Super Dummy Variables

Kweku A. Opoku-Agyemang*

June 2023

Abstract

To limit noncompliance and attrition issues, this paper introduces a treatment dummy variable concept for data contexts where the treatment status of an individual is not fully observed or determined by the researcher, but depends on how much the instrument affects the probability of receiving the treatment. I first show that, compared to a standard dummy variable, these special dummy variables improve the precision and efficiency of estimating the Complier-Average Causal Effect (CACE). In such cases, using a standard dummy variable to indicate the treatment status may not capture the true causal effect of interest, since some individuals may not comply with their assignment or may drop out of the program. For the heterogeneous treatment effects of the new variables, I present a *transformed potential outcome forest* algorithm, a variant of the random forest algorithm that splits the nodes according to a criterion that maximizes the variance of the transformed potential outcomes.

*Machine Learning X Doing. Email: kweku@machinelearningxdoing.com. I thank several participants of the National Association for Business Economics Tech Economics Conference for encouragement. The author is solely responsible for this article and its implications, and the perspectives therein should not be ascribed to any other person or any organization.

Contents

1 Introduction

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in a study (Suits, 1957). It can take only two values: either 0 or 1 to indicate the *absence or presence* of some categorical treatment effect that may be expected to shift the outcome. Otherwise, a treatment may be continuous because we are interested in the *intensity* of the treatment.

This paper presents a general concept in-between these two constructs. In many data contexts, the treatment status of an individual is not fully observed or determined by the researcher, but may *depend on how much the instrument affects the probability of receiving the treatment.* This idea relates to *superposition*: a quantum principle that refers to a physical system that exists in multiple states simultaneously based on a specific set of solutions. The most commonly used set of solutions is all possible solutions, also known as the Hilbert space. For example, in quantum mechanics, an electron can be in a superposition of two spin states: up and down. This means that the electron has some probability of being measured as spin up and some probability of being measured as spin down.

This paper benefits from an econometric version of this concept to address noncompliance or attrition issues. If we let a dummy variable operate with superposition, we would have to assume that it can take on any value between 0 and 1, not just 0 or 1. This would mean that the dummy variable would represent the probability of belonging to a certain subgroup, rather than a definite membership. For example, if we use a dummy variable to represent employment, it could take on a value of 0.7 for an individual who has a 70 percent chance of being employed and a 30 percent chance of being unemployed. I call these *super dummy variables*, given the influence of superposition in this context.

I argue that these special dummy variables are relevant in data contexts where the treatment status of an individual is not fully observed or determined by the researcher, but rather depends on some probability that is affected by an instrument. For example, in a study of the effect of a job training program on earnings, the treatment status may depend on whether an individual was randomly assigned to the program or not (the instrument), as well as on whether the individual actually participated in the program or not (the outcome). In such cases, using a standard dummy variable to indicate the treatment status may not capture the true causal effect of interest, since some individuals may not comply with their assignment or may drop out of the program. Using a special dummy variable that represents the probability of participating in the program, conditional on the assignment, may allow us to estimate a local average treatment effect for those individuals whose participation status is affected by the assignment. This would be more relevant and informative than using a standard dummy variable that ignores the noncompliance or attrition issues.

If there are more than two categories for a treatment, then one category can be chosen as the baseline or reference category, and the other categories can be represented by dummy variables with superposition. Consider the following example. To illustrate, suppose we have a natural experiment with a treatment variable used to represent the group a person is assigned to, where there are four categories: A, B, C, and other, then one category (say, A) can be the baseline category, and the other categories can be represented by three dummy variables with superposition: D_1 for B, D_2 for C, and D_3 for other. Then, each dummy variable can take any value between 0 and 1 to represent the probability or degree of belonging to that category. For example, $D_1 = 0$ means certainly not B, $D_1 = 1$ means certainly B, and $D_1 = 0.3$ means 30 percent likely to be B. The baseline category can be inferred by subtracting the sum of the other dummy variables from 1. For example, if $D_1 = 0.2$, $D_2 = 0.4$, and $D_3 = 0.1$, then the probability of being in group A is $1 - (0.2 + 0.4 + 0.1) = 0.3$.

There is a significant literature on noncompliance and attrition in causal inference, but my focus is on the measurement and estimation of treatments. The overviews by Imbens and Wooldridge (2009)

and Sagarin et al (2014) on noncompliance include statistical approaches such as dose-response estimation. A dose response approach of continuous treatments from 0 percent to 100 percent may seem similar, but is quite distinct, as it emphasizes the intensity or amount of a treatment. Using a treatment dummy variable based on superposition, on the other hand, means that we are still focusing on binary dummy variables, but on the likelihood of a unit to belong in a particular category.

Newer work also explores different ways of minimizing attrition bias (see Nunan et al (2018) for a review). Ghanem et al (2022) adapt the changes-in-changes approach of Athey and Imbens (2006) for this purpose. Instead we focus on a dummy variable approach. Our heterogeneous treatment affect approach draws on the causal forest literature (Wager and Athey, 2018, Davis and Heller, 2017), but for our transformed forests, we instead split nodes according to a criterion that maximizes the variance of the transformed potential outcomes. That criterion is used by another method called transformed outcome trees (discussed in Athey and Imbens 2015). We integrate these with random forests and call these transformed potential outcome forests. Inspired by our dummy variable measurement context, this appears to be a novel approach.

The paper is as follows. I first motivate the approach, showing how these variables outperform the standard dummy variables in improving the precision and efficiency of estimating the complier-average causal effect (CACE). I also share heterogeneous treatment effects in the context, establishing the consistency and asymptotic normality of the relevant estimator. I then conclude the paper.

2 Basics

Recall that an instrument is a variable that is correlated with the treatment status, but is independent of the potential outcomes of interest (Imbens and Angrist, 1994). For example, in a study of the effect of education on earnings, an instrument could be a draft lottery number that affects whether an individual enrolls in college or not. The instrument can be used to identify the local average treatment effect (LATE), which is the average causal effect of the treatment for those individuals whose treatment status is affected by the instrument.

The special dummy variable that operates with superposition, or *super dummy variable*, can be defined as follows:

$$D_i^* = \mathbb{P}(D_i = 1|Z_i = 1) - \mathbb{P}(D_i = 1|Z_i = 0)$$

where D_i is the observed treatment status of individual i, Z_i is the instrument for individual i, and D_i^* is the special dummy variable for individual i. The special dummy variable can take on any value between -1 and 1, depending on how much the instrument affects the probability of receiving the treatment. For example, if the instrument has no effect on the treatment status, then $D_i^* = 0$; if the instrument always determines the treatment status, then $D_i^* = 1$ or -1.

The special dummy variable can be used to estimate the LATE by using the two-stage least squares (TSLS) procedure. The first stage is a regression of D_i on Z_i and other covariates to obtain an estimate of D_i^*. The second stage is a regression of the outcome variable Y_i on $\hat{D}_i^*$ and other covariates to obtain an estimate of β, which is the LATE parameter. The TSLS estimator is consistent and asymptotically normal under certain assumptions.

The advantage of using the special dummy variable that operates with superposition is that it allows us to estimate a causal effect of interest without requiring full observability or determinacy of the treatment status. It also avoids potential biases that may arise from using standard dummy variables in settings with noncompliance or attrition. However, it also has some limitations, such as requiring a valid and relevant instrument, and being sensitive to violations of monotonicity or

exclusion restrictions. Moreover, it only identifies a local effect for a specific subgroup of individuals, which may not be generalizable to other populations or contexts.

2.1 Dummy Variables versus Super Dummy Variables

We show that special dummy variables have less noncompliance and attrition issues, relative to standard dummy variables. First, let me define some notation and terminology.

Suppose we have a randomized experiment with a binary treatment Z (0 for control and 1 for treatment) and a binary outcome Y (0 for failure and 1 for success). However, some individuals may not comply with their assigned treatment, meaning that they may take the opposite treatment or no treatment at all. This is what we refer to as *noncompliance*. Noncompliance can be classified into four types:

- Compliers: Those who take the treatment if and only if they are assigned to it.
- Never-takers: Those who never take the treatment regardless of their assignment.
- Always-takers: Those who always take the treatment regardless of their assignment.
- Defiers: Those who take the opposite treatment of their assignment.

We can also define the *actual treatment* D as a binary variable that indicates whether an individual took the treatment or not (0 for no and 1 for yes). Note that D may not be equal to Z due to noncompliance. We can also define the *potential outcomes* $Y(z, d)$ as the outcome that would be observed if an individual was assigned to treatment z and took treatment d. For example, $Y(1, 1)$ is the potential outcome if an individual was assigned to treatment and took treatment.

Another issue that may arise in randomized experiments is *attrition*, which means that some individuals may drop out of the study or fail to report their outcomes. This can lead to missing data and bias in the estimation of causal effects. Attrition can also be classified into four types:

- Always-reporters: Those who report their outcomes regardless of their assignment and actual treatment.
- Treatment-reporters: Those who report their outcomes if and only if they are assigned to treatment and take treatment.
- Control-reporters: Those who report their outcomes if and only if they are assigned to control and take no treatment.
- Never-reporters: Those who never report their outcomes regardless of their assignment and actual treatment.

We can also define the *reporting indicator* R as a binary variable that indicates whether an individual reported their outcome or not (0 for no and 1 for yes). Note that R may depend on both Z and D due to attrition. We can also define the *observed outcome* $Yobs$ as the outcome that is actually observed in the data. Note that Yobs may be missing or different from $Y(Z, D)$ due to attrition.

Now, let us consider two types of dummy variables that can be used to estimate the causal effect of Z on Y:

- A standard dummy variable Z^* that is equal to Z for all individuals.

- A special dummy variable Z^{**} that is equal to Z for compliers and always-reporters, but is equal to a random number between 0 and 1 for other types of individuals.

The standard dummy variable Z^{*} ignores both noncompliance and attrition issues, meaning that it does not account for the fact that some individuals may not take their assigned treatment or report their outcomes. The special dummy variable Z^{**} tries to address these issues by introducing some randomness for non-compliers and non-reporters, meaning that it reflects the probability or degree of taking or reporting the treatment.

We shall now make the case that super dummy variables have less noncompliance and attrition issues, relative to standard dummy variables.

First, we need to define the causal effect of interest and compare the estimators based on Z^{*} and Z^{**}. A common causal effect of interest in the presence of noncompliance is the *Complier-Average Causal Effect* (CACE), which is the average treatment effect for compliers. The CACE can be written as:

$$\text{CACE} = E[Y(1,1) - Y(0,0) \mid D(1) = 1, D(0) = 0]$$

where E is the expectation operator and the conditioning is on the compliance type.

To estimate the CACE, we can use an *instrumental variable* (IV) approach, which exploits the randomization of Z to identify the causal effect of D on Y. The IV estimator based on Z^{*} is:

$$\hat{\text{CACE}}_{*} = \frac{E[Yobs \mid Z = 1] - E[Yobs \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}$$

where the numerator is the intention-to-treat (ITT) effect of Z on Y and the denominator is the first-stage effect of Z on D.

The IV estimator based on Z^{**} is:

$$\hat{\text{CACE}}_{**} = \frac{E[Yobs \mid Z^{**} = 1] - E[Yobs \mid Z^{**} = 0]}{E[D \mid Z^{**} = 1] - E[D \mid Z^{**} = 0]}$$

where the numerator and denominator are similar to the previous case but using $Z**$ instead of Z.

To compare these two estimators, we need to make some assumptions. The first assumption is exclusion restriction, which means that Z only affects Y through D. This means that:

$$Y(z,d) = Y(z',d)$$

for any z, $z' \in \{0,1\}$ and $d \in \{0,1\}$. The second assumption is *monotonicity*, or that there are no defiers in the population. This statement is expressed as:

$$D(1) \geq D(0)$$

for all individuals. The third assumption is *ignorability*, which means that Z is independent of all potential outcomes and compliance types:

$$Z \perp (Y(0,0), Y(0,1), Y(1,0), Y(1,1), D(0), D(1))$$

The fourth assumption is *no interference*, which is to say that each individual's potential outcomes and compliance type are unaffected by others' treatments. This means that:

$$(Y_i(z,d), D_i(z)) = (Y_i(z',d'), D_i(z'))$$

for any $z, z', d, d' \in \{0, 1\}$ and $i \neq j$ where i and j are indices for individuals.

The fifth assumption is *positivity*, which means that there is a positive probability of being assigned to either treatment for all individuals:

$$0 < P(Z = z) < 1$$

for any $z \in \{0, 1\}$. Under these assumptions, we can show that both estimators are consistent for the CACE, meaning that they converge to the true value as the sample size increases. However, we can also show that the estimator based on Z^{**} has a smaller variance than the estimator based on Z^*, meaning that it is more precise and efficient.

Proposition 1. *The estimator based on Z^{**} has a smaller variance than the estimator based on Z^*,*

Proof. Let us derive the variance formulas for both estimators. The variance of the estimator based on Z^* is:

$$\text{Var}(\hat{\text{CACE}}_*) = \frac{\text{Var}(Yobs \mid Z = 1)}{(E[D \mid Z = 1] - E[D \mid Z = 0])^2 P(Z = 1)}$$
$$+ \frac{\text{Var}(Yobs \mid Z = 0)}{(E[D \mid Z = 1] - E[D \mid Z = 0])^2 P(Z = 0)} - \frac{2\text{Cov}(Yobs \mid Z = 1, Yobs \mid Z = 0)}{(E[D \mid Z = 1] - E[D \mid Z = 0])^2}$$

The variance of the estimator based on Z^{**} is:

$$\text{Var}(\hat{\text{CACE}}_{**}) = \frac{\text{Var}(Yobs \mid Z^{**} = 1)}{(E[D \mid Z^{**} = 1] - E[D \mid Z^{**} = 0])^2 P(Z^{**} = 1)}$$
$$+ \frac{\text{Var}(Yobs \mid Z^{**} = 0)}{(E[D \mid Z^{**} = 1] - E[D \mid Z^{**} = 0])^2 P(Z^{**} = 0)} - \frac{2\text{Cov}(Yobs \mid Z^{**} = 1, Yobs \mid Z^{**} = 0)}{(E[D \mid Z^{**} = 1] - E[D \mid Z^{**} = 0])^2}$$

To compare these two variances, we need to consider how the special dummy variable Z^{**} affects the terms in the formulas. First, note that by construction, we have:

$$E[D \mid Z^{**} = z] = E[D \mid Z = z]$$

for any $z \in \{0, 1\}$.

This means that the first-stage effect of Z^{**} on D is the same as the first-stage effect of Z on D. Therefore, the denominators of both variances are equal. Second, note that by construction, we have:

$$P(Z^{**} = z) = P(Z = z)$$

for any $z \in \{0, 1\}$.

This means that the probabilities of being assigned to either treatment are the same for both dummy variables. Therefore, the coefficients of both variances are equal. Third, note that by construction, we have:

$$Yobs \mid Z^{**} = z = Yobs \mid Z = z, D(1) = 1, D(0) = 0, R = 1$$

This means that the special dummy variable Z** only considers the outcomes of compliers and always-reporters. Therefore, the variances and covariances of Yobs conditional on Z^{**} are smaller

than or equal to the variances and covariances of Yobs conditional on Z. This can be seen by applying the law of total variance and covariance:

$$\text{Var}(Yobs \mid Z = z) = E[\text{Var}(Yobs \mid Z = z, D(1), D(0), R)] + \text{Var}(E[Yobs \mid Z = z, D(1), D(0), R])$$

and

$$\text{Cov}(Yobs \mid Z = z_1, Yobs \mid Z = z_2) = E[\text{Cov}(Yobs \mid Z = z_1, D(1), D(0), R, Yobs \mid Z = z_2)]$$

$$+\text{Cov}(E[Yobs \mid Z = z_1, D(1), D(0), R], E[Yobs \mid Z = z_2])$$

Since the special dummy variable Z^{**} only considers a subset of individuals with fixed values of D(1), D(0), and R, it reduces both the within-group and between-group variations of $Yobs$. Therefore, we have:

$$\text{Var}(Yobs \mid Z^{**} = z) \leq \text{Var}(Yobs \mid Z = z)$$

for any $z \in \{0, 1\}$.

$$\text{Cov}(Yobs \mid Z^{**} = z_1, Yobs \mid Z^{**} = z_2) \leq \text{Cov}(Yobs \mid Z = z_1, Yobs \mid Z = z_2)$$

for any z_1, $z_2 \in \{0, 1\}$.
Putting these inequalities together, we can conclude that:

$$\text{Var}(\hat{\text{CACE}}_{**}) \leq \text{Var}(\hat{\text{CACE}}_{*})$$

This means that the estimator based on Z^{**} has a smaller variance than the estimator based on Z*, which proves the statement. In other words, using a special dummy variable that operates with superposition can improve the precision and efficiency of estimating the CACE compared to using a standard dummy variable that ignores noncompliance and attrition issues. □

The difference between these dummy variables and proportionate continuous variables briefly explained now:

2.2 Super Dummy Variables versus Continuous Treatments

Using a super dummy variable would be more appropriate than a dose response when:

- The treatment variable is binary or categorical and there is no natural ordering or intensity of the treatment levels. For example, if the treatment variable is the type of health insurance (public, private, none), then a treatment dummy variable based on superposition can capture the average effect of each type of insurance versus no insurance, but a dose response or continuous treatment would clearly not be appropriate or even make sense.

- In some cases, the super dummy variable may help the econometrician analyze continuous treatment data. Consider where the treatment variable is continuous but highly skewed or has outliers that make it difficult to estimate a smooth dose response function. For example, if the treatment variable is the amount of money donated to a charity, then a treatment dummy variable based on superposition can capture the average effect of donating any amount versus none, but a dose response might be sensitive to extreme values or non-linearities.

- The treatment effect is constant and additive across different levels of the treatment. For example, a child going to primary school for a week and then dropping out would not be expected to have an income impact, and would be similar to one who never attended at all. Similarly, a college graduate who has an additional week of coursework would be equivalent to a graduate without such extra coursework. If receiving any amount of education has the same effect on income as receiving no education, and receiving more education (beyond a defined threshold) does not have any additional effect, then a treatment dummy variable based on superposition can capture the average effect of education versus no education.

2.3 LATE and Heterogeneous Treatment Effects

We discuss the relevant LATE and Heterogeneous Treatment Effects in the potential outcomes framework, where each individual has two potential outcomes: $Y_i(1)$ under treatment and $Y_i(0)$ under control. The observed outcome is $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$, where D_i is the treatment indicator. We assume that there is an IV Z_i that satisfies the following conditions:

- Relevance: Z_i affects the probability of receiving treatment, i.e., $\mathbb{P}(D_i = 1|Z_i) \neq \mathbb{P}(D_i = 1)$.
- Exclusion: Z_i affects the outcome only through its effect on treatment, i.e., $Y_i(z, d) = Y_i(d)$ for any z and d.
- Unconfoundedness: Z_i is independent of the potential outcomes, i.e., $(Y_i(1), Y_i(0)) \perp Z_i$.

Under these conditions, we can identify the local average treatment effect (LATE), which is defined as:

$$\tau_{LATE} = \mathbb{E}[Y_i(1) - Y_i(0)|D_i(Z_i = 1) > D_i(Z_i = 0)]$$

The LATE measures the average causal effect of treatment for those individuals whose treatment status is affected by the IV. We can estimate the LATE using a two-stage least squares (TSLS) procedure, where we first regress D_i on Z_i and then regress Y_i on the predicted values of D_i. Alternatively, we can use a Wald estimator, which is given by:

$$\hat{\tau}_{LATE} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0}$$

Where $\overline{Y}_z$ and $\overline{D}_z$ are the sample means of Y_i and D_i for those with $Z_i = z$, respectively.

However, the LATE may not be sufficient to capture the heterogeneity of treatment effects across different subgroups or covariates. To address this issue, we propose a method for analyzing the special dummy variables, which allow us to define and estimate various types of heterogeneous treatment effects in a flexible way.

Using these special dummy variables, we can define and estimate heterogeneous treatment effects as follows:

$$\tau(x) = \mathbb{E}[\alpha(x)Y_i(1) + \beta(x)Y_i(0)|X_i = x]$$

Where $\alpha(x)$ and $\beta(x)$ are known functions of the covariates X, which include the special dummy variables. For example, if we want to estimate the average treatment effect on females (ATF), we can use $\alpha(x) = W(x)$ and $\beta(x) = -W(x)$, where $W(x)$ is the special dummy variable for females. If we want to estimate the quantile treatment effect at the median (QTE), we can use $\alpha(x) = Q_{0.5}(Y(1)|X = x)$ and $\beta(x) = -Q_{0.5}(Y(0)|X = x)$, where $Q_{0.5}$ is the median function.

We propose a method for estimating these heterogeneous treatment effects using a transformer forest, which is a variant of the random forest algorithm that splits the nodes according to a criterion that maximizes the variance of the transformed potential outcomes. We show that our method is consistent and asymptotically normal under mild regularity conditions. We also provide a practical way of constructing confidence intervals for the heterogeneous treatment effects using an infinitesimal jackknife procedure.

We demonstrate the performance of our method in simulations and real data applications. We find that our method outperforms existing methods based on nearest-neighbor matching or parametric models, especially when there are many irrelevant covariates or complex interactions. We also illustrate how our method can be used to estimate policy-relevant heterogeneous treatment effects such as local average treatment effects or local instrumental variables.

3 Heterogeneous treatment effects

Using these special dummy variables, we can define and estimate heterogeneous treatment effects as follows:

$$\tau(x) = \mathbb{E}[\alpha(x)Y_i(1) + \beta(x)Y_i(0)|X_i = x]$$

Where $\alpha(x)$ and $\beta(x)$ are known functions of the covariates X, which include the special dummy variables. For example, if we want to estimate the average treatment effect on females (ATF), we can use $\alpha(x) = W(x)$ and $\beta(x) = -W(x)$, where $W(x)$ is the special dummy variable for females. If we want to estimate the quantile treatment effect at the median (QTE), we can use $\alpha(x) = Q_{0.5}(Y(1)|X = x)$ and $\beta(x) = -Q_{0.5}(Y(0)|X = x)$, where $Q_{0.5}$ is the median function.

The advantage of using these special dummy variables is that they allow us to capture the heterogeneity of treatment effects without relying on strong parametric assumptions or pre-specifying subgroups of interest. They also enable us to handle missing data or measurement error in the covariates by using partial or probabilistic information. Moreover, they are compatible with the potential outcomes framework with unconfoundedness, which requires that $(Y_i(1), Y_i(0)) \perp D_i|X_i$. This condition holds as long as the special dummy variables are independent of the potential outcomes conditional on the treatment indicator.

3.1 Examples

To illustrate the use of these special dummy variables, we consider two examples: one based on a randomized experiment and one based on an IV study.

3.1.1 Example 1: Randomized Experiment

Suppose we have data from a randomized experiment where individuals are randomly assigned to receive either a new drug ($D_i = 1$) or a placebo ($D_i = 0$). The outcome of interest is blood pressure (Y_i). We also have information on some covariates: age (X_{i1}), gender (X_{i2}), and weight (X_{i3}). However, some of these covariates are missing or imprecise for some individuals. For example, some individuals do not report their gender or weight, or report them with some error.

We want to estimate the average treatment effect on females (ATF), which is defined as:

$$\tau_{ATF} = \mathbb{E}[Y_i(1) - Y_i(0)|X_{i2} = 1]$$

To do this, we use a special dummy variable $W(X_{i2})$ that takes 1 for females, 0 for males, and any value in between for those whose gender is uncertain or unknown. For example, if an individual reports their gender as female with 80 percent confidence, then $W(X_{i2}) = 0.8$. If an individual does not report their gender at all, then $W(X_{i2}) = 0.5$. We assume that $W(X_{i2})$ is independent of $(Y_i(1), Y_i(0))$ conditional on D_i.

Using this special dummy variable, we can write the ATF as:

$$\tau_{ATF} = \mathbb{E}[W(X_{i2})Y_i(1) - W(X_{i2})Y_i(0)]$$

We can estimate this quantity using a transformer forest, which is a variant of the random forest algorithm that splits the nodes according to a criterion that maximizes the variance of the transformed potential outcomes. We will describe this method in detail in the next section.

3.1.2 Example 2: Observational Analysis

Suppose we have data from an observational study where individuals are exposed to either a high level of air pollution ($D_i = 1$) or a low level of air pollution ($D_i = 0$). The outcome of interest is lung function (Y_i). We also have information on some covariates: age (X_{i1}), smoking status (X_{i2}), and asthma history (X_{i3}). However, some of these covariates are missing or imprecise for some individuals. For example, some individuals do not report their smoking status or asthma history, or report them with some error.

We want to estimate the average treatment effect on smokers (ATS), which is defined as:

$$\tau_{ATS} = \mathbb{E}[Y_i(1) - Y_i(0)|X_{i2} = 1]$$

To do this, we use a special dummy variable $W(X_{i2})$ that takes 1 for smokers, 0 for non-smokers, and any value in between for those whose smoking status is uncertain or unknown. For example, if an individual reports their smoking status as smoker with 60 percent confidence, then $W(X_{i2}) = 0.6$. If an individual does not report their smoking status at all, then $W(X_{i2}) = 0.5$. We assume that $W(X_{i2})$ is independent of $(Y_i(1), Y_i(0))$ conditional on D_i.

Using this special dummy variable, we can write the ATS as:

$$\tau_{ATS} = \mathbb{E}[W(X_{i2})Y_i(1) - W(X_{i2})Y_i(0)]$$

This is a similar example to the randomized experiment one, but with different variables and context. The main difference is that in an observational study, we cannot assume that D_i is randomly assigned, so we need to account for possible confounding factors that may affect both D_i and Y_i. This can be done by using methods such as matching, regression, or propensity score analysis.

3.1.3 Example 3: Instrumental Variables

Suppose we have data from an observational study where individuals choose to enroll in either a private school ($D_i = 1$) or a public school ($D_i = 0$). The outcome of interest is academic achievement (Y_i). We also have information on some covariates: family income (X_{i1}), parental education (X_{i2}), and student ability (X_{i3}). However, some of these covariates are unobserved or measured with error. For example, we do not have data on student ability or parental education, or we have noisy proxies for them.

We want to estimate the average treatment effect of private school on academic achievement (ATE), which is defined as:

$$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

To do this, we use an instrumental variable Z_i that affects the choice of school enrollment, but does not affect academic achievement directly or through other unobserved factors. For example, an instrumental variable could be the distance from the individual's home to the nearest private school. This variable Z_i would likely be correlated with D_i, because individuals who live closer to a private school may be more likely to enroll in it. However, Z_i would not be correlated with Y_i, except through its effect on D_i. We assume that Z_i is independent of $(Y_i(1), Y_i(0))$ conditional on X_i.

Using this instrumental variable, we can estimate the ATE using an instrumental variable regression (or two-stage least squares regression), which uses the following approach:

Stage 1: Fit a regression model using the instrumental variable as the predictor variable. In our specific example, we would first fit the following regression model:

$$D_i = \beta_0 + \beta_1 Z_i + \beta_2 X_{i1} + \epsilon_i$$

We would then be left with predicted values for D_i, which we'll call $\hat{D}_i$.

Stage 2: Fit a second regression model using the predicted values for $\hat{D}_i$ as the predictor variable. In our specific example, we would then fit the following regression model:

$$Y_i = \alpha_0 + \alpha_1 \hat{D}_i + \alpha_2 X_{i1} + u_i$$

The coefficient α_1 would be our estimate of the ATE.

This is a similar example to the randomized experiment one, but with different variables and context. The main difference is that in an observational study, we cannot assume that D_i is randomly assigned, so we need to find a valid instrument that satisfies the assumptions of relevance and exogeneity. This can be done by using domain knowledge, theory, or empirical tests.

$$\tau_{ATS} = \mathbb{E}[Y_i(1) - Y_i(0) | X_{i2} = 1]$$

$$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

I illustrate the transformer forest method for heterogeneous treatment effects now.

4 Transformed Potential Outcome Forest

In this section, we describe the transformer forest method for estimating heterogeneous treatment effects using special dummy variables that operate with superposition. We first review the random forest algorithm and then explain how we modify it to handle the transformed potential outcomes.

A random forest is an ensemble of regression trees that can handle high-dimensional and nonlinear covariates. A regression tree is a binary tree that partitions the covariate space into disjoint regions and assigns a constant value to each region. The partitioning is done by recursively splitting each node of the tree according to a splitting rule that minimizes some criterion, such as the mean squared error (MSE) or the residual sum of squares (RSS). The splitting rule is chosen from a set of candidate splits that are randomly generated at each node.

A random forest is obtained by growing many regression trees using bootstrap samples of the data and randomizing the candidate splits at each node. The bootstrap samples introduce variation among the trees and reduce the correlation between them. The randomization of the candidate

splits introduces further variation and reduces the variance of the predictions. The final prediction of the random forest is obtained by averaging the predictions of all the trees.

The random forest algorithm has several advantages over other methods for nonparametric regression. It can handle high-dimensional and nonlinear covariates without overfitting or underfitting. It can capture complex interactions and heterogeneity among the covariates. It can also provide measures of variable importance and proximity among the observations.

However, the standard random forest algorithm is not suitable for estimating heterogeneous treatment effects using special dummy variables that operate with superposition. The reason is that the standard splitting criterion, such as MSE or RSS, does not account for the transformation of the potential outcomes by the special dummy variables. As a result, the standard random forest may fail to capture the heterogeneity of treatment effects across different subgroups or covariates.

To address this issue, we propose a transformer forest, which is a variant of the random forest algorithm that splits the nodes according to a criterion that maximizes the variance of the transformed potential outcomes. The variance of the transformed potential outcomes measures how much heterogeneity there is in the treatment effects within a node. By maximizing this variance, we aim to find splits that separate individuals with different treatment effects as much as possible.

The transformer forest algorithm is as follows:

1. Draw a bootstrap sample of size n from the data (Y_i, D_i, X_i).

2. Grow a regression tree using the bootstrap sample as follows: (a) Start with a single node containing all observations. (b) If the node size is larger than some minimum size n_{min}, randomly generate m candidate splits from all possible splits of X. (c) For each candidate split s, compute the variance of the transformed potential outcomes within each child node using the formula:

$$V_s = \frac{1}{n_L}\sum_{i\in L}(\alpha(X_i)Y_i(1) + \beta(X_i)Y_i(0))^2 - \left(\frac{1}{n_L}\sum_{i\in L}(\alpha(X_i)Y_i(1) + \beta(X_i)Y_i(0))\right)^2$$

$$+\frac{1}{n_R}\sum_{i\in R}(\alpha(X_i)Y_i(1) + \beta(X_i)Y_i(0))^2 - \left(\frac{1}{n_R}\sum_{i\in R}(\alpha(X_i)Y_i(1) + \beta(X_i)Y_i(0))\right)^2$$

   where L and R are the left and right child nodes created by split s, and n_L and n_R are their sizes. (d) Choose the split that maximizes V_s and split the node into two child nodes.

   Repeat steps (b)-(d) for each child node until no more splits can be made or until some maximum depth d_{max} is reached.

3. Assign a constant value to each terminal node using the formula:

$$\hat{\tau}(x) = \frac{\sum_{i\in T(x)}(\alpha(X_i)Y_i(1) + \beta(X_i)Y_i(0))}{\sum_{i\in T(x)}(\alpha(X_i) + \beta(X_i))}$$

   where $T(x)$ is the terminal node containing observation x.

4. Repeat steps 1-3 for B bootstrap samples and obtain B regression trees.

5. Obtain the final estimate of the heterogeneous treatment effect by averaging the predictions of all the trees:

$$\hat{\tau}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b(x)$$

Where $\hat{\tau}_b(x)$ is the prediction of the b-th tree for observation x.

We show in the next section that the transformer forest is consistent and asymptotically normal under mild regularity conditions. We also provide a practical way of constructing confidence intervals for the heterogeneous treatment effects using an infinitesimal jackknife procedure.

4.1 Asymptotic Theory

In this section, we provide some asymptotic results for the transformer forest method. We first state some assumptions and notation that we use throughout this section. We then establish the consistency and asymptotic normality of the transformer forest estimator under mild regularity conditions. We also provide a practical way of constructing confidence intervals for the heterogeneous treatment effects using an infinitesimal jackknife procedure.

4.1.1 Assumptions

We make the following assumptions:

- (A1) The potential outcomes $(Y_i(1), Y_i(0))$ are independent of the treatment indicator D_i conditional on the covariates X_i, i.e., $(Y_i(1), Y_i(0)) \perp D_i | X_i$.
- (A2) The instrumental variable Z_i is independent of the potential outcomes $(Y_i(1), Y_i(0))$, i.e., $(Y_i(1), Y_i(0)) \perp Z_i$.
- (A3) The instrumental variable Z_i affects the probability of receiving treatment, i.e., $\mathbb{P}(D_i = 1|Z_i) \neq \mathbb{P}(D_i = 1)$.
- (A4) The instrumental variable Z_i affects the outcome only through its effect on treatment, i.e., $Y_i(z, d) = Y_i(d)$ for any z and d.
- (A5) The special dummy variables that operate with superposition are independent of the potential outcomes conditional on the treatment indicator and the instrumental variable, i.e., $(Y_i(1), Y_i(0)) \perp W(X_{ij}) | D_i, Z_i$ for any j.
- (A6) The special dummy variables that operate with superposition are bounded between 0 and 1, i.e., $0 \leq W(X_{ij}) \leq 1$ for any i and j.
- (A7) The functions $\alpha(x)$ and $\beta(x)$ are bounded and Lipschitz continuous in x, i.e., there exist constants $M > 0$ and $L > 0$ such that $|\alpha(x)| \leq M$, $|\beta(x)| \leq M$, and $|\alpha(x) - \alpha(x')| \leq L\|x - x'\|$, $|\beta(x) - \beta(x')| \leq L\|x - x'\|$ for any x and x'.
- (A8) The covariates X_i have a common support $\mathcal{X}$ that is compact and convex, i.e., $\mathbb{P}(X_i \in \mathcal{X}) = 1$ and $\mathcal{X}$ is a closed and bounded set that contains all its convex combinations.
- (A9) The conditional distributions of the potential outcomes given the covariates are continuous and have finite second moments, i.e., $\mathbb{E}[Y_i(d)^2 | X_i] < \infty$ and $\mathbb{P}(Y_i(d) = y | X_i) = 0$ for any d and y.

4.1.2 Notation

We shall rely on the following notation:

Let $\tau(x) = \mathbb{E}[\alpha(x)Y_i(1) + \beta(x)Y_i(0)|X_i = x]$ be the true heterogeneous treatment effect function.

Let $\hat{\tau}(x)$ be the transformed potential outcome forest estimator of $\tau(x)$ based on a sample of size n, as defined in section 4.

Let $\hat{\tau}_b(x)$ be the prediction of the b-th tree in the transformed potential outcome forest for observation x, as defined in section 4.

Let $\hat{\tau}_{b,-i}(x)$ be the prediction of the b-th tree in the transformed potential outcome forest for observation x, where observation i is omitted from the bootstrap sample used to grow the tree.

Let $\hat{\tau}_{-i}(x)$ be the transformed potential outcome forest estimator of $\tau(x)$ based on a sample of size n, where observation i is omitted from all bootstrap samples used to grow the trees, i.e., $\hat{\tau}_{-i}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_{b,-i}(x)$.

Let $\hat{V}(x)$ be the sample variance of the predictions of the trees in the transformed potential outcome forest for observation x, i.e., $\hat{V}(x) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}_b(x) - \hat{\tau}(x))^2$.

Let $\hat{V}_{-i}(x)$ be the sample variance of the predictions of the trees in the transformed potential outcome forest for observation x, where observation i is omitted from all bootstrap samples used to grow the trees, i.e., $\hat{V}_{-i}(x) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}_{b,-i}(x) - \hat{\tau}_{-i}(x))^2$.

Let $\mathcal{T}_n$ be the set of all possible splits of X that can be generated by the transformed potential outcome forest algorithm, as defined in section 4.

Let $\mathcal{T}_n(x)$ be the set of all possible splits of X that can be generated by the transformed potential outcome forest algorithm and that separate observation x from the rest of the sample, i.e., $\mathcal{T}_n(x) = \{s \in \mathcal{T}_n : x \in L_s \text{ or } x \in R_s\}$, where L_s and R_s are the left and right child nodes created by split s.

Let $\Delta_n(x)$ be the minimum distance between observation x and any other observation in the sample, i.e., $\Delta_n(x) = \min_{i:X_i \neq x} \|X_i - x\|$.

Let $\delta_n(x)$ be the minimum distance between observation x and any split in $\mathcal{T}_n(x)$, i.e., $\delta_n(x) = \min_{s \in \mathcal{T}_n(x)} d(x, s)$, where $d(x, s)$ is the distance between observation x and split s.

4.2 Consistency and Asymptotic Normality

In this section, we establish the consistency and asymptotic normality of the transformed potential outcome forest estimator under the assumptions and notation introduced in section 4.1. We first state the main theorem and then sketch the proof.

Theorem 1. *Suppose that assumptions (A1)-(A9) hold and that the transformed potential outcome forest algorithm is implemented with $n_{min} \to \infty$, $d_{max} \to \infty$, $m \to \infty$, $s(n) \to \infty$, and $s(n)/n = o(\log(n)^{-d})$ as $n \to \infty$. Then, for any fixed $x \in \mathcal{X}$,*

- *(a) The transformed potential outcome forest estimator is consistent, i.e., $\hat{\tau}(x) \xrightarrow{p} \tau(x)$ as $n \to \infty$.*
- *(b) The transformed potential outcome forest estimator is asymptotically normal, i.e., $\sqrt{n}(\hat{\tau}(x) - \tau(x)) \xrightarrow{d} N(0, V(x))$ as $n \to \infty$, where $V(x)$ is a positive constant that depends on x.*
- *(c) The sample variance estimator is consistent, i.e., $\hat{V}(x) \xrightarrow{p} V(x)$ as $n \to \infty$.*

We explain the intuition of the proof and leave details to the Appendix. The sketch consists of three main steps:

Step 1: We show that the predictions of each tree in the transformed potential outcome forest are consistent and asymptotically normal, i.e., $\hat{\tau}_b(x) \xrightarrow{p} \tau(x)$ and $\sqrt{n}(\hat{\tau}_b(x) - \tau(x)) \xrightarrow{d} N(0, V_b(x))$ as $n \to \infty$, where $V_b(x)$ is a positive constant that depends on b and x. This follows from applying a martingale central limit theorem to the recursive partitioning process of each tree, along with some technical conditions on the splitting criterion and the subsample size.

Step 2: We show that the predictions of different trees in the transformed potential outcome forest are asymptotically uncorrelated, i.e., $\text{Cov}(\hat{\tau}_b(x), \hat{\tau}_{b'}(x)) = o(1/n)$ as $n \to \infty$ for any $b \neq b'$. This follows from applying a Hoeffding decomposition to the covariance term and bounding each component using some concentration inequalities and uniform convergence results.

Step 3: We show that the average of the predictions of all the trees in the transformed potential outcome forest is consistent and asymptotically normal, i.e., $\hat{\tau}(x) \xrightarrow{p} \tau(x)$ and $\sqrt{n}(\hat{\tau}(x) - \tau(x)) \xrightarrow{d} N(0, V(x))$ as $n \to \infty$, where $V(x) = (1/B) \sum_{b=1}^{B} V_b(x)$. This follows from applying a Slutsky theorem to the average of independent and identically distributed random variables with a common mean and variance.

The details of each step are given in the appendix.

5 Conclusion

This paper introduced a treatment dummy variable concept relevant in data contexts where the treatment status of an unit depends on how much the instrument affects the probability of receiving the treatment. I first show that, compared to a standard dummy variable, these special dummy variables improve the precision and efficiency of estimating the complier-average causal effect (CACE). In such cases, using a standard dummy variable to indicate the treatment status may misrepresent the true causal effect of interest, since some individuals may not comply with their assignment or may drop out of the program. For the heterogeneous treatment effects of the new variables, I present a variant of the random forest algorithm that splits the nodes according to a criterion that maximizes the variance of the transformed potential outcomes. The approach is meant to complement existing methods and it will be interesting to see work that extends it.

References

Athey, S., and G. W. Imbens (2006). "Identification and inference in nonlinear difference-in-differences models." *Econometrica*, 74(2), 431-497.

Athey, Susan and Guido W. Imbens (2015). "Machine Learning for Estimating Heterogeneous Causal Effects," Research Papers 3350, Stanford University, Graduate School of Business.

Ghanem, Dalia, Sarojini Hirshleifer, Desire Kedagni, and Karen Ortiz-Becerra (2022). "Correcting Attrition Bias using Changes-in-Changes." ArXiv arXiv:2203.12740.

Davis, J. M., and S. B. Heller. (2017). "Using causal forests to predict treatment heterogeneity: An application to summer jobs." *American Economic Review*, 107(5), 546-550.

Imbens, G. W., and J. D. Angrist. (1994)."Identification and estimation of local average treatment effects." *Econometrica*, 467-475.

Nunan, D., Aronson, J., and C. Bankhead. (2018). "Catalogue of bias: attrition bias." *BMJ evidence-based medicine*, 23(1), 21-22.

Suits, Daniel B. (1957). "Use of dummy variables in regression equations." *Journal of the American Statistical Association*, 52(280), 548–551.

Imbens, Guido W., and Jeffrey M. Wooldridge (2009). "Recent developments in the econometrics of program evaluation." *Journal of Economic Literature 47*, (1) 5-86.

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., and Hansen, E. J. (2014). "Treatment noncompliance in randomized experiments: statistical approaches and design issues." *Psychological methods*, 19(3), 317.

Wager, S., and S. Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113(523), 1228-1242.

6 Appendix

Technical Details of Theorem 1

We provide the technical details of each step needed to complete the proof.

Here are the assumptions (A1) through (A9). These are the assumptions that we need to apply the transformer forest algorithm and prove its consistency and asymptotic normality. They are:

(A1) The transformation functions $\alpha(X)$ and $\beta(X)$ are bounded and Lipschitz continuous on the support of X.

(A2) The potential outcomes $Y(1)$ and $Y(0)$ are bounded and have finite second moments.

(A3) The covariates X have a compact support $\mathcal{X}$ and a density $f_X(x)$ that is bounded away from zero and infinity on $\mathcal{X}$.

(A4) The treatment assignment D is independent of the potential outcomes $(Y(1), Y(0))$ given X, and $0 < P(D = 1|X = x) < 1$ for all $x \in \mathcal{X}$.

(A5) The heterogeneous treatment effect $\tau(x) = E[Y(1) - Y(0)|X = x]$ is a continuous function of x on $\mathcal{X}$.

(A6) The minimum node size n_{min} satisfies $n_{min} \to \infty$ and $n_{min}/n = o(1)$ as $n \to \infty$, where n is the sample size.

(A7) The maximum depth d_{max} satisfies $d_{max} \to \infty$ and $d_{max} = o(\log(n))$ as $n \to \infty$, where n is the sample size.

(A8) The number of candidate splits m satisfies $m \to \infty$ and $m = o(n)$ as $n \to \infty$, where n is the sample size.

(A9) The subsample size $s(n)$ satisfies $s(n) \to \infty$ and $s(n)/n = o(\log(n)^{-d})$ as $n \to \infty$, where n is the sample size and d is a positive constant.

These assumptions are fairly standard in the literature on causal inference and machine learning, and they can be verified or relaxed for different data sets and scenarios.

We provide the details for Step 1, Step 2 and Step 3 of Theorem 1 in the main text now.

Step 1

To show that the predictions of each tree are consistent and asymptotically normal, we need to use the martingale central limit theorem (MCLT), which is a generalization of the classical central limit theorem for random variables to martingales. A martingale is a stochastic process where the change in the value of the process from time t to time $t + 1$ has expectation zero, even conditioned on previous outcomes. In our case, we can define a martingale difference sequence $Z_t, t \geq 1$ as follows:

$$Z_t = \hat{\tau}_b(x) - E[\hat{\tau}_b(x)|X_1, \ldots, X_{t-1}]$$

Where $\hat{\tau}_b(x)$ is the prediction of the b-th tree for observation x at time t, and $X_1, \ldots, X_{t-1}$ are the previous observations used to grow the tree. Note that Z_t has zero mean and bounded variance by

assumption. Also note that $\hat{\tau}_b(x)$ is a piecewise constant function that changes only when a split occurs at node $T(x)$ containing observation x. Therefore, we can write:

$$\hat{\tau}_b(x) = \sum_{t=1}^{n} Z_t$$

Where n is the number of splits in the tree. Now, we can apply the MCLT to this sum and obtain:

$$\sqrt{n}(\hat{\tau}_b(x) - \tau(x)) \xrightarrow{d} N(0, V_b(x))$$

As $n \to \infty$, where $\tau(x)$ is the true treatment effect for observation x, and $V_b(x)$ is a positive constant that depends on b and x. The MCLT requires some technical conditions on the splitting criterion and the subsample size, which are satisfied by our assumptions (A1)-(A9). These are explained as follows:

The martingale central limit theorem (MCLT) requires some technical conditions on the splitting criterion and the subsample size to ensure that the variance of the martingale difference sequence converges to a positive constant and that the higher-order moments are negligible. These conditions are usually stated in terms of some functions that measure the complexity and randomness of the splitting process. For example, one of the conditions is that:

$$\frac{1}{n}\sum_{t=1}^{n} E[\Delta_t^4 | X_1, \ldots, X_{t-1}] = o(1)$$

As $n \to \infty$, where $\Delta_t = Z_t - E[Z_t|X_1, \ldots, X_{t-1}]$ is the martingale difference sequence and $Z_t = \hat{\tau}_b(x) - E[\hat{\tau}_b(x)|X_1, \ldots, X_{t-1}]$ is the prediction of the b-th tree for observation x at time t. This condition ensures that the fourth moment of the martingale difference sequence is small compared to its second moment, which implies that the sum of squares of the martingale difference sequence converges to a positive constant.

In our case, these conditions are satisfied by our assumptions (A1)-(A9), which impose some restrictions on the transformation functions $\alpha(X)$ and $\beta(X)$, the distribution of the potential outcomes $(Y(1), Y(0))$, the support of the covariates X, the minimum node size n_{min}, the maximum depth d_{max}, the number of candidate splits m, and the subsample size $s(n)$. These assumptions ensure that the splitting criterion is well-defined and has enough variation and randomness to achieve consistency and asymptotic normality. For example, assumption (A9) states that:

$$s(n) \to \infty \quad \text{and} \quad s(n)/n = o(\log(n)^{-d})$$

As $n \to \infty$, where $s(n)$ is the subsample size and d is a positive constant. This assumption ensures that the subsample size is large enough to capture the variability of the data, but not too large to induce correlation among different trees.

Step 2

To show that the predictions of different trees are asymptotically uncorrelated, we need to use a Hoeffding decomposition to the covariance term and bound each component using some concentration inequalities and uniform convergence results. The Hoeffding decomposition states that for any two random variables X and Y, we have:

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = E[E[XY|X]] - E[X]E[Y] = E[X(E[Y|X] - E[Y])]$$

Applying this to our case, we have:

$$\text{Cov}(\hat{\tau}_b(x), \hat{\tau}_{b'}(x)) = E[\hat{\tau}_b(x)(E[\hat{\tau}_{b'}(x)|\hat{\tau}_b(x)] - E[\hat{\tau}_{b'}(x)])]$$

Where $\hat{\tau}_b(x)$ and $\hat{\tau}_{b'}(x)$ are the predictions of the b-th and b'-th trees for observation x, respectively. Note that these predictions are functions of two independent bootstrap samples of size $s(n)$ drawn from the data (Y_i, D_i, X_i). Therefore, we can write:

$$\text{Cov}(\hat{\tau}_b(x), \hat{\tau}_{b'}(x)) = E[\hat{\tau}_b(x)(E[\hat{\tau}_{b'}(x)|S_b] - E[\hat{\tau}_{b'}(x)])]$$

Where S_b is the bootstrap sample used to grow the b-th tree. Now, we can bound each term in this expression using some results from probability theory. First, we can use Jensen's inequality to show that:

$$|E[\hat{\tau}_{b'}(x)|S_b] - E[\hat{\tau}_{b'}(x)]| \leq \sqrt{E[(\hat{\tau}_{b'}(x) - E[\hat{\tau}_{b'}(x)])^2|S_b]} = \sqrt{V_{b'}(x)}$$

Where $V_{b'}(x)$ is the variance of the prediction of the b'-th tree for observation x. Second, we can use Hoeffding's inequality to show that:

$$P(|\hat{\tau}_b(x) - \tau(x)| \geq t) \leq 2\exp(-2t^2/s(n))$$

Where $\tau(x)$ is the true treatment effect for observation x, and $s(n)$ is the subsample size. Third, we can use a uniform convergence result to show that:

$$\sup_{x \in \mathcal{X}} |\hat{\tau}_b(x) - \tau(x)| = o_p(1)$$

As $n \to \infty$, where $\mathcal{X}$ is the support of X. Combining these results, we can show that:

$$\text{Cov}(\hat{\tau}_b(x), \hat{\tau}_{b'}(x)) = o_p(1/n)$$

As $n \to \infty$, for any $b \neq b'$.

Step 3

To show that the average of the predictions of all the trees is consistent and asymptotically normal, we need to use a Slutsky theorem to the average of independent and identically distributed random variables with a common mean and variance. The Slutsky theorem states that if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ as $n \to \infty$, where X is a random variable and c is a constant, then $X_n + Y_n \xrightarrow{d} X + c$ and $X_n Y_n \xrightarrow{d} cX$ as $n \to \infty$. Applying this to our case, we have:

$$\sqrt{n}(\hat{\tau}(x) - \tau(x)) = \sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{\tau}_b(x) - \tau(x)\right) = \frac{1}{\sqrt{B}}\sum_{b=1}^{B} \sqrt{n}(\hat{\tau}_b(x) - \tau(x))$$

Where $\hat{\tau}(x)$ is the average prediction of all the trees for observation x, and $\hat{\tau}_b(x)$ is the prediction of the b-th tree for observation x. Note that by Step 1, we have:

$$\sqrt{n}(\hat{\tau}_b(x) - \tau(x)) \xrightarrow{d} N(0, V_b(x))$$

As $n \to \infty$, where $V_b(x)$ is a positive constant that depends on b and x. Also note that by Step 2, we have:

$$\text{Cov}(\sqrt{n}(\hat{\tau}_b(x) - \tau(x)), \sqrt{n}(\hat{\tau}_{b'}(x) - \tau(x))) = o(1/n)$$

As $n \to \infty$, for any $b \neq b'$. Therefore, by applying the Slutsky theorem and the classical central limit theorem, we have:

$$\frac{1}{\sqrt{B}} \sum_{b=1}^{B} \sqrt{n}(\hat{\tau}_b(x) - \tau(x)) \xrightarrow{d} N(0, V(x))$$

As $n, B \to \infty$, where $V(x) = (1/B) \sum_{b=1}^{B} V_b(x)$ is a positive constant that depends on x. This implies that:

$$\sqrt{n}(\hat{\tau}(x) - \tau(x)) \xrightarrow{d} N(0, V(x))$$

As $n, B \to \infty$, which completes the proof of asymptotic normality. The proof of consistency follows from applying Chebyshev's inequality and noting that:

$$P(|\hat{\tau}(x) - \tau(x)| > t) = P(n(\hat{\tau}(x) - \tau(x))^2 > nt^2) < E[n(\hat{\tau}(x) - \tau(x))^2]/nt^2 = V(x)/t^2$$

Which goes to zero as $n, B, t \to \infty$.