Final Group Project Proposal

Missing Data Imputation for Time Series Accelerometry Data

1. Objectives

The wearable device has emerged as an important means of assessing human behaviors and served to define outcome measures in health observational and experimental studies. However, instances such as non-wearing of the device or partial wear can lead to an underestimation of total activity. Moreover, there might be potential disparities in activity levels between individuals with complete and incomplete or even no data, thereby rendering the estimated summary statistics from complete days susceptible to selection bias. This study focuses on the measurement of sedentary time during the day (8am–10pm).

This research aims to address this challenge by pioneering and assessing various imputation methods for the missing time series data to mitigate underreporting bias, a crucial factor influencing the quality of wearable tracking data. This effort not only carries immediate relevance for ongoing investigations but also offers enduring value for future researchers to contemplate, extending beyond the confines of the present analysis result.

2. Data

We use data from the Physical Activity and Transit Survey (PAT). The survey asked adults in residential households in New York about their physical activity at work, home, commuting, and recreation. Among the 3811 respondents, 679 of them consented to wear accelerometer devices during all waking hours for one week. We created a concatenated dataset consisting of self-reported surveys and minute-level accelerometer data. The daytime raw accelerometer data has 4,277,700 observations, and the missing rate is about 22%.

3. Methods

We have two types of key measures to impute in the accelerometer data: 1) raw activity count; As an alternative, we also impute 2) a binary indicator of whether the missing minute is a sedentary minute, defined when the activity count falls between 0 and 99. There are several similar studies in the previous literature attempting to tackle the missing data problem in the time series accelerometry data. Two common methods are linear interpolation and zero-inflated Poisson models (R package ‘accelmissing’) (Lee et al, 2019).

In this research, we first analyze the relationship between missing patterns and time-varying variables. Then, starts with logical imputation and is followed by stochastic imputation based on two different methods: 1) Multilevel regression for imputation using the ‘mice’ package, given the multilevel data structure, time points nested into individuals, we can impute the binary indicator for a sedentary minute; 2) Long Short Term Memory Neural Network. One advantage of using the deep learning method is that it does not require feature selections allowing us to directly feed all available data to the model. For instance, times series forecasting for finance with this method (Chen et al., 2016; Borovykh et al. 2017).

Once we have the completed imputation using the four methods mentioned above, we will split the data into 75% and 25% train-test datasets, and evaluate the prediction accuracy. This study answers the following research questions:

1. How does the imputation result in different key estimates compared to the complete cases method?
2. How do the accelerometry-based estimates compare to the self-reports in the survey?
3. How do different modeling approaches perform?
4. Does imputing activity count and determining sedentary status, as opposed to imputing sedentary status, enhance the accuracy and reliability of activity data imputation

4. Computational challenge

Sensory digital traces are a new type of big data with a complex structure and relatively large size (approx. 1 GB). Multilevel regression can be challenging due to the nonconvergence because there are about 6,310 minute-level observations per participant and 678 participants. The computational intensity of these methods will be a focal point of our resource planning and execution strategy.

Reference

Ae Lee, J., & Gill, J. (2018). Missing value imputation for physical activity data measured by accelerometer. *Statistical methods in medical research*, *27*(2), 490-506.

Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*.

Chen, J. F., Chen, W. L., Huang, C. P., Huang, S. H., & Chen, A. P. (2016, November). Financial time-series data analysis using deep convolutional neural networks. In *2016 7th International conference on cloud computing and big data (CCBD)* (pp. 87-92). IEEE.