Modelling heterogeneous treatment effects by quantitle local polynomial decision tree and forest

Lai Xinglin\textsuperscript{1a}

\textsuperscript{a}Department of Economics, Guangdong university of Financial and Economics, Guangzhou, 510320, Guangdong, China

Abstract

To further develop the statistical inference problem for heterogeneous treatment effects, this paper builds on Breiman’s (2001)\textsuperscript{1} random forest tree (RFT) and Wager et al.’s (2018)\textsuperscript{2} causal tree to parameterize the nonparametric problem using the excellent statistical properties of classical OLS and the division of local linear intervals based on covariate quantile points, while preserving the random forest trees with the advantages of constructible confidence intervals and asymptotic normality properties [Athey and Imbens (2016)\textsuperscript{3}, Efron (2014)\textsuperscript{4}, Wager et al.(2014)\textsuperscript{5}], we propose a decision tree using quantile classification according to fixed rules combined with polynomial estimation of local samples, which we call the quantile local linear causal tree (QLPRT) and forest (QLPRF).

Keywords: Micro Empirical Econometrics, Decision Tree, Boost Forests algorithm

1. Introduction

From medicine to economics, the evaluation of a new drug trial or a new policy is crucial, and researchers try to get as much information as possible about the treatment effects by obtaining the results of two groups of randomised trials and comparing the differences between the treatment and control groups under the same conditions. For example, the economic or social effects of a new drug trial, a new policy or even the effects of a new feature in an advertisement or software are all areas of interest to researchers.

\textsuperscript{1}Email:Laixinglin@163.com

Preprint submitted to Journal of Multivariate Analysis December 1, 2021
arXiv:2111.15320v1 [econ.EM] 30 Nov 2021
An important but often lacking concern for treatment effect assessment is heterogeneity. Researchers generally choose to report average treatment effects rather than heterogeneous treatment effects for two reasons: insufficient sample size and the hidden risk of focusing too much on extreme treatment effects. The first problem is widespread across fields, such as pharmaceutical experiments and the greater time and monetary costs of policy evaluation and the voluntary nature of some social experiments, all of which contribute to the lack of observed samples. The existence of the first problem also increases the risk of the second problem, in that even if the sample size can distinguish heterogeneity, there is no guarantee that each heterogeneity will be treated objectively and accurately, i.e. there may be an interval of "heterogeneity that may be purely spurious". (see Assmann et al.(2000)\cite{6}, Cook et al.(2004)\cite{7})

However, heterogeneity may be indispensable for research purposes or even far from the reality if heterogeneity is not taken into account, leading to a large bias in the average treatment effect. On the one hand, heterogeneity may be the main purpose of research, for example, new drug trials need to take into account dose effects and the possibility of conflicts with other drugs, and changes in tax rates need to take into account the impact of the policy on businesses of different sizes and in different geographical areas, especially when considering industrial support or environmental and regulatory policies, and the fact that an advertisement will vary in its appearance at different times and on different types of websites. As a result, advertisers will want to know the specific time of day and website to be targeted in order to maximise ad revenue. On the other hand, ignoring heterogeneity can lead to biased estimates. For linear models, ignoring heterogeneity is essentially an omitted variable problem related to the treatment of variables, where the OLS estimates will be biased and inconsistent, which may lead us to the opposite of the truth. To address this problem, we propose an algorithm for regressing and testing the effects of heterogeneous treatments based on OLS(Chaisemartin et al.,2020)\cite{8} and a casual tree developed from the random forest tree proposed by breiman\cite{1}. With Unconfoundedness and the use of optimal quantile points for covariates, we can show that the algorithm is unbiased and asymptotically consistent for both individual variables corresponding to heterogeneous treatment effects and overall treatment effects. For the heterogeneous treatment effect, we obtain the variance of the heterogeneous treatment effect by using boostrap and a random forest tree to obtain its confidence interval.
Non-parametric methods such as K nearest-neighbor (K nearest-neighbor) with matching estimates and polynomial regression were used. Non-parametric methods do perform well with fewer covariates, but non-parametric methods generally suffer from dimensional disaster, as the number of covariates increases, the effectiveness of matching and computational speed decreases significantly. In addition, one problem with estimating heterogeneity treatment effects using non-parametric methods is that it is not possible to construct hypothesis tests for the estimates. [Stefan Wager al. (2018)] obtained confidence intervals for the heterogeneous treatment effects under nearest-neighbour estimation by building a causal tree, but their results lack statistical inference for the entire heterogeneous treatment effect function, and in addition, the strong dependence or sensitivity of the nearest-neighbour estimation itself on the sample feature space and distance measures may also lead to The model is not robust. To further develop the problem of statistical inference of heterogeneous treatment effects, this paper builds on Breiman’s (2001) random forest (RFF) tree and Wager et al.’s (2018) causal tree to parameterize the non-parametric problem using the excellent statistical properties of classical OLS and the division of local polynomial intervals based on covariate quantile points, while preserving the premise of constructible confidence intervals and asymptotic normality properties of random forest trees [Athey and Imbens (2016), Efron (2014), Wager et al. (2014)], we propose a decision tree using quantile classification according to fixed rules combined with polynomial estimation of local samples, which we call the quantile local polynomial causal tree (Quantile local polynomial regression tree, QLPRT) and the quantile local polynomial causal tree. (Quantile local polynomial regression forest, QLPRF).

1.1. Related Work

The random forest [Breiman, 2001b] is an effective forecasting tool and there have been many refinements as well as applications in the related literature. A further improvement is the work of Wager et al. (2018). Our proofs are based on Efron and Stein (1981), but a direct application of random forests to causal inference would face the problem (Wager et al., 2018) that we only get one-sided potential outcomes under the same characteristics and cannot get specific implied values, (Wager et al., 2018) proposed an asymptotic theory on random forest trees to address the problem of testing the heterogeneity treatment effect of random forest tree predictions. The causal forest (casual forest) consists of multiple causal trees.
(casual tree), and the specific causal effect estimation is based on the nearest neighbor estimation, however, the branching rule of the decision tree and the choice of tree depth (depth) will have an impact on the result is not easy to analyze, as well as the k-nearest neighbor estimation needs to determine the super params (super params), and the different super params will have significant differences for different sample sizes. Therefore, causal inference in the random forest framework should be further investigated in terms of superparameters and more reasonable branching rules for interpretation and inference.

Results on asymptotic normality theory for random forest trees (RFTs) can be found mainly in (Meinshausen (2006) [12], Sexton (2008) [13], Duan [14], Biau [2012] [15], Wager and Stefan [2014] [10], Scornet et al. [2015] [16], Mentch et al. [2016] [17]), These results basically establish an asymptotic normal theory based on parameter estimation of random forest trees, thus enabling machine learning methods to abate their black boxiness through hypothesis testing.

In recent years, the use of random forest tree algorithms to estimate heterogeneity treatment effects has gradually become a hot research topic, [Chipman et al. [2010] [18], Green and Kern [2012] [19], Hill [2011] [20], Hill and Su [2013] [21]], using Bayesian Additive Regression Tree (BART) to estimate heterogeneous treatment effects [Hahn, 2019] [22] as well as controlling for confounding effects (confounding). Non-parametric methods that also belong to the Bayesian causal framework include (Dirichlet Process Mixture Models) [Blei et al. (2011)] [23] and Gaussian Process Models [Kocijan et al. (2004)] [24]. The Bayesian causal inference framework has good model flexibility and relatively easy access to the target test distribution via the prior distribution, but is not fully data-driven and requires high sample size as it takes potentially unobservable outcomes and constructs posterior distributions based on them, while the validity of hypothesis testing based on Bayesian priors may be somewhat limited in practice, for example by not guaranteeing the MCMC algorithm’s convergence. Other approaches based on tree models for the analysis of causal effects are [Foster et al. (2011) [25], Wager et al. (2018) [2], Friedberg (2020) [26]].

In growing trees to build our forest, we follow most closely the approach of Athey and Imbens (2016) [3], who propose causal trees, and obtain valid confidence intervals for average treatment effects for each of the subpopulations (leaves) identified by the algorithm and Stefan Wager and Susan Athey (2018) [2], who applied honest tree to estimate heterogeneous treatment effect. Other related approaches in-
clude those of Su et al. (2009)[27], which build a tree for treatment effects in subgroups and use statistical tests to determine splits; however, these papers do not analyze bias or consistency properties, and Meinshausen (2006)[12], which came out regression forests based on Quantile.

Finally, we note a growing literature on estimating heterogeneous treatment effects using different machine learning methods. Imai and Ratkovic (2013)[28], and Weisberg and Pontes (2015)[29] develop lasso-like methods for causal inference in a sparse high-dimensional linear setting. Beygelzimer and Langford [2009][30], and others discuss procedures for transforming outcomes that enable off-the-shelf loss minimization methods to be used for optimal treatment policy estimation. In the econometrics literature, Bhattacharya and Dupas [2012], Dehejia[31], Hirano and Porter (2009)[32], Manski (2004)[33] estimate parametric or semi-parametric models for optimal policies, relying on regularization for covariate selection in the case of Bhattacharya and Dupas [2012][34]. Taddy et al. [2016][35] use Bayesian nonparametric methods with Dirichlet priors to flexibly estimate the data-generating process, and then project the estimates of heterogeneous treatment effects down onto the feature space using regularization methods or regression trees to get low-dimensional summaries of the heterogeneity; but again, there are no guarantees about asymptotic properties.

2. Heterogeneous Casual Forests

2.0.1. Assumptions

Suppose we have $N$ samples, and $p$ covariates,  \{\(Y_i \in \mathbb{R}, X_i^p \in \mathbb{R}^p, T_i \in \{0, 1\}\)\} and mark them one by one as \(i = 1, ..., n\), in which \(T_i\) is treatment indicator, indicates whether the \(i\)th individual is in the treatment group. Following the potential outcomes framework of Neyman (1923)[36] and Rubin (1974)[37], define the existence of potential outcomes given the \(i\)th individual’s eigenvector condition \(Y_i^{(1)}\) and \(Y_i^{(0)}\) corresponds to accepting the treatment or not accepting the treatment, respectively, and defines the conditional heterogeneity treatment effect as \(\tau(x_i^p) = E(Y_i^{(1)} - Y_i^{(0)} | X_i^p = x_i^p)\) in which, \(\tau(x_i^p)\) is a function on the multivariate eigenvector \(x_i^p\), and our proposed method of Quantile local polynomial regression tree is actually combined by estimating the heterogeneity treatment effect functions corresponding to each eigenvector separately, thus \(\tau(x_i^p) = \sum_{i=1}^{p} \tau(x_i)\) The equation is based on the assumption of independence between different features, which is also the assumption(1).
Assumption (1): Given any two of the feature variables $X_i X_j X_i \perp X_j$ holds. Since we can only observe one of $Y_i^{(1)}$ and $Y_i^{(0)}$, we follow the unconfoundedness assumption of [Rosenbaum and Rubin, 1983]

Assumption (2): $\{Y_i^{(0)}, Y_i^{(1)} \perp T_i | X_i\}$ The Unconfoundedness assumption allows us to reduce or even ignore large biases due to the lack of another potential outcome when estimating treatment effects based on similar features, so we only need to look for reasonable ways to find results that are as close as possible to $X_i$ that is as close as possible to $X_j$ to get a consistent estimate of the treatment effect.

Assumption (3): $T_i \perp X_i^p$ The assumption is in effect that the observed individuals receive the treatment completely randomly and are not influenced by human or self-selected factors. Studies by [Fortin et al.,(2011)38, Elder et al.,(2010)39] can show that when there is no selection puzzle, the OLS estimates are equivalent to the matching estimates, and the results are equally robust even in the presence of heterogeneity. However, matched estimators are difficult to test hypotheses directly, and their general methods of generating confidence intervals such as bootstrap or jacknife require larger sample sizes for statistical inference, which may lead to a higher probability of both types of estimation errors. These are the reasons why we have adopted OLS rather than matched estimators.

This assumption is made in the presence of observable variables $X_i^p, T_i$ when the following linear equation is generally used to estimate the treatment effect $Y_i = X_i^T \beta + \tau T_i + \epsilon_i$ where, $\epsilon_i \sim N(0, \sigma^2)$, $\tau$ is the homogeneous treatment effect, next we add the heterogeneous treatment effect to the linear model. For a given feature $X_i^p$, based on the assumption that (1) satisfies $\tau(x_i^p) = \sum_{i=k}^{p} \tau(x_k) = E(Y_i^{(1)} - Y_i^{(0)} | X_i^p = x_i^p)$.

Assumption (4) $\epsilon_i$ The overall variance of is a bounded positive real number, i.e.

$$D(\epsilon) < +\infty$$

Assumption (5) The heterogeneity treatment effect function for an arbitrary feature $k \tau(x_k)$ is the heterogeneity in $x_k$ that is Lipschitz continuous, first-order derivable and bounded over the domain of definition. To facilitate classification using the tree model, we use the features $x_k$ of the quantile $Q(x_k)$ as the target variable for split.
2.1. QLPRT and QLPRF

A random forest consists of a number of decision trees, each branch (leaf/leaves) of which has a pre-selected information gain criterion, e.g. Gini value, information entropy, and the classical decision tree selects branches with the aim of improving prediction accuracy, with the implicit assumption that we already know. However, in heterogeneity treatment effects analysis we do not know the classification of each subgroup, because we do not know the exact form of the heterogeneity function, and because the function is continuous by assumption (4), it is not straightforward to use a classification and regression tree for heterogeneity to predict heterogeneity itself. The honest casual tree of Wager et al. (2018)[2] was first trained using a decision tree as a measure of the distance of features between different samples, and then the trained decision tree was used to find the nearest K sample points and to estimate the average treatment effect (ATE) based on treatment variables distinguishing between treatment and control groups and the average treatment effect is respectively then estimated using the trained decision tree.

As in our hypothesis (1) and as in most cases in reality, the tree predictors (predictors) are obtained directly with the features \( x_p \). The first is that the weak correlation of the selected features leads to poor predictor performance and constrained regression, and the second is that strong predictors may account for strong correlations between features, which are essentially not much different from traditional K-nearest neighbor estimates. Honest casual tree yields data-driven estimates of the average treatment effect of \( x_p \). The non-parametric estimate of \( \mu(x_p) \) cannot be directly derived as a statistic to test for heterogeneity. Due to the shortcomings of the honest casual tree, the following adjustments were made to allow both the heterogeneity test and the regression of heterogeneous treatment effects to be formed by the training process of a single decision tree.

Suppose that for \( \tau(Q^{-1}(x_k)) \) at the quantile \( x_k = \frac{x_k{-}\alpha+k{+}\alpha}{2} \) at the first-order Taylor expansion, where \( Q^{-1}(x_k = x_k{-}\alpha \) and \( Q^{-1}(x_k = x_k{+}\alpha \) denote the difference of the quantile given each interval, respectively \( x_k = x_k{+} \) corresponding to the before and after quantile points.

\[
\tau(Q(x_k)) = \tau(Q(x_k{-}\alpha+Q(x_k{+}\alpha)) + \tau'(Q(x_k{+}\alpha)(Q(x_k)-Q(x_k{-}\alpha+Q(x_k{+}\alpha)^2))x_k{+}\alpha \in \frac{x_k{-}\alpha+\alpha}{2},x_k{+}\alpha
\]

where \( Q(x_k) \) denotes the quantile function for feature k. It can be shown (Appendix A) that in the sample size \( N \to +\infty \) and after selecting the optimal quantile according to the method of the 2.2, it is assumed that there
are a total of \( M \) quantile we define \( \alpha = \frac{1}{M+1} \), The Taylor expansion with
Lagrangian remainder term is as follows

\[
\lim_{N \to +\infty} \tau'(Q(x_k,\xi)) = 0, \quad \lim_{N \to +\infty} [Q(x_k) - \frac{Q(x_k^-,\frac{\alpha}{2}) + Q(x_k^+\frac{\alpha}{2})}{2}] = 0
\]

For \( \tau'(Q(x_k,\xi)) \) we define the independent variable as \( [Q(x_k) - \frac{Q(x_k^-,\frac{\alpha}{2}) + Q(x_k^+\frac{\alpha}{2})}{2}] \)
and estimated using general OLS, so that the corresponding t-statistic can be
obtained and statistical inference can be made without relying on asymptotic
normal theory, as for \( \tau(\frac{Q(x_k^-,\frac{\alpha}{2}) + Q(x_k^+\frac{\alpha}{2})}{2}) \), which is exactly \( [x_k^-,\frac{\alpha}{2}, x_k^+\frac{\alpha}{2}] \).
The interval on the interval about \( \tau(x_k) \) and the mathematical expectation on
the interval \( E(\tau(x_k)) \) can be proved, \( \lim_{N \to +\infty} E(\tau(x_k, x_k \in [x_k^-,\frac{\alpha}{2}, x_k^+\frac{\alpha}{2}])) \) \( \to \tau(\frac{Q(x_k^-,\frac{\alpha}{2}) + Q(x_k^+\frac{\alpha}{2})}{2}) \)
The aim of Assumption (4) is to find a constant that is
closest to the actual heterogeneity function at a suitable interval of quantile
points, so we would like to obtain the ideal situation where the heterogeneity
function is significantly non-zero and the Lagrangian remainder term with
first order derivatives is close to 0,

\[
P(E(\tau(x_k, x_k \in [x_k^-,\frac{\alpha}{2}, x_k^+\frac{\alpha}{2}]) = 0) \to 0
\]

\[
1 - P(\tau'(Q(x_k,\xi))) \to 0
\]

Where OLS is used, \( \tau(x_k, x_k \in [x_k^-,\frac{\alpha}{2}, x_k^+\frac{\alpha}{2}] \) It is in \( x_k \in x_k^-,\frac{\alpha}{2}, x_k^+\frac{\alpha}{2} \)
The interval is with respect to the processing variable \( T_i * I(Q(x_k, -\frac{\alpha}{2}) \leq Q(x_k) \leq Q(x_k, +\frac{\alpha}{2}) \) \( \leq 0,1 \) and the estimate of \( \tau'(Q(x_k,\xi)) \) is exactly the estimate of

\[
x_k \in x_k^-,\frac{\alpha}{2}, x_k^+\frac{\alpha}{2}
\]

The interval is with respect to the variable is \( T_i * I(Q(x_k, -\frac{\alpha}{2}) \leq Q(x_k) \leq Q(x_k, +\frac{\alpha}{2}) \) \( \leq[Q(x_k) - \frac{Q(x_k^-,\frac{\alpha}{2}) + Q(x_k^+\frac{\alpha}{2})}{2}] \),while the estimator of which \( I(.) \) denotes the demonstrative function.Thus the ideal case mentioned above can be
expressed as a P-value test on the estimated coefficient t-statistic, respectively,
and the optimal quantile should be the point that makes each interval
as close as possible to the above ideal case, so we need a unified criterion to
combine the cases of each interval, and since we have separately obtained in
a single interval about the constant term \( E(\tau(x_k, x_k) \) (hereafter abbreviated
as \( E \)) and Lagrangian residual coefficient \( \tau'(Q(x_k,\xi)) \) (hereinafter referred to
as \( L \)) and the t-statistics of \( t_E \) and \( t_L \), taking into account bilateral tests, we
can construct separately for individual intervals a $t^2_E$ or $t^2_L \sim \chi^2(1)$ statistic, under the assumption of the ideal case, we would like $t^2_E$ to be as large as possible, and $t^2_L$ as small as possible, and if all intervals are taken into account, then it is $Max: \sum_{l=1}^{M} t^2_{E,l}$, $Min: \sum_{l=1}^{M+1} t^2_{L,l}$, the objective function above will converge to the $F$ statistic, and since the degrees of freedom are different, we can compare different sample degrees of freedom $F$ statistic corresponding to the $p$ values for different sample degrees of freedom, and choose the hyperparameter that minimizes it.

$$Min_M: 1 - F^{-1}_{M,M} \left[ \frac{\sum_{l=1}^{M} t^2_{E,l} M}{\sum_{l=1}^{M} t^2_{L,l}} \right] = \frac{\sum_{l=1}^{M} t^2_{E,l}}{\sum_{l=1}^{M} t^2_{L,l}}$$

where $F^{-1}_{M,M}$ denotes the degrees of freedom respectively $[M, M]$ of the cumulative distribution function of the F-statistic. When the sample size tends to positive infinity, which is, $\lim_{N \to +\infty} \frac{\sum_{l=1}^{M+1} t^2_{E,l}}{\sum_{l=1}^{M+1} t^2_{L,l}} \sim F(M, M)$. However, the solution of the maximization function does not depend on asymptotic normal theory, and the above F-statistic can be used as a reference for the regression results of the heterogeneous treatment effect model. However, even if the sample size is not large enough for the maximization objective function to converge with the F-statistic corresponding to the standard normal, we can still maximize the above objective function and solve for the optimal value, i.e. the number of quantile points.

We differ from the honest casual tree in that our branching of the heterogeneity treatment effect number (split) uses a per-parameter $M$ which means that after determining $M$, the quantile points are $[0, \frac{1}{M+1}, \frac{2}{M+1}, \ldots, \frac{M}{M+1}, 1]$ for $M$. There is a problem with the optimization of $M$ in that $M$ itself is a discrete variable rather than a continuous one, and therefore it cannot be solved directly by optimization algorithms such as BFGS, etc. One solution is to give a reference range for $M$ such as $[2, \sqrt{N}]$ and then use grid search to solve for the maximum of the objective function. This ensures that every interval has at least $\sqrt{N}$ samples.

### 2.2. Forests for building confidence interval

The method in section 2.1 determines the hyperparameters $M$ as the number of quantile points and determines the use of an average distribution of quantile points by $M$. Thus the individual decision tree we obtain with respect to the heterogeneity treatment effect function $E(\tau(x_k))$ should be
written given the hyperparameters $M$ and also the conditional expectation under $E(\tau(x_k)\mid M)$. The classical tree models are given a maximum depth to avoid over-fitting, but in contrast to other machine learning methods such as decision tree and K-nearest neighbor algorithms, our proposed QLPRT method treats the hyperparameter $M$ as a random variable and samples it through a random forest model to obtain its sampling distribution, thus obtaining the expectation with respect to $M$.

$$E(\tau(x_k)) = \sum E(\tau(x_k)\mid M) \ast p(M)$$

The above analysis has two main benefits: Firstly, we are constructing the sampling distribution using a random forest tree model for the hyperparameters rather than the estimators themselves, even though the statistical inference on the treatment effects and other estimators has less impact compared to the latter with a smaller sample size, and the probability of making both types of errors decreases; and secondly, the conditional distribution through which we find $\tau(x_k)$ of the conditional expectation allows the estimator of the original single QLPRT at $x_k$. The conditional expectation of $\tau(x_k)$ can smooth out the estimates of individual QLPRT trees at different levels of $x_k$, thus transforming the stepwise function of individual LPQRT trees into a smoother continuous function and reflecting the heterogeneity treatment effect in a more scientific and reasonable way. The general sample generation approach for random forest tree models is bootstrap, assuming we generate $K$ groups of samples, with each group having a total of $D$ samples, each of which is derived from the original $N$ For each set of samples, the quantile hyperparameters are obtained by maximizing the objective function $M_l$ Thus we can also obtain the total number of samples with respect to the hyperparameter $M_l$ of the total number of samples $\{M_1, M_2, ..., M_D\}$ with the hyperparameter $M$ The corresponding quantile is also a sampling space, but its representation is more indirect, requiring a vector representation through the hyperparameters.

In summary, the complete QLPRT steps are as follows:

1. bracketed label: determine $K$ and number $D$ parameters Determine the number of zone groups on a case-by-case basis $K$ and the number of samples per block group $D$, and the variable for which the QLPRT regression is currently being performed $x_k$, in theory the number of groups and samples per group is higher, in practice it is determined by the amount of computational resources available.
2. Bootstrap Sampling
   From a total of $N$ samples $K \times D$ individuals and assign them sequentially to $K$ groups. Then we should...

3. Local polynomial regression
   For each tree $t$, according to the possible range of hyperparameters $[2, \sqrt{D}]$, enter the quantile points separately and perform a local polynomial regression with covariates to obtain a treatment effect function on heterogeneity $\tau(x_k)$ with constant values and t-statistics for the Lagrangian term coefficients, and construct the objective function
   $\text{sum}_{t=1}^{M_t} t_{E,t}^2, \text{Min :}
   \text{sum}_{t=1}^{M+1} t_{L,t}^2$. Compare different $M_t$ parameters corresponding to the objective function corresponding to the $p$ values, selecting the hyperparameters corresponding to the smallest of them and $\tau(x_k)$, which is the constant value of the coefficients of the Lagrangian term as the estimated value of this QLPRT tree $M_t$ sample values.

4. $x_k$
   Generate a smoothed heterogeneity treatment effect function on the variable $x_k$. Define the optimal per-tree constant term estimate for $\tau(x_k)$ as $c_{E,t}$ and an estimate of the Lagrangian term coefficients as $c_{L,t}$, so that
   $$ E[\tau(x_k)] = \sum_{t=1}^{K} c_{E,t} \times p(M = M_t) $$
   The above is the expectation without the Lagrangian term; to describe the heterogeneity function more accurately, the Lagrangian term can be added as follows.
   $$ E[\tau(x_k)] = \sum_{t=1}^{K} [c_{E,t} + c_{L,t} \times \{Q_{x_k} - \frac{Q(-\frac{\alpha}{2}) + Q(+\frac{\alpha}{2})}{2}\}] \times p(M = M_t) $$
   The choice of whether to include a Lagrangian term can be based on $M_t$ of the Lagrangian term of the tree corresponding to the plurality in $M_t$, $t_L$ statistics consisting of $\chi^2(M_t)$ statistic, if it is more significant, e.g. above 10% confidence level, then the Lagrangian term should be included, otherwise it can be excluded. Another simple but effective way to judge this is to choose both inclusion and exclusion, and to choose the class of function images obtained by regression that is as smooth and monotonic as possible, so that whether to include the Lagrangian term depends on the specific sample characteristics and the
heterogeneity function itself, according to our experience in section 5, for a given sample size, functions with more pronounced monotonicity do not need to include a Lagrangian term, while functions with more pronounced concavity need to include a Lagrangian term, and if the sample size is large enough, the inclusion or exclusion of a Lagrangian term does not affect the properties of the resulting function.

3. Inferring QLPRT

In section 2 we use Bootstrap and random forest trees for the hyperparameters of the QLPRT \( M \) to sample the QLPRT, thus passing the samples \( \{M_1, M_2, ..., M_t\} \) to obtain the the distribution of \( \hat{p}(M) \) QLPRT’s hyperparameter-based random forest sampling does not rely on asymptotic normal theory, which makes it more robust even with smaller samples, and we only need to ensure that the number of samples sampled is large enough to use the properties of Bootstrap. The probability distribution function of the sample is close to that of \( p(M) \). The QLPRT is used to obtain a distribution function that is close to the actual \( M \) probability distribution function is essentially a local polynomial regression based on the characteristic quantile with only a primary term, so its statistical properties are essentially the same as those of OLS, generalized linear models such as logistic, generalized moment estimation, and other econometric models, which essentially add methods on quantile point intervals and treatment of the two dummy variable interaction terms of the variable, although of course the same can be added to other covariates. The primary term for the current heterogeneous covariate \( x_k \) as a way of controlling the error, it follows that the double difference method (DID) is nothing more than a special case of QLPRT, where the double difference method contains the treatment variable dummy variable, the treatment time before and after dummy variable, and the interaction term between these two. The QLPRT, where the quantile point is the treatment time point divided by the total sample length, can be interpreted as a heterogeneous treatment effect over time and assumes the presence of a placebo effect at the time point (i.e. the inclusion of a pre- and post-treatment time dummy), i.e. a change in the mean effect due to a measure not receiving treatment, and the double difference method assumes a functional form of two stepwise.

The model form of the QLPRT can also be easily extended to cases where the dependent variable (response) is a discrete variable and takes advantage
of the well-developed experience of statistical inference from linear regression models, for which it is sufficient to change the model form for each subgroup from OLS to Logit or multinomial Logit. If the covariates contain endogenous variables, the QLPRT can consider two approaches, one is to replace the OLS estimator with a GMM, i.e. to add a moment condition for the exogeneity of the errors in addition to the first-order optimality condition for the general minimum mean squared error.

3.0.1. Sampling Bias

Since for \( \tau(x_p|M) \) the estimator for \( M \) and the conditional expectation of \( \tau(x_p) \) as the fully conditional expectation, and \( M \) itself is a discrete variable, then

\[
\lim_{N,K \to \infty} \sum_{t=1}^{K} E(\hat{\tau}(x_p)|M_t) * \hat{p}(M_t) = \lim_{N \to \infty} E(\hat{\tau}(x_p)) = \tau(x_p)
\]

(see the appendix B.1), as for the number of samples per subdivision group (subgroup) \( D \) (see the Appendix B.2), the number of samples per subgroup \( D \) is also theoretically required to satisfy the large sample requirement. The moment estimate of \( \tau(x_p) \) can be obtained by combining the above conditions, and extending the above conclusion to the estimation of \( \tau(x_p) \) and the variance \( D[\tau(x_p)] \).

\[
\lim_{N,K \to \infty} \sum_{t=1}^{M} D(\hat{\tau}(x_p)|M_t)p(M_t) = \lim_{N \to \infty} E(\sum_{t=1}^{K} [E(\hat{\tau}(x_p)) - \hat{\tau}(x_p)]^2|M_t) * \hat{p}(M_t) = D(\tau(x_p))
\]

The above way of moment estimation is based on the Bayesian (Bayes) formulation, i.e. the mean variance of the heterogeneity function is considered as a priori distributed by \( M_t \) as the posterior mean and variance determined by the prior distribution \( M_t \). Thus, for the heterogeneity function \( \tau(x_p) \) the estimates of the mean and variance are the weighted averages of the estimates corresponding to each hyperparameter, and the weights are the probability values of their hyperparameters, understood intuitively and simply as the weights are exactly the current hyperparameters \( M_t \). Based on the above equation, the Lindburg-Levy theorem and assumption (4), it can be deduced that

\[
\lim_{N \to \infty, K \to \infty} \frac{\sum_{t=1}^{K} E(\hat{\tau}(x_p)|M_t) * \hat{p}(M_t) - \tau(x_p)}{\sqrt{\sum_{t=1}^{M} D(\hat{h\tau}(x_p)|M_t)p(M_t)}} \sim N(0, 1)
\]
The proof given in the appendix C shows that the bias in the estimates for the mean and variance will increase with sample size, but remains robust even under small samples, allowing us to construct confidence intervals with fewer statistical errors of both types, defining $\eta$ as the confidence level, and $\Phi^{-1}(\cdot)$ denotes the inverse function of the cumulative distribution function of the standard normal distribution, and the construction of the confidence interval requires finding the bilateral test corresponding to the given confidence level $\Phi^{-1}(1 - \frac{\eta}{2})$ and so the value of $1 - \eta$. Confidence interval is

$$\left[ \sum_{t=1}^{K} E(\hat{\tau}(x_p)|M_t) \ast \text{hatp}(M_t) \pm \Phi^{-1}(1 - \frac{\eta}{2}) \ast \sqrt{\sum_{t=1}^{M} D(\tau(\hat{x}_p)|M_t) \ast \text{hatp}(M_t)} \right]$$

3.0.2. Modeling Bias

When the sample size is insufficient, the above linear assumptions about the first-order Taylor expansion will face a certain modeling error (Modeling Bias) problem, the modeling error is mainly related to the denseness of the quantile interval division, when the quantile interval width is extremely small, the difference between the linear expansion and the true heterogeneity function will be smaller, but blindly expanding the quantile distinction point in order to reduce the modeling error will lead to the sample sparsity problem (Sparsity), which is why we use the choice of optimal $M$ parameters.

3.1. Inferring Heterogeneous treatment effect

Point estimations and confidence intervals have been constructed above for specific treatment effect functions, but so far we have not answered how to distinguish between heterogeneous or homogeneous treatment effect functions and how to construct suitable statistics to give clear statistical inference to the above concepts, but it is clear that tests for heterogeneity are necessarily based on estimates and interval tests for treatment effect functions, and that heterogeneous treatment effects at different $x_k$. The levels of heterogeneous treatment effects must differ significantly across intervals, but we cannot conceptualize heterogeneity directly in section 2 and section 3.1, so first we need a clear definition of heterogeneity in the treatment effect function.

Heterogeneity is defined as the heterogeneity of the variable $x_p$ any two $x_{p,i}, x_{p,j}$ whose corresponding treatment effect functions obey different normal distribution with their params including location parameters (Location parameter) $\mu_{\tau(x_p)}$ and the scale parameter $\sigma_{\tau(x_p)}$. 
In general, most of the literature only considers the differences in the location parameters and tends to ignore the differences in the scale parameters, which are in essence as important as the location parameters for two main reasons: one is that the limitations of the sample size lead to possible bias in the estimation of the location parameters and therefore the scale parameters need to be taken into account, and the other is that even if the location parameters of the treatment effect function at any two points are equivalent, the existence of a large difference in the scale parameters. The second is that even if the location parameters of the treatment effect function at any two points are equivalent, the scale parameter is overestimated if there is a large difference between the true treatment effect function and the estimated treatment effect, and the proximity of the location parameter between the two points will be overestimated.

Given that both the location and scale parameters affect the heterogeneous treatment effect function, the heterogeneity of the treatment effect function is discussed below in two categories, one of which we call weak heterogeneity and the other strong heterogeneity.

Weak heterogeneity means that, for heterogeneous variables \( x_p \). Any two \( x_{p,i}, x_{p,j} \) whose corresponding location parameter for the treatment effect function \( \mu_{\tau}(x_p) \) is different, but the scale parameter \( \sigma_{\tau}(x_p) \) is the same. Strong heterogeneity means that, for heterogeneous variables \( x_p \). Any two \( x_{p,i}, x_{p,j} \) whose corresponding location parameter for the treatment effect function (Location parameter) \( \mu_{\tau}(x_p) \) and the scale parameter \( \sigma_{\tau}(x_p) \) are different. The difference between strong and weak heterogeneity is simply whether or not the scale parameter is assumed to be the same, and it is obviously easier to construct and test the statistic for weak heterogeneity. For weak heterogeneity, the Wilcoxon-Mann-whitney rank-sum test (W-M-W test) can be applied, which assumes that the distribution of the data is continuous and similar and the distribution of subgroups is similar. Since we can only obtain the estimated location and scale parameters for different subgroups, we need to obtain the treatment effects for different subgroups by generating samples through random simulations of \( \tau(x_{p,j}) \).

The zone groups are divided in the same way as in section 2, using quantile points for the division, and as for the hyperparameters \( M \) we recommend choosing the sample size of each block. \( n \) can be determined specifically depending on the amount of computational resources, e.g. 500, and once these parameters are determined we can assume that \( \tau(x_{p,j}) \) obeys a normal distribution and draws separately \( n \times M_{\text{Mode}} \) then mix all the samples and calculate.
the rank of each sample in order of size, taking the average rank for samples of the same size, then calculate the rank sum of squares for each group $R_j^2$, due to the H-value of the K-W test:

$$12 * \sum \frac{R_j^2}{n * n * M_{\text{Mode}}} - 3 * [n * M_{\text{Mode}} + 1] \sim \chi^2(M_{\text{Mode}} - 1)$$

Thus by looking up $\chi^2$ distribution table to test the original hypothesis. As for the test of $\sigma_{\tau(x_p)}$ test, if we only want to test one item of variance, then we can test for homoscedasticity by subjecting the variance corresponding to each sample to the K-W test mentioned above, and if the K-W test for the variance of the sample fails, i.e. the different $\tau(x_{p,i})$ variances are significantly different, then we should not continue to adhere to the conclusions of the mean K-W test and should consider the test proposed below that tests both the mean and the variance, notably by considering separately $\mu_{\tau(x_p)}$ and $\sigma_{\tau(x_p)}$ is not meaningful, the reason being that when the variances are significantly different but the means are close, treating the effect function $\tau(x_{p,i}) = \tau(x_{p,j})$ with no small probability, it is necessary to also consider $\mu_{\tau(x_p)}$ and $\sigma_{\tau(x_p)}$ and determine the probability value $P[\tau(x_{p,i}) = \tau(x_{p,j})] = \frac{12}{n^2 * M_{\text{Mode}}} - 3 * [n * M_{\text{Mode}} + 1] \sim \chi^2(M_{\text{Mode}} - 1)$ is a more straightforward and efficient method, when multiple values are obtained by sampling $\mu_{\tau(x_{p,1})}, \mu_{\tau(x_{p,2})}, \ldots, \mu_{\tau(x_{p,S})}$ and $\sigma_{\tau(x_{p,1})}, \sigma_{\tau(x_{p,2})}, \ldots, \sigma_{\tau(x_{p,S})}$ when the first one is taken as the other comparator. If the null hypothesis holds, then given the parameters $\mu_{\tau(x_{p,1})}$ and $\sigma_{\tau(x_{p,1})}$ and the corresponding sample either $\tau(x_{p,1})$ or $\tau(x_{p,j})$, the resulting probability values should be close, i.e.

$$H_0 : P(\tau(x_{p,1})|\mu = \mu_{\tau(x_{p,1})}) = P(\tau(x_{p,j})|\mu = \mu_{\tau(x_{p,1})}, \sigma = \sigma_{\tau(x_{p,1})})$$

is easy to compute when generating a total $S$ samples, the sample variance can be calculated based on the null hypothesis as follows:

$$D(P(\cdot)) = \frac{\sum_{s=2}^S (P(\tau(x_{p,1})|\mu = \mu_{\tau(x_{p,1})}, \sigma = \sigma_{\tau(x_{p,1})}) - P(\tau(x_{p,j})|\mu = \mu_{\tau(x_{p,1})}, \sigma = \sigma_{\tau(x_{p,1})}))^2}{S - 2}$$

We therefore only need to construct the following t-statistic, the

$$\sqrt{\frac{\sum_{s=2}^S (P(\tau(x_{p,1})|\mu = \mu_{\tau(x_{p,1})}, \sigma = \sigma_{\tau(x_{p,1})}) - P(\tau(x_{p,j})|\mu = \mu_{\tau(x_{p,1})}, \sigma = \sigma_{\tau(x_{p,1})}))^2}{S - 2}} \sim t(S - 2)$$

Then given the confidence level $\eta$ a t-test is sufficient.
4. Simulation Studies

4.1. Experimental Setup

We assume that a total of $n$ where for each individual the dependent variable satisfies the assumption of linearity and is affected by a total of four variables, of which variables 1 and 2 are heterogeneous, but whose heterogeneity treatment effect functions are unknown and we can only observe the covariates and the dependent variable (response)

$$Y_i = \beta_0 + \beta_1 * f_1(x_{1,i}) + \beta_2 * f_2(x_{2,i}) + \beta_3 * x_{3,i} + \beta_4 * W_i + \epsilon_i$$

where $W_i \in \{0, 1\}$ denotes the processing variable $\epsilon_i \sim N(0, \sigma^2)$, $\sigma^2 < +\infty$ for covariates $x_1, x_2, x_3$ each satisfying the condition of independent identical distribution (i.i.d.), specifically, $x_1 \sim N(10, 50), x_2 \sim N(20, 60), x_3 \sim Gamma(1, 30), W \sim Bernoulli(p = 0.5)$. For convenience, we define each coefficient separately as $\beta_0 = 20, \beta_1 = 10, \beta_2 = 20, \beta_3 = 30$. To show the robustness of the QLPRT to the estimation of heterogeneity treatment effect functions, we need to construct functions with significant monotonicity and functions with significant concavity as heterogeneity functions, respectively; an example of satisfying the former is to take Logistic function with mean 10 and standard deviation 50 of $x_1$ as parameters, also known as the Sigmoid function, while the latter can be taken as $x_2$ with a mean of 20 and a standard deviation of 60 as the parameters of the normal distribution probability density function.

$$f_1(.) = \frac{1}{\sqrt{2 \pi \cdot 50}} e^{-\frac{(x_1-10)^2}{2 \cdot 50^2}}$$

$$f_2(.) = \frac{e^{\frac{(x_2-20)}{60}}}{1 + \frac{(x_2-20)}{60}}$$

Our experiment is divided into two parts, with the first part testing the estimated heterogeneity treatment effect $\hat{\tau}(x)$ with the actual heterogeneity treatment effect $\tau(x)$ The correlation and asymptotic properties between the QLPRT and the actual heterogeneity treatment effect $\tau(x)$ are computed at sample sizes of 200, 500, 1000 and 2000 in order to show the robustness of the QLPRT at each sample size

$$\rho(\hat{\tau}(x), \tau(x))$$
Figure 1: QLPRT and QLPRT Variance

\[ \frac{\tau(x) - \hat{\tau}(x)}{\tilde{D} \left[ \tau(x) - \hat{\tau}(x) \right]} \]

where \( \rho \) denotes the Pearson’s correlation coefficient between the two variables. The first part focuses on the mitigation of modelling errors (Modeling Bias) by QLPRT, and we plot box plots (boxplots) of the distribution of errors for sample sizes of 500, 1000, 2000 and 5000. In addition, section 3.1 shows that the estimated variance of the true heterogeneity function is not mitigated by QLPRT, so in this section we generate separate sample sizes for QLPRT and QLPRT. By comparing the effects of sample size and adoption of QLPRT on the estimated confidence intervals for the heterogeneity treatment effects using the QLPRT-based model, we expect to see the conclusion that adopting QLPRT approximates the true confidence intervals more closely than it underestimates the size of the true intervals.

4.2. Results

Figure1 Plot the solid column part of the boxplot according to the scale of 25%-75% and the shadow part according to the scale of 1%-99%. Figure 1 shows the distribution of the residuals of QLPRT and QLPRT under different sample sizes, in general, the residuals of QLPRT and QLPRT basically conform to the symmetric distribution characteristics rather than fat-tailed or left-right skewed characteristics; %Table 1 shows the pearson correlation
Table 1: Correlation and Variance

| Sample Size | N = 200  | N = 500  | N = 1000 | N = 2000 | Convergence |
|-------------|----------|----------|----------|----------|-------------|
| QLPRF − Corr | 0.9287   | 0.9610   | 0.9831   | 0.9914   | 0.9363      |
| QLPRF − Std  | 416.3300 | 215.6105 | 133.5128 | 92.8160  | 0.6580      |
| QLPRF − EstimatedStd | 410.4338 | 213.6727 | 132.9188 | 92.5716  | 0.6527      |
| QLPRT − Corr | 0.9422   | 0.9643   | 0.9750   | 0.9856   | 0.5922      |
| QLPRT − Std  | 274.7046 | 210.2686 | 166.2238 | 119.6150 | 0.3565      |
| QLPRT − EstimatedStd | 270.3471 | 209.2423 | 165.6176 | 119.3302 | 0.3508      |

coefficients of the QLPRT and QLPRF with the true heterogeneity treatment effect function under different sample sizes and the standard deviations under the assumption of mean of 0, respectively, from the figure, the fitting performance of both QLPRT and QLPRF increases significantly with increasing sample size, and the results show that even with a small total sample size, a degree of linear correlation with the true function of more than 90% can be obtained, and when the sample size exceeds 500, the Pearson correlation coefficients of QLPRT and QLPRF with the true function are close to 0.96, basically restoring the true function, and the convergence rate item on the right-hand side indicates that the linear correlation coefficients and standard deviations of QLPRF and QLPRT. The convergence rate of both the linear correlation coefficient and standard deviation of QLPRF and QLPRT are around 0.5, which also shows that QLPTRF converge faster with the sample size. The difference between the true error and the prediction error decreases with increasing sample size, so both QLPRT and QLPRF have good ability to fit heterogeneous treatment effect functions.

The different characteristics of QLPRT and QLPRF are shown in both Figure 1 and Table 1. At smaller samples, i.e. sample size less than 500, QLPRT has stronger correlation with the true function and smaller prediction error, and conversely at larger samples, QLPRF has better fitting performance in terms of convergence rate, both in terms of correlation coefficient with the true function and prediction error, QLPRF has convergence rate is higher than that of QLPRT, but a point of possible concern is that even though the difference between the prediction and true errors decreases exponentially, the difference between the true and prediction errors is always greater than zero, suggesting that at small sample sizes, direct statistical inference using the variance estimated by the model will probably overestimate the significance of the coefficients and that this part of the excess error
actually arises from modelling error, i.e. we use a linear function to fit the curve with a higher order term that is easy to ignore when the sample is small. An important purpose of the QLPRF we propose in section 2.2, in addition to improving the fit performance, is to improve the robustness of statistical inference for the QLPRT in small samples, where the mean of the true function is predicted using the QLPRT and the variance of the QLPRF is used for statistical inference, as the QLPRF takes into account multiple functional forms and therefore its statistical inference is more reliable. A possible problem is that the prediction error of QLPRF in a small sample differs somewhat from the true error of QLPRT, as $N = 200$.

Is it reasonable to assume that the difference between QLPRT-Std=274.7 and $QLPRF - EstimatedStd = 410.4$ for $N=200$ is not a small amount? In fact we believe that this part of the difference is necessary, there are many possibilities for curves fitted with small samples, e.g. exponential functions, linear functions, etc. all have similar monotonicity characteristics as the logistic curves used in our simulations, and similarly the quadratic curves are similar in shape to the probability density functions of the logistic distributions. We use QLPRF by increasing the modelling error of estimates to obtain more robust estimates of the standard deviation, greatly reducing the two types of inference errors that can occur with small samples.

Therefore, from the results, QLPRF is more suitable for statistical inference under smaller samples, while QLPRT and QLPRF perform similarly under large samples, but QLPRF will consume a lot of computational resources, and therefore is not suitable for regressions with very large dataset sizes, such as new feature reviews of mobile applications that will face tens of millions of users. For example, a new feature review of a mobile application will involve tens of millions of users, where user heterogeneity is the purpose of the study but the dataset size is very large, in which case QLPRT can be used instead of QLPRF because the fitting performance of QLPRF and QLPRT is essentially the same at this data size, but QLPRT is much faster than QLPRF and other machine learning algorithm-based models.

5. Empirial Experiments

To demonstrate empirically the performance and validity of our proposed methodology, we consider QLPRT and QLPRF regressions using data from the National Supprted Work Project (NSW). The dataset we use consists of the Current Population Survey (CPS) and the Panel Study of Income
Dynamics (PSID). This dataset was used by LaLonde (1986)\cite{40} to study the differences in estimates of the actual effects of labour training programmes by different estimators (estimators) using artificially constructed non-experimental control group data to replace the original control group data. Dehejia and Wahba (1999)\cite{41} re-estimated the NSW data using propensity score estimation and compared the effects of matching and stratification on the estimation of causal effects. [Angrist and Pischke (2009)] used OLS estimation on the NSW-CPS data and concluded that the OLS estimate of the treatment effect for this labour item was close to 1794. [Tymon Sloczyński (1999)\cite{42}] found that OLS estimates of the treatment effect when there is heterogeneity in the treatment effect are a convex combination of the two parameters of the control and control group proportions and applied them to the NSW data.

National Supported Work (NSW) was a labour training programme run by the Manpower Demonstration Research Corporation (MDRC) in the mid-1970s with the primary aim of providing practical experience in a job that would help workers with no significant advantage in the labour market. Unlike other federally sponsored employment and training programmes in the US, the NSW programme uses random assignment to determine whether each eligible participant can receive training, rather than a blanket acceptance, which ensures that control and control groups can be randomised without interference from other factors. A total of 6,616 participants were recruited in ten cities in the USA, including Chicago, New Jersey and New York, and were placed in a control group and a treatment group. The participants are those who are enrolled in the project and the trainees are those who enter the project in the treatment and control groups respectively. Each team is assigned an NSW consultant to evaluate and advise them on their work. Salaries for the NSW programme are lower than the average labour market rate for equivalent work, but performance pay is paid when they over-achieve their work. At the end of the programme, participants will have to leave the job offered and look for a new job.

Participants in the NSW programme already have many different characteristics before they enter the programme, starting with geographical differences. For example, some participants live in Harvard (Hartford) generally having a work experience focused on petrol stations or print shops. Gender differences were also evident, with male participants more often in the manufacturing industry, while most of the female participants had experience in the service sector. In this paper, we focus on the heterogeneous treatment
effects of age and years of education on the difference in earnings before and after training. With a mean age of 33.37 and a standard deviation of around 7.43, it is clear that most participants in the NSW program are in their prime, but we are sometimes interested in the net effect of the program on different age groups of workers, or we want to estimate the treatment effect for a certain interval with the data available. For example, the net effect of labour training for female workers at the age of 45, although there is not much data on this age group in the NSW, we can use the heterogeneous treatment effect for age to make common sense assumptions and estimates of the treatment effect for the out-of-sample interval, where the ability to learn and the likelihood of being accepted in the labour market decreases with age, and therefore Based on this assumption we can easily obtain many valuable and more accurate answers based on the estimated heterogeneous treatment effect function; the average number of years of education of participants in the NSW project is around 10.30 years, with a standard deviation of 1.92, indicating that workers who need to participate in the NSW project to obtain a job are less educated. The heterogeneous treatment effect function for different years of education can also be used to estimate the treatment effect for workers with higher educational attainment based on the common sense assumption that the heterogeneous treatment effect function also conveniently provides us with more information about the function image, such as changes in slope, results that would not be possible by simply estimating participants in groups.

Table 2 shows a descriptive table of the NSW program data we used, where Treat indicates whether the individual entered the program, roughly 41% of individuals participated in the program. Age indicates the age of the individual and the age group of the individual is mostly at the youth stage, while the oldest individual is not more than 50 years old. The individual’s years of education is basically around 10 years and the standard deviation is small. Black indicates whether the individual is Black, the remaining ethnicity has a proportion of 10% of Hispanic. Black indicates an indicator of whether the individual is Black, and about 80% of the project’s members are black, with another 10% of the remaining ethnicity being Hispanic. Married indicates whether the individual is married, and nodegree indicates whether the individual has earned a college degree, re75 and re78 indicate the individual’s salary in 1975 and 1978 wages, respectively, and wage change indicates the wage differential in 1978 minus 1975.
The main dependent variable we use is the difference between the participant’s wage in 1978 and the wage in 1975 $\Delta wage$ and to demonstrate the compatibility of the QLPRT and QLPRF for different types of dependent variables (responses), we constructed dichotomous variables on whether program participants had higher wages in 1978 than in 1975, i.e., when $Re_{78} > Re_{75}$ takes 1 and 0 otherwise. For the dichotomous variable model we use Logit regression as the base model, and the other steps are the same as for QLPRT and QLPRF under general OLS. Based on the tests presented in section 3 we identify variables that may be heterogeneous, and in addition we plot the heterogeneity using OLS and Logit models under QLPRT and QLPRF respectively. The estimated treatment effect functions for heterogeneity, where the treatment effect functions based on the dichotomous dependent variable and the logit model represent the average probability that participation in the program leads to an increase in income after three years under that heterogeneous variable, and for variables with repeated values, we calculate the mean of their values taken at the same level and report their differences at the same level.

The main results are presented in table 2, figure2 and figure3, which show the results of the Mahn-Whitney test for the mean variance of the $M$ parameter, and for the mean and variance of the estimated heterogeneity treatment effect function, with the figures in the table corresponding to the statistics $p$ values, as well as the results of the likelihood ratio test for simultaneous
Table 3: QLR Regression on NSW data

| Continuous Model | M Mean | M Std | Mahn-Whitney Test | Variance Mahn-Whitney Test | Likelihood Ratio Test |
|------------------|--------|-------|-------------------|---------------------------|-----------------------|
| Education-QLPRT  | 3.22   | 1.4668| 0.0019            | 0.01113                   | 0.8271                |
| Age-QLPRF        | 10.32  | 4.7558| 0                  | 0                         | 1.0                   |

| Logit Model     | M Mean | M Std | Mahn-Whitney Test | Variance Mahn-Whitney Test | Likelihood Ratio Test |
|-----------------|--------|-------|-------------------|---------------------------|-----------------------|
| Education-QLPRT | 3.22   | 1.4668| 0.0019            | 0.01113                   | 0.8271                |
| Age-QLPRF       | 1.2    | 0.6   | 0                 | 0                         | 1.0                   |

In this paper, we consider age and years of education as the heterogeneity variables to be tested, with the other variables entering the model as control variables. The top of the table shows the results of the OLS-based regression model, while the bottom shows the results of the Logit-based regression model. At first, we have estimated the education using the QLPRT. The results of the M-W test for the mean and variance of the treatment effect function for Education indicate that there is a greater likelihood of significant differences in the estimated variance of the treatment effect across the population with different levels of education. As the sample size is only around 770, the QLPRT may be subject to modelling bias, and to avoid the problem of misjudgement, we use the QLPRF for further estimation, and as we expected, the level of heterogeneity testing for education was not very different from QLPRT, but the level of heterogeneity testing for age was quite different from QLPRT, both for the M-W statistic dealing with the mean and variance of the effect function alone and for the likelihood ratio statistic testing both the mean variance of $p$. The reason for the different findings on age heterogeneity between QLPRT and QLPRF is that, as described in our simulation study in section 3 and section 4, QLPRF has less modelling error and more accurate treatment effects than QLPRT for larger sample sizes. The test statistic based on its estimates is also more reliable.

The reliability of the QLPRF is not limited to continuous-valued dependent variable models, but is also well demonstrated for discrete variable models such as dichotomous variables. The results of the Logit model in Table 3 share common features with the OLS model, for example the estimated mean and variance heterogeneity M-W statistics for the treatment effect function for education both indicate possible heterogeneity, but the likelihood ratio
test rejects the heterogeneity hypothesis more strongly, which is more evident in the QLPRF model, and the p-values for both EDUCATION and AGE on the M-W and likelihood ratio statistics are relatively large, suggesting that the NSW program is not influenced by participants’ age or educational experience in terms of the likelihood of raising earnings, a finding that is not unexpected and, as we mention below, is related to treatment effects of the NSW project.

Figure 2: Continuous Response Model

Figure 3: Binary Response Model

Figure 2 and Figure 3 plot the treatment effect functions for education under the OLS and Logit models, respectively, where the red line shows the results for QLPRF, the black line shows the results for QLPRT, and the thick line shows the mean of the function, while the dashed part shows the mean of the function plus 1.96*standard deviation of the treatment effect for different individuals with the same number of years of education, indicating the heterogeneity of the treatment effect. The QLPRT and QLPRF estimated treatment effect functions for education in Figure 2 are basically
on an upward trend, but mostly at education greater than or equal to 15 years. This may indicate that the higher the number of years of education, the more pronounced the increase in earnings of the participant after receiving training from the programme, but there is a threshold effect for the treatment effect of years of education. The results in Figure 3 are the same as we expected in Table 2, where the number of years of education did not increase the probability that a participant would receive a salary increase or find a job from unemployment after participating in the NSW programme, especially for the QLPRF, and furthermore, the treatment effect estimated by the Logit-QLPRT fluctuated significantly after more than 15 years of education. However, the Logit-QLPRF estimation smoothed them out. The reason for the difference in findings due to OLS-QLPR and Logit-QLPR is related to the construction of the dependent variable. As seen in Figure 2, the treatment effects for education are essentially above zero, i.e. the NSW program essentially increases the wages of participants, and therefore the dichotomous dependent variable reflecting the probability of wage increases and employment does not reflect the treatment effects of the NSW program.

6. Discussion

In this paper, we propose a regression tree structure based on a linear model and scalable to generalized linear and binary discrete models for solving regression problems with heterogeneous treatment effects, using quantile of continuous variables as branching variables and conventional statistics as branching criteria, which we call QLPRT. QLPRT is essentially a local polynomial non-parametric method, but it is simpler and more robust than other non-parametric methods such as kernel methods and K-nearest neighbour estimation, and does not strictly restrict the types of distributions of dependent and covariates, so all QLPRT-based tests have a more solid theoretical foundation; the disadvantage of QLPRT is the potential for modelling error in small samples, so we further propose a random forest model QLPRF consisting of multiple QLPRTs based on bootstrap sampling. In addition to the modelling error, the bootstrap method of QLPRF is also beneficial in alleviating the possible selection bias problem (selection bias). In our simulation experiments, both QLPRT and QLPRT largely restore the heterogeneity function we assume and are asymptotically efficient in terms of their variance.

The main problem with the QLPRT and QLPRF is that its assumptions
are sometimes more stringent, one of which is the selection bias problem, see [Chernozhukov et al.(2015)](43), [Taylor and Tibshirani(2015)](44), when the sample is large enough and the QLPRF does reduce selection bias when the proportion of individuals with selection bias is small, but in real-world problems, completely randomised trials are almost impossible to exist due to technical or cost issues; the second is our assumption that any two heterogeneous variables are independent, which may not apply to treatment effects where two heterogeneous variables jointly affect each other, and therefore such independence violations need to be caution.

Follow-up research could consider the following: the former is to address or reduce the effect of selection bias on OLS estimators; the latter is to extend the focus on heterogeneity treatment effects for a single variable to the multivariate case to avoid possible omitted variable cases, and three is to extend the current model to a multivariate discrete selection model, using the multivariate Logit models, for example, which have more than one set of parameter vectors and where it is equally challenging to interpret treatment effects in a multivariate discrete choice model.

References

[1] L. Breiman, Consistency for a simple model of random forests (2004).

[2] S. Wager, S. Athey, Estimation and inference of heterogeneous treatment effects using random forests, Journal of the American Statistical Association 113 (2018) 1228–1242.

[3] S. Athey, G. Imbens, Recursive partitioning for heterogeneous causal effects, Proceedings of the National Academy of Sciences 113 (2016) 7353–7360.

[4] B. Efron, Estimation and accuracy after model selection, Journal of the American Statistical Association 109 (2014) 991–1007.

[5] S. Wager, Asymptotic theory for random forests, arXiv preprint arXiv:1405.0352 (2014).

[6] S. F. Assmann, S. J. Pocock, L. E. Enos, L. E. Kasten, Subgroup analysis and other (mis) uses of baseline data in clinical trials, The Lancet 355 (2000) 1064–1069.
[7] D. I. Cook, V. J. Gebski, A. C. Keech, Subgroup analysis in clinical trials, Medical Journal of Australia 180 (2004) 289.

[8] C. De Chaisemartin, X. d’Haultfoeuille, Two-way fixed effects estimators with heterogeneous treatment effects, American Economic Review 110 (2020) 2964–96.

[9] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[10] S. Wager, T. Hastie, B. Efron, Confidence intervals for random forests: The jackknife and the infinitesimal jackknife, The Journal of Machine Learning Research 15 (2014) 1625–1651.

[11] B. Efron, C. Stein, The jackknife estimate of variance, The Annals of Statistics (1981) 586–596.

[12] N. Meinshausen, G. Ridgeway, Quantile regression forests., Journal of Machine Learning Research 7 (2006).

[13] J. Sexton, P. Laake, Standard errors for bagged and random forest estimators, Computational Statistics & Data Analysis 53 (2009) 801–811.

[14] J. Duan, et al., Bootstrap-based variance estimators for a bagging predictor. (2011).

[15] G. Biau, Analysis of a random forests model, The Journal of Machine Learning Research 13 (2012) 1063–1095.

[16] E. Scornet, G. Biau, J.-P. Vert, Consistency of random forests, The Annals of Statistics 43 (2015) 1716–1741.

[17] L. Mentch, G. Hooker, Quantifying uncertainty in random forests via confidence intervals and hypothesis tests, The Journal of Machine Learning Research 17 (2016) 841–881.

[18] H. A. Chipman, E. I. George, R. E. McCulloch, Bart: Bayesian additive regression trees, The Annals of Applied Statistics 4 (2010) 266–298.

[19] D. P. Green, H. L. Kern, Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees, Public opinion quarterly 76 (2012) 491–511.
[20] J. L. Hill, Bayesian nonparametric modeling for causal inference, Journal of Computational and Graphical Statistics 20 (2011) 217–240.

[21] J. Hill, Y.-S. Su, Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children’s cognitive outcomes, The Annals of Applied Statistics (2013) 1386–1420.

[22] Y. V. Tan, J. Roy, Bayesian additive regression trees and the general bart model, Statistics in medicine 38 (2019) 5048–5069.

[23] Y. W. Teh, Dirichlet process, 2010.

[24] J. Kocijan, R. Murray-Smith, C. E. Rasmussen, A. Girard, Gaussian process model based predictive control, in: Proceedings of the 2004 American control conference, volume 3, IEEE, 2004, pp. 2214–2219.

[25] J. C. Foster, J. M. Taylor, S. J. Ruberg, Subgroup identification from randomized clinical trial data, Statistics in medicine 30 (2011) 2867–2880.

[26] R. Friedberg, J. Tibshirani, S. Athey, S. Wager, Local linear forests, arXiv preprint arXiv:1807.11408 (2018).

[27] X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, B. Li, Subgroup analysis via recursive partitioning, Journal of Machine Learning Research 10 (2009).

[28] K. Imai, M. Ratkovic, Estimating treatment effect heterogeneity in randomized program evaluation, The Annals of Applied Statistics 7 (2013) 443–470.

[29] H. I. Weisberg, V. P. Pontes, Post hoc subgroups in clinical trials: Anathema or analytics?, Clinical trials 12 (2015) 357–364.

[30] A. Beygelzimer, J. Langford, The offset tree for learning with partial labels, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 129–138.

[31] R. H. Dehejia, Program evaluation as a decision problem, Journal of Econometrics 125 (2005) 141–173.
[32] K. Hirano, J. R. Porter, Asymptotics for statistical treatment rules, Econometrica 77 (2009) 1683–1701.

[33] C. F. Manski, Statistical treatment rules for heterogeneous populations, Econometrica 72 (2004) 1221–1246.

[34] D. Bhattacharya, P. Dupas, Inferring welfare maximizing treatment assignment under budget constraints, Journal of Econometrics 167 (2012) 168–196.

[35] M. Taddy, M. Gardner, L. Chen, D. Draper, A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation, Journal of Business & Economic Statistics 34 (2016) 661–672.

[36] J. S. Neyman, On the application of probability theory to agricultural experiments. essay on principles. section 9. (translated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480), Annals of Agricultural Sciences 10 (1923) 1–51.

[37] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies., Journal of educational Psychology 66 (1974) 688.

[38] N. Fortin, T. Lemieux, S. Firpo, Decomposition methods in economics, in: Handbook of labor economics, volume 4, Elsevier, 2011, pp. 1–102.

[39] D. Elder-Vass, The causal power of social structures: Emergence, structure and agency, Cambridge University Press, 2010.

[40] R. J. LaLonde, Evaluating the econometric evaluations of training programs with experimental data, The American economic review (1986) 604–620.

[41] R. H. Dehejia, S. Wahba, Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, Journal of the American statistical Association 94 (1999) 1053–1062.

[42] T. Słoczyński, Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights, The Review of Economics and Statistics (2020) 1–27.
[43] V. Chernozhukov, C. Hansen, M. Spindler, Post-selection and post-regularization inference in linear models with many controls and instruments, American Economic Review 105 (2015) 486–90.

[44] J. Taylor, R. J. Tibshirani, Statistical learning and selective inference, Proceedings of the National Academy of Sciences 112 (2015) 7629–7634.