A Bayesian analysis of mixed-mode data collection

Lisette Bruin, Nino Mushkudiani, Barry Schouten

AAPOR, May 12 - 15, 2016, Austin

The Leverhulme Trust

Centraal Bureau voor de Statistiek

Summary

- Introduction;
- Bayesian analysis framework;
- Simulation study (goals, approach, results);
- Future work/discussion

Current case studies

Case studies at Statistics Netherlands:

- Health Survey (HS): Web-CAPI
- Labor Force Survey (LFS): Web-CATI-CAPI
- Travel survey (TS): Web-CATI-CAPI
- Survey on Income and Living Conditions (SILC): Web-CATI

TS will be redesigned in 2017, SILC started in a new design in 2016

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;
- To optimize (mixed-mode) survey design;

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;

Response propensities are split into contact and participation propensities for interviewer modes.

Functions of survey design parameters

We consider four functions of the design parameters:

- Response rate

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

- Total cost

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

- Coefficient of variation of propensities against relevant X

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

Functions of survey design parameters

Method effect

Outcome variable under strategy $s_{1,T}$

$$Y_{k,i}(s_{1,T}) = \frac{1}{\rho_i(s_{1,T})}\left(\rho_{1,i}(s_1)Y_{k,i}(s_1) + \sum_{t=2}^{T}\prod_{l=1}^{t-1}(1-\rho_{l,i}(s_{1,l}))\rho_{t,i}(s_{1,t})Y_{k,i}(s_t)\right)$$

Method effect relative to a benchmark BM

$$D_k(s_{1,T}; BM) = \left|\frac{\sum_{i=1}^{n} d_i\rho_i(s_{1,T})D_{k,i}(s_{1,T}; BM)}{\sum_{i=1}^{n} d_i\rho_i(s_{1,T})}\right|$$

with

$$D_{k,i}(s_{1,T}; BM) = Y_{k,i}(s_{1,T}) - Y_{k,i}(BM)$$

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. Assign prior distributions;
2. Derive likelihood functions;
3. Derive approximations to posterior distributions of design parameters using Gibbs samplers;
4. Derive approximations to posterior distributions of aggregate quality and cost measures (functions of design parameters).

Elicitation of parameters in prior distributions (hyperparameters):

- Expert knowledge
- Historic survey data

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

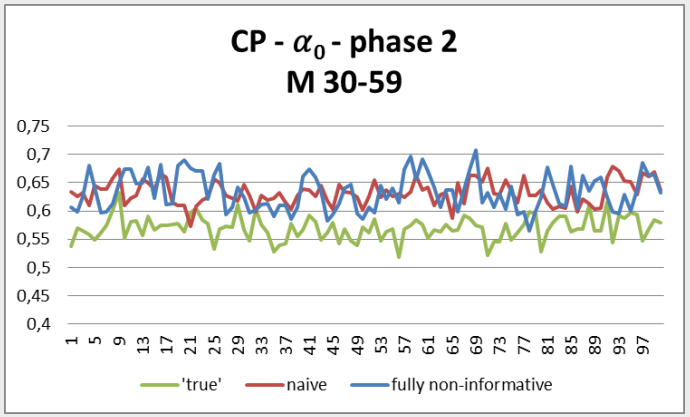
- Three phases: CAWI → CAPI → CAPI extended
- Simulation based on known parameters from the Health Survey

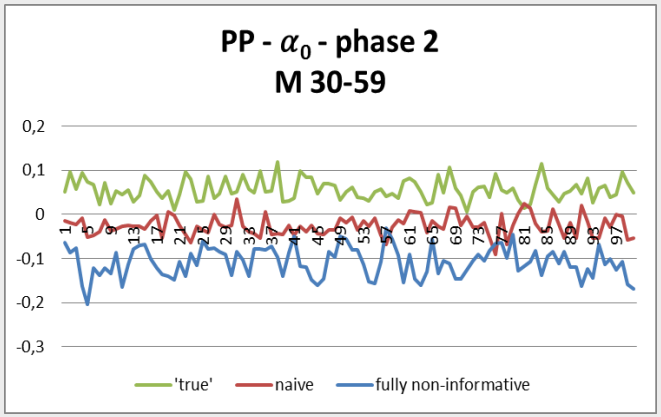
Prior specification

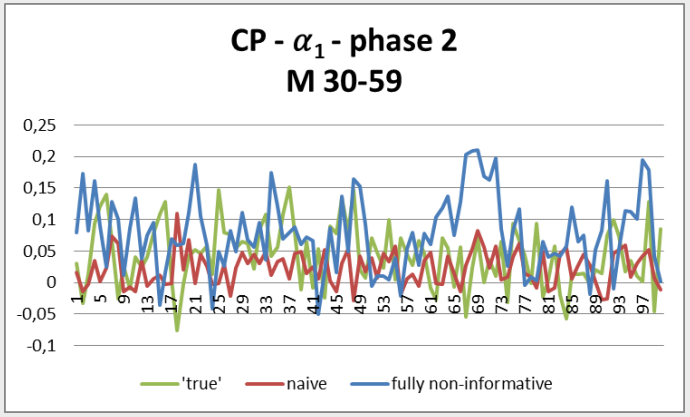
- **True:** Based on simulation model;
- **Naïve:** Prior distributions equal for all regression parameters;
- **Non-informative:** Like naïve but large variances

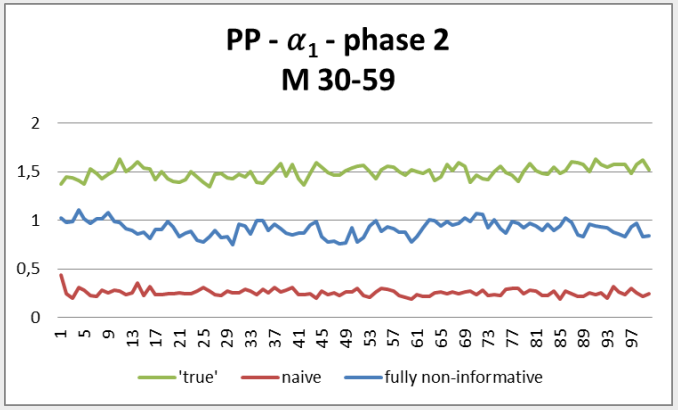
Simulation study – preliminary results

Gibbs sampler runs for phase 2 (CAPI) contact and participation equations



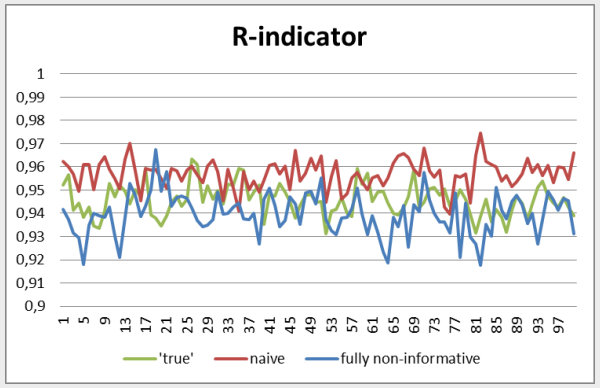


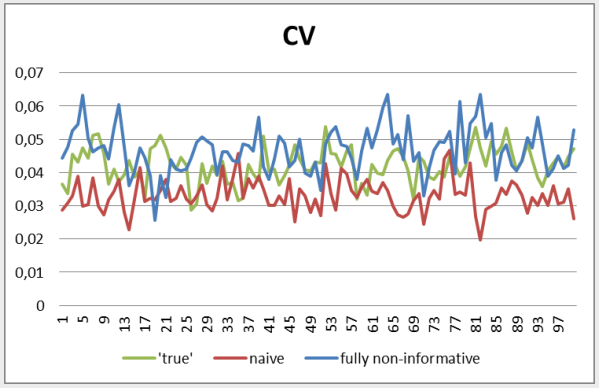




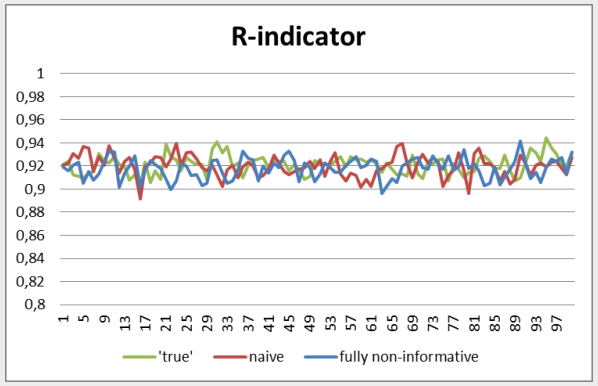
Simulation study – preliminary results

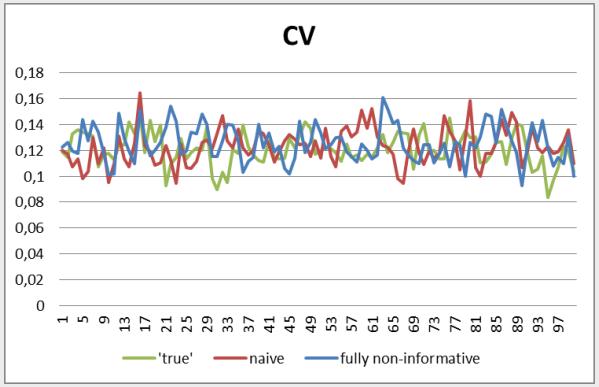
Gibbs sampler runs R and CV after phase 1





Gibbs sampler runs R and CV after phase 3





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;

Optimization

- Assess performance of Bayesian analysis
- Adaptation of strategies to quality and cost functions;

Contact: bstn@cbs.nl (Barry Schouten) or lbin@cbs.nl (Lisette Bruin)