Causal Inference with Bilibili Virtual-idol Dataset

Based on regression adjustment and causal forest

Jiahui Xin

Institute of Statistics and Big Data

Renmin University of China

April 18, 2023

Presentation Overview

Dataset

Variable	Description
Live	Liver name
Area	Live partition
DanmakusCount	Number of danmakus
StartDate	Start time of live
StopDate	End time of live
Title	Live title
TotalIncome	Total live income
WatchCount	Number of live viewers
InteractionCount	Number of live interactioners
SuperchatIncome	Live superchat income
SuperchatTimestamps	Live superchat timestamps
MembershipIncome	Live membership income
MembershipTimestamps	Live membership timestamps
affiliation	Company of liver
sex	Gender of liver
country	Country of liver
ip	IP address of liver

Figure: Variable Description.

Linearity & difference

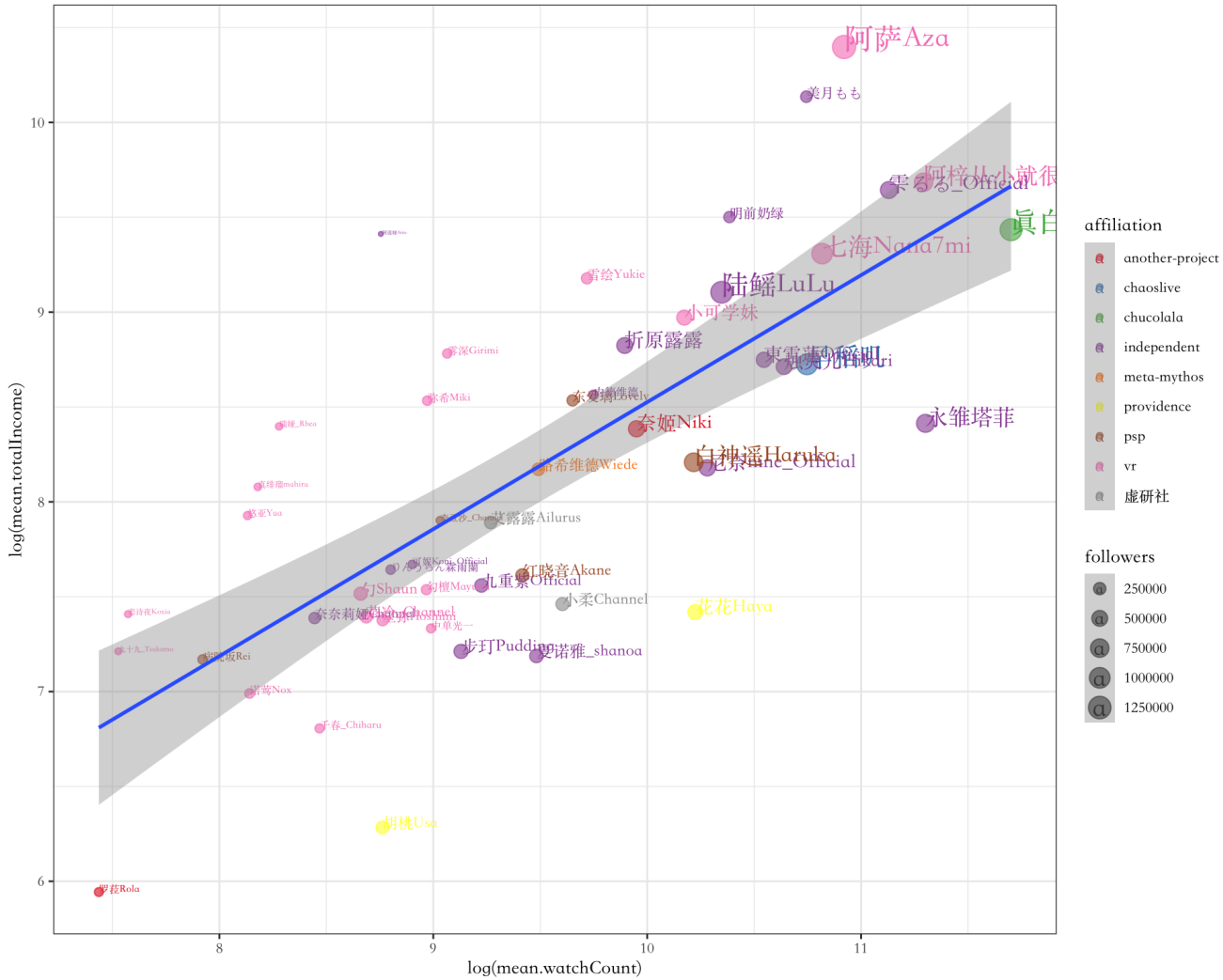


Figure: It seems that linear model works well.

Linearity & difference (Cont.)

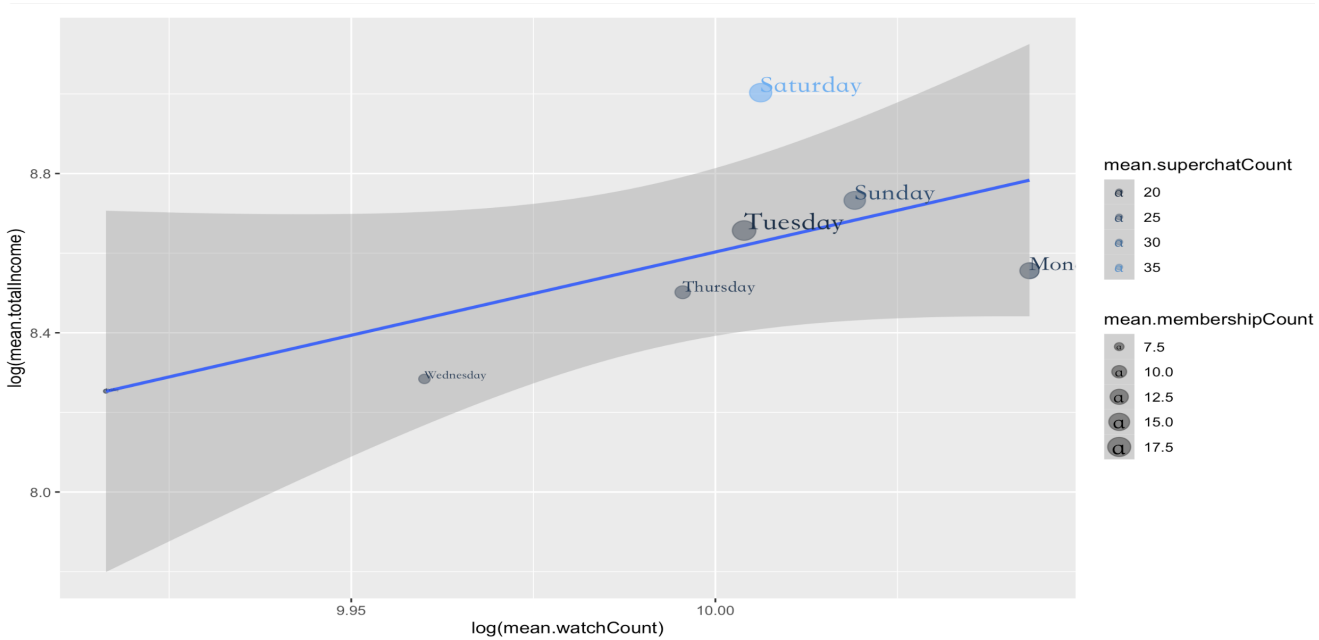


Figure: It seems that linear model works well .

Questions

- Is there difference between the income of independent vups and affiliated vups?
- Is there difference between the income of weekday lives and weekend lives?

Under the assumption that W_i is unconfounded given X_i (i.e., treatment is as good as random given covariates), we can posit the partially linear model:

$$Y_i = \tau(X_i)W_i + f(X_i) + \varepsilon_i, E[\varepsilon_i \mid X_i, W_i] = 0$$

where $\tau(X_i)$ is the conditional average treatment effect $E[Y(1) - Y(0) \mid X_i = x]$.

But how do we get around estimating τ when we do not know $f(X_i)$?

grf: Generalized Random Forests

If we define the following two intermediary objects:

$$e(x) = E\left[W_i \mid X_i = x\right]$$

the propensity score, and

$$m(x) = E\left[Y_i \mid X_i = x\right] = f(x) + \tau e(x)$$

the conditional mean of Y, then we can rewrite the above equation in "centered" form:

$$Y_i - m(x) = \tau(X_i)\left(W_i - e(x)\right) + \varepsilon_i$$

grf (Cont.)

If we imagine we had access to some neighborhood $\mathcal{N}(x)$ where τ was constant, we could proceed exactly as before, by doing a residual-on-residual regression on the samples belonging to $\mathcal{N}(x)$, i.e.:

$$\tau(x) := \text{lm}\left(Y_i - \hat{m}^{(-i)}(X_i) \sim W_i - \hat{e}^{(-i)}(X_i),\ \text{weights} = 1\{X_i \in \mathcal{N}(x)\}\right)$$

These weights play a crucial role and grf find them using Breiman's random forest as an adaptive neighborhood finder.

Regression adjustment in RCT

Robustness and efficiency

Papers

1. [Freedman, 2008] showed that even in RCTs, regression adjustment cannot be guaranteed to improve precision.
2. [Lin, 2013] pointed out that if using regression with interactions, it never hurts.
3. [Ma et al., 2022] further exploited regression adjustment under CAR setting.

Using the terms from [Ma et al., 2022], I will use 3 estimators.

1. difference-in-means
2. regression with additional covariates (without interaction)
3. regression with additional covariates (with interaction).

Observational data?

There are two key assumptions: unconfoundedness and overlap.

The former is plausible if controlling for detailed characteristics and the latter can be assessed by the estimated propensity score.

Diagnostics

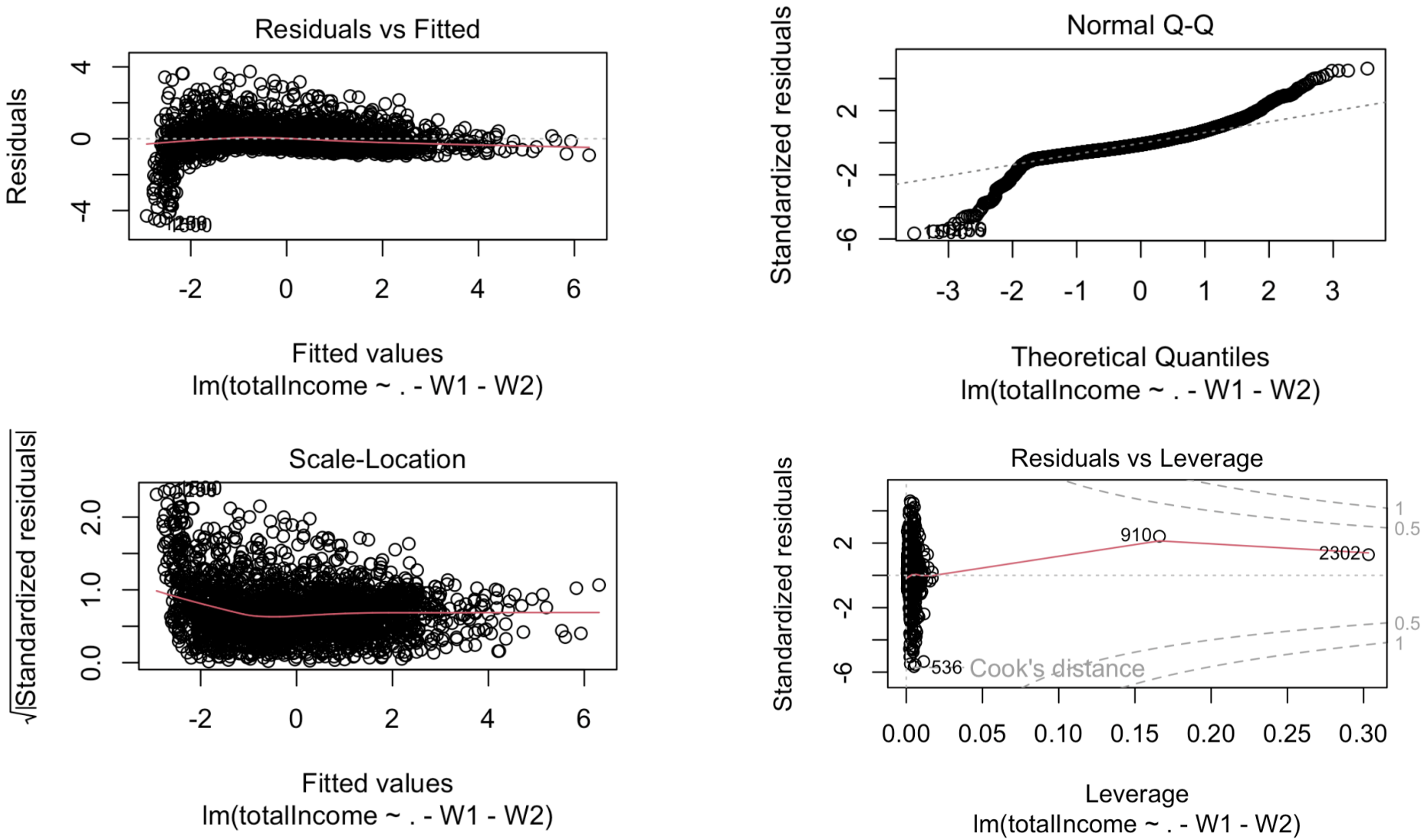
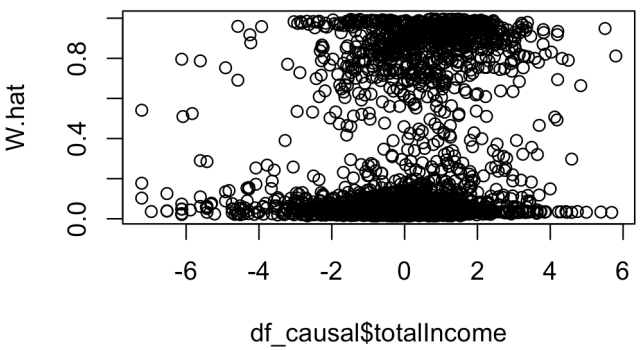
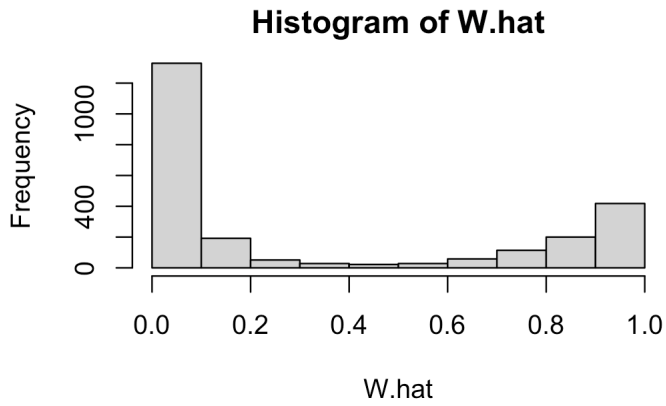


Figure: Random forest has almost same prediction error with OLS.

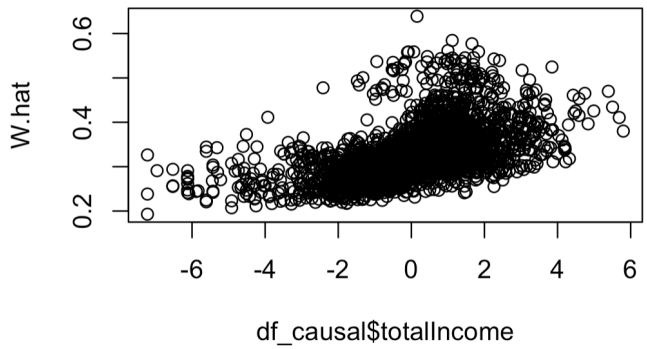
Diagnostics (Cont.)



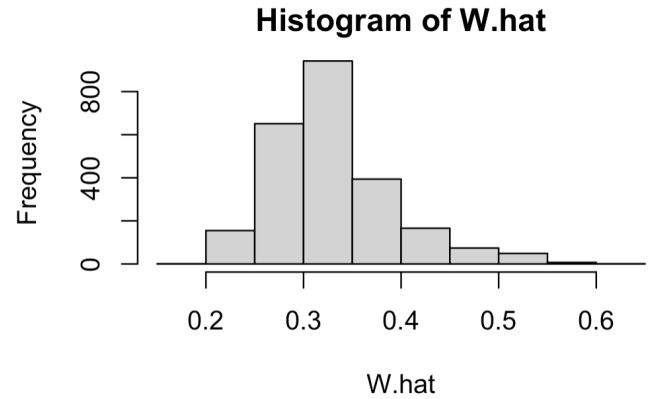
(a) Treatment **Independent**



(b) Treatment **Independent**



(c) Treatment **Weekend**



(d) Treatment **Weekend**

Figure: Estimated propensity score with random forest.

ATE

Point estimation

Method \ Treatment	Independent	Weekend
τ	$(0.513, 0.797)$	$(0.229, 0.518)$
τ^*	$(0.074, 0.220)$	$(-0.017, 0.121)$
$\tau^*_{\textbf{interact}}$	$(0.153, 0.302)$	$(-0.020, 0.119)$
Causal Forest	$(0.023, 0.442)$	$(0.004, 0.122)$

Table: Estimated Confidence Interval (point estimator $\pm$ 1.96 standard deviation)

CATE of treatment **Independent**

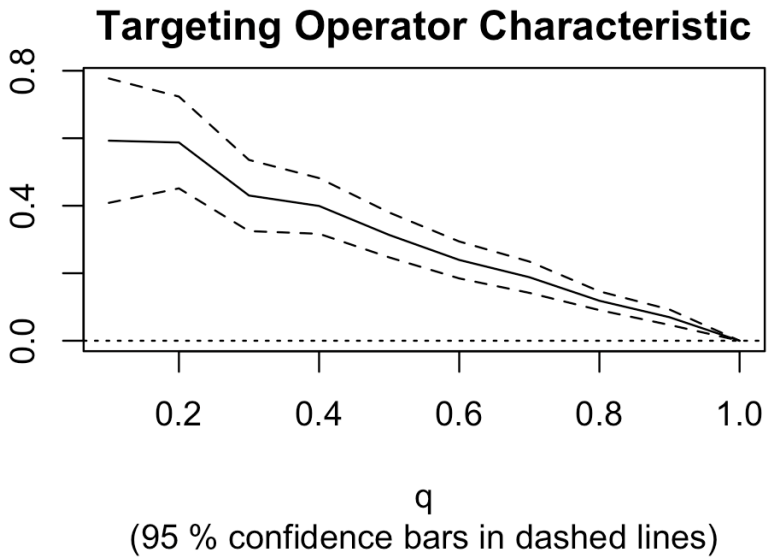
Best linear projection of the conditional average treatment effect. Confidence intervals are cluster- and heteroskedasticity-robust (HC3):

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	0.3623580	0.0201952	17.9428	$< 2.2e-16$
danmakusCount	−0.0019885	0.0635215	−0.0313	0.97503
timeDuration	0.4228089	0.0508939	8.3077	$< 2.2e-16$
watchCount	−0.0450065	0.0611101	−0.7365	0.46151
interactionCount	0.0182660	0.1024029	0.1784	0.85844
superchatCount	−0.2159035	0.0234864	−9.1927	$< 2.2e-16$
membershipCount	0.0407307	0.0193660	2.1032	0.03555
followers	−0.2810257	0.0451630	−6.2225	$5.744e-10$

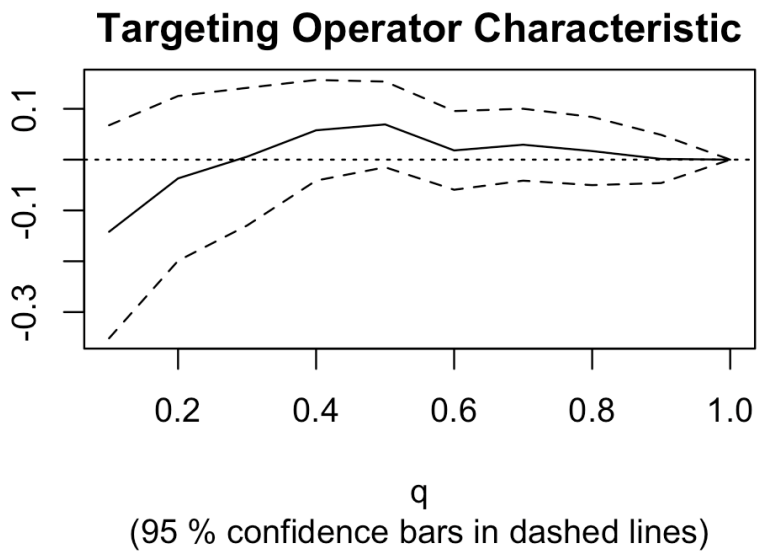
$$\text{TOC}(q) = E\left[Y_i(1) - Y_i(0) \mid \hat{\tau}(X_i) \geq F^{-1}_{\hat{\tau}(X_i)}(1-q)\right] - E\left[Y_i(1) - Y_i(0)\right]$$

($F(\cdot)$ is the distribution function).

I.e. at $q = 0.2$ the TOC quantifies the incremental benefit of treating only the 20% with the largest estimated CATEs compared to the overall ATE. We refer to the area under the TOC as the AUTOC.



(a) Treatment **Independent** has **AUTOC** : 0.33 ± 0.06



(b) Treatment **Weekend** has **AUTOC** : -0.01 ± 0.09

Figure: TOC curves

Conclusion

- With a real-world dataset, I used both regression adjustment and causal forest to exploit causal effect.
- Key assumptions are hard to verify but the results are similar.
- Given variables {**danmakusCount,timeDuration, watchCount, interactionCount,superchatCount, membershipCount,followers**}, treatment **Independent** has significant causal effect but treatment **Weekend** does not.

References

David A. Freedman (2008)
On regression adjustments to experimental data.
Advances in Applied Mathematics 40, no. 2 (2008): 180–193.

Winston Lin (2013)
Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique.
The Annals of Applied Statistics (2013): 295–318.

Wei Ma, Fuyi Tu, and Hanzhong Liu (2022)
Regression analysis for covariate-adaptive randomization: A robust and efficient inference perspective.
Statistics in Medicine 41, no. 29 (2022): 5645-5661.

Susan Athey, Julie Tibshirani, and Stefan Wager (2019)
Generalized random forests.
The Annals of Statistics 47, no. 2 (2019): 1148-1178.

Acknowledgements

Support

- NNNK-abaabatu@weibo
- sizukululu@bilibili

Dataset wandleshen@github

The End

Questions? Comments?