The Association Between the Social Screen Time and Pickups Times of Smartphone among the First-year Graduates in Biostatistics

Han Zhang, Xin Luo, Ruiqi Ren

Abstract

The prevalence of mobile touch screen devices, e.g. smartphones and tablets, causes the rise of mobile device addiction and a series of other social problems. This study aims to investigate the social screen time in relation to the frequency of smartphones' pickups and other factors influencing use.[1] For the security and privacy of the data sharing, the federated learning model of simple linear regression was implemented. The result shows that there doesn't exist a significant association between the social screen time and pickups. Next meta learning with multiple linear regression model was introduced to solve the case of unavailability of likelihood function by Best Linear Unbiased Predictor(BLUP). The final meta learning model is $y = -12.534 + 0.588 \times x_1 + 5.4231 \times x_2 + 4.2544 \times x_3$, where y denotes social app usage time and x_1, x_2, x_3 represents number of pickups, first pickup time and whether the day is weekend. Additionally, we conducted the confirmation analysis, which showed that the result from simple linear regression is identical to that from merged data, whereas traditional multiple linear regression output is completely divergent from meta learning. This range is due to the large gap between regression results of individuals and some other limitations. P-value of number of pickups in traditional linear model is 0.506, so in this population, under 95% confidence level, social app usage time is not significantly linearly associated with pickup times, adjusting confounders 'first pickup time' and 'if weekend'.

1 Introduction

Federated learning[5] and meta learning[4] are extremely popular techniques widely applied into a large number of fields including public health. The former plays an important role in data privacy and the latter takes advantage of self-learning ability through multi-tasks. This project utilizes these methods to achieve our goal: detecting the association between people's social time on the phone and the number of phone pickup times. Our hypothesis is that the more frequently do people pick up their phone, the more time they spend on social apps.[1] Our motivation originates from a mind to supervise ourselves to reduce the time spent on communicating with others on social media. The data is daily collected by three team members independently, with used features as 'social app usage time', 'pickup frequency', 'whether the day is weekend'[3] and 'first pickup time'. Workflow of the major analysis is from federated learning with simple linear regression to meta learning with multiple linear regression, which adds confounders of 'pickup number'. In addition we made confirmed analysis to compare results of linear models with these two methods, which shows that federated learning output is consistent with simple linear regression while outcome of meta learning has high discrepancy with multiple linear regression.

2 Data Description

Firstly, we collected 13 baseline covariates from three panelists(see appendix). Two of us are male and one is female. All of us are in our 20s and received our previous degree in China, no pets at home, own 1-3 mobile devices and 3-4 social apps per person, 13-15 credits this winter semester, and have slight procrastination. One of us has one sibling. What's more, we draw the correlation plots between social screen time, total pickups, first pickup time and whether the day is weekend or not.(See Figure 1)

As we can see from Figure 1 (d), 2 peaks in the first pick up time indicate we picked up our phones around 7am and 8am everyday. From the boxplots, total pickups are less and first pickup time is later on weekends. There is a positive correlation between

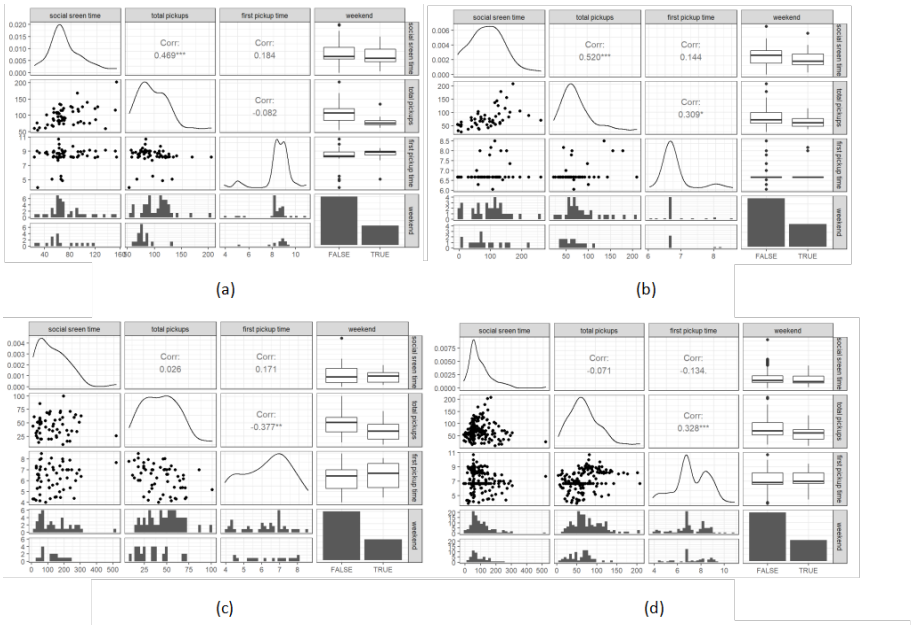

Figure 1 (a)~(d) represent the correlation plots from member 1, 2, 3 and merged data

social screen time and pickups, social time and first pickup from 3 individuals, but a negative correlation appears in the merged data, which is Simpson paradox. Also, the reason why the dots in plots between first pickup time and the other two continuous variables centered around a horizontal line is that members set the alarm clock every morning at the same time. All the characteristics from the correlation plots corresponds to Figure 2 and 3, where the black dots in Figure 2 indicate that day is weekend. Also, VIF values of 3 covariates, which is used to test multicollinearity, are about 1.1-1.2 in separate and combined dataset.(see appendix). Therefore, we can conclude that there's no highly-correlated covariates.

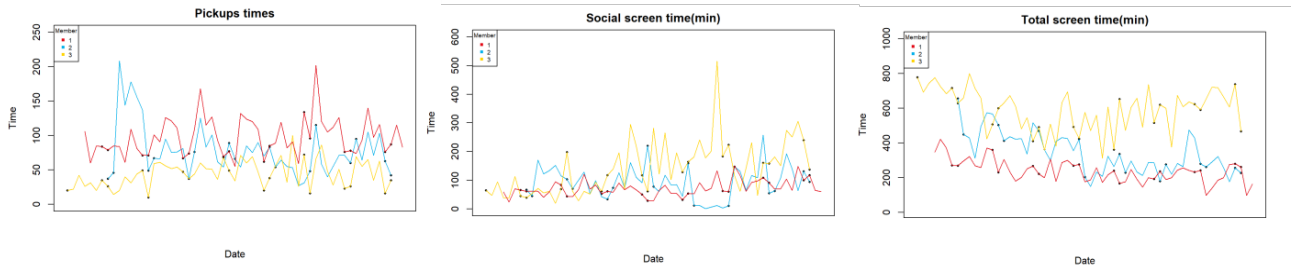

Figure 2 Pickup times, social screen time and total screen time

In addition, we use Cook's distance(cutoff=4/number of data) to measure the influence of data points(See appendix). It turns out that 4 data points are highly influential

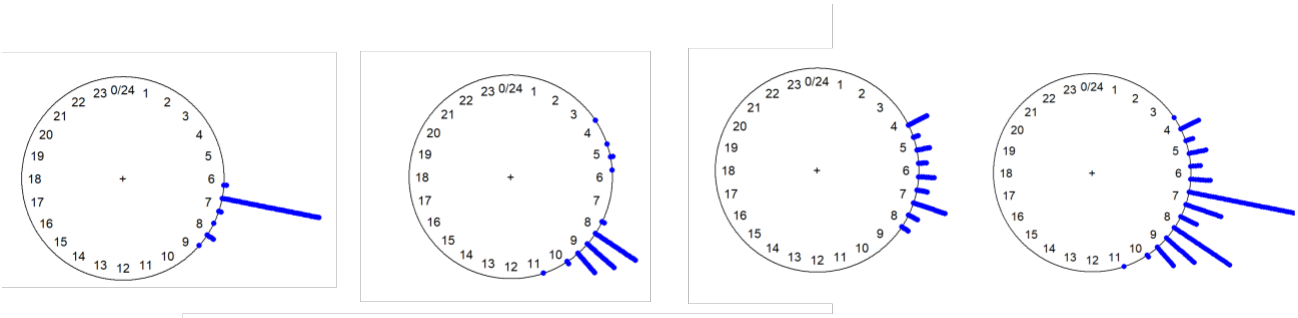

Figure 3 First pickup time of 3 members respectively and merged data

considering 3 covariates in merged data. Although they will impact the estimated β and variance of estimated β, we cannot discard them without support from literature.

3 Data Preprocessing

Before merging the data, we programmatically check whether the entered total screen time and social screen time are correct. When merging data, we added a variable named "id" to each dataset to distinguish data sources. We convert 1st pickup time into data points in hours, i.e, $X_{firstpickuptime} = hour + \frac{min}{60}$, in order to convert the time into numeric form, which could be fitted in the model.

4 Federated Learning

Federated learning[5] enables resource-constrained edge compute devices to learn a shared model for prediction, while keeping the training data local. In this work, we focus on the three self-reported datasets about the screen activity. (Step 1) Firstly we passed the summary statistics, including the number of observations (n_i) and the sample mean of the outcome (Y_i) and the interesting predictor (X_i), from the individual dataset into the central server. (Step 2) Based on the returned population mean of outcome and predictor from the central server, the local server computed the individual sum of square for X_i (SSX_i) and the individual sum of square for X_i and Y_i ($SSXY_i$), and then sent these two summary statistics to the central server which would accordingly calculate the slope and intercept of the model. (Step 3) After receiving the values of slope and intercept, the

local server was able to calculate the fitted value for every point, and the following sum of squared errors (SSE) was obtained. (Step 4) Once the central server integrated the individual SSE_i from the local server, the σ^2 can be estimated by mean squared error (MSE). (Step 5) Consequently, the variance of slope and intercept can be calculated.

$$Step1 : \bar{x} = \frac{\sum_{i=1}^{3} \bar{x}_i \times n_i}{n_1+n_2+n_3}, \bar{y} = frac\sum_{i=1}^{3} \bar{y}_i \times n_i n_1 + n_2 + n_3$$

$$Step2 : SSX_k = \sum_{i=1}^{n_k} (x_{ki} - \bar{x})^2, SSXY_k = \sum_{i=1}^{n_k} [(x_{ki} - \bar{x}) \times (y_{ki} - \bar{y})]$$

$$Step3 : \hat{\beta}_1 = \frac{\sum_{k=1}^{3} SSXY_k}{\sum_{k=1}^{3} SSX_k}, \hat{\beta}_0 = \bar{y} - \bar{x} \times \beta_1$$

$$Step4 : SSE_k = \sum_{i=1}^{n_k} (y_{ki} - \beta_0 - \bar{x}_{ki} \times \beta_1)^2 \; MSE = \hat{\sigma}^2 = \frac{\sum_{k=1}^{3} SSE_k}{n-2}$$

$$Step5 : Var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SSX}, Var(\hat{\beta}_0) = \hat{\sigma}^2 \times (\frac{1}{n} + \frac{\bar{x}^2}{SSX})$$

In practice, the laptops of group members acted as the distributed computing platform that stored the raw data and conducted the local computing, and the captain's laptop also served as the central server that integrated the summary statistics and calculated the parameters. The mutual interaction between the local server and the central server was realized via Github.

Table1

ID	Observations Number (n)	Average of Pickup Times (X)	Average of Social Screen Time (Y)	SSX	SSXY	SSE
1	56	97.89	73.04	77031.72	-21336.13	88041.85
2	50	78.20	93.10	67829.24	50787.70	177746.78
3	57	44.46	140.84	68467.57	-59367.93	560470.46

Table2

	Estimate	Std. Error	t value	P value
(Intercept)	113.1623	12.6597	8.9388	8.782e-16
Pickups	-0.1402	0.1551	-0.9041	1.6327

The sample summary statistics are shown in Table 1. And the summary of the final simple linear model is shown in Table 2. The slope of pickups is negative, which means there is a decrease of 0.1402 minutes in social screen time on average per unit increase in pickups, while the p value is greater than 0.05. So we conclude that within this cohort,

pickups is not significantly associated with social screen time.

5 Meta Learning

Meta learning [4] is an efficient method to build a model from several workers. It aims at learning the way that efficiency of learning systems can grow higher. It is designed to find the optimal learning strategy with an increasing number of tasks. To search association between social app usage time and number of pickups more deeply, we add two more variables 'first pickup time' and 'whether weekend' to be confounders of number of pickups, which change the simple linear regression into a multiple linear regression model. And meta learning is used to estimate coefficients of the model by BLUP(best linear unbiased predictor)[2].The process of calculating BLUP is as below:

(1)Assume $\hat{\beta}_{BLUP} = C_1 \times \hat{\beta}_1 + C_2 \times \hat{\beta}_2 + C_3 \times \hat{\beta}_3$

(2)Since $\hat{\beta}_{BLUP}$ is unbiased, $C_1 + C_2 + C_3 = 1$.

(3)Since $\hat{\beta}_{BLUP}$ has minimum variance of all unbiased estimators, we have to find: $\underset{C_1,C_2,C_3}{\text{argmax}}(Var(\hat{\beta}_{BLUP}))$.

(4)Since Through Lagrange Multiplier optimization algorithm, the $\hat{\beta}_p^{BLUP}$ is: $\frac{\frac{1}{var(\hat{\beta}_p^1)}\hat{\beta}_p^1+\frac{1}{var(\hat{\beta}_p^2)}\hat{\beta}_p^2+\frac{1}{var(\hat{\beta}_p^3)}\hat{\beta}_p^3}{\frac{1}{var(\hat{\beta}_p^1)}+\frac{1}{var(\hat{\beta}_p^2)}+\frac{1}{var(\hat{\beta}_p^3)}}$.

Here $\beta^1, \beta^2, \beta^3$ means 3 members of the group.

(5)The variance of $\hat{\beta}_{BLUP}$ is calculated as $\frac{1}{\frac{1}{var(\hat{\beta}^1)}+\frac{1}{var(\hat{\beta}^2)}+\frac{1}{var(\hat{\beta}^3)}}$.

After calculating, we get the related meta $\beta's$ and coefficients C, which are shown in table 3.

Table 3.

name	first $\hat{\beta}$	first $var(\hat{\beta})$	second $\hat{\beta}$	second $var(\hat{\beta})$	third $\hat{\beta}$	third $var(\hat{\beta})$
Intercept	-19.1822	727.8887	44.3858	10023.49	31.3864	7812.156
Pickups	0.5049	0.0166	0.8294	0.04489	0.3623	0.5368
first pickup time	5.0391	7.4230	-2.6160	232.43	15.8500	120.7953
if weekend	2.9078	62.5621	5.4941	248.3358	-18.1172	838.5326

Next, meta β coefficients are calculated and shown in table 4.

Table 4.

meta$\hat{\beta}_0$	meta$\hat{\beta}_1$	meta$\hat{\beta}_2$	meta$\hat{\beta}_3$
-11.1823	0.5875	5.4231	2.2165

Eventually, the multiple linear regression is built as: $Y = -11.1823 + 0.5875 \times X_1 + 5.4231 \times X_2 + 2.2165 \times X_3$

6 Confirmation Analysis

In the federated learning, we need to pass 5 different kinds summary statistics to establish a simple linear regression: n_i,$\bar{x}_i$,$\bar{y}_i$,SSX_i, $SSXY_i$,and SSE_i. We use the merged data to validate the result of the FL, which is shown in the Appendix Figure 1. As we see, the combined data produces the identical result to the FL. $y = 113.1623 - 0.1402 \times X_1$. This perfect match is supported by the linear decomposition of the total sum of those summary statistics, enabling to lead to an exact numeric solution.

Meta learning estimates overall β coefficients as their best unbiased linear unbiased predictor, which are calculated by those of each individual model and related variances, while least square estimation method is used in classic multiple linear regression models. In our data, The output of classic multiple linear regression is: $y = 156.0651 - 0.1137 \times X_1 - 5.6690 \times X_2 - 13.9038 \times X_3$, with p-value of X_1 is 0.506. Comparing the meta learning result with the linear model, we find the range is not tiny for all parameters.

7 Conclusion and Discussion

The result of the federated learning is the same as that of the simple linear regression with merged data. Outcome of meta learning is completely unlike multiple linear regression model. Such huge difference appears due to the limitation of meta learning. Meta learning should achieve efficiently when learning a large number of tasks so that bias among each task could be eliminated. Yet only 3 tasks from members are learned here and results of them clearly differ from each other.

Regard to linear models, there is not a significant linear association between the social screen time and the frequency of pickups in simple and multiple linear regression.Yet, the conclusion has several limitations. The first limit is the small sample size. Only 159 samples are collected in this project. Next, the assumption of constant variance is violated in this model. The observations come from three classes (person) and the variance of the outcome within each class differs. Last, the model has not included all confounders, which may result in a misleading summary.

8 Acknowledgement

We would like to express our deep gratitude to Prof Song and GSI Yuan Zhong who provided us with much support and suggestions for this project.

References

[1] Na Liu, Kevin Kuan, and Liyao Dong. The role of users impulsiveness in detecting mobile phone excessive dependence: A feature selection analysis. 2019.

[2] George K Robinson. That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32, 1991.

[3] Steven Van Canneyt, Marc Bron, Andy Haines, and Mounia Lalmas. Describing patterns and disruptions in large scale mobile app usage data. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1579–1584, 2017.

[4] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

[5] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207, 2019.

Appendix

Baseline covariates:

X1: Number of team members you have ever worked previously for any other group projects before (0,1,2)

X2: Number of team members you talk to regularly about academic matters (course notes,homework, exam, application for PhD, intern/job application etc).

X3: Number of team members you have ever talked to about topics other than academic matters (0,1,2). Examples of such topics include movie, concert, video game, sport, food, travel etc.

X4: Live with pets at home that you look after (Yes =1, No=0)

X5: Sex (female =0, male =1)

X6: Age (in year)

X7: Course credit hours in the winter semester

X8: Country where previous degree received (US=1, Non-US=0)

X9: Currently have a job (¿10 hours/week) such as RA/TA/Others (Yes =1, No=0)

X10: Number of siblings (e.g. 0, 1,2,...)

X11: Number of social apps installed on your major mobile device that you use regularly for communication and engaging virtual social activities (e.g, 0,1,2,...)

X12: Number of personal mobile devices possessed such as cell phone, iPad, iWatch

X13: The self-reported procrastination score

Appendix Table 1 Baseline variables

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
Han Zhang	0	2	2	0	1	27	16	0	0	1	4	2	24
Xin Luo	0	2	2	0	1	23	13	0	0	0	4	1	38
Ruiqi Ren	0	2	2	0	0	23	13	0	0	0	3	3	38

Appendix Table 2 VIF

	Pickups	First pickup time	If weekend
merged data	1.221658	1.146751	1.093744
member 1	1.193496	1.013761	1.199678
member 2	1.200579	1.117472	1.086351
member 3	1.297609	1.166649	1.125543

Appendix Figure 1 Cook's distance

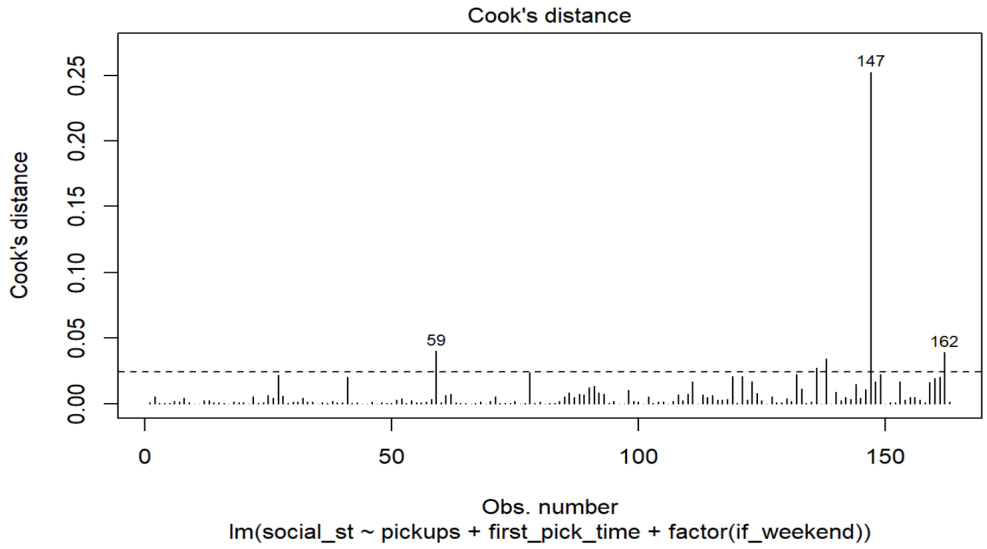

Figure 1: Appx

```
Call:
lm(formula = as.formula(paste(y, "~", x)), data = tt)

Residuals:
    Min      1Q  Median      3Q     Max
-105.81  -43.59  -18.64   28.89  405.48

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.1623    12.6597   8.939 8.78e-16 ***
pickups      -0.1402     0.1551  -0.904    0.367
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.64 on 161 degrees of freedom
Multiple R-squared:  0.005052,  Adjusted R-squared:  -0.001128
F-statistic: 0.8175 on 1 and 161 DF,  p-value: 0.3673
```

Figure 2: Appx

```
Call:
lm(formula = social_st ~ pickups + if_weekend + first_pick_time,
    data = tt)

Residuals:
    Min      1Q  Median      3Q     Max
-111.55  -45.44  -17.42   29.15  405.58

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     156.0651    29.6306   5.267 4.44e-07 ***
pickups          -0.1137     0.1707  -0.666    0.506
if_weekendTRUE  -13.9038    12.7417  -1.091    0.277
first_pick_time  -5.6990     4.3380  -1.314    0.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71.32 on 159 degrees of freedom
Multiple R-squared:  0.02606,   Adjusted R-squared:  0.007687
F-statistic: 1.418 on 3 and 159 DF,  p-value: 0.2395
```

Figure 3: Appx

```
Call:
lm(formula = ad$social_st ~ ad$pickups + ad$first_pick_time +
    ad$if_weekend)

Residuals:
    Min      1Q  Median      3Q     Max
-111.55  -45.44  -17.42   29.15  405.58

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       156.0651    29.6306   5.267 4.44e-07 ***
ad$pickups         -0.1137     0.1707  -0.666    0.506
ad$first_pick_time  -5.6990     4.3380  -1.314    0.191
ad$if_weekend      -13.9038    12.7417  -1.091    0.277
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71.32 on 159 degrees of freedom
Multiple R-squared:  0.02606,   Adjusted R-squared:  0.007687
F-statistic: 1.418 on 3 and 159 DF,  p-value: 0.2395
```