INVESTIGATING PREDICTION PERFORMANCE ON REGRESSIONS MODELS USING METHODS WITH MULTIPLE DECISION TREES

by

YUYUE LIAO

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Mathematics

College of Engineering and Physical Sciences

University of Birmingham

24 March 2022

Abstract

With the development of machine learning techniques, the prediction power of classical regression methods has been challenged. This thesis applies decision tree learning to design new algorithms for regression analysis. The objective of this thesis was to improve the prediction power of regression models by adding interaction terms that were generated by decision trees. More specifically, we designed new algorithms that allowed multi decision trees to be created and applied in constructing a regression model. These new algorithms were applied to analyse data from three different research fields.

Since the CART algorithm was developed in 1984 (Breiman et al., 2017), decision trees have become widely applied in both classification analysis and regression analysis. An early hybrid tree-logit regression method was designed by Stainberg et al. (1998), followed by other attempts to design hybrid tree-regression methods. In Chapter 1, we introduced both decision trees and existing hybrid tree-regression methods. We also introduced the overall research plan and the datasets applied in the study.

In Chapter 2, we applied hybrid tree-regression methods in meta-regression analysis and compared them with linear meta-regression. The results from model comparison demonstrate the capability of decision trees in optimising prediction performances of regression models, when all independent variables are binary. Random-effects meta-regression and weighted least squares (WLS) meta-regression are utilised in comparison. From both an analysis of the results of the distinct models and the results of previous studies, we have concluded that trade openness is beneficial for economic growth.

In Chapter 3, we applied linear and hybrid regression methods to analyse factors

that affect fundraising performances of crowdfunding projects. A new hybrid tree-regression method, called hybrid forest-linear regression (HFLR), is found to have much higher prediction power than other models applied in the study. By analysing both Monte-Carlo simulations and the crowdfunding data, it is proven that the HFLR method, which only applies categorical variables to construct decision trees, is able to deal with datasets that include both continuous and binary variables. From the results of the models, we discovered various factors that are influential to crowdfunding success.

In Chapter 4, we applied linear and hybrid regression methods to analyse a survey data about people's willingness to pay (WTP) to an environmental project. The HFLR method is proven to be capable with discovering joint effects between not only binary variables, but also non-binary ordinal variables. Meanwhile, a multi bounded model for the contingent valuation (CV) method is designed and compared with the single bounded model. From the results of the models, we estimated the scale of the starting-point bias of the CV method, as well as the median WTP.

Chapters 2, 3, and 4 contain summaries of their target studies, respectively. Chapter 5 presents the overall summary and discusses ideas for possible future developments of hybrid tree-regression methods.

Acknowledgements

Six and a half years ago, in September 2015, I arrived in the UK as an MSc student at the University of Birmingham. It was my first trip abroad. Two years later, I arrived here again as a PhD student and started this new adventure of four and a half years. As I approach the end of this journey, I would like to thank the ones that have offered me help, advice, and companionship through these years.

First of all, I would like to thank my supervisor, Dr Hui Li, for both the academic support and encouragement she had offered. The change from learning from textbooks to academic research was not easy. Through the years, she helped me constantly with learning from literature and academic writing. She also helped me to understand mathematical and practical meanings of statistical models in deeper ways. Meanwhile, when I felt stressed from the difficulties of research, her encouragement has always been helpful for me to keep my research progress on track. We also had nice conversations about shared hobbies including traveling, music, and art, which I enjoyed a lot.

I would also like to thank my co-supervisor, Professor Yemisi Takwoingi, for her academic support, especially her professional advice about meta-regression analysis for Chapter 2.

Apart from my supervisors, many academics have offered their help with my research. Specifically, I would like to thank Dr Mukti Upadhyay for his support for Chapter 2. I would also like to thank many academics at Huazhong University of Science and Technology for hosting the HUST-Birmingham workshops.

I would like to thank all academic and administration staffs (or ex-staffs) of the School of Mathematics for their help during my journey from 2015 to 2022, especially Professor Michal Kočvara, my master's supervisor; Dr Yunbin Zhao, my

master's course tutor; Dr Biman Chakraborty, who gave feedbacks of my first-year and second-year reports; Professor Chris Parker, who offered a lot of help as the Director of Graduate Students; Professor Michal Grove, who was running of the Maths Support Centre, where I enjoyed working there; and Ms Janette Lowe, for her administrative support, especially during my PhD application. I would also like to thank the lecturers of the modules that I have taken: Dr Sándor Németh, Professor David Smith.....

The collection of data is the cornerstone of statistical studies. I would like to thank Lingling who is the collector of the data applied in Chapter 3, and Yixin who is a good friend, a helpful coworker, and the collector of the data applied in Chapter 4. I would also like to thank Giovanni for our inspirational discussions about machine learning. Additionally, I would like to thank other friends and coworkers in the School of Mathematics: Matt, Nan, Alex, Gianmarco, Mark, Euan, Cara, Any, Tejas, Bradly, Norman.....

Life at the University of Birmingham is not just an academic life. Through the years, I feel both entertained and educated by concerts in Elgar Concert Hall and exhibitions in Braber's Institute of Fine Arts. Although I do not know their names, I would like to thank the musicians that played and sang in those concerts and the staffs of Barber's Institute.

I would also like to thank those people who organised events that enriched my social life. I will always remember the free meals and board games in Canvas Birmingham, the English culture I learned from Globe Cafe, the garden parties and house parties in places of various members of Friends International and D&D members, and events organised by Tea Society, Friends International Society, English Club, City English Corner, Global Buddies, Film Society, Sherlock Society, and other societies I have joined. There are so many people that I met in these organisations

that I shared beautiful memories with, and I could only list the names of a few of them, Tom and Jenna, Aimee, Derrik and Abby, Taylor, Luke, Mike and Rachel, Ella, Glenys, Nick and Pamela, Debs, Helina, Juan, Mini, Andy, Tim, Jess, Abigail, Martin, Harry, Anna, Luka.....

One of my favourite things is to explore the world, and I really appreciated those who shared and increased my enjoyment by travelling with me. I will always remember the moments when I watched the beautiful sunset in Swansea with Vicky, Jerrie, and Lin; felt the lovely combination of Chinese and Western cultures in Shanghai with Jun; learned fun facts from museums in London with Emma; visited the stunning wight cliffs near Brighton with Ke and Chloe; sozzled between the wonderful night lights of Paris with Evana and Shirley; faced the changeable Kent weather with Min; listened to the touching graduation songs in Coimbra with Xiaoyu; promenaded between colourful buildings in Warszawa with Weronika; talked cheerfully about the local and global history of Wroclaw with Vincent; appreciated timeless songs of the Beatles in Liverpool with Summer and Fancy; and admired the intoxicating starry night by the Algarve sea with Freya, Qin, and Xiaoshui.

Although my life in Birmingham is often full, there were hard times, especially during the Covid-19 pandemic. During these hard times, apart from a lot of people who were mentioned above, I am grateful to a few other groups of people. First of all, I caught up with some old friends on Zoom. Although I had not contacted lots of them for a while, we had some nice conversations during that isolating time. Secondly, some of my neighbours spent quality time with me, including Nora, Jasper, Iris, Jade, Ginger, Suki..... Thirdly, in 2021, I met a lot of amazing friends on Clubhouse, including some people mentioned above, Cecilia, Henry, Lola, Dr Zhang, Lans, Leon, and many others. We spent a lot of time talking about travelling, music, and films, and these conversations highly enriched my life during and after

the lockdowns.

The completion of a dissertation is not easy, and I have never imagined starting a relationship during the last few weeks of my PhD study. Nevertheless, during these weeks and through her companionship, my girlfriend gave me a lot of encouragement. Yijing, it is really nice to be with you, and I am looking forward to our graduation trip which I believe will take place very soon.

Last but not least, I would like to thank my family for bringing me to this amazing world, for witnessing and supporting my growth through the years, and for their financial and mental support during my PhD study. I would especially like to thank my mother for her friend-like educational approach, which helped me to gradually become independent through the years.

It is very hard to put everyone that I would thank in this acknowledgement, partly due to the limited length, and partly due to the easy-to-fade nature of memories. However, although some memories may become inaccessible, they are never truly forgotten.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to the Application of Decision Trees in Regression Models

1.1 Introduction

Machine learning techniques, such as decision trees, have become increasingly popular in various areas. Generally speaking, machine learning methods are able to deal with problems that do not satisfy the assumptions of classical statistical methods, and they often have superior prediction power. However, classical statistical methods have the advantage of more easily explainable results. As both machine learning methods and classical statistical methods have their own advantages in modelling, researchers have designed various hybrid methods in order to combine these advantages.

In this thesis, we have designed new hybrid methods between machine learning methods, in relation to decision tree learning, and certain classical regression methods. Compared to existing hybrid tree-regression methods, these new methods allow for the generation of interaction terms from multi decision trees. This enables their inclusion within the regression model. By applying various hybrid tree-regression methods in three specific problems, we are able to compare their prediction accuracies with existing statistical methods, including classic regression methods and previously designed hybrid tree-regression methods.

1.2 Classification and Regression Trees

1.2.1 Classification Trees

Decision tree is a machine learning method originally designed for classification analysis. A classification tree is a binary tree that predicts the value of a categorical dependent variable based on the values of independent variables.

Compared to classical classification methods such as logistic regression, minimum distance classifier (Cormack, 1971), and Fisher's linear discriminant analysis, classification trees apply broken lines instead of straight lines as the borders between different categories. As a result, although each split point of a decision tree is a linear classifier, it can be applied in cases where the categories are not linearly separable.

An example of a classification tree is shown in Figure 1.1, where $X1$ is a contin-

uous variable and $X2$ is a binomial variable. According to this classification tree, if the independent variable $X1 < 0.1$, we check the value of $X2$; if $X2 = 1$, we check the value of $X1$ again; if $X1 < -1.1$, the observation is classified in the first category. If the values of the independent variables do not satisfy some of these conditions, the observation is classified into another leaf node and likely another category.

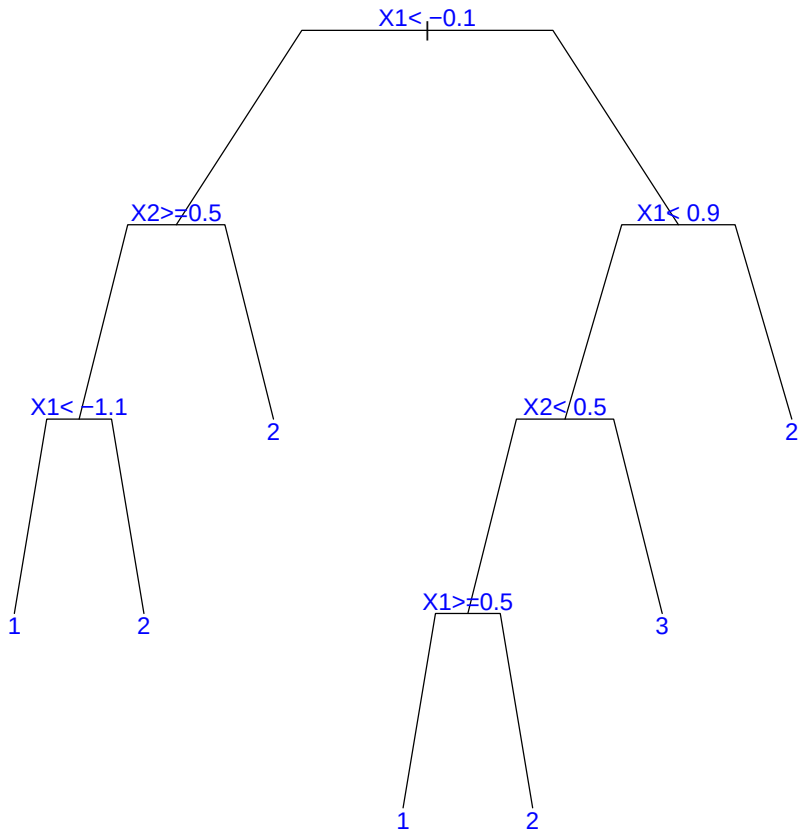


Figure 1.1: A Classification Tree

A brief concept of constructing classification trees is shown in Algorithm 1. Various algorithms with different splitting criteria were designed to build classification trees. For example, the splitting criterion of C4.5 (Quinlan, 2014) algorithm is entropy, whereas that of CART algorithm for classification trees (Breiman et al., 2017)

is the Gini index defined as

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2,$$

where P_i is the probability of class i. The Gini index is a generalisation of binomial variances.

Algorithm 1 Constructing a Decision Tree

Build a root node N containing all the observations used to build the model.
repeat
  $D = N$.
  if a stopping criterion is reached, **then**
    D is returned as a leaf.
  else
    Minimise a splitting criterion to split the observations in D into two categories D_1 and D_2.
    Build two new regression trees for $N_1 = D_1$ and $N_2 = D_2$.
  end if
until all leafs of the tree is returned.

1.2.2 Regression Trees

When applying decision trees in regression analysis, the dependent variable becomes continuous instead of discrete. The CART algorithm for regression trees (Breiman et al., 2017) applies the sum of squared deviations as the splitting criterion. While splitting a note into two categories, the splitting criteria is $SS(D) - (SS(D_1) + SS(D_2))$, where $SS(D) = \sum_D (y - \bar{y}_D)^2$. The CART algorithm can be applied using the "rpart" package (Therneau et al., 1997) in R.

An example of a regression tree is shown in Figure 1.2, where $X1$ is a continuous variable and $X2$ is a binomial variable. According to this regression tree, if the

independent variable $X1 < -0.1$, we check the value of $X2$; if $X2 = 0$, we check the value of $X1$ again; if $X1 < -1.1$, the dependent variable is estimated to be around -1.137. If the values of the independent variables do not satisfy some of these conditions, the observation is classified into another leaf node and the predicted value of the dependent variable may change.

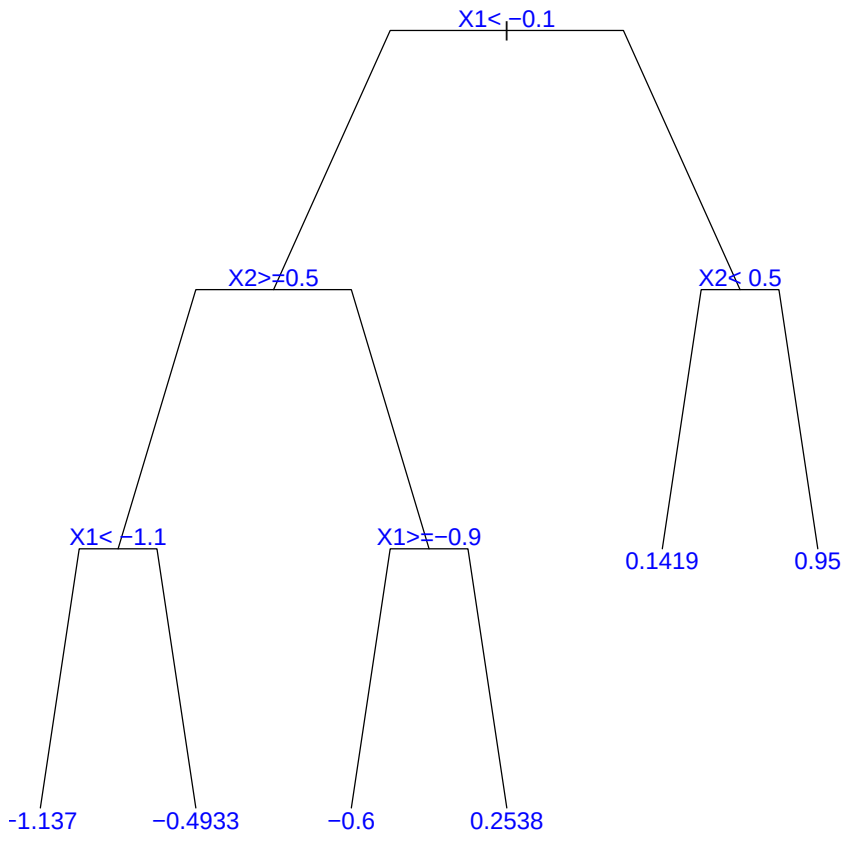


Figure 1.2: A Regression Tree

A complete decision tree constructed by Algorithm 1 classifies is often large and complex, while its prediction power is often limited due to overfitting. The way to solve these problems is pruning the decision tree after constructing it. A pruning criterion is used to decide if each subtree of the decision tree should be pruned. In the

CART algorithm, the pruning method is minimising the cost complexity functions

$$R_\alpha(T) = R(T) + \alpha|T|,$$

where $R(T)$ is the residual sum of squares, $|T|$ is the number of terminal nodes in the decision tree, and α is a given positive constant.

A regression tree is always equivalent to a regression equation in the form of Equation 1.1:

$$y = \sum \beta_i NODE_i + \epsilon, \tag{1.1}$$

where $NODE_i$ is a binomial variable representing the i-th leaf node of the decision tree, and β_i is the relevant coefficient, and ϵ is an error term. Notably, as each observation belongs to only one leaf node, β_i is also the predicted value of y for all observations in $NODE_i$. The regression tree in Figure 1.2 is equivalent to the regression equation

$$\begin{aligned} Y =& \beta_1 \times (X1 < -0.1) \times X2 \times (X1 < -1.1) + \\ & \beta_2 \times (X1 < -0.1) \times X2 \times (X1 \geq -1.1) + \\ & \beta_3 \times (X1 < -0.1) \times (1 - X2) \times (X1 \geq 0.9) + \\ & \beta_4 \times (X1 < -0.1) \times (1 - X2) \times (X1 < 0.9) + \\ & \beta_5 \times (X1 \geq -0.1) \times (1 - X2) + \beta_6 \times (X1 \geq -0.1) \times X2 + \epsilon. \end{aligned}$$

Rearranging Equation 1.1, a regression tree is also equivalent to a regression

equation in the form of Equation 1.2:

$$y = \tilde{\beta}_0 + \sum \tilde{\beta}_i SPLIT_i + \epsilon, \tag{1.2}$$

where $SPLIT_i$ is a binomial variable representing the i-th splitting point of the decision tree, and $\tilde{\beta}_i$ is the relevant coefficient, and ϵ is an error term. For example, the regression tree in Figure 1.2 is equivalent to the regression equation

$$\begin{aligned} Y =& \tilde{\beta}_0 + \tilde{\beta}_1 \times (X1 < -0.1) + \tilde{\beta}_2 \times (X1 < -0.1) \times X2 + \\ & \tilde{\beta}_3 \times (X1 \geq -0.1) \times X2 + \tilde{\beta}_4 \times (X1 < -0.1) \times X2 \times (X1 \geq -1.1) + \\ & \tilde{\beta}_5 \times (X1 < -0.1) \times (1 - X2) \times (X1 \geq 0.9) + \epsilon, \end{aligned}$$

where $\tilde{\beta}_0 = \beta_5$, $\tilde{\beta}_1 = \beta_4 - \beta_5$, $\tilde{\beta}_2 = \beta_2 - \beta_4$, $\tilde{\beta}_3 = \beta_6 - \beta_5$, $\tilde{\beta}_4 = \beta_1 - \beta_2$, and $\tilde{\beta}_5 = \beta_3 - \beta_4$.

Theorem 1.1. *Equation 1.1 is equivalent to Equation 1.2.*

Proof. Consider the case when there is only one splitting point in the regression tree. In this case, there are two leaf nodes $LEAF_1$ to the left and $LEAF_2$ to the right, with coefficients β_1 and β_2. Defining $SPLIT_1$ to be the condition of categorising an observation to the $LEAF_1$, we have $SPLIT_1 = LEAF_1$

$$\begin{aligned} y =& \beta_1 LEAF_1 + \beta_2 LEAF_2 + \epsilon \\ =& \beta_2 + (\beta_1 - \beta_2) LEAF_1 + \epsilon \\ =& \tilde{\beta}_0 + \tilde{\beta}_1 SPLIT_1 + \epsilon, \end{aligned}$$

where $\tilde{\beta}_0 = \beta_2$ and $\tilde{\beta}_1 = \beta_1 - \beta_2$. Thus, Theorem 1.1 is true when there is only one

splitting point in the regression tree.

Assuming that Theorem 1.1 is true for all regression trees with K splitting points. For a regression tree with $K+1$ splitting points, it can be considered as adding a splitting point on the $K+1$-th leaf node of a regression tree with K splitting points. Defining $LEAF'_{K+1}$ and $LEAF'_{K+2}$ to be the new two leafs with coefficients β'_{K+1} and β'_{K+2}, we have $LEAF_{K+1} = LEAF'_{K+1} + LEAF'_{K+2}$. Defining $SPLIT_{K+1}$ to be the new splitting point, we have

$$SPLIT_{K+1} = LEAF'_{K+1} - LEAF_{K+1} = -LEAF'_{K+2},$$

and

$$\begin{aligned} y &= \sum_{k=1}^{K} \beta_k LEAF_k + \beta'_{K+1} LEAF'_{K+1} + \beta'_{K+2} LEAF'_{K+2} + \epsilon \\ &= \sum_{k=1}^{K} \beta_k LEAF_k + \beta'_{K+1}(SPLIT_{K+1} + LEAF_{K+1}) - \beta'_{K+2} SPLIT_{K+1} + \epsilon \\ &= \sum_{k=1}^{K} \beta_k LEAF_k + \beta'_{K+1} LEAF_{K+1} + (\beta'_{K+1} - \beta'_{K+2}) SPLIT_{K+1} + \epsilon. \end{aligned}$$

Since Equation 1.1 is equivalent to Equation 1.2 for all regression trees with K splitting points, there must exist $\tilde{\beta}_k$, $k = 0, \cdots, K$ such that

$$\sum_{k=1}^{K} \beta_i LEAF_i + \beta'_{K+1} LEAF_{K+1} = \tilde{\beta}_0 + \sum_{k=1}^{K} \tilde{\beta}_k SPLIT_k.$$

Thus, we have

$$\begin{aligned}
y =& \sum_{k=1}^{K} \beta_k LEAF_k + \beta'_{K+1} LEAF'_{K+1} + \beta'_{K+2} LEAF'_{K+2} + \epsilon \\
=& \sum_{k=1}^{K} \beta_k LEAF_k + \beta'_{K+1} LEAF_{K+1} + (\beta'_{K+1} - \beta'_{K+2}) SPLIT_{K+1} + \epsilon \\
=& \tilde{\beta}_0 + \sum_{k=1}^{K} \tilde{\beta}_k + (\beta'_{K+1} - \beta'_{K+2}) SPLIT_{K+1} + \epsilon \\
=& \tilde{\beta}_0 + \sum_{k=1}^{K+1} \tilde{\beta}_k SPLIT_k + \epsilon,
\end{aligned}$$

where $\tilde{\beta}_{K+1} = \beta'_{K+1} - \beta'_{K+2}$. Therefore, Theorem 1 is true for all regression trees with $K+1$ splitting points.

Conclusively, as Theorem 1.1 is true for all regression trees with one splitting point, and it is true for all regression trees with $K+1$ splitting point if it is true for all regression trees with K splitting point, it is true for all regression trees. □

Terms in Equation 1.2 are shorter products than those in Equation 1.1. As a result, in practical statistical researches, coefficients in Equation 1.2 often have with clearer meanings. In Chapters 2 to 4, we always use splitting points of decision trees as added terms in regression models.

1.3 Hybrid Tree-regression Methods

1.3.1 Motivation

Hybrid tree-regression methods are algorithms based on both decision trees and regression models; they are developed in order to combine the strengths of both models and avoid their weaknesses.

Decision trees are capable of detecting high-order nonlinear relationships without prior assumptions and their results are less sensitive to outliers. Predictions made by a decision tree, however, are discontinuous, which raises the issue that a small change in a continuous independent variable may result in a dramatic change in the predicted dependent variable.

In comparison, classic regression models, such as linear regression, logistic regression, and probit regression, handle linear relations between variables very well. Without manual adjustments, however, they are not able to detect interaction effects between variables. Although interaction terms could be added manually, it is often difficult to decide which ones should be added. The model could become computationally infeasible if all of them were added. As an example, if thirty categorical variables were discussed in the regression model, 435 interaction terms of two variables and 4060 interaction terms of three variables can be created, which means even if the data contained 4000 observations and only interaction terms up to three-order are considered, the full model is still computationally infeasible.

1.3.2 Hybrid Tree-logit Method

Stainberg et al. (1998) designed a hybrid algorithm based on classification trees and logistic regression. The model generated by the algorithm is written as

$$logit(p) = \sum \alpha_i X_i + \sum \beta_i NODE_i + \epsilon, \tag{1.3}$$

where $logit(p) = \ln(\frac{p}{1-p})$, $p = P(y = 1)$, X_i is a linear independent variable with coefficient α_i, and $NODE_i$, α_i, and ϵ have the same definitions as in Equation 1.1. From the discussions in Section 1.2, Equation 1.3 is equivalent to

$$logit(p) = \tilde{\beta}_0 + \sum \alpha_i X_i + \sum \tilde{\beta}_i SPLIT_i + \epsilon. \tag{1.4}$$

The hybrid tree-logit method was applied in classification analysis in various topics. Zhu et al. (2011) applied the hybrid tree-logit method to analyse stock ranking, and it offered enhanced performance compared to decision trees, logistic regression, and random forest. Lapczynsky (2014) applied the same algorithm in churn analysis, where it performed better than logistic regression, but was outperformed by decision tree.

1.3.3 Further Developments of Hybrid Tree-regression Methods

Similar to the hybrid tree-logit method in classification analysis, Kim et al. (2017) applied a hybrid tree-regression method in regression analysis to combine the benefit of both models. The method is given in Algorithm 2.

Algorithm 2 Hybrid Tree-Regression Method by Kim et al. (2017)

1) Run a decision tree between the dependent variable y and all categorical independent variables.
2) Define $y_{reg} = y - \hat{y_{tree}}$, where $\hat{y_{tree}}$ is the predicted y given by the decision tree in step 1.
3) Run a distance-based regression model with y_{reg} as the dependent variable.
4) Calculate the final estimate $\hat{y} = \hat{y_{reg}} + \hat{y_{tree}}$.

The resulting regression model is written as

$$y = y_{reg} + y_{tree} + \epsilon = \sum \alpha_i X_i + \sum \beta_i NODE_i + \epsilon, \tag{1.5}$$

which is equivalent to

$$y = \tilde{\beta}_0 + \sum \alpha_i X_i + \sum \tilde{\beta}_i SPLIT_i + \epsilon. \tag{1.6}$$

The study applied ten mixed datasets to compare five distance-based regression models (OLS, ridge, lasso, k-NN regression, and SVR) and their hybrid counterparts. The results showed that hybrid methods provide smaller MSEs than their linear counterparts. The model is also interpretable, and the given algorithm is faster than previous ones such as M5.

Dumitrescu et al. (2018) is the first attempt, to our knowledge, to incorporate results of multi decision trees into one regression model. The resulting algorithm, referred to as penalized logit-tree regression (PLTR), is a divide-and-conquer algorithm which applies a decision tree for each continuous regressor. The PLTR method is given in Algorithm 3.

The resulting regression model is equivalent to Equations 1.3 and 1.4. Based on

Algorithm 3 Penalized Logit-tree Regression (PLTR) by Dumitrescu et al. (2018)

1.1) Run a decision tree between the dependent variable and each continuous independent variable.
1.2) Run a decision tree between the dependent variable and each pair of two continuous independent variables.
2) Define categorical variables based on all leaf nodes of all created decision trees in Step 1.
3) Build a regression model between the dependent variable and all independent variables including categorical variables defined in Step 2.
4) Apply lasso to select variables included in the regression model.

simulated and real datasets, PLTR models have much higher prediction accuracy than both linear and quadric logistic models. Although the prediction accuracy of PLTR models is slightly lower than that of random forests, they are much more explainable.

The benefit of PLTR models compared to linear and quadric logistic models is that PLTR allows for the existence of discontinuities in the relationship between the dependent variable and a continuous regressor. In comparison =, joint effects between different variables are also considered in the decision trees built in Step 1.2 of Algorithm 3.

1.3.4 New Methods in this Thesis

In this thesis, two new hybrid tree-regression algorithms were designed to improve the prediction accuracy of regression models. Both of them allow results from multi decision trees to be included in one regression model.

Compared to methods which only include results from one decision tree, those

that include results from multi decision trees allow more diversity of the form of terms. As an example, consider a simple function

$$f(X) = \beta_1 X_1 X_2 + \beta_2 X_1 X_3 + \beta_3 X_2 X_3, \tag{1.7}$$

where all independent variables are binomial. In order to apply only one decision tree to describe the relationship in Equation 1.7, it has to be rewritten as

$$\begin{aligned} f(X) =& X_1 f_l(X_2, X_3) + (1 - X_1) f_r(X_2, X_3) \\ =& X_1(\beta_1 X_2 + \beta_2 X_3 + \beta_3 X_2 X_3) + (1 - X_1)(\beta_3 X_2 X_3) \\ =& X_1 X_2 f_{ll}(X_3) + X_1(1 - X_2) f_{lr}(X_3) + \beta_3(1 - X_1) X_2 X_3 \\ =& X_1 X_2(\beta_1 + (\beta_2 + \beta_3) X_3) + X_1(1 - X_2)\beta_2 X_3 + \beta_3(1 - X_1) X_2 X_3 \\ =& \beta_1 X_1 X_2 + (\beta_2 + \beta_3) X_1 X_2 X_3 + \beta_2 X_1(1 - X_2) X_3 + 0 \cdot (1 - X_1) X_2 + \\ & \beta_3(1 - X_1) X_2 X_3, \end{aligned} \tag{1.8}$$

which is in the form of Equation 1.4 and is equivalent to the result of a decision tree. On the other hand, theoretically, Equation 1.7 can keep its original form when it is represented by the results of three decision trees.

Compared to Equation 1.7, Equation 1.8 have five parameters (or four parameters if a model selection process is applied to delete the zero term) instead of three. Meanwhile, although Equations 1.7 and 1.8 are mathematically equivalent, when the independent variables has practical meanings, the terms in Equation 1.7 is much more interpretable than those in Equation 1.8.

One of the new methods developed in this thesis, called hybrid groupwise tree-linear regression (HGTLR), is a divide-and-conquer algorithm. Compared to PLTR,

instead of applying a decision tree for each regressor, HGTLR applied a decision tree for each pre-defined subset of the dataset.

The other new method is a loop algorithm, called hybrid forest-linear regression (HFLR). Instead of applying a fixed amount of decision trees, the HFLR algorithm generated and added new decision trees continuously until a given loop ending condition is satisfied. The loop ending condition refers to a comparison between the model before and after a cycle.

1.4 Research Plan and Data

The objective of this thesis is to test the prediction accuracy of hybrid tree-regression methods, especially the new methods developed throughout the thesis. For each dataset, we compare the prediction results of a classic regression method, an existing hybrid tree-regression method which only include results from one decision tree, and a new hybrid tree-regression method which include results from multiple decision trees. To compare the prediction efficiencies of the methods in various cases, the dependent variable of each model can be applied to make meaningful predictions, while other aspects of the datasets are different to one other. Basic information on the datasets and models are listed in Table 1.1.

Chapter 2 applies meta-data collected from academic papers about the impact of trade openness to economic growth. The meta-data applied in the modelling process include 452 models gathered from sixty-five papers published between 1995 and 2016.

Table 1.1: Differences between Datasets applied in Chapters 2-4

Chapter number	2	3	4
Data source	Academic papers	Online platforms	Face-to-face survey
Dependent variable	Continuous	Continuous	Categorical
Independent variables	Categorical	Continuous and categorical	Continuous and categorical
Variables to generate decision trees	Categorical, binary	Categorical, binary	Categorical, binary and non-binary ordinal
Classic method	Random effects, linear WLS	Linear OLS	Probit

The dependent variable, called partial correlation coefficient (*pcc*), is applied to predict if an econometric model has positive significant, negative significant, or insignificant result. All independent variables are designed to be binary variables, and they are all applied to construct decision trees. Hybrid versions of random effects meta-regression and weighted least squares (WLS) meta-regression are constructed and compared with their linear counterparts.

The raw database included 295 papers that contained at least one of the keywords "trade", "openness" and "growth" in their titles. However, some studies were not applicable in the study due to various reasons. Firstly, some papers had not reported sufficient statistical results for calculating their effect sizes. Secondly, instead of considering the relationship between growth and openness, some studies discussed growth and openness separately, while some others only discussed one of the topics. Thirdly, plenty of different variables are applied in different studies as indicators of

growth and openness, which means that their results are not comparable.

To achieve higher comparability of the studies as well as a lower rate of missing data, the data was rearranged to select studies that satisfied three conditions: GDP, GDP per capita, or GDP per unit of labour was applied as the indicator of economic growth; the trade openness index was applied as the indicator of openness; and sufficient reported information to calculate the t-statistic and its degrees of freedom.

Chapter 3 applies a dataset collected from online crowdfunding platforms about fundraising performances of crowdfunding projects. By searching projects using the keywords "social enterprise" and "social entrepreneurship" and removing projects with extreme values, we obtained 236 projects completed between 2016 and 2018. Among these projects, fifty-two were successful and 184 were unsuccessful.

The dependent variable, which is the proportion of financing, is applied to predict if a crowdfunding project is successful or not. Independent variables are designed to be a mixture of continuous variables and categorical variables, and all categorical variables are binary. They include topics of social entrepreneurship projects, information about the founders, types of rewards for sponsors, and subjects of pictures displayed on the websites. Only categorical independent variables are applied to construct decision trees. Linear OLS regression and its hybrid tree-regression versions are constructed and their prediction performances are compared. Compared to Chapter 2, this chapter tested the performance of hybrid tree-regression methods with the existence of continuous regressors.

Chapter 4 applies data collected from a face-to-face survey in four Chinese cities. During the survey, participants were asked about their willingness to pay to a geo-

engineering project as well as their income, political and social views, and other personal characters. 1044 samples were collected in the survey and 778 were applied in the modelling process.

Probit regression and its hybrid tree-regression versions are applied to predict the yes or no response of a participant to a given bid, which indicates if the given bid is higher or lower than the willingness to pay (WTP) of a participant. Independent variables are designed to be a mixture of continuous variables and categorical variables, and categorical variables include both binary variables and non-binary ordinal ones. All categorical independent variables are applied to construct decision trees. Compared to Chapter 3, this chapter tested the performance of hybrid tree-regression methods when non-binary ordinal variables are applied to construct decision trees.

1.5 Chapter Summary

Decision tree is a machine learning method applied in both classification analysis and regression analysis. Hybrid methods of decision tree and classical regression models were applied in various studies to combine the strengths of both models. From the results of those studies, the hybrid algorithm outperformed classical regression models including linear regression and logistic regression. Therefore, the application and development of hybrid tree-regression algorithms will help researchers to build regression models with higher prediction accuracy and find out key interaction effects of independent variables to a dependent variable.

In this thesis, new hybrid tree-regression algorithms were designed. In the following chapters, they are applied in various datasets and are compared with linear regression and existing hybrid tree-regression algorithms. Chapter 2 applies HGTLR algorithm in a meta-regression analysis which discuss the effect of trade openness to economic growth. Chapter 3 compares HFLR with existing algorithms in simulated data as well as the crowdfunding data. Chapter 4 applies HFLR in probit regression models about environmental willingness to pay. Chapter 5 concludes the thesis and discusses future research plans.

Chapter 2

Exploring Trade Openness and Economic Growth – An Application of Hybrid Tree-regression Methods in Meta-Analysis

2.1 Introduction

The relationship between trade openness and economic growth is an important topic in development economics. It has been widely discussed in empirical studies. In recent years, with the growing trends of trade protection and trade conflicts around the world, the issue has held an even stronger practical significance than in previous years.

Several studies showed that the relationship between openness and growth of a country depends upon the stock and increment of its income. Based on data from seventy-nine countries over the period 1970-1998, Wang et al. (2004) discussed the relationship between economic growth and three openness-related variables: trade openness, foreign direct investment (FDI) and black market premium (BMP). The study concluded that different country groups benefit from different types of openness: FDI is more beneficial for high-income countries, while trade openness is more beneficial for developing countries.

Discussing both long-term and short-term effects, Dufrenot et al. (2010) analysed annual data for 1980-2006 from seventy-five developing countries. The results of these models show that among developing countries, trade openness has a higher positive effect on economic growth in low-growth countries than in high-growth countries. Ramanayake and Lee (2015) built OLS, fixed-effect and GMM regression models to various openness variables: trade openness, FDI, export diversification, and export growth. The impact of openness to economic growth is positive in most models, while the scale of this impact is larger in developing countries than in developed countries.

Based on the data over the period 1960-2007 from sixty-three countries, Kim et al (2012) discussed the interrelationship between economic growth, trade openness and financial development by building simultaneous equations models. The models have been run in various country groups, which are defined by the income levels, inflation rates and industrial structures of the countries. The study concluded that trade openness is beneficial for economic growth in high-income, low-inflation and non-agricultural countries. However, for countries with opposite characteristics, the

impact of trade openness on economic growth is negative.

In a relatively significant study on 169 countries from around the world, Huchet-Bourdon et al. (2018) analysed data from 1988-2014 to examine trade-growth relationship. They found the effect of openness on growth to be conditional of the quality and variety of export goods. Using GMM to study an endogenous growth model, these authors found that countries that export higher quality goods and newer varieties tend to grow faster. A non-linear pattern was also detected between the export ratio and the quality of the export basket, suggesting that trade openness may impact negatively on economic growth for countries that specialise in the production of low-quality products.

Rather than examining the non-linearity of trade growth relationship with respect to quality and variety of goods in trade, several researchers have analysed the relation in terms of the overall level of trade. Zahonogo (2016) used a dynamic growth model in a study of forty-two sub-Saharan African countries showing that greater trade is beneficial for growth up to a fairly high openness ratio of 134%. Purnama and Yao (2019) used Pedroni's panel cointegration method in the case of countries in the ASEAN region, and indicate that there is a pairwise bidirectional causality among trade, FDI and growth.

To numerically summarise the wide variety of findings in the literature on trade and growth, we analysed the relationship using meta-regression analysis in this study. Because of the differences in the size of the results and in degrees of their significance, a meta study can help to find out systematic variation between the effect size (partial correlation, *pcc*) and various regressors.

Meta-analysis has been used extensively in medical sciences (Itani et al, 2017; Lim et al, 2019; Hemilä et al, 2020) and social sciences (Benos et al, 2014; Cho et al, 2014; Abdullah et al, 2015) studies. It collects results of various individual studies to systematically synthesise results quantitatively. Meta-regression is one of the widely-used approach in modelling meta-analysis data where the observed effect sizes of interested topic is regression on relevant characteristics from individual studies. These characteristics include data source, publication date and key variables which are treated as regressors in the meta-regression models. While many of the regressors are represented by using dummy (binary) variables and its interaction terms in meta-regression models, there are limited discussions on how to accurately choose interaction terms in the meta-regression models.

We built a linear meta-regression model for the effect size and certain independent variables. We also proposed some new methods of using a decision tree in order to account for non-linear relationships between the effect size and the independent variables. These tree-based regression algorithms are developed to combine the strengths of decision trees and linear regression models. Explanatory prediction powers of hybrid tree-regression methods are compared with linear models.

The structure of the remainder of Chapter 2 is as follows. In Section 2.2, the process of rearranging the database is described, and variables in meta-regression models are chosen. In Section 2.3, both linear and tree-based meta-regression models are introduced. In Section 2.4, the results of these models are given and discussed. Section 2.5 comes to the conclusions about the study, and possible further works are also discussed.

2.2 Data

2.2.1 Data Collection

The data for this study was collected based on a raw database which included 295 papers. The titles of these papers contained at least one of the keywords: ”trade”, ”openness” or ”growth”. To achieve higher comparability of the studies as well as a lower rate of missing data, the data was rearranged to select studies that satisfied three conditions:

(1) the dependent variable of the model was a univariate function of GDP, GDP per capita or GDP per unit of labour;

(2) a function of the trade openness index, Trade, was included in the model as an independent variable; and

(3) the t-statistic of $Trade$ and its degrees of freedom was either reported or being able to be calculated from reported information.

The resulting meta-data included 452 models gathered from sixty-five papers published between 1995 and 2016. At the significance level of 0.05, there are 194 positive significant estimates and forty-eight negative significant estimates.

2.2.2 Indicator of Effect Sizes

The indicator of effect sizes is the partial correlation coefficient of trade openness, which is calculated by the formula

$$pcc_i = \frac{t_i}{\sqrt{t_i^2 + df_i}},$$

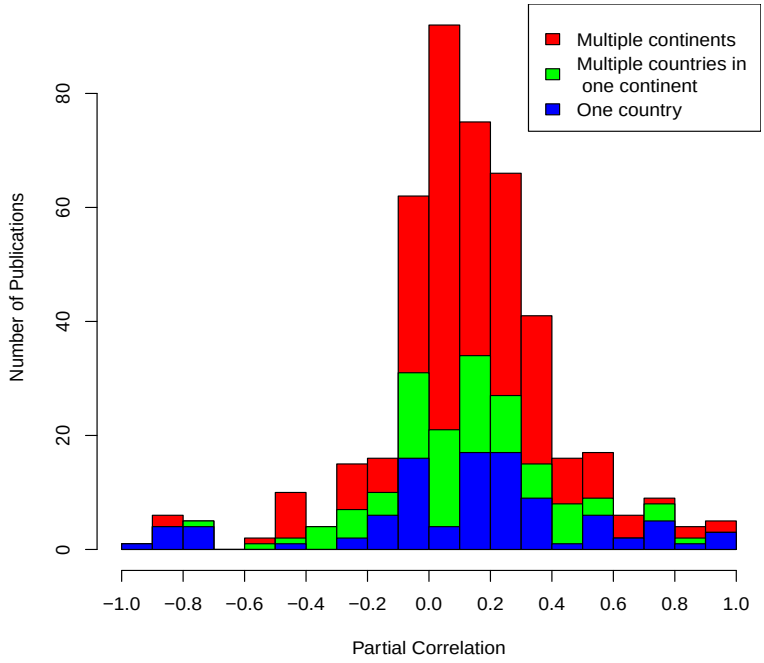
where t is the t-value of the indicator of trade openness, and df is the degrees of freedom of the regression model. The variance of pcc is given by

$$var(pcc_i) = \frac{1 - pcc_i^2}{df_i}.$$

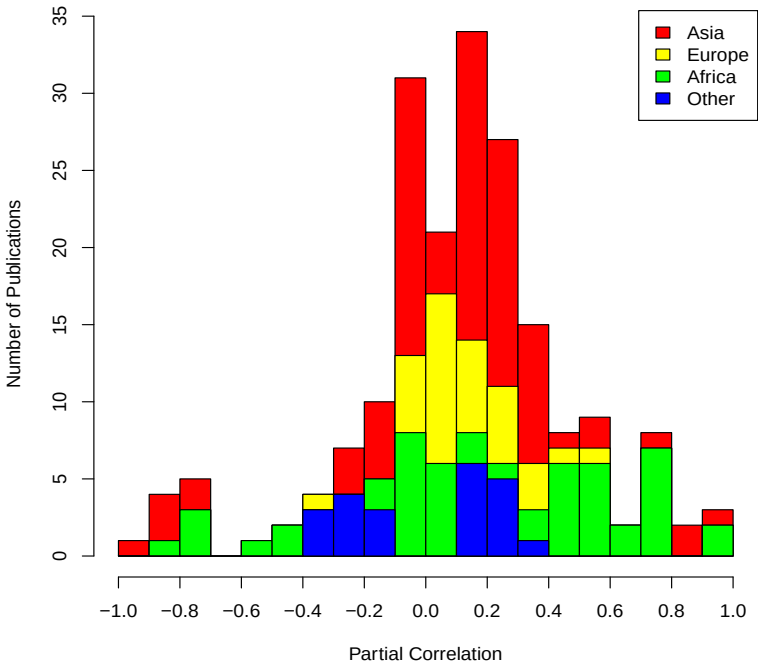
When analysing the effect of trade openness, some econometric studies applied the term for its long-run effect $TRADE$, and others applied the term for its short-run effect $\Delta TRADE$. When both terms are included in an econometric model, we calculate the effect size with the t-values of the long-run effect. If different lags of $TRADE$ are included as independent variables in an econometric model, we then calculate the effect size with the pcc of the term with the least lag.

Figure 2.1 shows the distribution of pcc in the whole meta-data as well as in certain subgroups. The weighted mean of pcc is 0.1572 with standard error 0.0028, where the weight of pcc_i is its reverse standard error $1/SE$.

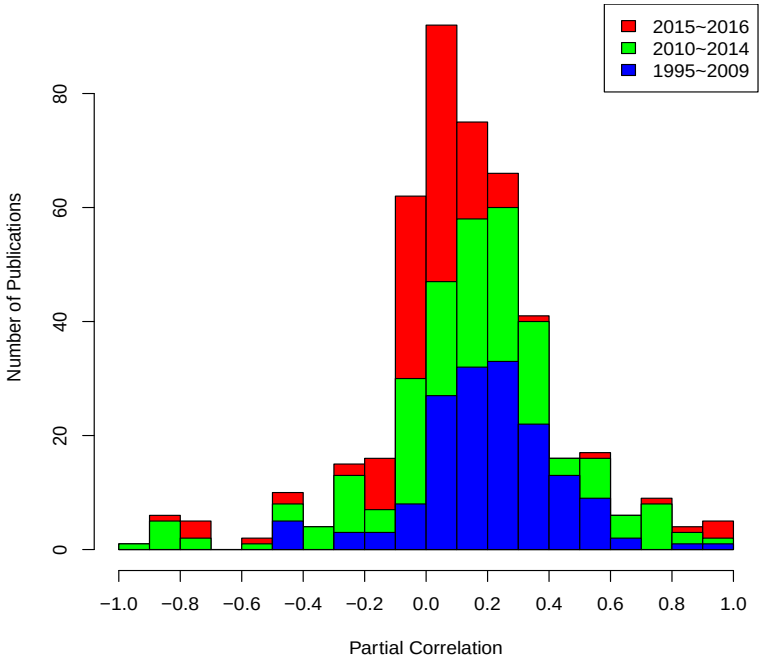
From Figure 2.1(a), it is determined that effect sizes are more likely to have a larger value in econometric models that focus on one continent, especially those that focus on only one country. Specifically, among those studies that focus on one



(a) Inclusion of Continents and Countries



(b) Continents (if only one is included)



(c) Publication Year

Figure 2.1: Distribution of Effect Sizes in Different Subgroups

continent, we noticed from Figure 2.1(b) that studies focusing on African countries are more likely to have larger effect sizes. Applying $1/SE$ as weights, the weighted mean of *pcc* for models focusing on Asia, Europe, and Africa are 0.1002, 0.1494, and 0.4052, respectively.

Apart from countries and continents included in econometric models, we also considered the relationship between publication years and effect sizes. From Figure 2.1(c), we can see that compared to publications before 2015, publications between 2015 and 2016 are more likely to have smaller estimates of the effect of trade openness to economic growth. The weighted mean of *pcc* for models published in 1995-2009, 2010-2014, and 2015-2016 are 0.1900, 0.1699, and 0.1096, respectively.

2.2.3 Independent Variables

Our meta-regression models included thirty independent variables. One of them was the standard deviation $SE = \sqrt{var(pcc)}$. The other twenty-nine independent variables were all categorical variables and are classified into four categories.

First, we had seven variables to show the differences among data: whether the model only applied data from one country ($UCTRY$); whether the model only applied data from one continent ($UCONT$); and whether the focused continent is Asia (AS), Europe (EU), Africa (AF), or others.

Second, we used two variables to distinguish the publication year of the studies: whether it was published in the period 1995-2009, 2010-2014 ($YA10$), or 2015-2016 ($YA15$).

Third, we used two variables to distinguish different definitions of the dependent and independent variables considered in our sample papers: whether the effect size was calculated by short-run effects of trade openness ($SHOEF$); and whether the study used GDP or GDP per capita as the dependent variable ($PCPT$).

It was determined from previous literature that the effect of trade openness on growth is different for various country groups. In this study, we have used eighteen variables to show whether the following variables or groups of variables were controlled: country effects and time effects ($CETE$); initial GDP ($IGDP$); openness variables other than trade openness index, including FDI (FDI), BMP (BMP), and others ($OOPEN$); external economic factors (EXT); investment (INV); inflation (INF); infrastructure ($INFR$); credit or financial depth ($CRED$); economic

policy, including government spending (GOV) and others ($OECOP$); political factors ($POLI$); social factors, including population (POP), life expectancy ($LIFE$), education (EDU), and others ($OSOC$); natural geographic factors ($NGEO$).

Definitions and descriptive statistics of all independent variables are listed in Table 1. Among all econometric models included in the meta-data, 42.92% only applied data from one continent, and 21.90% only applied data from one country. The most focused upon continent is Asia (19.47%), followed by Africa (11.28%). Publications before 2010, between 2010 and 2014, and between 2015 and 2016 provided 35.18%, 37.17% and 27.65% of all considered models. Most considered models analysed long-run effects (89.16%) of trade openness to GDP per capita (89.38%). The most frequently applied in the observations are initial GDP (67.70%) and education (65.04%), followed by investment (56.64%), country effects and time effects (45.58%), and government spending (44.91%).

Table 2.1: Independent variables in the meta-regression model

X	Meaning of $X = 1$	Mean (SD)
$UCTRY$	The model only applied data from one country.	0.2190 (0.4140)
$UCONT$	The model only applied data from one continent.	0.4292 (0.4955)
AS	The model only applied data from Asia.	0.1947 (0.3964)
EU	The model only applied data from Europe.	0.0730 (0.2604)
AF	The model only applied data from Africa.	0.1128 (0.3167)
$YA10$	The paper is published between 2010 and 2014.	0.3717 (0.4838)
$YA15$	The paper is published between 2015 and 2016.	0.2765 (0.4478)

Table 2.1: Independent variables in the meta-regression model (continued)

X	Meaning of $X = 1$	Mean (SD)
$SHOEF$	Short-run effect of trade openness is applied to calculate the effect size.	0.1084 (0.3112)
$PCPT$	The indicator of economic growth in the model is GDP per capita or GDP per labour.	0.8938 (0.3084)
$CETE$	The model is controlled for country effects and time effects.	0.4558 (0.4986)
$IGDP$	The model is controlled for initial GDP.	0.6770 (0.4681)
FDI	The model is controlled for FDI.	0.2854 (0.4521)
BMP	The model is controlled for BMP.	0.1637 (0.3704)
$OOPEN$	The model is controlled for other openness variables including export, regional trade agreements, years of openness, etc.	0.1681 (0.3744)
EXT	The model is controlled for external economic factors.	0.1084 (0.3112)
INV	The model is controlled for investment rate.	0.5664 (0.4961)
INF	The model is controlled for inflation rate.	0.3783 (0.4855)
$INFR$	The model is controlled for variables referring to infrastructure including telephone, railway, etc.	0.0774 (0.2676)

Table 2.1: Independent variables in the meta-regression model (continued)

X	Meaning of $X = 1$	Mean (SD)
$CRED$	The model is controlled for variables referring to credit including bank credit, private credit, financial depth, etc.	0.1836 (0.3876)
GOV	The model is controlled for government spending.	0.4491 (0.4980)
$OECOP$	The model is controlled for other variables referring to economic policy including monetary policy, capital control, state owned firms, etc.	0.2080 (0.4063)
$POLI$	The model is controlled for political factors including democracy, rule of law, corruption, etc.	0.2920 (0.4552)
POP	The model is controlled for variables referring to population.	0.3562 (0.4794)
$LIFE$	The model is controlled for life expectancy.	0.1173 (0.3221)
EDU	The model is controlled for variables referring to education.	0.6504 (0.4774)
$OSOC$	The model is controlled for other social factors including civil liberty, ethnic diversity, religions, etc.	0.1217 (0.3273)
$NGEO$	The model is controlled for natural geographic factors including latitude, distance from sea, etc.	0.0708 (0.2568)

2.3 Modelling considerations

2.3.1 Linear Meta-regression Models

Using the independent variables listed above, our aim was to explain the heterogeneity in existing estimates of t-statistics by constructing a regression model. While constructing meta-regression models, one usually applies a random-effects (RE) model which is written as:

$$pcc_i = pcc_0 + \beta_0 SE + \sum_{k=1}^{K} \beta_k x_{ik} + \nu_i + \epsilon_i, \tag{2.1}$$

where pcc_0 is the intercept term, SE is the standard error of pcc, $x_1, \cdots, x_J$ are independent variables with regression coefficients $\beta_1, \cdots, \beta_K$, ν is the vector of study-level random effects, and ϵ is the error term.

However, by applying multi groups of simulated data, Stanley et al. (2017) concluded that prediction results of WLS meta-regression have higher prediction efficiency compared to random-effects meta-regression, especially with the presence of publication selection, small sample biases, etc. The WLS meta-regression model is written as:

$$pcc_i = pcc_0 + \beta_0 SE_i + \sum_{k=1}^{K} \beta_k z_{ik} + \epsilon'_i, \tag{2.2}$$

where pcc_0 is the intercept term, SE is the standard error of pcc, $x_1, \cdots, x_J$ are independent variables with regression coefficients $\beta_1, \cdots, \beta_K$, and $\epsilon' \sim N(0, V)$ is

the error term where

$$V = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & & \cdots & \sigma_I^2 \end{bmatrix}.$$

The WLS estimates of the coefficients are given as

$$\hat{\beta} = (X^T V^{-1} X) X^T V^{-1} y.$$

In this study, we applied both RE and WLS approaches to estimate the linear meta-regression model. Backward stepwise regression is applied in both approaches for variable selection, while in WLS meta-regression, the weights are estimated twice: once when estimating the model with all independent variables, and once after the stepwise regression procedure.

2.3.2 Hybrid Tree-linear Meta-regression Models

The linear model has the limitation that it does not detect interaction effects between variables. Meanwhile, decision trees handle interaction effects between categorical variables very well. Based on the idea of the hybrid tree-logit model given by Stainberg & Cardell (1998), a hybrid tree-linear regression (HTLR) model can be formed by Algorithm 4.

The decision tree built for the HTLR model is shown in Figure 2.2. The categor-

Algorithm 4 Hybrid Tree-linear Regression (HTLR) Model

1) Build a regression tree model between the dependent variable and all categorical independent variables.
2) Create new categorical variables for splitting points in the regression tree.
3) Run a linear regression model between the dependent variable and all independent variables as well as categorical variables created in the previous step.
4) Apply stepwise regression to select variables included in the regression model.

ical variables created for the splitting points are $YA15 \times PCPT$, $YA15 \times PCPT \times IGDP$, $(1 - YA15) \times AF$, $(1 - YA15) \times (1 - AF) \times YA10$, etc. According to the definition of variables, some interaction terms can be simplified. As an example, $(1 - YA15) \times (1 - AF) \times YA10$ is equivalent to $YA10 \times (1 - AF)$.

From Figure 2.2, we noted that both $YA15$ and $YA10$ are key variables in the decision tree built for the HTLR model. All interaction terms derived by the decision tree can be written as either $YA15 \times f(X)$, $(1 - YA15) \times f(X)$, $YA10 \times f(X)$, or $(1-YA10-YA15)\times f(X)$. Noting that $YA15 = 1$, $YA10 = 1$, and $YA10+YA15 = 0$ represent econometric models published in three different time periods, we built three different decision trees for these subgroups, which are shown in Figure 2.3.

The categorical variables derived from these decision trees were in the form of products of a categorical variable representing the subsets ($YA15$, $YA10$, and $1 - YA10 - YA15$) and a categorical variable representing a splitting point in the tree ($PCPT$, $PCPT \times IGDP$, AF, $AF \times INF$, etc). It was easy to identify that some variables derived from these three decision trees were equivalent to those derived from the decision tree in Figure 2.2, such as $YA15 \times PCPT$, $YA15 \times PCPT \times IGDP$, etc. Applying all variables derived from decision trees in Figure 2.2, a hybrid groupwise tree-linear regression (HGTLR) model is given by Algorithm 5.

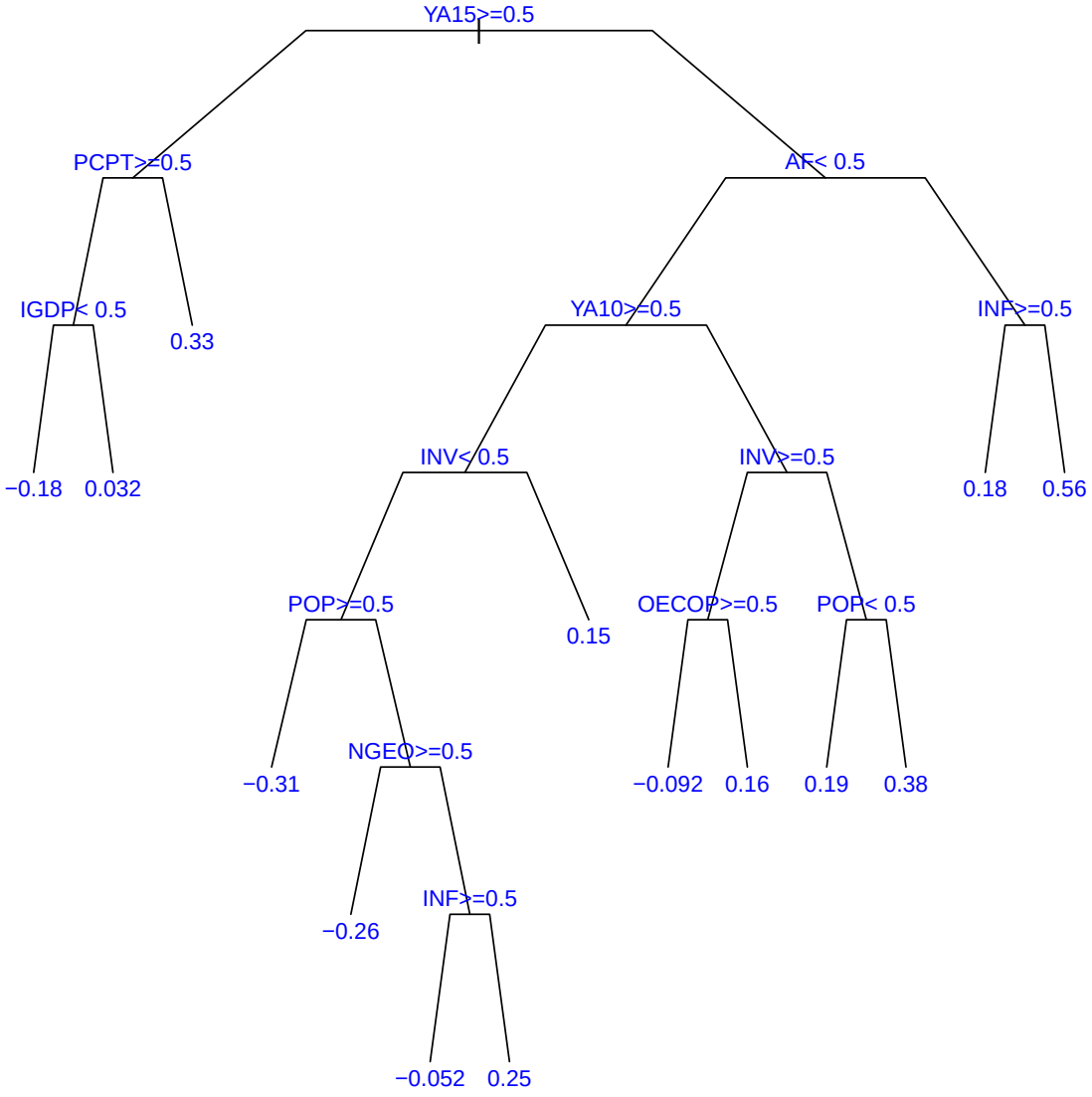
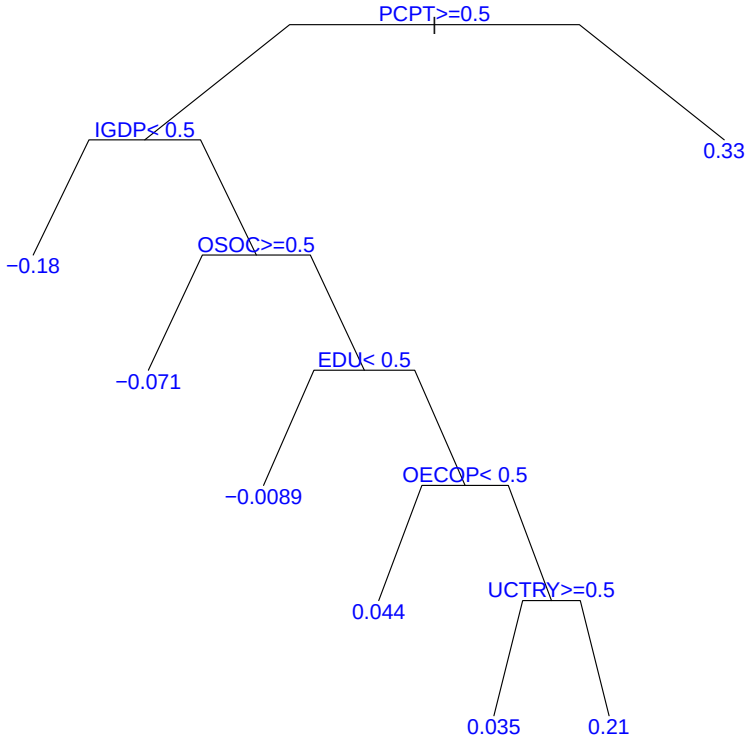
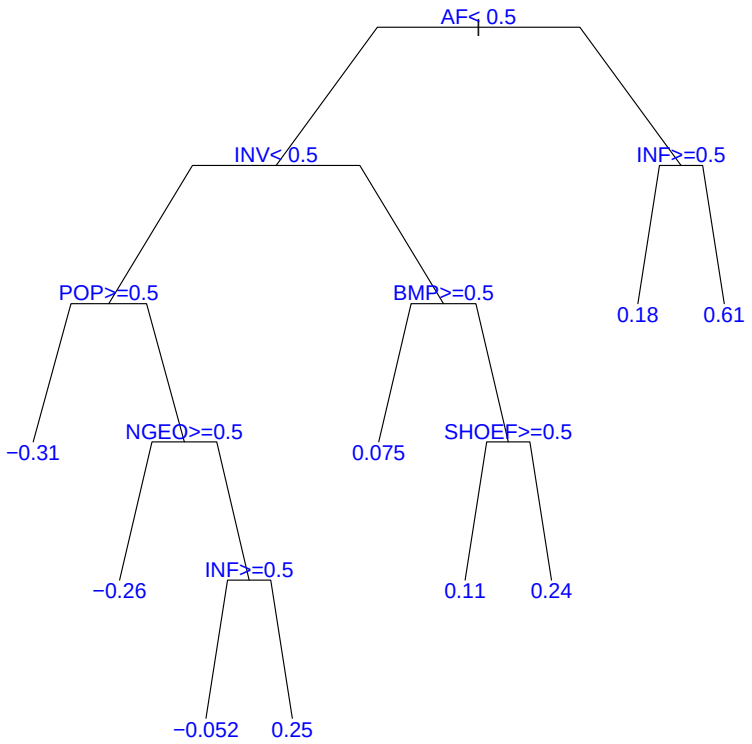


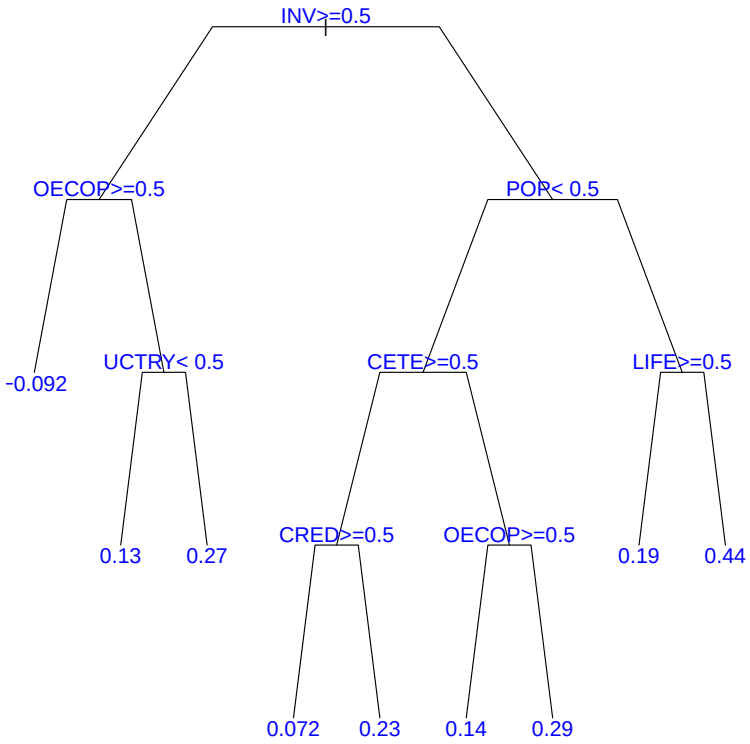
Figure 2.2: Decision Tree Built in HTLR Model



(a) 2015 ∼ 2016



(b) 2010 ∼ 2014



(c) $1995 \sim 2009$

Figure 2.3: Decision Tree Built for Three Subsets

Algorithm 5 Hybrid Groupwise Tree-linear Regression (HGTLR) Model

1) Split the data into a few meaningful subsets.
2) For each subset, build a regression tree model between the dependent variable and all categorical independent variables.
3) Create new categorical variables for splitting points in the regression trees.
4) Run a linear regression model between the dependent variable and all independent variables as well as categorical variables created in the previous step.

5) Apply stepwise regression to select variables included in the regression model.

Both HTLR and HGTLR models can be written as:

$$pcc_i = pcc_0 + \beta_0 SE + f(X_i) + \epsilon_i', \tag{2.3}$$

where $X = \{x_1, \cdots, x_J\}$ and $f(X)$ is a non-linear function. The model is estimated

by both RE and WLS approaches, where in random-effects meta-regression, $\epsilon' = \nu + \epsilon$. Backward stepwise regression is applied for variable selection.

Tree-based regression algorithms were developed because both decision trees and linear regression models have their strengths and weaknesses. Decision trees have the strength of detecting nonlinear relationships without prior assumptions of the underlying distribution of the data. However, all terms included in a decision tree are interaction terms between the variable on the root node and other variables, which may result in the missing of some key variables.

In comparison, linear regression models do not detect interaction effects between variables. If interaction terms are added manually, the estimation of the model could become highly computationally inefficient, or in extreme cases, infeasible. As an example, with twenty-three categorical variables discussed in the meta-regression model, 253 interaction terms of two variables and 1771 interaction terms of three variable can be created. If all these interaction terms were added in a linear regression model, there would be many more parameters than observations, which makes the construction of the model computationally infeasible.

2.4 Results and Discussion

2.4.1 Results of Meta-regression Models

Table 2.2 summarises the results of our RE meta-regression models, and Table 2.3 summarises the results of our WLS meta-regression models.

Table 2.2: Results of RE meta-regression models

Method	RE-LR	RE-HTLR	RE-HGTLR
Intercept	0.3682***	0.5782***	0.5082***
	(0.0779)	(0.0667)	(0.0871)
SE	-4.9335***	-4.8584***	-4.8112***
	(0.1843)	(0.1813)	(0.1830)
UCTRY	0.3091***	0.2634***	0.2618***
	(0.0481)	(0.0531)	(0.0535)
UCONT	0.1587***	0.1887	0.1951***
	(0.0467)	(0.0362)	(0.0363)
AS	0.0996 .		
	(0.0577)		
AF	-0.2327***		-0.2453***
	(0.0540)		(0.0476)
YA10		0.1442 .	0.1274
		(0.0798)	(0.0777)
SHOEF	-0.0876***	-0.0910***	
	(0.0151)	(0.0150)	
PCPT	0.0869 .		0.1041 .
	(0.0516)		(0.0571)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.2: Results of the models (continued)

Method	RE-LR	RE-HTLR	RE-HGTLR
IGDP	0.0475 .		0.0736
	(0.0288)		(0.0454)
FDI	-0.0771***	-0.0742***	-0.0705***
	(0.0188)	(0.0189)	(0.0190)
BMP	0.4118***		
	(0.1064)		
OOPEN	0.0446**	0.0407*	0.0400*
	(0.0160)	(0.0159)	(0.0159)
EXT	-0.1101***	-0.1054	-0.1039**
	(0.0332)	(0.0329)	(0.0328)
INV			0.0596*
			(0.0281)
INFR	-0.1736***	-0.2385***	-0.2107***
	(0.0383)	(0.0438)	(0.0440)
GOV	-0.0452**		
	(0.0143)		
OECOP	-0.0445*	-0.0366	-0.0579*
	(0.0223)	(0.0224)	(0.0256)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.2: Results of the models (continued)

Method	RE-LR	RE-HTLR	RE-HGTLR
$POLI$	0.0371	0.0487*	0.0461 .
	(0.0238)	(0.0241)	(0.0251)
POP	0.1200***		
	(0.0187)		
$LIFE$			-0.1104
			(0.0762)
EDU	-0.0589***		-0.0849**
	(0.0173)		(0.0264)
$OSOC$	0.0331 .		
	(0.0175)		
$YA15 \times PCPT$		-0.2759*	
		(0.1315)	
$YA15 \times PCPT \times IGDP$		0.1118***	-0.7229**
		(0.0316)	(0.2310)
$YA15 \times PCPT \times IGDP \times OSOC$			0.1061***
			(0.0271)
$YA15 \times PCPT \times IGDP \times$			0.7573**
$(1 - OSOC) \times EDU$			(0.2304)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.2: Results of the models (continued)

Method	RE-LR	RE-HTLR	RE-HGTLR
$YA15 \times PCPT \times IGDP \times$			0.0764
$(1 - OSOC) \times EDU \times OECOP$			(0.0520)
$YA15 \times PCPT \times IGDP \times$			-0.6935**
$(1 - OSOC) \times EDU \times$			(0.2338)
$OECOP \times UCTRY$			
$(1 - YA15) \times AF$		-0.2382***	
		(0.0484)	
$(1 - YA15) \times AF \times INF$		-0.5944***	
		(0.1524)	
$YA10 \times AF \times INF$			-0.6078***
			(0.1462)
$YA10 \times (1 - AF) \times INV$		-0.1876	-0.2366*
		(0.1165)	(0.1112)
$YA10 \times (1 - AF) \times INV \times$			-0.0963*
$(1 - BMP) \times SHOEF$			(0.0156)
$YA10 \times (1 - AF) \times (1 - INV) \times$		-1.2193***	-1.2327***
POP		(0.3467)	(0.3264)
$YA10 \times (1 - AF) \times (1 - INV) \times$		-0.8097*	-0.8350**
$(1 - POP) \times NGEO$		(0.3466)	(0.3229)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.2: Results of the models (continued)

Method	RE-LR	RE-HTLR	RE-HGTLR
$YA10 \times (1-AF) \times (1-INV) \times (1-POP) \times (1-NGEO) \times INF$		-0.4094*** (0.1182)	-0.4011*** (0.1112)
$(1-YA10-YA15) \times (1-AF) \times INV \times OECOP$		-0.5471*** (0.1156)	
$(1-YA10-YA15) \times (1-AF) \times (1-INV) \times POP$		0.3157 (0.2038)	
$(1-YA10-YA15) \times INV \times OECOP$			-0.5051*** (0.1275)
$(1-YA10-YA15) \times INV \times (1-OECOP) \times UCTRY$			-0.4942 (0.3197)
$(1-YA10-YA15) \times (1-INV) \times POP \times LIFE$			0.1560 (0.1001)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.3: Results of WLS meta-regression models

Method	WLS-LR	WLS-HTLR	WLS-HGTLR
Intercept	0.1790*** (0.0379)	0.1965*** (0.0277)	0.2410*** (0.0361)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.3: Results of the models (continued)

Method	WLS-LR	WLS-HTLR	WLS-HGTLR
SE			-0.3471
			(0.2198)
UCTRY			0.1029*
			(0.0412)
UCONT			0.1862***
			(0.0416)
AS			-0.2107***
			(0.0479)
EU			-0.3620***
			(0.0753)
AF	0.2386***	-0.2128**	
	(0.0487)	(0.0702)	
YA10	-0.0973**	0.1121*	
	(0.0348)	(0.0523)	
YA15	-0.2462***	0.2522***	
	(0.0382)	(0.0738)	
CETE		-0.1178***	
		(0.0298)	

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.3: Results of the models (continued)

Method	WLS-LR	WLS-HTLR	WLS-HGTLR
IGDP	0.1175**	0.0622*	0.1245***
	(0.0385)	(0.0308)	(0.0308)
INV	0.0471		0.1470**
	(0.0328)		(0.0452)
INF	-0.0615 .		
	(0.0359)		
INFR	0.0873		
	(0.0543)		
CRED	-0.1007*		
	(0.0433)		
GOV	0.0760*		
	(0.0296)		
POP	0.0553 .		0.1146***
	(0.0295)		(0.0333)
LIFE	-0.0739	-0.1408***	-0.1564***
	(0.0456)	(0.0389)	(0.0451)
EDU	-0.0965**		
	(0.0344)		

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.3: Results of the models (continued)

Method	WLS-LR	WLS-HTLR	WLS-HGTLR
$OSOC$	-0.1646***	-0.1189**	-0.1060*
	(0.0419)	(0.0387)	(0.0417)
$NGEO$	-0.1517**		
	(0.0560)		
$YA15 \times PCPT$		-0.3353***	-0.5321***
		(0.0750)	(0.0550)
$YA15 \times PCPT \times IGDP \times$			0.1099 .
$(1 - OSOC) \times EDU$			(0.0617)
$YA15 \times PCPT \times IGDP \times$			0.1957*
$(1 - OSOC) \times EDU \times OECOP$			(0.0812)
$YA15 \times PCPT \times IGDP \times$			-0.3072**
$(1 - OSOC) \times EDU \times$			(0.1152)
$OECOP \times UCTRY$			
$(1 - YA15) \times AF$		0.5110***	
		(0.0876)	
$(1 - YA15) \times AF \times INF$		-0.3925***	
		(0.0887)	
$YA10 \times AF \times INF$			-0.3995***
			(0.0953)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.3: Results of the models (continued)

Method	WLS-LR	WLS-HTLR	WLS-HGTLR
$YA10 \times (1 - AF) \times INV$		-0.1233*	-0.2332***
		(0.0550)	(0.0528)
$YA10 \times (1 - AF) \times INV \times BMP$			-0.2802***
			(0.0586)
$YA10 \times (1 - AF) \times INV \times$			-0.1241 .
$(1 - BMP) \times SHOEF$			(0.0722)
$YA10 \times (1 - AF) \times$		-0.6167***	-0.6928***
$(1 - INV) \times POP$		(0.1177)	(0.1094)
$YA10 \times (1 - AF) \times (1 - INV) \times$		-0.4532***	-0.6624***
$(1 - POP) \times NGEO$		(0.1204)	(0.1053)
$YA10 \times (1 - AF) \times (1 - INV) \times$		-0.2972***	0.3504***
$(1 - POP) \times (1 - NGEO) \times INF$		(0.0752)	(0.0590)
$(1 - YA10 - YA15) \times$		-0.2422**	
$(1 - AF) \times INV \times OECOP$		(0.0801)	
$(1 - YA10 - YA15) \times$		0.2088***	
$(1 - AF) \times (1 - INV) \times POP$		(0.0458)	
$(1 - YA10 - YA15) \times INV$			-0.3170***
			(0.0608)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 2.3: Results of the models (continued)

Method	WLS-LR	WLS-HTLR	WLS-HGTLR
$(1-YA10-YA15)\times INV\times OECOP$			-0.1938*
			(0.0859)
$(1-YA10-YA15)\times INV\times (1-OECOP)\times UCTRY$			0.2297*
			(0.1137)
$(1-YA10-YA15)\times(1-INV)\times (1-POP)\times CETE$			-0.0881
			(0.0546)
$(1-YA10-YA15)\times(1-INV)\times (1-POP)\times CETE\times CRED$			-0.1891**
			(0.0715)
$(1-YA10-YA15)\times(1-INV)\times (1-POP)\times(1-CETE)\times OECOP$			-0.2542**
			(0.0805)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

2.4.2 Comparison between Meta-regression Models

Table 2.4 shows summary statistics and values of certain model selection criteria of meta-regression models applied in the study. The prediction accuracy is calculated based on three categories: positive significant, negative significant, and insignificant at the significance level of 0.05.

Table 2.4: Comparison between meta-regression models

Method	RE-LR	RE-HTLR	RE-HGTLR	WLS-LR	WLS-HTLR	WLS-HGTLR
Amount of trees	0	1	3	0	1	3
Amount of parameters	21	23	32	15	17	28
AIC	2301.8	2236.8	2221.6	113.7	-41.5	-48.1*
BIC	2388.2	2331.4	2353.2	179.6	32.6*	71.2
MSE (cross validation)	0.0882	0.0659	0.0724	0.0759	0.0611	0.0602*
Prediction accuracy	51.77%	65.49%	59.51%	53.54%	67.26%*	66.59%

* represents the selected model from each model selection criteria.

According to the values of both AIC and BIC, all WLS meta-regression models have much higher explanation powers than any RE meta-regression models, and tree-based meta-regression estimates are better than their linear counterparts. The estimates with the highest explanation power is either the WLS-HGTLR estimate seleted by AIC, or the WLS-HTLR estimate selected by BIC. Meanwhile, the WLS-HGTLR also has the smallest MSE in 10-fold cross validation, followed by the WLS-HTLR model, which means that there is no sign of overfitting in both methods.

The prediction powers of all meta-regression models are similar when only the sign of the effect sizes are predicted. However, when both the sign and the significance of the effect sizes are considered, the prediction powers of WLS-HTLR and WLS-HGTLR estimates are dominant over other meta-regression models. As WLS-HTLR and WLS-HGTLR estimates are better than other estimates in both explanation and prediction powers, we are only applying their results in the following

discussions.

2.4.3 Economic Discussions

According to the WLS-HTLR estimate, compared to studies that considered data from multi continents, studies that focused on one continent other than Africa do not have significantly different results. However, if an econometric model focused on Africa, it is estimated that trade-growth partial correlation would either increase by about 0.30, if it is published before 2015, or decrease by about 0.21, if it is published after 2015.

On the other hand, according to the WLS-HGTLR estimate, compared to studies that considered data from multi continents, studies that focused on one continent had either higher or lower estimates of trade-growth partial correlation depending on the focused continent. If the focused continent is Asia or Europe then the estimated partial correlation would be expected to decrease by 0.02 or 0.18. However, the estimated partial correlation would be expected to increase by 0.19 if the focused continent is neither Asia nor Europe.

According to the WLS-HGTLR estimate, we also concluded that econometric models that focused on one country have higher estimates of trade-growth partial correlation than those that considered multi countries on a continent. Generally speaking, these results are consistent to the patterns discovered from Figure 2.1(a) and 2.1(b).

Based on both WLS-HTLR and WLS-HGTLR estimates, the relationship be-

tween estimated trade-growth partial correlation of an econometric model and its publication time depend on various characters of the model, including the data source, the indicator of economic growth, and the control variables applied.

Table 2.5 presents the fitted partial correlation coefficients for openness-growth econometric models in the time periods 1995-2009, 2010-2014, and 2015-2016. The fitted values are calculated based on the prediction when SE is set to be its median, the indicator of economic growth is GDP per capita, and three control variable groups: $CVG1 = \{IGDP, INV, EDU\}$ only include control variables applied in more than 50% of observations in the meta-data, $CVG2 = CVG1 \cup \{CETE, GOV, POP\}$ include control variables applied in more than 40% of observations, and $CVG3 = CVG2 \cup \{INF, POLI\}$ include control variables applied in more than 30% of observations.

According to the estimates in Table 2.5, under all circumstances, estimates of trade-growth partial correlation in publications in the period 2015-2016 were always lower than those in publications before 2015. Meanwhile, the difference between estimates in publications in the periods 1995-2009 and 2010-2014 depend upon the estimation method and the data source.

According to both WLS-HTLR and WLS-HGTLR estimates of the meta-regression model, the difference between long-run and short-run trade openness did not have a significant influence to the estimate of their effects on economic growth. According to both estimates, however, in publications in the period 2015-2016, econometric models that applied GDP per capita as the indicator of economic growth had much less positive estimates of the trade-growth partial correlation, compared to those

Table 2.5: Predictions made by meta-regression models

Method	WLS-HTLR	WLS-HTLR	WLS-HTLR	WLS-HGTLR	WLS-HGTLR	WLS-HGTLR
Publication time	1995-2009	2010-2014	2015-2016	1995-2009	2010-2014	2015-2016
Multi continent, $CVG1$	0.2588	0.2476	0.1757	0.1674	0.2512	0.0621
Multi continent, $CVG2$	0.1410	0.1298	0.0579	0.2820	0.3658	0.1767
Multi continent, $CVG3$	0.1410	0.1298	0.0579	0.2820	0.3658	0.1767
Asia, $CVG1$	0.2588	0.2476	0.1757	0.1429	0.2267	0.0376
Asia, $CVG2$	0.1410	0.1298	0.0579	0.2575	0.3413	0.1522
Asia, $CVG3$	0.1410	0.1298	0.0579	0.2575	0.3413	0.1522
Europe, $CVG1$	0.2588	0.2476	0.1757	-0.0085	0.0753	-0.1137
Europe, $CVG2$	0.1410	0.1298	0.0579	0.1062	0.1900	0.0009
Europe, $CVG3$	0.1410	0.1298	0.0579	0.1062	0.1900	0.0009
Africa, $CVG1$	0.5570	0.6691	-0.0371	0.3535	0.6705	0.2482
Africa, $CVG2$	0.4392	0.5513	-0.1549	0.4681	0.7851	0.3629
Africa, $CVG3$	0.0466	0.1588	-0.1549	0.4681	0.3857	0.3629

that applied GDP.

Both WLS-HTLR and WLS-HGTLR estimates of the meta-regression model showed that the use of certain control variables have significant influences on the effect size of openness-growth econometric models. Variables that are included in both estimates as linear terms include $IGDP$ with positive signs, and $LIFE$ and $OSOC$ with negative signs. In comparison, the variables $CETE$, INV, INF, $OECOP$, POP, EDU, and $NGEO$ are included in both estimates in either linear or inter-

action terms. These results showed that estimates of the trade-growth relationship varied across countries and time periods, and its estimates depend upon control variables included in econometric models such as initial GDP; investment; inflation; economic policy; social factors such as population, life expectancy, education; and natural geographic factors. Notably, some of those factors, such as life expectancy and natural geographic factors, are rarely applied as control variables in econometric studies included in the meta-data.

According to the estimates in Table 2.5, trade openness is estimated to be beneficial for economic growth in most cases, even in publications since 2015. Notably, according to the WLS-HTLR estimates, the trade-growth partial correlation in Asia-focused econometric models published since 2015 is estimated to be negative, regardless of the control variables applied in econometric models.

2.5 Conclusion

In this study, we investigated the relationship between trade openness and economic growth by building linear and tree-based meta-regression models. We found strong evidence that trade openness has positive effects on economic growth.

Specifically, models published before 2015 which focus on specific continents, especially Asia and Africa, tend to have more significantly positive effects. We also found that control variables such as initial GDP, investment, inflation, economic policy, social factors, and natural geographic factors were influential to the estimates of the trade-growth partial correlation in econometric models.

More importantly, when we considered the relationship between regressors in the meta-regression model and the effect sizes, the effects of different regressors are often dependent to the values of each other. We found clear evidence that tree-based meta-regression models are superior compared to linear models in terms of both explanatory power and predictive power. This attribute is evident in tree-based meta-regression models which stand out in all terms of model comparison measures including AIC, BIC, cross-validated MSE, and average prediction accuracy. Additionally, we also found evidence that WLS meta-regression models are superior to RE meta-regression models.

Chapter 3

What are the Key Factors that Affect the Fundraising Performance of Crowdfunding? Evidence from Hybrid Tree-regression Models

3.1 Introduction

Social entrepreneurial ventures alleviate social problems through market-based instruments. They bring positive externalities (Dees et al., 2004), increase social trust and social capital, and reduce the transaction costs and resource constraints of start-ups (Roundy et al., 2017). They have, to some extent, addressed market failures (Alter, 2007), alleviated poverty (Mail et al, 2009), reduced unemployment

(Catford, 1998), and pursued development opportunities for women and underrepresented groups (Nicholls, 2008).

In recent years, crowdfunding has garnered increasing attention as an emerging financing channel. The number of crowdfunding platforms worldwide is growing rapidly in a linear trend and their influence is only increasing. On large reward-based crowdfunding platforms such as Kickstarter and Indiegogo, dedicated sections for social or philanthropic entrepreneurship have emerged. Improving the success rate of crowdfunding financing is therefore very important for social entrepreneurial ventures.

Many factors affect crowdfunding project performance, including individual investor characteristics, crowdfunding platforms, crowdfunding projects and reward factors, and visualisation tools such as videos and charts. This paper aims to bring more attention to social entrepreneurship, defined as profitable marketing activities targeting the relief of socially disadvantaged groups (Leadbeater, 1997), and improve the performance of social entrepreneurship enterprises in raising funds through crowdfunding platforms, thereby creating greater social value. We collected data on social entrepreneurship porjects on Kickstarter and Indiegogo and explored the impact of various factors on crowdfunding performance.

We built a linear regression model for fundraising performance, measured by the proportion of financing, and certain independent variables and proposed new methods of using decision trees to account for non-linear relationships between fundraising performance and the independent variables; these tree-based regression algorithms combine the strengths of decision trees and linear regression models. By comparing

tree-based models with linear models, we found that tree-based models have higher explanatory and predictive power.

The structure of the remainder of Chapter 3 is as follows. Section 3.2 provides the background behind developing tree-based regression models; in Section 3.3, we describe the data collection process and choose variables in regression models; in Section 3.4, we introduce both linear and tree-based regression models applied in the study; Section 3.5 gives and discusses the results of these models; in Section 3.6 we draw our conclusions from the study, and discuss possible further research.

3.2 Background

In recent years, crowdfunding has quickly gained attentions from researchers. A systemic review by Bockel et al. (2021) pointed out that the number of publications related to crowdfunding increased from less than five per year at the beginning of 2010s to thirty-two in 2018, and quantitative research has gained dominance over time. In these studies, it was proven that crowdfunding has been helpful in various business sectors and social aspects, such as the arts and creative sector (Chiesa et al., 2021), and academic achievements of school children (Zhou et al, 2021).

Moritz et al. (2016) provided a systemic literature review on 127 articles which are categorised according to their research priorities among three crowdfunding factors: capital seekers, capital providers, and intermediaries. Among variables related to capital seekers, determining factors for crowdfunding success include non-profit oriented background, funding amount and duration, social network, product-related

videos, and geographical proximity to capital providers. On the other hand, capital providers have heterogeneous motivations including social reputation and intrinsic motives, and they are impacted by peer behaviour. However, it is also worth noting that as intermediaries, crowdfunding platforms offer benefits for both capital seekers and providers by reducing information asymmetries and building trust.

Another systemic review by Kaartemo (2017) analysed fifty-one studies published between 2011 and 2016. It was noted that thirty-seven studies only applied data from one crowdfunding platform, among which the most popular data source is Kickstarter (applied by nineteen studies), followed by Indiegogo (applied by five studies). In the systemic review, the factors that are considered by previous studies to affect crowdfunding performances are listed in four categories: (1) campaign-related factors, such as the fundraising target, various types of rewards, and application of visualisation tools; (2) crowdfunder-related factors, such as types of crowdfunders; (3) crowdfunding platform-related factors, such as platform design; and (4) fundseeker-related factors, such as various types of social capitals.

Hossain et al. (2017) discussed the definitions of common terms in crowdfunding-related studies and concluded that there are four important elements in crowdfunding: (1) the role of internet and social media, including personal social network, the judgment of other backers, individual groups representing special interests, and platform recommendations based on users' behaviors and preferences; (2) visual communication, including videos, technical drawings, photographs, and interactive chat forums; (3) motivation of founders and sponsors; and (4) trust and transparency, which is a key component in various stages of a project and a real-time feedback process.

Sauermann et al. (2019) analysed data from over 700 research-related crowdfunding campaigns before 2015. Variables related to creator characteristics, project characteristics, and campaign characteristics were applied to explain both the success and the size of funding. The results of linear and logistic regression both show that various factors, especially the position and gender of project creators, are significantly influential to the results of crowdfunding.

Huang et al. (2021) applied a dataset with sixty-two projects from 2013 to 2015 to analyse the influences on crowdfunding success of both project quality and experiences of entrepreneurs. Based on the results of qualitative comparative analysis, the study concluded that both groups of factors are influential on crowdfunding success.

Based on 167 responses of an online questionnaire, Moysidou et al. (2020) analysed the trust-building procedure of crowdfunding and concluded that crowdfunding platforms play the most important role in the procedure. Schraven et al. (2020) collected data of predictions made by a group of participants to success of crowdfunding campaigns and analysed their accuracy. It is concluded that negative information are more quickly recognised by participants.

Applying a panel data of 450 crowdfunding projects on a Chinese platform, Chen et al. (2020) built a regression model which showed that both static variables (social capital, descriptive information, and funding goal) and dynamic variables (project popularity) have significant relationships with the proportion of financing. Following this categorisation, Popescul et al. (2020) provided a systemic review which added more variables in both categories. Static variables, including reward-related variables, experiences of entrepreneurs, fundraiser related variables (gender, location,

and team size), risks, qualities of project plans, and platforms, are also considered to be influential to fundraising performances. In comparison, dynamic variables, such as previous crowdfunding performance, frequencies and qualities of updates, and other decisions made by entrepreneurs during the crowdfunding process, are also decisive to crowdfunding performances.

In previous studies, it was determined that the performances of crowdfunding projects are influenced by many factors. However, most studies are based on the assumption that the influences of these factors to crowdfunding performances are independent. By applying hybrid models of decision trees and linear regression, we discover joint effects between multi regressors to crowdfunding performances.

3.3 Data

3.3.1 Data Collection

The data of this paper was collected from two leading international crowdfunding websites: Kickstarter and Indiegogo. We began by using the keywords "social enterprise" and "social entrepreneurship" in our search for projects that were completed from 2016 to 2018. Then, we cleaned up the original data by removing projects with extreme values of variables such as funding ratio, funding target, and number of followers. In the end, 236 social entrepreneurship projects were obtained, of which fifty-two were from Kickstarter and 184 were from Indiegogo.

3.3.2 Indicator of Fundraising Performances

In previous studies, the fundraising performances of crowdfunding projects was measured with three variables: the success of financing, the size of financing, and the proportion of financing. The success of financing is a categorical variable that ignores the differences among unsuccessful projects, of which the proportion of financing could vary from 0 to 90.6%, based on the data. When considering the size of financing, we found that it did not directly show the success of financing. The proportion of financing, however, defined as the ratio between the size of fundraising and the fundraising target, is a continuous variable that measures different levels of success. Thus, we selected the proportion of financing (LPF) as a metric for fundraising performances.

Among 236 observations in this study, there were fifty-two successful and 184 unsuccessful social entrepreneurship projects. Among the unsuccessful projects, the proportion of financing was between 10% and 100% for 73 observations, between 1% and 10% for 56 observations, and less than 1% for 55 observations.

3.3.3 Independent Variables

Our regression models included thirty independent variables; two of which were continuous variables representing the fundraising target ($LTARG$) and the number of pictures displayed on the web page of the project ($LPICS$). We took natural logs of all continuous variables in the model.

We used one variable to represent the website where projects are published (*KICKS*), and three variables to show the topics of social entrepreneurship projects: (i) community activities (*COMACT*); (ii) design and technology (*DESTECH*); and (iii) food and craft (*FOODCR*). We used one variable to show if the project involved products (*PROD*).

Another eight variables held information about the founders of the crowdfunding project: whether they were individual persons (*PERSON*); their gender (*MALE*); whether a profile picture was displayed (*PROFL*) and if it was a picture of the founder (*SELF*); whether the founders have supported other projects (*SPT*); if they have prior experience in starting a business (*BUSIN*); and whether a job description (*JOBDES*), or a degree (*DEG*) is displayed on the crowdfunding website.

We used a further ten variables to describe the types of rewards for sponsors of the crowdfunding project: products produced by or related to the projects (*RWPROD*); experiences and techniques (*RWEXP*); VIP memberships (*RWVIP*); honor certificates (*RWHONOR*), letters of thanks (*RWLETTER*), or photos (*RWPHOTO*); chances to visit the project site (*RWVISIT*), join annual meetings (*RWANMET*), or be interviewed (*RWINT*); and the status of decision makers or shareholders (*RWDESH*).

Finally, we used five variables for subjects of pictures displayed on the project's crowdfunding website, namely pictures of: providers of the project (*PTPROV*), receivers of the project (*PTREC*), products produced by or related to the projects (*PTPROD*), the production procedure or the production environment (*PTPROC*), and project slogans, logos, missions, and plans (*PTSLMP*).

Definitions and descriptive statistics of all variables are listed in Table 3.1. The fundraising target of the observations varies from $120 to $631,000, and the number of displayed pictures varies from zero to sixty-six. The most popular category was community activities which contributed 42.37% of crowdfunding projects in the data. Common types of rewards included products of the project (71.19%), honor certificates (44.07%), and letters of thanks (41.95%). Common themes of pictures include providers (45.76%), receivers (38.14%), products (47.88%), production procedure or environment (42.37%), and project slogans, logos, missions, and plans (52.12%).

Table 3.1: Variables in the models

Variable	Meaning of variable	Mean (SD)	(Min,Max)
LPF	Natural logarithm of the ratio between the result and target of the crowdfunding project.	-2.8303 (3.1725)	(-17.0344, 0.7969)
$LTARG$	Natural logarithm of the fundraising target.	9.4078 (1.5014)	(4.7875, 13.3551)
$LPICS$	Natural logarithm of the number of pictures displayed on the webpage of the project. The amount is added by one.	1.5177 (1.1606)	(0,4.1897)
$KICKS$	Dummy variable, 1 indicates the project is published on Kickstarter, 0 otherwise.	0.2203 (0.4154)	(0,1)

Table 3.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$COMACT$	Dummy variable, 1 indicates the topic of the project is community activities, 0 otherwise.	0.4237 (0.4952)	(0,1)
$DESTECH$	Dummy variable, 1 indicates the topic of the project design and technology, 0 otherwise.	0.1653 (0.3722)	(0,1)
$FOODCR$	Dummy variable, 1 indicates the topic of the project is food and craft, 0 otherwise.	0.1568 (0.3644)	(0,1)
$PROD$	Dummy variable, 1 indicates the project involves products, 0 otherwise.	0.3517 (0.4785)	(0,1)
$PERSON$	Dummy variable, 1 indicates the founder of the project is an individual person, 0 otherwise.	0.8263 (0.3797)	(0,1)
$MALE$	Dummy variable, 1 indicates the founder of the project is male, 0 otherwise.	0.4025 (0.4915)	(0,1)
$PROFL$	Dummy variable, 1 indicates a profile picture is displayed, 0 otherwise.	0.7415 (0.4387)	(0,1)

Table 3.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$SELF$	Dummy variable, 1 indicates the profile picture is a picture of the founder, 0 otherwise.	0.4788 (0.5006)	(0,1)
SPT	Dummy variable, 1 indicates the founder has supported other projects, 0 otherwise.	0.4407 (0.4975)	(0,1)
$BUSIN$	Dummy variable, 1 indicates the founder has experiences in starting a business, 0 otherwise.	0.1441 (0.3519)	(0,1)
$JOBDES$	Dummy variable, 1 indicates a job description of the founder is displayed on the crowdfunding website, 0 otherwise.	0.3432 (0.4758)	(0,1)
DEG	Dummy variable, 1 indicates a degree of the founder is displayed on the crowdfunding website, 0 otherwise.	0.2797 (0.4498)	(0,1)
$RWPROD$	Dummy variable, 1 indicates sponsors of the project receive products produced by or related to the project, 0 otherwise.	0.7119 (0.4539)	(0,1)

Table 3.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$RWEXP$	Dummy variable, 1 indicates sponsors of the project receive experiences and techniques, 0 otherwise.	0.0763 (0.2660)	(0,1)
$RWVIP$	Dummy variable, 1 indicates sponsors of the project receive VIP memberships of the project, 0 otherwise.	0.0763 (0.2660)	(0,1)
$RWHONOR$	Dummy variable, 1 indicates sponsors of the project receive honor certificates, 0 otherwise.	0.4407 (0.4975)	(0,1)
$RWLET$	Dummy variable, 1 indicates sponsors of the project receive letters of thanks, 0 otherwise.	0.4195 (0.4945)	(0,1)
$RWPHOTO$	Dummy variable, 1 indicates sponsors of the project receive photos, 0 otherwise.	0.0975 (0.2972)	(0,1)
$RWVISIT$	Dummy variable, 1 indicates sponsors of the project have chances to visit the project site, 0 otherwise.	0.2246 (0.4182)	(0,1)
$RWANM$	Dummy variable, 1 indicates sponsors of the project have chances to join annual meetings, 0 otherwise.	0.1568 (0.3644)	(0,1)

Table 3.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$RWINT$	Dummy variable, 1 indicates sponsors of the project have chances to be interviewed, 0 otherwise.	0.0847 (0.2791)	(0,1)
$RWDESH$	Dummy variable, 1 indicates sponsors of the project become decision makers or shareholders, 0 otherwise.	0.0254 (0.1577)	(0,1)
$PTPROV$	Dummy variable, 1 indicates a picture about the providers of the project is displayed, 0 otherwise.	0.4576 (0.4993)	(0,1)
$PTREC$	Dummy variable, 1 indicates a picture about the receivers of the project is displayed, 0 otherwise.	0.3814 (0.4868)	(0,1)
$PTPROD$	Dummy variable, 1 indicates a picture of products produced by or related to the project is displayed, 0 otherwise.	0.4788 (0.5006)	(0,1)
$PTPROC$	Dummy variable, 1 indicates a picture about the production procedure or production environment is displayed, 0 otherwise.	0.4237 (0.4952)	(0,1)

Table 3.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$PTSLMP$	Dummy variable, 1 indicates the slogan, logo, mission, or plan of the project is displayed, 0 otherwise.	0.5212 (0.5006)	(0,1)

3.4 Methodology

3.4.1 Linear Regression

Using the independent variables listed above, our aim was to explain the heterogeneity in existing estimates of t-statistics by constructing a regression model. A linear regression model with both continuous and categorical independent variables can be written as:

$$y_i = \beta_0 + \sum_{j=1}^{J} \delta_j z_{ij} + \sum_{k=1}^{K} \beta_k x_{ik} + \epsilon_i, \tag{3.1}$$

where β_0 is the intercept term, $z_1, \cdots, z_J$ are categorical independent variables with regression coefficients $\delta_1, \cdots, \delta_J$, $x_1, \cdots, x_K$ are continuous independent variables with regression coefficients $\beta_1, \cdots, \beta_K$, and ϵ is the error term.

3.4.2 Hybrid Tree-linear Regression

Applying a similar idea to the hybrid models given by Kim et al. (2017), a hybrid tree-linear model (HFLR) can be formed by Algorithm 4 in Chapter 2. The HTLR model can be written as:

$$y_i = \beta_0 + f(Z_i) + \sum_{k=1}^{K} \beta_j x_{ik} + \epsilon_i, \tag{3.2}$$

where $Z = \{z_1, \cdots, z_J\}$ and $f(Z)$ is a non-linear function.

Instead of variables that refer to leaf nodes of decision trees, we applied variables that are associated to splitting points in decision trees as new independent variables in the regression model. Although the two ways of generating new variables are mathematically equivalent, variables generated by splitting points are products of fewer original variables, and are therefore easier to interpret. The decision tree built in the HTLR model is shown in Figure 3.1. The categorical variables created for the splitting points are $SPT \times PTREC$, $(1 - SPT) \times RWPROD$, $(1 - SPT) \times RWPROD \times MALE$, etc.

The HTLR method only allows results of one decision tree to be added, which means that all interaction terms included in the model are related to each other by sharing the same root node. However, in real data sets, there is often a wide range of nonlinear relationships that do not frequently share a common variable. To allow multiple interaction terms that do not share a common root variable to enter the tree-based regression model, we have to include the results of multiple decision trees.

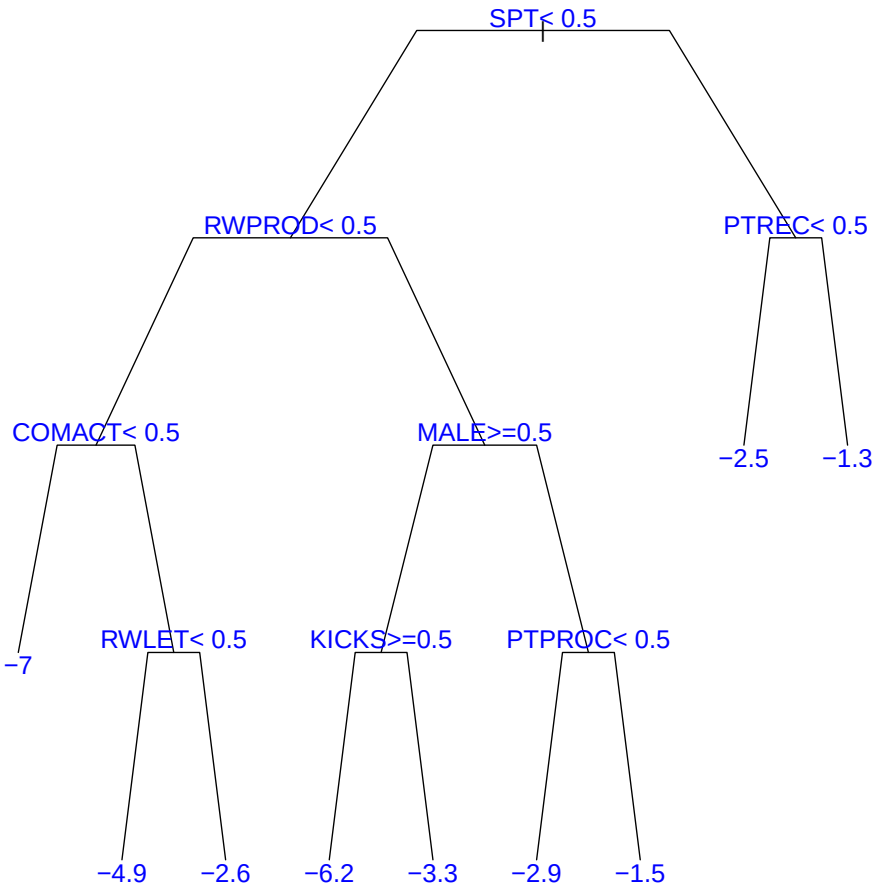


Figure 3.1: Decision Tree Built in HTLR Model

3.4.3 Hybrid Forest-linear Regression

Applying a loop algorithm, the hybrid forest-linear model (HFLR) given by Algorithm 6 generates a group of decision trees which can all be included in the same regression model. The results of HFLR models can also be written in the form of Equation 3.2.

Algorithm 6 Hybrid Forest-linear Regression (HFLR)

1.1) Create the set of categorical independent variables $Z_0 = \{z_1, \cdots, z_J\}$ the set of all independent variables $X_0 = \{z_1, \cdots, z_J, x_1, \cdots, x_K\}$. Create the set of active categorical variables $Z_1^a = Z_0$.
1.2) Run a linear regression model between y and all variables in X_0.
1.3) Apply stepwise regression to select variables included in the regression model. Calculate the BIC of the model.
repeat
2.1) Build a regression tree model between y and variables in Z_{ai}. Set the variable on the root node as r_i.
2.2) Create a set of new categorical variables Z_i based on splitting points in the decision tree build in 2.1.
2.3) Set $X_i = X_{i-1} \cup Z_i$, $Z_{i+1}^a = Z_i^a \setminus \{r_i\}$.
2.4) Run a linear regression model between y and all variables in X_i.
2.4) Apply an ANOVA test between the regression model build with X_i and X_{i-1}.
2.5) Apply stepwise regression to select variables included in the regression model in 2.4. Calculate the BIC of the model.
until the model build in 2.4 with X_i is not significantly better than the model build with X_{i-1}.
3) Select the model with the smallest BIC in the algorithm as the final result.

3.4.4 Monte-Carlo Simulations

We generated twelve simulated prediction problems to evaluate the performance of the linear regression (LR), hybrid tree-linear regression (HTLR), hybrid forest-linear regression (HFLR) which was applied in the crowdfunding analysis, and hybrid linear-forest regression (HLFR) which was applied in the meta-regression analysis.

The twelve simulated examples are all in the form of the following equation:

$$Y_{ij} = Y_{i0} + \epsilon_j, \tag{3.3}$$

where where $i = 1, 2, 3, 4$, $j = 1, 2, 3$, Y_{i0} are generated by independent variables,

and ϵ_j are error terms. The formula of Y_{i0} are given by the Equation 3.4:

$$\begin{aligned}
Y_{10} =& 10 + X_1 + X_4 + 5D_{11} + 3D_{51} + 2D_{71}, \\
Y_{20} =& 10 + X_1 + X_2 + X_5 + 5D_{12} + 6D_{52}D_{72} - 9D_{52}D_{54}D_{72} + 12D_{52}D_{54}D_{72}D_{74}, \\
Y_{30} =& 10 + X_3 + X_6 + 5D_{13} + 6D_{53}D_{73} - 9D_{55}D_{75}D_{95}, \\
Y_{40} =& 10 + X_4 + 5D_{14} + 6D_{52}D_{72} - 9D_{52}D_{54}D_{72} + 6D_{53}D_{73} - 9D_{55}D_{75}D_{95}.
\end{aligned} \tag{3.4}$$

Meanwhile, the error terms satisfies $\epsilon_j \sim N(0, \sigma_i^2)$ where $\sigma_1 = 2$, $\sigma_2 = 4$, and $\sigma_3 = 6$.

These twelve examples were obtained from linear and nonlinear models. Nonlinear relationships between Y_{20}, Y_{30}, and Y_{40} and categorical independent variables can be theoretically modeled as one, two, and three decision trees. Linear independent variables X_i, $i = 1, \cdots, 6$ are uniformly distributed in the interval [0,10]. On the other hand, distributions of categorical variables are $D_{ij} \sim B(P_i, 1)$, where $i \in \{1, 3, 5, 7, 9\}$, $j \in \{1, \cdots, 6\}$, and $P_i = i/10$. The number of coefficients was set to six, eight, six, or seven, and the sample size was set to 300.

Table 3.2 shows proportions of sources of variances of Y_{ij}. For example, when analysing the variance of Y_{31}, by calculating the sample variances, we have

$$\begin{aligned}
\frac{var(X_3 + X_6 + 5D_{13})}{var(Y_{31})} &= 0.389, \\
\frac{var(6D_{53}D_{73} - 9D_{55}D_{75}D_{95})}{var(Y_{31})} &= 0.528, \\
\frac{var(\epsilon_1)}{var(Y_{31})} &= 0.082,
\end{aligned}$$

which are the proportions of variances of Y_{31} from linear terms, interaction terms,

and the error term, respectively.

Table 3.2: Theoretical Proportions of Sources of Variances of Y_{ij}

Y_{ij}	Linear terms	Interaction terms	Error term
Y_{11}	0.846	0.000	0.154
Y_{12}	0.579	0.000	0.421
Y_{13}	0.379	0.000	0.621
Y_{21}	0.325	0.607	0.069
Y_{22}	0.269	0.503	0.228
Y_{23}	0.210	0.392	0.399
Y_{31}	0.389	0.528	0.082
Y_{32}	0.312	0.424	0.264
Y_{33}	0.235	0.319	0.447
Y_{41}	0.176	0.757	0.067
Y_{42}	0.147	0.631	0.222
Y_{43}	0.115	0.494	0.391

Each example was divided into a training set and a test set. The training set consisted of 200 observations, while the remainder was assigned to the test set. Our performance comparisons considered the standardised mean square errors (MSE/σ_j^2) of both the training set and the test set.

Table 3.3 shows the standardised MSE values within the training set for LR, HTLR, and HFLR models. For example, when running LR, HTLR, and HFLR models for Y_{11}, the MSE values within the training set are 1.246, 1.035, and 1.035, respectively.

From the results in Table 3.3, we concluded that hybrid tree-regression models performed consistently better than LR, while the differences between the standard-

Table 3.3: Standardised MSE of the Models within the Training Set

Method	LR	HTLR	HFLR
Y_{11}	1.246	1.035 *	1.035 *
Y_{12}	1.060	1.000	0.899 *
Y_{13}	0.947	0.891 *	0.891 *
Y_{21}	2.595	1.336 *	1.336 *
Y_{22}	1.494	1.061 *	1.061 *
Y_{23}	1.281	0.962	0.810 *
Y_{31}	2.990	1.304 *	1.304 *
Y_{32}	1.411	1.124 *	1.124 *
Y_{33}	1.134	0.977	0.924 *
Y_{41}	4.608	2.480	1.389 *
Y_{42}	1.781	1.662	1.280 *
Y_{43}	1.327	0.975 *	0.975 *
Average	1.823	1.234	1.086 *

For each Y_{ij}, * represents the method with the smallest MSE.

ised MSE values are larger when interaction terms were the major sources of variances of Y_{ij}. Furthermore, in all cases considered, HFLR was superior or equal to HTLR.

Table 3.4 shows the standardised MSE values within the test set for LR, HTLR, and HFLR models. For example, when running LR, HTLR, and HFLR models for Y_{41}, the MSE values within the training set are 4.700, 3.458, and 2.029, respectively.

From the results in Table 3.4, LR was the best model for Y_{11} and Y_{13} when only linear relationships existed between the dependent variable and the independent variables, and the differences between standardised MSEs are not large. HTLR was the best model for Y_{2j}, $j = 1, 2, 3$. where nonlinear relationships between the

Table 3.4: Standardised MSE of the Models within the Test Set

Method	LR	HTLR	HFLR
Y_{11}	1.236 *	1.552	1.552
Y_{12}	1.008	0.999 *	1.056
Y_{13}	1.022 *	1.072	1.072
Y_{21}	2.606	1.369 *	1.369 *
Y_{22}	1.540	1.043 *	1.043 *
Y_{23}	1.111	0.973 *	1.260
Y_{31}	3.531	1.660 *	1.660 *
Y_{32}	1.711	1.204 *	1.204 *
Y_{33}	1.338	1.111	1.105 *
Y_{41}	4.700	3.458	2.029 *
Y_{42}	2.258	2.336	1.924 *
Y_{43}	2.212	1.525 *	1.525 *
Average	2.023	1.525	1.400 *

For each Y_{ij}, * represents the method with the smallest MSE.

dependent variable and the independent variables could be modeled in one single decision tree. HFLR was the best model for Y_{3j} and Y_{4j}, $j = 1, 2, 3$, where nonlinear relationships were more complex. Additionally, for Y_{21} and Y_{22}, HFLR gave the same models as HTLR which were the best models.

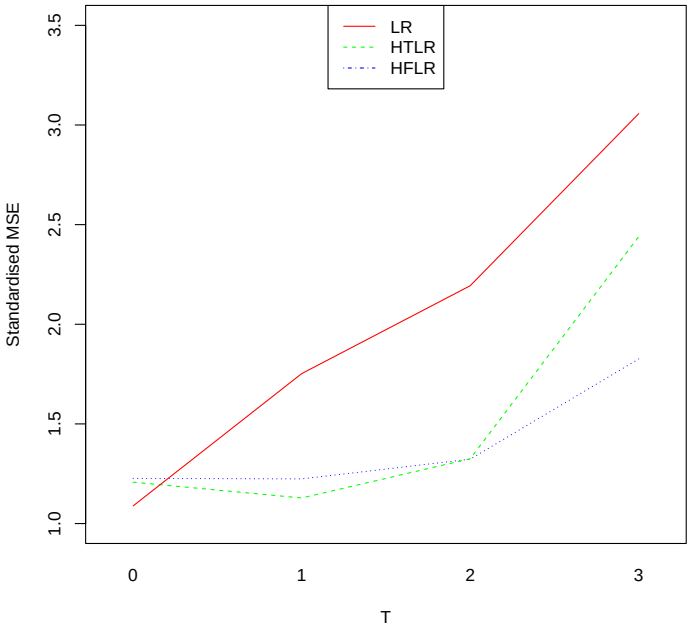
To access the effects of various parameters on standardised MSE in a more accurate manner, we applied two-way ANOVA to the experimental data given in Table 3.4. The experimental setting for each approach can be regarded as a full factorial design. The factors for the proposed approach included the prediction method (denoted by M with three levels of LR, HTLR, and HFLR), true number of trees (denoted by T with four levels of zero, one, two, and three), and standard deviations of the error term (denoted by S with three levels of two, four, and six).

Table 3.5 shows the ANOVA table for the test set. The three-way interaction effect ($M \times T \times S$) was assumed to be negligible. The main effects M, T, and S as well as for the interaction effect $M \times S$ were significant at 0.05 level, and the interaction effect $M \times T$ were significant at 0.1 level.

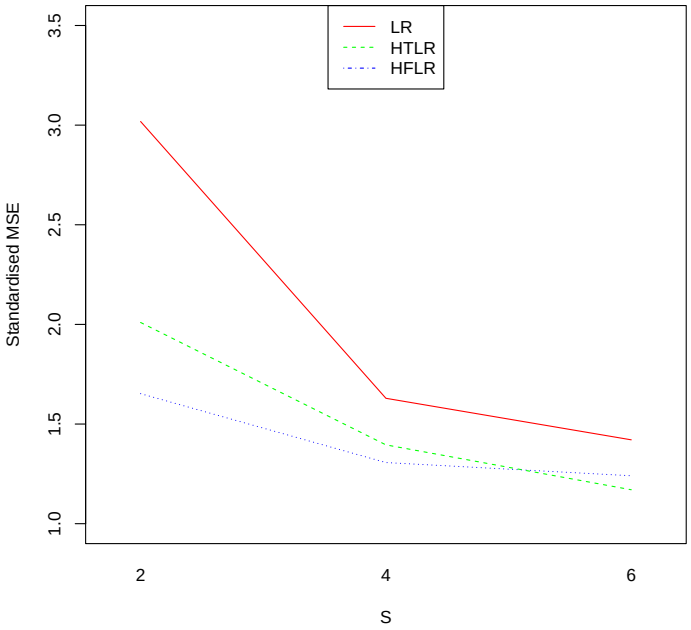
Table 3.5: ANOVA Table for Standardised MSE within the Test Set

	Df	Sum Sq	Mean Sq	F	p-value
M	2	2.605	1.3024	10.010	0.0028
T	3	8.391	2.7969	21.497	4.07×10^{-5}
S	2	6.171	3.0855	23.715	6.78×10^{-5}
$M \times T$	6	1.890	0.3150	2.421	0.0908
$M \times S$	4	1.764	0.4410	3.390	0.0448
$T \times S$	6	1.384	0.2307	1.774	0.1875
Error	12	1.561	0.1301		

Since all main effects were involved in the interaction effects, the effects of M, T, and S on standardised MSE should be assessed using their interaction plots. Figure 3.2 shows the interaction plots. The red real line, the green dashed line, and the blue dotted line represent the average standardised MSE for LR, HTLR and HFLR. According to the results, the performances of the all methods become worse as T increases and S decreases. HFLR was more robust than the other methods as T and S changes. Tree based models were superior to LR regardless of S.



(a) M and T



(b) M and S

Figure 3.2: Interaction Plots

3.5 Results and Discussion

3.5.1 Results of Regression Models

Table 3.6 summarises the results of our regression models. The HFLR model includes results from four decision trees which are displayed in Figures 3.1 and 3.3.

Table 3.6: Results of the models

Variable	LR	HTLR	HFLR
Intercept	5.482*** (1.097)	2.187* (1.058)	1.355 (0.933)
$LTARG$	-1.100*** (0.112)	-1.055*** (0.099)	-0.843*** (0.084)
$KICKS$	-1.236** (0.414)		
$PROD$	1.174** (0.361)	0.796* (0.314)	
$MALE$	-1.021** (0.327)		
SPT	1.340*** (0.321)	4.834*** (0.578)	2.366*** (0.584)
$RWPROD$	1.641*** (0.365)		2.653*** (0.640)
$RWINT$	1.657** (0.573)	1.551** (0.535)	1.098* (0.428)
$RWDESH$			-1.888* (0.778)
$PTREC$		0.783* (0.312)	
$PTPROV$			0.780** (0.256)
$PTPROC$	0.956** (0.337)		

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 3.6: Results of the models (continued)

Variable	LR	HTLR	HFLR
$PTSLMP$			0.656* (0.255)
$(1-SPT) \times RWPROD$		4.245*** (0.609)	1.709* (0.659)
$(1-SPT) \times RWPROD \times MALE \times KICKS$		-3.791*** (0.901)	-3.487*** (0.731)
$(1-SPT) \times RWPROD \times (1-MALE) \times PTPROC$		1.409** (0.519)	
$(1-SPT) \times (1-RWPROD) \times COMACT$		3.789*** (0.731)	2.401*** (0.594)
$(1-RWPROD) \times PTPROD$			2.568*** (0.722)
$(1-RWPROD) \times (1-PTPROD) \times RWHONOR$			3.395*** (0.689)
$(1-RWPROD) \times (1-PTPROD) \times RWHONOR \times PTPROV$			-3.062*** (0.824)
$(1-RWPROD) \times (1-PTPROD) \times (1-RWHONOR) \times JOBDES$			1.773* (0.730)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 3.6: Results of the models (continued)

Variable	LR	HTLR	HFLR
$(1 - PROD) \times RWLET$			-0.865** (0.285)
$(1 - PROD) \times (1 - RWLET) \times MALE$			-4.751*** (0.609)
$(1 - PROD) \times (1 - RWLET) \times MALE \times PTPROD$			4.480*** (0.755)
$(1 - PROD) \times (1 - RWLET) \times MALE \times (1 - PTPROD) \times JOBDES$			2.765** (0.852)
$(1 - PROD) \times (1 - RWLET) \times (1 - MALE) \times (1 - RWVISIT) \times PTSLMP$			-2.243*** (0.501)
$PTPROD \times COMACT \times MALE$			-1.809*** (0.454)
$PTPROD \times (1 - COMACT) \times (1 - RWLET) \times RWANM$			-2.321*** (0.688)
$(1 - PTPROD) \times (1 - RWANM) \times KICKS$			-2.381*** (0.524)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

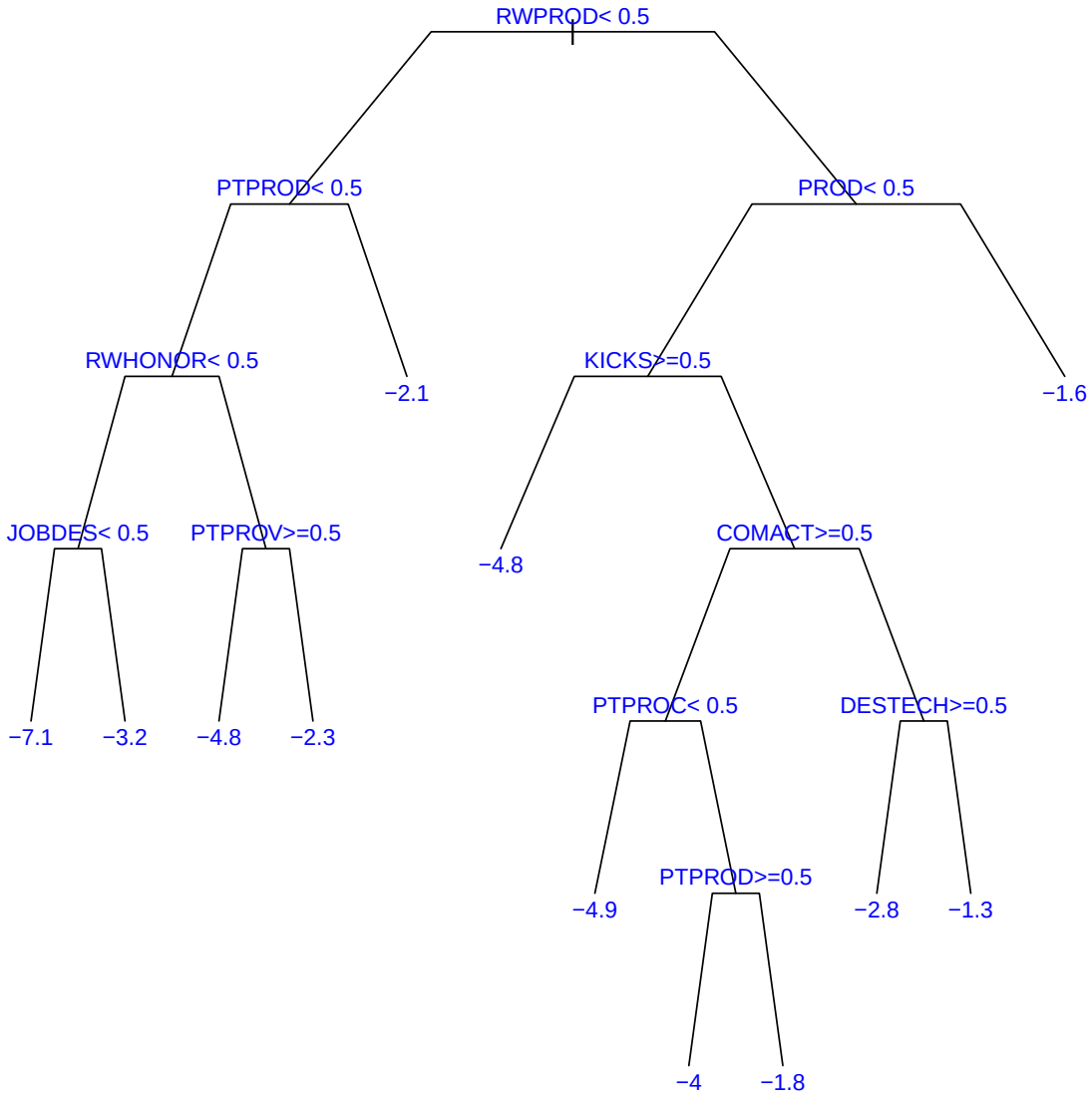
Table 3.6: Results of the models (continued)

Variable	LR	HTLR	HFLR
$(1 - PTPROD) \times (1 - RWANM) \times (1 - KICKS) \times (1 - PROFL) \times PTSLMP$			-2.480*** (0.716)

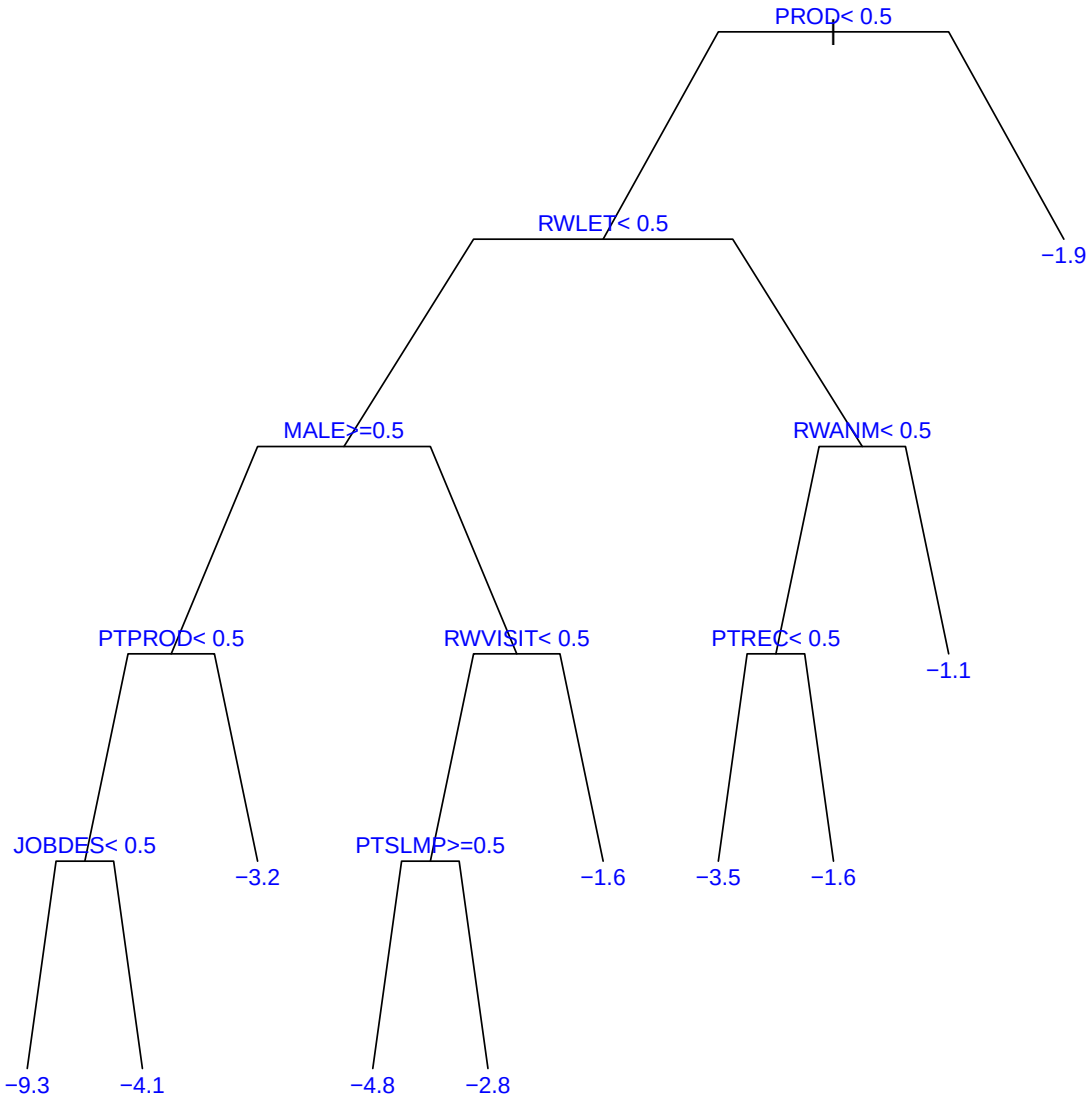
Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

The fundraising target had a significantly negative impact on the proportion of financing. The variable $LTARG$ is negative significant in all models. In all models, its coefficients are significantly less than 0, which means that holding all other variables constant, the proportion of financing decreases as the fundraising target rises. According to t-tests, in LR and HTLR models the coefficients of $LTARG$ are not significantly different from -1. However, in the HFLR model, the result of t-test show that the coefficients of $LTARG$ is significantly larger than -1 at the significance level 0.1, suggesting that the fundraising target have a positive impact on the size of financing. The variable $LPICS$ is not included in any models, which means that the number of pictures does not have a significant impact on the proportion of financing.

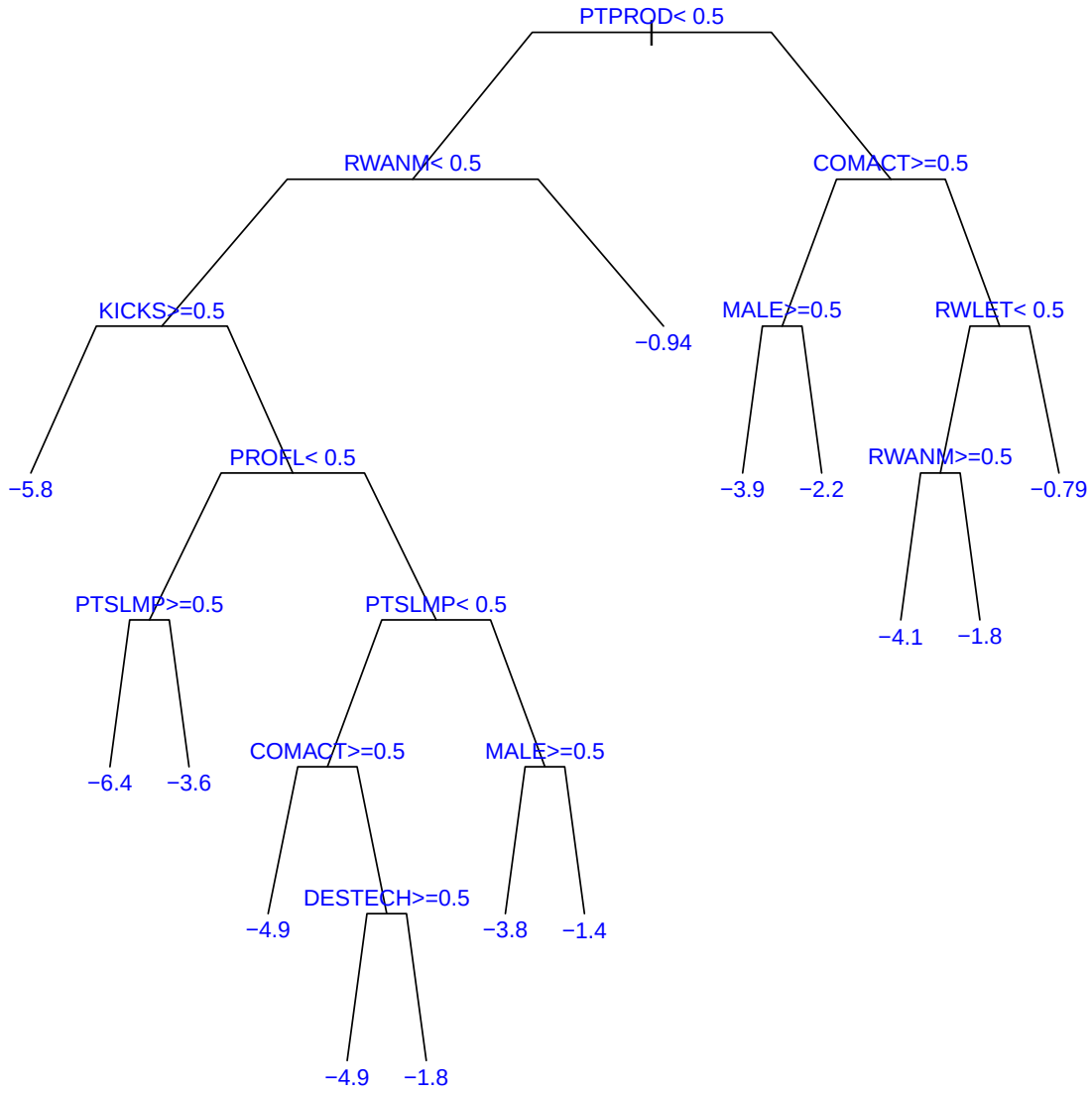
All models show that certain types of rewards, such as products and interviews, had positive significant influences on the proportion of financing, although the effect is not necessarily linear. According to the results of all models, the coefficient of $RWINT$ is positive significant and the value of the coefficient is between 1.098



(a) The Second Tree



(b) The Third Tree



(c) The Fourth Tree

Figure 3.3: Decision Trees Generated in the HFLR Algorithm

and 1.657, which means that a chance to be interviewed can boost the proportion of financing by between 3.0 and 5.2 times. In LR and HFLR models, the variable $RWPROD$ is included, while in HTLR and HFLR models, the interaction term $(1-SPT)\times RWPROD$ is included. All of these terms are positive significant, which means that the product produced by the project is also helpful for crowdfunding projects to achieve higher proportions of funding, especially when the founder has not supported other projects. Notably, the coefficients of SPT are positive significant in all models, and those coefficients of HTLR and HFLR models are much larger than that of the LR model. The interaction term suggests that, although a founder that has supported other projects and a reward that is a product are both useful factors for raising more money, their joint effects are smaller than the sum of their individual effects.

The results of the effects of visualisation tools differ between models. In the LR model, the variable $PTPROC$ is positive significant, suggesting that pictures of the production process or production environment had a positive effect on the proportion of funding. In the HTLR model, the positive significant variable about visualisation tools is $PTREC$, which are related to pictures of receivers of the product. In the HFLR model, both $PTPROV$ and $PTSLMP$ are positive significant, which means that visualisation tools about providers of the product and those about slogans, logos, missions, and plans are persuasive to the supporters to be more generous.

In the HFLR model, some interaction terms have very clear meanings. For example, the term $(1-RWPROD)\times PTPROD$ has a positive significant coefficient, suggesting that when there are no products as rewards for supporters of a project, pictures of products related to the project become persuasive to the supporters. In

comparison, the term $(1 - RWPROD) \times (1 - PTPROD) \times RWHONOR$ is also positive significant, suggesting that when products related to the project are neither given as rewards nor presented on the website, supporters of the projects may be more generous if they receive honor certificates as rewards.

3.5.2 Comparison between Regression Models

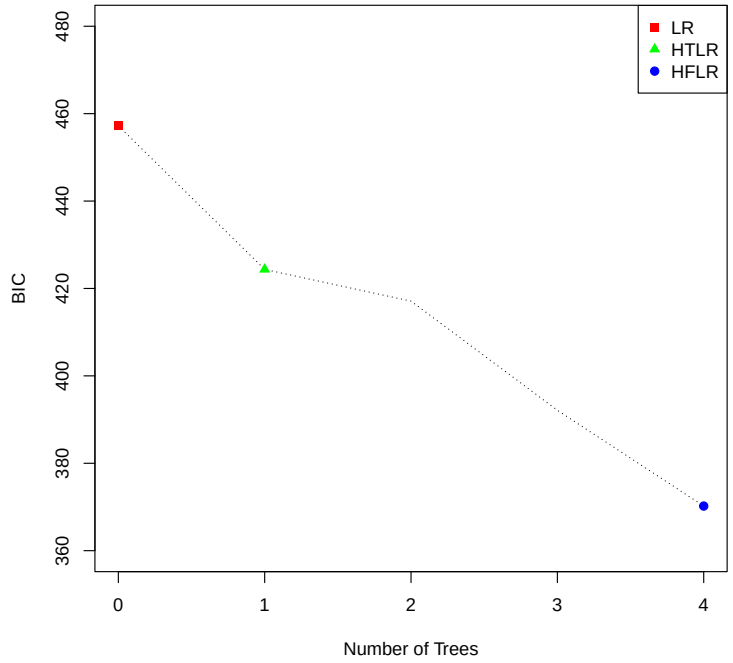
Table 3.7 shows summary statistics and certain model selection criteria of models built in our study; the prediction accuracy is based on predictions for a given project being successful or not. Figure 3.3 shows the change of BIC, cross validated MSE, and prediction accuracy during the HFLR algorithm until the loop terminates.

Table 3.7: Comparison of the Models

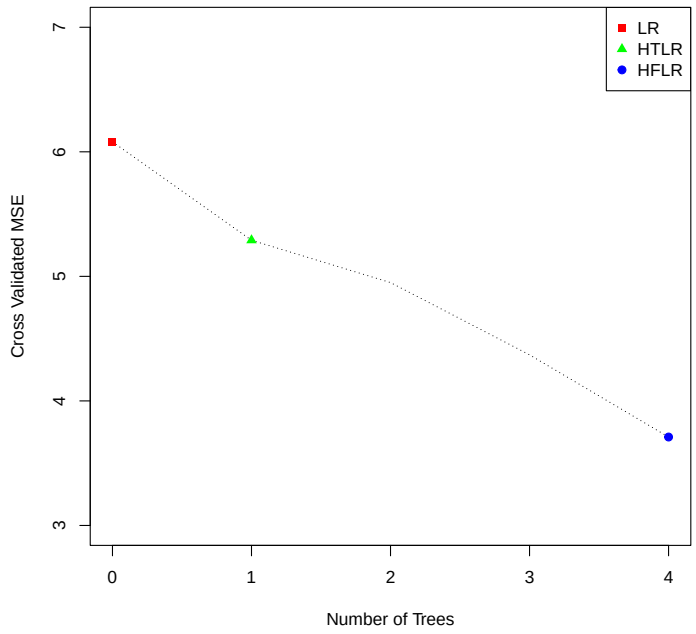
Method	LR	HTLR	HFLR
Amount of trees	0	1	4
Amount of parameters	9	10	24
R^2	0.4377	0.5220	0.7252
$\bar{R}^2$	0.4179	0.5030	0.6954
BIC	457.23	424.36	370.21*
Cross-validated MSE	6.08	5.29	3.71*
Prediction accuracy	78.81%	80.51%	81.36%*

* represents the selected model from each model selection criteria.

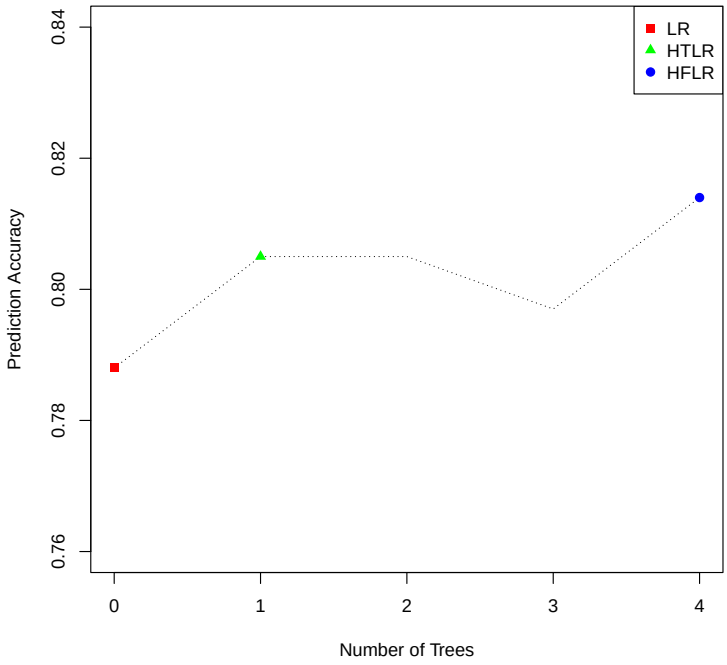
The explanation powers of all tree-based models were much higher than those



(a) BIC



(b) Cross Validated MSE



(c) Prediction Accuracy

Figure 3.4: Change of Model Efficiency during the HFLR Algorithm

of the linear model. According to the values of BIC, the model with the highest explanation power was the HFLR model, which also has the smallest MSE in 10-fold cross validation. According to the Figures 3.3(a) and 3.3(b), during the process of the HFLR algorithm, both BIC and cross validated MSE consistently decrease when significant variables from the first four decision trees are added. In addition, the HFLR model was the best model in terms of prediction powers. According to Figure 3.3(c), the HFLR model with four decision trees was the model with the highest prediction accuracy during the HFLR algorithm. In conclusion, the HFLR model is the best model according to all model selection criteria applied in our study; more precisely, the HFLR model is more powerful than the HTLR model, and the HTLR model is more powerful than the linear regression model.

Notably, as the best model selected in our study, the HFLR model included twenty-one interaction terms and only two linear terms ($LTARG$ and SPT). This result suggests that the relationship between the proportions of funding for crowdfunding projects and the factors considered in our study is often nonlinear.

Nonlinear terms in tree-based regression models can be divided into at least two categories: interpretable nonlinear terms and sparse nonlinear terms. Interpretable nonlinear terms, such as $(1 - SPT) \times RWPROD$, $(1 - RWPROD) \times PTPROD$, and $(1 - RWPROD) \times (1 - PTPROD) \times RWHONOR$ which are discussed above, can be easily interpreted in the real background. They are included in the model to deduce the joint effects of independent variables on the fundraising performance. Notably, these variables were discovered because we applied splitting points instead of leaf nodes in decision trees to generate new independent variables.

On the other hand, sparse nonlinear terms are interaction terms which are nonzero for only a small proportion of observations. In the HFLR (ANOVA) model which has twenty-four coefficients in total, ten terms are only nonzero for no more than 10% (23/236) of observations. Five terms, including $(1-SPT) \times RWPROD \times MALE \times KICKS$, $(1-RWPROD) \times (1-PTPROD) \times RWHONOR \times PTPROV$, $(1-PROD) \times (1-RWLET) \times MALE \times (1-PTPROD) \times JOBDES$, $PTPROD \times (1 - COMACT) \times (1 - RWLET) \times RWANM$, and $(1 - PTPROD) \times (1 - RWANM) \times (1 - KICKS) \times (1 - PROFL) \times PTSLMP$, are only nonzero for no more than 5% (11/236) of observations.

Although the meaning of these sparse nonlinear terms is often difficult to interpret, they play significant roles in tree-based regression models. For example, when

the five sparsest terms are deleted from the HLFR (ANOVA) model, the R^2 drops from 0.7249 to 0.6374, and a few other terms become insignificant. Stainberg & Cardell (1998) suggested that predictions made by decision trees are less sensitive to outliers. Outliers can be put in specific branches in decision trees, and therefore can be discussed separately. When these branches are included as nonlinear terms in tree-based regression models, the regression models also become less sensitive to related outliers.

3.6 Conclusion

In this paper, we applied regression models to explore the factors influencing the proportion of financing for crowdfunding projects. We demonstrated that various types of rewards, including the chance to be interviewed and to receive products, are helpful for crowdfunding projects to achieve higher proportions of financing. We also noted that the fundraising target does not have a significant effect on the size of financing.

Since decision tree learning algorithms were first developed, some papers have suggested that hybrid methods between classic regression models and decision trees are able to combine the strengths of both models to achieve higher explanation and prediction powers. This paper introduces two tree-based regression models: the HTLR model and the HFLR model. By comparing various statistical criteria, we observed that these tree-based regression models are better than the linear regression model in both explanation and prediction.

The reason for tree-based regression models outperforming the linear regression model is related to the inclusion of nonlinear terms: some of these terms clearly demonstrate the joint effects of various factors. As an example, in linear models, the effects of different types of rewards on fundraising performance are considered independently and are accumulative; but in reality, when various rewards are offered, their joint effect is often smaller than the sum of individual effects.

Meanwhile, some other nonlinear terms are likely associated with the effects of a specific group of outliers. When these terms are included, the models become less sensitive to outliers, and the effects of other factors are more accurately estimated.

We also noted that when applying an appropriate loop termination condition, the HFLR model has higher explanation and prediction powers than HTLR models. This is because, by applying a circular algorithm, HFLR models are able to combine results of multi decision trees. As a result, HFLR models are more inclusive to factor referring to joint effects.

Chapter 4

Predicting Environmental Willingness to Pay with Hybrid Tree-Regression Techniques

4.1 Introduction

Values of environmental projects, such as the treatment of air pollution, the control of water quality, and the protection of urban green spaces, are often unable to be estimated from market processes. Instead, they are commonly estimated by preferences stated in responses to survey questions. A target of this chapter is the application of the contingent valuation (CV) method to analyse people's willingness to pay (WTP) of income tax to fund an environmental geo-engineering project.

Tietenberg et al. (2018) indicated that the CV method is the only direct technique for measuring values from stated preferences. In CV surveys, participants are either asked directly for their maximum amount of WTP, or whether they would pay a possible bid value for the project. One of the major concerns of applying the CV method is the starting-point bias: the range of designed bids may affect the resulting estimate of WTP. In order to deal with the starting-point bias, we applied a new multi bounded model and it was compared with the single bounded model.

We built single bounded and multi bounded models for people's WTP and certain independent variables with linear regression (LR), hybrid tree-linear regression (HTLR), and hybrid forest-linear regression (HFLR) methods which were introduced in Chapter 3. By comparing performances of these methods, we are verifying the assumption that HFLR method have higher explanatory and prediction powers than LR and HTLR.

The structure of the remainder of Chapter 4 is as follows. In Section 4.2, the data collection process is briefly described, and variables in single bounded and multi bounded regression models are chosen. In Section 4.3, both single bounded and multi bounded regression models are introduced. In Section 4.4, the results of these models are given and discussed. In Section 4.5, we summarise the conclusions about the study.

4.2 Data

4.2.1 Data Collection

A WTP question contains two pieces of information: one is the setting of the beneficial scenario which, in our case, is our geo-engineering project which could allow people to live in a safer environment; the other is the description of the risk, which is the increase of their income tax. The question in our survey is given as follows:

> "Currently, China is facing serious environmental problems which affect people's life quality, including air pollution, water pollution, and the lack of green spaces. China has set up a special project to improve people's living environment, called 'geo-engineering'. The establishment of this project can greatly improve your living environment through comprehensively treating air pollution sources, controlling the quality of living water, and constructing green spaces. Would you be willing to increase your income tax by ____ every year in exchange for this project, which will greatly improve your life quality by making you live in a pollution-free environment?"

The bid is chosen from eight random levels: 5, 19, 57, 95, 285, 476, 952, and 1904.

Each participant who replied "YES" to the WTP question were given the bid to the next higher level. For those who replied "NO" were given the bid to the next lower level. This process was repeated until the answer of the participant changed,

and the last bid before the change of answer would be recorded.

1044 samples were collected from four Chinese cities (Zhengzhou, Harbin, Changsha, and Zhuhai) by nine trained interviewers. The raw database contained certain types of errors including typos, missing values, and logical mistakes. After correcting these errors and removing samples with missing values on thirty-two selected variables, 768 observations were used in the modelling process.

4.2.2 Variables in the Models

We applied two different models to analyse the data which are discussed in Section 4.4. In the single bounded model, the independent variable ($BID1$) is the first bid given to the participant, and the dependent variable ($YES1$) is the yes or no response given by a participant in the first round, which is defined by

$$YES1 = WTP \geq BID1.$$

In the multi bounded model, each participant is considered as eight observations with eight different bids $BID = 5, 19, 57, 95, 285, 476, 952, 1904$. Thus, the multi bounded model had 6144 observations. The dependent variable YES is the response of the participant to BID, which is defined by

$$YES = WTP \geq BID.$$

Additionally, $LBID1 = \ln(BID1)$ is also included as an independent variable in the

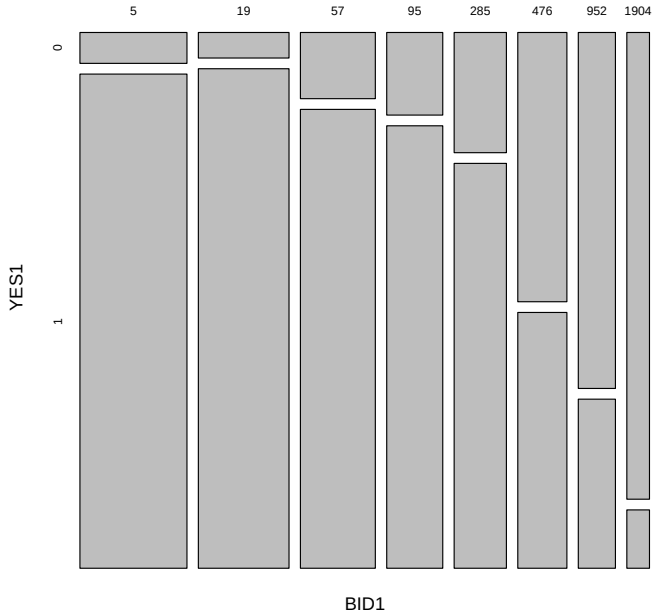
multi bounded model.

The distributions of replies of participants to the first given bid and those of all participant-bid combinations are shown in Figure 4.1. According to the results of the first round, the bid which is the closest to the median of WTP is 476: about 48.72% of participants who were given the bid 476 replied "yes". However, according to the final results, the bid which is the closest to the median of WTP is 285: the WTP of 49.74% of participants were larger than 285.

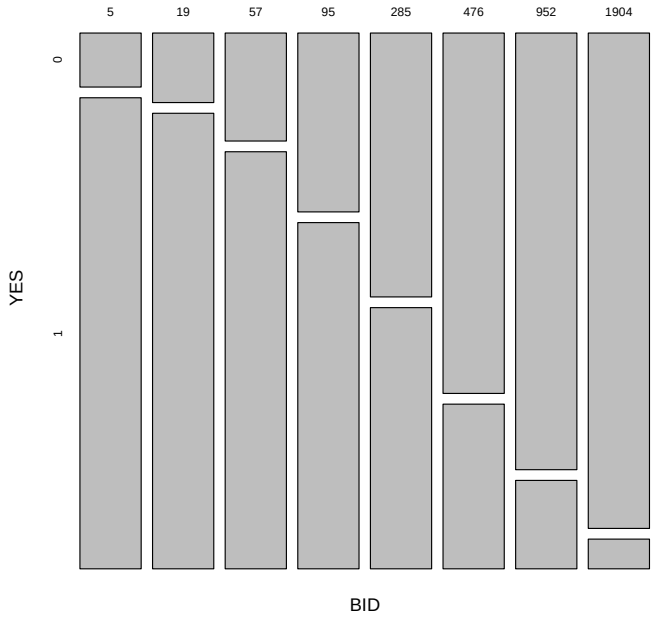
Both regression models include thirty-two control variables; one of which is a continuous variables the family income ($LFINC$) of participants. We took natural logs of all continuous variables in the model.

We used six variables to represent the attitude of participants about statements related to scientific studies, global warming, and geo-engineering projects: if scientists are biased ($STBIAS$), if more studies must be done before being relied on ($STMOR$), if global warming brings serious risks ($GWRISK$), if it is important to act against global warming ($GWACT$), if global warming is caused by human activities ($GWMANMD$), and if the geo-engineering project is beneficial ($BENEFIT$).

Another ten variables hold information relating to the attitude of the participants to the following social and political statements: individual rights should not depend on the attitude of others ($RIGHT$), a fairness revolution is needed ($FAIREV$), the best way of life is to do what is told ($TOLD$); people who do not obey authority cause problems ($AUTH$); life is largely determined by forces that are not controlled ($NOCONT$), people should be protected from hurting themselves by laws ($PROTLAW$), protecting people from hurting themselves is not a government business



(a) Distributions of Replies of Participants to the First Given Bid



(b) Distributions of Replies of all Participant-bid Combinations

Figure 4.1: Distribution of Replies of Participants to All Given Bids

(*PROTNG*), government should do more for social goals (*GOVGOAL*), there is too much push for equality (*TMEQL*), and the society has become too soft (*SOFSOC*).

We used a further nine variables to the risks that the participants are concerned with: chemical additives in food (*RCHEMA*), increased immigration (*RIMIG*), lawsuits against reporters and news media for libel (*RLMED*), speeches inciting racial hatred (*RRACHAT*), genetically modified food (*RGMFOOD*), illegal drug trafficking (*RDRUG*), government regulations of businesses (*RGOVREG*), childhood vaccinations (*RVAC*), and government budget deficits (*RGOVBUD*).

Finally, we used six variables for basic information and recent medical treatments of the participants: their age group (*AGE*), gender (*GEND*), interest in news and public affairs (*NEWS*), if they have seen a doctor (*DOCT*) or been hospitalised (*HOSP*) in the past six months, and their frequency of taking anti-aging drugs (*ANTIAGE*).

Definitions and descriptive statistics of all variables are listed in Table 4.1.

Table 4.1: Variables in the models

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$YES1$	Response of a participant to the first given bid.	0.7747 (0.4180)	(0,1)
$LBID1$	Natural logarithm of the bid given to the participant in the first round.	4.1117 (1.8547)	(1.6094, 7.5517)

Table 4.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
YES	Response of a participant to $BID = e^{LBID}$ based on the result of the final round.	0.5337 (0.4990)	(0,1)
$LBID$	Natural logarithm of an imaginative bid given to the participant.	4.9224 (1.8888)	(1.6094, 7.5517)
$LFINC$	Natural logarithm of family income of the participant.	12.0370 (1.5235)	(9.2103, 16.1181)
$STBIAS$	Agreement on "Scientists who did the study were biased", 1-6 scale.	3.2526 (1.1549)	(1,6)
$STMOR$	Agreement on "More studies must be done before policymakers rely on findings", 1-6 scale.	5.2461 (1.0164)	(1,6)
$GWRISK$	Agreement on "Global warming brings serious environmental risks", 1-6 scale.	5.2161 (1.0033)	(1,6)
$GWACT$	Agreement on "It is important to take actions to reduce global warming", 1-6 scale.	5.3607 (0.9100)	(1,6)
$GWMANMD$	Agreement on "Human activity causing global temperatures to rise", 1-6 scale.	4.7448 (1.1222)	(1,6)

Table 4.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$BENEFIT$	Agreement on "This geo-engineering project will benefit us", 1-6 scale.	4.7044 (1.0122)	(1,6)
$RIGHT$	Agreement on "Right of individual should not depend on how much others are willing to pay to avoid damage", 1-6 scale.	4.4167 (1.3446)	(1,6)
$FRIREV$	Agreement on "Need a fairness revolution", 1-7 scale.	5.3294 (1.2805)	(1,7)
$TOLD$	Agreement on "Best way to get ahead in life is to do what told to do", 1-7 scale.	3.2448 (1.6601)	(1,7)
$AUTH$	Agreement on "Society in trouble because people do not obey authorities", 1-7 scale.	3.3867 (1.5234)	(1,7)
$NOCONT$	Agreement on "Course of lives largely determined by forces beyond our control", 1-7 scale.	4.1732 (1.6236)	(1,7)
$PROTLAW$	Agreement on "Government needs to make laws that keep people from hurting themselves", 1-6 scale.	4.0052 (1.2539)	(1,6)

Table 4.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$PROTNG$	Agreement on "Not governments business to protect people from themselves", 1-6 scale.	2.6563 (1.3516)	(1,6)
$GOVGOAL$	Agreement on "Government should do more to advance society's goals, even if limiting freedom", 1-6 scale.	2.9505 (1.4307)	(1,6)
$TMEQL$	Agreement on "Gone too far in pushing equal rights in this country", 1-6 scale.	2.8203 (1.1851)	(1,6)
$SOFSOC$	Agreement on "Society has become too soft and feminine", 1-6 scale.	3.3151 (1.3422)	(1,6)
$RCHEMA$	Agreement on "Chemical additives in food is a risk", 0-10 scale.	7.4297 (2.3373)	(0,10)
$RIMIG$	Agreement on "Increased immigration is a risk", 0-10 scale.	5.5286 (2.7851)	(0,10)
$RLMED$	Agreement on "Lawsuits against reporters and news media for libel is a risk", 0-10 scale.	6.2956 (2.7945)	(0,10)
$RRACHAT$	Agreement on "Speeches inciting racial hatred is a risk", 0-10 scale.	8.0104 (2.3075)	(0,10)

Table 4.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$RGMFOOD$	Agreement on "Genetically modified foods is a risk", 0-10 scale.	6.0260 (2.7884)	(0,10)
$RDRUG$	Agreement on "Illegal drug trafficking is a risk", 0-10 scale.	9.0729 (1.7792)	(0,10)
$RGOVREG$	Agreement on "Government regulations of businesses is a risk", 0-10 scale.	6.5651 (3.4207)	(0,10)
$RVAC$	Agreement on "Childhood vaccinations is a risk", 0-10 scale.	5.6719 (3.5215)	(0,10)
$RGOVBUD$	Agreement on "Government budget deficits is a risk", 0-10 scale.	6.8164 (2.6717)	(0,10)
AGE	Age group of the participant, 1 if 0-17, 2 if 18-24, 3 if 25-29, 4 if 30-34, 5 if 35-39, 6 if 40-49, 7 if above 50.	3.9844 (1.4861)	(1,7)
$GEND$	Gender of the participant, 0 if male, 1 if female.	0.5130 (0.5002)	(0,1)
$NEWS$	Interest of the participant in news and public affairs, 1-4 scale.	1.9167 (0.8656)	(1,4)
$DOCT$	Response on "Have you seen a doctor in the past 6 months", 1 if yes.	0.2604 (0.4391)	(0,1)

Table 4.1: Variables in the models (continued)

Variable	Meaning of variable	Mean (SD)	(Min,Max)
$HOSP$	Response on "Have you been hospitalised in the past 6 months", 1 if yes.	0.0299 (0.1706)	(0,1)
$ANTIAGE$	Response on "How often do you take anti-aging drug or use related production", 0 if rarely or never, 1 if once per month, 2 if 2-3 times per month, 3 if once per week, 4 if 2-3 times per week, 5 if everyday or more.	0.3008 (1.0184)	(0,5)

4.3 Methodology

4.3.1 Underlying WTP Function

The WTP models in this study are based on the WTP function given by the contingent valuation model of Cameron et al. (1987):

$$WTP = \exp\{X\beta + \epsilon\}, \tag{4.1}$$

where X is the explanatory variables matrix, β is the vector of coefficients, and $\epsilon \sim N(0, \sigma)$ is the vector of random errors.

Since WTP values cannot be directly calculated from the results of the survey, in our WTP models, the binomial variables $YES1$ and YES defined in Section 4.2 are applied to observe ranges of true values of WTP. The variable $YES1$ is applied in the single bounded model, and the variable YES is applied in the multi bounded model.

4.3.2 Single Bounded Model

According to the definition of $YES1$, the single bounded model is derived from

$$P\{YES1 = 1\} = P\{WTP \geq BID1\} = 1 - \Phi(\frac{BID1 - X\beta}{\sigma}). \tag{4.2}$$

Rearranging Equation 4.2, we have

$$\Phi^{-1}(P\{YES1_i = 1\}) = \beta_0 + \alpha LBID1_i + \sum \beta_j x_{ij} + \epsilon_i, \tag{4.3}$$

which can be estimated by a probit model. Linear regression (LR), hybrid tree-linear regression (HTLR) given by Algorithm 4 in Chapter 2, and hybrid forest-linear regression (HFLR) methods given by Algorithm 6 in Chapter 3, are applied to estimate Equation 4.3.

Instead of the mean WTP, we estimated the median WTP in this study because it is more robust especially with the presence of outliers. Substituting $P\{YES = 1\} = \frac{1}{2}$ in the estimation of Equation 4.3, we have

$$0 = \Phi^{-1}(\frac{1}{2}) = \hat{\beta}_0 + \hat{\alpha}\hat{Q}_{\frac{1}{2}}(\ln WTP) + \sum \hat{\beta}_j x_{ij}. \tag{4.4}$$

Rearranging Equation 4.4, given a certain set of x_{ij}, the estimated median WTP is calculated by

$$\widehat{Q_{\frac{1}{2}}}(WTP) = \exp\{-\frac{1}{\hat{\alpha}}(\hat{\beta}_0 + \sum \hat{\beta}_j x_{ij})\}. \tag{4.5}$$

From Equation 5, when a control variable x_j is increased by 1, the median of WTP is estimated to change by the proportion $\exp\{-\frac{\hat{\beta}_j}{\hat{\alpha}_1}\}$.

In the single bounded model, since there is only one bid in the equation, two different effects of the bid are combined in one single coefficient α. One of them is when the bid becomes higher, it is more likely to be greater than the WTP of the participant. The other is the starting-point bias, which means that when participants are given a higher bid, their WTP may become higher because of psychological suggestions. In comparison, single bounded models only applies the results of the first round, while from the collected data, we were able to compare the WTP of each participant with eight given bids.

4.3.3 Multi Bounded Model

Hanemann et al. (1991) pointed out that the double bounded model is more statistically efficient compared to the single bounded model. The double bounded model applies the responses of participants to bids given in the first and second rounds of the survey.

In our study, according to the bids given to 764 participants that joined the

second round of the survey, we have

$$cor(LBID1, LBID2) = 0.9219,$$
$$cor(LBID1, \Delta LBID) = -0.7129,$$

where $LBID2$ is the natural logarithm of the bid given to the participant in the second round and $\Delta LBID = LBID2 - LBID1$. If both $LBID1$ and $LBID2$ were included in a regression model, no matter if we applied $LBID2$ directly or $\Delta LBID$, estimates of the model would become less robust due to collinearity. Although the double bounded model applies more information from the survey data compared to the single bounded model, the information from the third and further rounds of the survey is still unused.

Consider a bid chosen from the set $\{5, 19, 57, 95, 285, 476, 952, 1904\}$. For each participant, the bid is either given to the participant, smaller than a bid that the response of the participant is "yes", or larger than a bid that the response of the participant is "no". Thus, for each participant-bid combination, we can always know the response of the participant to the bid. To sufficiently apply information from the collected data, we consider each participant-bid combination as an observation to build a multi-bounded model.

According to the definition of YES in Section 4.2, the multi bounded model is derived from

$$P\{YES = 1\} = P\{WTP \geq BID\} = 1 - \Phi(\frac{BID - X\beta}{\sigma}). \tag{4.6}$$

Notably, in the multi bounded model, the variable $LBID1$ is also in the set of

explanatory variables. Rearranging Equation 4.6, we have

$$\Phi^{-1}(P\{YES_i = 1\}) = \beta_0 + \alpha_1 LBID_i + \alpha_2 LBID1_i + \sum \beta_j x_{ij} + \epsilon_i, \tag{4.7}$$

where α_1 and α_2 represent the two effects which are combined in Equation 1 as α. As we have

$$|cor(LBID, LBID1)| < 0.0001,$$

the multi bounded model does not suffer from the problem of collinearity.

Equation 4.4 is also estimated by LR, HTLR, and HFLR methods. Substituting $P\{YES = 1\} = \frac{1}{2}$ in the estimation, we have

$$0 = \hat{\beta}_0 + \hat{\alpha}_1 \hat{Q}_{\frac{1}{2}}(\ln WTP_i) + \hat{\alpha_2} LBID1_i + \sum \hat{\beta}_j x_{ij}. \tag{4.8}$$

Rearranging Equation 4.8, we calculate the estimated median of WTP by Equation 9:

$$\hat{Q}_{\frac{1}{2}}(WTP_i) = f(BID1_i) = \exp\{-\frac{1}{\hat{\alpha}_1}(\hat{\beta}_0 + \sum \hat{\beta}_j x_{ij})\} \times BID1_i^{-\frac{\hat{\alpha}_2}{\hat{\alpha}_1}}, \tag{4.9}$$

where $BID1 = \exp\{LBID1\}$. From Equation 4.9, we conclude that the when $BID1$ is doubled, the median of WTP is estimated to increase by the proportion $(2^{-\frac{\hat{\alpha_2}}{\hat{\alpha_1}}} - 1)$. The influences of control variables to the estimated median of WTP is similar to those in Equation 4.3.

In Equation 4.6, the estimated median of WTP is a function of $BID1$. The fixed

point of the function can be given by

$$\exp\{-\frac{1}{\hat{\alpha}_1 + \hat{\alpha}_2}(\hat{\beta}_0 + \sum \hat{\beta}_j x_{ij})\}.$$

The meaning of the fixed point is, if the value was given as a bid in the first round, the estimated median of WTP would be the same as the bid. The fixed point contains information about the participants' WTP which are uninterrupted by the bids given to them.

4.4 Results and Discussion

4.4.1 Results of the Single Bounded Model

The estimated single bounded model by three estimation methods are given in Table 4.2. Definitions of interaction terms generated by decision trees are given in Table 4.3.

Table 4.2: Estimates of the single bounded model

Variable	LR	HTLR	HFLR
Intercept	1.316** (0.448)	1.558*** (0.396)	0.379 (0.700)
$LBID1$	-0.486*** (0.040)	-0.467*** (0.042)	-0.466*** (0.044)
$LFINC$			0.136** (0.048)
$STMOR$	0.186** (0.058)	0.230*** (0.061)	0.325*** (0.070)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 4.2: Estimates of the single bounded model (continued)

Variable	LR	HTLR	HFLR
$BENEFIT$	0.319*** (0.062)		
$RIGHT$	-0.170*** (0.049)		
$NOCONT$		0.110* (0.040)	
$PROTLAW$	0.148** (0.052)		
$TMEQL$	-0.198*** (0.053)		
$M1T1V3$		-1.702*** (0.263)	-1.705*** (0.287)
$M1T1V6$		-1.497*** (0.242)	-1.418*** (0.267)
$M1T1V10$		-1.404** (0.446)	
$M1T1V11$		-1.992*** (0.567)	-2.043** (0.663)
$M1T1V12$		2.991*** (0.771)	2.686*** (0.752)
$M1T1V14$		-1.909** (0.666)	
$M1T2V7$			-0.573*** (0.174)
$M1T2V11$			-0.685** (0.255)
$M1T2V15$			-0.990** (0.350)
$M1T3V2$			-1.451*** (0.331)
$M1T3V11$			-1.853*** (0.427)
$M1T4V12$			-0.999** (0.340)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 4.2: Estimates of the single bounded model (continued)

Variable	LR	HTLR	HFLR
$M1T5V4$			-1.707*** (0.370)
$M1T5V12$			-2.363*** (0.607)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 4.3: Variables generated by the decision trees in the single bounded model

Variable V	Condition for $V = 1$	Mean (SD)
$M1T1V3$	$(BENEFIT \leq 3) \vee (RIMIG \geq 5) \vee (TOLD \geq 2)$	0.0521 (0.2223)
$M1T1V6$	$(BENEFIT \geq 4) \vee (TMEQL \geq 4) \vee (GWMANMD \leq 5) \vee (RRACHAT \geq 10)$	0.0508 (0.2197)
$M1T1V10$	$(BENEFIT \geq 4) \vee (TMEQL \leq 3) \vee (RDRUG \leq 4) \vee (NOCONT \geq 3)$	0.0156 (0.1241)
$M1T1V11$	$(BENEFIT \geq 4) \vee (TMEQL \leq 3) \vee (RDRUG \geq 5) \vee (RLMED \leq 2)$	0.0690 (0.2536)
$M1T1V12$	$(M1T1V13 = 1) \vee (RCHEMA \leq 9)$	0.0586 (0.2350)
$M1T1V14$	$(M1T1V14 = 1) \vee (GOVGOAL \leq 2) \vee (PROTLAW \leq 4)$	0.0169 (0.1291)
$M1T2V7$	$(STMOR \geq 5) \vee (RIGHT \geq 6)$	0.2331 (0.4231)

Table 4.3: Variables generated by the decision trees in the single bounded model (continued)

Variable	Formula	Mean (SD)
$M1T2V11$	$(STMOR \geq 5) \vee (RIGHT \geq 5) \vee (RIMIG \leq 3) \vee (GOVGOAL \leq 2)$	0.0833 (0.2766)
$M1T2V15$	$(STMOR \geq 5) \vee (RIGHT \geq 5) \vee (RIMIG \leq 3) \vee (GOVGOAL \geq 3) \vee (TOLD \geq 5) \vee (RRACHAT \geq 6)$	0.0404 (0.1969)
$M1T3V2$	$(RGMFOOD \geq 10) \vee (RCHEMA \leq 8)$	0.0378 (0.1907)
$M1T3V11$	$(RGMFOOD \leq 9) \vee (RDRUG \leq 9) \vee (STBIAS \geq 3) \vee (RGOVBUD \geq 4) \vee (NEWS \geq 3) \vee (PROTNG \geq 3)$	0.0260 (0.1594)
$M1T4V12$	$(GOVGOAL \geq 4) \vee (RLMED \leq 2) \vee (RIMIG \geq 3)$	0.0469 (0.2115)
$M1T5V4$	$(RDRUG \geq 10) \vee (2 \geq RGOVBUD \geq 6) \vee (STBIAS \leq 2)$	0.0326 (0.1776)
$M1T5V12$	$(RDRUG \geq 10) \vee (AGE \geq 5) \vee (AUTH \leq 1) \vee (SOFSOC \geq 4) \vee (RIMIG \leq 6)$	0.0156 (0.1241)

Substituting the means of all control variables, the estimated median of WTP given by LR, HTLR, and HFLR single bounded models are 552.78, 535.77, and 693.90. The closest given bid to these estimates is 476, which is consistent with the

pattern discovered in Figure 4.1(a).

4.4.2 Economic Discussion based on the Single Bounded Model

From the results in Table 4.2, the values of $\hat{\alpha}$ are always negative, which is consistent with the fact that when the given bid has a larger value, it is more likely to be larger than the WTP of a participant. However, the single bounded model cannot figure out the psychological effect of the given bid to the participant.

The control variable $STMOR$ appears in all estimates as positive significant linear terms, suggesting that participants have higher WTP values when they believe that more studies should be done before their findings are relied by policymakers. In the HFLR estimate, an interaction term $(STMOR \geq 5) \vee (RIGHT \geq 6)$ is negative significant, suggesting that this relationship may be smaller for participants who have strong belief in the independence of individual rights.

Similarly, the control variable $BENEFIT$ is a positive significant linear term in the LR estimate, while in HTLR and HFLR estimates, it is replaced by a few interaction terms generated by decision trees. This means that in general, participants have higher WTP values when they believe that the geo-engineering project is beneficial; however, this effect depends upon the participants' opinions on other topics, such as immigration and equal rights.

Table 4.4 features summary statistics and the values of certain model selection criteria of estimated single bounded models, and Figure 4.2 shows the change of model efficiency during the HFLR algorithm. According to these results, the HFLR

estimate had both the smallest BIC and the highest prediction accuracy. Furthermore, the changes of both BIC and prediction accuracy show that the estimate of the model is optimised gradually when more trees are added in the HFLR algorithm.

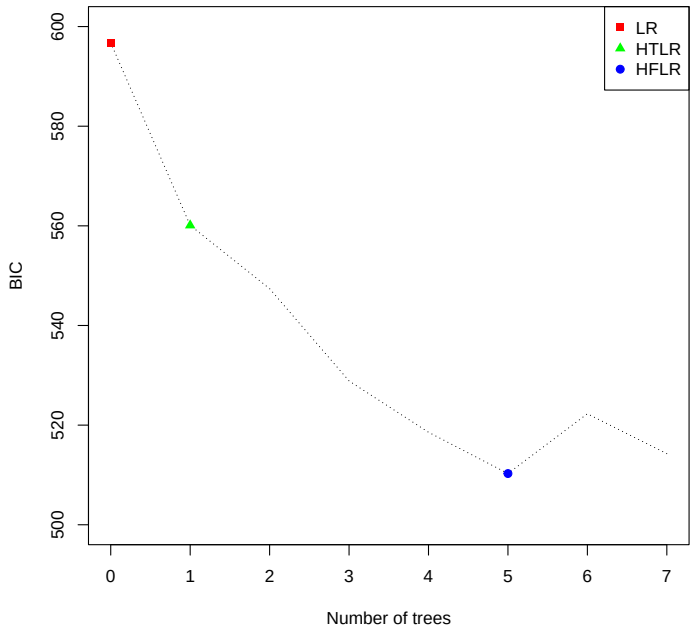
Table 4.4: Comparison between different estimates of the single bounded model

Method	LR	HTLR	HFLR
Amount of trees	0	1	8
Amount of parameters	7	9	22
BIC	601.34	568.00	489.21*
Prediction accuracy (comparison with $BID1$)	86.25%	84.96%	89.20%*

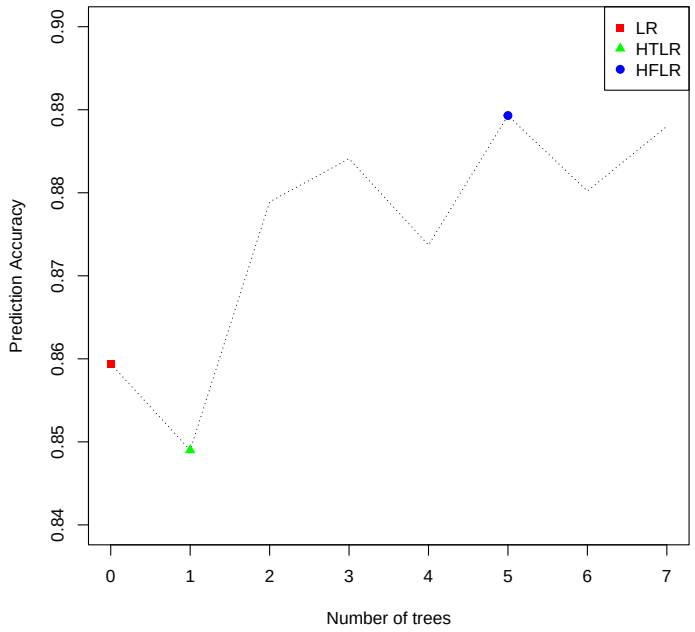
* represents the selected model by each model selection criteria.

As the HFLR estimate is better than other estimates in both explanation and prediction powers, we only applied its results in further discussions. Notably, there is no linear terms except $LBID1$ in the HFLR estimate. The 95% confidence interval of α_1 is (-0.665,-0.449), which means that when the given bid in the is doubled, the value of $\Phi^{-1}(P\{YES1 = 1\})$ is decreased by between 0.31 and 0.46.

It is observed from Figure 4.2(a) that the first and third decision trees bring the largest decrease of BIC during the HFLR algorithm, and the second and fifth decision trees bring the largest increase of prediction accuracy. These decision trees are displayed in Figure 4.3. Apart from $STMOR$ and $BENEFIT$ which were discussed above, $GWRISK$ and $RDRUG$ are also variables on root nodes of key decision trees, which means that the opinions of participants to the risks of global warming and drugs also have strong effects on their WTP values, although these relationships are dependent to their opinions on other topics.

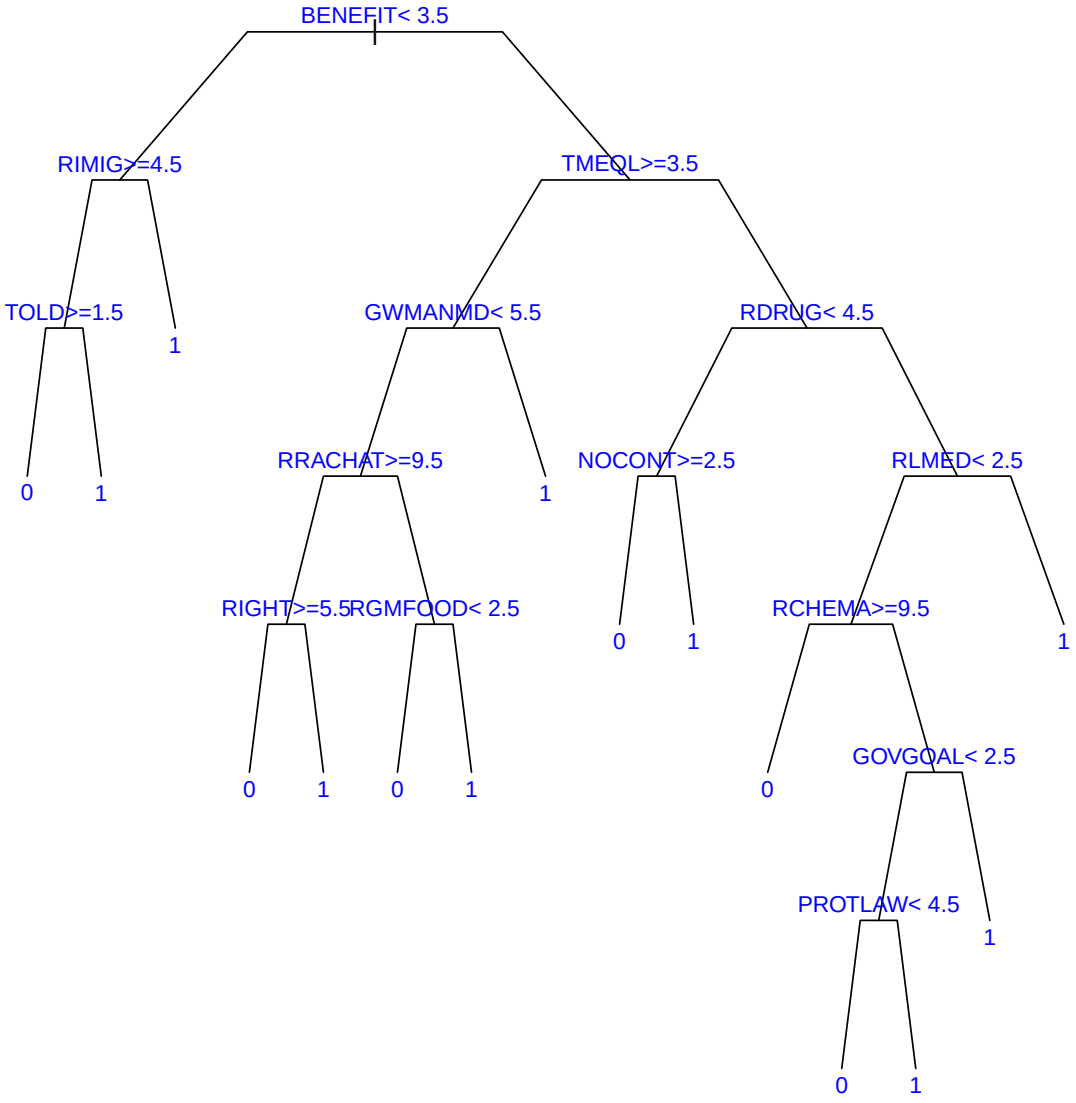


(a) BIC

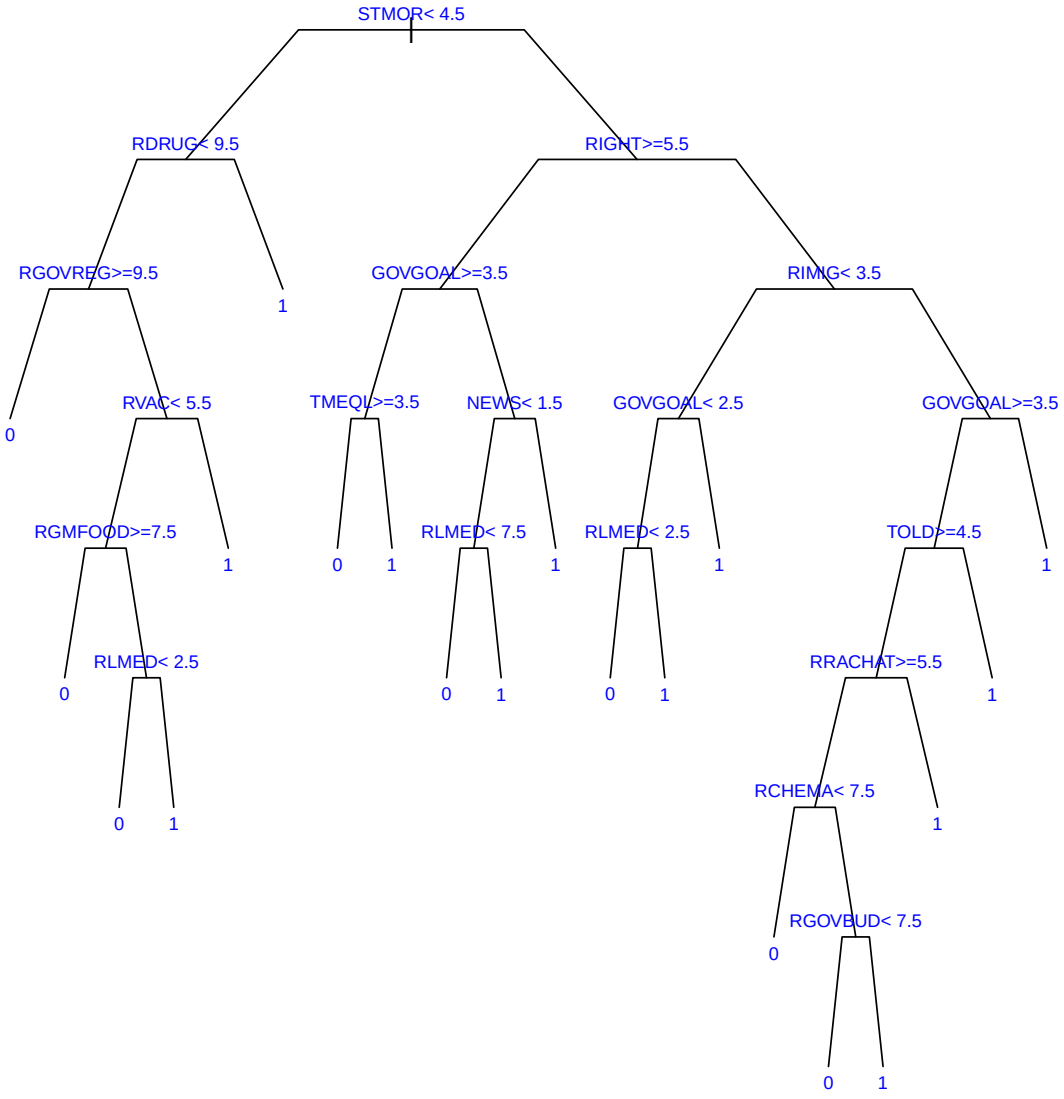


(b) Prediction Accuracy

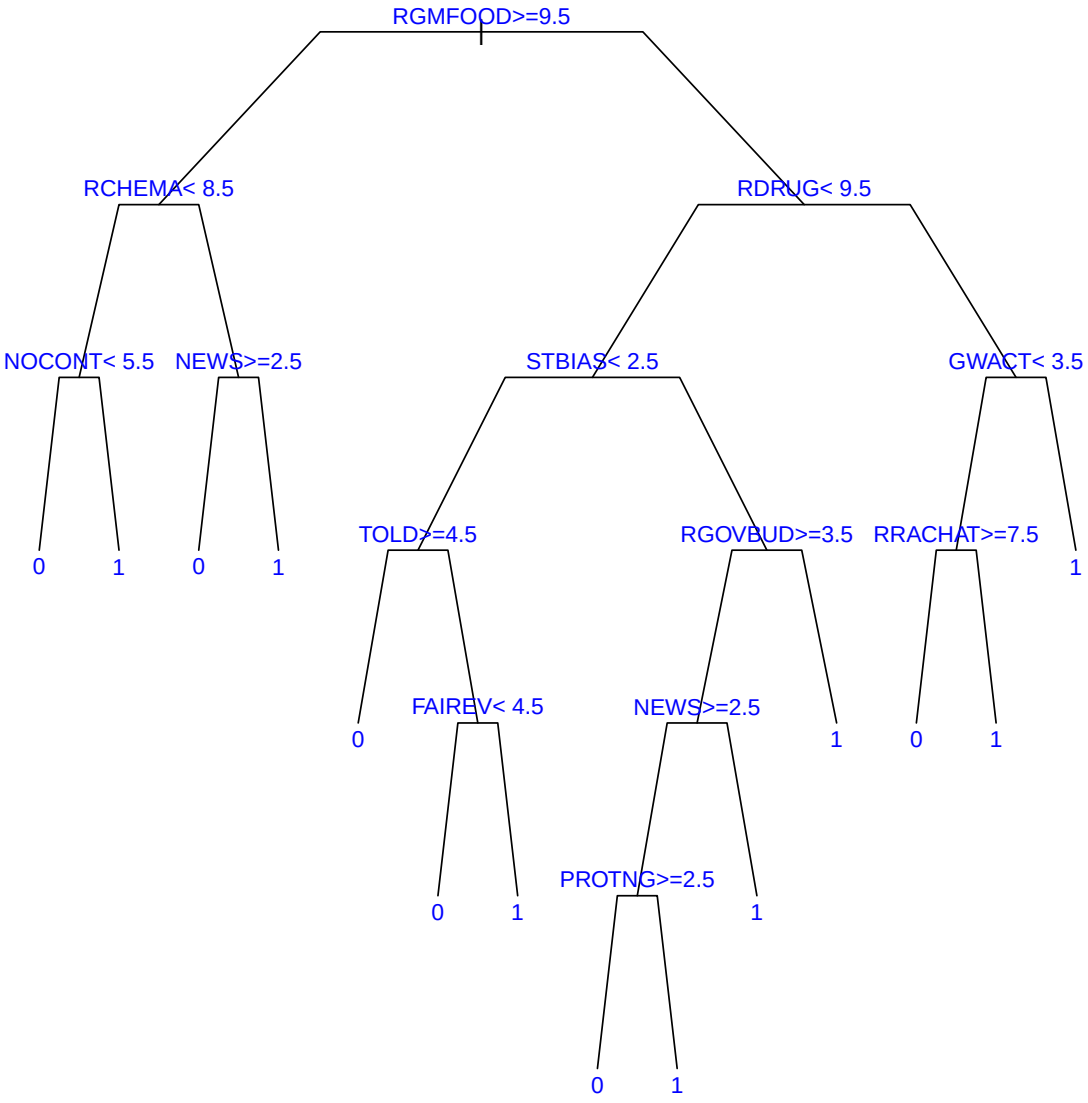
Figure 4.2: Change of Model Efficiency during the HFLR Algorithm (Single Bounded Model)



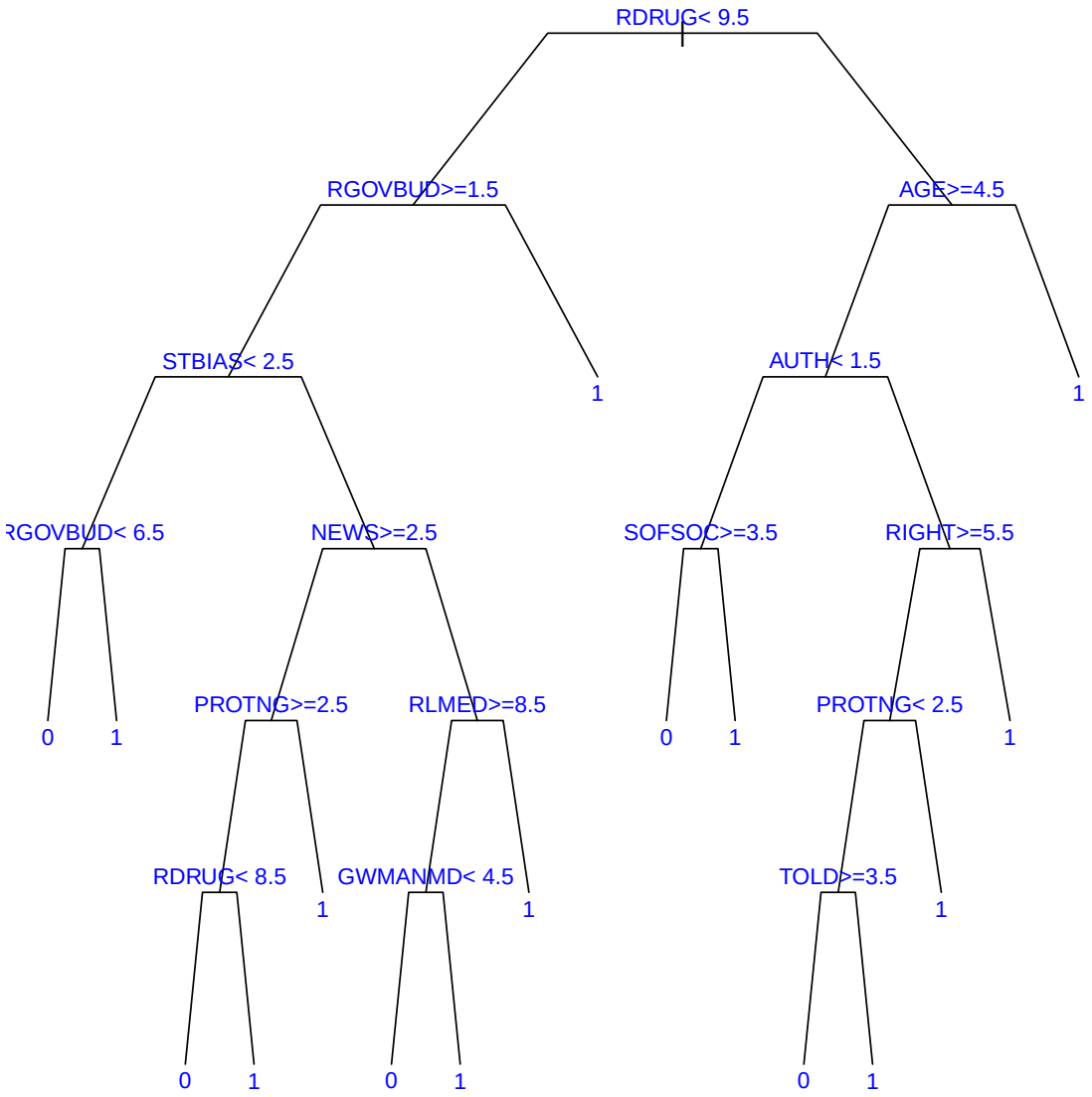
(a) The First Tree



(b) The Second Tree



(c) The Third Tree



(d) The Fifth Tree

Figure 4.3: Key Decision Trees Generated in the HFLR Algorithm (Single Bounded Model)

4.4.3 Results of the Multi Bounded Model

The estimated multi bounded model by three estimation methods are given in Table 4.5. Definitions of interaction terms generated by decision trees are given in Table 4.6.

Table 4.5: Estimates of the multi bounded model

Variable	LR	HTLR	HFLR
Intercept	0.873** (0.306)	1.174*** (0.292)	1.975*** (0.270)
$LBID$	-0.571*** (0.013)	-0.580*** (0.014)	-0.613*** (0.014)
$LBID1$	0.066*** (0.011)	0.067*** (0.011)	0.072*** (0.012)
$LFINC$	0.069*** (0.013)	0.064*** (0.013)	0.076*** (0.014)
$STBIAS$	-0.065*** (0.018)		
$STMOR$	0.088*** (0.022)	0.101*** (0.021)	
$GWACT$	0.102*** (0.024)		
$BENEFIT$	0.221*** (0.022)	0.224*** (0.021)	0.195*** (0.024)
$RIGHT$	-0.074*** (0.016)	-0.072*** (0.016)	-0.058*** (0.016)
$FAIREV$	-0.051** (0.017)	-0.059*** (0.017)	0.079*** (0.017)
$TOLD$	-0.044*** (0.013)	-0.046*** (0.013)	
$AUTH$	0.083*** (0.015)	0.088*** (0.015)	0.062*** (0.016)
$NOCONT$			0.056*** (0.014)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 4.5: Estimates of the multi bounded model (continued)

Variable	LR	HTLR	HFLR
$PROTLAW$	0.125*** (0.018)	0.131*** (0.018)	
$PROTNG$	-0.062*** (0.017)	-0.146*** (0.022)	
$TMEQL$	-0.108*** (0.019)	-0.109*** (0.019)	
$SOFSOC$	-0.079*** (0.015)	-0.084*** (0.015)	
$RCHEMA$	-0.059*** (0.009)	-0.045*** (0.010)	-0.041*** (0.010)
$RLMED$	-0.031*** (0.008)	-0.031*** (0.008)	
$RDRUG$	0.068*** (0.012)	0.073*** (0.012)	
$NEWS$	-0.086*** (0.024)	-0.080*** (0.024)	-0.082*** (0.025)
$DOCT$	0.209*** (0.047)	0.246*** (0.047)	
$ANTIAGE$	-0.066*** (0.020)	-0.067*** (0.020)	
$M2T1V2$		1.037*** (0.159)	0.484** (0.162)
$M2T1V3$		-1.007*** (0.163)	-0.858*** (0.168)
$M2T1V4$		0.558*** (0.076)	0.493*** (0.079)
$M2T1V5$		-0.368*** (0.095)	-0.352*** (0.102)
$M2T2V6$			-1.034*** (0.211)
$M2T2V7$			1.450*** (0.213)
$M2T2V8$			-0.567*** (0.107)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 4.5: Estimates of the multi bounded model (continued)

Variable	LR	HTLR	HFLR
$M2T3V1$			-0.607*** (0.081)
$M2T3V2$			0.635*** (0.102)
$M2T4V1$			-0.577*** (0.083)
$M2T4V2$			0.857*** (0.114)
$M2T4V3$			-1.303*** (0.140)
$M2T4V4$			0.556*** (0.126)
$M2T4V7$			-0.499*** (0.075)
$M2T4V9$			0.354*** (0.073)
$M2T5V2$			-0.505*** (0.087)
$M2T6V1$			0.343*** (0.073)
$M2T6V3$			-0.431*** (0.062)

Significance level: ".": 0.1; "*": 0.05; "**": 0.01; "***": 0.001.

Table 4.6: Variables generated by the decision trees in the multi bounded model

Variable V	Condition for $V = 1$	Mean (SD)
$M2T1V2$	$(PROTNG \geq 3) \vee (GWACT \geq 5)$	0.3854 (0.4870)
$M2T1V3$	$(M2T1V2 = 1) \vee (RCHEMA \geq 4)$	0.3646 (0.4816)
$M2T1V4$	$(M2T1V3 = 1) \vee (GWRISK \geq 6)$	0.1966 (0.3977)

Table 4.6: Variables generated by the decision trees in the multi bounded model (continued)

Variable V	Condition for $V = 1$	Mean (SD)
$M2T1V5$	$(M2T1V4 = 1) \vee (STBIAS \geq 4)$	0.0716 (0.2580)
$M2T2V6$	$(BENEFIT \geq 5) \vee (FAIREV \geq 4) \vee (RCHEMA \geq 4) \vee (PROTLAW \geq 3) \vee (SOFSOC \geq 4)$	0.2031 (0.4026)
$M2T2V7$	$(M2T2V6 = 1) \vee (RDRUG \geq 7)$	0.1901 (0.3926)
$M2T2V8$	$(M2T2V7 = 1) \vee (STBIAS \geq 4)$	0.0651 (0.2469)
$M2T3V1$	$STMOR \leq 4$	0.1771 (0.3820)
$M2T3V2$	$(M2T3V1 = 1) \vee (RDRUG \geq 10)$	0.0885 (0.2843)
$M2T4V1$	$GOVGOAL \geq 4$	0.3177 (0.4659)
$M2T4V2$	$(M2T4V1 = 1) \vee (GWACT \geq 6)$	0.1797 (0.3842)
$M2T4V3$	$(M2T4V2 = 1) \vee (TMEQL \geq 3)$	0.1276 (0.3339)
$M2T4V4$	$(M2T4V4 = 1) \vee (GWMANMD \geq 5)$	0.0859 (0.2805)
$M2T4V7$	$(GOVGOAL \leq 3) \vee (RCHEMA \geq 4) \vee (DOCT = 0)$	0.4518 (0.4980)
$M2T4V8$	$(M2T4V7 = 1) \vee (PROTLAW \geq 4)$	0.3151 (0.4649)
$M2T4V9$	$(M2T4V8 = 1) \vee (AUTH \leq 1)$	0.0326 (0.1776)
$M2T5V2$	$(GWACT \leq 4) \vee (RIMIG \geq 6)$	0.0938 (0.2917)
$M2T6V1$	$PROTLAW \geq 3$	0.8789 (0.3264)

Table 4.6: Variables generated by the decision trees in the multi bounded model (continued)

Variable V	Condition for $V = 1$	Mean (SD)
$M2T6V3$	$(M2T6V1 = 1) \vee (SOFSOC \geq 4) \vee (RLMED \geq 5)$	0.2865 (0.4524)

Substituting the log average family income $\ln(exp\{\bar{LFINC}\})$ and the means of all other control variables, the estimated medians of WTP given by LR, HTLR, and HFLR models are shown in Equations 4.10, 4.11, and 4.12:

$$\widehat{Q_{\frac{1}{2}}}(WTP) = 137.73 \times BID1^{0.1153}, \tag{4.10}$$

$$\widehat{Q_{\frac{1}{2}}}(WTP) = 134.11 \times BID1^{0.1163}, \tag{4.11}$$

$$\widehat{Q_{\frac{1}{2}}}(WTP) = 136.46 \times BID1^{0.1182}. \tag{4.12}$$

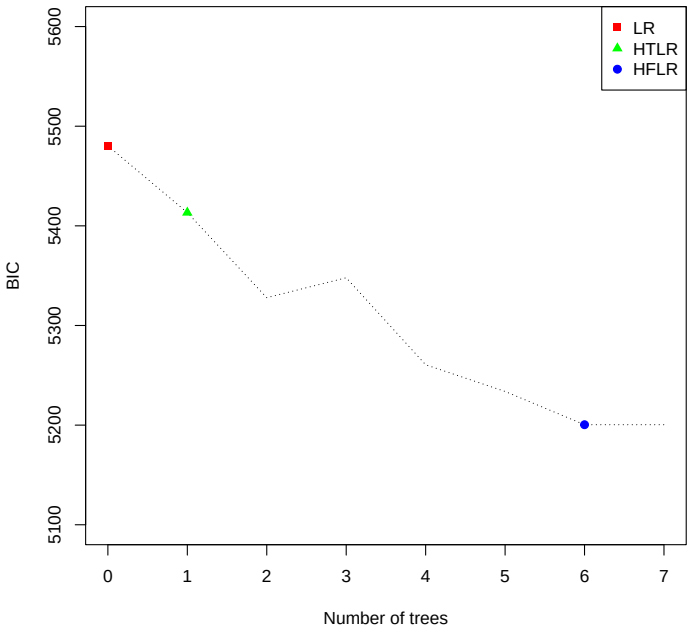
Substituting the log average first bid $\ln(\overline{BID1})$, the respective estimated medians of WTP given by LR, HTLR, and HFLR multi bounded model are 261.38, 255.60, and 263.72. Comparatively, the fixed points of the equations are 261.73, 255.96, and 263.12. The closest given bid to all of these estimates is 285, which is consistent with the pattern discovered in Figure 4.1(b). Notably, compared to the change of estimated medians of WTP in the single bounded model, the estimates of the multi bounded model were much more consistent.

4.4.4 Economic Discussion based on the Multi Bounded Model

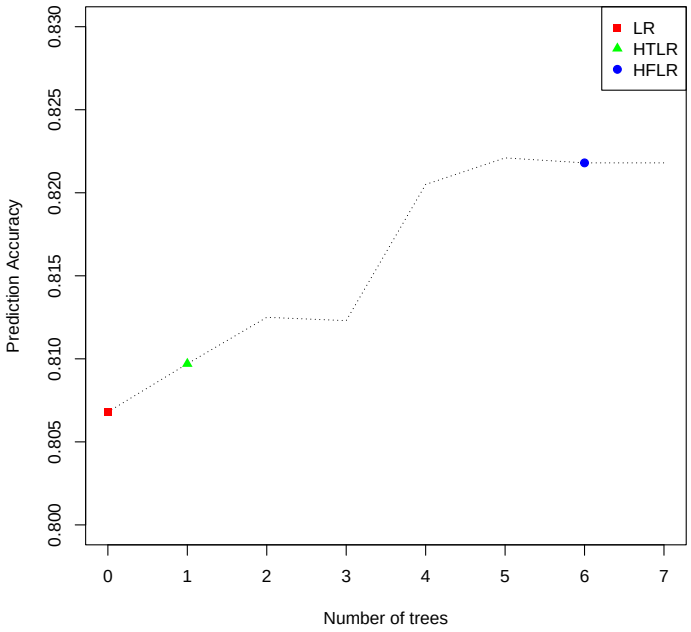
From the results in Table 4.5, the signs of $\hat{\alpha}_1$ are always negative, which is consistent with the fact that when the given bid has a larger value, it is more likely to be larger than the WTP of a participant. Meanwhile, the signs of $\hat{\alpha}_2$ is always positive, which suggest that an increase in the bid given to a participant would lead to an increase in the reported WTP of the participant. The reason of this phenomenon is probably because of positive psychological suggestions.

Notably, estimates of $(\alpha_1 + \alpha_2)$ given by LR, HTLR, and HFLR methods of the multi bounded model are -0.505, -0.513, and -0.541. Compared to $\hat{\alpha_1}$, they are closer to $\hat{\alpha}$ given by the single bounded model (-0.486, -0.467, and -0.466). This phenomenon shows that both effects of BID and $BID1$ to WTP values are combined in one term in the single bounded model. Based on these results, we concluded that the independence between median of WTP values and the given bid, which is a basic assumption of the single bounded model, is probably unreal.

Table 4.7 shows summary statistics and values of certain model selection criteria of estimated multi bounded models, and Figure 4.4 shows the change of model efficiency during the HFLR algorithm. Three different definitions of prediction accuracy are applied in Table 4.4: the average prediction accuracy for comparing WTP with a given bid; the probability that the predicted interval of WTP is accurately the real one; and the probability that the predicted interval of WTP is the real one or a neighbour of the real one. In Figure 4.4(b), the first definition of prediction accuracy is applied.

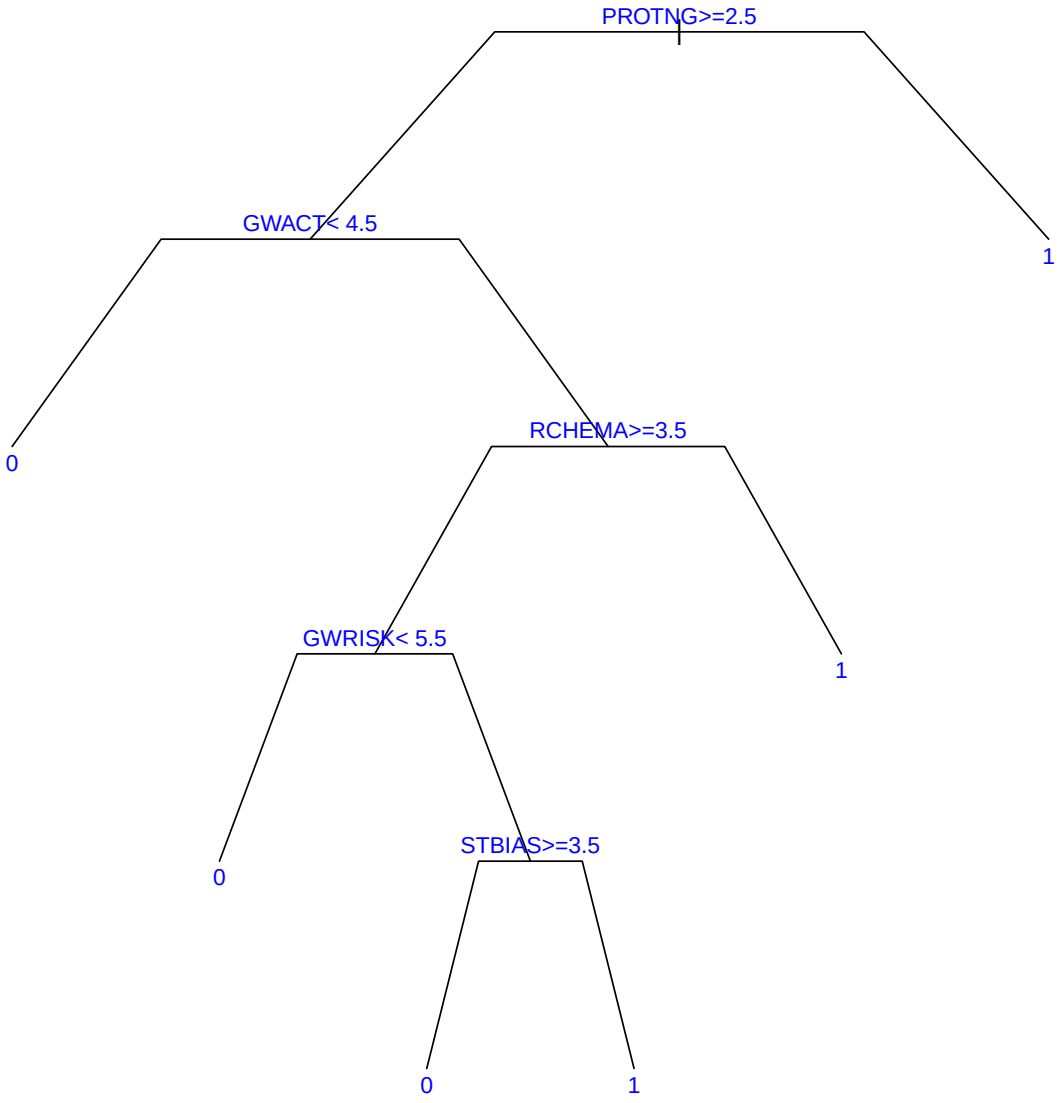


(a) BIC

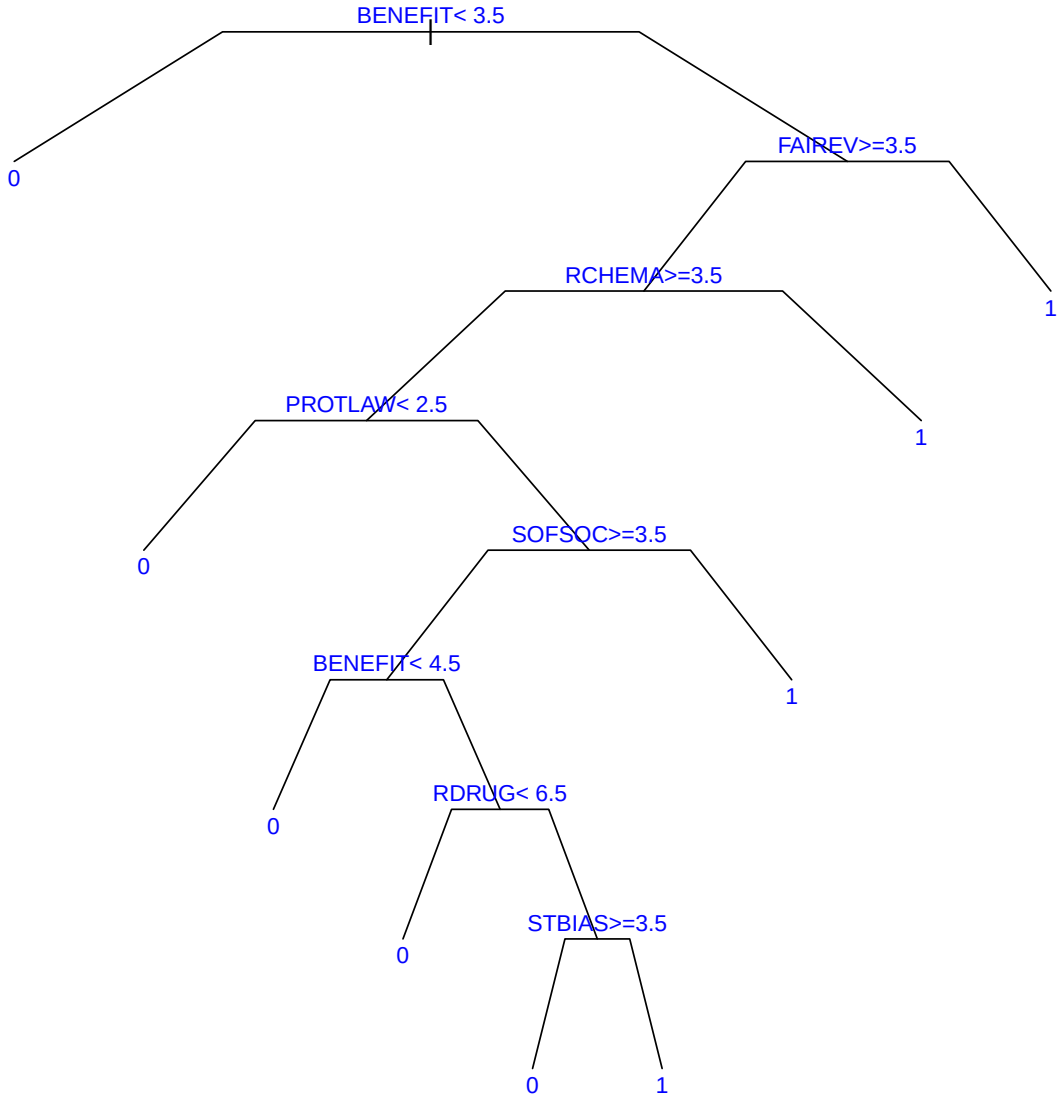


(b) Prediction Accuracy

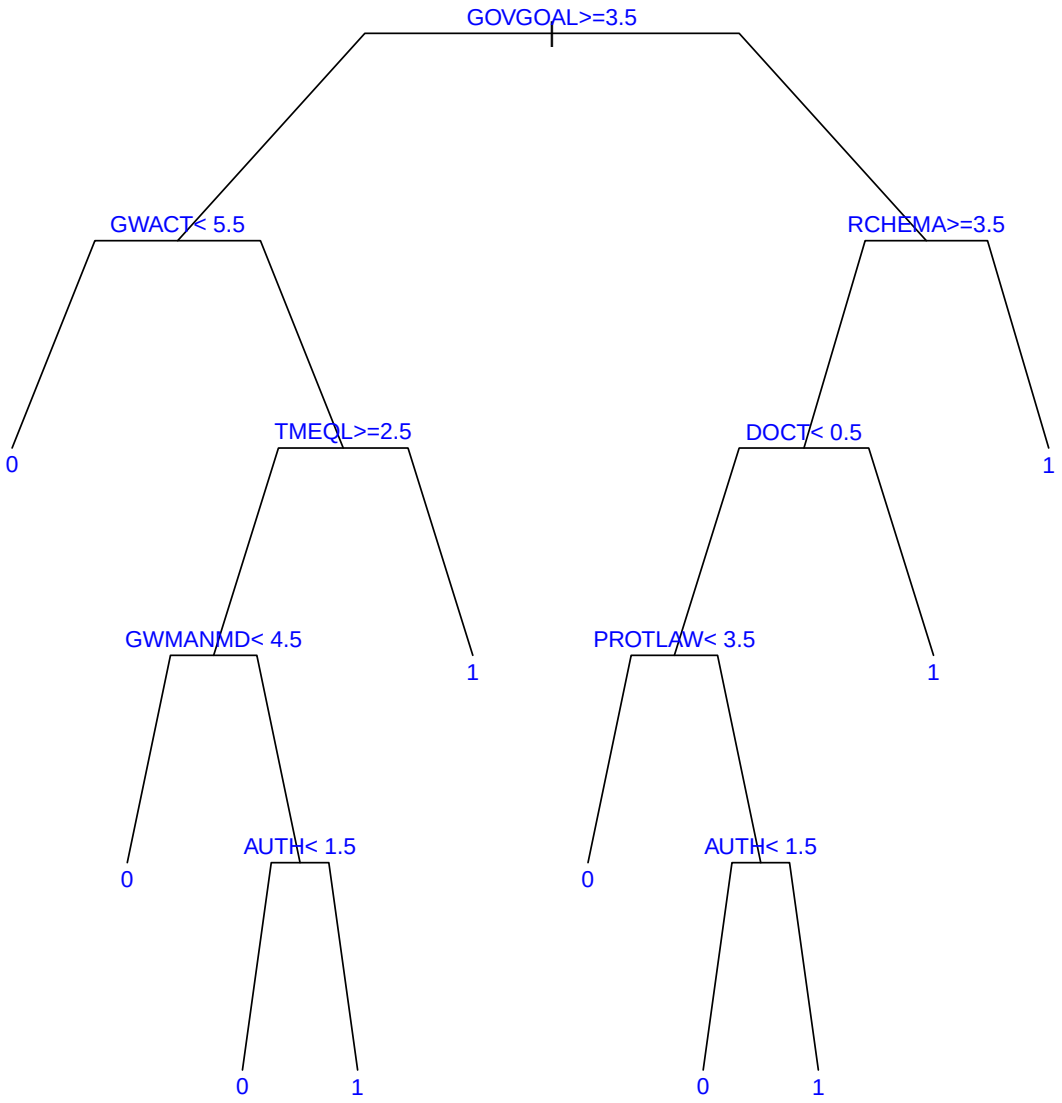
Figure 4.4: Change of Model Efficiency during the HFLR Algorithm (Multi Bounded Model)



(a) The First Tree



(b) The Second Tree



(c) The Fourth Tree

Figure 4.5: Key Decision Trees Generated in the HFLR Algorithm (Multi Bounded Model)

Table 4.7: Comparison between different estimates of the multi bounded model

Method	LR	HTLR	HFLR
Amount of trees	0	1	6
Amount of parameters	22	24	30
BIC	5480.3	5413.3	5200.4*
Prediction accuracy (comparison with BID)	80.68%	80.97%	82.17%*
Prediction accuracy (accurate interval)	34.11%*	34.11%*	33.72%
Prediction accuracy (neighbouring intervals)	58.20%	56.38%	58.72%*

* represents the selected model by each model selection criteria.

According to the results in Table 4.7, the HFLR estimate has both the smallest BIC and the highest prediction accuracy, according to two out of three definitions of prediction accuracy. Furthermore, according to Figure 4.4, the changes of both BIC and prediction accuracy show that the estimate of the multi bounded model is also optimised gradually when more trees are added in the HFLR algorithm. As the HFLR estimate is better than other estimates in both explanation and prediction powers, we are only utilising its results in further discussions.

From the estimated power of $BID1$ in Equation 4.12, we could estimate that when $BID1$ is doubled, the estimated median of WTP is increased by around 5.1%. From the results in Table 4.5, we could also calculate that according to the HFLR estimate, when family incomes of participants are doubled, the estimated median of WTP is increased by only around 5.4%. Notably, the coefficient of $LBID1$ is significantly larger than that of $LFINC$, suggesting that the first given bid is almost

as influential as family income to the WTP of participants.

The linear terms $BENEFIT$, $FAIREV$, $AUTH$, and $NOCONT$ are positive significant in the HFLR estimates. From these results, we concluded that people tended to have larger WTP values if they were in favour of a fairness revolution and obeying authorities, and if they believe that the geo-engineering project is beneficial and courses of lives are determined by factors beyond their control. On the other hand, the linear terms $RIGHT$, $RCHEMA$, and $NEWS$ are negative significant. From these results, we concluded that people tended to have smaller WTP values if they agreed that individual rights are independent on opinions of others and chemical additives in food are risky. Participants who are more interested in news and public affairs also tended to have smaller WTP values.

Out of thirty terms included in the HFLR estimate, eighteen are generated from six decision trees. It is observed from Figure 4.4 that the first, second, and fourth decision trees bring the largest decreases of BIC and the largest increases of prediction accuracy during the HFLR algorithm. These decision trees are displayed in Figure 4.5. The root node of these decision trees are $PROTNG$, $BENEFIT$, and $GOVGOAL$. This means that the influences to WTP of the participants' attitude towards protections provided by governments, the benefit of the geo-engineering project, and social goals of government, highly depend on other factors.

4.4.5 Comparison between the Models

To compare the prediction power of the multi model and the single bounded model, we calculated the prediction accuracy for estimates of the single bounded

model. All three indicators of prediction accuracy as defined in Table 4.7 are applied. The results are given in Table 4.8.

Table 4.8: Prediction accuracy criteria of estimates of the single bounded model

Method	LR	HTLR	HFLR
Prediction accuracy (comparison with BID)	78.61%	78.30%	79.88%
Prediction accuracy (accurate interval)	34.11%	33.33%	33.46%
Prediction accuracy (neighbouring intervals)	55.08%	56.90%	58.20%

Based on the results in Tables 4.7 and 4.8, we concluded that regardless of estimation methods and definitions of prediction accuracy, the multi bounded model always had higher prediction efficiency than the single bounded model. When the HFLR estimation method is applied, the gap between prediction efficiency between the multi bounded model and the single bounded model is larger. Based on the first and definition of prediction accuracy, the gap is estimated to be 2%-3%.

4.5 Conclusion

In this study, we applied single bounded and multi bounded probit regression models to explore the factors influencing the people's WTP to a geo-engineering project. By comparing their prediction accuracy, information utilisation, and economic meaning, we concluded that the multi bounded model is better than the single bounded model.

From results of the multi bounded model, we demonstrated that the first bids given to participants had strong influences upon their final WTP value in the survey. We also detected the relationship between WTP of participants and their family income, political and social views, and other personal characters. Notably, many of these relationships are nonlinear.

Since decision tree learning algorithms were first developed, researchers suggested that hybrid methods between classic regression models and decision trees were able to combine the strengths of both models to achieve higher explanation and prediction powers. This paper applied the HFLR algorithm developed in Chapter 3 to multi-level categorical variables. By comparing various statistical criteria, we observed that HFLR estimates of both single bounded and multi bounded models had higher explanatory and prediction powers than their LR and HTLR counterparts.

Similar to Chapter 3, the reason for HFLR estimates outperforming LR and HTLR estimates is related to the inclusion of nonlinear terms. Some of these terms have clear meanings in the WTP model. As an example, in the LR multi bounded model, the effect of recent visits to doctors is considered always positive to the participants' WTP, whereas in HFLR multi bounded model, it is discovered that recent visits to doctors may have either positive or negative effects, depending on the participants' political and social views.

Chapter 5

Conclusions and Further Research

5.1 Chapter Summaries

Following the hybrid classification method of logistic regression and decision trees given by Stainberg et al. (1998), various hybrid tree-regression methods were designed to combine the ability of classic regression models in dealing with continuous relationship and the strength of decision trees in figuring joint effects and dealing with outliers. This thesis introduced two new approaches of hybrid tree-regression methods: hybrid groupwise tree-linear regression (HGTLR), and hybrid forest-linear regression (HFLR). These new approaches were compared with linear regression (LR) and hybrid tree-linear regression (HTLR) which followed the approach of Stainberg et al. (1998). This research applied data from three different research fields, which were analysed in Chapters 2, 3, and 4.

Chapter 2 showed the ability of hybrid tree-regression methods in improving prediction power of regression models where all independent variables are binary. Based on 452 models gathered from sixty-five papers about the relationship between trade openness and economic growth, we applied LR, HTLR, and HGTLR to build random-effects and weighted least squares (WLS) meta-regression models. All applied model comparison criteria, including AIC, BIC, cross-validated MSE, and prediction accuracy, suggested that tree-based regression models (HTLR and HGTLR) have higher explanation and prediction powers than their linear counterparts. However, it is debatable where HGTLR is better than HTLR. Additionally, we verified the conclusion of Stanley et al. (2017) that WLS meta-regression models have higher prediction power than random-effects meta-regression models. From the predictions of both WLS-HTLR and WLS-HGTLR models, although the effect sizes of trade openness to economic growth are influenced by publication time, focused continent, and control variables of econometric studies, it is positive in most cases.

Chapter 3 showed that hybrid tree-regression methods, especially the newly developed HFLR method, improve prediction powers of regression models with both binary and continuous variables, although continuous variables were not applied to construct decision trees. We applied LR, HTLR, and HFLR to analyse factor that influences fundraising performances of reward-based crowdfunding projects based on 236 projects collected from Kickstarter and Indiegogo. By comparing BIC, cross-validated MSE, and prediction accuracy of their results, it was verified not only that the HTLR method is more powerful than LR in explanation and prediction, but also that the HFLR method is more powerful than HTLR in these aspects. These statements were also verified by applying those algorithms to 12 simulated prediction problems. By analysing the included interaction terms in hybrid tree-regression

models, they are more efficient because of their ability of discussing joint effects and dealing with outliers by adding interaction terms. From the result of the HFLR model, it is clear than variables such as fundraising target, rewards, and visualisation tools, have linear and joint influences on the fundraising performances.

Chapter 4 showed that the ability of decision trees of dealing with not only binary variables, but also non-binary ordinal variables. We applied LR, HTLR, and HFLR to predict the range of WTP of 778 participants from China to a geo-engineering project. Both single bounded and multi bounded models were built during the modelling process. Hybrid tree-regression methods showed their abilities of modelling with multi level categorical variables. In both models, HFLR estimates have smaller BIC and larger prediction accuracy than LR and HTLR estimates. Meanwhile, all estimates of the multi bounded model make better predictions compared to any estimates of the single bounded model. The probable reason of the better prediction performance of the multi bounded model is that it applied more information from the data. From the result of the multi bounded HFLR model, it is concluded that the given bid to participants has strong influences upon their resulting WTP. Meanwhile, WTP of participants are also influenced by their family income, political and social views, and other personal characters. The median WTP is estimated to be slightly lower than 285 which is one of the given bids.

5.2 Further Research

The results of this thesis have showed that hybrid tree-regression methods, especially the newly designed HFLR method, enable stronger prediction power than

classic regression models. Compared to LR, HTLR, and HGTLR methods, the HFLR method applies multi decision trees with different root nodes to the whole dataset, which allows for the consideration of various interaction terms without a common variable on the root node.

However, existing hybrid tree-regression methods have their limits. Firstly, methods applied in this thesis applied decision trees to categorical regressors only. Secondly, the meanings of some terms generated by decision trees are relatively hard to explain. Thirdly, as the method with the highest prediction power in this thesis, the HFLR algorithm has a relatively low time efficiency. These limits of existing hybrid tree-regression methods suggest several future research topics of applying multi decision trees to optimise prediction performances of regression models.

In this thesis, decision trees built during HTLR, HGTLR, and HFLR algorithms were only applying categorical variables. In future research, following the attempt of Dumitrescu et al. (2018), the application of decision trees in regression modelling can be broadened by building regression trees with not only categorical regressors, but also continuous regressors.

Additionaly, as discussed further within Chapter 3, some interaction terms added to regression models are explainable, while the meanings of some other terms are relatively unclear. In future research, instead of building decision trees based on groups of observations such as the HGTLR method, we could also build decision trees based on groups of variables in order to figure out more explainable joint effects.

Another inspiration of future research comes from machine learning methods such

as random forest and lasso. Although the HFLR method make better predictions than LR and HTLR, constructing a HFLR model takes much more time. The low time efficiently of the HFLR algorithm is due to two reasons.

Firstly, the HFLR algorithm is a serial algorithm instead of a parallel algorithm: the generation of each new decision tree is based on the results of previous decision trees. Random forest is a parallel algorithm that construct multi decision trees simultaneously. With inspiration from the random forest algorithm, hybrid tree-regression algorithms could be designed to be more time efficient.

Secondly, in the end of each cycle of the HFLR algorithm, stepwise regression is applied to select variables included in the resulting model of the cycle. Meanwhile, machine learning methods such as lasso offers more time efficient ways of the variable selection process. With inspiration taken from these methods, the time efficiency of hybrid tree-regression algorithms could also be improved.

Appendix A

Key R codes

A.1 Chapter 2

A.1.1 RE-LR

```
# READ DATA AND DEFINE INDEPENDENT VARIABLES
trade=read.csv("M:\\TradeGrowth.csv",header=T)
ctry=trade[,3]
time=trade[,7]
UCTRY=(ctry==1)*1
UTIME=(time==1)*1
.....
NGEO=(trade[,36]!="")*1
NUM=as.factor(trade[,1])
```

```
# DEFINE THE DEPENDENT VARIABLE
tsta=trade[,40]
sig=(tsta>qt(0.975,DF))-(tsta+qt(0.975,DF)<0)
pcc=tsta/sqrt(tstaˆ2+DF) # partial correlation coefficient
vpcc=(1-pccˆ2)ˆ2/DF
SE=sqrt(vpcc)


# SETUP THE DATAFRAME
MAV=cbind(pcc,vpcc,SE,UCTRY,UTIME,UCONT,AS,EU,AF,YA10,YA15,
    SHOEF,PCPT,CETE,IGDP,FDI,BMP,OOPEN,EXT,INV,INF,INFR,CRED,
    GOV,OECOP,POLI,POP,LIFE,EDU,OSOC,NGEO,NUM)
MAV=as.data.frame(MAV)


# RUN THE MODEL
library(mixmeta)
mamix=mixmeta(pcc∼.-vpcc-NUM,S=vpcc,data=MAV,random= 1|NUM,
    method="ml")
mamixa=step(mamix)
summary(mamixa)


# CROSS VALIDATION
cv.lm(data=MAV,mamixa$call,m=10)
```

```
mamixapr=cv.lm(data=MAV,mamixa$call,m=10,printit=FALSE)$cvpred
mamixapt=mamixapr*sqrt(DF/(1-mamixapr^2))
mamixapsig=(mamixapt>qt(0.975,DF))-(mamixapt+qt(0.975,DF)<0)
mean(sig==mamixapsig)
```

A.1.2 WLS-LR

```
# RUN THE MODEL
maols=lm(pcc~.-vpcc-NUM,data=MAV)
maolse=abs(resid(maols))
maolsp=predict(maols)
maolsep=predict(lm(maolse~maolsp))
wmaols=maolsep^(-2)
wmaols[wmaols>quantile(wmaols,0.75)]=quantile(wmaols,0.75)
mawls=lm(pcc~.-vpcc-NUM,weights=wmaols,data=MAV)
maolsa=step(mawls)
maolsae=abs(resid(maolsa))
maolsap=predict(maolsa)
maolsaep=predict(lm(maolsae~maolsap))
wmaolsa=maolsaep^(-2)
wmaolsa[wmaolsa>quantile(wmaolsa,0.75)]=quantile(wmaolsa,0.75)
mawlsa=lm(maolsa$call,weights=wmaolsa,data=MAV)
summary(mawlsa)
```

```
# CROSS VALIDATION
library(DAAG)
cv.lm(data=MAV,mawlsa$call,m=10)
mawlsapr=cv.lm(data=MAV,mawlsa$call,m=10,printit=FALSE)$cvpred
mawlsapt=mawlsapr*sqrt(DF/(1-mawlsapr^2))
mawlsapsig=(mawlsapt>qt(0.975,DF))-(mawlsapt+qt(0.975,DF)<0)
mean(sig==mawlsapsig)
```

A.1.3 RE-HTLR

```
# CONSTRUCT THE DECISION TREE
library(rpart)
mat1=rpart(pcc~.-vpcc-SE-NUM,data=MAV,method="anova")
par(mar=c(0.5,0.5,0.5,0.5))
plot(mat1,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(mat1,use.n=F,fancy=F,cex=0.8,digits=2,col="blue")


# DEFINE VARIABLES BASED ON THE DECISION TREE
mat1v1=YA15*PCPT
mat1v2=YA15*PCPT*IGDP
.....
mat1v11=(1-YA15)*(1-AF)*(1-YA10)*INV*OECOP
mat1v12=(1-YA15)*(1-AF)*(1-YA10)*(1-INV)*POP
```

```
MAV1=cbind(MAV,mat1v1,mat1v2,mat1v3,mat1v4,mat1v5,mat1v6,mat1v7,
    mat1v8,mat1v9,mat1v10,mat1v11,mat1v12)

# RUN THE MODEL
mamix1=mixmeta(pcc∼.-vpcc-NUM,S=vpcc,data=MAV1,random= 1|NUM,
    method="ml")
mamix1a=step(mamix1)
summary(mamix1a)


# CROSS VALIDATION
cv.lm(data=MAV1,mamix1a$call,m=10)
mamix1apr=cv.lm(data=MAV1,mamix1a$call,m=10,printit=FALSE)$cvpred
mamix1apt=mamix1apr*sqrt(DF/(1-mamix1apr^2))
mamix1apsig=(mamix1apt>qt(0.975,DF))-(mamix1apt+qt(0.975,DF)<0)
mean(sig==mamix1apsig)
```

A.1.4 WLS-HTLR

```
# RUN THE MODEL
maols1=lm(pcc∼.-vpcc-NUM,data=MAV1)
maols1e=abs(resid(maols1))
maols1p=predict(maols1)
maols1ep=predict(lm(maols1e∼maols1p))
wmaols1=maols1ep^(-2)
```

```
wmaols1[wmaols1>quantile(wmaols1,0.75)]=quantile(wmaols1,0.75)
mawls1=lm(pcc∼.-vpcc-NUM,weights=wmaols1,data=MAV1)
maols1a=step(mawls1)
maols1ae=abs(resid(maols1a))
maols1ap=predict(maols1a)
maols1aep=predict(lm(maols1ae∼maols1ap))
wmaols1a=maols1aepˆ(-2)
wmaols1a[wmaols1a>quantile(wmaols1a,0.75)]=quantile(wmaols1a,0.75)
mawls1a=lm(maols1a$call,weights=wmaols1a,data=MAV1)
summary(mawls1a)


# CROSS VALIDATION
cv.lm(data=MAV1,mawls1a$call,m=10)
mawls1apr=cv.lm(data=MAV1,mawls1a$call,m=10,printit=FALSE)$cvpred
mawls1apt=mawls1apr*sqrt(DF/(1-mawls1aprˆ2))
mawls1apsig=(mawls1apt>qt(0.975,DF))-(mawls1apt+qt(0.975,DF)<0)
mean(sig==mawls1apsig)
```

A.1.5 RE-HGTLR

```
# CONSTRUCT THE DECISION TREES
magt1=rpart(pcc∼.-vpcc-SE-NUM,data=MAV,subset=(YA10+YA15==0),method="anova")
par(mar=c(0.5,0.5,0.5,0.5))
```

```
plot(magt1,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(magt1,use.n=F,fancy=F,cex=1.2,digits=2,col="blue")
magt2=rpart(pcc~.-vpcc-SE-NUM,data=MAV,subset=(YA10==1),method="anova")
par(mar=c(0.5,0.5,0.5,0.5))
plot(magt2,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(magt2,use.n=F,fancy=F,cex=1.2,digits=2,col="blue")
magt3=rpart(pcc~.-vpcc-SE-NUM,data=MAV,subset=(YA15==1),method="anova")
par(mar=c(0.5,0.5,0.5,0.5))
plot(magt3,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(magt3,use.n=F,fancy=F,cex=1.2,digits=2,col="blue")


# DEFINE VARIABLES BASED ON THE DECISION TREE
magt1v1=(1-YA10-YA15)*INV
......
magt1v8=(1-YA10-YA15)*(1-INV)*(1-POP)*(1-CETE)*OECOP
magt2v1=YA10*AF
......
magt2v8=YA10*(1-AF)*(1-INV)*(1-POP)*(1-NGEO)*INF
magt3v1=YA15*PCPT
......
magt3v6=YA15*PCPT*IGDP*(1-OSOC)*EDU*OECOP*UCTRY
MAGV1=cbind(MAV,magt1v1,magt1v2,magt1v3,magt1v4,magt1v5,magt1v6,
    magt1v7,magt1v8,magt2v1,magt2v2,magt2v3,magt2v4,magt2v5,magt2v6,
    magt2v7,magt2v8,magt3v1,magt3v2,magt3v3,magt3v4,magt3v5,magt3v6)
```

```
# RUN THE MODEL
mamixg1=mixmeta(pcc∼.-vpcc-NUM,S=vpcc,data=MAGV1,random= 1|NUM,
    method="ml")
mamixg1a=step(mamixg1)
summary(mamixg1a)


# CROSS VALIDATION
cv.lm(data=MAGV1,mamixg1a$call,m=10)
mamixg1apr=cv.lm(data=MAGV1,mamixg1a$call,m=10,printit=FALSE)$cvpred
mamixg1apt=mamixg1apr*sqrt(DF/(1-mamixg1apr^2))
mamixg1apsig=(mamixg1apt>qt(0.975,DF))-(mamixg1apt+qt(0.975,DF)<0)
mean(sig==mamixg1apsig)
```

A.1.6 WLS-HGTLR

```
# RUN THE MODEL
magols1=lm(pcc∼.-vpcc-NUM,data=MAGV1)
magols1e=abs(resid(magols1))
magols1p=predict(magols1)
magols1ep=predict(lm(magols1e∼magols1p))
wmagols1=magols1ep^(-2)
wmagols1[wmagols1>quantile(wmagols1,0.75)]=quantile(wmagols1,0.75)
magwls1=lm(pcc∼.-vpcc-NUM,weights=wmagols1,data=MAGV1)
```

```
magols1a=step(magwls1)
magols1ae=abs(resid(magols1a))
magols1ap=predict(magols1a)
magols1aep=predict(lm(magols1ae~magols1ap))
wmagols1a=magols1aep^(-2)
wmagols1a[wmagols1a>quantile(wmagols1a,0.75)]=quantile(wmagols1a,0.75)
magwls1a=lm(magols1a$call,weights=wmagols1a,data=MAGV1)
summary(magwls1a)


# CROSS VALIDATION
cv.lm(data=MAGV1,magwls1a$call,m=10)
magwls1apr=cv.lm(data=MAGV1,magwls1a$call,m=10,printit=FALSE)$cvpred
magwls1apt=magwls1apr*sqrt(DF/(1-magwls1apr^2))
magwls1apsig=(magwls1apt>qt(0.975,DF))-(magwls1apt+qt(0.975,DF)<0)
mean(sig==magwls1apsig)
```

A.2 Chapter 3

A.2.1 LR

```
# READ DATA AND DEFINE VARIABLES
crowdfund=read.csv("M:\\HLL.csv",header=T)
```

```
crowdfund[is.na(crowdfund)]<-0
target=crowdfund[,13]
result=crowdfund[,14]+0.01
lrt=log(result/target) # proportion of funding
LTARG=log(target)
LPICS=log(crowdfund[,58]+1)
KICKS=(crowdfund[,3]==1)*1
.....
PTPROD=(crowdfund[,62]+crowdfund[,66]>0)*1
PTPROC=(crowdfund[,63]+crowdfund[,64]>0)*1
PTSLMP=(crowdfund[,68]+crowdfund[,69]+crowdfund[,70]+crowdfund[,71]+
    crowdfund[,72]>0)*1
hlvar0=cbind(lrt,LTARG,LPICS,KICKS,COMACT,DESTECH,FOODCR,PROD,
    PERSON,MALE,PROFL,SELF,SPT,BUSIN,JOBDES,DEG,RWPROD,
    RWEXP,RWVIP,RWHONOR,RWLET,RWPHOTO,RWVISIT,RWANM,
    RWINT,RWDESH,PTPROV,PTREC,PTPROD,PTPROC,PTSLMP)
hlvar0=as.data.frame(hlvar0)


# RUN THE MODEL
hllr=lm(lrt~.,data=hlvar0)
hllrb=step(hllr,k=log(236))
summary(hllrb)


# CROSS VALIDATION
```

```
library(DAAG)
cv.lm(hlvar0,hllrb$call,m=10)
hllrbp=cv.lm(hlvar0,hllrb$call,m=10,printit=FALSE)$cvpred
mean(hllrbp*lrt>0)
```

A.2.2 HTLR

```
# RUN THE FIRST TREE AND DEFINE NEW VARIABLES
library(rpart)
hldvar=hlvar0[,c(1,4:31)]
hlt1=rpart(lrt∼.,data=hldvar,method="anova")
par(mar=c(0.5,0.5,0.5,0.5))
plot(hlt1,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(hlt1,use.n=F,fancy=F,cex=1.2,digits=2,col="blue")
hlt1v1=SPT*PTREC
hlt1v2=(1-SPT)*RWPROD
.....
hlt1v6=(1-SPT)*(1-RWPROD)*COMACT
hlt1v7=(1-SPT)*(1-RWPROD)*COMACT*RWLET
hlvar1=cbind(hlvar0,hlt1v1,hlt1v2,hlt1v3,hlt1v4,hlt1v5,hlt1v6,hlt1v7)

# RUN THE MODEL
hlhtlr=lm(lrt∼.,data=hlvar1)
```

```
anova(hllr,hlhtlr)
hlhtlrb=step(hlhtlr,k=log(236))
summary(hlhtlrb)


# CROSS VALIDATION
cv.lm(hlvar1,hlhtlrb$call,m=10)
hltlbp=cv.lm(hlvar1,hlhtlr$call,m=10,printit=FALSE)$cvpred
mean(hltlbp*lrt>0)
```

A.2.3 HFLR

```
# RUN THE SECOND TREE AND DEFINE NEW VARIABLES
hlt2=rpart(lrt∼.-SPT,data=hldvar,method="anova")
par(mar=c(0.5,0.5,0.5,0.5))
plot(hlt2,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(hlt2,use.n=F,fancy=F,cex=0.8,digits=2,col="blue")
hlt2v1=RWPROD*PROD
.....
hlt2v10=(1-RWPROD)*(1-PTPROD)*(1-RWHONOR)*JOBDES
hlvar2=cbind(hlvar1,hlt2v1,hlt2v2,hlt2v3,hlt2v4,hlt2v5,hlt2v6,hlt2v7,hlt2v8,
    hlt2v9,hlt2v10)


# RUN THE MODEL WITH TWO TREES
```

```
hlhflr2=lm(lrt∼.,data=hlvar2)
anova(hlhtlr,hlhflr2) # Check the loop ending condition
hlhflr2b=step(hlhflr2,k=log(236))
summary(hlhflr2b)


# RUN THE THIRD TREE AND DEFINE NEW VARIABLES
hlt3=rpart(lrt∼.-SPT-RWPROD,data=hldvar,method="anova")
par(mar=c(0.5,0.5,0.5,0.5))
plot(hlt3,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(hlt3,use.n=F,fancy=F,cex=0.8,digits=2,col="blue")
hlt3v1=(1-PROD)*RWLET
.....
hlt3v8=(1-PROD)*(1-RWLET)*(1-MALE)*(1-RWVISIT)*PTSLMP
hlvar3=cbind(hlvar2,hlt3v1,hlt3v2,hlt3v3,hlt3v4,hlt3v5,hlt3v6,hlt3v7,hlt3v8)


# RUN THE MODEL WITH THREE TREES
hlhflr3=lm(lrt∼.,data=hlvar3)
anova(hlhflr2,hlhflr3)
hlhflr3b=step(hlhflr3,k=log(236))
summary(hlhflr3b)


# RUN THE FOURTH TREE AND DEFINE NEW VARIABLES
hlt4=rpart(lrt∼.-SPT-RWPROD-PROD,data=hldvar,method="anova")
```

```
par(mar=c(0.5,0.5,0.5,0.5))
plot(hlt4,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(hlt4,use.n=F,fancy=F,cex=0.8,digits=2,col="blue")
hlt4v1=PTPROD*COMACT
.....
hlt4v12=(1-PTPROD)*(1-RWANM)*(1-KICKS)*(1-PROFL)*PTSLMP
hlvar4=cbind(hlvar3,hlt4v1,hlt4v2,hlt4v3,hlt4v4,hlt4v5,hlt4v6,hlt4v7,hlt4v8,
    hlt4v9,hlt4v10,hlt4v11,hlt4v12)

# RUN THE MODEL WITH FOUR TREES
hlhflr4=lm(lrt~.,data=hlvar4)
anova(hlhflr3,hlhflr4)
hlhflr4b=step(hlhflr4,k=log(236))
summary(hlhflr4b)


# RUN THE FIFTH TREE AND DEFINE NEW VARIABLES
hlt5=rpart(lrt~.-SPT-RWPROD-PROD-PTPROD,data=hldvar,method="anova")
par(mar=c(0.5,0.5,0.5,0.5))
plot(hlt5,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(hlt5,use.n=F,fancy=F,cex=0.8,digits=2,col="blue")
hlt5v1=PTREC*PTSLMP
.....
hlt5v8=(1-PTREC)*(1-FOODCR)*(1-KICKS)*(1-RWANM)*MALE*PTPROC
hlvar5=cbind(hlvar4,hlt5v1,hlt5v2,hlt5v3,hlt5v4,hlt5v5,hlt5v6,hlt5v7,hlt5v8)
```

```
# RUN THE MODEL WITH FIVE TREES
hlhflr4=lm(lrt∼.,data=hlvar4)
anova(hlhflr3,hlhflr4) # The loop ending condition is satisfied


# CHOOSE THE REULSTING MODEL FROM THE HFLR PROCESS
extractAIC(hlhtlrb,k=log(236))
extractAIC(hlhflr2b,k=log(236))
extractAIC(hlhflr3b,k=log(236))
extractAIC(hlhflr4b,k=log(236)) # The selected model


# CROSS VALIDATION
cv.lm(hlvar4,hlhflr4b$call,m=10)
hlfl4bp=cv.lm(hlvar4,hlhflr4b$call,m=10,printit=FALSE)$cvpred
mean(hlfl4bp*lrt>0)
```

A.3 Chapter 4

A.3.1 Data Preparation

```
# READ DATA AND DEFINE VARIABLES
yxdata0=read.csv("M:\\yxproject.csv",header=T) lwtp=log(yxdata0[,1])
lbid1=log(yxdata0[,2])
```

```
yes1=(lwtp>lbid1)*1
lfinc=log(yxdata0[,6])
.....
antiage=yxdata0$antiage*yxdata0$tantiage
antiage=(antiage==1)+2*(antiage==2)+3*(antiage==4)+4*(antiage==10)+
    5*(antiage==30)


# DATA PREPARATION FOR SINGLE BOUNDED MODEL
sbmdata0=cbind(lwtp,yes1,lbid1,lfinc,stbias,stmor,gwrisk,gwact,gwmanmd,
    benefit,right,fairev,told,auth,nocont,protlaw,protng,govgoal,tmeql,sofsoc,
    rchema,rimig,rlmed,rrachat,rgmfood,rdrug,rgovreg,rvac,rgovbud,age,gend,
    news,doct,hosp,antiage)
sbmdata=na.omit(sbmdata0)
sbmdata=as.data.frame(sbmdata)
LWTP=sbmdata$lwtp
YES1=sbmdata$yes1
LBID1=sbmdata$lbid1
.....
DOCT=sbmdata$doct
HOSP=sbmdata$hosp
ANTIAGE=sbmdata$antiage
SBMDATA=cbind(YES1,LBID1,LFINC,STBIAS,STMOR,GWRISK,GWACT,
    GWMANMD,BENEFIT,RIGHT,FAIREV,TOLD,AUTH,NOCONT,
    PROTLAW,PROTNG,GOVGOAL,TMEQL,SOFSOC,RCHEMA,RIMIG,
    RLMED,RRACHAT,RGMFOOD,RDRUG,RGOVREG,RVAC,RGOVBUD,
```

```
    AGE,GEND,NEWS,DOCT,HOSP,ANTIAGE)
SBMDATA=as.data.frame(SBMDATA)


# DATA PREPARATION FOR MULTI BOUNDED MODEL
YES5=LWTP>log(5)
YES19=LWTP>log(19)
.....
YES952=LWTP>log(952)
YES1904=LWTP>log(1904)
FMDATA5=cbind(YES5,rep(log(5),778),LBID1,LFINC,STBIAS,STMOR,
    GWRISK,GWACT,GWMANMD,BENEFIT,RIGHT,FAIREV,TOLD,AUTH,
    NOCONT,PROTLAW,PROTNG,GOVGOAL,TMEQL,SOFSOC,RCHEMA,
    RIMIG,RLMED,RRACHAT,RGMFOOD,RDRUG,RGOVREG,RVAC,RGOVBUD,
    AGE,GEND,NEWS,DOCT,HOSP,ANTIAGE)
.....
FMDATA1904=cbind(YES1904,rep(log(1904),778),LBID1,LFINC,STBIAS,STMOR,
    GWRISK,GWACT,GWMANMD,BENEFIT,RIGHT,FAIREV,TOLD,AUTH,
    NOCONT,PROTLAW,PROTNG,GOVGOAL,TMEQL,SOFSOC,RCHEMA,
    RIMIG,RLMED,RRACHAT,RGMFOOD,RDRUG,RGOVREG,RVAC,RGOVBUD,
    AGE,GEND,NEWS,DOCT,HOSP,ANTIAGE)
FMDATA=as.data.frame(FMDATA)
```

A.3.2 Single Bounded Model

```
# LR
sbmlr=glm(YES1∼.,data=SBMDATA,family=”binomial”(link=”probit”))
sbmlrb=step(sbmlr,k=log(778))
summary(sbmlrb)


# HTLR
library(rpart)
sbmt1=rpart(YES1∼.,data=SBMDVAR,method=”class”)
par(mar=c(0.5,0.5,0.5,0.5))
plot(sbmt1,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(sbmt1,use.n=F,fancy=F,cex=0.8,col=”blue”)
sbmt1v1=(BENEFIT<3.5)*1
sbmt1v2=(BENEFIT<3.5)*(RIMIG>4.5)
.....
sbmt1v13=(BENEFIT>3.5)*(TMEQL<3.5)*(RDRUG>4.5)*(RLMED<2.5)*
    (RCHEMA<9.5)*(GOVGOAL<2.5)
sbmt1v14=(BENEFIT>3.5)*(TMEQL<3.5)*(RDRUG>4.5)*(RLMED<2.5)*
    (RCHEMA<9.5)*(GOVGOAL<2.5)*(PROTLAW<4.5)
SBMDATA1=cbind(SBMDATA,sbmt1v1,sbmt1v2,sbmt1v3,sbmt1v4,sbmt1v5,
    sbmt1v6,sbmt1v7,sbmt1v8,sbmt1v9,sbmt1v10,sbmt1v11,sbmt1v12,sbmt1v13,
    sbmt1v14)
sbmhtlr=glm(YES1∼.,data=SBMDATA1,family=”binomial”(link=”probit”))
anova(sbmlr,sbmhtlr,test=”Chisq”)
```

```
sbmhtlrb=step(sbmhtlr,k=log(768))
summary(sbmhtlrb)
sbmhtlrbs=update(sbmhtlrb,.∼.-sbmt1v7)
    # Delete a variable from a highly correlated pair of variables
summary(sbmhtlrbs)


# HFLR - THE SECOND TREE
sbmt2=rpart(YES1∼.-BENEFIT,data=SBMDVAR,method="class")
par(mar=c(0.5,0.5,0.5,0.5))
plot(sbmt2,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(sbmt2,use.n=F,fancy=F,cex=0.7,col="blue")
sbmt2v1=(STMOR<4.5)*1
.....
sbmt2v17=(STMOR>4.5)*(RIGHT<5.5)*(RIMIG>3.5)*(GOVGOAL>3.5)*
    (TOLD>4.5)*(RRACHAT>5.5)*(RCHEMA>7.5)*(RGOVBUD>7.5)
SBMDATA2=cbind(SBMDATA1s,sbmt2v1,sbmt2v2,sbmt2v3,sbmt2v4,sbmt2v5,
    sbmt2v6,sbmt2v7,sbmt2v8,sbmt2v9,sbmt2v10,sbmt2v11,sbmt2v12,sbmt2v13,
    sbmt2v14,sbmt2v15,sbmt2v16,sbmt2v17)
sbmhflr2=glm(YES1∼.,data=SBMDATA2,family="binomial"(link="probit"))
anova(sbmhtlrs,sbmhflr2,test="Chisq")
sbmhflr2b=step(sbmhflr2,k=log(768))
summary(sbmhflr2b)


# HFLR - THE THIRD TO SIXTH TREE
```

```
.....


# HFLR - THE SEVENTH TREE
sbmt7=rpart(YES1∼.-BENEFIT-STMOR-RGMFOOD-GOVGOAL-
    RDRUG-TMEQL,data=SBMDVAR,method="class")
par(mar=c(0.5,0.5,0.5,0.5))
plot(sbmt7,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(sbmt7,use.n=F,fancy=F,cex=0.7,col="blue")
sbmt7v1=(GWACT<3.5)*1
.....
sbmt7v7=(GWACT>3.5)*(NEWS>2.5)*(RIGHT>2.5)*(AGE>3.5)*
    (GWRISK>5.5)
SBMDATA10=cbind(SBMDATA6s,sbmt7v1,sbmt7v2,sbmt7v3,sbmt7v4,
    sbmt7v5,sbmt7v6,sbmt7v7)
sbmhflr7=glm(YES1∼.,data=SBMDATA7,family="binomial"(link="probit"))
anova(sbmhflr6s,sbmhflr7,test="Chisq") # The loop ending condition is satisfied


# HFLR - MODEL SELECTION
extractAIC(sbmhtlrbs,k=log(768))
extractAIC(sbmhflr2b,k=log(768))
.....
extractAIC(sbmhflr5b,k=log(768))
extractAIC(sbmhflr6b,k=log(768)) # The selected model
```

```
# CROSS VALIDATIONS
library(DAAG)
sbmp0=cv.lm(SBMDATA,sbmlrb$call,m=10,printit=FALSE)$cvpred
mean((sbmp0>0.5)==YES1)
sbmp1=cv.lm(SBMDATA1,sbmhtlrbs$call,m=10,printit=FALSE)$cvpred
mean((sbmp1>0.5)==YES1)
sbmp7=cv.lm(SBMDATA7,sbmhflr7b$call,m=10,printit=FALSE)$cvpred
mean((sbmp9>0.5)==YES1)
```

A.3.3 Multi Bounded Model

```
# LR
fmlr=glm(YES∼.,data=FMDATA,family="binomial"(link="probit"))
fmlrb=step(fmlr,k=log(6144))
summary(fmlrb)


# HTLR
fmt1=rpart(YES∼.,data=FMDVAR,method="class")
par(mar=c(0.5,0.5,0.5,0.5))
plot(fmt1,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(fmt1,use.n=F,fancy=F,cex=0.8,col="blue")
fmt1v1=(PROTNG>2.5)*1
.....
```

```
fmt1v5=(PROTNG>2.5)*(GWACT>4.5)*(RCHEMA>3.5)*(GWRISK>5.5)*
    (STBIAS>3.5)
FMDATA1=cbind(FMDATA,fmt1v1,fmt1v2,fmt1v3,fmt1v4,fmt1v5)
fmhtlr=glm(YES∼.,data=FMDATA1,family="binomial"(link="probit"))
anova(fmlr,fmhtlr,test="Chisq")
fmhtlrb=step(fmhtlr,k=log(6144))
summary(fmhtlrb)


# HFLR - THE SECOND TREE
FMDATA1s=cbind(FMDATA,fmt1v2,fmt1v3,fmt1v4,fmt1v5)
fmhtlrs=glm(YES∼.,data=FMDATA1s,family="binomial"(link="probit"))
fmt2=rpart(YES∼.-PROTNG,data=FMDVAR,method="class")
par(mar=c(0.5,0.5,0.5,0.5))
plot(fmt2,uniform=T,branch=0.5,margin=0,minbranch=0.5)
text(fmt2,use.n=F,fancy=F,cex=0.8,col="blue")
fmt2v1=(BENEFIT>3.5)*1
.....
fmt2v8=(BENEFIT>4.5)*(FAIREV>3.5)*(RCHEMA>3.5)*(PROTLAW>2.5)*
    (SOFSOC>3.5)*(RDRUG>6.5)*(STBIAS>3.5)
FMDATA2=cbind(FMDATA1s,fmt2v1,fmt2v2,fmt2v3,fmt2v4,fmt2v5,fmt2v6,
    fmt2v7,fmt2v8)
fmhflr2=glm(YES∼.,data=FMDATA2,family="binomial"(link="probit"))
anova(fmhtlrs,fmhflr2,test="Chisq")
fmhflr2b=step(fmhflr2,k=log(6144))
summary(fmhflr2b)
```

```
# HFLR - THE THIRD TO SIXTH TREE
.....


# HFLR - THE SEVENTH TREE
FMDATA6s=cbind(FMDATA14s,fmt15v1)
fmhflr6s=glm(YES∼.,data=FMDATA15s,family="binomial"(link="probit"))
fmt7=rpart(YES∼.-PROTNG-BENEFIT-STMOR-GOVGOAL-GWACT-PROTLAW,
    data=FMDVAR,method="class")
par(mar=c(0.5,0.5,0.5,0.5))
plot(fmt7,uniform=T,branch=0,margin=0)
text(fmt7,use.n=F,fancy=F,cex=0.8,col="blue")
fmt7v1=(STBIAS<4.5)*1
fmt7v2=(STBIAS<4.5)*(RIMIG>5.5)
fmt7v3=(STBIAS<4.5)*(RIMIG>5.5)*(GWRISK<5.5)
FMDATA7=cbind(FMDATA6s,fmt7v1,fmt7v2,fmt7v3)
fmhflr7=glm(YES∼.,data=FMDATA16,family="binomial"(link="probit"))
anova(fmhflr15s,fmhflr16,test="Chisq") fmhflr7b=step(fmhflr7,k=log(6144)) summary(fmhflr7b)
# The loop ending condition is satisfied


# HFLR - MODEL SELECTION
extractAIC(fmhtlrb,k=log(6144))
extractAIC(fmhflr2b,k=log(6144))
```

```
.....
extractAIC(fmhflr7b,k=log(6144)) # The selected model


# CROSS VALIDATION - LR
fmp0=cv.lm(FMDATA,fmlrb$call,m=10,printit=FALSE)$cvpred
fmc0=((fmp0>0.5)==YES)
mean(fmc0)
fmclv0=fmc0[8*1:768-7]+fmc0[8*1:768-6]+fmc0[8*1:768-5]+fmc0[8*1:768-4]+
    fmc0[8*1:768-3]+fmc0[8*1:768-2]+fmc0[8*1:768-1]+fmc0[8*1:768]
mean(fmclv0==8)
mean(fmclv0>=7)


# CROSS VALIDATION - HTLR
fmp1=cv.lm(FMDATA1,fmhtlrb$call,m=10,printit=FALSE)$cvpred
fmc1=((fmp1>0.5)==YES)
mean(fmc1)
fmclv1=fmc1[8*1:768-7]+fmc1[8*1:768-6]+fmc1[8*1:768-5]+fmc1[8*1:768-4]+
    fmc1[8*1:768-3]+fmc1[8*1:768-2]+fmc1[8*1:768-1]+fmc1[8*1:768]
mean(fmclv1==8)
mean(fmclv1>=7)


# CROSS VALIDATION - HFLR
fmp6=cv.lm(FMDATA6,fmhflr6b$call,m=10,printit=FALSE)$cvpred
```

```
fmc6=((fmp6>0.5)==YES)
mean(fmc6)
fmclv6=fmc6[8*1:768-7]+fmc6[8*1:768-6]+fmc6[8*1:768-5]+fmc6[8*1:768-4]+
    fmc6[8*1:768-3]+fmc6[8*1:768-2]+fmc6[8*1:768-1]+fmc6[8*1:768]
mean(fmclv6==8)
mean(fmclv6>=7)
```

Bibliography

[1] Abdullah, A., Doucouliagos, H., & Manning, E. (2015). Does education reduce income inequality? A meta-regression analysis. *Journal of Economic Surveys, 29*(2), 301-316.

[2] Alter, K. (2007). Social enterprise typology. *Virtue ventures LLC, 12*(1), 1-124.

[3] Benos, N., & Zotou, S. (2014). Education and economic growth: A meta-regression analysis. *World Development, 64*, 669-689.

[4] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees.* Routledge.

[5] Böckel, A., Hörisch, J., & Tenner, I. (2021). A systematic literature review of crowdfunding and sustainability: highlighting what really matters. *Management Review Quarterly, 71*, 433-453.

[6] Cameron, T. A., & Quiggin, J. (1994). Estimation using contingent valuation data from a" dichotomous choice with follow-up" questionnaire. *Journal of environmental economics and management, 27*(3), 218-234.

[7] Cawley, J. (2008). Contingent valuation analysis of willingness to pay to reduce childhood obesity. *Economics & Human Biology, 6*(2), 281-292.

[8] Catford, J. (1998). Social entrepreneurs are vital for health promotion—but they need supportive environments too. *Health promotion international, 13*(2), 95-97.

[9] Chen, Y., Zhang, W., Yan, X., & Jin, J. (2020). The life-cycle influence mechanism of the determinants of financing performance: an empirical study of a Chinese crowdfunding platform. *Review of Managerial Science, 14*(1), 287-309.

[10] Cho, Y., & Honorati, M. (2014). Entrepreneurship programs in developing countries: A meta regression analysis. *Labour Economics, 28*, 110-130.

[11] Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society: Series A (General), 134*(3), 321-353.

[12] Dalla Chiesa, C., & Dekker, E. (2021). Crowdfunding artists: beyond matchmaking on platforms. *Socio-Economic Review, 19*(4), 1265-1290.

[13] De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760-772.

[14] Dees, J. G., Economy, P., & Emerson, J. (2004). *Strategic tools for social entrepreneurs: Enhancing the performance of your enterprising nonprofit.* John Wiley & Sons.

[15] Dufrenot, G., Mignon, V., & Tsangarides, C. (2010). The trade-growth nexus in the developing countries: A quantile regression approach. *Review of World Economics, 146*(4), 731-761.

[16] Dumitrescu, E., Hue, S., Hurlin, C., & Tokpavi, S. (2018). *Machine learning for credit scoring: improving logistic regression with non linear decision tree effects*

(Doctoral dissertation, Ph. D. thesis, Paris Nanterre University, University of Orleans).

[17] Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems, 5*(4), 83-124.

[18] Hanemann, M., Loomis, J., & Kanninen, B. (1991). Statistical efficiency of double-bounded dichotomous choice contingent valuation. *American journal of agricultural economics, 73*(4), 1255-1263.

[19] Hemilä, H., & Chalker, E. (2020). Vitamin C may reduce the duration of mechanical ventilation in critically ill patients: a meta-regression analysis. *Journal of intensive care, 8*(1), 1-9.

[20] Hossain, M., & Oparaocha, G. O. (2017). Crowdfunding: Motives, definitions, typology and ethical challenges. *Entrepreneurship Research Journal, 7*(2).

[21] Huang, S., Pickernell, D., Battisti, M., & Nguyen, T. (2021). Signalling entrepreneurs' credibility and project quality for crowdfunding success: cases from the Kickstarter and Indiegogo environments. *Small Business Economics*, 1-21.

[22] Huchet-Bourdon, M., Le Mouël, C., & Vijil, M. (2018). The relationship between trade openness and economic growth: Some new insights on the openness measurement issue. *The World Economy, 41*(1), 59-76.

[23] Hue, S., Hurlin, C., & Tokpavi, S. (2017). Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects. *no. July,* 1-29.

[24] Itani, O., Jike, M., Watanabe, N., & Kaneita, Y. (2017). Short sleep duration and health outcomes: a systematic review, meta-analysis, and meta-regression. *Sleep medicine, 32*, 246-256.

[25] John, G. H. (1995, August). Robust Decision Trees: Removing Outliers from Databases. In *KDD* (Vol. 95, pp. 174-179).

[26] Kaartemo, V. (2017). The elements of a successful crowdfunding campaign: A systematic literature review of crowdfunding performance. *International Review of Entrepreneurship, 15*(3), 291-318.

[27] Kim, D. H., Lin, S. C., & Suen, Y. B. (2012). The simultaneous evolution of economic growth, financial development, and trade openness. *The Journal of International Trade & Economic Development, 21*(4), 513-537.

[28] Kim, K., & Hong, J. S. (2017). A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis. *Pattern Recognition Letters, 98,* 39-45.

[29] Kim, Y. S. (2008). Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Expert Systems with Applications, 34*(2), 1227-1234.

[30] Lapczynski, M. (2014). Hybrid C&RT-Logit Models In Churn Analysis. *Folia Oeconomica Stetinensia, 14*(2), 37-52.

[31] Leadbeater, C. (1997). *The rise of the social entrepreneur* (No. 25). Demos.

[32] Li, H., Berrens, R. P., Bohara, A. K., Jenkins-Smith, H. C., Silva, C. L., & Weimer, D. L. (2004). Would developing country commitments affect US house-

holds' support for a modified Kyoto Protocol?. *Ecological Economics, 48*(3), 329-343.

[33] Li, H., Berrens, R. P., Bohara, A. K., Jenkins-Smith, H. C., Silva, C. L., & Weimer, D. L. (2005). Exploring the beta model using proportional budget information in a contingent valuation study. *Economics Bulletin, 17*(8), 1-9.

[34] Liebe, U., Preisendörfer, P., & Meyerhoff, J. (2011). To pay or not to pay: Competing theories to explain individuals' willingness to pay for public environmental goods. *Environment and Behavior, 43*(1), 106-130.

[35] Lim, S. S., Kakoly, N. S., Tan, J. W. J., Fitzgerald, G., Bahri Khomami, M., Joham, A. E., ... & Moran, L. J. (2019). Metabolic syndrome in polycystic ovary syndrome: a systematic review, meta-analysis and meta-regression. *Obesity reviews, 20*(2), 339-352.

[36] Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery, 1*(1), 14-23.

[37] Mair, J., & Marti, I. (2009). Entrepreneurship in and around institutional voids: A case study from Bangladesh. *Journal of business venturing, 24*(5), 419-435.

[38] Moritz, A., & Block, J. H. (2016). Crowdfunding: A literature review and research directions. *Crowdfunding in Europe,* 25-53.

[39] Moysidou, K., & Hausberg, J. P. (2020). In crowdfunding we trust: A trust-building model in lending crowdfunding. *Journal of Small Business Management, 58*(3), 511-543.

[40] Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.

[41] Nicholls, A. (Ed.). (2008). *Social entrepreneurship: New models of sustainable social change.* OUP Oxford.

[42] Popescul, D., Radu, L. D., Păvăloaia, V. D., & Georgescu, M. R. (2020). Psychological Determinants of Investor Motivation in Social Media-Based Crowdfunding Projects: A Systematic Review. *Frontiers in Psychology, 11*, 3676.

[43] Purnama, P. D., & Yao, M. H. (2019). The Relationship between International Trade and Economic Growth. *International Journal of Applied Business Research, 1*(02), 112-123.

[44] Quinlan, J. R. (2014). *C4. 5: programs for machine learning.* Elsevier.

[45] Ramanayake, S. S., & Lee, K. (2015). Does openness lead to sustained economic growth? Export growth versus other variables as determinants of economic growth. *Journal of the Asia Pacific Economy, 20*(3), 345-368.

[46] Roundy, P. T., & Bonnal, M. (2017). The singularity of social entrepreneurship: Untangling its uniqueness and market function. *The Journal of Entrepreneurship, 26*(2), 137-162.

[47] Sauermann, H., Franzoni, C., & Shafi, K. (2019). Crowdfunding scientific research: Descriptive insights and correlates of funding success. *PloS one, 14*(1), e0208384.

[48] Schraven, E., van Burg, E., van Gelderen, M., & Masurel, E. (2020). Predictions of Crowdfunding Campaign Success: The Influence of First Impressions on Accuracy and Positivity. *Journal of Risk and Financial Management, 13*(12), 331.

[49] Stanley, T. D., & Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*. Routledge.

[50] Stanley, T. D., & Doucouliagos, H. (2017). Neither fixed nor random: weighted least squares meta-regression. *Research synthesis methods, 8*(1), 19-42.

[51] Steinberg, D., & Cardell, N. S. (1998). The hybrid CART-Logit model in classification and data mining. *Salford Systems White Paper.*

[52] Tietenberg, T., & Lewis, L. (2018). *Environmental and natural resource economics.* Routledge.

[53] Therneau, T. M., & Atkinson, E. J. (1997). *An introduction to recursive partitioning using the RPART routines* (Vol. 61, p. 452). Mayo Foundation: Technical report.

[54] Venkatachalam, L. (2004). The contingent valuation method: a review. *Environmental impact assessment review, 24*(1), 89-124.

[55] Wang, C., Liu, X., & Wei, Y. (2004). Impact of openness on growth in different country groups. *World Economy, 27*(4), 567-585.

[56] Zahonogo, P. (2016). Trade and economic growth in developing countries: Evidence from sub-Saharan Africa. *Journal of African Trade, 3*(1-2), 41-56.

[57] Zhou, C., Gill, M., & Liu, Q. (2021). Empowering Education with Crowdfunding: The Role of Crowdfunded Resources and Crowd Screening. *Journal of Marketing Research,* 00222437211033536.

[58] Zhu, M., Philpotts, D., Sparks, R., & Stevenson, M. J. (2011). A hybrid approach to combining CART and logistic regression for stock ranking. *The Journal of Portfolio Management, 38*(1), 100-109.