A Bayesian analysis of survey design parameters

Lisette Bruin, Nino Mushkudiani, Barry Schouten

Symposium StatCan, March 22-24, 2016, Ottawa

The Leverhulme Trust

Centraal Bureau voor de Statistiek

Summary

- Introduction;
    - Adaptive survey design
    - BADEN
    - Objectives;
- Design (definitions, notations, model);
- Bayesian analysis (general approach, priors and posteriors);
- Simulation study (goals, approach, results);
- Future work/discussion

BADEN

Bayesian **A**daptive survey **DE**sign **N**etwork

International network on targeted data collection strategies employing historical survey data

Jan 2015 – Dec 2017, funded by The Leverhulme Trust

- **Participating institutes:**
  - Universities of Manchester (PI), Michigan and Southampton,
  - CBS, RTI, Stat Sweden, US Census Bureau;
- **Goals:**
  - Add expert knowledge and knowledge from historic survey data in monitoring and analysis (phase 1);
  - Adjustment/adaptation (phase 2);

Adaptive survey design

Adaptive survey designs differentiate survey design features for different population subgroups based on auxiliary data about the sample obtained from frame data, registry data or paradata.

Instead of a single (uniform) strategy multiple candidate strategies can be drawn

Why adaptive survey designs?

- **Response:** persons have different preferences for communication and interview, i.e. respond differently to different data collection strategies;
- **Costs:** different strategies are associated with different costs per person;

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;
- $D_i\left(s_{1,T}\right)$: 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

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

- the total cost

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

- 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 $t = 1,2, \ldots, T$
- 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

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).

Parameters in prior distribution (hyperparameters):

- Expert knowledge
- Historic survey data

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)

Simulation study

Goals

Analyse the impact of:

- Misspecified prior distributions;
- Dispersion of prior distributions (non-informative vs informative);
- Sample size;

Additionally, investigate:

- Convergence properties and computation times;

Simulation study

Simulation

- Three phases: CAWI → CAPI3 → CAPI3+
- Simulation based on known parameters from the Health Survey

Decomposition

- Response per phase: $\rho_{t,i}\left(s_{1,t}\right) = \kappa_{t,i}\left(s_{1,t}\right) \cdot \lambda_{t,i}\left(s_{1,t}\right)$
- Costs per phase: $C_{t,i}(s_{1,t}) = C_{0,t,i}\left(s_{1,t}\right) + C_{R,t,i}\left(s_{1,t}\right) + C_{NR,t,i}(s_{1,t})$

Models

- Contact and participation: probit regression on age, gender and 0-1 indicator for web break-off;
- Costs: linear regression on age and gender;

Simulation study

Prior distributions (hyperpriors):

- Inverse Gamma: variance parameters in error terms of cost functions;
- Normal distribution: all other regression parameters;

Posterior distributions:

- Approximated using a Gibbs sampler with data augmentation;

Simulation study

Chosen priors:

- **‘True’;**

  Computed and estimated with regression from original response rates
- **Misspecified;**

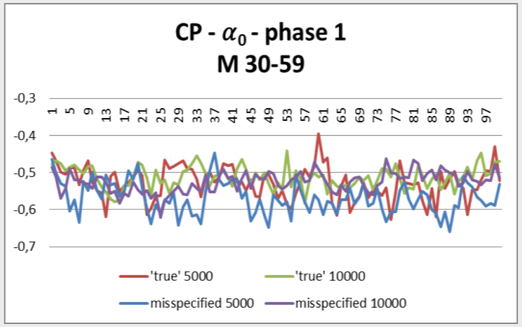
  ‘True’ prior with the mean of α_0 multiplied by 3
- **Non-informative;**

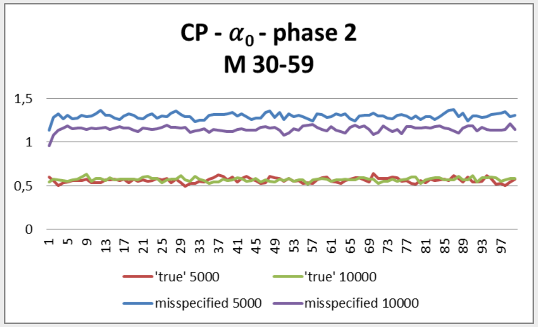
  Same standard deviations as ‘true’ prior, but equally distributed means
- **Non-informative with a larger variance;**

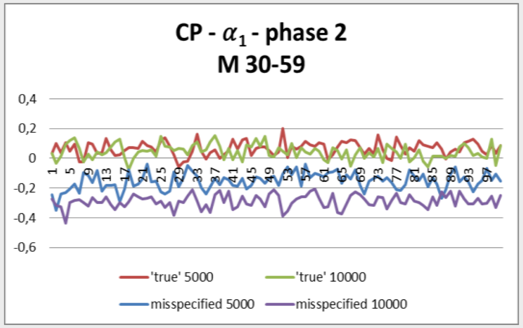
  Standard deviation multiplied by $\sqrt{10}$

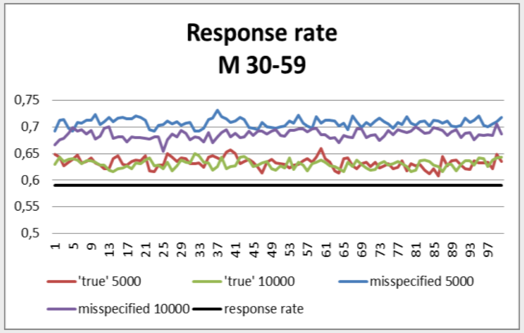
Simulation study - results

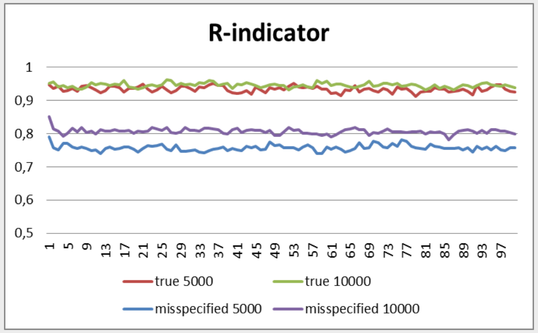
True prior vs misspecified prior (μ_{α_0} multiplied by 3)

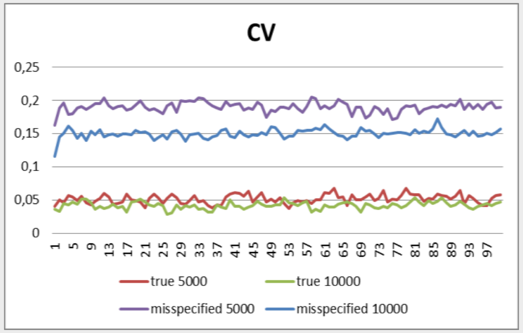

CP - α_0 - phase 1
M 30-59


CP - α_0 - phase 2
M 30-59


CP - α_1 - phase 2
M 30-59


Response rate
M 30-59

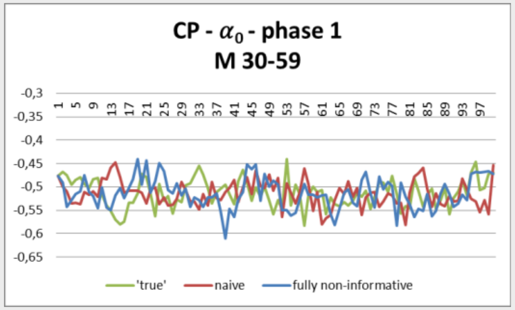

R-indicator

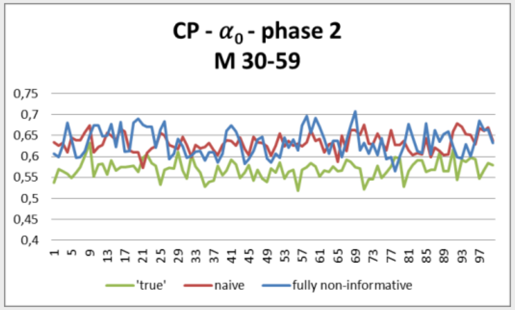

CV

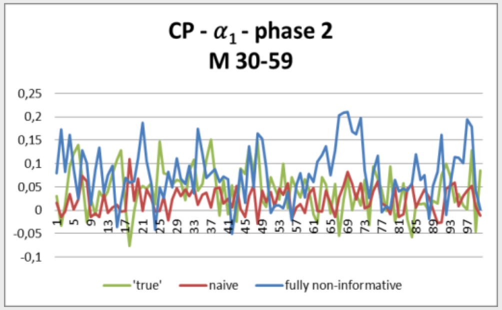
- A misspecified prior has no impact in the first phase with one covariate
- The impact of a misspecified prior is larger in a smaller dataset
- In a smaller dataset the variances of the posterior distributions are larger

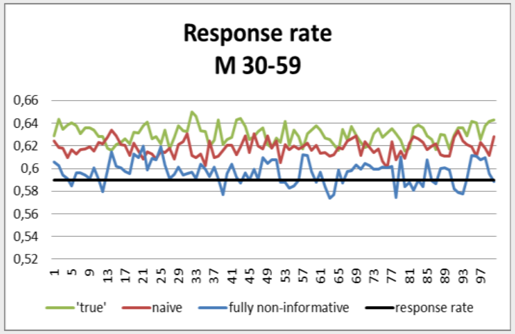
Simulation study - results

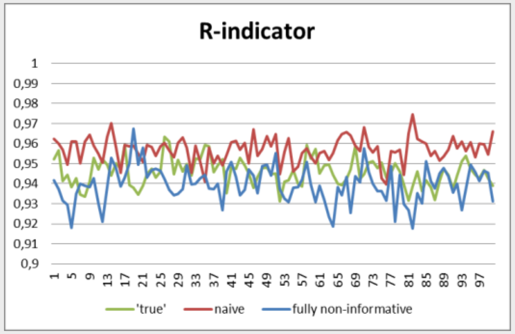
True prior vs non-informative priors

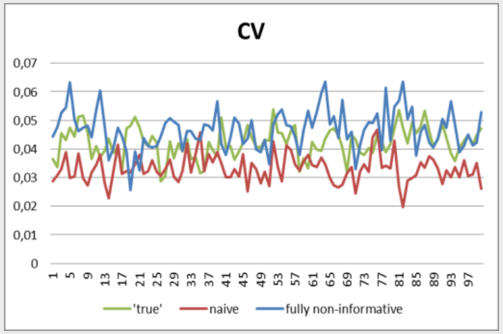

CP - α_0 - phase 1
M 30-59


CP - α_0 - phase 2
M 30-59


CP - α_1 - phase 2
M 30-59


Response rate
M 30-59


R-indicator


CV

- A non-informative prior has no clear impact in the first phase with one covariate
- Smaller differences with response rate when the variances are larger
- The R-indicator is smaller for non-informative priors with a larger variance

Simulation study - results

Conclusions

Misspecified prior distributions

- Larger impact in a smaller dataset

Non-informative vs informative

- The R-indicator is larger for non-informative priors with the same variance
- The R-indicator is smaller for non-informative priors with a large variance
- Smaller differences with true prior or response rate when the variance is larger

Sample size

- Impact on misspecified priors and posterior variance

Convergence

- Short burn-in period

Computation times

- Larger computation times with larger variances and misspecified priors

Future work/discussion

Future work

Priors

- Translation of expert knowledge and historic survey data to hyperparameters in prior distributions;
- Use of power priors to moderate impact of history;

Models

- Costs model;
- Dependency over phases (same mode in different phases);
- Inclusion of models for survey outcome variables and mode effects;