Embedded Editing: Is There a Limit?

Boris Lorenc[1]
Statistics Sweden, Stockholm[2]

Introduction

Data editing is an essential step in collecting and processing the quantitative information that forms the basis for producing summary statistics. In official statistics production, editing is a costly step as it tends to consume considerable proportion – as much as 40 percent in business surveys (Granqvist and Kovar, 1997) – of the total resources for a statistics product.

Production editing, performed by a statistics producer, may result in recontacts with the data provider, who is asked to confirm or verify a certain value that deviates from what the data collector/statistics producer considers a valid or expected value. Thus, alerting the data provider during data collection to deviations that are likely to cause an editing recontact later on may have two related positive consequences:
- it occurs in the course of the data collection itself, when the data provider likely has direct access to the information that enables verifying or editing (while subsequent recontacts generally happen outside the data provision context),
- it occurs automatically, so that the data collector does not need to engage directly with the data provider.

We refer to this as embedded editing: a form of editing where edit controls are embedded in the data collection instrument, which if triggered inform the data provider of a formal (formatting, etc) or substantive (magnitude, inconsistence, etc) deviation from an expectation of the data collector. The editing is thus performed by the data provider, who is asked to review and – if needed – edit the originally supplied value. A control that requires the data provider to edit (change) the response in order to finalise data provision is referred to as a hard editing control, whereas the one that informs but makes the edit optional as a soft editing control.

Embedded editing is thus expected to result in data of better quality, as it has been already edited when it arrives to the statistics producer, obtained with less respondent burden, as the data provider has been asked to verify and edit while in the context of data provision rather than at a later point in time, and obtained at a lesser cost, as the statistics producer has invested fewer resources in editing having assigned a part of it to the data provider.

Prior to this experiment, Statistics Sweden has had little empirical evidence regarding impact of increased embedded editing on data quality, response burden and cost. While the outlined theoretical expectations might hold regarding benefits of increased embedded editing on data quality, response burden and cost, it may also happen that increased embedded editing is experienced by data providers as an obstacle to data provision and thus lead to larger unit nonresponse and decreased preference for embedded editing features, it may increase perceived response burden as the amount of primary data provision work increases with embedded editing, or

1 This paper summarises the results of a project conducted at Statistics Sweden during 2016, entitled “Embedded Editing in SIV [Statistics Sweden’s Data Collection Vehicle]”. The project group consisted of Anette Björnram, Pia Hartwig, Boris Lorenc, Anders Norberg, and Magnus Ohlsson.

2 The opinions expressed in the paper are the author’s and are not to be seen as reflecting the position of Statistics Sweden or the project group.

it may fail to achieve the data quality improvements and thus leave the amount of production editing unchanged.

The experiment reported here was performed with the aim to give Statistics Sweden some guidance regarding embedding editing rules in its data collection tools.

Method

Context

The experiment was performed as an embedded experiment in the 2016 round of the annual survey "Wage and salary structures in the private sector (SLP)". Data collection for this survey started in the beginning of October 2016 and ended in the end of December 2016, by which time the overall response rate in the survey of 68% has been achieved.

Data editing for SLP consists of some embedded edit rules, some production (micro) data edit rules, as well as of macro (or output) editing inspection.

Design

Split-half experimental design, two conditions:
- control: standard level of embedded editing in SLP
- experimental: increased level of embedded editing in SLP.

A sampled business was randomly assigned to one of the two conditions, after stratification on: size (7 categories), activity (6 categories), whether seated in Stockholm county or not, whether sampled for 2015 SLP or not, whether responded in 2015 SLP or not (conditional on being sampled in 2015 SLP), data provision mode chosen for 2015 SLP (conditional on being a respondent in 2015 SLP).

After identification and removal of overcoverage, there were 3676 businesses in the sample, roughly half of them (1845) assigned to the experimental condition and the rest (1831) to the control condition. By the time of closure of data collection for the experiment in December 2016 2500 businesses have responded (1258 in the experimental condition and 1242 in control), for an average response rate of 68%.

Experimental manipulation

Experimental condition meant an increased number of embedded edit rules, compared to the control condition. In practice, it meant addition of six production edit rules to the existing embedded edit rules. The project team surmised that in order to be a useful embedded edit rule, a production edit rule has to:

a) have a high accuracy (i.e. when triggered in production editing, to often lead to a changed value after verification; if not, then it would mostly distract the data provider; information about the accuracy is obtained through process data on production editing),
b) in the course of verification lead to recontact with data provider (i.e. be a costly component of production data editing and in terms of respondent burden),
c) afford formulating a clear information message to the data provider as to what the issue might be with the originally submitted value.

Using these criteria, the six edit rules were selected. For reasons out of scope for this brief summary, it was deemed that this level of increase was sufficient for this first experiment on embedded editing at Statistics Sweden.

Explanatory variables

Following were the available explanatory variables. They refer to businesses sampled for 2016 SLP:

- Whether the sampled business was in the experimental group or control group
- Business's size (in terms of number of employment), 7 categories
- Business's area of activity, 6 groups of NACE codes
- Whether the sampled business was included or not in the 2015 SLP sample
- Whether the sampled business was a respondent or nonrespondent in 2015 SLP (given having been included in 2015 SLP sample)

Of them, in this short report, only analysis of the effect of the experimental manipulation (experimental/control group) is presented. Such an analysis answers the question of importance of the experimental variation for the dependent variables – what happens with such a variable when we increase the amount of embedded editing – whereby the aim of the experiment is fulfilled. Inclusion of the other available explanatory variables would answer the question of the relative importance of these variables, the experimental manipulation included, for the dependent variables. It answers the question of what explains a dependent variable best. While of research interest for understanding business response behaviour and for tailoring data collection procedures, in practice none of the other variables are under the statistics producer's control.

Dependent variables

These variables refer to a sampled business and its participation in 2016 SLP:

- Whether the sampled business was a respondent or not (Resp16)
- Number of edit rules that the sampled business triggered (#Trigg)
- Number of production edit comments, a proxy for amount of production editing that the sampled business required (#Prod edit com)
- Number of recontacts with the sampled business (#Recont)
- Length of time that the data provider within the business estimates that it took to provide data (Perc time) (from the question that generally follows submission of data to Statistics Sweden, asked to track changes in response burden caused by data collection to businesses)
- Preference for embedded editing versus recontacts later on (from an optional 3 questions survey after the SLP form has been submitted) (Pref emb edit)

Analysis

A generalised liner model $E(y)=g^{-1}(x'\beta)$ was fitted to the data, with the predictor $\eta=x'\beta$ and a link function $g^{-1}(.)$ which varied with the dependent variable analysed: logistic for a binary outcome, Poisson for a count, and identity for log-normal data (e.g. time). Specifically, the `glm` function of the `stats` package in R was used.

Results

Results of the analysis are presented in Table 1. They show that our experimental manipulation did not impact the response rates negatively, it did not reduce the preference for embedded editing, and it reduced a proxy for the amount of production editing. However, the manipulation did not reduce the number of recontacts with the businesses (another proxy for the amount of production editing), it increased the perceived length of time it takes to provide data for SLP and it increased the number of edit messages that the data provider has been exposed to.

Table 1 - Modelling of impact of the experimental manipulation on dependent variables. (Values in bold pertain to those dependent variables whose coefficients are significantly different from 0.)

y	$\hat{\beta}$	$Pr(\hat{\beta}=0)$	Effect if only		Effect of embedded editing (experim - control)
			control	experim	
Resp16	0,016	0,819	67,8%	68,2%	0,4%
#Trigg	**0,276**	**0,000**	**26787**	**35283**	**8496**
#Prod edit com	**-0,091**	**0,003**	**4597**	**4197**	**-400**
#Recont	-0,054	0,453	805	763	-42
Perc time[3]	**0,105**	**0,010**	**46**[4]	**51**[5]	**5**[6]
Pref emb edit[7]	-0,154	0,406	83,5%	81,2%	-2,2%

Discussion

The results that were obtained fall midway between what was hoped to be achieved and what could have been a less desirable outcome: while it apparently increased the respondent burden, the experimental manipulation did not lead to increased nonresponse; but it did not reduce the amount of recontacts either, which was an important expected outcome.

In setting the obtained results into a wider context, one should take into account that our experimental manipulation was relatively minimal, adding six embedded edits to about 30 embedded edits already existing in the data collection instrument.

The question that this research cannot answer is what would have happened if the experimental manipulation was much larger: would the increased burden still leave the response rate unimpacted, would the preference for embedded editing still be high, and so on. (Not treated here, an important question is also which edits should be hard and which soft.)

In addition to experiments, one needs a better understanding for the response process and of exactly in which way do messages triggered by edit rules support (or not support) the data provider in the task of providing correct data. Barely increasing the proportion of embedded edits does not guarantee that the theoretical promises of embedded editing will be fulfilled, if one first does not get a better grip on the role and impact of edit messages in the process. For understanding that, also qualitative studies will be needed.

3 With imputed values due to nonresponse on this variable; conditional independence of the nonresponse mechanism and the dependent variable is assumed).
4 Average per business, in minutes (’usual’ time, not logarithmised).
5 Average per business, in minutes (’usual’ time, not logarithmised).
6 Average per business, in minutes (’usual’ time, not logarithmised).
7 With imputed values due to nonresponse on this variable; conditional independence of the nonresponse mechanism and the dependent variable is assumed).

Another area seemingly lacking a better understanding is that concerning relations between the different edit rules and their respective edit messages. Is this a set of rules thrown in independently of each other, or is there some logical (and then also communicative) relation between them. Seeking these I believe can best be achieved in the context of cognitive aspects of survey methodology (possibly expanded for business surveys by approaches such as Willimack's, Bavdaz's or Lorenc's): as there is a four/five step model of responding to questions in surveys, the same model applies in principle also for processing edit messages in surveys. Theoretical understanding and empirical research on this topic is hugely missing in survey methodology: edit messages seem presently to be treated as an area of its own, outside any CASM methodology developed for questionnaires, questions and response alternatives.