Average Monthly Wage Means Difference Between Administrative Sources Vs. Business Survey, Case in Albania

Elsa Dhuli
Institute of Statistics, Albania

Doi:10.19044/esj.2018.v14n31p213 URL:http://dx.doi.org/10.19044/esj.2018.v14n31p213

Abstract

This paper examines the differences resulting from calculating the means of Pay Roll records and personal revenues used as secondary data with results from business survey in an empirical study taking a panel. The use of “secondary data” as primary source for producing the official indicators is a challenge worldwide. In the past decades has also been considered as the way forward for raising productivity and reducing burden on businesses. If the Short Term Survey is sample survey the Pay Roll records are administrative data. The purpose for what they are gathered is different. But both could be used for providing statistical indicators. In this paper the panel not weighted data are taken into consideration where the same business is analyzed from two related sources. The paired t-test is used to compare the values of means from two related sources. In those conditions the difference between the means of the two sources is unlikely to be equal to zero.

In this study the hypothesis test is designed to answer the question "Is the observed difference sufficiently large enough to indicate that the alternative hypothesis is true?" What does it mean in our case study the answer which comes in the form of a probability - the p-value? The paper shows some interesting findings about the means difference between the two sources within a year. The differences resulting from the conducted analysis come as a result of the definition used in both sources for the same indicator, errors in reporting and treatment of non-response in the survey and administrative data source, coding errors.

Keywords: Average monthly wage, Pay Roll records, Quarterly Business Survey

Introduction

In the large and long literature on measurement error, most studies begin with data that are believed to contain errors and look for a way to quantify those errors. The goal is to account for and remove the effects of those errors. Defining data errors fundamentally requires some measure of truth: an objective standard by which the accuracy of the data can be judged. A common approach to this measurement error problem is to find a second data source that contains the "truth" and define errors in the first data set as the difference between the two sources. It is believed, however, that the assumption that some data contain errors while other data do not is fundamentally flawed. While the error-generating process may be different in the two sources, no source is likely to be completely error free (Abowrd and Stinson, 2013)".The gross wage collected from the short term survey on businesses could be used for estimating average monthly wage for employee. As well from pay roll records and personal income the average monthly wage in business level could be calculated. Employment and wages indicators for statistical purposes are provided from statistical survey in business level and from administrative sources. The statistical survey is referred to the Quarterly business survey, while the administrative source is referred to the pay roll records and personal revenue.

The 2016th year is used for analyzing the average monthly wage estimated from two sources. The business survey was carried out with quarterly frequency, from statistical office. The variables contained in quarterly survey are turnover, average number of employees, gross wages and salaries for reference period for economic activities covered in the survey. The key variable is the unique code of businesses. The non-response units are imputed by using different imputation methods. The administrative sources are used as auxiliary source for editing and imputation the missing value, as one the usage of administrative sources (Brackstone, 1987). The pay roll records of the contributions for social and health insurance and income tax from employment obtained from the General Tax Department is taken into consideration. The data contained in the pay roll data file for the individuals are age, gender, citizenship, occupation, contributor category, unique contributor code, and unique enterprise code. Year by year the quality of Pay Roll list is improved based on some administrative actions taken from the administrative authorities. Data collected or created by administrative authorities can be of different nature apart of them were the authority wants to produce its own statistics (Walgren, 2011). For the purpose of this paper it is studied the Pay Roll list of 2016th. However, some other actions are taken to edit and control the administrative source with regard to errors at different reporting periods, duplication records existence, and non-identification of employee citizenship, salary differentials, and mistakes in economic code. The Pay Roll records in individuals' level for the purpose of this study are aggregated in business level. The key variable is the unique code of businesses.

1. Pay Roll records and quarterly survey

1.1 Pay Roll records on a panel data

Administrative records, such as pay roll records, personal income, gathered for taxation purposes may be less susceptible to underreporting of income and can be inconsistent in periods when the tax law has been changed (Strudler, Petska, and Hentz, 2004). Pay Roll records are the records maintained by the employer and declared to the Taxation office for about three types of tax liabilities:

- Payment of Social Security Contributions (defined Contributions to the Employer, Employee Contributions and Supplemental Contributions);
- Payment of Health Insurance Contributions
- Employee Revenues

The terms of the declaration and payment of contributions are dependent on the status of the person, who is obliged to pay the contributions, these are determined in the relevant legal framework (Decision of Council of Minister, DCM 77/2015).The purpose of pay Roll records it's not for statistical issues. Pay Roll records is information ensured by administrative way that can be used as well as secondary auxiliary source for several purposes, such as the average wage estimate, the imputation of non-response unit, editing and validating data. In the dataset used are matched three types of variables derived from: original data file from Taxation office, additional variables from statistical office and derived variables. The main interest variables in this data file are the number of employees and the gross salary paid by the employee by the section of economic activity.

The data set got from taxation office it's not in the form and quality to be used directly for statistical purpose. The register must fulfil many important quality requirements at micro level, such income for individual level. (Walgren and Walgren, 2014). They keep errors as while extensive work with editing and correcting is needed. Errors related to different process such reporting, capturing, and transmission are detected and reflected in detail as follows:

- Capturing Errors occur when employers fill in the electronic form of e-filling and information is thrown incorrectly. These types of errors are partially captured through checking methods or corrected by the employer in the next period of declaration.
- Conceptual errors are mistakes that cannot be captured by the system and are eliminated through manuals and instructions available at the tax office.
- Mistakes in Data Transmission are errors that occur when copying databases and exports to flat files.

These types of errors are eliminated by using different controlling methods: Some administrative sources can contain duplicates. I'ts important that those duplicates are founded and deleted, (Walgren and Walgren, 2014). Duplicates are found by selecting key variables that uniquely identify the records. The control of duplicates is done with key variables: Person ID, Business ID, and Reporting Month. Changes of reported variables are related to the consistency of the variables reported in each period from taxation office. They need to be considered and can be partially preceded by following the guidelines published by the taxation office.

- Reporting Time is another control related to sustainability of the data by the transmission period at the statistical office. This is related to the fact that transmission of the same data sets in different periods of time create risk for avoid corrections.
- Coding Process is made by taxation office for their purposes for each contributor's category which corresponds to the category description, as an important variable to avoid mistakes in the case of employment status determination based on the description and coding of economic activity made by NACE Rev.2.
- Imputation is a process needed for increasing the quality of administrative source data, by addressing the missing values.
- Identification of outliers is another type of control for identifying extremely large or small values.

The errors identified are corrected, measured and reduced in order to have a good quality. For some type of errors the quality indicators are measured as the rate of imputation was relatively low, 1.45 per cent, the synthetic indicator for the rate of imputation according to the records confirm the good performance of the imputation process, while these added values is 0.07 per cent and the rate of the changed values is 1.38 per cent. Related to the coding at the level of economic activity, NACE Rev.2, Imputation Rate was 4.185 per cent, Editing Rate was 0.199 per cent, Addition Rate was 3.986 per cent, and Desalination Rate was 0.00 per cent, (Annex 1, Table 1).

For the purpose of calculating the average wage by businesses the pay roll dataset as administrative sources was aggregated according to business ID code at the business level, called Pay Roll list by businesses (PRLB). To eliminate errors in the economic code and to add some identification variables of the businesses the data set created was matched with Statistical Business Register (SBR).Variables such as economic code, status of the enterprise, ownership, legal form were taken by the SBR.

1.2 Quarterly Business Survey (Short Term Statistics)

Population of enterprises for quarterly survey is based on the Statistical Business Register (SBR), where the main criterion is the coverage area by economic activity at 4 digit level and the total number of employees. Each unit has its own unique identifier called ID business code. The Business Register is nowadays primarily maintained and updated on the basis of administrative records and other sources.

The fields of coverage of the quarterly survey are: Industry, Construction, Transport, Hotels, Restaurant, Wholesale and Retail sale, Post-Communications, Computer Services, and Engineering Services, Travel agency, (Annex 1, Table 2). Sampling is selected on the group of enterprises with 1-19 employed, while the other with over 20 employees is listed. The survey is conducted with the enumerators and the data are captured through optical reading (scanner) by using verifier. The publication deadline for the reference quarter is 75 days after the reference quarter. The economic activity code after the survey for small enterprises remains the same with the code in the Statistical Business Register (SBR), and the changes are reflected in the following year.

The quarterly survey does not cover the entire economy. Information on Total Number of Employed includes the number of self-employed and number of employees. The quarterly survey does not identify the self-employed (it comes out as the total number of employees with paid employees), employee status, or other indicators such as gender, age, citizens.

The average monthly wage is calculated as the Pay Roll ratio and the number of employees. Monthly average wage is estimated by gross wage and the average number of employees. In the case of non-response, as a main source for imputation Pay Roll records at business level in monthly period. The mistakes in survey data sets are edited and checked.

2. Panel data File: Pay Roll list by businesses in Quarterly Survey

For the purpose of this study the two data sets were matched in one. The quality assessment in the context of matching of two data source needs a process approach. Each of the steps (the quality and the coherence of data sources, modelling techniques, matching/imputation algorithms) has a large impact on the quality of results. The integration process between the dataset created in business level PRLB and the SBR was somewhat complex because the Business Register has two parts, the Enterprise Register and the Local Unit Register. The Enterprise Register was taken for this study, as Quarterly Business Survey is conducted in enterprise level. For PRLB, the names and economic activity found on the Enterprise Business Register to correspond to the names and economic activity of business employers reported in the Pay roll records. One of the main controls done in the pay roll records was the direct link of the individuals with enterprise code. This was important for aggregating individual register in enterprise level. The matching process between two sources it's not convenient as the change in methodology has a direct impact on the estimates derived from each source. In fact, QBS estimates are not particularly focused on estimating average wages and the lack of detailed employee information (e.g. full-time or part-time employment, dual employment, contribution category etc.) makes it is impossible to apply the same methodology as the one used to estimate the average salary from the administrative source. But in our study the statistical units are taken into consideration with same behavior.

One major factor that facilitates the statistical use of administrative data records is the application of unified identification systems across different sources. Even though, enforcing rules and conducting controls was needed. Editing is the systematic work to find obvious and probable errors.(Walgren and Walgren, 2014). The new registry created is generated by using variables from three sources: Pay Roll records, Quarterly Business Survey (QBS) and Statistical Business Register (SBR). The new panel data created as register is cleared of missing values for both sources. The created panel is long as it's contains 4504 enterprises. The main fields are Business ID code, Economic Activity Code, Enterprise Status, Legal Form, from SBR. The main variables analyzed such as Number of employed, Number of employees, Gross Wages, number of self-employed, are generated from both sources for each quarter. The average monthly wage is calculated as a ratio of gross wage by number of employees for the reference quarter. As comparison base for the wage the minimum wage approved by law is taken, (No399/2017).

2.1 Empiric analyses of the monthly average wage from both sources

The objective of this study was to compare the means of difference of two paired samples, administrative data and survey data, to see if there is a difference between the averages (means) of the two measures. This study was performed on long series, 4504 cases. A paired *t*-test design is used as we have two measurements coming from the same panel. Observations which are measurements are often analyzed by the *t* test, a method which assumes that the data in the different groups come from populations where the observations have a normal distribution and the same variances. For this reason we compared the average wages of 4504 cases with values > 0. Also the outliers are studied. Three outliers have been identified in the Pay roll records compare with QBS with the scatter plot which are removed from the study (Annex1, Figure1 and Figure 2).

Some average wages are outside the border due to the economic activity where the presence of foreign citizens is operating or High tech activities were the average wage is larger than the other activities in the market. However, as the results are not too extreme to justify this cut off from the records it is considered better to keep them. To answer the research question on the differences of the means of average wages (from administrative source and quarterly survey) has been used the differences means controls in paired choices.

Difference between two averages, exists (Annex1, Table 3). The probability of the existence of differences in the 95 per cent confidence interval of the difference is 0.003, so the value of p<0.05.As the statistical significance value for two paired is determined by looking at the p-value, in our case this means that there are at least 0.003 chances that the difference means of the two populations are equal. The lower the p-value, the lower the probability of obtaining a result like the one that was observed if the null hypothesis was true. Thus, a low p-value = 0.003 indicates decreased support for the null hypothesis.

To draw a meaningful conclusions, the means of differences with the same method is done in aggregated level by economic activities in section level, (NACE Rev.2, Annex 1, Table 2) for the activities covered in QBS.

In practical significance the probability of the existence of differences in the 90 per cent confidence interval of the difference is 0.063, so the value of p< 0.10. It means that in the aggregated level the margin error is higher than in micro level when the two parametric values are compared. (Annex 1, Table 3).

This method was used for comparing the average wage from two different data sources in the same conditions where hypothetically should be the same. The variables definition even they came from different data sources they matched in order for them to be adequate.

Conclusion

The analysis conducted on the panel created with two data sources taking into consideration the average monthly salary, conclude that although in the same behavior of enterprises, the average monthly salary between the two sources is different. One of the reasons could be the free declaration on quarterly business survey of the respondents, while the information declared in administrative authorities is obligatory. Individuals who declare in tax authorities do not know about the economic activity and about the policy that businesses followed. **The first conclusion** is that the presence of significant differences between two sources gives us the possibility of using administrative sources for measuring average monthly wage from tax authorities even in panel data set. As the changes in the survey methodology and administrative data source methodology has a direct impact on the estimates derived from each source. The errors are present in both sources and in matched register, in detailed and in aggregated level. In addition, QBS estimates are not particularly focused on estimating average wages as the purpose of it is to measure the gross wage tendency in businesses and as well the lack of detailed employee information (e.g. full-time or part-time employment, self-employment, foreigner employment, etc.) makes it is impossible to apply the same methodology as the one used to estimate the average salary from the administrative source. And on **the second conclusion** comparison of two means differences in two pairs gives the possibility to measured informality in declaration of administrative information from businesses. **The third conclusion** is that as both sources are in system and some models for measuring errors are in place for better analyses and better understanding a metadata for each of them play important role. As the administrative data can be changed by administrative authority the metadata for administrative sources is more important than for survey data. The study should go further for measuring type of errors in both sources, differences of wages by economic activities, quality indicators of administrative sources.

References:

1. Albania, Decision of Council of Ministers decision no.77 date 28.01.2015, "*For compulsory contributions and benefits from the social security system and health care security*"; Published at:https://www.tatime.gov.al/eng/c/6/73/social-security-and-health-care-contributions (In Albanian)
2. Atkinson, Anthony, B. and Andrea Brandolini, (2001)."*Promise and Pitfalls in the Use of "Secondary"* Data-Sets: Income Inequality in OECD Countries As a Case Study."*Journal of Economic Literature,* 39(3):771-799.
3. Brackstone, G.J. (1987) "*Issues in the use of administrative records for statistical purposes*", Survey methodology, vol. 13, n. 1, pp 29-43
4. *Evaluating the foreign ownership wage premium using a difference-in-differences matching approach*, University of Nottingham, UK
5. John M. Abowd and Martha H. Stinsony, September 6, 2012, "*Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data*", Pages 97-112
6. Kirkman, T.W. (1996) Statistics to Use: Kolmogorov-Smirnov test.
7. H. B., J. M.Kennedy, S. J. Axford, and A. P. James, "*Automatic Linkage of Vital Records*," *Science* 130 (1959), 954–959.
8. Michael Strudler, Tom Petska, and Lori Hentz, 1979-2014, "*Analysis of the Distributions of Income, Taxes, and Pay Roll Taxes via Cross Section and Panel Data*, 1979-2004. "
9. N. Anderson, P. Hall, and D. Titterington. (1994). "*Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates*". *Journal of Multivariate Analysis,*. Pg.50:41–54
10. P. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society B*, 67(4):515–530, 2005.
11. S.Wallenstein, C. L. Zucker, J L Fleiss, "*Some statistical methods useful in circulation research*", July 1, 1980, pg.1-9

12. Statistics Notes: *Comparing several groups using analysis of variance*, BMJ 1996; pg.312: https://doi.org/10.1136/bmj.312.7044.1472 (Published 08 June 1996)

13. “*Use of Register and Administrative Data Sources for Statistical Purposes*”, Best Practices of Statistics Finland, Helsinki – Helsingfors 2004

14. INSTAT, (2016), Short term statistics publicationwww.instat.gov.al/Indystry;

15. Structural Business Survey publication, INSTAT, 2017, www.instat.gov.al/Indystry;

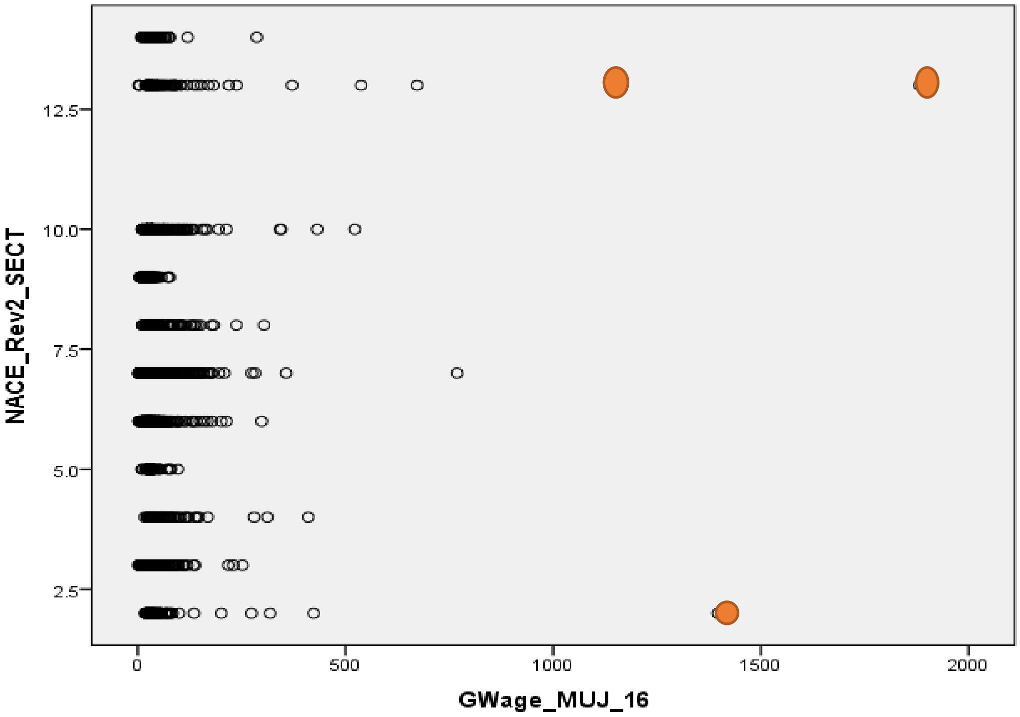
16. Statistical Business Register, INSTAT, 2017,www.instat.gov.al/en/themes/industry-trade-and-services/business-register;

17. R.Mytollari,(2017),Difference-between-survey-and-administrative-data“,INSTAT http://instat.gov.al/en/publications/research-magazine/

Appendix:

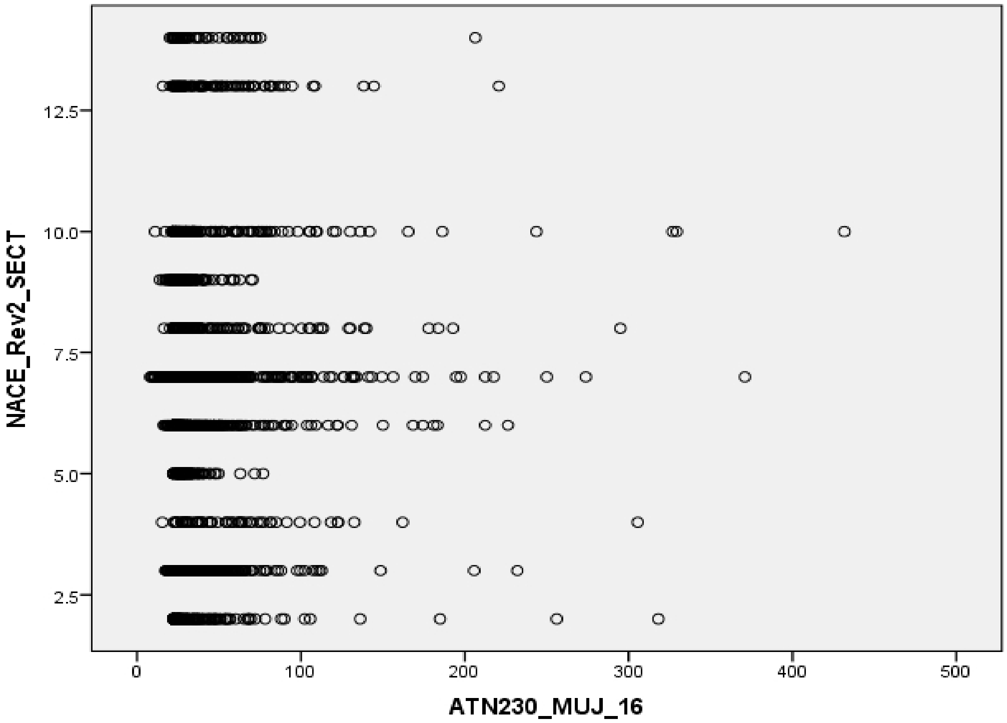
Annex 1

Figure 1 Scatter plot - Outliers detected in Gross monthly Wage in Pay roll list



***GWage_MUJ_16=GWage_PRL_16**

Figure 2 Scatter plot - Gross Monthly Wage in QBS



ATN230_MUJ_16=QBS_M_1 6

Table 1 Quality indicators of Pay roll records

Imputation scale (I) (%)	1.45
Addition imputation (Ia) (%)	0.07
Eliminated imputation (Ie) (%)	0
Measurement imputation (Im) (%)	1.38

Table 2 Economic activities covered in panel

NACE Code	Description
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam
E	Water supply; sewerage, waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I (55.1)	Hotels
J (58;61;62)	Information and communication
M (71)	Architectural and engineering activities
N (79)	Travel agency

Table3 Paired Samples Test Average Wage from PRLB compare with QBS

	Paired Differences (Mean)	t	p-value
Wage_PRLB_M_QBS_M_16(Thousand ALL)	0.822	2.949	0.003**
Wage_PRLB_M_QBS_M_16_Sections (Thousand ALL	2.763	2.094	0.063*

**5 percent
*10 percent

Note: The views herein are those of the individual and shall not to be taken to reflect the official opinion of Statistics Albania.