The use of administrative tax data as an estimation strategy

Sihle Khanyile
Statistics South Africa & University of Michigan
sihlek@umich.edu

Abstract

Statistics South Africa has only used tax administrative data in their imputation strategy and not as part of an estimation strategy and has therefore not kept pace with international best practice. This paper investigates how the "value added tax" administrative data affects the precision of estimates, the Motor Trade Survey is used as a prototype for other business Surveys in Statistics South Africa. Descriptive summaries of the coefficient of variation and design effect for a 3 months series were evaluated contrasting estimates that incorporated the admin data and those that did not. In line with theory It was found that the estimates that incorporated the tax administrative had improved precision.

Key Words: administrative data, value added tax data, Greg estimation, calibrate

Introduction

To draw its samples for economic surveys Statistics South Africa uses the business sampling frame (BSF) that contains businesses registered at the South African Revenue Service (SARS) for value added tax (VAT). The economic surveys in Statistics South Africa have only used the tax administrative data in their imputation strategy and not as part of an estimation strategy. Studies have shown that the use of calibration to incorporate tax data during estimation (estimation strategy) can improve the precisions of estimates given that the survey variable of interest is highly correlated to the auxiliary variable (Renaud & Larouche, 2016).

There are two factors that make the examination of the use of tax administration data important and urgent: One, the current context of decline in budgets due to austerity policies by the state consequently leading to decline in sample sizes and two, the fact that during the survey reference period the sample is not updated notwithstanding that in the market economy changes are taking place (e.g closure of some businesses).

This paper will explores if the use of VAT data during estimation using the GREG estimator will improve the total monthly sales estimate of the Motor Trade Survey. To evaluate the efficiency comparison will be made between monthly estimates of the months: April 2017 to June 2017 before and after calibration (GREG). The Motor Trade Survey (MTS) estimates are used to compile the Gross Domestic Product and for comparative industry performance.

Method

The MTS is conducted monthly, for the months April 2017-June 2017 Questionnaires were sent to a sample of 853 enterprises from a population of 10 857 enterprises. The sample was selected through a multistage selection process, the first stage is the stratification of the enterprises by their Standard Industrial classification of All Economic activities (SIC) at a four digit level. There are seven strata (domains) at the first stage: Wholesale sale of Motor vehicles (6311) making 7% of the sample, Retail Sales of Motor Vehicles (6312) making 35% of the sample, Maintenance and Repair of Motor Vehicles (6320) making 17% of the sample, sales of new parts and accessories (6331) making 13% of the sample, sales of used parts and accessories(6332) making 4% of the sample, sales, maintenance and repair of motor cycles and related parts and accessories(6340) making 6% of the sample and Retail sale of automotive fuel (6350) making 18% of the sample.

The second stage is a stratification by measure of size (Turnover) within each classification (Domain) and the third stage involves the simple random sampling of enterprise (Swedish Jales sampling technique) within each size group (stratum by size) the size group cut offs are presented in the table below:

Measure of size classes (Rand)

Enterprize Size	Size group	Lower limits	Upper limits
Very small	4	1 780 071	18 000 000

Small	3	18 000 001	85 500 000
Medium	2	85 500 001	175 500 000
Large	1	175 500 001	

Size group 1 enterprises were fully enumerated (self-representing) which make 54% of the total sample. The Non self-representing enterprises (NSR) make up 46% of the total sample.

Neyman optimum allocation was used to allocate the sample size to each stratum and to account for this complex sampling design, weights for those strata (NSR) are were calculated which are the inverse ratio of the sampling rate (design weights). These design weights which are constructed to make the sample representative of the population were adjusted to account for non-response, we use these weights in our design object for R.

The variables of interest for my analysis are the total sales and the VAT Turnover auxiliary variable from the frame.

My statistical analysis will first examine the correlation between the variable of interest (monthly sales) and the auxiliary variable (VAT) through a scatter plot and Pearson's correlation coefficient using svycor() (from jtools package) in R to account for the design features. I used the survey design object in r from survey package to create a design object which account for the complex sample design features.

To calibrate using the GREG estimate and To account for this calibration during variance estimation I used the function calibrate(). I created two subset groups in order to overcome the singularity problem I was encountering using the calibration function. One subset group entailed the self-representing enterprises (SR) the other the Non self-representing enterprises. Since SR enterprises are fully enumerated their variance is zero. I compare the estimates of the NSR enterprises before and after calibration.

To use the GREG estimation the two variables $\mathrm{y_i}$ and $\mathrm{x_i}$ must be available for each sample unit ($\mathrm{y_i}$ is measured during the survey, and $\mathrm{x_i}$ normally comes from administrative data). The distribution of $\mathrm{y_i}$

(Monthly sales) and $\mathrm{x_i}$ (VAT Turnover). In addition, the control totals of the variable ▯ must also be known.

The GREG estimation for the total of y can be written as

$$\mathrm{T_{yGREG} = \ t_y + (t_x - \ t_x)^T B}$$

where $\mathrm{t_y = \ \sum_s d_i\, y_i}$ is the estimator of the total based on input weights, the superscript T represents the transpose of the specified vector, $\mathrm{t_x = (t_{x1}, \dots \dots, t_{xp})^T}$ is the p X 1 vector of the population (or control) totals of the p auxiliaries using the number of rows by the number of columns matrix notation, $\mathrm{t_x = \ \sum_s d_i\, x_i}$ is the estimate of the totals of the x's based on $\mathrm{d_i}$ weights, $\mathrm{x_i}$ is the p x 1 vector of auxiliary values for the $\mathrm{i^{th}}$ sample unit (Valiant et al 2013: 361).

To answer my research question I evaluated the descriptive summaries of the estimates for the months April 2017-June2017, these estimates were calculated using the svytotal() function accounting for the complex design features stated above. I contrasted the computed estimates before calibration and the estimates after calibration, comparing the Coefficients of Variation (CV) and design effects.

The theory suggests that the observed CV should be lower after calibration, The GREG estimator takes advantage of the correlation that may exist between x and y. The greater the correlation, the more efficient the estimator will be in terms of variance (Valiant et al 2013:349)

Results

Figure 1 presents the scatter plot and Pearon's coefficient of variation between the outcome variable (sales) and auxiliary variable (VAT Turnover). The coefficient of variation is 0.82.

Figure 1: Scatter Plot between outcome variable (sales) and Auxiliary variable (VAT Turnover)

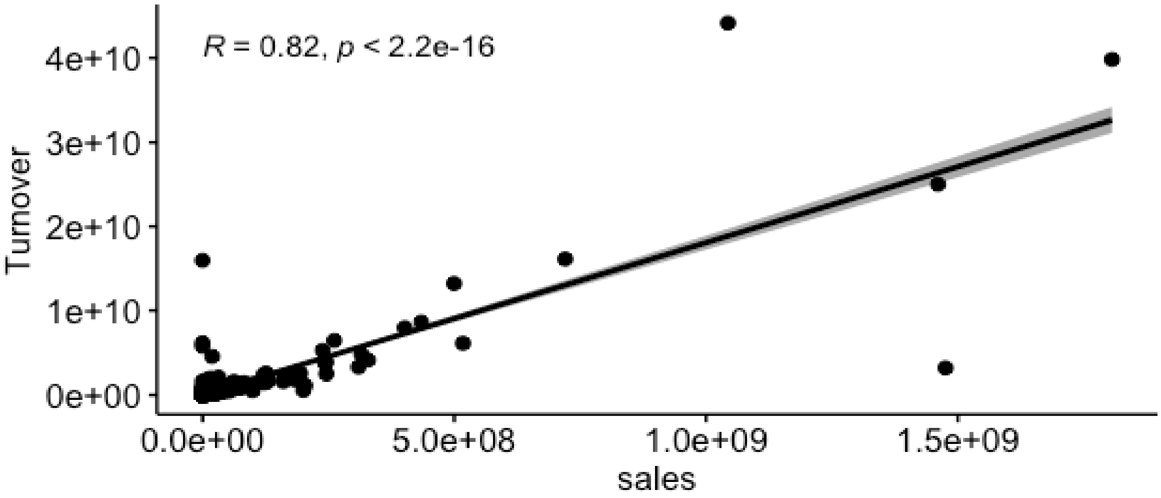

Table 1a and 1b present the descriptive summaries of the total population sales for the month of April 2017 before and after calibration respectively. The estimated sales in the population for the NSR enterprises is 2.0711e+10 before calibration and after 2.0815e+10 calibration. The total sales for the SR (fully enumerated) is 2.4475e+10. The CV dropped by 0.01 units from 0.06 before calibration to 0.05 after calibration.

As theory informs that stratification decreases variance, the complex design decreased the variance of the estimate by 0.4469 in contrast to simple random sampling design and calibration decreased it by 0.3303.

Table 1a: April 2017 Before calibration

Variable	Estimate	Standard Error(SE)	Coefficient of Variation (CV)	Design Effect
Total Sales(NSR)	2.0711e+10	1.3097e+09	0.06	0.4469
Total Sales(SR)	2.4475e+10			

Table 1b: April 2017 After calibration

Variable	Estimate	Standard Error(SE)	Coefficient of Variation (CV)	Design Effect
Total Sales(NSR)	2.0815e+10	1.1394e+09	0.05	0.3303
Total Sales(SR)				

Table 2a and 2b present the descriptive summaries of the total population sales for the month of May 2017 before and after calibration respectively. The estimated sales in the population for the NSR enterprises is 2.0295e+10before calibration and after 2.9333e+10 calibration. The CV dropped by 0.01 units from 0.07 before calibration to 0.06 after calibration.

As theory informs that stratification decreases variance, the complex design decreased the variance of the estimate by 0.5368 and calibration decreased it by 0.3346.

Table 2a: May 2017 Before calibration

Variable	Estimate	Standard Error(SE)	Coefficient of Variation (CV)	Design Effect
Total Sales(NSR)	2.0295e+10	1.3922e+09	0.07	0.5368
Total Sales(SR)	2.9333e+10			

Table 2b: May 2017 After calibration

Variable	Estimate	Standard Error(SE)	Coefficient of Variation (CV)	Design Effect
Total Sales(NSR)	2.2331e+10	1.2310e+09	0.06	0.3346
Total Sales(SR)				

Table 3a and 3b present the descriptive summaries of the total population sales for the month of June 2017 before and after calibration respectively. The estimated sales in the population for the NSR enterprises is 2.0319e+10 before calibration and after 2.2331e+10 calibration. The CV dropped by 0.02 units from 0.07 before calibration to 0.05 after calibration.

As theory informs that stratification decreases variance, the complex design decreased the variance of the estimate by 0.5561 and calibration decreased it by 0.3461.

Table 3a: June 2017 Before calibration

Variable	Estimate	Standard Error(SE)	Coefficient of Variation (CV)	Design Effect
Total Sales(NSR)	2.0319e+10	1.3691e+09	0.07	0.5561
Total Sales(SR)	2.8969e+10			

Table 2b: June 2017 After calibration

Variable	Estimate	Standard Error(SE)	Coefficient of Variation (CV)	Design Effect
Total Sales(NSR)	2.2331e+10	1.2037e+09	0.05	0.3461
Total Sales(SR)	2.2042e+10			

Discussion

The outcome variable and auxiliary variable are highly correlated as stated in the results section consequently the results of decreased variance are in line with the theory that has already been cited in the methods section. Incorporating tax data during estimating using calibration improved the estimate for all the 3 months(April 2017- June 2017), a 0.01 unit difference for April and May and 0.02 difference for June when comparing the CV before and after calibration. The design effects also corroborate the improved precision of the estimate. The study was limited by the inability to calibrate the SR enterprises although it does confirm the benefits that could be leveraged by the use of the already available auxiliary data in the frame.

An alternative study should explore the use of ratio estimation over a period of 12 months to assess the impact that the use of this admin data could as an estimation strategy (Renaud & Larouche, 2016).

References

Valliant Richard, Dever A. Jill & Kreuter Frauke, 2013, Practical Tools for designing and weighting Survey Samples, New York, Springer.

Renaud Martin & Laroche Richard, 2016, The use of Administrative Data in Business Surveys: The Statistics Canada Experience.