Investigating The Association Between The Time Of Homeworks Deadlines And The Social Screen Time Proportion Among Graduate Students

Project1 for BIOSTAT620 Department of Biostatistics, University of Michigan, Ann Arbor

Ruoer Bei, Zhengrui Huang, Mimi Li (Authors are named in alphabetical order)

Abstract

Digital devices and social media platforms have profoundly transformed lifestyles, particularly among students facing academic pressures. This study explores the relationship between homework deadlines and screen time related to social activities. The study cohort consists of three students enrolled in BIOSTAT620. The data was gathered using the "Screen Time" feature on participants' iPhones, covering the period from January 14th to February 13th, 2024. Homework due dates for this analysis were sourced from three courses—BIOSTAT602, BIOSTAT620, and BIOSTAT651. Federated Statistical Learning was used for data analysis to overcome data sharing challenges, and was compared and evaluated by the Centralized Analysis (Oracle) result. However, the study provides valuable insights into the dynamics of screen usage among students, particularly in relation to their academic stress, and also provides more evidence supporting the Federated Statistical Learning method.

Key phases

Screen time Usage, Homework deadlines, Academic stress, Federated Learning, Centralized Analysis, Linear Regression

Introduction

The prevalence of high screen time in modern life is an increasingly common phenomenon. Dr. Najmeh Khalili-Mahani's literature review (2019) highlights stressful self-considered screen-addictor are also more likely to use screens for entertainment and social networking. This pattern is particularly relevant in the rigorous academic demands of the University of Michigan's Biostatistics students, who have almost weekly homework assignments for each course, often requiring more than four hours to complete. Our study aims to investigate whether the stressful environment is significantly associated with more social screen time spending relative to the total screen time.

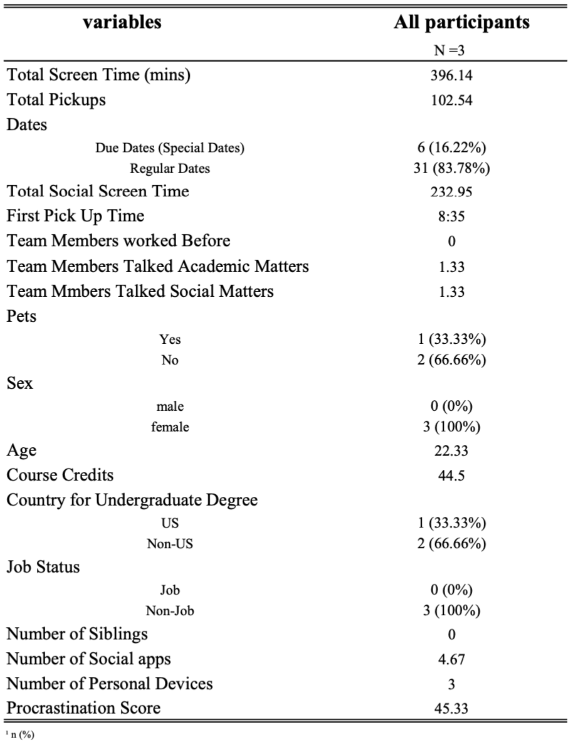
To achieve this, the study employs two distinct analytical perspectives: Federated Statistical Learning and Centralized Analysis. By comparing these methods, we aim to assess the association between homework due dates and screen time and evaluate the effectiveness and reliability of different analytical approaches in this context. The outcome of this study could provide valuable insights into students' digital behavior under academic stress and contribute to the broader discussion on managing screen time in high-pressure educational settings.

Data description

The self-reported data in this study was conducted from three individuals in the BIOSTAT620 class over a one-month period from January 14, 2024, to February 13, 2024. The participants collected screen usage statistics directly from the participants' personal devices, including daily total screen time, social media engagement, and pickup time (Table 1), providing a comprehensive view of each participant's digital interaction. Additionally, the collection of demographic and academic background information were also collected.

Outcome variable

The proportion of social screen time was chosen to be the outcome of interest, which was calculated by dividing the total screen time by the total social screen time. To be specific, the social screen time includes applications, such as instagram, Wechat, Weibo, Xiaohongshu, and etc, which provide both communication and entertainment needs.



variables	All participants N =3
Total Screen Time (mins)	396.14
Total Pickups	102.54
Dates	
Due Dates (Special Dates)	6 (16.22%)
Regular Dates	31 (83.78%)
Total Social Screen Time	232.95
First Pick Up Time	8:35
Team Members worked Before	0
Team Members Talked Academic Matters	1.33
Team Mmbers Talked Social Matters	1.33
Pets	
Yes	1 (33.33%)
No	2 (66.66%)
Sex	
male	0 (0%)
female	3 (100%)
Age	22.33
Course Credits	44.5
Country for Undergraduate Degree	
US	1 (33.33%)
Non-US	2 (66.66%)
Job Status	
Job	0 (0%)
Non-Job	3 (100%)
Number of Siblings	0
Number of Social apps	4.67
Number of Personal Devices	3
Procrastination Score	45.33

[1] n (%)

Table 1: Baseline characteristics

Predictors

The total screen time, total pickups, and a binary variable which indicates whether the day is an assignment due date (falling on February 1st, 6th, 8th, 12th, 15th, and 25th, 2024) in three major courses (Biostat602, Biostat651, Biostat620), are chosen to be the predictor of the model.

According to the statistics, the total screen time is measured as the cumulative amount of time spent on digital screens each day, starting from the first use of the phone in the morning and not extending beyond midnight.

Descriptive

Figure 1 features a time series plot that illustrates the daily proportion of social screen time, with particular focus on highlighting assignment due dates. This allows for a clear visual comparison of social media usage patterns on academically significant dates against those observed on typical days. A heatmap was also displayed for 5 screen usage metrics, offering insights into broader usage patterns over time (figure 2). Additionally, ACF plots for the social screen time proportion shows no meaningful autocorrelation between days, suggesting that daily records for the variable are statistically independent from preceding days' data (figure 3). Finally, figure 4 shows that there is a negative correlation between the total screen time and the total pickup times (-0.386).

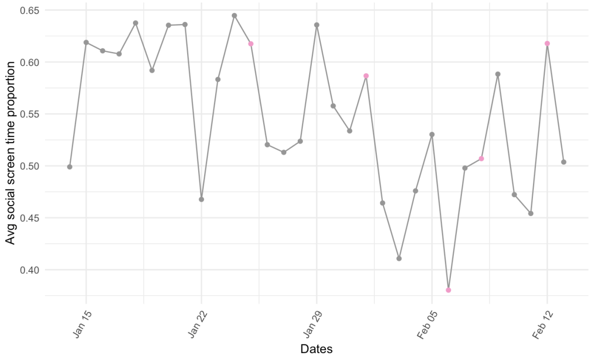


Figure 1. Time series plot of proportion of social screen emphasis on assignment due dates

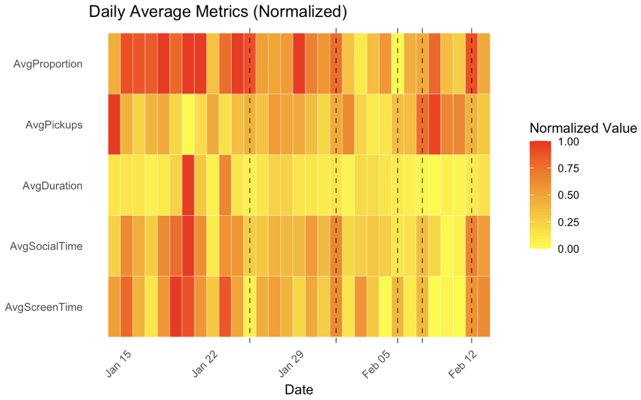


Figure 2. Heatmap of variables with emphasis on assignment due dates

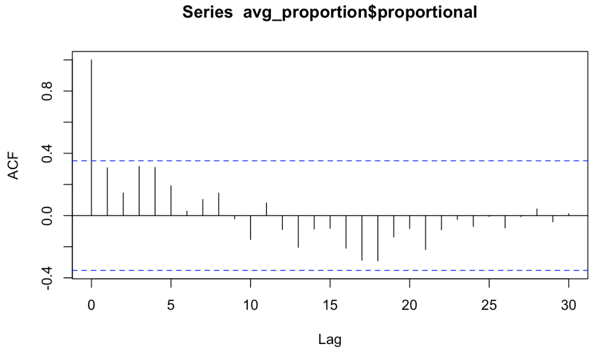


Figure 3. ACF plot for the social screen time proportion

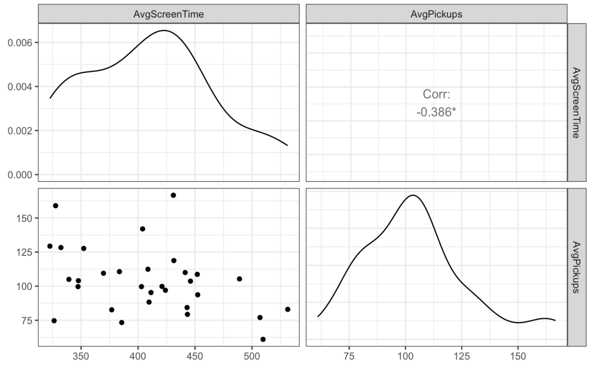


Figure 4. Pairwise plot between average total screen time and average pickups

Data preprocessing

The data collected from three participants were combined into a single dataset, with each participant's data stacked on top of one another. To distinguish between participants, a new column labeled "ID" has been added, with each entry corresponding to one of the participants. The start date for data collection was set based on the most recent date when all participants began recording their data.

Missing data

The missing data were addressed based on whether the entries fell on weekdays or weekends. Specifically, for any missing data on weekdays, the gaps were filled in by the average of each participant's recorded values on weekdays. Similarly, missing weekend data were filled in using the average of each participant's recorded values on weekends. This approach ensures that the imputed values reflect the typical usage patterns of each participant for weekdays and weekends, respectively.

Federated learning

Linear regression

The linear regression model used is

$$\begin{aligned} Y &= X^T\beta + \epsilon \\ &= \beta_1 \times \text{Total.ST.min} + \beta_2 \times \text{Pickups} + \beta_3 \times \text{Due} \end{aligned}$$

where Total.ST.min = total screen time/day, Pickups = total pickups/day, Due = whether the day has assignments due of class Biostat 602, Biostat 651, Biostat 620.

Hypothesis

Null Hypothesis (H0): there is an association between the due dates of homeworks and the proportion of social screen time. $\beta3 = 0$

Alternative Hypothesis (Ha): the due dates of homeworks would not influence the proportion of social screen time. β3 ≠ 0

Procedure of Federated Statistical learning

For data security and privacy protection, Federated Statistical Learning is applied to fit the linear regression model. In this approach, three groups of data from three users would be collected and processed separately. They are then transmitted to a centralized system for further analysis.

Let (Y1, X1) be the data collected from the first user, (Y2, X2) from the second user, and (Y3, X3) from the third user. As shown in the following formulas, we could use the summary statistics of three datasets to obtain the estimate of coefficients by original least squares (OLS).

- Matrix Form of Datasets:

$$X_{3\times 1} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \quad \text{and} \quad Y_{3\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$$

- The Estimate of Coefficients by OLS:

$$\hat{\beta} = (X^T X)^{-1}(X^T Y)$$

$$= \left\{ \begin{bmatrix} X_1^T \\ X_2^T \\ X_3^T \end{bmatrix} \begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix} \right\}^{-1} \left\{ \begin{bmatrix} X_1^T \\ X_2^T \\ X_3^T \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \right\}$$

$$= (X_1^T X_1 + X_2^T X_2 + X_3^T X_3)^{-1}(X_1^T Y_1 + X_2^T Y_2 + X_3^T Y_3)$$

- Standard errors:

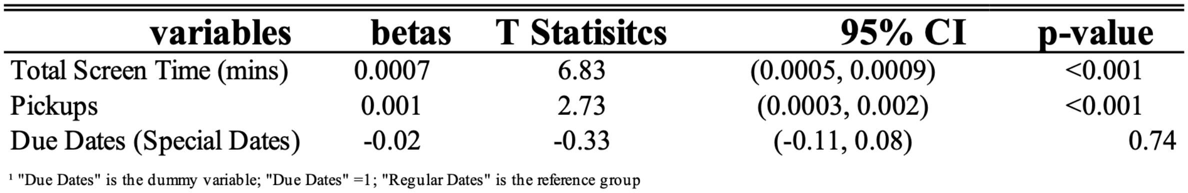
$$\text{se}(\hat{\beta}_j) = \hat{\sigma}\sqrt{((SSX_1 + SSX_2 + SSX_3)^{-1})_{jj}}, \quad j = 1, \ldots, p$$

$$\hat{\sigma}^2 = \frac{(SSY_1 + SSY_2 + SSY_3) - 2\hat{\beta}^T(SSXY_1 + SSXY_2 + SSXY_3) + \hat{\beta}^T(SSX_1 + SSX_2 + SSX_3)\hat{\beta}}{n - p}$$

$$SSX_i = X_i^T X_i; SSXY_i = X_i^T Y_i; SSY_i = Y_i^T Y_i$$

Distributed computing platform

During the R programming, each datasets were transformed to matrix form for the further calculation. The matrix $Xi^{\mathrm{T}}Xi$, $Xi^{\mathrm{T}}Yi$, $Yi^{\mathrm{T}}Yi$ and sample size ni $(i = 1,\ 2,\ 3)$ were calculated, transformed to the dataframe structure and saved as additional documents. Further, in the centralized system, $Xi^{\mathrm{T}}Xi$, $Xi^{\mathrm{T}}Yi$, $Yi^{\mathrm{T}}Yi$ and sample size ni $(i = 1,\ 2,\ 3)$ were extracted from each summary statistics to obtain the total SSX (SSX_total), total SSXY (SSXY_total), total SSY(SSY_total) and total sample size. In addition, the number of intercept and covariates indicated the degrees of freedom p. As a result, the estimate of coefficients, residuals and standard errors were able to be calculated. Finally, the necessary results of the linear regression such as t statistic, p-value, and 95% confidence interval could be attained. The Table2_1 below shows the statistics.



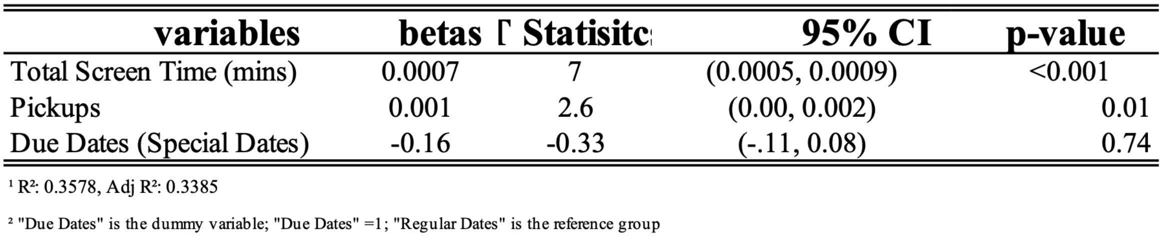
variables	betas	T Statisitcs	95% CI	p-value
Total Screen Time (mins)	0.0007	6.83	(0.0005, 0.0009)	<0.001
Pickups	0.001	2.73	(0.0003, 0.002)	<0.001
Due Dates (Special Dates)	-0.02	-0.33	(-0.11, 0.08)	0.74

[1] "Due Dates" is the dummy variable; "Due Dates" =1; "Regular Dates" is the reference group

Table 2_1: Results of the Linear Regression Model by Federated Statistical Learning

P-value for Due Date is bigger than 0.005, showing a non-significant association to Proportional Screen Time use. P-value for Total Screen Time is smaller than 0.001, showing that a one-unit increase in Total Screen Time is significantly associated with a 0.0007 unit increase in the Proportion of Social Screen Time, after adjusting for other variables. The P-value for Pickups is also smaller than 0.001, showing that a one unit increase in Pickups is significantly associated with a 0.001 unit increase of the Proportional Screen time, holding other variables constant.

Confirmation analysis



variables	betas Γ	Statisitc	95% CI	p-value
Total Screen Time (mins)	0.0007	7	(0.0005, 0.0009)	<0.001
Pickups	0.001	2.6	(0.00, 0.002)	0.01
Due Dates (Special Dates)	-0.16	-0.33	(-.11, 0.08)	0.74

[1] R²: 0.3578, Adj R²: 0.3385

[2] "Due Dates" is the dummy variable; "Due Dates" =1; "Regular Dates" is the reference group

Table 2_2: Results of the Linear Regression Model by Centralized Analysis

For the comparison between Federated Statistical Learning and Centralized Analysis, as the Table 2_2 shows, it is roughly the same as Table 2_1, showing the effectiveness of Federated Statistical Learning approach, and suggesting a way of modeling under data privacy and sharing barriers.

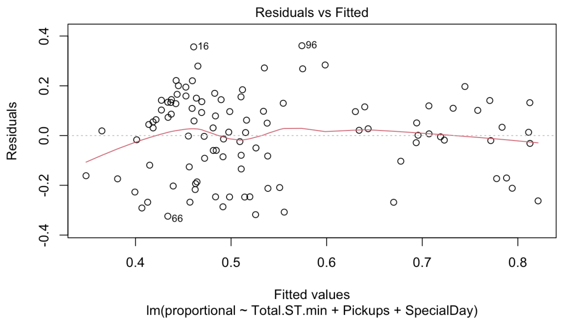


Figure 5. Residual VS Fitted plot: testing homoscedasticity

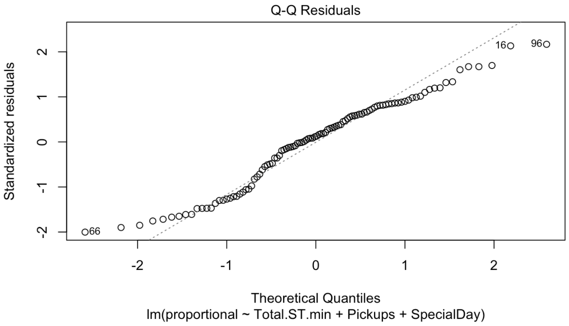


Figure 6. Q-Q plot: testing linearity

The residuals are randomly dispersed around the horizontal axis without any clear pattern of increasing or decreasing variance, which is an implication of homoscedasticity. For the Q-Q plot, although some points deviate from the line and there are some outliers, the majority of the points follow the line quite closely, showing the assumption of linearity holds.

The DW statistic (1.093) is significantly less than 2, suggesting positive autocorrelation among the residuals. At the same time, the Shapiro-Wilk test has a value of 0.96767 with a p-value of 0.01205, suggesting to reject the null hypothesis, indicating that the residuals are not normally distributed. Therefore, the model assumptions of normality and independence may be violated.

Conclusion and discussion

The analysis suggests that the relationship between Proportional Screen Time and the presence of assignment due dates is not statistically significant (p-value = 0.74), meaning that assignment deadlines do not significantly alter the amount of time participants spend on social media applications relative to other types of apps. Consequently, due dates appear to have no substantial impact on overall screen usage patterns among the participants.

The result also indicates a positive association between Total Screen Time and the proportion of time spent on social media (p-value < 0.001). Specifically, a one-unit increase in Total Screen Time corresponds to an increase of 0.0007 units in the Proportion of Social Screen Time, after adjusting for other variables. Therefore, as overall screen usage rises, there is a slight increase in the share of time dedicated to social media relative to total screen engagement.

For numbers of Pickups, one unit increase in Pickups is associated with a 0.001 unit increase of the Proportional Screen time (p-value < 0.001), holding other variables constant, showing that participants pick up their phones more frequently when using more social apps compared with using other apps.

Limitations

Given that the data violate the assumptions of normality and independence, the current model may not be ideally suited to accurately represent the underlying structure of the data. At the same time, the presence of correlation between predictors, specifically between Total Screen Time and Total Pickups, introduces multicollinearity, which could affect the model's accuracy.

This issue needs further analysis to better understand its impact and to improve the modeling approach accordingly.

Furthermore, concerns regarding the nature of the collected data warrant attention. Given the study's objective to understand students' screen usage behaviors in relation to academic stress, as implied by homework deadlines, incorporating questionnaires about their homework completion habits could be beneficial. This approach recognizes that not all students engage in homework activities strictly on the day of the deadline. By gathering self-reported data on when and how students allocate their time to homework, the study can provide a more comprehensive view of the relationship between screen time and academic responsibilities. Additionally, the limited number of features within the dataset might cause the predictors insufficient for accurately describing the model, which could explain the low R-squared value observed. More information should be collected to enhance the comprehensiveness of the analysis.

Acknowledge

We thank Dr. Peter Song (University of Michigan) for his generous help and excellent teaching skills.

Reference

Najmeh, Khalili-Mahani. “To Each Stress Its Own Screen: A Cross-Sectional Survey of the Patterns of Stress and Various Screen Uses in Relation to Self-Admitted Screen Addiction.” *National Library of Medicine*, April 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6465981/.

Github: https://github.com/huangzr1228/biostat620_project1