Estimation of Number of Persons Per Household Based on Characteristics of Consumption Items - utilization of big-data to improve the Consumption Trend Index in Japan-

Anri Mutoh (amuto@nstac.go.jp)
National Statistics Center, Japan

Masayo Yamashita (amuto@nstac.go.jp)
National Statistics Center, Japan

Yoshiyasu Tamura (amuto@nstac.go.jp)
National Statistics Center, Japan

Masahiro Matsumoto (amuto@nstac.go.jp)
National Statistics Center, Japan

ABSTRACT

The article suggests the possibility of utilizing big-data held by companies, integrating it with the data of official statistics. Official statistics agencies in Japan have sought to develop a Consumption Trend Index (CTI) by cooperating with academic researchers and companies as a provider of the big-data. One of the important roles of the CTI is to more accurately indicate the trend of one-person household consumption, therefore, the big-data is expected to reinforce existing official micro-data, especially one-person household. However, the obtainable big-data seldom includes the number of household members, and needs imputation of the missing value. Therefore, we estimate the number of members in each household according to the characteristics of consumption items in the Japanese traditional household expenditure survey. We used logistic regression with an L1 penalty (Lasso regression) for the analysis, with each type of household as the response variable and purchase items as the explanatory

variable. As a result, since one-person households and two-or-more-person households are identified by their purchasing tendencies, so the household characteristic become evident.

Keywords*: Consumption trend, household accounts, statistical imputation, logistic regression, LASSO, R package 'glmnet'*

JEL classification*: D13, D16, D90, P44, Z13*

1. INTRODUCTION

1.1 Big-data for Consumption Trend Index

We researched the utilization of big-data for official statistics. Since 2017, in Japan, the Statistics Bureau, Ministry of Internal Affairs and Communications, Statistical Research and Training Institute, and the National Statistics Centre have begun to research the development of a novel Consumption Trend Index (CTI) by cooperating with professors and commercial companies as the data holders (Statistics Bureau, 2017, 2018a). The CTI is an index that enables consumption trends to be grasped quickly and comprehensively. There are two types of CTI, for macro-level (CTI Macro) and micro-level (CTI Micro) (The Consumer Statistics Division of Statistics Bureau, 2018). The CTI Macro provides an early estimate of the monthly trend in the Household Final Consumption Expenditure of GDP. In contrast, the CTI Micro indicates the monthly trend in household average expenditure by major consumption items. In order to further improvement of the CTI, our research group plans to utilize big-data held by companies as a part of the input data of the CTI. We particular deal with the fusion of big-data and the source of CTI Micro in this paper.

The utilization of big-data as the source of CTI Micro is expected to reflect the consumption tendency of a one-person household more accurately. In Japan, even though of one-person households account for about 1/3 of the population (Statistics Bureau, 2018b), it is difficult to survey the one-person household in a Family Income and Expenditure Survey (FIES). One item of published evidence for the difficulty, which is a little old, is case of paradata research for the FIES by Hamasuna (Hamasuna, 1980). According to the paradata research, one-person households tended to be absent and need a lot of revisiting for the survey. This trend is considered to remain even in the 2010s. In order to deal with this difficulty, the source of CTI Micro consists of the Single Household Expenditure Monitor Survey, in addition to the FIES and the Survey of Household Economy. The big-data becomes a source of information to reinforce these surveys.

1.2 Details and issues of big-data for the CTI micro

Data of loyalty programs and data of online personal finance software are considered as usable big-data for the CTI Micro. Their advantages are 1) the ability to automatically and instantly obtain enormous amounts of data, 2) that the items of data correspond to a part of consumption items in the FIES (namely it is a proper subset), and 3) that the data consists of several samples whose unit is a user of loyalty program or personal finance software.

These big-data include information of the user's individual age and sex, however, they have the issue that they rarely include household information; the number of household members of the samples are unclear. Since the input data of CTI micro consists of the samples whose unit is household, the big-data need imputation of the missing value: the number of household members. As mentioned above, it is important but difficult to survey the one-person household for the FIES, thus at least data on one-person households have to be identified and used.

1.3 Purpose

The purpose of this paper is to estimate the number of persons per household by the consumption items and to clarify the characteristics of every consumption item of the household type in order to impute the big-data and suggest the possibility of utilization for the CTI Micro.

2. RELATED STUDY

2.1 Big-data and Official Statistics

The researches in the field of economic or social systems using big-data have increased in recent years (Japec et al., 2015). This is the same trend in official statistics. Struijs mentioned that the opportunity of collaboration between official statistics agency and business and universities was increased associating with big-data research in National Statistical Institutes (at the Netherlands); and reviewed issues and challenges about using big-data in official statistics (Struijs et al., 2014).

Research on Consumer Price Indices(CPI) is especially active among the studies using big-data for official statistics. For example, Office for National Statistics (at the UK) has reported several articles that estimated experimental CPI using web scraping data; and 10,000 price data on the web are collected automatically per month and utilized as `the harmonized index of consumer prices` in the Federal Statistical Office (at Germany) (Blaudow and Burg, 2018). However, few studies use big-data as a part of official micro data.

2.2 Stochastic regression imputation methods

Statistical imputation is a part of the most important field in official statistics. In recent years, multiple imputation of missing values has been commonly used and its software is large in variety (Takahashi and Ito, 2013). In this paper we do not deal with multiple imputation, but stochastic regression imputation. Because it is possible to design regression models for imputation. Unlike ordinary missing value, they have a full reason for missing, and also have a highly reliable reference data as the FIES. The imputation by stochastic regression is appropriate for the purpose of complementing a structure of the FIES.

In this paper, we use logistic regression with the L1 norm as a model for imputation, but there are few previous studies using such model for stochastic regression on imputation.

3. METHODOLOGY

3.1 FIES data

The data for analysis were retrieved from the January 2010 FIES conducted in Japan. The FIES had two types of survey, for one-person households and for two-or-more-person households. There was a total of 700 one-person households, along with approximately 7,800 two-or-more-person households. Although the two types of survey were different, their contents were almost the same, comprising the demographic characteristics of the householder and family members, and the purchased items as represented by price amount or frequency.

We consider the elements of the response variables of estimation to be one-person households, two-person households, three-person households, or four-or-more-person households, because 90 percent of the two-or-more-person households were occupied by 2–4 person households. Five-person households accounted for only 9 percent of the total (see Table 1), yet they show little difference from the four-person households in terms of the total spent.

Number and percentage of each type of household in the FIES data

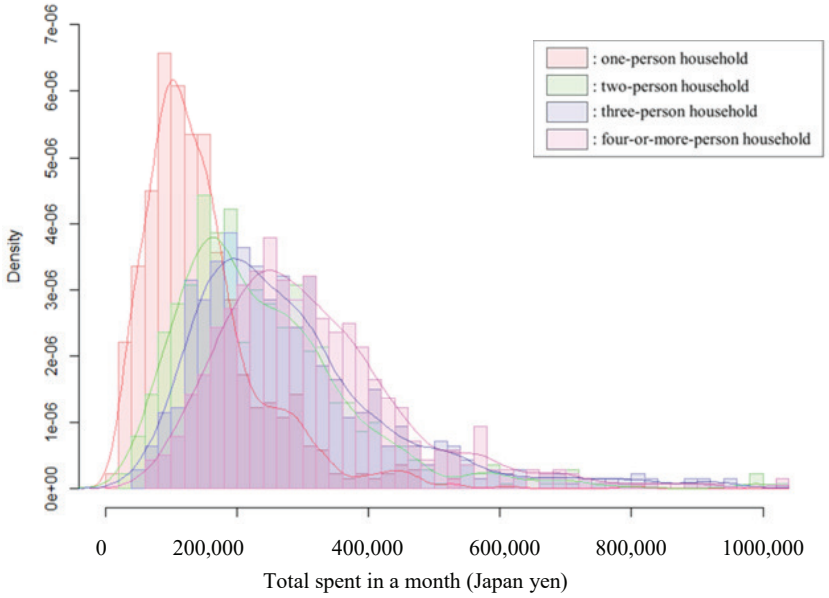
Table 1

one person	two-or-more persons	two persons	three persons	four persons	five persons	six persons	seven-or-more persons
700	7801	3165	2019	1719	676	182	40
(percentage)	100%	41%	26%	22%	9%	2%	1%

Although it is true that there is a positive correlation between the number of household members and the total amount spent, there is obvious overlap in the histograms based on the total spent by household size. Figure 1 shows the histograms and density plots with a uniform number of households. We are going to identify the items with less overlap among each household.

Histogram and density plot of total spent per household

Figure 1



3.2 Lasso regression

The FIES data contain almost 600 purchase items as explanatory variables, yet the actual observations contain many zero values. Therefore, we conducted a regression analysis that is proposed by Tibshirani (1996), so-called Lasso regression analysis. This performs simultaneous variable selection and minimization of the prediction error by adding L1 norm as a penalty. Since the L1 norm forms part of the parameters estimated as absolutely zero, it is possible to select the variables automatically for regression. Let y_i, X_{ji} $(i = 1, 2, \ldots, n,\ \ j = 1, 2, \ldots, p)$ be the response vector and a matrix of the explanatory variable, respectively, to give an $n \times (p + 1)$ data matrix. The problem thus takes the form of eqn [1]:

$$\arg\min_{\beta_0,\beta}(\boldsymbol{y}-\boldsymbol{\beta_0}-\boldsymbol{X\beta})^2 \; subject\ to\ t \geq ||\boldsymbol{\beta}|| := \sum_{j=1}^{p}|\beta|. \quad [1]$$

The t is a tu ing parameter. We are considering standardized data, and hence we omit β_0. Let the sum of the absolute values of regression coefficients become the L1 penalty. The $\lambda \geq 0$ is the regularization parameter by using the method of Lagrange's undetermined multipliers; thus, the Lasso regression model is defined as eqn [2],

$$\arg\min_{\beta}\{(\boldsymbol{y}-\boldsymbol{X\beta})^2+\lambda||\boldsymbol{\beta}||\}. \quad [2]$$

We aimed to estimate the types of households this time; thus, it uses logit as the link function. The environment of analysis is R 3.4.4 and we used the package 'glmnet' ver. 2.0-16. The estimation algorithm for the Lasso regression is the coordinate descent in this package, which is calculating differentiation for each numerical value of the norm and repeated updating (Friedman et al., 2010). The coefficient of the L1 norm was determined with 10-fold cross-validation so as to minimize misclassification error. The largest lambda, which minimizes misclassification error, was then selected within one standard error.

3.3 Data preprocessing

As data preprocessing, we first extracted the purchased items common to one-person households and two-or-more-person households. Next, we calculated the correlation between each item, and summarized those pairs with a correlation coefficient over 0.7. The reason for the preprocessing is that the variable selection by lasso regression become stable in the case of high correlation between explanatory variables. In addition, we applied the rank correlation as well as the linear correlation, but the linear correlation is better to summarize more items than the rank correlation. The pairs of highly correlated items result in 100 pairs, all of which have class and subclass relationships. For example, the pair {'Raw meat', 'Beef'} has a correlation coefficient of 0.726. In this case, the 'Raw meat' is a larger class including 'Beef'. In such pairs in a kind of hierarchical relationship, the subclass items are omitted for efficient modeling. If both the class and the subclass have similar behaviors, it is reasonable to leave the larger class that is affected by another subclass.

After the data preprocessing, the data still contained almost 500 purchased items as explanatory variables. This suggests that no class could

fully explain the features of all of its subclasses. Purchased items as represented by both the price amount and the frequency are processed in the same way.

4. RESULTS & DISCUSSION

4.1 Multinomial model

First, we consider a multinomial model in which the response variables are the four types of households: one-person households, two-person households, three-person households and four-or-more-person households. Table 2 shows that the prediction accuracy of the multinomial model is 0.657, which is poor. A similar level of accuracy is produced whether we use data represented by the price amount or by the frequency. This result suggests that it was difficult to identify items with less overlap even if estimated using the multinomial model.

The confusion matrix and the prediction accuracy of the multinomial model

Table 2

accuracy: 0.657		predicted			
		one-person	two-person	three-person	four-or-more -person
actual	one-person	288	409	2	1
	two-person	65	2833	173	94
	three-person	18	990	495	516
	four-or-more -person	12	379	253	1973

4.2 Binomial model

According to Section 1, we have to identify the data of one-person households. We consider binomial models whose response variables are dichotomous of one-person households and the others in order to indicate the items that are simply affected by the purchasing activity of multiple persons.

As a result, all the binomial models have prediction accuracy over 0.9, which is a similar result to the accuracy between the price amount and the frequency. Table 3 shows the confusion matrix and prediction accuracy. The

columns of the matrix show predicted items, while the rows show the actual items.

The confusion matrix and the prediction accuracy of the binomial model

Table 3

	predicted		accuracy
actual	two-or-more-person	one-person	
two-or-more-person	7547	254	0.942
one-person	240	460	
actual	three-or-more-person	one-person	
three-or-more-person	4552	84	0.969
one-person	84	616	
actual	four-or-more-person	one-person	
four-or-more-person	2572	45	0.972
one-person	48	652	

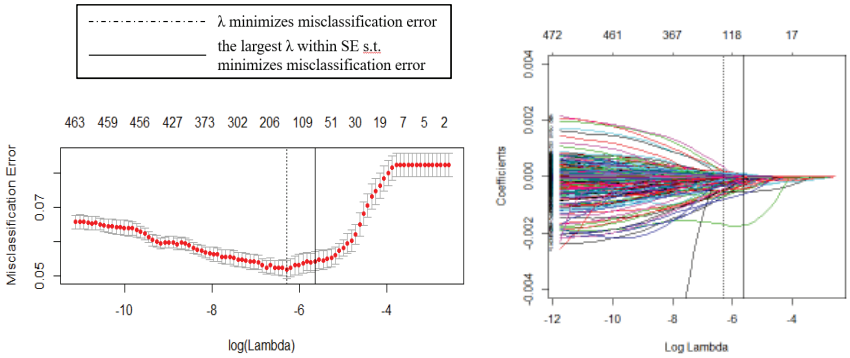
Figure 2 shows lambda coefficient plots in the one-person and two-or-more-person binomial model, and its solution paths. The lambda coefficient plots represent misclassification error by each lambda at the cross-validation; the solution paths represent the coefficients at the optimum lambda. There are two plots: plot-a is for the data represented by the price amount, and plot-b is for the frequency. The solid lines in the respective plots indicate the lambda that minimizes misclassification error. The dashed line indicates the largest lambda within one standard error that minimizes misclassification error. We select the optimal lambda as indicated by the dashed line, which is $\lambda = 0.0035$ for the model with the purchase price amount and $\lambda = 0.0042$ for the model with the frequency of purchased items. Each models left 84 and 114 variables.

The upper (or lower) 10 coefficients of the binomial models by one-person and two-or-more-person household are shown in Table 4 and 5. Each table shows the coefficients by the purchase price amount and the frequency of the purchased items. The dummy variables for the response variable are taken as 0 for two-or-more-person households and 1 for one-person households. Therefore, a positively loaded coefficient represents the items that characterize a one-person household, while a negatively loaded coefficient represents the items that characterize a multiple-person household.

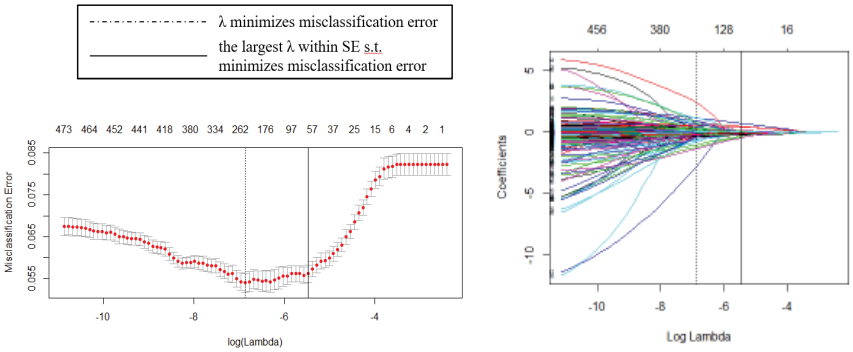
Lambda coefficient plots and its solution paths in the one-person and two-or-more-person binomial model

Figure 2

a. The lambda coefficient plots (left) and the solution paths (right) in the model with purchase price amount as the explanatory variable



b. The lambda coefficient plots (left) and the solution paths (right) in the model with frequency of purchased items as the explanatory variable



Based on Table 4, the items in third place and lower have a coefficient of less than 0.1. This suggests that the characteristics for identifying a single-person household are less obvious in the purchase price amount per item. However, focusing on the items with high coefficients, 'Drinking' has the largest coefficient in terms of the price amount for a one-person household, followed by 'Taxi fares'. This indicates that the relatively high unit prices for services and foods affect their identification.

On the other hand, 'Pocket money', 'Fuel, light & water charges' and 'Food' have large purchase price amount coefficients for two-or-more-person household, while 'Paper diapers' and 'Communication' also have large

coefficients. This indicates that the items proportional to the number of people and corresponding to the different stage of lives affect identification of multi-person household. For example, the variable 'Communication' represents a tendency for the number of contracts to increase as the number of household members increase, since, communication charges are fixed amounts and are proportional to the number of contract lines.

The coefficients of the binomial model by one-person and two-or-more-person household with the purchase price amount as the explanatory variable

Table 4

the one-person household		the two-or-more-persons household	
corresponding item	coefficient	corresponding item	coefficient
Drinking	0.15	Meat	-0.91
Taxi fares	0.11	Pocket money (Unexplained expenditure)	-0.88
Coffee beverages	0.08	Fuel, light & water charges	-0.62
Other admission fees & game charges	0.07	Paper diapers	-0.35
Women's nightwear	0.05	Gasoline	-0.33
Salad	0.05	Food	-0.33
Tea	0.04	Communication	-0.32
Other refreshments(Cafe)	0.03	Eggs	-0.30
Haircut charges	0.03	Oil, fats & seasonings	-0.30
Contact lenses	0.02	Soybean products	-0.26

According to Table 5, 'Rents for dwelling & land' and 'Rents for dwelling, issued houses' have large coefficient in terms of the frequency of items purchased by one-person households. This indicates the low rate of house ownership among one-person households. 'Coffee & cocoa', 'Salad' and 'Beer' have greater coefficients in the food category. It is nonessential grocery items with high unit prices in Japan.

On the other hand, with respect to two-or-more-person households, the items with large purchase price amount coefficients show similarly large coefficients in the purchase frequency. The daily necessities and items relating child care more affect the identification.

These variables are only a part of 84 of the model with the purchase price amount and 114 of the model with the frequency of purchased items. It means that at least 84 items are required to obtain the above estimation accuracy. Moreover, variables whose coefficients are estimated to be 0 by Lasso regression are unstable. It is not appropriate just because these 84 variables will be collected. There is still a long way for practical use.

The coefficients of the binomial model by one-person and two-or-more-person household with the frequency of purchased items as the explanatory variable

Table 5

the one-person household		the two-or-more-persons household	
corresponding item	coefficient	corresponding item	coefficient
Rents for dwelling & land	0.23	Pocket money (Unexplained expenditure)	-1.83
Coffee & cocoa	0.22	Meat	-1.46
Hospital charges	0.21	Food	-0.84
Cut flowers	0.19	Education	-0.82
Rents for dwelling, issued houses	0.17	Paper diapers	-0.60
Salad	0.12	Eggs	-0.46
Beer	0.11	Fish & shellfish	-0.27
"Onigiri" & others(rice ball)	0.09	Medical care	-0.26
Obligation fees related to dwelling	0.07	Furniture & household utensils	-0.23
Taxi fares	0.07	Private transportation	-0.21

4.3 Effect of age

In Table 3 and 4, the items 'Taxi fares', 'Cut flowers', and 'Hospital charges' of one-person households tend to be consumed more by elderly people. This reflects the experiential tendency of the FIES.

In fact, the over-65 category accounts for a large percentage of the age class among one-person households in the FIES (Statistics Bureau, 2005-2015). On the two-or-more-person households, middle age householders have a large percentage of the age class. Table 6 shows the age class of householders in the FIES. The proportion of elderly householders becomes larger as time goes on.

Householder distribution by age class in FIES (data from 'e-Stat' provided by Statistics Bureau (2005-2015))

Table 6

	one-person household			two-or-more-person household		
year	under 35	35-59	60 or more	under 35	35-59	60 or more
2005	26%	29%	45%	9%	50%	41%
2010	21%	28%	51%	7%	47%	45%
2015	18%	27%	55%	6%	42%	52%

One type of the big data that is planned to be provided by the cooperate companies is the data of online personal finance software. There is low utilization of online personal finance software among elderly people. Therefore, there is particular need to adjust the age class in the case of matching the FIES data to the big data.

From the above, it is possible to suggest that the age has the potential to be as great as the household type in affecting specific purchased items. Finally, we are going to describe below how the estimation accuracy of multinomial models can be improved by the demographic items, which are influential for the specific purchased items.

4.4 Improvement of prediction accuracy for the multinomial model

The binomial model of one-person and four-or-more-person household has acceptable accuracy, but the multinomial model does not. It is difficult to estimate household size based on their purchased items since the characteristics of one household must be analyzed as included in other households. For example, as it stands, some of the items bought by one-person households are also bought by two-or-more-person households. As a potential solution to this problem, we propose using the demographic items that are included in the big-data that have an antagonistic effect on the consumption items. Namely, we attempt to improve a prediction accuracy by using not only consumption items with a small degree of overlap among household sizes but also other items that are influenced by the demographic items included in the big-data, which are age and sex. Although age appeared in the previous section to be one of the influential demographic items for specific purchased items, in this section we discuss sex because it is distributed equally.

The equivalent for variable selection is to select particular purchased items affected by demographic items. Therefore, the generalized linear mixture model (glmm) was used to make the variable selection, with sex as the response variable and the consumption items, which were loaded on a single-person household in the multinomial model, as the explanatory variable. Here, age is used as the random effect.

As a result of actually performing the variable selection with the glmm using the R package ‘Ime4’, the items selecting by the binomial model (one-person and two-or-more-person) with significantly effecting by gender were ‘Drinking’ and ‘Apples’. Moreover, the coefficients are antagonistic by gender. If there are high purchase price amount of Drinking and Apples, this indicates that there are multiple individuals who purchased antagonistic products. These results may be useful if the probabilities of one-person and

two-person households are similar in the multinomial logistic Lasso regression model that simply estimates the number of people per household.

5. CONCLUSIONS

The purpose of this paper was to estimate the household size, then to indicate the consumption items that represent the household characteristics, in order to impute the missing information of the provided big-data and integrate to the source of the CTI Micro.

We analyzed the FIES micro-data using logistic Lasso regression analysis. The estimation conducted using the multinomial model, which distinguishes between one-, two-, three-, and four-or-more-persons, does not have good prediction accuracy. In contrast, the binomial model that distinguishes one and multiple-persons does have good accuracy. According to the coefficients of the binomial model, one-person households tend to consume high-unit-price nonessential grocery items and services, while four-or-more-person households tend to consume foods and daily necessities corresponding to the different stage of lives.

Though it was difficult to survey one-person household expenditures, the result implies that it is possible to obtain the one-person household consumption data in the big-data of loyalty programs and online personal finance software. Moreover, the items such as the sex and age included in the big-data with an antagonistic effect on the consumption items could improve poor prediction accuracy in the multinomial model.

However, variable selection by lasso regression is unstable. We should investigate the detailed relationship between the variables and prediction errors for the improvement of the stability in future work. We are considering using machine learning methods such as decision tree for interaction terms and a stability of variables. We should also consider carefully the semi-continuous data which is an explanatory variable of sparse estimation. Despite few studies having treated semi-continuous data as explanatory variables, these studies are important because the consumption items data is almost semi-continuous.

Official statistics agencies in Japan have summarized and combined official survey data into economic indicators, but they have done little analysis of the data for modeling. This study is rare among them because the FIES, which is often used as descriptive statistics so far, is analyzed for a mathematical model in anticipation of application to the big data. Thus, the CTI project is also meaningful as an attempt to develop official statistics in Japan. Since the FIES has a huge volume of data, and concerns surveying and summarizing as its first priority, it is difficult to identify consistent effects of

that data. However, the above analysis suggests the possibility of identifying characteristics that are important to merge the big data and the FIES.

Acknowledgements:

We are grateful to members of the CTI project and our research department for helpful discussions and thoughtful comments. The authors wish to thank for editors and referees for their fruitful suggestions. The views expressed here are those of the authors and not necessarily those of other members of the institute.

References:

1. **Blaudow, C., Burg, F.**, 2018, *"Dynamic Pricing as a Challenge for Consumer Price Statistics"*, EUROSTAT REVIEW ON NATIONAL ACCOUNTS AND MACROECONOMIC no. 1, 79-93.
2. **Breton, R., Flower, T., Mayhew, M., Metcalfe, E., Milliken, N., Payne, C., ... & Woods, A.**, 2016, *"Research indices using web scraped data: May 2016 update"*, Newport: Office for National Statistics.
3. **Friedman, J., Hastie, T., Tibshirani, R.**, 2010, *"Regularization paths for generalized linear models via coordinate descent"*, Journal of statistical software, 33(1), 1.
4. **Hamasuna, K**., 1980, *"Current Status of Statistical Survey"*, Hosei university Japan statistics research institute report, 05, 18-53. (Japanese only)
5. **Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., ... & Usher, A**. 2015, *"Big data in survey research: AAPOR task force report"*, Public Opinion Quarterly, 79(4), 839-880.
6. [electronic sources] **Statistics Bureau,** 2005-2015, *"Family Income and Expenditure Survey"*, one-person household annual data available from: *https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200561&tstat=000000330001&cycle=7&tclass1=000000330001&tclass2=000000330022&tclass3=000000330023* (Accessed 10.06.2019), two-or-more-person household annual data available from: *https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200561&tstat=000000330001&cycle=7&tclass1=000000330001&tclass2=000000330004&tclass3=000000330005&result_back=1*, e-Stat. Statistics Bureau, Ministry of internal affairs and communications.
7. **Statistics Bureau, Ministry of internal affairs and communications**, 2017, *"Establishment of Consumption Trend Index Research Council"*, available from: *https://www.stat.go.jp/data/cti/pdf/ho20170728.pdf* (Accessed 10.06.2019), Statistics Bureau, Ministry of internal affairs and communications. (Japanese only)
8. [web page] **Statistics Bureau, Ministry of internal affairs and communications**, 2018a, *"Statistics for Japan's Future'- A Quick Reference."*, available from: *https://www.stat.go.jp/english/info/guide/2018guide.html#p0201* (Accessed 10.06.2019), Statistics Bureau, Ministry of internal affairs and communications.
9. [web page] **Statistics Bureau, Ministry of internal affairs and communications**, 2018b, *"Telecommunications Annual Report 2018 – Part 1: Sustainable growth by ICT in the era of population decline"*, available from: *http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/html/nd141110.html* (Accessed 10.08.2019), Statistics Bureau, Ministry of internal affairs and communications. (Japanese only)
10. **Struijs, P., Braaksma, B., & Daas, P. J.**, 2014, *"Official statistics and big data"*, Big Data & Society, 1(1), 2053951714538417.

11. **Takahashi, M., Ito, T.**, 2013, *"Multiple imputation of missing values in economic surveys: Comparison of competing algorithms"*, In Proceedings of The 59th World Statistics Congress of the International Statistical Institute (ISI). Hong Kong, China, 3240-3245.
12. **The Consumer Statistics Division of Statistics Bureau**, 2018, *"The orientation of developing Consumption trend index(CTI)",* the 8th consumption Research Council, document No.2, available from: *https://www.stat.go.jp/info/kenkyu/skenkyu/pdf/20190122_02.pdf* (Accessed 10.06.2019), The Consumer Statistics Division of Statistics Bureau. (only Japanese)
13. **Tibshirani, R.**, 1996, *"Regression shrinkage and selection via the lasso"*, Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.