DILUTED TREATMENT EFFECT ESTIMATION FOR TRIGGER ANALYSIS IN ONLINE CONTROLLED EXPERIMENTS

Alex Deng and Victor Hu

February 2, 2015

Microsoft

- Trigger Analysis and The Dilution Problem
- Traditional Approach: Exact Dilution Formula
- Novel Approach: Dilution as Variance Reduction
- Empirical Results
- Illustrative Example (if time permits)

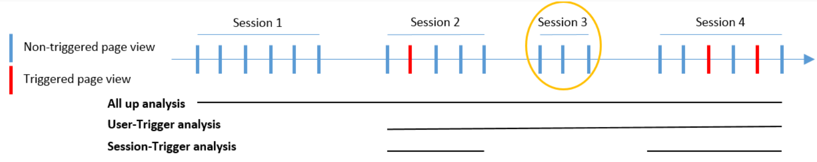
Paper and slides available at: http://alexdeng.github.io/

THE DILUTION PROBLEM

LOW COVERAGE FEATURE AND TRIGGER ANALYSIS

More and more low coverage features:

- A specific type of instant answer, e.g. weather, recipe, celebrity
- Personalized feature triggered on sophisticated criteria
- ...



Trigger analysis/Triggered Analysis: only use triggered data to have a better estimate of the treatment effect and higher statistical power

- Need to compare feature effect at the overall treatment effect level for return of investment (ROI) comparison.
- Should we invest $X to a feature only affecting a small user base or another feature with larger user base?
- Effect estimated in trigger analysis $\neq$ overall(all-up) treatment effect
- Problem: how do we translate estimation(and confidence interval) from trigger analysis to overall treatment effect so ROI calibration is possible?

DILUTION FORMULA

EASY CASE: ADDITIVE METRICS

Metrics like Clicks-per-user and Revenue-per-user are additive, because for each user, $X = TrX + UnTrX$. We proved

$$\Delta_{overall}(X) = \Delta_{Tr}(X) \times \frac{N_{Tr}}{N} \tag{1}$$

- N_{Tr} is the triggered user count. $\frac{N_{Tr}}{N}$ is user trigger rate.
- Effect diluted by the user trigger rate.
- Formula applies to user-trigger and also session-trigger analysis.
- Many people can guess this formula without derivation.

RATIO METRICS

Metrics like Click-Through-Rate (CTR), Session-Success-Rate(SSR) are called Ratio Metrics. Rratio metrics are as common as additive metrics

- Ratio metrics are typically defined as ratio per user, e.g. CTR-per-user, so users have equal weights
- For each user, $X_i = \frac{\text{Numerator}_i}{\text{Denominator}_i}$
- Asked people to guess the formula. It is much harder than additive metrics and the popular guess is:

$$\Delta_{\text{Overall}}(X) = \Delta_{\text{Tr}}(X) \times \frac{N_{\text{Tr}}}{N} \times \overline{\text{TR}},$$

  where $\text{TR} = \frac{\text{TrDenominator}}{\text{Denominator}}$ — "denominator trigger rate" and $\overline{\text{TR}}$ is the average over all triggered users.
- We learned it is WRONG and could be way off!

Theoretical derivation: Rubin Causal Model(potential outcome pairs)

For a ratio metric X, assuming for all users there is no treatment effect on the Denominator, then

$$\Delta_{\text{Overall}}(X) = \frac{1}{N}\sum_{\text{Tr}} \text{TR}_i \times (\text{TrX}_{iT} - \text{TrX}_{iC})$$

$$\stackrel{E}{=} \Delta(\text{TR} \times \text{TrX}) \quad \text{TR} = 0 \text{ and TrX} = 0 \text{ for untriggered user} \qquad (*)$$

where $(\text{TrX}_{iT}, \text{TrX}_{iC})$ is the potential outcome pair. Only when $\text{TrX}_{iT} - \text{TrX}_{iC}$ independent of TR_i, then

$$\Delta_{\text{Overall}}(X) \stackrel{E}{=} \Delta_{\text{Tr}}(X) \times \frac{N_{\text{Tr}}}{N} \times \overline{\text{TR}}. \qquad (2)$$

when effect $\text{TrX}_{iT} - \text{TrX}_{iC}$ has strong correlates with TR, (2) could be way off.

WE NEED BETTER SOLUTION

- Result (*) requires no effect on denominator assumption
- Derivation only guarantees that (*) is an unbiased estimator for the overall treatment effect. There is no guarantee that it will be better than $\Delta(X)$
- What about other types of metrics?

We have a better solution that is unified, elegant and better.

DILUTION VIA VARIANCE REDUCTION

TRIGGER ANALYSIS + DILUTION = VARIANCE REDUCTION

Q: What's the purpose of using trigger analysis and dilution?

1. Straightforward overall treatment effect in all-up analysis fail to use the information about triggering and the fact that un-triggered component of the data contains no information for the treatment effect but only noises.
2. The ultimate goal is to have a more accurate unbiased estimator for the overall treatment effect. Accuracy is measured by variance of the estimator.
3. We want to reduce the variance of the overall treatment effect estimator in the all-up analysis, by using side information contained in the triggering events. In other word, Trigger Analysis+Dilution = Variance Reduction!

VR VIA COVARIATE ADJUSTMENT

X is the metric of interest. Y is covariate(s) such that $E(Y^{(T)}) = E(Y^{(C)})$, i.e. $E(\Delta(Y)) = 0$ (no treatment effect on Y). Y represents side information.

Key Observation: for any θ,

$$E(\Delta(X)) = E\left(\Delta(X) - \theta \times \Delta(Y)\right) = E(\Delta(X)) - \theta \times \underbrace{E(\Delta(Y))}_{0}$$

So $\Delta(X) - \theta \times \Delta(Y)$ is an unbiased estimator for the treatment effect on X.

Task: choose a θ to minimize the variance

$$\underset{\theta}{\text{argmin}} \, \mathrm{Var}(\Delta(X) - \theta \times \Delta(Y))$$

DILUTION AS VR CONT'D

Closed-form optimal θ:

$$\theta^* = \frac{\mathrm{Cov}(\Delta(X), \Delta(Y))}{\mathrm{Var}(\Delta(Y))} = \frac{\mathrm{Cov}(\overline{X_T}, \overline{Y_T}) + \mathrm{Cov}(\overline{X_C}, \overline{Y_C})}{\mathrm{Var}(\overline{Y_T}) + \mathrm{Var}(\overline{Y_C})}$$

The second equation may vary if the metric is not in the form of an average, e.g. percentile metrics. θ is a vector when Y takes vector value and the above becomes matrix algebra

If we want want to estimate treatment effect on metric X

1. Identify a set of covariates Y.
2. θ^* can be first estimated from the data.
3. $\Delta(X) - \theta^* \Delta(Y)$ is an unbiased estimator and it has smaller variance!

For any metric X, trigger event information naturally implies UnTrX (the same metric X calculated using trigger-complement data) is a covariate!

For ratio metric, if we further can assume no effect on the denominator trigger rate, we might use TR (demoninator trigger rate) as additional covariate.

Technical remark: UnTrX might not be well defined, e.g. for ratio metrics when a user always trigger the feature. In this case we can define UnTrX=0 and add a binary indicator TR==1 as covariate (see paper for details)

VR VS. DILUTION FORMULA

- VR is a unified framework that works for all types of metrics. Dilution Formula approach need different formula for additive and ratio metrics.
- The ultimate purpose is to get a better estimation of the overall treatment effect. VR approach this in a straight line while trigger analysis + dilution is indirect
- VR quantifies the reduction of variance using the side information of feature triggering. It is strictly better than trigger analysis + dilution formula in most cases (Section 4.4). Using empirical results we show exact dilution formula (*) can perform worse than $\Delta(X)$ when trigger rate is high but VR always outperforms

COVARIATE ADJUSTMENT VS. LINEAR REGRESSION

- VR method mentioned here first published in Deng et. al. (WSDM 2013)
- It resembles linear regression but they are different! There is no linear model assumption whatsoever required. (David Freedman 2008: Randomization does not justify the assumption behind OLS model)
- Yang and Tsiatis et. al. (2001) showed a similar estimator (ANCOVA-II) based on semi-parametric theory. Also see Tsiatis, Davidian, Zhang and Lu (2008), Targeted Learning (Mark J. van der Laan and Sherri Rose, 2011). Deng et. al. 2013 is a much simpler derivation.

EMPIRICAL RESULTS

Experiments	Trigger Rate	
	%Triggered Users	%Triggered Sessions
ExpA	5.26%	1.27%
ExpB	33.46%	20.83%
ExpC	65.17%	60.35%

Table: Trigger User and Trigger Session Ratios.

Compare 3 methods:

1. Exact Dilution Formula $\Delta(\text{TR} \times \text{TrX})$ (*)
2. VR response $\text{TR} \times \text{TrX}$ and covariates $Y = (\text{UnTrX}, \text{TR}, \text{TR} == 1)$ (**)
3. VR response X and covariates $Y = (\text{UnTrX}, \text{TR}, \text{TR} == 1)$ (***)

VARIANCE REDUCTION RATE

Experiments	User Trigger		
	Exact Formula(*)	VR **	VR ***
	VR rate	VR rate	VR rate
ExpA	88.60% (11.40%)	**98.42% (1.58%)**	95.60% (4.40%)
ExpB	-17.14% (117.14%)	**84.80% (14.20%)**	78.57% (21.43%)
ExpC	-49.47% (149.47%)	**61.45% (38.55%)**	36.03% (63.97%)

Table: User Trigger Variance Reduction comparison.

Experiments	Session Trigger		
	Exact Formula(*)	VR **	VR ***
	VR rate	VR rate	VR rate
ExpA	97.12% (2.88%)	**99.44% (0.56%)**	98.25% (1.75%)
ExpB	28.12% (71.88%)	**89.85% (10.15%)**	85.99% (14.01%)
ExpC	-31.32% (131.32%)	**69.10% (30.90%)**	53.97% (46.03%)

Table: Session Trigger Variance Reduction Comparison.

Questions?

ILLUSTRATIVE EXAMPLE

EXAMPLE

Group	User	S1	S2	S3	S4	S5	X	TR	TrX	UnTrX	TR=1
T	A	1	**0**	0	0	1	2/5	1/5	0	1/2	0
T	B	**1**	**1**	**0**	**1**		3/4	1	3/4	0	1
T	C	**1**	0	0			1/3	1/3	1	0	0
T	D	0	0	0			0	0	0	0	0
C	E	0	1	**0**	1	1	3/5	1/5	0	3/4	0
C	F	**1**	**1**	**1**			1	1	1	0	1
C	G	0	0	1			1/3	0	0	1/3	0
C	H	0	**1**	0	0		1/4	1/4	1	0	0

1. Estimate θ^*. Note that $\theta^* = \mathrm{Var}(\mathsf{Y})^{-1} \times \mathrm{Cov}(\mathsf{X}, \mathsf{Y})$ where Y is the vector $(\mathsf{UnTrX}, \mathsf{TR}, \mathsf{IsTR} = 1)$ and Var and Cov here are matrices. For the control group, θ^* is

$$\begin{pmatrix} 0.127 & -0.081 & -0.090 \\ -0.081 & 0.192 & 0.213 \\ -0.090 & .213 & 0.250 \end{pmatrix}^{-1} \times \begin{pmatrix} -0.010 \\ 0.130 \\ 0.151 \end{pmatrix} = \begin{pmatrix} 0.488 \\ 0.317 \\ 0.512 \end{pmatrix}.$$

EXAMPLE

Group	User	S1	S2	S3	S4	S5	X	TR	TrX	UnTrX	TR=1
T	A	1	**0**	0	0	1	2/5	1/5	0	1/2	0
T	B	**1**	**1**	**0**	**1**		3/4	1	3/4	0	1
T	C	**1**	0	0			1/3	1/3	1	0	0
T	D	0	0	0			0	0	0	0	0
C	E	0	1	**0**	1	1	3/5	1/5	0	3/4	0
C	F	**1**	**1**	**1**			1	1	1	0	1
C	G	0	0	1			1/3	0	0	1/3	0
C	H	0	**1**	0	0		1/4	1/4	1	0	0

1. Estimate θ^*. Note that $\theta^* = \mathrm{Var}(\mathsf{Y})^{-1} \times \mathrm{Cov}(\mathsf{X}, \mathsf{Y})$ where Y is the vector $(\mathsf{UnTrX}, \mathsf{TR}, \mathsf{IsTR} = 1)$ and Var and Cov here are matrices. For the control group, θ^* is

$$\begin{pmatrix} 0.127 & -0.081 & -0.090 \\ -0.081 & 0.192 & 0.213 \\ -0.090 & .213 & 0.250 \end{pmatrix}^{-1} \times \begin{pmatrix} -0.010 \\ 0.130 \\ 0.151 \end{pmatrix} = \begin{pmatrix} 0.488 \\ 0.317 \\ 0.512 \end{pmatrix}.$$

2. VR Estimation:

$$\Delta_{\mathsf{VR}} = \Delta(\mathsf{X}) - 0.488 \times \Delta(\mathsf{UnTrX}) - 0.317 \times \Delta(\mathsf{TR}) - 0.512 \times \Delta(\mathsf{IsTR} = 1)$$
$$= -0.175 - 0.488 \times (-0.145) - 0.317 \times 0.021 - 0.512 \times 0 = -0.111$$

3. To get z-score, we also need to calculate the variance. which is

$$\begin{aligned}&\mathrm{Var}(\overline{X}_T) + \mathrm{Var}(\overline{X}_C) + (\theta^*)^\intercal \left(\mathrm{Cov}(\overline{Y}_T) + \mathrm{Cov}(\overline{Y}_C)\right) \theta^* \\ &- 2 \times (\theta^*)^\intercal \left(\mathrm{Cov}(\overline{X}, \overline{Y}_T) + \mathrm{Cov}(\overline{X}, \overline{Y}_C)\right) \\ &= 0.00435\end{aligned}$$

4. Z-score is then $-0.111/\sqrt{0.00434} = -1.685$.
5. Variance reduction rate: $1 - 0.00435/0.086 = 95.0\%$.
6. Confidence interval follows trivially