Statistical Analysis for SARA

Tianchen Qian, Mashfiqui Rabbi, Susan Murphy

2019.02.05

In this document, we describe the statistical methods to conduct the primary analysis, secondary analysis, and exploratory analysis for SARA. They are implemented in the attached R code; usage of the R code doesn't require understanding of the sections with *.

Throughout this document, we assume the treatment probability is a known constant (constant across time and across all subjects).

The wording of primary hypotheses, secondary hypotheses, and exploratory analysis are all extracted from Mash's working paper.

1 Statistical method for primary analysis

1.1 Primary hypothesis 1

H_0: The 4pm push notification with inspirational quote will **not** increase the full completion of survey and/or active task the same day as compared to no inspirational quote.

H_1: The 4pm push notification with inspirational quote will increase the full completion of survey and/or active task the same day as compared to no inspirational quote.

1.1.1 Test statistic and critical value

Data notation for testing this hypothesis. The data collected from participant i is denoted by $(Z_{i1}, A_{i1}, Y_{i1}, \ldots, Z_{i,T_i}, A_{i,T_i}, Y_{i,T_i})$, where the subscript $1, 2, \ldots, T_i$ denotes day in the study. Here we allow the length of observations T_i to be different for each individual. Y_{it} denotes the binary outcome (full completion of either the survey or completion of both active tasks) on day t. A_{it} is the indicator of treatment on day t ($A_{it} = 1$ if sent 4pm push notification, 0 if not). Z_{it} is the control variables on day t. The purpose of Z_{it} is to reduce variance in the analysis. The definition of which observations can be in Z_{it} depends on the

definition of A_{it}: Z_{it} can only include information that is available prior to A_{it}. In this particular example, Z_{it} can only include information that is available prior to 4pm on day t. For instance, Z_{it} can include variables such as: weather on day t (before 4pm), previous day adherence $Y_{i,t-1}$, and other information available before 4pm on day t. Let I_{it} denote the availability status indicator for day t: if unavailable, the treatment is not delivered.

Marginal treatment effect on log scale. To form the test statistic, we consider the marginal treatment effect on the log scale:

$$\begin{aligned}\beta_0 &:= \log\left(\frac{\sum_{t=1}^{T_i} E[Y_{it}|A_{it}=1, I_{it}=1]}{\sum_{t=1}^{T_i} E[Y_{it}|A_{it}=0, I_{it}=1]}\right) \\ &= \log\left(\frac{\sum_{t=1}^{T_i} e^{\beta_{0t}} E[Y_{it}|A_{it}=0, I_{it}=1]}{\sum_{t=1}^{T_i} E[Y_{it}|A_{it}=0, I_{it}=1]}\right), \end{aligned} \tag{1}$$

where $\beta_{0t} = \log\left(\frac{E[Y_{it}|A_{it}=1,I_{it}=1]}{E[Y_{it}|A_{it}=0,I_{it}=1]}\right)$. Thus the definition of β_0 given in (1) is marginal both accross time as well as across users.

Test statistic and critical value. In Section 1.1.2, we describe how to construct an estimator $\hat{\beta}$ for this marginal treatment effect β_0, and an estimate of its variance $\widehat{\text{Var}}(\hat{\beta})$. The test statistic T is defined by

$$T = \frac{\hat{\beta}}{\{\widehat{\text{Var}}(\hat{\beta})\}^{1/2}}. \tag{2}$$

To conduct two-sided hypothesis test with significance level η, the critical value is $c = t^{-1}_{n-1-q}(1-\eta/2)$. If $|T| > c$, we reject H_0. Here, n is the sample size, q is the length of vector Z_{it} (including the added intercept), and $t^{-1}_{n-1-q}(\gamma)$ denotes the γ-quantile of t-distribution with $(n-1-q)$ degrees of freedom. For example, suppose Z_{it} includes two variables: the previous day adherence $Y_{i,t-1}$ and the weather of day t. Then $q = 2+1 = 3$.

Software. R function to conduct primary hypothesis 1 is SARA_primary_hypothesis_1 in SARAanalysis.R.

1.1.2 Statistical details*

$\hat{\beta}$ in the test statistic (2) is obtained by solving the following estimating equation (simutaneously solving for $\hat{\alpha}$ and $\hat{\beta}$):

$$0 = \sum_{i=1}^{n} \sum_{t=1}^{T_i} I_{it} e^{-A_{it}\beta} \left(Y_{it} - e^{Z_{it}^T \alpha + A_{it}\beta} \right) \begin{pmatrix} (A_{it} - p_a) \\ e^{Z_{it}^T \alpha} Z_{it} \end{pmatrix}. \tag{3}$$

Here, $p_a = P(A_{it} = 1)$ is the randomization probability (assumed constant across all time and all subjects). The superscript T denotes matrix transpose. Note that as in linear regression, we add an intercept 1 to the vector Z_{it}. It turns out that $\hat{\beta}$ is consistent for β_0 in (1) even if $e^{Z_{it}^T \alpha}$ is a wrong model for $E(Y_{it} = 1 \mid A_{it} = 0)$. In order words, the choice of Z_{it} only affects the variance of $\hat{\beta}$, but does not affect the consistency of $\hat{\beta}$.

Mathematically, solving (3) for $\hat{\beta}$ in this primary hypothesis is a special case of the estimating equation for the exploratory analysis in Section 3.1.2. Therefore, the formula for computing $\widehat{\text{Var}}(\hat{\beta})$ is deferred to Section 4.

1.2 Primary hypothesis 2

H_0: Among individuals who complete the survey, providing a post-survey-completion meme or gif will **not** yield a higher rate of completion of the survey or active task the next day than not providing meme/gif reinforcement after survey completion.

H_1: Among individuals who complete the survey, providing a post-survey-completion meme or gif will not yield a higher rate of completion of the survey or active task the next day than not providing meme/gif reinforcement after survey completion.

1.2.1 Test statistic and critical value

Data notation for testing this hypothesis. The data collected from participant i is denoted by $(Z_{i1}, A_{i1}, Y_{i1}, \ldots, Z_{i,T_i}, A_{i,T_i}, Y_{i,T_i})$, where the subscript $1, 2, \ldots, T_i$ denotes day in the study. Y_{it} denotes the binary outcome (full completion of either the survey or completion of both active tasks) on day t. A_{it} is the indicator of treatment on day t ($A_{it} = 1$ if provided a post-survey-completion meme or gif, 0 if not). Z_{it} is the control variables on day t, which can only include information that is available prior to A_{it}. Let the survey-completion indicator $S_{it} = 1$ if participant i completed the survey on day t, and set $S_{it} = 0$ otherwise. Let I_{it} denote the availability status indicator for day t: if unavailable,

the treatment is not delivered. (Two notes: (i) $S_{it} = Y_{it}$. (ii) In the estimating equation we will multiply S_{it} with I_{it} to obtain the "overall" availability indicator, and the inference is conditional on these available dates. Here we separate those for conceptual clearness.)

Difference from primary hypothesis 1. (i) In testing this hypothesis, Z_{it} can include information that is available prior to the occurrence of post-survey-completion meme on day t. So here (unlike in primary hypothesis 1) we can include in the controls, Z_{it}, data collected by the survey on day t. (ii) In testing this hypothesis, we are interested in the effect of A_{it} on $Y_{i,t+1}$. In using the software, however, do not recode the outcome manually because the algorithm already takes this into account.

Marginal treatment effect on log scale. To form the test statistic, we consider the marginal treatment effect on the log scale for those who are available:

$$\begin{aligned}\beta_0 &:= \log\left(\frac{\sum_{t=1}^{T_i-1} E[Y_{i,t+1}|A_{it}=1, I_{it}=1, S_{it}=1]}{\sum_{t=1}^{T_i-1} E[Y_{i,t+1}|A_{it}=0, I_{it}=1, S_{it}=1]}\right) \\ &= \log\left(\frac{\sum_{t=1}^{T_i-1} e^{\beta_{0t}} E[Y_{i,t+1}|A_{it}=0, I_{it}=1, S_{it}=1]}{\sum_{t=1}^{T_i-1} E[Y_{i,t+1}|A_{it}=0, I_{it}=1, S_{it}=1]}\right), \end{aligned} \tag{4}$$

where $\beta_{0t} = \log\left(\frac{E[Y_{i,t+1}|A_{it}=1,I_{it}=1,S_{it}=1]}{E[Y_{i,t+1}|A_{it}=0,I_{it}=1,S_{it}=1]}\right)$. Thus the definition of β_0 given in (4) is marginal both accross time as well as across users.

Test statistic and critical value. In Section 1.2.2, we describe how to construct an estimator $\hat{\beta}$ for this marginal treatment effect β_0, and an estimate of its variance $\widehat{\text{Var}}(\hat{\beta})$. The test statistic T is defined by

$$T = \frac{\hat{\beta}}{\{\widehat{\text{Var}}(\hat{\beta})\}^{1/2}}. \tag{5}$$

To conduct two-sided hypothesis test with significance level η, the critical value is $c = t^{-1}_{n-1-q}(1-\eta/2)$. If $|T| > c$, we reject H_0. Here, n is the sample size, q is the length of vector Z_{it} (including the added intercept), and $t^{-1}_{n-1-q}(\gamma)$ denotes the γ-quantile of t-distribution with $(n-1-q)$ degrees of freedom. For example, suppose Z_{it} includes two variables: the previous day adherence $Y_{i,t-1}$ and the weather of day t. Then $q = 2+1 = 3$.

Software. R function to conduct primary hypothesis 2 is SARA_primary_hypothesis_2 in SARAanalysis.R.

1.2.2 Statistical details*

$\hat{\beta}$ in the test statistic (5) is obtained by solving the following estimating equation (simutaneously solving for $\hat{\alpha}$ and $\hat{\beta}$):

$$0 = \sum_{i=1}^{n} \sum_{t=1}^{T_i - 1} S_{it} I_{it} e^{-A_{it}\beta} \left(Y_{i,t+1} - e^{Z_{it}^T \alpha + A_{it}\beta} \right) \begin{pmatrix} (A_{it} - p_a) \\ e^{Z_{it}^T \alpha} Z_{it} \end{pmatrix}. \tag{6}$$

Here, $p_a = P(A_{it} = 1)$ is the randomization probability (assumed constant across all time and all subjects). The superscript T denotes matrix transpose. Note that as in linear regression, we add an intercept 1 to the vector Z_{it}. It turns out that $\hat{\beta}$ is consistent for β_0 in (4) even if $e^{Z_{it}^T \alpha}$ is a wrong model for $E(Y_{it} = 1 \mid A_{it} = 0)$. In order words, the choice of Z_{it} only affects the variance of $\hat{\beta}$, but does not affect the consistency of $\hat{\beta}$.

Mathematically, solving (6) for $\hat{\beta}$ in this primary hypothesis is a special case of the estimating equation for the exploratory analysis in Section 3.1.2. Therefore, the formula for computing $\widehat{\text{Var}}(\hat{\beta})$ is deferred to Section 4.

2 Statistical method for secondary analysis

2.1 Secondary hypothesis 1

H_0: The 6pm reminder notification with an extra persuasive message will not yield a higher rate of full completion of the survey or active task the same day than not providing the extra persuasive message.

H_1: The 6pm reminder notification with an extra persuasive message will not yield a higher rate of full completion of the survey or active task the same day than not providing the extra persuasive message.

Note: the statistical method for testing this secondary hypothesis is essentially the same as primary hypothesis 1 in Section 1.1. The differences are: (i) the treatment is defined differently; (ii) the variables allowed in Z_{it} are information available prior to 6pm on day t.

2.2 Secondary hypothesis 2

H_0: Among individuals who complete the active tasks, offering a post-active-task-completion life-insight will not yield a higher rate of the full completion of the survey or active task

the next day than not offering a life-insight after active tasks completion.

H_1: Among individuals who complete the active tasks, offering a post-active-task-completion life-insight will yield a higher rate of the full completion of the survey or active task the next day than not offering a life-insight after active tasks completion.

Note: the statistical method for testing this secondary hypothesis is essentially the same as primary hypothesis 2 in Section 1.2. The differences are: (i) the treatment is defined differently; (ii) the availability indicator $I_{it} = 1$ if participant i completed the active tasks on day t, and $I_{it} = 0$ otherwise.

3 Statistical method for exploratory analysis

3.1 Exploratory analysis

The following description of the exploratory analysis is an excerpt from Mash's working paper.

"We plan to run exploratory analyses to examine how the effectiveness of engagement strategies changes over time (we conjecture the effectiveness will decrease). We will also run additional exploratory analysis to assess effect moderation. We will examine how the effect of engagement strategies is moderated by gender, weekdays/weekends, and whether the day is Sunday vs other days of the week (we expect the longer Sunday's daily surveys' completion may be lower than other days)."

This section describes the statistical method for the second part of the exploratory analysis, about effect moderation.

3.1.1 Model and Estimator

Data notation for exploratory analysis. The data collected from participant i is denoted by $(Z_{i1}, A_{i1}, Y_{i1}, \ldots, Z_{i,T_i}, A_{i,T_i}, Y_{i,T_i})$, where the subscript $1, 2, \ldots, T_i$ denotes day in the study. Y_{it} denotes the binary outcome (full completion of either the survey or completion of both active tasks) on day t. A_{it} is the indicator of treatment on day t ($A_{it} = 1$ if sent 4pm push notification, 0 if not). Z_{it} is the control variables on day t. The purpose of Z_{it} is to reduce variance in the analysis. X_{it} is the potential effect moderators on day t, which is a subset of Z_{it}. For example, X_{it} could include gender, whether day t is a weekday, and whether day t is a Sunday. Z_{it} could include X_{it} and some other covariates

such as the weather of day t. Let I_{it} denote the availability status indicator for day t: if unavailable, the treatment is not delivered.

Moderated treatment effect on log scale. Denote by H_{it} the history of subject i up to day t prior to A_{it}: $H_{it} = \{Z_{i1}, A_{i1}, Y_{i1}, \ldots, Z_{i,t-1}, A_{i,t-1}, Y_{i,t-1}, Z_{it}\}$. To form the test statistic, we consider the moderated treatment effect on log scale:

$$\log\left(\frac{E\{E[Y_{it}|A_{it}=1, H_{it}, I_{it}=1] \mid X_{it}, I_{it}=1\}}{E\{E[Y_{it}|A_{it}=0, H_{it}, I_{it}=1] \mid X_{it}, I_{it}=1\}}\right) = X_{it}^T\beta, \qquad \text{for all } t = 1, 2, \ldots, T_i. \tag{7}$$

The superscript T denotes matrix transpose. As in usual linear regression, we add an intercept 1 to the vector X_{it}. β characterizes the magnitude of effect moderation, and it is a vector of length p, where p is the length of vector X_{it} (including the intercept). Notice that here we assume model (7) holds for all $t = 1, 2, \ldots, T_i$; that is, the definition of β in (7) is not marginal across time, unlike in (1) and (4).

Estimator and standard error. In Section 3.1.2, we describe how to construct an estimator $\hat{\beta}$ for this moderated treatment effect β, and an estimate of its variance-covariance matrix $\widehat{\mathbf{Var}}(\hat{\beta})$.

Software. R function to conduct exploratory analysis is SARA_exploratory_analysis in SARAanalysis.R.

3.1.2 Statistical details*

$\hat{\beta}$ in the test statistic (7) is obtained by solving the following estimating equation (simutaneously solving for $\hat{\alpha}$ and $\hat{\beta}$):

$$0 = \sum_{i=1}^{n}\sum_{t=1}^{T_i} I_{it} e^{-A_{it}X_{it}^T\beta}\left(Y_{it} - e^{Z_{it}^T\alpha + A_{it}X_{it}^T\beta}\right)\begin{pmatrix}(A_{it}-p_a)X_{it}\\ e^{Z_{it}^T\alpha}Z_{it}\end{pmatrix}. \tag{8}$$

Here, $p_a = P(A_{it} = 1)$ is the randomization probability (assumed constant across all time and all subjects). Note that as in linear regression, we add an intercept 1 to both the vector X_{it} and the vector Z_{it}. It turns out that $\hat{\beta}$ is consistent for β in (7) even if $e^{Z_{it}^T\alpha}$ is a wrong model for $E(Y_{it} = 1 \mid A_{it} = 0)$. In order words, the choice of Z_{it} only affects the variance of $\hat{\beta}$, but does not affect the consistency of $\hat{\beta}$.

We present the computation of $\widehat{\mathrm{Var}}(\hat{\beta})$ in the next section.

4 Technical details on computing $\widehat{\text{Var}}(\hat{\beta})$*

In this section, we describe the variance estimates $\widehat{\text{Var}}(\hat{\beta})$ in each of the hypothesis tests.

4.1 General form

Consider the following estimating equation, where n denotes sample size, and T denotes the total number of time points (here the number of days),

$$0 = \sum_{i=1}^{n}\sum_{t=1}^{T} I_{it} e^{-A_{it}X_{it}^T\beta}\left(Y_{it} - e^{Z_{it}^T\alpha + A_{it}X_{it}^T\beta}\right)\begin{pmatrix}(A_{it}-p_a)X_{it}\\ e^{Z_{it}^T\alpha}Z_{it}\end{pmatrix}. \tag{9}$$

For $(\hat{\beta},\hat{\alpha})$ that solves (9), its estimated variance-covariance matrix V equals

$$V = \frac{1}{n}M_n^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} D_i^T(I_i - H_i)^{-1}(Y_i-\mu_i)(Y_i-\mu_i)^T(I_i-H_i^T)^{-1}D_i\right\}\left(M_n^{-1}\right)^T. \tag{10}$$

The terms on the right hand side of (10) are defined as follows:

$$M_n := (1/n)\sum_{i=1}^{n}\sum_{t=1}^{T}\begin{bmatrix} -Y_{it}(A_{it}-p_a)A_{it}e^{-A_{it}X_{it}^T\beta}X_{it}X_{it}^T & -(A_{it}-p_a)e^{Z_{it}^T\alpha}X_{it}Z_{it}^T \\ -Y_{it}e^{-A_{it}X_{it}^T\beta}A_{it}e^{Z_{it}^T\alpha}Z_{it}X_{it}^T & Y_{it}e^{-A_{it}X_{it}^T\beta}e^{Z_{it}^T\alpha}Z_{it}Z_{it}^T - 2e^{2Z_{it}^T\alpha}Z_{it}Z_{it}^T\end{bmatrix},$$

$$D_i := \begin{bmatrix} e^{-A_{i1}X_{i1}^T\beta}(A_{i1}-p_a)X_{i1}^T & e^{-A_{i1}X_{i1}^T\beta}e^{Z_{i1}^T\alpha}Z_{i1}^T \\ \vdots & \vdots \\ e^{-A_{iT}X_{iT}^T\beta}(A_{iT}-p_a)X_{iT}^T & e^{-A_{iT}X_{iT}^T\beta}e^{Z_{iT}^T\alpha}Z_{iT}^T\end{bmatrix},$$

$$E_i := \begin{bmatrix} -e^{Z_{i1}^T\alpha + A_{i1}X_{i1}^T\beta}A_{i1}X_{i1}^T & -e^{Z_{i1}^T\alpha + A_{i1}X_{i1}^T\beta}Z_{i1}^T \\ \vdots & \vdots \\ -e^{Z_{iT}^T\alpha + A_{iT}X_{iT}^T\beta}A_{iT}X_{i1}^T & -e^{Z_{iT}^T\alpha + A_{iT}X_{iT}^T\beta}Z_{iT}^T\end{bmatrix},$$

$$H_i := \frac{1}{n}E_iM_n^{-1}D_i^T,$$

$$Y_i := \begin{bmatrix} I_{i1}Y_{i1}\\ \vdots \\ I_{iT}Y_{iT}\end{bmatrix}, \qquad \mu_i := \begin{bmatrix} I_{i1}e^{Z_{i1}^T\alpha + A_{i1}X_{i1}^T\beta}\\ \vdots \\ I_{iT}e^{Z_{iT}^T\alpha + A_{iT}X_{iT}^T\beta}\end{bmatrix},$$

and I_i is a $T\times T$ identity matrix. The superscript T denotes matrix transpose.

4.2 For primary hypothesis 1

The $\widehat{\text{Var}}(\hat{\beta})$ in the test statistic (2) of primary hypothesis 1 is obtained as follows. In (9), set $X_{it} = 1$ for all i, t, and set $T = T_i$. With these modification, (9) becomes equivalent to the estimating equation (3) presented in Section 1.1.2. Then the (1,1) entry of the matrix V in (10) is the estimated variance $\widehat{\text{Var}}(\hat{\beta})$ in (2).

4.3 For primary hypothesis 2

The $\widehat{\text{Var}}(\hat{\beta})$ in the test statistic (5) of primary hypothesis 2 is obtained as follows. In (9), set $I_{it} \leftarrow I_{it}S_{it}$, replace Y_{it} by $Y_{i,t+1}$, set $T = T_i - 1$. With these modification, (9) becomes equivalent to the estimating equation (6) presented in Section 1.2.2. Then the (1,1) entry of the matrix V in (10) is the estimated variance $\widehat{\text{Var}}(\hat{\beta})$ in (5).

4.4 For exploratory analysis

The $\widehat{\text{Var}}(\hat{\beta})$ in the test statistic (7) of exploratory analysis is obtained as follows. In (9), set $T = T_i$. Suppose p is the length of vector X_{it} (including the added intercept). Then the p-th principal submatrix (i.e., the top-left $p \times p$ submatrix) of the matrix V in (10) is the estimated variance-covariance matrix $\widehat{\text{Var}}(\hat{\beta})$ in (7).