GREG estimation with reciprocal transformation for a Polish business survey

Grażyna Dehnel[1]

Abstract

As a result of economic changes the scope of tasks for business statistics has expanded. There is a growing demand for short-term statistical data delivered on a monthly or quarterly basis, with improved accuracy and coherence. To meet this demand, it is necessary to develop estimation methods that can take advantage of administrative data. The purpose of these efforts is to increase the effectiveness of estimates and extend the scope of information in terms of the number of variables and cross-classifications.

The paper presents an attempt to estimate basic economic information about small enterprises by applying reciprocal transformation to one of the small area statistics methods - GREG estimation. Variables from administrative registers were used as auxiliary variables. The study was conducted for provinces cross-classified by categories of economic activity.

Keywords: *robust estimation, business statistics, outliers, GREG*

JEL Classification: C40, C51

1 Introduction

In recent years much attention has been devoted to business statistics. To meet the growing needs for information, it is necessary to conduct research aimed at expanding the scope of business statistics. The main difficulty facing official statistics in this respect is the rising nonresponse. In addition, the scope of economic information that can be collected is limited by survey costs and respondent burden resulting from statistical reporting. Under these circumstances, the growing demand for information can only be satisfied by exploiting administrative sources of data. It is expected that the adoption of new solutions will improve the efficiency of estimates and, above all, increase the number of cross-classifications available in statistical outputs (Markowicz, 2014). In the search for new approaches to estimating parameters of enterprises it is necessary to account for the specific characteristics of the target population. One characteristic feature of the population of enterprises is the presence of outliers (Schmid et al., 2016; Todorov et al., 2011). For this reason, the present article focuses on an estimation method used in small area estimation, which employs reciprocal transformation. The purpose of the study was to assess the possibility of applying

[1] Corresponding author: Poznań University of Economics and Business, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań, Poland, e-mail: grazyna.dehnel@ue.poznan.pl.

a modified generalised regression estimator (GREG) to estimate characteristics of small companies. In order to improve estimation quality, lagged auxiliary variables from administrative sources were used. Estimates were calculated using data about small companies classified into activity section: *Trade.*

2 The estimation method

One of the assumptions of a linear regression model is homoscedasticity. In the case of the population of enterprises characterised by strong asymmetry of distribution and strong variation, this assumption is often not satisfied (Zhang and Hagesaether, 2011). As a result, the variables are heteroscedastic, which results in inefficient estimates of parameters. For this reason, it is necessary to develop methods that minimize the impact of heteroscedasticity on the precision of estimates. One way of achieving this is by transforming variables. For example R. Chambers et al. (2001) propose a modification of the GREG estimator which involves an additional auxiliary variable z.

In its classic form the GREG estimator of the total of variable Y:

$$\hat{Y}_{GREG} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} w_i e_i \qquad (1)$$

where $\hat{y}_i = \mathbf{x}'_i \hat{\beta}$ in a population U

s - sample size,

model parameter $\hat{\beta}$ is estimated using a modified formula which includes additional variable z (Chambers et al., 2001):

$$\hat{\beta} = \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}'_i / z_i^{\gamma} \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{x}_i y_i / z_i^{\gamma} \right) \qquad (2)$$

where:

z, x - auxiliary variables,

γ - parameter selected depending on the degree of heteroscedasticity,

for $\gamma = 0$, estimator (1) has the classic form of the GREG estimator.

The estimator given by formula (1) can be expressed in the form which is identical to the classic formula for the GREG estimator:

$$\hat{Y}_{GREG} = \sum_{i \in s} w_i g_i y_i \ . \qquad (3)$$

The only difference is the definition of weight g_i, which depends on the value of auxiliary variable *x* for sampled units and is defined as:

$$g_i = 1 + \left(X_d - \hat{X}_{HT,d}\right)' \left(\sum_{i \in s_d} w_i \mathbf{x}_i \mathbf{x}'_i / z_i^{\gamma}\right)^{-1} \left(\mathbf{x}_i / z_i^{\gamma}\right) \quad (4)$$

where:

d - domain,

g_i - weight of the *i*-th unit,

$\hat{Y}_{GREG,d}$ - estimate of the total in domain *d* given by the GREG estimator,

$\hat{X}_{HT}$ - direct Horvitz-Thompson (HT) estimator of the total of auxiliary variable *x*,

X - total of auxiliary variable *x*.

In the modified version of the GREG estimator proposed by R. Chambers et al. (2001), DFBETA is strongly correlated with weight g_i. This means that the value of the distance measure affects the degree to which variable *y* is modified. The DFBETA measure for *i*-th unit can be calculated using the following formula (Chambers et al., 2001):

$$DFBETA_i = \left(\sum_{i \in s} \mathbf{x}_i \mathbf{x}'_i / z_i^{\gamma}\right)^{-1} \left(\frac{\mathbf{x}_i}{z_i^{\gamma}}\right) \left(\frac{e_i}{1 - h_i}\right) \quad (5)$$

where: $$h_i = \left(\mathbf{x}_i / z_i^{\gamma}\right)' \left(\sum_{i \in n} \mathbf{x}_i \mathbf{x}'_i / z_i^{\gamma}\right)^{-1} \left(\mathbf{x}_i / z_i^{\gamma}\right) \quad (6)$$

From previous surveys it is known that the value of parameter γ is included in the interval <1,2> (Särndal, 1992), which is why the following estimators were analysed in the study:

1. $\hat{Y}^{0}_{GREG}$ - $\gamma = 0 \Rightarrow z_k^0$ an estimator based on a linear regression model under homoscedasticity (classic GREG estimator),

2. $\hat{Y}^{1}_{GREG}$ - $\gamma = 1 \Rightarrow z_k^1$,

3. $\hat{Y}^{1,5}_{GREG}$ - $\gamma = 1{,}5 \Rightarrow z_k^{1,5}$,

4. $\hat{Y}^{2}_{GREG}$ - $\gamma = 2 \Rightarrow z_k^2$.

Numbers 2-4 denote regression estimators based on a linear regression model under homoscedasticity.

Estimation quality was assessed with reference to estimates obtained using classic direct estimators: HT and the GREG estimator (Gamrot, 2014):

- Horvitz-Thompson (HT) estimator $$\hat{Y}_{HT} = \frac{N}{n} \sum_{i \in s} y_i \quad (7)$$

where : $\hat{Y}_{HT}$ - estimator of the total of variable Y,

N - general population size,

y_i - value of the variable of interest for the *i*-th unit.

- GREG estimator $\hat{Y}_{GREG} = \hat{Y}_{HT} + \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)' \hat{\boldsymbol{\beta}}$ (8)

where: $\hat{\mathbf{X}}_{HT} = \sum_{i \in s} w_i x_i$ - vector of direct HT estimators of auxiliary variables *x*,

$\mathbf{X}$ - vector of totals of auxiliary variables *x*,

$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \left(\sum_{i \in s} w_i \mathbf{x}_i y_i\right)$ is estimated using the method of weighted least squares using design weights,

w_i - design weights.

Values of the classic GREG estimator differ from estimates obtained using the estimator of interest - $\hat{Y}^0_{GREG}$. The differences are due to the fact that estimator $\hat{Y}^0_{GREG}$ does not take into account all sampled units and ignores those for which auxiliary variable ‘z’ is zero.

If the constant is omitted, the resulting estimator is called a ratio estimator (Hedlin, 2004).

3 Assumptions of the study

The empirical study included small companies (10-49 employees) conducting activity classified into section Trade. The response variable estimated in the model was Revenue obtained in June 2012. Information about the response variable came from the DG1 survey. The survey is conducted in the form of monthly reports submitted by all large and medium-sized enterprises and a 10% sample of small enterprises. (Dehnel, 2014). The following variables were selected as auxiliary variables ‘x’ and ‘z’: the number of employees from the social insurance register (ZUS) and revenue form the tax register for December 2011. The decision to use lagged variables was motivated by technical limitations involved in surveys conducted by the Central Statistical Office, namely the delay between the release of administrative data for purposes of official statistics. In the analysed model it is assumed that each auxiliary variable can be used both as auxiliary ‘*x’* and ‘*z’* (on condition that auxiliary ‘z’ cannot take zero values). Given the above, in the final approach, auxiliary variable ‘z’ was taken to be the number of employees.

Estimation was conducted for domains defined by cross-classifying provinces with the section of business activity according to the Polish Business Classification (PKD).

4 The method of evaluating precision

Precision and accuracy of estimates were evaluated using the bootstrap method. 1000 bootstrap samples were drawn, which were then used to estimate *Revenue* for June 2012 for domains of interests. Estimation efficiency was assessed using the coefficient of variation of the estimator (Bracha, 2004):

$$CV\left(\hat{Y}_d\right)=\frac{\sqrt{Var\left(\hat{Y}_d\right)}}{E\left(\hat{Y}_d\right)}=\frac{\sqrt{\frac{1}{999}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d}-\hat{Y}_d\right)}}{E\left(\hat{Y}_d\right)}\,. \tag{9}$$

In order to estimate bias, it is necessary to know values of the estimated parameters for the general population. Since this information was not available in the survey it was estimated indirectly, based on data from the tax register for December 2012. It was assumed that the following relation holds true: the ratio of *Revenue* reported in tax statements submitted by companies at the domain level to the value of *Revenue* reported in the DG1 survey is constant.

$$\frac{Revenue_{XII2012}}{Revenue\_DG1_{XII2012}}=\frac{Revenue_{VI2012}}{Revenue\_DG1_{VI2012}} \tag{10}$$

$Revenue_{XII2012}$ ($Revenue_{VI2012}$) – value of revenue reported in the tax register in December 2012 (VI2012)

$Revenue\_DG1_{XII2012}$ ($Revenue\_DG1_{VI2012}$) – value of revenue reported in the DG1 survey in December 2012 (VI2012). This approach made it possible to determine approximate values of *Revenue* for June 2012.

5 Conditions of estimates and assessment of their quality

The first step of the analysis was to examine distributions of companies depending on the variables of interest. The coefficient of variation varied from 47% to 649%. The distributions were strongly asymmetric, with skewness coefficients ranging from 0.6 to 17.1.

The hypothesis of homoscedasticity was verified using the White test and the Breusch–Pagan test. For most domains of interest test results confirmed the validity of the hypothesis about the variability of the random component, see Table 1. This in turn justified the use of the above mentioned GREG estimators modified to account for variable ‘z’.

The estimates were assessed both in terms of accuracy and precision. The point of reference for the assessment of precision were the estimates obtained using classic, direct estimators: the HT estimator and the GREG estimator, including its ratio estimators. Based on CV values as a measure of efficiency, it can be seen that the HT estimator exhibits the greatest degree of variation (see Table 1). Classic GREG estimators and transformed GREG estimators are characterized by less variation.

Test:	**White**		**Breusch-Pagan**	
Province	statistic	p-value	statistic	p-value
dolnośląskie	121.9	0.0000	86.7	0.0000
kujaw.-pom.	46.13	0.0000	35.05	0.0000
lubelskie	49.7	0.0000	1.95	0.3771
lubuskie	80.34	0.0000	37.33	0.0000
łódzkie	82.18	0.0000	34.9	0.0000
małopolskie	136.1	0.0000	6.36	0.0417
mazowieckie	169.7	0.0000	142.6	0.0000
opolskie	42.65	0.0000	36.43	0.0000
podkarpackie	33.91	0.0000	18.09	0.0001
podlaskie	33.33	0.0000	16.64	0.0002
pomorskie	40.15	0.0000	36.32	0.0000
śląskie	97.41	0.0000	82.98	0.0000
świętokrzyskie	23.45	0.0003	6.18	0.0456
warm.-maz.	48.97	0.0000	25.96	0.0000
wielkopolskie	121.5	0.0000	74.83	0.0000
zachodniopom.	121.1	0.0000	113.9	0.0000

Table 1. Results of the White test and the Breusch–Pagan test for heteroscedasticity for section *Trade* across provinces.
Source: based on data from the DG-1 survey.

On closer analysis, it can be seen that estimation precision of the GREG estimator depends on the sample size. It is usually bigger in domains that have more representation in the sample. In most domains of interest, the CVs of the transformed estimators, regardless of the value of the γ coefficient, are more or less similar and slightly higher than the classic GREG estimator. Moreover, it can be noticed that, as variability and asymmetry in a given domain increases, the gain in precision resulting from using each of the GREG estimators improves.

The precision of estimating *Revenue* of companies was assessed in reference to ratio estimates given by formula 10. In addition, to provide a more complete assessment, transformed estimators were compared with the HT estimators and the classic GREG estimators, see Fig. 2. The results indicate that the inclusion of variable ‘z’ in the model yields a considerable improvement in estimation precision, especially compared to the HT estimator, but also with respect to the classic GREG estimator.

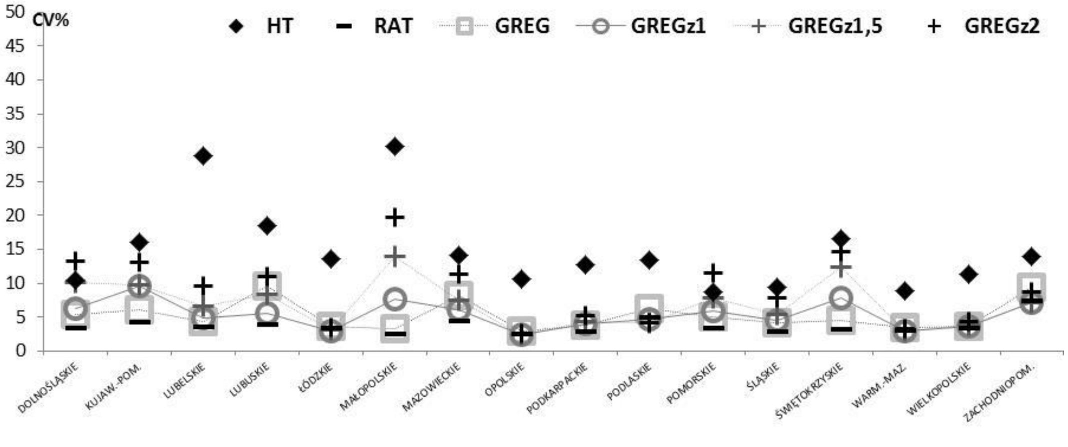

Fig. 1. Estimation precision (CV) for section *Trade* across provinces.

Source: based on data from the DG-1 survey and administrative registers.

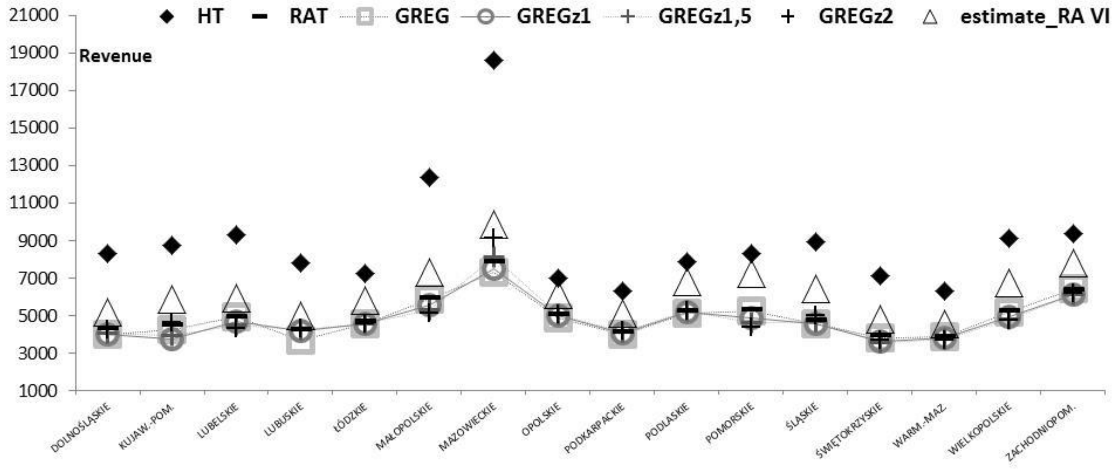

Fig. 2. Estimated revenue for June 2012 across provinces - section *Trade.*

Source: based on data from the DG-1 survey and administrative registers.

The HT estimator overestimated *Revenue* for nearly all domains of interest, while the GREG estimator tended to underestimate it. In contrast, estimates obtained by the transformed estimator are closest to the reference values. The largest discrepancies between parameter estimates produced by the different estimators can be observed for domains characterized by the largest variation and asymmetry of variables used in the model.

Figure 3 includes histograms for two selected provinces showing distributions of bootstrap estimates obtained by applying the studied estimators. The estimators using auxiliary variables from registers are less biased. As the γ parameter increases, the distribution of estimates tends to approximate the reference value marked in red.

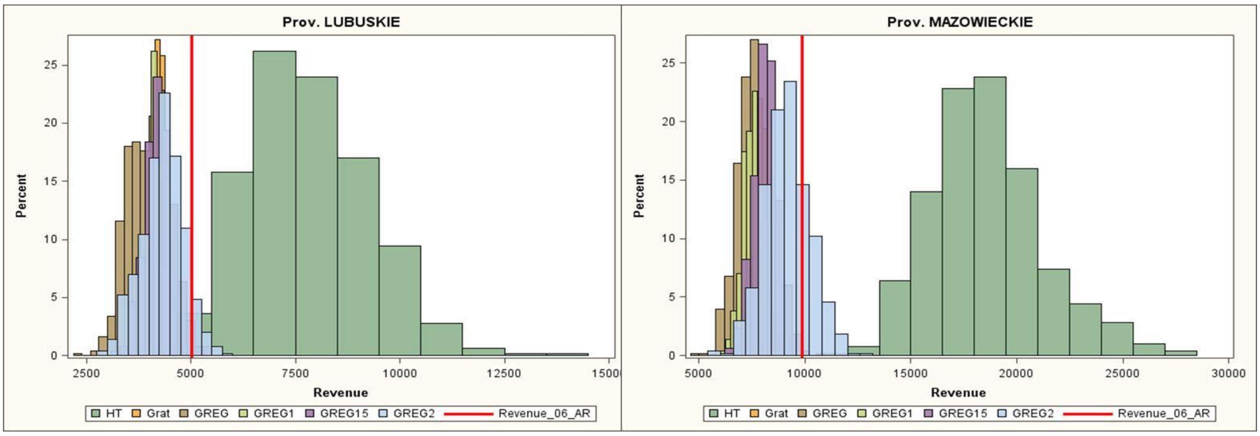

Fig. 3. Distribution of estimates for selected provinces for section *Trade*.

Source: based on data from the DG-1 survey and administrative registers.

Conclusion

The results obtained in the study lead to the following conclusions:

- the inclusion of auxiliary variables in the GREG estimator has considerably improved estimation precision compared to the HT estimator;
- the modified estimation method yields a greater gain in precision and accuracy for domains with greater variability and asymmetry;
- improvement in estimation precision for the estimators using reciprocal transformation depends on the value of the γ parameter. A considerable gain can be achieved if an appropriate model is selected for a given domain; the drawback of this approach, however, is that the application of modified GREG estimators for a very large number of domains is time-consuming and demanding.

Acknowledgements

The project is financed by the Polish National Science Centre, decision DEC-2015/17/B/HS4/00905.

References

Dehnel, G. (2014). Winsorization Methods in Polish Business Survey. *Sampling methods and estimation*, 97, 97.

Gamrot, W. (2014). Estimators for the Horvitz-Thompson statistic based on some posterior distributions. *Mathematical Population Studies*, *21*(1), 12-29.

Hedlin, D., Falvey, H., Chambers, R., & Kokic, P. (2001). Does the model matter for GREG estimation? A business survey example. *Journal of Official Statistics*, 17(4), 527-544.

Hedlin, D. (2004). Business survey estimation. *Statistiska centralbyrån*.

Lehtonen, R., Särndal, C. E., & Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7(3), 649-673.

Lehtonen, R., Särndal, C. E., & Veijanen, A. (2008, August). Generalized regression and model-calibration estimation for domains: Accuracy comparison. *In workshop on Survey Sampling Theory and Methodology* (pp. 25-29).

Markowicz, I. (2014). Business Demography-Statistical Analysis of Firm Duration. *Transformations in Business & Economics*, 13(2B), 801-817.

Rao, J. N., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Schmid, T., Tzavidis, N., Münnich, R., & Chambers, R. (2016). Outlier Robust Small-Area Estimation Under Spatial Correlation. *Scandinavian Journal of Statistics*.

Todorov, V., Templ, M., & Filzmoser, P. (2011). Detection of multivariate outliers in business survey data with incomplete information. *Advances in Data Analysis and Classification*, 5(1), 37-56.

Zhang, L. C., & Hagesæther, N. (2011). A domain outlier robust design and smooth estimation approach. *Canadian Journal of Statistics*, 39(1), 147-164.