# Improving precision of A/B experiments using trigger intensity

**Tanmoy Das**
Amazon, USA
tanmdas@amazon.com

**Dohyeon Lee**
Amazon, USA
dohnlee@amazon.com

**Arnab Sinha**
Amazon, USA
arsinha@amazon.com

## Abstract

Online randomized controlled experiments (A/B tests) measure causal changes in industry. While these experiments use incremental changes to minimize disruption, they often yield statistically insignificant results due to low signal-to-noise ratios. Precision improvement (or reducing standard error) traditionally focuses on trigger observations - where treatment and control outputs differ. Though effective, detecting all triggers (*full knowledge*) is prohibitively expensive. We propose a sampling-based approach (*partial knowledge*) where the bias in the evaluation outcome decreases inversely with sample size. Simulations (See Appendix C) show bias approaches zero with just $\leq 0.1\%$ of observations sampled. Empirical testing demonstrates a 38% variance reduction compared to CUPED methods [1].

## 1 Introduction

Online randomized controlled experiments (A/B tests) evaluate causal changes in industry [2, 3, 4, 5, 6, 7, 8, 9]. The main challenge is low signal-to-noise ratio [4, 10] as treatment models typically implement incremental changes that impact few observations. Consequently, these experiments often have higher standard error and lack statistical significance, resulting in missed opportunities to implement beneficial treatments. A general approach [11, 8] to improve precision focuses on *trigger* observations (Section 2). In a trigger observation, outputs of the control and treatment model differ. This approach assumes treatment effects are confined to these trigger observations. While full information (*full knowledge*) about trigger observations would be ideal, detecting all triggers is expensive as there can be billions of observations [12, 13] per day. In this paper, we propose an alternative solution. Instead of detecting all trigger observations, we use a sampling-based approach (*partial knowledge*) where we randomly select a subset of observations and determine their trigger status. We then use this trigger information for evaluation. Although this sampling-based approach introduces bias in the evaluation outcome, our theoretical analysis shows that this bias decreases linearly as the sample size increases, making it a promising approach. Our theoretical findings are supported by both simulated (See Appendix C) and empirical data. Empirical analysis shows our method reduces variance by 38% without any detectable bias. Key contributions are:

**1)** First implementation of A/B experiment evaluations using sampled trigger observations.

**2)** Theoretical proof that evaluation bias decreases *linearly* with sample size.

**3)** Performance analysis showing 38% variance reduction compared to CUPED [1].

## 2 System overview

The proposed method is applicable across diverse domains including recommendation systems, search, personalization, and others.

**Requirements for implementation:** The proposed evaluation method is designed to be applicable across various experimental settings that meet certain criteria: **1)** There must be a clear and well-defined mechanism to identify *trigger* observations - instances where control and treatment outputs

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

differ; **2)** The experiment must use proper randomization, ensuring that the trigger status of an observation is independent of treatment assignment; and **3)** While exhaustive trigger detection may be impossible or prohibitively expensive, it should be feasible to randomly sample a subset of observations from both the control and treatment groups to estimate the trigger status. In particular, two-sided trigger information should be available.

**An example use case:** An e-commerce retailer displays multiple product images in a specific order on product pages. Image ranking is crucial as optimal placement improves customer engagement, measured through visits and purchases. Each website visit constitutes an observation, with customer responses recorded for evaluation.

The ranking process considers product features, image metadata and content, customer behavior, and search intent. As some inputs have infinite possible values, pre-computing all ranking outputs is infeasible. The existing *control* model runs in production, while a new *treatment* model is proposed for evaluation. Before deployment, we must verify the treatment model's superiority through A/B testing [2]. When both models produce identical rankings, we have *non-trigger* observations. A *trigger* observation occurs when rankings differ. Figure 1 illustrates these cases.

The A/B experiment uses either **product randomization** or **customer randomization**. In product randomization, products are randomly assigned to control or treatment groups. Control groups use the control model for ranking, while treatment groups use the treatment model. Trigger status is determined by comparing actual rankings with counterfactual rankings generated in the backend. This comparison can be performed online (simultaneously) or offline (later). While customer randomization is possible, this paper focuses on product randomization, though our methods apply to both approaches.

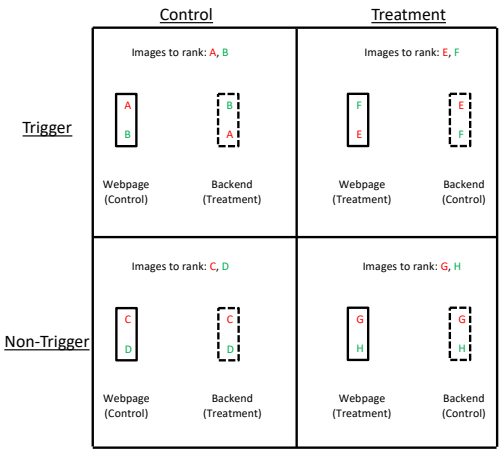



Figure 1: Trigger vs. non-trigger observations in a product randomized experiment that evaluates the impact of a change in product image ranking. We show four observations: two for products in control (left) and two for products in treatment (right). The top observations demonstrate triggers, while the bottom examples show non-triggers. *In top left*, there are two images (A and B) to rank. Because the control and treatment outputs differ, this observation is classified as a trigger. *In bottom left*, the control and treatment model outputs are identical, making this a non-trigger observation. Similar analysis applies to the treatment product examples (top right and bottom right).

## 3 Customer response model

We define our customer response model using two assumptions:

**Assumption 1:** Customer responses across observations are independent of each other.

**Assumption 2:** Treatment effects exist only in trigger observations where control and treatment outputs differ.

These simplifying assumptions enable theoretical analysis while maintaining practical utility as demonstrated by the empirical results. We use a linear model [1, 9, 14, 15] to capture the customer response because of its durability and interpretability. Also, random product assignment, ensures that a linear model can effectively estimate the treatment effect. Table 1 shows response variations across observations. $\mathcal{T}_i$ indicates treatment assignment ($\mathcal{T}_i = 1$ for treatment), and $r_{ij}$ indicates trigger status ($r_{ij} = 1$ for triggers).

For product $i$ with $n_i$ observations, the customer response for $j^{th}$ observation is $y_{ij} = \beta_0 + \beta_1 r_{ij} + \beta_2 \mathcal{T}_i r_{ij} + \eta_{ij}$, where $\eta_{ij}$ is noise. We use trigger intensity as the sole control variate.

After aggregating responses at product level (or customer level for customer-randomized experiments):

$$y_i = \beta_0 + \beta_1 r_i + \beta_2 \mathcal{T}_i r_i + \eta_i \tag{1}$$

where $y_i$ is average response and $r_i$ is the average trigger rate, henceforth known as *trigger intensity*.

| | Non-Trigger ($r_{ij} = 0$) | Trigger ($r_{ij} = 1$) |
|---|---|---|
| Control ($\mathcal{T}_i = 0$) | $\beta_0$ | $\beta_0 + \beta_1$ |
| Treatment ($\mathcal{T}_i = 1$) | $\beta_0$ | $\beta_0 + \beta_1 + \beta_2$ |

Table 1: $\beta_0$ represents the average customer response for non-trigger observations. For trigger control observations, which differ from non-trigger ones, the response is $\beta_0 + \beta_1$, where $\beta_1$ represents the additional change. Treatment trigger observations yield $\beta_0 + \beta_1 + \beta_2$, where $\beta_2$ captures the treatment-specific response difference.

# 4 Proposed evaluation methods

This section presents a theoretical analysis of two evaluation methods: 1) evaluation with *full knowledge* of trigger intensities; and 2) our proposed approach using *partial knowledge*. Our key contribution demonstrates that in the partial knowledge method, bias in estimated treatment effect and standard error decrease linearly with the number of samples used for trigger intensity estimation.

## 4.1 Evaluation: full knowledge of trigger intensity

Suppose, there are $N$ products and they typically have varying observation counts and trigger intensities. Research shows [16, 17] that incorporating treatment intensity variations reduces estimated treatment effect variance. We apply the linear model from Eq. (1) to determine the ATE, defined as $\rho$. Random assignment of products ensures trigger intensity $(r_i)$ is independent of treatment assignment $(\mathcal{T}_i)$.

**Lemma 4.1.** *Average treatment effect* $(\rho)$ *is* $\beta_2 E[r_i]$

*Proof.* Suppose, the treatment impact for the $i^{th}$ product is $\rho_i$. The average treatment effect for all products with trigger intensity $r_i$ is

$$E[\rho_i|r_i] = E[y_i|\mathcal{T}_i = 1, r_i] - E[y_i|\mathcal{T}_i = 0, r_i] = (\beta_0 + \beta_1 r_i + \beta_2 r_i) - (\beta_0 + \beta_1 r_i) = \beta_2 r_i \quad (2)$$

. The average treatment effect across all products is

$$\rho = E[\rho_i] = E[E[\rho_i|r_i]] = E[\beta_2 r_i] = \beta_2 E[r_i] \quad (3)$$

□

We estimate the value of $\beta_2$ using OLS.

**Corollary 4.1.1.** *With full knowledge of product trigger intensity,*
*a) the estimated ATE* $(\hat{\rho}) = \frac{\hat{E}[r_i]}{\hat{E}[r_i^2]} \left[\hat{E}[y_i r_i|\mathcal{T}_i = 1] - \hat{E}[y_i r_i|\mathcal{T}_i = 0]\right]$
*b) the variance of the estimated ATE* $\sigma^2(\hat{\rho}) = \frac{4\hat{E}[r_i]^2}{\hat{E}[r_i^2]} \frac{\sigma^2(\hat{\eta})}{N}$

*Proof.* The proof follows from Lemma 4.1 and E.1. □

The variance $\sigma^2(\hat{\rho})$ depends on two terms. First, the residuals $\sigma^2(\hat{\eta})$, which is smaller than the baseline method as long as there is a positive correlation between the trigger intensity $(r_i)$ and outcome variable $(y_i)$. This is already discussed in CUPED [1].

However, there is further reduction in variance from the $\frac{\hat{E}[r_i]^2}{\hat{E}[r_i^2]}$ term. This can be illustrated as follows. When $\hat{\sigma}^2(r_i) = \hat{E}[r_i^2] - \hat{E}[r_i]^2 > 0$, $\frac{\hat{E}[r_i]^2}{\hat{E}[r_i^2]} < 1$. Thus, the variance of the ATE will further reduce when $\hat{\sigma}^2(r_i) > 0$. This reduction is maximized when $\hat{\sigma}^2(r_i)$ is maximized.

The variance reduction stems from modeling the interaction between treatment and trigger intensity, rather than using trigger intensity as an independent covariate. This suggests that any covariate causing heterogeneous treatment effects should be incorporated into the treatment interaction term to improve estimation precision.

## 4.2 Evaluation: partial knowledge of trigger intensity

Due to the high cost of determining all trigger statuses $(r_{ij})$, we use a sampling-based approach on a small observation subset. We estimate trigger intensity as $r_i'$, where $r_i' = r_i + \epsilon_i$. Random product assignment ensures $r_i' \perp T_i$ and $\epsilon_i \perp T_i$. We assume error $(\epsilon_i)$ and true trigger intensity $(r_i)$ are uncorrelated $(E[\epsilon_i r_i] = 0)$. Unknown $\epsilon_i$ creates attenuation bias [18, 19] in coefficient estimates, denoted as $\hat{\beta}_2'$.

**Theorem 4.2.** *With partial knowledge of product trigger intensity,*
*a) the estimated value of parameter $\beta_2$ is*

$$\hat{\beta}_2' = \frac{1}{\hat{E}[(r_i')^2]} \left[ \hat{E}[y_i r_i' | \mathcal{T}_i = 1] - \hat{E}[y_i r_i' | \mathcal{T}_i = 0] \right] \tag{4}$$

*b) the bias in the estimated value is*

$$E[\hat{\beta}_2'] = \beta_2 \left[ 1 - \frac{\hat{E}[\epsilon_i r_i']}{\hat{E}[(r_i')^2]} \right] \tag{5}$$

*c) suppose, $\sigma^2(\hat{\eta'})$ is the residual variance and $\sigma^2(\eta)$ is the noise variance. When number of products $(N)$ is large*

$$\frac{1}{N}\sigma^2(\hat{\eta'}) \approx \frac{1}{N}\sigma^2(\eta) + \left[ \left( \beta_1 + \frac{\beta_2}{2} \right)^2 + \frac{\beta_2^2}{4} \right] \hat{E}[\epsilon_i^2] \tag{6}$$

*d) the variance of the estimated $\beta_2$ is*

$$\sigma^2 \left( \hat{\beta}_2' \right) = \frac{4}{\hat{E}[r_i'^2]} \frac{\sigma^2(\hat{\eta'})}{N} \tag{7}$$

We estimate ATE ($\hat{\rho'}$) with partial knowledge from $\hat{\beta}_2'$, including its bias and variance.

**Corollary 4.2.1.** *With partial knowledge of product trigger intensity,*
*a) the estimated ATE $\hat{\rho'} = \hat{\beta}_2' \hat{E}[r_i']$*
*b) the bias in the estimated ATE is $E[\hat{\rho'}] = \rho + \beta_2 \left( \hat{E}[\epsilon_i] - \frac{\hat{E}[\epsilon_i r_i'] \hat{E}[r_i']}{\hat{E}[r_i'^2]} \right)$*
*c) the variance of the estimated ATE is $\sigma^2 \left( \hat{\rho'} \right) = \frac{4\hat{E}[r_i']^2}{\hat{E}[r_i'^2]} \frac{\sigma^2(\hat{\eta'})}{N}$*

*Proof.* Proof follows from Theorem 4.2 and Lemma 4.1. □

### 4.3 Trigger intensity estimation method

In this section, we present a possible solution for computing the product trigger intensity and analyze its impact on the estimation bias of the ATE ($\hat{\rho'}$) with partial knowledge of product trigger intensity.

### 4.4 Independent sampling to estimate the trigger intensity

The $i^{th}$ product has $n_i$ observations. Suppose we randomly sample $m$ observations, where $m < n_i$, from these $n_i$ observations and determine their trigger status. As defined before, $r_{ij} = 1$ indicates the $j^{th}$ observation of the $i^{th}$ product is a trigger. The estimated trigger intensity for the $i^{th}$ product is $r_i' = \frac{\sum_{j=1}^{m} r_{ij}}{m}$.

The estimation error is $\epsilon_i = r_i' - r_i$. The mean is $E[\epsilon_i] = 0$ and variance is $\sigma^2(\epsilon_i) = E[\epsilon_i^2] = \frac{E[r_i] - E[r_i^2]}{m_i}$. With the help of this knowledge, it is possible to compute an upper bound on the estimation bias for ATE as defined in Theorem 4.2.

**Theorem 4.3.** *If $m$ (where $m \leq n_i, \forall i$ and $m > 1$) observations for all products are examined to estimate the product trigger intensity and $\sigma^2(r_i) > 0$,*
*a) there is a downward bias in the estimated ATE ($\hat{\rho'}$) and the bias is upper bounded as follows $\rho - E[\hat{\rho'}] < \frac{\beta_2}{m-1}$*
*b) the variance of the estimated ATE ($\sigma^2 \left( \hat{\rho'} \right)$) is larger than the variance of the estimated ATE ($\sigma^2(\hat{\rho})$) with full knowledge of trigger intensity. The difference has an upper bound as follows $\sigma^2 \left( \hat{\rho'} \right) - \sigma^2(\hat{\rho}) < \frac{1}{m} \left[ \left( \beta_1 + \frac{\beta_2}{2} \right)^2 + \frac{\beta_2^2}{4} \right]$*

*Proof.* Proof is in Appendix G □

There is a downward bias in the estimated ATE when we use the trigger intensity computed from independent sampling. On the other hand, it leads to an upward bias in the variance of the ATE. But these biases are inversely proportional to the number of observations $(m)$ used to estimate trigger intensity $(r_i')$.

## 5 Empirical evaluation using real A/B experiments

We implemented our partial knowledge evaluation method on an e-commerce A/B testing platform. Using 10-70 samples per product for trigger intensity estimation, the method reduces standard error without observable bias. We compare against two CUPED methods [1]: 1) stratification based on pre-experiment covariates, known as *Stratification*, and 2) pre-experiment covariate as a regressor, known as *Regression*. Following [4, 6, 1], we consider only participating products. Our five-month study included 23 experiments with 147 treatments (maximum 15 per experiment).

**Reduction in standard error:** The partial knowledge method achieves 21.2% lower standard error than stratification (Table 2 in Appendix), equivalent to 38% variance reduction. It increases statistically significant treatments (95%) by 145%, approaching perfect-knowledge performance. A paired t-test confirms statistical significance ($p \approx 0$).

**No bias in estimated ATE:** Confidence intervals (90% and 95%) overlap 100% with stratification and 99% with regression (Table 3 in Appendix). ATE signs match regression and stratification methods 74.1% and 71.4% of the time, confirming unbiased estimation.

| | Regression | Stratification | Partial knowledge |
|---|---|---|---|
| Avg standard Error | 0.38366 | 0.33343 | 0.25559 |
| Avg absolute t-val | 0.93977 | 0.88882 | 1.56887 |
| # statistically significant (90%) treatments | 22 | 22 | 54 |
| # statistically significant (95%) treatments | 17 | 17 | 46 |

Table 2: Comparison of absolute t-value, standard error, and statistically significant treatments for stratification, regression and partial knowledge evaluation methods from an online A/B experiment platform. The t-vale and standard error are aggregated over multiple treatments.

| | Regression | Stratification |
|---|---|---|
| % treatments where 95% confidence intervals overlap | 99% | 100% |
| % treatments where 90% confidence intervals overlap | 99% | 100% |
| % treatments with the same sign for ATE | 74.1% | 71.4% |

Table 3: Comparison of estimated ATE for partial knowledge evaluation method with the stratification and regression methods.

## 6 Conclusion

A/B experiments in industry typically have small treatment effects to minimize risks, often resulting in statistically insignificant results due to low signal-to-noise ratios. While focusing on trigger observations (where treatment and control outputs differ) can improve precision, identifying all triggers is resource-intensive. We propose a sampling-based evaluation method that reduces costs while maintaining effectiveness. Our theoretical analysis shows that sampling bias decreases inversely with sample size. Through simulations, we demonstrate that sampling just 0.1% of observations effectively eliminates bias, while empirical testing shows a 38% reduction in variance.

## References

[1] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International*

*Conference on Web Search and Data Mining*, WSDM '13, pages 123–132, New York, NY, USA, 2013.

[2] M. Luca and M. H. Bazerman. *The Power of Experiments: Decision Making in a Data-Driven World*. The MIT Press, 2021.

[3] Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. Designing and deploying online field experiments, 2014.

[4] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. From infrastructure to culture: A/b testing challenges in large scale social networks. KDD '15. Association for Computing Machinery, 2015.

[5] Raphael Lopez Kaufman, Jegar Pitchforth, and Lukas Vermeer. Democratizing online controlled experiments at booking.com, 2017.

[6] Ron Kohavi, Randal M. Henne, and Dan Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 959,Äì967, New York, NY, USA, 2007. Association for Computing Machinery.

[7] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings 16th Conference on Knowledge Discovery and Data Mining*, pages 17–26, 2010.

[8] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. Trustworthy online controlled experiments: five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 786,Äì794, New York, NY, USA, 2012. Association for Computing Machinery.

[9] Huizhi Xie and Juliette Aurisset. Improving the sensitivity of online controlled experiments: Case studies at netflix. KDD '16, page 645,Äì654, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Somit Gupta, Ronny Kohavi, and et al. Top challenges from the first practical online controlled experiments summit. In *KDD 2019*, 2019.

[11] Alex Deng and Victor Hu. Diluted treatment effect estimation for trigger analysis in online controlled experiments. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 349,Äì358, New York, NY, USA, 2015. Association for Computing Machinery.

[12] July 2024 traffic stats.

[13] Worldwide visits to amazon.com from july 2023 to december 2023.

[14] Susan Athey and Guido Imbens. Design-based analysis in difference-in-differences settings with staggered adoption, 2018.

[15] Michael Rosenblum and Mark J van der Laan. Using regression models to analyze randomized trials: asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3):937–945, February 2009.

[16] Amy S. Kelley, R. Sean Morrison, Neil S. Wenger, Susan L. Ettner, and Catherine A. Sarkisian. Determinants of treatment intensity for patients with serious illness: A new conceptual framework. *Journal of Palliative Medicine*, 13(7):807–813, 2010.

[17] P. J. Yoder and T. Woynaroski. How to study the influence of intensity of treatment on generalized skill and knowledge acquisition in students with disabilities. *ournal of behavioral education*, 24:152,Äì166, 2015.

[18] Steve Pischke. Lecture notes on measurement error.

[19] R. Hyslop and Guido W. Imbens. Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics*, 19(4):475–481, 2001.

[20] G.E.P. Box, J.S. Hunter, and W.G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley Series in Probability and Statistics. Wiley, 2005.

[21] A.S. Gerber and D.P. Green. *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton, 2012.

[22] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. Learning sensitive combinations of a/b test metrics. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 651,Äì659, New York, NY, USA, 2017. Association for Computing Machinery.

[23] Alex Deng and Xiaolin Shi. Data-driven metric development for online controlled experiments: Seven lessons learned. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 77,Äì86, New York, NY, USA, 2016. Association for Computing Machinery.

[24] Stephen Zakrewsky, Kamelia Aryafar, and Ali Shokoufandeh. Item popularity prediction in e-commerce using image quality feature vectors. *arXiv preprint arXiv:1605.03663*, 2016.

[25] Liwei Qian, Yajie Dou, Xiangqian Xu, Yufeng Ma, Shuo Wang, and Yi Yang. Product success evaluation model based on star ratings, reviews and product popularity. In *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, 2022.

# A Mathematical Symbols and their definitions

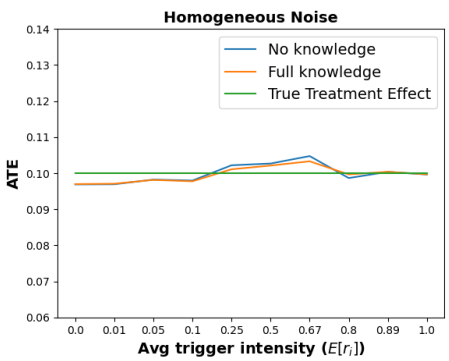
| Symbol | Definition |
|---|---|
| $n_i$ | number of observations for $i^{th}$ product |
| $y_{ij}$ | customer response for $i^{th}$ product and $j^{th}$ observation |
| $y_i$ | average customer response for $i^{th}$ product |
| $r_{ij}$ | trigger status for $i^{th}$ product and $j^{th}$ observation |
| $r_i$ | average trigger status for $i^{th}$ product known as *trigger intensity* |
| $r_i'$ | estimated trigger intensity for $i^{th}$ product |
| $\epsilon_i$ | error in estimated trigger intensity for $i^{th}$ product |
| $\mathcal{T}_i$ | treatment assignment for $i^{th}$ product |
| $\rho$ | treatment effect |
| $\rho'$ | treatment effect when using estimated trigger intensity |
| $\eta$, $\eta'$, $\eta_\alpha$ | indicates noise |
| $\hat{}$ | hat symbol indicates estimated parameters |
| $\mathbf{X}$ | bold symbols indicate matrix and vectors |

Table 4: Definition of mathematical symbols.

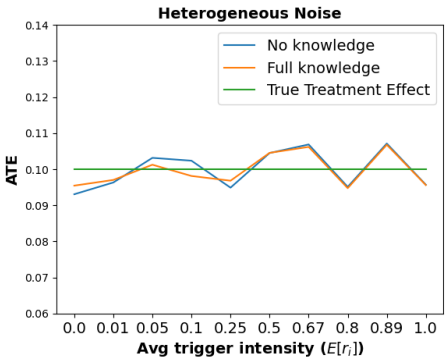
# B Related Work

The design and analysis of A/B experiments is a well-studied subject in statistics [20, 21]. Due to its effectiveness in detecting causal changes, A/B experimentation is widely used in industry for making data-driven decisions [2, 3, 4, 5, 6, 7, 8, 9]. Several works [4, 5, 6, 7] discuss proper guidelines for conducting A/B experiments in industrial settings, including methods for randomization, experiment design, engineering infrastructure, and choice of metrics. They also list numerous challenges that are difficult to solve.

One vexing problem for any A/B experiment is the issue of low precision, widely known as the *sensitivity* problem [10, 6, 4]. Precision is measured as the inverse of the variance of the evaluation results. Larger variance (lower precision) makes it harder to detect changes caused by the treatment model and prevents the launch of potentially beneficial features to production.




(a) Both evaluation methods are unbiased when noise is homogeneous.




(b) Both evaluation methods are unbiased when noise is heterogeneous.

Figure 2: Evaluations with no knowledge and full knowledge of trigger intensity are unbiased irrespective of the noise characteristics.

There are two typical solutions to address low precision: increasing the number of samples and choosing treatments with large impacts. However, increasing sample size is not always feasible due to multiple parallel experiments running simultaneously. Additionally, finding treatments with large effect sizes is often challenging because most changes are incremental in nature.

In the literature, three approaches are proposed to improve precision when conventional solutions fail: performing trigger analysis, employing better evaluation methods to reduce variance, and designing evaluation metrics with lower variance.

The concept of trigger analysis is proposed in [4, 6, 1]. These works consider only those products/users that actually participated in the experiment during evaluation.

In [11, 8], authors examine more granular trigger data, considering not only the products/customers participating in an experiment but also the number of trigger observations associated with each product/customer. Such granular information further reduces variance, but the cost of gathering such detailed information becomes prohibitive at scale with millions of experimental participants.

Two methods for improving precision are proposed in CUPED [1], which are also analyzed and extended by [9]. The first method involves clustering based on pre-experiment covariates to reduce between-cluster variance, leading to lower variance for aggregated results. The second method involves including pre-experiment covariates in the regression for evaluation, which can also minimize variance.

Our method combines both trigger-based [11] and covariate-based regression methods [1]. However, in comparison to [1], the novelty of our approach lies in the use of the covariate in regression. Instead of using it as a standalone feature, we interact the covariate with the treatment, which significantly reduces the variance of the estimated treatment effect when the covariate itself has high variance. Using the covariate as a standalone feature does not provide this benefit.

Furthermore, in comparison to [11], we extend the method and show that detailed trigger information is unnecessary; a sampling-based approach is sufficient to obtain an unbiased estimate, drastically reducing operational costs without sacrificing precision.

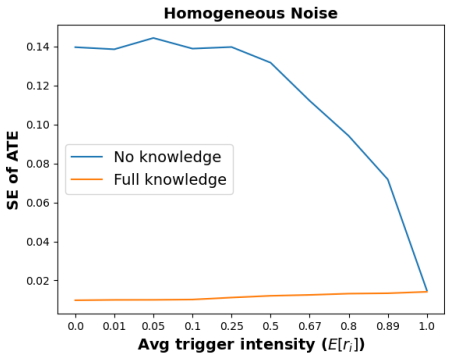
Methods for designing new evaluation metrics with lower variance are discussed in [22, 23]. Our evaluation method is agnostic to the choice of evaluation metric. However, the problem with any derived metric is low interpretability, which makes it challenging to explain experiment results.

## C Simulation study

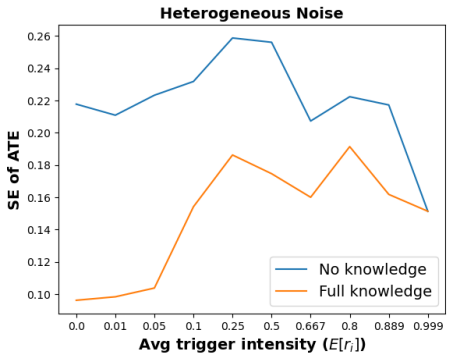
We use simulations to validate the theoretical analysis. Simulation analysis provides further evidence for the following two claims: 1) the standard error of evaluation with full knowledge is smaller than the standard error of the baseline method; and 2) the bias in the evaluation with partial knowledge decreases as the number of observations for sampling increases.

We use 2000 products, which are randomly divided into treatment and control groups. The number of observations for a product can be as high as 1 million. Here, only products that are part of the experiments are considered for analysis, as suggested by [4, 6, 1].

We also vary the noise characteristics (homogeneous vs heterogeneous) to check the robustness of the theoretical results. Homogeneous noise assumes measurements associated with all products have

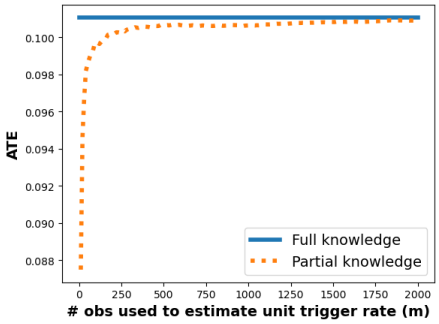



(a) The estimated ATE with full knowledge of triggers is more precise, as indicated by the smaller SE under homogeneous noise.

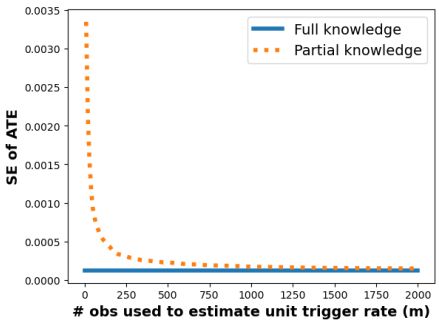



(b) The estimated ATE with full knowledge of triggers is more precise, as indicated by the smaller SE under heterogeneous noise.

Figure 3: Evaluation with full knowledge of trigger intensity is more precise in comparison to evaluation with no knowledge of trigger intensity irrespective of the noise characteristics.




(a) There is a negative bias in the estimated ATE with partial knowledge of trigger intensity. This bias is inversely proportional to the number of observations used to estimate trigger intensity increases.




(b) There is a positive bias in the variance of ATE with partial knowledge of trigger intensity. This bias is inversely proportional to the number of observations used to estimate trigger intensity increases.

Figure 4: Comparison of estimated values with full knowledge of trigger intensity and partial knowledge of trigger intensity.

the same noise variance, i.e., noise is i.i.d. For the heterogeneous case, noise is independent, but its variance can change based on products.

### C.1 Full knowledge vs baseline

Comparing baseline evaluation with full knowledge reveals both methods estimate ATE without bias (Figure 2). The difference between the estimated ATEs from these two methods decreases as the trigger intensity increases ($> 0.8$). These methods converge as trigger intensity increases above 0.8. With higher trigger rates, more observations are affected, reducing the bias in baseline estimates.

Evaluation with full knowledge achieves lower standard error (Figure 3), which increases precision and reduces Type II errors. This improved ability to detect small treatment effects is particularly valuable in industry settings, where subtle improvements in customer experience are common but difficult to validate statistically.

### C.2 Full knowledge vs partial knowledge

We also evaluate the performance of the evaluation method when only partial knowledge of trigger intensity is available. Here, the trigger intensity of a product is estimated by inspecting a sample of observations related to that product. The error in the estimated product trigger intensity ($r_i'$) depends on the number of observations ($m$) inspected per product.

Figure 4 shows the simulated results. As $m$ increases, the error in the estimated trigger intensity decreases, which generates more accurate (lower bias) and precise (lower standard error) estimates

of ATE. In this simulation setup, bias in the estimated ATE from evaluation with partial knowledge becomes very small when $m$ is larger than 20 samples. The same effect occurs for variance.

It should be mentioned that this threshold for the required number of samples will depend on the inherent noise variance and the true treatment effect.

# D Baseline evaluation: no knowledge of trigger intensity

In this section, we present a baseline evaluation method that does not use any information about trigger intensity. We use this model as a baseline to compare the performance of models that utilize trigger intensity. In particular, this model assumes all units are impacted by the treatment in the same way.

In this scenario, customer responses for all control products are the same. Likewise, customer responses for all treatment products are also the same. The difference in customer response between control and treatment products is caused by the treatment,

$$y_i = \alpha_0 + \alpha_1 \mathcal{T}_i + \eta_i \tag{8}$$

. Here, the customer response for control products is $\alpha_0$ and the additional change in customer response for treatment products is $\alpha_1$. Hence, $\alpha_1$ represents the average difference in customer response between the control and treatment products. This is commonly known as the *average treatment effect (ATE)*.

$\alpha_1$ is estimated using the Ordinary Least Square (OLS) method. This estimation is unbiased when the noise term $(\eta_i)$ is i.i.d and zero mean Gaussian.

**Lemma D.1.** *With no knowledge of trigger intensity a) The estimated average treatment effect is*

$$\hat{\alpha_1} = \hat{E}[y_i|\mathcal{T}_i = 1] - \hat{E}[y_i|\mathcal{T}_i = 0] \tag{9}$$

*b) Suppose, the variance of the residual is $\sigma^2(\hat{\eta}_\alpha)$. The variance of the estimated treatment effect is*

$$\sigma^2\left(\hat{\alpha_1}\right) = 4 \cdot \frac{\sigma^2(\hat{\eta}_\alpha)}{N} \tag{10}$$

*Proof.* We derive parameters of the baseline model. We use the matrix notation to represent the input parameters $\mathbf{X}$, which is a matrix of size $N \times 2$. $N$ is the number of observations and there are two parameters in the baseline model: a constant term and the treatment indicator. The treatment indicator is represented by a column vector $\mathbf{t}$ and the constant term is represented by $\mathbf{1}$. Both of them have the size of $N \times 1$. The observed outcome ($\mathbf{y}$) is a column vector of size $N \times 1$. The estimated parameters are denoted by vector $\hat{\boldsymbol{\alpha}} = [\hat{\alpha}_0 \quad \hat{\alpha}_1]$. The variance of the residuals is $\sigma^2(\hat{\eta_\alpha})$.

**Part a:**

$$\mathbf{X} = [\mathbf{1} \quad \mathbf{t}] \tag{11}$$

$$\frac{1}{N}\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & E[\mathcal{T}_i] \\ E[\mathcal{T}_i] & E[\mathcal{T}_i^2] \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{12}$$

$$\left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)^{-1} = 4\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \tag{13}$$

$$\frac{1}{N}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} E[y_i] \\ E[\mathcal{T}_i y_i] \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\left(E[y_i|\mathcal{T}_i = 1] + E[y_i|\mathcal{T}_i = 0]\right) \\ \frac{1}{2}E[y_i|\mathcal{T}_i = 1] \end{bmatrix} \tag{14}$$

$$\hat{\boldsymbol{\alpha}} = \left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)^{-1}\frac{1}{N}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} E[y_i|\mathcal{T}_i = 0] \\ E[y_i|\mathcal{T}_i = 1] - E[y_i|\mathcal{T}_i = 0] \end{bmatrix} \tag{15}$$

$$\tag{16}$$

Hence, $\hat{\alpha}_1 = E[y_i|\mathcal{T}_i = 1] - E[y_i|\mathcal{T}_i = 0]$. Note that the second equation in (27) holds exactly in population and approximately in the sample analog.

**Part b:**

$$\sigma^2(\hat{\boldsymbol{\alpha}}) = \text{diag}\left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)^{-1}\frac{\sigma^2(\hat{\eta_\alpha})}{N} = \begin{bmatrix} 2\frac{\sigma^2(\hat{\eta_\alpha})}{N} & 4\frac{\sigma^2(\hat{\eta_\alpha})}{N} \end{bmatrix} \tag{17}$$

Hence, $\sigma^2(\hat{\alpha}_1) = 4 \cdot \frac{\sigma^2(\hat{\eta_\alpha})}{N}$ □

The problem with the baseline method is that it assumes all treatment products are equally impacted by an experiment. However, this is not a realistic scenario.

For instance, there are popular products [24], [25] where the number of customer visits is much higher than for other products. It is reasonable to assume these products will have many trigger observations, thus more treatment impact. In the next section, we present an evaluation method that can overcome this limitation.

# E Full knowledge of trigger intensity

**Lemma E.1.** *With full knowledge of trigger intensity,*
*a) the estimated value of parameter $\beta_2$ is*

$$\hat{\beta}_2 = \frac{1}{\hat{E}[r_i^2]} \left[ \hat{E}[y_i r_i | \mathcal{T}_i = 1] - \hat{E}[y_i r_i | \mathcal{T}_i = 0] \right] \tag{18}$$

*b) suppose, the variance of the residuals is $\sigma^2(\hat{\eta})$, the variance of estimated $\beta_2$ is*

$$\sigma^2\left(\hat{\beta}_2\right) = \frac{4}{\hat{E}[r_i^2]} \frac{\sigma^2(\hat{\eta})}{N} \tag{19}$$

*Proof.* Trigger intensity is represented by a column vector $\mathbf{r}$ of size $N \times 1$. $\odot$ is a symbol for element wise multiplication. Assume, $E[r_i] = m$, $E[r_i^2] = s$, and $\sigma^2(r_i) = v$. The estimated parameters are $\hat{\beta}$.
**Part a:**

$$\mathbf{X} = [\mathbf{1} \quad \mathbf{r} \quad \mathbf{r} \odot \mathbf{t}] \tag{20}$$

$$\frac{1}{N}\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & E[r_i] & E[\mathcal{T}_i r_i] \\ E[r_i] & E[r_i^2] & E[\mathcal{T}_i r_i^2] \\ E[\mathcal{T}_i r_i] & E[\mathcal{T}_i r_i^2] & E[\mathcal{T}_i^2 r_i^2] \end{bmatrix} = \begin{bmatrix} 1 & m & \frac{m}{2} \\ m & s & \frac{s}{2} \\ \frac{m}{2} & \frac{s}{2} & \frac{s}{2} \end{bmatrix} \tag{21}$$

$$\left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} \frac{s}{v} & -\frac{m}{v} & 0 \\ -\frac{m}{v} & \frac{1}{v} + \frac{1}{s} & -\frac{2}{s} \\ 0 & -\frac{2}{s} & \frac{4}{s} \end{bmatrix} \tag{22}$$

$$\frac{1}{N}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} E[y_i] \\ E[r_i y_i] \\ E[\mathcal{T}_i r_i y_i] \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\left(E[y_i|\mathcal{T}_i = 1] + E[y_i|\mathcal{T}_i = 0]\right) \\ \frac{1}{2}\left(E[r_i y_i|\mathcal{T}_i = 1] + E[r_i y_i|\mathcal{T}_i = 0]\right) \\ \frac{1}{2}E[r_i y_i|\mathcal{T}_i = 1] \end{bmatrix} \tag{23}$$

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)^{-1} \frac{1}{N}\mathbf{X}^T\mathbf{y} \tag{24}$$

$$= \begin{bmatrix} \frac{s}{v}E[y_i] - \frac{m}{v}E[r_i y_i] \\ -\frac{m}{v}E[y_i] + \frac{1}{v}E[r_i y_i] - \frac{1}{2s}\left(E[r_i y_i|\mathcal{T}_i = 1] - E[r_i y_i|\mathcal{T}_i = 0]\right) \\ \frac{1}{s}\left(E[r_i y_i|\mathcal{T}_i = 1] - E[r_i y_i|\mathcal{T}_i = 0]\right) \end{bmatrix} \tag{25}$$

$$\tag{26}$$

**Part b:**

$$\sigma^2(\hat{\boldsymbol{\beta}}) = \text{diag}\left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)^{-1} \frac{\sigma^2(\hat{\eta}^2)}{N} \tag{27}$$

$$= \begin{bmatrix} \frac{s}{v}\frac{\sigma^2(\hat{\eta}^2)}{N} & \left(\frac{1}{v} + \frac{1}{s}\right)\frac{\sigma^2(\hat{\eta}^2)}{N} & \frac{4}{s}\frac{\sigma^2(\hat{\eta}^2)}{N} \end{bmatrix} \tag{28}$$

Again, (34) holds exactly in population and approximately in sample. □

This estimated $\hat{\beta}_2$ is unbiased as long as the noise term in Eq. (1) is i.i.d and zero mean Gaussian. Based on this information, we can estimate the ATE and its variance.

## F Partial knowledge of trigger intensity: Proof of theorem 4.2

Estimated product trigger intensity is represented by a column vector $\mathbf{r}'$ of size $N \times 1$. The estimation error for product trigger intensity is $\boldsymbol{\epsilon}$, which is a column vector of size $N \times 1$. Hence, $\mathbf{r}' = \mathbf{r} + \boldsymbol{\epsilon}$. We can write

$$\mathbf{E} = [\mathbf{0} \quad \mathbf{e} \quad \mathbf{e} \odot \mathbf{t}] \tag{29}$$

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{E} \tag{30}$$

Assume, $E[r'_i] = m'$, $E[(r'_i)^2] = s'$, $\sigma^2(r'_i) = v'$, $E[\epsilon_i] = e$, $E[\epsilon_i^2] = q$, and $E[\epsilon_i r'_i] = k$. The estimated parameters are $\hat{\beta}'$.

### F.1 Proof of Theorem 4.2: part a

*Proof.*

$$\tilde{\mathbf{X}} = [\mathbf{1} \quad \mathbf{r}' \quad \mathbf{r}' \odot \mathbf{t}] \tag{31}$$

$$\frac{1}{N}\tilde{\mathbf{X}^T}\tilde{\mathbf{X}} = \begin{bmatrix} 1 & E[r'_i] & E[\mathcal{T}_i r'_i] \\ E[r'_i] & E[{r'_i}^2] & E[\mathcal{T}_i {r'_i}^2] \\ E[\mathcal{T}_i r'_i] & E[\mathcal{T}_i {r'_i}^2] & E[\mathcal{T}_i^2 {r'_i}^2] \end{bmatrix} = \begin{bmatrix} 1 & m' & \frac{m'}{2} \\ m' & s' & \frac{s'}{2} \\ \frac{m'}{2} & \frac{s'}{2} & \frac{s'}{2} \end{bmatrix} \tag{32}$$

$$\left(\frac{1}{N}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1} = \begin{bmatrix} \frac{s'}{v'} & -\frac{m'}{v'} & 0 \\ -\frac{m'}{v'} & \frac{1}{v'} + \frac{1}{s'} & -\frac{2}{s'} \\ 0 & -\frac{2}{s'} & \frac{4}{s'} \end{bmatrix} \tag{33}$$

$$\hat{\boldsymbol{\beta}}' = \left(\frac{1}{N}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1} \frac{1}{N}\tilde{\mathbf{X}}^T\mathbf{y} \tag{34}$$

$$= \begin{bmatrix} \frac{s'}{v'}E[y_i] - \frac{m'}{v'}E[r'_i y_i] \\ -\frac{m'}{v'}E[y_i] + \frac{1}{v'}E[r'_i y_i] - \frac{1}{2s'}\left(E[r'_i y_i|\mathcal{T}_i = 1] - E[r'_i y_i|\mathcal{T}_i = 0]\right) \\ \frac{1}{s'}\left(E[r'_i y_i|\mathcal{T}_i = 1] - E[r'_i y_i|\mathcal{T}_i = 0]\right) \end{bmatrix} \tag{35}$$

From Eq. (34), we can say that

$$\hat{\beta}'_2 = \frac{1}{E[(r'_i)^2]}\left[E[y_i r'_i|\mathcal{T}_i = 1] - E[y_i r'_i|\mathcal{T}_i = 0]\right] \tag{36}$$

□

### F.2 Proof of Theorem 4.2: part b

*Proof.*

$$\frac{1}{N}\tilde{\mathbf{X}}^T\mathbf{E} = \begin{bmatrix} 0 & E[\epsilon_i] & E[\mathcal{T}_i\epsilon_i] \\ 0 & E[r'_i\epsilon_i] & E[\mathcal{T}_i r'_i\epsilon_i] \\ 0 & E[\mathcal{T}_i r'_i\epsilon_i] & E[\mathcal{T}_i r'_i\epsilon_i] \end{bmatrix} = \begin{bmatrix} 0 & e & \frac{e}{2} \\ 0 & k & \frac{k}{2} \\ 0 & \frac{k}{2} & \frac{k}{2} \end{bmatrix} \tag{37}$$

$$\left(\frac{1}{N}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1} \frac{1}{N}\tilde{\mathbf{X}}^T\mathbf{E} = \begin{bmatrix} 0 & \frac{s'e - m'k}{v'} & \frac{s'e - m'k}{2v'} \\ 0 & \frac{k - em'}{v'} & \frac{k - em'}{2v'} - \frac{k}{s'} \\ 0 & 0 & \frac{k}{s'} \end{bmatrix} \tag{38}$$

□

Hence, $E[\hat{\beta}'_2] = \beta_2\left[1 - \frac{E[\epsilon_i r'_i]}{E[(r'_i)^2]}\right]$

### F.3 Proof of Theorem 4.2: part c

*Proof.* The true parameters are $\boldsymbol{\beta}$ and the noise is $\boldsymbol{\eta}$.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} \tag{39}$$

$$\hat{\boldsymbol{\beta}}' = \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{y} \tag{40}$$

$$= \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}\right) \tag{41}$$

$$= \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\left((\tilde{\mathbf{X}} - \mathbf{E})\boldsymbol{\beta} + \boldsymbol{\eta}\right) \tag{42}$$

$$= \boldsymbol{\beta} - \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{E}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\eta} \tag{43}$$

We can use Eq. (38) to re-write Eq. (40) as

$$\hat{\boldsymbol{\beta}}' = \begin{bmatrix} 1 & -\frac{s'e - m'k}{v'} & -\frac{s'e - m'k}{2v'} \\ 0 & 1 - \frac{k - em'}{v'} & -\frac{k - em'}{2v'} + \frac{k}{s'} \\ 0 & 0 & 1 - \frac{k}{s'} \end{bmatrix}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\eta} \tag{44}$$

The residuals for OLS using estimated trigger intensity is

$$\hat{\boldsymbol{\eta}'} = \mathbf{y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}' \tag{45}$$

$$= \boldsymbol{\eta} + \mathbf{X}\boldsymbol{\beta} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}' \text{ from Eq. (39)} \tag{46}$$

$$= \boldsymbol{\eta} + \mathbf{X}\boldsymbol{\beta} - \tilde{\mathbf{X}}\left(\boldsymbol{\beta} - \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{E}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\eta}\right) (43) \tag{47}$$

$$= \boldsymbol{\eta} + \left(\tilde{\mathbf{X}} - \mathbf{E}\right)\boldsymbol{\beta} - \tilde{\mathbf{X}}\left(\boldsymbol{\beta} - \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{E}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\eta}\right) \tag{48}$$

$$= \boldsymbol{\eta} - \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\eta} - \mathbf{E}\boldsymbol{\beta} + \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{E}\boldsymbol{\beta} \tag{49}$$

$$= \left(\mathbf{I}_{N\times N} - \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\right)\boldsymbol{\eta} - \left(\mathbf{I}_{N\times N} - \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\right)\mathbf{E}\boldsymbol{\beta} \tag{50}$$

Suppose $\mathbf{A} = \left(\mathbf{I}_{N\times N} - \tilde{\mathbf{X}}\mathbf{M}\right)$. $\mathbf{A}$ is a special matrix. It satisfies the following properties.

- $\mathbf{A}$ is a symmetric matrix.
- $\mathbf{A}$ is idempotent i.e., $\mathbf{A}^T\mathbf{A} = \mathbf{A}$.
- $\mathbf{A}$ has $N - 3$ non-zero eigenvalues and 3 zero eigenvalues. All of its non-zero eigenvalues are 1. Hence $\text{tr}\left(\mathbf{A}\right) = N - 3$.

We can write Eq. (50) as

$$\hat{\boldsymbol{\eta}'} = \mathbf{A}\left(\boldsymbol{\eta} - \mathbf{E}\boldsymbol{\beta}\right) \tag{51}$$

Now, the expected value of the residuals is zero, so

$$\frac{1}{N}\sigma^2(\hat{\eta}') = \frac{1}{N}E[\hat{\boldsymbol{\eta}'}^T\hat{\boldsymbol{\eta}'}] \tag{52}$$

$$E[\hat{\boldsymbol{\eta}'}^T\hat{\boldsymbol{\eta}'}] = E[\left(\mathbf{A}\left(\boldsymbol{\eta} - \mathbf{E}\boldsymbol{\beta}\right)\right)^T\left(\mathbf{A}\left(\boldsymbol{\eta} - \mathbf{E}\boldsymbol{\beta}\right)\right)] \tag{53}$$

$$= E[\boldsymbol{\eta}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\eta}] + E[\boldsymbol{\beta}^T\mathbf{E}^T\mathbf{A}^T\mathbf{A}\mathbf{E}\boldsymbol{\beta}] \tag{54}$$

$$= E[\boldsymbol{\eta}^T\mathbf{A}\boldsymbol{\eta}] + \boldsymbol{\beta}^T E[\mathbf{E}^T\mathbf{A}\mathbf{E}]\boldsymbol{\beta} \tag{55}$$

We can show that

$$E[\mathbf{E}^T\mathbf{A}\mathbf{E}] = \text{tr}(A)\begin{bmatrix} 0 & 0 & 0 \\ 0 & q & \frac{q}{2} \\ 0 & \frac{q}{2} & \frac{q}{2} \end{bmatrix} \tag{56}$$

Based on Eq. (56), we can write Eq. (55) as

$$E[\hat{\boldsymbol{\eta}'}^T \hat{\boldsymbol{\eta}'}] = \sigma^2(\eta)\text{tr}(A) + \left[\left(\beta_1 + \frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right] q\text{tr}(A) \tag{57}$$

When $N$ is very large, we can approximate $\sigma^2(\hat{\eta'})$ as

$$\frac{1}{N}\sigma^2(\hat{\eta'}) \approx \frac{1}{N}\sigma^2(\eta) + \left[\left(\beta_1 + \frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right] q \tag{58}$$

□

### F.4 Proof of Theorem 4.2: part d

*Proof.* Suppose, $\sigma^2(\hat{\eta'})$ is the variance of the residuals, then

$$\sigma^2(\hat{\boldsymbol{\beta}'}) = \text{diag}\left(\frac{1}{N}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1} \frac{\sigma^2(\hat{\eta'})}{N} \tag{59}$$

$$= \begin{bmatrix} \frac{s'}{v'}\frac{\sigma^2(\hat{\eta'})}{N} & \left(\frac{1}{v'} + \frac{1}{s'}\right)\frac{\sigma^2(\hat{\eta'})}{N} & \frac{4}{s'}\frac{\sigma^2(\hat{\eta'})}{N} \end{bmatrix} \tag{60}$$

Hence, $\sigma^2\left(\hat{\beta}_2'\right) = \frac{4}{E[r_i'^2]}\frac{\sigma^2(\hat{\eta'})}{N}$ □

## G Bias in ATE and Variance using sampled estimate of trigger intensity: Proof of Theorem 4.3

Here, $E[\epsilon_i] = 0$ and $E[\epsilon_i r_i] = 0$. Thus, $E[r_i'] = E[r_i]$ and $E[r_i'^2] = E[r_i^2] + E[\epsilon_i^2]$

### G.1 Proof of Theorem 4.3: part a

*Proof.* We can re-write the bias in estimated $\hat{\rho'}$.

$$\rho - E[\hat{\rho'}] = \beta_2\left(\frac{E[\epsilon_i r_i']E[r_i']}{E[(r_i')^2]} - E[\epsilon_i]\right) \tag{61}$$

$$= \beta_2\frac{E[\epsilon_i r_i']E[r_i']}{E[(r_i')^2]} \tag{62}$$

$$= \beta_2\frac{E[\epsilon_i^2]E[r_i]}{E[r_i^2] + E[\epsilon_i^2]} \tag{63}$$

$$= \beta_2\frac{(E[r_i] - E[r_i^2])E[r_i]}{(m-1)E[r_i^2] + E[r_i]} \tag{64}$$

$$< \beta_2\frac{E[r_i]^2}{(m-1)E[r_i^2] + E[r_i]} \tag{65}$$

$$< \beta_2\frac{E[r_i]^2}{(m-1)E[r_i^2]} \tag{66}$$

$$< \frac{\beta_2}{m-1} \tag{67}$$

□

### G.2 Proof of Theorem 4.3: part b

$$\sigma^2\left(\hat{\rho}'\right) = \frac{4E[r_i']^2}{E[r_i'^2]}\frac{\sigma^2(\hat{\eta}')}{N} \tag{68}$$

$$= \frac{4E[r_i]^2}{E[r_i^2]+E[\epsilon_i^2]}\frac{\sigma^2(\hat{\eta}')}{N} \tag{69}$$

$$= \frac{4E[r_i]^2}{E[r_i^2]+E[\epsilon_i^2]}\left[\frac{\sigma^2(\eta)}{N} + \left[\left(\beta_1+\frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right]E[\epsilon_i^2]\right] \tag{70}$$

$$< \frac{4E[r_i]^2}{E[r_i^2]}\left[\frac{\sigma^2(\eta)}{N} + \left[\left(\beta_1+\frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right]E[\epsilon_i^2]\right] \tag{71}$$

$$= \sigma^2\left(\hat{\rho}\right) + \frac{4E[r_i]^2}{E[r_i^2]}\left[\left(\beta_1+\frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right]E[\epsilon_i^2] \tag{72}$$

$$< \sigma^2\left(\hat{\rho}\right) + 4E[\epsilon_i^2]\left[\left(\beta_1+\frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right], \text{ as } \frac{E[r_i]^2}{E[r_i^2]} < 1$$

$$\text{, when } \sigma^2(r_i) > 0 \tag{73}$$

$$= \sigma^2\left(\hat{\rho}\right) + 4\frac{E[r_i]-E[r_i^2]}{m}\left[\left(\beta_1+\frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right]$$

$$< \sigma^2\left(\hat{\rho}\right) + \frac{1}{m}\left[\left(\beta_1+\frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right], E[r_i]-E[r_i^2] \leq \frac{1}{4} \tag{74}$$

$$\sigma^2\left(\hat{\rho}'\right) - \sigma^2\left(\hat{\rho}\right) < \frac{1}{m}\left[\left(\beta_1+\frac{\beta_2}{2}\right)^2 + \frac{\beta_2^2}{4}\right] \tag{75}$$