Test of adaptive survey design towards a Bayesian perspective in a longitudinal survey

Peter Lundquist and Anton Johansson (Statistics Sweden)
Gabriele B. Durrant (University of Southampton)

JSM session Bayesian adaptive survey designs, July 30, 2017

Introduction and background

- Large survey resources are being spent on making unproductive calls e.g. contact attempts that do not lead to an interview.
- Unstable data collection and unclear collection strategy
- We have used work by Durrant et al. (2013, 2015) who assess the prediction of nonresponse models using paradata from previous and current wave
- The ambition is to find a new data collection strategy for the Swedish Labour Force Survey

Data used in our case study

The Swedish Labour Force Survey (LFS)

- Longitudinal survey with 8 waves
- "A new sample" every week
- Data collection mode: telephone only mode
- Two interviewer groups: field and call-center

- Data used: LFS in January 2016
  - Initial sample size 5,164; Week 4 sample

- Can we find a more stable (and cost reducing) data collection strategy?

Data collection in LFS

Approximately 100 Field + 100 call center interviewers sign up for the following shifts:

Shift/Day	Mon-Thu	Friday	Saturday	Sunday
09:00-12:59				
13:00-16:59				
17:00-18:59				
19:00-21:59		Contact by agreement		

- We assume that the resources could be better allocated to the different subsamples (Week1-Week4 in January 2016)
- In Choudhry et al. (2011) the **workload** was optimized to minimize the data collection costs
- We will initially use the **workload** for Week4 and try to reduce the number of unproductive calls

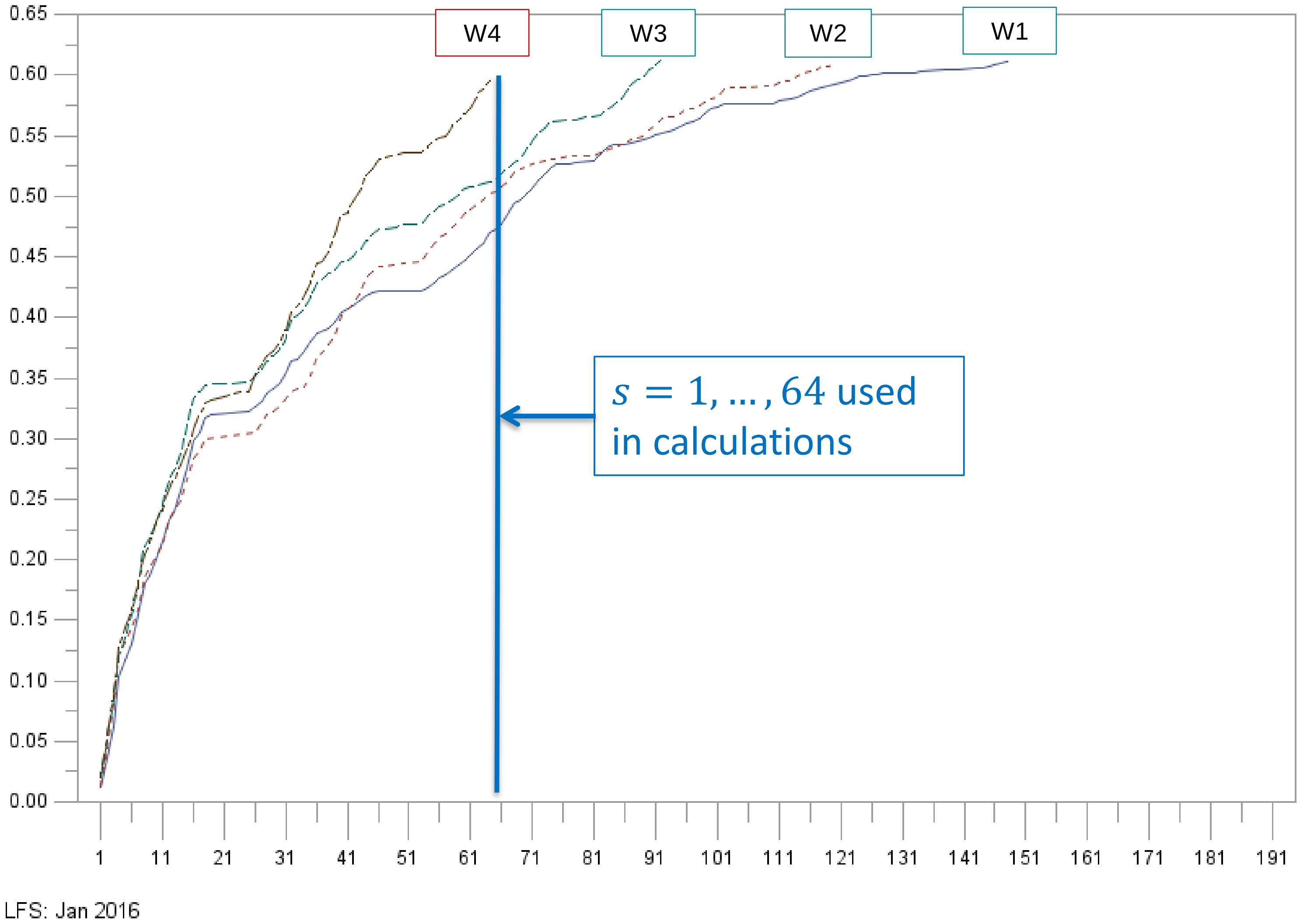
LFS – fieldwork January 2016
Number of call attempts

Week	1	2	3	4	5	6	(1)+(2)
W1	(1): 17,896		(2): 12,412				30,308
W2		(1): 20,680		(2): 9,767			30,447
W3			(1): 20,813		(2): 8,219		29,032
W4				(1): 28,992			28,992
Fieldday	1	8	15	22	29	37	

(1) Primary fieldwork 16 days (equal all samples W1-W4)

(2) Extended fieldwork

Response progression over time-slots for 4 LFS-weeks



Models for Phase 1+2 and 3

Logistic regression (binary), two different models

Model 1
Phase 1+2 time slots: $s = 1,\ 2, \ldots, 56$ (two weeks)

Model 2

Phase 3 time slots: $s = 57,\ \ldots, 64$ (day 15 and 16)

Dependent variable:

response (interview or **not** interview)

- Phase 1+2 ordinary fieldwork and Phase 3 follow-up
- Objects are individuals (not households)

Models for Phase 1+2 and 3

Logistic regression (binary), explanatory variables

Model 1: Phase 1+2 ($s = 1, 2, \ldots, 56$)	Model 2: Phase 3 ($s = 57, \ldots, 64$)
Register data: Age (16-54 or 55-74 yrs) Born in Sweden or not Education (high) or not Married or not House owner or not	**Register data:** Education (high) or not
1st wave 2nd-8th wave and interview last wave 2nd-8th wave and **no** interview last wave Day shift (9-12, 13-16, 17-18, 19-21) Time slot (1, 2, …, 56)	1st wave 2nd-8th wave and interview last wave 2nd-8th wave and **no** interview last wave 2nd-8th wave and LONG -1 (more than 6 call attempts or not **last wave**) Day shift (9-12, 13-16, 17-18, 19-21) LONG (more than 6 call attempts or not)

Data collection strategy

Phase	Description
1) Day 1-7 $(s = 1, 2, \ldots, 28)$	• Sort the predicted $\hat{p}_k$ objects in treatment (descending) according to **Model 1** (**Phase 1+2**) for each s, • n_s calls (based on decided capacity), • A maximum of 3 calls
2) Day 8-14 $(s = 29, \ldots, 56)$	• Sort the predicted $\hat{p}_k$ objects still in treatment (descending) according to **Model 1** for each s, • n_s calls (based on decided capacity), • A maximum of 9 calls • After 4 calls: stop individuals in wave 2-8 that refused to participate last wave
3) Day 15-16 $(s = 57, \ldots, 64)$	• Sort the predicted $\hat{p}_k$ objects still in treatment descending according to **Model 2** (**Phase 3**) for each s, • n_s calls (based on decided capacity, reduced number), • A maximum of 13 calls

Note: within each s is only one call attempt allowed to the objects

Simulation

- The response propensities are assumed to be $Be(\hat{p}_k)$ distributed in the two logistic regression models.
- For time slot 1: $5{,}164$ random selections were made from Model 1. The randomization corresponds to the outcome if all the individuals in the sample were contacted for $s = 1$. The n_1 highest response propensities are inspected and those who "respond" are set aside.
- The "nonresponders" continue to the 2nd time slot, $s = 2$ The n_2 highest response propensities are inspected and those who "respond" are set aside…
- The procedure continues until time slot 64, where the "data collection" ends.

The data collection is replicated 1,000 times for the described strategy.

Evaluation of the new strategy

$P =$ the weighted response rate in per cent

IMB = the imbalance measure measures the difference between the response set r and the selected sample s for a chosen **x**-vector. It could be demonstrated* that IMB is equal to the variance for the response propensities for the chosen **x**-vector

$$CV_s = \frac{\sqrt{IMB}}{P}$$

Measure of bias

Using *income* and *employed* from registers, available for the selected sample s, is it possible to estimate the difference between estimators based on the response set r and the selected sample s.

RDF_{exp} = the relative difference between an expansion estimator and the HT-estimator

RDF_{cal} = the relative difference between a calibration estimator and the HT-estimator

***Särndal & Lundquist 2014**

Note: auxiliary variables (register data) depends on available variables and the measures depends on the sample s.

Simulation results, Week 4

				Income		Employed	
LFS-January 2016	P	IMB	CV_s	RDF_{exp}	RDF_{cal}	RDF_{exp}	RDF_{cal}
Week 4 (INPUT)	57.5	1.61	9.45	12.83	3.85	6.91	3.00
				Income		**Employed**	
Simulation	P	IMB	CV_s	RDF_{exp}	RDF_{cal}	RDF_{exp}	RDF_{cal}
Strategy	64.0	1.67	8.78	10.03	1.42	6.93	2.49

The response rate P is weighted in percent, IMB, CV_s, RDF_{exp} and RDF_{cal} are multiplied with 100.

x-vector used in computations: Age, High Education, Owner, Origin, Civil, Gender
(3) (2) (2) (2) (2) (2)

The simulation manage to maintain the data quality and reduces the call attempts, this is extra clear in the follow-up (Phase 3)

Next steps

- **Work with the logistic regression models:**
  - **Factors to investigate:** *time slots,* the 2 first calls should be in a predefined time slot; *outcome of previous call*
  - Should Cox regression with time-varying coefficients be used?
  - Should Bayesian models (see Wagner and Raghunathan 2007) be implemented?
  - **Develop the tool:** include a simple cost function (e.g. time for interview, “not interview”) and maximum interviewer hours
- **Better strategies:**
  - The tool makes it possible to find better strategies with better control of the data collection (input to Schouten et al. 2017).
  - Experiments? -The possibility should be noted!

Thank you!

peter.lundquist@scb.se

anton.johansson@scb.se

References

Choudhry, Hidiroglou & Laflamme (2011). ”Optimizing CATI workload to minimize data collection cost.” Proceedings of the Survey Research Methods Section, 1904-1913, ASA.

Durrant, D’Arrigo & Müller (2013). “Modeling Call Record Data: Examples from Cross-Sectional and Longitudinal Surveys,” in Improving Surveys with Paradata: Analytic Use of Process Information, ed. Kreuter, F., 281–308, Hoboken, NJ: John Wiley and Sons.

Durrant, Maslovskaya & Smith (2015). “Modeling final outcome and length of call sequence to improve efficiency in interviewer call scheduling.” Journal of Survey Statistics and Methodology, 3, 397–424.

Johansson, Lundquist & Durrant (2016). ”Stopping rules in a longitudinal survey – impact on cost and survey quality.” Presented at AAPOR.

Särndal & Lundquist (2014). “Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation.” Journal of Survey Statistics and Methodology, 2, 361-387.

Schouten, Mushkudiani et al. (2017). “A Bayesian analysis of design parameters in survey data collection.” Paper submitted.

Thomas & Reyes (2014). “Tutorial: Survival estimation for Cox regression models with time-varying coefficients using SAS and R”. Journal of Statistical Software, 61, Code Snippet 1, 1-23.

Wagner & Raghunathan (2007). ”Bayesian approaches to sequential selection of survey design protocols.” Proceedings of the Survey Research Methods Section, 3333-3340, ASA.