Bayesian analysis within adaptive survey designs: Application to the Dutch Health Survey data

Barry Schouten

Statistics Netherlands, The Hague, The Netherlands – bstn@cbs.nl

Nino Mushkudiani*

Statistics Netherlands, The Hague, The Netherlands – nmsi@cbs.nl

Abstract: In this paper we apply a Bayesian framework for modelling the survey design parameters within the context of adaptive survey design to the Dutch Health Survey (DHS) data. The Bayesian framework is quite generic, however also new and complex. This framework has an advantage of making possible to incorporate prior knowledge or historical estimates into these models. In order to capture response or target variable trends we look at the DHS data of two years (2014 and 2015) with the fixed time window of three months and update the prior information by moving forward this window. We include the auxiliary and paradata and target variables of the DHS data in this framework. Our main goal is to define an optimal strategy allocation, in terms of quality and cost indicators, such as budget, response rate or measurement errors. For different strategies we derive overall quality and cost indicators using models defined through a Bayesian framework. This is a step forward to monitoring and adapting strategies during survey data collection based on prior knowledge.

1. Introduction

While response rates have been dropping gradually for last decades, it becomes urgent to move away from the traditional ways of data collection. Monitoring data and adapting survey strategies during data collection is a way to deal with the problem. This implies a strongly increasing interest in methodology for survey data collection monitoring, analysis, and intervention or adaptation; More specifically, into the direction of adaptive or responsive survey designs that adapt or tailor strategies and effort to known and relevant characteristics of sampled units from the target population, see Groves and Heeringa (2006), Wagner (2008) and Schouten, Calinescu and Luiten (2013). In order to adapt, accurate estimates of survey design parameters are not just needed at the overall population level, but also at the deeper level of population subgroups.

A natural approach to deduce inaccuracy of survey design parameters and to make analyses and design optimization more robust is to incorporate historic survey data and expert judgment through a Bayesian analysis. In this paper we apply a general model for survey design parameters and target variables. To these survey design parameters are then assigned prior distributions, which are updated and transformed to posterior distributions during data collection. We propose to use Gibbs samplers to obtain draws from the posterior distributions of response propensities, cost functions and target variables. We calculate quality and cost indicators and method effect based on these basic survey design parameters.

We carried out a simulation experiment on the Dutch Health Survey (DHS) data of two years (2014 and 2015). We want to including historic data through Bayesian analysis when deriving quality indicators and observe the trend in quality indicators in this period of time. We use the first three and six months of data as our historic data to define priors. For detecting a trend we update out prior by moving forward with the fixed time window of three months. We include the auxiliary variables (age, gender and income) and target variables (BMI and smoke) of the DHS data in this framework.

The paper is organised as follows: in the next section we introduce our notations and define the model. In Section 3 we describe Bayesian framework and introduce the quality indicators. In Section 4 we describe data and the simulation experiment and conclude with Discussion.

2. The model

First we introduce some notations. Let the survey design consist of a maximum of T phases that are labelled by $t = 1,2, \dots, T$. Define $\mathcal{S}_t$ as the collection of all possible actions in phase t and let s_t represent the action in phase t. For different phases, the collections of actions may be different. The action sets may contain $s_\emptyset$, which, if selected, implies that no attempt is made to obtain a response. We define the collection of survey strategies

$$\mathcal{S}_{1,T} := \{(s_1, \dots, s_T): s_t \in \mathcal{S}_t, t = 1,2, \dots, T\}$$

and let $s_{1,T} \in \mathcal{S}_{1,T}$ denote one possible strategy, i.e. sequence of actions. For a strategy $s_{1,T}$, we denote the actions in phase i til j by the vector $s_{i,j}$. For example if in phase 1 we carry out online data collection we will have that $\mathcal{S}_1 = \{CAWI\}$. If we would take two different actions for different groups of respondents, for example call one group and send a web questionnaire to the other we would have $\mathcal{S}_1 = \{CAWI, CATI\}$. In adaptive surveys, part of the design features may be implemented differently for different sample units, i.e. there is a set of strategies, see Groves and Heeringa (2006), Wagner (2008), Coffey, Reist and White (2013).

For a subject i, we let x_i be the vector of auxiliary variables that is linked from frame data, administrative data or paradata, x_i consists of the following entries

$$x_i = (x_{0,1,i}, \dots, x_{0,m_0,i}, \dots, x_{T,1,i}, \dots, x_{T,m_T,i})',$$

where $x_{0,i} = (x_{0,1,i}, \dots, x_{0,m_0,i})'$ contains the auxiliary variables available at the start of data collection, and $x_{t,i} = (x_{t,1,i}, \dots, x_{t,m_t,i})'$ are the auxiliary variables that are observed for the fielded sample units in phase t. In the optimization of the adaptive survey designs (ASD), actions in phase t can only be chosen based on $x_{0,i}$ to $x_{t-1,i}$.

The design of each survey has a range of features, e.g. advance letter, contact protocol, screener interview, number of phases, reminder protocol, use of incentive, mode of administration (web, telephone, face-to-face, mail), interviewer, refusal conversion procedure and type of questionnaire. The total of choices made for the design features (e.g. incentive, phases, first web mail then telephone interview) will define the data collection strategy or simply strategy.

On the other hand ASD either maximize a quality objective subject to cost constraints and other quality constraints or minimize a cost objective subject to quality constraints. The quality and cost constraints depend on the setting in which the survey is conducted. Three sets of survey design parameters suffice to compute most of the quality and cost constraints:

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

We first introduce basic models for response propensities and costs. Therefore, we break down these parameters into their basic components, like the contact and participation propensities. For these basic components we will, first, make some general assumptions. We assume that making contact, obtaining participation and the costs associated with an individual sample unit are independent of contact, participation and the costs of any other individual sample unit.

Here we define our model only for contact propensity, a part of response propensity. Models for the response propensity and cost functions can be defines similarly.
First introduce more notations:

- Let $\kappa_{t,i}(s_{1,t})$ be the propensity of a contact in phase t under strategy $s_{1,t}$ given that the unit did not respond in earlier phases and is eligible for follow-up. We assume that design features in subsequent phases have no impact on making contact. The outcome(s) of the previous phase(s) can be included in the auxiliary vector, when contact propensities are considered to be dependent on whether there was a noncontact or a refusal.

- $\lambda_{t,i}(s_{1,t})$ is the propensity of a participation in phase t of subject i under strategy $s_{1,t}$ given contact (and given that the unit did not respond in earlier phases and is eligible for follow-up).
- The response propensity in phase t of a subject i under strategy $s_{1,t}$, $\rho_{t,i}(s_{1,t})$, is:

$$\rho_{t,i}(s_{1,t}) = \kappa_{t,i}(s_{1,t}) \cdot \lambda_{t,i}(s_{1,t}).$$

When in subsequent phases all nonresponse receives a follow-up, then

$$\rho_i(s_{1,T}) = \kappa_{1,i}(s_1)\,\lambda_{1,i}(s_1) + \sum_{t=2}^{T}\left(\left(\prod_{l=1}^{t-1}\left(1-\kappa_{l,i}(s_{1,l})\,\lambda_{l,i}(s_{1,l})\right)\right)\kappa_{t,i}(s_{1,t})\,\lambda_{t,i}(s_{1,t})\right).$$

We model the propensities using a probit model, i.e. using a binomial link function. Each sample unit has a certain contactability represented as a latent variable $Z_{t,i}^{(\kappa)}(s_{1,t})$ and contact is obtained when this latent variable is larger than zero and $Z_{t,i}^{(\kappa)}(s_{1,t}) \sim N\left(\mu_{t,i}^{(\kappa)}(s_{1,t}), \sigma_{t,i}^{(\kappa)}(s_{1,t})\right)$, for some $\mu_{t,i}^{(\kappa)}(s_{1,t}), \sigma_{t,i}^{(\kappa)}(s_{1,t})$ so that

$$\kappa_{t,i}(s_{1,t}) = P\left(Z_{t,i}^{(\kappa)}(s_{1,t}) > 0\right).$$

Define $\beta_t^{(\kappa)}$to be the regression coefficients in phase t given that $s_{1,t}$ is applied to a unit i and X_t a matrix of auxiliary variables in phase *t*. The model could be written as

$$Z_{t,i}^{(\kappa)}(s_{1,t}) = X_t\beta_t^{(\kappa)} + \varepsilon_{t,i}^{(\kappa)},$$

where $\varepsilon_{t,i}^{(\kappa)} \sim N(0,1)$ is an error term for the uncertainty of contact of the subject.

This model has quite a few parameters. We want to simplify the model. To be able to include dynamic adaptive survey designs, we need to include paradata. However to keep the model simple, we assume that there is just one phase, say t_1, in which paradata is collected. Up to phase t_1 only the auxiliary variables in $x_{0,i}$ can be used to model the propensities. After phase t_1, the auxiliary variables obtained in phase t_1 can also be included in the model. Second, we consider the dependence on past actions. It is unrealistic to assume there is no such dependence in most settings. Past actions could be included by introducing a fixed or random effect per possible history. We add the history as a random effect here. Third, since we suggest to add a dependence on the history of actions as a random effect, the regression coefficients become necessarily dependent on the phase and not on the past. The model becomes

$$Z_{t,i}^{(\kappa)}(s_{1,t}) = \begin{cases} \beta_{t,0}^{(\kappa)}(s_t)x_{0,i} + \varepsilon_{t,i}^{(\kappa)} + \delta_t^{(\kappa)}(s_{1,t-1}), & t \le t_1, \\ \beta_{t,0}^{(\kappa)}(s_t)x_{0,i} + \beta_{t,1}^{(\kappa)}(s_t)x_{t_1,i} + \varepsilon_{t,i}^{(\kappa)} + \delta_t^{(\kappa)}(s_{1,t-1}), & t > t_1, \end{cases} \tag{1}$$

where $\delta_t^{(\kappa)}(s_{1,t-1})$ is a random effect.

For participation propensity $\lambda_{t,i}(s_{1,t})$ we will have the same model:

$$Z_{t,i}^{(\lambda)}(s_{1,t}) = \begin{cases} \beta_{t,0}^{(\lambda)}(s_t)x_{0,i} + \varepsilon_{t,i}^{(\lambda)} + \delta_t^{(\lambda)}(s_{1,t-1}), & t \le t_1, \\ \beta_{t,0}^{(\lambda)}(s_t)x_{0,i} + \beta_{t,1}^{(\lambda)}(s_t)x_{t_1,i} + \varepsilon_{t,i}^{(\kappa)} + \delta_t^{(\lambda)}(s_{1,t-1}), & t > t_1, \end{cases} \tag{2}$$

Next we define models for the costs. In general, the costs per sample depend on the phase, the sample unit and the strategy. For notational convenience here we drop the subscript t for phase. We define:

- $C_{t,i}^{(\kappa)}(s_{1,t})$ as the cost to make a contact attempt (visit or call) with a sample unit *i* in phase *t*, following strategy $s_{1,t} \in \mathcal{S}_{1,t}$;
- $C_{t,i}^{(\lambda)}(s_{1,t})$ as the cost for the response, of a sample unit *i* in phase *t*. These are defined in a similar way as the contact cost.

For some actions, these functions may be identical to zero, e.g. a response cost to a web survey. The cost parameters $C_i^{(\kappa)}(s_{1,t})$ can be written using these components and the contact and participation propensities

$$C_i^{(\kappa)}(s_{1,t}) \; = C_{0,i}^{(\kappa)}(s_{1,t}) + \kappa_{1,i}(s_1) C_{1,i}^{(\kappa)}(s_{1,t}) + \; \kappa_{1,i}(s_1) * \lambda_{1,i}(s_1) \; C_{1,i}^{(\lambda)}(s_{1,t}) +$$

$$+ \sum_{t=2}^{T} \left(\left(\prod_{l=1}^{t-1} \left(1 - \kappa_{l,i}(s_{1,l}) \lambda_{l,i}(s_{1,l})\right)\right) \left(\kappa_{t,i}(s_1) C_{t,i}^{(\kappa)}(s_{1,t}) + \kappa_{t,i}(s_{1,t}) \lambda_{t,i}(s_{1,t}) \; C_{t,i}^{(\lambda)}(s_{1,t})\right)\right).$$

In this paper, we make the simplification that cost functions do not depend on the phase and history of actions but only on the current action and hence cost functions do not depend on the phase and design features in previous phases, but only on the current phase and design features. We can make this simplification since we can assume that these dependences are negligible. For example we assume that breaking off in CAWI and the traveling costs of an interviewer in CAPI are independent. Costs are continuous variables and we use linear models for the costs functions

$$C_i^{(\kappa)}(s_{1,t}) = \gamma_i^{(\kappa)}(s) x_i + \zeta_i^{(\kappa)}(s), \quad C_i^{(\lambda)}(s_{1,t}) = \gamma_i^{(\lambda)}(s) x_i + \zeta_i^{(\lambda)}(s), \; s \in \mathcal{S} \tag{3}$$

where $\gamma_i^{(\kappa)}(s)$ and $\gamma_i^{(\lambda)}(s)$ are regression parameters allowing for interaction between the current action s and the auxiliary vector x_i and $\zeta_i^{(\kappa)}(s)$ and $\zeta_i^{(\lambda)}(s)$ are error terms that again allow for an interaction with the current action. The error terms are modelled as independent normal

$$\zeta_i^{(\kappa)}(s) \sim N\left(0, \sigma_{(\kappa)}^2(s)\right) \text{ and } \zeta_i^{(\lambda)}(s) \sim N\left(0, \sigma_{(\lambda)}^2(s)\right).$$

Suppose we have K target variables. Denote by $Y_{k,i}(s_{1,T})$ the outcome of the survey target variable Y_k, $(k = 1, \ldots, K)$ for population unit i when strategy $s_{1,T}$ is applied. We assume that the outcome of the survey variable depends only on the current action and not on the history of actions and not on the phase, i.e. a form of measurement equivalence, so that $Y_{t,k,i}(s_{1,t}) = Y_{k,i}(s_t)$.

for all possible histories and all phases. Now, it holds that

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

and, hence, the outcome for the strategy is modelled as a weighted mix of the outcomes under the possible actions.

The outcomes for the actions are modelled as

$$Y_{k,i}(s_t) = \theta_0(s_t) x_{0,i} + \theta_1(s_t) x_{,i} + \varepsilon_{Y,i}(s_t), \tag{4}$$

for continuous variables, where the error terms are modelled as independent normal

$\varepsilon_{Y,i}(s_t) \sim N\left(0, \sigma_{(Y)}^2(s)\right)$, and as

$$\tilde{Y}_{k,i}(s_t) = \theta_0(s_t) x_{0,i} + \theta_1(s_t) x_{1,i} + \varepsilon_{Y,i}(s_t), \tag{5}$$

for dichotomous variables, where $\tilde{Y}_{k,i}(s_t)$ is a latent variable and $Y_{k,i}(s_t) = 1$ when $\tilde{Y}_{k,i}(s_t) \geq 0$, and $Y_{k,i}(s_t) = 0$ otherwise and $\varepsilon_{Y,i}(s_t) \sim N(0,1)$. As usual, the regression parameters get assigned prior distributions. The type of distribution depends on the measurement level, but we follow the approach for response propensities and cost functions; we employ normal priors for regression slope parameters and inverse Gamma priors for dispersion parameters.

The observed data are extended by the matrix of survey variables over all sample units where entry $y_{k,i}$ is missing when unit i did not respond.

3. Bayesian analysis

The analysis become Bayesian by assigning prior distributions to the regression coefficients and random effects in (1) – (5). Our aim is to derive the posterior distributions of the individual response propensities $\rho_i(s_{1,T})$, the individual cost parameters $C_i(s_{1,T})$ and the target variables $Y_{k,i}(s_{1,T})$ per strategy given observed data. These overall parameters are, in general, complex functions of the underlying survey design parameters per phase. We derived expressions for the posterior distributions of the regression coefficients and random effects when it was possible, otherwise derived these numerical approximations and applied Markov Chain Monte Carlo methods to generate draws from the posterior distributions.

Below we use $\rho(s_{1,T})$ and $C(s_{1,T})$ for the vector of response propensities and cost parameters over all sample units for a particular strategy. In the same fashion, we use u_t, y, $c^{(\kappa)}$, $c^{(\lambda)}$ and x to denote the vectors of outcomes, realized costs components and auxiliary variables over sample units. Note that x may in fact be a matrix, when the auxiliary variables are a vector by themselves. With $\{s_{1,T}^i\}$ we denote the vector of used strategies for all sample units. To shorten expressions, we use $\beta^{(\kappa)}, \beta^{(\lambda)}, \delta^{(\kappa)}, \delta^{(\lambda)}, \gamma^{(\kappa)}, \gamma^{(\lambda)}, \sigma^2, \theta$ for the vectors of regression slope parameters, random effects and regression dispersion parameters over phases and actions, but elaborate when needed. Here p stands to express joint and marginal density functions; we omit the reference to the random variables to which they apply and ignore differences between discrete and continuous probability distributions. Finally, in the density functions, we omit the dependence on the hyperparameters. A straightforward solution is to perform a Gibbs sampler to the joint density of the regression parameters $\beta^{(\kappa)}, \beta^{(\lambda)}, \delta^{(\kappa)}, \delta^{(\lambda)}, \gamma^{(\kappa)}, \gamma^{(\lambda)}, \sigma^2, \theta$:

$$p\left(\beta^{(\kappa)}, \beta^{(\lambda)}, \delta^{(\kappa)}, \delta^{(\lambda)}, \gamma^{(\kappa)}, \gamma^{(\lambda)}, \sigma^2, \theta \middle| u_t, c^{(\kappa)}, c^{(\lambda)}, x, y, \{s_{1,T}^i\}\right).$$

A Gibbs sampler for this density function requires repeated draws from the conditional densities of each regression parameter given the observed data and the other regression parameters, the so-called full conditionals. Literature provides a range of options to sample from these conditional distributions, see Albert and Chib (1993) and Gelman et al (2003).
The Gibbs sampler has the following steps:

1. Set the random effects for the contact and participation equations to zero, $\delta_t^{(\kappa)} = 0$ and $\delta_t^{(\lambda)} = 0$, fit regression models to all contact, participation, cost and target variable equations and use the resulting estimated parameter values as starting values for the regression parameters $\beta^{(\kappa)}, \beta^{(\lambda)}, \delta^{(\kappa)}, \delta^{(\lambda)}, \gamma^{(\kappa)}, \gamma^{(\lambda)}, \sigma^2, \theta$;
2. For each unit (i) in each phase (t), sample the latent variables $Z_{t,i}^{(\kappa)}$ and $Z_{t,i}^{(\lambda)}$ from $p(Z_{t,i}^{(\kappa)} | \beta^{(\kappa)}, u_{t,i}, x_i, \{s_{1,T}^i\})$ and $p\left(Z_{t,i}^{(\lambda)} \middle| \beta^{(\lambda)}, u_{t,i}, x_i, \{s_{1,T}^i\}\right)$;
3. For each phase, sample the contact slope parameters $\beta^{(\kappa)}$ from $p(\beta^{(\kappa)} | Z_{t,i}^{(\kappa)}, x_i, \{s_{1,T}^i\})$;
4. Sample the random effects $\delta_t^{(\kappa)}$from $p(\delta_t^{(\kappa)} | Z_{t,i}^{(\kappa)}, \beta^{(\kappa)}, x_i, \{s_{1,T}^i\})$;
5. For each phase, sample the participation slope parameters $\beta^{(\lambda)}$ from $p\left(\beta^{(\lambda)} \middle| Z_{t,i}^{(\lambda)}, x_i, \{s_{1,T}^i\}\right)$;
6. Sample the random effects $\delta_t^{(\lambda)}$from $p\left(\delta_t^{(\lambda)} \middle| Z_{t,i}^{(\lambda)}, \beta^{(\lambda)}, x_i, \{s_{1,T}^i\}\right)$;
7. For the cost components, sample the variance parameters σ^2 from $p(\sigma^2 | \gamma, c^{(\kappa)}, c^{(\lambda)}, x, \{s_{1,T}^i\})$;
8. For the cost components sample the slope parameters γ from $p(\gamma | \sigma^2, c^{(\kappa)}, c^{(\lambda)}, x, \{s_{1,T}^i\})$;
9. For the categorical target variables, for each unit (i) in each phase (t), sample the latent variables $\tilde{Y}_{t,i}$ from $p(\tilde{Y}_{t,i} | \theta, Y_{t,i}, x_i, \{s_{1,T}^i\})$;
10. For each phase, sample the slope parameters θ from $p(\theta | \tilde{Y}_{t,i}, x_i, \{s_{1,T}^i\})$;
11. For the continuous target variables, sample the variance parameters $\sigma_{(Y)}^2$ from $p(\sigma_{(Y)}^2 | \theta, Y_{t,i}, x_i, \{s_{1,T}^i\})$;
12. For the continuous target variables, sample the slope parameters θ from $p(\theta | \sigma_{(Y)}^2, Y_{t,i}, x_i, u_{t,i}, x_i, \{s_{1,T}^i\})$;
13. Return to step 2 for another round of defining posteriors based on a new prior ;

In order to carry out the data augmentation, we did not make use of standard libraries in R but programmed the Gibbs sampler from scratch.

In the monitoring and optimization of data collection, the focus is on functions of the design parameters that correspond to overall quality or cost objectives. We consider three such functions here for the sake of brevity, the response rate, the total costs and the coefficient of variation of the response propensities; the analysis of other functions can often be done in an analogous way.

Let d_i represent the design or inclusion weight for sample unit i, $i = 1,2,\dots,n$. The response rate, *RR*, for strategy $s_{1,T}$ can be written as

$$RR(s_{1,T}) = \frac{1}{N}\sum_{i=1}^{n} d_i\rho_i\,(s_{1,T}), \tag{6}$$

the total costs, or required budget, B, associated with s_1^T are

$$B(s_{1,T}) = \sum_{i=1}^{n} c_i\,(s_{1,T}), \tag{7}$$

and the coefficient of variation, *CV,* is

$$CV(X, s_{1,T}) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{n} d_i(\rho_i(s_{1,T}) - RR(s_{1,T}))^2}}{RR(s_{1,T})}. \tag{8}$$

For the *CV*, we explicitly denote the dependence on the covariate vector X; for any other choice of auxiliary variables it will, generally, attain a different value. The response rate and total costs do not depend on the choice of X.

The another quality indicator, the adjusted absolute method effect is based on a benchmark strategy (BM). First define the expected measurement difference between strategy $s_{1,T}$ and the benchmark strategy:

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

Then the adjusted absolute method effect is

$$D_k(s_{1,T}; BM) = \left|\frac{\sum_{i=1}^{n} d_i\rho_i(s_{1,T})D_{k,i}(s_{1,T}; BM)}{\sum_{i=1}^{n} d_i\rho_i(s_{1,T})}\right|. \tag{9}$$

We will refer to (9) as simply the method effect of strategy $s_{1,T}$.

There are two important difference in the estimation with response propensity and cost design parameters. First, a sample unit may not respond and the outcome of survey variables may be unavailable. This means that the outcome posterior distribution for a sample unit may be based on the outcomes of similar sample units that did respond. Second, per sample unit at most one outcome will be observed. We assume that the sample is randomly allocated to different designs, so that all outcomes can be estimated. This implies that at least one outcome in (9) must be estimated from similar sample units.

Obviously, the prior and posterior distributions for these functions are determined by the prior and posterior distributions of the components of the response propensities, cost functions and target variables. They have even more complex forms than the individual response propensities, cost parameters and target variables. However, they can again be approximated as a by-product of the Gibbs sampler. For every draw of the individual response propensities and cost parameters, we compute (6) – (9).

4. Health Survey Data

In the simulation study, we investigate the impact of prior distribution specification and of survey sample size on the posterior distributions and investigate how much we profit from historic knowledge. We also observe the

trend in quality indicators in this period of time. We include historic data as a prior information in our Bayesian model. We carry out the simulation study on the Dutch Health Survey.
We have monthly data of DHS available to us from March 2014 till March 2016. There are large number of background variables available, which we obtained before or after data collection from the administrative sources and linked to the survey data. For all respondents and non-respondents we have these background information. To keep the number of parameters in the Gibbs sampler confined, we include three covariates: Age, Gender and Income:

- Gender: Male, Female;
- Age, 3 categories: [0, 30), [30, 60), [60, …);
- Income, 6 categories:[0, 1000), [1000, 2000), [2000, 3000), [3000, 4000), [4000, 5000), [5000, …).

After the first phase of data collection, paradata variable of web interview Brake-off is available. In this experiment this variable is not included. We will extend our model with this paradata variable in the future experiments.

We have two variables defining response; these are contact and participation. In phase 1 online data collection was carried out, the invitation letter was sent to all participant's addresses. Here participation is equal to 1 for everyone and we only consider contact that is defined by the response. For phase 2 and 3 we have face to face interviews and we define contact and participation. Phase 2 stops after 3 contact attempts. Phase 3 includes 4 and more contacts.

For each phase we also calculated costs made for each participant. Based on the information obtained from the data collection department we can calculate cost of a web participation and refusals. If the participant called to refuse, then the telephone costs are included. For face to face interviews we calculated costs based on number of visit, average travel distance of the interviewer, time duration of the interview, etc. As a result we can calculate cost per respondent per phase and the aggregated costs per phase for contact and participation.

As we mentioned above we consider two ways to define priors for the Gibbs sampler. We can define priors based on the expert opinion of subject matter specialists or use historic data. Here we first define priors using historic data. We consider data of the first three and six months from March till May and from March till August of 2014 and derive priors based on these data. To define fair priors we bootstrapped 1000 subsets of this prior – data, each subset of size 3000. As we work by phase we divide these bootstrapped data sets by phase as well. So for example, data for phase 2 will not include respondents of phase 1. In this way we consider only participants for each phase. For each phase we then fit the probit and linear models as defined in (1)-(5). That way we obtained 1000 estimates for regression coefficients for these models. Next we derive estimates of the hyperparameters: mean, standard error and covariance for these coefficients. These hyperparameters are the priors for the Gibbs sampler.
We consider three different size of data: a month, three months and one year. Using these priors and data we then apply the Gibbs sampler and obtain posterior distribution for the parameters.
Next step is to move the time window of a three months forward by one month. We derive priors as above for data of April – June 2014 and apply it using Gibbs sampler on a data of July 2014 to define the posterior distributions. Based on these posterior distributions we calculate quality indicators defined in (6) to (9).

5. Discussion

We apply a Bayesian model for survey design parameters related to response and costs. This model is fit for multiple data collection phases, different types of auxiliary data, multiple nonresponse outcomes and dependence on previous actions. The Gibbs sampler provides estimates for the posterior distributions of the contact and participation propensities and the costs per sample unit. From the runs of the Gibbs sampler, also the posterior distributions for overarching quality indicators, like the response rate or coefficient of variation of the response propensities, and cost indicators can easily be derived.

References

Albert, J.H., Chib, S. (1993), Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, 88, 669 – 680.

Coffey, S., Reist, B., White, M. (2013). Monitoring Methods for Adaptive Design in the National Survey of College Graduates, In JSM Proceedings, Survey Methods Research Section, Alexandria, VA: American Statistical Association, 3085-3099.

Gelman, A. (2006), Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, 1 (3), 515 – 534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. London: Chapman and Hall, second edition.

Groves, R.M., Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs, *Journal of the Royal Statistical Society: Series A*, 169, 439 – 457.

Kreuter, F. (2013), *Improving Surveys with Paradata. Analytic Uses of Process Informaton*, Edited monograph, John Wiley and Sons, Hoboken, New Jersey, USA.

Schouten, B., Calinescu, M., Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39 (1), 29 – 58.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias, PhD thesis, University of Michigan, USA.