Estimating Causal Effects of Exercise from Mood Logging Data

Dhanya Sridhar [1] **Aaron Springer** [1] **Victoria Hollis** [1] **Steve Whittaker** [1] **Lise Getoor** [1]

Abstract

Mood and activity logging applications empower users to monitor their daily well-being and make informed health choices. To provide users with useful feedback that can improve quality of life, a critical task is understanding the causal effects of daily activities on mood and other wellness markers. In this work, we analyze observational data from EmotiCal, a recently developed mood-logging web application, to explore the effects of exercise on mood. We investigate several methodological choices for estimating the conditional average treatment effect, and highlight a novel use of textual data to improve the significance of our results.

1. Introduction

Mood and activity logging applications play an important role in the larger, emerging trend of technologies that empower users to monitor and improve their quality of life (Konrad et al., 2016; Parks et al., 2012; LiKamWa et al., 2013). Notable platforms include Fitbit or Strava for exercise tracking, and Moodscape or Echo for reflection on emotion and mood (Konrad et al., 2016; LiKamWa et al., 2013). In mood logging applications, users track their activities together with markers of their mental state to promote psychological well-being. To facilitate positive outcomes, these applications must provide actionable feedback on how factors in users' daily life affect their mood. An important step to generating feedback is understanding the causal effects of these factors on mood, estimated by the *average treatment effect* (ATE). Estimating ATE requires several modeling assumptions and careful choices, especially on complex user-behavior data.

Early approaches to understanding factors that affect mood relied on traditional methods such as surveys and randomized intervention trials (Stone et al., 2006; Seligman et al., 2005). The findings suggest a link of exercise and socializing on mood. In recent years, mood logging applications such as Echo, Moodscape and iHappy have driven empirical research in user behavior (Isaacs et al., 2013; Konrad et al., 2016; LiKamWa et al., 2013; Parks et al., 2012). Some platforms perform direct interventions such as recommending activities to improve users' mood (Parks et al., 2012) while predictive applications like Moodscape infer mood changes from activity patterns.

Recently, a different line of work focuses on Twitter social media posts to find causal links on outcomes that span mood or emotion to significant life milestones (Olteanu et al., 2017; Dos Reis & Culotta, 2015). These approaches extract treatments such as exercising behavior and potential outcomes such as mood from tweets, and perform matching on text to eliminate confounds. Despite using non-conventional forms of observational data, both studies report findings validated in literature, such as the exercise and mood link.

Our work focuses on causal estimation using a unique mood logging application, EmotiCal (*Emoti*onal *Cal*endar), that features both recorded values for daily activities and mood, and text descriptions from users (Hollis et al., 2017; Springer et al., 2018). Motivated by promising results from the complementary studies of social media sites and task-specific mood logging platforms, we study the causal effect of exercise on mood by combining text and observational data. The link between exercise and mood is well-validated in literature, providing a benchmark for our analysis. To develop a robust methodology for estimating ATE from heterogeneous user data, we outline and investigate three important modeling questions in this paper:

1. What filtering or stratifying strategies on users are necessary to eliminate implicit confounds or outliers?
2. Should we perform our analysis per-user or treat all logged user entries as independent units of study?
3. What impact does including text sources in our analysis have on the estimated ATE?

Our findings highlight the importance of each modeling choice, and suggest that per-user analysis and incorporating text provide stronger causal results. We illustrate our empirical results with useful qualitative examples.

[1] University of California Santa Cruz. Correspondence to: Dhanya Sridhar <dsridhar@soe.ucsc.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

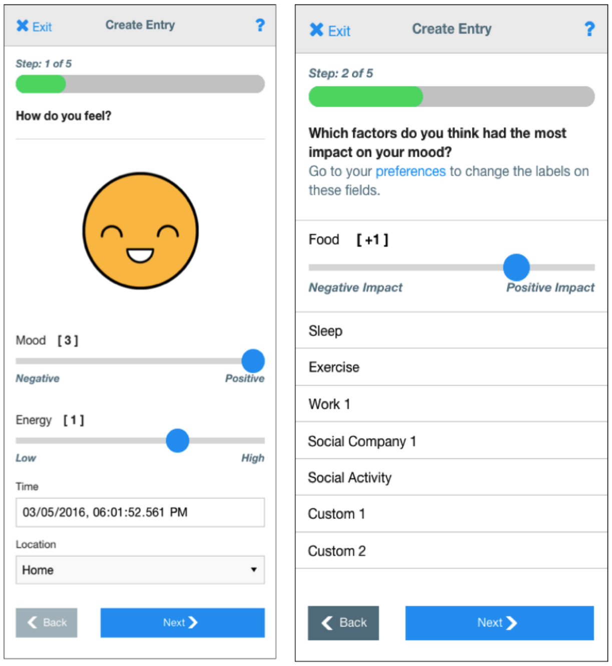
2. Data Description

We obtain our dataset from EmotiCal (Emotional Calendar), an application created to help people regulate and improve their mood and well-being (Hollis et al., 2017; Springer et al., 2018) [1]. EmotiCal users were asked to user the application at least twice a day, logging an entry each time. These entries consist of users' current mood, energy level, and up to 14 trigger activities that users believe have influenced their mood. For example, users can log social interactions (e.g., time spent with a friend or coworker), aspects of physical health (e.g., sleep or exercise), and work activities (e.g. meetings) to track these activities effects on mood. EmotiCal also prompts users to generate short textual explanations of how and why they think those activities have affected their mood.

Figure 1 shows the EmotiCal user interface for logging mood and energy levels (left panel), and activities that affect these factors (right panel). To create a mood entry, users first make a simple mood valence and strength decision, choosing a mood ranging from -3 (very negative) to +3 (very positive). Users also recorded energy levels ranging from -3 (low energy) to +3 (high energy). After selecting mood, users engage in active mood analysis. Users identify which of 14 possible trigger activities influenced their mood and rate that influence on a scale of -2 (negatively impacted mood) to +2 (positively impacted mood).

Users choose as many activities as they deem relevant, although most users choose relatively few per entry. Eight of these 14 trigger activities are constant across users: food, sleep, exercise, social activity, work, leisure, mood, and social company; the other 6 categories are customizable, allowing users to record triggers that are unique to their lives. After logging their trigger activities, users write short textual entries about the factors that affected their mood.

In this work, we focus on the eight non-customizable trigger activities that users logged in order to provide a consistent analysis across different users. In addition, we use the textual entries users wrote to improve the significance of our results. In total, the EmotiCal dataset consists of 6344 entries from 143 unique users. The EmotiCal data enjoys two important advantages over typical user-behavior modeling datasets: 1) participants provide real-time and longitudinal self-labels for all attributes without need for crowd-sourced annotations; and 2) users log their own perception of how activities influence mood and also include textual information, producing a reliable dataset.



Step: 1 of 5

How do you feel?

Mood [3]

Energy [1]

Time

03/05/2016, 06:01:52.561 PM

Location

Home

Step: 2 of 5

Which factors do you think had the most impact on your mood?

Go to your preferences to change the labels on these fields.

Food [+1]

Sleep

Exercise

Work 1

Social Company 1

Social Activity

Custom 1

Custom 2

Figure 1. EmotiCal System Components. The left screen shows the logging of mood and energy levels. The right screen shows the logging of different activities which affected the users mood

3. Problem Statement

We consider units $X = \{x_1 \dots x_n\}$ where each logged entry $x_i = (v_1 \dots v_8, t)$ includes measurements for the eight specified variables v_i and text denoted t. The treatment assignment for x_i is 1 if $v_{\text{exercise}} \geq 1$ and 0 otherwise, and is denoted by random variable T_i. For each x_i, we observe only the mood outcome under treatment $Y_1(x_i)$ when $T_i = 1$ or the outcome under control $Y_0(x_i)$ when $T_i = 0$. The goal is to estimate the average treatment effect (ATE), defined as:

$$\mathbb{E}_{p(x)}\left[\mathbb{E}_{p(Y_1|x)}[Y_1(x_i)] - \mathbb{E}_{p(Y_0|x)}[Y_0(x_i)]\right] \tag{1}$$

The expectation requires unobservable outcomes $Y_{1-t_i}(x_i)$, called counterfactuals. To overcome this hidden data problem, we perform nearest-neighbor matching which pairs each x_i with x_j such that $T_j = 1 - T_i$ and $x_j = \arg\min_{x_k \in X} d(x_i, x_k)$ where d is a measure of distance. Intuitively, each unit is matched to be compared against a unit which is most similar in covariate space, but has opposite treatment assignment. The matching produces a set of pairs $M = \{(x_i, x_j) | x_i, x_j \in X, T_i \neq T_j\}$ to estimate ATE as:

$$\frac{1}{|M|} \sum_{m \in M} (2T_i - 1)[Y_{T_i}(x_i) - Y_{T_j}(x_j)] \tag{2}$$

The summand is shorthand to indicate correct subtraction order, depending on whether $T_i = 1$ or not. In our analysis,

[1] The data collection process has IRB approval. All participants were directly informed of the research uses of their anonymized data and allowed to exclude private items from the final dataset. Strict procedures are in place to secure and protect users' privacy.

we investigate key experimental choices for estimating ATE with user-behavioral data.

4. Experimental Analysis

The goal of our analysis is to investigate three modeling questions to develop a methodology for estimating treatment effects from mood logging data:

- Q1: What filtering steps on the users help to control for implicit confounds and eliminate outliers in causal estimation from behavioral data?
- Q2: Should we aggregate the ATE estimate over treatment and control entries matched per-user, treating users as the key units of study, or should we treat entries as the unit of study?
- Q3: What signal does textual data contain to help us control for additional implicit confounding?

To validate key modeling decisions, we focus on estimating the ATE of exercise on mood, a link that has been well-studied in literature and found to have a significant positive effect. Studying the exercise-mood link allows us to interpret differences in ATE across the experimental conditions we use as better or worse performance. Our findings below suggest the importance of careful filtering and matching, and highlight the benefits of incorporating textual data sources.

4.1. Experimental Setup

For all experimental questions, we perform matching on treated units to obtain their nearest control unit. For Q1 and Q2, the metric used for matching is Euclidean distance over the eight other measured variables. We introduce a text-based propensity score matching (Rubin, 2006; Rosenbaum & Rubin, 1983) technique when we examine Q3. We also perform a Z-test to compare the mean mood across treatment and control samples to understand the significance of the effect. We introduce experimental conditions to evaluate each question that we investigate. Table 4.1 shows the ATE and hypothesis test p-value results for all conditions, and we provide a detailed discussion of these results below.

4.2. Q1: Filtering Users

Q1 investigates whether filtering users is necessary to mitigate noise in our analysis. Our baseline condition, which we call Baseline, includes all users who report the effects of exercise at least once as the sample population. The Baseline condition retains 114 users out of 143 total users. To compare against Baseline, we consider whether to target our study to 73 users who report having exercised at least 3 times. We refer to this experimental condition as Filtered.

Table 1. ATE and hypothesis testing results for experimental conditions across evaluation questions Q1 to Q3. The results suggest benefits to including textual data in matching methods.

Condition	ATE	Hypothesis Test P-value
Baseline	0.26	$3 \times 10^{-}5$
Filtered	0.31	$1.5 \times 10^{-}6$
User	0.49	$2 \times 10^{-}12$
Text, C=0.9	0.53	$6.7 \times 10^{-}13$
Text, C=0.01	0.61	0.0

Table 2. p-values from T-tests evaluating balance of other measured activities across control and treated groups. We compare the balance between three matching strategies. Text, C=0.9 improves balance over the User matching for three covariates.

Covariate	User	Text, C=0.01	Text, C=0.9
Food	0.058	$1.3 \times 10^{-}4$	0.023
Sleep	0.012	$3.8 \times 10^{-}5$	0.549
Work	0.163	0.017	0.586
Leisure	0.025	$6.1 \times 10^{-}5$	0.005
Social Company	0.072	0.014	0.068
Social Activity	0.437	0.005	0.649

Users that report exercise as a factor that effected mood only once may not consistently value this activity, introducing noise to the causal estimation.

Table 4.1 shows that the ATE estimated through the Baseline condition is significant. This result validates a previous regression analysis on EmotiCal data that showed significant correlations between exercise and mood (Springer et al., 2018). However, the Filtered condition increases both the ATE and its significance, which we expect to see for the well-validated causal link between exercise and mood. The comparison suggests that retaining only users that demonstrate consistent exercise patterns yields more robust ATE estimates.

4.3. Q2: User-specific Matching

Following findings from Q1, we adopt the Filtered condition throughout the analysis. Our next critical question is whether to perform matching for entries of the same user and aggregate across users for estimating ATE. To evaluate this, we introduce condition User which adds a constraint to the matching algorithm that matched entries must be from the same user. Formally, U_i is the user of entry x_i and the modified matched pairs are:

$$M_U = \{(x_i, x_j) | x_i, x_j \in X, T_i \neq T_j, U_i = U_j\}$$

The compute ATE given by equation 2, we now sum over $m \in M_u$. The User condition tests the impact of controlling for variations across users that potentially introduce noise and reduce significance of the ATE.

Table 3. Examples of matched treatment and control pairs that highlight differences between conditions TEXT and USER. TEXT results in more contextually similar pairs.

Treated Entry	**TEXT-Matched Control**	**USER-Matched Control**
"Really enjoyed a bike ride today to Pioneer park, an old timey park area. It was a fun new experience to explore it, it reminded me a bit of main street in Disney world. Before the bike ride I wrote in my diary too, nice. I feel good, but noah seems more distant today so my mood is more subdued and reflective."	"Visited the visitors center in Fairbanks for a few hours which I had never been to before. I chatted with some nice folks and it was fun. Also had dinner with a friend's family which I enjoyed."	"Slept well. Feeling relaxed."
"Got up early to do yoga class outside in morning ramped up my mood and energy"	"Content that I'm learning different things in excel but today's class required lots of focus"	"Going for walk to harvest garden event and shopping for gardening supplies impacted mood positively although energy could be higher but not due to oversleeping and running late and unhealthy breakfast"
"Went on a run which was good stress relief. Spending the day outside and getting sun also upped my mood. Ate something and got sick from it."	"Got to learn something new at work today which made me happy, but then the new tech that replaced me came in and I started feeling jealous/sad that I never had the interaction she has with my old boss. "	"Drinking makes me happy"

Table 4.1 shows that compared to the previous condition FILTERED, USER increases ATE to 0.49 and makes it more significant with a p-value of $2 \times 10^{-}12$. This finding substantiates approaches that estimate causal effects at the user-level, aggregating over tweets or entries (Olteanu et al., 2017; Dos Reis & Culotta, 2015). The goal of mood logging platforms is to personalize feedback or recommendations for each user, and the USER condition better captures this end goal.

4.4. Q3: Incorporating Text Data

For our final question about the additional benefits of incorporating textual data, we apply the USER condition and extend it with TEXT, which upgrades the matching strategy to support variables from text. In TEXT, we replace the distance metric $d(x_i, x_j)$ used in matching with the propensity score which includes text variables. Formally, given treatment assignments T_i and user entries x_i, the text-based propensity score is:

$$P(T_i|x_i, C_i)$$

where $C_i = \{c_1 \dots c_{|V|}\}$ is the set of count variables c_i for each unigram in the vocabulary across the textual entries t of user U_i. We model the propensity score with logistic regression. To reduce overfitting and select a sparser model, we include a L_1 penalty term with a cost parameter C that controls the degree of sparsity. Small values of C induce sparser models and might exclude the covariates that represent the other measured activities such as quality of sleep or food among others.

To evaluate this trade-off between sparsity and a propensity score that encourages better balance across measured covariates, we consider two variants of TEXT, with $C = 0.01$ and $C = 0.9$. We evaluate the balance across treated and control groups for each of the other measured activities by performing t-tests to assess the difference in means. Table 4.1 shows the p-values from these t-tests; larger values indicate better balancing of covariates between control and treated groups, which is an important criteria for causal inference.

Table 4.1 shows that condition TEXT, $C = 0.01$ gives the highest ATE of 0.61 and greatest significance, as the p-value is effectively 0.0. However, the p-values for TEXT, $C = 0.01$ in Table 4.1 suggests that the stricter L_1 penalty removes several other measured activity variables from the model, resulting in imbalance for these covariates. In contrast, by setting $C = 0.9$, the balance across these covariates remains comparable with that of the USER matching. Interestingly, for covariates sleep, work and social activity, the balance even improves when we include text-based attributes in the propensity score model. Additionally, the trade-off against estimating ATE remains desirable as TEXT, $C = 0.9$ still gains over USER in both significance of the ATE and its value.

The results from both estimation and balance suggest that users' language might encode other factors and variables that the study could not measure. However, the balance analysis points to the importance of carefully evaluating modeling choices and parameter settings to not violate key causal assumptions. To further understand the text-based approach, an in-depth exploration of the learned propensity score model is critical to indicate which signals were useful. Below, we follow this analysis up with several qualitative results that shed light on the usefulness of text in causal effect estimation.

4.5. Qualitative Results

We further study the TEXT (C=0.9 is used for the remainder of the analysis) propensity score model by examining its features and outputs. We first identify the unigrams used in TEXT which received the overall highest coefficients in logistic regression and find words such as: energized, drained, better, rest, great. These words indicate users' reflections on their internal states, which are inherently latent, but can be encoded or expressed through linguistic choices in text.

Next, we provide illustrative examples of treated entries that are matched with different control entries by TEXT and USER. Table 4.1 contrasts the differences in control entries chosen by each matching strategy. The examples suggest that condition TEXT, which models a text-based propensity score, yields matched pairs that are more lexically similar than those produced by the distance matching in User.

In the first treated entry, the user describes exploring a new area and interestingly, TEXT produces a matching control entry that also discusses travel and exploration. On the other hand, the control entry matched using USER is brief and less related, only discussing sleep. The second treated entry example conveys a relaxed, positive tone which is mirrored in the matched control entry chosen by TEXT. In contrast, USER produces a match that initially exudes a negative tone. The final example suggests a common tone of positivity combined with annoyance in the matched pairs by TEXT while the USER-selected control entry is semantically unrelated.

5. Discussion and Future Work

In this work, we introduce a text-based propensity score matching method to estimate the average treatment effect between exercise and mood using data from the recently developed EmotiCal mood-logging application. We develop our approach TEXT which incorporates text variables by carefully examining several modeling choices. Our findings highlight the importance of user-specific, stratified analysis when modeling user-behavior domains. Our preliminary results suggest several research directions for future work. We will explore more sophisticated regression methods that model dependencies between possible causal factors and latent variables. We will also investigate richer models of text.

6. Acknowledgments

This work is supported by NSF grants CCF-1740850 and NSF IIS-1703331.

References

Dos Reis, Virgile Landeiro and Culotta, Aron. Using matched samples to estimate the effects of exercise on mental health from twitter. In *AAAI Conference on Artificial Intelligence*, 2015.

Hollis, Victoria, Konrad, Artie, Springer, Aaron, Antoun, Matthew, Antoun, Christopher, Martin, Rob, and Whittaker, Steve. What does all this data mean for my future mood? actionable analytics and targeted reflection for emotional well-being. *Human–Computer Interaction*, 32 (5-6):208–267, 2017.

Isaacs, Ellen, Konrad, Artie, Walendowski, Alan, Lennig, Thomas, Hollis, Victoria, and Whittaker, Steve. Echoes from the past: how technology mediated reflection improves well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1071–1080. ACM, 2013.

Konrad, Artie, Tucker, Simon, Crane, John, and Whittaker, Steve. Technology and reflection: Mood and memory mechanisms for well-being. *Psychology of well-being*, 6 (1):5, 2016.

LiKamWa, Robert, Liu, Yunxin, Lane, Nicholas D, and Zhong, Lin. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pp. 389–402. ACM, 2013.

Olteanu, Alexandra, Varol, Onur, and Kiciman, Emre. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 370–386. ACM, 2017.

Parks, Acacia C, Della Porta, Matthew D, Pierce, Russell S, Zilca, Ran, and Lyubomirsky, Sonja. Pursuing happiness in everyday life: The characteristics and behaviors of online happiness seekers. *Emotion*, 12(6):1222, 2012.

Rosenbaum, Paul R and Rubin, Donald B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, Donald B. *Matched sampling for causal effects*. Cambridge University Press, 2006.

Seligman, Martin EP, Steen, Tracy A, Park, Nansook, and Peterson, Christopher. Positive psychology progress: empirical validation of interventions. *American psychologist*, 60(5):410, 2005.

Springer, Aaron, Hollis, Victoria, and Whittaker, Steve. Mood modeling: accuracy depends on active logging and reflection. *Personal and Ubiquitous Computing*, pp. 1–15, 2018.

Stone, Arthur A, Schwartz, Joseph E, Schkade, David, Schwarz, Norbert, Krueger, Alan, and Kahneman, Daniel. A population approach to the study of emotion: diurnal rhythms of a working day examined with the day reconstruction method. *Emotion*, 6(1):139, 2006.