566 features

He Ren, Kirara Kura, Matthieu Liger, Fumin Li

May 2024

1 observation = 1 person x given year-month ('Jon in March 2010') or 1 person x survey. Each feature associates a scalar to an observation. Call S the social features, D demographic, P physical, Z sleep, C for covid, W for well-being.

Ideally these feature definitions give an equation or procedure to go from the raw dataset, to a column in our pruned 20-dimensional dataset, for each observation, which as we discussed could be either (person × YYMM) or (person × YYMMDD EMA general survey)

e.g. if we have (Jon × 2202) or (Jon × EMA general survey 220217), we can say what 12.7 in that cell (((Jon × 2202)),X_i), how it is computed from the raw dataset.

1 Features

1.1 COVID

-Covid pandemic stage (could be binary: before March 2020/after March 2020, could be categorical, e.g. before onset of covid=0, between onset and vaccine=1, after vaccine but before end of lockdown=2 etc).

1.2 Well-being

- pam Definition: PAM measures affect from two dimensions: valence (unpleasurable to pleasurable or negative to positive) and arousal (low activation to high activation). It is highly positively correlated with Positive And Negative Affect Schedule (PANAS) Positive Affect (PA). The score is coded as

High Arousal

	6	8	14	16	
Negative Valence	5	7	13	15	Positive Valence
	2	4	10	12	
	1	3	9	11	

Low Arousal

- phq4_score
  - Definition: Total score on Patient Health Questionnaire-4, a four-item questionnaire, screening for depression and anxiety.
  - Calculation: The score range from 0 - 12 (a higher score refers to worse mental health).
- phq2_score
  - Definition: Total score of the first two items from phq4. They represent the depression level.
- gad2_score
  - Definition: Total score of the last two items from phq4. They represent the anxiety level.
- social_level
  - Definition: One question asking "Have you spent most of your time alone or with others today?".
  - Calculation: The score ranges from 1-5, where 1 represents almost always alone and 5 represents almost always with others.
- sse_score
  - Definition: Total score on four questions about the sense of self-esteem.

    - Calculation: The score ranges from 1 to 20, and a higher score refers to good self-esteem. (Note: I think the first question should be reversed-coded, and that is what I did)

- stress
    - Definition: One question asking ”Are you feeling stressed now?”
    - Calculation: The score ranges from 1-5, and a higher score refers to higher stress.

Overall, I created two files, one is File name - ”all_student_ema_data_cleaned.csv”; the other one is File namae - ”student_uid_class.csv”. Also, I created a files to identify student and the class (i.e., class = 2017 refers to student entering the college in 2017). It is consistent with the paper. 104 and 113 undergraduate students in each year.

1.3 Demographics

- gender
- race
- OS (iOS or Android)
- cohort year

1.4 Physical

(Kirara) Aim of this category: The aim of this category is to see how moving around physically and going to somewhere can affect stress.

- $P1$= average daily exercises
    - Definition: average amount of hours spent on walking&running(on foot) in a day over XXX days before observation
    - Aim: know how much moving around can affect mental health
    - Calculation:
        - **act_on_foot_ep_0**/count(prior XXX days) (Android)
        - (**act_running_ep_0**+**act_walking_ep_0**)/count(prior XXX days) (iOS)
    - Filling missing data:
- $P2$= studying time
    - Definition: On average, how much minutes did the observation study in a day before XXX days
    - Aim: To understand cofounding effect of studying had on mental health
    - Calculation: $(loc\_study\_still * loc\_study\_dur)/count(priorXXXdays)$ (for both iOS and Android
    - Filling missing data
- $P3$ = inside the house time
    - Definition: average amount of time spent at home in a day before XXX days
    - Aim: See relationship between staying home and mental health
    - Calculation: **loc_home_dur/count(prior XXX days)** (for both iOS and Android
    - Filling missing data:
- $P4$ = outside the house time
- $P5$ doing sport
    - Definition: The average time spent at sport/excersise location in a day before XXX days
    - Aim: To understand cofounding effect of playing sport and exercise effect on mental health
    - Calculation: **loc_workout_dur**/count(prior XXX days) (for both iOS and Android
    - Filling missing data

1.5 Social

(Matthieu)

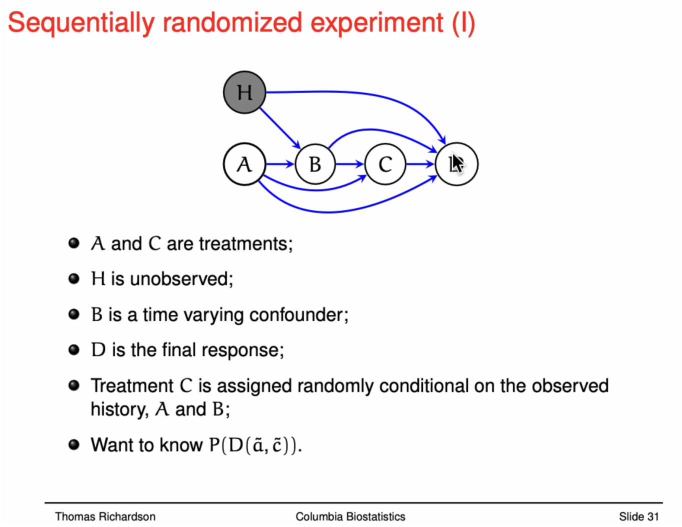
This category is intended to provide metrics of how socially active the user was during a given day. The word 'Social' is taken in a sense that includes how sedentary they are (how much do they travel, visit a variety places, move around when at a social venue).

The original dataset includes information about communications from the user: live conversations, phone conversations, SMS, but only for Android users. To avoid biasing the data, the features defined below are available for both iOS and Android users, hence do not include these three categories. Phone usage information was extracted from unlock duration and frequency. Cursory exploration of data suggests that phone usage while in social venues is especially relevant (either in a positive or a negative manner as it related to mental health). Note food venues are treated as social, as it is typically the case for college students. Large swings over the difference stages of the COVID pandemic are expected for these covariates.

Among all the features available in the raw dataset, features that had correlation (either negative or positive) with stress were preferred.

The metrics are defined in a time window before the well-being survey is taken by the user.

- S_1= Time spent travelling
  - Definition: Average daily seconds in a vehicle.
  - Aim: Metric of how 'out and about' people are and how that might be associated with mental health.
  - calculation: `act_in_vehicle_ep_0` (average over TK days prior to survey)
- S_2= Distance travelled.
  - Definition: Daily distance traveled in meters.
  - Aim: Metric of sedentarity.
  - Calculation: `loc_dist_ep_0` (average over TK days prior to survey)
- S_3= Time spent at social location.
  - Definition: Hours spent either at social or food locations.
  - Aim: Proxy for of number and frequency of social interactions.
  - Calculation: `loc_social_dur + loc_food_dur` (average over TK days prior to survey)
- S_4= Change of scenery
  - Definition: Daily number of locations visited.
  - Aim: Proxy to how many things and people a user is likely to interact with.
  - Calculation: `loc_visit_num_ep_0` (average over TK days prior to survey)
- S_5= Phone unlocked duration while at social and food venues
  - Definition: Average number of minutes the user's phone is unlocked per hour, when the user is either at a social location or at a food location.
  - Aim: Metric of how engaged or disengaged users are when in these type of locations and how that might be associated with mental health (one could imagine reasons for either).
  - Calculation: `loc_social_unlock_duration + loc_food_unlock_duration` (average over TK days prior to survey)
- S_6= Frequency of phone unlock while at social and food venues
  - Definition: Average number of times the user unlocks the phone within each hour spent at a social or food location.
  - Aim: Metric of how engaged or disengaged users are when in these location and how that might be associated with mental health (one could imagine reasons for either).
  - Calculation: `loc_social_unlock_num + loc_food_unlock_num` (average over TK days prior to survey)
- S_7= Activity while at social and food venues
  - Definition: How much the user sits still when at social places.
  - Aim: Association between not being 'social' when in a social venue and well-being.
  - Calculation: `loc_social_still + loc_food_still` (average over TK days prior to the survey)



Sequentially randomized experiment (I)

- A and C are treatments;
- H is unobserved;
- B is a time varying confounder;
- D is the final response;
- Treatment C is assigned randomly conditional on the observed history, A and B;
- Want to know $\mathrm{P}(\mathrm{D}(\tilde{a}, \tilde{c}))$.

Thomas Richardson Columbia Biostatistics Slide 31

Figure 1: Enter Caption

1.6 Sleep

We only have 4 columns in the sensing data talking about sleep, 1. sleep duration 2. sleep start 3. sleep end 4.sleep heathkit duration(ios). I guess we only need to consider 1, 2, 3 since they work for all participants. The third one can be calculated from 1 and 2, so we only use 1 and 2. Maybe need to delete the units that don't have enough records.

In this case, we have:

- Z_1 = sleeping time. `sleep_duration`
  - Definition: Average sleeping time per day(within the period that under our consideration(10 days before survey or average of a month)).
  - calculation`sleep_duration` average over 10 days prior to survey.
- Z_2 = start sleeping time.
  - Definition: Average time that students start to sleep time per day(within the period that under our consideration(10 days before survey or average of a month)).
  - calculation `sleep_start` average in 10 days prior to survey.
- Z_3 = end sleeping time `sleep_end` (redundant with sleep start and sleep duration but might be good to separate)
  - Definition: Average time that students start to sleep time per day(within the period that under our consideration(10 days before survey or average of a month)).
  - calculation `sleep_end` average in 10 days prior to survey.

1.7 Possible analyses

-for each level of the treatment, i.e. covid phase k, fit a model $W_i = f_k(\{X_j | j \in A\}\}$ where A are the covariates we adjust for. -use the collection of models to infer the counterfactuals $W(covidphase)$ -compute ATEs of for W_i using the inferred counterfactuals.

- Q1: Does COVID pandemic affect wellbeing?
  - treatment: COVID (after date 03/2020 1, before 03/2020 0
  - Output(result): wellbeing score (Either PHQ4 or Stress score)
  - Covariates (Z): Use covariates collected before the pandemic ONLY. This is because using covariates after the pandemic will cause the analysis to have a mixed effect of covariates being the effect of the pandemic and the cause of wellbeing and opens the path like figure 1.7
  - METHOD: Compare the time before and after the pandemic
    - ∗ DID
    - ∗ Synthetic Control
  - ∗ Do Y X + $\mathrm{Z}_1 + \mathrm{Z}_2$
- Q1?Q2?: How does Covid pandemic affect wellbeing through covariate(like sleep, physical activity,etc)

- treatment: COVID (after date 03/2020 1, before 03/2020 0
- Output(result): wellbeing score (Either PHQ4 or Stress score)
- Mediator: Sleep, social,physical,etc
- Covariate: Im not sure if we put this in...
- METHOD: Mediation Analysis