment group into compliers and non-compliers, estimating the treatment ratio from the clusters' mixing proportion.

We refer to γ_{11} as the treatment ratio. Using the treatment ratio, W_{DID} can be expressed without DID_D as follows:

$$W_{DID} = \frac{DID_Y}{DID_D} = \frac{DID_Y}{\gamma_{11}}. \quad (6)$$

According to the equation 6, the key point of this study is to estimate the treatment ratio only with the observed data (G, X_{igt} and Y_{igt}), so that we can estimate W_{DID} without individual-level treatment data (D_{igt}). In the following section, we explain the method to estimate the treatment ratio by clustering the treatment group into switchers and non-switchers only with observed data.

The Proposed Method

Here, we explain how to estimate W_{DID} when the treatment variable D_{igt} is unobserved, through the following two steps: unit-level ATT estimation and clustering of treatment group units for the treatment ratio estimation.

Unit-level ATT estimation

Before estimating unit-level ATT, we explain the assumed data generation process as background. We assume the outcome Y_{igt} is generated following the Two-Way Fixed Effect method (Angrist and Krueger 1999; Angrist and Pischke 2009; Heckman, LaLonde, and Smith 1999), common in causal estimation with panel data.

$$Y_{igt} = \alpha_i + \lambda_t + \mathbf{X_{igt}^T} * \boldsymbol{\beta} + \tau * (G_{igt} \times D_{igt}) + \epsilon_{it},$$

where α_i represents the unit fixed effect, λ_t is the time fixed effect, $\boldsymbol{\beta}$ is a d-dimensional vector for covariates' relationship with the outcome and τ is LATE. ϵ_{it} is the error term following normal distribution $\mathcal{N}(0, \sigma^2)$. With this foundation, we then estimate unit-level ATT for all treatment group units by matching each with a control group unit and conducting DID for each pair.

We perform nearest neighbor matching (Stuart 2010) of covariates between the treatment and control groups, and estimate the ATT for each unit by conducting DID at the unit level. The key idea is that the ATT for switcher units is estimated from a distribution with LATE τ as its mean, while the ATT for non-switcher units is estimated from a distribution with a mean of 0 since they did not receive the treatment (Figure 1). We present the DID estimator for a switcher unit i matched with a control unit j, selected through nearest neighbor matching based on covariates.

$$\begin{aligned}
&(Y_{i11} - Y_{i10}) - (Y_{j01} - Y_{j00}) = \\
&((\alpha_i + \lambda_1 + \mathbf{X_{i11}^T} * \boldsymbol{\beta} + \tau + \epsilon_{i1}) - (\alpha_i + \lambda_0 + \mathbf{X_{i10}^T} * \boldsymbol{\beta} + \epsilon_{i0})) \\
&\quad - ((\alpha_j + \lambda_1 + \mathbf{X_{j01}^T} * \boldsymbol{\beta} + \epsilon_{j1}) - (\alpha_j + \lambda_0 + \mathbf{X_{j00}^T} * \boldsymbol{\beta} + \epsilon_{j0})) \\
&\quad = \tau + \epsilon_{i1} - \epsilon_{i0} - \epsilon_{j1} + \epsilon_{j0}.
\end{aligned}$$

We assume the covariate effects for units i and j are removed through adjustment. Similarly, we present the DID estimator for a non-switcher unit i matched with a control unit j.

$$(Y_{i11} - Y_{i10}) - (Y_{j01} - Y_{j00}) = \epsilon_{i1} - \epsilon_{i0} - \epsilon_{j1} + \epsilon_{j0}.$$

From the above equation, the distribution of DID estimators for all non-switcher units follows a normal distribution with a mean of 0.

Clustering of treatment group units for the treatment ratio estimation

We estimate the treatment ratio γ_{11} by clustering DID estimates of the treatment group using GMM. This approach leverages the distinct distributions of ATTs for switchers and non-switchers, enabling us to estimate their numbers within the treatment group. The parameters to be estimated are $\Theta = (\gamma_{11}, \theta_1), (1 - \gamma_{11}, \theta_0)$, where $\theta_1 = (\mu_1, \sigma_1)$ and $\theta_0 = (\mu_0, \sigma_0)$ represent the mean and standard deviation of DID estimates for switcher and non-switcher units, respectively. In our setting, the treatment group has two patterns of $D_{igt} \in \{0, 1\}$, while the control group has $D_{igt} \in \{0\}$. Since non-switcher units in the treatment group do not receive the treatment effect, we expect two clusters ($2 \times 1 = 2$) in the ATT distribution, with non-switcher units having a mean of 0. Θ is estimated by maximizing the likelihood function with these initial values.

$$L(\Theta|N) = \prod_{i \subset N} ((1 - \gamma_{11}) P_0(ATT_i|0, \sigma_0) + \gamma_{11} P_1(ATT_i|\mu_1, \sigma_1)),$$

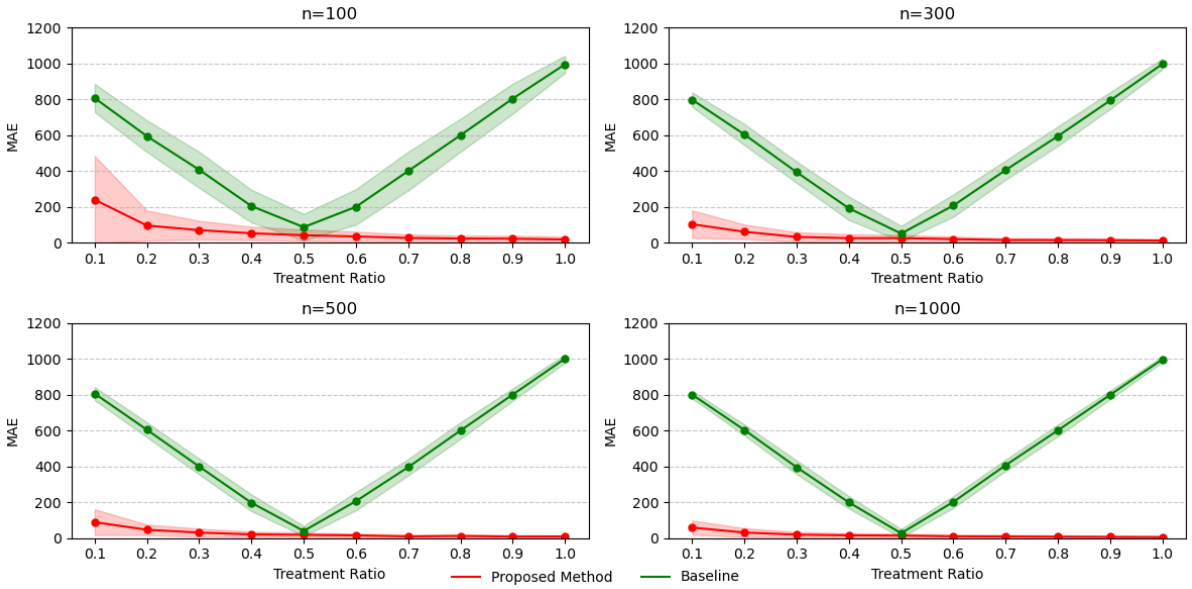

Figure 2: The graph shows the MAE for each estimator using simulation data. The line graph represents the MAE from 100 simulations for each condition. The shaded area surrounding the line graph represents the standard deviation.

Metrics	n	Treatment Ratio 0.1	0.3	0.5	0.7	0.9
MAE	100	0.025	0.037	0.039	0.037	0.024
	1000	0.007	0.012	0.012	0.011	0.006
std	100	0.020	0.029	0.032	0.030	0.021
	1000	0.005	0.010	0.010	0.008	0.005

Table 1: Accuracy of treatment ratios using the proposed method. The results show that the proposed method accurately estimates the true values under all conditions.

where ATT_i is the DID estimates for unit i. Since the outcome for each unit is observable and γ_{11}, which is equal to DID_D, can be estimated, we can calculate the Wald-DID estimator according to Equation 6.

Experiments

We conducted experiments using simulated datasets with known true causal effects to demonstrate the effectiveness of our method. We used the Wald-DID estimator with a 0.5 switcher proportion ($W_{DID} = DID_Y/0.5$) as a baseline for comparison.

Experimental Design

We conducted experiments using simulation datasets with varying treatment group sample sizes ($n = \{100, 300, 500, 1000\}$) and treatment ratios ($\gamma_{11} = \{0.1, 0.2, ..., 1.0\}$). The minimum treatment ratio was set to 0.1 as the Wald-DID estimator is undefined at 0. We averaged results over 100 data generations for each condition to eliminate bias arising from data generation variations. Performance was evaluated using Mean Absolute Error (MAE) $|y - \hat{y}|$, where y was the ground truth and $\hat{y}$ the estimated result.

Evaluation Results

We first verified our method's accuracy in estimating the treatment ratio. Table 1 demonstrate our method can accurately estimate true values across all conditions.

Figure 2 shows the MAE of Wald-DID estimates. Our method outperformed the baseline in all conditions, especially when the true ratio diverged from the baseline's assumed 0.5. While the baseline's MAE increased linearly with deviation from 0.5, our method maintained accuracy by adapting to the estimated treatment ratio. It also performed well with small samples of 100, demonstrating its efficacy for in-store advertising campaigns with limited targets.

These results show MAE and standard deviation increase as treatment ratio decreases. This occurs because our method divides DID_Y by the treatment ratio. Smaller ratios amplify errors, especially in smaller samples, causing greater Wald-DID estimates fluctuations.

Conclusion

This study proposes a method for accurately estimating causal effects on compliers in the treatment group when individual-level treatment data are unobserved. Our approach estimates the proportion of treated units in the treatment group, enabling stable causal effect estimation. Simulations demonstrate its superior accuracy compared to the baseline across various conditions. The method is particularly useful for in-store advertising campaigns where only a portion of visitors is exposed to advertisements. Its accuracy with small sample sizes makes it suitable for campaigns in low-traffic stores. This method can be applicable to other advertisement strategies which are difficult to acquire individual-level data for ad exposure, such as OOH and flyers. Future work will verify our method's performance using real in-store advertising data.

References

Adena, M.; Enikolopov, R.; Petrova, M.; Santarosa, V.; and Zhuravskaya, E. 2015. Radio and the Rise of the Nazis in Prewar Germany. *The Quarterly Journal of Economics*, 130(4): 1885–1939.

Angrist, J.; and Imbens, G. 1995. Identification and estimation of local average treatment effects.

Angrist, J. D.; Imbens, G. W.; and Rubin, D. B. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455.

Angrist, J. D.; and Krueger, A. B. 1999. Empirical strategies in labor economics. In *Handbook of labor economics*, volume 3, 1277–1366. Elsevier.

Angrist, J. D.; and Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Brakus, J. J.; Schmitt, B. H.; and Zarantonello, L. 2009. Brand experience: what is it? How is it measured? Does it affect loyalty? *Journal of marketing*, 73(3): 52–68.

De Chaisemartin, C.; and d ' Haultfoeuille, X. 2018. Fuzzy differences-in-differences. *The Review of Economic Studies*, 85(2): 999–1028.

De Chaisemartin, C.; and d ' Haultfoeuille, X. 2020. Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review*, 110(9): 2964–2996.

Han, Y.; Chandukala, S. R.; and Li, S. 2022. Impact of different types of in-store displays on consumer purchase behavior. *Journal of Retailing*, 98(3): 432–452.

Heckman, J. J.; LaLonde, R. J.; and Smith, J. A. 1999. The economics and econometrics of active labor market programs. In *Handbook of labor economics*, volume 3, 1865–2097. Elsevier.

Huddleston, P.; Behe, B. K.; Minahan, S.; and Fernandez, R. T. 2015. Seeking attention: an eye tracking study of in-store merchandise displays. *International Journal of Retail & Distribution Management*, 43(6): 561–574.

Puccinelli, N. M.; Goodstein, R. C.; Grewal, D.; Price, R.; Raghubir, P.; and Stewart, D. 2009. Customer experience management in retailing: understanding the buying process. *Journal of retailing*, 85(1): 15–30.

Reicks, M.; Splett, P.; and Fishman, A. 1997. Shelf Labeling Of Organic Foods: Effects On Customer Perceptions And Sales. Technical report.

Reynolds, D. A.; et al. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Sachse, M.; Oetzel, S.; and Klapper, D. 2023. I'll Try That, Too - A Field Experiment in Retailing on the Effect of Variety During Display Promotions. Rationality and Competition Discussion Paper Series 404, CRC TRR 190 Rationality and Competition.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1.

Vytlacil, E. 2002. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1): 331–341.