uRos(Use of R in Official Statistics) 2019
Bucharest, Romania, 20-21th, May.

Utilization of big data for improving Consumption Trend Index

- Estimation of the number of person per household based on the characteristics of purchase items-

Anri Mutoh[1], Masayo Yamashita[1], Yoshiyasu Tamura[1], Masahiro Matsumoto[1]

1 National Statistics Centre, Japan

Contents

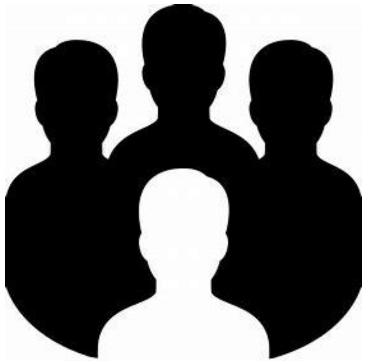
1. Background & Purpose

1.1 Background - What is CTI

Consumption Trend Index (CTI)

The index to grasp consumption trend quickly and comprehensively

being developed by Official statistics agencies in Japan



corporating with academic researchers and commercial companies(data holder)

CTI
- CTI macro
- CTI micro

We are engaged in it!

1.1 Background - Improvement of CTI micro

The CTI micro

- Its intention is to indicate the monthly trend of household average expenditure by the type of major items of households
- For compensating a possible bias in the **Family Income and Expenditure Survey (FIES**), it consists of the **Survey of Household Economy** and **the Single Household Expenditure Monitor Survey** in addition to FIES

Further improvement of the CTI Micro

In addition to the former data, utilizing big data obtained automatically by the corporate companies, such as the data of household accounts web service

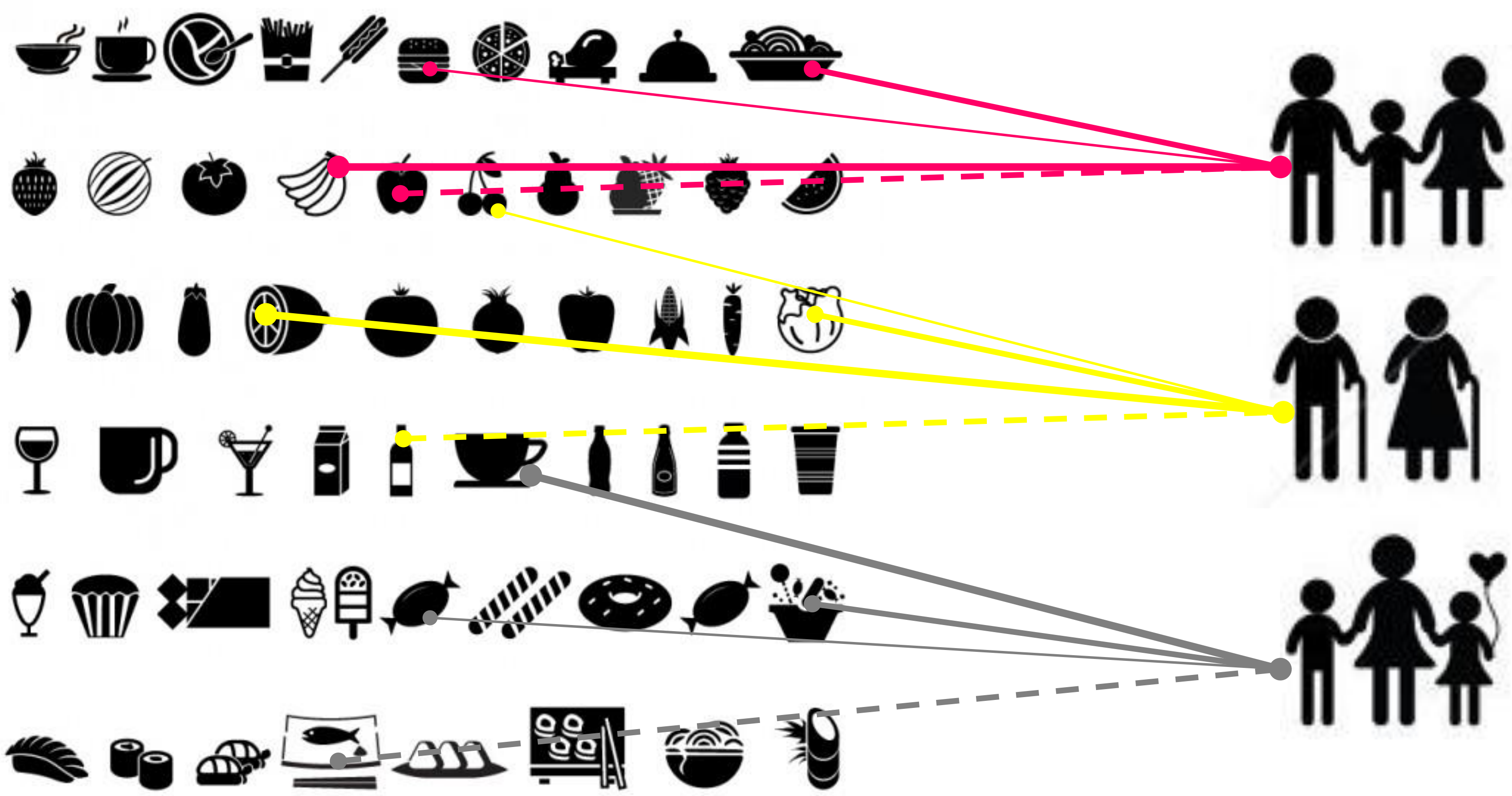
However

A problem in the improvement

The provided big data needs to correspond to corrective demographic items of FIES, such as the number of person per household, which tends to be missing

1.2 Purpose

The purpose is to estimate the number of person per household based on the characteristics of purchase items



How many person per household?

What items are purchased in each household?

Contents

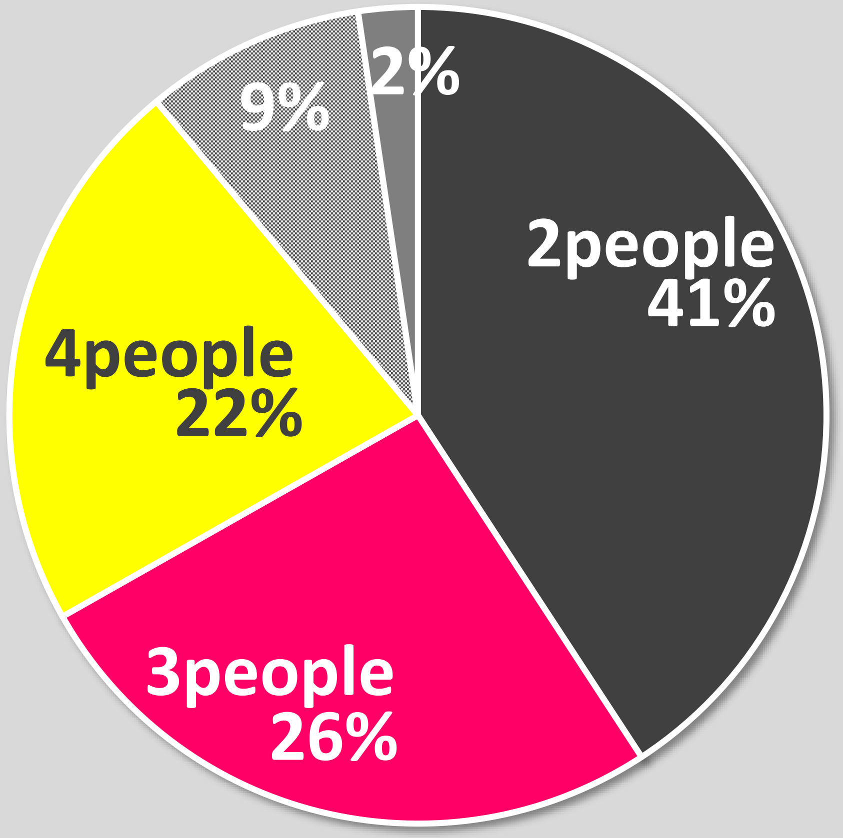
2. Methods

2.1 Methods

The analyzed data in this research are the results of the January 2010 the Family Income and Expenditure Survey (FIES) in Japan.

In the FIES, 700 one-person households and approximately 7,800 two–or-more people households has been surveyed.

The purchased items are very similar between the four-people and five-people households. Therefore we classify the data as one-person household, two-people household, three-people household, and four-or-more people households.



90% of the 2-or-more-people households are occupied by 2 to 4 people households.

2.1 Methods

The FIES data has almost 600 consumption expenditure items as explanatory variables and its observations are sparse (with a lot of zero values).

1	2	3	⋯	598	599	600
0	0	0	⋯	0	0	0
390	0	0	⋯	0	0	0
0	0	0	⋯	0	0	0
:	:	:	⋯	:	:	:
0	40	0	⋯	0	80	0
0	0	0	⋯	0	80	0
500	0	0	⋯	200	0	0
550	0	20	⋯	0	0	0
0	0	0	⋯	0	0	0
0	0	0	⋯	0	0	77

Therefore we employ a LASSO regression to investigate the factors of number of people per household. We use the **glmnet package** in R to sparse data.

2.1 Methods

LASSO regression : a regression analysis that uses a L1 regularization terms as a penalty for sparse data.
Let y_i, x_{ij} be a response and an explanatory variable $(i = 1 \sim n,\ j = 1 \sim p)$, so there are $\mathrm{n} \times (p+1)$ data matrix. Then let $\lambda \geq 0$ be the regularization parameter. The Lasso problem takes the below form(t is a free parameter) :

$$\min_{\beta_0, \beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i^T \beta \right)^2 \right\} \ subject\ to \sum_{j=1}^{p} |\beta_j| \leq t$$

$$\min_{\beta \in \mathbb{R}} \left\{ \frac{1}{n} ||y_i - X\beta||_2^2 + \lambda |\beta| \right\}$$

By the method of Lagrange's undetermined multipliers

2.1 Methods

1. Extract the consumption expenditure items common to one-person households and two-or-more households
2. Calculate the correlation of the purchase items, and reduce 100 items with a correlation coefficient over 0.7
3. Analysis by **logistic regression with L1 normalization term** whose explanatory value is the consumption expenditure items

Using the R package “glmnet”

①Multinomial logistic LASSO regression

cv.glmnet(observe, response,
family=“**multinomial**"), alpha=1)

household member**1~4** ～ the purchase items

prediction accuracy was low (accuracy0.47)

②Binomial logistic LASSO regression

cv.glmnet(observe, response,
family="binomial"), alpha=1)

household member**1or4** ～ the purchase items

Contents

3. Results

3.1 Results

As a result, an accuracy=71% in the
binomial logistic LASSO regression model,

one-person households were completely able to distinguish.

accuracy=0.71		predicted households	
		four-or-more-people	one-person
actual households	four-or-morepeople	1658	959
	one-person	0	700

choose the largest λ such that error is within 1 standard error of the minimum.

3.1 Results

The coefficients of **1 or 4 logistic regression model**
with **the amount of purchased price** as explanatory variable (λ=0.005085476)

the one-person household	
corresponding item	coefficient
Taxi fares	0.17
Drinking	0.13
Apples	0.11
Permanent wave charges	0.09
Railway fares	0.08
Women's stockings	0.08
Salad	0.08
Other citrus fruits	0.07
Other remittance	0.07
Quilts	0.07

the four-or-more-people household	
corresponding item	coefficient
Education	-1.72
Meat	-1.27
Fuel, light & water charges	-0.76
Pocket money	-0.70
Fried & salted snack crackers	-0.68
Paper diapers	-0.63
Communication	-0.52
Oil, fats & seasonings	-0.44
Chinese noodles	-0.37
Bean sprouts	-0.37

3.1 Results

The coefficients of **1 or 4 logistic regression model** with **the frequency of purchased items** as explanatory variable (λ=0.001911098)

the one-person household

corresponding item	coefficient
Rents for dwelling & land	0.40
Coffee & cocoa	0.19
Apples	0.17
Obligation fees related to dwelling	0.14
Family altar & gravestones	0.13
Hospital charges	0.11
Personal care and services	0.11
Other remittance	0.10
Permanent wave charges	0.10
Drinking	0.09

the four-or-more-people household

corresponding item	coefficient
Education	-2.09
Meat	-1.44
Pocket money	-1.03
Food	-0.90
Paper diapers	-0.74
Eggs	-0.46
Fried & salted snack crackers	-0.33
Furniture & household utensils	-0.27
Clothing	-0.24
Shampoo	-0.24

3.1 Results

Although we estimates with data of purchased items only, if the relationship between the demographic profiles and purchased items could be explained, we might be become distinguished one, two, and three-people households as well?

Generalized linear mixture model （GLMM）

- **Response variable** : sex
- **Explanatory variables** : the amount of purchased price of each items
- **Random effect**: age (related to intercept)

<dataframe>

- sex : 1=male, 2=female
- age : 18~93

(change to 3points categorical value)

- b5 : the amount of Drinking
- b6 : the amount of Apples
- b7 : the amount of Railway fares

3.1 Results

In the package “lme4“

glmer(sex ~ 1+ b5 + b6 + b7 + (1|age),
family=binomial(link = "logit"), data=dat_b)

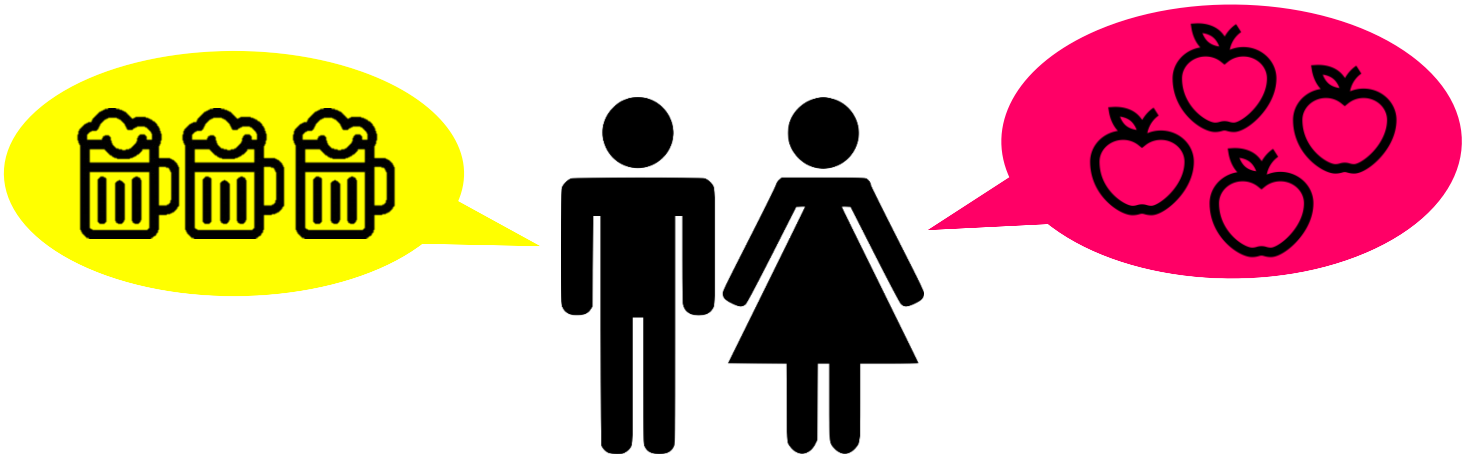
```
Random effects:
 Groups Name        Variance Std.Dev.
 year   (Intercept) 0.5624   0.7499
Number of obs: 700, groups:  age , 75

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6877     0.3099   2.219  0.02649 *
b5           -0.5235     0.1076  -4.863 1.16e-06 ***
b6            0.4562     0.1192   3.826  0.00013 ***
b7            0.1225     0.1103   1.110  0.26702
```

b5(Drinking), **b6(Apples)** are significantly effective

In addition, antagonism results are seen between men and women.

3.1 Results

If The amount of both purchase prices of drinking and apples are many, it indicates that **there are multiple individuals who purchased antagonistic products**.



These results could be useful in the case the probability of one-person household and two-person household are similar in the multinomial logistic LASSO regression model that estimate simply number of people per household.

Contents

4. Conclusion & Future work

4.1 Conclusion

- The estimation of number of household by the multinomial model, which distinguishes one, two, three, and four-or-more people, is not very good accuracy, but **the binomial model** distinguishing one and four-or-more people is good accuracy.

- Generally, in the one-person households, frequency of purchase of services and high unit price goods, in four-or-more-person households, foods and daily necessities which are comparatively reasonable and available in large quantities, seems to represent household characteristic.

- It is suggested that items with **antagonistic characteristic** on demographic profile items could improve the accuracy even in the multinomial model with poor prediction accuracy.

5.1 Future work

To increase the prediction accuracy of the multinomial models :

- It is too sparse for only one month data. Therefore consider summing up the data of several months or summing up the data in several types of items with less variance.

- Search for demographic profile items that have antagonistic feature on the purchased items besides age or sex.

Thank you for your attention

amuto@nstac.go.jp