**Cross Validation for Measuring Model Accuracy**

Our approach to evaluate the models under consideration is to simulate data from a known data generating process which incorporates treatment effect heterogeneity along one or more moderating variables. We fit the models using a part of the data and compute the error between true treatment effect and models’ prediction using the remaining data.

We rely on K-fold cross validation to estimate the average error of the models in correctly identifying treatment heterogeneity. Figure X illustrates the cross validation approach we will use.

**Exhibit X: Cross validation approach to compute the accuracy of the proposed models**

Testing

Training

Training

Testing

Training

3

4

.

.

K-1

Testing

Training

In K-fold cross validation the dataset is randomly split into K partitions. The models are built using K-1 partitions and evaluated on the testing partition. The testing partition is cycled through in K iterations as shown in Exhibit X. In each iteration the Mean Squared Error between true treatment effect and predicted treatment effect is computed from the observations in the testing partition. Finally, the MSE’s from all iterations provide an estimate of error of the model in correctly predicting individual unit treatment effect.

Formally, let be the full data sample generated by a data generating process . The data generating process consists of a treatment assignment function and potential outcome function . In the context of this proposal, we consider random assignment designs with marginal treatment probability of 0.5. Therefore, treatment assignment function randomly assigns half of the sample to treatment and half of the sample to control. The potential outcome function has a treatment effect function that determines the individual treatment effect for a given unit based on a number of moderating variables. In the K-fold cross validation, is randomly split into partitions . In iteration , units in the partition is designated as the testing data and the units in the remaining partitions are used as training data. The model built on the training data is applied on the the testing data to compute mean squared error between true treatment effect and predicted treatment effect for the units in testing data. The cross validated error of the model is the average MSE in all testing partitions and can be written as below.

Where is the predicted treatment effect for unit I from the model built using the training data and is the number of units in partition . is our primary criteria in comparing the performance of the models. Smaller the better the performance of a model in correctly capturing treatment effect heterogeneity.

**Simulation Analysis to Address Research Questions**

To assess the relative performance of the proposed algorithms we propose to carry out simulation studies. For each research question in Part I, we proposed a suitable simulation study design that helps to assess the accuracy of models under consideration. For all simulations we consider simulated datasets of size N = {1000, 5000, 10000} observations. We chose these samples sizes since typical medium sized RCTs involve subjects in those ranges. We vary the number of observations to assess the importance of sample size on the accuracy of the models. We propose to use K = 10 (10-fold) for cross validation following the common practice in machine learning literature. In all designs the marginal treatment probability is P = 0.5 and Q denotes the number of variables in the model. For RQ1-RQ4, we consider a model for the mean effect and for the treatment effect, and the potential outcomes are written, for ,

where , and the are independent of and one another.

In the following we describe the simulation designs we use to address each of the research questions in Part 1.

* **RQ1: How accurate are the heterogeneous treatment effect estimation models under different distributions of treatment effect moderating variable(s)?**

To answer RQ1, we consider a simple design with two variables. The mean function involves both of the variables but the treatment effect heterogeneity is defined only as function of one of the variables. The primary design specification is below:

Our goal is to evaluate how well the proposed methods can elicit the true structure of the heterogeneity in the treatment effect under different distributional specifications of the moderator variable . One important way the distribution of matters is it determines how much data is available at different levels of to learn the true function of treatment effect. For example, when is distributed according a uniform distribution, one has the same amount of data available to learn the relationship between treatment effect and at any level of . On the other hand, if is distributed according to a normal distribution, the available data is sparse at the tails of the distribution of , thus limiting the ability of the models to learn the relationship between treatment effect and . To systematically explore the relationship between the distribution of treatment effect moderator and the empirical ability of the models to recover the true treatment effect heterogeneity, we consider three distributional specifications of (note that all *x’s* are independent of each other):

* + Design 1: ;
  + Design 2: ;
  + Design 3: ;

**Exhibit 2: Distribution of under three design specifications**



Exhibit 2 illustrates the three specification of x1. While uniform distribution spreads data evenly for model fit, normal and beta distributions present more data in the center and in the tails respectively. Our primary metric to compare the performance of the methods is the cross validated MSE. The sparsity of examples to learn the true relationship between the moderator variable and treatment effect in the tails for normal distribution and center for beta distribution can lead to higher cross validated MSE. We present the for each of the Designs for each of the models for N = {1000, 5000, 10000}. This means in total we compute 36 values to compare models. For each design and the size of sample, we highlight best in bold. Exhibit 3 presents a sample illustration of the presentation of results from the simulation studies of RQ1.

**Exhibit 3: Cross validated MSE across four methods for designs varying the distribution of the moderator variable**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Design | Causal Tree | | | Causal Forest | | | Boosted Trees | | | X-Learner | | |
| N | **1000** | **5000** | **10000** | **1000** | **5000** | **10000** | **1000** | **5000** | **10000** | **1000** | **5000** | **10000** |
|  | 0.31 | 0.28 | 0.25 | **0.25** | 0.24 | 0.23 | 0.28 | 0.26 | 0.25 | 0.41 | **0.21** | **0.20** |
|  | 0.38 | 0.39 | 0.38 | 0.34 | **0.33** | **0.31** | **0.33** | 0.37 | 0.35 | 0.48 | 0.37 | 0.36 |
|  | 0.41 | 0.40 | 0.40 | **0.39** | **0.39** | **0.38** | 0.40 | 0.40 | 0.38 | 0.51 | 0.45 | 0.40 |

Note: Numbers are illustrative and are not based on real data.

Along with cross validated MSE values, we also produce visualizations that can inform the systematic presence of bias or lack of it. For example, Exhibit 4 presents scatterplots of predicted and true treatment effects for the Causal Tree model when the moderating variable is distributed according to a uniform distribution (in a) and a normal distribution (in b). Redlines are the 45 degree lines which stand for perfect prediction of the true treatment effect. It is easy to see that accuracy of the Causal Tree falls off significantly in the right chart as one move away from the mean of the distribution. This is expected since there is less data to learn the true relationship as one moves away from the mean of the distribution of the moderating variable. In the tails there is a downward bias in the magnitude of the predicted treatment effect.

**Exhibit 4: True vs predicted treatment effects under two different distributional specifications of the moderator variable**

1. **x1 ~ U[-2,2] (b) x1 ~ N(0,1.5)**



* **RQ2: How accurate are the heterogeneous treatment effect estimation models for different functional form specifications?**

In RQ2, our primary interest is in evaluating the effectiveness of the methods under different functional forms of the treatment effect function. Apriori it is difficult to know if the treatment effect heterogeneity is linear, non-linear, or has discontinuities. Discontinuities can exists in the relationship between a moderator variable and treatment effect if the individuals or units receiving policy can only realize the benefits when the moderating variable is above some threshold. For example, only those students who have access to certain technology resources may be able to realize the benefits of certain online tutorial intervention. Conditional on having technology resources, the treatment effect may be increasing in the quantity of resources. Thus, it is important to investigate if the proposed methods can accurately recover different types of functional relationships between treatment effects and moderator variables. To this end, we consider the following design specifications in which we systematically vary the specification of treatment effect functional form.

* + Design 4:
  + Design 5:
  + Design 6:
  + Design 7:
  + Design 8:

These designs include linear, non-linear, step functions, and mixed functional specifications for the treatment effect specification.

As in research question 1, we compare the performance of different models by comparing the cross validated MSE from each model for each of the designs. We also present the visualizations illustrating the the systematic biases in the prediction of

* **RQ3: How robust are the heterogeneous treatment effect estimation models in the presence of large number of nuisance parameters?**
  + Design 9:
  + Design 10:
  + Design 11:
  + Design 12:
* **RQ4: How accurate are the models when the treatment effect varies along multiple moderator variables?**

Research questions 1-3 considered designs where treatment effect varied only along one moderator variable. Research Questions 4 explores the accuracy of models when treatment effects are influenced by more than one moderator variables. Intuitively, this is a much harder test compared to the previous tests since data requirement grow significantly to lean multi-dimensional nature of the relationship between moderator variables and treatment effect.

* + Design 13:
  + Design 14:
  + Design 15:
* **RQ5: How robust are the models when a number of moderator variables are unobserved?**

It is well known social science research that we generally do not have access to the universal set of variables that might be influencing how an intervention might affect the outcome of interest. Often, researchers assume away the influences of these unmeasured variables as part of symmetric error terms. Due to the pervasiveness of the issue of unobservable variables, it is important to understand

* + Design 16:
  + Design 17:
  + Design 18: