2023-05-01

Flexible Models For The Estimation Of Treatment Effect

Habeeb Abolaji Bashir
University of Texas at El Paso

FLEXIBLE MODELS FOR THE ESTIMATION OF TREATMENT EFFECT

Habeeb Abolaji Bashir

Master's Program in Mathematical Sciences

APPROVED:

Suneel Babu Chatla , Ph.D., Chair

Abhijit Mandal, Ph.D.

Deepak K. Tosh, Ph.D.

Stephen Crites, Ph.D.
Dean of the Graduate School

I dedicate this thesis to my devoted family and friends who have supported, inspired, and encouraged me throughout my academic career.Thank you to my parents for their unwavering affection and advice, who gave me a passion for learning and who always saw my potential. I want to express my gratitude to my wife for always being by my side and providing me with infinite laughter and diversions when I needed them the most. Thank you to all my peers who supported me throughout the highs and lows of my research path with your words of wisdom, encouragement, and late-night studies. I also want to thank my thesis adviser Dr. Suneel Babu Chatla for the important advice, mentoring, and criticism he gave me. Your dedication to education and your enthusiasm for study have continuously motivated and inspired me.Without the affection,encouragement, and contributions of all these amazing individuals, this thesis would not have been feasible. Thank you all.

FLEXIBLE MODELS FOR THE ESTIMATION OF TREATMENT EFFECT

by

Habeeb Abolaji Bashir

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

May 2023

Acknowledgement

My sincere gratitude goes out to the members of my thesis committee, Dr. Suneel Babu Chatla, Dr. Deepak K. Tosh, and Dr. Abhijit Mandal, for their direction, criticism, and support throughout the research process. Their combined experience and knowledge have been crucial in determining the course of this thesis. I am also appreciative of my family and 'friends' unflagging support and inspiration. My inspiration has always come from their drive and affection. Finally, I want to thank my wife for always trusting in me and for her affection and support.

Abstract

Estimation of treatment effect is an important problem which is well studied in the literature. While the regression models are one of the most commonly used techniques for the estimation of treatment effect, they are prone to model misspecification. To minimize the model misspecification bias, flexible nonparametric models are introduced for the estimation. Continuing this line of research, we propose two flexible nonparametric models that allow the treatment effect to vary across different levels of covariates. We provide estimation algorithms for both these models. Using simulations and data analysis, we illustrate the usefulness of the proposed methods.

Keywords: Additive Model, Single Index Model, Propensity Score, Model-based Recursive partitioning

Contents

Page

Acknowledgement v

Abstract vi

Table of Contents vii

List of Tables viii

List of Figures ix

1 Introduction 1

2 Background 6

2.1 Single Index Model 6

2.2 Model Based Recurive Partitioning Tree 7

2.2.1 MOB Algorithm 9

2.2.2 Estimation 9

2.2.3 Parameter Testing Stability 10

3 Methodology 13

3.1 Single Index Treatment Model 13

3.2 Tree-based Treatment Model 14

4 Numerical Study 15

4.1 Simulation Results 15

4.2 House Price Data Analysis 17

4.2.1 Matching 17

4.2.2 Comparison of Baseline and Single Index Models 19

4.2.3 MOB Tree-based Model 22

Bibliography 24

Appendix 27

Curriculum Vitae 30

List of Tables

4.1 Simulated Results 16

4.2 Baseline Model: Parametric coefficients 20

4.3 Single Index Model: Parametric coefficients 20

4.4 Baseline: Nonparametric term 20

4.5 Single Index Model: Nonparametric term 21

4.6 Comparison of AIC, R-sq(adj) for Baseline, MOB and Single Index Model 23

List of Figures

1.1 Confounder 3

4.1 Plot for 500 samples size with Probability (0.2,0.5,0.8) respectively 16
4.2 Matching Plot 18
4.3 Matching Distribution 19
4.4 Single Index Model: Nonparametric term for the treatment 21
4.5 Single Index Model: Nonparametric term for the lotsize 21
4.6 Model Based Recursive Partitioning Tree (MOB) 22
4.7 GAM Plot 23

Chapter 1

Introduction

A "treatment effect" is the average direct impact of a binary (0-1) variable on a scientifically or politically relevant result variable (Greenewald et al. 2021). The genesis of the word "treatment effect" can be traced to the medical literature that examines the primary impacts of binary, yes-or-no "treatments," such as an experimental medication or a novel treatment option. But the phrase is now much more widely used (Robertson et al. 2021). For example, we may compare the salaries of people who did and did not participate in the program using a data collected that describes the trainees' and the control group's labour market conditions. Such straightforward comparisons would usually serve as the starting point for any empirical assessment of treatment effects. Regression techniques or matching may also be used to account for ethnic or historical factors. In reality, straightforward comparisons or even comparisons that account for regression might yield false estimates of causal effects. For instance, even after accounting for observable differences, participants in subsidised training programs are frequently found to earn less than seemingly comparable controls. Please see, for example, Ashenfelter (1978) and Heckman & Robb Jr (1985). This could be a result of an omitted variables bias, which is a result of unexplained variation in the two groups' earning potential. Generally speaking, the most important econometric issue that occurs in the estimation of treatment effects is omitted variables bias, sometimes referred to as selection bias. The potential-outcomes paradigm provides the best visual representation of the relationship between omitted variable bias, causation, and treatment effects.

Consider a situation where the treatment variable is binary: $Z = 0, 1$. The goal is to determine how Z affects the outcome variable Y. Potential outcome framework, which

is also called the Rubin-Causal-Model, is one popular approach to estimate the treatment effect. It requires the joint distribution of (Z, Y) augments with $(Y(1), Y(0))-$ the probable outcome pair of Y when Z is, respectively, 1 and 0 (Deng et al. 2023). Since Y is the observed outcome and by definition we have

$$Y = \begin{cases} Y(1) & \text{if } Z = 1, \\ Y(0) & \text{if } Z = 0 \end{cases} \tag{1.1}$$

When $Z = 1, Y(0)$ is the counterfactual and not the observed, and $Z = 0, Y(1)$ is the counterfactual. Then

$$\tau = Y(1) - Y(0), \tag{1.2}$$

be the treatment effect of switching X from the control to the treatment, or just the treatment effect of Z.

It should be noted that the treatment effect τ and prospective result pair are both random variables. For a particular subject or unit Y often indicates a specific measurement. The (population) average treatment effect (ATE) to be

$$\mathrm{E}^*(\tau) = \mathrm{E}^* \left(Y(1) - Y(0) \right), \tag{1.3}$$

where the expectation is considered under the joint distribution. If a sample of N units $(Y_i(1), Y_i(0)), i = 1, \ldots, N$, is provided, then the sample average treatment effect (SATE) is defined as

$$\sum_i \left(Y_i(1) - Y_i(0) \right) / N. \tag{1.4}$$

When the population is fixed to be the supplied sample, SATE is the population average treatment effect (PATE). SATE is still a widely used causal estimand in literature, and it was developed in the complete randomised trial scenario by pioneers in statistics like Fisher and Neyman. In this chapter, the population average treatment effect (PATE) is almost solely calculated using ATE. Also we could talk about the naive estimation for an unbiased

estimator of the average treatment which could be defined as

$$\widehat{\tau}_{\text{naive}} := \overline{Y_T} - \overline{Y_C} \tag{1.5}$$

where $\overline{Y}_T$ stands for the average for treatment and $\overline{Y}_C$ is the average for control. As a result, we can claim that naive estimator is an unbiased estimator of

$$\mathrm{E}(Y|Z=1) - \mathrm{E}(Y|Z=0). \tag{1.6}$$

The associated impact of Z on Y is represented by the equation above. The observed joint distribution can be used to characterise the association of naive estimation, in contrast to the causal effect. It is to be noted that other than the cause of interest Z, there are many other confounders that can influence the association. The following causal graphical model describe a causal relationship between Z and Y with a confounding factor called U that can affect both Z and Y at the same time.

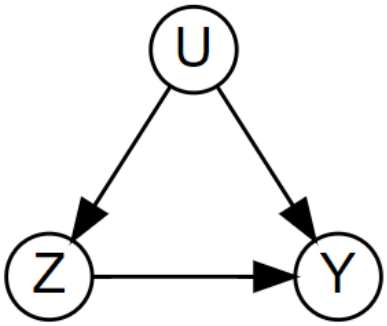

Figure 1.1: Confounder

Confounders are the main challenge of causal inference, which is why it is challenging to draw conclusions without randomised studies. The notion that smoking can cause lung cancer is now widely accepted. Numerous observational studies had long suggested a connection between smoking and lung cancer. However, none of them are as conclusive as a randomised experiment because it is simply not possible to randomly assign people to smoke or not smoke, so this method is not an option. Fisher, who popularised randomised experiments in statistics and was aware that a correlation does not necessarily suggest a cause, publicly criticised a 1950 study that linked tobacco use to lung cancer Fisher (1958). One of his points is the possibility of a hereditary predisposition to smoking, which is likely also connected to lung cancer. In other words, a gene or confounder may be present that increases a person's propensity for smoking as well as their risk of acquiring lung cancer. Unconfoundedness: The unconfoundedness has the following assumption

$$Y_i(0), Y_i(1) \perp Z_i | X_i,$$

which is the same as $Pr(Z_i|Y_i(0), Y_i(1), X_i) = Pr(Z_i|X_i)$. Below are some implications.

- It implies that treatment allocation is random within distinct populations defined by the outcomes of variables that can be observed.
- Excludes unknown confounding factors, often known as the "no unknown confounders" premise.
- Unconfoundedness is attained by random assignment.
- Exists in most observational studies although there is no formal mechanism to test.

Strategies for Estimating Confoundedness and Unconfoundedness: Consider estimation strategies for Average Treatment Effect such that

- Regression or (outcome) modelling (Levine & Rubin 1979):

$$\widehat{\tau}_{reg} = N^{-1} \sum_{i=1}^{N} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- Inverse Probability weighting(IPW) (Rosenbaum 1987):

$$\widehat{\tau}_{ipw} = \frac{\sum_{i=1}^{N} Z_i Y_i / e_i}{\sum_{i=1}^{N} Z_i / e_i} - \frac{\sum_{i=1}^{N} (1-Z_i) Y_i / (1-e_i)}{\sum_{i=1}^{N} (1-Z_i)/(1-e_i)},$$

  where $e_i = P(Z_i = 1|X)$.

Since the regression based estimators are sensitive to model misspecification, in this thesis, we propose two nonparametric models that are flexible and robust for the estimation of treatment effect. The proposed approaches are based on single index model (Ichimura 1987) and model-based recursive partitioning (Zeileis et al. 2008*a*). In the following chapter, we provide a brief introduction to both of these approaches.

Chapter 2

Background

This section discusses briefly about the single index model and the model-based recursive partitioning tree approach.

2.1 Single Index Model

The single index model (Ichimura 1987) is denoted as

$$y_i = h(x_i^T \theta_0) + \epsilon_i, \ \ i = 1, ..., n, \tag{2.1}$$

where:

- $y_i \in R$ is response and $x_i \in R^k$ is a covariate vector, and $\epsilon_i \in R$ is an unobserved disturbance, and $\theta_0 \in R^k$ is an unknown true parameter which needs to be estimated,
- The distribution of ϵ_i depend on x_i only through the index $h(x_i^T \theta_0)$,
- $E(\epsilon_i | x_i) = 0$ for $i = 1, \ldots, n$,
- The function of $h : R \longrightarrow R$ is known up to a parameter θ_0, and

For identifiability, we assume that $\|\theta_0\|_2 = 1$. The unknown link function $h(.)$ can be approximated by any smoothing method, however, we will concentrate on cubic B-splines. In particular, $h(\mu) \approx \eta^T Z(\mu)$, for some parameters $\eta_t \in R^d$ and $Z(\mu) = [\, B_1(\mu), \ldots, B_d(\mu)^T] \in R^d$ is a collection of d normalised cubic B-spline basis functions. The model now becomes

$$Y_i = \eta^T Z(x_i^T \theta_0) + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.2}$$

Suppose we have a $n \times n$ projection matrix of single index $S_\theta = Z_\theta(Z_\theta^\top Z_\theta)^{-1}Z_\theta^\top$. Then the profile loglikelihood, up to a constant, is defined as

$$Q(\theta) = \|Y - S_\theta Y\|_2^2 \tag{2.3}$$

The estimator for the indexing parameter is obtained as follows:

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, Q(\theta) \tag{2.4}$$

2.2 Model Based Recurive Partitioning Tree

Let's say we have a parameter model with a K-dimensional vector of parameters $\theta \in \Theta$. Suppose we have n observations $Y_i, i = 1, ..., n$, then the model can be fitted by minimizing some objective function $\Omega(Y, \theta)$ yielding the parameter estimate $\hat{\theta}$ (Zeileis et al. 2008*b*)

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_i^n \Omega(Y_i, \theta). \tag{2.5}$$

This category of estimators includes a number of well-known estimating methods, with maximum likelihood (ML) and ordinary least squares (OLS) being the most common. When using OLS, Ω typically represents the error sum of squares, whereas when using ML, it represents the negative log-likelihood. If the variable Y may be divided into dependent and explanatory variables, as in the case where $Y = (y, x)^T$, it may be the entire likelihood or the conditional likelihood. For instance, GLM, or generalised linear model. The model equation is given by $g(E(y)) = x^T\theta$, where y has a known distribution belonging to the exponential family, $g(.)$ is a known link function, and θ is the regression coefficient.

In general it is not reasonable to assume that a single global model perfectly fits all data. However, it might be possible to divide the observations according to some covariates in such a way that a locally a better model can be found in each part of the division. In this case, we can adaptively find a good approximation of this partition using a recursive

partitioning strategy based on partitioning variables $Z_j \in \mathcal{Z}_j$, $j = 1, \ldots, l$. Following the notation from Zeileis et al. (2008*a*), we let the the partition of $\{\mathcal{B}_b\}_{b=1,\ldots,B}$ of the space $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_l$ in such a way that in each $\mathcal{B}_b$ a local model is fitted. For example, classification and regression trees have numerous partitioning variables Z_j but with only a very simple model.

By computing the locally optimal parameter estimates $\hat{\theta}_b$ in each segment of the correct partition $\mathcal{B}_b$, it is straightforward to estimate the parameters that minimise the global objective function. However, if $\{B_b\}$ is unknown, the minimization of the overall objective function

$$\sum_{b=1}^{B} \sum_{i \in I_b} \Omega(Y_i, \theta_b) \longrightarrow \min, \tag{2.6}$$

over all partitions $\{B_b\}$ is complicated even if the number of segments B is fixed. The complexity of over all potential partitions B_b(with corresponding indexes I_b, $b = 1, \ldots, B$, is higher: The number of alternative partitions quickly exceeds the capacity of an exhaustive search if there is more than one partitioning variable ($l > 1$). Additionally, in this situation, precautions should be taken to prevent over-fitting with increasing B. In conclusion, even for fixed B, finding the best partition (with regard to Ω) is challenging. However, if there is just one partitioning variable (l = 1), it is simple to identify the best split(s): The literature on change point and structural change analysis in both statistics and econometrics discusses several algorithms for segmenting models over a single variable, often time. Zeileis et al. (2008*a*) propose a greedy forward search where the objective function ω can at least be locally optimised in each step to take use of this methodology for discovering a partition close to the optimal one in $l > 1$ dimensions. The next section contains a comprehensive description of this algorithm.

2.2.1 MOB Algorithm

Each node is associated to a single model. At each node, parameter instability fluctuation test is carried out to determine whether dividing the node is required. If any of the partitioning variables Z_j exhibits notable instability, divide the node into B locally optimum parts and repeat the process on each part. The recursion terminates if no other major instabilities can be discovered, returning a tree where each terminal node (or leaf) is linked to a model of type $M(\mathrm{Y}, \theta)$. Specifically, the steps of algorithm are described as follows:

- Estimate $\hat{\theta}$ by minimising the objective function Ω to fit the model once to all data in the current node.
- Examine the stability of the parameter estimations with regard to ordering of each of $Z_1, ..., Z_l$. Choose the variable Z_j linked to the highest parameter instability if there is any overall instability; otherwise, stop.
- For a fixed or adaptive number of splits, determine the split point(s) that locally optimise Ω.
- Repeat the process by dividing the node into daughter nodes.

For more algorithmic details, please see Zeileis et al. (2008*a*).

2.2.2 Estimation

This is a standard practice. We note that, under mild regularity requirements (see, for example, (Anderson et al. 1994)), the solutions to above can be obtained by solving the first order condition

$$\sum_{i=1}^{n} \Omega(Y_i, \hat{\theta}) = 0, \tag{2.7}$$

where

$$\psi(Y, \theta) = \tfrac{\partial \Omega(Y,\theta)}{\partial \theta},$$

is the estimating or score function that corresponds to $\Omega(Y, \theta)$. In many models of interest, well-established fitting algorithms are available for computing $\hat{\theta}$ (e.g., OLS estimation via QR decomposition for linear regression or ML via iterative weighted least squares for GLMs). Analytical closed form solutions for $\hat{\theta}$ are only available in a few special cases. The score function is next examined for systematic departures from its mean 0 at the estimated parameters $\hat{\psi} = \psi(Y_i, \hat{\theta})$.

2.2.3 Parameter Testing Stability

The goal of this algorithmic step is to determine whether the fitted model's parameters are stable over each specific ordering implied by the partitioning variables Z_j or whether slicing the sample in half with respect to one of the Z_j might be able to detect parameter instabilities and thus improve the fit. It makes sense to determine if the scores $\hat{\psi}$ show systematic deviations from 0 over Z_j or whether they fluctuate randomly around their mean 0 in order to evaluate parameter instability. The empirical fluctuation process can capture these variations.

$$W_j = \hat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \hat{\psi}_{\sigma(Z_{ij})}, \tag{2.8}$$

where the ordering permutation producing the antirank of the observation $\sigma(Z_{ij})$ in the vector $\sigma\ Z_j = (Z_j, ..., Z_{nj})^T$. Therefore, $W_j(t)$ is simply the partial sum process of the scores ordered by the variable Z_j, scaled by the number of observations n and a suitable estimate J of the covariance matrix $\text{cov}(\psi(Y, \hat{\theta}))$, for example, $\hat{J} = n^{-1} \sum_{i=1}^{n} \psi(Y_i, \hat{\theta}) \psi(Y_i, \hat{\theta})^T$, however, other reliable estimators are also relevant, such as the HC (heteroscedasticity consistent) and HAC (heteroscedasticity and autocorrelation consistent) estimators. According to a functional central limit theorem (Zeileis & Hornik 2007), If the parameters are stable, this experimental variation process merges to a Brownian bridge W^0. The use of a scalar operational facilitates the creation of a test statistic $\lambda(.)$. The limiting distribution that corresponds to the fluctuation process $\lambda(W_j(.))$ and in the empirical process, the restricting

process is just subject to the same functional (or its maximum counterpart), $\lambda(W^0(.))$. Zeileis & Hornik (2007) developed the generalised M-fluctuation test, a fairly broad framework for evaluating parameter stability. It has been demonstrated to include many structural change tests proposed in the economics and statistics literature, including OLS-based CUSUM and MOSUM tests (Chu et al. 1995), score-based tests (Merkle & Zeileis 2013), and statistics based on Lagrange multiplier statistics(Andrews & Ploberger 1994).
An overview of these tests was provided by Zeileis (2005). However, two distinct test statistics appear to be particularly appealing for evaluating numerical and categorical partitioning variables Z_j, respectively. Any test from this framework could theoretically be used in the recursive partitioning algorithm.

Model Splitting: The fitted model must be divided, with respect to the variable Z_j^*, in this algorithmic step into a segmented model with B segments, where B may be fixed or determined adaptively. Two competitor categorizations can be easily evaluated for a fixed number of splits by comparing the segmented objective function $\sum_{b=1}^{B}\sum_{i\in I_b}\omega(Y_i,\theta_b)$. The best partition can be found by performing an exhaustive search over all feasible partitions with B segments, but this can be time-consuming. Therefore, many search techniques for numerical and classifying partitioning variables are presented briefly.

Numerical Variable Splitting : It is possible to search thoroughly for splitting into $B = 2$ sections in $O(n)$ operations. A dynamic programming method of order $O(n^2)$ can be used to find the ideal partition for $B > 2$, whereas an exhaustive search would be of order $O(n^B - l)$; for $B > 2$. This is a Bellman's principle application that has been covered in several places in the literature on change point and structural change analysis (Zeileis et al. 2010). As an alternative, iterative algorithms that are known to lead to the most favourable result can be utilised (Muggeo 2003).A number of techniques are available if B is not fixed but rather needs to be selected adaptively (o'Brien et al. 2004). If the parameters are estimated by M, criteria for information can be utilised in particular.

Categorical Variable Splitting: When partitioning categorical variables, the number of segments cannot exceed the number of categories $B \leq C$. Either always splitting into all

levels of $B = C$, or alternately, always splitting into the fewest possible $B = 2$ segments, are two straightforward strategies. The search for the most suitable division in this last situation is of order $O(2^{C-l})$.

Chapter 3

Methodology

In this chapter, we present the contributions which includes single index and tree-based models.

3.1 Single Index Treatment Model

Suppose we have a sample of n units (Y_i, Z_i, X_i^T, W_i^T), $i = 1, \ldots, n$, where $X_i \in R^p$ and $W_i \in R^q$ are covariate vectors. We consider the following single index model:

$$Y_i = g(X_i^T \beta) Z_i + f(W_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $g(\cdot)$ and $f(\cdot)$ are unknown functions. For simplicity, take q=1. The proposed estimation approach includes the following steps:

- Set initial values for β with the first component is equal to 1
- Normalize β such that $\|\beta\|_2 = 1$
- Use optimization function by choosing the AIC or gcv.ubre from the following model

  $$gam(y \sim s(X^T \beta, by = Z) + s(W))$$

  as target
- Obtain the parametric estimator $\widehat{\beta}$
- Refit the above gam model with $\widehat{\beta}$

3.2 Tree-based Treatment Model

We now consider the following model

$$Y_i = g(X_i)Z_i + f(W_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $g(\cdot)$ and $f(\cdot)$ are unknown functions. To avoid the curse of dimensionality, we estimate $g(\cdot)$ using a tree-based method and $f(\cdot)$ using an additive structure. We use the following algorithm for estimating the tree-based model:

- Fit an additive model with only Y and W
- Compute the residuals from the previous model $\widehat{R}$
- Fit the following tree (Zeileis et al. 2008*a*) using the mob function from partykit package

$$mob(\widehat{R} \sim Z|X)$$

- Compute the fitted values from the mob $\widetilde{R}$ and using it compute the residuals $\widetilde{Y} = Y - \widetilde{R}$
- Refit the gam model in step 1 using $\widetilde{Y}$ as the response variable

Chapter 4

Numerical Study

4.1 Simulation Results

The data is generated as follows:

$$
\begin{aligned}
f(x) &= 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10} \\
m &= 3 \\
X &\sim U(0,1)^3 \\
W &\sim N(0,1) \\
\beta &= (1,-1,0.5) \\
Z &\sim \text{Binom}(p) \\
\epsilon &\sim N(0,1) \\
Y &= f((X^T\beta + 0.41)/4)Z + \sin(W) + \epsilon.
\end{aligned}
$$

For simulations, we consider different sample sizes $(100, 250, 500)$ and different probabilities $(0.2, 0.5, 0.8)$ for the treatment effect. For each combination, we generated 50 datasets. For each dataset, we computed the mean squared error for the treatment effect and the pearson correlation between the estimated and true treatment. The results are provided in Table 4.1. The numbers in parenthesis indicate standard errors. The results indicate that the proposed estimation algorithm works well in estimating the treatment effect. The true treatment function for different probabilities (p) is shown in 4.1.

Table 4.1: Simulated Results

No of observation	Probability	MSE(sd)	Correlation(sd)
100	0.2	0.1389(0.1473)	0.9909(0.00936)
	0.5	0.1037(0.0624)	0.9970(0.0021)
	0.8	0.1325(0.0704)	0.9967(0.0018)
250	0.2	0.0708(0.1672)	0.9947(0.0138)
	0.5	0.0362(0.0159)	0.9988(0.0006)
	0.8	0.0473(0.0318)	0.9987(0.0007)
500	0.2	0.0153(0.0077)	0.9991(0.0005)
	0.5	0.0206 (0.0092)	0.9993(0.0003)
	0.8	0.0320(0.0167)	0.9992(0.0003)

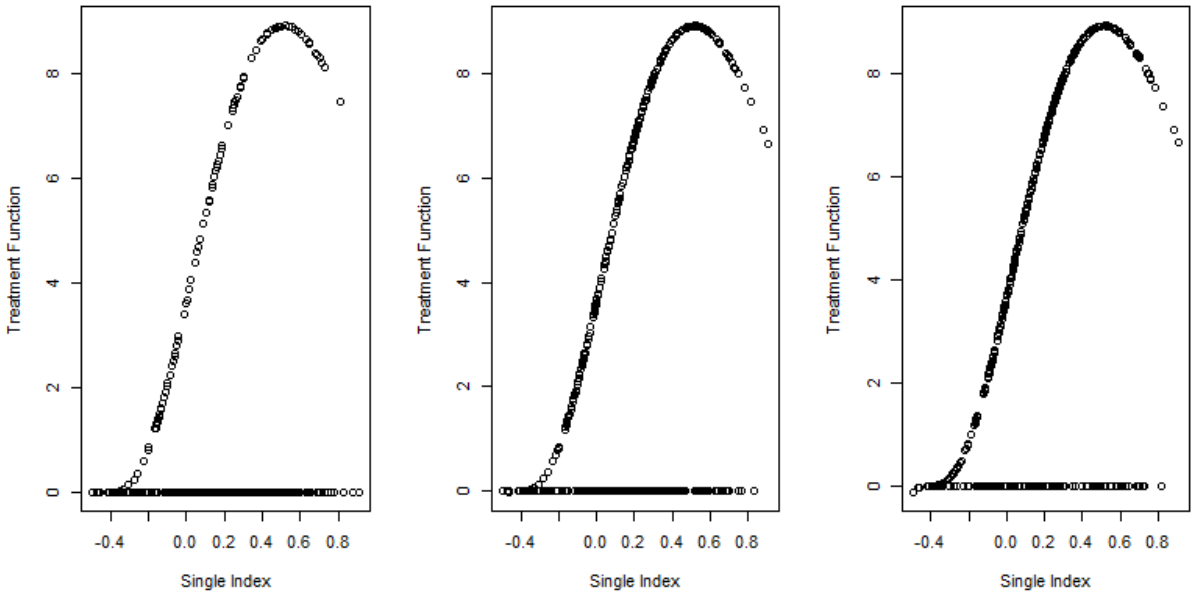

Figure 4.1: Plot for 500 samples size with Probability (0.2,0.5,0.8) respectively

4.2 House Price Data Analysis

Data Source and Description

Housing market is a key indicator for the general health of the economy. We consider a house price data from the Kaggle website. The data includes information on 546 homes across the city of Windsor, Canada. It has 12 different variables such as price, lot size, bedrooms, bathrooms, number of stories, presence of a driveway, recreation room, basement, gas heating, air conditioning, presence of a garage, and age of the home. The response variable is the price and treatment variable is the preferred area. Other predictors include bathrooms, stories, basement, gas heating, garage, driveway, recreation and age. Except price and lotsize which are continuous, the remaining variables are categorical in nature.

4.2.1 Matching

Propensity score analysis (PSA) balances the pretreatment covariates to produce an accurate inference for the causal effect from observational data. Due to missing randomness in the observational data, without balancing the estimated treatment effect may not be unbiased. The propensity score approach addresses this problem by matching the baseline characteristics between the treatment and control groups to make them as if they are observed from a random experiment. In this approach, the treatment variable is first regressed against the pretreatment characteristics using a logistic regression model. The propensity score, also known as the likelihood of assigning the course of treatment, can be determined using the fitted model. The next step is to deduce a causal effect using a propensity score. We note that the propensity score is still regarded as nonparametric even if it was obtained using a parametric regression model (Zhang 2013). The following are the specifications used for the matching procedure. The `matchit` function from `R` package is considered for matching Ho et al. (2018). It requires a formula as an input that specifies which pretreatment variables affect the treatment assignment. All the covariates are considered for the pretreatment variables. The matching is based on the propensity score

estimated with logistic regression. We use the nearest neighbor method with 1:1 matching and without replacement for matching. From the original 546 observations, the matched data is reduced to 256 observations. The histograms for the propensity scores before and matching are provided in Figures 4.2 and 4.3.

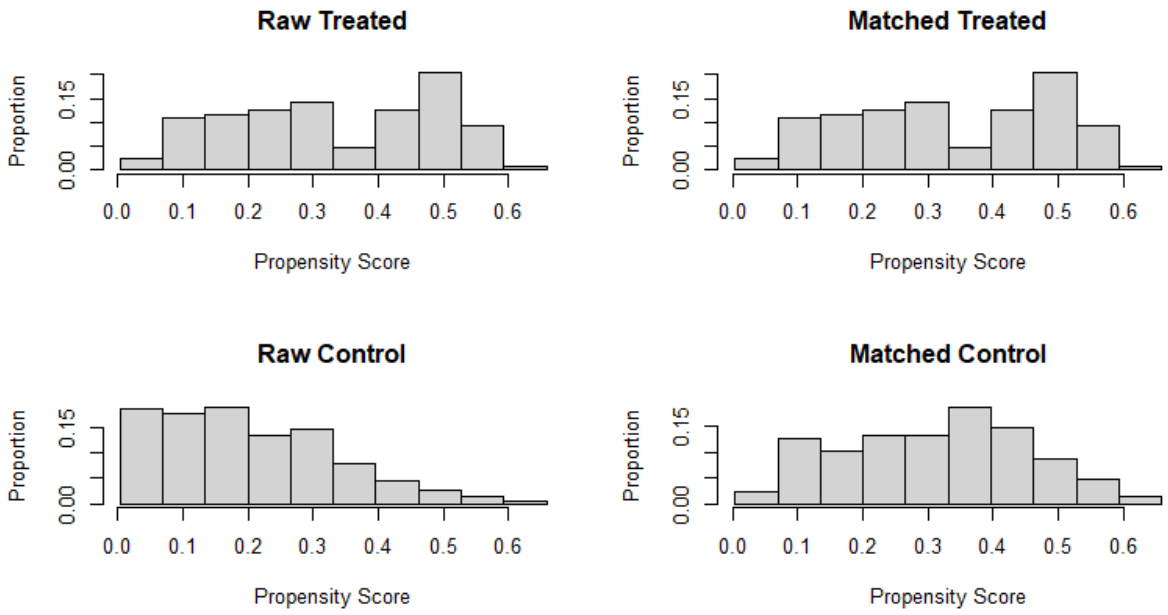

Figure 4.2: Matching Plot

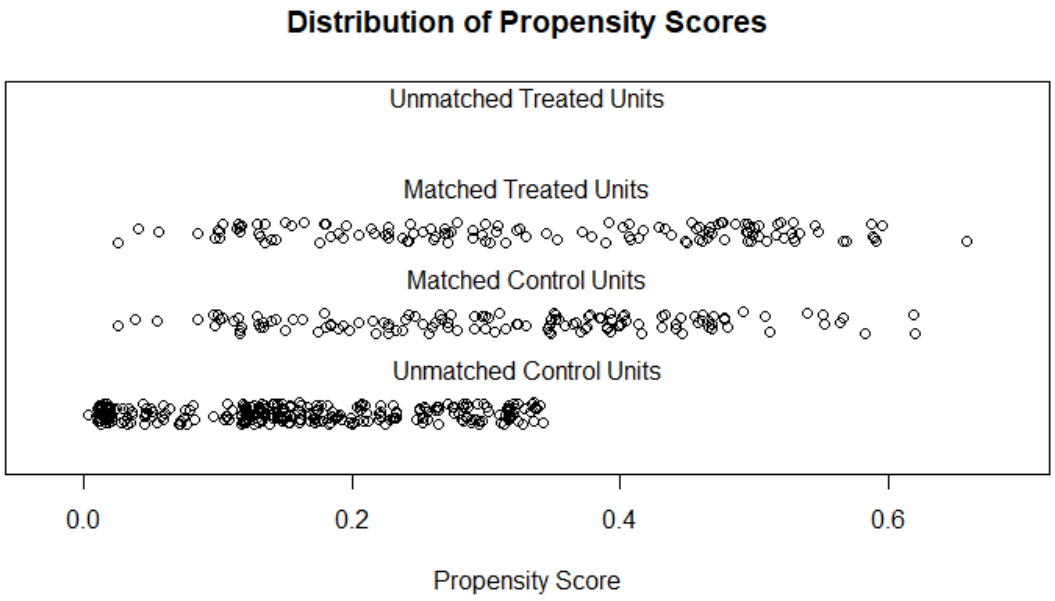

Figure 4.3: Matching Distribution

4.2.2 Comparison of Baseline and Single Index Models

For comparison, we fit a baseline model with all the categorical variables as parametric effects and the variable lotsize as a nonparametric effect. The coefficient significance results for the parametric part are shown in Table 4.2. Except bedrooms, recreation, and garage all the remaining variables are significant. From Table 4.4, we notice that the nonparametric effect for the lotsize is significant. It exhibits a positive and linear association with the price as shown in Figure 4.4.

The coefficient significance results from the proposed single index model are presented in Tables 4.3 and 4.5. The partial plots for the nonparametric terms are presented in Figures 4.4 and 4.5.

Table 4.2: Baseline Model: Parametric coefficients

Variables	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	10.2045	0.1185	86.145	$< 2e-16$***
preferyes	0.1513	0.0271	5.589	6.09e-08 ***
bedrooms	0.0287	0.0234	1.225	0.2219
stories	0.0859	0.0180	4.776	3.09e-06 ***
drivewayyes	0.2621	0.1002	2.615	9.40e-10 ***
bathrooms	0.1865	0.0292	6.369	9.40e-10 ***
recreationyes	0.0348	0.0333	1.045	0.29727
fullbaseyes	0.1064	0.0321	3.313	0.00106 **
gasheatyes	0.1588	0.0782	2.031	0.04331 *
airconyes	0.1484	0.0289	5.128	5.97e-07 ***
garage	0.0305	0.0163	1.876	0.06192.

Table 4.3: Single Index Model: Parametric coefficients

Variables	Estimate	Std. Error	t value
Stories	1.1192	0.0776	14.4237
drivewayyes	5.7191	0.1123	50.9256
bathrooms	2.1614	0.1319	16.3870
recreationyes	-0.2476	0.1061	2.3342
fullbaseyes	1.2865	0.1174	10.9624
gasheatyes	1.2848	0.2823	4.5503
airconyes	1.3908	0.1074	12.9463
garage	0.2514	0.0531	4.7331

Table 4.4: Baseline: Nonparametric term

Variables	Edf	Ref.df	F	p-value
s(lotsize)	1	1	82.56	$< 2e-16$ ***

Table 4.5: Single Index Model: Nonparametric term

Variables	Edf	Ref.df	F	p-value
s(a):A	2	2	47.68	$< 2e-16$ ***
s(z)	1	1	121.25	$< 2e-16$ ***

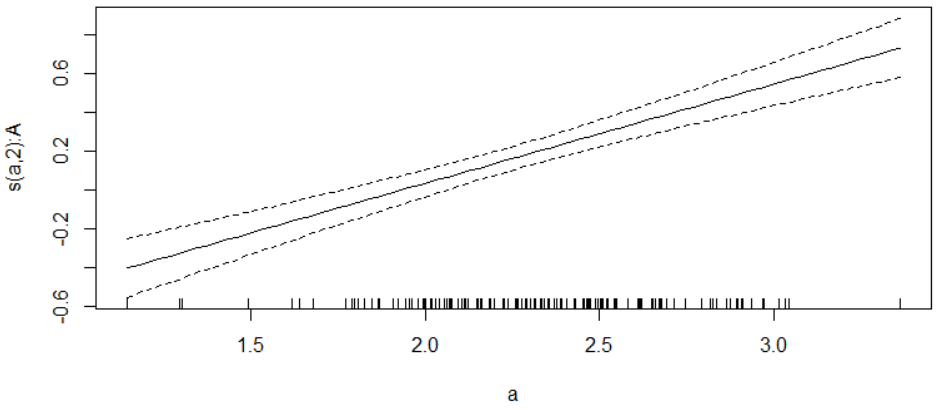

Figure 4.4: Single Index Model: Nonparametric term for the treatment

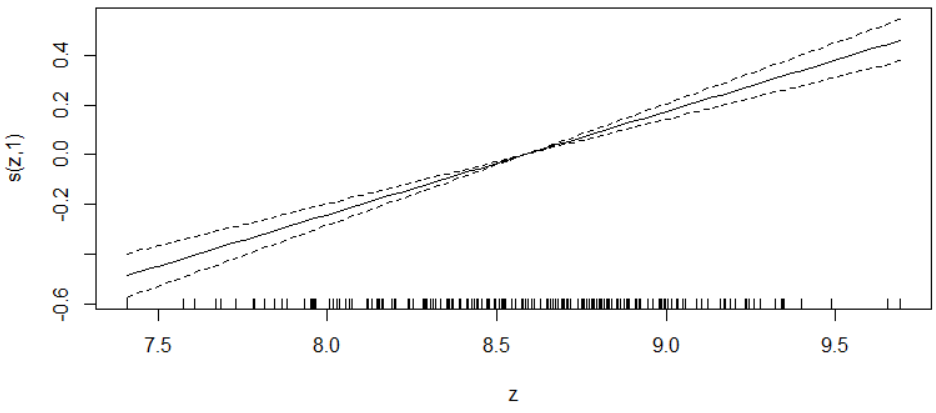

Figure 4.5: Single Index Model: Nonparametric term for the lotsize

4.2.3 MOB Tree-based Model

We now estimate the treatment using the MOB tree-based model. We still consider the modeling the variable lotsize as a fixed nonparametric function. At each terminal node of the MOB tree, a simple linear regression of preference on price is fitted to estimate the treatment effect. The fitted tree is shown in Figure 4.6. We notice that the magnitudes of the treatment effects in the terminal nodes are different. This model gives the flexibility of providing treatment effect for varying levels of covariates like single index model. Finally, Table 4.6 present the AIC values for all the models considered. While the AIC itself may not be good indicator for the model with causal effect, we presented these results for the sake of completeness.

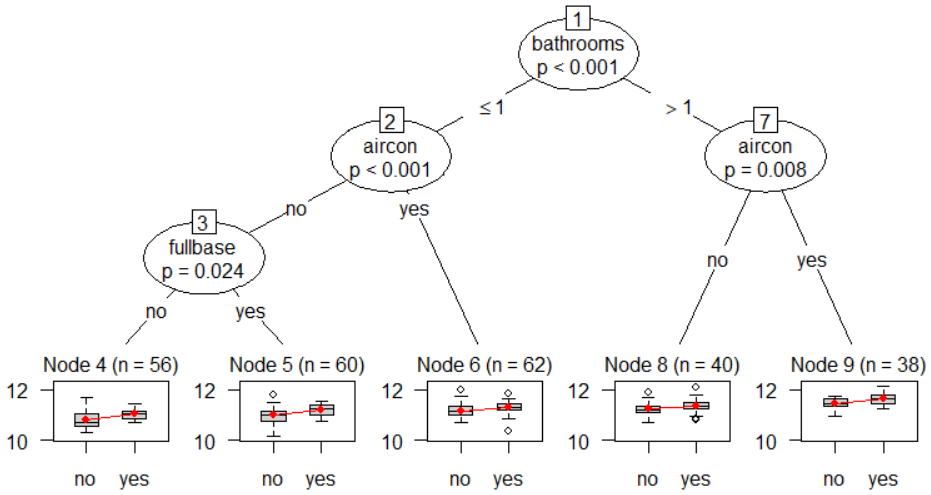

Figure 4.6: Model Based Recursive Partitioning Tree (MOB)

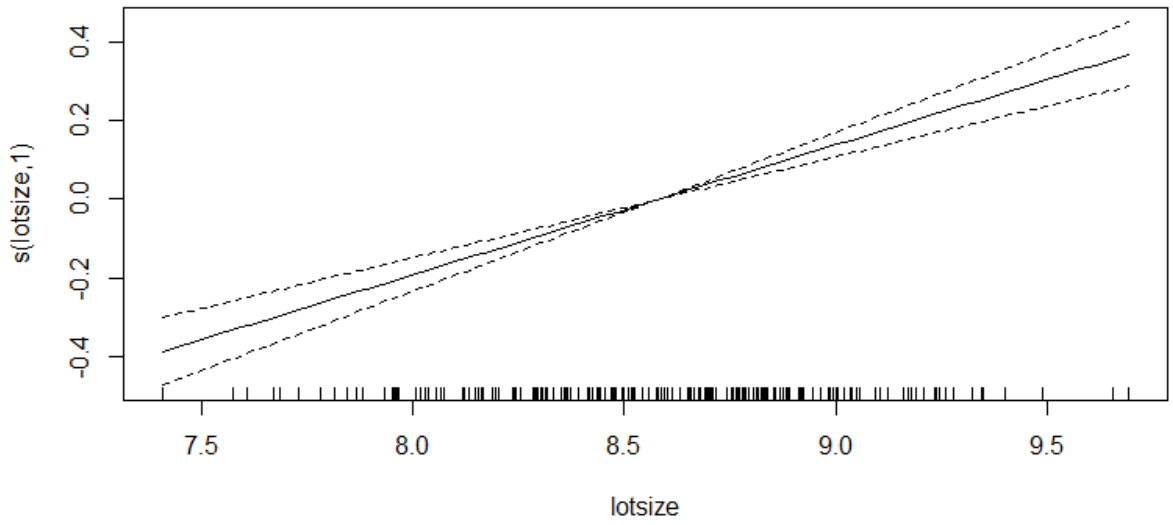

Figure 4.7: GAM Plot

Table 4.6: Comparison of AIC, R-sq(adj) for Baseline, MOB and Single Index Model

Model	AIC	R-sq.(adj)
Baseline	-48.0311	0.634
Single Index non-parametric fit	13.9465	0.519
MOB	99.2536	0.1417

In summary, the proposed methods provide a flexible way of estimating the treatment effects that varies across different levels of the covariates.

Bibliography

Anderson, D., Burnham, K. & White, G. (1994), 'Aic model selection in overdispersed capture-recapture data', *Ecology* **75**(6), 1780–1793.

Andrews, D. W. & Ploberger, W. (1994), 'Optimal tests when a nuisance parameter is present only under the alternative', *Econometrica: Journal of the Econometric Society* pp. 1383–1414.

Ashenfelter, O. (1978), 'Estimating the effect of training programs on earnings', *The Review of Economics and Statistics* pp. 47–57.

Chu, C.-S. J., Hornik, K. & Kaun, C.-M. (1995), 'Mosum tests for parameter constancy', *Biometrika* **82**(3), 603–617.

Deng, A., Yuan, L.-H., Kanai, N. & Salama-Manteau, A. (2023), Zero to hero: Exploiting null effects to achieve variance reduction in experiments with one-sided triggering, *in* 'Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining', pp. 823–831.

Fisher, R. (1958), 'Cigarettes, cancer, and statistics', *The Centennial Review of Arts & Science* **2**, 151–166.

Greenewald, K., Shanmugam, K. & Katz, D. (2021), High-dimensional feature selection for sample efficient treatment effect estimation, *in* 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 2224–2232.

Heckman, J. J. & Robb Jr, R. (1985), 'Alternative methods for evaluating the impact of interventions: An overview', *Journal of econometrics* **30**(1-2), 239–267.

Ho, D., Imai, K., King, G., Stuart, E. & Whitworth, A. (2018), 'Package 'matchit'', *Version.[Google Scholar]* .

Ichimura, H. (1987), Estimation of single index models, PhD thesis, Massachusetts Institute of Technology.

Levine, M. V. & Rubin, D. B. (1979), 'Measuring the appropriateness of multiple-choice test scores', *Journal of educational statistics* **4**(4), 269–290.

Merkle, E. C. & Zeileis, A. (2013), 'Tests of measurement invariance without subgroups: A generalization of classical methods', *Psychometrika* **78**, 59–82.

Muggeo, V. M. (2003), 'Estimating regression models with unknown break-points', *Statistics in medicine* **22**(19), 3055–3071.

o'Brien, K., Leichenko, R., Kelkar, U., Venema, H., Aandahl, G., Tompkins, H., Javed, A., Bhadwal, S., Barg, S., Nygaard, L. et al. (2004), 'Mapping vulnerability to multiple stressors: climate change and globalization in india', *Global environmental change* **14**(4), 303–313.

Robertson, S. E., Leith, A., Schmid, C. H. & Dahabreh, I. J. (2021), 'Assessing heterogeneity of treatment effects in observational studies', *American Journal of Epidemiology* **190**(6), 1088–1100.

Rosenbaum, P. R. (1987), 'Model-based direct adjustment', *Journal of the American statistical Association* **82**(398), 387–394.

Zeileis, A. (2005), 'A unified approach to structural change tests based on ml scores, f statistics, and ols residuals', *Econometric Reviews* **24**(4), 445–466.

Zeileis, A. & Hornik, K. (2007), 'Generalized m-fluctuation tests for parameter instability', *Statistica Neerlandica* **61**(4), 488–508.

Zeileis, A., Hothorn, T. & Hornik, K. (2008*a*), 'Model-based recursive partitioning', *Journal of Computational and Graphical Statistics* **17**(2), 492–514.

Zeileis, A., Hothorn, T. & Hornik, K. (2008*b*), 'Model-based recursive partitioning', *Journal of Computational and Graphical Statistics* **17**(2), 492–514.
URL: *http://www.jstor.org/stable/27594318*

Zeileis, A., Shah, A. & Patnaik, I. (2010), 'Testing, monitoring, and dating structural changes in exchange rate regimes', *Computational Statistics & Data Analysis* **54**(6), 1696–1706.

Zhang, J. (2013), A Comparison of Propensity Score Matching Methods in R with the MatchIt Package: A Simulation Study., PhD thesis, University of Cincinnati.

Appendix

R CODE

```
library(MatchIt)
source("single-index.R")

# read data
House_Prices <- read.table(file = "HousePrices.csv", sep = ",",
header = TRUE,na.strings = c("NA", "", " "), stringsAsFactors = TRUE)

# Transform price and lotsize variables by taking their logarithms
House_Prices$price <- log(House_Prices$price)
House_Prices$lotsize <- log(House_Prices$lotsize)

# Propensity score matching
matchmodel<-matchit(prefer ~ lotsize + stories +driveway+ bathrooms+
recreation+ fullbase + gasheat + aircon + garage, method="nearest"
,distance="glm",data = House_Prices)
matchmodel
plot(matchmodel, type="hist")
plot(matchmodel,type="jitter")

match_house_dt <- match.data(matchmodel)

basemodel <- gam(price ~ s(lotsize)+prefer+ bedrooms + stories +driveway+
```

```
                    bathrooms+ recreation+ fullbase + gasheat +
                    aircon + garage, data = match_house_dt)
# Baseline model
summary(basemodel)
AIC(basemodel)
plot(basemodel)
library(partykit)
library(party)
ctrl_1 <- mob_control(alpha = 0.05, bonferroni = TRUE, minsplit =20,
              trim = 0.1, breakties =TRUE, parm = NULL, verbose = FALSE)
mobmodel <- mob(price~prefer|+ bedrooms + stories +driveway+
                    bathrooms+ recreation+ fullbase + gasheat +
    aircon + garage,model=linearModel,control=ctrl_1,data = match_house_dt)
# Baseline model
summary(mobmodel)
AIC(mobmodel)
plot(mobmodel)
##
# list creation for data
y <- basemodel$y
xfull <- model.matrix(basemodel)
x <- xfull[, 3:11]
A <- xfull[,2]
z <- match_house_dt$lotsize
#
dlist <- list(y=y,x=x,z=z,A=A)
#
## fitting single index model
```

```
house_sifit <- si.fit(dlist)

### results
# summary of nonparametric fit + Treatment
summary(house_sifit$fit)
plot(house_sifit$fit, scale = FALSE)
AIC(house_sifit$fit)



# summary of parametric part + se
parcoef <- house_sifit$pcoef
names(parcoef) <- colnames(x)[2:9]
parcoef
#
parse <- house_sifit$se[2:9]
names(parse) <- colnames(x)[2:9]
#
tval <- parcoef/parse[-1]
names(tval) <- colnames(x)[2:9]
tval
```

Curriculum Vitae

My name is Habeeb Abolaji Bashir,a graduate student at the University of Texas in El Paso working towards a master's degree. Usmanu Dafodiyo University in Sokoto, Nigeria, is where I finished my undergraduate studies in statistics and found my passion for data science.

I got the chance to work on a number of research projects throughout my undergraduate studies, which ignited my interest in going on to pursue my study in data science. My enthusiasm for this subject of research was further stoked by the chance to collaborate closely with some of the top authorities in the field.

After earning my undergraduate degree in 2017, I went on to complete the one-year National Youth Service requirement.Thereafter, I started my graduate studies in statistics and data science at the University of Texas El Paso in 2021.Flexible models for estimating treatment effect are what I'm now working on. The chance to further investigate the areas of statistics and data science and also to add to the corpus of existing knowledge in this discipline excites me.

I am heavily involved in the University of Texas at El Paso graduate student community in addition to my research. As the Vice President of the Graduate Chapter of the American Mathematical Society, I have maintained a position of leadership in the graduate student organisation and have planned various activities aimed at building a feeling of community among graduate students. Additionally, I have volunteered to mentor new graduate students, helping them to navigate their graduate courses by providing advice and assistance. My long-term professional aim is to become a well-known Data Scientist, and I think the skills and information I acquire during my graduate studies will be crucial in assisting me in achieving this objective. I am committed to advancing knowledge in this field and am eager about continuing my education at Southern Methodist University, where I will begin my PhD program in Statistical Science in the autumn of 2023.

In my spare time, I like to watch football and spend time with my family. I also think that success in both academic and extracurricular endeavours depends on maintaining a strong work-life balance. I'm appreciative of the chance to pursue graduate studies at **SMU** and I am looking forward to the opportunities that lay ahead.

E-mail address: *bhabolaji@miners.utep.edu*