Module 5: Modeling and Hypothesis Testing

April 24, 2021

Contents

MODULE 5 GOAL: By the end of this module, you will be able to:

- Identify which rows to select in a 'Main Analysis' and a 'Sensitivity Analysis'.
- See an example of how the dataset created using the process discussed in Modules 1-4 can be used to test hypothesis.
- Through a live demo, know what to expect when using `module-05.R` independently.

1 Scientific Question

For convenience, let's display the scientific question we introduced in an earlier module.

SCIENTIFIC QUESTION 1: 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 the time variables we constructed in earlier modules also allow us to test whether the above associative effect varies with time, i.e.,

SCIENTIFIC QUESTION 2: Does association of self-efficacy *at the current time point* with *the proximal occurrence of cigarette smoking* vary across time during the post-quit period?

2 Dataset for Analysis

Let's now see how the Independent Variables, Dependent Variables, and Time Variables we have grown acquainted with in Modules 1-4 now come together into a dataset we may utilize to investigate the scientific questions above.

The dataset we now have (i.e., `dat_analysis` from Module 4) contains the following columns...

- `id`
- `sensitivity`
- `selfeff`
- `num_days_elapsed_since_quit`
- `num_hrs_elapsed_since_previous_ema`
- `count_within_bounds`

. . . and the information in these columns may be visualized in relation to each other.

Figure 1: Timeline

BREAK: Any questions?

3 Models for the Dependent Variable

In terms of variables in `dat_analysis` ...

MODEL FOR SCIENTIFIC QUESTION 1:

$$\log\left(E\left\{\frac{\text{count\_within\_bounds}}{\text{num\_hrs\_elapsed\_since\_previous\_ema}}\right\}\right) = \beta_0 + \beta_1 \text{selfeff} + \nu_i$$

MODEL FOR SCIENTIFIC QUESTION 2:

$$\begin{aligned}\log\left(E\left\{\frac{\text{count\_within\_bounds}}{\text{num\_hrs\_elapsed\_since\_previous\_ema}}\right\}\right) = \beta_0 &+ \beta_1 \text{selfeff} \\ &+ \beta_2 \text{num\_days\_elapsed\_since\_quit} \\ &+ \beta_3(\text{num\_days\_elapsed\_since\_quit} \times \text{selfeff}) \\ &+ \nu_i\end{aligned}$$

In terms of math ...

MODEL FOR SCIENTIFIC QUESTION 1:

$$\log\left(E\left\{\frac{Y_{i,t_j}}{L_{i,t_j}}\right\}\right) = \beta_0 + \beta_1 X_{i,t_j} + \nu_i$$

MODEL FOR SCIENTIFIC QUESTION 2:

$$\log\left(E\left\{\frac{Y_{i,t_j}}{L_{i,t_j}}\right\}\right) = \beta_0 + \beta_1 X_{i,t_j} + \beta_2 D_{i,t_j} + \beta_3 (D_{i,t_j} \times X_{i,t_j}) + \nu_i$$

4 Hypothesis Testing

Reference file with R code: `module-05.R`

4.1 Step 1

```
dat_analysis <- dat_analysis %>%
  select(id,
         sensitivity,
         selfeff,
         num_days_elapsed_since_quit,
         num_hrs_elapsed_since_previous_ema,
         count_within_bounds)

# Transform hours elapsed into log-scale
dat_analysis$logged_hrs_elapsed <- log(dat_analysis$num_hrs_elapsed_since_previous_ema)

# Round up totals;
# For example, 0.5 will be rounded up to 1; 1.5 will be rounded up to 2, etc.
dat_analysis$roundedup_count_within_bounds <- as.integer(ceiling(dat_analysis$count_within_bounds))
```

4.2 Step 2

```
# Create a new data frame, which is essentially dat_analysis copied
dat_main_analysis <- dat_analysis

# Now, using dat_main_analysis, take those rows which will be included
# in Sensitivity Analysis. In other words, drop all those rows which should
# be excluded from Sensitivity Analysis
dat_sensitivity_analysis <- dat_main_analysis %>% filter(sensitivity == 1)
```

We note that we will be using identical models for 'Main Analysis' and 'Sensitivity Analysis'. Both types of analyses only differ with respect to which participants will be used to estimate the two models discussed above:

- In 'Main Analysis', all participants in `dat_analysis` will be used to estimate both models
- In 'Sensitivity Analysis', only those participants in `dat_analysis` having `sensitivity = 1` will be used to estimate both models

4.3 Step 3

In the remaining steps, we will do a complete-case analysis. Rows having missing values in any of the dependent variables or independent variables will be omitted. In your analysis, you would have to consider how to address missing data in both of these variables, e.g., via an imputation procedure prior to running `glmer`.

```
# Estimate coefficients of the model using glmer
fit_main_1 <- glmer(roundedup_count_within_bounds ~ offset(logged_hrs_elapsed)
                                                  + 1 + selfeff
                                                  + (1 | id),
                    data = dat_main_analysis,
                    family = poisson(link="log"),
                    na.action = na.omit)
```

4.4 Step 4

```
# Estimate coefficients of the model using glmer
fit_sensitivity_1 <- glmer(roundedup_count_within_bounds ~ offset(logged_hrs_elapsed)
                                                          + 1 + selfeff
                                                          + (1 | id),
                           data = dat_sensitivity_analysis,
                           family = poisson(link="log"),
                           na.action = na.omit)
```

4.5 Step 5

A similar logic to Steps 3 and 4 above may be used to estimate the model for Scientific Question 2. To help with convergence of the estimation process, we will rescale the variable `num_days_elapsed_since_quit` prior to estimation (try removing the division by 100!). This rescaling has the effect of estimating the following model:

$$\log\left(E\left\{\frac{\text{count\_within\_bounds}}{\text{num\_hrs\_elapsed\_since\_previous\_ema}}\right\}\right) = \beta_0 + \beta_1 \text{selfeff} + \beta_2 \frac{\text{num\_days\_elapsed\_since\_quit}}{100} + \beta_3 \left(\frac{\text{num\_days\_elapsed\_since\_quit}}{100} \times \text{selfeff}\right)$$

```
# Estimate coefficients of the model using glmer
fit_main_2 <- glmer(roundedup_count_within_bounds ~ offset(logged_hrs_elapsed)
                                                  + 1 + selfeff
                                                  + I(num_days_elapsed_since_quit/100)
                                                  + selfeff:I(num_days_elapsed_since_quit/100)
                                                  + (1 | id),
                    data = dat_main_analysis,
                    family = poisson(link="log"),
                    na.action = na.omit)
```

4.6 Step 6

```
# Estimate coefficients of the model using glmer
fit_sensitivity_2 <- glmer(roundedup_count_within_bounds ~ offset(logged_hrs_elapsed)
                                                          + 1 + selfeff
                                                          + I(num_days_elapsed_since_quit/100)
                                                          + selfeff:I(num_days_elapsed_since_quit/100)
                                                          + (1 | id),
                           data = dat_sensitivity_analysis,
                           family = poisson(link="log"),
                           na.action = na.omit)
```

5 Live Demo of `module-05.R`

BREAK: Any questions?