A Bayesian analysis of survey design parameters

A paper for the BADEN Network

Lisette Bruin, Nino Mushkudiani, Barry Schouten

4th Workshop Advances in Adaptive and Responsive Survey Design, November 9 and 10, 2015, Manchester

The Leverhulme Trust

Centraal Bureau voor de Statistiek

Summary

- Objectives;
- Design (definitions, notations, model);
- Bayesian analysis
    - General approach to Bayesian analysis;
    - Prior distributions;
    - Posterior distributions;
- Discussion

Objectives

- To set up a general model for survey design parameters;
- To introduce a Bayesian analysis of survey design parameters;
- To introduce a Bayesian analysis of quality and cost indicators based on survey design parameters;

Survey design parameters

Three sets of survey design parameters suffice to compute most of the quality and cost constraints:

- $\rho_i(s_{1,T})$: Response propensities per unit per strategy;
- $C_i(s_{1,T})$: Expected costs per sample unit per strategy;
- $D_i(s_{1,T})$: Adjusted mode effects per unit per strategy;

We restrict to nonresponse error and leave the adjusted mode effects to future papers.

Functions of survey design parameters

We consider three functions of the design parameters:

- the response rate

$$RR(s_{1,T}) = \frac{1}{N}\sum_{i=1}^{n} d_i \rho_i\,(s_{1,T})$$

- the total cost

$$B(s_{1,T}) = \sum_{i=1}^{n} c_i\,(s_{1,T})$$

- the coefficient of variation

$$CV\left(X, s_{1,T}\right) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{n} d_i(\rho_i\left(s_{1,T}\right) - RR(s_{1,T}))^2}}{RR(s_{1,T})}$$

Definitions

- Actions
  - Choices for design features (number of calls, use of incentive, interview mode)
- Strategy
  - The total of choices made for the design features, denoted by $s_{1,T}$
- Phase
  - T phases of survey design $t = 1,2, \dots, T$
- Auxiliary data
  - A vector x_i that is linked from frame data, administrative data ($x_{0,i}$) or paradata ($x_{t,i}$)

If $x_i = x_{0,i}$, then the ASD is **static**. If for some t, $x_{t,i}$ is used to choose actions in a subsequent phase, then the ASD is **dynamic**.

Modeling survey design parameters

Goal:

A simple, but sufficiently general model including all potential features:

- more than 1 phase
- dynamic
- dependency on history of actions
- non-eligible nonresponse for follow-up

Modeling:

1. Decomposition of model parameters into their main components
2. General linear models that link these components to the available auxiliary variables
3. Assumption that cost, contact and participation per sample unit are independent of those of other sample units

Decomposition (response)

Components:

- $\kappa_{t,i}(s_{1,t})$ - propensity of a contact of subject i in phase t under strategy $s_{1,t}$.
- $\lambda_{t,i}(s_{1,t})$ - propensity of a participation of subject i in phase t given contact under strategy $s_{1,t}$.

Response per phase: $\rho_{t,i}(s_{1,t}) = \kappa_{t,i}(s_{1,t}) \cdot \lambda_{t,i}(s_{1,t})$

Total response (when in subsequent phases all nonresponse receives a follow-up):

$$\rho_i(s_{1,T}) = \underbrace{\kappa_{1,i}(s_1)\,\lambda_{1,i}(s_1)}_{\text{response in phase 1}} + \underbrace{\sum_{t=2}^{T}\left(\underbrace{\left(\prod_{l=1}^{t-1}\left(1-\kappa_{l,i}(s_{1,l})\,\lambda_{l,i}(s_{1,l})\right)\right)}_{\text{no response in phase t-1}}\kappa_{t,i}(s_{1,t})\,\lambda_{t,i}(s_{1,t})\right)}_{\text{response in phase t}}$$

Decomposition (cost)

Components:

- $C_{0,t,i}(s_{1,t})$ – expected costs to make a contact attempt;
- $C_{R,t,i}(s_{1,t})$ – expected costs for the response;
- $C_{NR,t,i}(s_{1,t})$ – expected costs for a nonresponse.

Expected costs per phase:

$$\underbrace{C_{0,t,i}(s_1)}_{\text{contact}} + \underbrace{\kappa_{t,i}(s_1)\ (1 - \lambda_{1,i}(s_1))\ C_{NR,t,i}(s_1)}_{\text{nonresponse}} + \underbrace{\kappa_{t,i}(s_1)\ \lambda_{t,i}(s_1)\ C_{R,t,i}(s_1)}_{\text{response}}$$

Example (phone)

- Costs for contact attempt for time to dial number
- Nonresponse costs only for the duration of the call
- Response costs for the duration of the interview and processing the responses

Modeling design parameter components

Model for contact propensity (similar for participation propensity):

$$h\left(\kappa_{t,i}\left(s_{1,t}\right)\right) = \begin{cases} \alpha_{t,0}(s_t)x_{0,i} + \delta_t^C\left(s_{1,t-1}\right), & t < t_1, \\ \alpha_{t,0}(s_t)x_{0,i} + \alpha_{t,t_1}(s_t)x_{t_1,i} + \delta_t^C\left(s_{1,t-1}\right), & t \geq t_1. \end{cases}$$

Model for expected response costs (similar for contact and nonresponse costs):

$$C_{R,i}(s) = \gamma_R(s)x_{0,i} + \varepsilon_{R,i}(s), \qquad s \in \mathcal{S}.$$

Examples

- $x_{0,i}$: the age (group) of the subject
- $\alpha_{t,0}(s_t)$: relation between age and response
- $x_{t_1,i}$: whether a (web) survey is started, but not finished
- $\gamma_R(s)$: a measure for the expected interview time per age (group).

Bayesian analysis

General approach:

1. Assume independency of parameters;
2. Assign prior distributions;
3. Derive likelihood functions;
4. Derive approximations to posterior distributions of design parameters;
5. Derive approximations to posterior distributions of aggregate quality and cost measures (functions of design parameters).

Bayesian analysis

Prior distributions (hyperpriors):

- Inverse Gamma: variance parameters in $\varepsilon_{0,i}(s), \varepsilon_{R,i}(s), \varepsilon_{NR,i}(s)$
- Normal distribution: all other regression parameters

Parameters prior distribution (hyperparameters)

derived from:

- Expert knowledge
- Historic survey data (empirical Bayes)

Bayesian analysis

Posterior distributions

Joint posterior distributions of interest:

1. Individual response propensities and costs – optimization parameters
2. Overall quality and cost indicators – monitoring analysis

Required observed data:

- Realized costs
- Response outcomes
- Used strategies
- Auxiliary data

Bayesian analysis

Posterior distributions

No closed forms: Posterior distributions of response propensities and costs (and overall quality and cost indicators) do not have closed forms.

Proposal: Draw MCMC samples from the posterior distributions of the regression parameters in the contact, participation and cost models.

Advantage: Posterior distributions of overall quality and cost indicators follow directly from the samples.

Obvious choice: Gibbs sampler to iterate draws for each parameter separately (some conditional distributions still without closed forms)

Discussion

Model

- Is the model sufficiently general/simple?

Prior distributions

- Acceptable choices?
- Is the assumption of the independency of the priors realistic?
- What are meaningful properties to investigate in a simulation study?
- How to translate knowledge to hyperparameters?

Posterior distributions

- Approximation using Gibbs sampler?
- How to deal with the non-linear link functions?