Module 2: Set-Up of Independent Variables

April 24, 2021

Contents

MODULE 2 GOAL: By the end of this module, you will be able to:

- Identify which rows and columns in the `merged.csv` file to select when one wishes to use responses from Pre-/Post- Quit Random EMA Questionnaires as Independent Variables (IV's) *exclusively from the post-quit period*
- Learn the logic of creating new time variables from existing time variables in the curated datasets

1 Scientific Question

We will begin Module 2 by introducing a scientific question which we will use to motivate our running illustrative example in Modules 2-5.

SCIENTIFIC QUESTION: On average, is self-efficacy *at the current time point* associated with *the proximal occurrence of cigarette smoking* during the post-quit period?

We note that what we will discuss today is a simplified illustrative example of an observational Intensive Longitudinal Data (ILD) analysis.

2 Potential Issues

What are potential issues we need to consider when investigating our scientific question? In this module, let's consider potential issues concerning our Independent Variable of interest.

Ideally, we would like to infer the associative relationship described in our scientific question of interest at all moments of time during the post-quit period. However,

- we do not have a numerical rating of self-efficacy at *all* time points during the post-quit period

- even if participants provide a rating of self-efficacy through the various different types of EMA Questionnaires, *ratings of self-efficacy provided in Self-Initiated EMA Questionnaires may potentially be influenced by a confounding variable* – namely, a variable simultaneously influencing both the Independent Variable and Dependent Variable in our scientific question

We then ask,

- How is information on self-efficacy collected during the post-quit period?

The question 'I *am* confident in my ability NOT TO SMOKE' is present in 8 out of the 9 types of EMA Questionnaires (i.e., all types of Questionnaires except for the Post-Quit Already Slipped EMA Questionnaire).

Figure 1: Rows and Columns in Toy Dataset

- Among the various types of EMA Questionnaires, which EMA Questionnaires may be utilized for our Independent Variable?

In most cases, we will utilize responses in Post-/Pre- Quit Random EMA Questionnaires for our Independent Variables.

Figure 2: Rows and Columns in Toy Dataset

BREAK: Any questions?

3 Trimming Down the Data Files

Reference file with R code: `module-02.R`

Let's see how these concepts are used as we trim down the data files `dat_postquit_random_ema` and `dat_prequit_random_ema` which we created back in Module 1.

3.1 Step 1

```r
subset_dat_postquit_random_ema <- dat_postquit_random_ema %>%
  # Exclude rows corresponding to EMAs having no response to any item
  filter(with_any_response == 1) %>%
  # Exclude rows corresponding to EMAs launched before Quit Date
  filter(use_as_postquit == 1) %>%
  # Select only the columns you will need;
  # this process will result in a smaller data file
  select(id:time_unixts, postquit_random_item_8) %>%
  rename(selfeff = postquit_random_item_8)

subset_dat_prequit_random_ema <- dat_prequit_random_ema %>%
  # Exclude rows corresponding to EMAs having no response to any item
  filter(with_any_response == 1) %>%
  # Exclude rows corresponding to EMAs launched before Quit Date
  filter(use_as_postquit == 1) %>%
  # Select only the columns you will need;
  # this process will result in a smaller data file
  select(id:time_unixts, prequit_random_item_8) %>%
  rename(selfeff = prequit_random_item_8)
```

The correspondence between the column names in `merged.csv` and the original column names in the codebook `PNS EMA Codebook 07202010.docx` can also be seen in `ema_item_names.csv` by comparing the `name_new` column against the `name_codebook` column.

Inspection of the codebook will show that variable names referring to the same question may be identical across different types of EMAs. As noted earlier, the question ‘I *am* confident in my ability NOT TO SMOKE’ is present in 8 out of the 9 types of EMA Questionnaires (i.e., all types of Questionnaires except for the Post-Quit Already Slipped EMA Questionnaire). These 8 types of EMA Questionnaires capture responses to this question in a variable named `AbsSelfEff`. There may be questions which were posed to participants only within specific kinds of EMA. For example, the question ‘Think about the specific CAUSE of your slip. The cause is controllable by me.’ is present in 2 out of the 9 kinds of EMA. These 2 kinds of EMA, namely the Post-Quit Already Slipped EMA and Post-Quit About to Slip Part Two EMA, captured responses to this question in a variable named `Attribution4`.

In the code snippet above, we have used the `ema_item_names.csv` file to pick out `postquit_random_item_8` and `prequit_random_item_8` by checking for item names under the `name_new` column which correspond to the name `AbsSelfEff` in the `name_codebook` column.

3.2 Step 2

```r
# rbind simply stacks the two data files on top of each other
# rbind will work only if both data files have identical column names
# Hence, there was a need to call the rename() function above prior
# to calling rbind()
dat_analysis <- rbind(subset_dat_postquit_random_ema, subset_dat_prequit_random_ema)
```

BREAK: Any questions?

4 Creating New Time Variables using Existing Time Variables in the Data Files

Reference file with R code: `module-02.R`

4.1 Step 1

```r
# Remember: order according to increasing participant ID
# and within each participant ID, according to increasing time
dat_analysis <- dat_analysis %>% arrange(id, time_unixts)
```

IMPORTANT: The calculations of the time variables in the following step will be incorrect if the data file in the current step has not yet been ordered according to increased participant ID and within each participant ID, according to increased time

4.2 Step 2

```r
# Using dat_analysis:
# - Calculate number of hours elapsed between the current EMA and the next EMA
# - Calculate number of days elapsed since the beginning of post-quit period

dat_analysis <- dat_analysis %>%
  # The group_by() function is needed to ensure that we do not accidentally
  # use data from another participant to calculate lagged time variables
  # for a particular participant
  group_by(id) %>%
  # When did the participant begin responding to the next EMA?
  # If there is no EMA that follows the current EMA,
  # we will initially set time_unixts_plusone to a missing value
  mutate(time_unixts_plusone = c(tail(time_unixts, n=-1), NA)) %>%
  # If there is no EMA that follows the current EMA, we will set time_unixts_plusone
  # to be equal to end_study_unixts
  mutate(time_unixts_plusone = if_else(is.na(time_unixts_plusone),
                                       end_study_unixts,
                                       time_unixts_plusone)) %>%
  # How many seconds elapsed between the current and next EMA?
  mutate(num_secs_elapsed_since_previous_ema = time_unixts_plusone - time_unixts) %>%
  # Now, make a conversion from seconds to hours
  mutate(num_hrs_elapsed_since_previous_ema = num_secs_elapsed_since_previous_ema/(60*60))

dat_analysis <- dat_analysis %>%
  # How many seconds elapsed between the current EMA and Quit Date?
  # Note that we consider 4AM on Quit Date to be the time when participants
  # quit smoking in the PNS study
  mutate(num_secs_elapsed_since_quit = time_unixts - quit_unixts) %>%
  # Now, make a conversion from seconds to hours
  mutate(num_hrs_elapsed_since_quit = num_secs_elapsed_since_quit/(60*60)) %>%
  # Now, make a conversion from hours to days
  mutate(num_days_elapsed_since_quit = num_hrs_elapsed_since_quit/24)

# Create a new time variable that captures the number of days elapsed
# since 12AM of start of study
dat_analysis <- dat_analysis %>%
  mutate(num_secs_elapsed_since_start_study = time_unixts - start_study_unixts) %>%
  mutate(num_hrs_elapsed_since_start_study = num_secs_elapsed_since_start_study/(60*60)) %>%
  mutate(num_days_elapsed_since_start_study = num_hrs_elapsed_since_start_study/24)
```

BREAK: Any questions?