An Analysis of Business Response Burden and Response Behaviour Using a Register of Data Provision

Boris Lorenc

Statistics Sweden

Acknowledgements

- Continuation of the work initiated under the BLUE-ETS project, that was funded by grant agreement n°244767 within the European Community's Seventh Framework Programme (FP7/2007-2013)
- The views expressed are those of the author and do not necessarily reflect policies of Statistics Sweden or of the EU
- I am grateful to Wim Kloek for suggesting what the Register of Data Provision (RDP) can be used for, and to Christina Rääf, Statistics Sweden, for facilitating access to data analysed in this study

Overview

- Introduction and research questions
- Data
- Methodology
- Results
- Conclusions and some proposals

Introduction

- If we bother the businesses more, are they less cooperative?
  - McCarthy et al. (2006)
    - Burden does not in general have a negative effect on survey response
    - Even in cases where negative effects were found, these were often small
- The research question: what is the relation between burden and nonresponse in business surveys

Introduction

- Opportunity: The Register of Data Provision (RDP)
  - Maintained by Statistics Sweden since 2009
  - Collects data from dedicated servers of ongoing surveys
  - The main content: whether the business provided the data or not (i.e. responded or not) for each cycle of each survey that the business has been sampled into
  - Additional data linked from the Business Register or from survey metadata
  - Currently used mainly to manage relations with businesses (especially, with businesses complaining of excessive burden)
  - Can be used to provide insights into relation between burden and response behaviour
    - total coverage of a national statistics producer's business data collection (all businesses, all surveys): enables broader generalisability

What we know and don’t know

- Business surveys
  - Less well researched than household surveys
  - “Response burden [in business surveys] is not a straight forward area to discuss, measure and manage” (Jones, 2012: p1)
  - Not obvious that increase in response burden is to lead to reduced participation
    - Ability to dedicate additional work force/efforts
    - However, doing so may hurt the bottom line
  - Data collection efforts by the surveying organisation are likely to differ within a survey depending on business size (e.g. large businesses may receive more efforts)

What we know and don’t know

- So
  - McCarthy et al. (2006): agricultural operations
    - Burden does not in general have a negative effect on survey response
    - Even in cases where negative effects were found, these were often small
  - Davis and Pihama (2009): mandatory annual survey
    - A statistically significant positive relationship between burden and the probability of nonresponse in the survey
    - The authors note that this effect was relatively small in magnitude in comparison to some other of the factors in the study
- What can we learn from RDP
  - Predictors of response
  - Impact of burden

RDP

- Collects data from dedicated servers of ongoing surveys
- The main content: whether the business provided the data or not (i.e. responded or not) for each cycle of each survey that the business has been sampled into
- Business ID, Survey ID
- Additional data linked from the Business Register or from survey metadata

Data

- Only 2013 data used
- Purged of records if
  - business was inactive;
  - a crucial variable was missing (e.g. business size or industry);
  - a data collection had less than 30 businesses as its sample size;
  - the response rate in a data collection was on the edge of the sample space (i.e. total response, 100%, or total nonresponse, 0%)
- Measures of the burden that was imposed on a business over time
  - number of different surveys they were requested to provide data for
  - number of different surveys instances they were requested to provide data in
  - total length of time, under compliance, that it took (or, would have taken) the business to provide data across all the survey instances they were requested to provide data for

Data

- Measures of burden reflect burden accrued over the first half of the year
- Effect (responded or not) investigated in the second part of the year

Preparation stage	Records	Businesses	Surveys
Before the reductions	1,124,610	187,510	135
After the reductions	1,023,500	176,128	108
Used for generating burden measures	450,277	58,461	94
Used for the analysis	421,290	58,461	66

Data

- Variables

Level	Variable
I. Business	1. ID
	2. Size (employees) (5 classes) [SizC]
	3. Industry (6 classes) [IndC]
	4. Accrued response burden: number of surveys [Surv]
	5. Accrued response burden: number of survey instances [Inst]
	6. Accrued response burden: time (log and 7 classes) [BurC]
	7. Response rate during accrual period [RespR]
II. Survey	1. ID
	2. Periodicity [Mont]
	3. Is a part of official statistics or not [Sos]
	4. Is mandatory or not [Vol]
	5. Is conducted due to EU regulation or not [Eu]
	6. Average length of time to provide data in a survey instance [ToT]
III. Record	Responded (1) or not (0) [y]

Data

- Variables - comments
  - Unbalanced
  - Single observations in some/many grouping cells as concerns businesses (a business taking part in only one survey, which in turn has only one instance)
  - May be strong correlation between some predictors
  - No record-level predictors

Methods

- Two approaches, both logistic

$$\Pr(y_i = 1) = \frac{e^{\boldsymbol{x}'_i\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}'_i\boldsymbol{\beta}}}$$

  - GLM with binomial link (i.e., a logistic regression)

    $\boldsymbol{x}$ is a vector of explanatory variables, and $\boldsymbol{\beta}$ a set of corresponding coefficients

    - function *glm* used (package *stats* in R)

  - A hierarchical (multi-level) cross-classified model, with the grouping variables:

    ○ Business ID

    ○ Survey ID

    so, $\boldsymbol{x}$ additionally includes indicators of belonging to a business and to a survey

    - function *glmer* used (package *lme4* in R)

Results

- Predictors (univariate analysis)

Variable	GLM		GLMER	
	AIC	Coefficient (sd)	AIC	Coefficient (sd)
1	374494	1.637	234539	2.536
Response rate	270822	4.172	205615	6.639
Size class 2	353772	0.464 (.014)	233238	1.020 (.057)
Size class 3		0.992 (.014)		1.841 (.058)
Size class 4		1.597 (.016)		2.289 (.073)
Size class 5		1.939 (.017)		2.100 (.100)
Industry class 2	366780	0.444	*233713*	*0.072*
Industry class 3		0.242		*0.163*
Industry class 4		-0.631		*-1.331*
Industry class 5		0.096		*0.376*
Industry class 6		0.131		*0.277*
Industry class 7		-0.078		*0.072*
VoluntaryYes	366060	-1.445	234541	-0.565
log(Total time)	354967	0.355	*233719*	*0.370*
BurC2	366325	0.210	*233628*	*0.156*
BurC3		-0.012		*-0.488*
BurC4		0.012		*-0.250*
BurC5		0.115		*-0.077*
BurC6		0.812		*1.111*
BurC7		1.146		*1.966*
Log(Surveys)	355834	0.580	233898	0.631
Log(Survey Instances)	357579	0.443	234093	0.410

Results

- Response surface of the predictions (multivariate models): “best” computable estimates involving a burden measure predictor

Results

- Response surface of the predictions (multivariate models): “best” computable estimates involving a burden measure predictor

GLM

```
y ~ RespR + Vol + IndC
    + SizC + BurC
    + RespR:Vol + IndC:SizC
    + RespRate:SizC
    + Vol:SizC + Vol:IndC
    + RespR:IndC + RespR:BurC
    + SizC:BurC + IndC:BurC
    + RespR:IndC:SizC
    + RespR:SizC:BurC
```

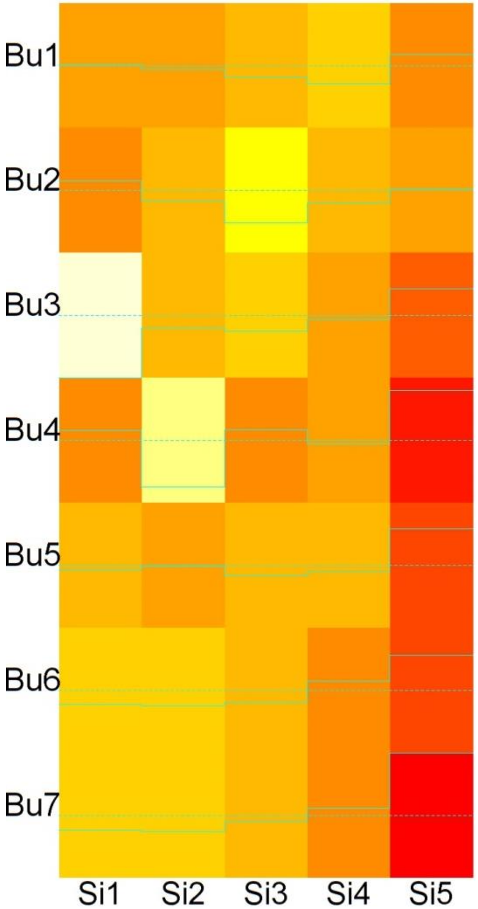
GLMER

```
y ~ RespR * BurC
    + SizC * BurC
    +(1 | BusID)
    +(1 | SurID)
```

Results

- Response surface of the predictions (multivariate models): “best” computable estimates involving a burden measure predictor

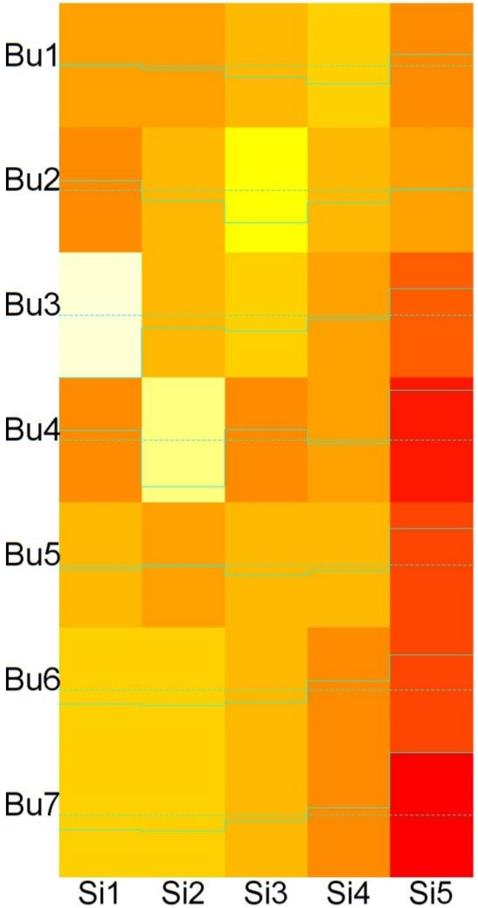
GLM (Pred RR 0.69 – 0.82)



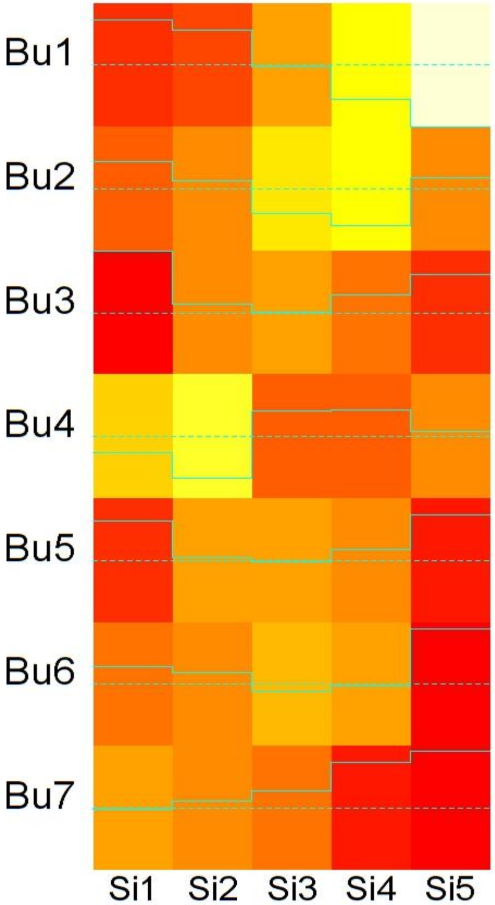
Results

- Response surface of the predictions (multivariate models): “best” <u>computable</u> estimates involving a burden measure predictor

GLM (Pred RR 0.69 – 0.82)



GLMER (Pred RR 0.69 – 0.82)



Conclusions: substantive

- Based on both models, for large businesses, increased burden during a period of time seems to reduce subsequent response
- Based on the hierarchical model only, level of burden “optimal” for subsequent response seems to vary between business sizes; loosely, in McCarthy’s terms,
  - if we bother small businesses more than their average, they will be more cooperative
  - if we bother big businesses more than their average, they will be less cooperative
- Caveats
  - It has not been calculated whether the differences are statistically significant
  - Models may be unstable

Conclusions: method

- GLM
  - Advantage: faster, computationally simpler
  - Disadvantage: might not reflect the data structure well (observations are actually clustered under businesses and under surveys)
- Hierarchical (GLMER)
  - Advantage: perhaps reflects the structure of the data better
  - Disadvantage: computationally complex, with large data sets (here: 400 K) it reaches fast hardware and software limits of a ‘normal’ contemporary computer, perhaps especially so with highly unbalanced data (as here)

Further work

- Include more years into the analysis (how?)
- Find out if possible to collect unit level data (estimate of actual time for each unit level observation)
- Tailor models further in e.g *bugs* or similar
- We will investigate if it is feasible, from a confidentiality perspective, to make the data set available for research purposes

References

- Davis, W.R., and N. Pihama (2009). Survey Response as Organisational Behaviour: An Analysis of the Annual Enterprise Survey, 2003–2007. Presented at New Zealand Association of Economists conference, Wellington, New Zealand, July 1–3 2009
- Jones, J. (2012). “Response Burden: Introductory Overview Lecture”. Presented at ICES-IV, Montréal, Québec, Canada, June 11–14, 2012
- McCarthy, J.S., Beckler, D.G., and Qualey, S.M. (2006). “An Analysis of the Relationship Between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative?” Journal of Official Statistics, 22(1), 97–112.