Causal Inference with Bilibili v-tuber data set: Does weekday streaming or v-tuber affiliation have causal effect on live streaming income?

Jiahui Xin

May 6, 2023

1 Introduction

Virtual idols have been a part of Japanese culture since the 1980s, drawing inspiration from anime and idol traditions. In recent years, the rise of the internet and advanced technology has given birth to a new form of virtual entertainer: the v-tuber or virtual YouTuber. These individuals use computer-generated avatars controlled by motion capture software to engage with their audiences in real-time.

One platform that has become particularly popular for v-tuber live streaming is Bilibili, where tens of thousands of virtual entertainers, also known as "vups," earn substantial incomes each month. According to the Danmakus API, some vups can earn up to 1 million CNY per month, making Bilibili's virtual idol market a billion-CNY industry.

Given the immense popularity of v-tubers and their potential for high earnings, it is natural to ask what factors contribute to their success. In this report, we focus on the impact of two key factors on v-tuber earnings: *weekday* streaming and v-tuber *affiliation*. Specifically, we analyze real-world data from various v-tubers to determine whether these factors have a causal effect on earnings. To do so, we employ a causal inference framework and sophisticated analytical tools, ultimately revealing that while *weekday* has no significant effect on earnings, *affiliation* does. We also present a well-fitted linear model using log-transformed variables to illustrate our findings.

2 Data exploration

2.1 Data set

The observational data is from wandleshen@github. They collect data with Danmakus from last 50 live streams of 50 livers each at 2023-02-10 16:42:54, all 2477 live streams. (Not every liver has 50 live streams.) Vtuber information data are manually collected from homepages, in which *liver,id,affiliation* are from wandleshen@github and *sex,country,ip* are from myself. The most interested variable is *totalIncome* and see Figure 1 for variable details.

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 1: Variable Description

2.2 Exploratory data analysis

First we illustrate the correlation between numerical variables as Figure 2. It shows 5 variable clusters: $\{totalIncome, membershipCount, membershipIncome\}$, $\{danmakusCount, interactionCount\}$, $\{superchatIncome, superchatIncome\}$, $\{timeDuration\}$ and $\{followers\}$.[1]

Even though correlation does not mean causation, it still helps to build model. From Figure 2 we notice that the linear dependence between *totalIncome* and *membershipIncome* (or *membershipCount*) is really strong and other variables contribute little. For factor variables of v-tubers we bi-plot the log-mean of *watchCount* versus log-mean of *totalIncome* with legends. Figure 3 and 4 reveals that linearity holds well and both *weekday* and *affiliation* relate with *totalIncome*.

We will not stop at the correlation. The question follows: Does weekday streaming or v-tuber affiliation have causal effect on live streaming income? In the following analysis we drop the two variables $\{SuperchatIncome, MembershipIncome\}$ because it is impossible to directly intervene on them.[2]

3 Method

Causal inference tools are rich in both randomized clinical trails (RCT) and observational data while they heavily rely on two key assumptions: unconfoundedness and overlap. In RCT the two assumptions hold naturally because we design the experiment. However, in observational case the two assumptions especially unconfoundedness generally cannot be verified.

[1] We transform the two variables $\{StartDate, StopDate\}$ to $\{weekday, timeDuration\}$.

[2] Actually we did PCA and kmeans clustering but found nothing interesting. We will also not use $Area, sex, country, ip$ by some findings from EDA.

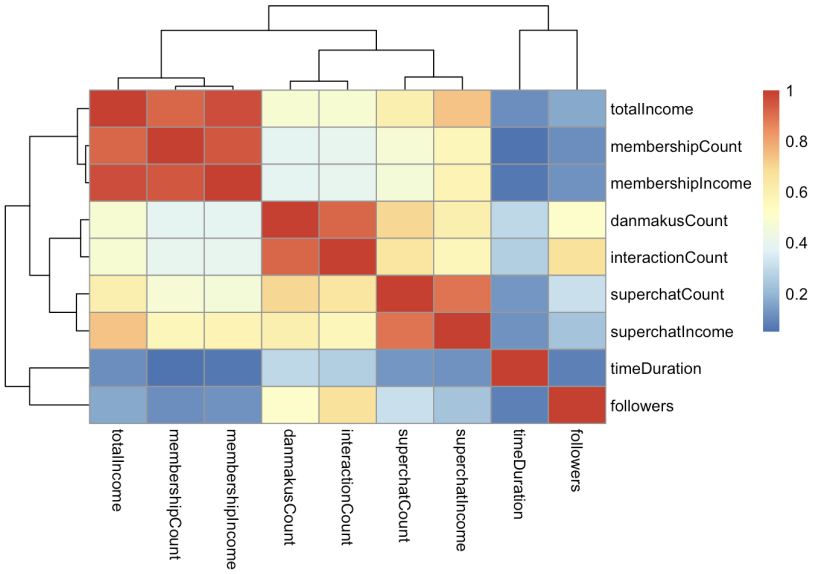

Figure 2: Correlation map with clustering.

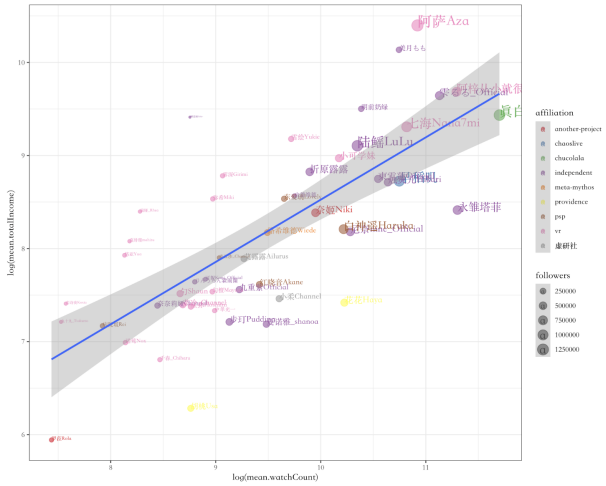

Figure 3: Income per live of different v-tubers. It seems that independent v-tubers (purple points) earn more than affiliated v-tubers.

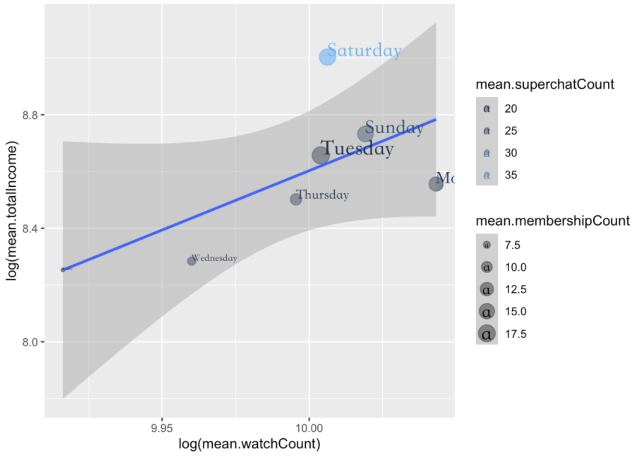

Figure 4: It seems that weekend streams are more profitable than weekdays.

Despite the risk of violating assumption, we can do some diagnostics based on observational data. Below we consider two common tools: regression adjustment ([Fre08, Lin13, MTL22]) and causal forest ([WA18]) as well as one non-parametric tool matching ([HIKS11]). With or without

matching, we do same regression adjustment and causal forest to estimate the average treatment effect (ATE).

3.1 Regression adjustment

Regression adjustment is almost basic procedure in RCT. However, [Fre08] showed that even in RCT, regression adjustment cannot be guaranteed to improve precision. On this critique, [Lin13] pointed out that if using regression with interactions, it never hurts. [MTL22] further exploited regression adjustment under CAR setting and showed its robustness and efficiency.

Linear model has great interpretability for relationship among variables. Although Bilibili v-tuber data set is observational, we still use the regression adjustment with R function $lm()$ to estimate ATE.

3.2 Causal forest

Causal forest ([WA18]) is designed to estimate conditional average treatment effect (CATE) with random forest. There are two steps: First estimate expected outcome $E[Y(x)]$ as $\hat{m}(x)$ and propensity score $E[W(x)]$ as $\hat{e}(x)$ and then evaluate $\tau(x)$ with one residual-residual weighted regression.

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.:

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

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

The R package grf is used to implement causal forest.

3.3 Matching

The R package $MatchIt$ ([HIKS11]) aims to reduce model dependence and make inferences less sensitive to modeling assumptions by preprocessing data with matching methods. Typically we can do any analysis after matching and compare the matching version with un-matching version. If the results are similar, we will be more confident in our model assumptions.

4 Results

We only use 10 variables: $\{danmakusCount, timeDuration, totalIncome, watchCount, interactionCount, superchatCount, membershipCount, followers, W_1, W_2\}$ in which $W_1 = 1$ (0) means independent (affiliated) v-tuber and $W_2 = 1$ (0) means weekend (workday) live, respectively.

All variables are in log-scale except the two treatment variables W_1, W_2. We de-mean all log-scale variables.

4.1 Linear model

We first regress $totalIncome$ on all other variables. Figure 5 shows that only 5 variables are significant: $danmakusCount, timeDuration, superchatCount, membershipCount$ and W_1. Among

significant variables, *membershipCount* and *superchatCount* have the largest coefficients. The simple linear model is written:

$$\begin{aligned}\log(totalIncome) \sim 0.147 * W_1 + 0.134 * \log(danmakusCount) + 0.117 * \log(timeDuration) \\ + 0.420 * \log(superchatCount) + 0.623 * \log(membershipCount)\end{aligned} \tag{1}$$

Standard diagnostic plots are passed and omitted here. However, diagnostic tests does not justify model assumption and we will use different tools to estimate ATE.[3]

```
##
## Call:
## lm(formula = totalIncome ~ ., data = df_causal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5903 -0.3885 -0.0908  0.3485  3.8241
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.0306519  0.0236359  -1.297  0.19481
## danmakusCount     0.1339164  0.0519730   2.577  0.01003 *
## timeDuration      0.1167450  0.0383133   3.047  0.00234 **
## watchCount        0.0125995  0.0397595   0.317  0.75135
## interactionCount -0.0639256  0.0755090  -0.847  0.39730
## superchatCount    0.4202298  0.0189408  22.186  < 2e-16 ***
## membershipCount   0.6225290  0.0157685  39.479  < 2e-16 ***
## followers        -0.0009806  0.0313762  -0.031  0.97507
## W1                0.1467601  0.0371184   3.954 7.91e-05 ***
## W2                0.0513868  0.0351373   1.462  0.14375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8073 on 2430 degrees of freedom
## Multiple R-squared:  0.7809, Adjusted R-squared:  0.7801
## F-statistic: 962.5 on 9 and 2430 DF,  p-value: < 2.2e-16
```

Figure 5: Summary of simple linear model.

4.2 ATE

We use 3 estimators: regression adjustment without interaction, regression adjustment with interaction and causal forest. The first two estimators are from [MTL22] and the last one is from [WA18]. Each estimator has matching version and un-matching version where matching method is from [HIKS11].

We only test the overlap assumption (See Figure 6,7,8,9). It seems that the treatment variable W_1 ($Weekend$) may violate overlap assumption but W_2 ($Independent$) does not. Whatsoever, even passing the test cannot guarantee assumptions. Let's see the final numerical result. Table 1 shows that the treatment *Independent* has significant causal effect, approximately 0.2. It means that

$$E[\log(Y(W_1 = 1)) - \log(Y(W_1 = 0))] = E[\log\left(\frac{Y(W_1 = 1)}{Y(W_1 = 0)}\right)] \approx 0.2 \tag{2}$$

[3]We considered random forest here and the prediction ability of random forest is almost exactly same as simple linear regression so we omit it here.

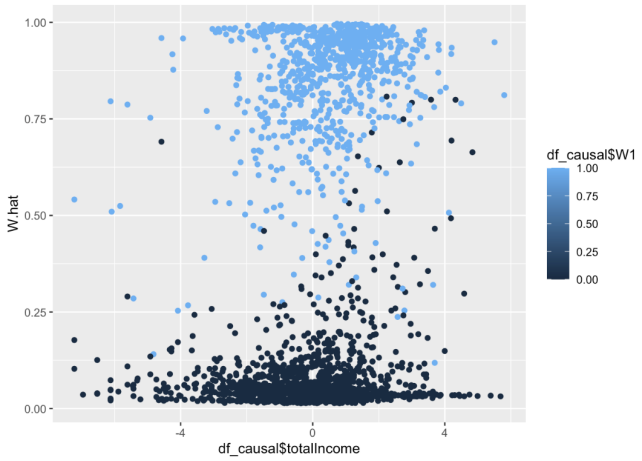

Figure 6: Bubble plot of estimated propensity scores of W_1 with grf.

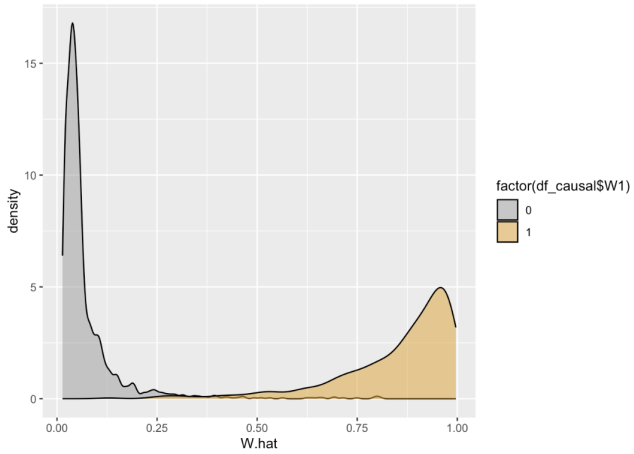

Figure 7: Density of estimated propensity scores of W_1 with grf.

Method \ Treatment	Independent	Weekend
τ^* without matching	$0.147,\ (0.074, 0.220)$	$0.052,\ (-0.017, 0.121)$
τ^* with matching	$0.213,\ (0.016, 0.320)$	$0.021,\ (-0.059, 0.100)$
$\tau^*_{\mathbf{interact}}$ without matching	$0.226,\ (0.151, 0.300)$	$0.048,\ (-0.021, 0.117)$
$\tau^*_{\mathbf{interact}}$ with matching	$0.217,\ (0.122, 0.311)$	$0.021,\ (-0.058, 0.100)$
Causal forest without matching	$0.237,\ (0.030, 0.443)$	$0.064,\ (0.005, 0.123)$
Causal forest with matching	$0.309,\ (0.000, 0.619)$	$0.021, (-0.058, 0.100)$

Table 1: Point estimator with 95% confidence interval (point estimator $\pm$ 1.96 standard deviation)

where we denote $totalIncome$ as Y. It is also notable that by Jensen's inequality,

$$\log\left(E\left[\frac{Y(W_1=1)}{Y(W_1=0)}\right]\right) \geq E\left[\log\left(\frac{Y(W_1=1)}{Y(W_1=0)}\right)\right] \approx 0.2$$

which means that

$$E\left[\frac{Y(W_1=1)}{Y(W_1=0)}\right] \geq \exp(0.2) \approx 1.221. \tag{3}$$

However, the treatment $Weekend$ has no significant causal effect.

In general results with and without matching version are similar. It gives our more confidence for our model assumption.

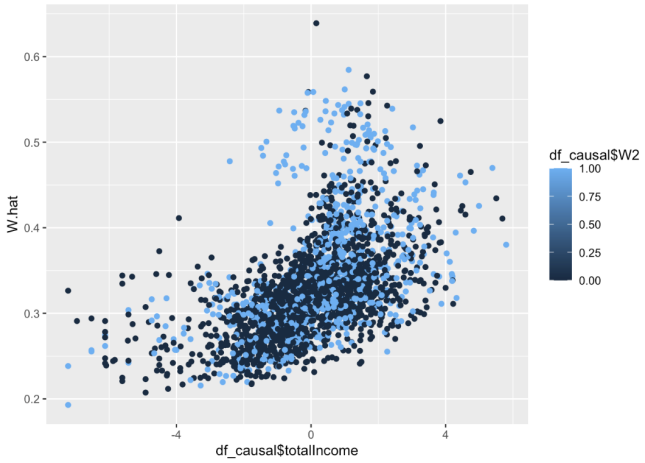

Figure 8: Bubble plot of estimated propensity scores of W_2 with grf.

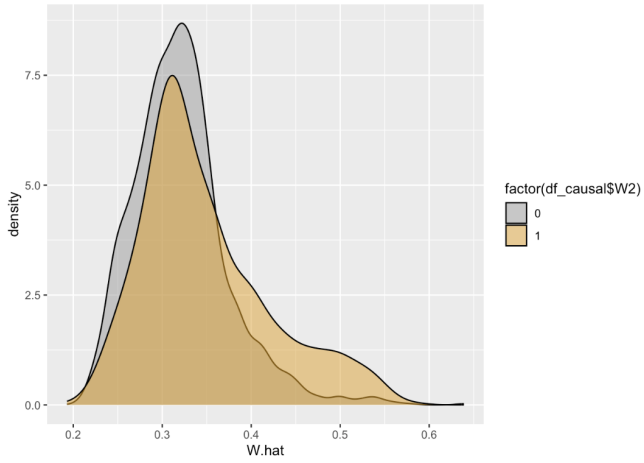

Figure 9: Density of estimated propensity scores of W_2 with grf.

5 Discussion

In this report, we analyze the causal effect of v-tuber affiliation on earnings using real-world data. Specifically, we investigate the difference in earnings between independent v-tubers and those who are affiliated with a larger organization. To do so, we employ both regression adjustment and causal forest techniques, as well as a matching method to ensure robustness and test assumptions.

The results of our analysis indicate that v-tuber affiliation has a significant causal effect on income. In particular, independent v-tubers earn at least 22.1% more than their affiliated counterparts, when other variables are held constant. This finding suggests that independent v-tubers enjoy higher levels of user loyalty and intention to pay compared to their affiliated counterparts, which is consistent with our intuition.

In contrast, our analysis shows that v-tubers streaming on weekends do not experience a significant causal effect on their income. While it may be true that certain times of the week generate more views, this does not necessarily translate into a change in the user structure or other factors that could affect income.

Overall, this analysis highlights the importance of understanding the causal effects of various factors on v-tuber income. By leveraging real-world data and advanced statistical techniques, we can gain a deeper understanding of the factors that drive success in this rapidly evolving field.

Acknowledgement

Many thanks to data collector wandleshen and Oshi of the author TuTu. The author also appreciate Prof. Ma and CAR group students. Their kindness, enthusiasm and selflessness are the driving force for me to move forward.

References

[Fre08] David A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

[HIKS11] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[Lin13] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295 – 318, 2013.

[MTL22] Wei Ma, Fuyi Tu, and Hanzhong Liu. Regression analysis for covariate-adaptive randomization: A robust and efficient inference perspective. *Statistics in Medicine*, 41(29):5645–5661, 2022.

[WA18] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.