Analyse bayésienne des paramètres du plan de sondage

Lisette Bruin, Nino Mushkudiani, Barry Schouten

Symposium de Statistique Canada, du 22 au 24 mars 2016, Ottawa

The Leverhulme Trust

Centraal Bureau voor de Statistiek

Résumé

- Introduction
  - Plan de sondage adaptatif
  - BADEN
  - Objectifs
- Plan (définitions, notations, modèle)
- Analyse bayésienne (approche générale, distributions a priori et a posteriori)
- Étude de simulation (objectifs, approche, résultats)
- Travaux futurs/discussion

BADEN

Bayesian **A**daptive Survey **DE**sign **N**etwork (Réseau du plan de sondage adaptatif bayésien)

Réseau international dédié aux stratégies de collecte de données ciblées employant des données d'enquête historiques

De janvier 2015 à décembre 2017, financé par The Leverhulme Trust

- **Établissements participants :**
  - Universités de Manchester (CP), du Michigan et de Southampton
  - Bureau de la statistique des Pays-Bas, RTI International, Statistique Suède, Bureau du recensement des É.-U.
- **Objectifs :**
  - Ajouter des connaissances spécialisées et des données d'enquête historiques lors de la surveillance et de l'analyse (phase 1)
  - Ajustement/adaptation (phase 2)

Plan de sondage adaptatif

Les plans de sondage adaptatifs différencient les caractéristiques du plan de sondage pour différents sous-groupes de population en fonction de données auxiliaires sur l'échantillon obtenues à partir de données de la base de sondage, de données du registre ou de paradonnées.

Au lieu d'une seule stratégie (uniforme), il est possible d'envisager de multiples stratégies.

Pourquoi des plans de sondage adaptatifs?

- **Réponse :** Les répondants ont différentes préférences concernant les communications et les interviews, c.-à-d. qu'ils réagissent différemment aux différentes stratégies de collecte des données.
- **Coûts :** Différentes stratégies sont associées à différents coûts par personne.

Objectifs

- Configurer un modèle général pour les paramètres du plan de sondage
- Présenter une analyse bayésienne des paramètres du plan de sondage
- Présenter une analyse bayésienne des indicateurs de coût et de qualité en fonction des paramètres du plan de sondage

Paramètres du plan de sondage

Trois ensembles de paramètres du plan de sondage suffisent pour calculer la plupart des contraintes de qualité et de coût :

- $\rho_i(s_{1,T})$: Propensions à répondre par unité et par stratégie
- $C_i(s_{1,T})$: Coûts attendus par unité d'échantillonnage et par stratégie
- $D_i(s_{1,T})$: Effets de mode ajustés par unité et par stratégie

Nous examinons uniquement l'erreur de non-réponse et laissons de côté les modes d'effet ajustés, qui pourront faire l'objet de futures études.

Fonctions des paramètres du plan de sondage

Nous examinons trois fonctions des paramètres du plan :

- le taux de réponse

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

- le coût total

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

- le coefficient de variation

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

Définitions

- Actions
  - Choix des caractéristiques du plan (nombre d'appels, utilisation d'un incitatif, mode d'interview)
- Stratégie
  - Nombre total de caractéristiques choisies pour le plan, désigné par $s_{1,T}$
- Phase
  - T phases du plan de sondage $t = 1,2, \ldots, T$
- Données auxiliaires
  - Vecteur x_i lié à partir des données du plan, des données administratives ($x_{0,i}$) ou des paradonnées ($x_{t,i}$)

Si $x_i = x_{0,i}$, l'analyse du plan de sondage (APS) est **statique**. Si, pour certains t, on utilise $x_{t,i}$ pour choisir les actions au cours d'une phase subséquente, l'APS est **dynamique**.

Modélisation des paramètres du plan de sondage

But :

Modèle simple, mais suffisamment général pour inclure toutes les caractéristiques potentielles :

- plus d'une phase
- dynamique
- dépendance à l'égard de l'historique des actions
- cas de non-réponse non admissibles aux fins du suivi

Modélisation :

1. Décomposition des paramètres du modèle en leurs composantes principales
2. Modèles linéaires généraux qui relient ces composantes aux variables auxiliaires disponibles
3. Hypothèse selon laquelle le coût, la méthode de contact et la participation d'une unité d'échantillonnage sont indépendants de ceux des autres unités

Analyse bayésienne

Approche générale :

1. Supposer que les paramètres sont indépendants
2. Attribuer des distributions a priori
3. Calculer les fonctions de vraisemblance
4. Calculer les distributions a posteriori des paramètres du plan
5. Calculer les distributions a posteriori des mesures agrégées de qualité et de coût (fonctions des paramètres du plan)

Paramètres de la distribution a priori (hyperparamètres) :

- Connaissances spécialisées
- Données d'enquête historiques

Analyse bayésienne

Distributions a posteriori

Distributions a posteriori conjointes d'intérêt :

1. Propensions à répondre et coûts individuels – paramètres d'optimisation
2. Indicateurs globaux de qualité et de coût – analyse de surveillance

Données observées requises :

- Coûts réalisés
- Résultats de la participation
- Stratégies employées
- Données auxiliaires

Analyse bayésienne

Distributions a posteriori

Aucune forme explicite : Les distributions a posteriori des propensions à répondre et des coûts (et les indicateurs globaux de qualité et de coût) n'ont pas de formes explicites.

Proposition : Prélever les échantillons MCMC à partir des distributions a posteriori des paramètres de régression dans les modèles de contact, de participation et de coût.

Avantage : Des distributions a posteriori des indicateurs globaux de qualité et de coût découlent directement des échantillons.

Choix évident : Un échantillonneur de Gibbs effectuera des prélèvements distincts pour chaque paramètre (certaines distributions conditionnelles n'ont toujours pas de formes explicites).

Étude de simulation

Objectifs

Analyser l'incidence :

- des distributions a priori incorrectement spécifiées
- de la dispersion des distributions a priori (non informative et informative)
- de la taille de l'échantillon

Examiner également :

- les propriétés de convergence et les temps de calcul

Étude de simulation

Simulation

- Trois phases : IWAO → IPAO3 → IPAO3+
- Simulation fondée sur des paramètres connus de l'Enquête sur la santé

Décomposition

- Réponse par phase : $\rho_{t,i}(s_{1,t}) = \kappa_{t,i}(s_{1,t}) \cdot \lambda_{t,i}(s_{1,t})$
- Coûts par phase : $C_{t,i}(s_{1,t}) = C_{0,t,i}(s_{1,t}) + C_{R,t,i}(s_{1,t}) + C_{NR,t,i}(s_{1,t})$

Modèles

- Contact et participation : régression probit pour l'âge, le sexe et l'indicateur 0-1 des interruptions sur le Web
- Coûts : régression linéaire pour l'âge et le sexe

Étude de simulation

Distributions a priori (hyperpriors):

- Gamma inverse : paramètres de variance dans les termes d'erreur des fonctions de coût
- Distribution normale : tous les autres paramètres de régression

Distributions a posteriori :

- Approximé au moyen d'un échantillonneur de Gibbs avec augmentation des données

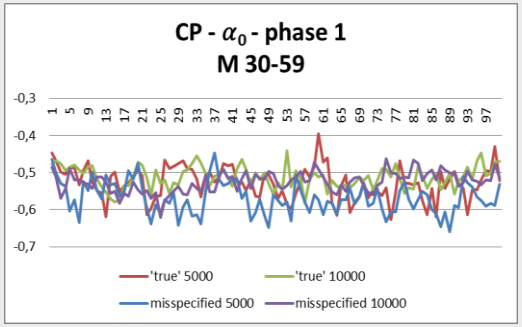
Étude de simulation

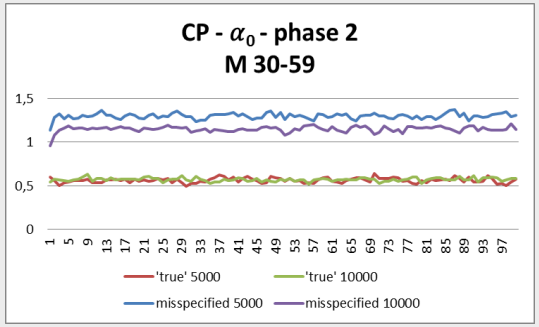
Choisi a priori:

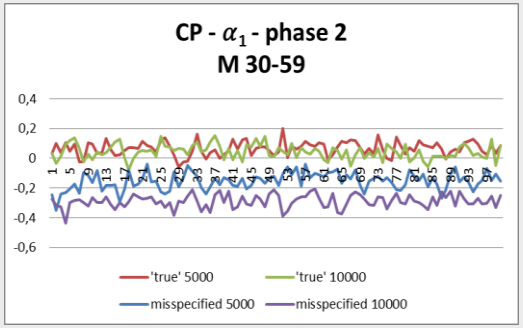
- **'Vrai';**
  Calculé et estimé par régression à partir des taux de réponse originaux
- **Incorrectement spécifié;**
  'Vrai' a priori avec une moyenne de α_0 multiplié par 3
- **Non-informative;**
  Mêmes écart-types que la 'vrai' a priori, mais de moyennes également distribuées
- **Non-informative avec une variance plus grande;**
  Écart-type multiplié par $\sqrt{10}$

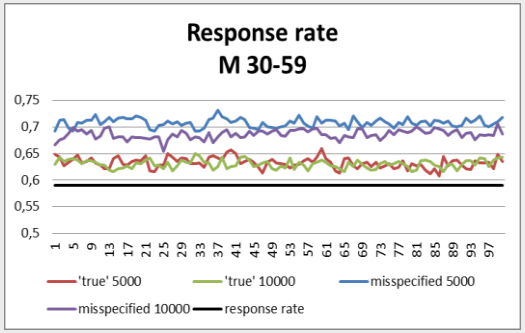
Simulation study - results

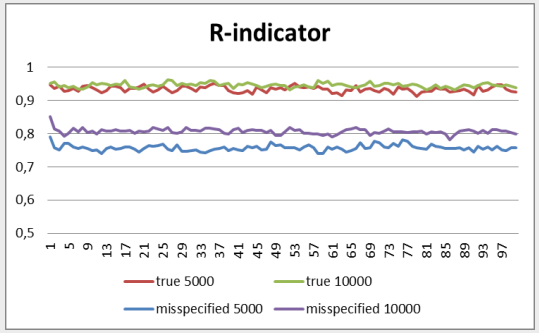
Vrai a priori vs incorrectement spécifié a priori (μ_{α_0} multiplié par 3)

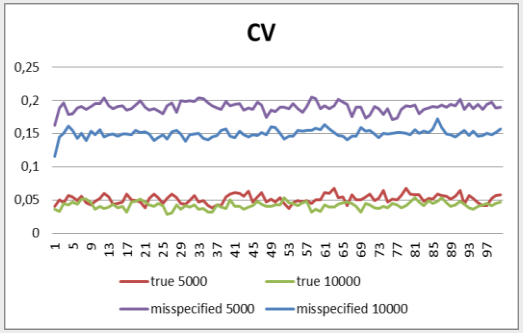








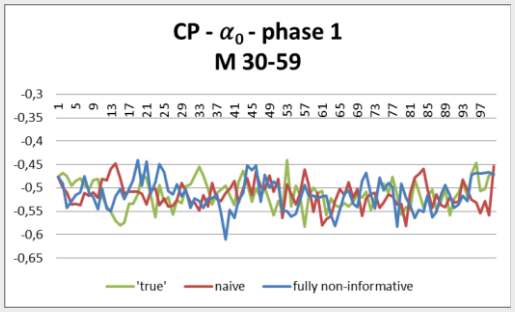


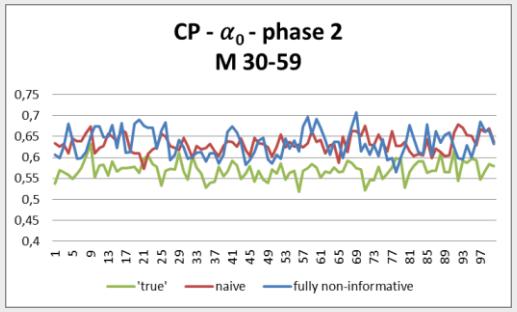


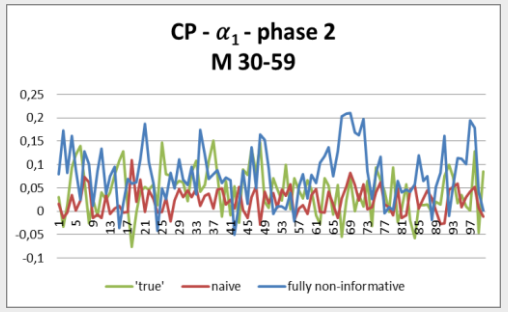
- Incorrectement spécifié a priori n'a pas d'impact à la première phase avec une covariable
- L'impact est plus grand sur les petits ensembles de données
- Dans les petits ensembles de données, les variances des distributions a postérion sont plus grandes

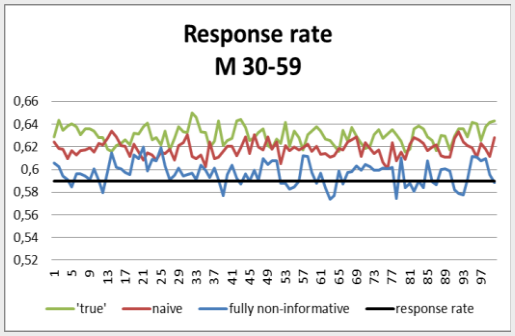
Simulation study - results

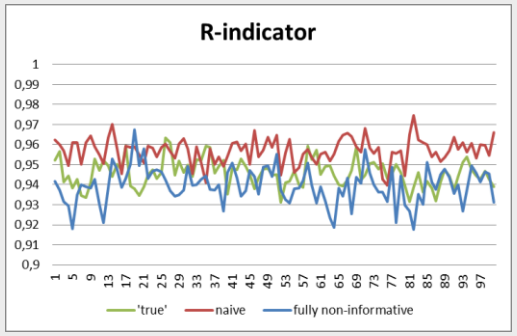
Vrai a priori vs non-informative a priori

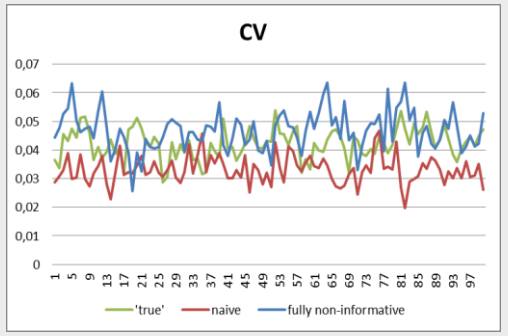












- Non-informative a priori n'a pas d'impact évident à la première phase avec une covariable
- Plus petites différences avec les taux de réponse quand les variances sont plus grandes
- L'indicateur R est plus petit pour les non-informative a priori avec variance plus grande

Étude de simulation - résultats

Conclusions

Distributions incorrectement spécifiées a priori

- Plus grand impact sur les plus petits ensembles de données

Non-informative vs informative

- L'indicateur R est plus grand pour les non-informatives a priori avec la même variance
- L'indicateur R est plus petit pour les non-informatives a priori avec une grande variance
- Plus petites différences si vrai a priori quand la variance est plus grande.

Taille d'échantillon

- Impact sur la variance si incorrectement spécifiés a priori ou a posteriori

Convergence

- Courte période de rodage

Temps de calcul

- Temps de calcul plus long lorsque les variances sont plus grandes et incorrectement spécifiées a priori

Travaux futurs/discussion

Travaux futurs

A priori

- Traduction de la connaissance spécialisée et des données d'enquêtes historiques vers les hyperparamètres des distributions a priori;
- Utilisation de puissance a priori pour modérer l'impact des antécédants;

Modèles

- Coûts des modèles;
- Dépendance entre les phases (même mode dans différentes phases);
- Inclusion de modèles pour les variables de résultats d'enquête et les effets de mode;