On calibration in DG-1 business survey

Maciej Beręsewicz (Poznan University of Economics, Statistical Office in Poznan)

Introduction

The goal of the presentation is to discuss the calibration approach in the context of short-term business statistics on the DG-1 survey example.

The following aspects will be presented:

- Description of the current sample selection scheme.
- (Self-)selection mechanism in relation to design and other variables.
- Correlation between propensity score and selected target variables.
- Weighting schemes and calibration approach.

DG-1 survey – the target population

- DG-1 is a monthly survey of establishments.
- The target population are establishments over 9 employees that are classified into two groups – big (over 49 employees) and medium (between 10-49 employees).
- In addition, establishments that are classified by European Classification of Economic Activities (NACE) starting with B to J, L, M (z excluding divisions 72 i 75), N, R and division 02, 95, 96 and class 03.11 take part in the survey.

DG-1 survey – scope of the survey

- Sales,
- Taxes and subsidies,
- Number of employees,
- Working time,
- Salaries,
- Base price indexes,
- Turnover,
- New orders/contracts,
- Transportation.

DG-1 survey – sample selection and allocation

How the sample is selected?

- Statistical Unit Database (pol. **Kartoteka**) – the sampling frame for DG-1 survey (updated on monthly basis $\sim$ 7 mln establishments from 2009-1 to 2014-9).
- All **big** establishments are obligated to take part in the DG-1 survey.
- At least 10% of all medium establishments stratified by ownership (private, public) section, division and group and section G defined in NACE (in total 453 strata) are selected.
- Minimum sample size for each strata is defined as follows

$$\left[\frac{\#\text{units in section/division/group NACE}}{10} + 1\right]$$

Where # denotes number of, [] denotes ceiling.

- Sample is drawn in the begining of January on each year.

Motivation – self-selection

We know that (Bethlehem 2010), in case of self-selection sample surveys bias of the mean of the target population is given by:

$$Bias(\bar{y}_s) = \frac{N_{ns}}{N}(\bar{Y}_s - \bar{Y}_{ns}) + \frac{C(\rho, Y)}{\bar{\rho}} = \frac{N_{ns}}{N}(\bar{Y}_s - \bar{Y}_{ns}) + \frac{R(\rho, Y)S(\rho)S(Y)}{\bar{\rho}}$$

where Y is a target variable, $\bar{y}$ denotes sample mean, s denotes sampled units, ns denotes not sampled units, N denotes number of units in population, N_{ns} denotes number of not sampled units, ρ denotes propensity score, $R(\rho, Y)$ denotes correlation between propensity scores and target variable(s), $S(\rho)$ is standard deviation of propensity scores and $S(Y)$ is standard deviation of target variable(s).

Motivation – self-selection

ρ denotes propensity score given by:

$$\rho(X) = P(r = 1|X)$$

where r denotes response to survey (1 answer, 0 refusal) and X variables that we consider as a explanatory for the response behavior of units. $\rho(X)$ can be estimated using various methods (e.g. logistic regression, random forest).

Motivation – self-selection

- Imputation or weighting adjustments can correct sample distribution of X to known population totals,
- However, when a strong correlation between ρ and Y is observed the bias in statistics may still be present.

$$|B_{max}| = S(Y)\sqrt{\frac{1}{\bar{\rho}} - 1}$$

- The self-selection (or a non-ignoble unit non-response) problem is common in business surveys.

Therefore, we will study the self-selection mechanisms before applying weighting procedures.

DG-1 survey – basic information about the DG-1 sample and population

Table 1: Sample count (in percent) by size of company

SIZE	Min	Mean	Median	Max
Big	96.55	98.05	98.12	98.55
Medium	16.77	17.79	17.45	18.94

Table 2: Population count by the size of company

SIZE	Min	Mean	Median	Max
Big	18008	18694	18766	19462
Medium	67087	76106	78806	83073

Table 3: Distribution between non-sampled, sampled and population count by business ownership

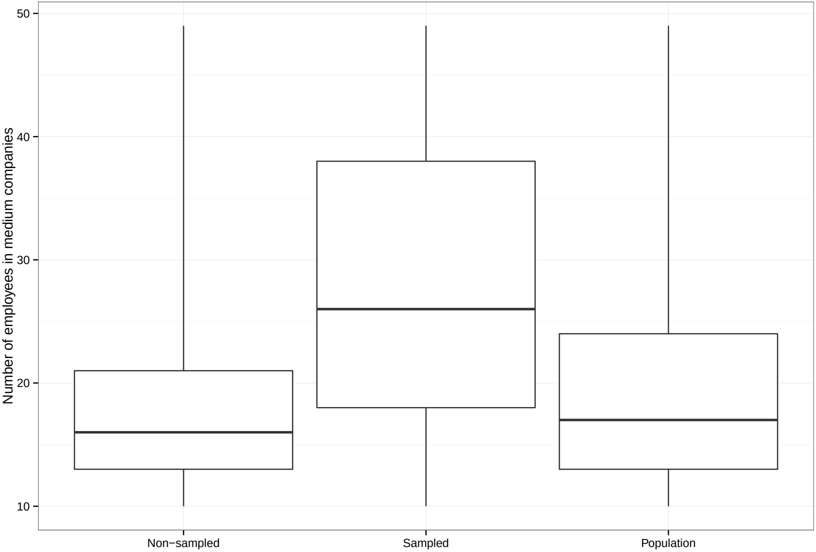
Ownership	Non-sampled	Sampled	Population
Public	2.59	6.93	3.36
Private	97.41	93.07	96.64

DG-1 survey – differences between sampled and non-sampled units

Table 4: Distribution between non-sampled, sampled and population count by NACE

NACE	Non-sampled	Sampled	Population
A	0.53	1.56	0.72
B	0.29	0.58	0.34
C	27.37	28.82	27.63
D	0.23	0.79	0.33
E	1.12	2.18	1.31
F	15.73	10.70	14.84
G	28.82	29.70	28.98
H	6.43	4.59	6.10
I	4.67	3.71	4.50
J	2.09	2.76	2.21
L	2.22	2.98	2.36
M	4.81	4.43	4.74
N	2.65	3.84	2.86
R	2.06	2.37	2.11
S	0.98	1.01	0.98

Details about the data – number of employees in medium companies



Calculation of propensies

The following models were considered:

- the model for each month separately
- the model for each month separately with additional information on previous month (took, or not took part in the survey)

The final model contained the following variables

- $Sampled_{t-1}$ – indicator whether a unit was in a $t-1$ sample (=1, else 0)
- VID – Voivodeship (16 levels)
- CITY - City (whether a company is from a city = 1, else = 0)
- NACE – NACE classification
- SIZE – Size of the company (2 levels, reflevel = 'BIG')
- OWN – Ownership Status (2 levels, reflevel = 'Public')

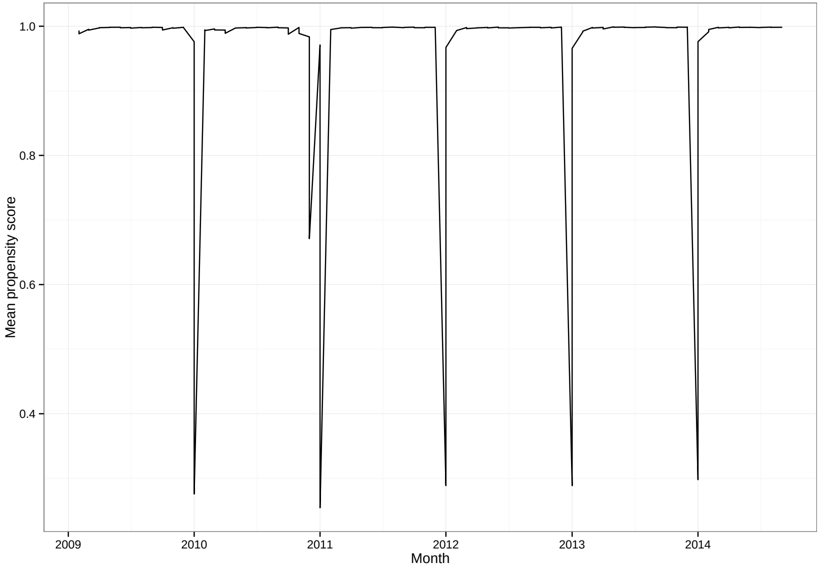
Number of Employees (NoE) was removed due to computational problems.

Propensity scores over time

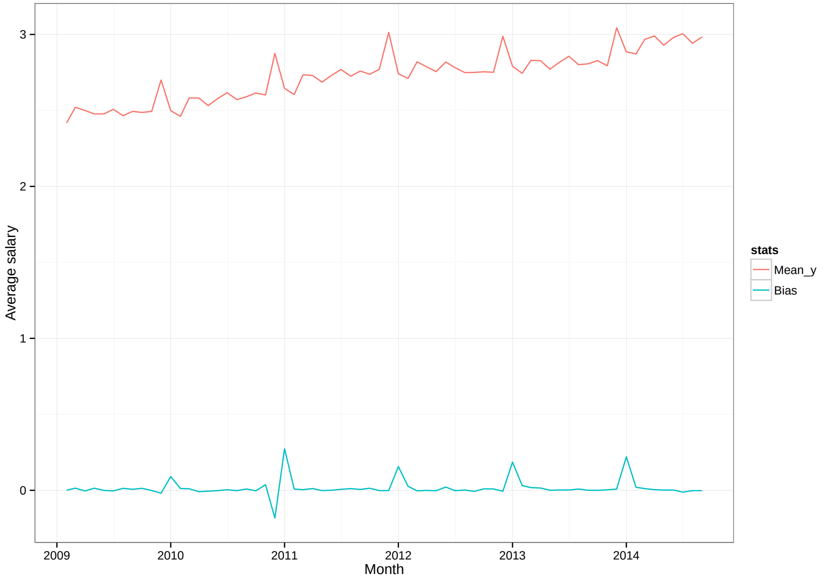
Date	SDy	Mean_y	SD_rho	Mean_rho	Corr	Bias
2009-02-01	1.714	2.417	0.087	0.988	0.006	0.001
2009-03-01	2.661	2.521	0.065	0.994	0.082	0.014
2009-04-01	2.409	2.500	0.041	0.998	-0.041	-0.004
2009-05-01	3.466	2.477	0.035	0.998	0.113	0.014
2009-06-01	2.264	2.477	0.044	0.998	-0.007	-0.001
2009-07-01	2.359	2.507	0.042	0.997	-0.037	-0.004

Date	SDy	Mean_y	SD_rho	Mean_rho	Corr	Bias
2014-04-01	2.424	2.990	0.033	0.998	0.056	0.004
2014-05-01	2.097	2.930	0.033	0.999	0.023	0.002
2014-06-01	2.332	2.980	0.038	0.998	0.018	0.002
2014-07-01	2.362	3.005	0.033	0.998	-0.152	-0.012
2014-08-01	2.181	2.942	0.036	0.999	-0.027	-0.002
2014-09-01	2.351	2.984	0.035	0.998	-0.022	-0.002

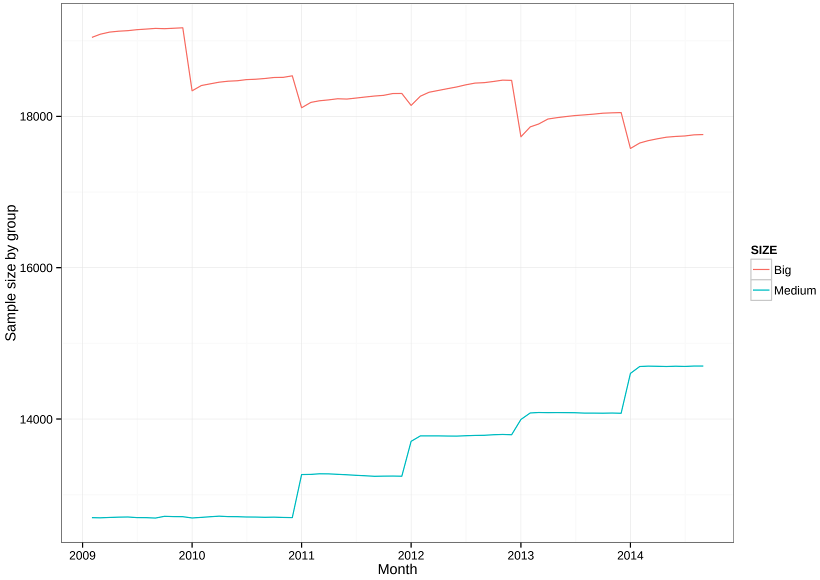
Propensity scores over time



Propensity scores over time



Propensity scores over time – why it happen?



The calibration approach

I have applied the standard calibration approach (Deville and Särndal 1992) given by

$$min \sum_{s} G_k(w_k, d_k)$$

with subject to the calibration equation

$$\sum_{s} d_k \boldsymbol{x}_k F(q_k \boldsymbol{x}'_k \boldsymbol{\lambda}) = \sum_{U} \boldsymbol{x}_k$$

using logit distance function given by

$$G(x) = x(log(x) - 1) + 1$$

The calibration equations

The following variables were considered:

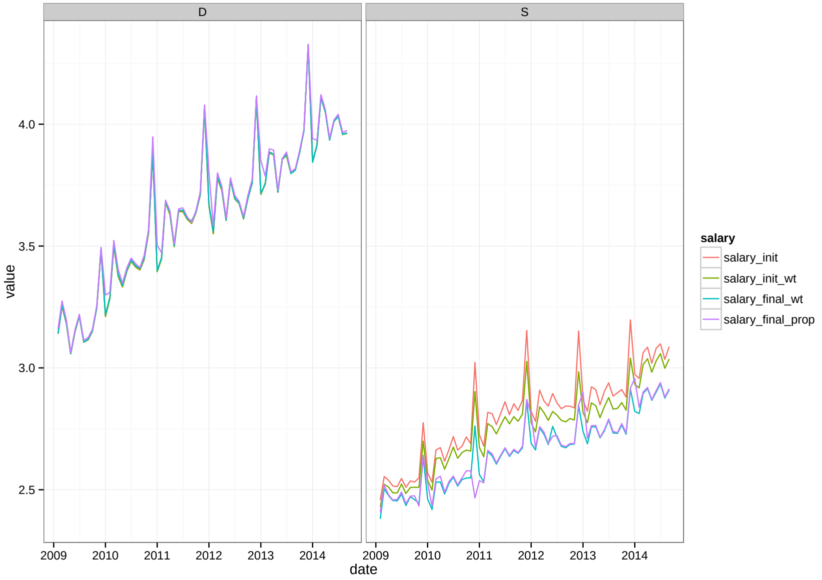
- VID – Voivodeship (16 levels)
- OWN – Ownership (2 levels)
- NoE – Number of employees
- NACE – Classification of company

and the following calibration equations were to met

$$
\begin{aligned}
&VID \times OWN \times SIZE \times NoE + \\
&\qquad NACE \times OWN \times SIZE \times NoE + \\
&\qquad NACE \times NoE + VID \times NoE
\end{aligned}
$$

where $\times$ denotes interaction between levels of variables

Average salary based on the proposed methods



Summary

- Differences between initial weights and calibrated are due to major changes in auxiliary variables (different from the ones that were used for sampling)
- There is a small correlation between propensity score and the selected target variables.
- Taking part in the survey in time $t - 1$ is the most influential variable in the propensity score model, however this model do not take into account certain (yearly) sampling schemes. Therefore, it should be further investigated.

Discussion

- Possible solution to the problem with breakdowns (in January) is to apply logistic mixed model to estimate propensities for each units taking into account auto-correlation in time.
- Unbalanced groups (sampled and non-sampled) indicates that logistic regression may be not suitable for the propensity score estimation; or re-sampling should be used to balance groups.
- Outliers/influential obs. caused overestimation of bias in target variables due to self-selection mechanism.

Extra information

I used R and RStudio with the following additional packages:

- data.table – for fast dataset summaries (much more faster than dplyr)
- tidyr – for transformation of datasets (wide to long, long to wide)
- laeken – for calibration
- Matrix – for sparse matrix manipulation
- speedglm – for speed logistic model computation (stats::glm is slooooow)
- parallel – for parallel computations
- ggplot2 – for visualisation
- knitr + rmarkdown – for the presentation

Thank you for your attention!

Literature

1. Bethlehem, J. (2010). Selection Bias in Web Surveys. International Statistical Review, 78(2), 161–188. doi:10.1111/j.1751-5823.2010.00112.x
2. Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382.

Calculation of propensies (example model for 2014.09)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.557486	2.461187	-0.227	0.8208	
r_2014_8	16.307614	0.501779	32.500	<2e-16	***
WON4	-0.174760	0.698542	-0.250	0.8025	
WON6	0.006921	0.865753	0.008	0.9936	
WON8	-0.137486	0.740247	-0.186	0.8527	
WON10	-0.974172	0.700088	-1.391	0.1641	
WON12	0.065273	0.605495	0.108	0.9142	
WON14	-0.746577	0.486814	-1.534	0.1251	
WON16	-0.316041	0.984851	-0.321	0.7483	
WON18	0.788754	0.636653	1.239	0.2154	
WON20	-1.315827	1.192993	-1.103	0.2700	
WON22	-0.568984	0.793099	-0.717	0.4731	
WON24	-0.118951	0.550186	-0.216	0.8288	
WON26	0.510224	0.921634	0.554	0.5798	
WON28	-0.412797	0.705315	-0.585	0.5584	
WON30	-0.950822	0.549360	-1.731	0.0835	.
WON32	-1.155491	0.863818	-1.338	0.1810	

Calculation of propensies (example model for 2014.09)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	5.57631	0.12184	45.767	< 2e-16	***
r_2013_12	4.79431	0.10414	46.037	< 2e-16	***
WON4	-0.03317	0.05450	-0.608	0.54286	
WON6	-0.04169	0.05961	-0.699	0.48433	
WON8	0.19847	0.06549	3.031	0.00244	**
WON10	-0.11838	0.05083	-2.329	0.01987	*
WON12	-0.20870	0.04775	-4.371	1.24e-05	***
WON14	-0.53442	0.04371	-12.228	< 2e-16	***
WON16	0.54664	0.06500	8.410	< 2e-16	***
WON18	0.12432	0.05564	2.234	0.02547	*
WON20	0.31284	0.06770	4.621	3.81e-06	***
WON22	-0.14672	0.05128	-2.861	0.00422	**
WON24	-0.21818	0.04508	-4.839	1.30e-06	***
WON26	0.08807	0.06569	1.341	0.18006	
WON28	0.03604	0.06223	0.579	0.56248	
WON30	-0.31382	0.04675	-6.712	1.92e-11	***
WON32	-0.02954	0.05888	-0.502	0.61590	

Initial result for mixed model

```
Generalized linear mixed model fit by maximum likelihood (I
 Family: binomial  ( logit )
Formula: sampled ~ WON + SEK + KLASA + miasta + PKD_sekcja
   Data: dg2014
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
111380.3 111802.2 -55654.1 111308.3   908161

Scaled residuals:
    Min      1Q  Median      3Q     Max
-109416       0       0       0    1937

Random effects:
 Groups Name        Variance Std.Dev. Corr
 ID     (Intercept)   1.167   1.08
        time        153.673  12.40     0.98
Number of obs: 908197, groups:  ID, 102509
```

Initial result for mixed model

```
(Intercept)    5.5225     0.8445     6.54 6.17e-11 ***
WON4          -0.8871     0.4156    -2.13 0.032797 *
WON6           2.1275     0.4915     4.33 1.50e-05 ***
WON8          -0.2733     0.5043    -0.54 0.587844
WON10          1.9342     0.3994     4.84 1.28e-06 ***
WON12         -0.6104     0.3571    -1.71 0.087398 .
WON14         -5.1787     0.3087   -16.78  < 2e-16 ***
WON16          2.3367     0.5176     4.51 6.35e-06 ***
WON18          0.9161     0.4299     2.13 0.033087 *
WON20          1.9420     0.5558     3.49 0.000476 ***
WON22          1.3054     0.4086     3.19 0.001400 **
WON24         -3.5269     0.3275   -10.77  < 2e-16 ***
WON26          0.2283     0.5399     0.42 0.672353
WON28          2.1195     0.4972     4.26 2.01e-05 ***
WON30         -2.1242     0.3303    -6.43 1.27e-10 ***
WON32          1.5831     0.4729     3.35 0.000814 ***
SEK2          -3.6060     0.4008    -9.00  < 2e-16 ***
KLASAS       -12.9964     0.1605   -80.97  < 2e-16 ***
```

Initial result for mixed model

Month	$\bar{\rho}$
1	0.9476
2	0.9984
3	0.9992
4	0.9994
5	0.9996
6	0.9997
7	0.9998
8	0.9998
9	0.9998