BETTER MODEL SELECTION FOR POVERTY TARGETING THROUGH MACHINE LEARNING: A CASE STUDY IN THAILAND

BY

MISS PISACHA KAMBUYA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF ECONOMICS

(INTERNATIONAL PROGRAM)

FACULTY OF ECONOMICS

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2017

BETTER MODEL SELECTION FOR POVERTY TARGETING THROUGH MACHINE LEARNING: A CASE STUDY IN THAILAND

BY

MISS PISACHA KAMBUYA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF ECONOMICS

(INTERNATIONAL PROGRAM)

FACULTY OF ECONOMICS

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2017

THAMMASAT UNIVERSITY

FACULTY OF ECONOMICS

THESIS

BY

MISS PISACHA KAMBUYA

ENTITLED

BETTER MODEL SELECTION FOR POVERTY TARGETING THROUGH MACHINE LEARNING: A CASE STUDY IN THAILAND

was approved as partial fulfillment of the requirements for the degree of Master of Economics (International Program)

on August 10, 2018

Chairman _____
(Dr. Chayanee Chawanote)

Member and Advisor _____
(Assoc. Prof. Dr. Chaiyuth Punyasavatsut)

Member _____
(Assoc. Prof. Dr. Tanapong Potipiti)

Dean _____
(Assoc. Prof. Dr. Chayun Tantivasadakarn)

Thesis Title	BETTER MODEL SELECTION FOR POVERTY TARGETING THROUGH MACHINE LEARNING: A CASE STUDY IN THAILAND
Author	Miss Pisacha Kambuya
Degree	Master of Economics (International Program)
Major Field/Faculty/University	Economics Faculty of Economics Thammasat University
Thesis Advisor	Assoc. Prof. Dr. Chaiyuth Punyasavatsut
Academic Years	2017

ABSTRACT

Proxy Means Test (PMT) is the method for targeting the poor who should obtain the benefit from social programs by estimating an income or expenditure with the Ordinary Least Square (OLS) regression using set of variables which are correlated with those welfare measurements because it is difficult to measure directly. However, the variable selection in OLS would require the stepwise regression which is time-consuming task when the set of variable is very large. Therefore, this study aims to propose the Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest (RF) algorithms which are part of Machine Learning field to improve PMT model in terms of variable selection and model performance by focusing on the out-of-sample targeting accuracy of poor household in Thailand. The data in this study comes from Thailand Social-economic survey (SES) in 2016.

The results show that PMT models based on selected variables from RF can reduce number of actual poor households that are classified as non-poor household (an exclusion error) and also increase poverty accuracy rate (target poor household as poor accurately) in the national, urban and rural levels, however, the inclusion error is still high. For the performance of PMT models based on selected variable from Stepwise regression and LASSO, are quite not different.

In addition, PMT models with Stepwise regression and LASSO selected variables outperform RF selected variables in terms of reduce in an inclusion error. On the other hand, an exclusion error for PMT models based on RF selected variables is significantly one time less than PMT models using Stepwise regression and LASSO selected variables. Since, there is a trade-off between the inclusion and exclusion errors, this study suggests that if the objective of social welfare program is to help the truly poor, PMT model based on variable selection of RF is more appropriate.

Keywords: Proxy Means Test, Poverty Targeting, Variable Selection, LASSO, Random Forest

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to Assoc. Prof. Dr. Chaiyuth Punyasavatsut, my kindly advisor, who constantly motivated me to keep on working and provided very valuable comments on my thesis. Without his guidance and persistence help, this thesis would not have been possible. Besides my advisor, I am indebted to Dr. Chayanee Chawanot, my committee chair, and Assoc. Prof. Dr. Tanapong Potipiti, my extra-departmental committee, for their supports and helpful suggestions.

I would like to express my entire grateful and the deepest appreciation to my parents and lovely sisters for their love, warmly care and financial support. I am also indebted to Ms. Wannah Vejbrahm and Ms. Siwaree Siriwong, faculty staff, for their support.

Furthermore, I would like to thank Miss Pimwilai Kijjanapanich who provides helpful suggestion to me when I got the problem. And I would like to thank other M.A. students that always help each other.

Miss Pisacha Kambuya

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Statement of the study

Over the past several years, the government of Thailand has been established many social programs to provide the subsidy in terms of money to the poor. One famous social program is the Register for state welfare program has been set up under responsibility of Ministry of Social Development and Human Security in 2016 which provides the subsidy 200 baht per month to person who is unemployment or income less than 100,000 baht per year and 300 baht per month if income less than 30,000 baht per year for purchasing goods through the state welfare card. However, an effectiveness of the program based on how many the truly poor receives the benefit. Therefore, there are 3 problems to concern: First, program will lose the benefit if registrants underreport their income. Second, if the budget of program is limited, program should provide the priority to the extremely poor. Third, to achieve high targeting accuracy, minimizing inclusion error (the non-poor are identified as poor) and exclusion error (the truly poor are classified as non-poor) are focused. In terms of an impact on poverty, the program implementers are likely to focus on reducing undercoverage rate, while the budget constraint aspect concerns about alleviation of inclusion error. Hence, the tools for targeting the poor are needed to consider two types of error.

One popular method to target the poor is called Proxy means test (PMT), which are based on the assumptions that household consumption expenditure or income is inappropriate measurements for individual's income due to unavailable and difficult to obtain directly since some of household or individual underreport their income or expenditure. Therefore, the estimation of household income or consumption through the ordinary least square (OLS) regression model has been implemented by using household characteristics as a proxy, such as age, quality of the dwelling, ownership of farm land and durable goods, or educational level of

household head as explanatory variables, which the variables that are significantly correlated with the income or expenditure will be consider as the selected variables in the model. In addition, the model that has the high R-Squared, in other words, the explanatory variables can explain the variation in an household income or expenditure very well will be selected as the preferred model to estimate an income. Whereas, a targeting accuracy is measured as inclusion and exclusion errors (M. Grosh & Baker, 1995). To verify the quality of visible standard of household living, a visit to household is needed. PMT can create the effective outcomes on poverty targeting among all targeting methods in Latin America (M. Grosh & Baker, 1995). Therefore, PMT has become common tool for targeting the poor in social programs because full means tests are costly and time consuming. Furthermore, the practitioners further improve PMT tool to classify households by using score to rank the degree of poor household and select the poorest households who are eligible and do receive the benefit from program.

Although, a proxy means test (PMT) is the tool that quickly and easily to target the poor households, however, it is time-consuming for OLS method to do the tasks of both variable selection and the process of running and comparing the performance of several models over the large set of variables, which would be required the stepwise regression to do these jobs. Thus, we will contribute to the literature on poverty targeting by considering an alternative algorithm for the selection of model and variable and prediction of poverty status to target the poor. The linear regression likes ordinary least square (OLS) will be based method for PMT tool, and the suggested alternative methods to select the best set of variable and predict the out-of-sample performance are the Random Forest (RF) method (Breiman, 2001) and Least Absolute Shrinkage and Selection Operator (LASSO).

The Random Forest method is part of the Machine Learning literature and has been applied for prediction in several areas of research. For example, Verikas, Gelzinis, and Bacauskiene (2011) reviewed the use of RF method in a wide range of research fields such as the prediction of distribution of animal species, the prediction of long disordered regions in the protein sequences, and classification of

agricultural practices based on satellite imagery. It can be noticed that this method is commonly used in the fields like science, medical and geographic fields, but lack in the economics field. Furthermore, Varian (2014) described the advantage of machine learning algorithm as the good at predictions, but cannot do estimation and hypothesis testing.

In the poverty prediction literature, the application of random forest method is still scant and very recent. Otok and Seftiana (2014) founded that the RF method is very accurate to identify the poor households who eligible for the social assistance packages in Indonesia. Thoplan (2014) used an RF method to predict poverty in Mauritius, the finding was that RF provided the accurate prediction in poverty. For the proxy means test method, McBride and Nichols (2015, 2016) was successful in apply the RF method to predict targeting performance compare with the linear regression based models for improving proxy means test targeting models. Their study compared the out-of-sample targeting accuracy in Malawi, Bolivia, and Timor-Leste, the result was that quantile RF is better at estimating a poor household as poor, or the undercoverage rate decline, while the leakage is still high. The conclusion of their study is that RF method can significantly improve out-of-sample performance about 2-18 percent.

Therefore, we will present whether improving the out-of-sample performance of PMT tools through LASSO and Random Forest (RF) can enhance an accuracy of poverty targeting on poor household in Thailand in terms of variable selection for PMT models compare to the Stepwise regression methods. In this study, we propose three models as follows; (1) PMT model based on selected variables from Stepwise regression (2) PMT model based on selected variables from LASSO, and (3) PMT model based on selected variables from RF. Then compare the performance of three PMT models in terms of an interpretation and targeting accuracy of the poor household in Thailand.

1.2 Objectives of the study

1.2.1 To improve the targeting accuracy of proxy means tests (PMT) model through the Stepwise regression, LASSO and Random Forest methods based on variable selection procedure.

1.2.2 To compare the targeting performance on poor household in terms of targeting accuracy between PMT model through the econometrics (Stepwise regression) and Machine Learning (LASSO and Random Forest) methods.

1.3 Scope of the study

This study focuses on the improve in poverty targeting of proxy means tests tool which employs LASSO and Random Forest that are a part of machine learning procedure by producing the forecasting model via stratification of appropriate household characteristic which is eligible to be the program beneficiaries compare with the traditional PMT with the selected variables by a stepwise regression. The empirical models are based on the Social – economic survey (SES) data of household in 76 provinces of Thailand in 2016, excepting Bangkok province because it has high variation of consumption expenditure.

1.4 Organization of the study

The study is organized in five chapters. The second chapter is review of related literature which provides the development of the proxy means tests (PMT) with several regression methods, variable selection, and the use of machine learning (ML) algorithms like the LASSO and Random Forest (RF) to develop PMT model. In the third chapter, the algorithm framework and research methodology are described. The empirical result of traditional PMT and new PMT which perform machine learning techniques and compares the performance of models between traditional PMT and new PMT are discussed in the fourth chapter. Finally, the fifth chapter

presents the conclusion and policy implication, and the recommendation for the future study.

CHAPTER 2
REVIEW OF LITERATURE

This chapter aims to present a literature review on research relating to the proxy means tests for targeting in social program and machine learning methods for improving the PMT tool. The organization is as follows; section 2.1 discusses the proxy means tests with several methods to estimate the consumption or income of household; 2.2 discuss the variable selection through machine learning algorithm and 2.3 discusses the improvement of proxy means tests by using machine learning procedures to develop the poverty targeting accuracy.

2.1 Development of Proxy Means Test (PMT)

Proxy Means Test (PMT) model is adopted to predict the household income or consumption based on observable characteristics that are correlated since the reported income is difficult to identify directly. PMTs have successfully been implemented to measure the household income and provided the best outcomes among other targeting tools in Latin America (M. E. Grosh, 1994).

In the past, Bidani and Ravallion (1993) used a reference food bundle cost as a food poverty line to construct the regional poverty targeting in Indonesia. The finding provided the greater poverty rate in rural areas than in urban areas but the process to measure the impact on poverty was obscure. Then, Ravallion (1996) constructed the poverty targeting in terms of regression form of the individual poverty which was measured by using the variety of household's characteristics as a proxy. The advantage of poverty regression is that the policy makers will know which region (A or B) should get priority in the social program. Therefore, the PMTs based on the ordinary least square (OLS) regression model using the log of income or expenditure as the dependent variable have become the common tool to target the poor in developing countries. For example, Ahmed and Bouis (2002) implemented the OLS regression model for constructing PMT to target the needy on food subsidy

program in Egypt using consumption expenditure per capita as a welfare measurement. This study used the household size, education of members and ownership of durable goods as a proxy, the results showed that the government can save the budgetary about 74 percent, which more than the saving from the selected practical model. However, the OLS regression method for PMT contributes two problems; First, the OLS minimizes the squared between the true and predicted outcomes, but it is different from the minimized poverty problem; second, variables on the right hand side of equation (explanatory variable) face an endogenous problem (M. Grosh & Baker, 1995).

Most studies of PMT have been classified the variables that are correlated with income or expenditure into several categories such as household demographic, ownership of asset, characteristics of dwelling, education of household head, and location variables. The question such which method they use to choose these variables to be the best indicator in final model is raised. For large set of candidate variables, the stepwise procedure is preferred to select these variables in PMT tool (Brown, Ravallion, & Van De Walle, 2016; M. Grosh & Baker, 1995; Narayan & Yoshida, 2005; Nguyen & Lo, 2016). However, the study of James and McCulloch (1990) suggested that stepwise procedure cannot rank and provide the best variables using their importance.

Over the past decade, there are several studies proposed alternative methods beside OLS regression to improve robustness of PMT model. The quantile regression was suggested by Koenker and Bassett Jr (1978) showed that this method is more robust outliers than OLS model. Nevertheless, Houssou, Zeller, Alcaraz, Schwarze, and Johannsen (2007) argued that the OLS are more robust than the quantile regression method. In addition to OLS regression, this study also employed the Linear Probability Model (LPM), probit, and quantile regression methods to test the robustness and out-of-sample validity of the model. The results showed that the quantile regression performs moderately accuracy in-sample predictions of poverty and provides less robust, while the OLS and probit perform better out-of-sample.

This study suggested that the probit method provides the optimized accuracy and robustness for PMT model.

In case of Thailand, Healy and Jitsuchon (2007) used the stepwise regression to construct the poverty map in Thailand using the SES and census data. The objective of this study is to improve the policies that aim to allocate resources for reducing in the poverty gap and inequality through the poverty map. Their study estimated the income and consumption of household from the household assets, demographic, and occupational variables that are correlated with in both of household and local level. In conclusion, the study can improve the targeting poor household and poverty gap was reduced in both amphoe and tumbon levels.

For the implementing PMT in Thailand, it is the common method for targeting the eligible poor in the cash transfer program. For instance, Thailand's Child Support Grant program (2015) used the PMT to target the newborn and pregnant woman in poor household to receive the grant under following five conditions; household monthly income lower than 3,000 baht per person, having dependency member, housing conditions, not owning a car or truck, and farmer owning less than one Rai of land (about 1,600 square meters). The other program that used the PMT for targeting the poor is the grant for poor students program in Thailand (Punyasavatsut, 2017), this study employed the same condition with a Child Support Grant program. However, Punyasavatsut (2017) targeted the poor in provincial level while Child Support Grant program targeted the poor in national level.

As we mention earlier about the weakness of OLS regression for predict the poverty, OLS is not appropriate since an endogeneity problem. This problem is that set of explanatory variables are not only correlated with a dependent variable but also related with each other. Hence, the algorithm of machine learning like Random Forest (RF) will be discussed about its special procedure of variable selection which is known as "Variable Importance (VI)".

2.2 Variables selection

An importance step of PMT is a variable selection. Which variable should be included in the final model for PMT is interesting topic. For PMT with OLS regression, the stepwise regression is used to select the set of variable by eliminating the variable that is not statistically significant with dependent variable and also not decrease in the explanatory power of model (R-squared) when the variables are added or excluded in the model. In practice, the set of variable in PMT should easy for staff to collect and calculate the PMT score; therefore the small set of variables will be better than many of variables. However, stepwise regression is time-consuming task for OLS when the set of variable is large and also has an endogeneity problem. Therefore, there are many studies propose the alternative algorithms in the machine learning field to study the variable selection and try to capture the pattern of data to understand the variable in dataset, especially non-linear variable that linear regression such OLS cannot capture directly.

Tibshirani (1996) introduced the Least Absolute Shrinkage and Selection Operator (LASSO) method which is one of machine learning algorithm that propose the penalized term called " loss function" to regularize (shrink) the coefficient in OLS estimator to zero for the variable that provide less correlation with dependent variable. This LASSO make the OLS become a spare model, in other words, LASSO can eliminate variables by shrinking those variables to zero inside its algorithm, thus we can obtain a small set of variable. The coefficient from LASSO will converge or diverge from coefficient of OLS is depended on the lambda parameter that we choose which make LASSO coefficients bias from OLS. Therefore, Belloni and Chernozhukov (2013) proposed the OLS-post LASSO method by proposing the LASSO to select variable and model at the first step and then estimate these LASSO selected variables using OLS. The study described the results by deriving the theoretical properties of post-model selection of LASSO estimator; they founded that if LASSO can capture the true variables in the model, then after propose OLS will make the error smaller than propose only LASSO and performance of OLS-post

LASSO in terms of converge coefficient close to zero is as good as the LASSO. Similarly, the study of Hastie, Tibshirani, and Wainwright (2015) also provided example of the LASSO as first step for variable selection and then propose OLS. The result showed that propose OLS-post LASSO make the model is sparsity.

However, LASSO can fail in terms of missing the true variable that provides the main effect to dependent variable. Therefore, Random Forest or RF (Breiman, 2001) is one of algorithm that provide an variable selection within its algorithm, is called the variable importance. Random Forest runs an algorithm based on the aggregate bootstrap, growing the several tree and then come up the prediction result by averaging outcomes from each tree, this method can reduce the variance and improve in an accuracy of the model. RF algorithm also performs better in out-of-sample prediction and can capture the non-linear variable. RF ranks the order of selected variable using importance of each variable; high importance value means variable has highly effect on dependent variable if we exclude high importance variable, the accuracy of model will decrease as well. However, the variable importance of RF cannot provide the important information such coefficient that leads to the crucial problem that we cannot obtain and realize the magnitude of main effect of selected variable from RF; hence, RF is known as a non-interpreted model. The recent study of Welling, Refsgaard, Brockhoff, and Clemmensen (2016) introduced the new method to visual the variable contribution of RF which is called Forest Floor Visualization. This method proposes Goodness-of-Visualization (GOV) tests to provide the visualization of variable make us understand whether variable has main effect on dependent variable or an interaction term with explanatory variable or not. Moreover, this test can provide R-squared value of each variable to understand the explanatory power of the variable on dependent variable which is better than consider only the rank of variable from their importance.

2.3 Improvement of PMT through Machine Learning

Machine learning (ML) algorithms have been developed in the fields of computer science and statistics. The econometrics literature, such as Ordinary least square (OLS) regression focuses on the unbiased estimator but the predictions are not optimized, while the ML algorithm sheds light on the minimization of the out-of-sample prediction error. In other words, the ML algorithms provide powerful techniques to improve out-of-sample prediction accuracy. This study focuses on improvement of the PMT tool to obtain precise targeting accuracy with out-of-sample prediction, thus we will review the studies that implement and develop the ML techniques to improve the PMT method.

Recently, the application of random forest method has become popular method in poverty study, Otok and Seftiana (2014) found that the RF method is very accurate to identify the poor households who eligible for the social assistance program in Indonesia. Thoplan (2014) also used RF method to predict poverty in Mauritius, the finding was that RF improves an accuracy in poverty prediction.

In the improvement of poverty targeting via machine learning, McBride and Nichols (2015) employed stochastic ensemble (SE) methods to develop PMT targeting in the country level by improving out-of-sample prediction performance based on the Living Standards Measurement Study (LSMS) data which collected by IRIS center. Their study compared the out-of-sample targeting accuracy in Malawi, Bolivia, and Timor-Leste. Firstly, they used the PMT method to produce targeting accuracy by identifying a subset of household characteristics, which are 15 approximately from 70-125 variables. They then identified the parameters which are related to household income level by the statistical methods, such as OLS, Logit, Probit and Quantile regression that provided the highest prediction accuracy (in-sample prediction). Secondly, they applied Random Forest and Quantile regression forest algorithms that aim to improve out-of-sample targeting accuracy on previous PMT procedure of IRIS center. The results showed that stochastic ensemble methods provided significant increase in poverty targeting accuracy, a significant reduction in

exclusion error rate, and overall gain in balance poverty accuracy criterion (BPAC) in comparison to PMT with linear regression methods.

Moreover, McBride and Nichols (2016) also extended the machine learning methods by using the cross-validation (CV) to minimize prediction error and stochastic ensemble (SE) procedures for improvement of poverty targeting accuracy of PMT targeting. Their study divided into two steps; first, producing the PMT methods, then performs the k-fold cross-validation in training sample and select preferred model that produces the best BPAC[1] in cross-validation; second, performing the stochastic ensemble approaches via the same training sample with the cross-validation approach by using the random forest and quantile regression forests model. The study found that the CV and SE methods produces a gain in poverty targeting accuracy, reduces the undercoverage rates, and improves BPAC as the same results with their previous study. However, in terms of variable selection are not well demonstrate in this analysis since they used selected subset of variables, approximately 15 variables from full set of variables.

For the recent study, Sohnesen and Stender (2017) used the LASSO and RF methods to predict poverty using one year of data for prediction within same year and two years of data to predict poverty over time. They found that the random forest method is a good predictor for poverty and provide a better robust predictor than linear regression methods; although RF models provide highly accurate poverty prediction in urban and rural but do not provide more accurate prediction compare with LASSO and linear regression models in national level. However, RF method can predict the poverty with accuracy even though they use the small selected variables in model instead of full set of variables. This study concluded that RF method is simple and easy to use. Furthermore, Kshirsagar, Wieczorek, Ramanathan, and Wells (2017) used the bootstrap LASSO for selecting a subset of variables that provide an

[1] BPAC is the balanced poverty accuracy criterion, which is the correctly predicted very poor as a percentage of the true very poor minus the absolute difference between the undercoverage and leakage rates.

accurate prediction of poverty rate. Similarly, the study of Knippenberg, Jensen, and Constas (2017) captured the food insecurity dynamic of household using the Coping Strategy Index (CSI) as a measurement and implemented LASSO and RF algorithms to choose the ten best selected variables. The results indicated that the predictive accuracy of CSI between LASSO and RF methods is quite not different. LASSO provided more accuracy than RF only 0.8 percent.

In case of Thailand, there has been no machine learning algorithm like the Random Forest (RF) and LASSO for targeting the poor in Thailand, thus this study will propose the LASSO and RF to improve the proxy means test (PMT) method in terms of variable selection and model selection, compare the performance of these models with the econometrics method such OLS regression in terms of interpretation and accuracy of the models for targeting the poor household in Thailand.

This study will improve the proxy means test (PMT) model through the Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest (RF) algorithms using its variable selection procedure for selecting the set of variable to estimate PMT models and then construct the PMT score to target the poor Thai household in terms of out-of-sample. The contribution of the study will try to propose OLS-post RF to study the performance of RF selected variable whether improve the targeting accuracy of PMT or not.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Algorithm frameworks

The goal is to improve the proxy means test (PMT) method through machine learning algorithm using the LASSO and RF methods. Thus, the algorithm implication in this study is associated with the LASSO and Random Forest framework and targeting performance.

3.1.1 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is developed based on least square estimator by adding the penalty term, the LASSO estimator can be shown as equation (3.1)

$$\min_{\beta_0,\beta_j}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{P}x_{ij}\beta_j)^2\right\}, \tag{3.1}$$

$$\text{subject to } \sum_{j=1}^{P}\left|\beta_j\right|\le t, \tag{3.2}$$

equation (3.1) is a problem optimization in terms of least square functional form , which subjective equation (3. 2) , where $i=1,...,N$ denotes the number of observations, $j=1,...,P$ denotes the number of explanatory variables and t is the parameter that defines a regularization size.

To obtain β_{LASSO}, the function of equation (3.1) aims to optimize the problem by minimizing the residual sum of squares (RSS). The LASSO estimator, β_{LASSO}, can be solved by equation (3.3)

$$\beta_{LASSO}=\min_{\beta_0,\beta_j}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{P}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{P}\left|\beta_j\right|\right\}, \tag{3.3}$$

where the second term is "l_1 loss function". This loss function will sum across all absolute coefficients and then multiplied by λ, which is the parameter that define Bayesian shrinkage degree of the problem. Since we have to pick β_{LASSO} to minimize RSS, LASSO estimator thus allows us to tune the parameter, λ. In other words, the residual sum of square (RSS) will increase or decrease is depends on the λ, which is the advantage of LASSO to reduce an error by tuning parameter.

To select the λ to solve the problem, unfortunately there is no the theory for supporting the choice of λ. For the relationship between λ and the coefficient, if λ converges to zero ($\lambda \to 0$) the objective function then becomes an OLS estimator and β_{LASSO} is equal to β_{OLS}. Nevertheless, if the value of λ is any positive then the coefficient of β_{LASSO} will divert from the coefficient of β_{OLS}. Moreover, if λ converges toward infinity ($\lambda \to \infty$) the coefficients of β_{LASSO} will tend to close to zero, in the other words, the coefficients will have been shrunk to zero. Therefore, all coefficient estimates depend on the chosen λ.

In practice, we select the value of λ through cross validation (CV) method. Initially, we will have the untransformed coefficients, which the value of λ will rely on between the mean of zero ($\lambda.\min$) and standard deviation of one ($\lambda.1se$).

The LASSO estimator can select the variable by penalizing model based on the sum of an absolute value of coefficients. Some variable will be zero after optimizing the objective function, therefore the coefficients that remain non-zero will be considers as variable selection.

3.1.2 Random Forest

The model of random forests was first introduced by Ho (1995), who proposed the stochastic modeling to construct decision tree - based classifiers which can be randomly expanded for increase in accuracy for training and testing (unseen) data. In other words, this method constructs the multiple trees in a random feature subspace (set of variables). Amit and Geman (1997) then studied the new approach that aimed to shape classification and illustrated performance in high

dimensions in terms of number of shaped classes and the degree of variability within classes by defined a large number of geometric arrangements in the split at each node, which is based on the growing binary classification trees.

Random Forest (RF) grows the trees based on the decision tree that is used to predict outcome in terms of classification and regression trees. The Classification and Regression Trees (CART) procedure is one class of supervised learning methods in machine learning algorithm that predict observations from the data in terms of characteristics (classification trees) and continued variables (regression trees), which split a space into the regions following the binary decision rule. This study sheds light on the regression trees model, in particular, the random forests for making predictions of household expenditure.

Regression trees models are constructed by building a tree and each node following the recursive binary tree, which splitting algorithm as follows (Hastie, Tibshirani, & Friedman, 2009):

The algorithm decides on the splitting variable, X_j, and the splitting point, $X_j = s$, then define the half planes of R_1 and R_2, which can be shown as:

$$R_1(j,s) = \{X \mid X_j \le s\} \text{ and } R_2(j,s) = \{X \mid X_j > s\}, \tag{3.4}$$

we then select X_j and *s* to solve the minimization problem,

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right], \tag{3.5}$$

for any X_j and *s* can be solved by

$$\hat{c}_1 = n^{-1}\sum_i (y_i \mid x_i \in R_1(j,s)) \text{ and } \hat{c}_2 = n^{-1}\sum_i (y_i \mid x_i \in R_2(j,s)). \tag{3.6}$$

For the best split, this algorithm divides the data into the two results of region and repeats the splitting process at each of the two regions. Then,

repeat on all of the resulting regions. In addition, the optimal size of growing tree depends on the data, which the very large tree will be confronted the over-fitting problem, while the small size of tree cannot capture the structure (under-fitting problem). The algorithm then stops the process when each branch meets the terminal node.

However, the problem of regression trees model is that the small change in data can be affected the split of trees, high variance, which the error can spread from the top of tree to below. To alleviate the variance, Bootstrap aggregation or Bagging (Breiman, 1996) is adapted.

To increase the prediction accuracy of a model with high- variance, we build the prediction models separately $\hat{f}(x), \hat{f}^{2}(x), \mathrm{K}, \hat{f}^{B}(x)$ on B separate training dataset, and then average the resulting predictions. Bagging generates the new training set using random bootstrap sampling from an original dataset with replacement. The set of tree models then can be trained independently by applying the regression tree algorithm on the new training dataset. The predicted responses are calculated by averaging all the models $\hat{f}^{*b}(x)$, which can be written by:

$$\hat{f}_{bag}(x) = B^{-1} \sum_{b=1}^{B} \hat{f}^{*b}(x). \tag{3.7}$$

Unfortunately, even we can reduce the variance, but the constant term of variance is remained. The idea is that we give a set of *B* identically distribution and regression trees are correlated with variance, σ^2 . Give an example, let ρ represents the pairwise correlation between the trees, and then the average of set of *B* independent observations is $\rho\sigma^2 + \frac{(1-\rho)}{B}\sigma^2$. The $\frac{(1-\rho)}{B}\sigma^2$ term will converge toward zero while *B* grows large, but the term $\rho\sigma^2$ is still persisted (McBride & Nichols, 2016). The next model that can improve the variance alleviation based on regression trees model is called the Random Forests model.

The extension of Random Forests (RF) model was proposed by (Breiman, 2001), this version of RF reduce variance by using bagging to improve the

classification accuracy, which combine the resulting classifications of randomly generated training sets. In addition, the Out-of-bag (OOB) method was implemented to gain accuracy in the model by measuring the generalization error (or out-of-sample error), in other words, to measure how the accuracy of algorithm can predict outcome values for an unseen data. Avoiding over-fitting problem can be minimized the generalization error.

Random Forests algorithm (Breiman, 2001) is closely related to bagging method since we construct a large number of decision trees on bootstrapped training samples. Every time Random Forest splits the tree, we begin with the prediction of *single tree, B* $\{T_1(X),...,T_B(X)\}$, where $X=\{x_1,...,x_M\}$ is the full set of *M*-dimensional vector of predictors (independent variables) and then we take a random sample of *m* predictors from this full set.

Ensemble produces *b* outputs, $\{\hat{f}_1(x)=T_1(X),..., \hat{f}_B(x)=T_B(X)\}$, where $\hat{f}_b(x)$, *b* = 1, ..., *B* is the prediction of a training data by the b^{th} tree. Outputs of all trees are aggregates to perform one final prediction, $\hat{f}_b^{*}(x)$. In classification problem, $\hat{f}^{*}(x)$ is the class predicted by the majority of trees, and the average of the individual tree predictions for regression problem. Then, the Random Forest predictor is constructed in equation (3.8):

$$\hat{f}^{*}(x)=B^{-1}\sum_{b=1}^{B}\{T_1(X),...,T_B(X)\}. \tag{3.8}$$

3.2 Research methodology

This study aims to predict consumption expenditure for constructing the PMT through Stepwise regression, LASSO and Random Forest variable selection methods for Thai households. To make the prediction, we find the set of variables that are present in the SES and use these variables to estimate models of monthly per capita consumption expenditure in terms of natural logarithm. For the Stepwise regression, LASSO and RF, we use these methods to select the best set of variables in the process of variable selection based on the best model's performance. Note

that, this study emphasizes the out-of-sample performance of PMT to improve the targeting accuracy of the poor households in Thailand.

Then, estimates the per capita consumption expenditure using the selected variables from Stepwise regression, LASSO and RF through OLS regression model and then obtains the coefficients to construct the PMT scores. To assess targeting performance of PMT, we evaluate all PMT model using exclusion (Type I) and inclusion (Type II) errors as criteria.

3.2.1 Data

The consumption expenditure data in this study comes from the 2016 Socio-economic survey (SES) was conducted by the National Statistical Office of Thailand. The SES is a stratified random sample of 43,887 households in Thailand. There are 77 strata, one for each changwat (province). Each of these strata is separated into two categories: municipal and non-municipal areas.

This survey contained important information on social economic aspects of household such as income, expenditures, debt, assets, demographics, and characteristics of dwelling. This study uses the household observation in 76 provinces, excepting Bangkok province because it has a high variation of consumption expenditure in this province. Hence, the observation in this study is 41,488 observations.

For out-of-sample prediction, the data is divided into two sets. The initial SES data with 41,488 observations is partitioned into two sub-samples in ratio 50:50. The first sample or Training sample (20,744 observations) is employed to train or fit the model for identifying the best model and also the best set of selected variables. Another sample is test sample or validation sample (20,744 observations) is used to test out-of-sample prediction accuracy of the constructed models.

3.2.2 Set of variables

The first step is to identify the variables present in the SES. These variables will be chosen for variable selection process in Stepwise regression, LASSO

and RF approaches and used to construct the PMT in OLS model. From the literatures, we consider household asset, demographic, dwelling's characteristics, and dependency variables that are correlated with consumption expenditure and we begin the model with 47 variables that are closely related to per capita consumption expenditure (see table B.1 in appendix B). The condition can classify into 5 categories as follows:

3.2.2.1 Household characteristics

Household head characteristics such sex, age, marital status and education levels which are composed of primary education, lower secondary education, upper secondary education, vocational education and higher education. Furthermore, number of household member and working member.

3.2.2.2 Dependency

Dependency is the household status that has the elderly person at aged upper 60-year-old, the children at aged below 15-year-old, and the disabled person.

3.2.2.3 Housing conditions

The housing conditions are characteristics of dwelling, which are composed of dwelling status as free rent, live with the others, and dwelling that is constructed by non-permanent or local material such as bamboo.

3.2.2.4 Ownership of assets

The assets that household is owned such as car, truck, van, tractor, microwave oven and etc.

3.2.2.5 Location

We also add dummy variable for regions for eradicating the difference of location. However, these dummy variables will not include in the variable selection process of Stepwise regression, LASSO and RF.

To consider which condition should be used to screen the poor households, the PMT model will estimate consumption expenditure per capita and find the poor households compare with the poverty line. The poverty line of Thailand is showed in table B.2 in appendix B, household that has the consumption

expenditure below poverty line or 2,667, 2,902 and 2,425 baht per month for the national, urban and rural area respectively (NESDB, 2016) is classified as poor and non-poor for above poverty line.

3.2.3 Variable selection process

To select the set of variables that can predict accurate consumption expenditure, we propose three procedures, the Stepwise regression, Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest (RF) to select subset of variables for constructing the PMT score in OLS regression. For these algorithms, use the training set to calibrate the models and then using selected variables from calibrated model to predict the outcomes (log of monthly per capita consumption expenditure) in testing set using OLS. The performance of these algorithms is considered from its predictive accuracy which is measured by mean square error (MSE). In the process of model's calibration, the model can be adjusted iteratively to obtain the best performance and the algorithm will identify subset of variables that are opted as the best selected variables within this process.

Stepwise regression, LASSO and RF are used to identify the best selected variables of monthly per capita consumption expenditure. We start the variable selection process of Stepwise regression, LASSO and RF with 47 variables. After subset selection procedure we retain only a subset of the variables that Stepwise regression, LASSO and RF select, and eliminate the rest from the model. OLS is used to estimate the coefficients of the variables that are retained.

3.2.3.1 Stepwise regression

There are two methods of the stepwise regression: the forward and backward methods.

1. Forward method

Start the model with no candidate variable. First step, select variable that provides the highest R-squared for model (provides p-value less than some cut-off, e.g. 0.05). Stop adding the variables into model when there is no variable that is significant at cut-off level.

2. Backward method

Begin with all variables (47 variables) in the model. Delete variable that provides the p-value more than cut-off, and then refit the model with remaining variables that is deleted one of variable (p-value larger than cut-off). Continue the process until there is no variable that is significant at cut-off level.

To obtain smaller model (smaller selected variables) from stepwise regression procedure, the results are based on a 0.01 significant level for addition variable to the model and 0.05 significant level for removal variable from the model. We run the forward stepwise regression in the STATA program. Begin with empty model in training set, if the most-significant removal term is significant, then add it into model and refit model again; if not, stop. Continue the process, if the least-significant additional term is "insignificant", then remove it and refit model again and if the most-significant removal term is "significant", add it and refit the model. Repeat these steps until there is no variable for addition and deletion. Final step is to run OLS for estimating the coefficients of stepwise regression selected variables.

3.2.3.2 LASSO

The variable selection process of Lasso is defined as following steps;

1. Run the LASSO algorithm in training set using "glmnet" function and assign alpha value is equal to 1 that is defined as the LASSO function. Training set contains the expected output value (log of consumption expenditure per capita) and 47 candidate variables at initial step.

2. The model has been trained in training set will be assessed the accuracy of model, in this case, we use the mean square error (MSE) as the criteria. Moreover, use the parameter tuning, lambda, to choose the best of lambda by using cross-validation method which 10 folds-cross validation is considered in this case.

3. Tuning the model in validation set to select the best lambda with the lowest mean square error (MSE). We train the LASSO model with k-folds cross validation, say 10-folds cross validation using cv.glmnet function. Then, we will obtain several values of lambda with different number of selected variables. The variables that coefficient is not equal to zero will be identified as selected variables in this step.

4. Perform the model with selected lambda that is obtained from training set to predict the output in test set as out-of-sample prediction to evaluate performance of model.

5. Using OLS to estimate the coefficients of LASSO selected variable.

3.2.3.3 Random Forest

Empirical approach on Random Forest, we partition the data into two sets, with size of data set. Therefore the number of observation in training and test sets is 20,744 observations. Training set is used to construct the model, while testing data is used to test the predictive outcome of the model as the same with LASSO procedure.

In the training set, we have 20,744 observations (number of household in SES 2016) with 47 variables (explanatory variables). Random forest will be built the multiple models (CART) with different sample and different initial variables. In this case, it will take n observations and m randomly selected variables to build the model in 2/3 of training set. For the remaining 1/3 of training set that left out for constructing the model is called Out-of-Bag sample (OOB) . It will use to select the variables that provide the lowest OOB error, in other words, the lowest mean square error (MSE) for regression. Therefore, RF use OOB sample to select the variable that provide the preferred model with the lowest prediction error. Then, it will repeat the process (say) 500 times and then takes the model that is constructed in training set to prediction the log of consumption expenditure per capita in test set (out-of-sample) and assess the prediction accuracy of model (see the random forest algorithm process in figure 3.1).

To see the steps of random forest obviously, we will show the random forest working in practice using R program and package's random forest by Breiman.

1. Randomly split the data into 2 sets; training set for constructing model by 20,744 observations, and test set for predicting the model's performance by 20,744 observations. In this case, we have 41,488 households (observations) in SES data. From 20,744 observations in training set, we randomly pick 13,829 observations to construct the model in training set and remaining 6,915 observations to assess the performance of model in OOB procedure to select number of variables tried at each split of trees (mtry) that provide the lowest MSE value.

2. To run the random forest algorithm in a training set with the best mtry, we use the "library(randomForest)" code in R. Then, create the random forest with 500 trees.

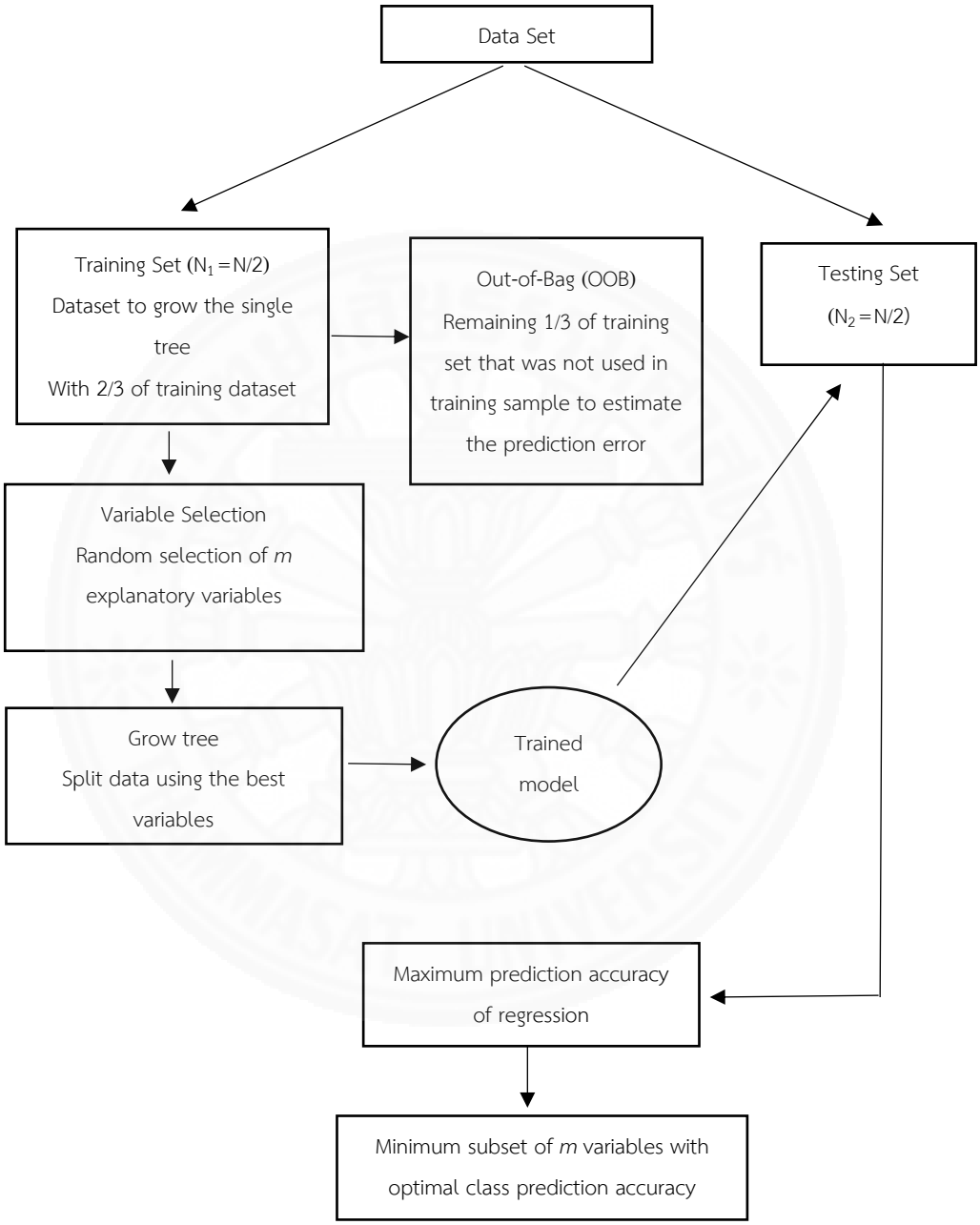
3. The Random Forest has its own variable selection which is called “Variable Importance”. The variable importance process will provide % inc mse value for the regression process. The higher of this value, the more importance of variable, which provide more effect on dependent variables (log of consumption expenditure per capita).

4. Predict and evaluate the accuracy of model in test set using the model that was trained in training set.

5. Using OLS to estimate the coefficients of RF selected variable.

Figure 3.1

Random Forest Algorithm Flowchart



Source: Author's summary.

3.2.4 OLS estimation for constructing PMT

Linear regression is the simplest and earliest predictive method for the proxy means test tool, typically the ordinary least square (OLS) regression. Using a linear combination of predictors (independent variables) like household characteristics, household ownership of assets, and characteristics of dwelling to estimate a continuous outcome (dependent variable), as consumption expenditure of household in terms of natural logarithm. The objective of OLS regression model is to estimate the regression coefficient vector β in a way that the mean squared error (MSE) is minimized.

Given the dataset of *n* household observations, OLS regression model with *k* explanatory variables is expressed as:

$$y_i = \alpha_i + \beta_k X_{ik} + \varepsilon_i \; i = 1,2,\mathrm{K}\,,n, \tag{3.9}$$

where y_i represents the household consumption expenditure per capita for i^{th} household, α_i is a constant term, β_k is the regression coefficient for the k^{th} variable, X_{ik} denotes the set of explanatory variables that are obtained from Stepwise regression, LASSO and RF for k^{th} variable of i^{th} household, and ε_i is the random error term. Then the PMT is based on:

$$\hat{y}_i = \hat{\alpha}_i + \hat{\beta}_k X_{ik}\,. \tag{3.10}$$

In practice, the OLS method for estimating α_i and β_k implements the log of consumption expenditure per capita as the dependent variable. The log of per capita expenditure of household can be expressed as:

$$\log(\hat{y}_i) = \hat{\alpha}_i + \hat{\beta}_k X_{ik}\,. \tag{3.11}$$

The selected variables from Stepwise regression, LASSO and RF that are statistically significant with log of consumption expenditure per capita on OLS

procedure will be considered as the selected variables in the final model. After running OLS estimation, the coefficients of each variable is used to construct the variable weight. Then, the household is assigned an aggregate score (predicted consumption expenditure of household is also called PMT score) that is a weighted combination of variables X_{ik} and calculated as the regression constant plus or minus the weighted variables, each coefficient is multiplied by 100 and rounded nearest the integer.

3.2.5 Assessing targeting accuracy of PMT

The targeting error is adopted to evaluate the targeting accuracy. M. Grosh and Baker (1995) proposed Type I and Type II errors to measure inclusion and exclusion error rates by categorizing the household into four groups which are following whether their true and predicted (by the regression model) consumption expenditure levels fall above or below the cutoff point. From table 3. 1, the households which are likely to exclude from beneficial program, are classified as a case of Type I error. In contrast, the households which are incorrectly identified as eligible are classified as a case of Type II error.

Exclusion error rate or Undercoverage is calculated by dividing the number of type I error by the total number of households that should be obtaining the benefit (*E1/N1*). Whereas, inclusion error rate or leakage is calculated by dividing the number of type II error category by the number of households that are selected by the program to be beneficiary (*E2/M1*). Tradeoffs between inclusion and exclusion error rates are concerned since if the objective is to reduce the cost of budgetary, the decreasing of inclusion error rate is preferred. Conversely, if the objective is to increase the welfare of the poor, the alleviation of exclusion error rate is favored.

Table 3.1

Type I and Type II errors

	Truly poor $(p=1)$	Non-poor $(p=0)$	Total
Eligible $(\hat{p}=1)$	Targeting success (S_1)	Type II error Leakage (E_2)	M_1
Ineligible $(\hat{p}=0)$	Type I error Undercoverage (E_1)	Targeting success (S_2)	M_2
Total	N_1	N_2	

Source: Summarized by author.

However, the study of McBride and Nichols (2016) provided the interesting targeting performance that is different from M. Grosh and Baker (1995). Their study categorizes the targeting measurement into 5 forms as follows:

1) Leakage (LE) = (E_2 / N_1),

2) Undercoverage (UC) = (E_1 / N_1),

3) Poverty accuracy (PA) is the correctly predicted poor divided by the total of true poor, which is calculated by (S_1 / N_1),

4) Total accuracy (TA) is the sum of the correctly predicted the poor and non-poor divided by total sample, which is calculated by $(S_1 + S_2 / S_1 + S_2 + E_1 + E_2)$, and

5) Balanced poverty accuracy criterion (BPAC) is the correctly predicted poor divided by the true poor minus the absolute difference between the undercoverage and leakage, which is calculated by $(S_1 / N_1) - \left|(E_1 / N_1) - (E_2 / N_1)\right|$.

Generally, the calculation of the leakage is (E_2 / M_1) but this case the leakage is computed as (E_2 / N_1). McBride and Nichols (2016) suggested that denominator of leakage is adjusted because it is easy to compute BPAC when the denominator of poverty accuracy, leakage and undercoverage are constant.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Data partition

In sampling households for training and test samples, one has to concern with the original design of the survey. Households from SES were sampled for a two-stage procedure. Firstly, primary sampling units (PSUs) were randomly selected. In a second step, households within PSUs were sampled (NSO, 2016). We randomly sample PSUs in order to obtain training and test samples. It is obvious that urban households are over-sampled both in the training and test sets since the urban household samples in initial data are greater than the household samples in rural. Stepwise regression, LASSO and RF are used to identify best selected variables to determine monthly per capita consumption expenditure. We start the variable selection process of Stepwise regression, LASSO and RF with 47 variables. With subset selection we retain only a subset of the variables, and eliminate the rest from the model. Least squares regression is used to estimate coefficients of the inputs that are retained.

Table 4.1 shows the number of urban and rural household observation in initial, training and test sets. In the initial set, the proportions of household live in urban and rural area are 58.80 and 41.20 percent, respectively. While the results of data partition both in training and test sets provide the proportion of household living in urban and rural areas correspond to the proportion of initial set. Therefore, we ensure to use this data partition to estimate the model.

Table 4.1

Number of urban and rural household observations in initial, training and test sets

	Initial Set		Training Set		Test Set	
	Urban	Rural	Urban	Rural	Urban	Rural
Observations	24,394	17,094	12,226	8,518	12,168	8,576
Percent	58.80	41.20	58.94	41.06	58.66	41.34
N	41,488		20,744		20,744	

Source: Author's calculation based on SES data in 2016.

4.2 Variable selection results

4.2.1 Stepwise regression selected variables

After running forward stepwise regression with a 0.01 significant level for addition variable to the model and 0.05 significant level for removal variable from the model. In the national level, 41 variables out of 47 variables are selected by stepwise regression (household head is female, ownership of cooking stove using gas, refrigerator, electric cooking pot, washing machine and fluorescence are excluded from national model). In the urban level, 38 variables out of 47 variables are selected by stepwise regression (household head is female, dwelling that has no toilet, ownership of cooking stove using gas and electricity, refrigerator, electric cooking pot, TV, washing machine and water boiler are excluded from urban model). In the rural level, 38 variables out of 47 variables are selected by stepwise regression (household head is female, dwelling with electricity, dwelling with free rental, ownership of refrigerator, electric cooking pot, TV, washing machine, telephone, and fluorescence are excluded from rural model).

4.2.2 LASSO selected variables

After running 10-folds cross-validation to find the value of lambda in training set, we select the variables from LASSO with 41 and 38 variables out of 47

variables for national and rural levels respectively. For the urban model, we obtain 37 variables since there is no appropriate lambda value to shrink the coefficients to 38 variables.

The lambda values with lowest MSE that provide the above number of selected variables for the national, urban and rural models are 0.0064, 0.0109 and 0.0088 respectively.

In the national level, 41 variables out of 47 variables are selected by LASSO variable selection (household head is female, household head with primary education, ownership of refrigerator, electric cooking pot, water boiler and fluorescence are excluded from national model). In the urban level, 37 variables out of 47 variables are selected by LASSO variable selection (household head is female, number of household working member, household head with lower secondary education, ownership of motorcycle, electric cooking stove using gas, refrigerator, electric cooking pot, washing machine, water boiler and fluorescence are excluded from urban model). In the rural level, 38 variables out of 47 variables are selected by stepwise regression (household head is female, number of household working member, household head with primary secondary education, dwelling with electricity, ownership of refrigerator, electric cooking pot, TV, water boiler and fluorescence are excluded from rural model).

4.2.3 Random Forest selected variables

After tuning the number of variables tried at each split of tree (mtry) in training set, the best mtry values which provide the lowest MSE for the national, urban and rural models are 10, 15 and 10 respectively. From the selected mtry value, RF algorithm is not excluded any variable out of model, however, all 47 variables are ranked by their importance values. The ranking of selected variables depend on its increase in MSE value. The variable with high an increase in MSE value is considered to be more importance.

As the same with variable selection of Stepwise regression and LASSO, we select the same number of selected variables to estimate log of

consumption expenditure with OLS by selecting 41, 38 and 38 variables based on the highest importance score from RF variable selection to compare with Stepwise regression and LASSO.

In the national level, there are 41 variables out of 47 variables that are selected by RF variable selection (household head is female, household head with lower education, dwelling with electricity, dwelling that has no toilet, ownership of telephone and fluorescence are excluded from national model). In the urban level, there are 38 variables out of 47 variables that are selected by RF variable selection (household head with lower and vocational education, dwelling with electricity, dwelling constructs with local material, dwelling with free rental, dwelling that has no toilet, ownership of electric cooking stove using electricity, fluorescence and compact fluorescence are excluded are excluded from urban model). In the rural level, there are 38 variables out of 47 variables that are selected by RF variable selection (household head with lower education, dwelling with electricity, dwelling constructs with local material, dwelling with free rental, drinking water from underground water, dwelling that has no toilet, ownership of telephone, light bulb and fluorescence are excluded from rural model).

4.3 OLS estimation for constructing PMT results

In this section, we run OLS regression to estimate the coefficients of variables selected by Stepwise regression, LASSO and RF. The variables that are statistically insignificant will be dropped from the models. To capture variables from different location, we also add dummy variables of north, northeast and south regions into the models which central region (exclude Bangkok) is the base region.

After running the selected variables from Stepwise regression procedure on OLS regression, the number of variables has statistically significant at 0.01, 0.05 and 0.1 levels are 33, 30 and 32 variables in national, urban and rural models respectively. For the selected variable of LASSO, the number of variables has statistically significant at 0.01, 0.05 and 0.1 levels are 32, 30, 30 variables for national,

urban and rural models respectively. For the selected variable of RF, the number of variables has statistically significant at 0.01, 0.05 and 0.1 levels are 33, 30, 27 variables for national, urban and rural models respectively. Note that these numbers of selected variables are still not included dummy of region and constant value (see table B.3 in appendix B).

4.3.1 OLS regression results for national level

The results of OLS regression based on selected variables from Stepwise regression, LASSO and RF are clarified in table 4.2. There are 3 models; first, *Model I* that run on 33 variables based on selected variables by Stepwise regression procedure.

1. Household characteristics condition, the number of household member has a negative relationship with consumption expenditure per capita. It means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 17 percent. Household head that has married status tends to have the consumption expenditure per capita lower than rear of other status by 14 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases by 1 year, then consumption expenditure per capita will decrease by 0.32 percent. For the number of working household member, if the number of household member that has working status increases 1 person, then the consumption expenditure per capita tends to higher than household that has no member that has working status by 2.2 percent. Educational level of household head clarifies that household with household head has completed with higher level of education tends to has a higher consumption expenditure per capita than household head that has completed with lower educational level. If household head has completed with primary educational level, the consumption expenditure per capita increases by 7 percent, while household head has completed with higher educational level, the consumption expenditure per capita increases by 36 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members tend to have low consumption expenditure per capita. If the proportion of dependency increases by 1 unit, then consumption expenditure per capita will decrease by 1 1, 19 and 3 5 percent respectively for elderly, disable and children members.

3. Housing condition, the household that has more number of rooms in the dwelling is likely to have higher per capita consumption expenditure. If number of room increases by 1 room, the consumption expenditure will increase by 2.3 percent. Dwelling constructed with local material tends to have lower per capita consumption expenditure than dwelling constructed with rear of other material by 22 percent. A household that drink water from river is likely to have consumption expenditure per capita lower than drink water from other source by 9 percent, while drink water from underground source tends to has consumption per capita lower than other source by 12 percent. Dwelling that has no toilet has the consumption expenditure per capita lower than dwelling that has toilet by 24 percent and household using squat are also likely to have lower per capita consumption expenditure than household that use other type of squat by 11 percent.

4. Ownership of assets condition, household that owns motor cycle tends to have consumption expenditure per capita lower than household that does not own by 5.1 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 28 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 25 percent. Household that has a cooking stove using electricity tends to have consumption expenditure per capita higher than household that does not own by 5.3 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 8.2 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 4 percent. Household that has an electric iron tends to have

consumption expenditure per capita higher than household that does not own by 11 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.6 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 6 percent. Household that has an air-conditioner tends to have consumption expenditure per capita higher than household that does not own by 12 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 8.5 percent. Household that has a telephone tends to have consumption expenditure per capita higher than household that does not own by 6.6 percent. Household that has a mobile phone tends to have consumption expenditure per capita higher than household that does not own by 14 percent. Household that has a light bulb tends to have consumption expenditure per capita higher than household that does not own by 3.7 percent. Household that has a compact fluorescence tends to have consumption expenditure per capita higher than household that does not own by 2.5 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is a base region. The households in north, northeast and south regions have per capita consumption expenditure lower than households in the central region by 24, 14 and 4.8 percent on average.

Second, *Model II* that contains 32 variables based on LASSO selected variables.

1. Household characteristics condition, the number of household member has a negative relationship with consumption expenditure per capita. It means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 17 percent. Household head that has married status tends to have the consumption expenditure per capita lower than other status by 14 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases 1 year, then consumption expenditure per capita will decrease by 0.32 percent. For the number of working

household member, if the number of household member that has a working status increases by 1 person, then the consumption expenditure per capita tends to higher than household that has no member that has a working status by 2.2 percent. Educational level of household head clarifies that household with household head has completed with higher level of education tends to has a higher consumption expenditure per capita than household head that has completed with lower educational level. If household head has completed with lower secondary level, the consumption expenditure per capita increases by 9.9 percent, while household head completed with higher educational level, the consumption expenditure per capita increases by 29 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members tend to have low consumption expenditure per capita. If proportion of dependency increases by 1 unit, then consumption expenditure per capita will decrease by 10, 19 and 35 percent, respectively for elderly, disable and children members.

3. Housing condition, the household that has more number of rooms in the dwelling is likely to have higher per capita consumption expenditure. If the number of room increases by 1 room, the consumption expenditure will increase by 2.3 percent. Dwelling constructed with local material tends to have lower per capita consumption expenditure than dwelling constructed with rear of other material by 23 percent. A household that drink water from river is likely to have consumption expenditure per capita lower than drink water from other source by 8.9 percent, while drink water from underground source tends to has consumption per capita lower than other source by 12 percent. Dwelling that has no toilet has the consumption expenditure per capita lower than dwelling that has toilet by 24 percent and household using squat are also likely to have lower per capita consumption expenditure than household that use other type of squat by 11 percent.

4. Ownership of assets condition, the household that owns motor cycle tends to have consumption expenditure per capita lower than household that

does not own by 4.7 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 28 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 25 percent. Household that has a cooking stove using electricity tends to have consumption expenditure per capita higher than household that does not own by 5.2 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 8.2 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 4 percent. Household that has an electric iron tends to have consumption expenditure per capita higher than household that does not own by 11 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.6 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 6 percent. Household that has an air-conditioner tends to have consumption expenditure per capita higher than household that does not own by 12 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 8.5 percent. Household that has a telephone tends to have consumption expenditure per capita higher than household that does not own by 6.7 percent. Household that has a mobile phone tends to have consumption expenditure per capita higher than household that does not own by 14 percent. Household that has a light bulb tends to have consumption expenditure per capita higher than household that does not own by 3.6 percent. Household that has a compact fluorescence tends to have consumption expenditure per capita higher than household that does not own by 2.6 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is a base region. The households in north, northeast and south regions have per capita consumption expenditure lower than households in the central region by 25, 14 and 4.9 percent on average.

Third, *Model III* that contains 33 variables based on RF selected variables.

1. Household characteristics condition, the number of household member has a negative relationship with consumption expenditure per capita. It means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 1 7 percent. Household head that has married status tends to have the consumption expenditure per capita lower than rear of other status by 14 percent. Household head is female is likely to have the consumption per capita lower than household head is male by 1.7 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases by 1 year, then consumption expenditure per capita will be decreased 0.36 percent. For the number of working household member, if number of household member that has working status increases by 1 person, then the consumption expenditure per capita tends to higher than household that has no member that has working status by 2 percent. If household head has completed with primary educational level, the consumption expenditure per capita less than household head has completed with other level by 2.8 percent, while household head has completed with higher educational level, the consumption expenditure per capita increases by 24 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members tend to have low consumption expenditure per capita. If proportion of dependency increases by 1 unit, then consumption expenditure per capita will decrease by 10, 19 and 36 percent respectively for elderly, disable and children members.

3. Housing condition, the household that has more number of rooms in the dwelling is likely to have higher per capita consumption expenditure. If number of room increases by 1 room, the consumption expenditure will increase by 2.3 percent. Dwelling constructed with local material tends to have lower per capita consumption expenditure than dwelling constructed with other material by 25 percent. A household that drink water from river is likely to have consumption

expenditure per capita lower than drink water from other source by 9.1 percent, while drink water from underground source tends to has consumption per capita lower than other source by 12 percent. Household using squat is also likely to have lower per capita consumption expenditure than household that use other type of squat by 11 percent.

4. Ownership of assets condition, the household that owns motor cycle tends to have consumption expenditure per capita lower than household that does not own by 4.4 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 28 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 25 percent. Household that has a cooking stove using electricity tends to have consumption expenditure per capita higher than household that does not own by 5.2 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 8.4 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 4.7 percent. Household that has an electric iron tends to have consumption expenditure per capita higher than household that does not own by 12 percent. Household that has a radio tends to have consumption expenditure per capita higher than household that does not own by 2.2 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.7 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 6.3 percent. Household that has an air-conditioner tends to have consumption expenditure per capita higher than household that does not own by 12 percent. Household that has a water boiler tends to have consumption expenditure per capita higher than household that does not own by 3 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 8.7 percent. Household that has a mobile phone tends to have consumption expenditure per capita higher than

household that does not own by 14 percent. Household that has a light bulb tends to have consumption expenditure per capita higher than household that does not own by 4.2 percent. Household that has a compact fluorescence tends to have consumption expenditure per capita higher than household that does not own by 2.8 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is a base region. The households in north, northeast and south regions have per capita consumption expenditure lower than households in the central region by 26, 14 and 5.4 percent on average.

For the explanatory power of model, *Model I* (Stepwise regression selected variables) has the highest value of R-squared, 0.6641, following by *Model II* (LASSO selected variables) and *Model III* (RF selected variables) with 0.634 and 0.6610 respectively.

Table 4.2

Regression results from OLS estimations for national level

Variable	Coefficient		
	Model I	*Model II*	*Model III*
Number of HH member	-0.1656***	-0.1664***	-0.1668***
	(-30.27)	(-30.49)	(-30.05)
HHH is female			-0.0174*
			-2.08
HHH is married	-0.1373***	-0.1361***	-0.1408***
	(-13.49)	(-13.37)	(-13.63)
Age of HHH (Year)	-0.0032***	-0.0032***	-0.0036***
	(-9.27)	(-9.25)	(-10.76)
Number of working HH member	0.0218***	0.0219***	0.0204***
	(3.97)	(3.95)	(3.66)

Table 4.2 (continued)

Variable	Coefficient		
	Model I	*Model II*	*Model III*
Proportion of HHM aged < 15	-0.3540***	-0.3531***	-0.3566***
	(-13.29)	(-13.24)	(-13.29)
Proportion of HHM aged >= 60	-0.1053***	-0.1041***	-0.0965***
	(-6.44)	(-6.35)	(-6.07)
Proportion of HHM is disable	-0.1923***	-0.1926***	-0.1978***
	(-7.11)	(-7.14)	(-7.41)
HHH with primary education	0.0732***		-0.0281*
	(4.70)		(-2.58)
HHH with lower secondary	0.1668***	0.0985***	
	(8.33)	(8.11)	
HHH with upper secondary	0.2036***	0.1348***	0.0881***
	(10.34)	(9.98)	(6.11)
HHH with vocational education	0.2185***	0.1499***	0.1011***
	(11.32)	(12.31)	(6.74)
HHH with higher education	0.3617***	0.2928***	0.2427***
	(15.14)	(16.86)	(13.13)
Number of rooms	0.0225***	0.0226***	0.0232***
	(3.60)	(3.63)	(3.71)
Electricity in dwelling			
Dwelling constructs with local material	-0.2208**	-0.2328**	-0.2505**
	(-3.22)	(-3.24)	(-3.3)
Rent paid by other			
Drinking water from underground water	-0.1157***	-0.1185***	-0.1242***
	(-5.50)	(-5.71)	(-5.95)

Table 4.2 (continued)

Variable	Coefficient		
	Model I	*Model II*	*Model III*
Drinking water from the river etc.	-0.0895***	-0.0892***	-0.0907***
	(-4.43)	(-4.37)	(-4.49)
Dwelling has no toilet	-0.2357***	-0.2421***	
	(-5.65)	(-5.69)	
Using squat	-0.1070***	-0.1081***	-0.1050***
	(-9.28)	(-9.39)	(-9.21)
Bicycle			
Motorcycle	-0.0512***	-0.0473***	-0.0440***
	(-4.95)	(-4.56)	(-4.08)
Car	0.2757***	0.2757***	0.2785***
	(21.64)	(21.59)	(21.29)
Van or mini-truck	0.2484***	0.2496***	0.2507***
	(19.80)	(19.87)	(19.46)
Other mini-truck			
Cooking stove using gas			
Cooking stove using electricity	0.0537***	0.0522***	0.0516***
	(5.10)	(4.97)	(4.77)
Microwave oven	0.0819***	0.0818***	0.0835***
	(7.35)	(7.32)	(7.6)
Electric pot	0.0397***	0.0419***	0.0466***
	(3.46)	(3.65)	(4.03)
Refrigerator			

Table 4.2 (continued)

Variable	Coefficient		
	Model I	*Model II*	*Model III*
Electric iron	0.1104***	0.1144***	0.1216***
	(9.02)	(9.16)	(9.58)
Electric cooking pot			
Electric fan			
Radio			-0.0222*
			(-2.22)
TV			
LCD or LED or PLASMA	0.0660***	0.0663***	0.0671***
	(8.68)	(8.72)	(8.76)
Video player	0.0602***	0.0604***	0.0628***
	(6.59)	(6.55)	(6.62)
Washing machine			
Air-conditioner	0.1193***	0.1193***	0.1190***
	(10.67)	(10.62)	(9.55)
Water boiler			0.0297*
			(2.04)
Computer	0.0847***	0.0847***	0.0871***
	(8.68)	(8.7)	(8.85)
Telephone	0.0661***	0.0667***	
	(4.05)	(4.06)	
Mobile phone	0.1365***	0.1401***	0.1411***
	(8.47)	(8.74)	(8.69)

Table 4.2 (continued)

Variable	Coefficient		
	Model I	*Model II*	*Model III*
Fluorescence			
Light bulb	0.0365*	0.0364*	0.0415**
	(2.49)	(2.48)	(2.79)
Compact fluorescence	0.0253*	0.0261*	0.0277*
	(2.15)	(2.22)	(2.3)
North	-0.2440***	-0.2470***	-0.2548***
	(-9.67)	(-9.66)	(-9.29)
Northeast	-0.1415***	-0.1374***	-0.1403***
	(-5.82)	(-5.64)	(-5.65)
South	-0.0484	-0.0493	-0.0544
	(-1.42)	(-1.44)	(-1.56)
Constant	8.9142***	8.9719***	9.0342***
	(243.39)	(290.53)	(270.02)
R-squared	0.6641	0.6634	0.6610
Observations	20,744	20,744	20,744

Source: Author's calculation based on training dataset.

Note: t statistics in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.3.2 OLS regression results for urban level

The results of OLS regression based on selected variables from Stepwise regression, LASSO and RF are clarified in table 4.3. There are 3 models; first, Model IV that runs on 30 variables based on selected variables by Stepwise regression procedure.

1. Household characteristics condition, the number of household member has a negative relationship with consumption expenditure per capita. It

means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 18 percent. Household head that has a married status tends to have the consumption expenditure per capita lower than other status by 11 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases by 1 year, then consumption expenditure per capita will decrease by 0.23 percent. For the number of working household member, if the number of household member that has a working status increases by 1 person, then the consumption expenditure per capita tends to higher than household that has no member that has a working status by 2.6 percent. Educational level of household head clarifies that household with household head has completed with higher level of education tends to has a higher consumption expenditure per capita than household head that has completed with lower educational level. If household head completed with primary level, the consumption expenditure per capita increases by 7.4 percent, while household head completed with higher educational level, the consumption expenditure per capita increases by 38 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members tend to have low consumption expenditure per capita. If proportion of dependency increases by 1 unit, then consumption expenditure per capita will be decreased 12, 14 and 27 percent respectively for elderly, disable and children members.

3. Housing condition, the dwelling that use electricity tends to have consumption expenditure per capita higher than household that has no electricity by 48 percent. A household that drink water from river is likely to have consumption expenditure per capita lower than drink water from other source by 14 percent, while drink water from underground source tends to has consumption per capita lower than other source by 18 percent. Household using squat is likely to have lower per capita consumption expenditure than household that use other type of squat by 11 percent.

4. Ownership of assets condition, household that owns bicycle tends to have consumption expenditure per capita lower than household that does not own by 3. 3 percent. Household that owns motor cycle tends to have consumption expenditure per capita lower than household that does not own by 4 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 27 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 23 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 9.2 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 4. 8 percent. Household that has an electric iron tends to have consumption expenditure per capita higher than household that does not own by 12 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.4 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 4.3 percent. Household that has an air-conditioner tends to have consumption expenditure per capita higher than household that does not own by 12 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 8.7 percent. Household that has a telephone tends to have consumption expenditure per capita higher than household that does not own by 9.4 percent. Household that has a mobile phone tends to have consumption expenditure per capita higher than household that does not own by 13 percent. Household that has a light bulb tends to have consumption expenditure per capita higher than household that does not own by 6.3 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is the base region. The households in north and northeast regions have per capita consumption expenditure lower than households in the central region by 2 2, 1 1 percent on average. In contrast, the

households in south region tend to have consumption per capita higher than households in central region 1.6 percent.

Second, *Model V* that contains 3 0 variables based on LASSO selected variables.

1. Household characteristics condition, the number of household member has a negative relationship with consumption expenditure per capita. It means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 1 7 percent. Household head that has married status tends to have the consumption expenditure per capita lower than other status by 10 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases by 1 year, then consumption expenditure per capita will decrease by 0.3 percent. For the education level of household head, if household head has completed with primary educational level, the consumption expenditure per capita tends to lower than household head has completed with other education levels by 4.7 percent, while household head has completed with higher educational level, the consumption expenditure per capita increases by 23 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members tend to have low consumption expenditure per capita. If the proportion of dependency increases by 1 unit, then consumption expenditure per capita will be decreased 1 3, 14 and 3 1 percent respectively for elderly, disable and children members.

3. Housing condition, the household that has more number of rooms in the dwelling is likely to have higher per capita consumption expenditure. If number of room increases by 1 room, the consumption expenditure will increase by 1.7 percent. The dwelling that use electricity tends to have consumption expenditure per capita higher than household that has no electricity by 48 percent. Housing with free rental status is likely to have consumption expenditure per capita higher than housing with other status by 4.6 percent. A household that drink water from river is

likely to have consumption expenditure per capita lower than drink water from other source by 13.5 percent, while drink water from underground source tends to has consumption per capita lower than other source by 18.3 percent. Household using squat is likely to have lower per capita consumption expenditure than household that use other type of squat by 11 percent.

4. Ownership of assets condition, a household that owns bicycle tends to have consumption expenditure per capita lower than household that does not own by 3.3 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 27 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 23 percent. Household that has a cooking stove using electricity tends to have consumption expenditure per capita higher than household that does not own by 4.6 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 8.6 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 4.8 percent. Household that has an electric iron tends to have consumption expenditure per capita higher than household that does not own by 12.4 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.6 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 4.3 percent. Household that has an air-conditioner tends to have consumption expenditure per capita higher than household that does not own by 11 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 7.9 percent. Household that has a telephone tends to have consumption expenditure per capita higher than household that does not own by 9.5 percent. Household that has a mobile phone tends to have consumption expenditure per capita higher than household that does not own by 13 percent.

Household that has a light bulb tends to have consumption expenditure per capita higher than household that does not own by 5.7 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is the base region. The households in north, northeast and south regions have per capita consumption expenditure lower than households in the central region by 25, 13 and 0.47 percent on average.

Third, *Model VI* that contains 30 variables based on RF selected variables.

1. Household characteristics condition, the number of household member has a negative relationship with consumption expenditure per capita. It means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 1 8 percent. Household head that has married status tends to have the consumption expenditure per capita lower than rear of other status by 10 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases by 1 year, then consumption expenditure per capita will decrease by 0. 3 percent. For the number of working household member, if number of household member that has a working status increases by 1 person, then the consumption expenditure per capita tends to higher than household that has no member that has a working status by 2.5 percent. For the education level of household head, if household head has completed with primary educational level, the consumption expenditure per capita tends to lower than household head has completed with other education levels by 7.6 percent, while household head has completed with higher educational level, the consumption expenditure per capita increases by 19 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members tend to have low consumption expenditure per capita. If the proportion of dependency increases by 1 unit, then consumption expenditure per capita will be decreased 1 1, 15 and 26 percent respectively for elderly, disable and children members.

3. Housing condition, the household that has more number of rooms in the dwelling is likely to have higher per capita consumption expenditure. If the number of room increases by 1 room, the consumption expenditure will increase by 1.7 percent. A household that drink water from river is likely to have consumption expenditure per capita lower than drink water from other source by 13 percent, while drink water from underground source tends to has consumption per capita lower than other source by 19 percent. Household using squat is also likely to have lower per capita consumption expenditure than household that use other type of squat by 11 percent.

4. Ownership of assets condition, a household that owns bicycle tends to have consumption expenditure per capita lower than household that does not own by 3. 9 percent. Household that owns motorcycle tends to have consumption expenditure per capita higher than household that does not own by 2.9 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 27 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 23 percent. Household that has a cooking stove using gas tends to have consumption expenditure per capita less than household that does not own by 5.2 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 8. 6 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 5. 2 percent. Household that has an electric iron tends to have consumption expenditure per capita higher than household that does not own by 14 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.4 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 4.2 percent. Household that has an air-conditioner tends to have consumption expenditure per capita higher than household that does not own by 10 percent. Household that has a water boiler

tends to have consumption expenditure per capita higher than household that does not own by 6.5 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 8.5 percent. Household that has a telephone tends to have consumption expenditure per capita higher than household that does not own by 8.7 percent. Household that has a mobile phone tends to have consumption expenditure per capita higher than household that does not own by 15 percent. Household that has a light bulb tends to have consumption expenditure per capita higher than household that does not own by 6 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is the base region. The households in north and northeast regions have per capita consumption expenditure lower than households in the central region by 2 6, 1 3 percent on average. In contrast, the households in south region tend to have consumption per capita higher than households in central region 0.45 percent.

For the explanatory power of model, *Model IV* (Stepwise regression selected variables) has the highest value of R-squared, 0.6718, following by *Model VI* (RF selected variables) and *Model V* (LASSO selected variables) with 0.6698 and 0.6696 respectively.

Table 4.3

Regression results from OLS estimations for urban level

Variable	Coefficient		
	Model IV	*Model V*	*Model VI*
Number of HH member	-0.1765***	-0.1711***	-0.1817***
	(-21.28)	(-25.81)	(-22.55)
HHH is female			
HHH is married	-0.1111***	-0.1034***	-0.1026***
	(-8.84)	(-8.17)	(-7.51)

Table 4.3 (continued)

Variable	Coefficient		
	Model IV	*Model V*	*Model VI*
Age of HHH (Year)	-0.0023***	-0.0030***	-0.0029***
	(-4.88)	(-6.2)	(-6.19)
Number of working HH member	0.0261***		0.0254**
	(3.5)		(3.23)
Proportion of HHM aged < 15	-0.2730***	-0.3121***	-0.2622***
	(-9.42)	(-11.5)	(-8.67)
Proportion of HHM aged >= 60	-0.1164***	-0.1260***	-0.1144***
	(-6.11)	(-6.88)	(-5.98)
Proportion of HHM is disable	-0.1369***	-0.1443***	-0.1463***
	(-4.18)	(-4.57)	(-4.52)
HHH with primary education	0.0743**	-0.0474**	-0.0756***
	(3.39)	(-3)	(-5.64)
HHH with lower secondary	0.1805***		
	(6.64)		
HHH with upper secondary	0.2333***	0.0967***	0.0603***
	(9.59)	(5.34)	(4.02)
HHH with vocational education	0.2266***	0.0913***	
	(9.58)	(4.57)	
HHH with higher education	0.3768***	0.2334***	0.1929***
	(11.68)	(10.27)	(8.34)
Number of rooms		0.0174*	0.0173*
		(2.11)	(2.14)
Electricity in dwelling	0.4750**	0.4758**	
	(2.97)	(2.68)	
Dwelling constructs with local material			

Table 4.3 (continued)

Variable	Coefficient		
	Model IV	*Model V*	*Model VI*
Rent paid by other		0.0457*	
		(2.15)	
Drinking water from underground water	-0.1836***	-0.1828***	-0.1890***
	(-5.9)1	(-5.76)	(-5.98)
Drinking water from the river etc.	-0.1369***	-0.1345***	-0.1333***
	(-5.58)	(-5.33)	(-5.43)
Dwelling has no toilet			
Using squat	-0.1113***	-0.1121***	-0.1137***
	(-8.92)	(-8.45)	(-8.73)
Bicycle	-0.0326**	-0.0330**	-0.0392**
	(-2.75)	(-2.84)	(-3.41)
Motorcycle	-0.0398**		-0.0288*
	(-2.86)		(-2.12)
Car	0.2684***	0.2706***	0.2667***
	(13.89)	(13.72)	(12.53)
Van or mini-truck	0.2267***	0.2326***	0.2299***
	(14.52)	(14.49)	(14.72)
Other mini-truck			
Cooking stove using gas			-0.0515**
			(-2.73)
Cooking stove using electricity		0.0459**	
		(3.05)	
Microwave oven	0.0923***	0.0857***	0.0864***
	(5.49)	(5.2)	(5.32)

Table 4.3 (continued)

Variable	Coefficient		
	Model IV	*Model V*	*Model VI*
Electric pot	0.0476***	0.0484***	0.0523***
	(3.82)	(4.01)	(4.34)
Refrigerator			
Electric iron	0.1198***	0.1243***	0.1395***
	(8.21)	(8.12)	(8.75)
Electric cooking pot			
Electric fan			
Radio			
TV			
LCD or LED or PLASMA	0.0644***	0.0664***	0.0643***
	(4.93)	(4.79)	(4.82)
Video player	0.0434***	0.0425**	0.0424**
	(3.44)	(3.19)	(3.21)
Washing machine			
Air-conditioner	0.1188***	0.1098***	0.1022***
	(8.09)	(7.42)	(6.77)
Water boiler			0.0652**
			(3.12)
Computer	0.0874***	0.0789***	0.0847***
	(7.48)	(6.77)	(6.86)

Table 4.3 (continued)

Variable	Coefficient		
	Model IV	*Model V*	*Model VI*
Telephone	0.0941***	0.0953***	0.0868***
	(4.33)	(4.44)	(4.26)
Mobile phone	0.1248***	0.1289***	0.1487***
	(3.48)	(3.64)	(3.99)
Fluorescence			
Light bulb	0.0627**	0.0574**	0.0598**
	(3.02)	(2.66)	(2.88)
Compact fluorescence			
North	-0.2243***	-0.2493***	-0.2569***
	(-7.33)	(-7.22)	(-7.12)
Northeast	-0.1131**	-0.1274**	-0.1282**
	(-3.18)	(-3.34)	(-3.35)
South	0.0158	-0.0047	0.0045
	(0.32)	(-0.09)	(0.09)
Constant	8.4784***	8.5887***	9.1032***
	(44.09)	(42.6)	(167.67)
R-squared	0.6718	0.6696	0.6698
Observations	12,226	12,226	12,226

Source: Author's calculation based on training dataset.

Note: t statistics in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.3.3 OLS regression results for rural level

The results of OLS regression based on selected variables from Stepwise regression, LASSO and RF are clarified in table 4.4. There are 3 models; first,

Model VII that runs on 32 variables based on selected variables by Stepwise regression procedure.

1. Household characteristics condition, the number of household member has a negative relationship with consumption expenditure per capita. It means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 1 6 percent. Household head that has a married status tends to have the consumption expenditure per capita lower than rear of other status by 15 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases by 1 year, then consumption expenditure per capita will decrease by 0.32 percent. For the number of working household member, if the number of household member that has a working status increases by 1 person, then the consumption expenditure per capita tends to higher than household that has no member that has a working status by 2 percent. Educational level of household head clarifies that household with household head has completed with higher level of education tends to has a higher consumption expenditure per capita than household head that has completed with lower educational level. If the household head has completed with primary educational level, the consumption expenditure per capita increases by 7.7 percent, while household head has completed with higher educational level, the consumption expenditure per capita increases by 34 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members tend to have low consumption expenditure per capita. If proportion of dependency increases by 1 unit, then consumption expenditure per capita will be decreased 9.8, 22 and 37 percent respectively for elderly, disable and children members.

3. Housing condition, the dwelling constructed with local material tends to have consumption expenditure per capita less than dwelling that constructed with other material by 24 percent. Household that has more number of rooms in the dwelling is likely to have higher per capita consumption expenditure. If the number of room increases 1 room, the consumption expenditure will increase by

2.8 percent. A household that drink water from river is likely to have consumption expenditure per capita lower than drink water from other source by 6.8 percent, while drink water from underground source tends to has consumption per capita lower than other source by 8.2 percent. Household that has no toilet tends to have consumption expenditure lower that household that has toilet by 28 percent. Household using squat is likely to have lower per capita consumption expenditure than household that use other type of squat by 9.6 percent.

4. Ownership of assets condition, household that owns motor cycle tends to have consumption expenditure per capita lower than household that does not own by 6.1 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 29 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 27 percent. Household that has a cooking stove using gas tends to have consumption expenditure per capita higher than household that does not own by 3.7 percent. Household that has a cooking stove using electricity tends to have consumption expenditure per capita higher than household that does not own by 6.5 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 8.4 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 3.5 percent. Household that has an electric iron tends to have consumption expenditure per capita higher than household that does not own by 9.6 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.8 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 7.3 percent. Household that has an air-conditioner tends to have consumption expenditure per capita higher than household that does not own by 13 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 9 percent. Household that has a mobile phone tends to have consumption

expenditure per capita higher than household that does not own by 13 percent. Household that has a compact fluorescence tends to have consumption expenditure per capita higher than household that does not own by 3.3 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is the base region. The households in north, northeast and south regions have per capita consumption expenditure lower than households in the central region by 25, 15 and 8 percent on average.

Second, *Model VIII* that contains 3 0 variables based on LASSO selected variables.

1. Household characteristics condition, the number of household member has a negative relationship with consumption expenditure per capita. It means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 1 5 percent. Household head that has married status tends to have the consumption expenditure per capita lower than other status by 14 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases by 1 year, then consumption expenditure per capita will decrease by 0.35 percent. For the education level of household head, if household head has completed with lower education level, the consumption expenditure per capita tends to higher than household head has completed with other education levels by 8 percent, while household head has completed with higher educational level, the consumption expenditure per capita increases by 26 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members tend to have low consumption expenditure per capita. If the proportion of dependency increases by 1 unit, then consumption expenditure per capita will decrease by 1 1, 23 and 42 percent respectively for elderly, disable and children members.

3. Housing condition, the household that has more number of rooms in the dwelling is likely to have higher per capita consumption expenditure. If

the number of room increases 1 room, the consumption expenditure will increase by 2.7 percent. The dwelling constructed with local material tends to have consumption expenditure per capita less than dwelling that constructed with other material by 25 percent. A household that drink water from river is likely to have consumption expenditure per capita lower than drink water from other source by 6.5 percent, while drink water from underground source tends to has consumption per capita lower than other source by 8.5 percent. Household that has no toilet tends to have consumption expenditure lower that household that has toilet by 30 percent. Household using squat is likely to have lower per capita consumption expenditure than household that use other type of squat by 9.7 percent.

4. Ownership of assets condition, a household that owns motorcycle tends to have consumption expenditure per capita lower than household that does not own by 5.5 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 29 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 27 percent. Household that has a cooking stove using electricity tends to have consumption expenditure per capita higher than household that does not own by 6.2 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 8.2 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 3.8 percent. Household that has an electric iron tends to have consumption expenditure per capita higher than household that does not own by 9.8 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.7 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 7.2 percent. Household that has a washing machine tends to have consumption expenditure per capita higher than household that does not own by 3.6 percent. Household that has an air-conditioner tends to have consumption expenditure per capita higher than

household that does not own by 13 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 9 percent. Household that has a mobile phone tends to have consumption expenditure per capita higher than household that does not own by 13.6 percent. Household that has a compact fluorescence tends to have consumption expenditure per capita higher than household that does not own by 3.4 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is the base region. The households in north, northeast and south regions have per capita consumption expenditure lower than households in the central region by 26, 15 and 7.8 percent on average.

Third, *Model IX* that contains 27 variables based on RF selected variables.

1. Household characteristics condition, the number of household member has negative relationship with consumption expenditure per capita. It means that if the number of household member increases by 1 person, then their consumption expenditure per capita will decrease by 1 6 percent. Household head that has a married status tends to have the consumption expenditure per capita lower than other status by 15 percent. Aged of household head has a negative relationship with consumption expenditure per capita. If age increases by 1 year, then consumption expenditure per capita will decrease by 0.38 percent. For the number of working household member, if the number of household member that has a working status increases by 1 person, then the consumption expenditure per capita tends to higher than household that has no member that has a working status by 1.8 percent. For the education level of household head, if a household head has completed with upper secondary level, the consumption expenditure per capita tends to higher than household head completed with other education levels by 8.3 percent, while household head has completed with higher educational level, the consumption expenditure per capita increases by 25 percent.

2. Dependency condition, the household with a large proportion of children members with aged below 15 years, elderly members, and disable members

tend to have low consumption expenditure per capita. If the proportion of dependency increases by 1 unit, then consumption expenditure per capita will decrease by 8.9, 22 and 38 percent respectively for elderly, disable and children members.

3. Housing condition, the household that has more number of rooms in the dwelling is likely to have higher per capita consumption expenditure. If the number of room increases by 1 room, the consumption expenditure will increase by 2.7 percent. A household that drink water from river is likely to have consumption expenditure per capita lower than drink water from other source by 6.2 percent. Household using squat is also likely to have lower per capita consumption expenditure than household that use other type of squat by 9.6 percent.

4. Ownership of assets condition, a household that owns motorcycle tends to have consumption expenditure per capita lower than household that does not own by 5.2 percent. Household that owns car tends to have consumption expenditure per capita higher than household that does not own by 29 percent. Household that owns van or mini-truck tends to have consumption expenditure per capita higher than household that does not own by 27 percent. Household that has a cooking stove using electricity tends to have consumption expenditure per capita less than household that does not own by 6.3 percent. Household that has a microwave oven tends to have consumption expenditure per capita higher than household that does not own by 8 percent. Household that has an electric pot tends to have consumption expenditure per capita higher than household that does not own by 4.4 percent. Household that has an electric iron tends to have consumption expenditure per capita higher than household that does not own by 11 percent. Household that has a LCD or LED or PLASMA tends to have consumption expenditure per capita higher than household that does not own by 6.9 percent. Household that has a video player tends to have consumption expenditure per capita higher than household that does not own by 7.3 percent. Household that has a washing machine tends to have consumption expenditure per capita higher than household that does not own by 3.9 percent. Household that has

an air-conditioner tends to have consumption expenditure per capita higher than household that does not own by 14 percent. Household that has a computer tends to have consumption expenditure per capita higher than household that does not own by 9.6 percent. Household that has a mobile phone tends to have consumption expenditure per capita higher than household that does not own by 14 percent. Household that has a compact fluorescence tends to have consumption expenditure per capita higher than household that does not own by 3.5 percent.

5. Location condition, regional dummy variables are included in model which the central region (exclude Bangkok) is the base region. The households in north, northeast and south regions have per capita consumption expenditure lower than households in the central region by 27, 15 and 9.3 percent on average.

For the explanatory power of model, *Model VII* (Stepwise regression selected variables) has the highest value of R-squared, 0.6261, following by *Model VIII* (LASSO selected variables) and *Model IX* (RF selected variables) with 0.6248 and 0.6202 respectively.

Table 4.4

Regression results from OLS estimations for rural level

Variable	Coefficient		
	Model VII	*Model VIII*	*Model IX*
Number of HH member	-0.1576***	-0.1485***	-0.1597***
	(-23.39)	(-23.97)	(-23.2)
HHH is female			
HHH is married	-0.1474***	-0.1425***	-0.1475***
	(-10.16)	(-10.01)	(-9.82)
Age of HHH (Year)	-0.0032***	-0.0035***	-0.0038***
	(-6.95)	(-7.55)	(-8.51)
Number of working HH member	0.0202**		0.0183*
	(2.85)		(2.54)

Table 4.4 (continued)

Variable	Coefficient		
	Model VII	*Model VIII*	*Model IX*
Proportion of HHM aged < 15	-0.3720***	-0.4182***	-0.3782***
	(-9.63)	(-10.78)	(-9.98)
Proportion of HHM aged >= 60	-0.0984***	-0.1053***	-0.0891***
	(-4)	(-4.26)	(-3.73)
Proportion of HHM is disable	-0.2152***	-0.2262***	-0.2230***
	(-5.82)	(-6.19)	(-6.16)
HHH with primary education	0.0769***		
	(3.73)		
HHH with lower secondary	0.1574***	0.0802***	
	(5.8)	(4.93)	
HHH with upper secondary	0.1743***	0.0942***	0.0825***
	(6.22)	(4.92)	(4.47)
HHH with vocational education	0.2044***	0.1250***	0.1128***
	(7.17)	(6.9)	(6.31)
HHH with higher education	0.3443***	0.2619***	0.2458***
	(9.23)	(9.08)	(8.52)
Number of rooms	0.0283**	0.0272**	0.0270**
	(3.27)	(3.12)	(3.06)
Electricity in dwelling			
Dwelling constructs with local material	-0.2352**	-0.2472**	
	(-3.36)	(-3.24)	
Rent paid by other			
Drinking water from underground water	-0.0822***	-0.0847***	
	(-3.68)	(-3.84)	

Table 4.4 (continued)

Variable	Coefficient		
	Model VII	*Model VIII*	*Model IX*
Drinking water from the river etc.	-0.0682**	-0.0652**	-0.0620**
	(-3.13)	(-2.89)	(-2.76)
Dwelling has no toilet	-0.2819***	-0.3015***	
	(-6.25)	(-6.42)	
Using squat	-0.0961***	-0.0969***	-0.0958***
	(-5.41)	(-5.55)	(-5.52)
Bicycle			
Motorcycle	-0.0611***	-0.0548***	-0.0517**
	(-3.91)	(-3.64)	(-3.2)
Car	0.2897***	0.2911***	0.2922***
	(17.48)	(17.69)	(17.35)
Van or mini-truck	0.2647***	0.2673***	0.2687***
	(15.51)	(15.93)	(16.01)
Other mini-truck			
Cooking stove using gas	0.0369*		
	(2.05)		
Cooking stove using electricity	0.0649***	0.0619***	0.0625***
	(4.11)	(4.04)	(4.16)
Microwave oven	0.0835***	0.0815***	0.0806***
	(5.42)	(5.36)	(5.26)
Electric pot	0.0352*	0.0382*	0.0437*
	(2.17)	(2.27)	(2.56)
Refrigerator			

Table 4.4 (continued)

Variable	Coefficient		
	Model VII	*Model VIII*	*Model IX*
Electric iron	0.0961***	0.0983***	0.1059***
	(5.93)	(5.86)	(6.31)
Electric cooking pot			
Electric fan			
Radio			
TV			
LCD or LED or PLASMA	0.0683***	0.0670***	0.0685***
	(7.7)	(7.45)	(7.54)
Video player	0.0728***	0.0717***	0.0732***
	(5.85)	(5.8)	(5.68)
Washing machine		0.0362**	0.0392**
		(2.75)	(2.85)
Air-conditioner	0.1300***	0.1278***	0.1351***
	(8.15)	(8)	(8.64)
Water boiler			
Computer	0.0902***	0.0895***	0.0960***
	(6.27)	(6.25)	(6.36)
Telephone			
Mobile phone	0.1268***	0.1360***	0.1392***
	(6.91)	(7.46)	(7.27)

Table 4.4 (continued)

Variable	Coefficient		
	Model VII	*Model VIII*	*Model IX*
Fluorescence			
Light bulb			
Compact fluorescence	0.0327*	0.0342*	0.0352*
	(2)	(2.05)	(2.12)
North	-0.2452***	-0.2553***	-0.2663***
	(-7.07)	(-7.09)	(-7.23)
Northeast	-0.1468***	-0.1448***	-0.1517***
	(-4.58)	(-4.41)	(-4.54)
South	-0.0799	-0.0783	-0.0928*
	(-1.97)	(-1.91)	(-2.16)
Constant	8.8560***	8.9489***	8.9493***
	(177.61)	(202.41)	(189.78)
R-squared	0.6261	0.6248	0.6202
Observations	8,518	8,518	8,518

Source: Author's calculation based on training dataset.

Note: t statistics in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.4 Assessing targeting accuracy of PMT results

Before calculating Type I and Type II errors, we rank the actual per capita consumption expenditure of sample household in descending order in national, urban and rural models and rank the cumulative household members of the corresponding households; we define 2,667, 2,902 and 2,425 baht as the cut-off point of national, urban and rural areas respectively to classify the poor household.

Next, we predict per capita consumption expenditure from the OLS regression models[2]. This predicted per capita consumption of household is defined as the PMT score of the households. We rank the household PMT score in descending order and pick the PMT score of household that has the actual consumption expenditure equal to 2,667, 2,902 and 2,425 baht for national, urban and rural areas respectively. Then sum their PMT score to obtain average score as a cut-off PMT. For example in case of national level, there are 4 households that have consumption expenditure equal to 2,667 baht and their PMT scores are 0.42, 0.39, 0.39 and 0.49. Three values of PMT scores are summed and divided by 4; the result is 0.4 as the PMT cut-off score for targeting. After we have finished this process for 9 models, the PMT cut-off score is 0.4 for all models. Any household with the PMT score below 0.4 is considered as the poor household in terms of PMT criteria. Finally, we evaluate the targeting accuracy of PMT by calculating the undercoverage (Type I) and leakage (Type II) errors. For example, if households are defined PMT score higher than 0.4, but the actual consumption per capita less than poverty line. It means that they are truly poor household but PMT score determine them as non-poor household which is undercoverage (or exclusion error) rate.

The targeting performance result of PMT in national is showed in table 4.5. The cut-off score is 0.4 for PMT and 2,667 baht for actual consumption expenditure.

1. Targeting accuracy, to target the number of truly poor and non-poor household, *Model I* (Stepwise regression selected variables) provides the highest percentage of targeting accuracy with 89.11 percent, following by *Model II* (LASSO selected variables) and *Model III* (RF selected variables) with 84.92 and 73.03.

---

[2] To obtain the weight of each variable, its coefficients are multiplied by 100 and rounded the number nearest an integer. Then, multiply value of explanatory variable by its weight. Finally, sum these value with the constant term value that is also multiplied by 100, then we obtain the PMT score for each household (see table B.4-B.6 in appendix B).

2. Poverty accuracy, *Model III* provides the highest poverty accuracy rate compare with *Model I* and *II*. This result indicates that *Model III* targets the truly poor household accurately by 73.83 percent.

3. Leakage rate or Inclusion error, *Model I* has the lowest leakage rate with 50. 29 percent, implies that *Model I* can reduce the number of non-poor household that is classified as the poor household.

4. Undercoverage rate or Exclusion error, *Model III* provides the lowest value of undercoverage rate which has the performance to reduce the number of truly poor household that is classified as non-poor household. While the undercoverage rate of *Model I* and *II* is one time as many as *Model III*.

For assessing PMT performance in the national level, *Model I* with stepwise regression selected variables outperforms *Model II* and *III* in terms of increase targeting accuracy and reduces in an inclusion error. In case of increasing poverty accuracy and reduce in exclusion error, we found that *Model III* with RF selected variables performs better than *Model I* and *II*. When we compare *Model II* with *Model III* that construct PMT model using selected variables from LASSO and RF. The results indicate that *Model II* can increase targeting accuracy and decreases in an inclusion error, while *Model III* increases in poverty accuracy and reduces in an exclusion error. For the poverty accuracy and exclusion error, *Model II* and *III* outperform Model *I* (stepwise regression selected variable). In the overall, it seems that *Model III* (based on RF variables) is appropriate for targeting the poor household in national level in case of reducing exclusion error (undercoverage rate) and gains in number of poor household (poverty accuracy).

Table 4.5

PMT performance in national level

	National		
	Model I	*Model II*	*Model III*
Targeting Accuracy (TA)	89.11	88.82	84.59
Poverty Accuracy (PA)	50.29	51.92	73.83
Inclusion Error (IE)	46.00	47.45	58.65
Exclusion Error (EE)	49.71	48.08	26.17
PMT cut-off score	0.4	0.4	0.4
Poverty line (Baht)	2,667	2,667	2,667
N	20,744	20,744	20,744

Source: Author's calculation.

Note: Using test dataset (out-of-sample) for evaluating the targeting performance.

The targeting performance result of PMT in urban area is showed in table 4.6. The cut-off score is 0.4 for PMT and 2,902 baht for actual consumption expenditure.

1. Targeting accuracy, to target the number of truly poor and non-poor household, *Model VI* (Stepwise regression selected variables) provides the highest percentage of targeting accuracy with 85.50 percent, following by *Model V* (LASSO selected variables) and *Model IV* (RF selected variables) with 84.92 and 73.03.

2. Poverty accuracy, *Model VI* provides the highest poverty accuracy rate compare with *Model IV* and *V*. This result indicates that *Model VI* targets the truly poor household accurately by 87.16 percent.

3. Leakage rate or Inclusion error, *Model IV* has the lowest leakage rate with 56.26 percent, implies that *Model IV* can reduce the number of non-poor household that is classified as the poor household.

4. Undercoverage rate or Exclusion error, *Model VI* provides the lowest value of undercoverage rate which has the performance to reduce the number of

truly poor household that is classified as non-poor household. While the undercoverage rate of *Model IV* and *V* is one time as many as *Model VI*.

For assessing PMT performance in the urban level, *Model IV* with stepwise regression selected variables outperforms *Model V* and *VI* in terms of increases in targeting accuracy and reduces in an inclusion error. In case of increasing poverty accuracy and reduce in exclusion error, we found that *Model VI* with RF selected variables performs better than *Model IV* and *V*. When we compare *Model V* with *Model VI* that construct PMT model using selected variables from LASSO and RF. The results indicate that *Model V* can increase targeting accuracy and decreases in an inclusion error, while *Model VI* increases in poverty accuracy and reduces in an exclusion error. For the poverty accuracy and exclusion error, *Model V* and *VI* outperform Model *IV* (stepwise regression selected variable). In the overall, it seems that *Model VI* (RF selected variables) is appropriate for targeting the poor household in urban level in case of reducing exclusion error (undercoverage rate) and gains in number of the poor household (poverty accuracy).

Table 4.6

PMT performance in urban level

	Urban		
	Model IV	*Model V*	*Model VI*
Targeting Accuracy (TA)	85.50	84.92	73.03
Poverty Accuracy (PA)	73.90	76.65	87.16
Inclusion Error (IE)	56.26	57.26	70.92
Exclusion Error (EE)	26.10	23.35	12.84
PMT cut-off score	0.4	0.4	0.4
Poverty line (Baht)	2,902	2,902	2,902
N	12,168	12,168	12,168

Source: Author's calculation.

Note: Using test dataset (out-of-sample) for evaluating the targeting performance.

The targeting performance result of PMT in rural area is showed in table 4.7. The cut-off score is 0.4 for PMT and 2,425 baht for actual consumption expenditure.

1. Targeting accuracy, to target the number of truly poor and non-poor household, *Model VIII* (LASSO selected variables) provides the highest percentage of targeting accuracy with 88.14 percent, following by *Model VII* (Stepwise regression selected variables) and *Model IX* (RF selected variables) with 87.86 and 79.71 respectively.

2. Poverty accuracy, *Model IX* provides the highest poverty accuracy rate compare with *Model VII* and *VIII*. This result indicates that *Model IX* targets the truly poor household accurately by 79.75 percent.

3. Leakage rate or Inclusion error, *Model VIII* has the lowest leakage rate with 53.31 percent, implies that *Model VIII* can reduce the number of non-poor household that is classified as the poor household.

4. Undercoverage rate or Exclusion error, *Model IX* provides the lowest value of undercoverage rate which has the performance to reduce the number of truly poor household that is classified as non-poor household. While the undercoverage rate of *Model VII* and *VIII* is one time as many as *Model IX*.

For assessing PMT performance in the rural level, *Model VIII* with LASSO selected variables outperforms *Model VII* and *IX* in terms of increases in targeting accuracy and reduces in an inclusion error. In case of increasing poverty accuracy and reduce in exclusion error, we found that *Model IX* with RF selected variables performs better than *Model VI* and *VII*. When we compare *Model VIII* with *Model IX* that construct PMT model using selected variables from LASSO and RF. The results indicate that *Model VIII* can increase targeting accuracy and decreases in an inclusion error, while *Model IX* increases in poverty accuracy and reduces in an exclusion error. In the overall, it seems that *Model IX* (RF selected variables) is appropriate for targeting the poor household in urban level in case of reducing exclusion error (undercoverage rate) and gains in number of poor household (poverty accuracy).

Table 4.7

PMT performance in rural level

	Rural		
	Model VII	*Model VIII*	*Model IX*
Targeting Accuracy (TA)	87.86	88.14	79.71
Poverty Accuracy (PA)	52.99	47.43	79.75
Inclusion Error (IE)	54.01	53.31	67.06
Exclusion Error (EE)	47.01	52.57	20.25
PMT cut-off score	0.4	0.4	0.4
Poverty line (Baht)	2,425	2,425	2,425
N	8,576	8,576	8,576

Source: Author's calculation.

Note: Using test dataset (out-of-sample) for evaluating the targeting performance.

The above results show the number of selected variables from three variable selection processes: Stepwise regression, LASSO and RF which construct the model and select variable in training dataset by evaluate the model performance in test dataset. In the process of PMT construction, we run OLS to estimate the coefficients of selected variables from Stepwise regression, LASSO and RF in training dataset and then calculate the PMT score in test dataset using the coefficient of each variable as weight. Finally, we evaluate the PMT performance by calculating the targeting errors: an inclusion and exclusion errors.

For reducing the number of truly poor households that are classified as non-poor household (an exclusion error) and increasing the number of correctly target the poor household (the poverty accuracy), we found that PMT models that are constructed by selected variables from Random Forest's variable selection process can perform better in all areas (national, urban and rural areas). The exclusion error less than one time of PMT models with selected variables from Stepwise regression and LASSO. In contrast, PMT model based on RF selected

variables do not improve the total accuracy in all case (national, urban and rural), this finding is supported by the study of McBride and Nichols (2016) which found that the Quantile RF and RF cannot improve the total accuracy at all country (Bolivia, Malawi and East-Timor Leste).

Then we consider the variables in the model that can explain the per capita consumption expenditure, most of variable in all PMT models contains the set of variables as the same with the study of Punyasavatsut (2017). In national area, *Model III* (RF selected variables) provides the different variable from *Model I* and *II* which are household head is female, ownership of radio and water boiler. It implies that poor household tends to have the female as a household head (the per capita monthly consumption expenditure less than the male as household head) and household that has the radio and water boiler is likely to be the poor household. For the urban area, *Model VI* (RF selected variables) provides the different variable from *Model IV* and *V* which are the ownership of cooking stove using gas and water boiler. It implies that household that uses the cooking stove with gas and water boiler is likely to be the poor household in urban area. In the rural area, *Model IX* (RF selected variables) provides the set of variables as the same with *Model VII* (Stepwise regression selected variables) and *VIII* (LASSO selected variables). However, *Model IX* provide the number of selected variables (significance) less than other models even through start the variable selection with the same number. The variables that appeared in Model VII and VIII but excluded from Model IX consists of household head has completed with lower secondary education level, dwelling constructed with local material, dwelling has no toilet. It indicates that these variables are quite not good enough to target the poor household in rural area.

In conclusion, PMT models with RF selected variables can accurately target the number of actual poor household in the national, urban and rural levels. When comparing between PMT models with LASSO and RF selected variables, the results show that there is trade-off between two errors: decreasing the exclusion error tends to increase the inclusion error. This study suggest that using LASSO and RF in terms of variable selection for constructing PMT models provide the best

results of reducing exclusion error and increasing poverty targeting which are better than PMT models with Stepwise regression selected variables.

CHAPTER 5
CONCLUSIONS AND RECOMMENDATIONS

Based on Thailand household survey-data (SES) in 2016, this study stipulates and evaluates a series of multiple regression-based PMTs in terms of model's performance in out-of-sample prediction and targeting accuracy performance. LASSO model can provide the sign of explanatory variables that are relative to the per capita consumption expenditure, however, the statistically significant of the set of variable cannot observe. In the same with Random Forest, It cannot interpret the set of variables that are obtained from this model since the coefficients of explanatory variable do not exist for Random Forest. The variable selection of RF can only tell us that which variable has the most influence to the dependent variable, known as "Variable Importance", so we can select the best variables for predicting but cannot interpret how these variables affect the dependent variable. Therefore, the two-step procedure is proposed to solve the problem of variable interpretation. By running OLS regression with only set of selected variables from LASSO and RF algorithm, then select the variables that are significant with log of per capita consumption expenditure to construct the PMT scores. The targeting accuracy performance reveals that when comparing PMT model based on variable selection of LASSO and RF, the PMT model based on LASSO's variable selection outperforms in terms of total accuracy and inclusion error rate. On the other hand, PMT models based on variable selection of RF can perform better in terms of poverty accuracy and exclusion error rate. However, PMT based on variable selection of LASSO and RF performs better PMT based on variable selection of Stepwise regression in some case. For example, PMT model with selected variables from Stepwise regression performs poverty accuracy and inclusion error better than PMT with LASSO and RF in the national and urban level.

For the policy implication, the implementation of proxy means test model for targeting the poor household should concern about difference of area of poverty. Additionally, we suggest PMT models based on variable selection with RF

are more appropriate because the overall performance can perform to target the actual poor household and reduce undercoverage rate better than PMT based on selected variable from LASSO and Stepwise regression. However, when we compare the PMT performance based on LASSO and RF, we will confront with trade-off between undercoverage and leakage error. If policy maker concerns about the budget burden of the program, PMT based on variable selection of LASSO is the suitable choice since a leakage of budget issue is lessen. In contrast, if policy maker would like to coverage the poor household, PMT based on RF's variable selection is appropriate because it can reduce an under-coverage rate and also target the poor household accurately.

For the future study, we would like to improve PMT model based on LASSO and RF algorithm by using two years of household data to examine the overtime results of poor household targeting and also includes the remittance variables to investigate how the remittance effect on income or expenditure.

REFERENCES

Books and Book Articles

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations: CRC press.

Articles

Ahmed, A. U., & Bouis, H. E. (2002). Weighing what's practical: proxy means tests for targeting food subsidies in Egypt. *Food Policy, 27*(5), 519-540.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation, 9*(7), 1545-1588.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli, 19*(2), 521-547.

Bidani, B., & Ravallion, M. (1993). A regional poverty profile for Indonesia. *Bulletin of Indonesian Economic Studies, 29*(3), 37-68.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brown, C., Ravallion, M., & Van De Walle, D. (2016). *A poor means test? econometric targeting in Africa*.

Chan-Lau, M. J. A. (2017). *Lasso Regressions and Forecasting Models in Applied Stress Testing*: International Monetary Fund.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1.

Grosh, M., & Baker, J. L. (1995). Proxy means tests for targeting social programs. *Living Standards Measurement Study Working Paper, 118*, 1-49.

Grosh, M. E. (1994). *Administering targeted social programs in Latin America: From platitudes to practice* (Vol. 94): World Bank Publications.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning* (pp. 9-41): Springer.

Healy, A. J., & Jitsuchon, S. (2007). Finding the poor in Thailand. *Journal of Asian Economics, 18*(5), 739-759.

Ho, T. K. (1995). *Random decision forests.* Paper presented at the Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on.

Houssou, N., Zeller, M., Alcaraz, G., Schwarze, S., & Johannsen, J. (2007). Proxy means tests for targeting the poorest households: Applications to Uganda. *Propoor development in low income countries: Food, agriculture, trade, and environment*.

James, F. C., & McCulloch, C. E. (1990). Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annual review of Ecology and Systematics, 21*(1), 129-166.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *The American economic review, 105*(5), 491-495.

Knippenberg, E., Jensen, N., & Constas, M. (2017). Resilience, Shocks, and the Dynamics of Food Insecurity Evidence from Malawi.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the econometric society*, 33-50.

Kshirsagar, V., Wieczorek, J., Ramanathan, S., & Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach. *arXiv preprint arXiv:1711.06813*.

McBride, L., & Nichols, A. (2015). Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools. *econthatmatters. com/wp-content/uploads/2015/01/improvedtargeting_21jan2015. pdf, retrieved, 4*.

McBride, L., & Nichols, A. (2016). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, lhw056.

Narayan, A., & Yoshida, N. (2005). Proxy Means Tests for Targeting Welfare Benefits in Sri Lanka. *Report No. SASPR–7, Washington, DC: World Bank, http://siteresources. worldbank. org/EXTSAREGTOPPOVRED/Resources/493440-1102216396155/572861-1102221461685/Proxy+ Means+ Test+ for+ Targeting+ Welfare+ Benefits. pdf, accessed February, 5*, 2009.

Nguyen, C., & Lo, D. (2016). Testing Proxy Means Tests in the Field: Evidence from Vietnam.

Otok, B., & Seftiana, D. (2014). The classification of poor households in Jombang with random forest classification and regression trees (RF-CART) approach as the solution in achieving the 2015 Indonesian MDGs' targets. *International Journal of Science and Research (IJSR) Volume, 3*.

Ravallion, M. (1996). Issues in measuring and modeling poverty.

Sohnesen, T. P., & Stender, N. (2017). Is random forest a superior methodology for predicting poverty? an empirical assessment. *Poverty & Public Policy, 9*(1), 118-133.

Thoplan, R. (2014). Random forests for poverty classification. *International Journal of Sciences: Basic and Applied Research (IJSBAR), North America, 17*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives, 28*(2), 3-27.

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*(2), 330-349.

Welling, S. H., Refsgaard, H. H., Brockhoff, P. B., & Clemmensen, L. H. (2016). Forest floor visualizations of random forests. *arXiv preprint arXiv:1605.09196*.

Other Materials

Chanmorchan, P., Pornwalai, T., & Popivanova, C. (n.d.). Thailand's child grant support programme. Retrieved November 23, 2017, from https://transfer.cpc.unc.edu/wp-content/uploads/2016/04/18-Thailands-Child-Grant-Programme.pdf

NSEDB. (n.d.). Report of Poverty and Inequality Circumstance in Thailand 2016. Retrieved May 20, 2018, from https://transfer.cpc.unc.edu/wp-content/uploads/2016/04/18-Thailands-Child-Grant-Programme.pdf

NSO. (n.d.). The 2016 Household Social-Economics Survey in Thailand. Retrieved August 7, 2018, from http://ddi.nso.go.th/index.php/catalog/220

Punyasavatsut, C. (2017). The development of information system for learning opportunities insurance. *Economic Research and Training Center (ERTC), Faculty of Economics, Thammasar University.*

APPENDICES

APPENDIX A

RANDOM FOREST ALGORITHM, OUT-OF-BAG ESTIMATION AND CROSS-VALIDATION PROCESS

A1. Random Forest Algorithm (Breiman, 2001; Hastie et al., 2009)

1. Grow B trees, $\{T_1(X),...,T_B(X)\}$, by recursively repeating step (a)-(c):

a. Randomly select m variables from the total set of M variables.

b. Select variable x_j and split point $x_{ij} = s$ to solve the minimization problem as shown in equation (1)-(3).

c. Split the data into the resulting regions.

2. Output ensemble of trees $\{T_b\}_1^B$.

3. To make prediction at new dataset (test set), drop observation down all trees and calculate $\hat{f}_{RF}(x) = B^{-1}\sum_{b=1}^{B} T_b(x)$, where $b = 1,\mathrm{K}\ ,B$.

A2. Out-of-Bag Estimate of Performance

In theory, the assessment of performance for a prediction algorithm should be completed using a large independent test data set that is not used in training data. In practice, when the data is limited, some type of cross-validation is usually used. The random forest performs the cross-validation in parallel with the training proceed by using the Out-of-Bag (OOB) samples. In the training procedure, each growing tree is using a bootstrap sample since bootstrapping is sampling with replacement from the observation in training data, some of the observations will be left out of the sample while some observations will be repeated in the sample. The left of training set, D^{OOB} is performed in terms of the Out-of-Bag (OOB) sample. Let D_b^{OOB} be the OOB part of the data for b^{th} tree. Then, use the b^{th} tree to predict D_b^{OOB} . Since each training set, X_i, is in an OOB sample, on average, about one thirds of the training set (Breiman, 2001), then calculate the ensemble prediction $\hat{Y}^{OOB}(X_i)$ by aggregating only its OOB predictions. Calculate the estimate of an error rate (ER) for classification or mean square error (MSE) for regression by

$$ER \approx ER^{OOB} = n^{-1}\sum_{i=1}^{n} I(\hat{Y}^{OOB}(X_i) - Y_i)$$

$$MSE \approx MSE^{OOB} = n^{-1}\sum_{i=1}^{n} \{\hat{Y}^{OOB}(X_i) - Y_i\}^2$$

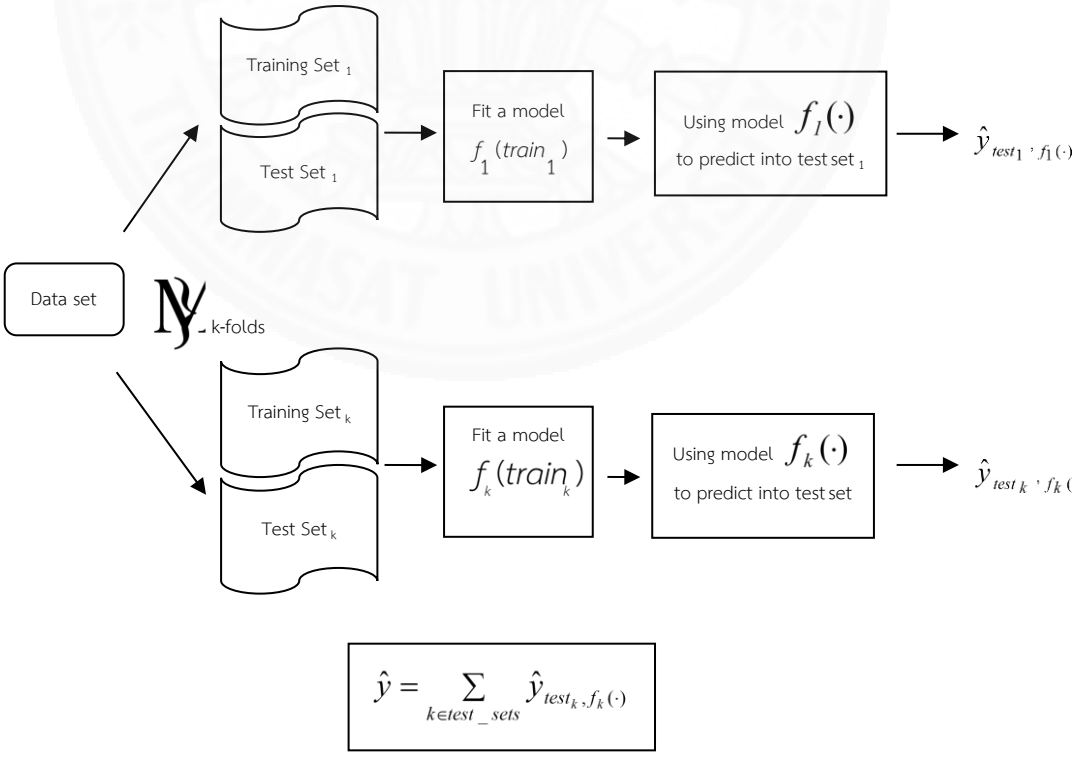
$I(\cdot)$ is defined as an indicator function.

A3. Cross-Validation for Evaluating Out-of-Sample Performance

To assess the performance of each modeling method, we use the k-fold cross-validation method to fit predicted income from each modeling approach, since the good in-sample fit does not guarantee a high performance of out-of-sample. This method partitions the data into several training and test folds, fitting a model on the training set and predicting into test set, and then repeating the process until all folds have been used for prediction.

Figure A.1

Algorithm of Cross-Validation method



Source: Author's summarization.

Figure A. 1 shows the algorithm of k-fold cross-validation that is as follows: firstly, setting k = 10, we divide the data into 10 folds with equal size. Start with fold 1, fitting a model using folds 2 through k of the data, estimating the model $f_1(X_{2\mathrm{K}\,k})$. Using this estimated model to predict into the fold X_1 that is not used for train model (test set), then generating predicted values $\hat{y}_{test_1, f_1(\cdot)}$. Repeat until all folds have been restrained and we have predicted values (per capita consumption expenditure) for all households in *X*.

APPENDIX B

THE RESULTS

Table B.1

Descriptive statistics of variable set

Variable Name	Variable Description	Mean	Std. Dev.	Min	Max
hhsize	Number of HH member	2.90	1.56	1	14
female_head	HHH is female	0.40	0.49	0	1
hhmarried	HHH is married	0.65	0.48	0	1
hhage	Age of HHH (Year)	54.01	15.23	12	99
workingmem	Number of working HH member	1.69	1.05	0	8
prop_lower15	Proportion of HHM aged < 15	0.13	0.19	0	1
prop_upper60	Proportion of HHM aged >= 60	0.25	0.35	0	2
prop_disable	Proportion of HHM is disable	0.04	0.14	0	1
primaryeduc	HHH with primary education	0.59	0.49	0	1
lower_secondary	HHH with lower secondary	0.10	0.30	0	1
upper_secondary	HHH with upper secondary	0.09	0.28	0	1
vocational	HHH with vocational education	0.06	0.24	0	1
highereduc	HHH with higher education	0.11	0.31	0	1
num_rooms	Number of rooms	2.83	1.23	1	9
electric_dwelling	Electricity in dwelling	0.99	0.04	0	1
localmatrl	Dwelling constructs with local material	0.004	0.06	0	1
free_rent	Rental paid by others	0.06	0.24	0	1
drinkwtr_undergr	Drinking water from the well or underground water	0.05	0.21	0	1

Table B.1 (continued)

Variable Name	Variable Description	Mean	Std. Dev.	Min	Max
drinkwtr_river	Drinking water from the river, steam, rain water, etc.	0.13	0.34	0	1
notoilet	Dwelling has no toilet	0.004	0.07	0	1
squat	Using squat	0.59	0.49	0	1
bicycle	Bicycle ownership	0.41	0.49	0	1
motorcycle	Motorcycle ownership	0.82	0.38	0	1
car	Car ownership	0.17	0.37	0	1
van_minitruck	Van or mini-truck ownership	0.29	0.45	0	1
other_minitruck	Other mini-truck ownership	0.11	0.31	0	1
cookingstove_gas	Cooking stove using gas ownership	0.81	0.39	0	1
cookingstove_elec	Cooking stove using electricity ownership	0.15	0.36	0	1
microwave_oven	Microwave oven ownership	0.23	0.42	0	1
electric_pot	Electric pot ownership	0.73	0.44	0	1
refrigerator	Refrigerator ownership	0.92	0.28	0	1
electric_iron	Electric iron ownership	0.82	0.38	0	1
electric_cookpot	Electric cooking pot ownership	0.90	0.31	0	1
electric_fan	Electric fan ownership	0.98	0.14	0	1
radio	Radio ownership	0.44	0.50	0	1
TV	TV ownership	0.77	0.42	0	1
LCD_LED_PLASMA	LCD or LED or PLASMA ownership	0.34	0.47	0	1
video_player	Video player ownership	0.36	0.48	0	1
washingmachine	Washing machine ownership Yes)	0.69	0.46	0	1

Table B.1 (continued)

Variable Name	Variable Description	Mean	Std. Dev.	Min	Max
airconditioner	Air conditioner ownership	0.24	0.43	0	1
waterboiler	Water boiler ownership	0.20	0.40	0	1
computer	Computer ownership	0.21	0.41	0	1
telephone	Telephone ownership	0.06	0.24	0	1
mobilephone	Mobile phone ownership	0.96	0.20	0	1
fluorescences	Fluorescence ownership	0.96	0.20	0	1
lightbulb	Light bulb ownership	0.10	0.30	0	1
compact_fluoresc	Compact fluorescent ownership	0.40	0.49	0	1
N		41,488			

Source: Author's calculation based on SES 2016.

Note: HH = household, HHH = household head, and HHM = household member.

For dummy variable; 0 = No and 1 = Yes.

Table B.2

Thailand poverty line in 2016

	National	Urban	Rural
Poverty line (Baht)	2,667	2,902	2,425
Poor household	2,442	1,456	953
Non-poor household	18,302	10,712	7,623
N	20,744	12,168	8,576

Source: Author's calculation based on test dataset (out-of-sample).

Note: Consumption expenditure below the poverty line classifies as poor.

Table B.3

Set of variables in PMT models

Variable	National			Urban			Rural		
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
Household Characteristics									
Number of HH member	X	X	X	X	X	X	X	X	X
HHH is female			X						
HHH is married	X	X	X	X	X	X	X	X	X
Age of HHH (Year)	X	X	X	X	X	X	X	X	X
Number of working HH member	X	X	X	X		X	X		X
Proportion of HHM aged < 15	X	X	X	X	X	X	X	X	X
Proportion of HHM aged >= 60	X	X	X	X	X	X	X	X	X
Proportion of HHM is disable	X	X	X	X	X	X	X	X	X
HHH with primary education	X		X	X	X	X	X		
HHH with lower secondary	X	X		X			X	X	
HHH with upper secondary	X	X	X	X	X	X	X	X	X
HHH with vocational education	X	X	X	X	X		X	X	X
HHH with higher education	X	X	X	X	X	X	X	X	X
Housing Characteristics									
Number of rooms	X	X	X		X	X	X	X	X
Electricity in dwelling				X	X				
Dwelling constructs with local material	X	X	X				X	X	
Rent paid by other					X				
Drinking water from underground	X	X	X	X	X	X	X	X	
Drinking water from the river, etc.	X	X	X	X	X	X	X	X	X
Dwelling has no toilet	X	X					X	X	
Using squat	X	X	X	X	X	X	X	X	X

Table B.3 (continued)

Variable	National			Urban			Rural		
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
Ownership of Assets									
Bicycle				X	X	X			
Motorcycle	X	X	X	X		X	X	X	X
Car	X	X	X	X	X	X	X	X	X
Van or mini-truck	X	X	X	X	X	X	X	X	X
Other mini-truck									
Cooking stove using gas						X	X		
Cooking stove using electricity	X	X	X		X		X	X	X
Microwave oven	X	X	X	X	X	X	X	X	X
Electric pot	X	X	X	X	X	X	X	X	X
Refrigerator									
Electric iron	X	X	X	X	X	X	X	X	X
Electric cooking pot									
Electric fan									
Radio			X						
TV									
LCD or LED or PLASMA	X	X	X	X	X	X	X	X	X
Video player	X	X	X	X	X	X	X	X	X
Washing machine								X	X
Air-conditioner	X	X	X	X	X	X	X	X	X
Water boiler			X			X			
Computer	X	X	X	X	X	X	X	X	X
Telephone	X	X		X	X	X			
Mobile phone	X	X	X	X	X	X	X	X	X
Fluorescence									

Table B.3 (continued)

Variable	National			Urban			Rural		
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
Light bulb	X	X	X	X	X	X			
Compact fluorescence	X	X	X				X	X	X
Number of variables*	33	32	33	30	30	30	32	30	27

Source: Author's summarization.

Note: * is the number of variables that included in PMT models with significance level of 99 percent and above.

Table B.4

Weight on each variable of OLS estimation results in national level

Variable	Dummy	Weight on each variable		
		I	*II*	*III*
Household Characteristics				
Number of HH member		-17	-17	-17
HHH is female	*	0	0	-2
HHH is married	*	-14	-14	-14
Age of HHH (Year)		0	0	0
Number of working HH member		2	2	2
Proportion of HHM aged < 15		-35	-35	-36
Proportion of HHM aged >= 60		-11	-10	-10
Proportion of HHM is disable		-19	-19	-20
HHH with primary education	*	7	0	-3
HHH with lower secondary	*	17	10	0
HHH with upper secondary	*	20	13	9

Table B.4 (continued)

Variable	Dummy	Weight on each variable		
		I	*II*	*III*
Household Characteristics				
HHH with vocational education	*	22	15	10
HHH with higher education	*	36	29	24
Housing Characteristics				
Number of rooms		2	2	2
Electricity in dwelling	*	0	0	0
Dwelling constructs with local material	*	-22	-23	-25
Rent paid by other	*	0	0	0
Drinking water from underground water	*	-12	-12	-12
Drinking water from the river, etc.	*	-9	-9	-9
Dwelling has no toilet	*	-24	-24	0
Using squat	*	-11	-11	-11
Ownership of Assets				
Bicycle	*	0	0	0
Motorcycle	*	-5	-5	-4
Car	*	28	28	28
Van or mini-truck	*	25	25	25
Other mini-truck	*	0	0	0
Cooking stove using gas	*	0	0	0
Cooking stove using electricity	*	5	5	5
Microwave oven	*	8	8	8
Electric pot	*	4	4	5
Refrigerator	*	0	0	0
Electric iron	*	11	11	12
Electric cooking pot	*	0	0	0

Table B.4 (continued)

Variable	Dummy	Weight on each variable		
		I	*II*	*III*
Household Characteristics				
Electric fan	*	0	0	0
Radio	*	0	0	-2
TV	*	0	0	0
LCD or LED or PLASMA	*	7	7	7
Video player	*	6	6	6
Washing machine	*	0	0	0
Air-conditioner	*	12	12	12
Water boiler	*	0	0	3
Computer	*	8	8	9
Telephone	*	7	7	0
Mobile phone	*	14	14	14
Fluorescence	*	0	0	0
Light bulb	*	4	4	4
Compact fluorescence	*	3	3	3
Location				
North	*	-24	-25	-25
Northeast	*	-14	-14	-14
South	*	-5	-5	-5
Constant	*	891	897	903

Source: Author’s calculation.

Table B.5

Weight on each variable of OLS estimation results in urban level

Variable	Dummy	Weight on each variable		
		IV	*V*	*VI*
Household Characteristics				
Number of HH member		-18	-17	-18
HHH is female	*	0	0	0
HHH is married	*	-11	-10	-10
Age of HHH (Year)		0	0	0
Number of working HH member		3	0	3
Proportion of HHM aged < 15		-27	-31	-26
Proportion of HHM aged >= 60		-12	-13	-11
Proportion of HHM is disable		-14	-14	-15
HHH with primary education	*	7	-5	-8
HHH with lower secondary	*	18	0	0
HHH with upper secondary	*	23	10	6
HHH with vocational education	*	23	9	0
HHH with higher education	*	38	23	19
Housing Characteristics				
Number of rooms		0	2	2
Electricity in dwelling	*	48	48	0
Dwelling constructs with local material	*	0	0	0
Rent paid by other	*	0	5	0
Drinking water from underground water	*	-18	-18	-19
Drinking water from the river, etc.	*	-14	-13	-13
Dwelling has no toilet	*	0	0	0
Using squat	*	-11	-11	-11

Table B.5 (continued)

Variable	Dummy	Weight on each variable		
		IV	*V*	*VI*
Ownership of Assets				
Bicycle	*	-3	-3	-4
Motorcycle	*	-4	0	-3
Car	*	27	27	27
Van or mini-truck	*	23	23	23
Other mini-truck	*	0	0	0
Cooking stove using gas	*	0	0	-5
Cooking stove using electricity	*	0	5	0
Microwave oven	*	9	9	9
Electric pot	*	5	5	5
Refrigerator	*	0	0	0
Electric iron	*	12	12	14
Electric cooking pot	*	0	0	0
Electric fan	*	0	0	0
Radio	*	0	0	0
TV	*	0	0	0
LCD or LED or PLASMA	*	6	7	6
Video player	*	4	4	4
Washing machine	*	0	0	0
Air-conditioner	*	12	11	10
Water boiler	*	0	0	7
Computer	*	9	8	8
Telephone	*	9	10	9
Mobile phone	*	12	13	15
Fluorescence	*	0	0	0

Table B.5 (continued)

Variable	Dummy	Weight on each variable		
		IV	*V*	*VI*
Light bulb	*	6	6	6
Compact fluorescence	*	0	0	0
Location				
North	*	-22	-25	-26
Northeast	*	-11	-13	-13
South	*	2	0	0
Constant	*	848	859	910

Source: Author's calculation.

Table B.6

Weight on each variable of OLS estimation results in rural level

Variable	Dummy	Weight on each variable		
		VII	*VIII*	*IX*
Household Characteristics				
Number of HH member		-16	-15	-16
HHH is female	*	0	0	0
HHH is married	*	-15	-14	-15
Age of HHH (Year)		0	0	0
Number of working HH member		2	0	2
Proportion of HHM aged < 15		-37	-42	-38
Proportion of HHM aged >= 60		-10	-11	-9
Proportion of HHM is disable		-22	-23	-22
HHH with primary education	*	8	0	0
HHH with lower secondary	*	16	8	0
HHH with upper secondary	*	17	9	8
HHH with vocational education	*	20	13	11
HHH with higher education	*	34	26	25
Housing Characteristics				
Number of rooms		3	3	3
Electricity in dwelling	*	0	0	0
Dwelling constructs with local material	*	-24	-25	0
Rent paid by other	*	0	0	0
Drinking water from underground water	*	-8	-8	0
Drinking water from the river, etc.	*	-7	-7	-6
Dwelling has no toilet	*	-28	-30	0
Using squat	*	-10	-10	-10

Table B.6 (continued)

Variable	Dummy	Weight on each variable		
		VII	*VIII*	*IX*
Ownership of Assets				
Bicycle	*	0	0	0
Motorcycle	*	-6	-5	-5
Car	*	29	29	29
Van or mini-truck	*	26	27	27
Other mini-truck	*	0	0	0
Cooking stove using gas	*	4	0	0
Cooking stove using electricity	*	6	6	6
Microwave oven	*	8	8	8
Electric pot	*	4	4	4
Refrigerator	*	0	0	0
Electric iron	*	10	10	11
Electric cooking pot	*	0	0	0
Electric fan	*	0	0	0
Radio	*	0	0	0
TV	*	0	0	0
LCD or LED or PLASMA	*	7	7	7
Video player	*	7	7	7
Washing machine	*	0	4	4
Air-conditioner	*	13	13	14
Water boiler	*	0	0	0
Computer	*	9	9	10
Telephone	*	0	0	0
Mobile phone	*	13	14	14
Fluorescence	*	0	0	0

Table B.6 (continued)

Variable	Dummy	Weight on each variable		
		VII	*VIII*	*IX*
Light bulb	*	0	0	0
Compact fluorescence	*	3	3	4
Location				
North	*	-25	-26	-27
Northeast	*	-15	-14	-15
South	*	-8	-8	-9
Constant	*	886	895	895

Source: Author's calculation.

APPENDIX C

THE ESTIMATION OF PMT MODELS

Table C.1

PMT regression result with Stepwise selected variables in national level

```
Survey: Linear regression

Number of strata   =         2                  Number of obs     =     20,744
Number of PSUs     =       152                  Population size   =  9,330,896
                                                Design df         =        150
                                                F(  36,    115)   =     303.43
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6641
```

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.165554	.0054688	-30.27	0.000	-.1763599	-.1547481
hhage	-.0032043	.0003457	-9.27	0.000	-.0038874	-.0025212
hhmarried	-.1373447	.0101814	-13.49	0.000	-.1574623	-.1172272
workingmem	.0218333	.0054994	3.97	0.000	.010967	.0326996
prop_lower15	-.3539775	.0266278	-13.29	0.000	-.4065916	-.3013634
prop_upper60	-.1052803	.0163569	-6.44	0.000	-.1376	-.0729606
prop_disable	-.1923002	.0270287	-7.11	0.000	-.2457063	-.1388942
primaryeduc	.0732227	.0155729	4.70	0.000	.0424522	.1039932
lower_secondary	.1667596	.0200127	8.33	0.000	.1272164	.2063028
upper_secondary	.2036496	.019695	10.34	0.000	.1647341	.2425652
vocational	.2184651	.0192948	11.32	0.000	.1803404	.2565897
highereduc	.3617146	.0238971	15.14	0.000	.3144962	.4089329
num_rooms	.0224723	.0062453	3.60	0.000	.0101321	.0348125
localmatrl	-.2208263	.0686435	-3.22	0.002	-.3564594	-.0851932
drinkwtr_undergr	-.1156954	.0210438	-5.50	0.000	-.1572759	-.0741149
drinkwtr_river	-.089539	.0202018	-4.43	0.000	-.1294558	-.0496222
notoilet	-.2357447	.0417412	-5.65	0.000	-.3182215	-.153268
squat	-.106979	.0115314	-9.28	0.000	-.129764	-.084194
motorcycle	-.0512015	.0103427	-4.95	0.000	-.0716377	-.0307652
car	.2756588	.012739	21.64	0.000	.2504878	.3008299
van_minitruck	.2484067	.0125469	19.80	0.000	.2236153	.2731981
cookingstove_elec	.0536603	.0105224	5.10	0.000	.032869	.0744517
microwave_oven	.0819042	.0111474	7.35	0.000	.059878	.1039304
electric_pot	.0396507	.0114451	3.46	0.001	.0170363	.0622651
electric_iron	.1104491	.0122437	9.02	0.000	.0862567	.1346414
LCD_LED_PLASMA	.0659861	.0075988	8.68	0.000	.0509716	.0810006
video_player	.0602313	.009145	6.59	0.000	.0421617	.078301
airconditioner	.1192602	.0111765	10.67	0.000	.0971766	.1413439
computer	.084713	.009764	8.68	0.000	.0654202	.1040058
telephone	.0660976	.0163366	4.05	0.000	.0338182	.0983771
mobilephone	.136507	.0161127	8.47	0.000	.1046698	.1683442
lightbulb	.0364816	.0146503	2.49	0.014	.0075341	.0654292
compact_fluoresc	.0252963	.0117555	2.15	0.033	.0020685	.0485241
north	-.2440439	.0252384	-9.67	0.000	-.2939127	-.1941752
northeast	-.1415047	.0243132	-5.82	0.000	-.1895453	-.093464
south	-.0483555	.0340989	-1.42	0.158	-.1157317	.0190206
_cons	8.914182	.0366254	243.39	0.000	8.841814	8.986551

Source: Author's calculation in STATA.

Table C.2

PMT regression result with LASSO selected variables in national level

Survey: Linear regression

Number of strata	=	2	Number of obs	=	20,744
Number of PSUs	=	152	Population size	=	9,330,896
			Design df	=	150
			F(35, 116)	=	282.90
			Prob > F	=	0.0000
			R-squared	=	0.6634

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.166406	.0054572	-30.49	0.000	-.177189	-.155623
hhage	-.0032289	.0003492	-9.25	0.000	-.0039189	-.0025388
hhmarried	-.1360861	.0101815	-13.37	0.000	-.1562037	-.1159685
workingmem	.0219181	.0055466	3.95	0.000	.0109586	.0328776
prop_lower15	-.3531107	.0266797	-13.24	0.000	-.4058272	-.3003942
prop_upper60	-.104051	.0163752	-6.35	0.000	-.136407	-.0716951
prop_disable	-.1925719	.0269543	-7.14	0.000	-.245831	-.1393128
lower_secondary	.0985388	.012145	8.11	0.000	.0745414	.1225361
upper_secondary	.1348325	.0135141	9.98	0.000	.1081298	.1615351
vocational	.1498732	.0121733	12.31	0.000	.1258199	.1739265
highereduc	.2927668	.0173682	16.86	0.000	.2584489	.3270847
num_rooms	.0226155	.0062332	3.63	0.000	.0102994	.0349317
localmatrl	-.2328135	.0719276	-3.24	0.001	-.3749357	-.0906914
drinkwtr_undergr	-.1184704	.0207427	-5.71	0.000	-.1594561	-.0774847
drinkwtr_river	-.0891527	.0204229	-4.37	0.000	-.1295066	-.0487989
notoilet	-.2420899	.0425513	-5.69	0.000	-.3261672	-.1580126
squat	-.1081409	.0115135	-9.39	0.000	-.1308906	-.0853912
motorcycle	-.0472824	.0103761	-4.56	0.000	-.0677846	-.0267802
car	.2756746	.012767	21.59	0.000	.2504482	.300901
van_minitruck	.2495512	.0125586	19.87	0.000	.2247366	.2743657
cookingstove_elec	.0522054	.0105101	4.97	0.000	.0314384	.0729724
microwave_oven	.0818228	.0111794	7.32	0.000	.0597334	.1039121
electric_pot	.0418581	.0114661	3.65	0.000	.0192023	.0645139
electric_iron	.1144031	.0124959	9.16	0.000	.0897124	.1390937
LCD_LED_PLASMA	.0663251	.0076081	8.72	0.000	.0512922	.081358
video_player	.0604287	.0092217	6.55	0.000	.0422076	.0786499
airconditioner	.1193396	.0112384	10.62	0.000	.0971337	.1415456
computer	.0846701	.0097274	8.70	0.000	.0654497	.1038904
telephone	.0667006	.016428	4.06	0.000	.0342405	.0991607
mobilephone	.1401318	.0160316	8.74	0.000	.1084549	.1718088
lightbulb	.0364207	.0147139	2.48	0.014	.0073474	.065494
compact_fluoresc	.0261158	.011788	2.22	0.028	.0028238	.0494078
north	-.2469975	.0255788	-9.66	0.000	-.2975387	-.1964562
northeast	-.137353	.0243524	-5.64	0.000	-.185471	-.0892349
south	-.0492868	.0342235	-1.44	0.152	-.1169092	.0183356
_cons	8.9719	.0308813	290.53	0.000	8.910882	9.032918

Source: Author's calculation in STATA.

Table C.3

PMT regression result with RF selected variables in national level

Survey: Linear regression

Number of strata	=	2	Number of obs	=	20,744
Number of PSUs	=	152	Population size	=	9,330,896
			Design df	=	150
			F(36, 115)	=	324.57
			Prob > F	=	0.0000
			R-squared	=	0.6610

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.1667504	.00555	-30.05	0.000	-.1777167	-.1557841
hhage	-.0036195	.0003365	-10.76	0.000	-.0042843	-.0029546
female_head	-.0174312	.0083744	-2.08	0.039	-.0339782	-.0008843
hhmarried	-.1407695	.0103275	-13.63	0.000	-.1611756	-.1203634
workingmem	.0204323	.0055768	3.66	0.000	.0094132	.0314515
prop_lower15	-.3566278	.0268313	-13.29	0.000	-.4096439	-.3036116
prop_upper60	-.0965353	.0159015	-6.07	0.000	-.1279552	-.0651154
prop_disable	-.1978254	.0266865	-7.41	0.000	-.2505554	-.1450953
primaryeduc	-.0281453	.0109047	-2.58	0.011	-.049692	-.0065986
upper_secondary	.0881165	.0144287	6.11	0.000	.0596069	.1166262
vocational	.1011296	.0150118	6.74	0.000	.0714677	.1307916
highereduc	.2426724	.0184846	13.13	0.000	.2061486	.2791962
num_rooms	.0232006	.0062593	3.71	0.000	.0108328	.0355685
localmatrl	-.250506	.0759477	-3.30	0.001	-.4005714	-.1004406
drinkwtr_undergr	-.1241883	.0208686	-5.95	0.000	-.1654227	-.0829539
drinkwtr_river	-.0906553	.0201999	-4.49	0.000	-.1305685	-.0507422
squat	-.1050046	.0113961	-9.21	0.000	-.1275223	-.0824869
motorcycle	-.0439872	.0107723	-4.08	0.000	-.0652724	-.0227021
car	.2785289	.0130833	21.29	0.000	.2526776	.3043802
van_minitruck	.2506992	.0128824	19.46	0.000	.2252449	.2761535
cookingstove_elec	.0516466	.0108382	4.77	0.000	.0302313	.0730619
microwave_oven	.0834777	.0109824	7.60	0.000	.0617774	.1051779
electric_pot	.0465744	.0115559	4.03	0.000	.0237411	.0694077
electric_iron	.1216041	.0126944	9.58	0.000	.0965212	.146687
radio	-.0222092	.0100137	-2.22	0.028	-.0419953	-.0024231
LCD_LED_PLASMA	.0671112	.0076627	8.76	0.000	.0519705	.082252
video_player	.062829	.0094848	6.62	0.000	.044088	.08157
airconditioner	.1189524	.0124573	9.55	0.000	.094338	.1435668
waterboiler	.0296755	.0145818	2.04	0.044	.0008631	.0584878
computer	.0871438	.009842	8.85	0.000	.067697	.1065906
mobilephone	.1411243	.0162484	8.69	0.000	.1090191	.1732295
lightbulb	.0414825	.0148841	2.79	0.006	.0120728	.0708921
compact_fluoresc	.0276684	.0120273	2.30	0.023	.0039036	.0514331
north	-.2547575	.0274164	-9.29	0.000	-.3089297	-.2005852
northeast	-.1403004	.0248206	-5.65	0.000	-.1893435	-.0912572
south	-.0544349	.0348993	-1.56	0.121	-.1233926	.0145228
_cons	9.034187	.0334569	270.02	0.000	8.96808	9.100295

Source: Author's calculation in STATA.

Table C.4

PMT regression result with Stepwise selected variables in urban level

Survey: Linear regression

Number of strata	=	1	Number of obs	=	12,226
Number of PSUs	=	76	Population size	=	3,631,263
			Design df	=	75
			F(33, 43)	=	534.77
			Prob > F	=	0.0000
			R-squared	=	0.6718

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.1764741	.0082935	-21.28	0.000	-.1929956	-.1599526
hhage	-.0023492	.0004812	-4.88	0.000	-.0033077	-.0013907
hhmarried	-.1111193	.0125769	-8.84	0.000	-.1361737	-.0860648
workingmem	.0261438	.0074673	3.50	0.001	.0112682	.0410193
prop_lower15	-.2729862	.0289854	-9.42	0.000	-.330728	-.2152444
prop_upper60	-.1164089	.0190426	-6.11	0.000	-.1543438	-.0784741
prop_disable	-.136936	.0327799	-4.18	0.000	-.2022369	-.0716351
primaryeduc	.0742717	.0219349	3.39	0.001	.0305752	.1179683
lower_secondary	.1804823	.0271915	6.64	0.000	.1263141	.2346505
upper_secondary	.2332695	.0243365	9.59	0.000	.1847888	.2817502
vocational	.2265525	.0236467	9.58	0.000	.1794459	.2736591
highereduc	.3768392	.0322643	11.68	0.000	.3125655	.441113
electric_dwelling	.4750134	.1596877	2.97	0.004	.1568992	.7931275
drinkwtr_undergr	-.1835501	.0310552	-5.91	0.000	-.2454152	-.1216851
drinkwtr_river	-.1369229	.0245369	-5.58	0.000	-.1858029	-.0880429
squat	-.111279	.0124693	-8.92	0.000	-.1361192	-.0864389
bicycle	-.0326295	.0118669	-2.75	0.007	-.0562697	-.0089893
motorcycle	-.0397846	.0138908	-2.86	0.005	-.0674565	-.0121127
car	.2684051	.0193233	13.89	0.000	.2299111	.3068992
van_minitruck	.2266852	.0156102	14.52	0.000	.1955881	.2577824
microwave_oven	.0922881	.0168103	5.49	0.000	.0588002	.125776
electric_pot	.0476256	.0124695	3.82	0.000	.0227851	.0724661
electric_iron	.1197687	.0145888	8.21	0.000	.0907063	.1488312
LCD_LED_PLASMA	.0644157	.0130595	4.93	0.000	.0383998	.0904316
video_player	.0434168	.0126184	3.44	0.001	.0182796	.068554
airconditioner	.1187702	.0146756	8.09	0.000	.0895349	.1480055
computer	.0873654	.0116741	7.48	0.000	.0641095	.1106213
telephone	.094051	.021701	4.33	0.000	.0508204	.1372816
mobilephone	.1247721	.0358577	3.48	0.001	.0533399	.1962044
lightbulb	.0626875	.020757	3.02	0.003	.0213373	.1040376
north	-.2242675	.0305855	-7.33	0.000	-.285197	-.163338
northeast	-.113142	.0355937	-3.18	0.002	-.1840482	-.0422358
south	.0157628	.0488135	0.32	0.748	-.0814786	.1130043
_cons	8.478428	.1923091	44.09	0.000	8.095328	8.861527

Source: Author's calculation in STATA.

Table C.5

PMT regression result with LASSO selected variables in urban level

Survey: Linear regression

Number of strata	=	1	Number of obs	=	12,226
Number of PSUs	=	76	Population size	=	3,631,263
			Design df	=	75
			F(33, 43)	=	535.07
			Prob > F	=	0.0000
			R-squared	=	0.6696

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.1710828	.0066291	-25.81	0.000	-.1842887	-.157877
hhage	-.0030051	.000485	-6.20	0.000	-.0039712	-.0020389
hhmarried	-.1034095	.012662	-8.17	0.000	-.1286336	-.0781854
prop_lower15	-.3120925	.0271501	-11.50	0.000	-.3661782	-.2580068
prop_upper60	-.1260184	.0183203	-6.88	0.000	-.1625142	-.0895226
prop_disable	-.1443191	.0316121	-4.57	0.000	-.2072936	-.0813446
primaryeduc	-.0474127	.0158112	-3.00	0.004	-.0789102	-.0159153
upper_secondary	.0967459	.0181206	5.34	0.000	.0606478	.1328441
vocational	.0912743	.0199661	4.57	0.000	.0514998	.1310489
highereduc	.2334415	.0227253	10.27	0.000	.1881705	.2787125
num_rooms	.0174456	.0082814	2.11	0.038	.0009481	.033943
electric_dwelling	.4758237	.1776135	2.68	0.009	.1219994	.8296479
free_rent	.0456995	.0212559	2.15	0.035	.0033557	.0880433
drinkwtr_undergr	-.1827974	.0317209	-5.76	0.000	-.2459887	-.119606
drinkwtr_river	-.1345265	.0252371	-5.33	0.000	-.1848013	-.0842517
squat	-.1120909	.013261	-8.45	0.000	-.1385083	-.0856736
bicycle	-.0330312	.0116485	-2.84	0.006	-.0562362	-.0098262
car	.270601	.0197237	13.72	0.000	.2313093	.3098927
van_minitruck	.232624	.0160586	14.49	0.000	.2006336	.2646144
cookingstove_elec	.0459165	.0150657	3.05	0.003	.0159041	.0759289
microwave_oven	.0857095	.0164748	5.20	0.000	.05289	.1185291
electric_pot	.0484002	.0120779	4.01	0.000	.0243397	.0724606
electric_iron	.1243272	.0153147	8.12	0.000	.0938188	.1548357
LCD_LED_PLASMA	.0664242	.0138801	4.79	0.000	.0387736	.0940748
video_player	.0425097	.0133345	3.19	0.002	.015946	.0690735
airconditioner	.1097606	.0147835	7.42	0.000	.0803102	.1392109
computer	.0788986	.0116518	6.77	0.000	.055687	.1021102
telephone	.095323	.0214762	4.44	0.000	.0525403	.1381058
mobilephone	.1288576	.0354113	3.64	0.001	.0583146	.1994005
lightbulb	.0574095	.0215678	2.66	0.010	.0144443	.1003747
north	-.2492677	.0345121	-7.22	0.000	-.3180193	-.180516
northeast	-.1273982	.0381704	-3.34	0.001	-.2034376	-.0513588
south	-.0047126	.0534482	-0.09	0.930	-.111187	.1017617
_cons	8.588741	.2016175	42.60	0.000	8.187099	8.990384

Source: Author's calculation in STATA.

Table C.6

PMT regression result with RF selected variables in urban level

Survey: Linear regression

Number of strata	=	1	
Number of PSUs	=	76	
Number of obs	=	12,226	
Population size	=	3,631,263	
Design df	=	75	
F(33, 43)	=	498.70	
Prob > F	=	0.0000	
R-squared	=	0.6698	

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.181735	.0080604	-22.55	0.000	-.197792	-.1656779
hhage	-.0029479	.0004764	-6.19	0.000	-.0038969	-.0019989
hhmarried	-.1026193	.0136728	-7.51	0.000	-.129857	-.0753817
workingmem	.0254231	.0078632	3.23	0.002	.0097588	.0410874
prop_lower15	-.2621669	.0302363	-8.67	0.000	-.3224008	-.2019331
prop_upper60	-.1144245	.0191441	-5.98	0.000	-.1525615	-.0762876
prop_disable	-.1462696	.0323694	-4.52	0.000	-.2107527	-.0817865
primaryeduc	-.0756216	.0134048	-5.64	0.000	-.1023253	-.0489178
upper_secondary	.0603086	.0149985	4.02	0.000	.0304301	.0901872
highereduc	.1929226	.0231289	8.34	0.000	.1468476	.2389977
num_rooms	.0172943	.0080877	2.14	0.036	.0011828	.0334058
drinkwtr_undergr	-.1889546	.0315784	-5.98	0.000	-.251862	-.1260471
drinkwtr_river	-.1332541	.0245582	-5.43	0.000	-.1821766	-.0843316
squat	-.1137051	.0130209	-8.73	0.000	-.1396441	-.0877662
bicycle	-.0392023	.0115039	-3.41	0.001	-.0621193	-.0162854
motorcycle	-.0288081	.0135621	-2.12	0.037	-.0558252	-.0017911
car	.2666642	.0212817	12.53	0.000	.2242689	.3090595
van_minitruck	.2299297	.0156231	14.72	0.000	.1988069	.2610525
cookingstove_gas	-.0514887	.0188356	-2.73	0.008	-.0890112	-.0139663
microwave_oven	.0863538	.0162337	5.32	0.000	.0540146	.1186931
electric_pot	.0523357	.0120636	4.34	0.000	.0283038	.0763675
electric_iron	.1395268	.0159438	8.75	0.000	.1077652	.1712884
LCD_LED_PLASMA	.0643297	.0133496	4.82	0.000	.037736	.0909235
video_player	.0423735	.0132024	3.21	0.002	.0160729	.0686741
airconditioner	.1022127	.0150928	6.77	0.000	.0721463	.1322791
waterboiler	.0652485	.020931	3.12	0.003	.0235518	.1069452
computer	.0846848	.0123405	6.86	0.000	.0601012	.1092684
telephone	.0867555	.0203688	4.26	0.000	.0461787	.1273323
mobilephone	.1487027	.0372595	3.99	0.000	.074478	.2229274
lightbulb	.0597709	.0207355	2.88	0.005	.0184636	.1010782
north	-.2569234	.0360681	-7.12	0.000	-.3287747	-.185072
northeast	-.1281684	.0382796	-3.35	0.001	-.2044254	-.0519115
south	.0044999	.0519562	0.09	0.931	-.0990021	.1080019
_cons	9.103244	.0542928	167.67	0.000	8.995088	9.211401

Source: Author's calculation in STATA.

Table C.7

PMT regression result with Stepwise selected variables in rural level

```
Survey: Linear regression

Number of strata   =         1                  Number of obs     =      8,518
Number of PSUs     =        76                  Population size   =  5,699,634
                                                Design df         =         75
                                                F(  35,     41)   =     165.63
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6261
```

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.157615	.0067375	-23.39	0.000	-.1710368	-.1441933
hhage	-.0032183	.0004629	-6.95	0.000	-.0041404	-.0022962
hhmarried	-.1473668	.0144988	-10.16	0.000	-.1762498	-.1184837
workingmem	.0202133	.0070976	2.85	0.006	.0060742	.0343524
prop_lower15	-.3719599	.0386168	-9.63	0.000	-.4488885	-.2950313
prop_upper60	-.0983893	.0245974	-4.00	0.000	-.1473899	-.0493888
prop_disable	-.2151577	.0369478	-5.82	0.000	-.2887614	-.141554
primaryeduc	.0768772	.020618	3.73	0.000	.035804	.1179504
lower_secondary	.1573936	.0271472	5.80	0.000	.1033137	.2114736
upper_secondary	.1743253	.0280394	6.22	0.000	.118468	.2301827
vocational	.2043567	.028516	7.17	0.000	.1475499	.2611635
highereduc	.3443228	.0372947	9.23	0.000	.2700279	.4186177
num_rooms	.0282756	.0086599	3.27	0.002	.0110241	.0455271
localmatrl	-.2352142	.0700926	-3.36	0.001	-.3748457	-.0955827
drinkwtr_undergr	-.0821974	.0223098	-3.68	0.000	-.1266409	-.037754
drinkwtr_river	-.0682377	.0218244	-3.13	0.003	-.1117141	-.0247614
notoilet	-.2819218	.0451396	-6.25	0.000	-.3718446	-.1919991
squat	-.096073	.017744	-5.41	0.000	-.1314208	-.0607252
motorcycle	-.0610838	.0156288	-3.91	0.000	-.092218	-.0299497
car	.2896659	.0165682	17.48	0.000	.2566604	.3226713
van_minitruck	.2647377	.0170707	15.51	0.000	.2307312	.2987443
cookingstove_gas	.0368784	.0179994	2.05	0.044	.0010217	.0727352
cookingstove_elec	.0648951	.0158041	4.11	0.000	.0334118	.0963784
microwave_oven	.0834662	.015397	5.42	0.000	.0527938	.1141385
electric_pot	.0352305	.0162	2.17	0.033	.0029585	.0675026
electric_iron	.096145	.0162175	5.93	0.000	.0638381	.1284519
LCD_LED_PLASMA	.0682843	.0088733	7.70	0.000	.0506078	.0859608
video_player	.0727793	.0124429	5.85	0.000	.0479918	.0975668
airconditioner	.1299785	.0159439	8.15	0.000	.0982167	.1617404
computer	.0902448	.0143844	6.27	0.000	.0615896	.1189001
mobilephone	.1268186	.0183448	6.91	0.000	.090274	.1633632
compact_fluoresc	.0327224	.0163967	2.00	0.050	.0000586	.0653863
north	-.2451939	.034705	-7.07	0.000	-.3143297	-.176058
northeast	-.1468423	.0320386	-4.58	0.000	-.2106665	-.0830181
south	-.0799363	.0406317	-1.97	0.053	-.1608787	.0010061
_cons	8.856022	.0498617	177.61	0.000	8.756693	8.955352

Source: Author's calculation in STATA.

Table C.8

PMT regression result with LASSO selected variables in rural level

Survey: Linear regression

Number of strata	=	1	Number of obs	=	8,518
Number of PSUs	=	76	Population size	=	5,699,634
			Design df	=	75
			F(33, 43)	=	154.46
			Prob > F	=	0.0000
			R-squared	=	0.6248

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.1485443	.0061965	-23.97	0.000	-.1608883	-.1362003
hhage	-.0034549	.0004575	-7.55	0.000	-.0043663	-.0025434
hhmarried	-.1424939	.0142349	-10.01	0.000	-.1708513	-.1141365
prop_lower15	-.4182189	.0387968	-10.78	0.000	-.495506	-.3409318
prop_upper60	-.1053236	.0247486	-4.26	0.000	-.1546253	-.0560219
prop_disable	-.2261535	.0365293	-6.19	0.000	-.2989236	-.1533833
lower_secondary	.080179	.0162646	4.93	0.000	.0477782	.1125799
upper_secondary	.094234	.0191475	4.92	0.000	.0560901	.1323779
vocational	.125028	.0181081	6.90	0.000	.0889549	.1611012
highereduc	.2618836	.0288391	9.08	0.000	.2044331	.3193341
num_rooms	.0272461	.0087373	3.12	0.003	.0098406	.0446516
localmatrl	-.2471914	.0763033	-3.24	0.002	-.3991954	-.0951874
drinkwtr_undergr	-.0847338	.022092	-3.84	0.000	-.1287433	-.0407243
drinkwtr_river	-.0651571	.0225203	-2.89	0.005	-.1100198	-.0202943
notoilet	-.3014512	.046991	-6.42	0.000	-.395062	-.2078404
squat	-.096948	.0174696	-5.55	0.000	-.1317492	-.0621467
motorcycle	-.054817	.0150471	-3.64	0.000	-.0847923	-.0248416
car	.2910802	.0164542	17.69	0.000	.2583018	.3238585
van_minitruck	.2672841	.0167788	15.93	0.000	.233859	.3007092
cookingstove_elec	.0618715	.0153183	4.04	0.000	.031356	.0923871
microwave_oven	.0815177	.0152203	5.36	0.000	.0511972	.1118381
electric_pot	.0381929	.0168357	2.27	0.026	.0046544	.0717313
electric_iron	.0982564	.016764	5.86	0.000	.0648607	.131652
LCD_LED_PLASMA	.0669601	.0089822	7.45	0.000	.0490666	.0848536
video_player	.0717141	.0123625	5.80	0.000	.0470868	.0963413
washingmachine	.0361633	.0131593	2.75	0.008	.0099487	.062378
airconditioner	.1277852	.0159698	8.00	0.000	.0959717	.1595987
computer	.0895442	.0143186	6.25	0.000	.06102	.1180684
mobilephone	.135969	.0182204	7.46	0.000	.0996722	.1722659
compact_fluoresc	.0342026	.0167115	2.05	0.044	.0009116	.0674936
north	-.2552579	.0360252	-7.09	0.000	-.3270238	-.1834919
northeast	-.1448178	.0328595	-4.41	0.000	-.2102773	-.0793583
south	-.078265	.0408988	-1.91	0.059	-.1597395	.0032095
_cons	8.948924	.0442119	202.41	0.000	8.860849	9.036998

Source: Author's calculation in STATA.

Table C.9

PMT regression result with RF selected variables in rural level

```
Survey: Linear regression

Number of strata   =         1                  Number of obs     =      8,518
Number of PSUs     =        76                  Population size   =  5,699,634
                                                Design df         =         75
                                                F(  30,     46)   =     174.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6202
```

log_exp	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	-.1597473	.0068867	-23.20	0.000	-.1734663	-.1460284
hhage	-.0038172	.0004487	-8.51	0.000	-.0047111	-.0029233
hhmarried	-.1474823	.0150188	-9.82	0.000	-.1774012	-.1175634
workingmem	.0182856	.0071912	2.54	0.013	.0039601	.0326112
prop_lower15	-.3782347	.0379153	-9.98	0.000	-.4537658	-.3027035
prop_upper60	-.089061	.0239037	-3.73	0.000	-.1366797	-.0414423
prop_disable	-.2230146	.0362139	-6.16	0.000	-.2951563	-.1508729
upper_secondary	.0825325	.0184474	4.47	0.000	.0457835	.1192815
vocational	.1128251	.017878	6.31	0.000	.0772104	.1484399
highereduc	.2457883	.0288368	8.52	0.000	.1883425	.303234
num_rooms	.027003	.0088128	3.06	0.003	.009447	.0445591
drinkwtr_river	-.0620047	.0224585	-2.76	0.007	-.1067443	-.0172651
squat	-.0958055	.017348	-5.52	0.000	-.1303645	-.0612466
motorcycle	-.0516935	.0161604	-3.20	0.002	-.0838867	-.0195002
car	.2922349	.0168437	17.35	0.000	.2586805	.3257893
van_minitruck	.2687069	.0167856	16.01	0.000	.2352682	.3021456
cookingstove_elec	.0625291	.0150476	4.16	0.000	.0325527	.0925056
microwave_oven	.0805975	.0153147	5.26	0.000	.0500891	.1111058
electric_pot	.0437117	.0170691	2.56	0.012	.0097083	.0777152
electric_iron	.1058919	.0167693	6.31	0.000	.0724858	.139298
LCD_LED_PLASMA	.0684531	.0090784	7.54	0.000	.0503679	.0865382
video_player	.0731539	.0128843	5.68	0.000	.0474871	.0988208
washingmachine	.039217	.0137808	2.85	0.006	.0117643	.0666697
airconditioner	.1350553	.0156258	8.64	0.000	.1039271	.1661835
computer	.0959943	.0150911	6.36	0.000	.0659311	.1260574
mobilephone	.1391876	.0191356	7.27	0.000	.1010674	.1773077
compact_fluoresc	.0352047	.0165864	2.12	0.037	.0021628	.0682466
north	-.2663046	.0368545	-7.23	0.000	-.3397225	-.1928866
northeast	-.1516779	.0334128	-4.54	0.000	-.2182396	-.0851162
south	-.0927727	.0429086	-2.16	0.034	-.178251	-.0072944
_cons	8.949345	.0471552	189.78	0.000	8.855407	9.043283

Source: Author's calculation in STATA.

BIOGRAPHY

Name	Miss Pisacha Kambuya
Date of Birth	October 16, 1993
Educational Attainment	2016: Bachelor of Economics (First-Class Honors) Srinakharinwirot University, Thailand
Scholarship	2017: Faculty of Economics scholarship, Thammasat University