BIOS 6643 Project Data

Samantha Bothwell

10/30/2020

Data Description

Individuals who classified as overweight or obese were enrolled into the study to understand factors that contribute to weight loss. Participants were asked to step on a bluetooth scale once a day over the course of the study. Within the study, there are 3 cohorts. These cohorts indicate participants who started the study around the same time.

Cohort 1 (N = 29)

- 29 people beginning 04/10/2018 – 07/01/2018 (All but 1 began on in April and May)
- Ending between 11/14/2018 – 02/20/2020
- Mean study days 546.6 (95% CI: 324.8, 600.0)

Cohort 2 (N = 27)

- 27 people beginning 02/13/2019 – 09/30/2019 (All but 2 began between 02/13/2020 – 02/16/2020)
- Ending between 12/22/2019 – 04/20/2020 (All but 2 ended between 04/05/2020 – 04/20/2020)
- Mean study days 412.85 (95% CI: 233.4, 433.0)

Cohort 3 (N = 37)

- 37 people beginning 09/30/2019 – 11/05/2020
- Ending between 12/02/2019 – 04/20/2020 (All but 2 ended in March or April)
- Mean study days 178.35 (95% CI: 79.7, 202.0)

Goal

The research questions of interest are :

- What is the trajectory of weight over the duration of time in the study?
- Is there a relationship between month of study and weight loss, when accounting for sex and age?

Data Analysis

For this project I am considering two different approaches :

- A Functional Data Analysis approach - this would incorporate fitting a curve to the longitudinal data and analyzing its behavior.
- A Spline Model - this would be similar to the FDA approach, however the functions will not smoothed as they are in FDA.

Correlation

There are many ways the data are correlated.

1. Recordings by individual
2. Cohort 1 weights were all taken before COVID. Cohorts 2 and 3 had their weights recorded before and during COVID.
3. There may be a correlation based on the month. For example, individuals may be more incentized to lose weight during January because of New Years. Or, warmer months may encourage people to workout than during colder months.

Interesting Data Factors

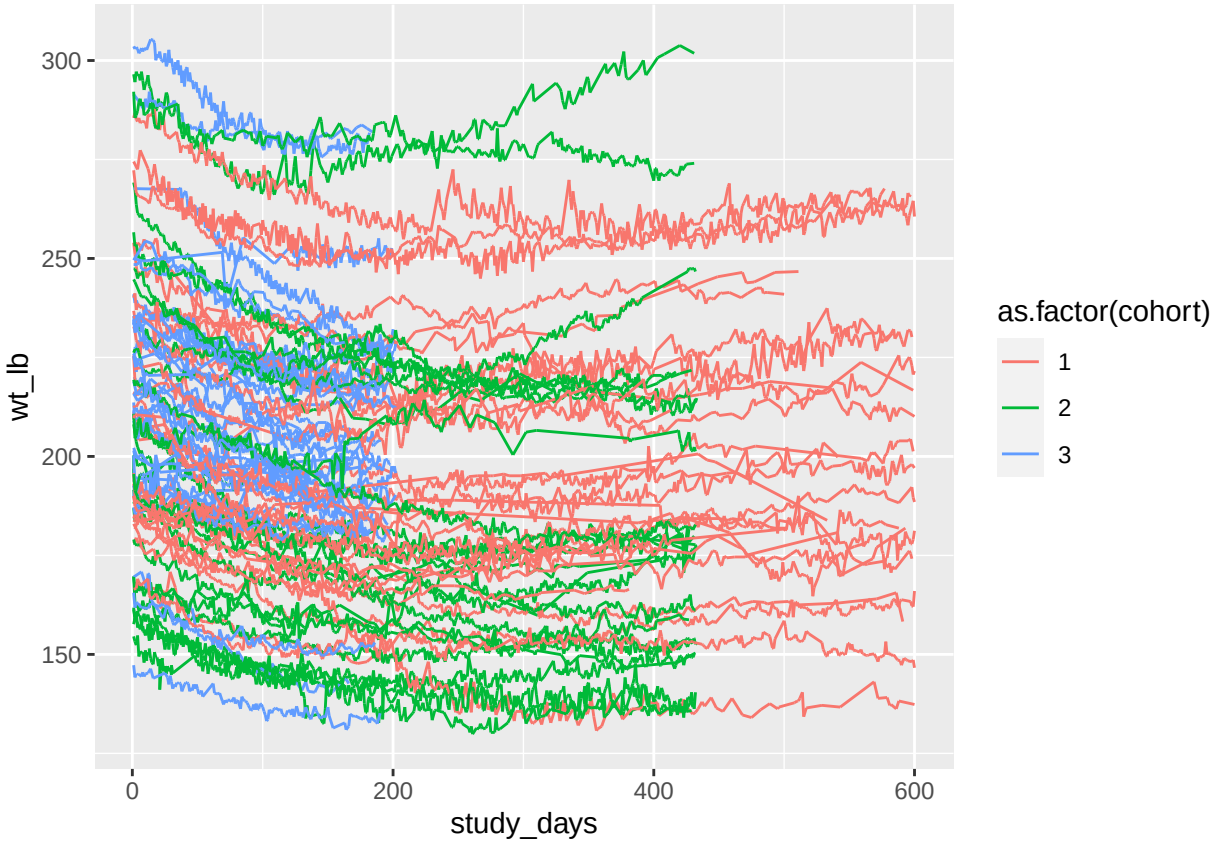
When exploring data in the functional data analysis framework, it is ideal to have regular data. Regular data are data where measurements are taken on a consistent data, without gaps. In this dataset, individuals have differing days of data and skip recording on some days. To account for this, I will need to research how to properly handle irregular data for analysis.

Messy/Unique Data Factors

Cohorts were grouped by start date as well as length of time in the study. Both cohorts 2 and 3 ended at the same time, due to COVID-19, but started at varying times. The given cohorts are not clearly defined by start date. For example, in cohort 2 25/27 participants started the study between 2/13/2020 - 2/16/2020 and 25/27 participants ended the study between 4/5/2020 - 4/20/2020. For the individuals who do not fall within the same interval as the majority, it is not clear yet how they should be grouped.
It may be more meaningful to simply look at month in the study, rather than the cohort groupings. This is something I plan to investigate further.

Preliminary Analysis

Based on the spaghetti plot, printed below, the biggest difference between the 3 cohorts is the length of study. Cohort 3 has an average length in study of 178.35 days. Cohort 2 has an average length in study of 412.85 days. Cohort 1 has the longest average length in study of 546.6 days. A reason why cohort may be important is because all the cohorts seem to show a decline in weights from 1 - 200 days. Cohorts 1 and 2 seem to flatten out after 200 days. Cohort 3 does not seem to flatten out due to its short study length. An additonal reason to consider cohort, as mentioned before, is because COVID-19 could confound the weight loss trajectory since stay at home orders could inhibit someone's ability to lose weight.


300
250
200
150
wt_lb
0
200
400
600
study_days
as.factor(cohort)
1
2
3